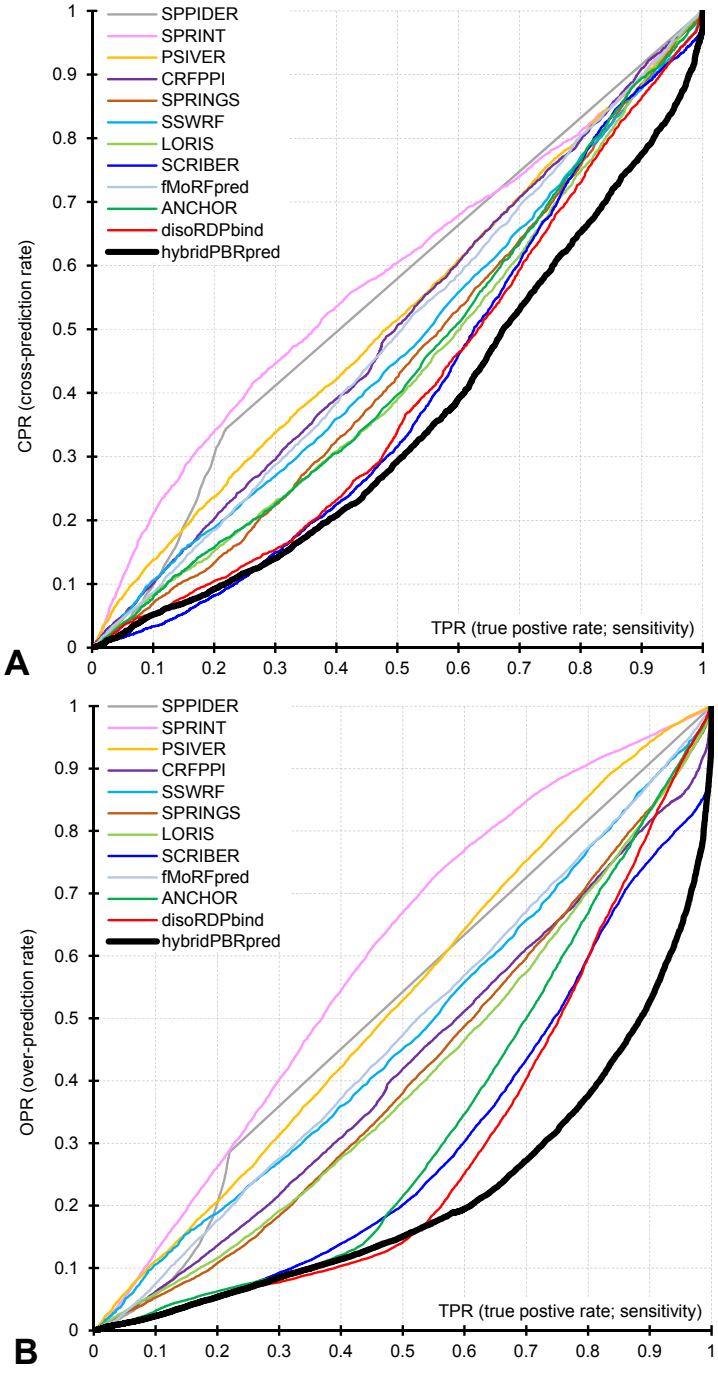
Supplement

Prediction of protein-binding residues: dichotomy of sequence-based methods developed using structured complexes vs. disordered proteins

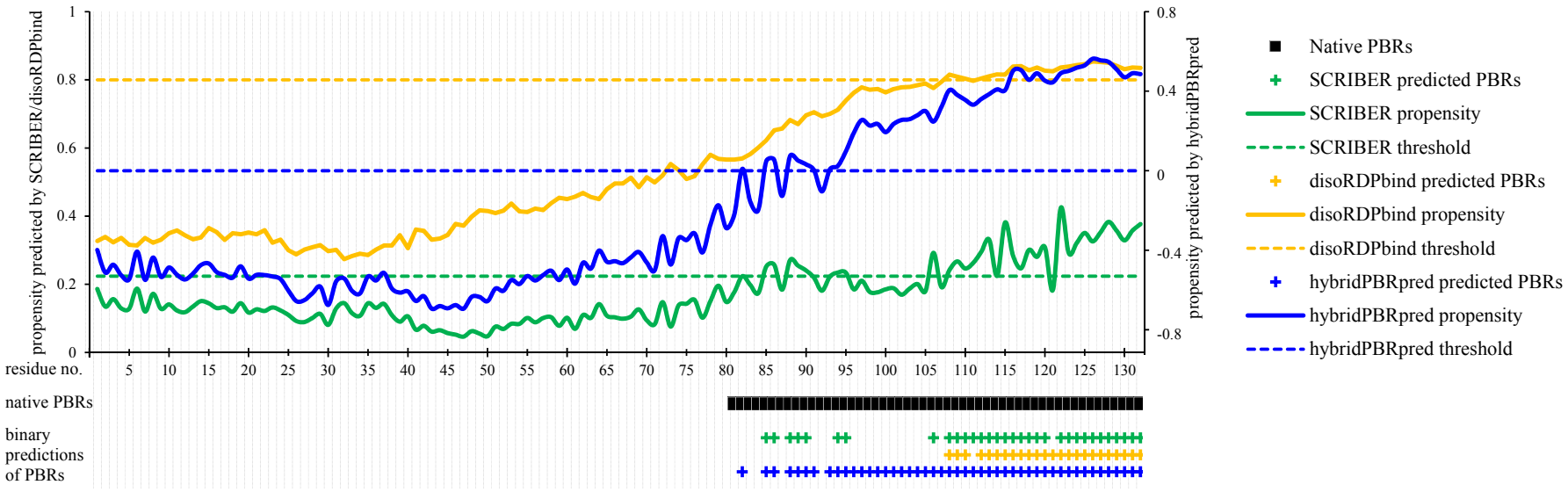
Jian Zhang[1], Sina Ghadermarzi[2], and Lukasz Kurgan[2*]

