Electronic Supplementary Material (ESI) for Molecular BioSystems.
This journal is © The Royal Society of Chemistry 2015

Supplement

Molecular Recognition Features (MoRFs) in three domains of life

Jing Yan,[1] A. Keith Dunker,[2,3*] Vladimir N. Uversky,[4,5,6*] and Lukasz Kurgan[7,1*]

[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Canada, T6G 2V4
[2]Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, USA, 46202
[3]Indiana University School of Informatics, Indianapolis, USA, 46202
[4]Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL, USA, 33612
[5]Institute for Biological Instrumentation, Russian Academy of Sciences, Pushchino, Moscow Region, Russian Federation, 142292
[6]Biology Department, Faculty of Science, King Abdulaziz University, P.O. Box 80203, Jeddah, Kingdom of Saudi Arabia, 21589
[7]Department of Computer Science, Virginia Commonwealth University, Richmond, U.S.A., 23284

[*]Corresponding authors

LK: Virginia Commonwealth University, 401 West Main Street, Room E4225, Richmond, VA, USA 23284
E-mail: lkurgan@vcu.edu; phone: 804-827-3986; fax: 804-828-2771

VNU: University of South Florida, 12901 Bruce B. Downs Blvd. MDC07, Tampa, FL, USA 33647
E-mail: vuversky@health.usf.edu; Tel.: 813-974-5816; Fax: 813-974-3757

AKD: Indiana University, 410 W. 10th Street, Suite 5000, Indianapolis, IN, USA 46202
E-mail: kedunker@iu.edu; Tel.: 317-278-9220; Fax: 317-278-9217

Supplementary Methods

We develop and benchmark an accurate and high-throughput predictor of MoRFs regions, called fMoRFpred, and apply it to characterize MoRFs on the genomic scale. This is motivated by the observation that the existing and widely-used predictor MoRFpred [1] is too slow to be used to perform large-scale predictions. In spite of this limitation MoRFpred enjoys a wide-spread use with over 1100 unique users from 58 countries since 2012 when the method was released. This method was referenced in dozens of studies, with a few recent examples including characterization of MoRFs in PTEN protein, membrane proteins, ribosomal proteins, kinases, viral proteomes, plant proteomes, and their contribution to prediction of protein function and formation of scaffolds [2-9]. Our aim is to obtain predictive performance that is similar to that of MoRFpred while securing a substantial speedup.

Datasets

We use the same datasets as in Disfani *et al.* [1] to design and benchmark our predictive model. The design was performed using the TRAINING dataset with 421 chains with the annotated MoRFs. We test our model on four test datasets proposed in Disfani *et al.* [1]: TEST (419 proteins), TEST2012 (45 proteins which include new depositions when compared with the TRAINING and TEST datasets), EXPER2008-12 (8 proteins) and the NEGATIVE dataset (28 proteins with no MoRFs). The EXPER2008-12 includes proteins with MoRFs in regions that were experimentally verified to be disordered in isolation. The four test datasets share <=30% sequence similarity with the TRAINING dataset.

Evaluation protocols and measures

The design and test protocol and evaluation criteria follow the work in Disfani *et al*. [1] The evaluation criteria include Predicted Positive Rate (PPR), Success Rate (SR) and Area Under the Receiver Operating Characteristic (AUC) curve. The PPR measure is defined as (TP+FP)/(TP+FN) where TP is the number of true positive samples (correctly predicted native MoRF residues), FN is the number of false negatives (MoRF residues predicted as non-MoRFs), FP is the number of false positives (non-MoRF residues predicted as MoRFs), and TN is the number of true negatives (correctly predicted native non-MoRF residues). The Receiver operating (ROC) curve is a plot of false positive rate (FPR) = FP/(FP+TN) vs. true positive rate (TPR) = TP/(TP+FN) that is obtained by changing a cutoff to binarize predictions (MoRF vs. non-MoRF residue) from the numeric propensity generated by the predictive model. We also compute the SR measure that was design to estimate predictive quality in cases where the annotations are incomplete. A given protein is assumed to be predicted correctly if the average propensity of residues to form MoRFs predicted by the model for the native MoRFs is higher than for all residues in the protein. The SR value is the fraction of the correctly predicted proteins in a given dataset.

