Bioinformatics, 35(10), 2019, 1692–1700
doi: 10.1093/bioinformatics/bty881
Advance Access Publication Date: 17 October 2018
Original Paper

Structural bioinformatics

Quality assessment for the putative intrinsic disorder in proteins

Gang Hu[1], **Zhonghua Wu**[1], **Christopher J. Oldfield**[2], **Chen Wang**[2] **and Lukasz Kurgan**[2,*]

[1]School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, People's Republic of China and [2]Department of Computer Science, Virginia Commonwealth University, Richmond, VA 23284, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on June 7, 2018; revised on September 19, 2018; editorial decision on October 14, 2018; accepted on October 15, 2018

Abstract

Motivation: While putative intrinsic disorder is widely used, none of the predictors provides quality assessment (QA) scores. QA scores estimate the likelihood that predictions are correct at a residue level and have been applied in other bioinformatics areas. We recently reported that QA scores derived from putative disorder propensities perform relatively poorly for native disordered residues. Here we design and validate a general approach to construct QA predictors for disorder predictions.

Results: The QUARTER (QUality Assessment for pRotein inTrinsic disordEr pRedictions) toolbox of methods accommodates a diverse set of ten disorder predictors. It builds upon several innovative design elements including use and scaling of selected physicochemical properties of the input sequence, post-processing of disorder propensity scores, and a feature selection that optimizes the predictive models to a specific disorder predictor. We empirically establish that each one of these elements contributes to the overall predictive performance of our tool and that QUARTER's outputs significantly outperform QA scores derived from the outputs generated the disorder predictors. The best performing QA scores for a single disorder predictor identify 13% of residues that are predicted with 98% precision. QA scores computed by combining results of the ten disorder predictors cover 40% of residues with 95% precision. Case studies are used to show how to interpret the QA scores. QA scores based on the high precision combined predictions are applied to analyze disorder in the human proteome.

Availability and implementation: http://biomine.cs.vcu.edu/servers/QUARTER/

Contact: lkurgan@vcu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Intrinsic disorder is characterized by lack of stable tertiary structure under physiological conditions (van der Lee *et al.*, 2014). Recent estimates reveal that up to 50% of eukaryotic proteins have at least one long ($\geq$ 30 consecutive residues) intrinsically disordered region (IDR) (Ward *et al.*, 2004; Xue *et al.*, 2012a) and that approximately 19% of residues in these proteins are disordered (Peng *et al.*, 2015). As a part of their very diverse functional repertoire, proteins with IDRs are enriched in cellular functions that involve protein–protein, protein–nucleic acids and virus–host interactions (Dyson, 2012; Fan *et al.*, 2014; Fuxreiter *et al.*, 2014; Hu *et al.*, 2017; Meng *et al.*, 2016; Peng *et al.*, 2014b, 2015; van der Lee *et al.*, 2014; Wang *et al.*, 2016; Xue *et al.*, 2014; Xue *et al.*, 2012b). Given the high abundance, functional importance and the fact that IDRs can be accurately predicted from the protein sequences (Monastyrskyy *et al.*, 2014; Necci *et al.*, 2017a; Peng and Kurgan, 2012; Walsh *et al.*, 2015), many sequence-based computational disorder predictors were developed (Atkins *et al.*, 2015; Deng *et al.*, 2012; Kozlowski and Bujnicki, 2012; Meng *et al.*, 2017). Their predictions are used to support and plan experimental studies and to investigate prevalence and functions of disorder on a large, genomic scale (Peng *et al.*, 2013, 2014a, 2015; Pentony and Jones, 2010; Wang *et al.*, 2016; Xue *et al.*, 2012a). The intrinsic disorder predictions are also used in other areas, such as structural genomics (Oldfield *et al.*, 2013). Two large databases of putative intrinsic disorder: MobiDB (Di Domenico *et al.*, 2012; Piovesan *et al.*, 2018; Potenza *et al.*, 2015) and D^2P^2 (Oates *et al.*, 2012), enjoy a substantial amount of interest. The MobiDB and D^2P^2 articles were cited 223 and 200 times, respectively (source: Google Scholar on June 7, 2018).

Perhaps surprisingly, the easy to collect disorder predictions do not include quality assessment (QA) scores. We recently defined QA for the prediction of disorder (Wu *et al.*, 2017). The QA scores are not disorder predictions but rather they are produced separately to accompany these predictions. They quantify correctness (confidence) of the disorder predictions at a residue level to reveal which predictions are more likely to be correct. High values of QA scores correspond to correctly predicted native disordered and structured residues. On the other hand, residues that are incorrectly predicted (native disordered predicted as structured or vice versa) should have low QA scores. Importantly, the QA scores must be optimized for specific predictors of disorder since these methods rely on different types of disorder annotations (Meng *et al.*, 2017; Piovesan *et al.*, 2018). While QA scores would be very useful to provide context for the disorder predictions, to date there are no QA predictors for the intrinsic disorder. In contrast, QA of putative tertiary protein structures has been intensely researched (Cao *et al.*, 2016, 2017; Kihara *et al.*, 2009; McGuffin *et al.*, 2013; Skwark and Elofsson, 2013).

In Wu et al. (2017), we used a large benchmark set of 26 thousand proteins to empirically assess whether the propensities of intrinsic disorder generated by ten popular and computationally efficient disorder predictors (i.e. methods that process such large dataset in no more than several days) can be used as QA scores. We name them default quality assessment (DQA) scores; they are defined in Section 2.2. The ten predictors include three versions of ESpritz that predict disorder annotated using X-ray structures, NMR structures and DisProt database (Walsh *et al.*, 2012), two versions of IUPred (Dosztanyi *et al.*, 2005), two versions of DisEMBL (Linding *et al.*, 2003), RONN (Yang *et al.*, 2005), VSL2B (Peng *et al.*, 2006) and GlobPlot (Linding *et al.*, 2003). Our results reveal that DQA scores are inaccurate, especially for the native disordered residues. Supplementary Figure S1 shows that the native disordered residues that are incorrectly predicted as structured have high DQA values (and low putative propensities for disorder) for 9 out of the 10 methods. This is why their ROC curves lie close to or even below the diagonal line, in particular for the low false positive rates. The only predictor for which DQA scores are reasonable is VSL2B (thick black line in Supplementary Fig. S1). However, the predictive quality of VSL2B's DQA scores is rather modest, with the AUC = 0.66 and $\text{AUC}_{\text{lowFPR}} = 0.005$ (AUC for the FPR range < 0.05; Supplementary Fig. S1B) for the native disordered residues.

