Protein Structure & Function

Supplement for article entitled

"MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins"

Fatemeh Miri Disfani[1], Wei-Lun Hsu[2], Marcin J. Mizianty[1], Christopher J. Oldfield[2], Bin Xue[3], A. Keith Dunker[2], Vladimir N. Uversky[3,4], and Lukasz Kurgan[1*]

[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, CANADA

[2]Center for Computational Biology and Bioinformatics and Department of Biochemistry and Molecular Biology, Indiana University, Indianapolis, USA

[3]Department of Molecular Medicine, University of South Florida, Tampa, USA

[4]Institute for Biological Instrumentation, Russian Academy of Sciences, Pushchino, RUSSIA

1 METHODS

1.1 Test and evaluation protocols

To calculate ***success rate***, we compare the average predicted probability/propensity p of residues in the native MoRF region to the average probability of the whole sequence, and we assign a score to each sequence. For i_{th} sequence in a dataset, the success rate S_i is calculated as follows:

$$Ave_{MoRF} = \frac{\sum_{j \in MoRFs} p_j}{number\ of\ MoRF\ residues}$$

$$Ave_{non-MoRF} = \frac{\sum_{j \in non-MoRFs} p_j}{number\ of\ non-MoRF\ residues}$$

$$\begin{cases} if\ \ Ave_{MoRF} > Ave_{nonMoRF} & then\ S_i = 1 \\ othwerwise & then\ S_i = 0 \end{cases}$$

Since probabilities of the predicted MoRFs should be higher than the non-MoRFs, a correctly predicted sequence should have S_i = 1. Total success rate S is calculated by averaging the per sequence scores over all sequences in a given dataset:

$$S = \frac{\sum_i S_i}{number\ of\ sequences}$$

The ***accuracy***, ***true positive rate*** and ***false positive rate*** are defined as follows:

Accuracy = (TP+TN)/(TP+FP+TN+FN)

True positive rate = TPR = TP / (TP + FN) = TP / N_{MoRF}

False Negative rate = FPR = TN / (TN + FP) = TN / $N_{nonMoRF}$

where *TP* is the number of true positives (correctly predicted MoRF residues), *FP* denotes false positives (non-MoRF residues that were predicted as MoRF), *TN* denotes true negatives (correctly predicted non-MoRF residues), *FN* stands for false negatives (MoRF residues that were predicted non-MoRFs), N_{MoRF} is the number of native MoRF residues and $N_{non\text{-}MoRF}$ is the number of native non-MoRF residues. The accuracy values range between 0 and 1 and it is equal one when all residues are predicted correctly.

To generate the receiver operating characteristic (ROC) curve, the probabilities p (between 0 and 1) generated by a given prediction method are binarized such that all residues with probability equal or greater than a given threshold are set as MoRFs and all other residues are set as non-MoRFs. The thresholds are varied between 0 and 1 (they are set to each of the values of p) and for each threshold the TPR and the FPR are calculated. We use the ***area under the corresponding ROC curve*** (AUC), i.e., curve created by adjacent TPR vs. FPR points, to quantify the predictive quality.

To perform ***5-fold cross validation*** we divide the training set into 5 equal-sized subsets of protein chains. We use four of these subsets to form a training dataset that is utilized to compute the model and the fifth subset constitutes a test set that is used to perform the evaluation. This procedure is repeated five times, each time choosing a different fold as the test set. Finally, the results from the 5 test folds are averaged to estimate the performance. We note that sequence in that training set are clustered based on their similarity, using procedure explained in the last paragraph in section 2.1 in the main text. When selecting the five folds, the sequences in the same cluster are kept together. This assures that sequences between the folds share low similarity below 30%, which is also true when comparing training and test datasets.

We also use a modified version of the 5-fold cross validation, which we call ***4+1-fold cross validation***. The modification is meant to prevent overfitting (due to the large number of features that are considered) and to simulate predictions on the independent test dataset when using the training set. To implement the 4+1-fold cross validation, we use 4 of the 5 folds to implement the 4 fold cross validation and we keep the 5th fold as an independent test set.

1.2 Feature selection

Biserial correlation (Tate, 1954) is used to measure correlation of two quantities where one is binary and the other is continuous. Given binary variable X, we divide values of the continuous variable Y to two groups: 0 and 1, based on their corresponding values of X. The biserial correlation is calculated as:

$$corr(x, y) = \frac{M_0 - M_1}{S_n} \sqrt{\frac{n_0 n_1}{n^2}}$$

where S_i is the standard deviation of X and M_0 and M_1 are mean values for group 0 and group 1 with sizes n_0 and n_1 respectively.

We use biserial correlation when designing our method to perform feature selection i.e., to quantify the correlation of a given input feature with the native (binary) annotation of MoRFs. We perform this by calculating an average biserial correlation over 5 training folds using the training dataset. We use this average to sort the features in the descending order.

