Journal of Theoretical Biology 437 (2018) 9–16

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/jtbi

MoRFPred-plus: Computational Identification of MoRFs in Protein Sequences using Physicochemical Properties and HMM profiles

Ronesh Sharma[a,b,*], Maitsetseg Bayarjargal[c], Tatsuhiko Tsunoda[d,e], Ashwini Patil[f], Alok Sharma[b,d,e,g,*]

[a] Department of Electronics Engineering, Fiji National University, Suva, Fiji
[b] Department of Engineering and Physics, the University of the South Pacific, Suva, Fiji
[c] Department of Health Science, Fiji National University, Fiji
[d] CREST, JST, Yokohama 230-0045, Japan
[e] RIKEN Center for Integrative Medical Science, Yokohama 230-0045, Japan
[f] Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan
[g] Griffith University, Australia

ARTICLE INFO

Article history:
Received 25 January 2017
Revised 12 October 2017
Accepted 13 October 2017
Available online 16 October 2017

Keywords:
Intrinsically disordered proteins
Molecular recognition feature
Hidden Markov model
Support vector machine

ABSTRACT

Motivation. Intrinsically Disordered Proteins (IDPs) lack stable tertiary structure and they actively participate in performing various biological functions. These IDPs expose short binding regions called Molecular Recognition Features (MoRFs) that permit interaction with structured protein regions. Upon interaction they undergo a disorder-to-order transition as a result of which their functionality arises. Predicting these MoRFs in disordered protein sequences is a challenging task.

Method. In this study, we present MoRFpred-plus, an improved predictor over our previous proposed predictor to identify MoRFs in disordered protein sequences. Two separate independent propensity scores are computed via incorporating physicochemical properties and HMM profiles, these scores are combined to predict final MoRF propensity score for a given residue. The first score reflects the characteristics of a query residue to be part of MoRF region based on the composition and similarity of assumed MoRF and flank regions. The second score reflects the characteristics of a query residue to be part of MoRF region based on the properties of flanks associated around the given residue in the query protein sequence. The propensity scores are processed and common averaging is applied to generate the final prediction score of MoRFpred-plus.

Results. Performance of the proposed predictor is compared with available MoRF predictors, MoRFchibi, MoRFpred, and ANCHOR. Using previously collected training and test sets used to evaluate the mentioned predictors, the proposed predictor outperforms these predictors and generates lower false positive rate. In addition, MoRFpred-plus is a downloadable predictor, which makes it useful as it can be used as input to other computational tools.

Availability. https://github.com/roneshsharma/MoRFpred-plus/wiki/MoRFpred-plus:-Download

1. Introduction

In the traditional view of protein structure-function paradigm, the notion is that the function critically depends on the stable three-dimensional structure, however, recent findings revealed that most of the functional regions do not adopt a well-defined tertiary structure (Dyson and Wright, 2005; Lee et al., 2014; Uversky, 2014; Wright and Dyson, 2015). These protein regions are called Intrinsically disordered proteins (IDPs) or Intrinsically disordered regions (IDRs) (Dyson and Wright, 2005; Tompa, 2011). The functional importance of these regions is associated with signal transduction and cell-cycle regulation (Lee et al., 2014; Uversky, 2014). Recently, many different types of functional regions have been investigated and analyzed to understand IDRs (Lee et al., 2014). Of particular interest, first, are the linear motifs that are enriched in IDRs. Second, are the disordered segments that provide disorder-to-order transition upon binding, these segments are called molecular recognition features (MoRFs) (Disfani et al., 2012; Lee et al., 2014; Malhis and Gsponer, 2015). Third, are the interaction domains that are identified using crystallography and sequence analysis methods (Lee et al., 2014).

* Corresponding authors.
E-mail addresses: ronesh.sharma@fnu.ac.fj (R. Sharma), alok.sharma@griffith.edu.au (A. Sharma).

https://doi.org/10.1016/j.jtbi.2017.10.015
0022-5193/

Linear motifs are known as short linear motifs (SLiMs) and are of 3 to 10 amino acids in length (Edwards et al., 2007; Lee et al., 2014; Wright and Dyson, 2015). On the other hand, MoRFs are peptide segment of length 10 to 70 amino acids. Similar to our previous study (Sharma et al., 2016), we focus on MoRFs of size 5 to 25 amino acids located within long IDPs (Disfani et al., 2012). Several computational methods have been recently outlined to predict functional sites in IDPs. Of a particular interest to predict SLiMs and MoRFs, recently predictors MoRFchibi (Malhis and Gsponer, 2015), MoRFpred (Disfani et al., 2012), ANCHOR (Dosztányi et al., 2009; Mészáros et al., 2009), MFSPSSMpred (Fang et al., 2013), γ-MoRF-PredII (Cheng et al., 2007), SliMpred (Mooney et al., 2012), SLiMDis (Davey et al., 2006) and SliMFinder (Edwards et al., 2007) have been developed. It is observed that SLiMs and MoRFs overlap each other, but the method of identifying their locations are very different. Due to the short lengths of SLiMs, identifying them is difficult compared to the identification of MoRFs and prediction results in detection of high false positive rate (FPR). Overall, with the overlapping feature of SLiMs and MoRFs, it is a challenging task to computationally identify the location of these functional sites.

Recently developed MoRF predictors have been mostly benchmarked by comparing their performance with that of MoRFpred and ANCHOR. ANCHOR is available as downloadable software whereas MoRFpred (Disfani et al., 2012) is a web based predictor, the prediction approach of both the predictors are very different and are described in detail in our previous study (Sharma et al., 2016). For prediction of MoRF regions of size 5 to 25 residues, recently MoRFchibi (Malhis and Gsponer, 2015) predictor has been introduced. MoRFchibi uses local physicochemical properties of amino acids for prediction of MoRF regions by employing two support vector machine (SVM) models. The first model uses composition contrast information of training with no similarity information and the second model mainly targets similarity information and the final propensity score is processed by using Bayes rule.

