

Published in final edited form as:

Biochemistry. 2007 November 27; 46(47): 13468–13477. doi:10.1021/bi7012273.

## Mining $\alpha$ -helix-forming molecular recognition features ( $\alpha$ -MoRFs) with cross species sequence alignments<sup>†</sup>

Yugong Cheng<sup>‡,§</sup>, Christopher J. Oldfield<sup>‡</sup>, Jingwei Meng<sup>‡</sup>, Pedro Romero<sup>\*,#</sup>, Vladimir N. Uversky<sup>\*,‡,§,¶</sup>, and A. Keith Dunker<sup>\*,‡,§</sup>

<sup>‡</sup> Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, IN 46202

<sup>§</sup> Molecular Kinetics, Inc., 6201 La Pas Trail, Suite 160, Indianapolis, IN 46268

<sup>#</sup> School of Informatics, Indiana University–Purdue University at Indianapolis, 535 West Michigan St., Indianapolis, IN 46202

<sup>¶</sup> Institute for Biological Instrumentation, Russian Academy of Sciences, 142292 Pushchino, Moscow Region, Russia

### Abstract

Previously described algorithms for mining  $\alpha$ -helix-forming molecular recognition elements (MoREs, described in Oldfield *et al.* (2005) *Biochemistry* **44** (6) 1989–2000), known also as molecular recognition features (MoRFs, Mohan *et al.* (2006) *J. Mol. Biol.* **362** (5) 1043–1059), revealed that regions undergoing disorder-to-order transition are involved in many molecular recognition events and are crucial for protein-protein interactions. However, these algorithms were developed using training dataset of a limited size. Here we propose to improve the prediction algorithms by (1) including additional  $\alpha$ -MoRF examples and their cross species homologues in the positive training set; (2) careful extracting monomer structure chains from PDB as the negative training set; (3) including attributes from recently developed disorder predictors, secondary structure predictions, and amino acid indices as attributes; and (4) constructing neural network based predictors and performing validation. Over 50 regions which undergo disorder-to-order transition regions that were identified in PDB together with a set of corresponding cross species homologues of each structure-based example were included in new positive training set. Over 1500 attributes, including disorder predictions, secondary structure predictions and amino acid indices were evaluated by conditional probability method. The top attributes, including VSL2 and VL3 disorder predictions and several physicochemical propensities of amino acid residues, were used to develop the feed forward neural networks. The sensitivity, specificity and accuracy of the resulting predictor,  $\alpha$ -MoRF-PredII, were  $0.87 \pm 0.10$ ,  $0.87 \pm 0.11$ , and  $0.87 \pm 0.08$  over 10-cross validation, respectively. We present the results of these analyses and validation examples to discuss the potential improvement of the  $\alpha$ -MoRF-PredII prediction accuracy.

<sup>†</sup>This work was supported in part by the grants R01 LM007688-01A1 (to A.K.D.) and GM071714-01A2 (A.K.D and V.N.U.) from the National Institutes of Health and the Programs of the Russian Academy of Sciences for the “Molecular and cellular biology” and “Fundamental science for medicine” (to V.N.U.)

CORRESPONDING AUTHOR FOOTNOTE To whom correspondence should be addressed at Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, 410 W. 10th Street, HS 5000, Indianapolis, IN 46202. Phone: 317-278-9650; fax: 317-278-9217; E-mail: vuffersky@iupui.edu (V.N.U.) or kedunker@iupui.edu (A.K.D.), or promero@compbio.iupui.edu (P.R.).

AUTHOR EMAIL ADDRESS: vuffersky@iupui.edu

## Keywords

Intrinsically disordered protein; molecular recognition; cell signaling; protein; protein interaction; MoRF; PONDR

---

Interactions between proteins and their partners are crucial for biological functions. Identification and predicting such interactions would provide insights and guides for laboratory experimental efforts to understand the mechanisms of signaling and regulation within biological systems. Further, based on such knowledge, small molecule therapies could be developed to target human diseases (1,2).

Molecular recognition serves as the initial step for protein-protein interactions. The mechanisms of signaling and regulatory molecular recognition include high specificity with low affinity, and binding diversity in terms of various structural accommodations at binding surface. Coupled binding and folding has been found in several well-characterized protein-protein interactions during molecular recognition; one of the partners in each case undergoes a disorder-to-order transition upon binding to its structured complement (3–7). A large decrease in conformation entropy accompanies disorder-to-order transition, which uncouples specificity from binding strength. This phenomenon has the effect of making highly specific interactions easily reversible, which is beneficial for cells, especially in the inducible responses typically involved in signaling and regulation. Recent computational studies of such binding illustrated that the disordered partner contains a “conformational preference” for the structure it will take upon binding, and that these so-called “preformed elements” tend to be helices (8–11). These studies validated previous findings for individual protein-protein interactions, such as p27<sup>Kip1</sup> (12,13) and p53 (14), both of which have disordered regions with significant helical character that form  $\alpha$ -helices upon binding to their partners.

These foldable partners of protein-protein interactions are the members of the recently discovered class of intrinsically disordered (ID) proteins, which lack rigid 3-D structure under physiological conditions *in vitro*. Bioinformatics studies indicated that about 25 to 30% of eukaryotic proteins are mostly disordered (15), that more than half of eukaryotic proteins have long regions of disorder (15–17), and that more than 70% of signaling proteins have long disordered regions (18). Despite the fact that intrinsically disordered proteins fail to form fixed 3D-structures under physiological conditions, they carry out numerous crucial biological functions (3–7,18–35).

It has been emphasized that signaling and regulation are among the most important functions of intrinsically disordered proteins (18,21,25). Qualitatively, it seems reasonable that highly mobile proteins would provide a better basis for signaling and recognition. For example, disordered regions can bind partners with both high specificity and low affinity (36). This means that the regulatory interactions can be specific and also can be easily dispersed. Obviously this represents a keystone of signaling –turning a signal off is as important as turning it on (23). Another crucial property of ID proteins for their function in signaling networks is binding diversity; i.e., their ability to partner with many other proteins and other ligands, such as nucleic acids (37). This opens a unique possibility for one regulatory region or one regulatory protein to bind to many different partners. In agreement with this hypothesis it has been shown that proteins making multiple interactions are more likely to lead to lethality if deleted (38). An interesting consequence of the capability of ID regions to interact with different binding partners is their polymorphism in bound state; i.e., an ID protein (or ID region) might have completely different geometries in the rigidified structures induced by the binding to its partner, depending on the nature of the bound partner (39).

Recently, a concept of molecular recognition fragment (MoRF, because such regions “morph” from disorder to order upon binding) was introduced for specific, short (around 20 residues) structural element that mediates certain classes of binding events of disordered regions (8–10). This short fragment is found within a region of disorder and undergoes a disorder-to-order transition that is stabilized by binding to its partner. An algorithm to identify protein regions having  $\alpha$ -helix-forming MoRF signatures was developed based on the patterns of Predictors Of Naturally Disordered Regions (PONDRs<sup>®</sup>), secondary structure predictions, and hydrophobic cluster analysis (9). The application of this algorithm to databases of genomics functionally annotated proteins indicates that such features are highly abundant and are likely to play important roles in protein-protein interactions involved in signaling events (9). Others have used this order/disorder plots to predict binding sites that were subsequently verified by laboratory experiments (40,41). For some of these predicted examples, the regions did indeed form helix upon binding to their partners (42,43). Alternatively, a sequence-based approach was developed to identify short, conserved recognition sites, called eukaryotic linear motifs (ELMs) (44–46). While MoRFs are identified by general order/disorder tendencies and while ELMs are identified by motif discovery from sequence analysis, the resulting binding sites identified by both methods share several features (47).

The current MoRF algorithm was trained on a small number of  $\alpha$ -MoRF examples (14 regions from 12 proteins). All of the training examples are correctly identified by the current algorithm, suggesting over-fitting. Here we represent a novel algorithm which is improved by (1) including updated  $\alpha$ -MoRF examples and their cross species homologues in the positive training set; (2) extracting monomer structure chains from PDB as the negative training set; (3) mining attributes from newly developed disorder predictions, secondary structure predictions, and amino acid index (48); and (4) constructing neural network based predictors and performing validation.

## Materials and Methods

### Data sets

$\alpha$ -MoRF examples were retrieved by the following procedures from structures in Protein Data Bank (PDB) as of Aug 31, 2005: (1) Chains of 30 residues or less were selected from structures with valid reference identification (i.e., SwissProt, PIR or GB ID) that were bound to another protein chain longer than 85 residues. (2) Only chains with helical content were kept. (3) Only chains bound to a different molecule were kept.

Cross species homologues of these examples were retrieved from SwissProt. The parent sequences of MoRF containing proteins from PDB were aligned with their homologues using ClustalW (49) with a disorder based similarity matrix (50). One homologue that aligned well with the known MoRF region was randomly selected and also included in the positive training set.

For a control set, monomeric chains were retrieved from Macromolecular Structure Database (<http://pqs.ebi.ac.uk/>) by selecting “monomeric” in Protein Quaternary Structure query. These chains were further filtered for sequence redundancy at 30% identity.