For binary input features we use ***φ coefficient*** (Ernest, 1991), which quantifies correlation when both variables are binary. Using notation from Figure S7 in the Supplement we define the φ coefficient as follows:

$$\varphi = \frac{P_{00}P_{11} - P_{10}P_{01}}{\sqrt{P_1 Q_1 P_2 Q_2}}$$

We scale φ to [-1, 1] range as φ/φ_{max} where φ_{max} is defined as:

$$\varphi_{\max} = \frac{\sqrt{Q_2 P_1}}{\sqrt{Q_1 P_2}} \quad for\ P_2 \geq P_1$$

2 RESULTS

2.1 Probability scores generated by MoRFpred identify higher quality predictions

We demonstrate that probabilities that are generated by MoRFpred can be used to select predictions that have higher quality. Figure S8 in the Supplement plots positive predictive value (PPV) for MoRF predictions (probability > 0.5) and negative predictive value (NPV) for non-MoRF predictions (probability < 0.5) against binned prediction probabilities generated by MoRFpred on the test dataset. The PPV is the percentage of correctly predicted MoRF residues and NPV is the percentage of correctly predicted non-MoRF residues and they quantify the predictive performance of MoRFpred when it predicts MoRF and non-MoRF residues, respectively. The non-MoRF (negative) predictions for the low probabilities between 0 and 0.25, which account for 20% of all predictions, have substantially higher NPV when compared with the predictions with higher probabilities, e.g. in 0.4 to 0.5 range. The same is true for the MoRF (positive) predictions. We observe that for high probabilities between 0.7 and 1, our method provides a much higher PPV when compared with the predictions for probabilities closer to 0.5 (between 0.5 and 0.6). To sum up, we show that predictions with probabilities farther away from the 0.5, which is the threshold to differentiate between MoRF and non-MoRF residues, are characterized by higher predictive quality. This means that a user should be more confident with the predictions associated with either low or high probabilities.

Suppl. Table S1. Summary of datasets.

Dataset name	Number of proteins	Number of MoRF residues	Number of non-MoRF residues	Notes
TRAINING	421	5396	240588	Dataset used to develop the method (to perform feature selection and parameterize the prediction algorithm) based on 5-fold cross validation protocol.
TEST	419	5153	253676	Dataset developed using PDB depositions from before April 2008, which is used to evaluate and compare our method with the existing predictors. Shares up to 30% similarity with the training dataset.
TEST2012	45	626	36907	Dataset developed using PDB depositions from 2012, which is used to evaluate and compare our method with the existing predictors. Shares up to 30% similarity with the training dataset.
EXPER2008-12	8	210	2479	Dataset developed using experimentally validated data extracted from publications between 2008 and 2012 (Nagulapalli et al., 2012; Ganguly et al., 2012; Matsumura et al., 2011; Reingewertz et al., 2011; Serrière et al., 2011; Wang et al., 2011a; Wang et al., 2011b; Garcia-Pino et al., 2008). This dataset is used to evaluate and compare our method with the existing predictors. Shares up to 30% similarity with the training dataset.
NEGATIVE	28	Not applicable	9211	Dataset developed using PDB depositions between January 2010 and March 2012, which is used to evaluate and compare our method with the existing predictors.

Suppl. Table S2. Description of features considered in building the proposed MoRFpred method. The features are grouped into two types: per residue and aggregated. Each of these types is further sub-divided based on the type of information they utilize. For features which calculate the difference between the outside and inner windows, the size of the inner window is specified by parameter *w* and size of the outside window = 25–*w*. The difference is calculated by subtracting the value for the inner window from the value for the outside window.

Feature type	Input type		Description	Window size	Number of features
Per residue	Disorder, RSA, B-factor		For each prediction method, we include binary values and probabilities in a window. 7 (methods: 5 disorder + RSA + B-factor) * 25 (window size) * 2 (binary and probability) = 350 features.	$w = 25$	350
	PSSM generated with PSI-BLAST		For each residue a matrix of size 7*20 = 140 is included in the features where each row is a window of size 7 centered on the main residue and each column contains values corresponding to different amino acids.	$w = 7$	140
Aggregated	Disorder	Average probability	Average of probability over the window of size *w*.	$w = \{2*n+1 \mid n=2,..,12\}$	170
		Content	Content of binary prediction over the window of size *w*.	$w = \{2*n+1 \mid n=2,..,12\}$	
		Average difference	Difference of probability averages in an inside window of size *w* and an outside window of size 25.	$w = \{2*n+1 \mid n=2,..,7\}$	
		MinMax average	Difference of minimum average in an inside window of size *w* from maximum average in an outside window of size 25.	$w = \{2*n+1 \mid n=2,..,7\}$	
	Relative solvent accessibility (RSA)	Average RSA	Average of RSA values over the window of size *w*.	$w = \{2*n+1 \mid n=2,..,7\}$	24
		Standard deviation (stdv)	Standard deviation of RSA values over the window of size *w*.	$w = \{2*n+1 \mid n=2,..,7\}$	
		Content	Content of binary prediction over the window of size *w*.	$w = \{2*n+1 \mid n=2,..,7\}$	
		Stdv difference	Difference of standard deviation in an inside window of size *w* and an outside window of size 25.	$w = \{2*n+1 \mid n=2,..,7\}$	
	B-values	Minimal B-factor	Minimum of normalized B-factor over the window of size *w*.	$w = \{2*n+1 \mid n=2,..,7\}$	18
		Content	Content of binary prediction over the window of size *w*.	$w = \{2*n+1 \mid n=2,..,7\}$	
		Content difference	Difference of content in an inside window of size *w* and an outside window of size 25.	$w = \{2*n+1 \mid n=2,..,7\}$	
	AA Indices	Average	Average of amino acid index over a window of size *w*.	$w = 15$	1062
		Average difference	Difference of averages in an inside window of size *w* and outside window of size 25.	$w = 15$	

Suppl. Table S3. Comparison of results of MoRF prediction using different feature selection methods and different sampling strategies. The results are based on the cross validation on the training dataset. Rows list individual setups, which consider three sampling strategies and 3 feature selection approaches. We also use a combined feature set which implements a union of the features selected by the three selection approaches. The columns list results when evaluation is performed using the whole chain, using only the flanking region (see Section 2.2 in the main text), and the average of the two.