Since MoRFchibi does not rely on any component predictors, this feature of MoRFchibi makes it a very good MoRF predictor in term of processing speed and can be utilized as a component predictor for MoRF prediction. However, with the complexity and importance of MoRF regions, the prediction accuracy is limited. Performance evaluation using the benchmark dataset introduced in Disfani et al. (2012) provided area under the receiver operating characteristics (ROC) curve of 74 percent for MoRFchibi, 68 percent for MoRFpred and 61 percent for ANCHOR.

In this study, we present MoRFpred-plus predictor, an improved predictor over our previous published predictor (Sharma et al., 2016). Here, we utilize hidden Markov model (HMM) profiles with local physicochemical properties of amino acids for identifying MoRFs in IDR sequences, whereas in our previous predictor (Sharma et al., 2016) we only used HMM profiles. Feature vector is extracted to represent query protein sequence and an SVM model is used to generate propensity score for each query residue. Two novel aspects are incorporated in the proposed predictor, first, we use comprehensive set of features encoded in HMM profiles and physicochemical properties. Second, we select and combine suitable SVM models to predict the propensity scores. In terms of performance measure, the proposed predictor is more accurate than ANCHOR, MoRFpred and MoRFchibi. MoRFpred-plus achieved AUC of 75.5 percent, which is 15.5 percent greater than ANCHOR, 8.2 percent greater than MoRFpred and 1.5 percent greater than MoRFchibi. Moreover, the proposed predictor outperforms the best accurate predictor and generates lower false positive rate.

2. Materials and Methods

2.1. Benchmark dataset

In order to benchmark the proposed predictor, we used the training and test sets that were previously used to benchmark MoRFchibi (Malhis and Gsponer, 2015), MoRFpred (Disfani et al., 2012) and ANCHOR (Dosztányi et al., 2009) predictors. The data set was initially created by Disfani et al. (2012). They collected structures with protein-peptide interaction from protein data bank (PDB) and identified peptide regions of 5 to 25 residues which were supposed to be MoRF regions. From 840 protein sequences they obtained, further to analyze MoRF predictors they divided these into 421 training sequences and 419 test sequences. The training set contains 245, 984 residues, in which 240, 588 are non-MoRF residues and the test set contains 258, 829 residues, in which 253, 676 are non-MoRF residues. To validate MoRF predictors, Malhis et al. (2015) filtered and assembled a test set (EXP53). This test set has 53 non-redundant protein sequences which contain MoRF regions that are experimentally verified to be disordered in isolation. Within 53 protein sequences, there are 2,432 MoRF residues and 22,754 non-MoRF residues. From 2,432 MoRF residues, 729 are from sections of short MoRFs (up to 30 residues) and 1,703 are from sections of long MoRFs (longer than 30 residues). The second test set was used to validate and compare MoRFpred-plus. Each of the sequence in training and test set are annotated with single MoRF of length 5 to 25 residues, therefore, bias in the dataset is introduced as there are more non-MoRF residues compared to MoRF residues in the data set, i.e., training set has 5396 MoRF residues compared to 240,588 non-MoRF residues, test set has 5153 MoRF residues compared to 253,676 non-MoRF residues. To reduce the risk of over prediction, Disfani et al. (2012) filtered sequence in the data sets such that no more than 30 percentage sequence similarity exists between any of the sequences.

2.2. Feature extraction techniques

Features from protein sequence can be captured from many different sources of information. These could be structural information of protein sequence (Dehzangi et al., 2013), syntactical and physicochemical properties of amino acids (Dubchak et al., 1997; Sharma et al., 2015), gene ontology information (Wang et al., 2015) and evolutionary information (Dehzangi et al., 2013; Lyons et al., 2016; Sharma et al., 2013). Recent findings focus on the use of evolutionary information for improving prediction accuracies. To use evolutionary information as a source for feature extraction, either position specific scoring matrix can be utilized (generated using PSI-BLAST (Altschul et al., 1997)) or hidden Markov model (HMM) profiles (generated using HHblits Remmert et al., 2011) can be utilized. Both are sequence profiles. For a given query protein sequence, PSI-BLAST or HHblits searches a protein database, performs multiple sequence alignments (MSAs) to find similar protein sequences and extracts a profile that provides a substitution probability of each query residue in the protein sequences. In this study, features are extracted from physicochemical properties encoded in amino acid indexes and from evolutionary profiles of protein sequences.

2.3. Overview of the proposed method

Fig 1 shows the overview of the proposed method, two different methods are used to extract useful features from amino acid indexes and HMM profiles of protein sequences. The two methods are named as MoRF region flank method and MoRF residue