[1]School of Computer and Information Technology, Xinyang Normal University, Xinyang, China, 464000; [2]Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA, 23284

*corresponding author: phone: +1-804-827-3986; fax: +1-804-828-2771; email: lkurgan@vcu.edu



Supplementary Figure S1. The cross-prediction curves (panel A) and the over-prediction curves (panel B) computed on the entire benchmark dataset. The cross-predictions consider the residues that interact with all non-protein partners (nucleic acids and small ligands).



Supplementary Figure S2. Comparison of predictions generated by hybridPBRpred, SCRIBER and disoRDPbind for art v 1 protein, a pollen allergen from *Artemisia vulgaris* (UniProt ID: Q84ZX5; DisProt ID: DP00807). The solid lines at the top represent the putative propensities for protein binding generated by SCRIBER (in green), disoRDPbind (orange) and hybridPBRpred (blue). The propensities are converted into the binary predictions (PBRs vs. non-PBRs) using the predictor specific color-coded thresholds represented by the dashed horizontal lines. The binary predictions are shown using the color-coded horizontal lines at the bottom of the figure. The black markers denote the native annotations of the protein-binding residues.

Supplementary Table S1. The cross-prediction rates (AUCPC) and over-prediction rates (AUOPC) for the 11 predictors of PBRs and the new hybridPBRpred. We note that lower values of these metrics indicate stronger predictive performance. We test robustness of predictions across diverse protein sets by bootstrapping 10 sets of 50% randomly selected proteins from a given dataset; we report the corresponding averages ± standard deviations. For each dataset, we evaluate statistical significance of differences between the predictor shown in bold font (the best overall predictor) and the other predictors based on the 10 bootstrapped tests; ++ and + mean that the best predictor is significantly better with p-value < 0.001 and p-value < 0.05, respectively; = means that the difference is not significant (p-value ≥ 0.05); – – and – mean that the best predictor is significantly worse with p-value < 0.001 and p-value < 0.05, respectively.

