BIOINFORMATICS

Vol. 28 ISMB 2012, pages i75–i83
doi:10.1093/bioinformatics/bts209

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

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

[1]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, T6G 2V4, Canada, [2]Center for Computational Biology and Bioinformatics and Department of Biochemistry and Molecular Biology, Indiana University, Indianapolis, 46202, USA [3]Department of Molecular Medicine, University of South Florida, Tampa, 33612, USA and [4]Institute for Biological Instrumentation, Russian Academy of Sciences, Pushchino, 142290, Russia

ABSTRACT

Motivation: Molecular recognition features (MoRFs) are short binding regions located within longer intrinsically disordered regions that bind to protein partners via disorder-to-order transitions. MoRFs are implicated in important processes including signaling and regulation. However, only a limited number of experimentally validated MoRFs is known, which motivates development of computational methods that predict MoRFs from protein chains.

Results: We introduce a new MoRF predictor, MoRFpred, which identifies all MoRF types (α, β, coil and complex). We develop a comprehensive dataset of annotated MoRFs to build and empirically compare our method. MoRFpred utilizes a novel design in which annotations generated by sequence alignment are fused with predictions generated by a Support Vector Machine (SVM), which uses a custom designed set of sequence-derived features. The features provide information about evolutionary profiles, selected physiochemical properties of amino acids, and predicted disorder, solvent accessibility and B-factors. Empirical evaluation on several datasets shows that MoRFpred outperforms related methods: α-MoRF-Pred that predicts α-MoRFs and ANCHOR which finds disordered regions that become ordered when bound to a globular partner. We show that our predicted (new) MoRF regions have non-random sequence similarity with native MoRFs. We use this observation along with the fact that predictions with higher probability are more accurate to identify putative MoRF regions. We also identify a few sequence-derived hallmarks of MoRFs. They are characterized by dips in the disorder predictions and higher hydrophobicity and stability when compared to adjacent (in the chain) residues.

Availability: http://biomine.ece.ualberta.ca/MoRFpred/; http://biomine.ece.ualberta.ca/MoRFpred/Supplement.pdf

Contact: lkurgan@ece.ualberta.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The existence of disordered proteins challenges the classical structure-to-function paradigm, which states that a unique 3D conformation of a given protein determines its interactions with other molecules. While this paradigm is true for many proteins, disordered proteins (i.e. proteins without a defined structure in isolation) can also be involved in complex interaction networks. There are several examples of protein–protein and protein–nucleic acid interactions that involve coupled folding and binding, i.e. disorder-to-order transition upon binding. Such interactions are significant since they often enable binding diversity, and yet they are specific and reversible due to lower binding strength compared with classical binding. This is especially beneficial in signaling and regulation where highly specific yet dispensable/weak interactions are needed (Uversky and Dunker, 2010). Here, we focus on molecular recognition features (MoRFs), which are short (5–25 residues) binding regions located within longer intrinsically disordered regions. Although in their unbound state MoRFs might or might not have residual structure, they bind to protein partners typically via disorder-to-order transitions resulting in α-helix (α-MoRFs), β-strand (β-MoRFs), coil (γ-MoRFs) or mixtures of these (complex–MoRFs) (Mohan *et al*. 2006) often with partner-dependent conformational differences (Oldfield *et al*., 2008).

These short binding regions have been studied using two computational approaches: as features observed in disorder predictions (Callaghan *et al*., 2004; Cheng *et al*., 2007; Dosztanyi *et al*., 2009; Garner *et al*., 1999; Mészáros *et al*., 2009; Oldfield *et al*., 2005) and as sequence patterns called short sequence motifs (Obenauer *et al*. 2003) or linear motifs (Davey *et al*., 2006; Puntervoll *et al*., 2003).

Long disordered binding regions (more than 30 residues) can also associate with globular partners. These long disordered binding sequences are typically conserved, so they often show up in databases derived from hidden Markov models such as Pfam or SMART (Chen *et al*., 2006a, b). The terms ‘conserved predicted disorder regions’ (Chen *et al*., 2006a, b) and ‘disordered domains’ (Tompa *et al*., 2009) have been used to describe these long disordered binding regions.

Only two related predictors are available: α-MoRF-PredII (Cheng *et al*., 2007) that supersedes α-MoRF-PredI (Oldfield *et al*., 2005) and ANCHOR (Dosztányi *et al*., 2009; Mészáros *et al*., 2009). α-MoRF-PredII is a neural network-based predictor which is limited to prediction of α-MoRFs. ANCHOR predicts binding regions that undergo a disorder-to-order transition upon binding to a globular protein partner, some of which include MoRFs regions.

To this end, we develop a novel sequence-based MoRF predictor. Our approach converts an input sequence into a feature vector that represents various relevant characteristics of this chain, which

*To whom correspondence should be addressed.

is next inputted into a machine learning classifier that generates predictions. Three novel aspects contribute to a good predictive performance offered by our method. First, we use a large dataset of annotated MoRF regions to design and to empirically compare our predictor. Second, we use a comprehensive set of features that encode previously unexplored characteristics of the protein chain and we utilize a well-performing Support Vector Machine (SVM) classifier. Third, we extend our design by combining the SVM-based predictions with annotations generated using sequence alignment. We use our model to describe several sequence-derived hallmarks of MoRF regions and we show that it generates plausible putative short disordered binding sites.

2 METHODS

2.1 Datasets

We collected 4289 protein complexes, which concern interaction between a protein and a small peptide, from Protein Data Bank (PDB) (Berman *et al*., 2007) of March 2008. Each peptide is a putative MoRF that has between 5 and 25 residues, which is consistent with the related work (Cheng *et al*., 2007; Oldfield *et al*., 2005). Next, we removed complexes for which the interaction between the two amino acid chains is not significant enough to be considered as biologically relevant. We measured whether a biologically relevant interaction occurs by calculating the change of accessible surface area (ΔASA) between unbound and bound states. We used BALL library (http://www.bioinf.uni-sb.de/OK/BALL/) to calculate ΔASA and we considered an interaction as spurious if its ΔASA <400 Å^2 (Jones and Thornton, 1996; Vacic *et al*., 2007). The cutoff is intended to be small enough to catch small interfaces between the two chains and, at the same time, large enough to remove spurious contacts. As a result, 452 complexes were removed. Of the remaining complexes, 3148 that include globular partners with more than 70 amino acids (Jones *et al*., 1998) were kept. The cutoff at 70 amino acids was chosen to avoid discarding shorter folded domains. The remaining MoRFs were mapped to the UniProtKB/Swiss-Prot v56.8 and UniProtKB/TrEMBL v39.8 (Jain *et al*., 2009) with FASTA algorithm (http://fasta.bioch.virginia. edu/fasta_www2/fasta_down.shtml) using e-value of 1000 (Pearson and Lipman, 1988). 1805 MoRF segments were successfully mapped to their parent sequences; in the remaining cases the MoRFs were too short to uniquely map to the UniProt or could not be found. A total of 842 MoRFs were left after removing duplicates and MoRFs that include ambiguous amino acids, such as X. We evaluated whether these 842 MoRFs are disordered when unbound. The analysis based on the protocol described in (Gunasekaran *et al*., 2004) shows that all MoRFs are disordered in isolation, see Figure S1 in the Supplement. The amino acids that form these MoRFs were annotated in the parent sequences and these sequences were used to develop and assess our predictor. We divided the dataset into two parts: TRAINING and TEST set, such that chains in the training and test sets share low sequence similarity. We used CD-Hit (Huang *et al*., 2010) with default parameters and sequence identity cutoff at 30% to cluster sequences in the entire dataset. This resulted in 427 clusters where 274 of them include only one sequence. We then assigned each cluster to either training or test set at random. This assures that training and test sets have similar number of chains and that similarity between sequences in these two datasets is $<30\%$. The training dataset was used to develop the method (to perform feature selection and parameterize the prediction algorithm) and test dataset was used to evaluate and compare our method with the existing predictors. Two chains were removed from the test dataset since we could not generate DISOclust predictions, which are needed as inputs for our method, for them. Thus, the training and test datasets have 421 and 419 chains, respectively.

