BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Computational prediction of MoRFs based on protein sequences and minimax probability machine

Hao He, Jiaxiang Zhao* and Guiling Sun

Abstract

Background: Molecular recognition features (MoRFs) are one important type of disordered segments that can promote specific protein-protein interactions. They are located within longer intrinsically disordered regions (IDRs), and undergo disorder-to-order transitions upon binding to their interaction partners. The functional importance of MoRFs and the limitation of experimental identification make it necessary to predict MoRFs accurately with computational methods.

Results: In this study, a new sequence-based method, named as MoRF$_{MPM}$, is proposed for predicting MoRFs. MoRF$_{MPM}$ uses minimax probability machine (MPM) to predict MoRFs based on 16 features and 3 different windows, which neither relying on other predictors nor calculating the properties of the surrounding regions of MoRFs separately. Comparing with ANCHOR, MoRFpred and MoRF$_{CHiBi}$ on the same test sets, MoRF$_{MPM}$ not only obtains higher AUC, but also obtains higher TPR at low FPR.

Conclusions: The features used in MoRF$_{MPM}$ can effectively predict MoRFs, especially after preprocessing. Besides, MoRF$_{MPM}$ uses a linear classification algorithm and does not rely on results of other predictors which makes it accessible and repeatable.

Keywords: Molecular recognition features, Intrinsically disordered proteins, Minimax probability machine

Background

Intrinsically disordered proteins (IDPs) are protein sequences that contain at least one region lacking a unique 3-D structure [1]. Although not being folded, IDPs perform a variety of important functions such as molecular recognition, transport catalysis, signaling regulation, entropic chain activities, and so on [2]. Furthermore, a single protein may contain several disordered regions that possess different functions [3]. The functions of disordered regions usually stem from their ability to bind to partner molecules [4]. Disordered regions can provide malleable interfaces which can recognize molecules through increase complementarity via induced fit or offer alternative interaction upon variable conditions and more complex cellular responses [5]. These recognition regions may form folded and complementary interfaces, while the neighboring regions, often denoted as fuzzy, can maintain their disordered state [6]. The notion of fuzziness implies that conformational heterogeneity can be maintained upon interactions of IDPs [7]. The disordered regions mainly contain two types of binding motifs: short linear motifs (SLiMs) and MoRFs. SLiMs are enriched in IDRs. They are generally conserved and 3-10 residues long, and thus may not fall into regular secondary structures [7]. MoRFs generally locate within longer IDRs and are up to 70 residues long [8]. They promote specific protein-protein interactions, and undergo disorder-to-order transitions upon binding their partners [4]. According to the structures they adopt in bound state, MoRFs can be classified into four subtypes: α-MoRFs, β-MoRFs, ι-MoRFs and complex-MoRFs [9]. The first three types form α-helix, β-strand, irregular secondary structure and the last one contains multiple secondary structures when bound [9].

* Correspondence: zhaojx@nankai.edu.cn
College of Electronic Information and Optical Engineering, Nankai University, Tianjin, China

Because of the functional importance of MoRFs and the limitation of experimental identification, several computational methods have been produced in recent years, such as α-MoRF-Pred I [10], α-MoRF-PredII [11], ANCHOR [12, 13], MoRFpred [14], MSPSSMpred [15] and $\mathrm{MoRF_{CHiBi}}$ [16]. α-MoRF-PredII is an improved method for α-MoRF-Pred I, which is limited to predict α-MoRFs. ANCHOR and MoRFpred are the most used comparison methods in recent years. ANCHOR is a web based method, which predicts protein binding regions that are disordered in isolation but can undergo disorder-to-order transition upon binding by using the energy estimation approach of IUPred [17]. MoRFpred is also a web based method, which is a comprehensive method. It calculates a MoRF propensity score using a linear kernel support vector machine (SVM) based on nine sets of features: physicochemical properties in Amino Acid Index [18], Position Specific Scoring Matrices (PSSM), predicted relative solvent accessibility [19], predicted B-factors [20] and the results of five different intrinsic disorder predictors. Then, using PSI-BLAST [21] to align the input sequence with the training sequence to gain an alignment e-value, which is used to adjust the calculated MoRF propensity score. MSPSSMpred using a radial basis function (RBF) kernel SVM model to predict MoRFs based on calculated conservation scores. This method does not use predicted results from other predictors as input, and the performance in AUC is approximate to MoRFpred. $\mathrm{MoRF_{CHiBi}}$ uses two SVM models to predict MoRFs based on physicochemical properties of amino acids. The first model use a sigmoid kernel SVM to predict MoRF propensities, which target direct similarities between MoRF sequences. The second model focus on the general contrast of amino acid composition of MoRFs, Flanks and the general protein population using a RBF Gaussian kernel SVM. Finally, join the results of the two SVM models and compute the propensity score using Bayes rule. $\mathrm{MoRF_{CHiBi}}$ is a very good MoRF predictor that does not rely on other predictors.

In this paper, we propose a novel sequence-based method, $\mathrm{MoRF_{MPM}}$, for predicting MoRFs. First, simulated annealing algorithm is utilized for selecting candidate feature sets from Amino Acid Index (AA Index) [18]. Then, five structural features from our previous study [22] about IDPs prediction are put into candidate sets for further selection, which contain Shannon entropy and topological entropy calculated directly from protein sequences, as well as three amino acid propensities from GlobPlot NAR paper [23]. Finally, we select 16 features and 3 different windows to preprocess the protein sequences and use MPM [24] which is a linear classification algorithm to predict MoRFs. The simulation results show that even though $\mathrm{MoRF_{MPM}}$ just uses 16 features, 3 different windows and a linear classification, it obtains higher AUC and TPR than ANCHOR, MoRFpred and $\mathrm{MoRF_{CHiBi}}$.

Results

Datasets

In order to compare our method with ANCHOR, MoRFpred and $\mathrm{MoRF_{CHiBi}}$, we use the datasets collected by Disfani et al. [14], which are also used to train and test MoRFpred and $\mathrm{MoRF_{CHiBi}}$. Disfani et al. collected a lot of protein complexes concerning interactions of protein-peptide from Protein Data Bank (PDB) [25] of March 2008 and filtered them on several principles to identify peptide regions of 5 to 25 residues which were presumed to be MoRFs. The obtained 840 protein sequences are divided into a training set (TRAINING) and a test set (TEST). There are 181 helical, 34 strand, 595 coil and 30 complex MoRF regions on the two sets. TRAINING contains 421 sequences which consists of 245,984 residues with 5396 MoRF residues. TEST contains 419 sequences which consists of 258,829 residues with 5153 MoRF residues. Besides, using the same protocol [26, 27], they also collected TESTNEW set from PDB entries deposited between January 1 and March 11, 2012. TEST2012 contains 45 sequences which consists of 37,533 residues with 626 MoRF residues. In addition, we use the EXP53 collected by Malhis et al. [28] as the third test set. The test set contains 53 non-redundant sequences possessing MoRFs, which are collected from four publicly available experimentally validated sets. EXP53 includes 2432 MoRF residues which consist of 729 residues from short MoRF regions (up to 30 residues) and 1703 residues from long MoRF regions (longer than 30 residues). For more intuitive description of the four datasets, Table 1 lists their specific information.

