Improved Protein Disorder Predictor by Smoothing Output

Sumaiya Iqbal, Md Nasrul Islam and Md Tamjidul Hoque
Computer Science, University of New Orleans
2000 Lakeshore Drive, New Orleans, LA 70148, USA
Email: {siqbal1, mnislam, thoque}@uno.edu

***Abstract*—Intrinsically disorder regions (IDRs) or, proteins (IDPs) are associated with important biological functions, while lacking stable structure in their native state. The phenomena of disordered proteins or residues are abundant in nature and are extensively involved in critical human diseases and hence impacting drug discovery. Thus, the study using disorder prediction is becoming crucial in the proteomic research. The large scale growth of genome database demands high performance computational methods for identification of protein disorder. We developed a canonical support vector machine based disorder predictor, DisPredict by integrating RBF kernel. It employs novel feature set for accurate characterization of disorder which outperformed two leading predictors: the neural network based SPINE-D and Meta predictor MFDp based on ten-fold cross validation. We propose a post processing of probabilities to further improve the accuracy, named DisPredict1.1 which yields outstanding performance further both in binary annotation and real valued probability prediction per residue in both short and long disordered regions. It provides highest Mathews Correlation Coefficient (MCC), competitive Area Under receiver operating characteristic Curve (AUC) and lowest Mean Absolute Error (MAE) when compared with twenty existing predictors of several kinds on independent benchmark dataset. DisPredict is available online.**

Keywords*—*Intrinsic disorder, Protein prediction, Pattern recognition, SVM, RBF kernel, Monogram, Bigram, Probability smoothing, Cross validation.

I. Introduction

Intrinsically disordered proteins (IDPs) or, disordered regions (IDRs) do not adopt well-defined and stable three-dimensional (3D) structures in their native state [1]. These proteins or partial regions of proteins are also known as natively unstructured, denatured or unfolded. The disordered proteins or residues have been extensively observed in nature. For the last two decades, numerous experiments have been done in evidence that many proteins do not follow the well known paradigm of protein function: *sequence* $\rightarrow$ *structure* $\rightarrow$ *function*; and requires the unstructured state for some functions [2], [3]. IDPs and IDRs are nevertheless frequently involved in essential biological activities, such as cell cycle control and cellular signal transduction, transcriptional and translational regulation, membrane fusion and control pathways [4]. Intrinsic disorder enables a number of capabilities of a protein and therefore participates in molecular recognition, molecular assembly and protein modification [5], [6] via protein-protein, protein-nucleic acid and protein-ligand interactions [7], [8]. Disorder proteins are also found to play key role in critical human diseases [9], [10], such as cancer, AIDS, amyloidoses, cardiovascular and neurodegenerative diseases, genetic diseases, as well as in drug development [11]. Thus, further theoretical and experimental exploration are required for locating protein disorder for better understanding of protein functions related to disorder.

Intrinsically disordered proteins (or regions) form variable conformations as the coordinates of their backbone atoms have no specific equilibrium states and thus adopt dynamic structural ensembles. Structurally, IDPs (or IDRs) encompass proteins or protein-regions with extended disorder (*i.e.*, random coils, intrinsic coils), collapsed disorder (*i.e.*, molten globules) and semi-collapsed disorder (*i.e.*, pre-molten globules, polyglutamine regions and polar sequences) [2], [12]. They are characterized by, alone or in combination, low hydrophobicity, high net charge, low level of stable secondary structure and highly dynamic side chains. Recognition of this protein disorder is important for appropriate protein structure prediction, disease causing protein identification, proper annotation of function, induced folding and binding region prediction.

There are several methods that experimentally define the residues of IDRs or, IDPs. These methods include X-ray crystallography [13], NMR spectroscopy [14], near or far ultraviolet circular dichroism (CD) [15] etc. A curated database of disordered proteins, called DisProt [16] contains annotation for 694 protein sequences and PDB [17] database which gives provision of finding disordered regions in the solved secondary or tertiary structure incorporates $83,452$ protein entries. To compare, the overall number of non-redundant protein sequences is $38,633,935$ according to the most recent 65 release of RefSeq database [18]. On the contrary, the computational methods are capable to produce high throughput predicted annotation of disordered residues in IDPs/IDRs, providing a reasonable solution to fill up the time consuming and costly experimental annotation gap with the rapid growth rate of known protein sequences. Many of these predictors are developed using different pattern recognition methods such as Logistic Regression, Artificial Neural Network (ANN), Support Vector Machine (SVM), Bayesian Classifier, Random Forest (RF) etc. Several other methods utilize the predicted three-dimensional structural characteristics, relative composition and propensity of amino acids or, combination of individual self complementary methods. However, the Critical Assessment of protein Structure Prediction (CASP) competitions[1] further signifies the importance in this area as well as necessity of new and accurate disorder predictors.

