Note: Published in PLOS One Journal in 2015,
see http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0141551

Dispredict: A Predictor of Disordered Protein using Optimized RBF Kernel

Sumaiya Iqbal and Md Tamjidul Hoque*
Computer Science, University of New Orleans, LA 70148, USA

Authors contributed equally to this work.
* thoque@uno.edu

Abstract

Intrinsically disordered proteins or, regions perform important biological functions through their dynamic conformations during binding. Thus accurate identification of these disordered regions have significant implications in proper annotation of function, induced fold prediction and drug design to combat critical diseases. We have innovated DisPredict, a disorder predictor that employs a single support vector machine with RBF kernel and novel features for reliable characterization of protein structure. DisPredict yields outstanding performance. In addition to 10-fold cross validation, training and testing of DisPredict was conducted with independent test datasets. The Results were consistent with both the training and test error minimal. The use of multiple data sources, makes the predictor generic. The datasets used in developing the model include disordered regions of various length which are categorized as short and long having different compositions, different types of disorder, ranging from fully to partially disordered regions as well as completely ordered regions. Through comparison with other state of the art approaches and case studies, DisPredict is found to be a useful tool with superior performance. DisPredict is available at `http://cs.uno.edu/~tamjid/Software.html`.

1 INTRODUCTION

Many protein regions and some entire proteins do not adopt well-defined, stable three-dimensional (3D) structures in an isolated state and under different non native environments [1–3]. These proteins or partial regions of proteins are called intrinsically disordered proteins (IDPs) or disordered regions in proteins (IDRs), also known as natively unstructured, denatured or unfolded. The coordinates of their backbone atoms have no specific equilibrium states and can vary largely due to variable physiological conditions, and thus adopt dynamic structural ensembles. Structurally, IDPs (or IDRs) encompass proteins or protein-regions with extended disorder, collapsed disorder and semi-collapsed disorder. These reflect differences in the underlying biophysical characteristics including low hydrophobicity and high net charge, marginal level of residual secondary structure [5, 6], dynamic side chains and secondary structures [4, 7], rapidly exchanging backbone side-chain hydrogen bonds which make a region unable to form specific secondary structure [5]. Recognition of these protein disordered regions is important for appropriate protein structure prediction, disease causing protein identification, proper annotation of function, induced folding and binding region prediction.

For the last two decades, many works have been presented in evidence that many proteins do not follow the well-known paradigm of sequence to stable structure to function. Rather these proteins adopt disordered state for complex and essential biological functions [1,4,27,35] such as cell cycle control and cellular signal transduction, transcriptional and translational regulation, membrane fusion and control pathways [1,36,37]. They participate in molecular recognition, molecular assembly and protein modification [38,39] via protein-protein, protein-nucleic acid and protein-ligand interactions as well. Disorder proteins are found to be highly associated with critical human diseases [40–42], such as cancer, amyloidoses, cardiovascular and neurodegenerative diseases, genetic diseases. Thus, identifying them assists in effective drug development [43,44].

In reality the IDPs are abundant. Approximately 70% of the structures released by Protein Data Bank (PDB) [8] contain some disordered residues [45,46]. A curated database of disordered proteins, called DisProt [9] contains annotation for 694 protein sequences and 1539 disordered regions in its current version 6.02. The IDEAL [47,48] and MobiDB [81,82] databases also provide useful collections for annotation of intrinsic disorder. PDB [8] database, which gives provision of finding disordered regions in the solved secondary or tertiary structure incorporates 105,097 protein entries. To compare, the overall number of non-redundant protein sequences is 46,968,574 according to the most recent 68 release of RefSeq database [49]. However, due to highly flexible characteristics of the residues of IDRs or, IDPs [19]), experimentally verified annotation of intrinsic disorder is growing slowly. Thus to keep pace with this large-scale increase in protein database, effective computational methods for correct identification of disordered residues in IDPs or, IDRs are necessary.

Several computational methods have been developed to fulfill the fast annotation requirements for the rapidly growing known protein sequences. Machine learning based some of these well-known approaches are PONDR series [45,50,51], DISOPRED [17], DISOPRED2 [52], DisEMBL [21], DISpro [53], RONN [54], Spritz [55], PROFbval [19,56], DisPSSMP [57,58], PrDOS [59], POODLE series [60,61], NORSnet [12], IUP [62], OnD-CRFs [63], PreDisOrder [64], SPINE-D [14] and ESpritz [65]. Several existing tools, for instance GlobPlot [66], IUPred [67], FoldIndex [68] and Ucon [11], usage knowledge such as the relative composition and propensity of amino acids. On the other hand, DISOclust [69] is based on the analysis of how disorder is related with protein folding and uses predicted three-dimensional structural characteristics. Combination of individual methods in a complementary method gave raise to effective disorder predictors, such as metaPrDOS [70], MD [22], MFDp [13], PONRD-FIT [46] and very recent MFDp2 [71].

In this article, we propose a new disorder predictor, named "DisPredict (Disorder Predictor)" [79]. DisPredict classifies ordered and disordered residues in a protein sequence with higher accuracy, specifically in terms of Mathews Correlation Coefficient (MCC) and Area Under the receiver operating characteristics Curve (AUC). Dispredict is based on novel pattern recognition algorithm, Support Vector Machine (SVM) using Radial Basis Function (RBF) as kernel. We further strengthened the classification performance of DisPredict by selecting optimized parameters of SVM which significantly improved the performance. We utilized a comprehensive set of 56 features to characterize disorder in protein sequence. We compared DisPredict's performance with existing predictors, SPINE-D [14] and MFDp [13], followed by an analysis of its performance with respect to different types of amino acid, length of disorder region and datasets.

2 MATERIALS AND METHODS

2.1 Preliminary Disordered Data Sources

In the prior studies, DisProt [9] and PDB [8] are considered as the primary repositories of IDPs. Disorder regions are composed of residues with missing coordinates in structure solved by X-ray crystallography, whereas the residues show highly variable coordinates within ensemble solved by NMR. We selected two datasets which combine sequences from PDB having disordered residues without coordinates (recorded in REMARK 465) and sequences from DisProt, having curated annotations of disorder regions including properties such as short ($\leq$ 30 residues) and long ($>$ 30 residues) disordered regions, partial as well as fully ordered or disordered chains.

2.2 Datasets

We used two different datasets, MxD and SL, to train, test and cross-validate our proposed DisPredict. MxD and SL datasets were used to train top two disorder predictors, MFDp [13] and SPINE-D [14], respectively. We collected and utilized these datasets to be able to consistently compare DisPredict with these two predictors.

The Mixed Disorder (MxD) dataset is a combination of protein sequences with disordered residues from both PDB and DisProt. Originally developed MxD dataset [13] has 514 protein sequences including 205 chains from PDB and 309 chains from DisProt. We carried out further purification by removing sequences with unknown amino acid (X-tag) since they do not have specific physicochemical properties to get corresponding features in our methodology. This led to the MxD444 dataset, with 444 chains and $214,054$ residues, that mixes $49,090$ (about 23%) disordered residues and $164,964$ (about 77%) ordered residues.

SL477 dataset was prepared by the developers of SPINE-D predictor from the benchmark SL (Short Long) dataset [10]. The SL dataset encompasses short and long disordered regions as well as ordered regions. It was built by re-annotating the sequences extracted from DisProt to include reliable order and disorder contents. Among the annotated regions in the SL dataset, 50% of the regions are of the short-disordered category. The short regions in SL dataset are of length 20 residues or less [10]. It is important to incorporate this disorder annotation in a dataset since these short disordered regions are found functionally important as they obtains induced folding with the close proximity of an appropriate partners. SL477 also includes very long disordered regions as well as completely disordered proteins, called intrinsically disordered proteins (IDPs). SL dataset comprises of proteins with disorder regions annotated by NMR experimental method as well. To achieve combination of sequences with low sequence identity, SL dataset's sequences were clustered and filtered using BLASTCLUST [16] which resulted in 477 chains with $< 25\%$ sequence identity between each pair. SL477 has total $215,343$ residues, of which $56,887$ (about 25%), $72,808$ (about 34%) and $85,648$ (about 40%) residues are annotated as disorder, order and unknown, respectively. Unknown residues are those which are marked unknown in the source datasets (such as marked, 'X' in PDB [8]). We disregarded the residues with unknown annotation during both in training and in evaluating our proposed approach.

Moreover, to test our predictor with less overlapped sequences from training dataset, we extracted two independent test datasets from two training datasets using BLASTCLUST [16]. We filtered 171 protein chains from SL477 datasets with less than 10% similarity with any sequence from MxD444 dataset. We call these 171 protein chains with $42,572$ residues as SL171. We used SL171 as test dataset to independently test our predictor's performance while it is trained by MxD444 dataset. Similarly, we extracted 134 sequences from MxD444 dataset that are independent from SL477 dataset

at 10% identity cut off. We call these 134 protein sequences with 38,823 residues as MxD134. We utilized MxD134 as test dataset to independently test our predictor's performance while it is trained by SL477 dataset.

Further, we prepared a completely new dataset that is completely independent of the training sets of DisPredict, SPINE-D [14] and MFDp [13]. We collected 48 new protein chains from DisProt [9] released after version 5.1 upto current version of 6.02. These protein sequences were combined with another 25 protein chains culled from PDB [8]. Altogether, it gave us 73 protein sequences which is a combination of 37 full disorder chains, 23 full ordered chains and 13 protein chains with disordered and ordered regions. We call this Disorder Dataset as DD73. DD73 dataset allows us to perform a robust comparison among DisPredict, SPINE-D [14] and MFDp [13], as it is independent of both SL and MxD dataset.

2.3 Input Features

Table 1. List of features used in DisPredict

Feature Category	Feature Count
Amino Acid (AA)	1
Physicochemical Property (PP)	7
PSSM Profile (PSSM)	20
Secondary Structure Content (SS)	3
Accessible Surface Area (ASA)	1
Torsion Angle Fluctuation (Φ,Ψ)	2
Monogram (MG)	1
Bigram (BG)	20
Terminal Indicator (T)	1
Total	56

Our inputs features were carefully chosen to be able to include useful properties such as the sequence information, evolutionary information as well as the structural information (listed in Table 1). Studies suggest that necessary information for the correct folding of a protein is encoded in its amino acid sequence including disorder contents [17]. Moreover, disordered regions are abundant in low complexity regions and in regions with low content of hydrophobic amino acids [19, 20]. The physicochemical properties [80] of amino acid are also found to have some degree of correlation with the length of disordered regions; as short disordered regions are mainly negatively charged while long disordered regions are nearly neutral [19, 22]. These observations motivated us to use amino acid type (AA), indicated by one numerical value out of twenty and seven physicochemical properties (PP) as features to predict disordered residues in our proposed approach.

Disordered regions and their related functions are conserved within the sequence during evolution [23], thus we considered position specific scoring matrix (PSSM) as input features to capture evolutionary information. PSSM (sequence length $\times$ 20) was generated for each sequence by executing three iterations of PSI-BLAST against NCBI's non-redundant database and were normalized further using numeric value nine [24]. We employed sequence based predicted secondary structure (SS) probabilities for helix, sheet and coil residues [24], predicted solvent accessibility (ASA) [26] and predicted backbone dihedral torsion angles (Φ and Ψ) fluctuations [25] as features. We included these six features since disordered residues can be characterized by the lack of stable secondary structure [7, 19, 27] and also the unstructured regions are found to have large solvent accessible area [22].

A feature extraction technique [28] shows that the conserved evolutionary information given by PSSM can be transformed from primary structure (amino acid

sequence) level to three dimensional structure level by computing monograms and bigrams from PSSM values, which motivated us to utilize them as features. We computed monogram feature matrix (1×20) and bigram feature matrix (20×20) for each sequence from its PSSM. Monogram feature matrix consists of one monogram value (MG) for each type of amino acid and bigram feature matrix consists of one bigram value (BG) for each pair of 20 possible amino acids, respectively. Further, our analysis based on multiple datasets collected from PDB and DisProt shows that both the monograms and bigrams follows a normal density distribution in their logarithmic space with approximately consistent median value equals to 6.0 within any dataset (Figure 1). Therefore, we used exp(6.0) to normalize these values and reduce the noise. To understand the relevance of these new features with protein disorder, we separately evaluated DisPredict's performance with and without monograns and bigrams. We found 3.2% improved balanced accuracy after including monograms and bigrams. To distinguish the terminal residues for their position specific disorder like behavior, we included terminal indicator feature (T) by encoding five residues of N-terminal as $\{-1.0, -0.8, -0.6, -0.4, -0.2\}$ and C-terminal as $\{+1.0, +0.8, +0.6, +0.4, +0.2\}$ respectively, whereas rest of the residues were labeled 0.0.

Figure 1. Density distribution curves of monograms and bigrams for (A) SL477 and (B) MxD444 dataset. The x-axis and y-axis shows the monograms/bigrams in logarithmic scale and density index of the distribution, respectively. For each figure, the dotted (*red*) and solid (*blue*) vertical lines correspond to median values of the distribution for monograms (MG) and bigrams (BG), respectively.