		Whole Sequence					Flanking Region					Average (whole and flanking)	
Sampling	Feature selection	ACC	TPR	FPR	Success rate	AUC	ACC	TPR	FPR	Success rate	AUC	avg. AUC	avg. success rate
local	Complete ranking	0.948	0.183	0.034	0.665	0.642	0.682	0.183	0.063	0.637	0.616	0.629	0.651
	Local ranking	0.788	0.391	0.203	0.748	0.632	0.650	0.391	0.218	0.696	0.632	0.632	0.722
	Success rate ranking	0.503	0.596	0.499	0.720	0.564	0.566	0.596	0.450	0.705	0.598	0.581	0.713
	Combined	0.920	0.245	0.064	0.703	0.654	0.686	0.245	0.088	0.703	0.665	0.660	0.703
random 3:1	Complete ranking	0.929	0.205	0.055	0.696	0.664	0.660	0.205	0.106	0.632	0.584	0.624	0.664
	Local ranking	0.503	0.637	0.500	0.722	0.599	0.559	0.637	0.481	0.694	0.620	0.609	0.708
	Success rate ranking	0.740	0.428	0.253	0.751	0.630	0.614	0.428	0.291	0.663	0.579	0.604	0.707
	Combined	0.931	0.225	0.053	0.696	0.674	0.679	0.225	0.088	0.691	0.611	0.643	0.694
random 2:1	Complete ranking	0.456	0.767	0.551	0.774	0.672	0.447	0.767	0.716	0.679	0.570	0.621	0.727
	Local ranking	0.504	0.599	0.498	0.698	0.572	0.577	0.599	0.434	0.698	0.614	0.593	0.698
	Success rate ranking	0.178	0.947	0.839	0.765	0.636	0.378	0.947	0.914	0.615	0.548	0.592	0.690
	Combined	0.454	0.768	0.553	0.762	0.653	0.442	0.768	0.725	0.601	0.539	0.596	0.682

Suppl. Table S4. Comparison of performance of MoRFpred before and after the addition of the alignment-based predictions. We use the best selected (using training dataset) SVM model and we train it on the training dataset. The alignment is performed against the training dataset. The results are based on the independent test dataset. Alignment generates only binary predictions and thus its AUC cannot be calculated. The two main columns list results when evaluation is performed using the whole chain and using only the flanking region (see Section 2.2 in the main text).

	Whole Sequence					Flanking Regions				
Predictor	ACC	TPR	FPR	Success rate	AUC	ACC	TPR	FPR	Success rate	AUC
SVM	0.937	0.226	0.048	0.714	0.663	0.706	0.226	0.059	0.752	0.678
SVM + Alignment	0.937	0.254	0.049	0.718	0.673	0.711	0.254	0.065	0.754	0.684
Alignment	0.980	0.039	0.001	0.043	NA	0.679	0.039	0.008	0.038	NA

Suppl. Table S5. Comparison of prediction results for the disorder predictors on the test dataset. The two main columns list results when evaluation is performed using the whole chain and using only the flanking region (see Section 2.2 in the main text for details). Statistical significance of the differences in the success rates and AUC between the MoRFpred and the disorder predictors is shown next to the success rate and AUC values, where ++ and + denote that the improvement is significant at the *p*-value < 0.01 and < 0.05, respectively. The methods are sorted in the descending order by their AUC values when evaluating on the whole sequences.

	Whole Sequence						Flanking Region					
Predictor	ACC	TPR	FPR	Success rate		AUC	ACC	TPR	FPR	Success rate		AUC
IUPredS	0.675	0.338	0.318	0.537	++	0.541 ++	0.519	0.338	0.393	0.427	++	0.471 ++
MFDp	0.385	0.720	0.622	0.592	++	0.535 ++	0.425	0.720	0.719	0.329	++	0.460 ++
Spine-D	0.496	0.631	0.507	0.513	++	0.532 ++	0.438	0.631	0.656	0.337	++	0.449 ++
IUPredL	0.607	0.416	0.389	0.499	++	0.522 ++	0.486	0.416	0.48	0.372	++	0.454 ++
DISOPRED2	0.55	0.456	0.448	0.296	++	0.507 ++	0.435	0.456	0.575	0.265	++	0.429 ++
DISOclust	0.405	0.648	0.600	0.449	++	0.499 ++	0.404	0.648	0.715	0.310	++	0.423 ++

Suppl. Table S6. Comparison of prediction results (including disorder predictors) on the test dataset when using only the predicted disordered residues (excluding MoRF residues) as the negatives. We use a majority-vote based on the predictions from Spine-D, MD, and MFDp to annotate disordered residues. The methods are sorted in the descending order by their AUC values.

Predictor	ACC	TPR	FPR	Success rate	AUC
MoRFpred	0.904	0.267	0.070	0.683	0.650
ANCHOR	0.540	0.389	0.454	0.621	0.404
MD	0.249	0.485	0.761	0.277	0.362
IUPredS	0.377	0.338	0.621	0.313	0.303
IUPredL	0.264	0.416	0.743	0.308	0.270
MFDp	0.032	0.720	0.996	0.372	0.267
DISOPRED2	0.167	0.456	0.845	0.181	0.243
DISOclust	0.072	0.648	0.952	0.229	0.237
Spine-D	0.047	0.631	0.977	0.284	0.223
α-MoRF-PredI	0.899	0.123	0.070	0.153	NA
α-MoRF-PredII	0.796	0.258	0.182	0.296	NA

Suppl. Table S7. Comparison of prediction results (including disorder predictors) on the test2012 and exper2008-12 datasets. The two main columns list results when evaluation is performed using the whole chain and using only the flanking region (see Section 2.2 in the main text for details). The methods are sorted in the descending order by their AUC values when evaluating on the whole sequences.