Performance evaluation

We use AUC to evaluate the performance of different candidate feature sets and different windows. It is also utilized to compare our method with other methods. AUC is the area under the ROC curve, which can provide an overall assessment about the prediction. In order to compare the performance of each method in detail, we also calculate ACC and FPR at different TPR. ACC

Table 1 Datasets used in this paper

	TRAINING	TEST	TESTNEW	EXP53
Number of Sequences	421	419	45	53
Number of MoRFs Residues	5396	5153	626	2432
Number of non-MoRFs Residues	240,588	253,676	36,907	22,754
Total Residues	245,984	258,829	37,533	25,186

The detail information of four datasets

describes the total number of residues that are correctly predicted, FPR is the false positive rate and TPR is the true positive rate. They are defined as:

$$\mathrm{ACC} = \frac{TP + TN}{N_{\mathrm{MoRF}} + N_{\mathrm{non}}}, \quad \mathrm{FPR} = \frac{TN}{N_{\mathrm{non}}}, \quad \mathrm{TPR} = \frac{TP}{N_{\mathrm{MoRF}}}, \tag{1}$$

Where TP and TN are the numbers of accurately predicted MoRFs residues and non-MoRFs residues, N_{MoRF} and N_{non} are the total numbers of MoRFs residues and non-MoRFs residues, respectively.

Selecting the optimal feature set

Firstly, we use simulated annealing algorithm to select several candidate sets of different feature number based on the TRAINING from 544 amino acid index. Then, we use MPM [24, 29] to predict MoRFs based on these candidate feature sets, and select the feature set with the best performance. Figure 1 shows the predictive results on TRAINING and TEST with different candidate feature sets. The blue line represents the AUC values on TRAINING, the red line represents the AUC values on TEST. The distances between AUC values on the two sets reflect the over-fitting situation of each candidate set, and the shorter the distance, the more robust the predictive performance. Because MPM is a linear classification algorithm, the over-fitting is not serious in all of these candidate sets. However, it is obvious that when the feature number in the candidate set is 12 or 13, the predictor gains more robust performance and better AUC value on TEST at the same time.

When the feature number of candidate set is 12 or 13, the predictive performance is approximate. Thus, to further compare their performance, the ROC curves are shown on Fig. 2. The left one shows the full ROC curves of them, which almost overlap. Since we are more concerned about the predictive performance at low FPR, the right figure shows the ROC curves at FPR < 0.1. Obviously, in this area, predictive performance on 13 is much better. Thus, we select the candidate set with 13 features as the final candidate feature set from AA Index, which is listed with the AA Index accession numbers in Table 2.

After that, we put the five structural properties which selected by our previous study [22] about IDPs prediction into the candidate feature set. Then, we change the number of structural properties in the candidate feature set and use MPM to predict MoRFs. Since there are only five structural features in total, we use the enumeration method to select structural properties for each candidate feature set with different number of structural properties. Figure 3 shows the best AUC values with different numbers of structural properties. Obviously, when the number is between 2 and 4, the performance is similar and obviously better than other cases. To further compare their performance, the ROC curves are shown on Fig. 4. Though the full ROC curves of them almost overlap as shown in the left figure, 3 and 4 obtain better performance at FPR < 0.1 as shown in the right figure. Considering that the AUC value of 3 is slightly higher than that

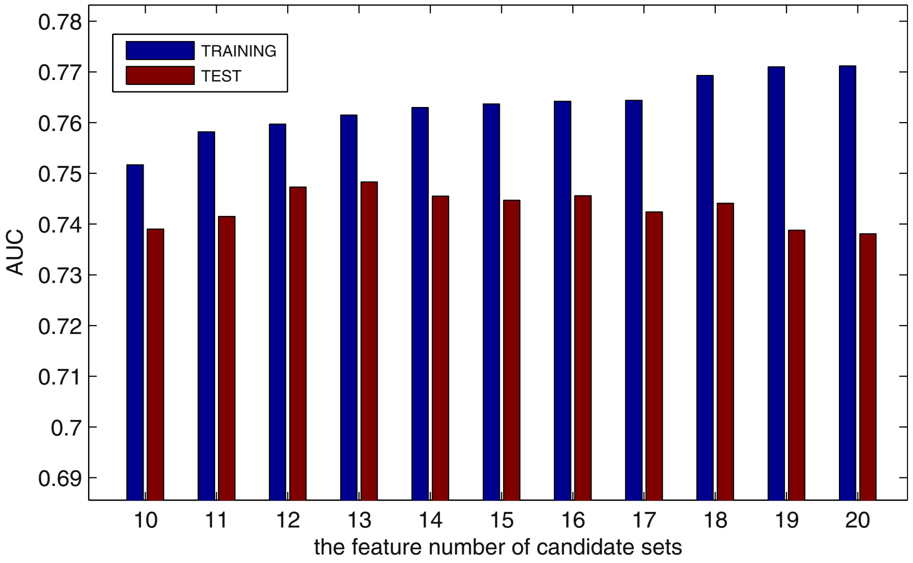

Fig. 1 Predictive performance with different number of properties from AA Index. The blue line is the AUC values on TRAINING set, and the red line is the AUC values on TEST set

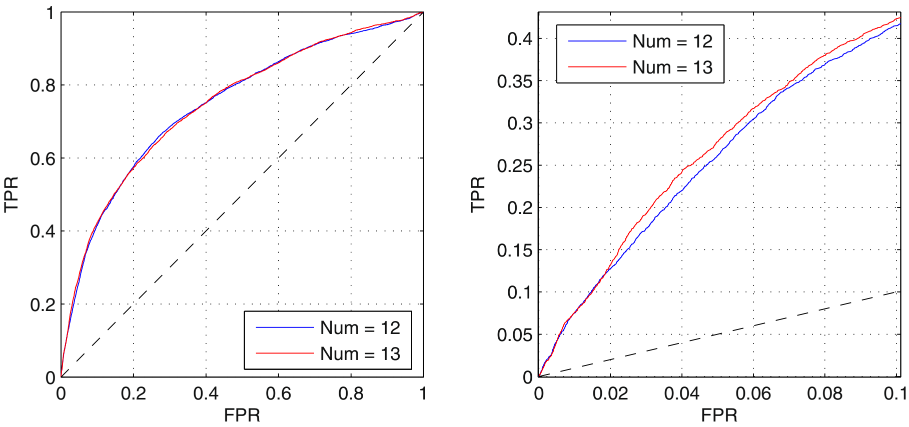

Fig. 2 ROC curves when the feature numbers in candidate sets are 12 and 13. The left figure is the full ROV curves. The right figure is the ROC curves at low FPR

of 4 on TEST set, we finally select the three structural properties which contain topological entropy calculated directly from protein sequences, as well as the Remark 465 and Deleage/Roux propensities from GlobPlot NAR paper [23].