We address the lack of methods that produce QA scores with desirable levels of predictive performance. We have devised a comprehensive and innovative solution that:

- provides ten QA methods for the ten corresponding disorder predictors that were investigated in Wu *et al.* (2017). We focus on this comprehensive group of methods because: (i) they are computationally efficient, i.e. they can predict our large dataset (6271 proteins and 1 778 616 residues) and genome-scale protein sets in no more than several days; (ii) they are included in the popular MobiDB database (Piovesan *et al.*, 2018); and (iii) they were empirically shown to provide accurate predictions (Walsh *et al.*, 2015) and their consensus accurately predicts long IDRs (Necci *et al.*, 2017b).
- optimizes each of the ten QA methods for the specific disorder predictor. This way we demonstrate that accurate QA scores can be produced for methods that rely on different sources of the disorder annotations, including X-ray structures, NMR structures and a variety of other experimental techniques that are covered in the DisProt database.
- uses a novel approach to design inputs for the models that generate QA scores. We rely on three ideas: (i) scaling of the sequence-derived physicochemical properties of the input protein chain; (ii) use of DQA scores; and (iii) empirical selection of predictive inputs. The latter allows us to optimize these models for the specific disorder predictors. Consequently, our QA values outperform DQA values.
- provides the resulting ten QA methods as an easy to use and publicly available webserver.

2 Materials and methods

2.1 Datasets

Disorder predictions were extracted from the MobiDB resource (Piovesan *et al.*, 2018). MobiDB provides the predictions from ten methods: $\text{DisEMBL}_{\text{remark456}}$ and $\text{DisEMBL}_{\text{HotLoops}}$ (Linding *et al.*, 2003), $\text{Espritz}_{\text{Disprot}}$, $\text{Espritz-}_{\text{NMR}}$ and $\text{Espritz}_{\text{X-ray}}$ (Walsh *et al.*, 2012), Globplot (Linding *et al.*, 2003), $\text{IUPred}_{\text{long}}$ and $\text{IUPred}_{\text{short}}$ (Dosztanyi *et al.*, 2005), RONN (Yang *et al.*, 2005) and VSL2B (Obradovic *et al.*, 2005; Peng *et al.*, 2006). We utilize a benchmark dataset with native annotation of disorder from (Walsh *et al.*, 2015) that originally includes 25 717 proteins. We removed proteins with sequences that have unknown/undetermined amino acid (AA) types. Next, we reduced pairwise sequence similarity to 25% using BLASTCLUST with the other parameters set to default (Camacho *et al.*, 2009) using the remaining 12 129 proteins. The resulting set of 6271 protein chains shares < 25% similarity and includes 105 709 disordered and 1 672 907 structured residues. We selected at random 999 proteins to establish a training dataset (to have an equal number of chains for 3-fold cross-validation); the other 5272 sequences constitute the test dataset. We opted to use 3 folds instead of the more commonly used setups with 5- or 10-folds to reduce the amount of runtime necessary to process the cross-validation. This is motivated by the relatively large size of the training dataset and the need to repeat the whole process for each of the 10 predictors. This size of the training dataset provides enough data to perform empirical design, while the larger test dataset allows for a reliable statistical analysis of differences between different approaches to the QA prediction. The training dataset is used in the 3-fold cross-validation to design and empirically parametrize the new QA methods. Once the modelling is completed, we use the test dataset that shares <25% similarity to the training proteins to comparatively evaluate these methods. The datasets are available at http://biomine.cs.vcu.edu/servers/QUARTER/.

We empirically evaluated predictive performance of the ten disorder predictors on the test dataset and, as expected, we found that these results are consistent with the results published in (Walsh *et al.*, 2015), see Supplementary Figure S2. The native disorder content in the test dataset is 5.98%, while the putative content generated by the ten predictors ranges between 4% (for Espritz$_{\mathrm{Disprot}}$) and 21% (for DisEMBL$_{\mathrm{HotLoops}}$).

2.2 DQA scores for the prediction of intrinsic disorder

Disorder predictors typically generate two outputs for each residue in the input protein sequence: a real-valued propensity score that quantifies likelihood for disorder and a binary prediction (disordered versus structured residue). The binary prediction is obtained from the propensities using a threshold, such that residues with propensities greater than the threshold are assumed to be disordered and the remaining residues are assumed to be structured. Given the maximal propensity value $V_{\max}$, minimal propensity value $V_{\min}$ and threshold C, the DQA scores are computed from the output propensities as follows:

$$DQA = \begin{cases} \dfrac{P - C}{V_{\max} - C} & if \quad P > C \\[2ex] \dfrac{P - C}{V_{\min} - C} & otherwise \end{cases}$$

Supplementary Figure S3 explains how the propensities for disorder generated with VSL2B are converted into the DQA scores. For this method, $V_{\max} = 1$, $V_{\min} = 0$ and $C = 0.5$, respectively. Essentially, the propensities that are further away from the threshold (i.e. the value where the binary prediction changes) are associated with higher DQA values.

2.3 Evaluation setup

QA scores for the disorder prediction range between 0 and 1, where higher value correspond to a higher quality prediction. They represent the quality of the binary (disordered versus structured residue) prediction. In principle, the correctly predicted residues should be associated with higher QA scores than the incorrectly predicted residues.

Supplementary Section 1 provides definitions of the measures that we use to empirically assess the predictive quality of the QA scores. Here, we just name and briefly explain these measure. We use receiver operating characteristic (ROC) curve and the area under ROC (AUC) to evaluate the QA scores. Ratio of native disordered versus structured residues is skewed, with only about 5.98% natively disordered AAs. Thus, similar to other studies that consider similarly unbalanced data (Meng and Kurgan, 2016; Yan and Kurgan, 2017; Zhang *et al.*, 2017) we compute AUCs for the low range FPR values (AUC$_{\mathrm{lowFPR}}$ for the ROC curve where FPR < 0.05). The binary predictions (accurate/high quality versus inaccurate/low quality prediction) are evaluated with MCC (Matthews correlation coefficient), true positive rate (TPR) and F_1 score at false positive rate (FPR) $= 0.05$. Moreover, to balance the evaluation between the unevenly distributed disordered and structured residues, we evaluate them separately and use a product of the two corresponding measures to quantify the results across the entire test dataset. Statistical significance of differences in the predictive performance, in particular between the new QA scores and the corresponding DQA scores, is assessed with *t*-test for normal measures, otherwise we use the Wilcoxon rank test. We verify normality with the Anderson-Darling test at 0.05 significance.

