Supplementary material for OPAL+: Length-specific MoRF prediction in intrinsically disordered protein sequences

Ronesh Sharma[1,2, §], Alok Sharma[1,3,4,5, §] , Gaurav Raicar[1], Tatsuhiko Tsunoda[3,4,6] and Ashwini Patil[7, *]

[1]School of Engineering and Physics, The University of the South Pacific, Suva, Fiji, [2]School of Electrical and Electronics Engineering, Fiji National University, Suva, Fiji, [3]Laboratory of Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama 230-0045, Japan, [4]Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University (TMDU), Tokyo 113-8510, Japan, [5]Institute for Integrated and Intelligent Systems, Griffith University, Nathan, Brisbane, QLD, Australia, [6]CREST, JST, Tokyo 113-8510, Japan, [7]Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo 108-8639, Japan.

Supplementary Materials and Methods

An overview of the length-specific MoRF prediction scheme is given in Figures 1 and S2. Four different models were constructed to predict MoRFs in disordered protein sequences, each trained to target different MoRF lengths. We partitioned MoRFs into 4 groups based on their lengths, from 5 to 9 residues, 10 to 14 residues, 15 to 19 residues and 20 to 24 residues. Table S1 gives the number of MoRFs in the training and test sets for each group.

In the training step, features were computed from MoRFs and non-MoRFs. Since the TRAIN set has a single MoRF region and the number of non-MoRF residues is greater compared to the number of MoRF residues, balanced sampling is required. To enable balanced sampling, we extracted upstream/downstream flanking amino acid residues along with the MoRF region as a positive sample. We then extracted the same size of the negative sample from a non-MoRF region. Suppose P is a protein database with n protein sequences, where $P = \{p_1, p_2, p_3, \cdots \cdots p_n\}$.

These n protein sequences have MoRF regions of lengths given as:

$$5i \leq m_j^{g_i} \leq 5i + 4 \quad (1)$$

where i varies from 1 to 4, $m_j^{g_i}$ refers to the j-th MoRF in group $g_i$and MoRF groups are defined as:

$$M_{g_i} = \{m_1^{g_i}, m_2^{g_i}, m_3^{g_i}, \cdots \cdots m_{n_i}^{g_i}\} \quad (2)$$

where n_i is the total number of MoRFs in the group. From equations (1) and (2), MoRF groups are interpreted as:

$$\text{MoRF groups} = \{M_{g_1}, M_{g_2}, M_{g_3}, M_{g_4}\} \quad (3)$$

Similarly, non-MoRF groups are given as:

Non-MoRF groups $= \{N_{g_1}, N_{g_2}, N_{g_3}, N_{g_4}\}$ (4)

where M_{g_i} and N_{g_i} refers to the MoRF and nonMoRF groups, respectively, for i values ranging from 1 to 4.

For each length-specific model, we computed bigram feature vectors [1] from each MoRF and non-MoRF group by utilizing steps of BigramMoRF method described in Sharma *et al*., [2] and using the structural attributes predicted using spider2 [3]. The bigram features from *k*-th attribute to *l*-th attribute of a protein sequence is computed as follows:

$$B_{k,l} = \frac{1}{L}\sum_{i=1}^{L-1} S_{i,k}\ S_{i+1,l}\ (1 \le k \le q \text{ and } 1 \le l \le q) \quad (5)$$

where $S_{i,k}$ is the element of structural matrix *S* of size *L* by *q*, *L* is the length of a protein region and q is the number of structural attributes. Computing the bigram frequencies $B_{k,l}$ for $k = 1,2,\dots q$ and $l = 1,2,\dots,q$ would give a bigram matrix B of size $q \times q$. This matrix B can be represented as a vector form by reshaping the $q \times q$ matrix into a vector of length q^2.

Each length-specific model is trained independently as illustrated in Figure S2. During the test phase, all four length-specific models are used for scoring and the output scores are combined by taking the minimum score as the output score. Suppose the length-specific model scores for a query protein sequence of length *L* is given as:

$$Length\text{-}specific\ model\ scores_i = \{S_1^i, S_2^i, \cdots, S_j^i, \cdots, S_{L-1}^i, S_L^i\} \quad (6)$$

where S_j^i is the score of j-th residue in the query protein sequence for i-th length-specific model and i varies from 1 to 4. The combined score for j-th residue is taken as:

$$Combined\ scores_j = min\{S_j^1, S_j^2, S_j^3, S_j^4\} \quad (7)$$

Model parameters and performance measures were chosen as previously described [2]. We selected SVM classifier with RBFkernel. The C and Gamma values of the kernel were selected as 1000 and 0.0038, respectively. To select the structural attributes for each model in Figure S2, we performed successive feature selection scheme in the forward direction [4] and observed the AUC performance measure to select the highly ranked attributes for each model.

To further improve the model performance, we combined MoRFpred-plus and MoRFchibi with the proposed model, since they were constructed using complementary features and learning algorithms. To calculate the scores for each residue, we applied the common averaging principle where all scores are added and divided by the number of models used (Figure S3).

The final score calculation was performed for each residue by taking a window of scores consisting of the residue score itself and the score of its z flanking residues on either side [2]. Suppose the scores of the residues in a query protein sequence is given as:

Query sequence scores $= \{S_1, S_2, \cdots, S_j, \cdots, S_{L-1}, S_L\}$ (8)

where S_j is the score of j-th residue in the query protein sequence of length L. The window scores are defined as:

$$Window\ scores_j = \begin{cases} \{S_1, S_2, \cdots\cdots S_{j+z}\}, & j \leq z \\ \{S_{j-z}, \cdots\cdots, S_{j-1}, S_j, S_{j+1}, \cdots\cdots, S_{j+z}\}, & z < j \leq L - z \\ \{S_{j-z}, \cdots\cdots, S_{L-1}, S_L\}, & j > L - z \end{cases} \quad (9)$$

where z is the flank size and j varies from 1 to L. Schematic illustration of extracting window scores from the query protein sequence is shown in Figure S4. The output score is computed as follows:

$$\text{Output score}_j = (\max(\text{Window scores}_j) + \text{median}(\text{Window scores}_j))/2 \quad (10)$$

where $j = 1,2, \ldots, L$ and L is the length of the query protein sequence. The value of flank size z for each for the model used in this study was evaluated for its effect on the prediction performance. Using the output scores of the query protein sequences, the MoRF regions are defined as:

$$\text{MoRF regions} = \text{Output score}_j > T, \quad (j = 1,2, \ldots, L) \quad (11)$$

where T is the threshold score.