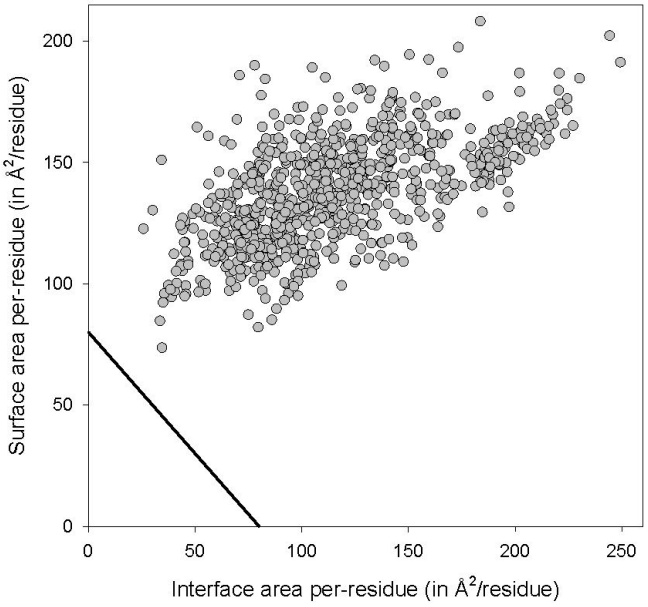
Dataset	Predictor	Whole Sequence					Flanking Region				
		ACC	TPR	FPR	Success rate	AUC	ACC	TPR	FPR	Success rate	AUC
TEST2012	MoRFpred	0.943	0.236	0.045	0.756	0.697	0.691	0.236	0.074	0.733	0.686
	MD	0.565	0.613	0.436	0.578	0.679	0.465	0.613	0.612	0.467	0.520
	ANCHOR	0.759	0.433	0.236	0.578	0.638	0.571	0.433	0.358	0.511	0.551
	IUPredS	0.708	0.449	0.287	0.600	0.634	0.529	0.449	0.430	0.422	0.498
	IUPredL	0.618	0.572	0.382	0.600	0.620	0.444	0.572	0.622	0.356	0.476
	MFDp	0.450	0.754	0.556	0.556	0.620	0.433	0.754	0.734	0.600	0.493
	Spine-D	0.482	0.720	0.522	0.467	0.605	0.413	0.720	0.746	0.467	0.476
	DISOPRED2	0.545	0.534	0.455	0.244	0.548	0.428	0.534	0.626	0.289	0.431
	DISOclust	0.411	0.653	0.593	0.556	0.512	0.407	0.653	0.721	0.400	0.455
	α-MoRF-PredI	0.955	0.091	0.030	0.133	NA	0.655	0.091	0.053	0.111	NA
	α-MoRF-PredII	0.894	0.291	0.096	0.311	NA	0.700	0.291	0.088	0.289	NA
EXPER2008-12	MoRFpred	0.867	0.210	0.077	0.750	0.636	0.647	0.210	0.071	0.875	0.637
	MD	0.328	0.690	0.702	0.500	0.616	0.412	0.69	0.767	0.375	0.525
	ANCHOR	0.521	0.548	0.481	0.500	0.556	0.440	0.548	0.629	0.750	0.492
	IUPredL	0.321	0.724	0.714	0.375	0.471	0.440	0.724	0.742	0.250	0.435
	IUPredS	0.449	0.486	0.554	0.250	0.451	0.435	0.486	0.598	0.500	0.427
	MFDp	0.221	0.919	0.839	0.500	0.337	0.431	0.919	0.883	0.500	0.353
	Spine-D	0.256	0.710	0.783	0.250	0.330	0.388	0.710	0.819	0.500	0.297
	DISOPRED2	0.295	0.481	0.720	0.125	0.310	0.369	0.481	0.702	0.250	0.298
	DISOclust	0.238	0.581	0.791	0.250	0.290	0.360	0.581	0.782	0.500	0.307
	α-MoRF-PredI	0.858	0.000	0.069	0.000	NA	0.608	0.000	0.000	0.000	NA
	α-MoRF-PredII	0.792	0.238	0.161	0.250	NA	0.586	0.238	0.190	0.250	NA

Suppl. Table S8. Comparison of prediction results for different MoRF types on the test dataset. The two main columns list results when evaluation is performed using the whole chain and using only the flanking region (see Section 2.2 in the main text). α-MorfPredI and α-MorfPredII generate only binary predictions and thus their AUC cannot be calculated. Statistical significance of the differences in the success rates and AUC between the MoRFpred and the other three methods is shown next to the success rate and AUC values, where ++, +, and = denote that the improvement is significant at the *p*-value < 0.01, at *p*-value < 0.05, and that the difference is not significant, respectively.