2.4 Design

The QUARTER (QUality Assessment for pRotein inTrinsic disordEr pRedictions) model generates QA scores at the residue level and is designed specifically for a given disorder predictor. However, we used the same set of steps to design QUARTER for each of the ten considered disorder predictors. The overall design consists of three layers (Fig. 1):

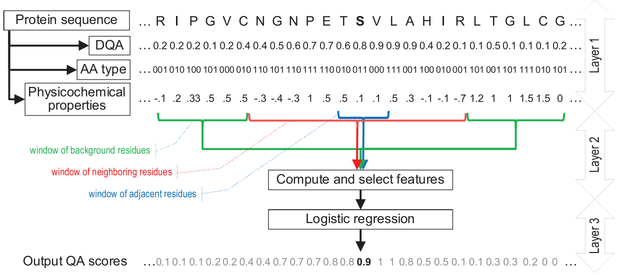

Fig. 1. Architecture of the QUARTER method

- The first layer uses the input protein chain to derive a rich profile of properties that are relevant to the QA prediction. The profile includes the sequence, DQA scores produced by a given disorder predictor, and a set of selected physicochemical properties of the input protein.
- The second layer converts the profile into a fixed number of custom-designed numerical features. We utilize information about the predicted residue and its neighbors in the sequence to compute these features. We use three types of sliding windows that are color-coded in Figure 1: window of 3 adjacent residues (in blue); window of 13 neighboring residues (in red); and window of background residues (in green). We use the background residues as a contrast to the physicochemical properties of the neighboring residues. We empirically selected 13 as the size of the window of neighboring residues because 80% of native disordered segments in the training dataset ≤ this size. The residues in the background window extend this window by another 12 residues (6 on each side). We pad the windows at either terminus of the sequence with the residue from the center of the window. In total, we consider over 150 features that are computed from the profile using these three types of windows. We optimize QUARTER for a specific disorder predictor by empirically selecting a subset of these features that maximizes predictive performance.
- The third layer inputs the selected (disorder predictor-specific) features into a logistic regression model that outputs the QA scores.

We selected the logistic regression as our predictive model based on several factors: i) this model outputs real numbers in the 0 to 1 range that intuitively correspond to the QA scores; ii) simplicity of this linear model reduces likelihood of overfitting it into the training dataset; iii) logistic regression has been successfully used to predict disorder and functions of disorder (Meng and Kurgan, 2016; Obradovic *et al.*, 2005; Peng *et al.*, 2014a, 2017), protein functions (Zhang *et al.*, 2018), protease cleavage sites (Song *et al.*, 2018) and post-translational modification sites (Li *et al.*, 2018); and iv) compared to other popular models, like SVMs and neutral networks, this model is more computationally efficient to train from the training dataset and to produce predictions. Runtime efficient training has allowed us to execute wrapper-based feature selection to optimize design of our models for specific disorder predictors. Fast predictions are important when the model is applied on a genomic scale, and facilitate our analysis of results on human proteome. The prediction is simply computed as a sum of multiplications, which requires relatively few calculations when the number of features is low.

Detailed description of the three-layered design of QUARTER is included in Supplementary Section 2. The first layer introduces on a wide range of relevant physicochemical properties that include putative relative solvent accessibility (RSA), sequence complexity, hydrophobicity, charge, flexibility and propensity for the intrinsic disorder. They are estimated/predicted from the input sequence using several fast tools that have sub-second runtime. We scale values of the physicochemical properties to ensure that they are compatible with the QA scores. The second layer generates three groups of features from the properties collected in the first layer using the three sliding window types:

1. **DQA-based features** that are computed from DQA values generated by the corresponding disorder predictor.
2. **Sequence-based features** that are based on composition of AAs and length of the input protein chain.
3. **Physicochemical properties-based features** that encode information about scaled putative RSA, scaled hydrophobicity, scaled disorder propensity, scaled flexibility, net charge and sequence complexity.

We also attempted to use features computed directly from propensities predicted by the disorder predictors instead of the corresponding DQA scores. The corresponding models performed poorly given that the putative disorder propensities and QA scores are not compatible (both low and high values of propensity correspond to high QA values). Altogether, we consider 17 (DQA-based) + 81 (Sequence-based) + 70 (property-based) = 168 features. We use wrapper-based feature selection to select a subset of relevant features that optimizes the predictive model for a specific disorder predictor. This approach was selected based on its successful use in related studies (Mizianty *et al.*, 2010; Yan *et al.*, 2016; Yan and Kurgan, 2017) and the fact that it directly optimizes the selected feature set for the predictive models that are ultimately used to make the predictions. We use the wrapper selection to maximize predictive performance measured with $\text{AUC}_{\text{lowFPR}}$ in the 3-fold cross-validation on the training dataset. We specifically focus on the $\text{AUC}_{\text{lowFPR}}$ of the native disordered residues since DQA values for these residues have low predictive performance (Supplementary Fig. S1). Supplementary Table S1 summarizes results of the feature selection across the 10 disorder predictors while Supplementary Table S2 provides ranking of features for each considered disorder predictor. The selected features vary substantially between these predictors. For instance, QA score predictor for $\text{ESpritz}_{\text{X-Ray}}$ is based primarily on the DQA-based features, while the predictors for RONN and VSL2B rely mostly on the physiochemical features. Possible reasons why the selected feature sets are so diverse are that the underlying disorder predictors use different sources of disorder annotations, have different architectures and use different predictive inputs (Meng *et al.*, 2017; Peng and Kurgan, 2012). However, each disorder predictor-specific version of QUARTER uses features from each of the three feature groups. The GlobPlot's version includes the smallest set of 7 features, while the version for $\text{ESpritz}_{\text{NMR}}$ boasts the largest set of 21 features. Moreover, each selected features set is relative small, allowing for the runtime efficient predictions.

3 Results

3.1 Empirical analysis of The Quarter model

QUARTER relies on several key ideas: (i) scaling of the values of the input physicochemical properties; (ii) inclusion of the DQA-based features; and (iii) use of the feature selection. We empirically study contributions of these three design factors. To do that, we compare the predictive performance of QUARTER with the following five of its versions where we remove some of these factors:

Version 1. QA predictor with feature selection but without DQA-based features and scaling of the physicochemical properties (no scaling, no DQA-based features and with feature selection)

Version 2. QA predictor without feature selection and using only DQA-based features (with scaling, only DQA features and no feature selection)

Version 3. QA predictor with feature selection using only DQA-based features (with scaling, only DQA features and with feature selection)

Version 4. QA predictor without feature selection and using all 168 features (with scaling, all features and no feature selection)

Version 5. QA predictor without feature selection, DQA-based features and scaling of the physicochemical properties (no scaling, no DQA-based features and no feature selection)

We retrained the corresponding five models for each of the 10 disorder predictors in the 3-fold cross validation on the training dataset. Figure 2 compares results between these five versions (the five yellow boxplots) and the original QUARTER (red boxplots) on the test dataset. Boxplots summarize results over the ten disorder predictors where the error bars give the lowest and highest values and the thick black line is the median of the ten results. To ease comparisons, the plots show ratio of the value of a given measure to the median for the DQA scores (thick black line in the blue boxplots). For instance, QUARTER's median $\text{AUC}_{\text{lowFPR}}$=3 means that it is three times better than the median for the DQA scores. The numbers at the top of the yellow boxes summarize statistical significance of differences between QUARTER and the five setups in the (x+, y-, z=) format, where x, y and z denote the number of disorder predictors for which the QA scores from QUARTER are significantly better, worse and not significantly different at P-value < 0.05, respectively. For example, (7+, 1-, 2=) in Figure 2A means that when considering the results for the ten disorder predictors QUARTER was seven times significantly better, once significantly worse and twice the difference was not significant when compared to the second version (QA predictor with scaling, only DQA features and no feature selection).