### MoRF predictor architecture

The MoRF predictor was designed with a stacked architecture, similar to the one used previously (9), where multiple prediction algorithms are applied in serial. First, a heuristic is used to detect potential MoRF regions from predictions of intrinsic disorder. The heuristic used here is similar to the one used previously (9), which identifies short regions of order within longer regions of disorder – or “dips” – in disorder prediction profiles. Second, a discrimination

algorithm is applied to these potential MoRF regions to distinguish between actual MoRFs (true MoRFs) and other sources of “dips” (false MoRFs). In this work, a neural network was developed for this second stage prediction. Inputs to the neural network consisted of sequence features calculated for sequence regions relative to the potential MoRF region: binding, which is the potential MoRF region; flanking, which is two regions of residues abutting the potential MoRF region in terms of sequence; and whole, which is the union of binding and flanking regions.

### Sequence features

Besides the features used previously (9), features included in this study were the average scores of newly developed disorder predictions (VL3 (51), VSL2 (52), DisEMBL REMARK-465 (53)), secondary structure propensities by GOR-IV (54), and the greater than 400 scales in the amino acid index database (48). Each feature was calculated for each of the binding, flanking and whole regions of each training example.

### Feature selection

Feature selection was performed in two stages. First, the total set of features was reduced to the 30 features best correlated with MoRFs using the area ratio (AR) test. Second, the set of features for neural network training was selected by forward selection, branch and bound, or the best features according to the AR test. The AR test has been described previously (55). Briefly, the conditional probability of an observation  $Y$  given the prior knowledge that a variable has the value  $x$ , is given by  $P(Y|x)$ . In this work, the observation  $Y$  would correspond to whether a region belongs to the true MoRF set or the false MoRF set given prior knowledge that the sequence characteristic has a value of  $x$ . When plotting the conditional probability versus the value of the attribute, the greater the separation of the two curves, the better a given attribute distinguishes between the positive and control samples. This separation can be quantified by dividing the area bounded by the two curves by the total area to give the area ratio.

The top 30 attributes from the AR test were subjected to further selection using TOOLDIAG (56), using both sequential forward selection and branch and bound, with the Mahalanobis distance as the criterion distance metric for both strategies. The Mahalanobis distance is a statistical distance, which is defined in terms of the distance between two sample means in units of standard deviation, based on the assumption of equal variance of the two samples. In sequential forward selection, attributes are added in rounds, where all attributes are evaluated in conjunction with all attributes selected in previous rounds and the best attribute is retained for the next round of selection. In contrast, branch and bound selection begins with all features and divides them into subsets. Since the Mahalanobis distance of a subset of features is necessarily less than or equal to its superset, many feature subsets never need to be evaluated if their superset is worse than the current bound. This significantly reduces computation time and is guaranteed to find the optimal set of features – in terms of the Mahalanobis distance – for a given number of features.

### Construction and training of neural networks

Feed-forward neural networks were constructed with one hidden layer and trained using a supervised learning algorithm in Matlab's Neural Networks toolbox. A 10-cross validation scheme was performed, where the dataset is divided into 10 subsets and training is repeated 10 times using each set for validation in turn and the remain sets for training. For each cross-validation cycle, 10 experiments were performed using different initializations of the neural network. The reported results are the average of the testing results over these 100 trained neural networks.

## Evaluation of neural networks

The results of the neural network training were evaluated by multiple methods. Several of these –including positive predict value (PPV), negative predict value (NPV), sensitivity ( $S_n$ ), specificity ( $S_p$ ), and accuracy (Acc) – are defined in Table 1. These measures depend on the particular choice of decision threshold applied to the neural network output. To obtain a more general measure of predictor performance that is independent of the selected threshold, receiver operating curves (ROC) were used.

A ROC curve is a two-dimensional measure of classification performance. The ROC curve is defined as a plot of the true positive rate as a function of the false positive rate. An empirical ROC curve can be generated by calculating TP and FP for all relevant thresholds. We approximated the ROC curve by simply connecting the data points ( $S_n$ ,  $1-S_p$ ) with straight lines. The full area under the ROC curve (AUC) is the most commonly used ROC index (57). Conceptually, it has several interpretations: (1) the probability that the test will produce a value for a randomly chosen true MoRF that is greater than the value for a randomly chosen false MoRF, (2) the average sensitivity for all values of specificity, and (3) the average specificity for all values of sensitivity. A perfect predictor has an AUC of 1.0, whereas random class assignment gives an AUC of 0.5.

## Results and Discussion

### Selection of the disorder predictor for MoRF prediction

The first stage of the stacked predictor architecture is the identification of potential MoRF regions from disorder prediction profiles. In previous work (9), PONDR VL-XT was selected for this purpose based on previous observations (58). Here, we re-examine this choice by examining if indication of binding regions is a feature specific to PONDR® VLXT profiles or if other predictors of intrinsic disorder can be used for the MoRF prediction purposes. To answer this question we compared disorder plots produced by several predictors for proteins with archetypal MoRFs: 4E-BP1, p53, and RNase E. Specifically, we examined whether or not each predictor produced “dips” – or short regions of predicted order within longer regions of predicted disorder – corresponding to known binding regions. This behavior is required for successful MoRF prediction.

4E-BP1 is a human phosphoprotein of 118 residues with a critical role in controlling protein synthesis, and hence, in cell survival and proliferation through the phosphorylation of eukaryotic initiation factor 4E (eIF4E). Phosphorylation of 4E-BP1 results in the release of eIF4E and activation of cell protein synthesis (59). Deletion and site-directed mutagenesis identified 4E-BP1 central region (residues 49-66) as a motif essential for eIF4E binding (60). NMR and CD experiments indicated that 4E-BP1 is completely unstructured in the absence of eIF4E (61). However, upon complex formation, 4E-BP1  $^{15}\text{N}$  HSQC spectrum showed a small number of weak new peaks dispersed upfield from the majority of other signals. The analysis of a 20 residue peptide fragment of 4E-BP1 (residues 49-68) containing the eIF4E binding motif revealed that this peptide was able to bind to eIF4E, producing similar chemical shift changes to the full-length 4E-BP1, and inhibited translation in reticulocyte lysate. Together, these results suggested that a short central region of the 4E-BPs is responsible for eIF4E binding and translation inhibition while the flanking regions are unfolded and flexible (61,62). Figure 1A shows that the whole 4E-BP1 is predicted to be disordered by PONDR® VLXT, whereas there is a downward spike in the central region of the disordered prediction plot, which overlaps with the experimentally verified binding region for eIF4E. This feature suggested that the peculiarities of PONDR® VLXT plots can be used to visualize regions in disordered proteins important for protein-protein interactions (58). Additional work has further validated the use of these distinctive downward spikes in PONDR® VLXT curves to locate functional binding

regions. Later this pattern was used as the fundamental brick for the development of algorithm for identifying  $\alpha$ -MoRFs (9).

The results of disorder prediction for 4E-BP1 are shown in Figure 1, where disorder profiles produced by different predictors are grouped by their overall appearance and shape. The vast majority of the analyzed algorithms (with the except for DisPro (63), DRIPPRED (<http://www.forcasp.org/paper2127.html>) and DISOPRED (17), see Figure 1D) correctly predicted 4E-BP1 as mostly disordered protein. Furthermore, many predictors produced “dips” in the central region of 4e-BP1. IUPred (64) and RONN (65) gave shallow “dips” which matched in their positions with “dip” predicted by PONDR<sup>®</sup> VLXT (Figure 1A). Members of the VL3 (including VL3, VL3H, and VL3E) (51) and VL2 families (including VL2 and VL2-S) (66) showed shifted shallow “dips” covering a broad region (~60 residues) centered around residue Gly40. The behavior of these predictors is illustrated by plots for VL3 and VL2 (Figure 1B). The “dips” produced by the predictors of VSL2 family [VSL2B and VSL2P, (67)] and by PONDR<sup>®</sup> VL3BA (51) were very shallow, their plots were located above the threshold of 0.5 suggesting that according to these predictors 4e-BP1 does not have ordered residues at all (Figure 1C).

Next we analyzed the tumor suppressor protein p53, which is at the center of a large signaling network, regulating expression of genes involved in many cellular processes such as cell cycle progression, apoptosis induction, DNA repair, and response to cellular stress (68). When p53 function is lost, either directly through mutation or indirectly through several other mechanisms, the cell often undergoes oncogenesis (69). Tumors showing mutations in p53 are found in colon, lung, esophagus, breast, liver, brain, reticuloendothelial tissues and hemopoietic tissues (69). It has been shown that p53 induces or inhibits over 150 genes, including *p21*, *GADD45*, *MDM2*, *IGFBP3*, and *BAX* (70). There are three structural domains in p53: N-terminal translational activation domain, central DNA binding domain, and C-terminal tetramerization and regulatory domain. At the transactivation region, it interacts with TFIID, TFIIF, Mdm2, RPA, CBP/p300 and CSN5/Jab1 (68). At the C-terminal domain, it interacts with GSK3 $\beta$ , PARP-1, TAF1, TRRAP, hGcn5, TAF, 14-3-3, and S100B( $\beta\beta$ ).