We collected two more test sets: TEST2012 that includes MoRFs annotated using the same protocol, PDB entries deposited between January 1 and March 11, 2012 and UniProtKB release from February 22, 2012; and EXPER2008-12 that includes proteins with MoRFs in regions that were experimentally verified to be disordered in isolation. Exper2008-12 was created using works published between 2008 and 2012. The initial 74 chains from test2012 were screened using (Gunasekaran *et al*., 2004) and 72 of them with MoRFs that were determined as disordered when unbound were kept. Next, test2012 and exper2008-12 were filtered with CD-Hit to remove chains that share $>30\%$ similarity with sequences in the training set. As a result, exper2008-12 and test2012 include 8 and 45 proteins, respectively.

We also developed a negative dataset that includes proteins that (are likely to) have no MoRFs. We selected crystal structures deposited to PDB between January 1, 2010 and March 19, 2012 that have resolution <2.5 Å. We selected only single chain apo structures to eliminate targets which could be disordered in unbound state. Next, we filtered the dataset to keep chains that are at most 30% similar with each other (to evenly sample the sequence space) and removed proteins that had any REMARK 465 annotation, i.e. any annotated disorder. We removed one chain for which disorder predictors failed to generate predictions and proteins for which the amount of disorder predicted with MD (Schlessinger *et al*., 2009) or Spine-D (Zhang *et al*., 2012) was greater than 30%. These two disorder predictors are not used in our MoRFpred. The final set consists of 28 proteins.

The datasets are summarized in Table S1 in the Supplement and are available at http://biomine.ece.ualberta.ca/MoRFpred/

MoRFs were classified into one of the four types: α (helix), β (strand), γ (coil) or complex based on the largest percentage value of their secondary structure types assigned by DSSP (Kabsch and Sander, 1983). MoRFs that have no clear preponderance of any secondary structure type (at least 1% greater than the other two types) are categorized as a complex MoRFs. Only the residues in the interface were counted in the secondary structure classification. Among the 840 MoRFs in the training and test datasets, there are 181 helical, 34 strand, 595 coil and 30 complex MoRF regions. We also annotated MoRF segments in the training and test datasets as those associated with an immune response and others. We used text mining of HEADER, TITLE and KEYWDS records in each corresponding PDB entry to find the following keywords: histocompatibility, MHC, IgG, antigen, antibody, HLA, T cell, B cell, heavy chain, light chain, FAB fragment and cycophilin. We identified 120 immune response-related MoRFs.

The annotated MoRF regions in the training, test and test2012 sets were used to identify full chains in the UniProt and the remaining amino acids in these chains (all amino acids except the residues that compose MoRFs) were by default assumed as non-MoRFs. We anticipate that some of the non-MoRFs could in fact correspond to MoRFs, i.e. our MoRF annotations are incomplete. We address this issue when we design and evaluate our method.

2.2 Test and evaluation protocols

Prediction of MoRFs is performed for each amino acid in the input chain. Each prediction includes a numerical score P that quantifies propensity (probability) of a given amino acid to form a MoRF segment and a binary value that categorizes this amino acid as MoRF or non-MoRF. We compare predictions for a given sequence with its native annotation using two types of assessment: (i) per residue assessment which evaluates predictions for individual amino acids; and (ii) per sequence assessment that looks at the sequence as a whole. Detailed definitions of all evaluation measures are provided in the Supplement.

We use success rate as the per sequence measure. This is motivated by the fact that there might be some un-annotated MoRF regions in our dataset which, if predicted, would count as false positives. The success rate was designed to deal with a similar incompleteness of B-cell epitope predictions (Rubinstein *et al*., 2009). This measure is a real value in the [0, 1] range where 0 and 1 mean that all proteins were predicted incorrectly and correctly, respectively. A higher value indicates a better prediction quality.

For the per residue evaluations, we use three criteria to assess binary predictions: accuracy, true-positive rate (TPR) and false-positive rate (FPR). Another per residue measure is based on the receiver operating characteristic

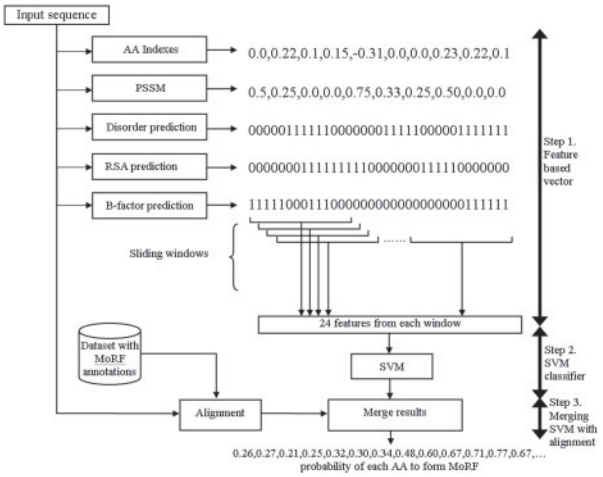

Fig. 1. Architecture of the MoRFpred method

(ROC) curve that examines the predicted probabilities/propensities. We use the area under the ROC curve (AUC) to quantify the predictive quality.

We calculate the measures (success rate, accuracy, TPR, FPR and AUC) using full protein chains. To accommodate the potential incompleteness of the annotations of MoRFs, we also perform the same evaluation on specific regions in a chain that are arguably less likely to contain unannotated MoRF residues. Since MoRF regions are defined as small segments in a larger segment of disorder, we hypothesize that the sequence surrounding a given MoRF region is less likely to contain unannotated MoRFs when compared to the remaining parts of the chain. Figure S2 in the Supplement shows that the difference between the amino acid composition in these flanking regions and MoRF residues is larger than the difference between all non-MoRF and MoRF residues, which supports our hypothesis. Consequently, we perform a 'second' evaluation using a fragment of protein sequence that consists of the MoRF region that has n amino acids and n flanking amino acids on each side of this region. We call this evaluation on the *flanking region*.