[1] http://predictioncenter.org/

We proposed a SVM based disorder predictor, named "*DisPredict*" [19] to classify ordered and disordered residues in a protein sequences as well as assign a disorder confidence score with higher accuracy. DisPredict is a unification of the classical Support Vector Machine (SVM) with *radial basis function (RBF)* as the kernel and a comprehensive set of features per residue. The performance of DisPredict is also strengthened by selection of optimal parameters for RBF kernel and SVM. Finally, we improve the accuracy of its initial version (DisPredic1.0)[2] in this article with window based averaging of per residue disorder probability, resulting in DisPredict1.1 which shows competitive performance when compared with 20 other existing predictors.

The remainder of the paper is organized as follows. Section II, materials and methods, defines disorder proteins, describes the formations of the training and test datasets, selected features and their properties, architecture of the predictor and evaluation metrics. The section confirms that the training and test encompass residues from disordered regions of various length and disorder annotation derived from different sources, such as PDB and DisProt databases which eventually assist DisPredict in identifying various types of disorder correctly. Section III covers different test results and comparison with existing predictors. Finally, we briefly conclude in section IV.

II. Materials and Methods

A. Data Sources and Collection

PDB [17] or DisProt [16] databases include disordered residues or regions assigned by several experimental methods. X-ray crystallography can identify disordered residues with missing coordinates in structure and NMR can show disordered residues with highly variable coordinates within ensemble. The annotation of residues should be done in consistent way for better evaluation of a predictor's performance [20], [21]. We selected two datasets which combine sequences from PDB having disordered residues without coordinates (recorded in REMARK 465) and sequences from DisProt to separately train our predictor. Then, we tested the performance by three independent datasets to generalize our method's insensitivity to annotation technique. The datasets accumulate various types of disorder, including disordered regions of all sizes, both short ($\leq$ 30 residues) and long ($>$ 30 residues) and chains having only ordered or disordered residues. This combination shows that our method is robust in recognizing both short and long disorder regions in proteins.

1) Training Datasets (SL477 and MxD444): SL477 dataset was prepared from the benchmark SL (Short Long) dataset [22] which was built by re-annotating the sequences extracted from DisProt. SL dataset contains disordered regions which are short, with length less than 20 residues [22] which are found functionally and structurally important as well as very long disordered regions. SL dataset's sequences were further clustered and filtered using BLASTCLUST 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. Note that, we disregarded the residues with unknown annotation during training.

The Mixed Disorder (MxD) dataset is a combination of diordered protein sequences from both PDB and DisProt databases. Originally developed MxD dataset [23] has 514 protein sequences including 205 chains from PDB and 309 chains from DisProt. Later, we purified the dataset to discard any sequence containing unknown amino acid (X-tag). 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.

2) Test Datasets (SL171, MxD134 and DP_NEW): SL171 dataset is generated from SL477 dataset by executing BLASTCLUST from NCBI-BLAST [24] package. We filtered SL477 dataset to discard any sequence with greater than 10% sequence identity with MxD444 dataset which gave us a test set of 171 chains with $42,572$ residues, named as SL171. We utilized SL171 as an independent test set to evaluate the DisPredict model when it was trained on MxD444 dataset. Another distinction between our two test datasets is, MxD134 was ensured to contain sequences with disordered regions defined by PDB. On the other hand, SL171 contains protein sequences with disorder annotation only from DisProt.

MxD134 dataset is prepared to facilitate the process of independently train our model by SL477 dataset and test the resulting model with MxD134 dataset. We extracted this independent test dataset from MxD444 by removing sequences with similarity greater than 10% to any sequence from SL477 dataset using BLASTCLUST package, retrieving a set of 134 protein chains with $38,823$ residues. MxD134 dataset was employed to evaluate our predictor while training was performed on SL477 dataset.

In addition, we downloaded the benchmark DP_NEW dataset [25]. This dataset encompasses disorder annotation from PDB REMARK 465 as well as curated annotation from DisProt. The standard CASP protocol was followed [**?**] for the PDB annotation and the DisProt annotations were further enriched by with help of PDB REMARK 465 [25]. Moreover, BLATCLUST was used to filter the resulting dataset so that no sequence is more than 25% similar to MxD dataset which resulted another independent test dataset of 105 protein chains. DP_NEW dataset comprises of $31,511$ residues that combines 4640 (about 14.7%) disordered residues, $17,798$ ordered residues (about 56.4%) and $9,073$ unknown residues(about 28.7%).