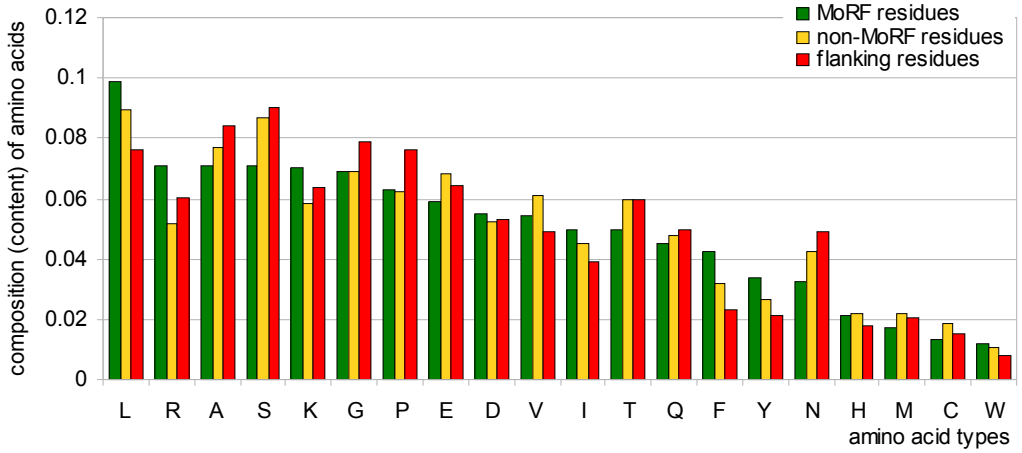
MoRF type # (%) of MoRF segments	Predictor	Whole Sequence							Flanking Region						
		ACC	TPR	FPR	Success rate		AUC		ACC	TPR	FPR	Success Rate		AUC	
Helix 97 (23%)	α-MorfPredI	0.930	0.176	0.056	0.320	++	NA		0.648	0.176	0.115	0.258	++	NA	
	α-MorfPredII	0.847	0.403	0.144	0.598	++	NA		0.677	0.403	0.186	0.546	++	NA	
	ANCHOR	0.623	0.545	0.376	0.866	+	0.635	++	0.657	0.545	0.286	0.876	=	0.662	++
	MoRFpred	0.937	0.357	0.052	0.907		0.747		0.741	0.357	0.066	0.907		0.763	
	Alignment only	0.982	0.063	0	0.093		NA		0.68	0.063	0.010	0.093		NA	
Sheet 15 (4%)	α-MorfPredI	0.961	0.099	0.018	0.067	++	NA		0.697	0.099	0.009	0.067	++	NA	
	α-MorfPredII	0.936	0.224	0.046	0.200	++	NA		0.706	0.224	0.058	0.200	++	NA	
	ANCHOR	0.866	0.168	0.117	0.333	++	0.506	+	0.681	0.168	0.067	0.600	++	0.554	++
	MoRFpred	0.934	0.149	0.047	0.600		0.654		0.685	0.149	0.052	0.733		0.698	
	Alignment only	0.974	0.043	0.004	0.067		NA		0.681	0.043	0.006	0.067		NA	
Coil 288 (69%)	α-MorfPredI	0.954	0.084	0.027	0.094	++	NA		0.677	0.084	0.039	0.08	++	NA	
	α-MorfPredII	0.912	0.175	0.073	0.198	++	NA		0.667	0.175	0.096	0.156	++	NA	
	ANCHOR	0.811	0.308	0.178	0.528	++	0.595	++	0.630	0.308	0.216	0.583	++	0.555	++
	MoRFpred	0.937	0.206	0.048	0.653		0.634		0.697	0.206	0.067	0.701		0.638	
	Alignment only	0.978	0.029	0.002	0.028		NA		0.68	0.029	0.008	0.021		NA	
Complex 19 (4%)	α-MorfPredI	0.946	0.332	0.043	0.389	++	NA		0.663	0.332	0.157	0.278	++	NA	
	α-MorfPredII	0.860	0.467	0.133	0.500	++	NA		0.708	0.467	0.162	0.500	++	NA	
	ANCHOR	0.590	0.640	0.411	0.833	++	0.658	+	0.645	0.640	0.352	0.722	++	0.692	++
	MoRFpred	0.940	0.369	0.050	0.889		0.760		0.736	0.369	0.066	0.833		0.767	
	Alignment only	0.982	0	0.001	0		NA		0.649	0	0	0		NA	

Suppl. Table S9. Comparison of prediction results for immune response-related and other proteins on the test dataset. The two main columns list results when evaluation is performed using the whole chain and using only the flanking region (see Section 2.2 in the main text). α-MorfPredI and α-MorfPredII generate only binary predictions and thus their AUC cannot be calculated. Statistical significance of the differences in the success rates and AUC between the MoRFpred and the other three methods is shown next to the success rate and AUC values, where ++, +, and = denote that the improvement is significant at the *p*-value < 0.01, at *p*-value < 0.05, and that the difference is not significant, respectively.

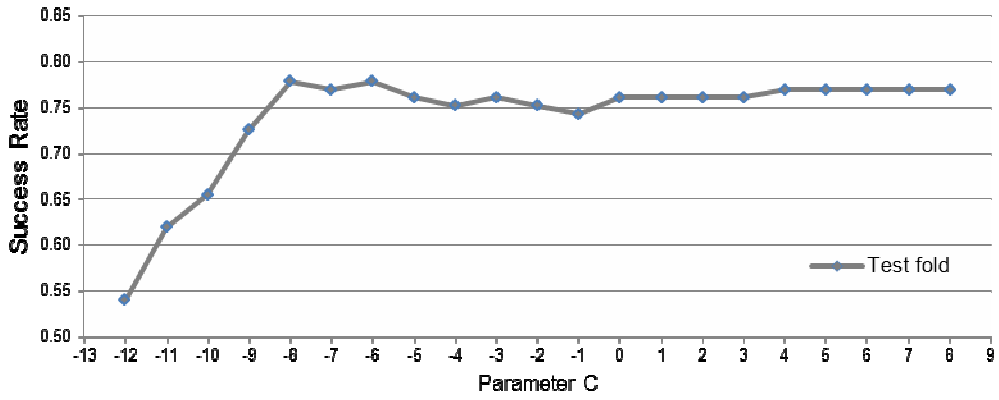
MoRF type # (%) of MoRF segments	Predictor	Whole Sequence							Flanking Region						
		ACC	TPR	FPR	Success rate		AUC		ACC	TPR	FPR	Success rate		AUC	
Immune response-related 74 (18%)	α-MoRF-PredI	0.958	0	0.019	0	++	NA		0.691	0	0	0	++	NA	
	α-MoRF-PredII	0.921	0.016	0.057	0.027	++	NA		0.681	0.016	0.021	0.014	++	NA	
	ANCHOR	0.824	0.214	0.161	0.5	=	0.573	++	0.654	0.214	0.149	0.635	=	0.569	+
	MoRFpred	0.932	0.156	0.049	0.581		0.568		0.716	0.156	0.033	0.662		0.583	
	Alignment Only	0.976	0	0	0		NA		0.691	0	0	0		NA	
Other 345 (82%)	α-MoRF-PredI	0.945	0.143	0.039	0.191	++	NA		0.664	0.143	0.077	0.157	++	NA	
	α-MoRF-PredII	0.885	0.298	0.104	0.362	++	NA		0.672	0.298	0.143	0.316	++	NA	
	ANCHOR	0.729	0.419	0.265	0.635	++	0.608	++	0.638	0.419	0.253	0.664	++	0.595	++
	MoRFpred	0.937	0.273	0.049	0.748		0.692		0.711	0.273	0.072	0.774		0.701	
	Alignment Only	0.98	0.045	0.001	0.052		NA		0.677	0.045	0.009	0.046		NA	