To guarantee an unbiased assessment of our method we use the training dataset for design and the independent test sets for the evaluation. The test datasets include sequences that are dissimilar, i.e. they share <30% identity, to sequences in the training dataset. The design, which includes feature selection, parameterization of the SVM and selection of the final method, uses the training set with the 5-fold cross validation protocol. We also use 4+1-fold cross validation on the training dataset. These cross validations are explained in the Supplement.

2.3 Overall architecture

Figure 1 shows the overall architecture of our method, which is called MoRFpred. The first step calculates a feature set that represents each residue in the input sequence using a sliding window, i.e. the calculation of the features is based on a segment of residues centered over the input (to-be-predicted) residue. The use of the window is a popular approach in the design of similar sequence-based predictors (Kurgan and Disfani, 2011). In the second step, a vector of 24 features is fed into a linear SVM to calculate propensity of a given input residue to form a MoRF region. The choice of SVM is motivated by its applications in prediction of disordered regions (Ishida and Kinoshita, 2008; Mizianty *et al*. 2010) and B-factors (Chen *et al*., 2007). Due to a large number of amino acids that need to be predicted, we use a fast liblinear SVM (Fan *et al*. 2008) that was previously utilized in MFDp (Mizianty *et al*. 2010), which is one of the top disorder predictors (Peng and Kurgan, 2011). The output of our method is a real value that quantifies probability of a given residue to form a MoRF region. These values are binarized using a cutoff of 0.5; i.e. amino acids with $P > 0.5$ are assumed to form MoRFs. Finally, in the third step, these propensities are merged with results of alignment of the input protein against the MoRF-annotated proteins in the training dataset to produce the final propensities.

2.4 Feature-based sequence representation

We calculate five types of features that are based on the alignment, amino acid indices and predicted disorder, solvent accessibility and flexibility (measured using B-factor). We utilize IUPredL and IUPredS (Dosztányi *et al*., 2005), DISOPRED2 (Ward *et al*., 2004), DISOclust (McGuffin, 2008) and MFDp (Mizianty *et al*. 2010) to predict disorder. Real-SPINE3 (Faraggi *et al*., 2009) and PROFbval (Schlessinger *et al*., 2006) are used to predict relative solvent accessibility and B-factors, respectively. The choice of the disorder predictors is based on the results from a recent review (Kurgan and Disfani, 2011). These predictions are based on standalone implementations using default parameters. Since training data used by these methods may intersect with our dataset, we validate our method on the test 2012 set that includes depositions from 2012. We also calculate Position Specific Scoring Matrix (PSSM) profiles generated with PSI-BLAST (Altschul *et al*., 1997). These profiles were generated using the non-redundant (nr) database from NCBI that was filtered using PFILT (Jones and Swindells, 2002) to remove low-complexity regions, transmembrane regions and coiled-coil segments. Finally, we represent various physiochemical properties of amino acids with indices collected from the amino acid index database (Kawashima *et al*., 2008). For each type we compute several per residue and aggregated features as explained below. The total number of considered features is 1764.

For each residue in a given input chain, we include information about the residue itself and its neighbors. To do so, we create a sliding window of size 25 that is centered on the input residue and we extract information from this window to calculate the feature set. For the residues on both termini (ends) of the sequence where there are no neighboring residues on the right or left side, we fill these positions with default values. Calculations of the features for each position in the window was inspired by the previous methods, α-MoRF-PredI and α-MoRF-PredII, which used predicted disorder and secondary structure. However, we also use previously unexplored inputs (relative solvent accessibility, B-factors, selected amino acid indices). When calculating the features, we use the predictions in two forms: the probabilities (propensities) and the corresponding binary values.

We also generate another, novel group of features which provide information about a segment in the sequence rather than an individual residue (a position in the window). These features are created by aggregating raw values over a window of a certain size. Simple aggregations include averaging over the window for real valued data or calculating the content for binary valued predictions. We also aggregate by calculating a difference between an average value in a smaller (inner) window and a larger (outside) window. We utilize this aggregation to contrast the values calculated using amino acids that are close to the input (to-be-predicted) residue against the values associated with residues in a wider neighborhood in a sequence. This is motivated by the fact that MoRF segments are usually surrounded by larger disordered segments. While the size of the entire/sliding window is fixed at 25, the size of the inner window is adjusted.

Table S2 in the Supplement defines and summarizes the per-residue (calculated for each position in the sliding window) and the aggregated features for each feature type. We calculate the disorder-based features for each of the five disorder predictors.

2.5 Feature selection and parameterization of SVM

Our datasets are heavily unbalanced, i.e. there is only one MoRF residue for about more than 40 non-MoRF residues. This imbalance is likely to bias a prediction method to under-predict or completely ignore the MoRF regions. To avoid this, we undersample the non-MoRF residues when performing feature selection, parameterization and training. We consider three ways to undersample. As motivated in Section 2.2, in the first strategy we use the non-MoRF residues that are the flanking the MoRF residues (*local sampling*); this results in 2:1 ratio between non-MoRF and MoRF residues. We also use

random sampling with the same 2:1 (two non-MoRFs for each MoRF) and higher 3:1 ratios using the entire chain to select non-MoRFs.

Feature selection is used to select a subset of features relevant to the prediction of MoRF regions. We perform feature selection in three steps using the training dataset. First, a correlation-based score is used to rank the features based on their relevance/relation to MoRF annotations in training dataset. Second, features with lower ranks (below a certain threshold) are removed. Third, a best first search is implemented to pick features that improve predictive results based on cross validation on the training dataset.

We use biserial and ϕ coefficients, which are defined in the Supplement, to rank the features in Step 1. These coefficients quantify the correlation of a given input feature with the native (binary) annotation of MoRFs. We perform this by calculating an average biserial/ϕ correlation over five training folds using the training dataset. We use this average to sort the features in the descending order.

We repeat feature selection nine times, considering three ranking functions executed for the three sampling strategies. We rank the features in three different ways based on:

- Their average (over five training folds) biserial correlations with annotation of MoRFs using the complete training set, i.e. using all residues in the training set (referred to as *complete correlation* ranking)
- Their average (over five training folds) biserial correlations with annotation of MoRFs for the MoRF and the flanking residues, i.e. using the same residues as in the local sampling (*local correlation* ranking)
- Their average success rate calculated when using a single feature on training set to predict the annotation of MoRFs in 5-fold cross validation on training dataset (*success rate* ranking). The predictions are performed using a linear kernel SVM classifier with the default complexity parameter $C=5$ (Fan *et al*. 2008).