Therefore, both N- and C-terminal domains of p53 are involved in numerous protein-protein interactions, some of which involve disorder-to-order transitions. For example, Mdm2 was shown to interact with a short stretch of p53, residues 13-29. As this region of p53 is within the transactivation domain, p53 cannot activate or inhibit other genes when Mdm2 is bound. Although X-ray crystallographic studies of the p53-Mdm2 bimolecular complex reveal that the Mdm2 binding region of p53 forms a helical structure that binds into a deep groove on the surface of Mdm2 (71), NMR studies of p53 show that the unbound N-terminal region lacks fixed structure, although it does possess an amphipathic helix that forms secondary structure part of the time (14). It has been shown that interaction of S100B( $\beta\beta$ ) with p53 inhibits its PKC-dependent phosphorylation and tetramer formation (72). Interaction occurs in a Ca<sup>2+</sup>-dependent manner and involves a peptide located in the C-terminal regulatory domain of p53 (residues 367-388) (73). In the absence of S100B( $\beta\beta$ ), the p53 peptide (S367-E388) exists as a random coil as determined by NMR. However, much of this C-terminal peptide (residues S376-T387) adopts a helical conformation when bound to Ca<sup>2+</sup> loaded S100B( $\beta\beta$ ) (74).

Thus, fragments from the N-terminal transactivation and the C-terminal tetramerization/regulatory domains undergo disorder-to-order transition upon binding to their partners. The PONDR<sup>®</sup> VLXT plot shown in Figure 2A illustrates such a predisposition for the disorder-to-order transition as sharp dips within the disordered regions. Next, we analyzed p53 by several other disorder predictors. Among more than 15 predictors analyzed only DRIPPRED (<http://www.forcasp.org/paper2127.html>) possessed both “dips” at N- and C-terminals as PONDR<sup>®</sup> VLXT (see Figure 2A). Predictions by IUPred (64) and RONN (65) showed a

matched “dip” at N-terminal and a shallower “dip” at C-terminal, compared to that of VLXT (Figure 2A). All VL3 (including BA, E, and H) (51), VL2 (including VL2, VL2C, S, and V) (66), and VSL2 (VSL2B and P) predictors (67) showed a shifted shallow “dip” in disordered N-terminus, and predicted C-terminus to be totally disordered, without “dips”. Their prediction patterns are represented in Figure 2B. Finally, Figure 2C shows that DisPro (63) and DISOPRED (17) were able to predict a “dip” in the disordered C-terminus of p53.

Similarly, we applied various disorder predictors to the *E. coli* ribonuclease RNase E. The endoribonucleases operate under tight cellular regulation and are involved in the modification, maturation and degradation of different RNAs (75). RNase E is an important member of this family, and is responsible for controlling the levels of many different transcripts that encode enzymes of fundamental metabolic pathways, including glycolysis (76). The protein can be divided into two roughly equal fragments that are involved in different functions: the N-terminal domain (NTD; residues 1–498) hosts catalytic function and the C-terminal domain (CTD; residues 499–1061) preserves the biologically significant ability to interact with other degradosome components and with structured RNA (40).

CTD was shown to be highly disordered by experiments and PONDR<sup>®</sup> VLXT prediction (40). However, VLXT prediction showed four sharp downward spikes within the entirely disordered CTD. These “dips” were referred as “regions of increased structural propensity” RISPs (40), and labeled in Figure 3A as A, B, C, and D. RISP A appears to be a protein–RNA interaction site whereas the other segments possibly correspond to sites of self-recognition (segment B, segment of potential coiled coil) and to sites of interaction with the other degradosome proteins (segments C and D interact with enolase and PNPase, respectively) (40). The crystal structure of the complex between the enolase and fragment C has been determined, which showed that region C forms  $\alpha$ -helix in the complex (Luisi, personal communication). Therefore, all C-terminal regions in RNase E with the predisposition for the disorder-to-order transition were correctly visualized by PONDR<sup>®</sup> VLXT (40).

Analysis of RNase E by other disorder predictors is described below. All VL2 and VL3 predictions together with RONN and IUPred plots have shallow “dips” at “A”, “C”, and “D” positions (Figure 3B), while VSL2 and VL3BA predictions almost do not have dip at position “A”. However, they have pronounced dips “C” and “D” and VL3BA has a very broad dip centered at dip “C” (Figure 3C). Figure 3D shows that DRIPPRED, DISOPRED and DisPro predict site “A” as a sharp “dip” and do not have specific disorder-based features at “A”, “C”, and “D” positions. In fact, DRIPPRED predicts that last 450 residues of RNaseE are completely disordered, whereas according to DISOPRED, fragment 830-1061 is mostly ordered. Figure 3 illustrates that none of the predictors analyzed show a sharp dip at the “B” position as PONDR<sup>®</sup> VLXT does.

Overall, data presented above show that many predictors gave similar general disorder/order predictions on the three examples. However, PONDR<sup>®</sup> VLXT was more sensitive for features associated with regions potentially undergoing disorder-to-order transition than other predictors and therefore it was selected for the identification of potential MoRFs for the first stage of the prediction algorithm. The formerly defined basic MoRF pattern (9) was used with little modification.

## Data sets

54 basic MoRF regions (from 51 proteins) were retrieved from PDB structures. Of these MoRF regions, 48 have at least one cross species homologue sequences that could be retrieved from SwissProt and the remaining six examples do not have any obvious cross species homologue. This gave a positive training set containing 102 regions from 99 proteins. The sequence identities for the paired proteins were from 5% to 100%.

For the control set, structured monomers were used. These sequences of these proteins are known to be ordered in isolation and therefore cannot contain MoRF regions. PONDR VL-XT predictions were made for the structured, monomeric proteins and their prediction profiles scanned for the basic MoRF pattern. The basic MoRF pattern was found 236 times in 120 of the structured monomers.

For both MoRF-associated PONDR patterns and control PONDR patterns, all features were generated for the MoRF pattern region, flanking regions, and the whole region.

### Feature selection

Over 1,500 attributes of binding, flanking and whole region were evaluated by the AR test. The best 30 AR values yielded from attributes of VSL2, VL3 and others over the whole, binding and flanking regions, in a range of 0.64 to 0.54. As shown in Figure 4, the Mahalanobis distances were compared for top 10 features selected from top 30 AR attributes by Sequential Forward Selection, or Branch and Bound to direct top 10 AR attributes. Branch and Bound, and Forward Selection yielded higher Mahalanobis distances than direct AR for three to 10 features selected, indicating those feature combinations may yield better separations between the positive and negative training sets. Furthermore, the Mahalanobis distance reached plateau when six features were selected. Thus, top 6 and top 10 feature combinations from all three methods were used to train neural networks.

### Neural networks training and evaluation

Feed-forward neural networks were constructed with one hidden layer with 10 neurons using the combined features selected by all three methods. The results from top 6 features from Forward Selection were used as an illustration in Figure 5A: sensitivity, specificity, and accuracy after 10-cross validation were  $0.87 \pm 0.10$ ,  $0.87 \pm 0.11$ , and  $0.87 \pm 0.08$ , respectively. When increased the threshold for prediction from  $-0.9$  to  $0.9$ , PPV and specificity increased from  $0.63 (\pm 0.10)$  to  $0.95 (\pm 0.09)$  and from  $0.45 (\pm 0.19)$  to  $0.98 (\pm 0.07)$ , respectively; while NPV and sensitivity decreased from  $0.92 (\pm 0.10)$ , to  $0.67 (\pm 0.08)$ , and from  $0.96 (\pm 0.09)$  to  $0.53 (\pm 0.15)$ , respectively (Figure 5B). However, accuracy reached maximum (0.87) when threshold was set to 0 or  $-0.1$ .

ROC curves can be used to quantify predictor performance, because they do not require determination of an optimal threshold and provide information concerning predictor performance over a range of thresholds. The curves of better predictors lie above and to the left of the curves produced by worse predictors. The area under the ROC curve is commonly used for quantification of predictor performance. This area is defined between 1 and 0, where a value of 0.5 (along the diagonal in Figure 6) would be expected for random classification. ROC analysis (Figure 6) of all six neural networks constructions showed that all of them were better than random, whereas curves for top 10 features from Forward Selection, top 10 features from AR, and top 6 features from Forward Selection were on top of others. Thus, neural network constructed based on top six features from Forward Selection was chosen over the top 10 features.

### Examination of specific $\alpha$ -MoRF-PredIII predictions

The performance of the  $\alpha$ -MoRF-PredIII identifier was analyzed using a set of proteins with known MoRFs.

**E. coli ribonuclease RNase E**—Current algorithm predicted all four RISPs as MoRFs (see Figure 2), whereas the original predictor failed to identify any these dips as molecular recognition features.

**p53 from different species**—Both N- and C-terminal regions of human p53 were used as positive training examples in the development of previous and present algorithms. A group of 33 sequences of p53 were collected from SwissProt.  $\alpha$ -MoRF-PredII identified 27 N-terminal regions and 11 C-terminal regions as MoRFs from cross species alignments, compared to 4 N-terminal regions and 9 C-terminal regions identified by the original predictor.

**PDB\_Select25**—A set of structured chains of over 30 residues was retrieved from PDB\_Select25 (<http://bioinfo.tg.fh-giessen.de/pdbselect/>) as a validation (negative) set. Only 18% of over 1,500 basic MoRFs (“dips”) patterns were predicted as MoRFs, i.e., 82% accuracy is achieved for this set.