Selecting the appropriate windows sizes

We select three windows to preprocess protein sequences. Based on each window, we calculate the 16 selected features. Thus, each residue can obtain a 48 dimensional feature vector. Then, we change the sizes of three windows, and use MPM to predict MoRFs. The appropriate size of three windows are set by comparing their predictive performance on TRAINING and TEST. Figure 5 shows the predictive performance with different windows sizes. The middle window is always set to the half size of the long window. In the left figure, we fix the size of the long and middle window to 90 and 45, and change the size of the short window from 5 to 11. Obviously, when the short window is set to 10, the AUC is better on TEST set.

Then, we fix the short window to 10 and change the size of the long and middle windows as shown in the right figure of Fig. 4. The long window size is varied from 50 to 110, and the middle window size is changed following the long window. At the beginning, as the long window size increases, the AUC of both data sets increases, and the distance between them decreases. But when the size is larger than 80, the AUC of the two data sets grows slowly, and the distance between them increases. Moreover, when the size is larger than 90, the AUC of TEST tends to be stable. Figure 6 shows the ROC curves on TEST set with the long window size between 90 and 110. In the left figure, the ROC curves of the three sizes almost overlap. However, the ROC curve of 90 is better at low FPR as shown in the right figure. Considering that the proportion of MoRF residues is only about 2% in the TRAINING and TEST sets, we pay more attention to the predictive performance at low FPR. Thus, the long and middle windows are eventually set to 90 and 45.

Considering that researchers may require different precision depending on the applications, we do not set a standard threshold value. However, if one needs a binary categorical prediction, Table 3 provides three threshold values and their predictive results for reference, according to the FPRs on TRAINGING set. The threshold value can be selected in (– 0.5, 0.5), and the larger the value is, the larger the FPR.

Table 2 AA Index accession numbers of selected features

CIDH920101	ROBB760101	CORJ870103	MIYS990104
EISD860103	ROBB760108	CORJ870106	–
NISK860101	ROBB760112	CORJ870107	–
QIAN880105	ROBB760113	CORJ870108	–

These 13 features are collected by simulated annealing algorithm from AA Index

Comparing with other prediction methods

In this part, we compare our method MoRF_{MPM} with ANCHOR, MoRFpred and $\text{MoRF}_{\text{CHiBi}}$ for three test sets TEST, TESTNEW and EXP53. The results of other methods on these three sets are adopted from [16, 28]. Table 4 shows the AUC values for the four methods on TEST and TESTNEW sets. Obviously, MoRF_{MPM}

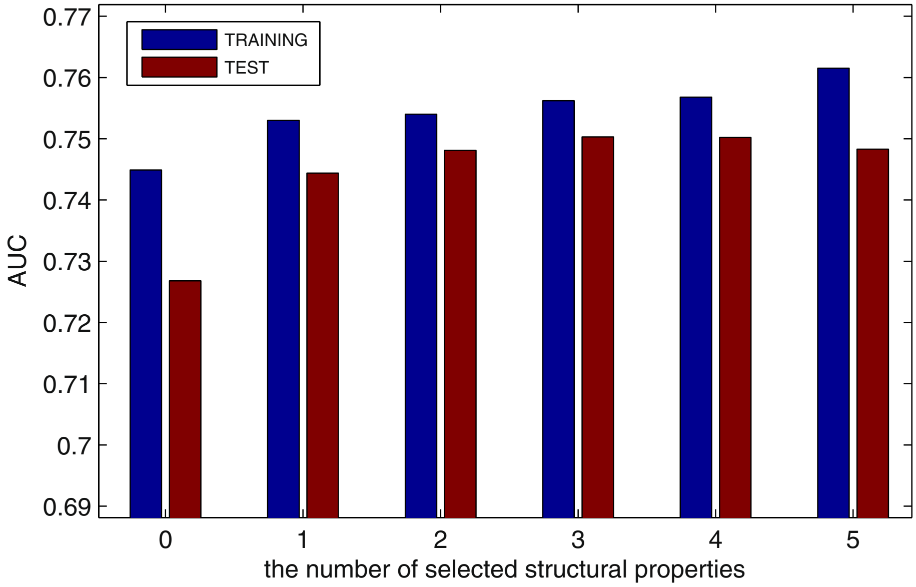

Fig. 3 Predictive performance with different number of structural properties. The blue line is the AUC values on TRAINING set, and the red line is the AUC values on TEST set

achieves higher AUC than ANCHOR, MoRFpred and $\mathrm{MoRF}_{\mathrm{CHiB}}$ on both TEST and TESTNEW sets.

On TEST set, we also compare ACC and FPR at different TPR with other methods, as shown in Table 5. $\mathrm{MoRF}_{\mathrm{MPM}}$ achieves the lower FPRs and higher ACCs on the three TPRs compared with ANCHOR, MoRFpred and $\mathrm{MoRF}_{\mathrm{CHiBi}}$. In other words, $\mathrm{MoRF}_{\mathrm{MPM}}$ can obtain higher TPR at low FPR.

In addition, Table 6 shows the AUC results of these four methods on EXP53 set. In EXP53_short set, only MoRF regions with up to 30 residues are considered, while longer MoRF regions are masked out. In EXP53_long set, only MoRF regions longer than 30 residues are considered, while shorter MoRF regions are masked out [28]. From Table 6, $\mathrm{MoRF}_{\mathrm{MPM}}$ also obtains higher AUC on EXP53_all, EXP53_short and EXP53_long sets.

