SUPPLEMENT FOR ARTICLE ENTITLED

Computational Identification of MoRFs in Protein Sequences

Nawar Malhis[1,*], and Jörg Gsponer[1,2,*]

[1]Centre for High-Throughput Biology, University of British Columbia, Vancouver, BC V6T 1Z4, Canada.

[2]Department of Biochemistry and Molecular Biology, University of British Columbia, Vancouver, BC V6T 1Z4, Canada.

1 METHODS

1.1 Initial feature selection algorithm

The initial feature selection algorithm is a multi-step iterative algorithm that selects increasing number of features from the initial set of 1,100 features. For training, the algorithm utilizes a balanced set of 8,000 samples (4,000 * 2) from SYN4000. And for testing, it relies on a balanced set of 842 samples (421 * 2) from TRAINNG, we call it TsS.

Starting with an empty set of selected features (SSF), for each step, the algorithm iterates 5,000 times to find the "best" set of 3 features that complement those in SSF. For each of these 5,000 iterations:

- Three features are selected at random from the ISF 1,100 features, excluding those in SSF.
- These 3 selected features are used with those already in SSF to train a RBF SVM with default parameters (C, Gamma).
- TsS is scored and an AUC is computed.

At the end of each step, the three features with the highest AUC are considered the "best" and permanently added to SSF.

The AUC value increases with each three added features until a point where it starts to drop. The algorithm terminates at that point.

We selected an initial set of 39 features in 13 steps.

1.2 Feature selection algorithm

The feature selection algorithm is a 5 steps iterative algorithm designed to select exactly N features out of the ISF 1,100 features.

- With SVM_T, the algorithm utilizes TRAINING in cross-validation X5 for training and testing.
- With SVM_S, it utilizes SYN4000 for training and TRAINNG for testing.

For both models, a RBF SVM is trained on a balanced sample, and then sequences are scored with either the F1 or F2[1] scoring function.

- In the first step, j=1: for each iteration i <= X[1], a set S of N features is selected at random from ISF. The model is trained, test data is scored, and a wAUC is computed. At the end of the first step, the set of features S with the highest wAUC, $wAUC_{max}$, is retained as S_{max}.
- In each of the following steps, 1< j <= 5: at each iterations i <= X[j], a number of features NF[j] <= N is selected at random from ISF, excluding those in the current S_{max}, and the remaining N – NF[j] features are selected at random from S_{max}. If the resulting wAUC is greater than $wAUC_{max}$, S_{max} is replaced by S and $wAUC_{max}$ by wAUC.
- At the end of the fifth step, the set of features with the highest wAUC, S_{max}, is projected as the selected set of features.

F1 residue scoring is used in all steps j < 5, and F2 is used in the fifth step. NF values are set such that:

$$NF(j) = \begin{cases} N, for\ j = 1 \\ NF(j-1), for\ j > 1 \end{cases}$$

The values of NF and X are assigned properly according to N, for example, at N = 12:

○ X[1] = 4,000; X[2] = 3,000; X[3] = 2,000; X[4] = 1,000; and X[5] = 20,000.

○ NF[1] = 12, NF[2] = 8, NF[3] = 5, NF[4] = 4, and NF[5] = 3.

1.3 Identifying the appropriate model complexity using TEST

Sequences in TOTAL enclose patterns that are more frequent than random. Some of these patterns, we call *information patterns*, are concentrated in MoRFs and Flanks. Additional patterns, we call *noise patterns*, are enriched in non-MoRF sections. Information patterns are used by the training process to identify MoRFs.

The effect of splitting of TOTAL on homology: In the construction of balanced training data (section 1.1 and main paper section 2.3), positive samples are always chosen from MoRFs and their surrounding Flanks. And negative samples are chosen at random from the remaining sections of the proteins. Since two thirds of the sequences in TOTAL are homologous to one or more sequences in TOTAL splitting TOTAL on homology creates a significant imbalance in the frequencies of information patterns between TRAINING and TEST. It is a zero sum with respect to TOTAL; those underrepresented information patterns in TRAINING/TEST, are overrepresented in TEST/TRAINING. We are going to call those patterns that are underrepresented in TRAINING and overrepresented in TEST *tsPatterns*. While maximizing its objective on TRAINING, the feature selection process drifts towards to or away from tsPatterns. Giving that each set of features is selected with about 30K iterations on TRAINING, even at extreme drifts towards tsPatterns, tsPatterns will continue to be under-fitted with respect to TOTAL, perhaps to a lesser degree.

