ON THE COMPLEMENTARITY OF THE CONSENSUS-BASED DISORDER PREDICTION

ZHENLING PENG AND LUKASZ KURGAN
Electrical and Computer Engineering Department, University of Alberta, Edmonton, AB, Canada
Emails: zhenling@ualberta.ca, lkurgan@ece.ualberta.ca

Intrinsic disorder in proteins plays important roles in transcriptional regulation, translation, and cellular signal transduction. The experimental annotation of the disorder lags behind the rapidly accumulating number of known protein chains, which motivates the development of computational predictors of disorder. Some of these methods address predictions of certain types/flavors of the disorder and recent years show that consensus-based predictors provide a viable way to improve predictive performance. However, the selection of the base predictors in a given consensus is usually performed in an ad-hock manner, based on their availability and with a premise that more is better. We perform first-of-its-kind investigation that analyzes complementarity among a dozen recent predictors to identify characteristics of (future) predictors that would lead to further consensus-based improvements in the predictive quality. The complementarity of a given set of three base predictors is expressed by the differences in their predictions when compared with each other and with their majority vote consensus. We propose a regression-based model that quantifies/predicts quality of the majority-vote consensus of a given triplet of predictors based on their individual predictive performance and their complementarity measured at the residue and the disorder segment levels. Our model shows that improved performance is associated with higher (lower) similarity between the three base predictors at the residue (segment) level and to their consensus prediction at the segment (residue) level. We also show that better consensuses utilize higher quality base methods. We use our model to predict the best-performing consensus on an independent test dataset and our empirical evaluation shows that this consensus outperforms individual methods and other consensus-based predictors based on the area under the ROC curve measure. Our study provides insights that could lead to the development of a new generation of the consensus-based disorder predictors.

1. Introduction

The intrinsically disordered proteins (IDPs) are characterized by the lack of stable tertiary structure when their isolated chains are under physiological conditions *in vitro*.[1] IDPs include random coil-like regions, partially folded or molten/pre-molten globule-like domains with poorly packed side chains, and dynamic structural ensembles.[2,3] They implement important functional roles in transcriptional regulation, translation, and cellular signal transduction,[4] and they are relatively common.[5] The prevalence of disorder was implicated in various human diseases and they were suggested as important targets for drug discovery.[6,7] The above motivates research towards improved understanding of the principles and mechanisms of IDPs. Some studies show that IDPs are characterized by relatively unique sequence signatures. For example, they often have

a low content of bulky hydrophobic amino acids and a high proportion of polar and charged residues, a low content of predicted secondary structure, low complexity, and unique evolutionary and solvent accessibility profiles.[8-13] This implies that the disorder is predictable from the sequence. Therefore, a number of computational predictors were developed over the past decade. These efforts intensified after the disorder prediction was introduced into the biannual CASP experiments in 2002.[14-16] The disorder predictors are categorized into four types:[17]

1. *propensity-based* methods based on relative propensity of amino acids to form disorder/ordered regions: GlobPlot,[18] FoldIndex,[19] IUPred,[20] and Ucon;[21]
2. *machine learning-based* predictors: DISOPRED2,[22] DISpro,[23] RONN,[24] ProfBval,[25,26] PONDR predictors,[9,14,27-30] PreDisorder,[23,32] NORSnet,[21] DisEMBL,[18] and Spritz;[31]
3. *consensus-based* methods that combine predictions from multiple base predictors: metaPrDOS,[33] GS-MetaServer,[34] MD,[35] PONDR-FIT,[36] and MFDp;[17]
4. *structural models-based* approaches that make use of predicted tertiary structure models: PrDOS [37] and DISOCLUST.[38]

The results from a recent comparative review [39] and the CASP8 competition,[16] demonstrate that the consensus-based methods, such as the GS-MetaServer,[34] MD,[35] and MFDp,[17] generally outperform other methods. This is perhaps due to the fact that certain disorder predictors target specific flavors/types of the disorder and thus they are suitable to form a well-performing consensus.[35] One of the desired characteristics of the consensus-based methods is that the base predictors that are being combined should be complementary to each other.[17,35,40] However, the existing methods select the base predictors in an ad hock manner: based on their availability, differences in the architecture of the base methods and/or their targeted type of the disorder, and utilizing a premise that inclusion of a larger number of base predictors is beneficial.[17,34-36] We perform a first-of-its-kind empirical investigation into the complementarity among disorder predictors. For a given triplet of methods, we measure complementarity based on the differences in their predictions, by comparing them with each other and to their majority vote consensus (MVC). We build a regression-based model that predicts the quality of the MVC using the information concerning complementarity measured at the residue- and segment-level and the predictive performance of the base methods. We empirically demonstrate that a proper selection of the base methods leads to the improvements in the disorder predictions, when compared with the current single and consensus-based disorder predictors.

