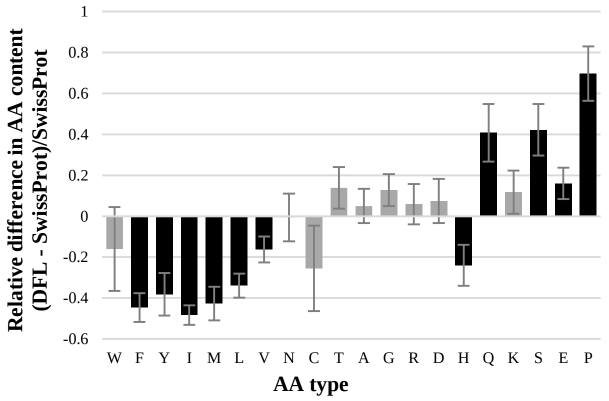
Supplement for the article entitled “DFLpred: High throughput prediction of disordered flexible linker regions in proteins sequences”

Fanchi Meng[1] and Lukasz Kurgan[2,1]*

[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, T6G 2V4, Canada.

[2]Department of Computer Science, Virginia Commonwealth University, Richmond, 23284, U.S.A.



Supplementary Figure S1. Enrichment and depletion of amino acids types in the DFL regions. The amino acid composition of residues in the DFL regions was compared against a generic set of residues collected from SwissProt ver. 51 using Composition Profiler. Residues on the *x*-axis are sorted according to the propensity for intrinsic disorder based on the TOP-IDP scale, from low to high. Black bars indicate that the depletion or enrichment is significant with *p*-value < 0.05.

1 Amino acid composition in DFL regions

Supplementary Figure S1 shows relative difference in the composition of each of the 20 AA types between residues in DFL regions and a reference population of residues collected from ver. 51 of SwissProt (Bairoch, et al., 2005). The relative difference is defined as (AA content AA in DFLs – AA content in SwissProt) / AA content in SwissProt. Values > 0 denote that a given AA type is enriched in DFLs while values < 0 denote that it is depleted in DFLs. The analysis was performed using Composition Profiler (Vacic, et al., 2007). AAs are sorted according to their propensity for intrinsic disorder based on the TOP-IDP scale (Campen, et al., 2008), from the most order-promoting AAs on the left to the most disorder promoting residues on the right. The residues significantly enriched (*p*-value < 0.05) in DFLs include Q, S, E, and P and they are also among the residues with the highest propensity for disorder. The residues significantly depleted (*p*-value < 0.05) in DFLs are primarily biased to be order-promoting and include F, Y, I, M, L, and V. The one exception is H which is depleted in DFLs and disorder promoting at the same time. The depletion in the DFLs for this residue is potentially due to fact that H has low propensity for flexibility in contrast to Q, S, E, and P (Bhaskaran and Ponnuswamy, 1988). Overall, as expected, the pattern of depletion and enrichment of AA in DFLs is primarily driven by the disordered nature of these regions.

2 List of considered features for the predictive model

Features from the amino acid (AA) sequence (40 features)

- CENT_AA$_{\{\text{AA type}\}}$: binary coding for the type of AA of the residue in the center (CENT) of the window (20 features).
- WIN_AA_content$_{\{\text{AA type}\}}$: number of residues of a given type of AA in the sliding window (WIN), divided by the length of the window (20 features).

Features based physicochemical properties of AAs quantified based on the 531 amino acid indices from the AAindex database (AAind, 2124 features):

- CENT_AAind_val$_{\{\text{index name}\}}$: value of a given AAindex for the type of AA of the residue in the center of the window (531 features).
- WIN_AAind_avg$_{\{\text{index name}\}}$: average value of a given AAindex for all residues in the sliding window (531 features).
- WIN_AAind_std$_{\{\text{index name}\}}$: standard deviation of values of a given AAindex for all residues in the sliding window (531 features).
- WIN_AAind_dif$_{\{\text{index name}\}}$: difference between average value of a given AAindex for all residues in the sliding window and average value for residues on segments that flank the window on both sides; the number of these flanking residues equals to the half of the window size (i.e., eight residues that extend the original window on side are used). These features were inspired by ref. (Disfani, et al., 2012) (531 features).