The fMoRFpred was designed utilizing the 5 fold-cross validation on the TRAINING dataset. The final model was build using the entire TRAINING dataset and was tested on the four test datasets. We evaluated statistical significance of the differences in the predictive quality of fMoRFpred and other relevant predictors. For each test dataset, we randomly select 50% proteins

10 times and calculate the 10 corresponding SR and AUC values. We compare the 10 paired results for each measurement between a given pair of predictors. Given that the measurements follow normal distribution, tested using Anderson-Darling test with the 0.05 significance level, we apply the paired *t*-test; otherwise we use the Wilcoxon-Mann-Whitney test. Differences are assumed statistically significant when p-value < 0.05.

Design of fMoRFpred

The design consists of several steps. First, every residues in a given protein sequence is encoded using a set of numerical features representing its physicochemical and biochemical properties and properties of its neighboring residues. Next, feature selection is used to select a subset of these features that are relevant to the prediction of MoRF residues. Then, we parameterize a linear Support Vector Machine (SVM) model that takes these selected features as an input to generate propensities for formation of MoRFs. Finally, we devise a filter to remove MoRFs that are predicted outside of intrinsically disordered regions (IDRs).

Feature-based encoding of the input sequence

We consider a comprehensive set of numerical features including amino acid composition, information derived from intrinsic disorder predicted by the IUPred [10] and Espritz [11] methods, secondary structure (SS) predicted with PSIPRED [12], and over 500 amino acid (AA) indices that quantify physicochemical properties of residues that were collected from the AAindex database [13]. This information is processed using sliding windows centered on the predicted residues. We use the same window size of 25 residues as in Disfani *et al*. [1], which means that predictions for residues at position k in the sequence use information from residues at positions k-12, k-11, …, k, …, k+12. We use predictions of disorder and secondary structure for each position in the window and we also aggregate this and other information, such as AA indices and composition, in the window. More specifically, we calculate content of disorder and secondary structures and average values of AA indices over the whole window and smaller windows inside with sizes of 3, 5, 7,…, 23 residues. We also compute difference between average values of the AA indices between predicted intrinsically disordered and ordered regions, coil and non-coil regions, helical and non-helical regions, and strand and non-strand regions in the whole window. Moreover, we calculate the difference between inner part of the window (residues in the center of the window) and their flanking residues, as explained in Disfani *et al*. [1]. While the total number of features considered in the design of the MoRFpred was 1764, here we consider a much larger set of 7036 features. The new features are related to the differences between regions of putative disorder and secondary structures and the features that consider smaller inner windows. Moreover, we also assure that the features can be computed rapidly, which means that we use only high-throughput predictors of disorder and secondary structure.

Feature selection

We use the two-step feature selection to select a subset of relevant (step 1) and non-redundant (step 2) features, which is inspired by the methodology used in [1]. The selection is performed exclusively using the TRAINING dataset. The first step combines results of three methods to rank the considered 7036 features. Two methods rank features based on their correlation with the native MoRF annotations quantified using the point-biserial and Phi correlation coefficients for continuous and dichotomous features, respectively. This calculation is done in two different

ways: using all the residues in the TRAINING dataset (referred as complete-data) and using a sampled subset of residues from the TRAINING dataset (referred as local-data). In the local-data case, we undersampled the non-MoRF residues in each protein to obtain 2:1 ratio with the MoRF residues. Since, as argued in [1], annotations of MoRF regions are incomplete and the residues surrounding the MoRF regions are less likely to be MoRFs, we selected the residues that flank MoRF regions. The values were computed as averages on the 5 training folds based on the 5-fold cross validation on the TRAINING dataset. The third method is a wrapper-based selection which ranks features based on their predictive performance measured by SR values when used individually with an SVM model with linear kernel and default parameters (complexity parameter $C = 5$) on 5-fold cross validation on the TRAINING dataset. Features with their absolute values of correlation coefficients and the SR values < 0.05 are removed and we maintain three lists of features for each of the three rankings.

In the second step, we use the wrapper-based sequential forward feature selection on the features obtained from the first step for each of the three rankings. Starting with the top ranked feature, we accept the next ranked feature into our feature set only if this feature improves the SR values by at least 0.01 when compared with the prediction obtained using the feature set before the addition. We go through the sorted list of features once. The predictions are based on linear SVM model with default parameters ($C = 5$) using the 4+1 fold cross validation protocol, which was introduced in [1]. This protocol includes 4-fold cross validation on 4 out of the 5 original folds and additional test in which these four folds are used together to build a model that is tested on the set-aside 5th fold. The use of the 4+1-fold cross validation helps to reduce over fitting into the TRAINING dataset. This sequential forward feature selection is performed on the sampled subset of residues from the TRAINING dataset (local-data). As a result, 13, 3, and 4 features were selected for the complete-data, local-data and the SR values based ranking.