2. Methods

2.1. *Considered disorder predictors*

We consider 17 disorder predictors, which are accessible to the end user via web server or a standalone program, including 16 that were discussed in a recent review [39] and DRIPPRED. Since our study requires that a given predictor outputs real-value disorder propensity, we could not include the predictors that only produce the binary outputs, namely GlobPlot,[18] FoldIndex,[19] DisEMBL,[18] and Spritz.[31] Furthermore, the IUPred method is used in both of its modes, one for the prediction of short disordered segments, named IUPredS, and the other for long segments, named IUPredL. Consequently, the 14 disorder predictors are considered; see Table 1.

Table 1. Summary of disorder predictors included in this study. Methods are sorted by the year of publication. Acronyms in column "Type": propensity-based (PB), machine learning-based (MLB), consensus-based (CB), and structural models-based (SB); in column "Algorithm": support vector machine (SVM), neural network (NN), scoring function (SF) and self-organizing map (SOM).

Prediction method				URL
Name	Published	Type	Algorithm	
MFDp	2010	CB	SVM	http://biomine-ws.ece.ualberta.ca/MFDp.html
PONDR-FIT	2010	CB	NN	http://www.disprot.org/pondr-fit.php
MD	2009	CB	NN	https://rostlab.org/owiki/index.php/Metadisorder
DISOCLUST	2008	SB	SF	http://www.reading.ac.uk/bioinf/DISOclust/
NORSnet	2007	MLB	NN	https://rostlab.org/owiki/index.php/Norsnet
Ucon	2007	PB	SF	https://rostlab.org/owiki/index.php/UCON
ProfBval	2006	MLB	NN	https://rostlab.org/owiki/index.php/Profbval
VSL2B	2006	MLB	SVM	http://www.dabi.temple.edu/disprot/Predictors.html
DISpro	2005	MLB	NN	http://scratch.proteomics.ics.uci.edu/
IUPred	2005	PB	SF	http://iupred.enzim.hu/
RONN	2005	MLB	NN	http://www.strubi.ox.ac.uk/RONN
DISOPRED2	2004	MLB	SVM	http://bioinf.cs.ucl.ac.uk/disopred/
DRIPPRED	2004	PB	SOM	http://www.sbc.su.se/~maccallr/disorder/

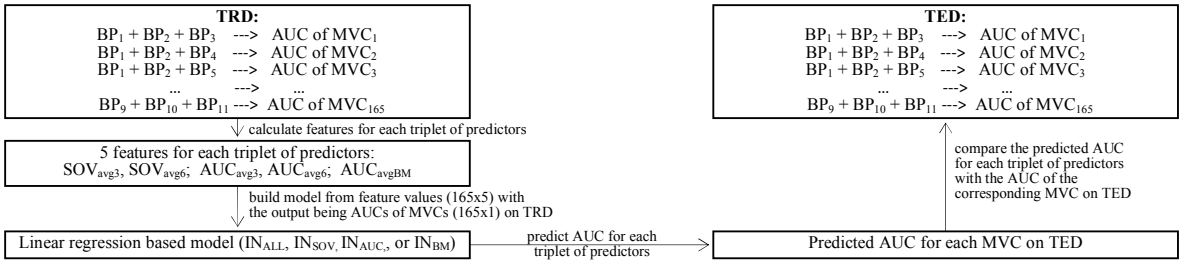

Figure 1. Overview of the prediction model. BP_i and *MVC* denote the i^{th} base predictor and the majority vote consensus, respectively. Sections 2.4 and 2.5 provide detailed explanations.

2.2. *Datasets*

We utilize a recently-built MxD dataset that was developed by Mizianty *et al.*,[17] and we further improve the disorder annotation using the SL dataset-based procedure.[41] The protein chains were originally collected from the Protein Data Bank (PDB)[42] and the release 4.9 of the DisProt database[43]. The resulting MxD_SL dataset includes 494 out of 514 proteins from the MxD; the remaining 20 chains could not be predicted by the MD and DISpro methods. The SL dataset-based procedure combines the disorder annotations from the DisProt with the disorder/order annotations based on the corresponding structural domains in PDB. Consequently, the DisProt derived chains include the DisProt's annotation of disorder, while the remaining residues that were not annotated in this database are annotated using the PDB domains. This means that some of these residues are left without any annotation (if they cannot be found in PDB); we did not use them to perform design and evaluations. However, we use all annotated residues, even if they are in the partially

annotated chains. The MxD_SL dataset includes 196,434 residues, where 51,733 are without annotation. Similarly as in CASP8, we discard the native disordered segments with 3 or fewer consecutive disordered residues.[16] The MxD_SL dataset is randomly partitioned into equal-size training dataset (TRD) and test dataset (TED), which have 23,406 and 23,074 disordered residues, respectively. Importantly, chains in MxD_SL are characterized by pair-wise sequence identity below 25%,[17] which means that TRD and TED are independent at the 25% similarity level. The datasets can be downloaded from http://biomine.ece.ualberta.ca/MVCdisorder/MVCdisorder.htm.