Discussion

We propose a new method, $\mathrm{MoRF}_{\mathrm{MPM}}$, to predict MoRFs within protein sequences. It uses MPM to train the predictor based on 16 features and 3

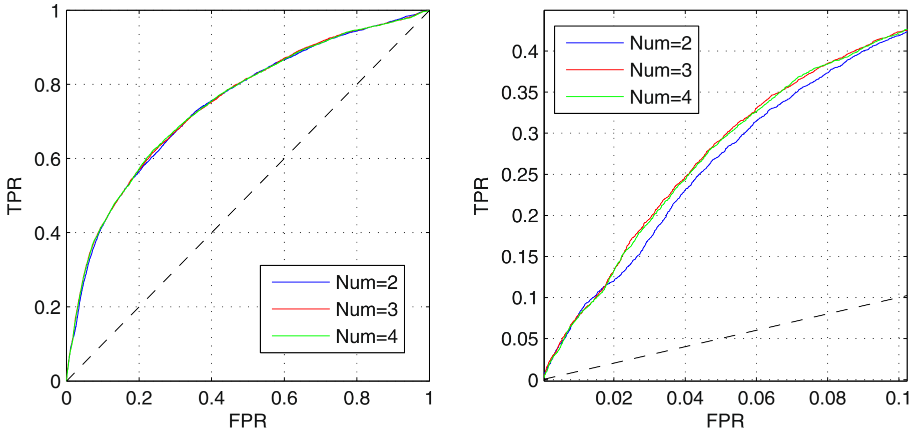

Fig. 4 ROC curves with different number of structural properties. The left figure is the full ROV curves. The right figure is the ROC curves at low FPR

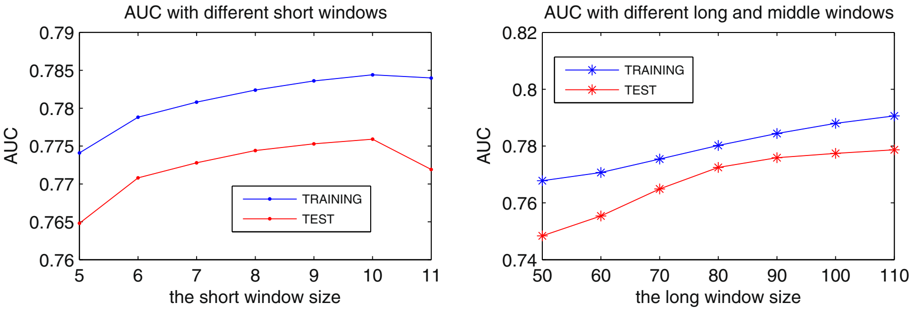

Fig. 5 Predictive performance with different windows sizes. The left figure is the AUC values with different short windows. The right figure is the AUC values with different long and middle windows. The size of the middle window is always the half size of the long window

different windows. The feature set contains 13 physicochemical properties selected from Amino Acid Index and 3 structural properties selected from our previous study [22] about IDPs prediction including topological entropy and two amino acid propensities in GlobPlot NAR paper [23]. We compare MoRF_{MPM} with ANCHOR, MoRFpred and $\text{MoRF}_{\text{CHiBi}}$ on three different test sets: TEST, TESTNEW and EXP53. The results show that MoRFMPM obtains better performance on these test sets.

To further illustrate the predictive performance of MoRF_{MPM}, the protein p53 is predicted as an example, as shown in Fig. 7. The protein p53 is a master protein in tumor regulation, which is one of the most extensively studied IDPs [30, 31]. The N-terminal and C-terminal regions of this protein are confirmed to contain MoRFs [32–34] which are enclosed by the red lines in Fig. 7. The blue line is the predictive results of MoRF_{MPM} for each residue. From Fig. 7, MoRF_{MPM} can effectively identify MoRFs of the protein p53.

The following points enable MoRF_{MPM} to achieve such good performance. First, the appropriate preprocessing highlights the relationship between the residue and its surrounding residues. Second, the feature set used in MoRF_{MPM} is highly effective for predicting MoRFs, especially after preprocessing. Third, instead of considering the properties of Flanks with fix length, MoRF_{MPM} uses a long window of 90 to describe the influence of adjacent areas on MoRFs, and uses a short window of 10 to highlight the properties of MoRFs. Though the long window may contain much non-MoRFs information when calculating the feature vector of MoRF residues, MoRF_{MPM} uses a middle window of 45 to reduce the noise brought by the long window. Finally, although MPM is a linear classification algorithm, it is efficient and robust, especially when there are not too many features used.

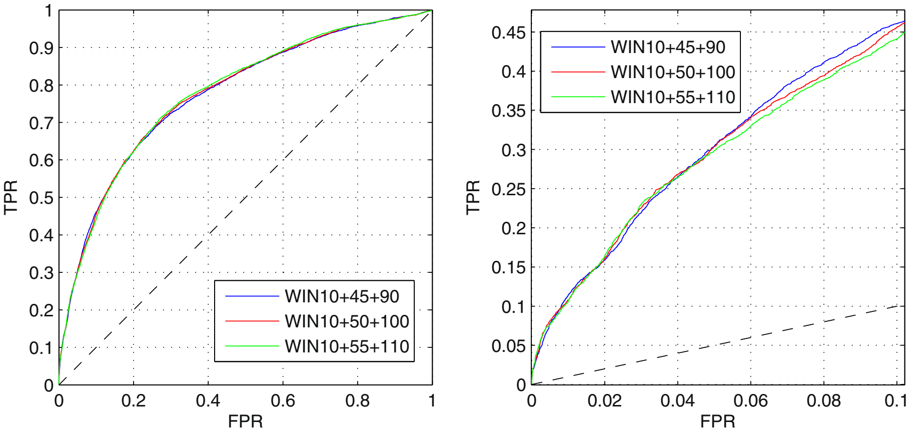

Fig. 6 ROC curves for with different long and middle windows. The short window is set to 10. The left figure is the full ROV curves. The right figure is the ROC curves at low FPR

Table 3 Three threshold values and their predictive results

TRAINING_FPRs	FPR = 0.05		FPR = 0.1		FPR = 0.15	
Thresholds	−0.12		− 0.0735		−0.0438	
	TPR	FPR	TPR	FPR	TPR	FPR
TEST	0.313	0.052	0.458	0.098	0.535	0.141
TESTNEW	0.300	0.037	0.401	0.074	0.470	0.116

The thresholds are calculated by the fixed FPR values on TRAINING set. The default value of the threshold is 0

Table 5 ACC and FPR at different TPR on TEST set

	TPR = 0.222		TPR = 0.254		TPR = 0.389	
	FPR	ACC	FPR	ACC	FPR	ACC
MoRF$_{\mathrm{MPM}}$	0.030	0.955	0.038	0.948	0.072	0.917
MoRF$_{\mathrm{CHiBi}}$	0.035	0.951	0.045	0.942	0.098	0.893
MoRFpred	0.037	0.948	0.049	0.937	0.137	0.854
ANCHOR	0.092	0.894	0.125	0.863	0.253	0.740