We included the information of neighboring residues within the features of each residue by using a sliding window, keeping the target residue at the center of the window. The motivation was to incorporate the native interactions and contacts of neighboring residues which are found to play essential roles in determining protein structures and protein folding dynamics [72, 73]. We determined the 10-fold cross-validation performance of DisPredict for 13 different window sizes (1, 3, 5, ..., 23, 25) to find the optimal window size 21. Thus, there were 1176 (since, $window\ size \times total\ feature\ count = (21 \times 56) = 1176$) features used for each residues and each features were finally scaled within the range $[-1, +1]$ before using.

2.4 SVM Design and Parameterization

DisPredict is a two-layer disorder predictor that integrates *optimization-layer* and *classification-layer*. The classification-layer is developed using a single support vector machine (SVM), namely LIBSVM [29]. Due to the working principle of SVM of simultaneously minimizing the empirical classification error (training error) and generalized error (test error) by maximizing the geometric margin of the separating hyperplane, it can be regarded as an effective technique in hard classification problems specially in bioinformatics and computational biology area. We used Gaussian or, radial basis function (RBF) kernel for the SVM to extend its capability to handle non-linearly separable classes. RBF transforms the input feature space into infinite dimension space (*i.e.* Hilbert space), which results in a linear separating hyperplane. On the other hand, in the optimization-layer of DisPredict, we selected two parameters, C and γ, where C is the cost of misclassification and γ is the parameter of fitting best mode of RBF. The optimal values for the parameters C and γ are determined by grid search using 5 fold cross validation. However, in our case the grid search turned out to be computationally very intensive. Thus, we used 5% of the training dataset to determine the optimal parameters instead. The DisPredict output classes such as disordered or ordered residue, in terms of probability, is optimized by another round of 5-fold cross validation. Using the threshold value 0.5, the probabilities are converted into binary decision variables, where probability ranges $0.5 \leq range_d \leq 1.0$ is considered as disordered and

$0.0 \leq range_o < 0.5$ is considered as ordered. Figure 2 shows the detail paradigm of DisPredict.

Figure 2. Overview of feature aggregation, optimization-layer and classification-layer in DisPredict. In the feature aggregation step, features are shown in their abbreviated form according to Table 1 and the arrows are labeled by the number of features involved. The classification-layer receives final feature set from the feature aggregation step and optimal parameters from the optimization-layer. Then, it generates the predictor model and outputs both binary annotation and real-valued class probabilities.

2.5 Performance Evaluation and Statistical Test Criteria

The performance of DisPredict is evaluated using the criteria followed in the past Critical Assessment of protein Structure Prediction (CASP) competitions [30–32]. The measures and procedures used in CASP experiments are comprehensive. The predictions are done in two levels:

1. Binary value, defining whether a residue is disorder or not ("+1" for disorder and "−1" for order) and
2. Real value, quantifying the probability of a residue being disorder ("≥ 0.5" for disorder and "< 0.5" for order).

Binary Prediction Evaluation In binary (two-class) prediction of disorder, *TP* (True Positive) = number of correctly predicted disordered residues, *TN* (True Negative) = number of correctly predicted ordered residues, *FP* (False Positive) = number of incorrectly predicted disordered residues and *FN* (False Negative) = number of incorrectly predicted ordered residues. To determine the total number of correct prediction (both ordered and disordered), $N_{correct} = TP + TN$ is calculated. Sensitivity *(SENS)* and specificity *(SPEC)* are two complementary statistical measures identifying the proportionate values of correct prediction of disordered (positive class) and ordered (negative class) residues, respectively.

$$SENS = \frac{TP}{TP + FN} = \frac{TP}{N_d} \text{ and } SPEC = \frac{TN}{TN + FP} = \frac{TN}{N_o}$$

Here, N_d and N_o are the total number of disordered and ordered residues, respectively. As increment of one of these measures (SENS and SPEC) usually leads towards the decrement of another measure, neither of these two measures is a suitable indicator of performance for an imbalanced dataset. On the contrary, the balanced accuracy (ACC), weighted score (S_w) and Mathews correlation coefficient (MCC) are the measures that take all four components of prediction quality (TP, TN, FP and FN) into account and thus can be regarded as more important indicators.

$$ACC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right)$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$S_w = \frac{w_d \times TP - w_o \times FP + w_o \times TN - w_d \times FN}{w_d \times N_d + w_o \times N_o}$$

where, $w_d = \frac{N_d}{N_o + N_d}$ is the percentage of disordered residues and $w_o = \frac{N_o}{N_o + N_d}$ is the percentage of ordered residues in the dataset. The S_w measure includes weight to address the imbalance in the ratio of ordered and disordered residues and rewards

correct disorder classification over correct classification of ordered residues, which is later found to have a linear relationship with ACC ($S_w = 2 \times ACC - 1$). Since both of these measures (ACC and S_w) have been used in CASP assessment, we have also included both of them in our paper instead of just one. MCC score, another measure that accounts for all four parameters of the prediction quality, is the most reasonable and consistent measure for disorder prediction assessment because of not being favorable to over prediction of any class (order/disorder). MCC and S_w scores vary from -1 to 1, where -1 and 1 represent perfect misclassification and classification, respectively with a random classification scoring by 0. More recently, precision ($PPV = \frac{TP}{TP+FP}$) has been appeared as a good measure for binary disorder prediction as it is totally insensitive to the prediction of the dominant class (i.e., here the order state), is therefore computed to evaluate DisPredict. As the prediction becomes better, the values of these metrics also get higher.

We calculated Mean Absolute Error (MAE) = $\frac{\sum_{i=1}^{n} |c_d^a(i) - c_d^p(i)|}{n}$ to quantify the error of disorder prediction in content level. Here, n is the total number of protein chains, and $c_d^a(i)$ and $c_d^p(i)$ are the actual and predicted disorder content (fraction of disordered residues) for the i^{th} protein chain, respectively. The lower value of MAE corresponds to better prediction.

Evaluation of Predicted Probability The SVM model of DisPredict generates a predicted probability value for each residue which signifies the disorder confidence of that residue. This probability value is then binarized using a threshold of 0.5 to generate class annotation. If the probability is greater than or equal to 0.5, the predicted class is 'disorder' and if the probability is less than 0.5, the predicted class is 'order'. Assessment of the predicted probability by a DisPredict is performed by receiver operating characteristic (ROC) curve, which depicts the correlation between the true positive rate (TPR or, $SENS$) and false positive rate ($FPR = 1$ - $SPEC$) for a probability threshold. The area under the ROC curve (AUC) quantifies the predictive quality of a classifier, where the AUC value equal to 1 indicates a perfect prediction and 0.5 corresponds to a random prediction. Moreover, 95% confidence interval (CI) for the AUC score is evaluated using DeLong's [33] variance estimate by bootstrapping with sample size of 2000. The evaluation of AUC and CI are performed using the statistical R package with the pROC [83] library.

3 Test Procedures and Results

3.1 Performance of 10-Fold Cross Validation

We evaluated the 10-fold cross validation performance of DisPredict separately on SL477 and MxD444 dataset. Regarding the optimum selection of the window size, we ran cross validation individually for 13 different windows, shown in Table 2, for both of the SL477 and MxD444 dataset with default parameters for SVM. The best result for window size 25 was found with ACC, MCC and AUC values equal to 0.82, 0.65 and 0.91, respectively for SL477 dataset, whereas for MxD444 dataset the values are 0.77, 0.48 and 0.85, respectively. The gradual increase in performance becomes a plateau as window goes higher above size 23 (Figure 3). Table 2 also depicts the inverse relationship between SENS and SPEC scores with increasing window size for MxD444 dataset. The best SENS (0.74) is achieved by window size 25 while the best SPEC (0.81) is achieved at window size 5. Overall, the consistent increment in balanced accuracy (ACC) and PPV prove our methodology to be well balanced.

Note that, this preliminary extensive analysis of performance with multiple window sizes is done without selection of optimal parameters for SVM. For a specific window size (W_{size}) and total number of residues ($Residue_{total}$) in a dataset, we have a feature

Table 2. 10-fold Cross Validation Performance of DisPredict (Default Parameter)

W_{size}	TP	TN	FP	FN	$N_{correct}(Residue_{total})$[1]	SENS	SPEC	ACC	S_w	PPV	MCC	AUC [95% CI][2]
SL477 Dataset												
1	4440	5804	1469	1260	10244 (12973)	0.779	0.798	0.788	0.577	0.751	0.574	0.869 [0.862 , 0.876]
3	4467	5954	1319	1233	10421 (12973)	0.784	0.819	0.801	0.602	0.772	0.601	0.884 [0.877 , 0.890]
5	4457	6020	1253	1243	10477 (12973)	0.782	0.828	0.805	0.609	0.781	0.609	0.889 [0.882 , 0.896]
7	4441	6076	1197	1259	10517 (12973)	0.779	0.835	0.807	0.614	0.787	0.615	0.893 [0.886 , 0.899]
9	4457	6086	1187	1243	10543 (12973)	0.782	0.837	0.809	0.618	0.789	0.619	0.895 [0.888 , 0.902]
11	4483	6113	1160	1217	10596 (12973)	0.786	0.841	0.813	0.627	0.794	0.628	0.898 [0.891 , 0.905]
13	4502	6114	1159	1198	10616 (12973)	0.790	0.841	0.815	0.630	0.795	0.631	0.899 [0.892 , 0.905]
15	4513	6150	1123	1187	10663 (12973)	0.792	0.845	0.819	0.637	0.801	0.638	0.902 [0.896 , 0.909]
17	4540	6133	1140	1160	10673 (12973)	0.796	0.843	0.820	0.640	0.799	0.640	0.902 [0.895 , 0.902]
19	4545	6148	1125	1155	10693 (12973)	0.797	0.845	0.821	0.643	0.802	0.643	0.903 [0.896 , 0.910]
21	4548	6148	1125	1152	10696 (12973)	0.798	0.845	0.822	0.643	0.802	0.643	0.903 [0.896 , 0.910]
23	4555	6167	1106	1145	10722 (12973)	0.800	0.847	0.823	0.647	0.804	0.647	0.904 [0.898 , 0.911]
25	4564	6164	1109	1136	**10728** (12973)	**0.801**	**0.847**	**0.824**	**0.648**	**0.804**	**0.648**	**0.905 [0.898 , 0.911]**
MxD444 Dataset												
1	3284	13093	3397	1632	16377 (21406)	0.668	0.793	0.731	0.462	0.491	0.419	0.817 [0.810 , 0.825]
3	3369	13241	3249	1547	16610 (21406)	0.685	0.803	0.744	0.488	0.509	0.444	0.832 [0.826 , 0.840]
5	3410	13302	3188	1506	16712 (21406)	0.694	**0.807**	0.750	0.500	0.516	0.456	0.839 [0.833 , 0.847]
7	3419	13275	3215	1497	16694 (21406)	0.695	0.804	0.750	0.501	0.515	0.455	0.840 [0.833 , 0.847]
9	3446	13253	3237	1470	16699 (21406)	0.700	0.805	0.752	0.505	0.516	0.458	0.842 [0.834 , 0.849]
11	3503	13232	3258	1413	16735 (21406)	0.712	0.802	0.757	0.515	0.517	0.466	0.846 [0.839 , 0.853]
13	3523	13188	3302	1393	16711 (21406)	0.717	0.800	0.758	0.516	0.516	0.466	0.847 [0.839 , 0.853]
15	3564	13145	3345	1352	16709 (21406)	0.725	0.797	0.761	0.522	0.515	0.469	0.848 [0.842 , 0.855]
17	3578	13097	3393	1338	16675 (21406)	0.728	0.794	0.761	0.522	0.513	0.469	0.848 [0.841 , 0.855]
19	3607	13068	3422	1309	16675 (21406)	0.734	0.792	0.763	0.526	0.513	0.471	0.849 [0.842 , 0.856]
21	3613	13078	3412	1303	16691 (21406)	0.735	0.793	0.764	0.528	0.514	0.473	0.850 [0.843 , 0.857]
23	3640	13059	3431	1276	16699 (21406)	0.740	0.792	0.766	0.532	0.515	0.476	0.851 [0.845 , 0.859]
25	3658	13064	3426	1258	**16722** (21406)	**0.744**	0.792	**0.768**	**0.536**	**0.517**	**0.479**	**0.852 [0.847 , 0.861]**

W_{size} indicates the size of sliding window.
Best values of each metric are marked in bold for each dataset separately.
[1] $N_{correct}$ is reported with total number of residues ($Residue_{total}$) to be predicted in parentheses. Both of the counts correspond to one subset (fold) of the full dataset which is generated for performing cross validation.
[2] For AUC, the values within bracket indicate 95% confidence interval with 2000 stratified bootstrap replicas.
As the window size continues to increase, the rate of increase in scores becomes slow. Increase of scores is ≤ 0.001, as the windows size grows from 23 to 25 for SL477 dataset and ≤ 0.004 for MxD444 dataset, respectively.