Figure 2 reveals that the five reduced versions produce on average (over the 10 disorder predictors) much lower predictive quality when compared to QUARTER. The QUARTER's median values of $\text{AUC}_{\text{lowFPR}}$, MCC, TPR and F_1 are higher than the corresponding results for the five reduced versions. QUARTER also secures the second highest median AUC, behind only the version that is based on all 168 features. However, the same *version 4* secures much lower $\text{AUC}_{\text{lowFPR}}$, which is arguably a more adequate measure given the unbalanced nature of the dataset (i.e. the native disordered residues make up about 5% of the test dataset). As expected, the biggest drop in the predictive quality measured with $\text{AUC}_{\text{lowFPR}}$ is for the *version 5* where we exclude all three key design factors. Interestingly, *version 4* that only excludes the feature selection has the second worst range of the $\text{AUC}_{\text{lowFPR}}$ values (Fig. 2A). When compared to this version, the QA scores produced by QUARTER are significantly better for nine of the ten disorder predictors. Moreover, *version 1* that includes only the feature selection and removes the other two design factors (scaling and DQA-based features) produces the second best (after QUARTER) median $\text{AUC}_{\text{lowFPR}}$. These two observations suggest that feature selection substantially contributes to the predictive performance of

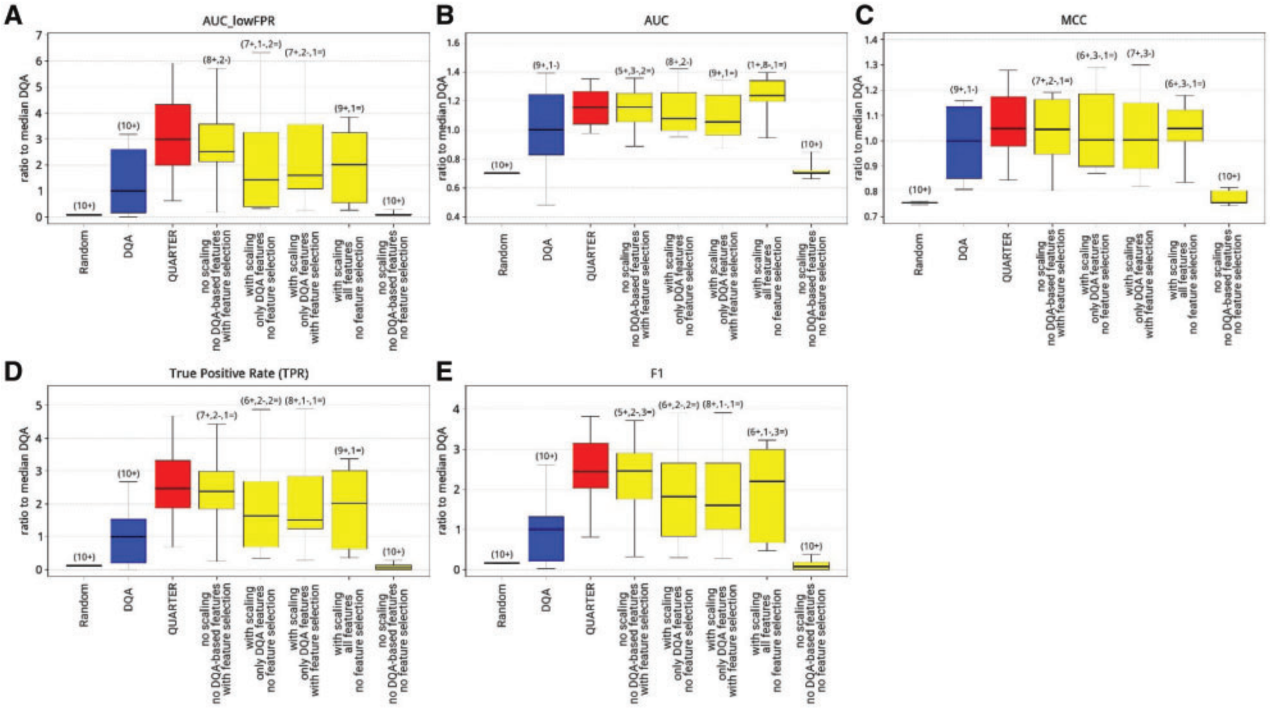

Fig. 2. Evaluation of predictive performance measured with $\text{AUC}_{\text{lowFPR}}$ (**A**), AUC (**B**), MCC (**C**), TPR (**D**) and F_1 (**E**) on the test dataset. We compare QA scores generated with QUARTER (in blue), five versions of QUARTER where specific design ideas were removed (yellow), DQA scores (blue) and random scores (black). Each boxplots summarizes results over the ten disorder predictors where whiskers denote the lowest and highest results, box defines the range between the first and third quartile, and the thick black line is the median result. To ease comparison, the plots show ratio of a given measure to the median result for the DQA scores, e.g. QUARTER's median $\text{AUC}_{\text{lowFPR}}=3$ means that it is three times better than the median for the DQA scores. The numbers at the top of the boxes summarize statistical significance of differences between QUARTER and the other methods using *x*+, *y*-, *z*= format, where *x*, *y* and *z* are the number of disorder predictors for which the QA scores from QUARTER are significantly better, worse and not significantly different, respectively. Details of the test are given in the Supplementary Material; we assume that differences are significant when P-value < 0.05. For example, (7+, 1-, 2=) means that QUARTER is seven times significantly better, once significantly worse and twice the difference between QUARTER and the other method is not significant. (Color version of this figure is available at *Bioinformatics online.*)

QUARTER. This stems from the fact that we use feature selection to optimize the QUARTER's design for specific and different disorder predictors. Furthermore, *version 3* that uses solely the DQA-based features is also much worse that QUARTER; the differences in $\text{AUC}_{\text{lowFPR}}$, MCC, TPR and F_1 are statistically significant for 7, 7, 8 and 8 out of the 10 disorder predictors, respectively. This clearly demonstrates the importance of the physicochemical properties and sequence-based features that we use to implement QUARTER. Finally, removal of the scaling of physicochemical properties (*versions 1* and *5*) similarly results in a much lower predictive performance. Altogether, these results suggest that each of the three design elements provides strong contribution to the QUARTER's predictive performance.

3.2 Empirical comparison with other approaches

Figure 2 also compares the QA scores computed by QUARTER (red boxplots) with the DQA scores (blue boxplots) and scores that are generated at random (gray boxplots). The latter scores are random numbers in the 0 to 1 range, generated such that they produce the same number of predicted positives (putative correct predictions) as the DQA scores for a given disorder predictor. Results produced by QUARTER significantly outperform both DQA and random scores. Specifically, QUARTER's $\text{AUC}_{\text{lowFPR}}$ (Fig. 2A), TPR (Fig. 2D) and F_1 (Fig. 2E) are significantly higher (P-value < 0.05) than the corresponding measures for DQA and random scores for each of the ten disorder predictors. Similarly, the QUARTER's AUC (Fig. 2B) and MCC (Fig. 2C) significantly outperform random predictions and DQA for 10 and 9 disorder predictors, respectively.