B. Input Features

We gathered the most comprehensive and independent set of residue level input features, which is capable of capturing the sequence information, evolutionary information as well as the structural information. The residue level information includes: (a) single valued amino acid type (all the necessary information for the correct folding of a protein is encoded in its amino acid sequence [26]); (b) seven physicochemical properties of amino acid (different types, short or long, disordered regions in protein are found to have distinguished physicochemical properties); (c) twenty PSSM's (position specific scoring matrix) indicating the evolutionary information accumulated in each residue position of a protein sequence; (d) three predicted

[2]DisPredict is available at https://www.dropbox.com/s/dz8tzoej692vs0l/SuppMaterial_DisPredict.zip

secondary structure (helix, strand and coil) probabilities from SPINE-X [27], one predicted accessible surface area (ASA) normalized by the ASA of an extended conformation (Ala-X-Ala) [28] and two predicted backbone torsion angle (phi, psi) fluctuations [29] since disordered residues are characterized by lack of stable secondary structure [30], highly exposed area and angle fluctuations; (e) one monogram and twenty bigrams computed from PSSM [31] representing the conserved evolutionary information of PSSM transformed from primary structure level to three dimensional structure level, which are normalized by the median of normal density distribution of monogram and bigram values in their logarithmic space; (f) one indicator for terminal residues (five residues from N-terminal as $\{-1.0, -0.8, -0.6, -0.4, -0.2\}$, five residue from C-terminal from $\{+1.0, +0.8, +0.6, +0.4, +0.2\}$ respectively, with the rest as 0.0). Finally, before feeding the features into the classifier, neighboring residue's information is aggregated using a sliding window of 21 residues (10 residues on each residue to be predicted), resulting in $21 \times 56 = 1176$ features per residue. The window size 21 was found to be optimal both in terms of accuracy and speed of prediction. This is also motivated 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, making our methodology biologically significant.

C. Predictor Framework

DisPredict1.1 follows our initially designed SVM based classifier model [19] for prediction of per residue binary annotation (order or disorder) and assigning two real values as probability score of being order or disorder. SVM with RBF kernel simultaneously minimizes the empirical classification error (training error) and generalized error (test error) by maximizing the geometric margin of the separating hyperplane. The predictor consists of two layers. The optimization layer determines the optimal values of two parameters, C and γ, where C is the cost of misclassification which softly penalizes the feature space points that lie on the wrong side of the decision boundary and γ is the parameter involved in radial basis function (RBF). The parameter selection is done with optimization on accuracy (fraction of correctly predicted residues) by grid search, which is guided by 5-fold cross validation. The classifier layer of the predictor generates the binary and real valued prediction. The real values are binarized using a natural threshold equal to 0.5, $0.5 \leq$ range ≤ 1.0 is considered as disordered probability and $0.0 \leq$ range < 0.5 is considered as ordered probability. We utilized LIBSVM [32] for SVM parameterization and model generation. Finally, we processed the probabilities by taking the average of the resulting probabilities with a sliding window of 29 residues (14 residues on either side of the target residue) and converted the scores into binary annotation using the same threshold of 0.5. We selected the window size which provided us the highest MCC scores in performance evaluation. With this post processing step, DisPredict1.1 applies a smoothing on the probabilities to take the impact of relative type (order or disorder) of the neighboring residues while assigning the score for a target residue which improves both MCC and AUC scores achieved by DisPredict1.0. However, we have not applied this smoothing of probability for the N and C terminal region due to their highly flexible and dynamic conformation.

D. Performance Evaluation

We followed the evaluation criteria of CASP [33] to determine the performance of our predictor. The binary prediction is assessed by using the following criteria:

$$\text{Sensitivity } (SENS) = TP/(TP + FN)$$

$$\text{Specificity } (SPEC) = TN/(TN + FP)$$

$$\text{Balanced Accuracy } (ACC) = (SENS + SPEC)/2$$

$$\text{Precision } (PPV) = TP/(TP + FP)$$

$$\text{Weighted Score } (S_w) = \frac{w_d \times TP - w_o \times FP + w_o \times TN - w_d \times FN}{w_d \times N_d + w_o \times N_o}$$

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

Here, *TP* (True Positive), *TN* (True Negative), *FP* (False Positive) and *FN* (False Negative) denote the number of correctly predicted disordered residues, correctly predicted ordered residues, incorrectly predicted disordered residues and incorrectly predicted ordered residues, respectively. N_d and N_o are the total number of disordered and ordered residues in the dataset whereas $w_d = \frac{N_d}{N_o+N_d}$ and $w_o = \frac{N_o}{N_o+N_d}$ are the percentage of disordered and ordered residues in the dataset, respectively. MCC score accounts for all four parameters of the prediction quality and is regarded as is the most reasonable 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. 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 for the i^{th} protein chain, respectively. The lower value of MAE corresponds to better prediction.

The probability prediction performance of a predictor can be analyzed by the receiver operating characteristic (ROC) curve. A ROC curve 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. We calculate the AUC to assess the performance of DisPredict using the trapezoid rule.