FPR and ACC as functions of TPR are calculated on TEST set

Conclusions

In this paper, a new sequence-based method, named as MoRF$_{\mathrm{MPM}}$, is proposed to predict MoRFs. MoRF$_{\mathrm{MPM}}$ calculate 16 features for each residue through preprocessing with 3 different windows, and use MPM to predict MoRFs. MoRF$_{\mathrm{MPM}}$ does not depend on results of other predictors. Comparing with ANCHOR, MoRFpred and MoRF$_{\mathrm{CHiBi}}$ on three different test sets: TEST, TESTNEW and EXP53, MoRF$_{\mathrm{MPM}}$ obtains the best AUC on these test sets. In addition, on TEST set, MoRF$_{\mathrm{MPM}}$ achieves lower FPR and higher ACC when TPR is set to 0.222, 0.254 and 0.389. The predicting code of MoRF$_{\mathrm{MPM}}$ are available at https://github.com/HHJHgithub/MoRFs_MPM, where we also provide an example with the protein p53.

Methods

Preprocessing

To highlight the interrelation between residues, the protein sequences are preprocessed. For a general protein sequence w with length L, we select a window with the length of $N(N < L)$ and fill $N_0 = \lfloor (N-1)/2 \rfloor$ zeros at the beginning and end of the sequence. Then we slide the window to intercept regions of length N successively with step of length 1. At this point, the sequence length becomes $L_0 = L + 2N_0$, and the intercepted region can be denoted as:

$$w_i = w_0(i) \cdots w_0(i+N-1), \quad 1 \le i \le L_0 - N + 1 \ , \tag{2}$$

where w_0 represents the sequence after zero-padding. For each w_i, the values corresponding to the selected features are calculated as following:

$$\mathbf{v}_i = [M_1(w_i)\, M_2(w_i) \cdots\ M_k(w_i) \cdots]^{\mathrm{T}} \ , \quad 1 \le i \le L_0 - N + 1. \tag{3}$$

$M_k(w_i)$ denotes the value of k-th feature calculated on w_i. For one amino acid property, $M_k(w_i)$ denotes the average value of w_i mapped by the scale of the property. For Shannon entropy or topological entropy, $M_k(w_i)$ denotes the value calculated on w_i by their respective formulas [22]. After that, we assign $\mathbf{v}_i$ to each residue in w_i. For each residue, add up all $\mathbf{v}_i$ of them and divide by their respective cumulative number. The feature vector $\mathbf{x}_j$ $(1 \le j \le L)$ of each residue can be expressed as:

$$\mathbf{x}_j = \begin{cases} \dfrac{1}{j+N_0} \displaystyle\sum_{i=1}^{j+N_0} \mathbf{v}_i \ , & 1 \le j \le N_0 \\ \dfrac{1}{N} \displaystyle\sum_{i=j+N_0-N+1}^{j+N_0} \mathbf{v}_i \ , & N_0 < j \le L - N_0 \\ \dfrac{1}{L_0 - j - N_0 + 1} \displaystyle\sum_{i=j+N_0-N+1}^{L_0-N+1} \mathbf{v}_i \ , & L - N_0 < j \le L \end{cases} \tag{4}$$

Feature selection

As mentioned, our feature set contains two parts: properties from AA Index [18] and structural properties. We first select properties from AA Index using simulated annealing algorithm, as shown in Fig. 8.

The detailed steps are as follows:

Table 4 AUC on TEST and TESTNEW sets

	MoRF$_{\mathrm{MPM}}$	MoRF$_{\mathrm{CHiBi}}$	MoRFpred	ANCHOR
TEST	0.777	0.746	0.673	0.600
TESTNEW	0.790	0.770	0.697	0.638

The AUC values of four methods on TEST and TESTNEW sets

Table 6 AUC on EXP53 set

	MoRF$_{\mathrm{MPM}}$	MoRF$_{\mathrm{CHiBi}}$	MoRFpred	ANCHOR
EXP53_all	0.761	0.714	0.620	0.615
EXP53_short	0.814	0.790	0.673	0.683
EXP53_long	0.739	0.681	0.598	0.586

The AUC values of four methods on EXP53_all, EXP53_short and EXP53_long sets

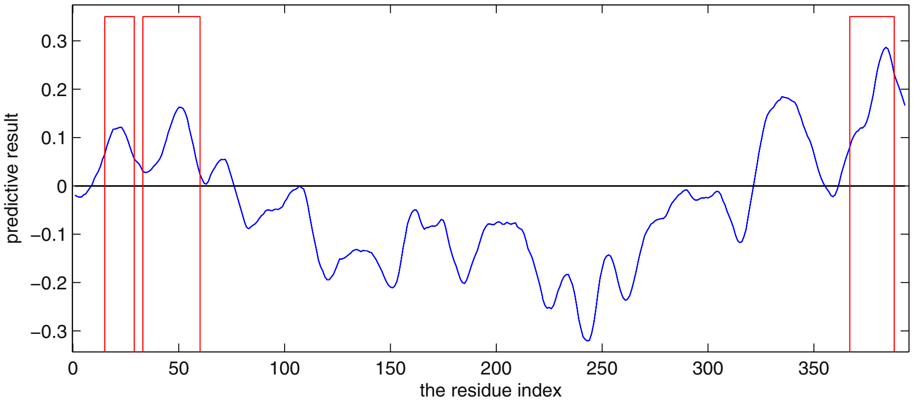

Fig. 7 Predictive results for the protein p53. The blue line is the predictive results of our method. The red lines indicate confirmed MoRFs. The threshold is 0, which is shown as the black line. If the regions in blue line are higher than the black line, they are predicted to be MoRFs

