Articles

Mining α-Helix-Forming Molecular Recognition Features with Cross Species Sequence Alignments[†]

Yugong Cheng,[‡,§] Christopher J. Oldfield,[‡] Jingwei Meng,[‡] Pedro Romero,*,[||]
Vladimir N. Uversky,*,[‡,§,⊥] and A. Keith Dunker*,[‡,§]

Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, Indiana 46202, Molecular Kinetics, Inc., 6201 La Pas Trail, Suite 160, Indianapolis, Indiana 46268, School of Informatics, Indiana University-Purdue University at Indianapolis, 535 West Michigan Street, Indianapolis, Indiana 46202, and Institute for Biological Instrumentation, Russian Academy of Sciences, 142292 Pushchino, Moscow Region, Russia

Received June 21, 2007; Revised Manuscript Received August 22, 2007

ABSTRACT: Previously described algorithms for mining α-helix-forming molecular recognition elements (MoREs), described by Oldfield et al. (Oldfield, C. J., Cheng, Y., Cortese, M. S., Brown, C. J., Uversky, V. N., and Dunker, A. K. (2005) Comparing and combining predictors of mostly disordered proteins, *Biochemistry 44*, 1989−2000), also known as molecular recognition features (MoRFs) (Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. S., Dunker, A. K., and Uversky, V. N. (2006) Analysis of Molecular Recognition Features (MoRFs), *J. Mol. Biol. 362*, 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 a training data set of a limited size. Here we propose to improve the prediction algorithms by (1) including additional α-MoRF examples and their cross species homologues in the positive training set, (2) carefully extracting monomer structure chains from the Protein Data Bank (PDB) as the negative training set, (3) including attributes from recently developed disorder predictors, secondary structure predictions, and amino acid indices, and (4) constructing neural network based predictors and performing validation. Over 50 regions which undergo disorder-to-order transition that were identified in the PDB together with a set of corresponding cross species homologues of each structure-based example were included in a new positive training set. Over 1500 attributes, including disorder predictions, secondary structure predictions, and amino acid indices, were evaluated by the 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, α-MoRF-PredII, were 0.87 ± 0.10, 0.87 ± 0.11, and 0.87 ± 0.08 over 10 cross validations, respectively. We present the results of these analyses and validation examples to discuss the potential improvement of the α-MoRF-PredII prediction accuracy.

Interactions between proteins and their partners are crucial for biological functions. Identification and prediction of such interactions would provide insights and guides for laboratory experimental efforts to understand the mechanisms of signaling and regulation within biological systems. Further, on the basis of 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 the 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 the disorder-to-order transition, which uncouples specificity from binding strength. This phenomenon has the effect of making highly specific interactions easily

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

* To whom correspondence should be addressed. Phone: (317) 278-9650. Fax: (317) 278-9217. E-mail: vuversky@iupui.edu (V.N.U.); kedunker@iupui.edu (A.K.D); promero@compbio.iupui.edu (P.R.).

[‡] Indiana University School of Medicine.

[§] Molecular Kinetics, Inc.

[||] Indiana University-Purdue University at Indianapolis.

[⊥] Russian Academy of Sciences.

10.1021/bi7012273 CCC: $37.00
Published on Web 11/01/2007

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 for p27$^{\mathrm{Kip1}}$ (*12*, *13*) and p53 (*14*), both of which have disordered regions with significant helical character that form α-helices upon binding to their partners.

These foldable partners of protein−protein interactions are members of the recently discovered class of intrinsically disordered (ID) proteins, which lack rigid 3D structure under physiological conditions in vitro. Bioinformatics studies indicated that about 25−30% 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 the 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, the concept of a molecular recognition feature (MoRF;[1] because such regions "morph" from disorder to order upon binding) was introduced for a 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 α-helix-forming MoRF signatures was developed on the basis of the patterns of predictors of naturally disordered regions (PONDRs), 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 plot to predict binding sites that were subsequently verified by laboratory experiments (*40*, *41*). For some of these predicted examples, the regions did indeed form a 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 α-MoRF examples (14 regions from 12 proteins). All of the training examples are correctly identified by the current algorithm, suggesting overfitting. Here we represent a novel algorithm which is improved by (1) including updated α-MoRF examples and their cross species homologues in the positive training set, (2) extracting monomer structure chains from the Protein Data Bank (PDB) as the negative training set, (3) mining attributes from newly developed disorder predictions, secondary structure predictions, and amino acid indices (*48*), and (4) constructing neural network based predictors and performing validation.

MATERIALS AND METHODS

Data Sets. α-MoRF examples were retrieved by the following procedures from structures in the 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 the 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 the Macromolecular Structure Database (http://pqs.ebi.ac.uk/) by selecting "monomeric" in the 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 fashion. 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

[1] Abbreviations: MoRF, molecular recognition feature; MoRE, molecular recognition element; PONDR, predictor of naturally disordered regions; ROC, receiver operating curve; AR, area ratio; PPV, positive prediction value; NPV, negative prediction value; Sn, sensitivity; Sp, specificity; Acc, accuracy; AUC, area under the ROC; ELM, eukaryotic linear motif.