Figure 3. Increase of performance in 10-fold cross validation (default parameter) according to ACC, MCC and AUC scores with the increase of window size for (A) SL477 and (B) MxD444 dataset. The x-axis and y-axis represents the window sizes and scores, respectively.

matrix of dimension, $Residue_{total} \times (W_{size} \times 56)$. Therefore, the increase in window size leads towards the increase in the dimensions of the feature space, which in turn makes the time expensive grid search for parameters slower. To trade off between performance with optimization and time complexity of parameter selection along with model generation, we determined the optimal values of parameters with a 5% randomly selected subset of residues from training dataset for 3 window sizes (15, 21 and 25). The optimal parameters (C and γ) found from grid search are reported in Table 3. Furthermore, we inserted repeated disordered residue information only in case of training to balance the dataset as the support vector points for the slammer class may not be sufficient to determine the optimal SVM margin. Specifically, duplicates (2 times for SL477 dataset and 3 times for MxD444 dataset) of disorder samples were provided during generation of predictor model. However, in case of testing, no repeated information was inserted. Table 4 illustrates the detail of the cross validation results with optimized parameters for 3 different window sizes.

The improvement of performance with optimized parameters over non-optimized one was significant. To compare, for SL477 dataset (window size 21), FP and FN values are reduced to $1,002$ and $1,083$ from $1,125$ and $1,152$ due to optimization. In case of MxD dataset (window size 21), the FN value is increased by 133 residues. However, the FP value is also decreased by $1,812$ residues which maintains the overall increase in the total number of correctly predicted residues from $16,691$ to $18,370$. The improvement of prediction, both in terms of increased correct classification and decreased

Table 3. Optimized Parameters used to build DisPredict Models

	SL477 Dataset		MxD444 Dataset	
W_{size}	C	γ	C	γ
15	8.0	0.001953125	8.0	0.0312500
21	2.0	0.007812500	2.0	0.0078125
25	0.5	0.007812500	2.0	0.0078125

C is the soft penalty parameter to handle overlapped class.
γ is the parameter for radial basis kernel for SVM.

Table 4. 10-fold Cross Validation Performance of DisPredict (Optimized Parameter)

W_{size}	TP	TN	FP	FN	$N_{correct}(Residue_{total})$[1]	SENS	SPEC	ACC	S_w	PPV	MCC	AUC [95% CI][2]
SL477 Dataset												
15	4655	6056	1217	1045	10711 (12973)	**0.817**	0.833	0.825	0.649	0.793	0.647	0.898 [0.890 , 0.904]
21*	4617	6271	1002	1083	**10888** (12973)	0.810	**0.862**	**0.836**	**0.672**	**0.822**	**0.673**	**0.956 [0.950 , 0.963]**
25	4624	6234	1039	1076	10858 (12973)	0.810	0.857	0.834	0.668	0.816	0.669	0.911 [0.904 , 0.917]
MxD444 Dataset												
15	2590	15590	900	2326	18180 (21406)	0.527	**0.945**	0.736	0.472	**0.742**	0.538	0.838 [0.831 , 0.845]
21*	3480	14890	1600	1436	**18370** (21406)	**0.708**	0.903	**0.805**	**0.611**	0.685	**0.600**	**0.853 [0.847 , 0.859]**
25	3367	3367	1635	1549	18222 (21406)	0.685	0.901	0.793	0.586	0.673	0.582	0.850 [0.843 , 0.858]

W_{size} indicates the size of sliding window and * mark represents window size with overall optimal (best) performance.
Best values of each metric are marked in bold for each dataset separately.
[1] $N_{correct}$ is reported with total number of residues ($Residue_{total}$) to be predicted in parentheses. Both of the counts correspond to one subset (fold) of the full dataset which is generated for performing cross validation.
[2] For AUC, the values within bracket indicate 95% confidence interval with 2000 stratified bootstrap replicas.

misclassification, is also visible from both the sensitivity and specificity scores. For window size 21, the values of S_w, precision and MCC are improved by 4.5%, 2.5% and 4.5% respectively due to optimized training on SL477 dataset. At the same time, for MxD444 dataset, these progresses are 15.7%, 33.3% and 26.8% respectively. Note that, this significant improvement in MCC strongly supports our method's capability in handling the imbalance ratio of ordered and disordered residues. Further, the AUC score is also increased by 4.4% and 0.4% as the result of optimization for SL477 and MxD444 dataset, respectively. A comparative analysis on Table 2 and 4 also shows that optimized DisPredict model with window size 21 outperforms all the other models of its own kind. Thus we select 21 as the optimal window size for our proposed DisPredict.

To uniformly distribute the residues into ten subsets for cross validation, we applied modular arithmetic operation to split the dataset in residue level. As the residues are already included within the neighboring information based on the window, they are detachable from their original sequence. However, this inclusion of residue information within window may yield overlap of information between training and test sets in case of residue level splitting of dataset for cross validation. We analyzed the probability of this residual overlap between training and test sets. Let, there are N sequences in the dataset and the expected length of the sequence is $\mathcal{L}$. Then, the possibility of picking two residues for training and test subsets of 10 fold cross validation which belongs to same sequence is $(\frac{1}{\frac{9N}{10}} \times \frac{1}{\frac{N}{10}}) = \frac{100}{9N^2}$. Since the expected length of a sequence is $\mathcal{L}$, the chance of training and test overlap for a specific window size (W_{size}) is $\frac{W_{size}-1}{\mathcal{L}}$. Altogether, the probability of a train and test residue overlap from the same sequence is $(\frac{100}{9N^2} \times \frac{W_{size}-1}{\mathcal{L}}) = (\frac{100}{9})\frac{W_{size}-1}{N^2\mathcal{L}}$. For SL477 dataset with $N = 477$, approximate $\mathcal{L} = 400$ and $W_{size} = 21$, the probability of the overlap is 2.44×10^{-06}, which is significantly low and thus can be safely ignored. Further, we reevaluated DisPredict's 10 fold cross validation performance with sequence level sampling by modular operation for SL477 dataset to generate training and test subsets. Table 5 quantifies the difference in performance between residue level and state of the art practice of sequence level

splitting of dataset for cross validation with window size 21 and default parameters for SVM. It justifies that DisPredict's performance remains consistent without any significant over prediction in terms of all the metrics.

Table 5. DisPredict's cross validation performance with residue level and sequence level splitting of SL477 dataset.

Splitting Method	SENS	SPEC	ACC	S_w	PPV	MCC	AUC [95% CI]
Residue Level	0.798	0.845	0.822	0.643	0.802	0.643	0.903 [0.896 , 0.910]
Sequence Level	0.784	0.844	0.814	0.628	0.793	0.627	0.892 [0.886 , 0.898]

Default values of C and γ are applied for SVM.
Window size 21 is used.

3.2 Evaluation of Independent Training and Testing

With optimized parameters and balanced dataset, we carried out independent training on SL477 and MxD444 datasets followed by testing the resulting predictor model with MxD134 and SL171 dataset, respectively. Note that, these independent test datasets (MxD134 and SL171) were generated at low sequence identity (10%) with the corresponding training datasets (SL477 and MxD444). The consistent results of these two tests done through cross validation and independent test confirm the usage of robust technique and effective feature set in DisPredict as well as training efficacy avoiding possible over-fittings. Table 6 further illustrates the results of these tests, where we reported the average of the scores computed for equally divided 10 subsets of the full dataset along with the corresponding standard deviation (STDEV). Table 6 reveals that training by SL477 dataset gives consistent performance regardless of test datasets and test procedures (cross validation or independent test) in terms of ACC: 0.836, 0.833 and S_w: 0.672, 0.667. These consistencies are also evident in case of training with MxD444 dataset while tested by different datasets and the evaluations are, ACC: 0.805, 0.789 and S_w: 0.611, 0.577. We calculated the Mean Absolute Error (MAE) which is also reported along with its corresponding STDEV from mean. The score indicates that the error does not increase from cross validation to independent test as the test-results were robust.

To analyze the probability prediction, the ROC curves given by DisPredict are plotted in Figure 4 in continuous scale between 0.0 and 1.0. In each figure, two ROCs are plotted keeping the training dataset same with varying test datasets and evaluation procedure. Finally, we reported the AUC values which are found consistent for cross validation and independent test indicating our predictor's capability to avoid over-fitting.

Figure 4. ROC curves given by DisPredict for the per residue probability prediction while the training is performed with (A) SL477 and (B) MxD444 dataset. In each figure, the solid (*blue*) curve corresponds to the cross validation test on the same dataset and the dotted (*red*) curve corresponds to the independent test. The AUC values given in each figure correspond to the values in Table 6. The x-axis and y-axis shows the Specificity and Sensitivity, respectively.

3.3 Comparison with Existing Predictors

The performance of DisPredict is compared with the state-of-the-art disorder predictor, MFDp [13] and SPINE-D [14]. To avoid any inconsistencies while comparing DisPredict with each of the above two predictors separately, we used similar datasets. DisPredict is compared with MFDp based on dataset MxD444, while dataset SL477 used to compare DisPredict with SPINE-D (Table 7).

In particular, MFDp [13] is a meta predictor that combines the predictions from three disorder predictors (DISOPRED2 [52], DISOclust [69] and IUPred [67]). Further,

Table 6. Performance Comparison of Cross Validation and Independent Tests

Model Evaluation Procedure[1]	SENS (STDEV)	SPEC (STDEV)	ACC (STDEV)	S_w (STDEV)	PPV (STDEV)	MCC (STDEV)	AUC (STDEV)	MAE (STDEV)
10-fold cross validation on SL477	0.810 (0.004)	0.862 (0.001)	0.836 (0.002)	0.672 (0.005)	0.822 (0.002)	0.673 (0.004)	0.956 (0.007)	0.032 (0.002)
Train by SL477, Test on MxD134	0.744 (0.002)	0.923 (0.002)	0.833 (0.002)	0.667 (0.003)	0.574 (0.002)	0.598 (0.004)	0.906 (0.001)	0.023 (0.001)
10-fold cross validation on MxD444	0.708 (0.006)	0.903 (0.001)	0.805 (0.003)	0.611 (0.006)	0.685 (0.002)	0.600 (0.004)	0.853 (0.007)	0.208 (0.001)
Train by MxD444, Test on SL171	0.718 (0.003)	0.860 (0.001)	0.789 (0.001)	0.577 (0.003)	0.748 (0.001)	0.583 (0.002)	0.872 (0.007)	0.151 (0.001)

[1] All the evaluations are carried out using a sliding window of length 21 and optimized parameters for SVM.

MFDp combines the output from three SVMs with linear kernel using a threshold of 0.37, used to output binary prediction. In contrast, we utilized single SVM with RBF kernel and optimized parameters combined with a comprehensive set of features to develop the standalone predictor. However, the performance of MFDp in Table 7 is of 5-fold cross validation whereas DisPredict is evaluated by 10-fold cross validation and hence to be considered reliable rather than over-fitted by chance. In terms of MCC, DisPredict improved significantly, which is 26.67% better than MFDp. The improvement in S_w score is also 16.4%. DisPredict showed lower sensitivity (7%) than MFDp while at the same time improved specificity by 16.7%, which in turn improved the balanced accuracy by 6.25%. Moreover, DisPredict outperformed MFDp in AUC score by 4.7% which is used to assess the probability based prediction.

Table 7. Comparative predictive quality of DisPredict with MFDp on MxD444 dataset and SPINE-D on SL477 dataset.

Method	SENS	SPEC	ACC	S_w	MCC	AUC
DisPredict[2]	0.71	**0.90**	**0.80**	**0.61**	**0.60**	**0.85**
MFDp[1]	**0.76**	0.75	0.75	0.51	0.44	0.81
DisPredict[3]	**0.81**	**0.86**	**0.84**	**0.67**	**0.67**	**0.96**
SPINE-D[4]	0.77	0.85	0.81	0.62	0.63	0.87

[1] 5-fold cross validation performance of MFDp on MxD dataset of 514 protein chains [13].
[2] 10-fold cross validation performance of DisPredict on MxD444 which is a subset of 444 chains out of 514 chains with no X-tag.
[3] 10-fold cross validation performance of DisPredict on SL477.
[4] 10-fold cross validation performance of SPINE-D [14] on SL477.

The other state of the art predictor, SPINE-D [14] utilizes ANN technique which was at first developed to output three state prediction and later reduced into two state predictor of ordered and disordered residues. SPINE-D employs a disorder probability threshold of 0.06 that was optimized to achieve maximum S_w score. On the contrary, DisPredict is a SVM based two state disorder predictor using a more meaningful threshold for two-class classification of value 0.5. DisPredict outperformed SPINE-D in terms of sensitivity as well as specificity by 5% and 1.4% respectively which leads to 3.1% improvement in overall accuracy. DisPredict also outperformed SPINE-D in terms of S_w, MCC and AUC by 7.7%, 6.4% and 9% respectively.

In addition to the comparison on cross validation test, we evaluated DisPredict, SPINE-D [14] and MFDp [13] on independent DD73 dataset. The comparison among these three methods is illustrated in Table 8. It shows that DisPredict gives superior performance among three predictors except in case of sensitivity. SPINE-D [14] yielded 2.63% higher sensitivity than that of DisPredict, whereas DisPredict gave 4.25% higher specificity than that of SPINE-D [14]. Table 8 also shows that DisPredict outperformed SPINE-D [14] and MFDp [13] in terms of MCC by 0.73% and 3.69%, respectively. At the same time, DisPredict gave 1.2% and 5.3% improved precision (PPV) than

MFDp [13] and SPINE-D [14], respectively. However, DisPredict resulted slightly lower sensitivity than those of SPINE-D [14] and MFDp [13]. At the same time, both SPINE-D [14] and MFDp [13] gave lower specificity than that of DisPredict. Figure 5 and Figure 6 compare the ROC curves and precision-recall curves, respectively, given by DisPredict, SPINE-D [14] and MFDp [13]. Figure 5 shows that the ROC curves given by the three predictors are comparative. At the same time, the precision-recall curves (Figure 6) depicts that DisPredict achieves consistently higher precision upto less than 65% sensitivity (recall).

