Analysis of Structured and Intrinsically Disordered Regions of Transmembrane Proteins

Bin Xue[1,2], Samy O. Meroueh[1,2], Vladimir N. Uversky[1-3], and A. Keith Dunker[1,2]

[1]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202, USA;

[2]Institute for Intrinsically Disordered Protein Research, Indiana University School of Medicine, Indianapolis, IN 46202, USA;

[3]Institute for Biological Instrumentation, Russian Academy of Sciences, 142290 Pushchino, Moscow Region, Russia;

Supplementary Material

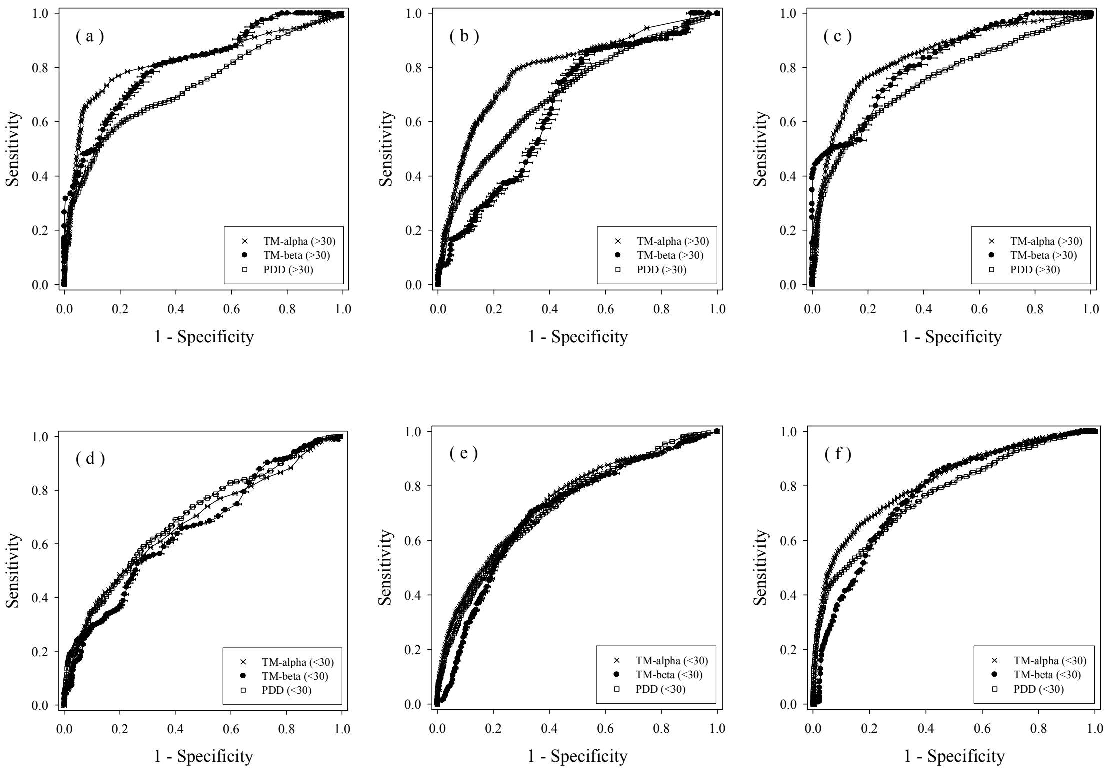

Figure S1. Balanced ROC curves for disorder predictions on subsets of PDD, TM-alpha, and TM-beta. All the sequences are classified into short and long subsets by a critical length of 30 amino acids. (a), (b), and (c) are for longer sequences while (d), (e), and (f) for shorter sequences. (a) and (d) are the results from VL3, (b) and (e) are from VLXT, (c) and (f) are the results of VSL2, respectively. The open squares represent data for PDD, the filled circles show data for TM-beta, and thin cross indicate results for TM-alpha. There are also 1000 times of bootstrapping to calculate the average and error.