Supplementary Figures

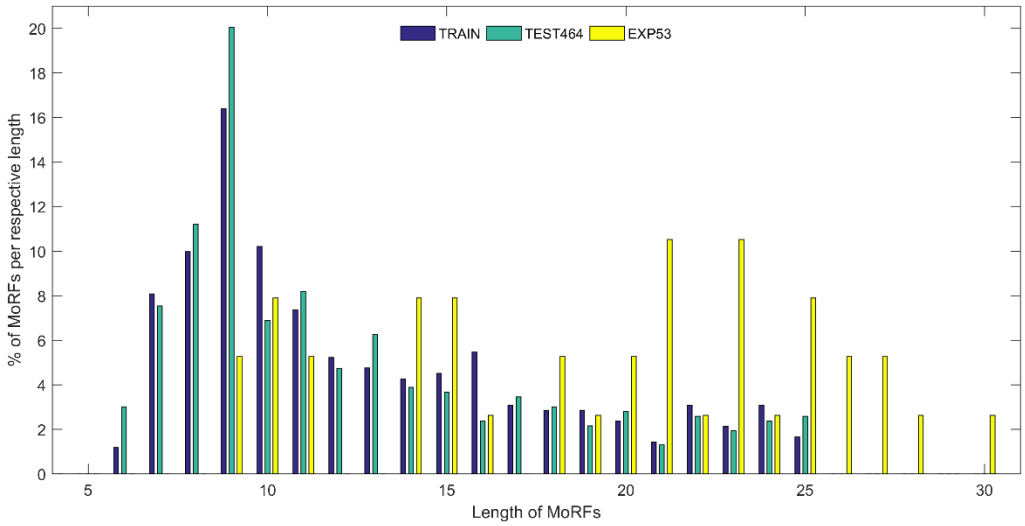

Figure S1: Percentage of MoRFs for MoRFs of specific length in TRAIN, TEST464 and EXP53SHORT sets.

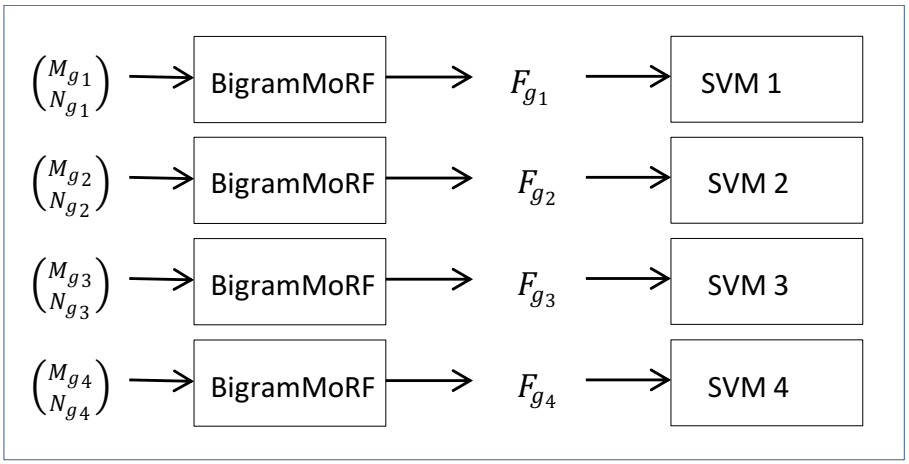

Figure S2: Training length-specific model independently. Bigram feature vectors (F_{g_1}, F_{g_2}, F_3, F_{g_4}) for each MoRF (M_{g_i}) and non-MoRF (N_{g_i}) group is extracted using BigramMoRF method and 4 different models are trained.

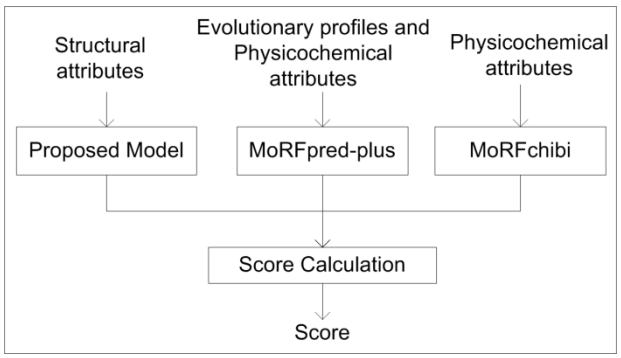

Figure S3: Combined MoRF model. In the score calculation, all the model scores are added and divided by the number of models used.