Dataset	Type of Methods	Predictors	AUOPC	AUCPC residues binding non-protein partners (nucleic acids and small ligands)	$AUCPC_{NA}$ nucleic acid-binding residues	$AUCPC_{SL}$ small ligand binding residues
Structure-annotated proteins	Trained on proteins annotated from structure	SPPIDER	0.486±0.012 ++	0.626±0.011 ++	0.605±0.039 ++	0.632±0.012 ++
		SPRINT	0.431±0.014 ++	0.684±0.021 ++	0.537±0.040 ++	0.706±0.015 ++
		PSIVER	0.441±0.010 ++	0.568±0.024 ++	0.663±0.109 ++	0.557±0.030 ++
		CRF-PPI	0.320±0.010 ++	0.443±0.024 ++	0.479±0.044 ++	0.444±0.024 ++
		SSWRF	0.329±0.012 ++	0.439±0.030 ++	0.500±0.059 ++	0.436±0.030 ++
		SPRINGS	0.392±0.007 ++	0.522±0.013 ++	0.492±0.088 ++	0.527±0.012 ++
		LORIS	0.363±0.010 ++	0.459±0.019 ++	0.461±0.083 ++	0.461±0.020 ++
		SCRIBER	**0.278±0.016**	**0.265±0.013**	**0.161±0.038**	**0.279±0.016**
	Trained on proteins annotated with disordered PBRs	fMoRFpred	0.498±0.010 ++	0.447±0.018 ++	0.437±0.038 ++	0.446±0.018 ++
		ANCHOR	0.524±0.024 ++	0.497±0.046 ++	0.597±0.100 ++	0.485±0.038 ++
		disoRDPbind	0.477±0.028 ++	0.465±0.040 ++	0.507±0.123 ++	0.458±0.035 ++
	Hybrid predictor	hybridPBRpred	0.294±0.021 =	0.283±0.018 +	0.217±0.088 =	0.292±0.017 =
Disorder-annotated proteins	Trained on proteins annotated from structure	SPPIDER	0.539±0.027 ++	0.508±0.031 ++	0.508±0.031 ++	0.537±0.076 ++
		SPRINT	0.654±0.035 ++	0.484±0.041 ++	0.455±0.041 +	0.661±0.082 ++
		PSIVER	0.580±0.040 ++	0.478±0.050 ++	0.447±0.045 =	0.645±0.045 ++
		CRF-PPI	0.470±0.038 ++	0.503±0.051 ++	0.488±0.054 +	0.589±0.041 ++
		SSWRF	0.453±0.043 ++	0.465±0.069 +	0.449±0.076 =	0.566±0.042 ++
		SPRINGS	0.425±0.063 ++	0.389±0.070 =	0.363±0.063 =	0.515±0.068 ++
		LORIS	0.429±0.058 ++	0.402±0.073 =	0.377±0.071 =	0.527±0.060 ++
		SCRIBER	0.272±0.051 +	0.454±0.090 +	0.464±0.094 =	0.422±0.095 +
	Trained on proteins annotated with disordered PBRs	fMoRFpred	0.458±0.013 ++	0.514±0.015 ++	0.521±0.013 ++	0.479±0.035 ++
		ANCHOR	0.265±0.029 +	0.445±0.038 +	0.464±0.044 +	0.311±0.095 +
		disoRDPbind	**0.212±0.033**	**0.381±0.052**	**0.402±0.062**	**0.227±0.116**
	Hybrid predictor	hybridPBRpred	0.192±0.017 =	0.386±0.066 =	0.406±0.082 =	0.276±0.090 =
All proteins	Trained on proteins annotated from structure	SPPIDER	0.525±0.019 ++	0.546±0.027 ++	0.506±0.030 =	0.620±0.048 ++
		SPRINT	0.593±0.038 ++	0.537±0.044 ++	0.406±0.043 -	0.752±0.062 ++
		PSIVER	0.536±0.030 ++	0.504±0.039 ++	0.432±0.038 =	0.625±0.037 ++
		CRF-PPI	0.417±0.029 ++	0.483±0.041 ++	0.461±0.051 =	0.527±0.033 ++
		SSWRF	0.406±0.030 ++	0.455±0.051 +	0.444±0.070 =	0.482±0.038 ++
		SPRINGS	0.410±0.040 ++	0.429±0.048 +	0.382±0.052 -	0.504±0.038 ++
		LORIS	0.403±0.036 ++	0.418±0.050 =	0.386±0.061 -	0.474±0.040 ++
		SCRIBER	0.293±0.041 ++	0.400±0.061 =	0.425±0.068 =	0.366±0.060 ++
	Trained on proteins annotated with disordered PBRs	fMoRFpred	0.472±0.006 ++	0.493±0.010 ++	0.526±0.009 +	0.441±0.009 ++
		ANCHOR	0.341±0.035 ++	0.467±0.041 ++	0.565±0.042 ++	0.310±0.068 +
		disoRDPbind	0.292±0.038 ++	0.422±0.049 +	0.525±0.033 +	0.272±0.028 +
	Hybrid predictor	**hybridPBRpred**	**0.211±0.023**	**0.376±0.037**	**0.469±0.057**	**0.228±0.047**

Supplementary Table S2. Predictive performance of several approaches used to combine the scores generated by the structure-based predictor (SCRIBER) and the disordered-based predictor (disoRDPbind) on the complete benchmark dataset. We test robustness of predictions across diverse protein sets by bootstrapping 10 sets of 50% randomly selected proteins from the benchmark dataset; we report the corresponding averages ± standard deviations. The binary predictions were generated from the propensity scores using threshold that ensures that the number of predicted PBRs equals to the number of native PBRs, allowing direct comparison between predictors. We evaluate statistical significance of differences between the best overall method shown in bold font (the scheme used in hybridPBRpred) and the other three approaches based on the 10 bootstrapped tests; ++ and + mean that the best predictor is significantly better with p-value < 0.001 and p-value < 0.05, respectively; = means that the difference is not significant (p-value ≥ 0.05); – – and – mean that the best predictor is significantly worse with p-value < 0.001 and p-value < 0.05, respectively.

Methodology to combine predictors	Sensitivity	Specificity	F1	MCC	AUC	AULCratio	AUPRC
Maximum of the normalized predictions	0.223±0.035 ++	0.873±0.022 ++	0.223±0.035 ++	0.096±0.031 ++	0.639±0.018 ++	1.977±0.422 ++	0.206±0.035 ++
Minimum of the normalized predictions	0.168±0.019 ++	0.863±0.025 ++	0.168±0.019 ++	0.031±0.038 ++	0.474±0.018 ++	1.744±0.424 ++	0.152±0.017 ++
Mean of the normalized predictions	0.254±0.037 ++	0.878±0.017 ++	0.254±0.037 ++	0.132±0.028 ++	0.646±0.029 ++	2.311±0.250 ++	0.225±0.036 ++
hybridPBRpred Maximum of the normalized predictions for residues which at least one method predicts as PBRs; otherwise mean of the normalized predictions	**0.567±0.053**	**0.812±0.017**	**0.418±0.053**	**0.309±0.049**	**0.779±0.023**	**3.314±0.490**	**0.322±0.062**