Table 8. Performane comparison among DisPredict, SPINE-D and MFDp on independent DD73 dataset.

Predictor	SENS	SPEC	ACC	S_w	PPV	MCC	AUC [95% CI]
DisPredict*	0.775	**0.883**	**0.829**	**0.658**	**0.806**	**0.663**	**0.89 [0.88 , 0.90]**
SPINE-D	**0.796**	0.847	0.822	0.644	0.765	0.639	**0.89 [0.88 , 0.90]**
MFDp	0.780	0.875	0.828	0.656	0.796	0.658	0.88 [0.87 , 0.89]

Best results are marked in bold.
* Window size = 21, $C = 2.0$ and $\gamma = 0.0078125$.

Figure 5. ROC curves for disorder prediction on DD73 dataset given by DisPredict(*blue*), SPINE-D(*green*) and MFDp(*red*). The AUC values shown in the figure correspond to the values in Table 8. The x-axis and y-axis shows the Specificity and Sensitivity, respectively.

Figure 6. Precision-Recall curves for disorder prediction on DD73 dataset given by DisPredict(*blue*), SPINE-D(*green*) and MFDp(*red*). The x-axis and y-axis shows the Recall(Sensitivity) and Precision (PPV), respectively.

MFDp and SPINE-D has been established as the best disorder predictor among 8 and 11 existing disorder predictors [13, 14] covering different approaches in their relevant publication and in this article, our predictor is shown to outrank both these top methods. Therefore, DisPredict can be considered to be one of the finest disorder predictor and can be utilized to produce more reliable annotation of disorder versus order residues.

3.4 Case Studies: Characteristic Region and Protein Function

Proteins with disordered regions are found to contain several regions of interest, such as self-stabilizing folded regions, DNA or, nucleotide binding regions, short (up to 20 amino acids) conserved regions of biological significance (known as motif), mediating regions for protein interaction with different partners etc. These characteristic regions undergo various conformational changes, gain structure and affect many important biological functions. We selected three proteins as cases (UniProt IDs: P41212, P01116 and P04637) with experimentally verified regions of interest to analyze per residue disorder confidence score assigned by DisPredict, SPINE-D and MFDp. Figure 7 illustrates the disorder probability of each residue with respect to residue index. P41212

Figure 7. Disorder probability plot for (A) human ETV6 (P41212), (B) human KRAS (P01116) and (C) human p53 (P04637) proteins, given by DisPredict(*red*), SPINE-D (*blue*) and MFDp (*green*). In (P41212, A), the yellow ($40-124$ residues) and pink bar ($339-420$ residues) represents to the PNT domain [74] and ETS DNA binding region [75], respectively. In (P01116, B), the orange ($10-17$ residues), cyan ($32-40$ residues) and purple bar ($166-185$ residues) corresponds to the GTP binding region [76], effector region and hypervariable region, respectively. In (P04637, C), the dark green ($17-25$ residues), red ($325-356$ residues) and gray bar ($370-372$ residues) highlights to the TADI motif [77], oligomer region and [KR]-[STA]-K binding motif, respectively.

(Figure 7(A)) is a human ETV6 protein for transcriptional repressor function, which is also involved in several kinds of leukemia and syndrome. For this protein, DisPredict

and SPINE-D showed comparable performance in detecting the highly conserved region of PNT (pointed) domain [74] [residues $40-124$] and ETS (E26 transformation-specific) DNA binding region [75] [residues $339-420$], respectively, while MFDp outperformed both of them with relatively less noise. P01116 (Figure 7(B)) is a human KRAS protein with intrinsic GTPase activity (binds GDP/GTP [76]) and related to several diseases, such as gastric cancer (GASC), acute myelogenous leukemia (AML), cardiofaciocutaneous syndrome 2 (CFC2) etc. DisPredict could identify its GTP (guanosine triphosphate) binding region [residues $10-17$] and effector region [residues $32-40$] respectively, with close to cut-off (0.5) probabilities. Note that, these two regions are experimentally verified unstructured regions, which are strongly suggested as structured by both SPINE-D and MFDp. However, the C-terminal hypervariable region [residues $166-185$] is consistently detected by all three of these predictors. P04637 corresponds to human p53 protein which acts as a tumor suppressor. Figure 7(C) illustrates that DisPredict and MFDp outperformed SPINE-D with relatively sharp detection of N-terminal TADI (transcriptional repression domain-I) motif [77] [residues $17-25$]. On the other hand, DisPredict and SPINE-D outperformed MFDp in determining oligomerization domain [78] of residues $325-356$. Figure 7(C) also shows that both SPINE-D and MFDp missed the very short, 3 residue ($370-372$) long [KR]-[STA]-K binding motif at C-terminal, while DisPredict detected it correctly. The overall comparison depicts that DisPredict's performance is more biologically relevant with correct identification of these short regions. Therefore, it would be interesting to utilize DisPredict in a broader scope in near future.

4 Discussion

In this article, we proposed a canonical support vector machine which uses a RBF kernel and includes useful and advanced features for predicting disordered residues, called DisPredict. DisPredict not only generates the binary class annotation for ordered and disordered residues but also provides order-disorder probabilities that can be treated as the confidence level of the prediction too. The DisPredict outperforms other existing top performing predictors both in predicting binary annotation and probability. The superior performance of DisPredict is mainly due to the use of a novel methodology that incorporates firstly, radial basis kernel function (RBF) that can implicitly map the feature space in infinite dimension, secondly and most importantly the optimization of the parameters and thirdly, the novel features monogram (MG) and bigram (BG) assisted in determining an optimal as well as effective class separating hyperplane.

This overall performance of DisPredict is also persuaded by the use of a comprehensive set of features that well captures the sequential (amino acid composition) and structural characterization of ordered and disordered residues or, proteins. We used SPINE X [24] to generate the secondary structure related fine features. The distinguishing property of our feature set in comparison with existing predictors is the inclusion of monogram (MG) and bigram (BG), computed from PSSM. When a region of a protein is evolutionary conserved in a fold, then all the proteins within that fold are likely to have a conserved group of MGs and BGs. As some intrinsic disordered regions are conserved, addition of these features provides important structural evolutionary characteristics. Involvement of MGs and BGs along with PSSM leads to a further increase in accuracy of binary prediction in terms of MCC from 0.651 to 0.673 (i.e., improved 3.8%) and S_w score from 0.621 to 0.672 (i.e., improved 8.21%). By determining the appropriate window size, we have also included the effect of optimal interactions due to the contacts among neighboring residues.

The robust performance of DisPredict is also justified by training and testing the predictor with multiple datasets: SL477, SL171 and MxD444, MxD134. The datasets

used to train DisPredict encompasses disorder annotation from several complementary sources (X-ray and NMR defined disorder from PDB and DisProt) as well as disorder region of various lengths. The SL dataset comprises of 81 full disordered proteins (IDPs) while rest of the chains contain 928 disordered regions (IDRs). On the other hand, the MxD dataset is composed of 55 full disordered chains, 4 full ordered chains and 385 chains, sharing both structured and disordered regions, which include 730 disordered regions (IDRs). Furthermore, 70% of the IDRs included within partially disordered proteins are short ($\leq$ 30 residues) and 30% of them are long ($>$ 30 residues). This combination of several length disordered regions (Figure 8) included within training confirms the consistent performance of DisPredict for disordered regions of all sizes as well as different types of disordered residues.

Figure 8. Distribution of disordered regions of different lengths in MxD444 (*left*) and SL477 (*right*) dataset. Legends are shown for different range of lengths (with interval size 15) and each bar is labeled with total number of occurrence of a disordered region of this specific length.

It is interesting to note that, regardless of cross validation or independent test, DisPredict's performance is relatively better while it is trained on SL477 dataset than that of MxD444 (Table 6). To further insight into this discrepancy, we investigated the correlation of true annotation provided in the dataset with the actual structural characterization of disordered and ordered residues. Disordered residues are distinguished from ordered residues by low content of secondary structure [19, 27], therefore high probability of coil residues than helical or beta strand residues and disordered regions are likely to have large solvent accessible (exposed) area [22]. We represented the correlation of the fraction of secondary structure content and fraction of exposed residues for disordered and ordered region of all length in Figure 9. We employed the predicted probability of each residue to be coil and predicted per residue solvent accessibility provided by SPINE-X [24] since all residues do not have defined coordinates (structure) to compute secondary structure and solvent accessibility.

Figure 9. Correlation plot between structural characterizations of ordered (*blue*) and disordered (*black*) regions within (A) SL477 and (B) MxD444 dataset. The x-axis and y-axis corresponds to the probability of having well defined secondary structure (in terms of probability being coil) and fraction of exposed residues of that region, respectively.

We calculated the average coil probability (P_{coil}) for each ordered or disordered region and computed the fraction of exposed residues with greater than 25% solvent accessibility ($F_{exposed}$) of that region. In this analysis, we discarded 5 residues from N and C-terminal regions of each protein sequence as they are mostly found on the surface of a protein chain (not buried in the core) and more likely to be affected by the interaction with nearby structured protein, yielding to a highly flexible and dynamic conformation. The plots for both datasets show that the ordered regions are mostly concentrated in the portion with relatively low coil probability, $0.3 \leq P_{coil} < 0.5$ (high content of well defined helical or strand secondary structured residues) and low exposure, $0.2 \leq F_{exposed} < 0.5$. While on the contrary, the disorder regions are found abundant in the area of high coil probability, $0.5 \leq P_{coil} \leq 0.9$ (low content of helical or strand secondary structured residues) and high exposure, $0.5 \leq F_{exposed} \leq 1.0$. However, we found the intrinsic difference between these two datasets according to their annotation of residues as order and disorder, which is also evident from the top right location of the correlation plot, $0.6 \leq P_{coil} \leq 0.8$ and $0.4 \leq P_{coil} \leq 0.9$, designated for disordered regions. For SL477 dataset (Figure 9(A)), the number disordered regions are predominant over the number of ordered regions in this location. In contrast, the same location of the plot is overlapped by both ordered and disordered regions in case of MxD444. We further quantified the difference as 13% of the data in MxD444's ordered set are more likely to be coil as well as highly exposed while 6% of the data in SL477's

ordered set are exposed as well as coil. This higher proportion of misleading annotation in MxD444 dataset contributes relatively lower signal to noise ratio (SNR) of 87/13 compared to 94/6 for SL477 which is the most compelling reason of the better performance of DisPredict in case of training dataset SL477 over MxD444. As the prediction produced by DisPredict is well capable of detecting such discrepancies in the native annotation of the datasets, it can be utilized as a reliable source of correct annotation of the ordered and disordered residues. We should also focus that, a similar proportion of 11% and 13% of the disordered data are also mixed with the ordered residues in the low coil probability region of the plot for both MxD444 and SL477 dataset, respectively.

We would like to highlight that the amino acid residue compositions may vary in different datasets as well as within short (≤ 30 residues) and long (> 30 residues) disordered regions [19, 50]. Specifically, short disordered regions are enriched with aspartic acid (D), glycine (G) and serine (S). On the contrary, glutamic acid (E), lysine (K) and proline (P) are likely to be abundant in long disordered regions. To give further insight into this residue composition and confirm the ability of DisPredict to detect the residue preferences of short and long disordered regions, we determined the residual composition profile for our two test datasets, SL171 (Figure 10(A)) and MxD134 (Figure 10(B)). It is to be noted that, these two datasets contain experimentally annotated disorder from two different sources. SL171 contains sequences with disorder annotation from DisProt while MxD134 contains that from PDB. The composition profile consists of the actual ratio (r_a) and predicted ratio (r_p) of each amino acid type out of total annotated and predicted disordered residues.

Figure 10. Percentage of amino acid type residues in actual composition (*blue, or left adjacent bar*) and predicted composition (*red, or right adjacent bar*) of (A) SL171 and (B) MxD134 dataset. The x-axis and y-axis represents the 20 different amino acids and their relation proportions in the composition.

The composition profile demonstrates that SL171's disordered residue set accommodates relatively higher ratio of amino acid type E (10%) and K (9%), which are long disorder prone residues. In contrast, MxD134's disordered residue set is enriched with high ratio of amino acid type S (11%), G (10%) and D (9%), known as short disorder prone residues. Another significant difference between the intrinsic compositions of these two datasets is in the proportion of histidine (H). Disorder annotation from PDB includes higher ratio of H-tag (8% in MxD134, compared to 2% in SL171), which is sometimes used for protein purification. The predicted proportion of all these amino acids given by DisPredict ensures its capability of detecting residues in disordered region of all length accurately with no significant over prediction. Moreover, DisPredict could also accurately predict methionine (M) at highly flexible N-terminal region. To further quantify DisPredict's performance in detecting residue composition, we evaluated the Root Mean Square Difference (RMSE) and Pearson Correlation Coefficient (PCC) between actual and predicted ratio (r_a and r_p) for each amino acid type. For MxD134 test dataset, we found RMSE of 0.0046, which was comparatively higher than the RMSE value computed for SL171 which equals to 0.0018. However, the correspondence between actual composition and predicted composition by DisPredict measured with PCC (P-Value $< 10^{-5}$) was found equally positive, 0.9976 and 0.9897 for SL171 and MxD134 dataset, respectively. It is important to note that, this consistent result is corresponding to the independent test where the dataset used to train DisPredict shared significantly low sequence identity (at most 10%) with test dataset, which once again implicates the strength of the classification methodology of DisPredict.