Features from the putative secondary structure (SS) derived from the input sequence using PSIPRED (SS, 22 features):

- CENT_SS_is$_{\{\text{H, E, C}\}}$: binary coding for the type of SS of the residue in the center (CENT) of the window (3 features).
- WIN_SS_content$_{\{\text{H, E, C}\}}$: number of helix, strand and coil residues in the sliding window divided by the length of the window (3 features).
- WIN_SS_sum$_{\{\text{HE, HC, EC}\}}$: sum of number of helix and strand residues, helix and coil residues, and strand and coil residues in the sliding window, normalized by the length of the window (3 features).
- WIN_SS_num_region$_{\{\text{H, E, C}\}}$: number of helix, strand and coil regions in the sliding window, normalized by the length of the win-

dow. Each region consists of a segment of consecutive helix/strand/coil residues; the minimal length is 3/1/2, which is the size of the shortest helix/strand (beta bridge)/coil. (3 features).

- WIN_SS_sum_region$_{HEC}$: sum of the number of helix, strand and coil regions in the sliding window, normalized by the length of the window (1 feature).
- WIN_SS_{longest, shortest, avg}_region$_{\{H, E, C\}}$: longest, shortest and average length of helix, strand and coil regions in the sliding window, normalized by the length of the window (3 × 3 = 9 features).

Features from the putative intrinsically disordered and structured regions derived from the input sequence using IUPred (IUP, 40 features):

- CENT_IUP_is$_{\{L, S, D\}}$: binary encoding of the prediction of long disordered regions with IUPred_long, short disordered regions with IUPred_short and structured regions with IUPred_struct for the residue in the center of the window (3 features).
- CENT_IUP_val$_{\{L, S\}}$: propensity score for disorder predicted with IUPred_long and IUPred_short for the residue in the center of the window (2 features).
- WIN_IUP_content$_{\{L, S\}\_\{0, 1\}}$: number of ordered and disorder residues predicted with IUPred_long and IUPred_short in the sliding window, divided by the length of the window (2 × 2 = 4 features).
- WIN_IUP_num_region$_{\{L, S\}\_\{0, 1\}}$: number of ordered and disordered regions predicted with IUPred_long and IUPred_short in the sliding window, normalized by the length of the window. Each region consists of a segment of consecutive disordered or ordered residues; the minimal length of disordered regions is 4 (Monastyrskyy, et al., 2011; Monastyrskyy, et al., 2014) (2 × 2 = 4 features).
- WIN_IUP_sum_region_$_{\{L, S\}\_01}$: sum of the number of ordered and disorder regions predicted with IUPred_long and IUPred_short in the sliding window, normalized by the length of the window (2 features).
- WIN_IUP_{longest, shortest, avg}_region$_{\{L, S\}\_\{0, 1\}}$: longest, shortest and average length of ordered and disorder regions predicted with IUPred_long and IUPred_short in the sliding window, normalized by the length of the window (3 × 2 × 2 = 12 features).
- WIN_IUP_{avg, std}$_{\{L, S\}}$: average and standard deviation of propensity scores predicted with IUPred_long and IUPred_short for residues in the sliding window. (2 × 2 = 4 features).
- WIN_IUP_fractionD$_{\{0, 1\}}$: number of residues in structured regions and other regions (not located in structured regions) predicted with IUPred_struct in the sliding window, divided by the length of the window (2 features).
- WIN_IUP_{longest, shortest, avg}_regionD$_{\{0, 1\}}$: longest, shortest and average length of structured regions and other regions (not located in structured regions) predicted with IUPred_struct in the sliding window, normalized by the length of the window. Each region consists of a segment of consecutive structured or non-structured residues (3 × 2 = 6 features).
- WIN_IUP_sum_regionD$_{01}$: sum of the number of structured regions and other regions (not located in structured regions) predicted with IUPred_struct in the sliding window, normalized by the length of the window (1 feature).

Features based on the sequence complexity derived from the input sequence using SEG (SEG, 10 features):

- CENT_SEG_is$_{H}$: binary encoding of the high vs. low complexity computed with SEG of residue in the center of the window (1 feature).
- WIN_SEG_content$_{\{L, H\}}$: number of residues in the sliding window in low and high complexity regions, divided by the length of the window (2 features).
- WIN_SEG_{longest, shortest, avg}_region$_{\{L, H\}}$: longest, shortest and average length of low and high complexity regions in the sliding window, normalized by the length of the window (3 × 2 = 6 features).
- WIN_SEG_sum_region$_{LH}$: sum of the number of low and high complexity regions in the sliding window, normalized by the length of the window (1 feature).