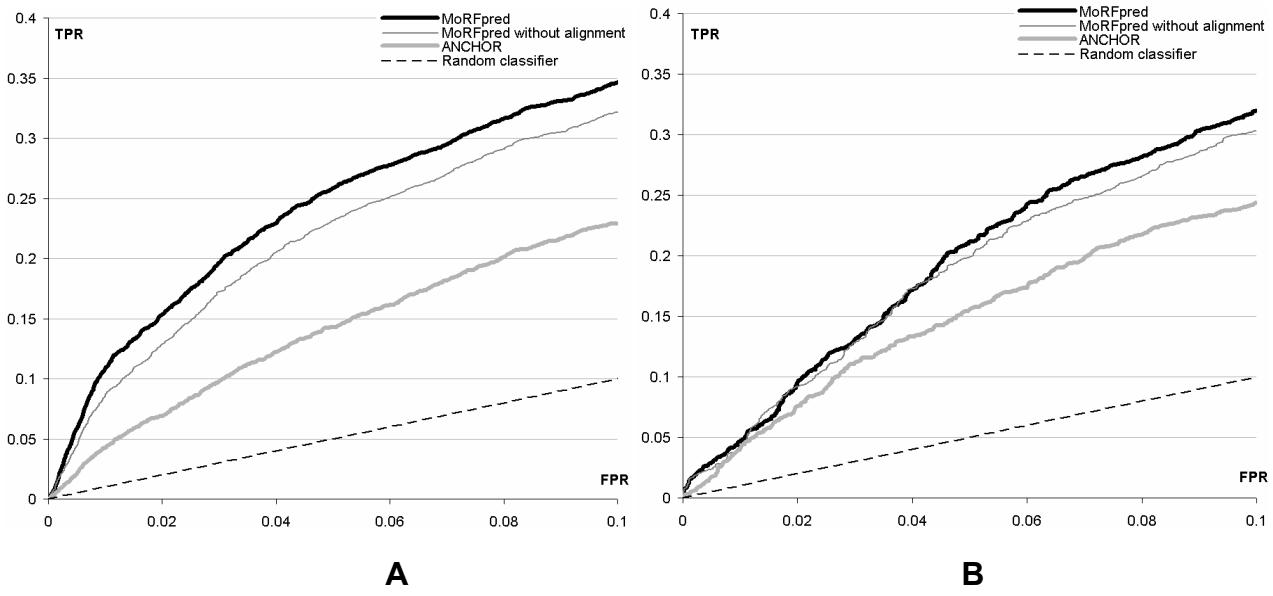
Suppl. Fig. S1. Gunasekaran-Tsai-Nussinov (Gunasekaran et al., 2004) graph for the 842 MoRFs. The plot provides a scale that measures confidence with which one can say whether a protein is ordered or disordered. The farther the point, which corresponds to a given chain, is from the dividing black line (boundary), the greater the confidence with which a protein can be classified into either of the classes. Points above the line correspond to disordered chains.



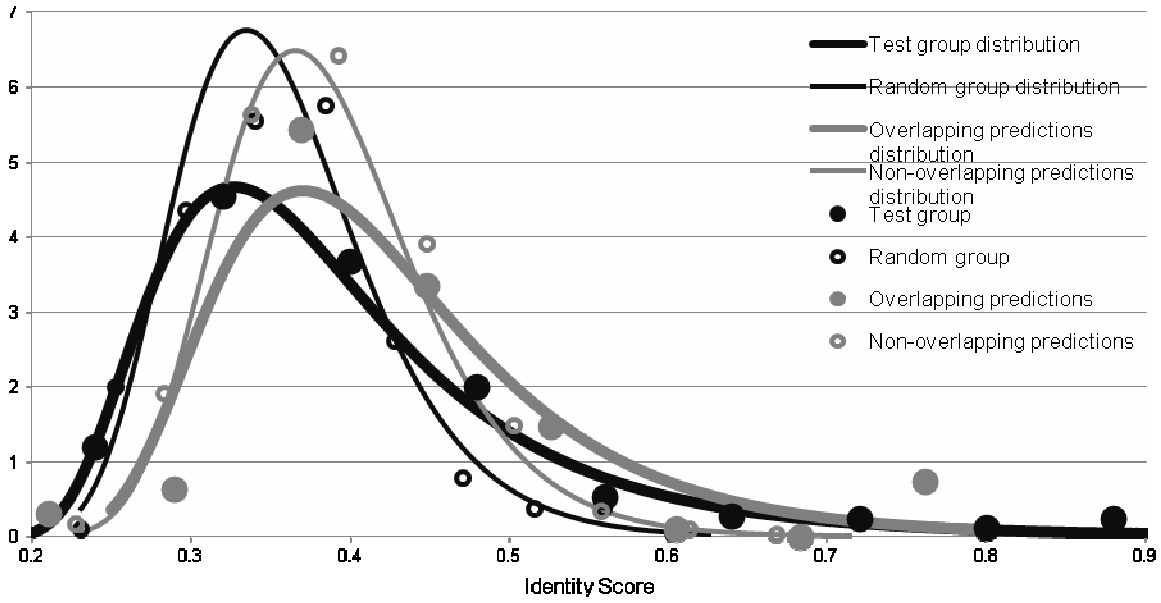
Suppl. Fig. S2. Amino acids composition (fraction of AA of a given type) among the MoRFs residues (green bars), non-MoRF residues (orange bars), and flanking residues (red bars) on the training dataset. The amino acids are sorted in descending order by the composition for MoRF residues.