We sort the features in the descending order for each of the three rankings and we remove features with correlation <0.05 for the complete and local correlation rankings and with success rate <0.5 in case of the success rate ranking. We selected these thresholds to remove only the irrelevant/poorly performing features. We then execute the best first search on the sorted list of the remaining features. In this search we start with the top ranked feature and we continue by adding one (next-ranked) feature at a time. A given feature is added into the current feature set if it results in improved prediction quality when compared with the methods that uses the current feature set. The predictions are based on a linear kernel SVM classifier with the default parameter $C=5$ and using our modified $4+1$-fold cross validation on the training dataset (see Supplement for details). We calculate the success rates for both the 4-fold cross validation and the independent fifth fold in the $4+1$-cross validation and compare these with the currently best success rates. If the newly added feature improves the success rate by at least 0.01 on both tests then we add the feature. If the success rate improves in only one or none of the tests we discard the feature and move on to the next ranked feature.

Table S3 in the Supplement summarizes the results of the cross validation on the training set for the nine feature selection setups; 3 (sampling strategies) $\times$ 3 (ranking methods). For each sampling, the last row of the table also presents results of a model that uses a feature set that combines the features selected by all three feature selection methods. We select the best performing setup by considering predictive performance on both the flanking region and the whole sequence. Considering the average (over the flanking region and the whole sequence) AUCs and success rates (the last two columns in Table S3 in the Supplement) we note that the model based on the local sampling and combined features have the highest average AUC and a reasonably high success rate. Thus, we select this setup to implement our method.

Next, the selected feature set is used to parameterize the SVM model, i.e. to optimize value of parameter C, utilizing the $4+1$-fold cross validation on the training dataset. We consider $C=2^x$, where $x=-13, -12, \ldots, 8, 9$ and select $C=2^{-6}$ which has the highest success rate on the independent test fold, see Figure S3 in the Supplement. SVM generates results with similarly high-quality for a wide range of values of C, between 2^{-8} and 2^8.

2.6 Alignment-based MoRF prediction

We align target proteins in a given test set (a test fold when using cross validation on the training set) against chains from the training set, which are annotated with MoRF regions, using PSI-BLAST with default parameters. For each target chain we get a number of matching/similar sequences in the training set, with their associated e-values that quantify similarity. These matches indicate sequences in the training set that are (partly) aligned with our target sequence. If the amino acids that are aligned between the target and the training sequence contain a MoRF region, then we transfer these annotations into the target sequence. We tested different thresholds for the e-value using the training dataset by merging the results of the SVM with the annotations transferred through the alignment. We picked e-value $= 0.5$ which provides the best AUC and success rate. We use the sequences with e-value < 0.5 and discard the remaining alignments.

We add the annotations acquired from the alignment to the SVM predictions by updating the probabilities generated by the SVM. For the residues that are predicted as MoRFs by alignment and as non-MoRFs by SVM (SVM generates $P < 0.5$), we add 1 to the probabilities generated by SVM and divide the result by 2; consequently, these residues will be predicted as MoRF residues. We use the probability generated by the SVM for the remaining residues.

2.7 Prediction of MoRF regions by merging SVM and alignment

We test our best selected setup, i.e. the SVM model with $C=2^{-6}$ on the locally sampled training set using the combined feature set, on the independent test dataset (using the model built on the training dataset). Table S4 in the Supplement presents results of prediction before and after merging the alignment-based predictions. The results are slightly improved after incorporating alignment; the AUC is higher by 0.01 and TPR by 3%. We also evaluate the alignment-only-based results in the last row. Although alignment helps to improve the predictive performance of the SVM, it cannot be used alone as an accurate MoRF predictor. The potential reasons are that the MoRF regions are relatively short, compared to the length of the entire chain, and the fact that our training and test datasets share $< 30\%$ similarity.

3 RESULTS

3.1 Comparison with existing predictors

We empirically compare our MoRFpred method with the three available related predictors: α-MoRF-PredI, α-MoRF-PredII and ANCHOR. We evaluate results on the test set on the whole sequences and the flanking region (MoRF region and its flanking areas; see Section 2.2 for details). The α-MoRF-PredI and α-MoRF-PredII predictors provide only the binary values, which means that their AUC cannot be calculated. We also evaluate the statistical significance of the improvements offered by MoRFpred. We compare 10 paired results for success rate and AUC obtained using the bootstrapping with 50% of randomly selected test chains. We determine normality of a given measurement with Anderson–Darling test at the 0.05 significance. For normal distributions, we use paired t-test and otherwise we use Wilcoxon rank-sum test. Finally, we measure significance of the differences at the 0.01 and 0.05 levels. The results are summarized in Table 1.

We observe a relatively large gap between the success rates of α-MoRF-PredI and α-MoRF-PredII predictors and the results

Table 1. Comparison of prediction results on the test dataset

Predictor	Whole sequence					Flanking region				
	ACC	TPR	FPR	Success rate	AUC	ACC	TPR	FPR	Success rate	AUC
α-MoRF-PredI	0.946	0.123	0.037	0.158 ++	NA	0.668	0.123	0.065	0.129 ++	NA
α-MoRF-PredII	0.889	0.258	0.098	0.303 ++	NA	0.673	0.258	0.124	0.263 ++	NA
ANCHOR	0.740	0.389	0.253	0.611 ++	0.600 ++	0.640	0.389	0.237	0.659 ++	0.590 ++
MD	0.612	0.485	0.386	0.480 ++	0.598 ++	0.437	0.485	0.586	0.308 ++	0.449 ++
MoRFpred (SVM + alignment)	0.937	0.254	0.049	0.718	0.673	0.711	0.254	0.065	0.754	0.684
MoRFpred (to match the highest TPR)	0.854	0.389	0.137	0.718	0.673	0.696	0.389	0.153	0.754	0.684
MoRFpred (to match the lowest FPR)	0.948	0.222	0.037	0.718	0.673	0.711	0.254	0.065	0.754	0.684

The last two rows show results for MoRFpred where the binary predictions were calculated (by adjusting the threshold on probabilities) to match the highest TRP and FPR generated by ANCHOR and α-MoRF-PredI and II. The two main columns list results when using the whole chain and only the flanking region (see Section 2.2). α-MorfPredII and II generate only binary predictions and thus their AUC cannot be computed. Statistical significance of the differences in the success rates and AUC between the MoRFpred and the other three methods is shown next to success rate and AUC values, where ++ and + denote that the improvement is significant at the *P*-value <0.01 and <0.05, respectively.

generated by ANCHOR and MoRFpred. This is since the former two predictors were developed to identify MoRF regions that form only α-helixes upon binding. Table 1 shows that MoRFpred outperforms ANCHOR in terms of both AUC and success rate by 0.07 and 10%, respectively. These are relatively substantial improvements considering that AUC ranges between 0.5 (which corresponds to random predictor) and 1 and success rate between 0 and 1 (100%). Both improvements are statistically significant at 0.01 for both evaluations; i.e. using the whole chains and the flanking regions. We note that ANCHOR was designed to find a different type of binding regions that undergo a disorder-to-order transition, which explains its lower performance on our MoRF-based datasets. The binary predictions generated by our method are characterized by low FPR and relatively high TPR. To compare the binary predictions side-by-side, we added last two rows in Table 1 where we match MoRFpred's TPR and FPR to the highest TPR and lowest FPR of the other three methods, respectively. We match these by adjusting the thresholds on the predicted probabilities and we perform that separately for the evaluations on the whole sequence and the flanking regions. These results demonstrate that MoRFpred outperforms the related solutions by providing substantially higher TPRs given similar FPRs, and lower FPRs for comparable TPRs.