III. Test Results and Discussion

We first performed the ten fold cross validation on our two training dataset, SL477 and MxD444 separately [19] to select the window size which yields optimal performance. To appropriately partition the dataset into completely non overlapping subsets, we employed modular arithmetic operation to split the dataset in residue level. After that, we carried out the grid search guided by another round of internal cross validation, for optimizing the parameters for RBF kernel and SVM. With the optimal window size (21) and parameters (C and γ), DisPredict achieved the ACC, MCC and AUC score of 0.836, 0.673 and 0.956 as a result of ten fold cross validation on SL477 dataset. Training on SL477 dataset and applying to the independent test set of MxD134 with lower than 10%

sequence similarity resulted consistent ACC, MCC and AUC of values 0.833, 0.598 and 0.906 respectively [19]. On the other hand, as a result of ten fold cross validation on MxD444 dataset, DisPredict predicted disorder with ACC, MCC and AUC values of 0.805, 0.600 and 0.853. The result of training on MxD444 dataset and test on the independent set SL171 is found equally promising with ACC, MCC and AUC values of 0.789, 0.583 and 0.872 [19]. The consistency of the two tests indicated robust training. Moreover, the accuracy in prediction of disorder for two different types of datasets in balanced accuracy (ACC) and PPV values prove that our methodology is accurate as well as precise.

With the additional correction of predicted probabilities by sliding window based averaging and transforming the resulting probabilities into binary annotation, DisPredict1.1 outperforms DisPredict1.0 [19] both in binary annotation and probability prediction. Table I further illustrates this comparison of results in case of independent tests of the predictor two (MxD134 and SL171) datasets. DisPredict1.1 improved the performance for binary disorder or order prediction by 0.48%, 0.89%, 2.96%, 2.17% in terms of accuracy, S_w, precision and MCC, respectively during the test by MxD134 dataset. On the other hand, while testing with SL171 dataset, there are significant increase by 6.38% and 4.63% in precision and MCC, respectively. However, the accuracy decreased slightly which caused by the decrease of SENS along with significant increase in SPEC. DisPredict1.1 also provided consistent improvement in assigning per residue confidence score with 0.55% and 1.83% increase in AUC score for MxD134 and SL171 datasets. This improvement is further analyzed with the ROC curves in Figure 1 which depicts better correlation between sensitivity and specificity with smoothing. Overall, the consistent performance for two different test sets justifies rigorous training and precise methodology.

TABLE I: PERFORMANCE COMPARISON OF DISPREDICT1.0 AND DISPREDICT1.1

Predictor[a]	Test Set	SENS	SPEC	ACC	S_w	PPV	MCC	AUC	MAE
DisPredict1.1 (SL477)[b]	MxD134	0.745	0.928	0.837	0.673	0.591	0.611	0.911	0.083
DisPredict1.0 (SL477)[c]	MxD134	0.744	0.923	0.833	0.667	0.574	0.598	0.906	0.023
DisPredict1.1 (MxD444)[b]	SL171	0.644	0.926	0.785	0.570	0.834	0.610	0.888	0.032
DisPredict1.0 (MxD444)[c]	SL171	0.718	0.860	0.789	0.577	0.748	0.583	0.872	0.151

[a] The predictor name is specified with the corresponding training dataset in parenthesis. The training was done with window size 21 and optimal SVM parameters.
[b] Probabilities smoothed with a sliding window size 29.
[c] No probability smoothing.

We compared the performance of DisPredict1.0 and DisPredict1.1 against twenty existing methods (including sub versions of some tools for different types of disorder) which cover all four categories of disorder prediction methods discussed in Section I. The methods include DISOPRED [34], 3 versions of ESpritz (X, N and D) [35], PROFbval [36], PrDOS [37], NORSnet [21], PreDisOrder [38], 2 versions of IUPred (short and long) [39], Ucon [20], DISOclust [40], 2 versions of CSpritz (short and long) [41], MD [42], SPINE-D [43], MFDp [23], PONRD-FIT [44] and very recent 2 versions of MFDp2 (with and without BLAST) [25]. To compare consistently, we collected the performances of these methods on DP_NEW benchmark dataset from MFDp2 article [25] and evaluated the performance of DisPredict1.0 and DisPredict1.1 on same dataset. Note that, DP_NEW dataset