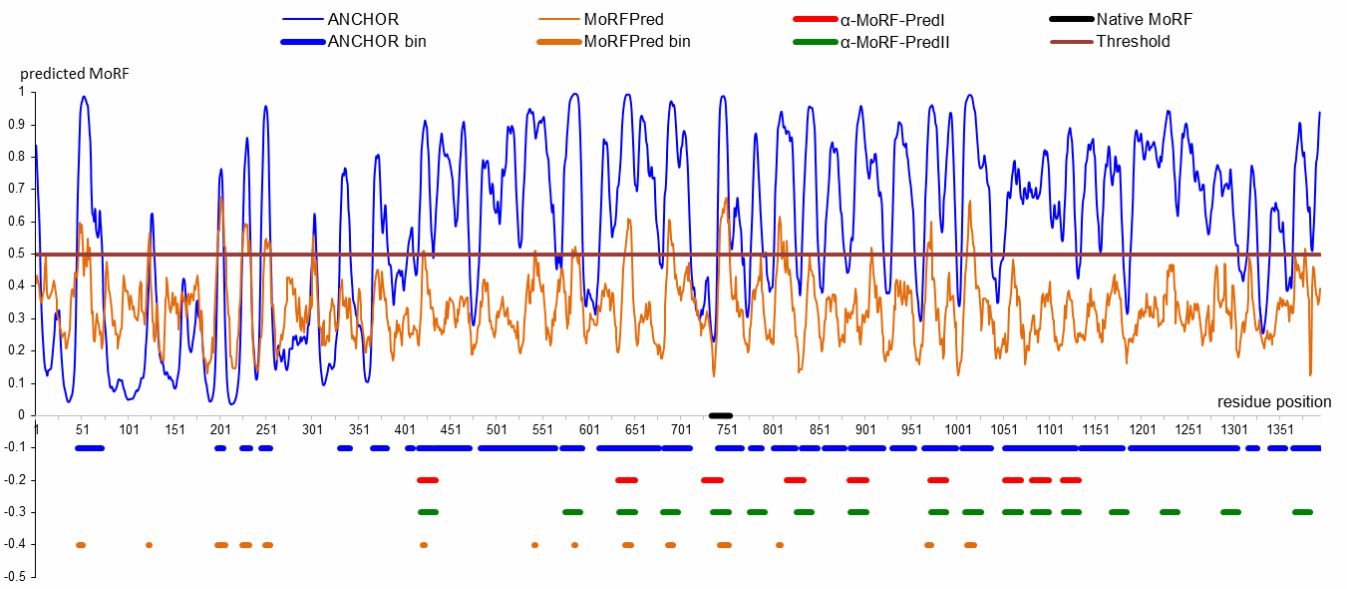
Suppl. Fig. S3. Results of parameterization of parameter *C* for the SVM classifier that uses the combined feature set selected based on the local sampling, which are based on the 4+1-fold cross validation on the training dataset. The vertical axs represent success rate and horizontal axis shows $\log_2 C$.



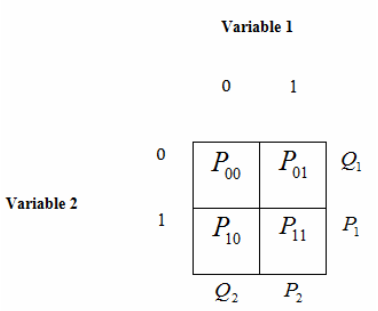
Suppl. Fig. S4. Comparison of ROCs for MoRFpred and ANCHOR on the test dataset. Panel A compares ROCs for when evaluations is performed using the whole sequences (the same as Figure 2 in the main text) and panel B when using the flanking region. The ROC curves are provided for the FPR < 0.1.