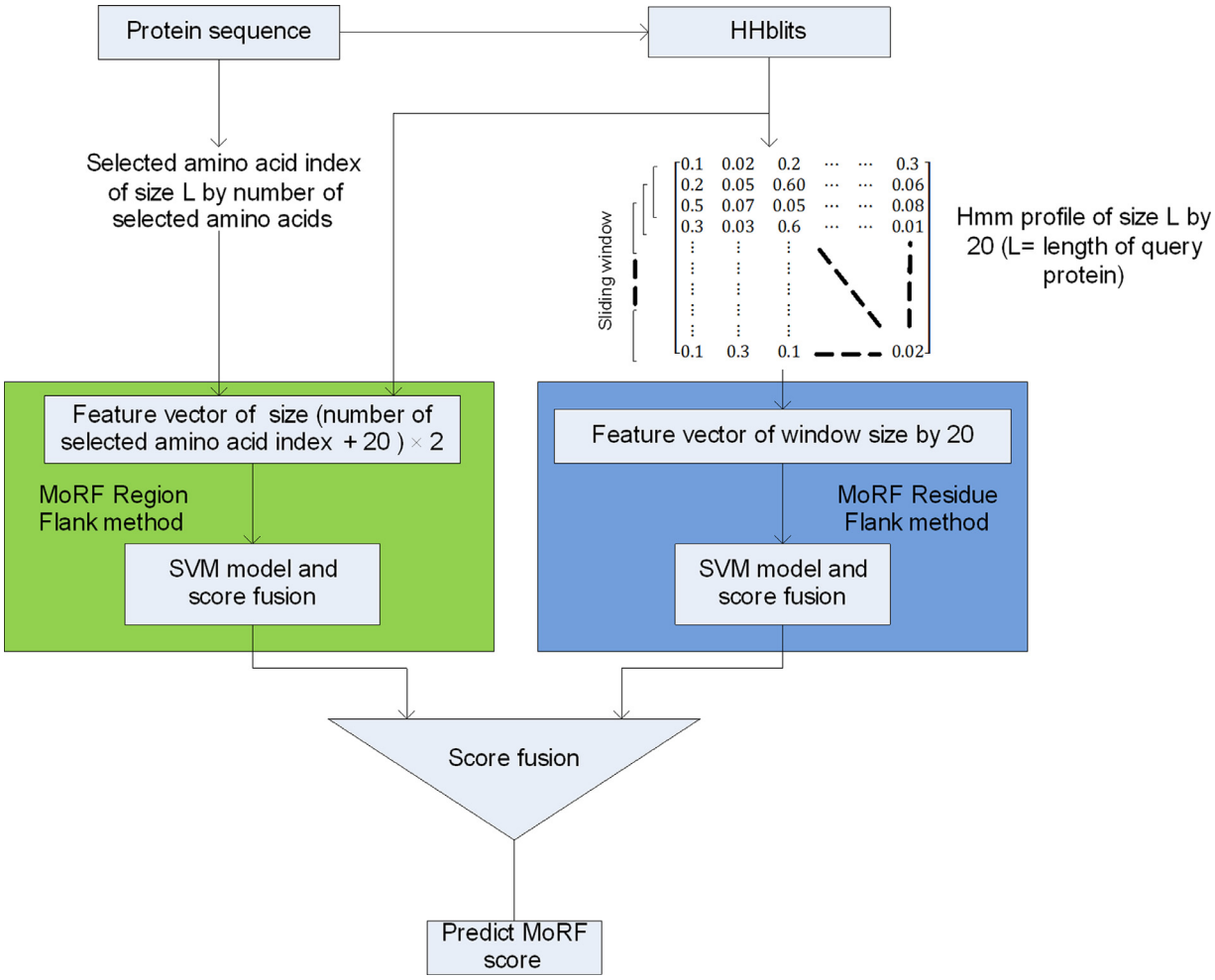

Fig. 1. Overview of the proposed predictor (MoRFpred-plus). The two methods used are Region flank method and Residue flank method. Output score of SVM models are fused using the common averaging strategy. In common averaging strategy, score of all models are added and is divided by the number of models used.

flank method. For rest of the paper, we refer to these two methods as RegionMoRF and ResidueMoRF methods, respectively. Using the first method, feature vectors are extracted to represent composition and sequence similarity information of assumed MoRF and flank regions. A feature vector of size (number of amino acid index + 20)× 2 is fed into a LibSVM classifier (Chang and Lin, 2011) to predict the propensity score of each query residue to form a MoRF region.

In the second method, sliding window is used to extract feature vector of size w × 20 (where w refers to the sliding window size and number 20 refers to selected columns of HMM profile) from HMM profile of input protein sequence. These features represent the flank properties of MoRF residues and are used to predict propensity score of each query residue to form a MoRF region. Common averaging is applied to the scores of each method to provide the final MoRF prediction score (the detailed analyses of the two methods are given later).

2.4. Amino acid indexes

We used two sets of physicochemical properties which are included in standard 544 amino acid indexes (Kawashima et al., 2008), these indexes are available at the web-link: ftp://ftp.genome.jp/pub/db/community/aaindex/. These two sets have shown significant importance in relation to MoRF prediction in Malhis and Gsponer (Malhis and Gsponer, 2015). The first set contains 14 amino acid indexes and the second set contains 13 amino acid indexes. The details of these indexes are given in supporting information S1 Text.

2.5. HMM profiles

To generate HMM profiles, HHblits searches a protein database to find significant similar protein sequence to build multiple sequence alignments (MSAs). Using this MSAs after each iterative search, HHblits computes HMM profiles. For each protein sequence, HMM profile contains substitution probabilities of each common amino acid based on its position within the protein sequence. HMM profiles provide extra information compared to other sequences profiles, it has extra 10 columns to represent information of insertion, match and deletion during MSAs. Using NR20 protein database, we computed the HMM profile of each protein sequence using HHblits with its cut off value (E) set to 0.001 in four iterations. HHblits generates HMM profile of size $L \times 30$ matrix for a given query protein sequence of length L. Using the equation $p = 2^{(-N/1000)}$, the output values in HMM profiles are converted to linear scores. For this study, we only use the first 20 columns of HMM profile.

2.6. Training

We took similar approach as in our previous study (Sharma et al., 2016) and divided the training sequence into two segments. From one of the segment we extract positive

samples for training and from the other segment we extract negative samples for training. Feature vectors are generated for the samples and are used for training the model.