### MoRF predictions across genomes and functional groups

$\alpha$ -MoRF-PredII algorithm was applied to sequences from 82 genomes in the three kingdoms of life to estimate the prevalence of regions having  $\alpha$ -MoRF propensities. The results from 1000-resampling were compared to that from the previous method as shown in Figure 7 (9). The average eukaryotic genome has a greater than 3- and 4-fold higher fraction of proteins with  $\alpha$ -MoRF propensities than the average bacterial and archaeal genome, respectively.  $\alpha$ -MoRFs are indicated to occur with 4- and 6-fold higher frequency in the average eukaryotic genome than in the average bacterial and archaeal genome, respectively. Furthermore, all eukaryotic genomes have higher fractions of proteins with  $\alpha$ -MoRF propensities and higher frequencies of  $\alpha$ -MoRF indications than all bacterial and archaeal genomes.

$\alpha$ -MoRF-PredII algorithm was also applied to functional classes of human proteins retrieved from SwissProt as described previously (18). As shown in Table 2, human proteins involved in regulation, cell division, cytoskeleton, as well as ribosomal proteins contain more  $\alpha$ -MoRF than proteins in the average eukaryotic genome. Membrane, transport and inhibitory proteins have similar  $\alpha$ -MoRF propensities as proteins from the average eukaryotic genome. Finally, proteins associated with biosynthesis, protease, G protein coupled receptor, metabolism, degradation, and kinase activities have lower  $\alpha$ -MoRF propensities than the average eukaryotic protein.

Summarizing, we elaborated a novel neural network-based algorithm for mining  $\alpha$ -helix-forming molecular recognition features,  $\alpha$ -MoRFs, which are intrinsically disordered regions undergoing disorder-to-order transition as a result of interaction with their binding partners. In comparison with the original  $\alpha$ -MoRE identifier (9), this algorithm was improved by using the extended set of newly identified  $\alpha$ -MoRF and their homologues, by extracting monomer chains from PDB as the negative training set and via mining novel attributes related to disorder and secondary structure predictions as well as amino acid index. The top attributes were used to develop the feed forward neural networks. The performance of the resulting tool,  $\alpha$ -MoRF-PredII predictor, was validated based on a set of proteins with known  $\alpha$ -MoRFs as a positive control and a set of ordered proteins that do not contain  $\alpha$ -MoRFs as a negative control. The sensitivity, specificity, and accuracy of  $\alpha$ -MoRF-PredII predictor were close to 0.9. The usefulness of this new predictor is illustrated via its application for analysis of 82 genomes in the three kingdoms of life and for analysis functional classes of human proteins.

### ABBREVIATIONS

<b>MoRF</b>	molecular recognition features
<b>MoRE</b>	molecular recognition element

<b>PONDR®</b>	Predictor of naturally disordered regions
<b>ROC</b>	receiver operating curves
<b>AR</b>	area ratio
<b>PPV</b>	positive predict value
<b>NPV</b>	negative predict value
<b>S<sub>n</sub></b>	sensitivity
<b>S<sub>p</sub></b>	specificity
<b>Acc</b>	accuracy
<b>AUC</b>	area under the ROC curve
<b>ELM</b>	eukaryotic linear motifs

## References

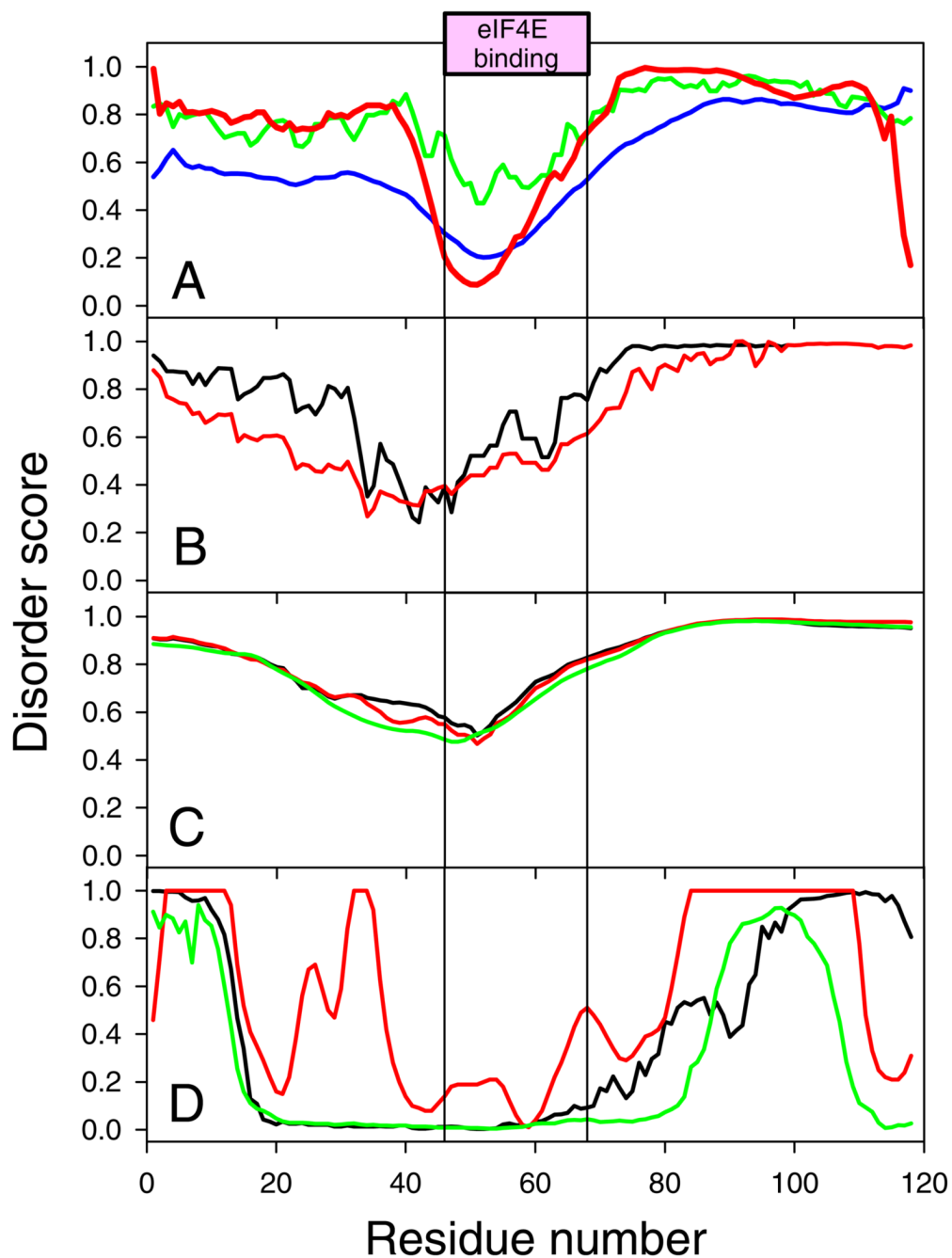
1. Fry DC, Vassilev LT. Targeting protein-protein interactions for cancer therapy. *J Mol Med* 2005;83:955–963. [PubMed: 16283145]
2. Arkin M. Protein-protein interactions and cancer: small molecules going in for the kill. *Curr Opin Chem Biol* 2005;9:317–324. [PubMed: 15939335]
3. Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol* 2002;12:54–60. [PubMed: 11839490]
4. Uversky VN, Gillespie JR, Fink AL. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins* 2000;41:415–427. [PubMed: 11025552]
5. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. *Biochemistry* 2002;41:6573–6582. [PubMed: 12022860]
6. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 2005;6:197–208. [PubMed: 15738986]
7. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 1999;293:321–331. [PubMed: 10550212]
8. Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, Dunker AK, Uversky VN. Analysis of Molecular Recognition Features (MoRFs). *J Mol Biol* 2006;362:1043–1059. [PubMed: 16935303]
9. 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;44:12454–12470. [PubMed: 16156658]
10. Vacic V, Oldfield CJ, Mohan A, Radivojac P, Cortese MS, Uversky VN, Dunker AK. Characterization of molecular recognition features, MoRFs, and their binding partners. *J Proteome Res* 2007;6:2351–2366. [PubMed: 17488107]