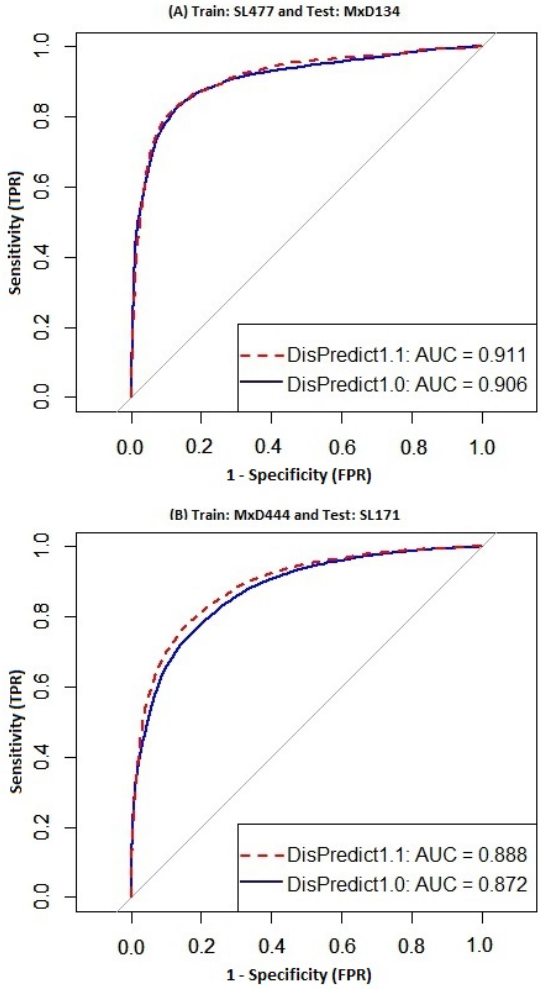

Fig. 1: ROC curves for (A) Train by SL477, test by MxD134 and (B) Train by MxD444, test by SL171. In each figure, the solid (*blue*) and dotted (*red*) curve corresponds to the performance of DisPredict1.0 and DisPredict1.1, respectively. The AUC values are given in the legend according to the respective ROC.

contains about 28.7% residues annotated as unknown. To remain consistent, we also evaluated our predictors assuming the unknown residues as order at first and then discarding the unknown residues. Comparisons among different predictors at both level are presented quantitatively in Table II and Table III in terms of SENS, SPEC, MCC, AUC, MAE and PCC. Here, SENS, SPEC, MCC and AUC are used to determine the performance in binary annotation prediction and probability prediction at residue level, while MAE indicates the performance of disorder prediction in content level. Table II shows that DisPredict1.1 results highest MCC among all the other methods and outperforms the previous best result given by MFDp2 [25] by 4.18% when trained on SL477 dataset and and by 0.63% when trained on MxD444 dataset. The AUC score of DisPredict1.1 is also found competitive. The best score of specificity was given by Espritz D at the cost of very low sensitivity. However, both sensitivity and specificity given by DisPredict1.1 are comparable. Table III shows that all the scores provided by DisPredict are competitive and outperform 18 existing predictors in terms of MCC and AUC except MFDp2. However, MFDp2 does not consider relatively short

TABLE II: PERFORMANCE COMPARISON OF DISPREDICT1.0 AND DISPREDICT1.1 WITH 20 EXISTING PREDICTORS WHEN RESIDUES WITHOUT ANNOTATION ARE ASSUMED AS ORDERED

Method[a]	SENS	SPEC	MCC	AUC	MAE
DisPredict1.1 (SL477) [This Work]	77.3	83.8	**0.499**	0.857	0.081
DisPredict1.1 (MxD444) [This Work]	66.2	**88.1**	0.482	**0.862**	**0.04**
MFDp2 [25]	75.9	83.2	0.479	0.862	0.153
DisPredict1.0 (SL477) [19]	**77.4**	82.2	0.478	0.85	0.092
MFDp2 (no blast) [25]	75.4	83.2	0.475	0.86	0.153
MFDp [23]	80.9	79.3	0.466	0.85	0.174
DisPredict1.0 (MxD444) [19]	67.8	86.3	0.466	0.845	0.054
Cspritz L [41]	83.5	77.5	0.463	0.87	0.242
MD [42]	72.6	79.9	0.414	0.829	0.235
Espritz X [35]	53.8	88.7	0.394	0.801	0.139
Cspritz S [41]	73.5	77.2	0.39	0.823	0.209
PrDos [37]*	55.8	86.8	0.388	0.818	0.137
PONDR-FIT [44]	66.3	81.5	0.387	0.8	0.162
SPINE-D [43]	78.4	72.9	0.381	0.823	0.204
IUPreD L [39]	60.4	84.4	0.38	0.788	0.13
PreDisorder [38]*	74.5	74.1	0.374	0.797	0.234
DISOPRED2 [34]	65.6	80.5	0.37	0.797	0.153
IUPreD S [39]	54.5	86.7	0.368	0.782	0.133
Espritz D [35]	40.9	92.0	0.349	0.827	0.186
DISOCLUST [40]	75.3	71.3	0.343	0.803	0.19
Espritz N [35]	60.2	80.5	0.329	0.785	0.168
NORSnet [21]	47.3	87.6	0.323	0.761	0.172
UCON [20]	60.5	76.6	0.289	0.732	0.179
PROFBVaL [36]	52.8	65.1	0.13	0.631	0.307

[a] The methods are sorted according to MCC.
For each metric, our best result is marked in bold and previously found best result is underlined.
* According to MFDp2 [25], PrDos and PreDisorder failed for one chain and were evaluated on 104 chains.