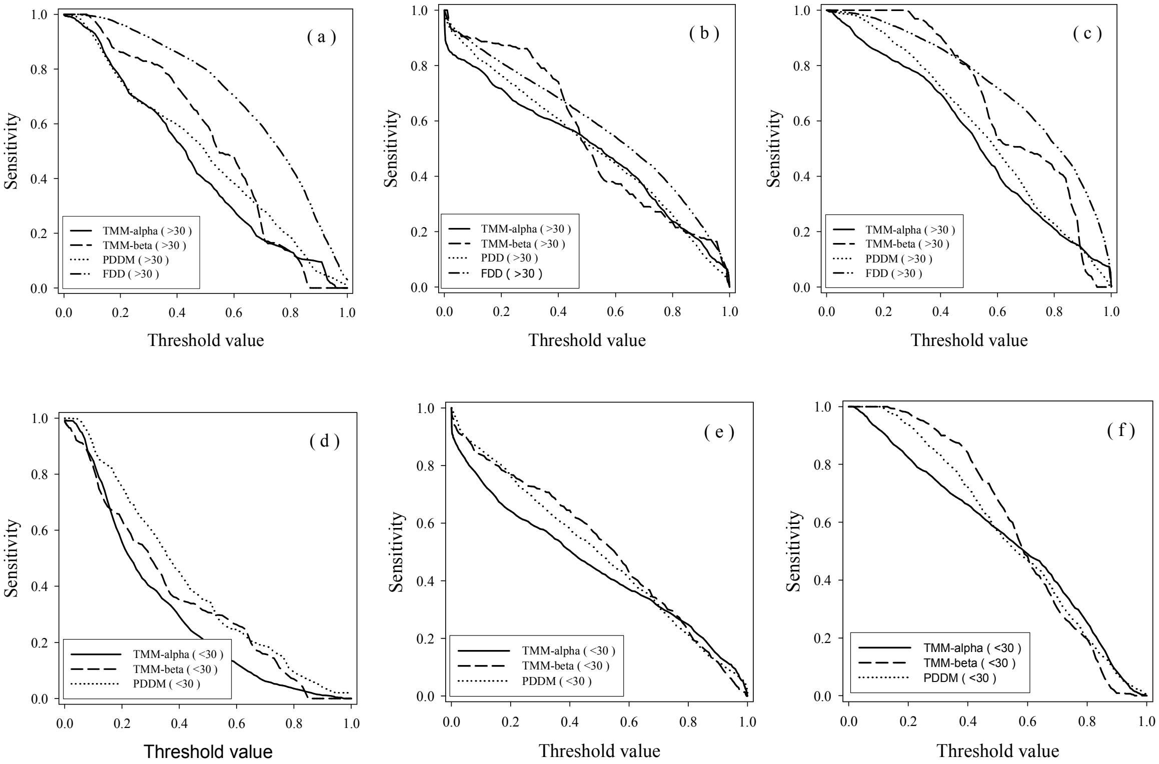

Figure S2. The change of disorder prediction accuracy of three predictors in the subsets as a function of the threshold value for disordered predictions. (a), (b) and (c) are the results of VL3, VLXT and VSL2 for the longer subsets of TM-alpha, TM-beta, and PDD, respectively. (d), (e) and (f) are of VL3, VLXT, and VSL2, for the shorter ones, accordingly.