2.3. *Overview of the prediction model*

We investigate complementarity between disorder predictors based on a majority vote consensus (MVC). We combine three base predictors to model the relation between the complementarity in their predictions, which is measured using an approach described in section 2.4, and the predictive quality of the resulting MVC using linear regression on the TRD. Next, we apply regression to predict the best combination of methods on the TED, and we compare this consensus with the existing solutions (see Figure 1). We do not use cross-validation or another sampling since we do not perform parameterizations that could lead to overfitting. Furthermore, we have sufficient amount of data in the training dataset, and training and test datasets are independent.

2.4. *Inputs for the prediction model*

We assess the complementarity of the base predictors at the residue and disordered segment levels, i.e. we study the complementarity between the individual predicted propensities and between the binary predictions that form segments of predicted consecutive disordered residues:

1. The area under the ROC curve (AUC) [26] is usually used to examine the predicted residue-level propensities. For a pair of predictions, we quantify their residue-level complementarity by computing two AUC values between them, i.e., the first AUC when one prediction is assumed to be the true outcome and the other to be the prediction, and another AUC when the second prediction is assumed to be the true outcome. Consequently, higher AUC values indicate lower complementarity. Given three base methods, we compute the average of the six corresponding AUC values, and we call this input AUC_{avg6}. We also quantify complementarity between the three base predictions and their MVC. We calculate the average of the three AUC values when using the consensus as the true outcome, and we refer to this input as AUC_{avg3}.
2. The segment overlap (SOV)[45] measures the amount of overlap between the disordered segments.[39] Given three base methods, we compute average of the six SOV values (SOV_{avg6}), i.e., two SOV values for each pair of methods when one prediction is assumed to be the true outcome and the other to be the prediction, and when second prediction is assumed to be the true outcome. Similarly as for the AUC, we also compute the average SOV between the base predictions and their MVC (SOV_{avg3}). Higher SOV values correspond to predictions with more overlapping disordered segments, i.e., weaker complementarity at the segment-level.

The performance of the individual base methods is also likely to have impact on the predictive quality of their consensus. We quantify their performance based on the average of three AUC values (AUC_{avgBM}), where predictions from each method are compared against the native disorder.

We group these inputs into four sets: IN_{ALL} (with all 5 inputs), IN_{SOV} (SOV_{avg6} and SOV_{avg3}), IN_{AUC} (AUC_{avg6} and AUC_{avg3}), and IN_{BM} (AUC_{avgBM}), to investigate whether combining the 3 input types (segment-, residue-, and base method-based) provides a better model when compared with the use of the individual input types. The calculation of the AUC_{avgBM} involves the native annotations and we use AUC_{avgBM} values calculated on TRD to perform predictions on the TED.

2.5. *Majority vote and linear regression*

We normalize the propensities generated by the 14 considered methods so that they use the same threshold = 0.5 to binarize their predictions. Specifically, for a given method that uses threshold p, we linearly map the probability within the range [0, p) and [p, 1] to [0, 0.5) and [0.5, 1], respectively. We next build the MVC for every combination of three methods. Since MD,[35] MFDp,[17] and PONDR-FIT[36] are already consensuses, they are not utilized as the base methods. Consequently, using the remaining 11 methods we generate $(11\times10\times9)/6 = 165$ combinations. The predictions from the base methods are combined using majority vote; as shown in Table 2.
Following Figure 1, we build a linear regression-based model on TRD to express the relation between the AUC of the MVC and the complementarity and quality of the three base predictors, which is quantified using IN_{ALL} feature set. The model outputs real values that predict the AUC of a given MVC. The regression coefficients demonstrate the relation between each of the inputs and the predictive quality (expressed with AUC), and can be used to investigate how the residue- and segment-level complementarity among the base methods affects the performance of the consensus.

Table 2. Implementation of MVC with 3 base predictors BP_i where i=1,2,3. The outputs include the propensity calculated from normalized propensities $p(BP_i)$ generated by BP_i, and binary prediction, where D and O are disordered and ordered residues, respectively.

BP_1	binary	O	D	O	O	D	D	O	D
BP_2	binary	O	O	D	O	D	O	D	D
BP_3	binary	O	O	O	D	O	D	D	D
MVC	binary	O				D			
	probability	$\min_{i=1,2,3}\{p(BP_i)\}$				$\max_{i=1,2,3}\{p(BP_i)\}$			

2.6. *Evaluation criteria*

We compare AUC values predicted by the linear regression with AUCs of the MVC using Pearson Correlation Coefficient (*PCC*), Mean Absolute Error (*MAE*), and Mean Squared Error (*MSE*).