3 PREDICTIONS GENERATED BY THE CLOSEST ALTERNATIVE METHODS

3.1 The UMA method

Based on the description in ref. (Udwary, et al., 2002), first, we found homologous sequences for the query proteins in the test dataset by running BLAST with *e*-value equals 1e-20 against the non-redundant (NR) database. We filtered the corresponding hits to insure that they have similar length compared to the length of the query sequence (± 20% of the query sequence length). We selected the first 14 hit sequences (ranked by the *e*-value) as the homologs of the query sequence. According to the authors of UMA, 5 or more sequences in the multiple sequences alignment (MSA) are suggested, and in their software package they use 14 sequences for the multiple sequence alignment (MSA), and thus we maintained the same setup. If the number of filtered hits < 14 then we selected all hits. If the number of filtered hits < 7 then we relax the threshold on the *e*-value to 1e-10. We generated the MSA profile with ClustalX (Thompson, et al., 1997) and putative secondary structure with PHDsec (Rost, 1996; Rost and Sander, 1993); these tools are suggested by the authors of UMA. Since a low UMA score indicates that a residue is more likely to be a flexible linker, we used 1-UMA score (normalized to the range between 0 and 1 using the min-max normalization) as the propensity score of a residue being in a DFL, to make it consistent with results generated by DFLpred and the other methods. UMA cannot predict the first 20 and last 20 residues in the input sequence due to the use of sliding windows and thus we exclude these residues from the evaluation of this method, i.e., we calculate the predictive quality using the remaining residues.

3.2 Predictors of flexible residues

The flexibility scores generated by FlexPred and PredBF were collected from their webservers; flexibility scores of Predyflexy and DynaMine were derived by running their standalone packages. All methods were run with their default parameters. Since different flexibility predictors output scores in different ranges, we normalized their outputs to the 0 to 1 range using the min-max normalization. For the DynaMine method that outputs S^2 valued, we use (1 – S^2 value) as the flexibility score because unlike the other flexibility predictors, smaller DynaMine output corresponds to residues that are predicted as more likely to be flexible. Predyflexy cannot predict the first 10 and last 10 residues due to the use of sliding windows and thus we exclude these residues from the evaluation of this method, i.e., we calculated the predictive quality using the remaining residues.

3.3 Domain Predictor

We collected the domain conservation scores (DCS) from the ThreaDom webserver. We scaled the output score from 0 to 1 and since a higher DCS indicates a higher propensity of domain we use an inversed score (1 – normalized DCS) as the propensity of a given residue for DFLs. Five sequences in our test set are shorter than the required length of ThreaDom (≥ 80 residues) and they are excluded from our evaluations.

3.4 Combination of UMA and disorder predictors and of flexibility and disorder predictors

We applied two approaches to combine predictions: 1) by multiplying the scores predicted with UMA and flexibility predictors by the binary disorder predictions; and 2) by multiplying the scores predicted with UMA and flexibility predictors by the predicted real-valued propensity for the disorder. In the first scenario, if a given residues is predicted as disordered then we use the UMA score/flexibility score to quantify its predicted propensity, otherwise (residues was not predicted as disordered) we set this propensity score to zero. The second scenario uses product of the putative disorder scores and UMA score/flexibility scores as the predicted propensity score. Disorder scores of IUPred (long and short version) and Espritz (NMR, X-Ray and Disprot flavors) were derived from running standalone versions of the predictors. The disorder scores of MFDp were collected from its webserver.

Supplementary Table S1. Cross validation results for the three types of classifiers on the training dataset.