The predictive performance for the final selected set of features generated by the three feature selection methods and the combined set of 20 features based on the 5-fold cross validation on TRAINING dataset is shown in Supplementary Table S1. The complete-data method secures the best AUC and PPR, the local-data method achieves the best TPR, and the SR based ranking approach obtains the highest success rate. Combing the three sets of selected features together results in a design that outperforms the individual methods by improving success rate by at least 0.068 and AUC by at least 0.01. Consequently, we use the combined set of 20 selected features.

Parameterization of predictive model

The 20 selected features are used together with a linear SVM classifier to implement our predictor. The selection of the SVM model is motivated by its use in the MoRFpred method and low runtime requirements. We attempted to parameterize the SVM model by performing a grid search over its complexity parameter $C = 2^n$ where $n = -7,-6,\ldots ,7$ based on cross-validation on the TRAINING dataset. The predictive performance of this model did not improve when compared with the default value of $C = 5$ and thus we use the default value.

Filtering of raw fMoRFpred predictions

Since MoRF regions are embedded in usually longer disordered regions, we filter out our raw MoRF predictions that are not located in a putative disordered region as follows. If a predicted

MoRF residue is classified as structured (ordered) by a given disorder predictor or a consensus of disorder predictors (i.e., $P_d > P_{Td}$ where P_d is the propensity for disorder generated by the disorder predictor and P_{Td} is the threshold used to convert this propensity to binary prediction: disordered vs. structured residue), then we decrease the propensity P_m generated by fMoRFpred as follows:

$$P_m = \begin{cases} P_m - \left(P_{Tm} - \left(\frac{P_{Td} - P_d}{P_{Td}} \right) \right) & when\ P_d < P_{Td} \\ P_m & otherwise \end{cases}$$

where P_{Tm} represents the threshold used to convert the propensity generated by fMoRFpred to binary prediction: MoRF vs. non-MoRF residue. If after applying the deduction the propensity P_m is still above the threshold P_{Tm} then we lower it to a value just below the threshold. The formula reveals that we do not change the propensity P_m if prediction is inside a putative disordered region. Note that the P_{Td} values for Espritz were set either based on the FPR or S_w indices [11] and we considered both options.

We empirically test using each of the five high-throughput disorder predictors individually (IUPred long regions, UIPred short regions, Espritz X-ray, Espritz NRM, and Espritz Disprot), and in consensus where the propensity P_d is computed as either the maximal, minimal or average value of the propensities generated by the five methods. The effects of the resulting 11 approaches (each Espritz version uses two P_{Td} values) when filtering the predictions of fMoRFpred on the TRAINING dataset are summarized in Supplementary Table S1. We observe an improvement when compared with the results before the filtering (see "Combination of the three methods (20 features)" row). We selected the filter based on the consensus of the five methods with the propensity calculated as the maximal value since this approach provides the highest SR value, favorable reduction of PPR from 2.696 to 0.723 (overprediction of MoRF residues is reduced to a rate comparable with the rate of native MoRF annotations) and relatively high value of AUC.

Benchmarking results

We benchmark fMoRFpred on three test datasets that are independent from the TRAINING dataset (i.e., they share sequence similarity below 30%) and compare the results with a comprehensive set of three MoRF predictors, ANCHOR[14] that predicts disordered protein binding regions, and 13 modern disorder predictors, see Supplementary Table S2. The predictions of disorder were used to predict MoRFs since the latter are usually located inside IDRs. The MoRF predictors include α-MoRF-PredI [15], α-MoRF-PredII [16], and MoRFpred [1]. The Anchor method does not predicts MoRFs (short recognition motifs involved in protein binding) but longer protein binding regions that also undergo a disorder-to-order transition. The considered disorder predictors are the three version of Espritz (each using two thresholds to compute binary predictions of disorder based on S_w and FPR measures) [11], the two versions of IUPred [10], MFDp [17], SPINED [18], MD [19], DISOCLUST [20] and DISOPRED2 [21]. We rank all methods according to the SR values on each test dataset and compute the average ranking. The fMoRFpred method achieves the second best SR value on each test datasets and overall. Only MoRFpred provides higher SR values, however this is at the expense of worse PPR values. Note that PPR = 1 when the number of predicted MoRF residues equals to the number of native MoRF residues and value higher (lower) than 1 indicates overprediction (underprediction). PPR values