We use the regression to predict the top performing MVC on the TED. The quality of disorder predictions using this MVC, two runner-up predicted MVCs, and existing disorder predictors is evaluated using Matthews Correlation Coefficient (MCC), S_w, AUC, and SOV. The first 3 criteria, which were used in CASP,[15,16] assess the per-residue predictions; SOV evaluates prediction of disordered segments.[39] The binary per-residue predictions are assessed with MCC and S_w:

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

$$S_w = \frac{w_{disorder} \cdot TP - w_{order} \cdot FP + w_{order} \cdot TN - w_{disorder} \cdot FN}{w_{order} \cdot N_{disorder} + w_{disorder} \cdot N_{order}}$$

where *TP* is *true positive* (number of correctly predicted disordered residues), *FP* denotes *false positive* (number of ordered residues predicted as disordered), *TN* denotes *true negative* (number of correctly predicted ordered residues), *FN* stands for *false negative* (number of disordered residues predicted as ordered), $W_{disorder}$ (N_{order}) and W_{order} ($N_{disorder}$) are the percentages (numbers) of the ordered and the disordered residues, respectively. The S_w and MCC values range between -1 and 1; they are equal to zero when all residues are predicted as ordered or disordered and their higher values indicate better predictions.

The quality of the predicted per-residue propensities is assessed with the ROC curve. For each value of propensities *p* achieved by a given method (between 0 and 1), the residues with propensities ≥ *p* are set as disordered, and all other residues are set as ordered. Next, the *TP-rate* = *TP* / (*TP* + *FN*) and the *FP-rate* = *FP* / (*FP* + *TN*) are calculated and we use the area under the resulting ROC curve (*AUC*) to quantify the predictive quality.

The segment-level evaluation is based on the SOV,[39,45] which quantifies the overlap between the segments formed by the binary per-residue predictions and the native disorder segments. We compute the SOV values only for the chains with the complete native disorder annotation.

Table 3. The coefficients of the linear regression models that predict AUC of the majority-vote consensuses which are based on the 5 input features (IN_{ALL}), the 2 SOV-based inputs (IN_{SOV}), the 2 AUC-based inputs (IN_{AUC}), and the average AUC of the base methods (IN_{BM})

Model	SOV_{avg3}	SOV_{avg6}	AUC_{avg3}	AUC_{avg6}	AUC_{avgBM}	Constant
IN_{ALL}	0.013	-0.023	-0.089	0.024	0.109	0.823
IN_{SOV}	0.055	-0.037				0.835
IN_{AUC}			-0.051	0.070		0.837
IN_{BM}					0.039	0.820

3. Results

3.1. *Modeling the complementarity in disorder prediction*

We generate linear regression-based model on the training dataset (TRD), using the inputs defined in section 2.4 and following Figure 1. We compare the models based on all 5 inputs (IN_{ALL}) with the models based on the IN_{SOV}, IN_{AUC}, and IN_{BM} features; see Table 3. We observe that signs of the coefficients are consistent between different input sets, i.e., the SOV_{avg3}, AUC_{avg6}, and AUC_{avgBM} are positively correlated with the AUC of the MVCs while SOV_{avg6} and AUC_{avg3} are negatively correlated. The model reveals that the improved performance of a MVC is associated with: (1) higher quality of the base methods (positive coefficient for AUC_{avgBM}); (2) lower complementarity/higher similarity between the base predictors (positive coefficients for AUC_{avg6}) and their higher complementarity/lower similarity to the consensus prediction (negative coefficient for AUC_{avg3}) at the residue level; and (3) their lower complementarity/higher similarity to the consensus prediction (positive coefficient for SOV_{avg3}) and higher complementarity/lower similarity between them (negative coefficient for SOV_{avg6}) at the segment level.

Table 4. Comparison of the predictive quality of the regression models that utilize 5 input features (IN_{ALL}), the 2 SOV-based inputs (IN_{SOV}), the 2 AUC-based inputs (IN_{AUC}), and the average AUC of the base methods (IN_{BM}), and the random model on the training and the independent test sets.

Model	Evaluation on training dataset			Evaluation on test dataset		
	MAE	MSE	PCC	MAE	MSE	PCC
IN_{ALL}	0.007	0.105	0.813	0.007	0.112	0.785
IN_{SOV}	0.011	0.167	0.381	0.009	0.144	0.444
IN_{AUC}	0.010	0.158	0.476	0.008	0.133	0.571
IN_{BM}	0.009	0.145	0.597	0.008	0.125	0.640
random	0.019	0.290	0.008	0.018	0.282	0.009

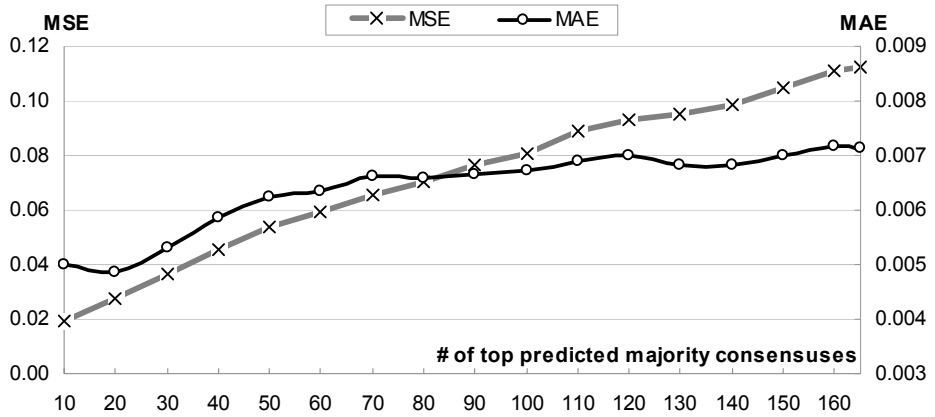

Figure 2. Relation between the number of evaluated consensuses that are sorted by their predicted AUC on TED (*x*-axis), and MSE and MAE (*y*-axis) which are based on the predictions from the regression that utilizes 5 inputs (IN_{ALL}) generated using TRD and tested on TED.