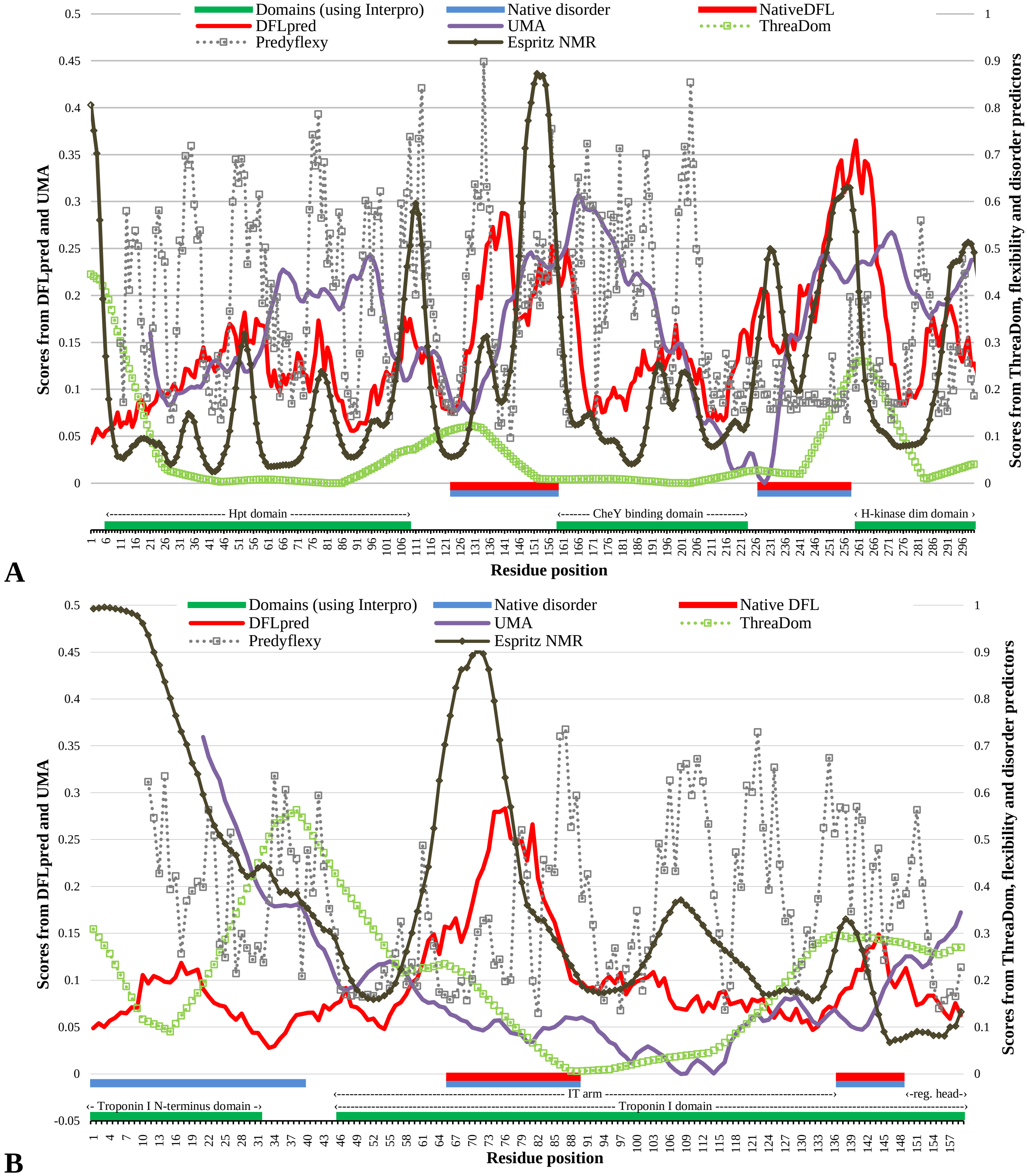
Classifier	T_{step1}	T_{step2}	Number of the selected features	Param.	AUC	AUC$_{lowFPR}$	Ratio
Logistic regression	0.50	0.35	4	$r = 100$	0.702	0.016	3.270
Naive Bayes	0.50	0.35	4	N/A	0.680 +	0.014 +	2.812 +
k-nearest neighbor	0.45	0.35	5	$k = 500$	0.677 +	0.015 +	2.933 +

T_{step1}: threshold for normalized r_{pb} or φ; T_{step2}: threshold for r_{pc}; Param.: parameters selected for individual classifiers where r is the ridge for logistic regression and k is the number of nearest neighbors; AUC: area under the ROC; AUC$_{lowFPR}$: area of a part of the ROC for FPR between 0 and 0.1; Ratio = AUC$_{lowFPR}$/AUC$_{random}$ where AUC$_{random}$ is the AUC of random predictor assessed for FPR between 0 and 0.1. The AUC, AUC$_{lowFPR}$ and ratio values were calculated over the 4 combined test folds in the cross validation, and thus they represent results on the entire training dataset. + indicates that difference in predictive quality between LR and another classifier is statistically significant at *p*-value < 0.01.