Finally, accurate prediction of disorder has significant implication in proteomic studies due to its direct involvement in the proper function of a protein. Successful detection of disordered region of a protein is considered to be the first step in drug

design to combat critical diseases. We have built DisPredict using the canonical SVM classifier with RBF kernel and established it as a successful fine predictor of disorder by utilizing the benchmark datasets. In addition to that, our case studies ensure biologically relevant performances of DisPredict.

Acknowledgments

We gratefully acknowledge the Louisiana Board of Regents through the Board of Regents Support Fund, LEQSF (2013-16)-RD-A-19. We also acknowledge the discussion with Md Nasrul Islam, Avdesh Mishra and Denson Smith. Special thanks to Denson Smith for critically reviewing the paper.

References

1. Wright,P.E. and Dyson, H.J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, **293(2)**, 321 – 331.

2. Uversky,V.N. and Dunker,A.K. (2010) Understanding protein non-folding. *Biochimica Et Biophysica Acta (BBA) - Proteins And Proteomics*, **1804(6)**, 1231 – 1264.

3. Uversky,V.N., Gillespie.,J.R., and Fink,A.L. (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins*, **41(3)**, 415 – 427.

4. Dunker,A.K. and Obradovic,Z. (2001) The protein trinity–linking function and disorder. *Nat Biotechnol*, **19(6)**, 805 – 806.

5. Uversky,V.N. (2002) Natively unfolded proteins: A point where biology waits for physics. *Protein Science*, **11(4)**, 739 – 756.

6. Tompa,P. (2002) Intrinsically unstructured proteins. *TRENDS in Biochemical Sciences*, **10(1)**, 527 – 533.

7. Vucetic,S., Brown,C.J., Dunker,A.K. and Obradovic,Z. (2003) Flavors of protein disorder. *Proteins: Structure, Function, and Bioinformatics*, **52(4)**, 573 – 584.

8. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (1999) The Protein Data Bank. *Nucleic Acids Res*, **28**, 235 – 242.

9. Sickmeier,M., Hamilton,J.A., LeGall,T., Vacic,V., Cortese,M.S., Tantos,A., Szabo,B., Tompa,P., Chen,J., Uversky.V.N., Obradovic,Z. and Dunker,A.K. (2007) DisProt: the Database of Disordered Proteins. *Nucleic Acids Res*, **35**, 786 – 793.

10. Sirota,F.L., Ooi,H.S., Gattermayer,T., Schneider,G., Eisenhaber,F., and Maurer-Stroh,S. (2010) Parameterization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset. *BMC Genomics*, **11(Suppl 1)**, S15.

11. Schlessinger,A., Punta,M. and Rost,B. (2007) Natively unstructured regions in proteins identified from contact predictions. *Bioinformatics*, **23(18)**, 2376 – 2384.

12. Schlessingera,A., Liu,J. and Rost,B. (2007) Natively Unstructured Loops Differ from Other Loops. *Bioinformatics*, **3(7)**, e140 – e151.

13. Mizianty,M.J, Stach,W., Chen,K., Kedarisetti,K.D., Disfani,F.M. and Kurgan,L. (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics*, **26(18)**, i489 – i496.

14. Zhang,T., Faraggi,E., Xue,B., Dunker,A.K., Uversky,V.N. and Zhou,Y. (2012) SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. *J Biomol Struct Dyn*, **29(4)**, 799 – 813.

15. Wang,G. and Dunbrack,R.J. (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19(12)**, 589 – 591.

16. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J Mol Biol.*, **215**, 403 – 410.

17. Jones,D.T. and Ward,J.J. (2003) Prediction of disordered regions in proteins from position specific score matrices. *Proteins*, **53(Suppl 6)**, 573 – 578.

18. Anfinsen,C.B. (1973) Principles that govern the folding of protein chains. *Science*, **181(96)**, 223 – 230.

19. Radivojac,P., Obradovic,Z., Smith,D.K., Zhu,G., Vucetic,S., Brown,C.J., Lawson,J.D. and Dunker,A.K. (2004) Protein flexibility and intrinsic disorder. *Protein Sci*, **10**, 71 – 80.

20. Peng,K., Radivojac,P., Vucetic,S., Dunker,A.K. and Obradovic,Z. (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, **7**, 208.

21. Linding,R., Jensen,L.J., Diella,F., Bork,P., Gibson,T.J. and Russell,R.B. (2003) Protein disorder prediction: implications for structural proteomics. *Structure*, **11(11)**, 1453 – 1459.

22. Schlessinger,A., Punta,M., Yachdav,G., Kajan,L. and Rost,B. (2009) Improved Disorder Prediction by Combination of Orthogonal Approaches. *PLoS One*, **4**, e4433 – e4442.

23. Su,C.T., Chen,C.Y. and Ou,Y.Y. (2006) Protein disorder prediction by condensed PSSM considering propensity for order or disorder. *BMC Bioinformatics*, **7**, 319 – 334.

24. Faraggi,E., Zhang,T., Yang,Y., Kurgan,L. and Zhou,Y. (2012) SPINE X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *J Comput Chem.*, **33(3)**, 259 – 267.

25. Zhang, T., Faraggi E. and Zhou Y. (2010) Fluctuations of backbone torsion angles obtained from NMR-determined structures and their prediction. *Proteins.*, **78(16)**, 3353 – 3362.

26. Faraggi.E., Xue,B. and Zhou,Y. (2009) Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. *Proteins.*, **74(4)**, 847 – 856.

27. Radivojac,P., Iakoucheva,L.M., Oldfield,C.J., Obradovic,Z., Uversky,V.N. and Dunker,A.K. (2007) Intrinsic Disorder and Functional Proteomics. *Biophysical Journal*, **92(5)**, 1493 – 1456.

28. Sharma,A., Lyons,J., Dehzangi,A. and Paliwal,K.K. (2013) A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. *J Theor Biol.*, **320**, 41 – 46.

29. Chang,C.-C. and Lin,C.-J. (2011) LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, **2(3)**, 27:1–27:27.

30. Noivirt-Brik,O., Prilusky,J. and Sussman,J.L. (2009) LAssessment of disorder predictions in CASP8. *Proteins*, **77(Suppl 9)**, 210 – 216.

31. Monastyrskyy,B., Fidelis,K., Moult,J., Tramontano,A. and Kryshtafovych,A. (2011) Evaluation of disorder predictions in CASP9. *Proteins*, **79(Suppl 10)**, 107 – 118.

32. Monastyrskyy,B., Kryshtafovych,A., Moult,J., Tramontano,A. and Fidelis,K. (2014) Assessment of protein disorder region predictions in CASP10. *Proteins*, **82(Suppl 2)**, 127 – 137.

33. DeLong,E.R., DeLong,D.M. and Clarke-Pearson,D.L. (1988) Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, **44(3)**, 837 – 845.

34. Petsko,G.A. and Ringe,D. (2004) *Protein Structure and Function.*

35. Whitford,P.C. (2013) Disorder guides protein function. *Proc Natl Acad Sci USA*, **110(18)**, 7114 – 7115.

36. Dyson,H.J. and Wright,P.E. (2002) Coupling of folding and binding for unstructured proteins. *Current opinion in structural biology*, **12(1)**, 54 – 60.

37. Uversky,V.N., Oldfield,C.J. and Dunker,A.K. (2005) Showing your ID : intrinsic disorder as an ID for recognition, regulation, and cell signaling. *J. Mol. Recogn.*, **18(5)**, 343 – 384.

38. Dunker,A.K., Brown,C.J. and Obradovic,Z. (2002) Identification and functions of usefully disordered proteins. *Adv. Protein Chem*, **62**, 25 – 49.

39. Dunker,A.K., Brown,C.J., Lawson,J.D., Iakoucheva,L.M. and Obradovic,Z. (2002) Intrinsic disorder and protein function. *Biochemistry*, **41**, 6573 – 6582.

40. Xue,B., Dunker,A.K. and Uversky,V.N. (2012) The Roles of Intrinsic Disorder in Orchestrating the Wnt-Pathway. *Journal of Biomolecular Structure and Dynamics*, **29(5)**, 843 – 861.

41. Kulkarni,P., Rajagopalan,K., Yeater,D. and Getzenberg,R.H. (2011) Protein folding and the order/disorder paradox. *J Cell Biochem*, **112(7)**, 1949 – 1952.

42. Uversky,V.N., Oldfield,C.J., Midic,U., Xie,H., Xue,B., Vucetic,S., Iakoucheva,L.M., Obradovic,Z. and Dunker,A.K. (2009) Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. *BMC Genomics*, **10**, S1 – S7.

43. Babu,M.M., Lee,R., Groot,N.S. and Gsponer,J. (2011) Intrinsically disordered proteins: regulation and disease. *Current Opinion in Structural Biology*, **21(3)**, 432 – 440.

44. Cheng,Y., LeGall,T., Oldfield,C.J., Mueller,J.P., Van,Y.-Y.J., Romero,P., Cortese,M.S., Uversky,V.N. and Dunker,A.K. (2006) Rational drug design via intrinsically disordered protein. *Trends Biotechnol.*, **24(10)**, 435 – 442.

45. Obradovic,Z., Peng,K., Vucetic,S., Radivojac,P., Brown,C.J. and Dunker,A.K. (2003) Predicting intrinsic disorder from amino acid sequence. *Proteins*, **53**, 566 – 572.

46. Xue,B., Dunbrack,R.L., Williams,R.W., Dunker,A.K. and Uversky,V.N. (2010) PONDR-FIT: A Meta-Predictor of Intrinsically Disordered Amino Acids. *Biochim Biophys Acta*, **1804(4)**, 996 – 101.

47. Fukuchi,S., Amemiya,T., Sakamoto,S., Nobe,Y., Hosoda,K., Kado,Y., Murakami,S.D., Koike,R., Hiroaki,H. and Ota,M. (2014) IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. *Nucleic Acids Res.*, **42**, D320 – D325.

48. Fukuchi,S., Sakamoto,S., Nobe,Y., Murakami,S.D., Amemiya,T., Hosoda,K., Koike,R., Hiroaki,H. and Ota,M., (2012) IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature. *Nucleic Acids Res.*, **40**, D507 – D511.

49. Pruitt,K.D., Tatusova,T., Klimke,W. and Maglott,D.R. (2009) NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, **37**, D32 – D35.

50. Peng,K., Vucetic,S., Radivojac,P., Brown,C.J., Dunker,A.K. and Obradovic,Z. (2005) Optimizing long intrinsic disorder predictors with protein evolutionary information. *J Bioinform Comput Biol.*, **3(1)**, 35 – 60.

51. Peng,K., Radivojac,P., Vucetic,S., Dunker,A.K. and Obradovic,Z. (2006) Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, **7**, 208.

52. Ward,J.J., McGuffin,L.J., Bryson,K., Buxton,B.F. and Jones,D.T. (2004) The DISOPRED server for the prediction of protein disorder. *Bioinformatics*, **20(13)**, 2138 – 2139.

53. Cheng,J., Sweredoski,M.J. and Baldi,P. (2005) Accurate Prediction of Protein Disordered Regions by Mining Protein Structure Data. *Data Mining and Knowledge Discovery*, **11(3)**, 213 – 222.

54. Yang,Z.R., Thomson,R., McNeil,P. and Esnouf,R.M. (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*, **21(16)**, 3369 – 3376.

55. Vullo,A., Bortolami,O., Pollastri,G. and Tosatto,S.C. (2006) Spritz: a server for the prediction of intrinsically disordered regions in protein sequences using kernel machines. *Nucleic Acids Res.*, **34**, W164 – W168.

56. Schlessinger,A., Yachdav,G. and Rost,B. (2006) PROFbval: predict flexible and rigid residues in proteins. *Bioinformatics*, **22(7)**, 891 – 893.

57. Su,C.T., Chen,C.Y. and Ou,Y.Y. (2006) Protein disorder prediction by condensed PSSM considering propensity for order or disorder. *BMC Bioinformatics*, **7**, 319 – 334.

58. Su,C.T., Chen,C.Y. and Hsu,C.M. (2007) iPDA: integrated protein disorder analyzer. *Nucleic Acids Res*, **35**, W465 – W472.

59. Ishida,T. and Kinoshita,K. (2007) PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res*, **35**, W460 – W464.

60. Shimizu,K., Muraoka,Y., Hirose,S., Tomii,K. and Noguchi,T. (2007) Predicting mostly disordered proteins by using structure-unknown protein data. *BMC Bioinformatics*, **8**, 78 – 92.

61. Hirose,S., Shimizu,K., Kanai,S., Kuroda,Y. and Noguchi,T. (2007) POODLE-L: a two-level SVM prediction system for reliably predicting long disordered regions. *Bioinformatics*, **23(16)**, 2046 – 2053.