Next, we evaluate each regressions model (based on the IN_{ALL}, IN_{SOV}, IN_{AUC}, and IN_{BM} features) and compare them with a model that generates AUC values in the actual range of these values in the TRD at random. The random prediction is repeated 100 times and we report the average results; we use that as a baseline to examine the performance of our regression models. The predictive quality is measured by MAE, MSE, and PCC, on both TRD and the independent test dataset (TED); see Table 4. The four regression models outperform the random predictor by a wide margin. The model that utilizes the five inputs (IN_{ALL}) improves over the other regressions that use a subset of inputs, which demonstrates that all inputs contribute to the predictions. The IN_{ALL}-based model obtains the lowest MAE (0.007 on TRD and TED), lowest MSE (0.105 on TRD and 0.112 on TED) and the highest PCC (0.81 and 0.78 on TRD and TED, respectively). Moreover, the quality of the regression predictions is better for the consensuses with higher predicted AUCs and it progressively gets worse for the lower predicted AUCs; see Figure 2. The MAE and MSE are at about 0.005 and 0.02, respectively for the top 10 predicted consensuses, which indicates that our predictions are quite accurate when predicting the top-performing consensuses. The model successfully predicts the top-performing consensus, which consists of DISOCLUT, DISpro, and VSL2B, on the TED; the predicted and actual AUCs for this consensus are 0.871 and 0.872, respectively. To compare, the worst performing consensus with DRIPPRED,

NORSnet, and ProfBval, which obtains the AUC = 0.809, is predicted to obtain AUC = 0.819. The relatively high predictive quality of our regression model suggests that the observations made based on this model should be accurate.

For the top predicted MVC, the SOV_{avg3} and SOV_{avg6} values are 3.7% and 10.7% smaller than the average values of SOV_{avg3} (0.521) and SOV_{avg6} (0.546), respectively, which are calculated over all 165 MVCs; The AUC_{avg3}, AUC_{avg6}, and AUC_{avgBM} values are 8.4%, 18.2%, and 28.3% higher than the corresponding averages (0.582, 0.510, and 0.59), respectively. The feature values were normalized to make these calculations, i.e., the original feature x was normalized as $(x-a)/(b-a)$, where a and b are the minimum and maximum value of x on TRD, respectively; if (for TED) $x<a$, we set $x=a$; if $x>b$, we set $x=b$. The above differences indicate that the corresponding three base predictors have high predictive quality (high AUC_{avgBM}), above average complementarity at the segment level (low SOV_{avg}), and below average complementarity at the residue level (high AUC_{avg}). This shows that complementarity between base predictors affects performance of their majority vote consensus. The 2nd and 3rd top predicted MVCs have all feature values above the corresponding averages, with their AUC_{avgBM} values higher by wide, 39% and 41%, margins, respectively. This indicates that the high quality of these MVCs stems mostly from the good predictive performance of their base methods.

We also investigate whether the quality of a given MVC depends on the types of its base predictors (see Table 1). The top predicted MVC includes structural models-based and two different machine learning-based predictors, one that utilizes neural networks (NNs) and another that uses Support Vector Machines (SVMs). The 2nd top predicted MVC includes three different machine learning-based predictors, two of which use SVMs and one based on NN. The 3rd top predicted consensus uses one propensity-based method and two machine learning-based predictors. Our analysis suggests that use of diverse types of base predictors may have an effect on the complementarity and may lead to an improved consensus.

3.2. *Evaluation of consensus-based and individual disorder predictors*

We evaluate disorder predictions, which were generated by the 14 considered predictors and using the top 3 MVCs predicted by the linear regression (MVCs with the highest predicted AUCs), on the independent test dataset (TED), see Table 5. MD[35] and MFDp[17] are based on a consensus approach; however, both of these methods utilize complex, non-linear predictors to combine the outputs of base classifiers with additional inputs, including multiple alignment, predicted flexibility, secondary structure, solvent accessibility, etc. Since such complex consensuses were shown to outperform the simple MVC,[35] Table 5 includes results of the MVCs which use the same base methods as in MD and MFDp to allow for a fair comparison to the MVCs that were predicted with the regression. The MD-based MVC includes NORSnet, ProfBval, Ucon, and DISOPRED2, while the MFDp-based MVC incorporates DISOPRED2, IUPredL, IUPredS, and DISOCLUST. In both cases the consensus has 4 methods, and thus we resolve the ties by using the prediction from the DISOPRED2 that has the highest AUC on the TRD. We also rebuilt the MFDp method by substituting the original base predictors with the methods included in the top MVC predicted by the regression, i.e., DISOCLUST, DISpro, and VSL2B. The other consensus method is PONDR-

FIT,[36] which uses a relatively simple approach to combine the predictions from six base methods and does not use additional inputs. Thus, we decided to directly compare with this method, without building the MVC. We note that our 3rd best predicted MVC corresponds to the combination of three methods (DISOPRED2, IUPredL, and VSL2B) that obtain the highest AUC and MCC values on TRD.