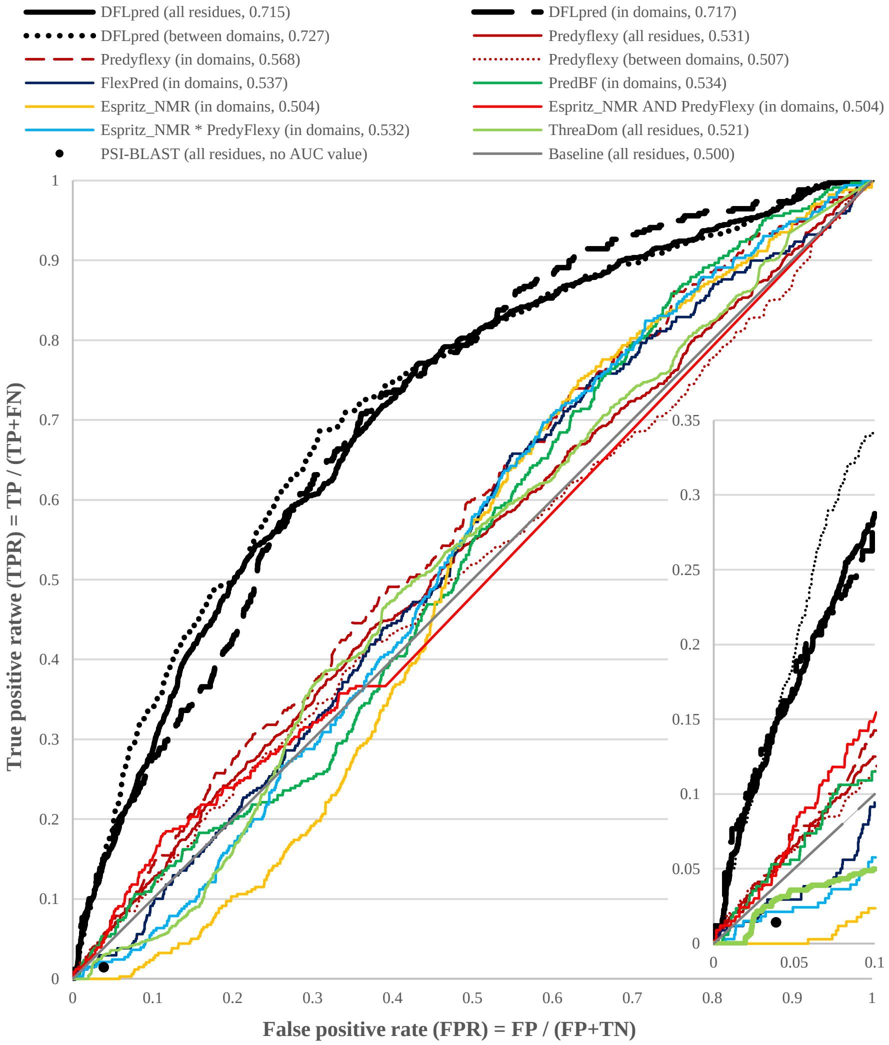
Supplementary Table S2. Comparison of predictive quality on the test dataset for residues localized in domains and outside of domains.