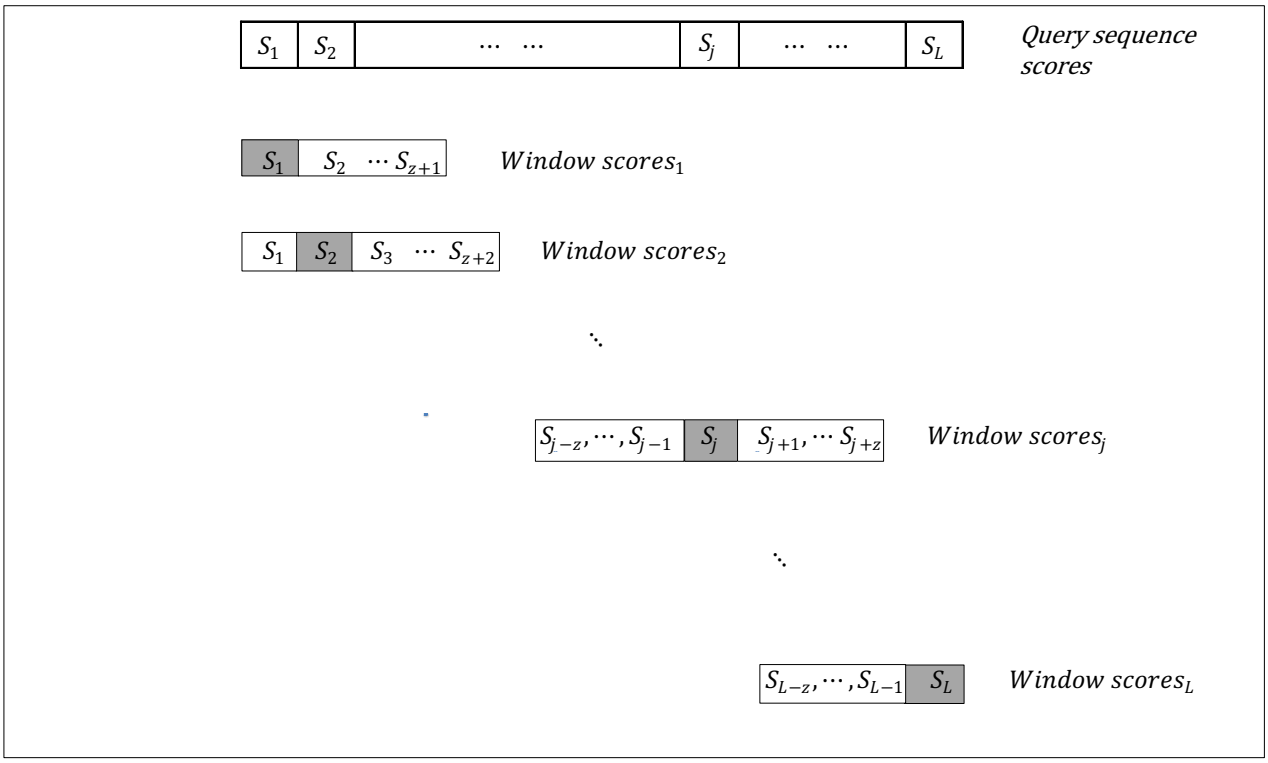

Figure S4: Schematic illustration of extracting window scores from a query sequence. S_j is the score of the j-th residue in the query sequence and L refers to the length of the query protein sequence.