These stochastic drifts generate randomness, reflected in the form of discrepancy in trends of the model performance between the wAUC of TEST and TRAINING (main paper, figures 2).

2 RESULTS

2.1 Feature selection and appropriate model complexity identification

For SVM_S, we have two candidate sets to choose from:

- First, the set with only two features, it is the smallest in size, its wAUC on TEST is the second highest and its wAUC on TRAINING is the lowest. It is interesting to note that the performance of this set on TEST is higher than on TRAINING. Theoretically, the performance on the training set should always be higher than that on the test set because the training process fits some of the training noise patterns. However, since TRAINING and TEST are not selected from the general population using the same distribution function, MoRFs in TEST happened to be more identifiable than those in TRAINING. Thus, when only few

[1] F1 and F2 were introduced in the main paper section 2.6

features are used, the training process would not be able to fit much of TRAINING noise, and its performance on TEST became higher than that on TRAINING.

- Second, the set with fourteen features, which happened to have high fitting quality. Its wAUC on TEST is the highest with a small gap between wAUC on TEST and wAUC on TRAINING.

Thus, which of these two sets performance generalizes better? After analyzing the performance of different sets of features using multiple query sequences, we concluded that sets with very few features, although can generate high overall performances, their outcome is uneven and subjective to the query sequences having appropriate values of those features in use. Therefore we chose the second set.

2.2 Selected Features

- SVM$_S$ Features

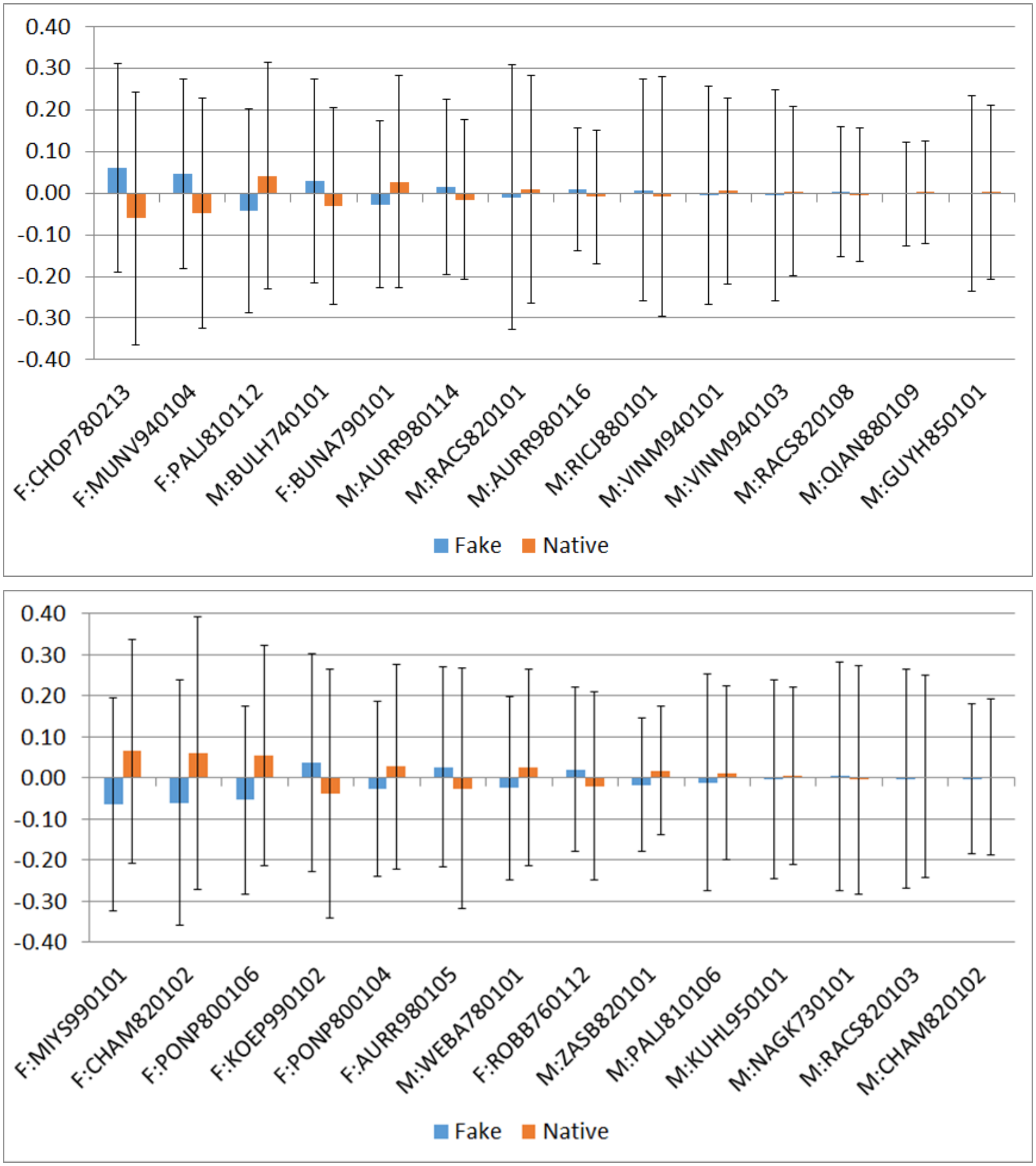
CHOP780213: Frequency of the 2nd residue in turn.
MUNV940104: Free energy in beta-strand region.
PALJ810112: Normalized frequency of beta-sheet in alpha/beta class.
BULH740101: Transfer free energy to surface.
BUNA790101: Alpha-NH chemical shifts.
AURR980114: Normalized positional residue frequency at helix termini C2.
RACS820101: Average relative fractional occurrence in A0(i).
AURR980116: Normalized positional residue frequency at helix termini Cc.
RICJ880101: Relative preference value at N".
VINM940101: Normalized flexibility parameters (B-values), average.
VINM940103: Normalized flexibility parameters (B-values) for each residue surrounded by one rigid neighbor.
RACS820108: Average relative fractional occurrence in AR(i-1).
QIAN880109: Weights for alpha-helix at the window position of 2.
GUYH850101: Partition energy.