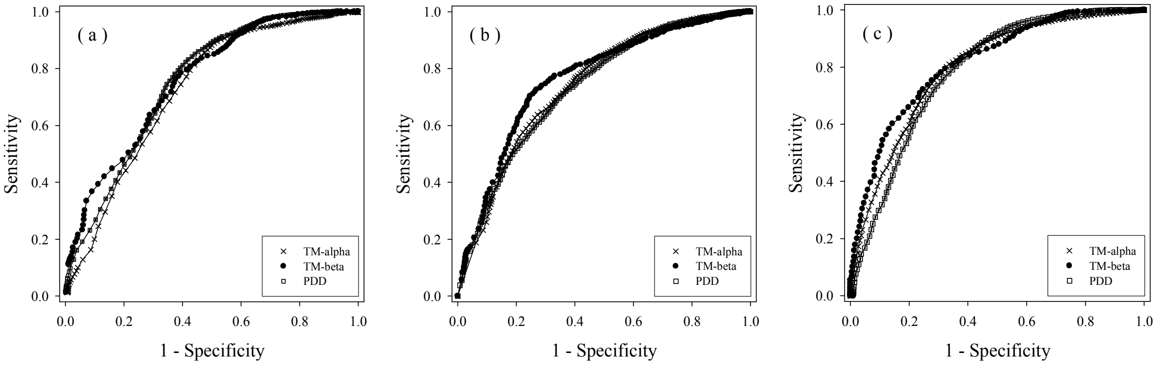

Figure S3. Balanced ROC plots for order predictions. (a), (b), and (c) are for VL3, VLXT, and VSL2, respectively. The blank squares, filled circles, and thin crosses are calculated from PDD, TM-beta, and TM-alpha, accordingly. The bars on each symbol are the statistical error of specificity for that point from 1000 times of bootstrapping sampling while the symbol itself is the bootstrapping average.

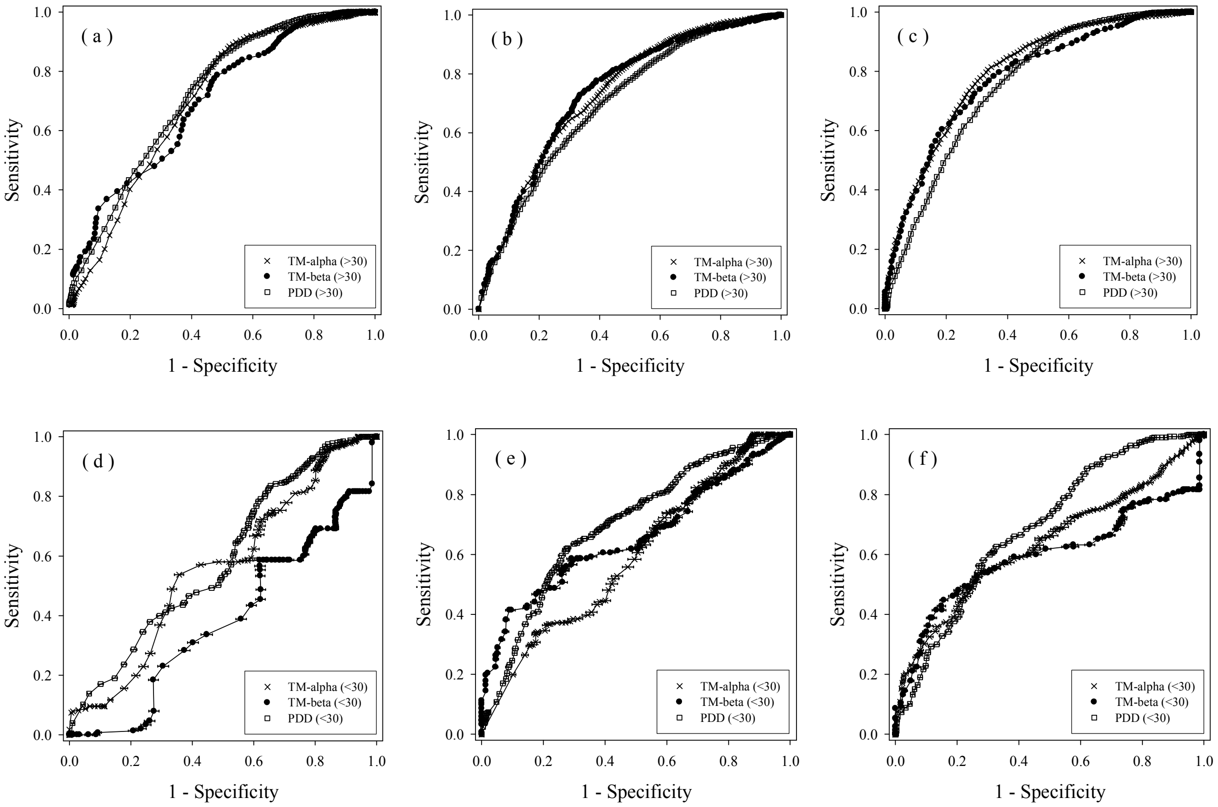

Figure S4. Balanced ROC of order predictions for subsets of PDD, TM-alpha, and TM-beta, of which all the sequences are classified into short and long subsets by a critical length of 30 amino acids. (a), (b), and (c) are for short sequences, while (d), (e), and (f) for longer ones. (a) and (d) are the results from VL3, (b) and (e) from VLXT, (c) and (f) from VSL2, respectively. The blank squares are results for PDD, filled circles for TM-beta, and thin cross for TM-alpha. The symbol is the average over 1000 times of bootstrapping and the bars on each symbol are the statistical error from the bootstrapping.