For RegionMoRF method, features are generated from indexes and HMM profiles such that each index (or each profile column) generates two features: the first one is computed as the average value of scores over the supposed amino acids of MoRF region and the second one is computed as the average value of scores over up to 24(12×2) amino acids of flanks surrounding the MoRF region (each flank of 12 amino acids, unless if MoRF region is present at the start or at the end of the protein sequence). For each sequence, same number of positive samples are selected as the number of MoRF regions per sequence.

For ResidueMoRF method, sliding window technique is used to extract features from HMM profiles. For each MoRF residue present in the segment, its residue information is taken together with the information of left and right neighbor regions (maximum of 12 amino acids). The number of positive sample is equal to the number of MoRF residues present in each training sequences. The details of the above two methods are described in supporting information S1 Text.

The dataset used in this study is unbalanced. There are more non-MoRF residues present in the sequence compared with the number of MoRF residues. Thus, it could lead to unbiased prediction. To address these, first we take non-MoRF residues that are not overlapping the flanks of MoRF region and we randomly select same number of negative samples. Second, we increase the ratio between positive and negative samples to 1:2, i.e. for each MoRF sample we select 2 non-MoRF samples. Further, this ratio is increased to 1:3 and the best performing models are selected. Moreover, to avoid over scoring of the training data, non-MoRF samples for each model are randomly selected.

2.7. Testing

To predict scores for a query sequence, features are extracted using a sliding window. It would be easier to select sliding window size, if MoRF sizes are known for the query sequence. However, since MoRF sizes are not known, for RegionMoRF method, 19 different sizes of sliding windows are used to analyze each of the query sequences. These sizes have shown significant enhancement in Malhis and Gsponer (2015) for MoRF prediction. The sizes are from 6 to 24, since the proposed predictor is limited to predict MoRFs of size 5 to 25 residues. With size ranging from 6 to 24, each residue in the query sequence will receive a total of 285 scores except those at the start or end of the sequence. Each of these scores are processed and propensity score for each residue is evaluated as either maximum or minimum of 285 scores. For ResidueMoRF method, the window is centered on the query residue and the flank size is varied on both sides to extract features. These features are then processed using an SVM classifier.

2.8. SVM model and score fusion

SVM classifier with Radial basis function (RBF) and Sigmoid kernels were used to evaluate the features generated using the above two methods.

For RegionMoRF method, using each of the SVM kernels with different C and gamma values, the features generated from amino acid indexes and HMM profiles are evaluated to meet the performance criteria. For ResidueMoRF method, using window size of 7 (w is used as 7 due to the limitation on processing speed), SVM classifier is parameterized to obtain best AUC, success rate and FPR. Furthermore, these parameters are then used to evaluate and rank the features generated by varying the window size. Finally, the proposed method uses common averaging (add all model scores and divide it by the number of models used) at different stages to fuse the propensity scores of multiple best performing model.

2.9. Performance measure

We used the evaluation metrics that were previously used to analyze MoRF predictors (Disfani et al., 2012; Malhis and Gsponer, 2015; Malhis et al., 2015). These are AUC, success rate and accuracy. AUC is the area under the receiver operating characteristics curve, success rate compares the average predicted propensity scores between actual MoRF residues and non-MoRF residues. Accuracy shows the total number of residues that are correctly predicted. These metrics are defined in Disfani et al. (2012).

3. Results

The performance of the proposed predictor is evaluated using a test set and is compared with other MoRF predictors, ANCHOR (Dosztányi et al., 2009), MoRFpred (Disfani et al., 2012) and MoRFchibi (Malhis and Gsponer, 2015).

3.1. SVM model and feature selection

To achieve high AUC, success rate and FPR for RegionMoRF method, we selected two high noise tolerance kernels with high gamma value of 5 and low gamma value of 0.0038 to evaluate and rank appropriate features generated from amino acid indexes and HMM profiles. Feature vectors generated are individually evaluated first and then are concatenated in order to meet the performance criteria. For ResidueMoRF method, the window size is varied to select features generated from HMM profiles and appropriate SVM models with different kernels and gamma values are selected.

For both methods, the sampling ratio between MoRF and non-MoRF sample is increased and similar evaluation is carried out to select 3 best performing models. Common averaging is applied to the output scores of each model to provide the final MoRF prediction score for each method (for more details on model selection for both methods please see supporting information S1 Text). Observing performance during model selection, it is noted that the models alone tend to over score MoRFs (they have high FPR, low success rates and AUCs). Thus, this indicates that the models individually are not able to identify MoRFs accurately. Therefore, we apply common averaging at different stages to combine the scores generated by different models.

3.2. Comparison with available predictors

The performance metrics of the proposed predictor and the available predictors are compared and are outlined in Table 1. It is noted that MoRFpred-plus achieves improved performance in terms of AUC, success rate, FPR and accuracy. Compared to 8 component predictors of MoRFpred, MoRFpred-plus only uses one component predictor and achieves 8.2 percent increase in AUC. As shown in Fig 2, we generate AUC curves for the available predictors and the proposed predictor using test set. It is observed that MoRFpred-plus achieves lower FPR at any given TPR when compared with ANCHOR, MoRFpred and MoRFchibi. This is also demonstrated in Table 2. The superior performance of MoRFpred-plus lies in the combination of HMM profiles with amino acid indexes and ranking of appropriate SVM kernels. Moreover, we validate and compare the proposed predictor with the available MoRF predictors using the second test set (EXP53). Since the proposed and available MoRF predictors are trained to predict MoRFs of sizes up to 30 residues, however, EXP53 set contains MoRFs longer than 30 residues. Therefore, we provide performance metrics for short

Table 1
Overall Comparison of results with other predictors using test set.