Prediction target	Method	Evaluation on residues inside domains						Evaluation on residues between domains						AUC$_{average}$
		AUC	*p*-value	AUC$_{lowFPR}$	*p*-value	Ratio	*p*-value	AUC	*p*-value	AUC$_{lowFPR}$	*p*-value	Ratio	*p*-value	
DFLs	DFLpred	0.717		0.017		3.288		0.727		0.019		3.843		0.722
Flexible linkers	UMA	0.290	0.000	0.006	0.002	1.240	0.002	0.415	0.000	0.001	0.002	0.107	0.002	0.352
Flexible residues	Predyflexy	0.568	0.001	0.007	0.002	1.364	0.002	0.507	0.000	0.007	0.000	1.233	0.000	0.538
	FlexPred	0.537	0.002	0.004	0.002	0.694	0.002	0.476	0.001	0.003	0.001	0.639	0.001	0.506
	PredBF	0.534	0.002	0.006	0.002	1.239	0.002	0.399	0.000	0.005	0.000	0.902	0.000	0.466
	PROFbval	0.436	0.002	0.006	0.002	0.662	0.002	0.446	0.000	0.008	0.004	0.300	0.000	0.441
	Dynamine	0.455	0.002	0.001	0.002	0.165	0.002	0.354	0.000	0.003	0.002	0.628	0.002	0.404
Disordered residues	Espritz NMR	0.504	0.000	0.000	0.002	0.093	0.002	0.338	0.000	0.002	0.002	0.289	0.002	0.421
	IUPred_short	0.387	0.000	0.000	0.002	0.001	0.002	0.323	0.002	0.000	0.000	0.062	0.000	0.355
	MFDp	0.273	0.000	0.000	0.002	0.000	0.002	0.316	0.000	0.000	0.002	0.000	0.002	0.294
DFLs	Espritz NMR & Predyflexy(best combination using binary disorder)	0.504	0.000	0.008	0.000	1.486	0.000	0.424	0.000	0.005	0.001	1.024	0.001	0.464
	Espritz NMR & Predyflexy (best combination using disorder propensity)	0.532	0.002	0.002	0.002	0.485	0.002	0.371	0.000	0.004	0.002	0.772	0.002	0.452

The methods were ranked by AUC value in each category; *p*-values quantify significance of the differences in predictive quality when compared with DFLpred.



Supplementary Figure S2. Predictions and native annotations for the C-terminus of the chemotaxis cheA protein (panel A) and the N-terminus of the troponin I protein (panel B) from the test dataset. We include annotations and names of domains (green horizontal line at the bottom with names above the line), disordered regions (blue horizontal line at the bottom), DFLs (red horizontal line at the bottom), and predictions from DFLpred (thick red plot), UMA (thick violet plot), best performing disorder predictor Espritz (black line with diamond markers), best performing flexibility predictor PredyFlexy (dotted gray line with square markers), and the domain predictor TreaDom (dotted green line with square markers).



Supplementary Figure S3. ROC curves on the test dataset for methods that achieved AUC > 0.5 in Table 2 (on the whole test dataset) or in Supplementary Table S2 (we use average of the results for the intra- and inter-domain residues). Insert in the bottom right corner zooms on the ROCs for FPR between 0 and 0.1. The scope of the prediction (all residues, in domains or between domains) and AUC values are shown inside brackets next to the names of methods in the figure legend.

References

Bairoch, A., *et al.* The Universal Protein Resource (UniProt). *Nucleic Acids Research* 2005;33(suppl 1):D154-D159.

Bhaskaran, R. and Ponnuswamy, P.K. Positional Flexibilities of Amino-Acid Residues in Globular-Proteins. *Int J Pept Prot Res* 1988;32(4):241-255.

Campen, A., *et al.* TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. *Protein and peptide letters* 2008;15(9):956-963.

Disfani, F.M., *et al.* MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics* 2012;28(12):i75-83.

Monastyrskyy, B., *et al.* Evaluation of disorder predictions in CASP9. *Proteins* 2011;79 Suppl 10:107-118.

Monastyrskyy, B., *et al.* Assessment of protein disorder region predictions in CASP10. *Proteins* 2014;82 Suppl 2:127-137.

Rost, B. PHD: Predicting one-dimensional protein structure by profile-based neural networks. In: Russell, F.D., editor, *Methods in Enzymology*. Academic Press; 1996. p. 525-539.

Rost, B. and Sander, C. Prediction of Protein Secondary Structure at Better than 70% Accuracy. *Journal of Molecular Biology* 1993;232(2):584-599.

Thompson, J.D., *et al.* The CLUSTAL_X Windows Interface: Flexible Strategies for Multiple Sequence Alignment Aided by Quality Analysis Tools. *Nucleic Acids Research* 1997;25(24):4876-4882.

Udwary, D.W., Merski, M. and Townsend, C.A. A Method for Prediction of the Locations of Linker Regions within Large Multifunctional Proteins, and Application to a Type I Polyketide Synthase. *Journal of Molecular Biology* 2002;323(3):585-598.

Vacic, V., *et al.* Composition Profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics* 2007;8(1):211.