SVM$_T$ Features

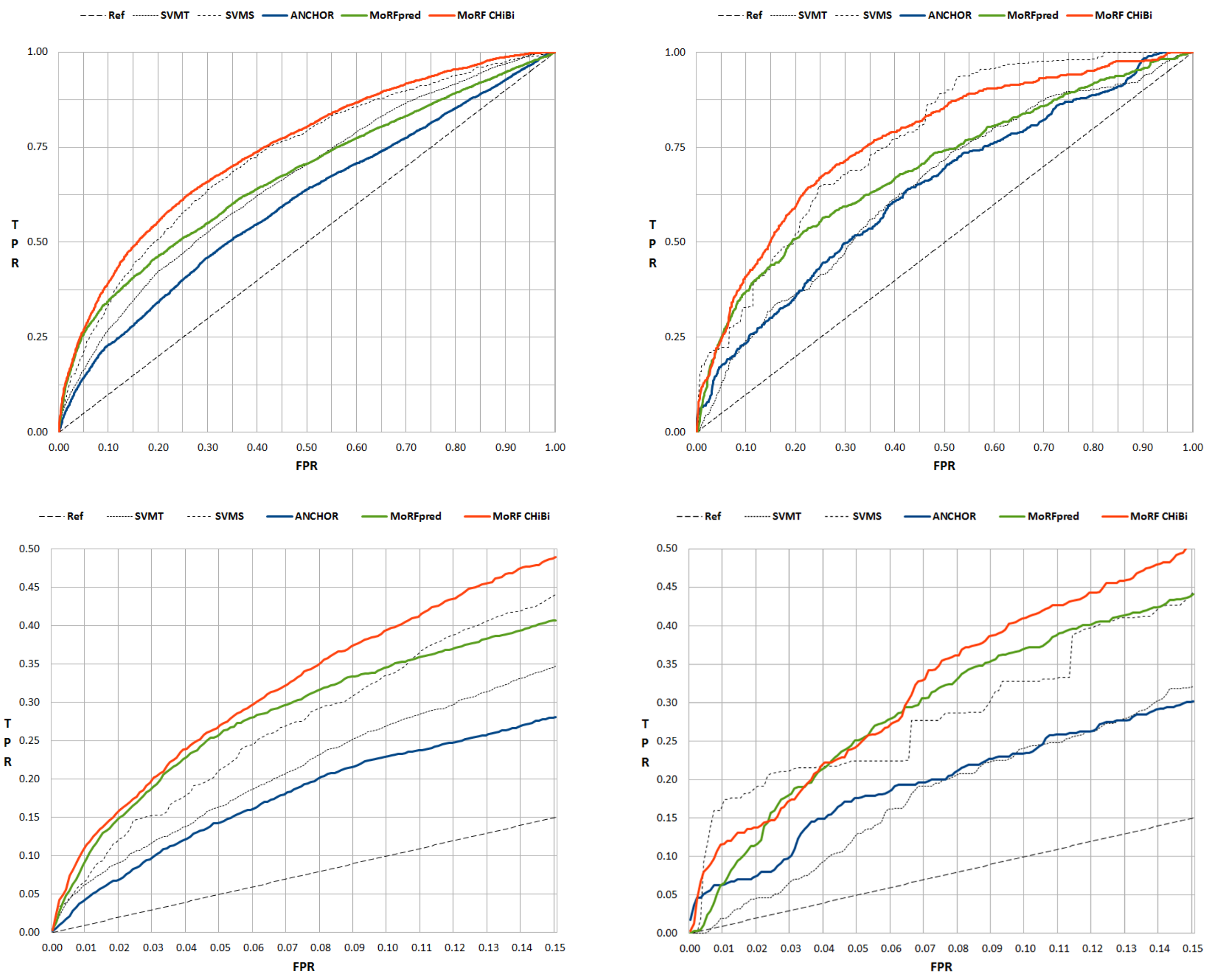
MIYS990101: Relative partition energies derived by the Bethe approximation.
CHAM820102: Free energy of solution in water, kcal/mole.
PONP800106: Surrounding hydrophobicity in turn.
KOEP990102: Beta-sheet propensity derived from designed sequences.
PONP800104: Surrounding hydrophobicity in alpha-helix.
AURR980105: Normalized positional residue frequency at helix termini Nc.
WEBA780101: RF value in high salt chromatography.
ROBB760112: Information measure for coil.
ZASB820101: Dependence of partition coefficient on ionic strength.
PALJ810106: Normalized frequency of turn from CF.
KUHL950101: Hydrophilicity scale.
NAGK730101: Normalized frequency of alpha-helix.
RACS820103: Average relative fractional occurrence in AL(i).
CHAM820102: Free energy of solution in water, kcal/mole.