Figure 2 presents the ROC curves for MoRFpred and ANCHOR on the test set. The curves zoom on FPR values <0.1, which is motivated by the imbalanced nature of our dataset; i.e. higher FPR would result in substantial over-prediction of MoRFs. We note a large separation between MoRFpred and ANCHOR and the fact that addition of the alignment into MoRFpred results in improvements for all values of FPRs. Figure S4 in the Supplement shows the ROC curve for the flanking region where MoRFpred also obtains favorable results.

We also compare with modern disorder predictors, including the five methods used by MoRFpred (IUPredL, IUPredS, DISOPRED2, DISOclust and MFDp), MD (Schlessinger *et al*., 2009) and Spine-D (Zhang *et al*., 2012) assuming that all predicted disordered residues are MoRFs. Table 1 lists the results for the MD method that has the highest AUC on the whole chains; Table S5 in the Supplement gives the remaining results. As expected these methods overpredict MoRFs; they have relatively high FPR. Based on their relatively low success rates and AUCs we conclude that they could not be used to accurately identify MoRFs.

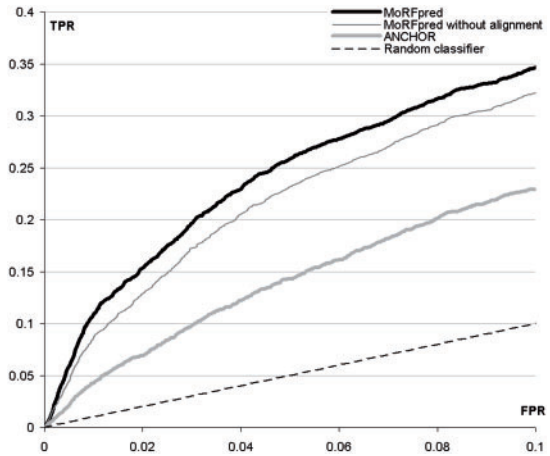

Fig. 2. Comparison of ROCs for MoRFpred and ANCHOR on the test dataset. The ROC curves are provided for the FPR < 0.1

Table S6 in the Supplement summarizes results on the test dataset when using only the predicted disordered residues (excluding MoRF residues) as the negatives. This allows us to evaluate how well the MoRFs can be distinguished from other disordered amino acids. We use a majority-vote based on the predictions from recent, well-performing Spine-D, MD and MFDp to annotate disordered residues. MoRFpred provides the most accurate results with success rate and AUC = 0.68 and 0.65, respectively. The other methods have AUCs below 0.5 and relatively high FPRs, except for MoRF-PredI which has the same FPR and substantially lower TPR when compared to MoRFpred. Overall, the results suggest that our predictor can find MoRFs among disordered residues.

The comparison on the test2012 and exper2008-12 datasets is presented in Table S7 in the Supplement. The results are in agreement with the predictive performance on the test dataset. When evaluated on the entire chains, MoRFpred obtains success rates of 0.76 and 0.75 and AUC of 0.7 and 0.64 on the test 2012 and exper2008-12 datasets, respectively. To compare, ANCHOR and best performing disorder predictor MD have AUCs of 0.64 and 0.68 on the test2012 and 0.56 and 0.62 on exper2008-12, respectively.

We note that the improvements (relative differences between predictors) are consistent for the evaluations with the whole sequences and the flanking regions. The differences in absolute values of accuracy between these two evaluations are due to different ratios of non-MoRF to MoRF residues.

Finally, we estimate FPR of different predictors on the negative dataset. MoRFpred predicts 6.5% of residues in this dataset as MoRFs compared to 0% for MoRF-PredI and α-MoRF-PredII and 0.5% for ANCHOR. We note that Spine-D and MD predict 7.4 and 3.1% residues in this dataset as disordered; there were no chains where both MD and Spine-D predicted no disorder and only 8 for which they predicted less than 5% of disordered amino acids. The relatively high FPR of MoRFpred could be explained by these residual amounts of predicted disorder and the fact that the conformations of ordered segments are similar to the MoRFs when they become ordered upon binding.

3.2 Evaluation for different MoRF types

We evaluate the considered methods separately for each MoRF type, see Table S8 in the Supplement. The results show that MoRFpred outperforms the other three methods with respect to the success rates and AUC for the four MoRF types. The evaluation on the α-MoRFs shows a visible improvement for α-MoRF-PredI and α-MoRF-PredII when compared to their predictions on the other MoRF types. This is expected since these methods were designed for the prediction of the α-MoRFs. However, ANCHOR and MoRFpred are still better than α-MoRF-Pred methods for all four MoRF types. The success rates of MoRFpred are higher by 4%, 12% and >5% than ANCHOR for the prediction of the α-, coil- and complex-MoRFs, respectively. All methods perform relatively poorly for the predictions of β-MoRFs, although MoRFpred still outperforms the other solutions. However, we note relatively low numbers of the β- and complex–MoRFs in our dataset, which could affect the validity of our conclusions.

The alignment only-based predictions have low TPRs coupled with very low ($\sim$0) FPRs for all MoRF types. This shows that alignment predicts a few MoRFs but with high quality. The alignment contributes 3 to 6% to the TPR of the MoRFpred for the α-, β- and coil MoRF types.

MoRFpred produces the most accurate results (high AUC and success rate values) for the α-MoRFs. This likely originates from the fact that helices are local (in the sequence) and thus they are easier to capture using a window-based approach that is implemented by our method. β-sheets can span over large stretches of sequence and thus the window may not be sufficient to find them.

Table S9 in the Supplement contains evaluations for the immune response-related MoRFs (which account for 18% of MoRFs in our test dataset) versus the remaining MoRFs. Our method outperforms the other approaches for the non-immune MoRFs. However, for the immune response-related MoRFs, ANCHOR and MoRFpred offer similar predictive performance. We note that all considered methods perform relatively poorly for these MoRFs, which motivates further research in this area.

3.3 Similarity analysis

We investigate a hypothesis that MoRF regions have non-random similarity to each other. If true, this could be used to support our claim that some of our 'false positive' MoRF predictions might correspond to true MoRF regions. We create four different sets of protein chain segments which are used to investigate the similarity:

- The native MoRF segments in the test dataset.
- Random segments generated from test dataset that have the same length distribution and number when compared to the set of the native MoRF segments.
- MoRF segments predicted by MoRFpred on the test set that have at least 50% overlap with the native MoRFs in test set.
- MoRF segments predicted by MoRFpred on the test set that have no overlap with the native MoRFs (predicted 'false positive' MoRFs).