Method/predictors	TPR	AUC	Success rate	FPR	Accuracy
ANCHOR	0.222	0.600	0.611	0.092	0.894
MoRFPred	0.222	0.673	0.718	0.038	0.948
MoRFchibi	0.222	0.740	0.730	0.035	0.951
MoRFpred-plus (proposed)	**0.222**	**0.755**	**0.745**	**0.027**	**0.958**

AUC, Success rate and Accuracy for proposed and available predictors. Bold numbers indicate the best performance metrics.

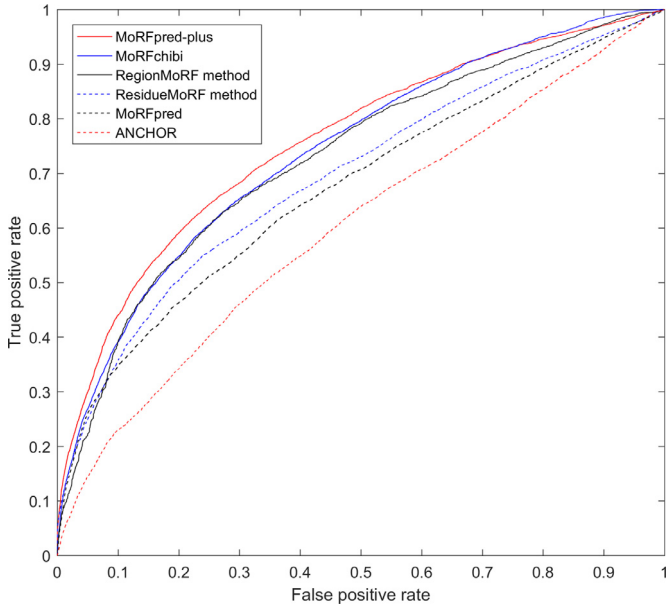

Fig. 2. AUC curves for the available predictors and the proposed predictor generated using test set. AUC curves for predictors: MoRFpred-plus; MoRFchibi; RegionMoRF method; ResidueMoRF method; MoRFpred and ANCHOR.

Table 2
FPR as a function of TPR using test set.

TPR	MoRFpred-plus	MoRFchibi	MoRFPred	ANCHOR
0.1	**0.005**	0.009	0.011	0.031
0.2	**0.022**	0.030	0.033	0.075
0.3	**0.046**	0.062	0.072	0.165
0.4	**0.076**	0.105	0.145	0.248
0.5	**0.127**	0.162	0.241	0.341

FPR for TPR values of 0.1, 0.2, 0.3, 0.4 and 0.5. Bold numbers indicate the best performance metrics.

Table 3
AUC values for test and EXP53 (short and long) sets.

Test sets	MoRFpred-plus	MoRFchibi	MoRFPred	ANCHOR
Test	**0.755**	0.740	0.673	0.600
EXP53	**0.821**, 0.670	0.790, 0.679	0.673, 0.598	0.683, 0.586

AUC values of MoRFpred-plus predictor compared to those of MoRFchibi, MoRFpred and ANCHOR using test and EXP53 sets. For EXP53 set, MoRF prediction was evaluated for short MoRFs (up to 30 residues) and long MoRFs (more than 30 residues), AUC values are in the form short, long. Bold numbers indicate the best performance metrics. Underline values are taken from Malhis et al. (2015).

and long MoRFs separately. Table 3 shows the achieved AUC values. It is noted that there is consistent improvement in the performance when compared with performance of available predictors. This symbolizes that the performance improvement is not due to over fitting.

4. Discussion

We presented MoRFpred-plus predictor which utilizes two methods named as RegionMoRF method and ResidueMoRF method to extract important features from amino acid indexes and HMM profiles to predict MoRFs in protein sequences. Compared with available predictors, the proposed predictor clearly demonstrates significant improvement in terms of AUC, success rate, accuracy and FPR. To compare the proposed predictor in terms of its processing speed, we tested MoRFchibi and ANCHOR using the entire test set on i5, 3.5GHz computer, since both do not require multiple sequences alignments. Using a single sequence from test set (Uniprot:Q38087) with 903 residues, we tested MoRFpred-plus using i5, 3.5GHz computer and since MoRFpred is not downloadable, we submitted single sequence (Uniprot:Q38087) to the MoRFpred prediction server.

Prediction time for ANCHOR and MoRFchibi, both do not require generation of evolutionary profiles, therefore were fastest with speed 3.9×10^6 residues/minute (r/m) and 10.5×10^3 r/m, respectively. The proposed predictor came third with 526 r/m and MoRFpred came slowest at 48r/m. Table 4 shows the overall comparison. The overall comparison may not be entirely fair, since MoRFpred server hardware processor is unknown and hence ANCHOR and MoRFchibi do not rely on evolutionary information such as PSI-BLAST or HHblits.

To provide analyses on the average length of MoRFs predicted by MoRFpred-plus, we threshold the predicted scores at values of 0.50, 0.55, 0.60 and 0.65, respectively. At these thresholds, we show TPR and FPR by MoRFpred-plus in Table 5 and plot of length of MoRFs versus percentage of correctly predicted residues in Fig 3 and Fig 4. At a threshold value of 0.50 in Fig 3, it is observed that MoRF of length 14 is predicted very well from test set followed by a good performance obtained for MoRFs of length 10, 12, 16, 21 and 22. In Fig 4, it is noticed that MoRFs length up to 30 residues are predicted very well compared to MoRFs length greater than 30 residues, this confirms that MoRFpred-plus is trained to predict short MoRFs and here we are able to evaluate how it reacts to long MoRFs.