of fMoRFpred are close to 1 while MoRFpred overpredicts MoRFs by 180% on the TEST and TEST2012 datasets. This suggests that fMoRFpred can be used to correctly estimate the overall abundance of MoRF residues, although their location in the sequence is predicted better by MoRFpred based on its higher SR and AUC values. Although the success rates and AUC values of fMoRFpred are lower than of MoRFpred, they are still relatively high in the 62 to 67% range for SR and 0.59 to 0.67 for AUC, depending on the dataset used. Anchor provides lower predictive performance which is expected as it targets disordered protein binding regions which could be seen as a superset of MoRFs. This explains why its PPR values are higher and the overall performance is lower but still substantially better then random. The α-MoRF-PredI and α-MoRF-PredII methods were designed to predict MoRF that fold into helices upon binding and as such cannot predict other types of MoRFs. This explains their low success rates. The disorder predictors overpredict the MoRFs by large margin between 340% (Espritz Disprot on the EXPER2008-12 dataset) and 3560% (DISOCLUST on TEST2012). This is expected as MoRFs constitute only a small fraction of the IDRs that these methods predict. The success rates of Espritz designed using Disprot-based annotations of disorder are relatively good and are ranked 3rd best behind fMoRFpred.

We also test fMoRFpred and the other considered methods on the NEGATIVE dataset which includes only structured proteins that do not have MoRFs; see Supplementary Table S3. We measure FPR values which quantify how many residues were incorrectly predicted (overpredicted) as MoRFs. The fMoRFpred makes mistakes for only 0.9% of residues compared to 6.3% obtained by MoRFpred. Only a few other methods, such as α-MoRF-PredI, α-MoRF-PredII, Anchor, and Espritz Disprot have lower FPR values.

We conclude that fMoRFpred can be used to accurately estimate abundance of MoRFs and provides reasonably good predictions of their location in the protein sequence, although these predictions are inferior to the predictions provided by MoRFpred. However, MoRFpred overpredicts MoRFs, which is evident based on its relatively high PPR values on the three test datasets and high FPR on the NEGATIVE dataset. The relatively good predictive quality of fMoRFpred could be attributed to the novel aspects of its design, in particular use of a comprehensive feature set and the disorder-based filter.

Runtime analysis

The high computational cost of MoRFpred was the main motivation behind the design of fMoRFpred. We compare the runtime of fMoRFpred, MoRFpred and Anchor based on predictions on the largest TEST dataset. We perform predictions with these methods on the same hardware using a Linux system with kernel version 3.5.0-28-generic x86_64. We divide proteins according to their length into ten equally-sized subsets and calculate average runtime over proteins in these subsets. Results are shown in Supplementary Figure S1. We observe that runtime is linearly proportional to the size of protein chain for the three methods, i.e., the runtime values are fit well by a linear function. Anchor is the fastest method and is two orders of magnitude faster than fMoRFpred, while fMoRFpred is over two orders of magnitude faster than MoRFpred. The new methods generates predictions in a second for an average sized protein with about 300 residues and in up to 10 seconds for a long protein with about 2000 residues. To compare MoRFpred takes about 1 hour to predict such long chains. Extrapolating from the runtime of fMoRFpred, prediction of MoRFs for the entire human genome that includes

approximately 70,000 proteins would take 70,000 seconds, which translates into about 20 hours. We conclude that fMoRFpred implements a reasonable trade-off between relatively low runtime, which allows for genome-scale predictions on a single desktop computer, and good predictive performance, being more accurate than Anchor and substantially faster than MoRFpred. A webserver-based implementation of fMoRFpred is freely available at http://biomine.ece.ualberta.ca/fMoRFpred/.

Supplementary Tables

Supplementary Table S1. Comparison of prediction results using different feature selection and filter methods on the TRAINING dataset. Highest success rate (SR) values and AUCs for each design step are shown in bold font.