Figures 3A, 4B, C and D provide side-by-side comparison of results for individual disorder predictors on the test dataset. They include $\text{AUC}_{\text{lowFPR}}$ and F_1 values for the native disordered and all residues. QUARTER outperforms DQA scores and random scores on both measures when tested across the ten disorder predictors on all residues (Fig. 3C and D). The results on the native disordered residues (Fig. 3A and B) show that the QA scores produced by QUARTER are always better than the random scores and almost always better than the DQA scores. The only exception are the results for the $\text{DisEMBL}_{\text{remark456}}$ where QUARTER provides lower predictive quality for the native disordered residues (Fig. 3A and B) but much better results overall (Figs 3C and 4D); this suggests that QUARTER's QA scores for $\text{DisEMBL}_{\text{remark456}}$ are much better for the native structured residues. We note the relatively bad results of the DQA scores generated by majority of the disorder predictors (except $\text{DisEMBL}_{\text{remark456}}$, RONN and VSL2B) for the native disordered residues; see blue and gray bars in Figure 3A and B. These results are a consequence of the trend that we discuss in the Introduction (Supplementary Fig. S1). Given this trend, we compare the ROC curves for QA scores from QUARTER (Fig. 3E) and with the corresponding DQA scores (Supplementary Fig. S1B). The QUARTER' scores secure TPRs that are much above the random predictor levels (diagonal line where TPRs equal FPRs) for all disorder predictors. This is stark contrast to most of the DQA scores (7 out of 10 disorder predictors) that are below the random levels

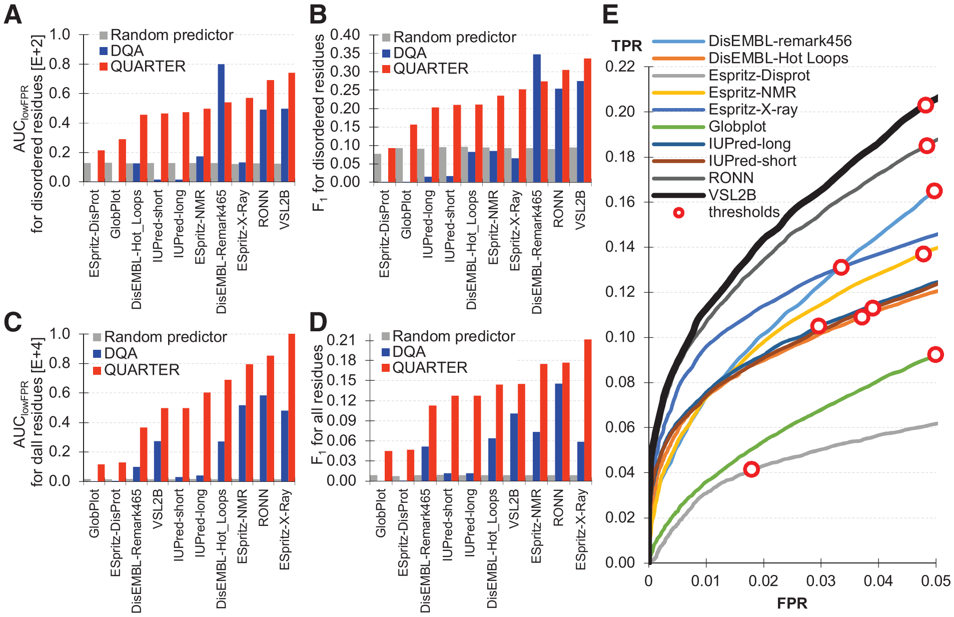

Fig. 3. Comparison of predictive performance between QUARTER, DQA scores and scores generated at random for individual disorder predictors on the test dataset. (**A**) and (**B**) compare AUC$_{lowFPR}$ and F$_1$ for the native disordered residues, respectively. (**C**) and (**D**) compare AUC$_{lowFPR}$ and F$_1$ for all residues (computed as a product of these values for the disordered and the structured residues; see Supplementary Material for details). (**E**) gives the ROC curves for the low FPR range (FPR $\leq$ 0.05) for the QA scores generated with QUARTER for the native disordered residues. Red circles correspond to disorder predictor-specific thresholds that we established to select a subset of high quality predictions based on the QA scores from QUARTER

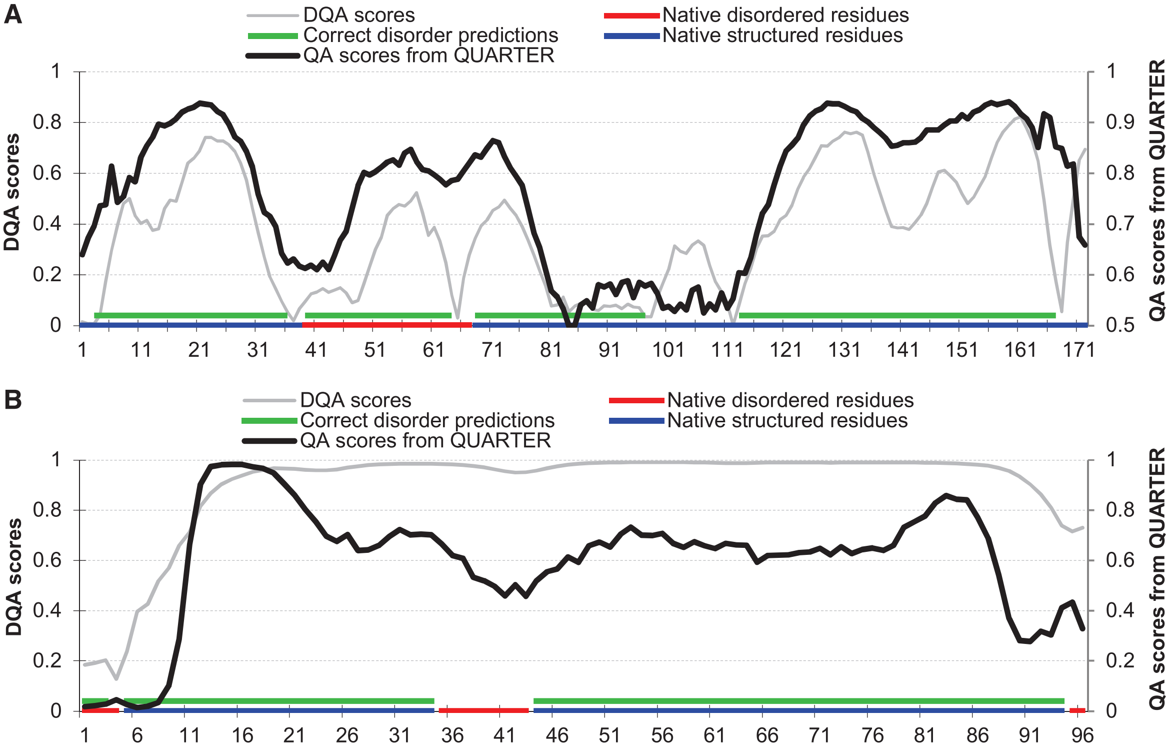

Fig. 4. Comparison between DQA scores (gray curves) and QA scores generated by QUARTER (black curves). (**A**) shows results for the predictions with VSL2B for the p23 protein (UniProt ID: P13693). (**B**) gives results for the predictions with ESpritz$_{DisProt}$ for the nuclease YbcO (UniProt ID: P68661). The color-coded horizontal lines on the *x*-axis identify the native disordered residues (in red), native structured residues (in blue) and correct predictions generated by a given predictor (in green). (Color version of this figure is available at *Bioinformatics online*.)