62. Yang,J.Y. and Yang,M.Q. (2008) Predicting protein disorder by analyzing amino acid sequence. *BMC Genomics*, **9(suppl 2)**, S8 – S15.

63. Wang,L. and Sauer,U.H. (2008) OnD-CRF: predicting order and disorder in proteins using [corrected] conditional random fields. *Bioinformatics*, **24(11)**, 1401 – 1402.

64. Deng,X., Eickholt,J. and Cheng,J. (2009) PreDisorder: ab initio sequence-based prediction of protein disordered regions. *BMC Bioinformatics*, **10**, 436 – 441.

65. Walsh,I., Martin,A.J.M., Domenico,T.D. and Tosatto,S.C.E. (2012) ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*, **28(4)**, 503 – 509.

66. Linding,R., Russell,R.B., Neduva,V. and Gibson,T.J. (2003) GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res*, **31(13)**, 3701 – 3708.

67. Dosztányi,Z., Csizmok,V., Tompa,P. and Simon,I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21(16)**, 3433 – 3434.

68. Prilusky,J., Felder,C.E., Zeev-Ben-Mordehai,T., Rydberg,E.H., Man,O., Beckmann,J.S., Silman,I. and Sussman,J.L. (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*, **21(16)**, 3435 – 3438.

69. McGuffin,L.J. (2008) Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics*, **24(16)**, 1798 – 1804.

70. Ishida,T. and Kinoshita,K. (2008) Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics*, **24(11)**, 1344 – 1348.

71. Mizianty,M.J., Peng,Z. and Kurgan,L. (2013) MFDp2: Accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles. *Intrinsically Disordered Proteins*, **1**, e24428.

72. Sun,Y. and Ming,D. (2014) Energetic Frustrations in Protein Folding at Residue Resolution: A Homologous Simulation Study of Im9 Proteins. *PLoS ONE*, **9(5)**, e97982.

73. Vendruscolo,M., Paci,E., Dobson,C.M. and Karplus,M. (2000) Three key residues form a critical contact network in a protein folding transition state. *Letters to Nature*, **409**, 641 – 645.

74. Slupsky,C.M., Lisa,N.G., Donaldson,L.W., Mackereth,C.D., Seidel,J.J., Graves,B.J. and McIntosh,L.P. (1998) Structure of the Ets-1 pointed domain and mitogen-activated protein kinase phosphorylation site. *Proc. Natl. Acad. Sci. USA*, **95(25)**, 12129 – 12134.

75. Baens,M., Peeters,P., Guo,C., Aerssens,J. and Marynen,P. (1996) Genomic organization of TEL: the human ETS-variant gene 6. *Genome Res.*, **6(5)**, 404 – 413.

76. Colicelli,J. (2004) Human RAS Superfamily Proteins and Related GTPases. *Sci. STKE*, **2004(250)**, re13.

77. Piskacek,S., Gregor,M., Nemethova,M., Grabner,M., Kovarik,P. and Piskacek,M. (2007) Nine-amino-acid transactivation domain: establishment and prediction utilities. *Genomics*, **89**, 756 – 768.

78. McCoy.M., Stavridi,E.S., Waterman,J.L., Wieczorek,A.M., Opella,S.J. and Halazonetis,T.D. (1997) Hydrophobic side-chain size is a determinant of the three-dimensional structure of the p53 oligomerization domain. *EMBO J*, **16**, 6230 – 6236.

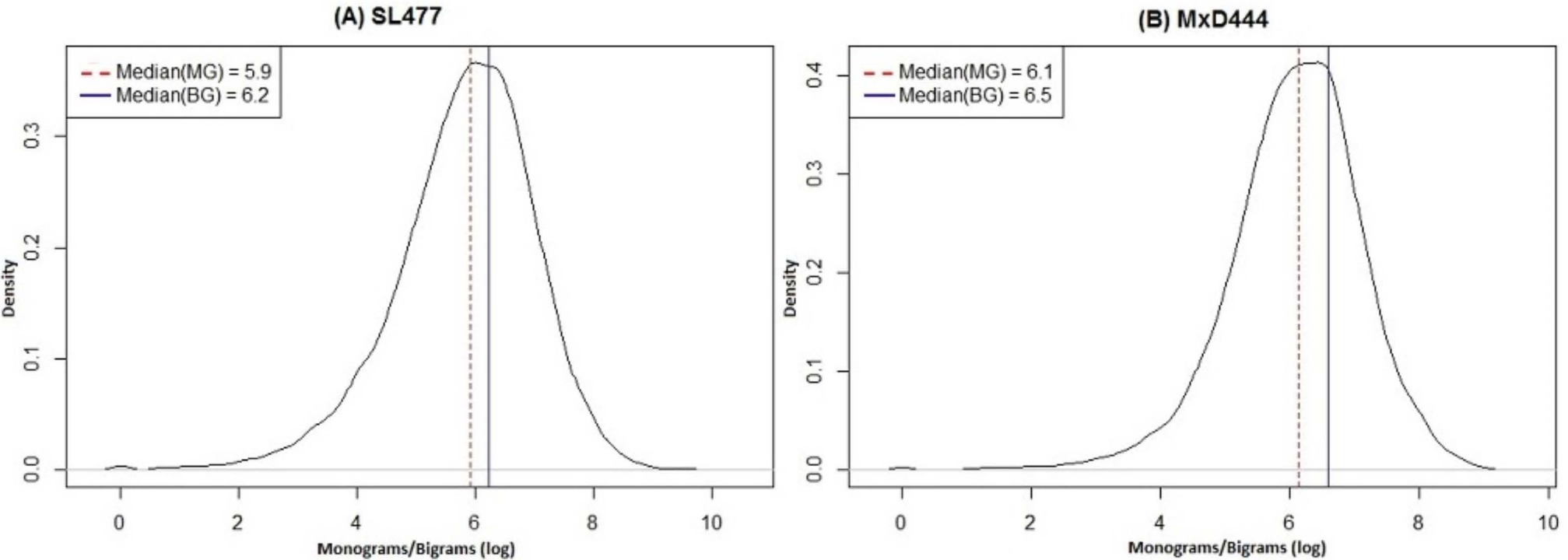
79. Iqbal,S. and Hoque,M.T. (2014) DisPredict: A Fine Disorder-Protein Predictor. *Tech. Report TR-2014/1*, `http://cs.uno.edu/~tamjid/TechReport/DisPredict.pdf`.

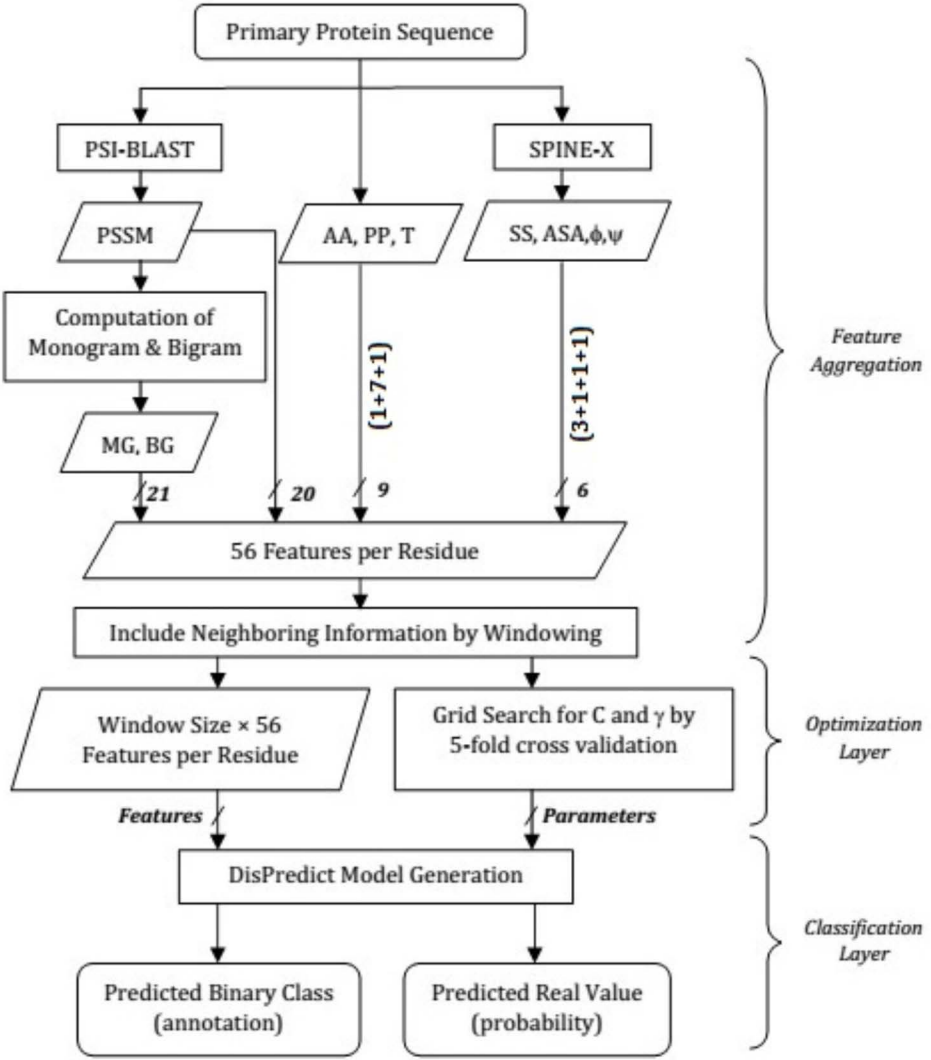
80. Meiler,J., Muller,M., Zeidler,A. and Schmäschke,F. (2001) Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *J Mol Model*, **7**, 360 – 369.

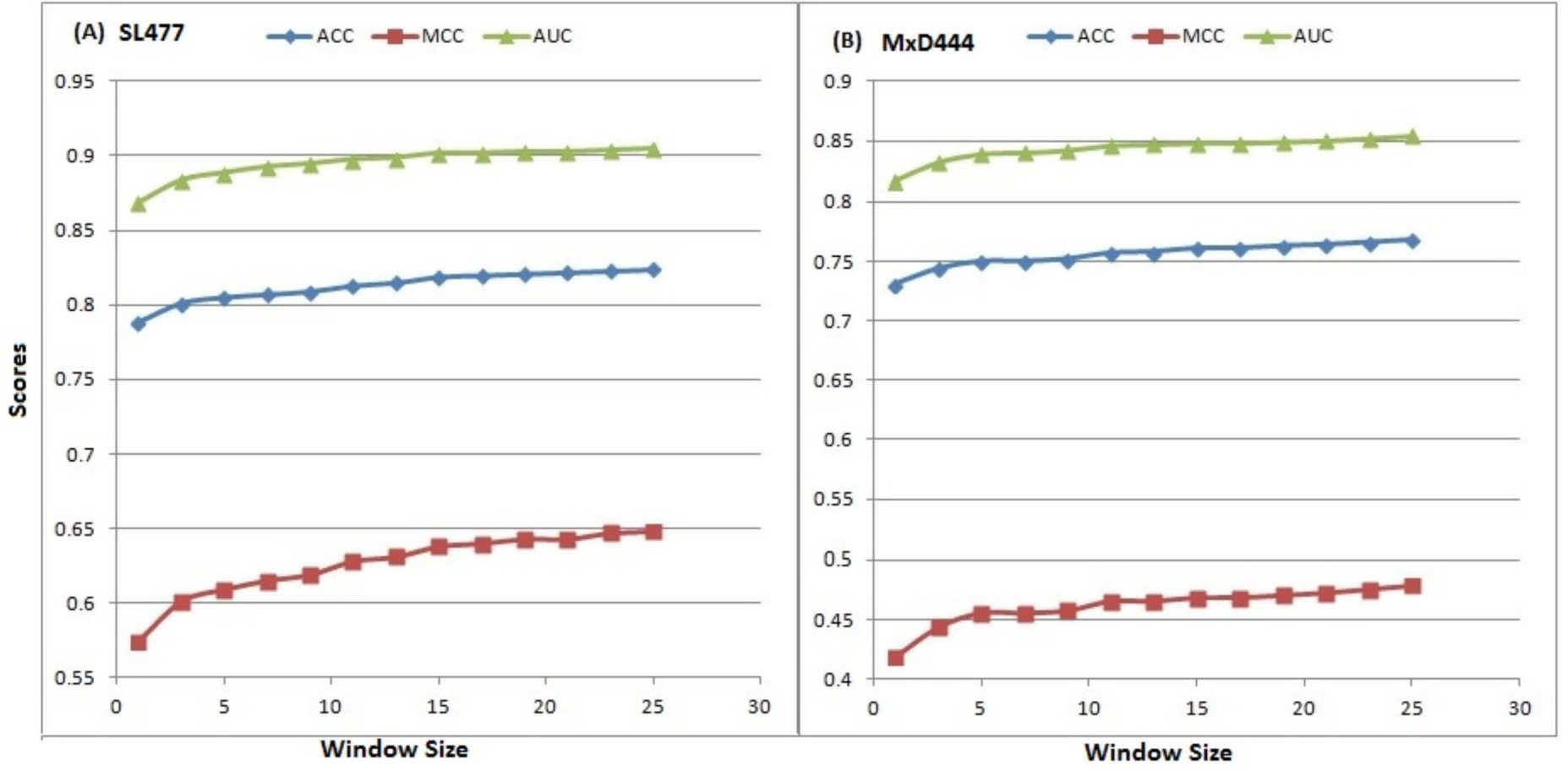
81. Potenza,E., Domenico,T.D., Walsh,I., and Tosatto,S.C.E. (2014) MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucl. Acids Res.*.