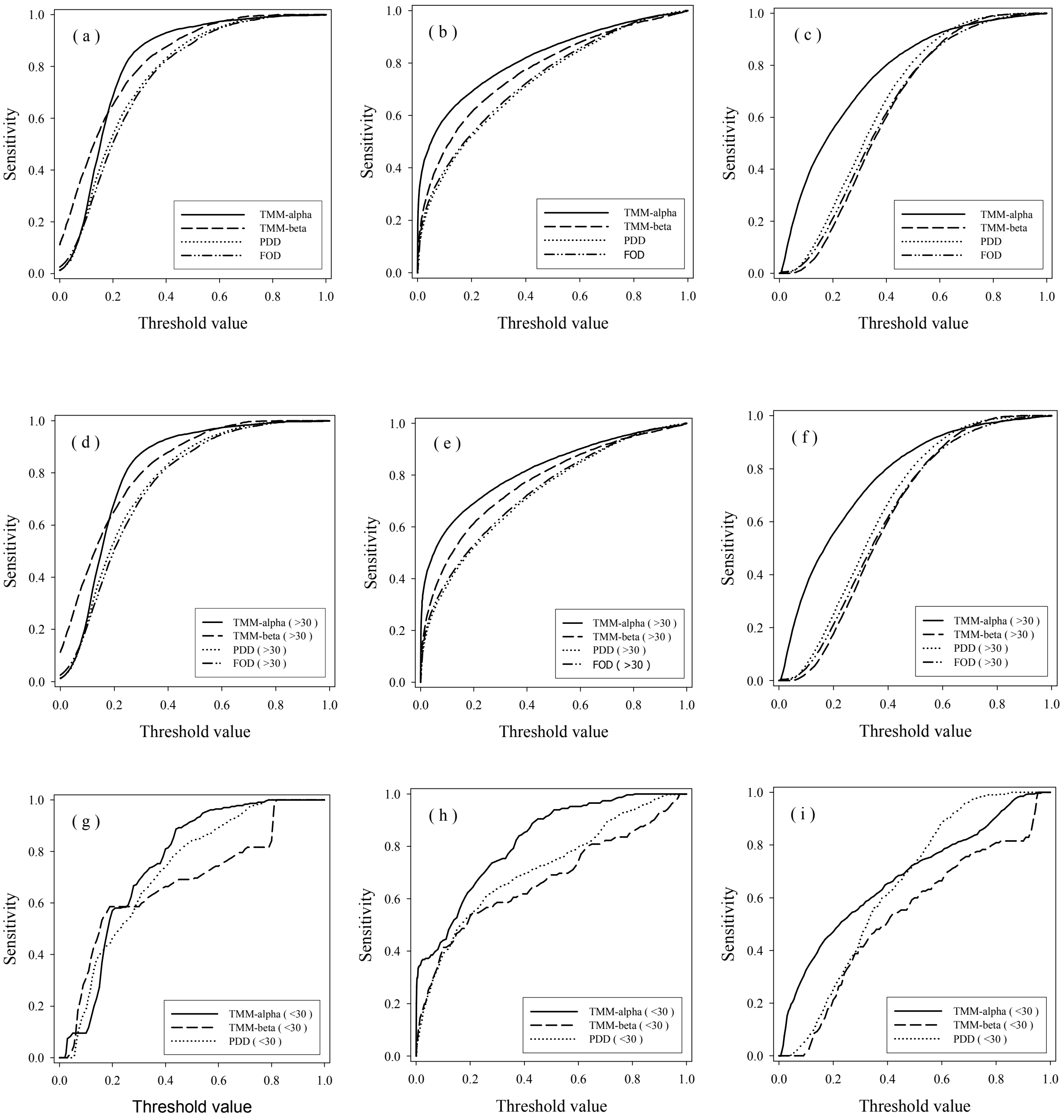

Figure S5. The change of prediction accuracy of three predictors in three datasets as a function of the threshold value for order predictions. (a), (b), and (c) are for VL3, VLXT, and VSL2, respectively. FDD represents fully disordered dataset, PDDM is the disordered regions of partially disordered proteins, TMM-alpha stands for disordered regions of helical transmembrane proteins, while TMM-beta is the disordered regions of beta transmembrane dataset. (d), (e) and (f) are the results of VL3, VLXT and VSL2 for the longer subsets of TM-alpha, TM-beta, and PDD, respectively. (g), (h) and (i) are of VL3, VLXT, and VSL2, for the shorter ones, accordingly.

Table S1. Summary of ROC Curves of Prediction on Structured Residues

	Area			Threshold value			Accuracy		
	VL3	VLXT	VSL2	VL3	VLXT	VSL2	VL3	VLXT	VSL2
TM-alpha	0.723 ± 0.001	0.743 ± 0.001	0.798 ± 0.001	0.19 ± 0	0.18 ± 0	0.33 ± 0	66.9% ± 0.1%	67.9% ± 0.1%	73.2% ± 0.1%
TM-beta	0.757 ± 0	0.763 ± 0	0.815 ± 0	0.21 ± 0	0.33 ± 0	0.48 ± 0	68.0% ± 0	73.1% ± 0	73.6% ± 0
PDD	0.735 ± 0.001	0.738 ± 0.002	0.784 ± 0.002	0.28 ± 0.01	0.36 ± 0.01	0.43 ± 0	68.0% ± 0	67.3% ± 0	71.5% ± 0.2%

The areas under ROC curve, threshold values, and accuracies at break-even points of VL3, VLXT, and VSL2 for TM-alpha, TM-beta and PDD in order prediction by applying the balanced bootstrapping.