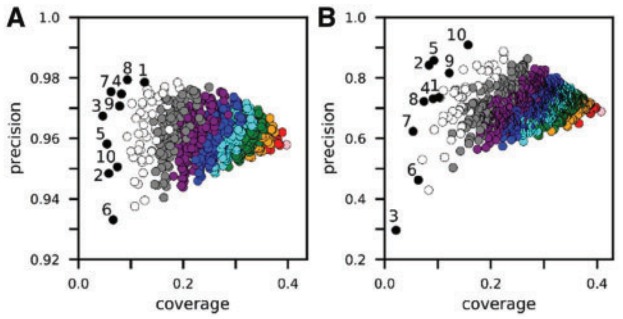

Fig. 5. Comparison of coverage and precision for selected high quality thresholds (black points) and all combinations of 2 to 10 disorder predictors (white, grey, purple, blue, cyan, green, orange, red and pink point, respectively). (**A**) compares results for the entire test dataset, where labels correspond to individual predictors (coverage, precision): 1, DisEMBL$_{\text{remark456}}$ (0.13, 0.98); 2, DisEMBL$_{\text{HotLoops}}$ (0.058, 0.95); 3, ESpritz$_{\text{DisProt}}$ (0.046, 0.97); 4, ESpritz$_{\text{NMR}}$ (0.082, 0.97); 5, ESpritz$_{\text{X-Ray}}$ (0.055, 0.96); 6, GlobPlot (0.066, 0.93); 7, IUPred$_{\text{long}}$ (0.062, 0.98); 8, IUPred$_{\text{short}}$ (0.093, 0.98); 9, RONN (0.079, 0.97); 10, VSL2B (0.074, 0.95). (**B**) compares results for the disordered residues from the test dataset, where labels correspond to individual predictors (coverage, precision): 1, DisEMBL$_{\text{remark456}}$ (0.1, 0.74); 2, DisEMBL$_{\text{HotLoops}}$ (0.084, 0.84); 3, ESpritz$_{\text{DisProt}}$ (0.021, 0.3); 4, ESpritz$_{\text{NMR}}$ (0.091, 0.73); 5, ESpritz$_{\text{X-Ray}}$ (0.093, 0.86); 6, GlobPlot (0.063, 0.46); 7, IUPred$_{\text{long}}$ (0.054, 0.62); 8, IUPred$_{\text{short}}$ (0.074, 0.72); 9, RONN (0.12, 0.82); 10, VSL2B (0.16, 0.91). (Color version of this figure is available at *Bioinformatics online*.)

(Supplementary Fig. S1B). In particular, the QUARTER's scores secure TPRs ≥ 0.12 at the FPR $= 0.05$ for 8 out the 10 disorder predictors, and TPRs > 0.19 for RONN and VSL2B. We focus on the results for the low FPRs given the low, 5% content of the native disordered residues in the test dataset. The QUARTER's ability to perform well on the native disorder residues stems from the feature selection that specifically targets improvements on these residues.

To sum up, when compared to DQA and random scores, QUARTER provides significantly better predictive performance on the disordered residues while also maintaining favorable predictive quality on all residues.

3.3 Case studies

We use two proteins from the test dataset to demonstrate practical value of the putative QA scores generated by QUARTER. The first protein (Fig. 4A) was predicted with VSL2B that produces reasonably good DQA scores for the disordered residues (Supplementary Fig. S1). Predictions at the N-terminus are incorrect and both QUARTER and DQA correctly identify these problems. However, only QUARTER suggests that the VSL2B's predictions are in fact incorrect at the C-terminus. Similarly, a longer region between the positions 98 and 112 that is incorrectly predicted by VSL2B coincides with an appropriate dip in the QUARTER scores, while DQA scores crest there. The putative disorder for the second protein (Fig. 4B) was predicted by ESpritz$_{\text{DisProt}}$ that offers lower quality DQA scores (Supplementary Fig. S1). Both termini of this protein were incorrectly predicted and correspondingly both QUARTER and DQA scores identify these problematic areas. However, only QUARTER finds the low quality prediction of the disordered region between positions 35 and 43. This means that ESpritz$_{\text{DisProt}}$ entirely misses this disordered regions and QUARTER identifies this issue.

3.4 Quality assessment of disorder predictions in the human proteome

We use QUARTER to select and characterize a subset of high quality disorder predictions in the human proteome. First, we setup selection of the high quality predictions on the test dataset in two steps: (i) for each of the ten disorder predictors; and (ii) we combine the results from individual predictors to improve overall coverage. Finally, we use the best setup to characterize the high quality putative disorder in the human proteins.

In the first setup step we use the QUARTER's QA scores to identify a subset of high quality (i.e. high precision) predictions for each disorder predictor. The corresponding optimal QA score threshold (i.e. furthest from random) or FPR $= 0.05$ threshold was selected on the test dataset, whichever had a lower FPR (Fig. 3E, red circles). The black markers in Figure 5A show precision and coverage of these high quality predictions. The best performing high quality predictions for DisEMBL$_{\text{remark456}}$ (marked as method 1) secure 98% precision and cover 13% of residues.

In the second setup step, we combine the high quality predictions from multiple disorder predictors to increase coverage. The individual high quality predictions were combined residue-by-residue by defining a high quality prediction if any QA predictor gave a high quality prediction. For the high quality residues, the predictor with the greatest QA score defines the order-disorder prediction. For the remaining low quality residues, disorder predictions were combined by majority vote. We evaluate the combined predictions by precision and coverage of the high quality predictions (Fig. 5). We assess all possible combinations of the ten methods. In general, coverage of the test dataset with high quality predictions increased with larger predictor combinations, accompanied by a narrowing range of precision. Maximum coverage and adequate 95.5% precision was achieved by combining all ten predictors (Fig. 5A, pink point). This best result covers 40% of the test dataset and offers proportional coverage of the native ordered and disordered residues, at 40 and 41%, respectively.