Table 5. Predictive quality, measured with MCC, S_w, and AUC on TED for the 14 considered disorder predictors and the top 3 majority-vote consensuses (MVCs) predicted by the linear regression. The consensus-based predictors include MFDp, MD, PONDR-FIT, and the MVCs based on our top 3 predictions and using the base methods from MFDp and MD. The consensus-based and the single predictors are sorted by their AUC in the descending order, respectively. The SOV is calculated based on 176 proteins in TED that are fully annotated.

Type	Predictor	MCC	S_w	SOV	AUC
Consensus predictors	MFDp rebuild using the top predicted base methods	0.637	0.636	0.620	0.8895
	MFDp	0.633	0.625	0.638	0.8875
	MD	0.625	0.596	0.478	0.8784
	top predicted MVC (DISOCLUST, DISpro, VSL2B)	0.613	0.602	0.609	0.8716
	3rd predicted MVC (DISOPRED2, IUPredL, VSL2B)	0.634	0.604	0.558	0.8633
	2nd predicted MVC (DISOPRED2, DISpro, VSL2B)	0.629	0.568	0.524	0.8607
	MVC using 4 base methods from MFDp	0.627	0.600	0.544	0.8606
	MVC using 4 base methods from MD	0.618	0.582	0.462	0.8569
	PONDR-FIT	0.592	0.556	0.590	0.8549
Single predictors	DISOPRED2	0.608	0.570	0.514	0.8574
	VSL2B	0.557	0.576	0.644	0.8568
	DISOCLUST	0.504	0.533	0.608	0.8497
	DISpro	0.458	0.304	0.361	0.8391
	IUPredL	0.575	0.514	0.341	0.8390
	IUPredS	0.544	0.472	0.511	0.8314
	RONN	0.524	0.517	0.517	0.8190
	NORSnet	0.569	0.496	0.184	0.8146
	DRIPPRED	0.480	0.470	0.500	0.7874
	Ucon	0.423	0.405	0.297	0.7844
	ProfBval	0.273	0.279	0.527	0.7353

Table 5 shows that MFDp and MD have the highest AUC, but this is since they use the complex consensuses. The MVCs built using their base predictors have AUC equal 0.86 and 0.85, respectively, which is lower than the AUCs of the top three regression predicted MCVs. Among these three predictions, the highest AUC = 0.87 is obtained by our top prediction. Moreover, the rebuilt MFDp that uses the base methods from the top prediction slightly improves over the original MFDp in spite of the fact that it uses fewer base methods (3 vs. 4). Importantly, the top three predicted MCVs improve over all other methods, including the simple consensus-based PONDR-FIT and the 11 modern disorder predictors. The AUC is higher by 1.5% when comparing the top predicted MVC and the best result from among these 12 methods. Similarly, the MCC and S_w values are higher and they equal 0.61 and 0.6 for the top prediction generated by the regression.

However, the top 3 predicted ensembles are outperformed by VSL2B in the disordered segment prediction measured with SOV; this is consistent with prior work that shows that segment predictions from VSL2B outperform other methods, including consensus approaches.[39] At the same time, our top prediction improves over VSL2B by 5.5% in MCC, 2% in S_w, and 1.5% in AUC. Moreover, the segment predictions from our top predicted consensus are competitive or better than the predictions from the remaining 11 methods, MD, and the MVCs based on the base predictors from the MD and from the MFDp; this is likely since our consensus includes VSL2B.

The top three predicted MVCs are better than their base methods by a relatively wide margin. Table 5 shows that the top predicted ensemble improves the MCC, S_w, and AUC by at least 5.6%, 2.7%, and 1.5%, respectively, when compared to the best base method. The 3rd best predicted ensemble is better by 2.7% in MCC, 2.8% in S_w, and 0.6% in AUC.