11. Fuxreiter M, Simon I, Friedrich P, Tompa P. Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. *J Mol Biol* 2004;338:1015–1026. [PubMed: 15111064]
12. Bienkiewicz EA, Adkins JN, Lumb KJ. Functional consequences of preorganized helical structure in the intrinsically disordered cell-cycle inhibitor p27(Kip1). *Biochemistry* 2002;41:752–759. [PubMed: 11790096]
13. Lacy ER, Filippov I, Lewis WS, Otieno S, Xiao L, Weiss S, Hengst L, Kriwacki RW. p27 binds cyclin-CDK complexes through a sequential mechanism involving binding-induced protein folding. *Nat Struct Mol Biol* 2004;11:358–364. [PubMed: 15024385]
14. Lee H, Mok KH, Muhandiram R, Park KH, Suk JE, Kim DH, Chang J, Sung YC, Choi KY, Han KH. Local structural elements in the mostly unstructured transcriptional activation domain of human p53. *J Biol Chem* 2000;275:29426–29432. [PubMed: 10884388]
15. Oldfield CJ, Cheng Y, Cortese MS, Brown CJ, Uversky VN, Dunker AK. Comparing and combining predictors of mostly disordered proteins. *Biochemistry* 2005;44:1989–2000. [PubMed: 15697224]
16. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform* 2000;11:161–171.
17. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 2004;337:635–645. [PubMed: 15019783]
18. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol* 2002;323:573–584. [PubMed: 12381310]
19. Daughdrill, GW.; Pielak, GJ.; Uversky, VN.; Cortese, MS.; Dunker, AK. Natively Disordered Proteins. In: Buchner, J.; Kiefhaber, T., editors. *Protein Folding Handbook*. Wiley-VCH; 2005. p. 275-357.
20. Dunker AK, Brown CJ, Obradovic Z. Identification and functions of usefully disordered proteins. *Adv Protein Chem* 2002;62:25–49. [PubMed: 12418100]
21. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. Flexible nets. The roles of intrinsic disorder in protein interaction networks. *Febs J* 2005;272:5129–5148. [PubMed: 16218947]
22. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Higgs KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z. Intrinsically disordered protein. *J Mol Graph Model* 2001;19:26–59. [PubMed: 11381529]
23. Dunker AK, Obradovic Z. The protein trinity--linking function and disorder. *Nat Biotechnol* 2001;19:805–806. [PubMed: 11533628]
24. Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN, Dunker AK. Intrinsic disorder in transcription factors. *Biochemistry* 2006;45:6873–6888. [PubMed: 16734424]
25. Uversky VN, Oldfield CJ, Dunker AK. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J Mol Recognit* 2005;18:343–384. [PubMed: 16094605]
26. Uversky VN, Roman A, Oldfield CJ, Dunker AK. Protein intrinsic disorder and human papillomaviruses: increased amount of disorder in E6 and E7 oncoproteins from high risk HPVs. *J Proteome Res* 2006;5:1829–1842. [PubMed: 16889404]
27. Uversky VN. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci* 2002;11:739–756. [PubMed: 11910019]
28. Uversky VN. What does it mean to be natively unfolded? *Eur J Biochem* 2002;269:2–12. [PubMed: 11784292]
29. Uversky VN. Protein folding revisited. A polypeptide chain at the folding-misfolding-nonfolding cross-roads: which way to go? *Cell Mol Life Sci* 2003;60:1852–1871. [PubMed: 14523548]
30. Tompa P. Intrinsically unstructured proteins. *Trends Biochem Sci* 2002;27:527–533. [PubMed: 12368089]
31. Fink AL. Natively unfolded proteins. *Curr Opin Struct Biol* 2005;15:35–41. [PubMed: 15718131]
32. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN. Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins. *J Proteome Res* 2007;6:1917–1932. [PubMed: 17391016]

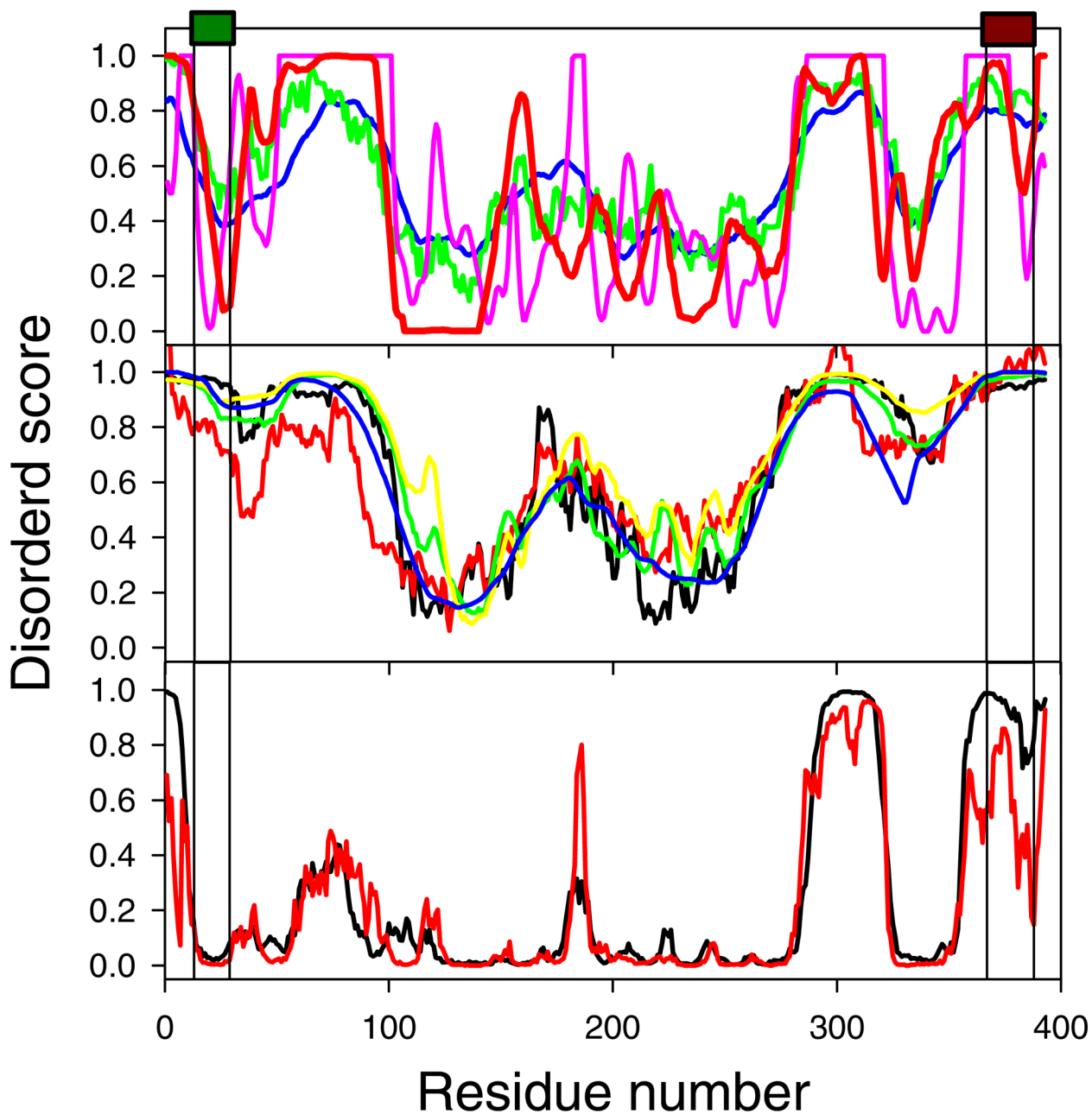
33. Vucetic S, Xie H, Iakoucheva LM, Oldfield CJ, Dunker AK, Obradovic Z, Uversky VN. Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions. *J Proteome Res* 2007;6:1899–1916. [PubMed: 17391015]
34. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z. Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. *J Proteome Res* 2007;6:1882–1898. [PubMed: 17391014]
35. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker AK. Intrinsic disorder and functional proteomics. *Biophys J* 2007;92:1439–1456. [PubMed: 17158572]
36. Schulz, GE. Nucleotide binding proteins. In: Balaban, M., editor. *Molecular mechanism of biological recognition*. Elsevier/North-Holland Biomedical Press; New York: 1979. p. 79-94.
37. Kriwacki RW, Hengst L, Tennant L, Reed SI, Wright PE. Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. *Proc Natl Acad Sci U S A* 1996;93:11504–11509. [PubMed: 8876165]
38. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. *Nature* 2001;411:41–42. [PubMed: 11333967]
39. Dajani R, Fraser E, Roe SM, Yeo M, Good VM, Thompson V, Dale TC, Pearl LH. Structural basis for recruitment of glycogen synthase kinase 3beta to the axin-APC scaffold complex. *Embo J* 2003;22:494–501. [PubMed: 12554650]
40. Callaghan AJ, Aurikko JP, Ilag LL, Gunter Grossmann J, Chandran V, Kuhnel K, Poljak L, Carpousis AJ, Robinson CV, Symmons MF, Luisi BF. Studies of the RNA degradosome-organizing domain of the *Escherichia coli* ribonuclease RNase E. *J Mol Biol* 2004;340:965–979. [PubMed: 15236960]
41. Bourhis JM, Johansson K, Receveur-Brechot V, Oldfield CJ, Dunker AK, Canard B, Longhi S. The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner. *Virus Res* 2004;99:157–167. [PubMed: 14749181]
42. Kingston RL, Hamel DJ, Gay LS, Dahlquist FW, Matthews BW. Structural basis for the attachment of a paramyxoviral polymerase to its template. *Proc Natl Acad Sci U S A* 2004;101:8301–8306. [PubMed: 15159535]
43. Chandran V, Luisi BF. Recognition of enolase in the *Escherichia coli* RNA degradosome. *J Mol Biol* 2006;358:8–15. [PubMed: 16516921]
44. Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A, Ferre F, Maselli V, Via A, Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Kuster B, Helmer-Citterich M, Hunter WN, Aasland R, Gibson TJ. ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res* 2003;31:3625–3630. [PubMed: 12824381]
45. Neduva V, Linding R, Su-Angrand I, Stark A, de Masi F, Gibson TJ, Lewis J, Serrano L, Russell RB. Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol* 2005;3:e405. [PubMed: 16279839]
46. Neduva V, Russell RB. Linear motifs: evolutionary interaction switches. *FEBS Lett* 2005;579:3342–3345. [PubMed: 15943979]
47. Fuxreiter M, Tompa P, Simon I. Local structural disorder imparts plasticity on linear motifs. *Bioinformatics* 2007;23:950–956. [PubMed: 17387114]
48. Kawashima S, Kanehisa M. AAindex: amino acid index database. *Nucleic Acids Res* 2000;28:374. [PubMed: 10592278]
49. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;22:4673–4680. [PubMed: 7984417]
50. Radivojac P, Obradovic Z, Brown CJ, Dunker AK. Improving sequence alignments for intrinsically disordered proteins. *Pac Symp Biocomput* 2002:589–600. [PubMed: 11928510]
51. Obradovic Z, Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK. Predicting intrinsic disorder from amino acid sequence. *Proteins* 2003;53(Suppl 6):566–572. [PubMed: 14579347]

52. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics* 2006;7:208. [PubMed: 16618368]
53. Linding R, Russell RB, Neduva V, Gibson TJ. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic Acids Res* 2003;31:3701–3708. [PubMed: 12824398]
54. Garnier J, Gibrat JF, Robson B. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol* 1996;266:540–553. [PubMed: 8743705]
55. Arnold GE, Dunker AK, Johns SJ, Douthart RJ. Use of conditional probabilities for determining relationships between amino acid sequence and protein secondary structure. *Proteins* 1992;12:382–399. [PubMed: 1579571]
56. Rauber, TW.; Barata, MM.; Steiger-Garcia, AS. The International Conference on Fault Diagnosis; Toulouse, France. 1993.
57. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. *J Biomed Inform* 2005;38:404–415. [PubMed: 16198999]
58. Garner E, Romero P, Dunker AK, Brown C, Obradovic Z. Predicting Binding Regions within Disordered Proteins. *Genome Inform Ser Workshop Genome Inform* 1999;10:41–50.
59. Heesom KJ, Gampel A, Mellor H, Denton RM. Cell cycle-dependent phosphorylation of the translational repressor eIF-4E binding protein-1 (4E-BP1). *Curr Biol* 2001;11:1374–1379. [PubMed: 11553333]
60. Mader S, Lee H, Pause A, Sonenberg N. The translation initiation factor eIF-4E binds to a common motif shared by the translation factor eIF-4 gamma and the translational repressors 4E-binding proteins. *Mol Cell Biol* 1995;15:4990–4997. [PubMed: 7651417]
61. Fletcher CM, Wagner G. The interaction of eIF4E with 4E-BP1 is an induced fit to a completely disordered protein. *Protein Sci* 1998;7:1639–1642. [PubMed: 9684899]
62. Fletcher CM, McGuire AM, Gingras AC, Li H, Matsuo H, Sonenberg N, Wagner G. 4E binding proteins inhibit the translation factor eIF4E without folded structure. *Biochemistry* 1998;37:9–15. [PubMed: 9453748]
63. Cheng J, Sweredoski M, Baldi P. Accurate prediction of protein disordered regions by mining protein structure data. *Data Mining and Knowledge Discovery* 2005;11
64. Dosztanyi 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;21:3433–3434. [PubMed: 15955779]
65. Yang ZR, Thomson R, McNeil P, Esnouf RM. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics* 2005;21:3369–3376. [PubMed: 15947016]
66. Vucetic S, Brown CJ, Dunker AK, Obradovic Z. Flavors of protein disorder. *Proteins* 2003;52:573–584. [PubMed: 12910457]
67. Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins* 2005;61(Suppl 7):176–182. [PubMed: 16187360]
68. Anderson, CW.; Appella, E. Signaling to the p53 tumor suppressor through pathways activated by genotoxic and nongenotoxic stress. In: Bradshaw, RA.; Dennis, EA., editors. *Handbook of Cell Signaling*. Academic Press; New York: 2004. p. 237-247.
69. Hollstein M, Sidransky D, Vogelstein B, Harris CC. p53 mutations in human cancers. *Science* 1991;253:49–53. [PubMed: 1905840]
70. Zhao R, Gish K, Murphy M, Yin Y, Notterman D, Hoffman WH, Tom E, Mack DH, Levine AJ. Analysis of p53-regulated gene expression patterns using oligonucleotide arrays. *Genes Dev* 2000;14:981–993. [PubMed: 10783169]
71. 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;274:948–953. [PubMed: 8875929]
72. Baudier J, Delphin C, Grunwald D, Khochbin S, Lawrence JJ. Characterization of the tumor suppressor protein p53 as a protein kinase C substrate and a S100b-binding protein. *Proc Natl Acad Sci U S A* 1992;89:11627–11631. [PubMed: 1454855]

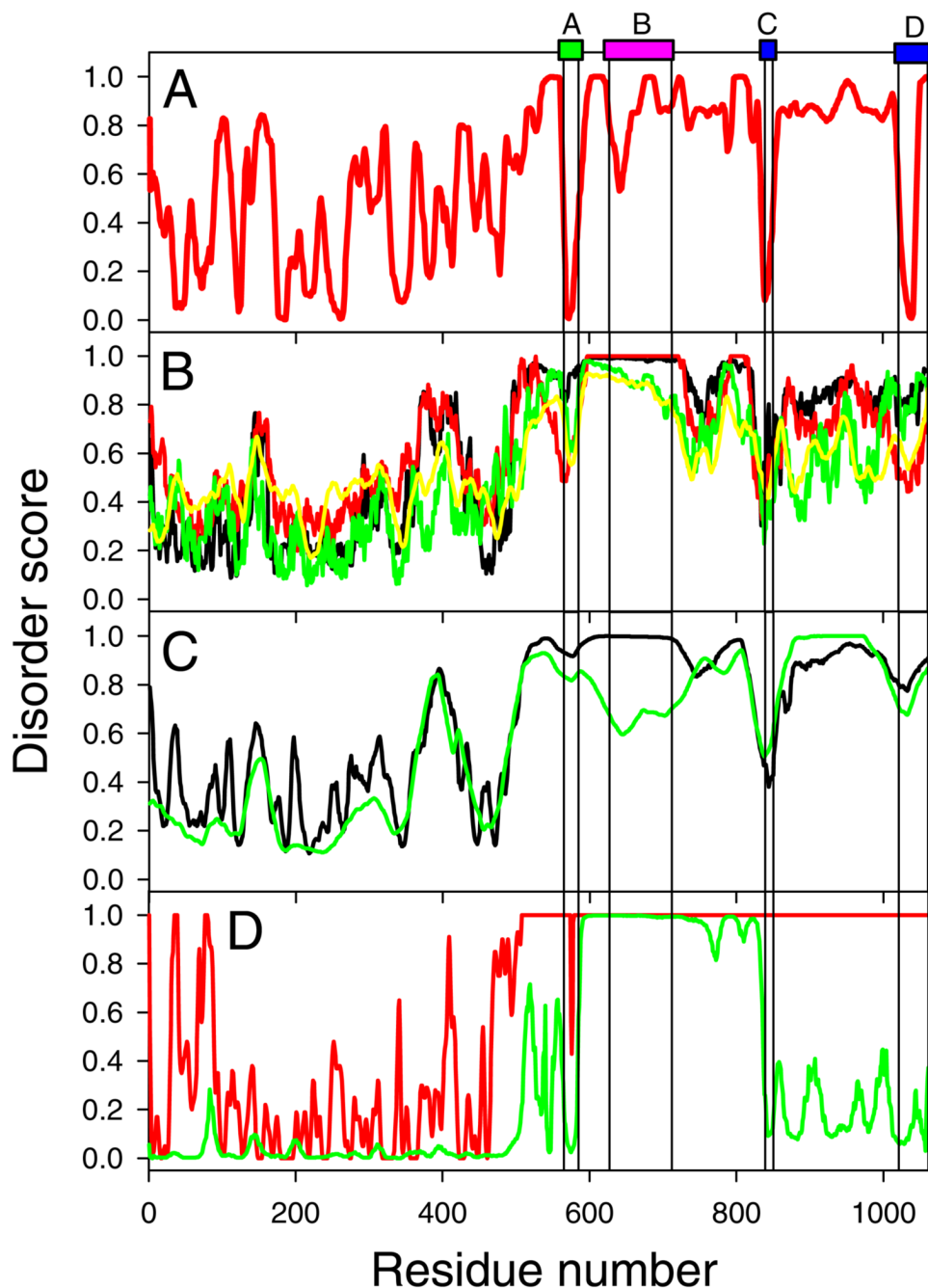
73. Rustandi RR, Drohat AC, Baldisseri DM, Wilder PT, Weber DJ. The Ca<sup>2+</sup>-dependent interaction of S100B(beta beta) with a peptide derived from p53. *Biochemistry* 1998;37:1951–1960. [PubMed: 9485322]
74. Rustandi RR, Baldisseri DM, Weber DJ. Structure of the negative regulatory domain of p53 bound to S100B(beta beta). *Nat Struct Biol* 2000;7:570–574. [PubMed: 10876243]
75. Ehretsmann CP, Carpousis AJ, Krisch HM. Specificity of Escherichia coli endoribonuclease RNase E: in vivo and in vitro analysis of mutants in a bacteriophage T4 mRNA processing site. *Genes Dev* 1992;6:149–159. [PubMed: 1730408]
76. Lee K, Bernstein JA, Cohen SN. RNase G complementation of me null mutation identifies functional interrelationships with RNase E in Escherichia coli. *Mol Microbiol* 2002;43:1445–1456. [PubMed: 11952897]