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; 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 the conditional probability is plotted 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 bound 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, on the basis of 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 data set 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.

Table 1: Statistics Performed on Neural Networks Results[a]

evaluation	formula	evaluation	formula
PPV	TP/(TP + FP)	Sp	TN/(TN + FP)
NPV	TN/(TN + FN)	Acc	(TP + TN)/(TP + TN + FP + FN)
Sn	TP/(TP + FN)		

[a] Key: PPV, positive prediction value; NPV, negative prediction value; Sn, sensitivity; Sp, specificity; TP, true positive; FP, false positive; FN, false negative; TN, true negative.

Evaluation of Neural Networks. The results of the neural network training were evaluated by multiple methods. Several of these—including the positive prediction value (PPV), negative prediction value (NPV), sensitivity (Sn), specificity (Sp), 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 (ROCs) were used.

An ROC is a two-dimensional measure of classification performance. The ROC is defined as a plot of the true positive rate as a function of the false positive rate. An empirical ROC can be generated by calculating TP and FP for all relevant thresholds. We approximated the ROC by simply connecting the data points (Sn, 1 − Sp) with straight lines. The full area under the ROC (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 on the basis of previous observations (*58*). Here, we reexamine this choice by examining whether indication of binding regions is a feature specific to PONDR VLXT profiles or 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 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 mu-

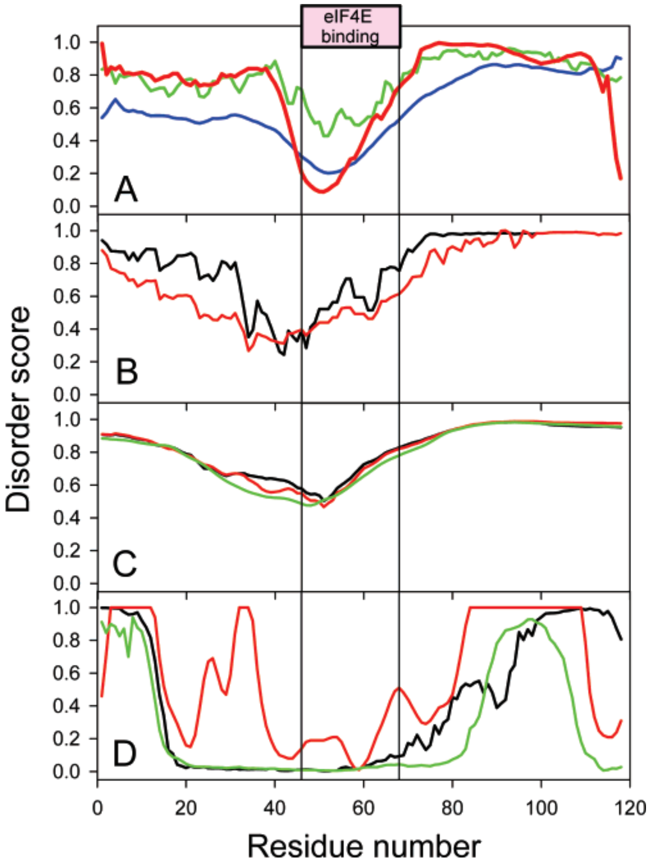

FIGURE 1: Analysis of 4E-BP1 disorder propensity by different predictors of intrinsic disorder. The plots produced by different predictors are grouped by their overall appearance and shape. (A) PONDR VLXT (red curve), RONN (blue curve) and IUPred (green curve). (B) VL3 (black line) and VL2 (red line).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

tagenesis identified the 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, the 4E-BP1 ^{15}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 chemical shift changes similar to those of 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 an algorithm for identifying α-MoRFs (*9*).