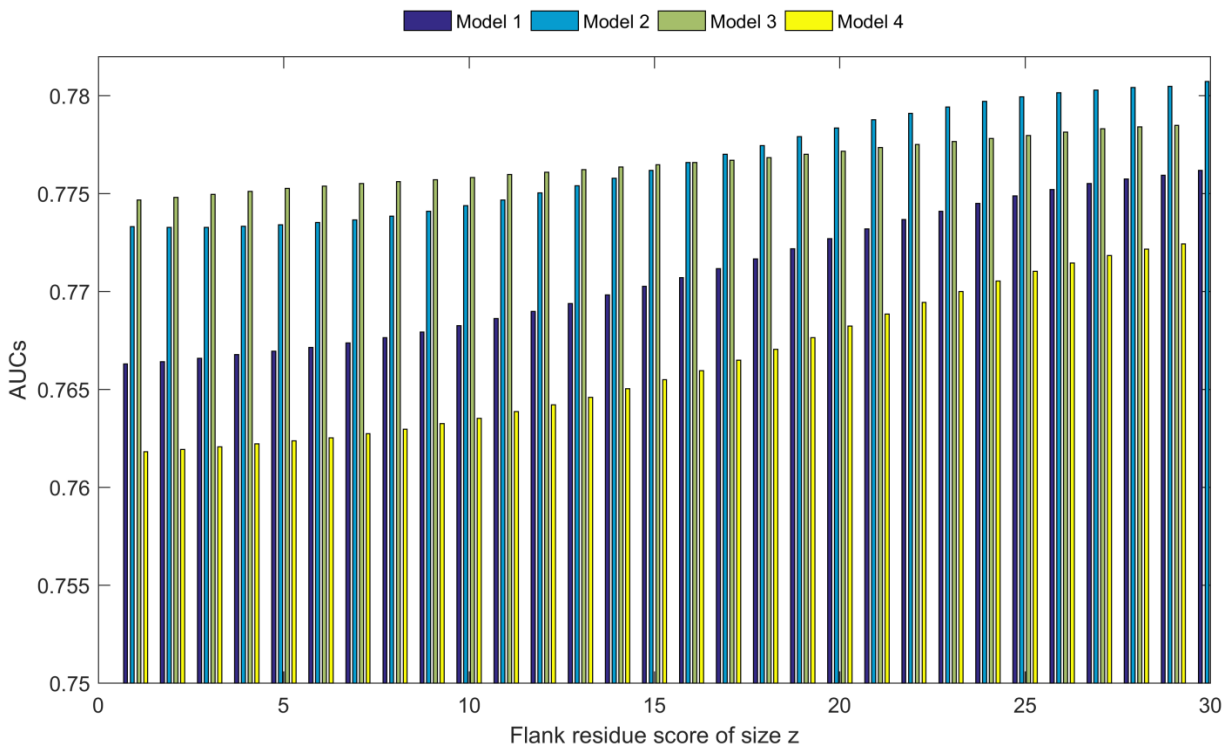

Figure S5: AUCs for varying the value of flank size, *z*, from 1 to 30

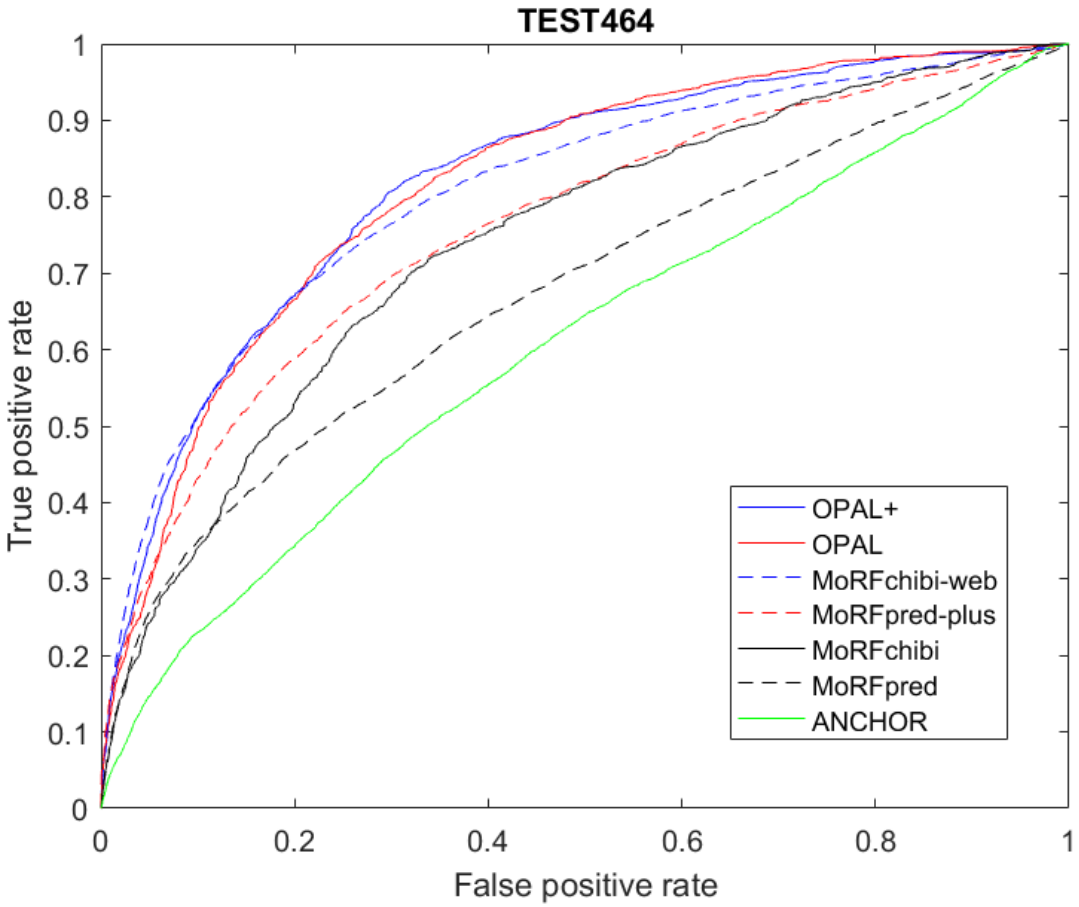

Figure S6: AUC curves generated using TEST464 set for various MoRF predictors.

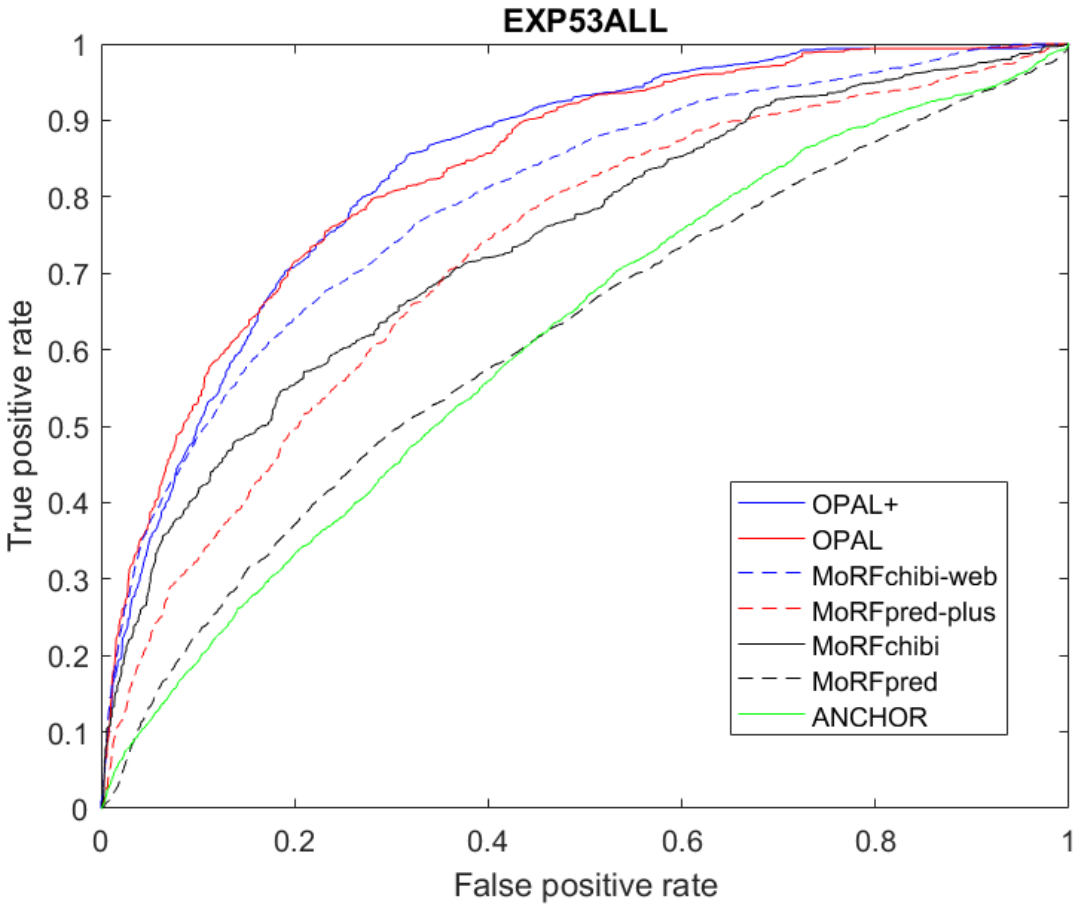

Figure S7: AUC curves generated using EXP53ALL for various MoRF predictors.