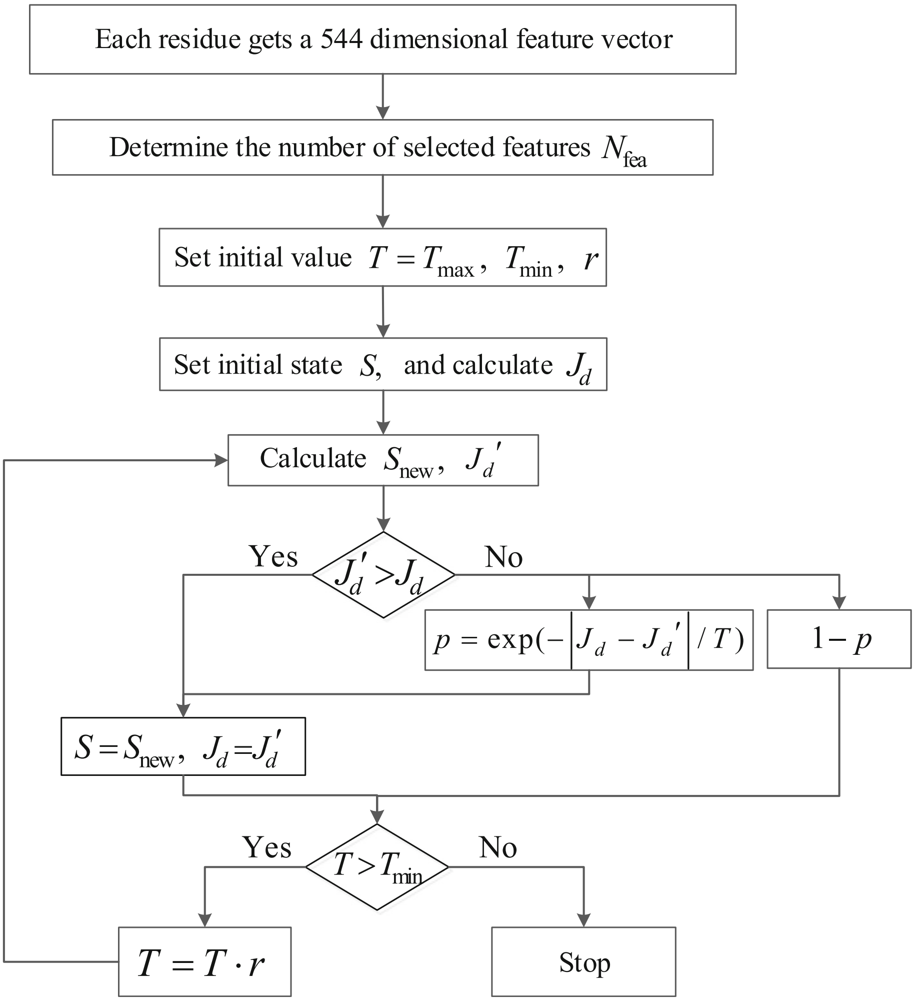
Each residue gets a 544 dimensional feature vector
Determine the number of selected features N_{fea}
Set initial value $T = T_{\max}$, $T_{\min}$, r
Set initial state S, and calculate J_d
Calculate S_{new}, J_d'
$J_d' > J_d$
Yes
No
$p = \exp(-|J_d - J_d'|/T)$
$1-p$
$S = S_{\mathrm{new}}$, $J_d = J_d'$
$T > T_{\min}$
Yes
No
$T = T \cdot r$
Stop

Fig. 8 The process of feature selection by simulated annealing algorithm. Using simulated annealing algorithm, we select properties from AA Index

(1) According to the section of preprocessing, the sequences in TRAINING set are preprocessed based on the 544 amino acid scales from AA Index. Then, each residue can obtain a 544 dimensional feature vector.

(2) Set the number of selected features N_{fea}.

(3) Set the initial temperature $T = T_{max}$, the lower limit temperature T_{min} and the annealing rate r.

(4) N_{fea} features are selected randomly from 544 scales as the initial state S. Then, the distance between MoRF residues and non-MoRF residues is denoted as J_d and calculated using the selected N_{fea} feature vector. J_d can be expressed by $J_d = \text{tr}(\mathbf{S}_w + \mathbf{S}_b)$, where $\mathbf{S}_b$ denotes the between-class scatter matrix $\mathbf{S}_b = \sum_{i=1}^{2} P_i(\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T$ and $\mathbf{S}_w$ is the within-class scatter matrix $\mathbf{S}_w = \sum_{i=1}^{2} P_i \frac{1}{N_i} \sum_{j=1,\ x_j \in X_i}^{N_i} (\mathbf{x}_j - \mathbf{m}_i)(\mathbf{x}_j - \mathbf{m}_i)^T$. Besides, $\mathbf{m}_i$ represents the mean vector of the i-th class and $\mathbf{m}$ represents the total mean vector. Thus, the larger J_d is, the more separable the two types of samples are.

(5) Randomly select a feature that does not belong to state S from 544 scales, and make it replace any one of S to form a new state S_{new}. Calculate the distance $J_d^{'}$ in the new state.

(6) If $J_d^{'} > J_d$, go to (7). Otherwise, calculate $p = \exp(-|J_d - J_d^{'}|/T)$, then go to (7) with probability p and go to (8) with probability $1 - p$.

(7) Set $S = S_{new}$, $J_d = J_d^{'}$.

(8) If $T > T_{min}$, set $T = T \cdot r$ and go to (5). Otherwise, stop iteration.

In this paper, we set $T_{max} = 1$, $T_{min} = 0.0001$, $r = 0.9995$. The parameter N_{fea} is set from 10 to 20, and thus we obtain 11 candidate feature sets. Then, we use the 11 candidate feature sets to train MPM respectively, and select the feature set with the best prediction performance.

In addition, we select structure properties from five features used by our previous research [22] about IDPs prediction which contain Shannon entropy, topological entropy and three propensities from GlobPlot NAR paper [23] (http://globplot.embl.de/html/propensities.html) including the Deleage/Roux, Remark 465 and Bfactor (2STD) propensities. From [22], it has been shown that these five features can effectively predict IDPs. In addition, MoRFs generally locate within longer IDRs. Thus, we add these five features to the feature set obtained from AA index for further selection.

Since MoRFs generally locate within longer IDRs, the protein sequences with MoRFs usually contain three types of residues: MoRF residues, residues flanking (Flanks) the MoRFs and general non-MoRF residues. In other words, the Flanks represent other disordered residues on both sides of MoRFs, and general non-MoRF residues represent the ordered residues in the sequence. The properties of the three types of residues are different from each other. Thus MSPSSMpred and MoRF$_{\text{CHiBi}}$ calculate the properties of Flanks separately, and select 5 and 8 residues on both sides of MoRFs as Flanks respectively. However, the number of Flank residues in each MoRF region is different, and even the number on both sides of one MoRF region is also different. Therefore, instead of calculating the properties of Flanks separately, we consider the impact of Flanks by choosing three different windows. The first window is shorter to highlight the properties of MoRFs, and the second window is longer to highlight the influence of Flanks. The third window is between them to reduce the noise generated by the longer window. The short window is selected from 5 to 11. Meanwhile, since MoRFs generally locate within longer IDRs, we select the long window no less than 50. If the long window is very long, it may contain much non-MoRFs information when calculating the feature vectors of MoRF residues. These non-MoRFs information will reduce the predictive accuracy of MoRFs at low FPR that we are most concerned about, even if we have used a short window. Therefore, we select a middle window half the length of the long window to improve the performance at low FPR.