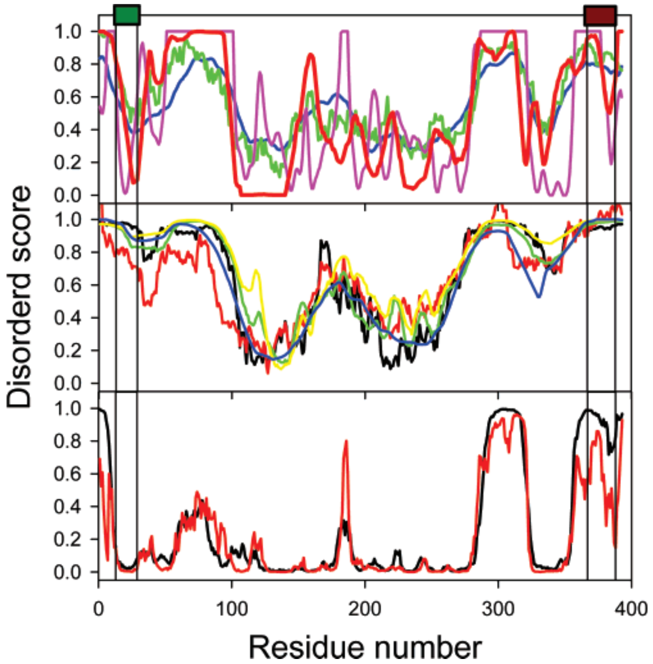

FIGURE 2: Analysis of p53 disorder propensity by different predictors of intrinsic disorder. The plots produced by different predictors are grouped by their overall appearance and shape. (Top) PONDR VLXT (red curve), RONN (blue curve), IUPred (green curve), and DRIPPRED (pink line). (Middle**)** VL3 (black line), VL2 (red line), VSL2B (green line), VSL2P (yellow line), and VL3BA (blue line). (Bottom**)** 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(*ββ*), respectively.

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 (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 the dip predicted by PONDR VLXT (Figure 1A). Members of the VL3 (including VL3, VL3H, and VL3E) (*51*) and VL2 (including VL2 and VL2-S) families (*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 the VSL2 family [VSL2B and VSL2P (*67*)] and by PONDR VL3BA (*51*) were very shallow and 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

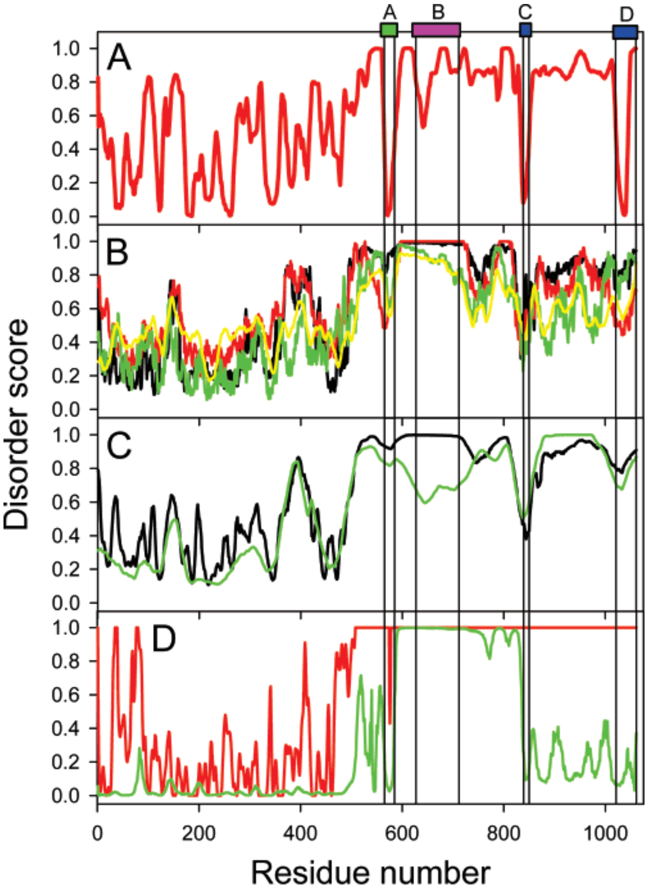

FIGURE 3: Analysis of RNase E disorder propensity by different predictors of intrinsic disorder. The plots produced by different predictors are grouped by their overall appearance and shape. (A) PONDR VLXT (red curve). (B) VL3 (black line), VL2 (red line), RONN (yellow curve), and IUPred (green curve). (C) VSL2B (black line), VSL2P (red line), and VL3BA (green line). (D) DisPro (black line), DRIPPRED (red line), and DISOPRED (green line). The bars at the top of panel A indicate RISPs 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; D (residues 1021−1061), PNPase binding site.

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: the N-terminal translational activation domain, central DNA binding domain, and C-terminal tetramerization and regulatory domain. At the transactivation region, it interacts with TFIID, TFIIH, Mdm2, RPA, CBP/p300, and CSN5/Jab1 (*68*). At the C-terminal domain, it interacts with GSK3β, 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^{2+}-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^{2+}-loaded S100B($\beta\beta$) (*74*).

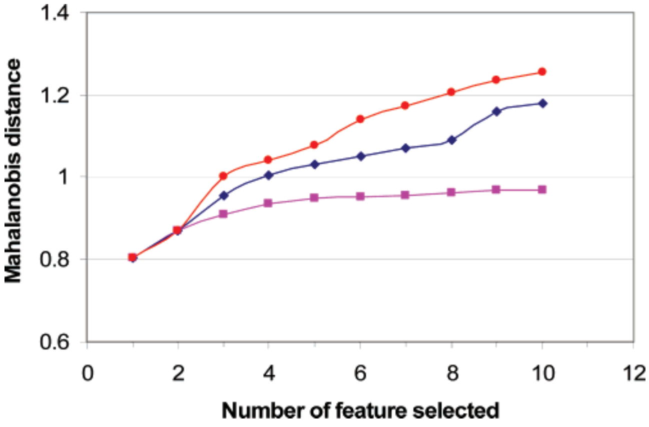

FIGURE 4: Feature selection. Mahalanobis distances for various numbers of feature combinations selected were plotted for branch and bound (circles), forward selection (tilted squares), and AR (squares).