We use the native MoRFs in the training dataset as our reference population against which we align the four abovementioned sets. We measure the similarity using EMBOSS needle (Rice *et al.*, 2000) with default parameters. Each random, native, and predicted MoRF segment is aligned against the 421 native MoRFs in the reference population and we use the maximum score from the 421 similarities. We obtain four sets of scores for the native MoRFs on the test set (called *test*), random set (called *random*), predicted overlapping MoRFs (called *overlapping predictions*), and predicted non-overlapping MoRFs (called *non-overlapping predictions*). Using these scores, we generate distributions that are fitted into the data using EasyFit (http://www.mathwave.com/products/easyfit.html). We tried six commonly used distributions: normal, log-normal, γ, β, Pearson 5 and Pearson 6. Their fit into the data was evaluated using the Kolmogorov-Smirnov goodness of fit test. We use the Pearson 5 distribution which provided the best rank when considering the four sets of similarity scores.

Figure S5 in the Supplement depicts the distributions of the four sets of similarities. The distribution of the similarities for the native test group (using the native MoRFs) has a higher and longer right tail when compared to the distribution of the random group (for the randomly generated segments). This means that the native MoRFs have higher similarity to each other when compared to their similarity with randomly selected segments. This motivates the use of the alignment in our MoRFpred. Moreover, both distributions for the predicted MoRFs are also characterized by higher than random, (i.e. higher than for the randomly chosen segments) similarity. The shift to the right of the distribution for the overlapping (with the native MoRFs) predicted MoRFs, when compared with the distribution for the native test group, means that our overlapping predictions tend to focus on MoRFs that are similar to the MoRFs in the training set. This is likely due to the use of the sequence alignment. However, this bias is relatively minor considering that the two distributions are shifted by only 0.05. Importantly, the distribution for the predicted non-overlapping MoRFs (predicted 'false positives') is shifted toward higher similarity when compared with the random group. This suggests that some of our 'false positives'/putative MoRFs may correspond to native MoRFs.

3.4 Sequence-derived hallmarks of MoRFs

We describe a few sequence-derived markers of MoRF residues based on the features that were selected to implement the MoRFpred. MoRFpred uses 24 features which were selected using three feature selection methods. To analyze the selected features, we sort them in

the ascending order based on their average ranking generated by the three selection protocols; see Section 2.5 for details. We calculate the average values of the top ranked features for the native MoRF and non-MoRF residues (in the flanking region), respectively. These averages for the five top-ranked features are compared in Figure 3. The values of these features have opposite signs for the native MoRF and non-MoRF residues. However, the large and overlapping standard deviations (denoted by error bars) show that they could not be used individually to accurately find MoRFs. This is why we employ multiple features in our predictor.

The three left-most sets of bars in Figure 3 represent the same type of features, which is based on the average difference of disorder probabilities (Table S2 in the Supplement) calculated using predictions from IUPred with $w = 15$ and from DISOPRED2 with $w = 5$ and $w = 15$, respectively. These features were designed to contrast the value of the predicted disorder propensities in a MoRF region (inner window) and the flanking segments (outside window). According to Oldfield *et al.*, (2005), MoRFs are short segments inside a larger disordered segment. Therefore, we expect a higher average of predicted disorder propensities in the outside window when compared to inner window, which should result in a positive value for our features for the native MoRF residues. This is confirmed in Figure 3 where the three features have, on average, positive values for the MoRF residues and negative (near zero) values for the non-MoRF residues. This shows that MoRFs are characterized by dips (lower values) in the predicted disorder probabilities when compared with the surrounding residues.

The right-most two sets of bars in Figure 3 show average difference-based features (features based on the differences in values between the inner and outside windows) calculated using two amino acid indices, which quantify stability (Zhou and Zhou, 2004) and transfer energy that translates into hydrophobicity (Nozaki and Tanford, 1971), respectively. The stability scale is a quantity used to characterize the contributions of individual residues to stability of a protein fold where higher values mean higher stability. We observe

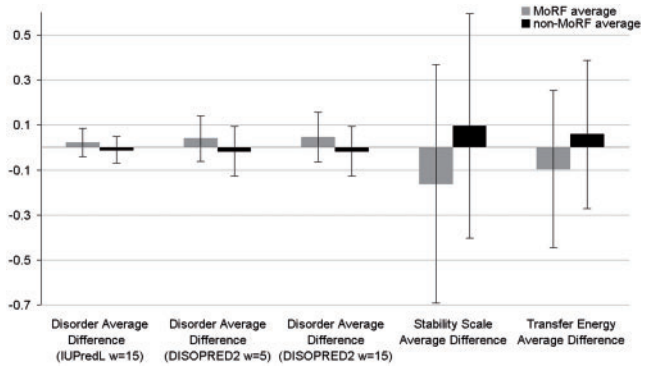

Fig. 3. Analysis of the top-ranked features that serve as sequence-derived markers of MoRFs. The average values of the top five ranked features used by MoRFpred, which are shown on the *x*-axis, for the native MoRF residues (light gray bars) and native non-MoRF residues (dark gray bars) are compared. The corresponding standard deviations are shown using the error bars. The selected five features represent an average difference of a given quantity (predicted disorder, stability or transfer energy). Negative values mean that average in the inner window of size w was higher than the average in the flanking segments

that the average difference for residues in the native MoRF region for the stability-based feature is negative. That means that residues in the MoRF region have higher stability when compared to the surrounding residues. This agrees with the underlying biology, since MoRF residues should be more stable to transition into the structured state when compared to the flanking residues that are likely to be (more) disordered. The last feature is based on the hydrophobicity. The negative value of this feature for the native MoRF residues indicates that these residues are, on average, more hydrophobic than the surrounding residues.

Our features reveal a few interesting sequence-derived markers of MoRF residues. These residues are less disordered, more stable and more hydrophobic when compared to the disordered residues that surround them in a protein chain. This is in line with the observations in (Mészáros *et al.*, 2007; Vacic *et al.*, 2007) that the local increase in the hydrophobicity in a disordered segment is a characteristic of binding sites in IDPs. Bastolla *et al.*, (2005) show a strong positive correlation between a hydrophobicity profile and a contact matrix (i.e. the residue-residue contacts that quantify stability), which describes stability of the protein structure. This supports our result that also shows that increase in both stability and hydrophobicity are indicative of MoRF residues. Importantly, our model shows that these markers can be derived from sequence based on the predicted (with IUPred and DISOPRED2) disorder and two amino acid scales (Nozaki and Tanford, 1971; Zhou and Zhou, 2004).

3.5 Case studies

3.5.1 Case studies for true positive predictions Two case studies are used to demonstrate MoRFpred predictions. They were selected from the test dataset to represent two situations: when the MoRFs are under-predicted and when they are potentially over-predicted. Moreover, the first case concerns a long native MoRF segment, while the second concerns a short segment.