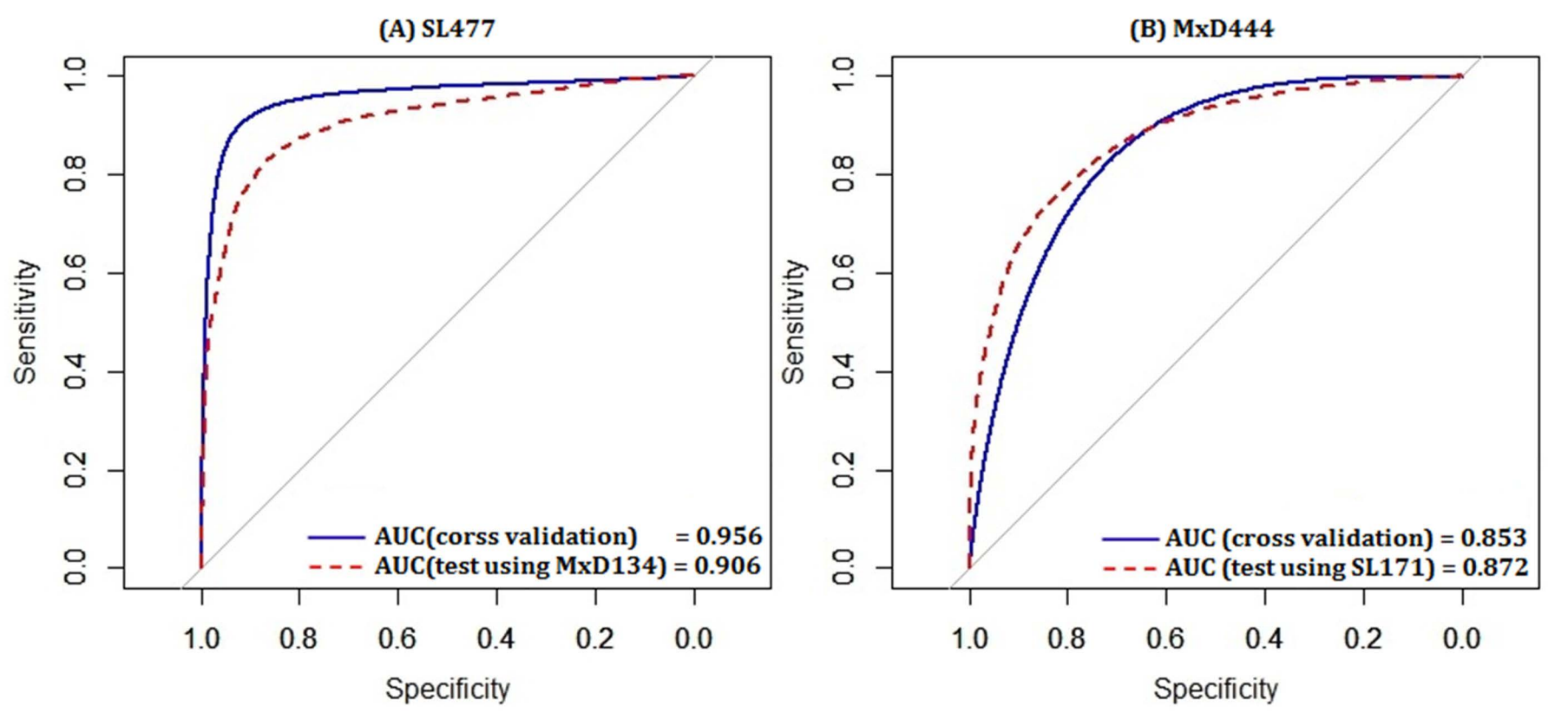
82. Domenico,T.D., Walsh,I., Martin,A.J.M. and Tosatto,S.C.E. (2012) MobiDB: a comprehensive database of intrinsic protein disorder annotations. *Bioinformatics*, **28(15)**, 2080 – 2081.

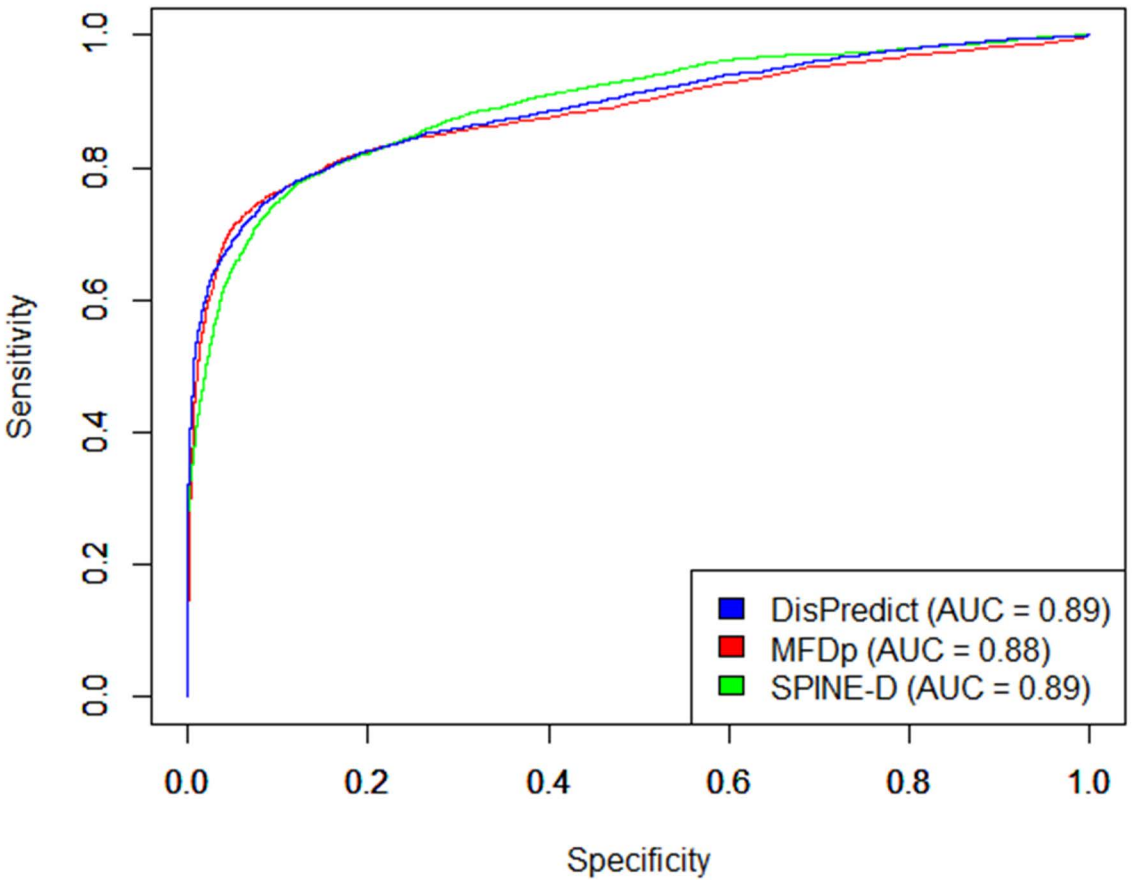
83. Robin,X., Turck,N., Hainard,A., Tiberti,N., Lisacek,F., Sanchez,J.-C. and Müller,M. (2011) pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, **12**, 77.

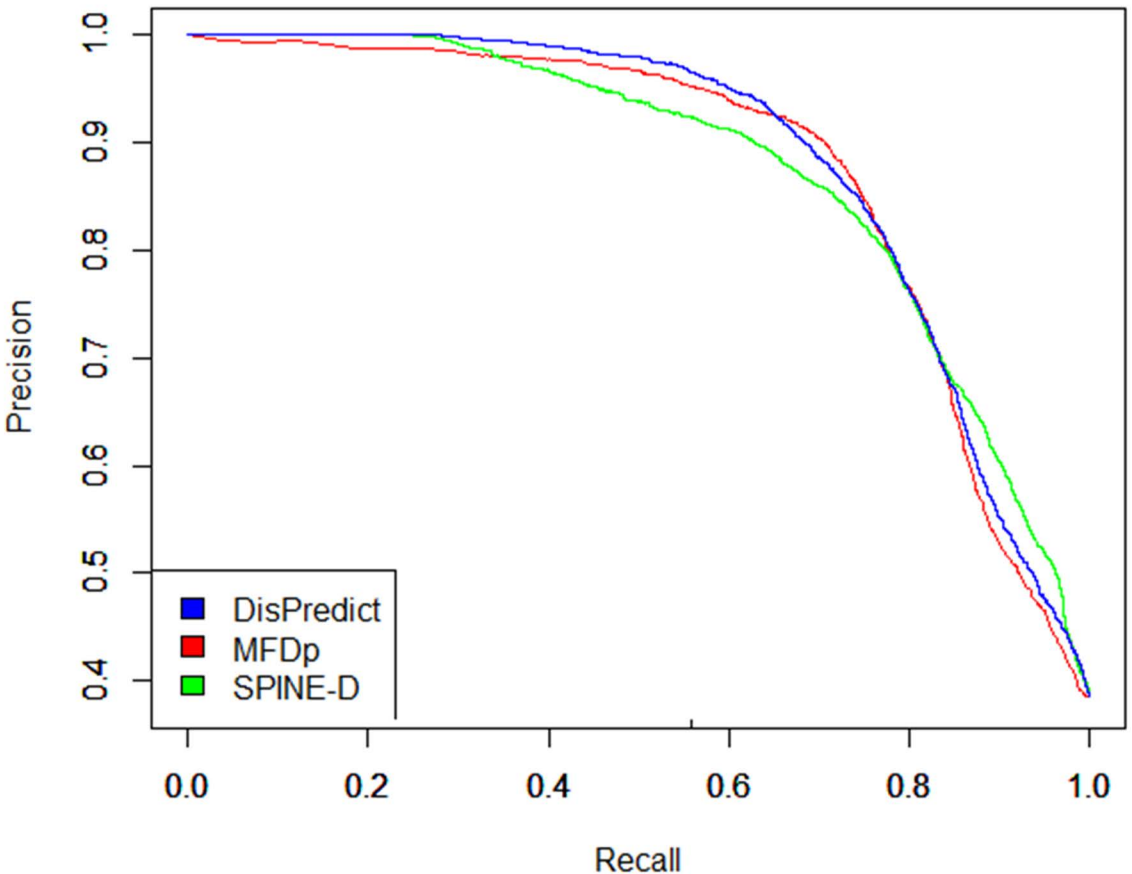

(A) SL477
Median(MG) = 5.9
Median(BG) = 6.2
Density
0.0
0.1
0.2
0.3
0
2
4
6
8
10
Monograms/Bigrams (log)
(B) MxD444
Median(MG) = 6.1
Median(BG) = 6.5
Density
0.0
0.1
0.2
0.3
0.4
0
2
4
6
8
10
Monograms/Bigrams (log)