Design step tested	Method used	PPR	SR	AUC
Feature selection	Complete-data ranking (13 features)	0.937	0.589	0.634
	Local-data ranking (3 features)	3.286	0.584	0.543
	Success rate ranking (4 features)	0.020	0.596	0.596
	Combination of the three methods (20 features)	2.696	**0.664**	**0.644**
Filters of the raw predictions	Consensus of 5 methods where P_d is set to maximal propensity generated by the five methods	0.723	**0.676**	0.640
	Espritz NMR (P_{Td} value based on FPR)	1.170	0.671	0.642
	Espritz Disprot (P_{Td} value based on S_W)	1.408	0.671	0.645
	IUPred short regions	0.831	0.669	0.628
	Espritz NM (P_{Td} value based on S_W)	1.841	0.669	0.649
	Espritz X-ray (P_{Td} value based on S_W)	2.344	0.668	0.646
	Espritz X-ray (P_{Td} value based on FPR)	2.033	0.666	**0.650**
	Consensus of 5 methods where P_d is set to average propensity generated by the 5 methods	0.723	0.658	0.624
	IUPred long regions	0.747	0.652	0.620
	Espritz Disprot (P_{Td} value based on FPR)	0.472	0.652	0.627
	Consensus of 5 methods where P_d is set to minimal propensity generated by the five methods	0.723	0.632	0.604

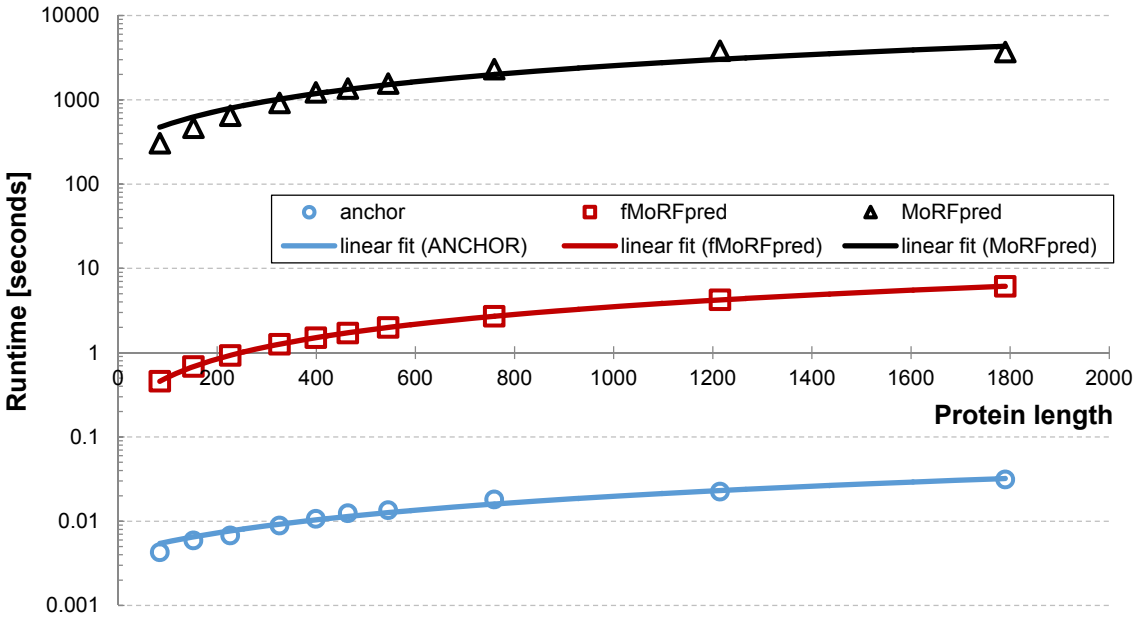
Supplementary Table S2. Predictive performance on the TEST, TEST2012 and EXPER2008-12 datasets. Results are ranked by success rate (SR) values. We consider 5 MoRF and 13 disorder predictors. All methods are ranked according to the SR value on each dataset and the average ranking is shown in the last column. Significance of the differences in the success rate (SR) and AUC values between fMoRFpred and each other method is shown in the "p" and "test" columns; +/=/– means that the predictive performance is significantly higher/ is not significantly different/ is significantly lower (*t*-test (t); Wilcoxon test (w); degrees of freedom = 9; *p*-value <0.05; details in the Methods section) when compared with a method in a given row.