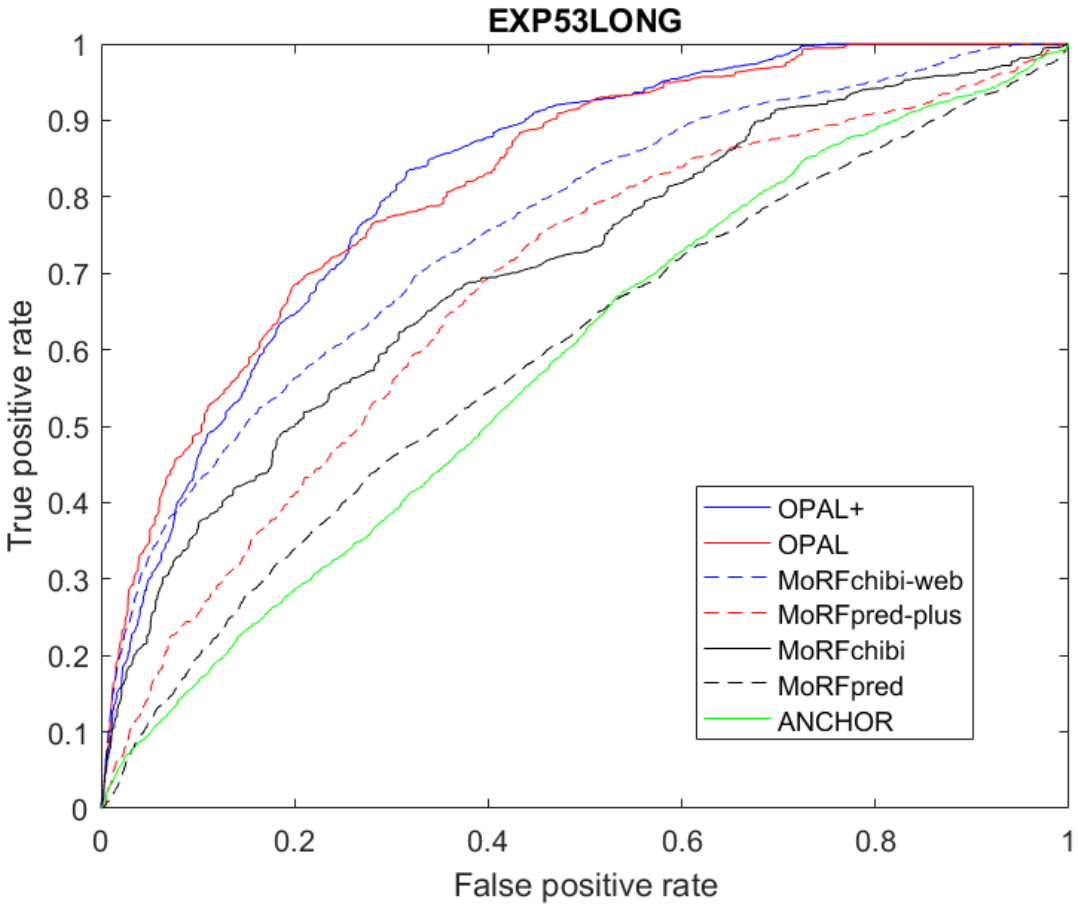

Figure S8: AUC curves generated using EXP53LONG for various MoRF predictors.