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 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 dips at both the N- and C-termini as PONDR VLXT (see Figure 2A). Predictions by IUPred (*64*) and RONN (*65*) showed a matched dip at the N-terminus and a shallower dip at the C-terminus, compared to those 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 the disordered N-terminus and predicted the 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

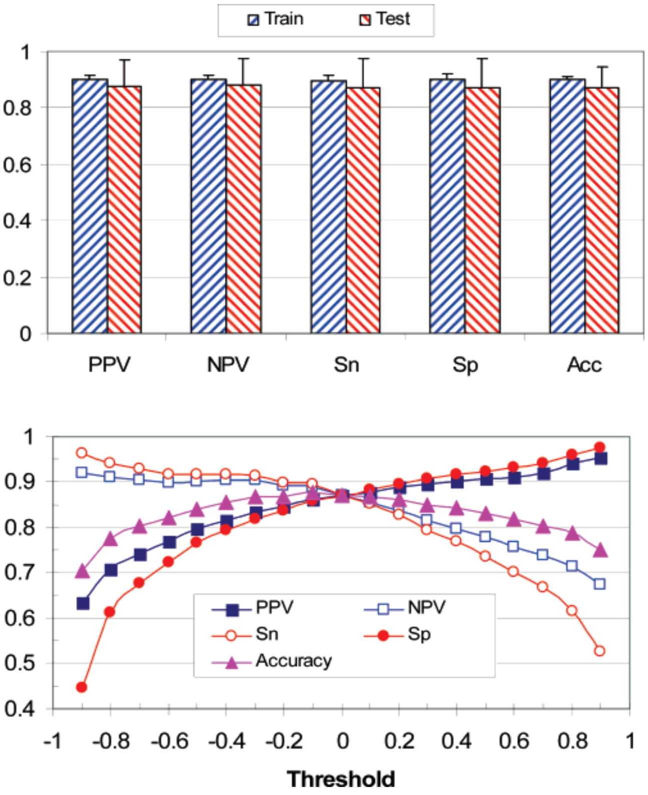

FIGURE 5: Neural network training. (A) The 10 cross validation results of neural networks constructed using the top six attribute combinations from forward selection were plotted. (B) Evaluation parameters from (A) were plotted for various threshold values. Key: PPV, positive prediction value; NPV, negative prediction value; Sn, sensitivity; Sp, specificity; Acc, accuracy (Table 1).

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 VLXT prediction (*40*). However, VLXT prediction showed four sharp downward spikes within the entirely disordered CTD. These dips were referred to as “regions of increased structural propensity” (RISPs) (*40*) and are 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 a 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 an α-helix in the complex (*77*). Therefore, all C-terminal regions in RNase E with the predisposition for the disorder-to-order transition were correctly visualized by PONDR 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 the A, C, and D positions (Figure 3B), while VSL2 and VL3BA predictions almost do not have a dip at position A. However, they have pronounced dips at positions C and D, and VL3BA has a very broad dip centered at position 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 the B, C, and D positions. In fact, DRIPPRED predicts that the 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 VLXT does.

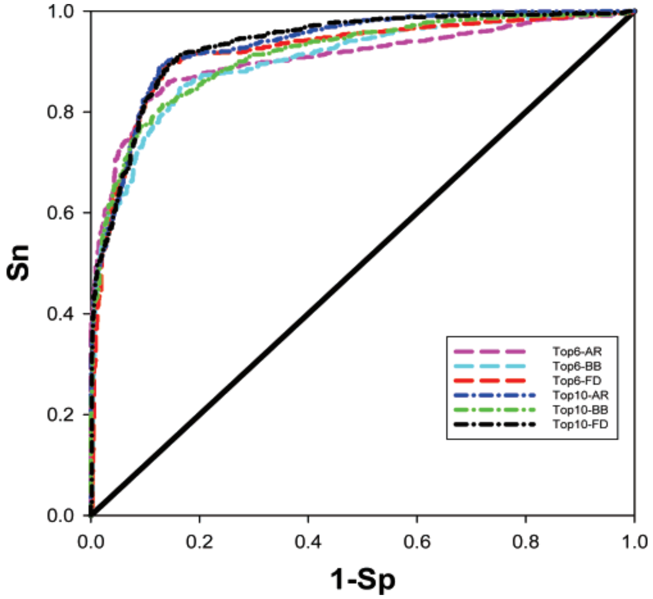

FIGURE 6: ROCs from different constructions of neural networks were plotted. Key: Top6-AR or Top10-AR, the top 6 or top 10 attribute combination from AR was used for neural network construction, respectively; Top6-BB or Top10-BB, the top 6 or top 10 attribute combination from branch and bound was used for neural network construction, respectively; Top6-FD or Top10-FD, the top 6 or top 10 attribute combination was used for neural network construction, respectively.