2.3 MoRFs vs SLiMs

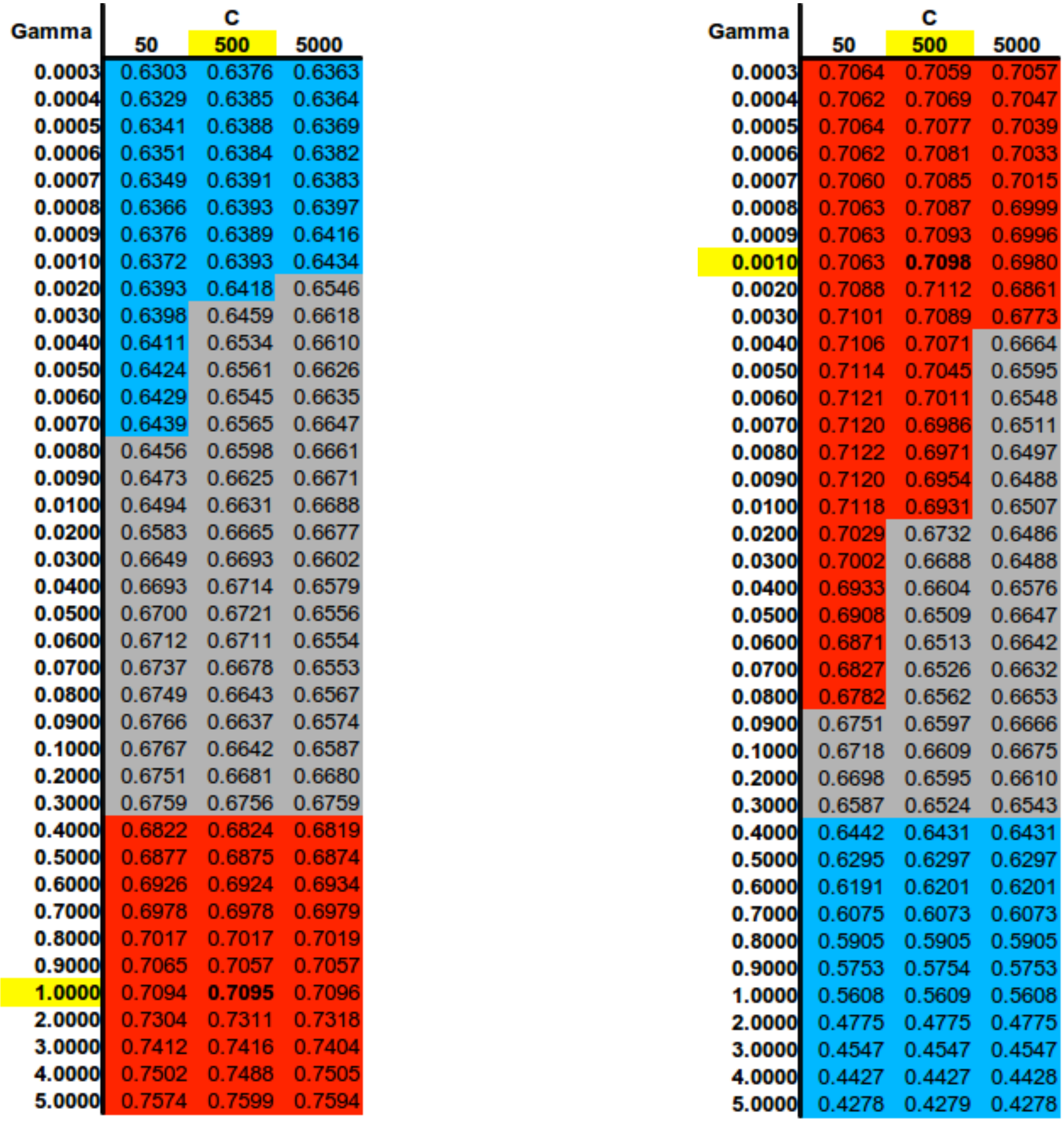
Short linear motifs SLiMs are defined as conserved sequence stretches of 3 to 10 amino acids that are enriched in IDRs (Weatheritt, et al 2012) and promote interactions with specific domains. **Molecular recognition features (MoRFs)** are 10–70 residues loosely structured protein regions within IDRs that bind to structured proteins (Mohan et al. 2006). Hence, MoRFs are on average longer than SLiMs. Figure S4 shows a histogram of the size distribution of the MoRFs in TRAINIG/TEST and ELMs from the ELM database (Dinkel, H. et al. 2013). While MoRFs and SLiMs generally share many features, e.g. they are more hydrophobic than their surroundings, and more conserved (Fuxreiter et al. 2007; Mészáros et al. 2009; Disfani et al. 2012), utilizing these features in identifying candidate binding locations is only feasible with MoRFs. SLiMs' shorter size increases the noise to signal level and renders these features unusable. Thus, SLiMs are modeled with regular expressions and computationally identified using direct alignment tools (Neduva and Russel 2005) such as blast (Altschul, S. et al. 1997) and HMMer (Eddy, S.R. 1998).



Suppl. Fig. S1. *The average values and standard deviation (as error bars) for selected features. (Top) the SVM_S features, and (bottom) the SVM_T features. Feature names are the Amino Acid Index Accession number preceded by two characters "F:" indicating it was generated from flanks, and M:" indicating it was generated from a MoRF. Features are sorted from left to right on their descending average values.*