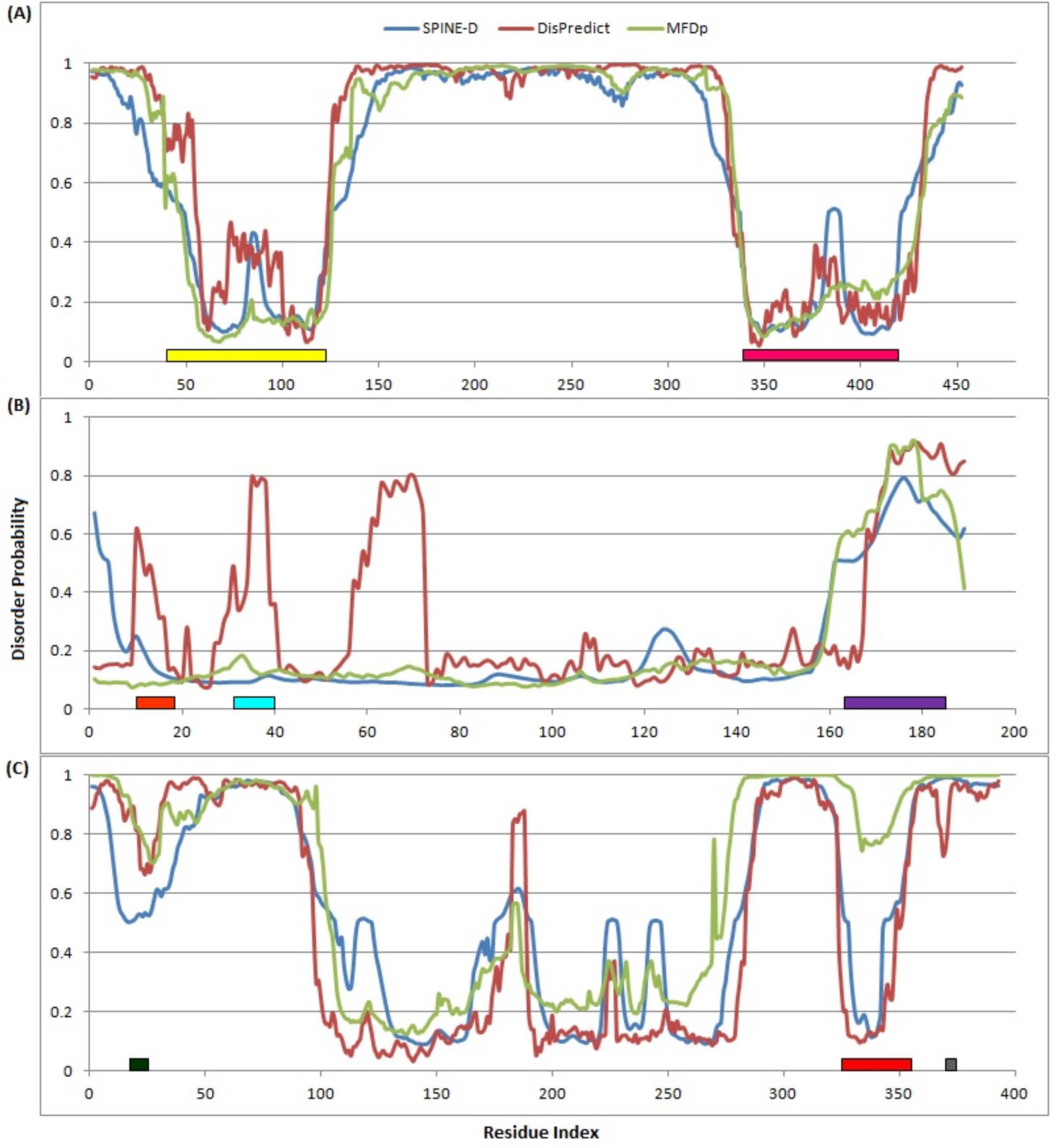

Primary Protein Sequence
PSI-BLAST
SPINE-X
PSSM
AA, PP, T
SS, ASA,ϕ,ψ
Computation of Monogram & Bigram
MG, BG
(1+7+1)
(3+1+1+1)
21
20
9
6
56 Features per Residue
Include Neighboring Information by Windowing
Feature Aggregation
Window Size × 56 Features per Residue
Grid Search for C and γ by 5-fold cross validation
Features
Parameters
Optimization Layer
DisPredict Model Generation
Predicted Binary Class (annotation)
Predicted Real Value (probability)
Classification Layer

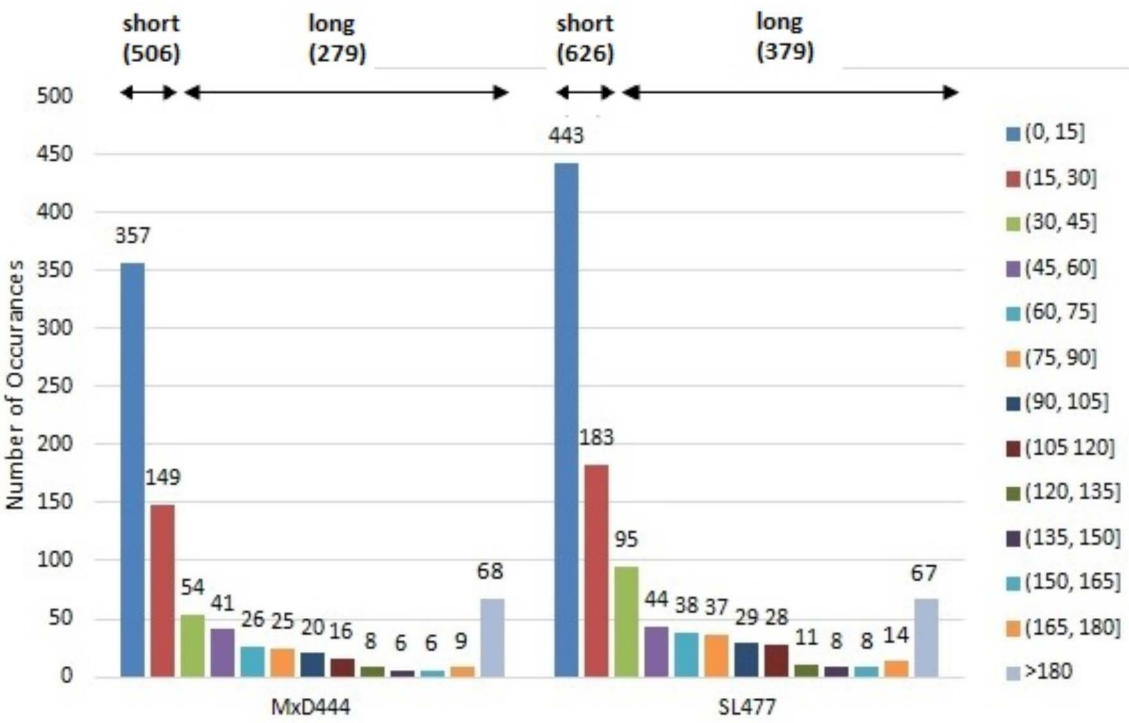

(A) SL477
ACC
MCC
AUC
Scores
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0
5
10
15
20
25
30
Window Size
(B) MxD444
ACC
MCC
AUC
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0
5
10
15
20
25
30
Window Size

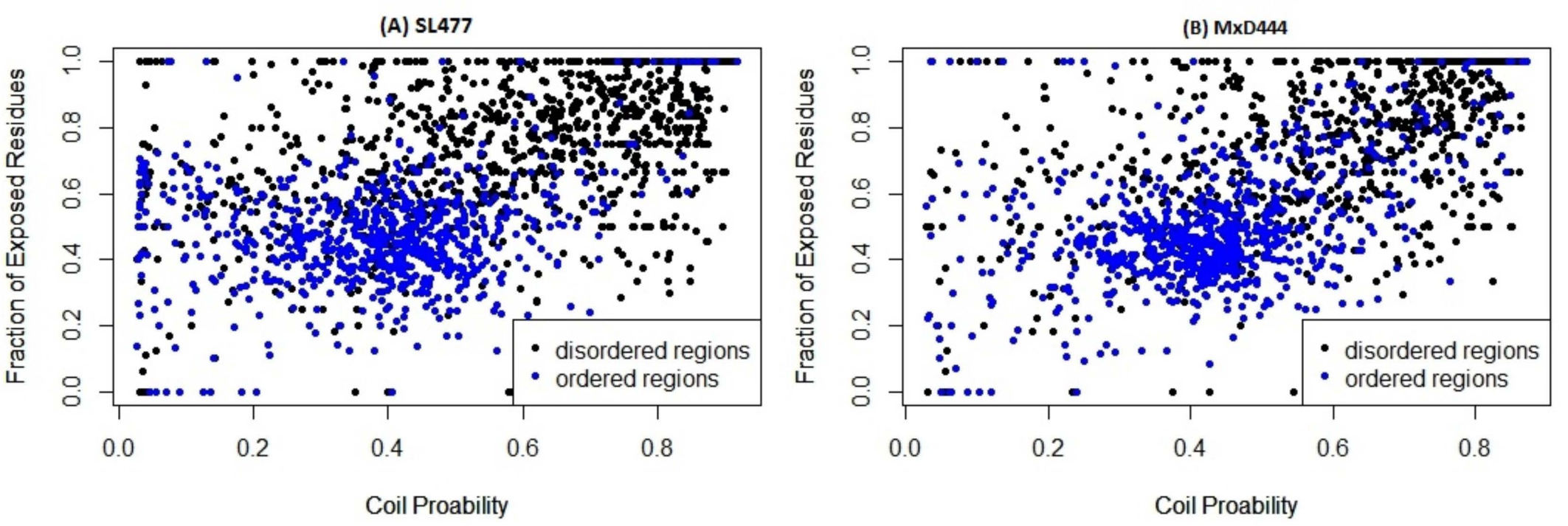

(A) SL477
Sensitivity
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
AUC(corss validation) = 0.956
AUC(test using MxD134) = 0.906
(B) MxD444
Sensitivity
Specificity
AUC (cross validation) = 0.853
AUC (test using SL171) = 0.872

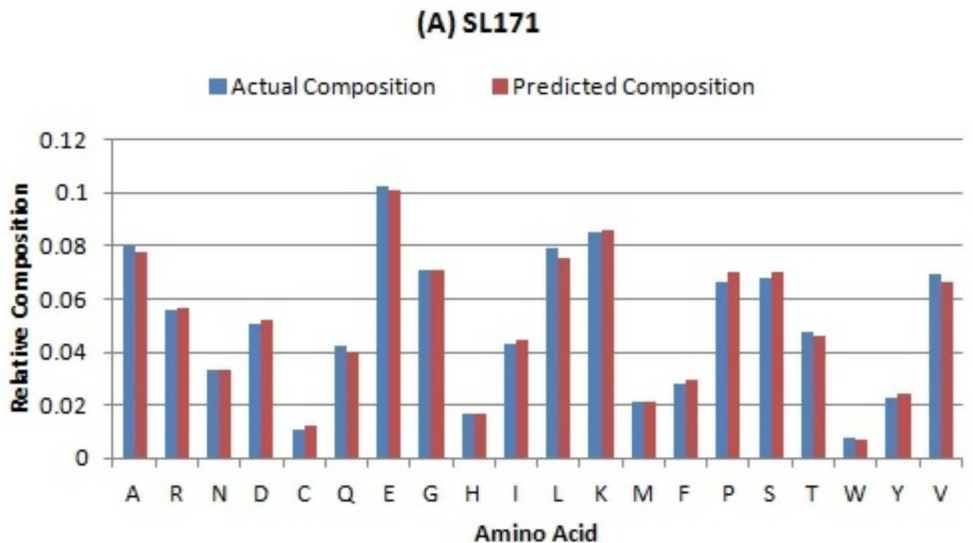

Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
DisPredict (AUC = 0.89)
MFDp (AUC = 0.88)
SPINE-D (AUC = 0.89)

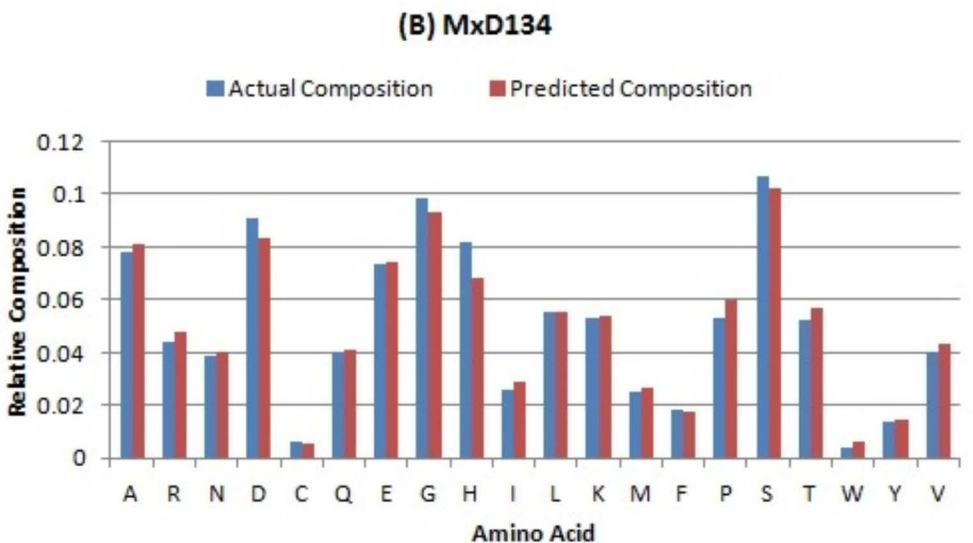

Precision
1.0
0.9
0.8
0.7
0.6
0.5
0.4
DisPredict
MFDp
SPINE-D
0.0
0.2
0.4
0.6
0.8
1.0
Recall


(A)
SPINE-D
DisPredict
MFDp
1
0.8
0.6
0.4
0.2
0
0
50
100
150
200
250
300
350
400
450
(B)
Disorder Probability
1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100
120
140
160
180
200
(C)
1
0.8
0.6
0.4
0.2
0
0
50
100
150
200
250
300
350
400
Residue Index


short
(506)
long
(279)
short
(626)
long
(379)
Number of Occurances
500
450
400
350
300
250
200
150
100
50
0
357
149
54
41
26
25
20
16
8
6
6
9
68
443
183
95
44
38
37
29
28
11
8
8
14
67
MxD444
SL477
(0, 15]
(15, 30]
(30, 45]
(45, 60]
(60, 75]
(75, 90]
(90, 105]
(105 120]
(120, 135]
(135, 150]
(150, 165]
(165, 180]
>180


(A) SL477
Fraction of Exposed Residues
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
Coil Proability
disordered regions
ordered regions
(B) MxD444
Fraction of Exposed Residues
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
Coil Proability
disordered regions
ordered regions


(A) SL171
Actual Composition
Predicted Composition
Relative Composition
0.12
0.1
0.08
0.06
0.04
0.02
0
A R N D C Q E G H I L K M F P S T W Y V
Amino Acid


(B) MxD134
Actual Composition
Predicted Composition
Relative Composition
0.12
0.1
0.08
0.06
0.04
0.02
0
A R N D C Q E G H I L K M F P S T W Y V
Amino Acid