Overall, our analysis shows that the top MVC predicted using linear regression outperforms or is at least competitive when compared with the existing single and consensus-based predictors.

```
Native      ------------------*--------------*-----------------------------*--DDDDDDDDDDDDD-------*--------DDDDDDDDD
Top1        DDDDD-------------*---DDDDD------*-----------------------------*---DDDDDDDDDDDDDDDDD---*-DDDDDDDDDDDDDDDDD
VSL2B       DDDDD-------DDDDD-*---DDDDD------*-----------------------------*---DDDDDDDDDDDDDDDDDD-*----------DDDDDDD
DISOCL      DDDDDDDDDDD-------*-DDDDDDDDDDDD-*-D----DDDDDDDDDDDDDDDDDDDDDD-*-DDDDDDDDDDDDDDDDDD---*-DDDDDDDDDDDDDDDDD
DISpro      DD----------------*--------------*-----------------------------*----------------------*-DDDDDDDDDDDDDDDDD
MVC_MD      ------------------*--------------*-----------------------------*---------DDDDDD-------*-----------------
MVC_MFDp    DDDD---D----------*--------------*-----------------------------*---------DDDDDD-------*---------DDDDDDDD
PONDRF      DDDDDDDDDDDDDDDDD--*--------------*-----------------------------*----DDDDD-------------*---------DDDDDDDD
```

Figure 3. Comparison of the predictions for the Glutamyl-Q tRNA(Asp) synthetase protein (PDB id: 1NZJA) based on the top MVC predicted by the regression (Top1 line), its base methods DISOCLUST (DISOCL line), DISpro, and VSL2B, the MVC using the base methods from MD (MVC_{MD} line) and from MFDp (MVC_{MFDp} line), and PONDR-FIT (PONDRF line). 'D' and '–' denote the disordered and ordered residues, respectively, and '*' stands for a segment of ordered residues. The first line shows the native annotation.

3.3. *Case study*

We use the Glutamyl-Q tRNA(Asp) synthetase protein (PDB id: 1NZJA) from the TED that has two relatively short disordered segments, including one at the C-terminus and one inner-chain segment, as a case study to compare disorder predictions. We compare predictions from the top MVC predicted by the regression, its three base methods: DISOCLUST, DISpro, and VSL2B, and based on the three most relevant other consensuses: the MVCs using the base methods from MD and from MFDp, and PONDR-FIT; see Figure 3. Majority of the predictors, except for the DISpro and the MVC using the base methods from MD, find both native disordered segments; however, most of the methods find additional segments towards and at the N-terminus. Our top MVC reduces the over-prediction of the disorder when compared with two of its base methods, VSL2B and DISOCLUST, and finds one native segments that was missed by the third base method DISpro. The MVC using the base methods from MFDp under-predicts the disorder, especially the inner segment, and still produces disorder predictions at the N-terminus. PONDR-FIT also under-predicts the inner-chain segment and generates a relatively long disordered segment at the N-terminus. Although the predictions in this case study should not be assumed typical, we conclude that they demonstrate that the proper selection of the base methods can result in improvements.

4. Conclusion

Our study provides insights that could help in development of a new generation of consensus-based disorder predictors. We use linear regression to model relation between certain aspects that quantify complementarity between the base methods and the predictive quality of the resulting majority-vote based consensus. Our modeling reveals that the complementarity has to be evaluated at the residue and the disordered segment levels, i.e., models that ignore one of these aspects are shown to provide inferior results. The model shows that improved predictive performance is associated with inclusion of accurate base predictors that are similar to each other at the residue level and which complement each other at the segment level, and which complement the consensus at the residue level. We also observe that use of different types of base predictors to implement the consensus seems to be beneficial. We empirically demonstrate the top majority-based consensuses predicted by our model on an independent test dataset outperform existing predictors, including consensus-based methods, which suggests that our observations have practical value.

Acknowledgments

This work was supported by the Alberta Innovates Scholarship in Omics to ZP and the NSERC Discovery grant to LK. We thank Marcin Mizianty for help with the datasets and MFDp method.

References

1. V. N. Uversky, A. K. Dunker, *Biochim Biophys Acta* **1804**(6), 1231(2010).
2. A. K. Dunker, J. D. Lawson, C. J. Brown, et al., *J Mol Graph* **19**(1), 26 (2001).
3. V. N. Uversky, *Prot Sci* **11**, 739 (2002).
4. A. K. Dunker, C. J. Oldfield, J. W. Meng, P. Romero, J. Y. Yang, J. W. Chen, V. Vacic, Z. Obradovic and V. N. Uversky, *BMC Genomics* **9**(S2), S1 (2008).
5. A. K. Dunker, Z. Obradovic, P. Romero, E. C. Garner and C. J. Brown, *Genome Inform Ser Workshop Genome Inform* **11**, 161 (2000).
6. V. N. Uversky, C. J. Oldfield, U. Midic, H. Xie, B. Xue, S. Vucetic, L. M. Iakoucheva, Z. Obradovic and A. K. Dunker, *BMC Genomics* **10**(S1), S7 (2009).
7. Y. Cheng, T. LeGall, C. J. Oldfield, J. P. Mueller, Y. Y. Van, P. Romero, M. S. Cortese, V. N. Uversky and A. K. Dunker, *Trends Biotechnol* 24(10):435 (2006).
8. V. N. Uversky, J. R. Gillespie and A. L. Fink, *Proteins* **41**, 415 (2000).
9. P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown and A. K. Dunker, *Proteins* **42**, 38 (2001).
10. J. Liu, H. Tan and B. Rost, *J Mol Biol* **322**, 53 (2002).
11. H. J. Dyson and P. E. Wright, *Nat. Rev. Mol. Cell Bio*. **6**, 197 (2005).
12. Z. Dosztányi, B. Mészáros and I. Simon, *Brief. Bioinform* **11**(2), 225 (2010).
13. M. J. Mizianty, T. Zhang, B. Xue, Y. Zhou, A. Keith Dunker, V.N. Uversky and L. Kurgan, *BMC Bioinformatics* **12**, 245 (2011).
14. S. Vucetic, C. J. Brown, A. K. Dunker and Z. Obradovic, *Proteins* **52**, 573 (2003).
15. L. Bordoli, F. Kiefer and T. Schwede, *Proteins* **69**(S8), 129 (2007).