The first case study is a 89 residues long H2A class histone protein for which MoRF region was extracted from the 1ydp_P complex from PDB. The native MoRF region in this protein folds into irregular structure (coil), which is located near the C-termini and is 9 residues long. Figure 4 shows that α-Morf-PredI and α-Morf-PredII did not predict any MoRF residues in this sequence, which is correct since these methods are designed to predict α-MoRFs. MoRFpred predicted the native MoRF region and another 'false positive' MoRF region, which was also predicted by ANCHOR. The probability profiles of ANCHOR and MoRFpred are very similar except for the C-terminus where the native MoRF is located. The ANCHOR outputs probabilities that are > 0.5 threshold in a vicinity of 55th position, but these predictions were removed through post-processing used by this method. MoRFpred generated $P < 0.5$ in that region.

The second case study is the transcriptional intermediary factor-2 isoform 2 protein that was collected from UniProt and for which MoRF was extracted from the PDB complex 1m2z_B. This protein has 1394 residues and contains a coil MoRF region which is 21 residues long. Figure S6 in the Supplement visualizes predictions for this protein from the four considered predictors: α-MoRF-PredI, α-MoRF-PredII, ANCHOR and MoRFpred. All methods were able to (partially) identify the native MoRF region; however, they also predicted additional MoRFs in this chain. MoRFpred

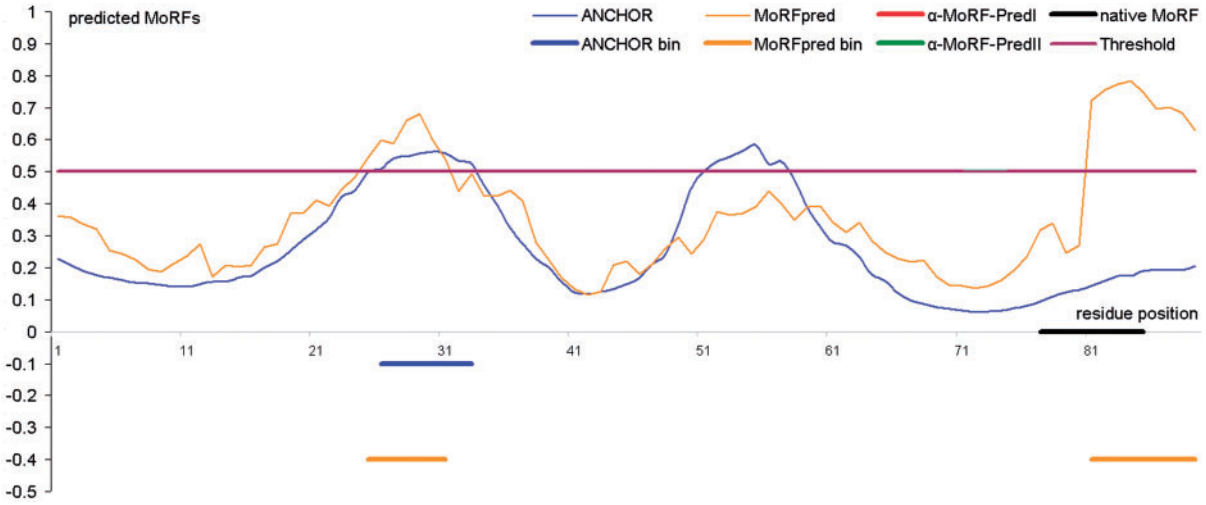

Fig. 4. Prediction of MoRF residues for the Histone H2A protein by ANCHOR (blue lines), MoRFpred (orange lines), α-MoRF-PredI (thick red line) and α-MoRF-PredI I (thick green line) predictors. The x-axis shows positions in the protein sequence. Probability values are only available for ANCHOR and MoRFpred and are shown by thin blue and orange lines, respectively, at the top of the figure. The cutoff of 0.5 to convert probabilities into binary predictions for ANCHOR and MoRFpred is shown using a brown horizontal line. The native MoRF regions are annotated using black horizontal line. The binary predictions from ANCHOR, α-MoRF-PredI, α-MoRF-PredI I and MoRFpred are denoted using blue (at the -0.1 point on the y-axis), red (at the -0.2), green (at the -0.3) and orange (at the -0.4) horizontal lines. Lack of red and green lines means that α-MoRF-PredI and α-MoRF-PredII did not predict MoRFs

predicts the smallest number of 89 of MoRF residues when compared to α-MoRF-PredI with 171, α-MoRF-PredII with 306 and ANCHOR with 876 MoRF predictions. MoRFpred predictions in these case studies were generated by the SVM (alignment did not find MoRF regions), which confirms that the machine learning classifier contributes beyond what can be found based on sequence similarity.

3.5.2 Case studies for 'false positive' predictions In Section 2.1 we argue that our datasets may contain unannotated MoRFs and thus some of the 'false positive' MoRF predictions might correspond to native MoRFs. We also show that MoRFs predicted by MoRFpred that have no overlap with the native MoRFs ('false positive' predictions) have above-random similarity to the native MoRFs. This led us to investigate the strongest 'false positive' MoRF predictions, i.e. predictions with the highest probability (see Section 2.1 in the Supplement). We use average (over all residues in the predicted segment) probability generated by MoRFpred to rank 'false positive' MoRFs. We use UniProt to annotate binding sites in these regions. The following two cases (among others) have binding regions in the predicted MoRFs.

The first case is P-selectin glycoprotein ligand 1 (PSGL-1) protein (UniProt ID Q14242) for which the predicted 'false positive' MoRF has average probability of 0.85. This region (DDLTLHSFLP, residues 393 to 402) is predicted by the SVM and was not found by alignment. It implements a few interaction sites:

- The predicted MoRF region includes a site phosphorylated by Polo-like kinase (DDLTLHS, residues 393–399).
- A part of the MAPK docking motif (REDREGDDLTL, residues 387–397) that helps to regulate a specific interaction in the MAPK cascade overlaps with the predicted MoRF region.
- Our prediction is also close to the TRAF6-binding site (PEPREDREG, residues 384–392), that acts as intracellular adaptor recruited to different receptors through its C-terminal TRAF domain.

ANCHOR predicts binding regions for about half of this protein, which includes the above region; this could be correct since this protein is known to bind multiple partners. α-MoRF-PredII predicts this region as MoRF with only a moderate number of additional MoRF predictions. This region was predicted to be disordered by the majority-vote consensus of MFDp, MD and Spine-D.

The second case is putative uncharacterized protein DKFZp459P0162 (UniProt ID Q5RDR1) for which the 'false-positive' MoRF was predicted with average probability of 0.65. The predicted region (SPAVPNKEVTP, residues 212–222) is associated with the following binding sites:

- This region includes the subtilisin/kexin isozyme-1 (SKI1) cleavage site (KEVTP, residues 218–222)
- The PAVPNK sub-segment (residues 213–218) is potentially recognized by class II SH3 domains and is involved in protein–protein interaction mediated by SH3 domains.