Overall, data presented above show that many predictors gave similar general disorder/order predictions on the three examples. However, PONDR 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. A total of 54 basic MoRF regions (from 51 proteins) were retrieved from PDB structures. Of these MoRF regions, 48 have at least one cross species homologue sequence that could be retrieved from SwissProt and the remaining 6 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. The 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.

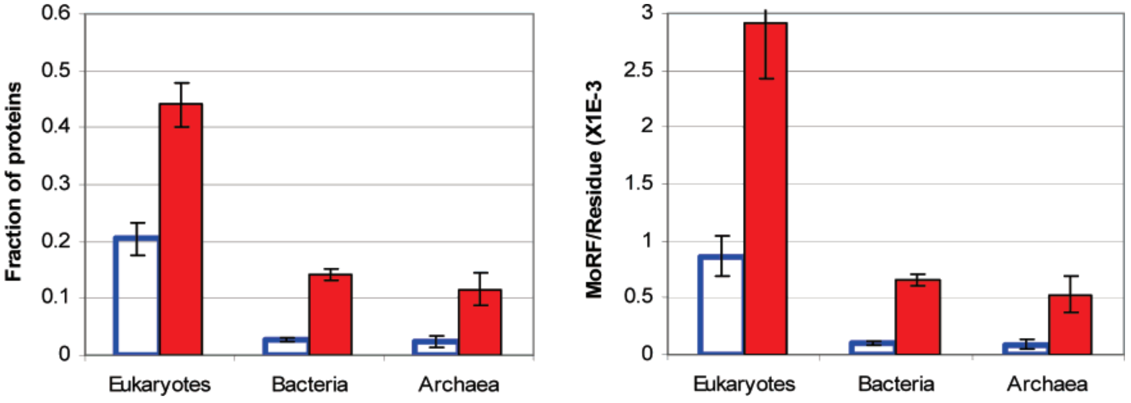

FIGURE 7: α-MoRF predictions across genomes of three kingdoms. (A) Fractions of proteins in 9 eukaryotic, 57 bacterial, and 16 archaeal genomes predicted to contain α-MoRFs by previous (open bars) and present (closed bars) methods. The error bars indicate the 95% confidence interval over 1000 resamplings. (B) Frequency of α-MoRFs in 9 eukaryotic, 57 bacterial, and 16 archaeal genomes by previous (open bars) and present (closed bars) methods. The error bars indicate the 95% confidence interval over 1000 resampling.

Table 2: α-MoRF Prediction in Functional Classes

function class	proteins containing predicted MoRFs (%)	no. of predicted MoRFs per 1000 residues	function class	proteins containing predicted MoRFs (%)	no. of predicted MoRFs per 1000 residues
regulation	82 (±3)	5.9	inhibitor	42 (±9)	3
differentiation	78 (±11)	4.2	biosynthetic	32 (±6)	1.2
cell division	72 (±11)	5.4	protease	32 (±14)	1.8
cytoskeleton	71 (±7)	4.8	G protein coupled receptors	32 (±5)	1.2
ribosomal	58 (±10)	3.8	metabolism	28 (±9)	0.9
membrane	52 (±7)	3.1	degradation	27 (±11)	1.5
transport	44 (±4)	2.2	kinase	26 (±9)	1.1

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

Feature Selection. Over 1500 attributes of the binding, flanking, and whole regions 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 were in the range of 0.64−0.54. As shown in Figure 4, the Mahalanobis distances were compared for the top 10 features selected from the top 30 AR attributes by sequential forward selection or branch and bound to direct the top 10 AR attributes. Branch and bound and forward selection yielded higher Mahalanobis distances than direct AR for 3−10 features selected, indicating those feature combinations may yield better separations between the positive and negative training sets. Furthermore, the Mahalanobis distance reached a plateau when six features were selected. Thus, the top 6 and top 10 feature combinations from all three methods were used to train neural networks.

Neural Network 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 the top 6 features from forward selection were used as an illustration in Figure 5A: sensitivity, specificity, and accuracy after 10 cross validations were 0.87 ± 0.10, 0.87 ± 0.11, and 0.87 ± 0.08, respectively. When the threshold for prediction increased from −0.9 to 0.9, PPV and specificity increased from 0.63 (±0.10) to 0.95 (±0.09) and from 0.45 (±0.19) to 0.98 (±0.07), respectively, while NPV and sensitivity decreased from 0.92 (±0.10) to 0.67 (±0.08) and from 0.96 (±0.09) to 0.53 (±0.15), respectively (Figure 5B). However, accuracy reached a maximum (0.87) when the threshold was set to 0 or −0.1. ROCs 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 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 network constructions showed that all of them were better than random, whereas curves for the top 10 features from forward selection, the top 10 features from AR, and the top 6 features from forward selection were on top of the others. Thus, a neural network constructed on the basis of the top 6 features from forward selection was chosen over that constructed on the basis of the top 10 features.