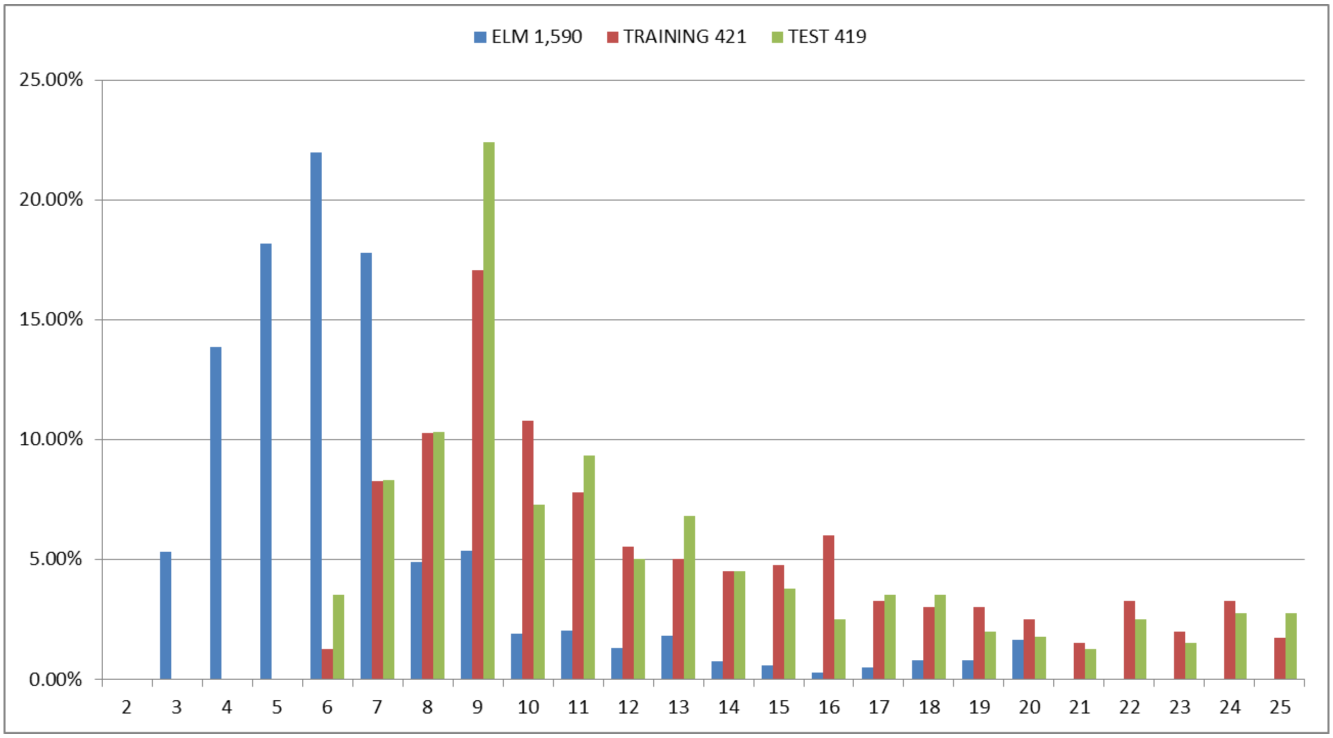
Suppl. Fig. S2 *ROC curves of TEST (left) and NEW (right) datasets for the three predictors; MoRF$_{CHiBi}$ in red, MoRFpred in green, and ANCHOR in blue. The two dotted lines show the performance of each of MoRF$_{CHiBi}$ two component predictors. The full ROC curve is at the top, and the lower left corner of the curve is at the bottom. Vertical axis is the true positive rate TPR and horizontal axis is the false positive rate FPR.*



Gamma	C		
	50	500	5000
0.0003	0.6303	0.6376	0.6363
0.0004	0.6329	0.6385	0.6364
0.0005	0.6341	0.6388	0.6369
0.0006	0.6351	0.6384	0.6382
0.0007	0.6349	0.6391	0.6383
0.0008	0.6366	0.6393	0.6397
0.0009	0.6376	0.6389	0.6416
0.0010	0.6372	0.6393	0.6434
0.0020	0.6393	0.6418	0.6546
0.0030	0.6398	0.6459	0.6618
0.0040	0.6411	0.6534	0.6610
0.0050	0.6424	0.6561	0.6626
0.0060	0.6429	0.6545	0.6635
0.0070	0.6439	0.6565	0.6647
0.0080	0.6456	0.6598	0.6661
0.0090	0.6473	0.6625	0.6671
0.0100	0.6494	0.6631	0.6688
0.0200	0.6583	0.6665	0.6677
0.0300	0.6649	0.6693	0.6602
0.0400	0.6693	0.6714	0.6579
0.0500	0.6700	0.6721	0.6556
0.0600	0.6712	0.6711	0.6554
0.0700	0.6737	0.6678	0.6553
0.0800	0.6749	0.6643	0.6567
0.0900	0.6766	0.6637	0.6574
0.1000	0.6767	0.6642	0.6587
0.2000	0.6751	0.6681	0.6680
0.3000	0.6759	0.6756	0.6759
0.4000	0.6822	0.6824	0.6819
0.5000	0.6877	0.6875	0.6874
0.6000	0.6926	0.6924	0.6934
0.7000	0.6978	0.6978	0.6979
0.8000	0.7017	0.7017	0.7019
0.9000	0.7065	0.7057	0.7057
1.0000	0.7094	**0.7095**	0.7096
2.0000	0.7304	0.7311	0.7318
3.0000	0.7412	0.7416	0.7404
4.0000	0.7502	0.7488	0.7505
5.0000	0.7574	0.7599	0.7594