In contrast to the first case, this region is predicted by alignment and none of the existing predictors were able to (fully) predict it. ANCHOR, which predicts about one-third of this protein as binding regions, predicts only parts of this region. The majority-vote consensus of MFDp, MD and Spine-D predicts this region as disordered. This supports our claim that this is a strong putative MoRF.

These two case studies demonstrate that some of the 'false positives' generated by MoRFpred may implement important binding events that require a structured conformation.

4 SUMMARY AND DISCUSSION

We introduce a new and accurate sequence-based predictor of MoRFs. The improvements offered by MoRFpred, when compared to prior methods, are due to the use of large dataset and novel architecture that combines SVM-based predictions with alignment and which uses a comprehensive and well designed feature-based sequence representation. We utilize multiple disorder predictions, predicted B-factors and RSA, evolutionary profiles and selected amino acid indices to encode inputs to the SVM. We designed a successful class of features to contrast the properties in the immediate neighborhood of the predicted residues with its flanking regions.

Our analysis reveals a few interesting sequence-derived markers of MoRF regions. We also demonstrate that MoRFpred can be used to identify putative MoRFs.

ACKNOWLEDGEMENTS

We acknowledge contributions of Dr Istvan Simon group in providing access to their ANCHOR server and constructive criticism provided by the anonymous reviewers.

Funding: This work was supported in part by the NSERC Discovery grant 298328-07 to LK, by the Russian Academy of Sciences 'Molecular and Cellular Biology' program to VU, and by NIH grant R01 LM007688-01A1 and NSF grant EF 0849803 to AKD and VU. MJM is the recipient of the Izaak Walton Killam Memorial Scholarship.

Conflict of Interest: none declared.

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.

Bastolla,U. *et al*. (2005) Principal eigenvector of contact matrices and hydrophobicity profiles in proteins. *Proteins*, **58**, 22–30.

Berman,H. *et al*. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.

Callaghan,A.J. *et al*. (2004) Studies of the RNA degradosome-organizing domain of the Escherichia coli ribonucleaseRNase E. *J. Mol. Biol*., **340**, 965–979.

Chen,J.W. *et al*. (2006a) Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. *J. Proteome Res*., **5**, 879–887.

Chen,J.W. *et al*. (2006b) Conservation of intrinsic disorder in protein domains and families: II. functions of conserved disorder. *J. Proteome Res*., **5**, 888-898.

Chen,P. *et al*. (2007) Prediction of protein B-factors using multi-class bounded SVM. *Protein Pept. Lett*., **14**, 185–190.

Cheng,Y. *et al*. (2007) Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry*, **46**, 13468–13477.

Davey,N.E. *et al*. (2006) SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent. *Nucleic Acids Res*., **34**, 3546–3554.

Dosztányi,Z. *et al*. (2009) ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics*, **25**, 2745–2746.

Dosztányi,Z. *et al*. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.

Fan,R.E. *et al*. (2008) LIBLINEAR: a library for large linear classification. *J. Mach. Learn. Res.*, **9**, 1871–1874.

Faraggi,E. *et al*. (2009) 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*, **74**, 847–856.

Garner,E., *et al*. (1999) Predicting binding regions within disordered proteins. *Genome Informatics*, **10**, 41–50.

Gunasekaran,K. *et al*. (2004) Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. *J. Mol. Biol.*, **341**, 1327–1341.

Huang,Y. *et al*. (2010) CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–2.

Ishida,T. and Kinoshita,K. (2008) Prediction of disordered regions in proteins based on the meta approach. *Bioinformatics*, **24**, 1344–1348.

Jain,E. *et al*. (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics*, **10**, 136.

Jones,D. and Swindells,M. (2002) Getting the most from PSI-BLAST. *Trends Biochem. Sci*., **27**, 161–164.

Jones,S. *et al*. (1998) Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Sci.*, **7**, 233–242.

Jones,S. and Thornton,J. (1996) Principles of protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kawashima,S. *et al*. (2008) AAindex: amino acid index database, progress report. *Nucleic Acids Res.*, **36**, D202–D205.

Kurgan,L. and Disfani,F.M. (2011) Structural protein descriptors in 1-dimension and their sequence-based predictions. *Curr. Protein Pept. Sci.*, **12**, 470–489.

McGuffin,L. (2008) Intrinsic disorder prediction from the analysis of multiple protein fold recognition models. *Bioinformatics*, **24**, 1798–1804.

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

Mészáros,B. *et al*. (2007) Molecular principles of the interactions of disordered proteins. *J. Mol. Biol.*, **372**, 549–561.

Mizianty,M.J. *et al*. (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. *Bioinformatics*, **26**, i489–i496.

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

Nozaki,Y. and Tanford,C. (1971) The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. *J. Biol. Chem.*, **246**, 2211–2217.

Obenauer,J.C. *et al*. (2003) Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*., **31**, 3635–3641.

Oldfield,C.J. *et al*. (2005) Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry*, **44**, 12454–12470.

Oldfield,C.J. *et al*. (2008) Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics*, **9(Suppl 1)**, S1.

Pearson,W. and Lipman,D. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–8.

Peng, Z.L. and Kurgan, L. (2012) Comprehensive comparative assessment of in-silico predictors of disordered regions. *Curr. Protein Pept. Sci.*, **13**, 6–18.

Puntervoll,P. *et al*. (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res*., **31**, 3625–3630.

Rice,P. *et al*. (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet.*, **16**, 276–277.

Rubinstein,N.D. *et al*. (2009) Epitopia: a web-server for predicting B-cell epitopes. *BMC Bioinformatics*, **10**, 287.

Schlessinger,A. *et al*. (2009) Improved disorder prediction by combination of orthogonal approaches. *PLoS One*, **4**, e4433.

Schlessinger,A. *et al*. (2006) PROFbval: predict flexible and rigid residues in proteins. *Bioinformatics*, **22**, 891–893.

Tompa,P. *et al*. (2009) Close encounters of the third kind: disordered domains and the interactions of proteins. *Bioessays*, **31**, 328–335.

Uversky,V.N. and Dunker,A.K. (2010) Understanding protein non-folding. *Biochim Biophys Acta*, **1804**, 1231–1264.

Vacic,V, *et al*. (2007) Characterization of molecular recognition features, MoRFs, and their binding partners. *J. Proteome Res.*, **6**, 2351–2366.

Ward,J. *et al*. (2004) The DISOPRED server for the prediction of protein disorder. *Bioinformatics*, **20**, 2138–2139.

Zhang,T. *et al*. (2012) SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. *J. Biomol. Struct. Dyn*., **29**, 799–813.

Zhou,H. and Zhou,Y. (2004) Quantifying the effect of burial of amino acid residues on protein stability. *Proteins*, **54**, 315–322.