TABLE III: PERFORMANCE COMPARISON OF DISPREDICT1.0 AND DISPREDICT1.1 WITH 20 EXISTING PREDICTORS WHEN RESIDUES WITHOUT ANNOTATION ARE DISCARDED

Method[a]	SENS	SPEC	MCC	AUC
MFDp2 [25]	75.9	95.3	0.729	0.94
MFDp2 (no blast) [25]	75.4	95.3	0.725	0.938
DisPredict1.1 (SL477) [This Work]	77.3	94	**0.711**	**0.925**
MFDp [23]	80.9	92.2	0.704	0.925
DisPredict1.1 (MxD444) [This Work]	66.2	**96.4**	0.683	0.912
DisPredict1.0 (SL477) [19]	**77.4**	92.2	0.677	0.914
DisPredict1.0 (MxD444) [19]	67.9	94.0	0.642	0.89
CSpritz L [41]	83.5	85.9	0.621	0.909
DISOPRED2 [34]	65.6	93.6	0.614	0.88
IUPreD L [39]	60.4	94.3	0.588	0.851
DISOCLUST [40]	75.3	87.4	0.581	0.904
MD [42]	72.6	88.4	0.576	0.873
PrDos [37]*	55.8	95.4	0.576	0.883
SPINE-D [43]	78.4	85.4	0.575	0.893
PONDR-FIT [44]	66.3	90.3	0.558	0.85
NOrSnet [21]	47.3	96.7	0.54	0.834
ESPrITZ X [35]	53.8	94.5	0.54	0.845
IUPreD S [39]	54.5	93.6	0.525	0.83
Cspritz S [41]	73.5	83.6	0.512	0.857
PreDisorder [38]*	74.5	82.4	0.503	0.85
ESPrITZ N [35]	60.2	89.4	0.492	0.844
ESPrITZ D [35]	40.9	94.4	0.426	0.866
UCON [20]	60.5	84.4	0.42	0.78
PROFBVal [36]	52.8	67.2	0.167	0.647

[a] The methods are sorted according to MCC.
For each metric, our best result is marked in bold and previously found best result is underlined.
* According to MFDp2 [25], PrDos and PreDisorder failed for one chain and were evaluated on 104 chains.
MAE is not reported since content level evaluation is not consistent when residues are discarded from a protein chain.

disordered regions (less than 4 residues) in the evaluation, while DisPredict is evaluated for all types and length of disordered regions. We consider the short disordered regions since they are biologically significant and our result provides us with evidence that the methodology of our predictor gives promising performance for all types of disorder.

IV. CONCLUSION

In this article, we proposed an improvement over our classical support vector machine based disorder predictor, called DisPredict1.1, which usages a RBF kernel and includes useful and advanced features for predicting disordered residues. The superior performance of our predictor is mainly due to the use of a novel methodology that incorporates highly effective radial basis kernel function (RBF) in generating classifier model capable of dealing with non linearly separable classes, along with optimized set of parameters for handling overlapped classes. The distinguishing property of our feature set in comparison with existing predictors is the inclusion of monogram (MG) and bigram (BG) which can identify evolutionary conserved fold. An additional post processing of probabilities with window based averaging and correcting the binary order or disorder annotation accordingly is found effective to reduce the noise in prediction, as such averaging captures the impact of the relative structured or unstructured status of neighboring residues.

DisPredict1.1 outperforms its own predecessor during test by two independent dataset, MxD134 and SL171, both in binary annotation and real valued confidence score prediction. The datasets used to train and test our predictor model encompass 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 benchmark dataset DP_NEW, used to compare DisPredict's methodology with twenty existing state-of-the-art disorder predictors, combines 43 protein chains with curated annotation of DisProt and 62 chains annotated by PDB. Moreover, this dataset contains 115 short disordered regions (less than 30 residues) and 28 long disordered regions (greater than or equal to 30 residues) combined with 17 full ordered and disordered proteins. This combination of several length disordered regions included within training and testing confirms the consistent performance for all sizes of disordered region as well as different types of disordered residues. An extensive comparison with twenty other methods using DP_NEW dataset reveals that our predictor achieves highest MCC, which is regarded as the most reasonable measure for disorder prediction, as MCC does not favor over prediction of any classes. Moreover, DisPredict1.1 results in comparable AUC scores which indicates the quality of real valued probability prediction. Our predictor is available online as a standalone software tool.

ACKNOWLEDGMENT

The authors would like to 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 Avdesh Mishra and Denson Smith.

REFERENCES

[1] V. Uversky and A. Dunker, "Understanding protein non-folding." *Biochimica Et Biophysica Acta (BBA) - Proteins And Proteomics*, vol. 1804, no. 6, pp. 1231 – 1264, June 2010. [Online]. Available: http://dx.doi.org/10.1016/j.bbapap.2010.01.017

[2] A. Dunker and Z. Obradovic, "The protein trinity–linking function and disorder." *Nat Biotechnol*, vol. 19, no. 6, pp. 805 – 806, Sep 2001.