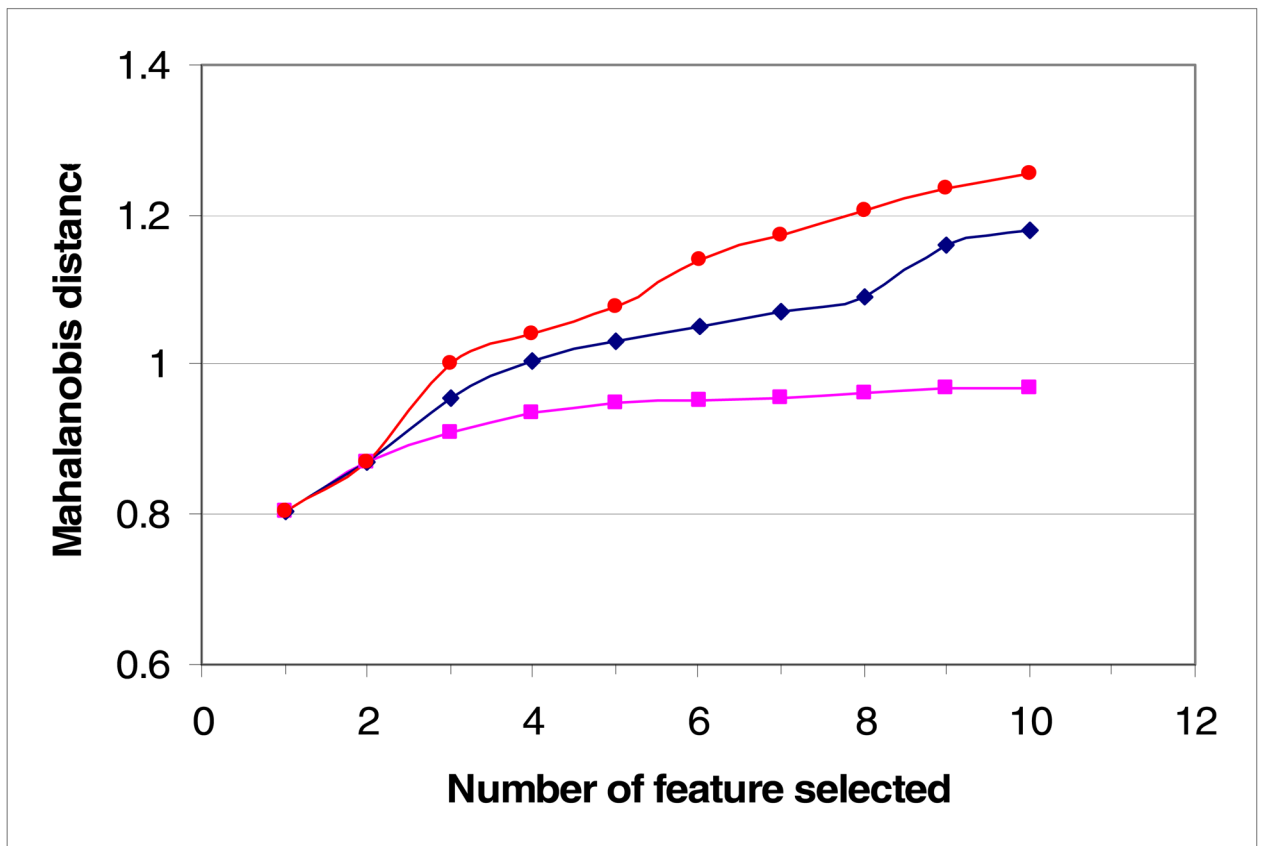
**Figure 1.** Analysis of 4E-BP1 disorder propensity by different predictors of intrinsic disorder. The plots produced by different predictors are grouped by their overall appearance and shape: **A.** PONDRL<sup>®</sup> VLXT (red curve), RONN (blue curve) and IUPred (green curve). **B.** VL3 (black line) and VL2 (red line). **C.** VSL2B (black line), VSL2P (red line) and VL3BA (green line). **D.** DisPro (black line), DRIPPRED (red line) and DISOPRED (green line). Pink bar at the top of panel A indicates the region involved in binding of the eukaryotic initiation factor 4E (eIF4E).



**Figure 2.** Analysis of p53 disorder propensity by different predictors of intrinsic disorder. The plots produced by different predictors are grouped by their overall appearance and shape: **Top panel.** PONDRL<sup>®</sup> VLXT (red curve), RONN (blue curve), IUPred (green curve), and DRIPPRED (pink line). **Middle panel.** VL3 (black line) and VL2 (red line), VSL2B (green line), VSL2P (yellow line), and VL3BA (blue line). **Bottom panel.** DisPro (black line) and DISOPRED (red line). Dark green and dark red bars at the top of panel A indicate the regions involved in binding of Mdm2 and S100B( $\beta\beta$ ), respectively.

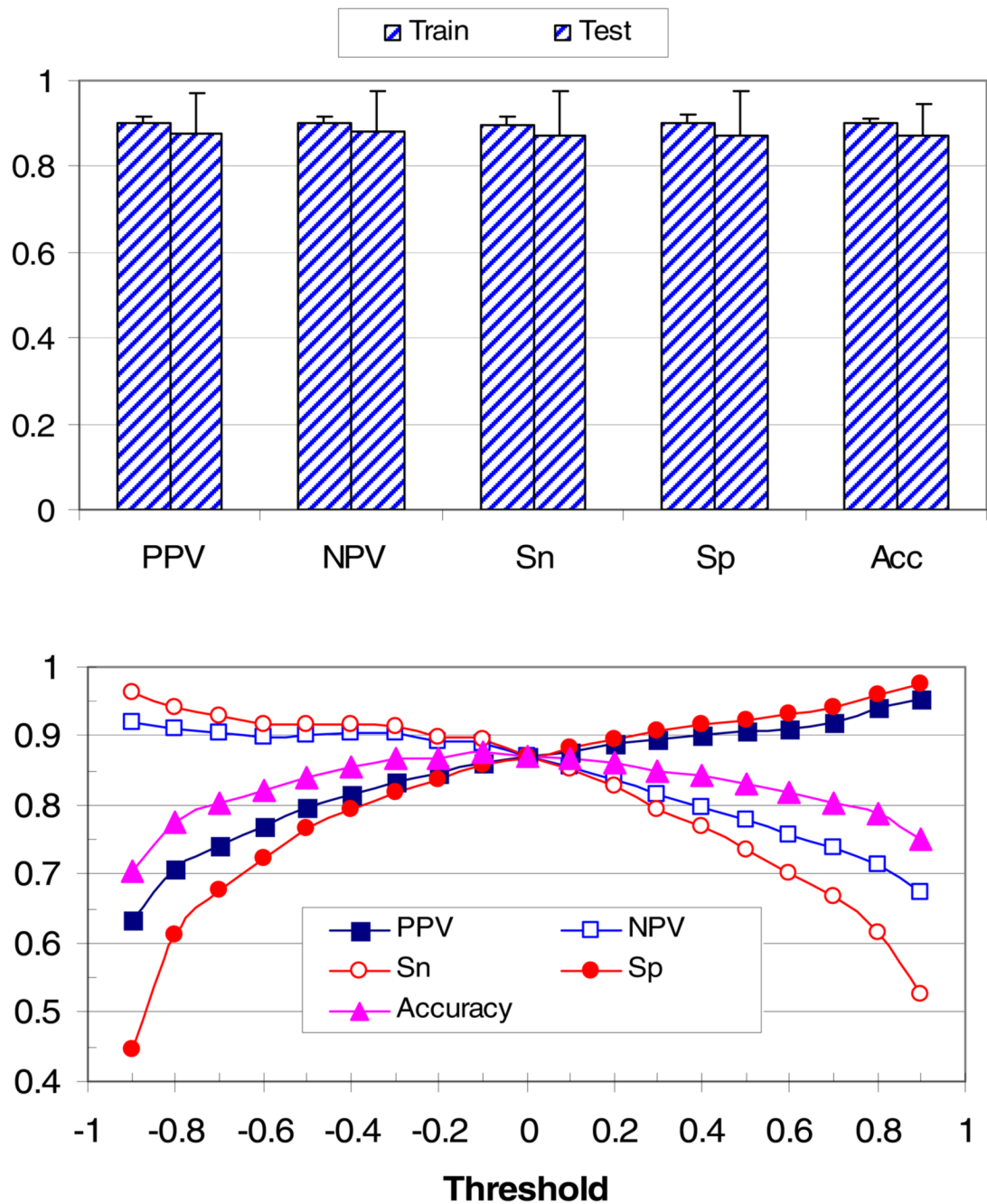


**Figure 3.** Analysis of RNase E disorder propensity by different predictors of intrinsic disorder. The plots produced by different predictors are grouped by their overall appearance and shape: **A.** PONDRL<sup>®</sup> VLXT (red curve). **B.** VL3 (black line), VL2 (red line), RONN (yellow curve), and IUPred (green curve). **C.** VSL2B (black line), VSL2P (red line) and VL3BA (green line). **D.** DisPro (black line), DRIPPRED (red line) and DISOPRED (green line). Bars at the top of panel A indicate RISP regions responsible for RNase E interaction with different binding partners: A (residues 565-585), protein–RNA interaction site; B (residues 633-712), self-recognition region; C (residues 839-850), enolase binding site; and D (residues 1021-1061), PNPase binding site.