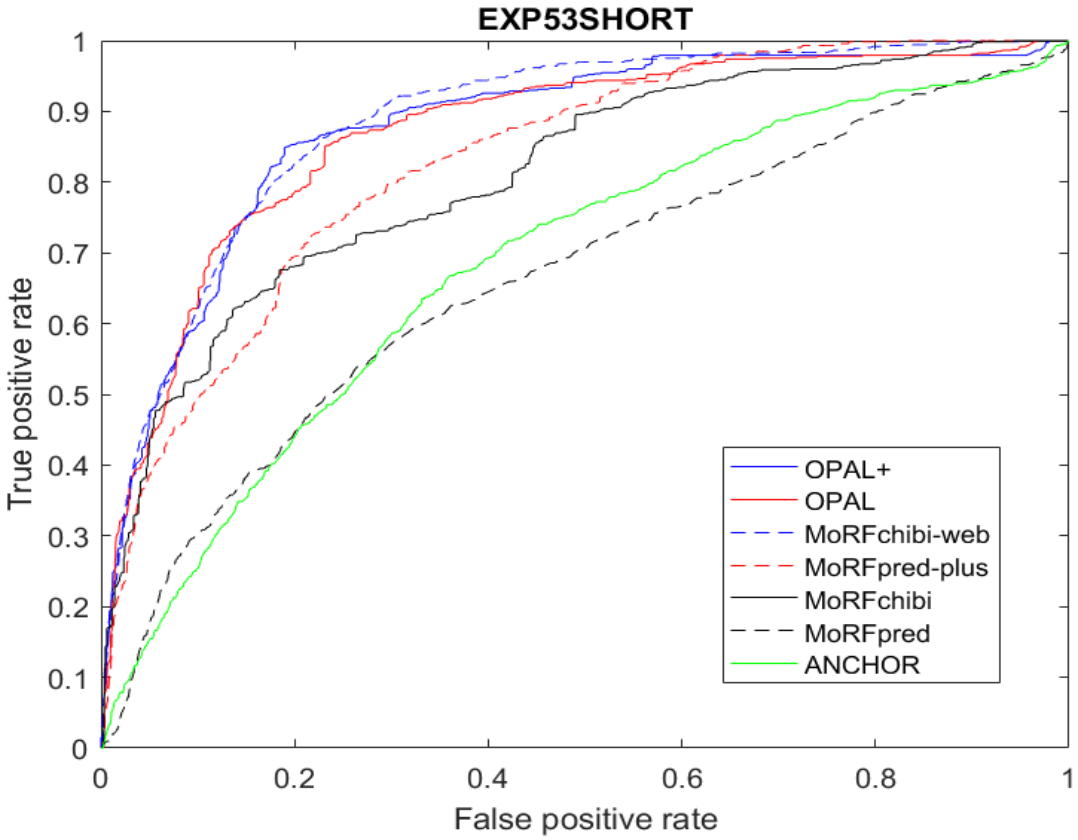

Figure S9: AUC curves generated using EXP53SHORT for various MoRF predictors.

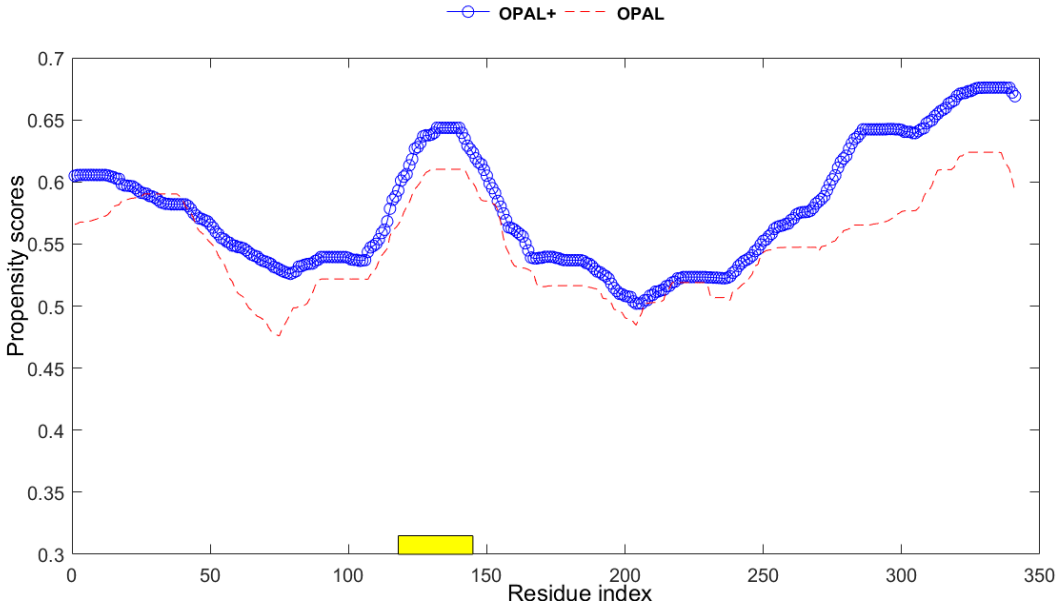

Figure S10: Propensity scores for rat the protein, Creb1 (P15337). The MoRF position is marked in yellow. OPAL+ scores are higher in the verified MoRF region.

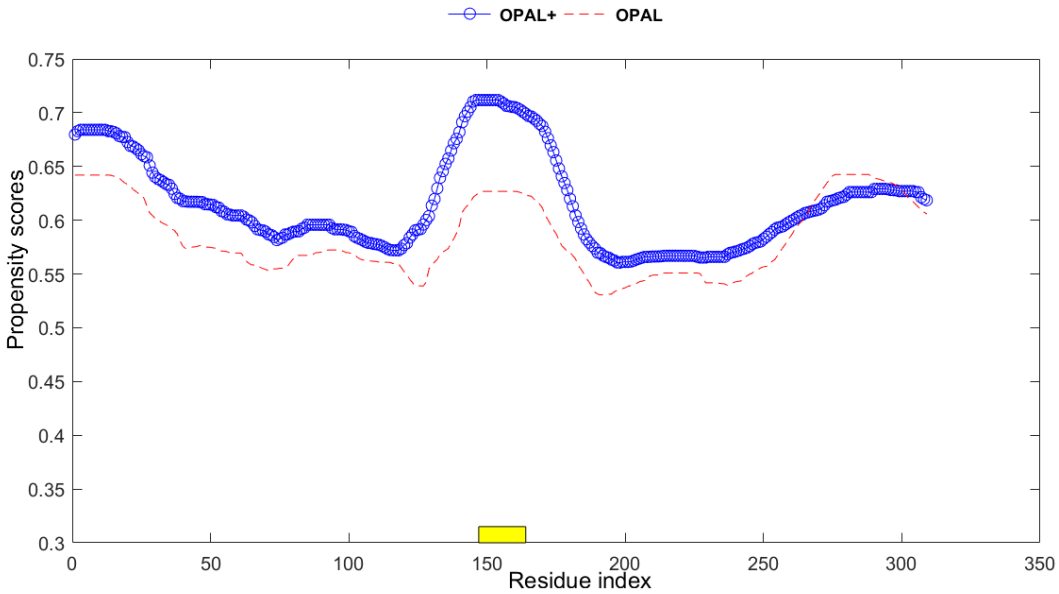

Figure S11: Propensity scores for the mouse protein, Marcs (P26645). The MoRF position is marked in yellow. OPAL+ scores are higher in the verified MoRF region, and lower at the C terminal region where MoRFs are not present.