16. O. Noivirt-Brik, J. Prilusky and J. L. Sussman, *Proteins* 77(S9), 210 (2009).
17. M. J. Mizianty, W. Stach, K. Chen, K. D. Kedarisetti, F. M. Disfani and L. Kurgan, *Bioinformatics* **26**(18), i489 (2010).
18. R. Linding, L. J. Jensen, F. Diella, P. Bork, T. J. Gibson and R. B. Russell, *Structure* **11**, 1453 (2003).
19. J. Prilusky, C. E. Felder, T. Zeev-Ben-Mordehai, E. H. Rydberg, O. Man, J. S. Beckmann, I. Silman and J. L. Sussman, *Bioinformatics* **21**, 3435 (2005).
20. Z. Dosztányi, V. Csizmok, P. Tompa and I. Simon, *Bioinformatics* **21**, 3433 (2005).
21. A. Schlessinger, J. Liu and B. Rost, *PLoS Comput Biol* **3**, e140 (2007).
22. J. J. Ward, L. J. McGuffin and Bryson K, *Bioinformatics* 20, 2138 (2004).
23. J. Cheng, M. Sweredoski and P. Baldi, *Data Min Knowl Disc* **11**(3), 213 (2005).
24. Z. R. Yang, R. Thomson, P. McNeil and R. M. Esnouf, *Bioinformatics* 21, 3369 (2005).
25. A. Schlessinger and B. Rost, *Proteins* **61**(1), 115 (2005).
26. A. Schlessinger, G. Yachdav and B. Rost, *Bioinformatics* 22, 891 (2006).
27. Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac, C. J. Brown and A. K. Dunker, *Proteins* 53(S6), 566 (2003).
28. K. Peng, S. Vucetic, P. Radivojac, C. J. Brown, A. K. Dunker and Z. Obradovic, *J Bioinform Comput Biol* 3, 35 (2005).
29. Z. Obradovic, K. Peng, S. Vucetic, P. Radivojac and A. K. Dunker, *Proteins* 61(S7), 176 (2005).
30. K. Peng, P. Radivojac, S. Vucetic, A. K. Dunker and Z. Obradovic, *BMC Bioinformatics* **7**, 208 (2006).
31. A. Vullo, O. Bortolami, G. Pollastri and S. C. Tosatto, *Nucleic Acids Res* **34,** W164 (2006).
32. X. Deng, J. Eickholt and J. Cheng, *BMC Bioinformatics* 10, 436 (2009).
33. T. Ishida and K. Kinoshita, *Bioinformatics* 24, 1344 (2008).
34. M.A. Kurowski and J.M. Bujnicki, *Nucleic Acids Res* 31, 3305-3307 (2003).
35. A. Schlessinger, M. Punta, G. Yachdav, L. Kajan and B. Rost, *PLoS One* **4**, e4433 (2009).
36. B. Xue, R. L. Dunbrack, R. W. Williams, A. K. Dunker and V. N. Uversky, *Biochim Biophys Acta* **1804**, 996 (2010).
37. T. Ishida and K. Kinoshita, Nucleic Acids T. Ishida and K. Kinoshita, *Nucleic Acids Res* 35,W460 (2007).
38. L. J. McGuffin, *Bioinformatics* 24, 1798 (2008).
39. Z. L. Peng and L. Kurgan, *Curr. Prot. Pept. Sci*., accepted on October 2010.
40. C. J. Oldfield, *Biochemistry* 44, 1989 (2005).
41. F. L. Sirota, H. S. Ooi, T. Gattermayer, G. Schneider, F. Eisenhaber and S. Maurer-Stroh, *BMC Genomics* 11(S1), S15 (2010).
42. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, *Nucleic Acids Res*. **28**, 235-242 (2000).
43. M. Sickmeier, J.A. Hamilton, T. LeGall, V. Vacic, M.S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V.N. Uversky, et al., *Nucleic Acids Res*. **35**, D786-793 (2007).
44. T. Fawcett, *Pattern Recogn Lett* 27, 861 (2006).
45. A. Zemla, C. Venclovas, K. Fidelis and B. Rost, *Proteins* **34**(2), 220 (1999).