Examination of Specific α-MoRF-PredII Predictions. The performance of the α-MoRF-PredII identifier was analyzed using a set of proteins with known MoRFs.

1. Escherichia coli Ribonuclease RNase E. The current algorithm predicted all four RISPs as MoRFs (see Figure 2), whereas the original predictor failed to identify any of these dips as molecular recognition features.

2. 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. α-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.

3. 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 1500 basic MoRF (dip) patterns were predicted as MoRFs; i.e., 82% accuracy is achieved for this set.

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

The α-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, and the cytoskeleton, as well as ribosomal proteins, contain more α-MoRF than proteins in the average eukaryotic genome. Membrane, transport, and inhibitory proteins have α-MoRF propensities similar to those of proteins from the average eukaryotic genome. Finally, proteins associated with biosynthesis, protease activities, G protein coupled receptors, metabolism, degradation, and kinase activities have lower α-MoRF propensities than the average eukaryotic protein.

Summarizing, we elaborated a novel neural network based algorithm for mining α-helix-forming molecular recognition features, α-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 α-MoRE identifier (*9*), this algorithm was improved by using an extended set of newly identified α-MoRFs and their homologues, by extracting monomer chains from the PDB as the negative training set and via mining novel attributes related to disorder and secondary structure predictions as well as amino acid indices. The top attributes were used to develop the feed forward neural networks. The performance of the resulting tool, α-MoRF-PredII predictor, was validated on the basis of a set of proteins with known α-MoRFs as a positive control and a set of ordered proteins that do not contain α-MoRFs as a negative control. The sensitivity, specificity, and accuracy of the α-MoRF-PredII predictor were all 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 of functional classes of human proteins.

REFERENCES

1. Fry, D. C., and Vassilev, L. T. (2005) Targeting protein-protein interactions for cancer therapy, *J. Mol. Med. 83*, 955−963.
2. Arkin, M. (2005) Protein-protein interactions and cancer: small molecules going in for the kill, *Curr. Opin. Chem. Biol. 9*, 317−324.
3. Dyson, H. J., and Wright, P. E. (2002) Coupling of folding and binding for unstructured proteins, *Curr. Opin. Struct. Biol. 12*, 54−60.
4. Uversky, V. N., Gillespie, J. R., and Fink, A. L. (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions?, *Proteins 41*, 415−427.
5. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., and Obradovic, Z. (2002) Intrinsic disorder and protein function, *Biochemistry 41*, 6573−6582.
6. Dyson, H. J., and Wright, P. E. (2005) Intrinsically unstructured proteins and their functions, *Nat. Rev. Mol. Cell Biol. 6*, 197−208.
7. Wright, P. E., and Dyson, H. J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm, *J. Mol. Biol. 293*, 321−331.
8. Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. S., Dunker, A. K., and Uversky, V. N. (2006) Analysis of molecular recognition features (MoRFs), *J. Mol. Biol. 362*, 1043−1059.
9. Oldfield, C. J., Cheng, Y., Cortese, M. S., Romero, P., Uversky, V. N., and Dunker, A. K. (2005) Coupled folding and binding with alpha-helix-forming molecular recognition elements, *Biochemistry 44*, 12454−12470.
10. Vacic, V., Oldfield, C. J., Mohan, A., Radivojac, P., Cortese, M. S., Uversky, V. N., and Dunker, A. K. (2007) Characterization of molecular recognition features, MoRFs, and their binding partners, *J. Proteome Res. 6*, 2351−2366.
11. Fuxreiter, M., Simon, I., Friedrich, P., and Tompa, P. (2004) Preformed structural elements feature in partner recognition by intrinsically unstructured proteins, *J. Mol. Biol. 338*, 1015−1026.
12. Bienkiewicz, E. A., Adkins, J. N., and Lumb, K. J. (2002) Functional consequences of preorganized helical structure in the intrinsically disordered cell-cycle inhibitor p27(Kip1), *Biochemistry 41*, 752−759.
13. Lacy, E. R., Filippov, I., Lewis, W. S., Otieno, S., Xiao, L., Weiss, S., Hengst, L., and Kriwacki, R. W. (2004) p27 binds cyclin-CDK complexes through a sequential mechanism involving binding-induced protein folding, *Nat. Struct. Mol. Biol. 11*, 358−364.
14. Lee, H., Mok, K. H., Muhandiram, R., Park, K. H., Suk, J. E., Kim, D. H., Chang, J., Sung, Y. C., Choi, K. Y., and Han, K. H. (2000) Local structural elements in the mostly unstructured transcriptional activation domain of human p53, *J. Biol. Chem. 275*, 29426−29432.
15. Oldfield, C. J., Cheng, Y., Cortese, M. S., Brown, C. J., Uversky, V. N., and Dunker, A. K. (2005) Comparing and combining predictors of mostly disordered proteins, *Biochemistry 44*, 1989−2000.
16. Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C., and Brown, C. J. (2000) Intrinsic protein disorder in complete genomes, *Genome Inf. Ser. Workshop Genome Inf. 11*, 161−171.
17. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life, *J. Mol. Biol. 337*, 635−645.
18. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z., and Dunker, A. K. (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins, *J. Mol. Biol. 323*, 573−584.
19. Daughdrill, G. W., Pielak, G. J., Uversky, V. N., Cortese, M. S., and Dunker, A. K. (2005) Natively disordered proteins, in *Protein Folding Handbook* (Buchner, J., and Kiefhaber, T., Eds.) pp 275−357, Wiley-VCH, New York.
20. Dunker, A. K., Brown, C. J., and Obradovic, Z. (2002) Identification and functions of usefully disordered proteins, *Adv. Protein Chem. 62*, 25−49.
21. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M., and Uversky, V. N. (2005) Flexible nets. The roles of intrinsic disorder in protein interaction networks, *FEBS J. 272*, 5129−5148.
22. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., and Obradovic, Z. (2001) Intrinsically disordered protein, *J. Mol. Graphics Modell. 19*, 26−59.
23. Dunker, A. K., and Obradovic, Z. (2001) The protein trinity−linking function and disorder, *Nat. Biotechnol. 19*, 805−806.
24. Liu, J., Perumal, N. B., Oldfield, C. J., Su, E. W., Uversky, V. N., and Dunker, A. K. (2006) Intrinsic disorder in transcription factors, *Biochemistry 45*, 6873−6888.