For evolutionary profiles, MoRFpred-plus relies on HHblits, which is faster than PSI-BLAST and generates more accurate alignments. Extracting features from HMM profiles and concatenating it with features extracted from amino acid indexes, MoRFpred-plus offered higher predicting speed compared with MoRFpred. Though ANCHOR and MoRFchibi are the fastest in terms of predicting speed, the results show that MoRFpred-plus is more accurate. The prediction time for MoRFpred-plus mainly depends on the generation of HMM profiles.

Overall, MoRFpred-plus is a new MoRF predictor and its success relies on the use of HMM profiles computed from MSAs and the use of amino acid properties encoded in 544 common amino acid indexes. The improved performance is firstly, the result of adopting a suitable architecture that combines multiple models score at two different stages and secondly, uses different source of features for each model with different classification parameters. The use of ResidueMoRF method for feature extraction provided a com-

Table 4
Overall Comparison with available predictors.

		Model predicting	speed (r/m)		
Predictors	AUC values using test set	i5 3.5GHz computer	Predicting Server	Multiple sequence alignments	downloadable
ANCHOR	0.600	3.9×10^6	-	×	✓
MoRFchibi	0.740	10.5×10^3	-	×	✓
MoRFpred-plus	0.755	526	-	✓	✓
MoRFpred	0.673	-	48	✓	×

Predicting speed: residues/minute (r/m). The server hardware processor for MoRFpred is unknown.

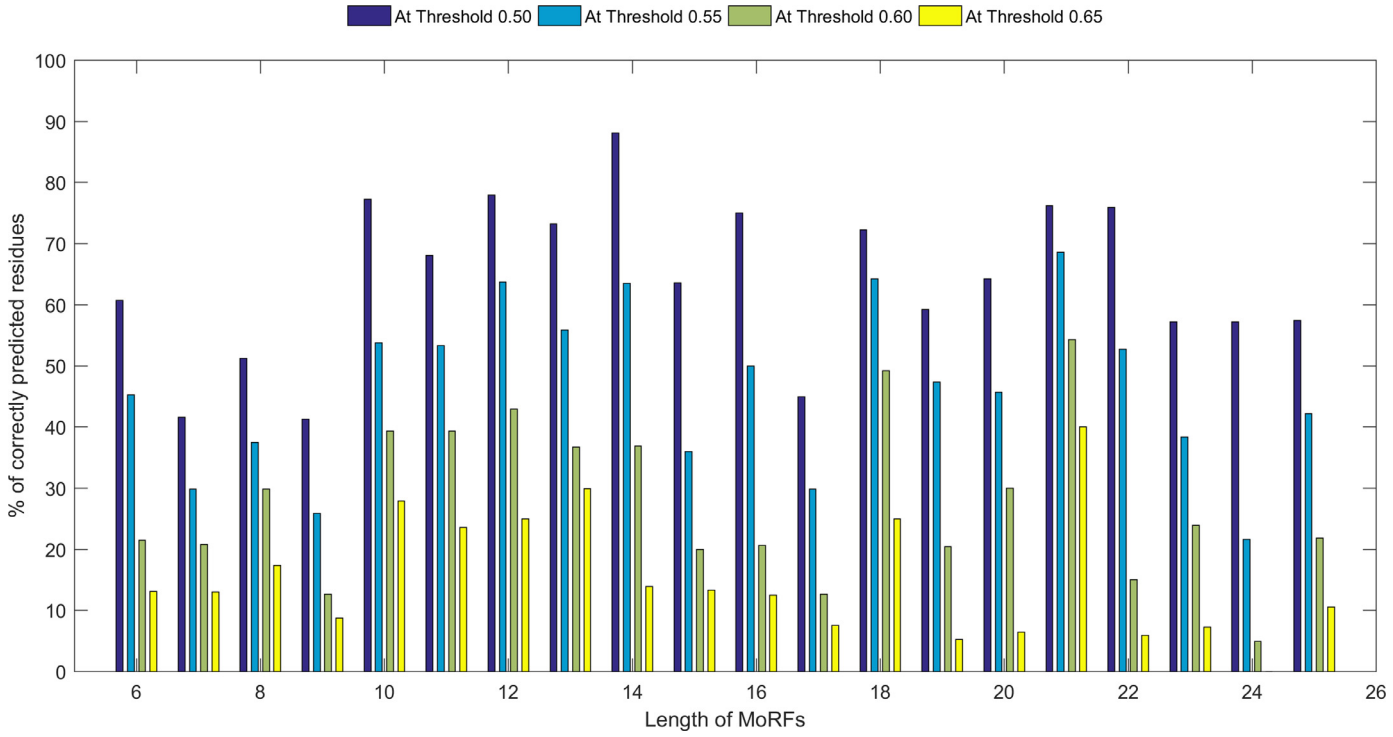

Fig. 3. Length of MoRFs versus percentage of correctly predicted residues using test set, where MoRF length is from 6 to 25 residues. Percentage of correctly predicted residues is shown for threshold values of 0.50, 0.55, 0.60 and 0.65.