[3] P. C. Whitford, "Disorder guides protein function." *Proc Natl Acad Sci USA*, vol. 110, no. 18, pp. 7114 – 7115, 2013.

[4] V. Uversky, C. Oldfield, and A. K. Dunker, "Showing your ID : intrinsic disorder as an ID for recognition, regulation, and cell signaling." *J. Mol. Recogn.*, vol. 18, no. 5, pp. 343 – 384, Aug 2005.

[5] A. Dunker, C. Brown, and Z. Obradovic, " Identification and functions of usefully disordered proteins." *Adv. Protein Chem*, vol. 62, pp. 25 – 49, 2002.

[6] A. K. Dunker, C. J. Brown, J. D. Lawson, L. M. Iakoucheva, and Z. Obradovic, "Intrinsic disorder and protein function." *Biochemistry*, vol. 41, pp. 6573 – 6582, 2002.

[7] P. Tompa, "The interplay between structure and function in intrinsically unstructured proteins." *FEBS Lett*, vol. 15, no. 13, pp. 3346 – 3354, Jun 2005.

[8] M. Fuxreiter, I. Simon, and S. Bondos, "Dynamic protein - DNA recognition : beyond what can be seen." *Trends Biochem Sci*, vol. 36, no. 8, pp. 415 – 423, Aug 2011.

[9] B. Xue, A. Dunker, and V. Uversky, "The Roles of Intrinsic Disorder in Orchestrating the Wnt-Pathway." *Journal of Biomolecular Structure and Dynamics*, vol. 29, no. 5, pp. 843 – 861, 2012.

[10] B. Xue, M. Mizianty, L. Kurgan, and V. Uversky, "Protein intrinsic disorder as a flexible armor and a weapon of HIV-1." *Cell Mol Life Sci*, vol. 69, no. 8, pp. 1211 – 59, Apr 2012.

[11] M. Babu, R. van der Lee, N. de Groot, and J. Gsponer, "Intrinsically disordered proteins: regulation and disease." *Current Opinion in Structural Biology*, vol. 21, no. 3, pp. 432 – 440, 2011.

[12] A. H. Mao, S. L. Crick, A. Vitalis, C. L. Chicoine, and R. V. Pappu, "Net charge per residue modulates conformational ensembles of intrinsically disordered proteins." *Proc Natl Acad Sci USA*, vol. 107, no. 18, pp. 8183 – 8188, Feb 2010.

[13] D. Ringe and G. Petsko, "Study of protein dynamics by X-ray diffraction." *Methods in Enzymology*, vol. 131, pp. 389 – 433, 1986.

[14] S. Kosol, S. Contreras-Martos, C. C. no, and P. Tompa, "Structural Characterization of Intrinsically Disordered Proteins by NMR Spectroscopy." *Molecules*, vol. 18, no. 9, pp. 10 802 – 10 828, 2013.

[15] G. Fasman, "Circular dichroism and the conformational analysis of biomolecules." Plenum Press, 1996.

[16] M. Sickmeier, J. Hamilton, T. LeGall, V. Vacic, M. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V. Uversky, Z. Obradovic, and A. Dunker, "DisProt: the Database of Disordered Proteins." *Nucleic Acids Res*, vol. 35, pp. 786 – 793, Jan 2007.

[17] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank." *Nucleic Acids Res*, vol. 28, pp. 235 – 242, Oct 1999. [Online]. Available: http://dx.doi.org/10.1093/nar/28.1.235

[18] K. Pruitt, T. Tatusova, W. Klimke, and D. Maglott, "NCBI Reference Sequences: current status, policy and new initiatives." *Nucleic Acids Res.*, vol. 37, pp. D32 – D35, 2009.

[19] S. Iqbal and M. Hoque, "DisPredict: A Fine Disorder-Protein Predictor." *Tech. Report TR-2014/1*, July 2014. [Online]. Available: https://www.dropbox.com/s/9n2hah5h0ieifpc/DisPredict_RBF.pdf

[20] A. Schlessinger, M. P. M, and B. Rost, "Natively unstructured regions in proteins identified from contact predictions." *Bioinformatics*, vol. 23, no. 18, pp. 2376 – 2384, Sep 2007.

[21] A. Schlessingera, J. Liu, and B. Rost, "Natively Unstructured Loops Differ from Other Loops." *Bioinformatics*, vol. 3, no. 7, pp. e140 – e151, Jul 2007.

[22] F. L. Sirota, H. S. Ooi, T. Gattermayer, G. Schneider, F. Eisenhaber, and S. Maurer-Stroh., "Parameterization of disorder predictors for large-scale applications requiring high specificity by using an extended benchmark dataset." *BMC Genomics*, vol. 11, no. Suppl 1, p. S15, 2010.