For selecting the optimum features from 544 amino acid indexes, we just use the short window and set the length to 10, firstly. Through preprocessing the TRAINING set, each residue gets a 544 × 1 feature vector. Then, using simulated annealing algorithm, we select several feature sets with different feature numbers as candidate feature sets. After that, we put the five structural properties into them, and predict MoRFs based on MPM algorithm with the short window of 10 and the long window of 50 to select the best feature set. Finally, we change the number of structural properties to further optimize the feature set.

MPM prediction model

MPM is a machine learning method of statistical learning proposed by Lanckriet et al. [24]. The main idea is to analyze the upper bound of classification error rate and make it as small as possible. Given a feature matrix to be classified $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_{N_s}]$, where N_s denotes the number of samples and $\mathbf{x}_j (1 \leq j \leq N_s)$ denotes the feature vector of the j-th sample. Suppose that these samples are divided into two groups $\mathbf{X}_1$, $\mathbf{X}_2 \in \mathbf{X}$, and $\mathbf{X}_1 \sim (\boldsymbol{\mu}_1, \mathbf{R}_1)$, $\mathbf{X}_2 \sim (\boldsymbol{\mu}_2, \mathbf{R}_2)$. MPM is expected to build a classification surface $\mathbf{W}^{\mathrm{T}}\mathbf{X} = b$, which make the upper bound of the classification error rate as small as possible. Make an assumption that the correct classification satisfies $\mathbf{W}^{\mathrm{T}}\mathbf{X}_1 > b$ for the first group and $\mathbf{W}^{\mathrm{T}}\mathbf{X}_2 < b$ for the second group. The classification error rate is $P\{\mathbf{W}^{\mathrm{T}}\mathbf{X}_1 \le b\}$ for the first group and $P\{\mathbf{W}^{\mathrm{T}}\mathbf{X}_2 \ge b\}$ for the second group. Then the classification surface constructed by MPM should satisfy the following requirements:

$$\min\left[Sup\ P\{\mathbf{W}^{\mathrm{T}}\mathbf{X}_1 \le b\}\right] \quad and \quad min\left[Sup\ P\{\mathbf{W}^{\mathrm{T}}\mathbf{X}_2 \ge b\}\right]. \tag{5}$$

Through a series of solutions, the optimization problem becomes:

$$\max_{\mathbf{W},b} \ \kappa$$

$$s.t. \quad \frac{1}{\kappa} \ge \left(\sqrt{\mathbf{W}^{\mathrm{T}}\mathbf{R}_1\mathbf{W}} + \sqrt{\mathbf{W}^{\mathrm{T}}\mathbf{R}_2\mathbf{W}}\right), \quad \mathbf{W}^{\mathrm{T}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mathbf{1}. \tag{6}$$

Since κ is only an intermediate variable, the optimization problem can be expressed as:

$$\min_{\mathbf{W}} \sqrt{\mathbf{W}^{\mathrm{T}}\mathbf{R}_1\mathbf{W}} + \sqrt{\mathbf{W}^{\mathrm{T}}\mathbf{R}_2\mathbf{W}} \qquad s.t. \quad \mathbf{W}^{\mathrm{T}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \mathbf{1}. \tag{7}$$

The classification surface of MPM is finally reduced to solution formula Eq.7. It is a second order cone program problem, which can be solved by iterative least square method and interior point method. In this paper, we use the iterative least square method given in the reference [29]. Assuming that $\mathbf{W}_*$ is the calculated optimal value, then the optimal κ and b can calculated by:

$$\kappa_* = \frac{1}{\left(\sqrt{\mathbf{W}_*{}^{\mathrm{T}}\mathbf{R}_1\mathbf{W}_*} + \sqrt{\mathbf{W}_*{}^{\mathrm{T}}\mathbf{R}_2\mathbf{W}_*}\right)} \quad , \tag{8}$$

$$\begin{aligned} b_* &= \mathbf{W}_*{}^{\mathrm{T}}\boldsymbol{\mu}_2 + \kappa_*\sqrt{\mathbf{W}_*{}^{\mathrm{T}}\mathbf{R}_2\mathbf{W}_*} \\ &= \mathbf{W}_*{}^{\mathrm{T}}\boldsymbol{\mu}_1 - \kappa_*\sqrt{\mathbf{W}_*{}^{\mathrm{T}}\mathbf{R}_1\mathbf{W}_*} \ . \end{aligned} \tag{9}$$

Prediction process

For a protein sequence to be predicted, the specific prediction process is shown in the Fig. 9. First, the sequence is preprocessed by the selected feature set with three different windows. Then, the calculated feature matrix is input into the trained MPM, and the predicted result is obtained.

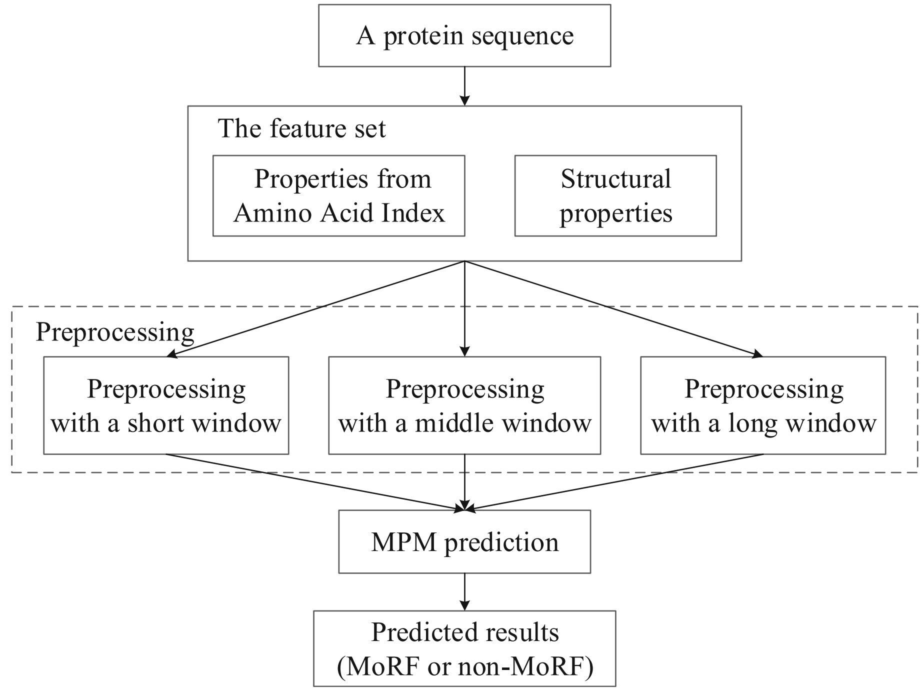
A protein sequence
The feature set
Properties from Amino Acid Index
Structural properties
Preprocessing
Preprocessing with a short window
Preprocessing with a middle window
Preprocessing with a long window
MPM prediction
Predicted results (MoRF or non-MoRF)

Fig. 9 Specific prediction process. Based on the selected feature set, the protein sequence is preprocessed by three different windows, and then is predicted by MPM

Abbreviations

AA Index: Amino Acid Index; IDP: Intrinsically disordered protein; IDR: Intrinsically disordered region; MoRFs: Molecular recognition features; MPM: Minimax probability machine; PDB: Protein Data Bank; PSSM: Position Specific Scoring Matrices; RBF: Radial basis function; SLiMs: Short linear motifs; SVM: Support vector machine

Acknowledgements

We would like to thank Disfani et al. for publicity providing the three data sets for the MoRFs prediction.