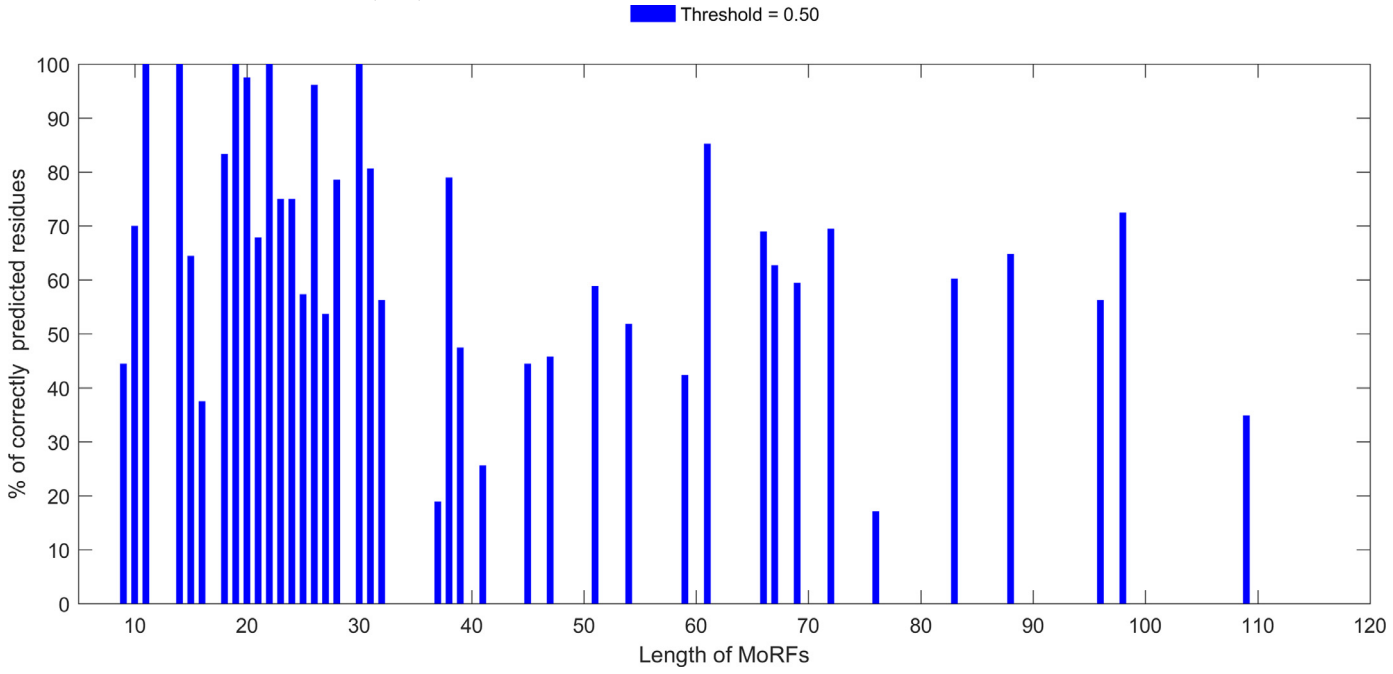

Fig. 4. Length of MoRFs versus percentage of correctly predicted residues using EXP53 set. This set contains both short MoRFs (up to 30 residues) and long MoRFs (greater than 30 residues). Percentage of correctly predicted residues is shown for threshold value of 0.50.

Table 5
TPR and FPR with different threshold for test and EXP53 sets.

Score Threshold	TPR (test, EXP53)	FPR (test, EXP53)
0.50	0.617, 0.595	0.223, 0.278
0.55	0.440, 0.366	0.099, 0.129
0.60	0.265, 0.203	0.039, 0.047
0.65	0.150, 0.068	0.012, 0.014

TPR, FPR is given in the form test, EXP53.

prehensive set of information to distinguish a MoRF residue along its flank region. Furthermore, using RegionMoRF method and concatenating features generated from HMM profiles and amino acid indexes provided composition and similarity information between MoRF region and its surrounding regions resulting in performance improvement.

To predict MoRF scores in query protein sequence, predictors are supposed to be consistent on the entire sequence length. However, if the regions in the query sequence have similar properties to that of training sequence region, this could result in biased prediction and misclassify MoRF residues. To avoid biased prediction, MoRFpred-plus uses several approaches, these are: two different methods of feature extraction; two sources for feature extraction; SVM models with different parameters; suitable sampling ratios between MoRF and non-MoRF samples; selecting non-MoRF segments that are not flanks of MoRF region and randomly selecting non-MoRF samples for each model. From the result, it is noted that different models may illustrate different biasing with same training data, i.e. SVM model with sigmoid kernel avoids over scoring whereas RBF kernel turns to over score their training data. This is demonstrated in the results of ResidueMoRF method (please refer to supporting information S1 Text). Moreover, applying common averaging at two stages, MoRFpred-plus avoids producing biased scores.

The proposed predictor only utilizes evolutionary and physicochemical information; therefore, we compare this predictor with predictors of similar approaches. Recently, MoRFchibi-web (Malhis et al., 2015) and MoRFchibi-light (Malhis et al., 2016) predictors has been proposed. It uses previously published predictors MoRFchibi, other disordered predictors and conservation information to combine prediction scores at several stages to predict MoRFs. Nonetheless, incorporating number of predictors and combining their scores will significantly improve the overall performance of predicting MoRFs in protein sequence, i.e. MoRFchibi-web achieved AUC of 0.800, and MoRFchibi-light achieved AUC of 0.777 evaluated on test set. Using single sequence (Uniprot:Q38087), we tested both predictors on i5 3.5GHz computer and MoRFchibi-web was also tested on its prediction server. The predicting speed of MoRFchibi-light remains almost same as the speed of MoRFchibi, however, the speed for MoRFchibi-web is significantly reduced to 80 r/m compared with 526 r/m for the proposed predictor. MoRFchibi-web server predicting speed came to 588 r/m, however, the server hardware processor is unknown. The proposed predictor builds predicting models using primary protein information, therefore, does not rely on any other disordered predictors.