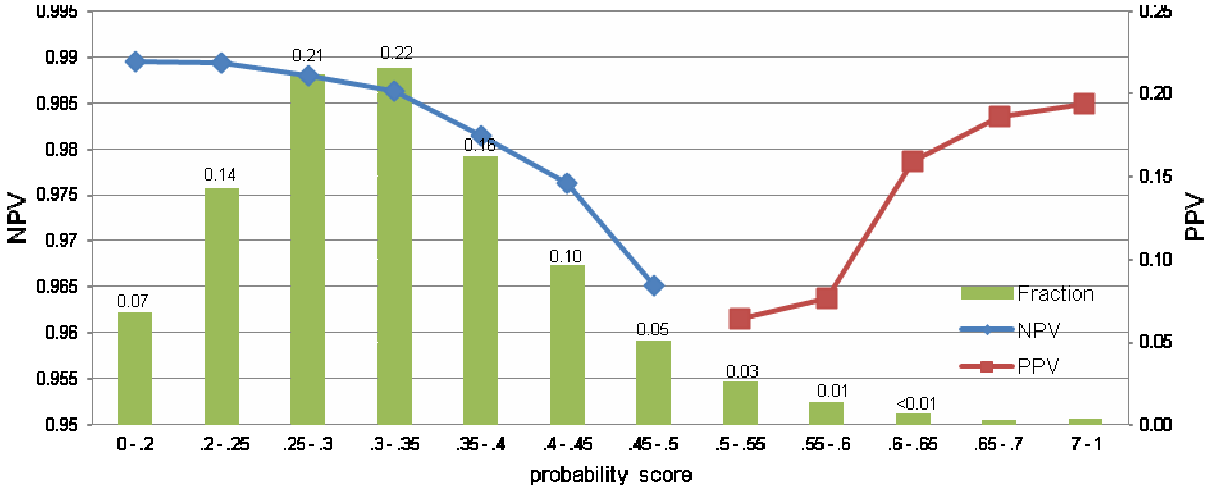
Suppl. Fig. S5. Comparison of sequence similarity between native and predicted MoRF segments. The figure includes similarity between the native MoRFs in the test dataset (test group), the random segments in the test dataset (random group), the MoRFs predicted by MoRFpred in the test dataset which overlap with the native MoRFs (overlapping predictions), the MoRFs predicted by MoRFpred in the test dataset which do not overlap with the native MoRFs (non-overlapping predictions that correspond to false positive predictions) and the native MoRFs in the training dataset. The distributions, which are based on the Pearson 5 function, were fitted using EasyFit. The *x*-axis shows the similarity between the segments measured with EMBOSS needle and *y*-axis shows the relative number of segments.



Suppl. Fig. S6. Prediction of MoRF residues for the transcriptional intermediary factor-2 isoform 2 protein by ANCHOR (blue lines), MoRFpred (orange lines), α-MoRF-PredI (thick red line), and α-MoRF-PredII (thick green line) predictors. The *x*-axis shows positions in the protein sequence. Probability values are only available for ANCHOR and MoRFpred and are shown by thin blue and orange lines, respectively, at the top of the figure. The cut-off of 0.5 to convert probabilities into binary predictions for ANCHOR and MoRFpred is shown using a brown horizontal line. The native MoRF regions are annotated using black horizontal line. The binary predictions from ANCHOR, α-MoRF-PredI, α-MoRF-PredII and MoRFpred are denoted using horizontal lines at the bottom of the figure in blue (at the -0.1 point on the *y*-axis), red (at the -0.2), green (at the -0.3), and orange (at the -0.4), respectively.



Suppl. Fig. S7. Matrix that defines combinations of values of two binary variables. In case of the MoRF prediction, variable 1 corresponds to the native MoRF annotations and variable 2 could be an input feature or a binary MoRF prediction.



Suppl. Fig. S8. Relation between predictive quality and the magnitude of the probabilities generated by MoRFpred on the test dataset. Values of probabilities are binned and shown on the *x*-axis. The left *y*-axis shows the percentage of correctly predicted non MoRF residues (NPV), which quantifies predictive quality when probabilities are below 0.5. The right *y*-axis corresponds to the percentage of correctly predicted MoRF residues (PPV), which evaluates predictive quality when probabilities are above 0.5. The bars indicate the fraction of all residues in the test dataset for a given range of the probability. We note that majority of the residues are non-MoRFs and thus the bars for the probabilities above 0.5 are low.

REFERENCES

Ernest, C., et al. (1991) Phi/Phimax: review and synthesis. *Educational and Psychological Measurement* 51, 821-828

Ganguly, D., et al. (2012) Synergistic folding of two intrinsically disordered proteins: searching for conformational selection. *Mol Biosyst*. 8(1), 198-209.

Garcia-Pino, A., et al., (2008). Doc of prophage P1 is inhibited by its antitoxin partner Phd through fold complementation. *J Biol Chem*. 283(45), 30821-7.

Gunasekaran, K., et al. (2004) Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. *J Mol Biol* 341, 1327-41

Nagulapalli, M., et al. (2012) Recognition pliability is coupled to structural heterogeneity: a calmodulin intrinsically disordered binding region complex. *Structure* 20(3), 522-33.

Matsumura, H., et al. (2011) Structure basis for the regulation of glyceraldehyde-3-phosphate dehydrogenase activity via the intrinsically disordered protein CP12. *Structure* 19(12), 1846-54.

Reingewertz, T.H., et al. (2011) Mechanism of the interaction between the intrinsically disordered C-terminus of the pro-apoptotic ARTS protein and the Bir3 domain of XIAP. *PLoS One* 6(9), e24655.

Serrière, J., et al. (2011) Fab'-induced folding of antigenic N-terminal peptides from intrinsically disordered HIV-1 Tat revealed by X-ray crystallography. *J Mol Biol.* 405(1), 33-42.

Tate, R.F. (1954). Correlation between a discrete and a continuous variable. Point-biserial correlation. *Annals of Mathematical Statistics* 25, 603-7.

Wang, T., et al. (2010a) Binding-induced folding of prokaryotic ubiquitin-like protein on the Mycobacterium proteasomal ATPase targets substrates for degradation. *Nat Struct Mol Biol.* 17(11), 1352-7.

Wang, X., et al. (2010b) A large intrinsically disordered region in SKIP and its disorder-order transition induced by PPIL1 binding revealed by NMR. *J Biol Chem*. 285(7), 4951-63.