	Predictor	TEST dataset								TEST2012 dataset								EXPER2008-12 dataset								
		PPR	SR			AUC			Rank	PPR	SR			AUC			Rank	PPR	SR			AUC			Rank	Avg Rank
			value	p	test	value	p	test			value	p	test	value	p	test			value	p	test	value	p	test		
MoRF predictors	MoRFpred	2.857	0.718	–	t	0.672	–	t	1	2.843	0.756	–	t	0.697	=	t	1	1.114	0.750	–	w	0.636	–	w	1	1
	fMoRFpred	0.956	0.663			0.648			2	0.995	0.667			0.671			2	0.643	0.625			0.590			2	2
	Anchor	12.820	0.611	+	t	0.599	+	t	5	14.339	0.578	+	t	0.634	+	t	5	6.229	0.500	+	w	0.556	+	w	5	5
	α-MoRF-PredII	5.096	0.303	+	t	N/A		w	16	5.923	0.311	+	t	N/A			16	2.143	0.250	+	w	N/A		w	9	16
	α-MoRF-PredI	1.947	0.158	+	w	N/A		w	18	1.851	0.133	+	t	N/A			18	0.814	0.000	+	w	N/A		w	18	18
Disorder predictors	Espritz Disprot (S_W)	18.402	0.616	+	t	0.704	–	t	3	22.193	0.533	+	t	0.734	–	t	9	9.033	0.625	+	w	0.652	–	w	2	3
	Espritz Dispot (FPR)	5.216	0.616	+	t	0.704	–	t	3	6.655	0.533	+	t	0.734	–	t	9	3.424	0.625	+	w	0.652	–	w	2	3
	MFDp	31.324	0.592	+	t	0.535	+	t	6	33.511	0.556	+	t	0.620	+	t	7	10.819	0.500	+	w	0.337	+	w	5	6
	IUPred short regions	16.015	0.537	+	t	0.521	+	t	7	17.375	0.600	+	t	0.612	+	t	3	7.024	0.250	+	w	0.446	+	w	9	7
	SPINED	25.575	0.513	+	t	0.532	+	t	8	31.479	0.467	+	t	0.605	+	t	11	9.948	0.250	+	w	0.330	+	w	9	11
	IUPred long regions	19.573	0.499	+	t	0.521	+	t	9	23.075	0.600	+	t	0.618	+	t	3	9.148	0.375	+	w	0.471	+	w	8	8
	MD	19.481	0.480	+	t	0.598	+	t	10	26.324	0.578	+	t	0.679	=	t	5	8.981	0.500	+	w	0.616	–	w	5	8
	DISOCLUST	30.202	0.449	+	t	0.499	+	t	11	35.636	0.556	+	t	0.512	+	t	7	9.914	0.250	+	w	0.290	+	w	9	10
	Espritz NMR (S_W)	29.163	0.379	+	t	0.493	+	t	12	33.521	0.400	+	t	0.549	+	t	14	10.624	0.250	+	w	0.350	+	w	9	12
	Espritz NMR (FPR)	22.089	0.379	+	t	0.493	+	t	12	24.296	0.400	+	t	0.549	+	t	14	8.829	0.250	+	w	0.350	+	w	9	12
	Espritz X-ray (S_W)	17.713	0.344	+	w	0.597	+	t	14	20.920	0.444	+	t	0.687	=	t	12	7.586	0.250	+	w	0.481	+	w	9	12
	Espritz X-ray (FPR)	12.632	0.344	+	w	0.597	+	t	14	14.800	0.444	+	t	0.687	=	t	12	5.548	0.250	+	w	0.481	+	w	9	12
	DISOPRED2	22.491	0.296	+	t	0.506	+	t	17	27.331	0.244	+	t	0.547	+	t	17	8.986	0.125	+	w	0.310	+	w	17	17

Supplementary Table S3. Predictive performance on the NEGATIVE dataset.