Gamma	C		
	50	500	5000
0.0003	0.7064	0.7059	0.7057
0.0004	0.7062	0.7069	0.7047
0.0005	0.7064	0.7077	0.7039
0.0006	0.7062	0.7081	0.7033
0.0007	0.7060	0.7085	0.7015
0.0008	0.7063	0.7087	0.6999
0.0009	0.7063	0.7093	0.6996
0.0010	0.7063	**0.7098**	0.6980
0.0020	0.7088	0.7112	0.6861
0.0030	0.7101	0.7089	0.6773
0.0040	0.7106	0.7071	0.6664
0.0050	0.7114	0.7045	0.6595
0.0060	0.7121	0.7011	0.6548
0.0070	0.7120	0.6986	0.6511
0.0080	0.7122	0.6971	0.6497
0.0090	0.7120	0.6954	0.6488
0.0100	0.7118	0.6931	0.6507
0.0200	0.7029	0.6732	0.6486
0.0300	0.7002	0.6688	0.6488
0.0400	0.6933	0.6604	0.6576
0.0500	0.6908	0.6509	0.6647
0.0600	0.6871	0.6513	0.6642
0.0700	0.6827	0.6526	0.6632
0.0800	0.6782	0.6562	0.6653
0.0900	0.6751	0.6597	0.6666
0.1000	0.6718	0.6609	0.6675
0.2000	0.6698	0.6595	0.6610
0.3000	0.6587	0.6524	0.6543
0.4000	0.6442	0.6431	0.6431
0.5000	0.6295	0.6297	0.6297
0.6000	0.6191	0.6201	0.6201
0.7000	0.6075	0.6073	0.6073
0.8000	0.5905	0.5905	0.5905
0.9000	0.5753	0.5754	0.5753
1.0000	0.5608	0.5609	0.5608
2.0000	0.4775	0.4775	0.4775
3.0000	0.4547	0.4547	0.4547
4.0000	0.4427	0.4427	0.4428
5.0000	0.4278	0.4279	0.4278

Suppl. Fig S3. *The Grids for SVM$_T$ (left) and SVM$_S$ (right) generated using the initial set of 39 selected features. Each line starts with its Gamma, and each column starts with its C value. Cells in each grid holds AUC values and are divided into three groups, blue cells have the lowest AUC values, gray are with mid-range values and red cells holds the highest values. In each grid, the selected cell is in bold and its parameters are highlighted in yellow.*



Suppl. Fig S4. *Size distribution of SLiMs and MoRFs. SLiMs (1,590 instances) were taken from the Eukaryotics Linear Motif database (Dinkel, H. et al. 2013) and MoRFs from TRAINING and TEST. The x axis represents the length and the y axis the percentage of occurrence.*

REFERENCES

Altschul, S. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25, 3389–3402.

Babu, M. et al. (2011) Intrinsically disordered proteins: regulation and disease. Current Opinion in Structural Biology 2011,21:1–9.

Dinkel, H. et al. (2013) The eukaryotic linear motif resource ELM: 10 years and counting Nucl. Acids Res. (2013) doi: 10.1093/nar/gkt1047.

Disfani, F. M. et al. (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics 28 (12): i75-i83.

Eddy, S.R. (1998) Profile hidden Markov models. Bioinformatics14 (9), 755–763.

Fuxreiter M, Tompa P, Simon I (2007) Local structural disorder imparts plasticity on linear motifs. Bioinformatics 23: 950–956. doi: 10.1093/bioinformatics/btm035.

Mészáros, B. et al. (2009) Prediction of protein binding regions in disordered proteins. PLoS Comput. Biol., 5, e1000376.

Mohan, A. et al. (2006) Analysis of molecular recognition features (MoRFs). J. Mol. Biol., 362, 1043–1059.

Neduva V, Russell R (2005) Linear motifs: evolutionary interaction switches. FEBS lett 579: 3342. doi: 10.1016/j.febslet.2005.04.005.

Weatheritt , R.J. and Gibson, T.J. (2012) Linear motifs: lost in (pre)translation Trends in Biochemical Sciences, August 2012, Vol. 37, No. 8.