The proposed predictor is downloadable and the output scores provided are numerical, since it is assumed that different protein sequences in different applications might require different levels of threshold values. Overall, MoRFpred-plus is available without any limitation and can be easily integrated as input to other applications.

Authors contributions

RS, AP and AS conceived the project. RS performed the analysis and wrote the manuscript under the guidance of MB, AP and AS. TT provided computational resources.

Acknowledgments

We would like to acknowledge the authors of MoRFpred predictor for publicity providing the training and test sequence data for MoRF prediction. In addition, we thank Dr. Nawar Malhis and Dr. Lukasz Kurgan for providing guidelines on MoRF prediction.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.jtbi.2017.10.015.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Research 25 (17), 3389–3402.
Chang, C.C., Lin, C.J., 2011. Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2 (3), 27:1–27:27.
Cheng, Y., Oldfield, C.J., Meng, J., Romero, P., Uversky, V.N., Dunker, A.K., 2007. Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. Biochemistry 46, 13468–13477.
Davey, N.E., Shields, D.C., Edwards, R.J., 2006. Slimdisc: short, linear motif discovery, correcting for common evolutionary descent. Nucleic Acids Research 34 (12), 3546–3554.
Dehzangi, A., Paliwal, K.K., Lyons, J., Sharma, A., Scattar, A., 2013. A segmentation-based method to extract structural and evolutionary features for protein fold recognition. IEEE/ACM Transactions on Computational Biology and Bioinformatics 11 (3), 510–519.
Disfani, F.M., Hsu, W.L., Mizianty, M.J., Oldfield, C.J., Xue, B., Dunker, A.K., Uversky, V.N., Kurgan, L., 2012. Morfpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics 28, i75–i83.
Dosztányi, Z., Mészáros, B., Simon, I., 2009. Anchor: web server for predicting protein binding regions in disordered proteins. Bioinformatics 25 (20), 2745–2746.
Dubchak, I., Muchnik, I., Kim, S.H., 1997. Protein folding class predictor for scop: approach based on global descriptors. In: ISMB-97 Proceedings Int Conf Intell Syst Mil Biol, 5, pp. 104–107.
Dyson, H.J., Wright, E.P., 2005. Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol 6, 197–208.
Edwards, R.J., Davey, N.E., Shields, D.C., 2007. Slimfinder: A probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. PLos one 2 (10), e967.
Fang, C., Noguchi, T., Tominaga, D., Yamana, H., 2013. Mfspssmpred: identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation. BMC Bioinformatics 14 (300).
Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., Kanehisa, M., 2008. Aaindex: amino acid index database, progress report 2008. Nucleic Acids Res 36 (D202-D205).
Lee, R.V.D., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D.T., Kim, P.M., Kriwacki, R.W., Oldfield, C.J., Pappu, R.V., Tompa, P., Uversky, V.N., Wright, P.E., Babu, M.M., 2014. Classification of intrinsically disordered regions and proteins. Chemical Reviews 114, 6589–6631.
Lyons, J., Paliwal, K.K., Dehzangi, A., Hefferman, R., TatsuhikoTsunoda, Sharma, A., 2016. Protein fold recognition using hmm–hmm alignment and dynamic programming. Journal of Theoretical Biology 393, 67–74.
Malhis, N., Gsponer, J., 2015. Computational identification of morfs in protein sequences. Bioinformatics 31 (11), 1738–1744.
Malhis, N., Jacobson, M., Gsponer, J., 2016. Morfchibi system: software tools for the identification of morfs in protein sequences. Nucleic Acids Research 44, W488–W493.
Malhis, N., Wong, E.T.C., Nassar, R., Gsponer, J., 2015. Computational identification of morfs in protein sequences using hierarchical application of bayes rule. PLoS ONE 10 (e0141603), e0141603.
Mooney, C., Pollastri, G., Shields, D.C., Haslam, N.J., 2012. Prediction of short linear protein binding regions. Mol Biol 415 (1), 193–204.
Mészáros, B., Simon, I., Dosztányi, Z., 2009. Prediction of protein binding regions in disordered proteins. PLoS Comput Biol 5 (5:e1000376), e1000376.
Remmert, M., Biegert, A., Hauser, A., Söding, J., 2011. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. Nature Methods 9 (2), 173–175.

Sharma, A., Lyons, J., Dehzangi, A., Paliwai, K.K., 2013. A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition. Theoretical Biology 320 (7), 41–46.
Sharma, R., Dehzangi, A., Lyons, J., Paliwal, K., Tsunoda, T., Sharma, A., 2015. Predict gram-positive and gram-negative subcellular localization via incorporating evolutionary information and physicochemical features into chou's general pseaac. IEEE transactions on nanobioscience 14 (8), 915–926.
Sharma, R., Kumar, S., Tsunoda, T., Patil, A., Sharma, A., 2016. Predicting morfs in protein sequences using hmm profiles. BMC Bioinformatics 17 Suppl 19, 504, 252–258.
Tompa, T., 2011. Unstructural biology coming of age. Curr. Opin. Struct. Biol 2011, 419–425.
Uversky, V., 2014. Introduction to intrinsically disordered proteins (idps). Chemical Reviews 114, 6557–6560.
Wang, X., Zhang, J., Li, G.Z., 2015. Gpos-ecc-mploc and gneg-ecc-mploc: Multi-location gram-positive and gram-negative bacterial protein subcellular localization using gene ontology and multi-label classifier ensemble. BMC Bioinformatics 16, Suppl 12.
Wright, P.E., Dyson, H.J., 2015. Intrinsically disordered proteins in cellular signalling and regulation. Nature Reviews:molecular cell biology 16, 18–29.