	Predictor	ACC	FPR
MoRF predictors	α-MoRF-PredI	1.000	0.000
	α-MoRF-PredII	1.000	0.000
	Anchor	0.995	0.005
	fMoRFpred	0.991	0.009
	MoRFpred	0.937	0.063
Disorder predictors	Espritz_DP_FPR	1.000	0.000
	MFDp	0.984	0.016
	IUPredL	0.974	0.026
	MD	0.969	0.031
	DISOPRED2	0.967	0.033
	Espritz_Cx_FPR	0.956	0.044
	IUPredS	0.949	0.051
	SPINED	0.926	0.074
	DISOCLUST	0.925	0.075
	Espritz_NMR_FPR	0.913	0.087
	Espritz_DP_SW	0.910	0.090
	Espritz_Cx_SW	0.903	0.097
	Espritz_NMR_SW	0.792	0.208

Supplementary Figures



Supplementary Figure S1. Comparison of runtime on the TEST dataset for fMoRFpred, MoRFpred, and ANCHOR. Proteins were sorted by their sequence length and divided into 10 sets of equal size by their size. The plot reports average runtime and average sequence size for each of the 10 sets. All predictors were run on the same hardware.

References

1. F. M. Disfani, W. L. Hsu, M. J. Mizianty, C. J. Oldfield, B. Xue, A. K. Dunker, V. N. Uversky and L. Kurgan, *Bioinformatics*, 2012, **28**, i75-83.
2. Z. Peng, C. J. Oldfield, B. Xue, M. J. Mizianty, A. K. Dunker, L. Kurgan and V. N. Uversky, *Cell Mol Life Sci*, 2014, **71**, 1477-1504.
3. J. J. Kathiriya, R. R. Pathak, E. Clayman, B. Xue, V. N. Uversky and V. Dave, *Molecular bioSystems*, 2014, **10**, 2876-2888.
4. X. Fan, B. Xue, P. T. Dolan, D. J. LaCount, L. Kurgan and V. N. Uversky, *Mol Biosyst*, 2014, **10**, 1345-1363.
5. D. Cozzetto and D. T. Jones, *Curr Opin Struct Biol*, 2013, **23**, 467-472.
6. X. Sun, E. H. Rikkerink, W. T. Jones and V. N. Uversky, *The Plant cell*, 2013, **25**, 38-55.
7. P. Malaney, R. R. Pathak, B. Xue, V. N. Uversky and V. Dave, *Scientific reports*, 2013, **3**, 2035.
8. I. Kotta-Loizou, G. N. Tsaousis and S. J. Hamodrakas, *Biochimica et biophysica acta*, 2013, **1834**, 798-807.

9. B. Xue, P. R. Romero, M. Noutsou, M. M. Maurice, S. G. Rudiger, A. M. William, Jr., M. J. Mizianty, L. Kurgan, V. N. Uversky and A. K. Dunker, *FEBS Lett*, 2013, **587**, 1587-1591.
10. Z. Dosztanyi, V. Csizmok, P. Tompa and I. Simon, *J Mol Biol*, 2005, **347**, 827-839.
11. I. Walsh, A. J. Martin, T. Di Domenico and S. C. Tosatto, *Bioinformatics*, 2012, **28**, 503-509.
12. D. T. Jones, *J Mol Biol*, 1999, **292**, 195-202.
13. S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama and M. Kanehisa, *Nucleic acids research*, 2008, **36**, D202-205.
14. B. Meszaros, I. Simon and Z. Dosztanyi, *PLoS computational biology*, 2009, **5**, e1000376.
15. C. J. Oldfield, Y. Cheng, M. S. Cortese, P. Romero, V. N. Uversky and A. K. Dunker, *Biochemistry*, 2005, **44**, 12454-12470.
16. Y. Cheng, C. J. Oldfield, J. Meng, P. Romero, V. N. Uversky and A. K. Dunker, *Biochemistry*, 2007, **46**, 13468-13477.
17. M. J. Mizianty, W. Stach, K. Chen, K. D. Kedarisetti, F. M. Disfani and L. Kurgan, *Bioinformatics*, 2010, **26**, i489-496.
18. T. Zhang, E. Faraggi, B. Xue, A. K. Dunker, V. N. Uversky and Y. Zhou, *Journal of biomolecular structure & dynamics*, 2012, **29**, 799-813.
19. A. Schlessinger, M. Punta, G. Yachdav, L. Kajan and B. Rost, *PloS one*, 2009, **4**, e4433.
20. L. J. McGuffin, *Bioinformatics*, 2008, **24**, 1798-1804.
21. J. J. Ward, L. J. McGuffin, K. Bryson, B. F. Buxton and D. T. Jones, *Bioinformatics*, 2004, **20**, 2138-2139.