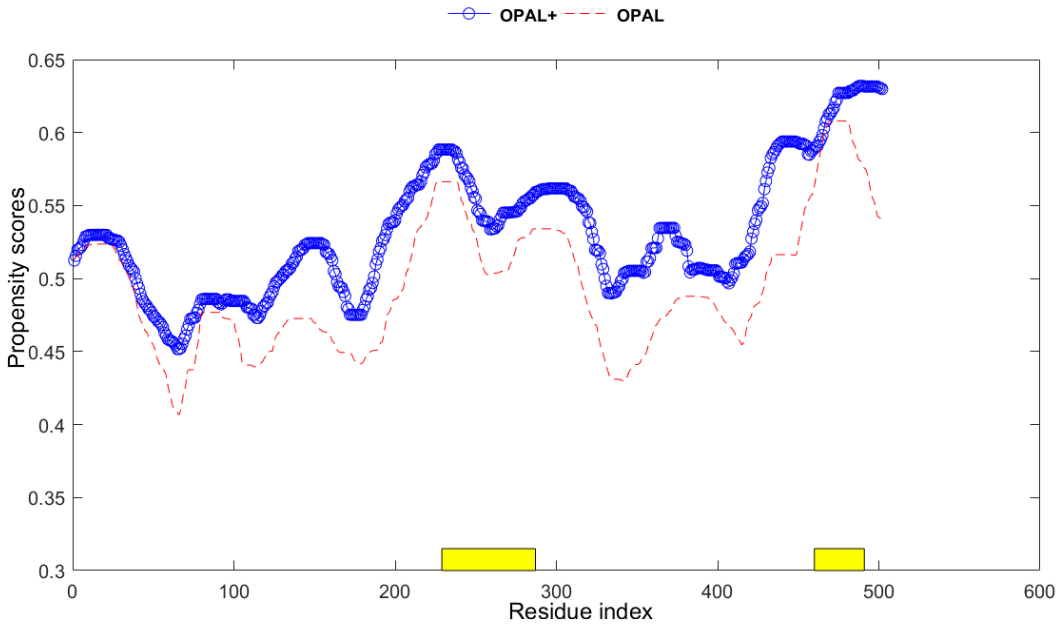

Figure S12: Propensity scores for the human protein, Wasp (P42768). The MoRF positions are marked in yellow. OPAL+ predicts MoRFs more accurately.

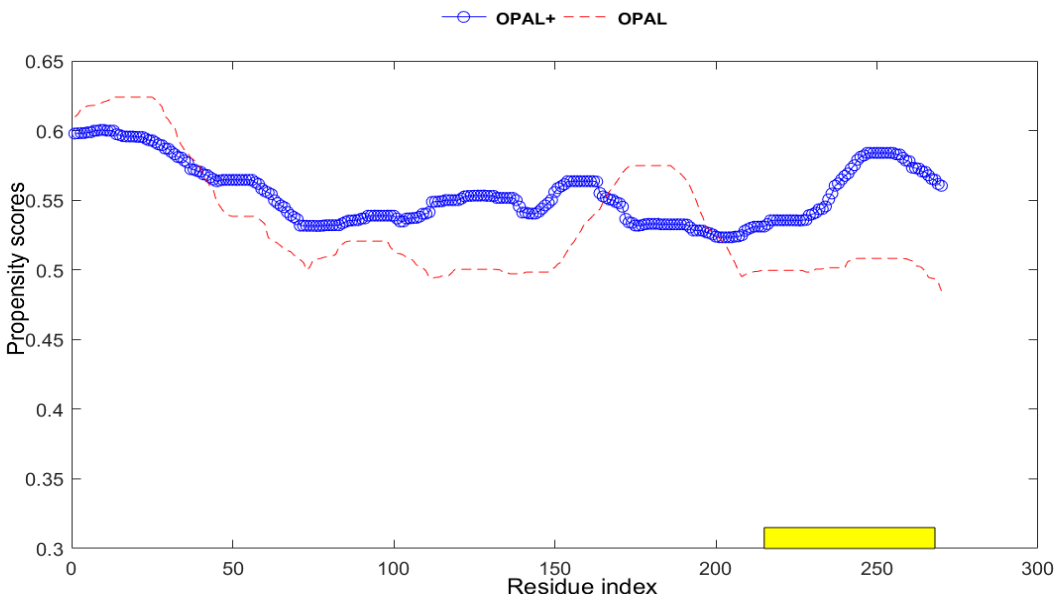

Figure S13: Propensity scores for the human protein, Cited2 (Q99967). The MoRF position is marked in yellow. OPAL scores are higher at the N terminal region though this region does not contain MoRFs.

Supplementary Tables

Table S1: Details of training and test sets

Data sets		No. of Sequences	Total residues	No. of MoRF residues	No. of non-MoRF residues	Number of MoRFs for lengths:			
						5 to 9 residues	10 to 14 residues	15 to 19 residues	20 to 24 residues
Train set	TRAIN	421	245,984	5,396	240,588	150	134	79	58
Test sets	TEST	419	258,829	5,153	253,676	177	131	61	39
	TEST464	464	296,362	5,779	290,583	194	139	68	51
	EXP53	53	25,186	ALL: 2,432 SHORT: 729 (MoRF length up to 30 residues) LONG: 1703 (MoRF length greater than 30 residues)	22,754	2	8	7	12

Table S2: Precision, F-measure, accuracy and false positive rate (FPR) is given for TPR values of 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.85 for TEST464 set. Bold numbers indicate the best performance for OPAL+, compared with OPAL.