[23] M. Mizianty, W. Stach, K. Chen, K. Kedarisetti, F. Disfani, and L. Kurgan, "Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources." *Bioinformatics*, vol. 26, no. 18, pp. 489 – 496, Sep 2010.

[24] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool." *J Mol Biol.*, vol. 215, pp. 403 – 410, 1990. [Online]. Available: http://ncbi.nlm.nih.gov/BLAST

[25] M. J. Mizianty, Z. Peng, and L. Kurgan, "MFDp2: Accurate predictor of disorder in proteins by fusion of disorder probabilities, content and profiles." *Intrinsically Disordered Proteins*, vol. 1, p. e24428, 2013.

[26] C. Anfinsen, "Principles that govern the folding of protein chains." *Science*, vol. 181, no. 96, pp. 223 – 230, 1973.

[27] E. Faraggi, T. Zhang, Y. Y. Y, L. Kurgan, and Y. Zhou, "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.*, vol. 33, no. 3, pp. 259 – 267, Jan 2012.

[28] S. Ahmad, M. Gromiha, and A. Sarai, "Real value prediction of solvent accessibility from amino acid sequence." *Proteins.*, vol. 50, no. 4, pp. 629 – 635, Mar 2003.

[29] Z. Y. Zhang T, Faraggi E, "Fluctuations of backbone torsion angles obtained from NMR-determined structures and their prediction." *Proteins.*, vol. 78, no. 16, pp. 3353 – 3362, Dec 2010.

[30] P. Radivojac, Z. Obradovic, D. Smith, G. Zhu, S. Vucetic, C. Brown, J. Lawson, and A. Dunker, "Protein flexibility and intrinsic disorder." *Protein Sci*, vol. 10, pp. 71 – 80, Jan 2004.

[31] A. Sharma, J. Lyons, A. Dehzangi, and K. Paliwal, "A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition." *J Theor Biol.*, vol. 320, pp. 41 – 46, Mar 2013.

[32] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

[33] B. Monastyrskyy, A. Kryshtafovych, J. Moult, A. Tramontano, and K. Fidelis, "Assessment of protein disorder region predictions in CASP10." *Proteins*, vol. 82, no. Suppl 2, pp. 127 – 137, Feb 2014.

[34] J. J. Ward, L. J. McGuffin, K. Bryson, B. F. Buxton, and D. T. Jones, "The DISOPRED server for the prediction of protein disorder." *Bioinformatics*, vol. 20, no. 13, pp. 2138 – 2139, 2004.

[35] I. Walsh, A. Martin, T. D. Domenico, and S. Tosatto, "ESpritz: accurate and fast prediction of protein disorder." *Bioinformatics*, vol. 28, no. 4, pp. 503 – 509, 2012.

[36] A. Schlessinger, G. Yachdav, and B. Rost, "PROFbval: predict flexible and rigid residues in proteins." *Bioinformatics*, vol. 22, no. 7, pp. 891 – 893, Apr 2006.

[37] T. Ishida and K. Kinoshita, "PrDOS: prediction of disordered protein regions from amino acid sequence." *Nucleic Acids Res*, vol. 35, no. Web Server issue, pp. W460 – W464, Jul 2007.

[38] X. Deng, J. Eickholt, and J. Cheng, "PreDisorder: ab initio sequence-based prediction of protein disordered regions." *BMC Bioinformatics*, vol. 10, pp. 436 – 441, 2009.

[39] Z. Dosztányi, V. Csizmok, P. Tompa, and I. Simon, "IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content." *Bioinformatics*, vol. 21, no. 16, pp. 3433 – 3434, Aug 2005.

[40] L. McGuffin, "Intrinsic disorder prediction from the analysis of multiple protein fold recognition models." *Bioinformatics*, vol. 24, no. 16, pp. 1798 – 1804, Aug 2008.

[41] I. Walsh, A. Martin, T. D. Domenico, A. Vullo, G. Pollastri, and S. Tosatto, "CSpritz: accurate prediction of protein disorder segments with annotation for homology, secondary structure and linear motifs." *Nucleic Acids Res.*, vol. 39, pp. W190 – W196, Jul 2011.

[42] A. Schlessinger, M. Punta, G. Yachdav, L. Kajan, and B. Rost, "Improved Disorder Prediction by Combination of Orthogonal Approaches." *PLoS One*, vol. 4, pp. e4433 – e4442, 2009.

[43] T. Zhang, E. Faraggi, B. Xue, A. Dunker, V. Uversky, and Y. Zhou, "SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method." *J Biomol Struct Dyn*, vol. 29, no. 4, pp. 799 – 813, 2012.

[44] B. Xue, R. Dunbrack, R. Williams, A. Dunker, and V. Uversky, "PONDR-FIT: a meta-predictor of intrinsically disordered amino acids." *Biochim Biophys Acta*, vol. 1804, no. 4, pp. 996 – 1010, Apr 2010.