We apply the ten-way combination of predictions to the reviewed human proteome collected from the UniProt database (The UniProt Consortium, 2017), which includes 20 737 proteins. Individual and combination predictions for these human proteins are available at http://biomine.cs.vcu.edu/servers/QUARTER/. The overall coverage by the high quality predictions among human residues, 42%, is similar to the coverage in the test dataset. This suggests that the results on the human proteome are characterized by similar (to the test dataset) predictive performance. However, the proportion of predicted disorder is much different for human residues than the test dataset, which are 38 and 6.5%, respectively. This is not unexpected. The proportion of the disordered residues in the test dataset, 6.0%, is comparable to the predicted estimate. Moreover, previous estimates of the proportion of disordered residues in human proteins are much larger than in the test dataset (Xue *et al.*, 2012a,b). The high quality estimate of the proportion of intrinsic disorder agrees with a previous estimate of the frequency of disordered residues in eukaryotes which is between 35 to 45% (Xue *et al.*, 2012a). We note that the high quality predictions cover less than half of human residues, so this estimate may not accurately represent the entire proteome.

The high quality predictions are distributed broadly among human proteins (Fig. 6, black line). A median protein has 40% of residues with the high quality predictions. The distribution has a long upper tail, with 16% of proteins having over 60% of residues with the high quality predictions. Partitioning residues into predicted disorder and structure (Fig. 6, red and blue lines, respectively) shows a narrower distribution of the high quality order predictions and a broader distribution for the disorder predictions. This indicates a larger variance in the amount of the high quality disorder predictions from protein to protein.

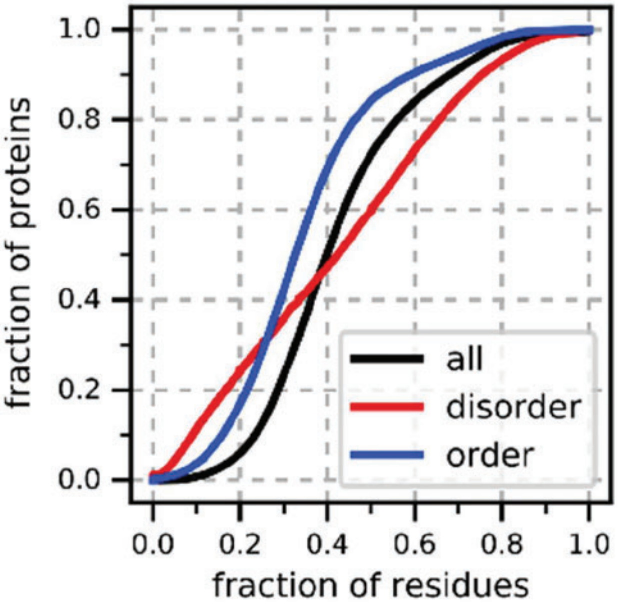

Fig. 6. Empirical cumulative distributions of high quality predictions in human proteins based on the 10-way combination. The black line shows the fraction of human proteins with at least the given fraction of residues with high quality predictions. The blue (red) lines show distribution of high quality predictions for the predicted structured/ordered (disordered) residues. (Color version of this figure is available at *Bioinformatics online.*)

3.5 Functional characterization of human proteins predicted with high quality

Proteins with over 60% of residues that have high quality predictions were examined for bias in functional annotations to investigate if there are distinguishing features of these 3280 accurately predicted proteins. This set was further divided into two groups: putative disordered proteins that have majority of the high quality disorder predictions (HQD, 1236 proteins) and putative structured proteins with majority of high quality structure predictions (HQS, 2042 proteins). These two protein sets were submitted to the PANTHER server (Mi *et al.*, 2017) for analysis of overrepresentation of Gene Ontology (GO) annotations relative to what would be expected from a random sample of the human proteome.

Relative enrichment was found for HQD and HQS proteins for several terms in the molecular function and cellular component ontologies (Supplementary Fig. S4). Enrichment found for one set of proteins often corresponds with depletion for the other, indicating that these annotations are structure and disorder specific. HQD proteins share several distinguishing GO terms with previous characterizations of intrinsic disorder. Binding to nucleic acids, both RNA and DNA, is a well-known function of many proteins with IDRs (Dyson, 2012; Wang *et al.*, 2016). Also, many proteins with IDRs are known to localize to the nucleus (Frege and Uversky, 2015). Proteins with IDRs are also known to play a role in the extracellular matrix and associate with the cytoskeleton (Dunker *et al.*, 2015). For HQS proteins, enriched molecular functions and cellular components all indicate integral or peripheral membrane involvement. Transmembrane domains are often well structured, but transmembrane proteins often contain cytoplasmic or extracellular intrinsically disordered domains (Xue *et al.*, 2009). These results do not contradict this; HQS proteins are not necessarily predicted to be entirely ordered; only the majority of their high quality predictions are ordered. Overall, both the HQD and HQS term biases are consistent with previous results for disordered and ordered protein biases.

4 Summary and conclusions

We design, implement, empirically test and release QUARTER, first-of-its-kind tool that provides accurate QA scores for ten popular disorder predictors. The QA scores are individually optimized for each of the ten disorder predictors. This optimization is guided by an empirical feature selection and takes advantage of the disorder predictor-specific DQA scores.

Empirical tests on a large test dataset reveal that QUARTER generates high quality results. These tests show that our novel tool produces accurate QA scores for a wide range of disorder predictors that utilize different sources of disorder annotations. The QUARTER's $\text{AUC}_{\text{lowFPR}}$ values are on average 300% higher than the corresponding values for the DQA scores (Fig. 2A). We numerically demonstrate that our QA scores secure significantly better predictive performance on the native disordered residues when compared to DQA and random scores. The empirical tests also highlight the main reasons for the high predictive quality. The three key contributing advancements are: scaling of the input physicochemical properties; inclusion of the DQA-based features; and use of the empirical feature selection.

The primary application of the QA scores is annotation/selection of a high-quality subset of residue-level disorder predictions. We show that the best QA scores for a single disorder predictor can be used to identify 13% of amino acids for which disorder is predicted with a very high, 98% precision. When combining predictions of the ten considered disorder predictors, the QA scores can be used to find 40% of residues for which disorder is predicted with 95% precision in the test dataset. Application of QUARTER to the human proteome tells that 42% of human residues are predicted with high quality (95% precision) and that about 38% of these residues are disordered. This high disorder content and our functional analysis of human proteins that are enriched in the high-quality disorder predictions are in good agreement with the existing literature.

We release QUARTER as a freely available webserver at http://biomine.cs.vcu.edu/servers/QUARTER/. Details about the webserver are provided in Section S3 in the Supplementary Material.

Acknowledgements

We thank Dr. Silvio Tosatto and his research group from the University of Padova for sharing datasets and predictions of disorder.

Funding

This research was supported in part by the National Science Foundation [grant 1617369 to L.K.], Natural Science Foundation of Tianjin [grant 18JCYBJC24900 to G.H.]

Conflict of Interest: none declared.

References

Atkins,J.D. *et al.* (2015) Disorder prediction methods, their applicability to different protein targets and their usefulness for guiding experimental studies. *Int. J. Mol. Sci.*, **16**, 19040–19054.

Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Cao,R. *et al.* (2017) QAcon: single model quality assessment using protein structural and contact information with machine learning techniques. *Bioinformatics*, **33**, 586–588.

Cao,R. *et al.* (2016) Massive integration of diverse protein quality assessment methods to improve template based modeling in CASP11. *Proteins*, **84**, 247–259.