	Precision for TPR values of :						
Predictors	0.3	0.4	0.5	0.6	0.7	0.8	0.85
MoRFchibi-web	0.158	0.128	0.096	0.074	0.057	0.044	0.037
OPAL	0.102	0.096	0.088	0.072	0.060	0.047	0.043
OPAL+	**0.132**	**0.114**	**0.095**	**0.076**	0.058	**0.051**	**0.044**
	F-measure for TPR values of :						
	0.3	0.4	0.5	0.6	0.7	0.8	0.85
MoRFchibi-web	0.207	0.194	0.161	0.132	0.105	0.083	0.071
OPAL	0.152	0.155	0.150	0.129	0.110	0.089	0.081
OPAL+	**0.184**	**0.177**	**0.160**	**0.135**	0.108	**0.097**	**0.0834**
	Accuracy for TPR values of :						
	0.3	0.4	0.5	0.6	0.7	0.8	0.85
MoRFchibi-web	0.855	0.935	0.898	0.847	0.769	0.659	0.567
OPAL	0.935	0.915	0.889	0.842	0.779	0.681	0.625
OPAL+	**0.948**	**0.928**	**0.897**	**0.850**	0.774	**0.708**	**0.636**
	FPR for TPR values of :						
	0.3	0.4	0.5	0.6	0.7	0.8	0.85
MoRFchibi-web	0.031	0.054	0.093	0.147	0.229	0.347	0.438
OPAL	0.052	0.074	0.102	0.152	0.218	0.320	0.380
OPAL+	0.039	0.062	0.094	0.145	0.224	**0.293**	**0.368**

Table S3: Precision, F-measure, accuracy and FPR is given for TPR values of 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.85 for EXP53ALL set. Bold numbers indicate the best performance for OPAL+, compared with OPAL.

	Precision for TPR values of :						
Predictors	0.3	0.4	0.5	0.6	0.7	0.8	0.85
MoRFchibi-web	0.494	0.410	0.332	0.278	0.222	0.182	0.164
OPAL	0.530	0.435	0.386	0.335	0.279	0.230	0.190
OPAL+	0.444	0.390	0.350	0.307	**0.283**	**0.241**	**0.224**
	F-measure for TPR values of :						
	0.3	0.4	0.5	0.6	0.7	0.8	0.85
MoRFchibi-web	0.373	0.404	0.399	0.379	0.338	0.297	0.275
OPAL	0.384	0.416	0.436	0.429	0.399	0.380	0.310
OPAL+	0.358	0.394	0.411	0.406	**0.404**	**0.370**	**0.355**
	Accuracy for TPR values of :						
	0.3	0.4	0.5	0.6	0.7	0.8	0.85
MoRFchibi-web	0.902	0.886	0.854	0.811	0.735	0.635	0.567
OPAL	0.906	0.891	0.875	0.846	0.797	0.722	0.636
OPAL+	0.896	0.882	0.862	0.831	**0.80**	**0.737**	**0.701**
	FPR for TPR values of :						
	0.3	0.4	0.5	0.6	0.7	0.8	0.85
MoRFchibi-web	0.033	0.061	0.107	0.166	0.261	0.382	0.463
OPAL	0.029	0.056	0.085	0.128	0.193	0.285	0.387
OPAL+	0.040	0.067	0.099	0.144	**0.189**	**0.269**	**0.315**

Table S4: Precision, F-measure, accuracy and FPR is given for TPR values of 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.85 for EXP53SHORT set. Bold numbers indicate best performance for OPAL+, compared with OPAL.

	Precision for TPR values of :						
Predictors	0.3	0.4	0.5	0.6	0.7	0.8	0.85
MoRFchibi-web	0.322	0.275	0.207	0.170	0.150	0.126	0.109
OPAL	0.378	0.237	0.19	0.175	0.163	0.107	0.105
OPAL+	0.312	**0.266**	**0.214**	0.160	0.147	**0.132**	**0.125**
	F-measure for TPR values of :						
	0.3	0.4	0.5	0.6	0.7	0.8	0.85
MoRFchibi-web	0.311	0.326	0.293	0.265	0.247	0.218	0.193
OPAL	0.334	0.298	0.276	0.272	0.265	0.190	0.187
OPAL+	0.310	**0.319**	**0.30**	0.252	0.243	**0.227**	**0.217**
	Accuracy for TPR values of :						
	0.3	0.4	0.5	0.6	0.7	0.8	0.85
MoRFchibi-web	0.958	0.948	0.925	0.896	0.867	0.822	0.779
OPAL	0.963	0.941	0.917	0.900	0.879	0.784	0.771
OPAL+	0.857	**0.947**	**0.927**	0.890	0.865	**0.831**	**0.810**
	FPR for TPR values of :						
	0.3	0.4	0.5	0.6	0.7	0.8	0.85
MoRFchibi-web	0.020	0.033	0.061	0.094	0.126	0.018	0.222
OPAL	0.015	0.041	0.069	0.090	0.114	0.216	0.231
OPAL+	0.022	**0.035**	**0.058**	0.101	0.130	**0.167**	**0.191**

Table S5: Efficiency for various MoRF predictors.

Predictors	AUC in test sets (TEST464, EXP53ALL, EXP53LONG, EXP53SHORT)	Predictor speed residues/minute (r/m)		Multiple sequence alignments	Combined component predictors
		i5 4 core 3.50GHz desktop	Server		
ANCHOR	0.605,0.615,0.586,0.683	3.9*10^6	-	×	×
MoRFchibi	0.743,0.712,0.679,0.790	10.5*10^3	-	×	×
MoRFpred	0.675,0.620,0.598,0.673	-	48	√	×
MoRFpred-plus	0.724,0.712,0.670,0.821	526	-	√	×
PROMIS	0.790,0.818,0.815,0.823	220	-	√	×
MoRFchibi-web	0.805,0.797,0.758,0.886	80	588	√	√
OPAL	0.816,0.836,0.822,0.870	215	-	√	√
OPAL+	0.820,0.838,0.822,0.876	152	-	√	√

References

[1] A. Sharma, J. Lyons, A. Dehzangi, K. K. Paliwai, *Theoretical Biology* 2013, *320*, 41.
[2] R. Sharma, G. Raicar, T. Tsunoda, A. Patil, A. Sharma, *Bioinformatics* 2018, *34*, 1850.
[3] Y. Yang, R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Zhou, *Methods Mol Biol* 2017, *1484*, 55.
[4] A. Sharma, K. K. Paliwal, A. Dehzangi, J. Lyons, S. Imoto, S. Miyano, *BMC Bioinformatics* 2013, *14*, 1.