25. Uversky, V. N., Oldfield, C. J., and Dunker, A. K. (2005) Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling, *J. Mol. Recognit. 18*, 343−384.
26. Uversky, V. N., Roman, A., Oldfield, C. J., and Dunker, A. K. (2006) Protein intrinsic disorder and human papillomaviruses: increased amount of disorder in E6 and E7 oncoproteins from high risk HPVs, *J. Proteome Res. 5*, 1829−1842.
27. Uversky, V. N. (2002) Natively unfolded proteins: a point where biology waits for physics, *Protein Sci. 11*, 739−756.
28. Uversky, V. N. (2002) What does it mean to be natively unfolded?, *Eur. J. Biochem. 269*, 2−12.
29. Uversky, V. N. (2003) Protein folding revisited. A polypeptide chain at the folding-misfolding-nonfolding cross-roads: which way to go?, *Cell. Mol. Life Sci. 60*, 1852−1871.
30. Tompa, P. (2002) Intrinsically unstructured proteins, *Trends Biochem. Sci. 27*, 527−533.
31. Fink, A. L. (2005) Natively unfolded proteins, *Curr. Opin. Struct. Biol. 15*, 35−41.
32. Xie, H., Vucetic, S., Iakoucheva, L. M., Oldfield, C. J., Dunker, A. K., Obradovic, Z., and Uversky, V. N. (2007) Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins, *J. Proteome Res. 6*, 1917−1932.
33. Vucetic, S., Xie, H., Iakoucheva, L. M., Oldfield, C. J., Dunker, A. K., Obradovic, Z., and Uversky, V. N. (2007) 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. 6*, 1899−1916.
34. Xie, H., Vucetic, S., Iakoucheva, L. M., Oldfield, C. J., Dunker, A. K., Uversky, V. N., and Obradovic, Z. (2007) Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions, *J. Proteome Res. 6*, 1882−1898.
35. Radivojac, P., Iakoucheva, L. M., Oldfield, C. J., Obradovic, Z., Uversky, V. N., and Dunker, A. K. (2007) Intrinsic disorder and functional proteomics, *Biophys. J. 92*, 1439−1456.
36. Schulz, G. E. (1979) Nucleotide binding proteins, in *Molecular Mechanism of Biological Recognition* (Balaban, M., Ed.) pp 79−94, Elsevier/North-Holland Biomedical Press, New York.
37. Kriwacki, R. W., Hengst, L., Tennant, L., Reed, S. I., and Wright, P. E. (1996) 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. 93*, 11504−11509.
38. Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. (2001) Lethality and centrality in protein networks, *Nature 411*, 41−42.
39. Dajani, R., Fraser, E., Roe, S. M., Yeo, M., Good, V. M., Thompson, V., Dale, T. C., and Pearl, L. H. (2003) Structural basis for recruitment of glycogen synthase kinase 3beta to the axin-APC scaffold complex, *EMBO J. 22*, 494−501.
40. Callaghan, A. J., Aurikko, J. P., Ilag, L. L., Gunter, Grossmann, J., Chandran, V., Kuhnel, K., Poljak, L., Carpousis, A. J., Robinson, C. V., Symmons, M. F., and Luisi, B. F. (2004) Studies of the RNA degradosome-organizing domain of the Escherichia coli ribonuclease RNase E, *J. Mol. Biol. 340*, 965−979.
41. Bourhis, J.-M., Johansson, K., Receveur-Brechot, V., Oldfield, C. J., Dunker, A. K., Canard, B., and Longhi, S. (2004) 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. 99*, 157−167.
42. Kingston, R. L., Hamel, D. J., Gay, L. S., Dahlquist, F. W., and Matthews, B. W. (2004) Structural basis for the attachment of a paramyxoviral polymerase to its template, *Proc. Natl. Acad. Sci. U.S.A. 101*, 8301−8306.
43. Chandran, V., and Luisi, B. F. (2006) Recognition of enolase in the Escherichia coli RNA degradosome, *J. Mol. Biol. 358*, 8−15.
44. Puntervoll, P., Linding, R., Gemund, C., Chabanis-Davidson, S., Mattingsdal, M., Cameron, S., Martin, D. M., 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, W. N., Aasland, R., and Gibson, T. J. (2003) ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins, *Nucleic Acids Res. 31*, 3625−3630.