Deng,X. *et al.* (2012) A comprehensive overview of computational protein disorder prediction methods. *Mol. Biosyst.*, **8**, 114–121.

Di Domenico,T. *et al.* (2012) MobiDB: a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics*, **28**, 2080–2081.

Dosztanyi,Z. *et al.* (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.

Dunker,A.K. *et al.* (2015) Intrinsically disordered proteins and multicellular organisms. *Semin. Cell Dev. Biol.*, **37**, 44–55.

Dyson,H.J. (2012) Roles of intrinsic disorder in protein–nucleic acid interactions. *Mol. Biosyst.*, **8**, 97–104.

Fan,X. *et al.* (2014) The intrinsic disorder status of the human hepatitis C virus proteome. *Mol. Biosyst.*, **10**, 1345–1363.

Frege,T. and Uversky,V.N. (2015) Intrinsically disordered proteins in the nucleus of human cells. *Biochem. Biophys. Rep.*, **1**, 33–51.

Fuxreiter,M. *et al.* (2014) Disordered proteinaceous machines. *Chem. Rev.*, **114**, 6806–6843.

Hu,G. *et al.* (2017) Functional analysis of human hub proteins and their interactors involved in the intrinsic disorder-enriched interactions. *Int. J. Mol. Sci.*, **18**, 2761.

Kihara,D. *et al.* (2009) Quality assessment of protein structure models. *Curr. Protein Pept. Sci.*, **10**, 216–228.

Kozlowski,L.P. and Bujnicki,J.M. (2012) MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. *BMC Bioinformatics*, **13**, 1–11.

Li,F. *et al.* (2018) Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome. *Bioinformatics*, https://doi.org/10.1093/bioinformatics/bty522.

Linding,R. *et al.* (2003) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459.

Linding,R. *et al.* (2003) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701–3708.

McGuffin,L.J. *et al.* (2013) The ModFOLD4 server for the quality assessment of 3D protein models. *Nucleic Acids Res.*, **41**, W368–W372.

Meng,F. and Kurgan,L. (2016) DFLpred: high-throughput prediction of disordered flexible linker regions in protein sequences. *Bioinformatics*, **32**, i341–i350.

Meng,F. *et al.* (2016) Compartmentalization and functionality of nuclear disorder: intrinsic disorder and protein–protein interactions in intra-nuclear compartments. *Int. J. Mol. Sci.*, **17**, 24.

Meng,F. *et al.* (2017) Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. *Cell Mol. Life Sci.*, **74**, 3069–3090.

Mi,H. *et al.* (2017) PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res.*, **45**, D183–D189. (Database issue):

Mizianty,M.J. *et al.* (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics*, **26**, i489–i496.

Monastyrskyy,B. *et al.* (2014) Assessment of protein disorder region predictions in CASP10. *Proteins*, **82**, 127–137.

Necci,M. *et al.* (2017a) A comprehensive assessment of long intrinsic protein disorder from the DisProt database. *Bioinformatics*, **34**, 445–452.

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

Oates,M.E. *et al.* (2012) D2P2: database of disordered protein predictions. *Nucleic Acids Res.*, **41**, D508–D516.

Obradovic,Z. *et al.* (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins*, **61**, 176–182.

Oldfield,C.J. *et al.* (2013) Utilization of protein intrinsic disorder knowledge in structural proteomics. *Biochim. Biophys. Acta*, **1834**, 487–498.

Peng,K. *et al.* (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, **7**, 208.

Peng,Z. *et al.* (2014a) Genome-scale prediction of proteins with long intrinsically disordered regions. *Proteins*, **82**, 145–158.

Peng,Z. *et al.* (2014b) A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome. *Cell Mol. Life Sci.*, **71**, 1477–1504.

Peng,Z. *et al.* (2017) Prediction of disordered RNA, DNA, and protein binding regions using DisoRDPbind. *Methods Mol. Biol.*, **1484**, 187–203.

Peng,Z. *et al.* (2013) Resilience of death: intrinsic disorder in proteins involved in the programmed cell death. *Cell Death Differ.*, **20**, 1257–1267.

Peng,Z. *et al.* (2015) Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell Mol. Life Sci.*, **72**, 137–151.

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

Pentony,M.M. and Jones,D.T. (2010) Modularity of intrinsic disorder in the human proteome. *Proteins*, **78**, 212–221.

Piovesan,D. *et al.* (2018) MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. *Nucleic Acids Res.*, **46**, D471–D476.

Potenza,E. *et al.* (2015) MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Res.*, **43**, D315–D320.

Skwark,M.J. and Elofsson,A. (2013) PconsD: ultra rapid, accurate model quality assessment for protein structure prediction. *Bioinformatics*, **29**, 1817–1818.

Song,J. *et al.* (2018) PROSPERous: high-throughput prediction of substrate cleavage sites for 90 proteases with improved accuracy. *Bioinformatics*, **34**, 684–687.

The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.

van der Lee,R. *et al.* (2014) Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.*, **114**, 6589–6631.

Walsh,I. *et al.* (2015) Comprehensive large-scale assessment of intrinsic protein disorder. *Bioinformatics*, **31**, 201–208.

Walsh,I. *et al.* (2012) ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*, **28**, 503–509.

Wang,C. *et al.* (2016) Disordered nucleiome: abundance of intrinsic disorder in the DNA- and RNA-binding proteins in 1121 species from Eukaryota, Bacteria and Archaea. *Proteomics*, **16**, 1486–1498.

Ward,J.J. *et al.* (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.*, **337**, 635–645.

Wu,Z. *et al.* (2017) Exploratory analysis of quality assessment of putative intrinsic disorder in proteins. In *6th International Conference on Artificial Intelligence and Soft Computing*. Zakopane, Poland. pp. 722–732.

Xue,B. *et al.* (2014) Structural disorder in viral proteins. *Chem. Rev.*, **114**, 6880–6911.

Xue,B. *et al.* (2009) Analysis of structured and intrinsically disordered regions of transmembrane proteins. *Mol. Biosyst.*, **5**, 1688–1702.

Xue,B. *et al.* (2012a) Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J. Biomol. Struct. Dyn.*, **30**, 137–149.

Xue,B. *et al.* (2012b) Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. *Cell Mol. Life Sci.*, **69**, 1211–1259.

Yan,J. *et al.* (2016) Molecular recognition features (MoRFs) in three domains of life. *Mol. Biosyst.*, **12**, 697–710.

Yan,J. and Kurgan,L. (2017) DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. *Nucleic Acids Res.*, **45**, e84.

Yang,Z.R. *et al.* (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, **21**, 3369–3376.

Zhang,C. *et al.* (2018) MetaGO: predicting gene ontology of non-homologous proteins through low-resolution protein structure prediction and protein–-protein network mapping. *J. Mol. Biol.*, **430**, 2256–2265.

Zhang,J. *et al.* (2017) Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. *Brief. Bioinform*, https://doi.org/10.1093/bib/bbx168.