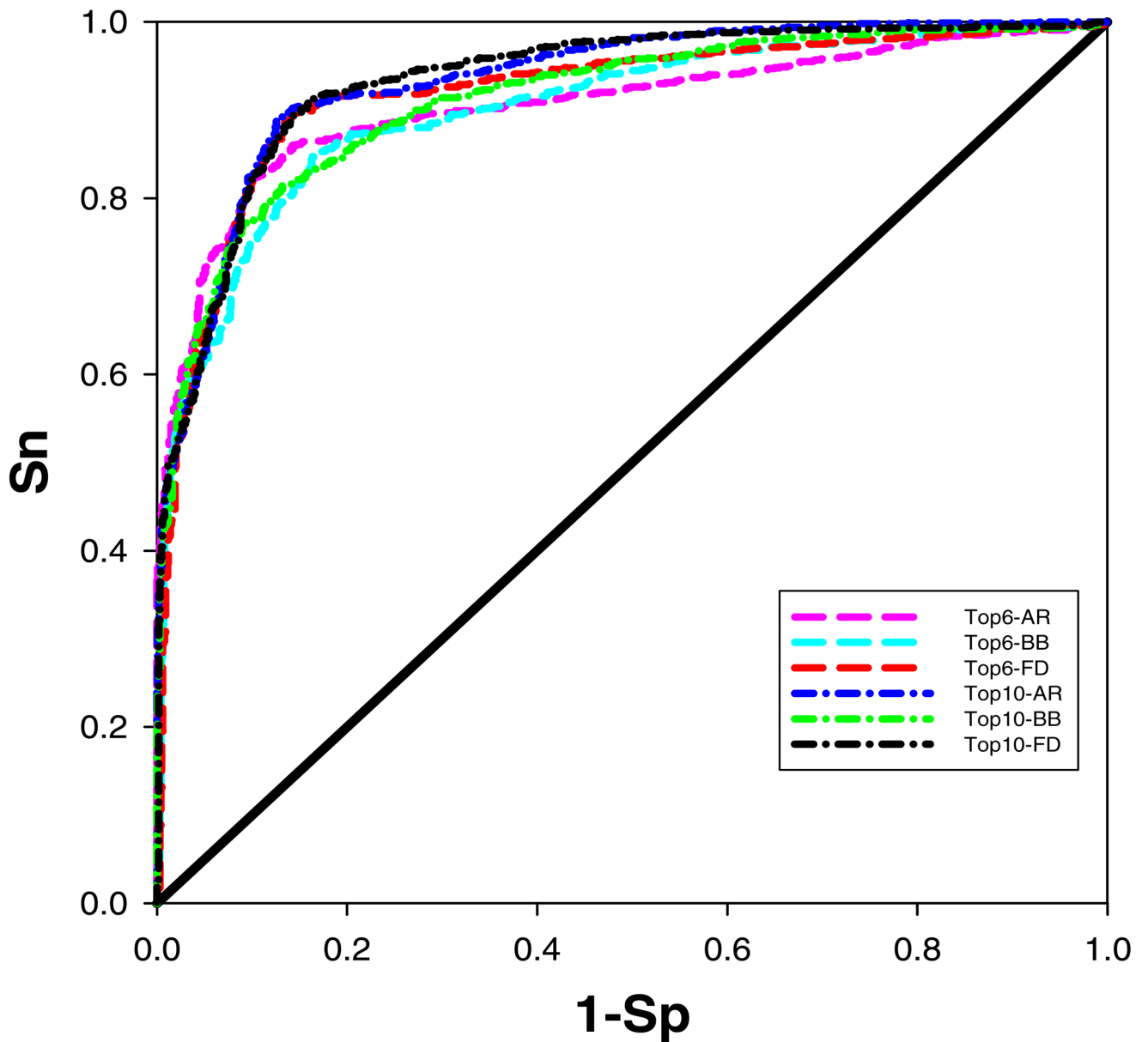
**Figure 4. Feature selection**

Mahalanobis distances for various numbers of feature combinations selected were plotted for Branch and Bound (circle), Forwards Selection (diamond), and AR (square).



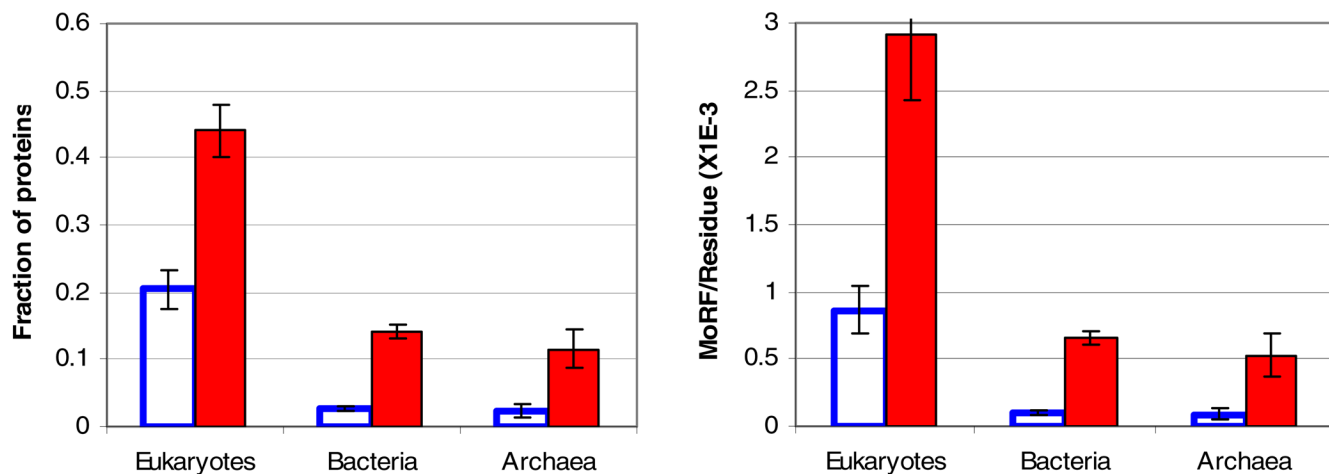
**Figure 5. Neural networks training**

A: The 10 cross validation results of neural networks constructed using top six attribute combination from Forward Selection were plotted. B: Evaluation parameters from A were plotted for various thresholds values. PPV: positive prediction value; NPV: negative prediction value; Sn: sensitivity; Sp: specificity; Acc: accuracy (Table 1).



**Figure 6. ROC curves**

Receiver operating characteristic (ROC) curves from different constructions of neural networks were plotted. Top6, Top10-AR: top 6 or top 10 attribute combination from AR was used for neural network construction, respectively; Top6, Top10-BB: top 6 or top 10 attributes combination from Branch and Bound was used for neural network construction, respectively; Top6, Top10-FD: top 6 or top 10 attributes combination was used for neural network construction, respectively.



**Figure 7.  $\alpha$ -MoRF predictions across genomes of three Kingdoms**

A: Fractions of proteins in 9 eukaryotic, 57 bacterial, and 16 archaeal genomes predicted to contain  $\alpha$ -MoRF by previous (open bar) and present method (closed bar). The error bars indicated 95% confidence interval over 1000-resampling. B: Frequency of  $\alpha$ -MoRF in 9 eukaryotic, 57 bacterial, and 16 archaeal genomes by previous (open bar) and present method (closed bar). The error bars indicated 95% confidence interval over 1000-resampling.

**Table 1**

Statistics performed on neural networks results. PPV: positive predictive value. NPV: negative prediction value. Sn: sensitivity. Sp: specificity. TP: true positive. FP: false positive. FN: false negative. TN: true negative.

<b>Evaluation</b>	<b>Formula</b>
PPV	$TP/(TP+FP)$
NPV	$TN/(TN+FN)$
Sn	$TP/(TP+FN)$
Sp	$TN/(TN+FP)$
Accuracy	$(TP+TN)/(TP+TN+FP+FN)$

**Table 2** $\alpha$ -MoRF prediction in functional classes.

Function class	% proteins contain predicted MoRF	Numbers of MoRF predicted per 1000 residues
Regulation	82 ( $\pm$ 3)	5.9
Differentiation	78 ( $\pm$ 11)	4.2
Cell division	72 ( $\pm$ 11)	5.4
Cytoskeleton	71 ( $\pm$ 7)	4.8
Ribosomal	58 ( $\pm$ 10)	3.8
Membrane	52 ( $\pm$ 7)	3.1
Transport	44 ( $\pm$ 4)	2.2
Inhibitor	42 ( $\pm$ 9)	3
Biosynthetic	32 ( $\pm$ 6)	1.2
Protease	32 ( $\pm$ 14)	1.8
G protein coupled receptors	32 ( $\pm$ 5)	1.2
Metabolism	28 ( $\pm$ 9)	0.9
Degradation	27 ( $\pm$ 11)	1.5
Kinase	26 ( $\pm$ 9)	1.1