Authors' contributions

HH designed the study, carried out data analysis and drafted the manuscript. JZ participated in the design of the study and helped to draft the manuscript. GS participated in the coordination of the study and helped to draft the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China [grant numbers 61771262]. The funding body has no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The datasets supporting the conclusions of this article are available on the references [14, 28].

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 22 November 2018 Accepted: 20 September 2019
Published online: 28 October 2019

References

1. Uversky VN. The mysterious unfoldome: structureless, underappreciated, yet vital part of any given proteome. J Biomed Biotechnol. 2010.
2. Uversky VN. Functional roles of transiently and intrinsically disordered regions within proteins. FEBS J. 2015;282:1182–9.
3. Uversky VN. The most important thing is the tail: multitudinous functionalities of intrinsically disordered protein termini. FEBS Lett. 2013;13: 1891–901.
4. Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT, Kim PM, Kriwacki RW, Oldfield CJ, Pappu RV, Tompa P, Uversky VN, Wright PE, Babu MM. Classification of intrinsically disordered regions and proteins. Chem Rev. 2014;114:6589–631.
5. Fuxreiter M. Fold or not to fold upon binding - does it really matter? Curr Opin Struct Biol. 2018;54:19–25.
6. Pancsa R, Fuxreiter M. Interactions via intrinsically disordered regions: what kind of motifs? IUBMB Life. 2012;64:513–20.
7. Fuxreiter M. Fuzziness in protein interactions-a historical perspective. J Mol Biol. 2018;430:2278–87.
8. Cumberworth A, Lamour G, Babu MM, Gsponer J. Promiscuity as a functional trait: intrinsically disordered regions as central players of interactomes. Biochem J. 2013;454:361–9.
9. Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, Uversky VN. Analysis of molecular recognition features (MoRFs). J Mol Biol. 2006;362: 1043–59.
10. Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, Dunker AK. Coupled folding and binding with alpha-helix-forming molecular recognition elements. Biochemistry. 2005;44:12454–70.
11. Cheng Y, Oldfield CJ, Meng J, Romero P, Uversky VN, Dunker AK. Mining α-helix-forming molecular recognition features with cross species sequence alignments. Biochemistry. 2007;46(47):13468–77.
12. Dosztányi Z, Mészáros B, Simon I. ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics. 2009;25(20): 2745–6.
13. Mészáros B, Simon I, Dosztányi Z. Prediction of protein binding regions in disordered proteins. PLoS Comput Biol. 2009;5:e1000376.
14. Disfani FM, Hsu WL, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, Uversky VN, Kurgan L. MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics. 2012;28(12):i75–83.
15. Fang C, Noguchi T, Tominaga D, Yamana H. MFSPSSMpred identifying short disorder-to-order binding regions in disordered proteins based on contextual local evolutionary conservation. BMC Bioinformatics. 2013;14:300.
16. Malhis N, Gsponer J. Computational identification of MoRFs in protein sequences. Bioinformatics. 2015;31(11):1738–44.
17. Dosztányi Z, Csizmok V, Tompa P, Simon I. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics. 2005;21:3433–4.
18. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 2008;36:D202–5.
19. Faraggi E, Xue B, Zhou Y. Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by fast guided learning through a two-layer neural network. Proteins. 2009;74: 847–56.
20. Schlessinger A, Yachdav G, Rost B. PROFbval: predict flexible and rigid residues in proteins. Bioinformatics. 2006;22:891–3.
21. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.
22. He H, Zhao JX. A low computational complexity scheme for the prediction of intrinsically disordered protein regions. Math Probl Eng. 2018.
23. Linding R, Russell RB, Neduva V, Gibson TJ. Globplot: exploring protein sequences for globularity and disorder. Nucleic Acids Res. 2003;31(13):3701–8.
24. Lanckriet GRG, El GL, Bhattacharyya C, Jordan MI. Minimax probability machine. Neural information processing systems (NIPS) 14. Cambridge: MIT Press; 2002.
25. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide protein data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res. 2007;35:D301–3.
26. Gunasekaran K, Tsai GJ, Nussinov R. Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. J Mol Biol. 2004;341:1327–41.
27. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT suite: a web server for clustering and comparing biological sequences. Bioinformatics. 2010;26:680–2.
28. Malhis N, Wong ETC, Nassar R, Gsponer J. Computational identification of MoRFs in protein sequences using hierarchical application of Bayes rule. PLoS One. 2015. https://doi.org/10.1371/journal.pone.0141603.
29. Lanckriet GRG, Ghaoui LE, Bhattacharyya C, Jordan MI. A robust minimax approach to classification. J Mach Learn Res. 2002;3:555–82.
30. Signorelli S, Cannistraro S, Bizzarri AR. Structural characterization of the intrinsically disordered protein p53 using Raman spectroscopy. Appl Spectrosc. 2016. https://doi.org/10.1177/0003702816651891.
31. Kannan S, Lane DP, Verma CS. Long range recognition and selection in IDPs: the interactions of the C-terminus of p53. Sci Rep. 2016. https://doi.org/10.1038/srep23750.
32. Kussie PH, Gorina S, Marechal V, Elenbaas B, Moreau J, Levine AJ, Pavletich NP. Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. Science. 1996;274:948–53.
33. Bochkareva E, Kaustov L, Ayed A, Yi GS, Lu Y, Pineda-Lucena A, Liao JC, Okorokov AL, Milner J, Arrowsmith CH, Bochkarev A. Single-stranded DNA mimicry in the p53 transactivation domain interaction with replication protein a. Proc Natl Acad Sci U S A. 2005;102:15412–7.
34. Rustandi RR, Baldisseri DM, Weber DJ. Structure of the negative regulatory domain of p53 bound to S100B(ββ). Nat Struct Biol. 2000;7:570–4.