Table S2. Summary of ROC Curve Results Using Short and Long Regions of Structured Residues

	Area			Threshold value			Accuracy		
	VL3	VLXT	VSL2	VL3	VLXT	VSL2	VL3	VLXT	VSL2
TM-alpha$^{(L)}$	0.702 ± 0.002	0.731 ± 0.002	0.797 ± 0.002	0.19 ± 0	0.17 ± 0	0.33 ± 0.01	63.7% ± 0.2	66.2% ± 0.2%	73.0% ± 0.3%
TM-beta$^{(L)}$	0.697 ± 0.001	0.734 ± 0.001	0.776 ± 0	0.19 ± 0.01	0.28 ± 0	0.46 ± 0	62.7% ± 0.1%	68.7% ± 0.1%	71.4% ± 0
PDD$^{(L)}$	0.711 ± 0.001	0.702 ± 0.001	0.755 ± 0.001	0.25 ± 0	0.33 ± 0.01	0.41 ± 0	65.4% ± 0.1%	64.4% ± 0.4%	68.8% ± 0.1%
TM-alpha$^{(S)}$	0.566 ± 0.013	0.614 ± 0.025	0.639 ± 0.006	0.20 ± 0	0.16 ± 0.01	0.33 ± 0.01	57.2% ± 1.5%	54.4% ± 0.1%	59.1% ± 0.6%
TM-beta$^{(S)}$	0.406 ± 0.012	0.697 ± 0.008	0.602 ± 0.006	0.13 ± 0.01	0.37 ± 0.01	0.50 ± 0.01	42.4% ± 0	59.4% ± 0.4%	59.2% ± 0.7%
PDD$^{(S)}$	0.589 ± 0.009	0.701 ± 0.010	0.691 ± 0.009	0.24 ± 0.01	0.34 ± 0.01	0.43 ± 0.01	51.2% ± 0.6%	65.6% ± 0.2%	64.1% ± 0.2%

Areas, threshold values, and accuracies at break-even points of VL3, VLXT, and VSL2 for length-dependent subsets of TM-alpha, TM-beta, and PDD in order prediction. All the results are from balanced bootstrapping. Superscript (L) and (S) indicate the longer and shorter subsets.