45. Neduva, V., Linding, R., Su-Angrand, I., Stark, A., de Masi, F., Gibson, T. J., Lewis, J., Serrano, L., and Russell, R. B. (2005) Systematic discovery of new recognition peptides mediating protein interaction networks, *PLoS Biol. 3*, e405.
46. Neduva, V., and Russell, R. B. (2005) Linear motifs: evolutionary interaction switches, *FEBS Lett. 579*, 3342−3345.
47. Fuxreiter, M., Tompa, P., and Simon, I. (2007) Local structural disorder imparts plasticity on linear motifs, *Bioinformatics 23*, 950−956.
48. Kawashima, S., and Kanehisa, M. (2000) AAindex: amino acid index database, *Nucleic Acids Res. 28*, 374.
49. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res. 22*, 4673−4680.
50. Radivojac, P., Obradovic, Z., Brown, C. J., and Dunker, A. K. (2002) Improving sequence alignments for intrinsically disordered proteins, *Pac. Symp. Biocomput.* 589−600.
51. Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C. J., and Dunker, A. K. (2003) Predicting intrinsic disorder from amino acid sequence, *Proteins 53* (Suppl. 6), 566−572.
52. Peng, K., Radivojac, P., Vucetic, S., Dunker, A. K., and Obradovic, Z. (2006) Length-dependent prediction of protein intrinsic disorder, *BMC Bioinf. 7*, 208.
53. Linding, R., Russell, R. B., Neduva, V., and Gibson, T. J. (2003) GlobPlot: Exploring protein sequences for globularity and disorder, *Nucleic Acids Res. 31*, 3701−3708.
54. Garnier, J., Gibrat, J. F., and Robson, B. (1996) GOR method for predicting protein secondary structure from amino acid sequence, *Methods Enzymol. 266*, 540−553.
55. Arnold, G. E., Dunker, A. K., Johns, S. J., and Douthart, R. J. (1992) Use of conditional probabilities for determining relationships between amino acid sequence and protein secondary structure, *Proteins 12*, 382−399.
56. Rauber, T. W., Barata, M. M., and Steiger-Garcao, A. S. (1993) *The International Conference on Fault Diagnosis*, Toulouse, France.
57. Lasko, T. A., Bhagwat, J. G., Zou, K. H., and Ohno-Machado, L. (2005) The use of receiver operating characteristic curves in biomedical informatics, *J. Biomed. Inf. 38*, 404−415.
58. Garner, E., Romero, P., Dunker, A. K., Brown, C., and Obradovic, Z. (1999) Predicting Binding Regions within Disordered Proteins, *Genome Inf. Ser. Workshop Genome Inf. 10*, 41−50.
59. Heesom, K. J., Gampel, A., Mellor, H., and Denton, R. M. (2001) Cell cycle-dependent phosphorylation of the translational repressor eIF-4E binding protein-1 (4E-BP1), *Curr. Biol. 11*, 1374−1379.
60. Mader, S., Lee, H., Pause, A., and Sonenberg, N. (1995) 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. 15*, 4990−4997.
61. Fletcher, C. M., and Wagner, G. (1998) The interaction of eIF4E with 4E-BP1 is an induced fit to a completely disordered protein, *Protein Sci. 7*, 1639−1642.
62. Fletcher, C. M., McGuire, A. M., Gingras, A. C., Li, H., Matsuo, H., Sonenberg, N., and Wagner, G. (1998) 4E binding proteins inhibit the translation factor eIF4E without folded structure, *Biochemistry 37*, 9−15.
63. Cheng, J., Sweredoski, M., and Baldi, P. (2005) Accurate prediction of protein disordered regions by mining protein structure data, *Data Min. Knowl. Discovery 11*.
64. Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content, *Bioinformatics 21*, 3433−3434.
65. Yang, Z. R., Thomson, R., McNeil, P., and Esnouf, R. M. (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins, *Bioinformatics 21*, 3369−3376.
66. Vucetic, S., Brown, C. J., Dunker, A. K., and Obradovic, Z. (2003) Flavors of protein disorder, *Proteins 52*, 573−584.
67. Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., and Dunker, A. K. (2005) Exploiting heterogeneous sequence properties improves prediction of protein disorder, *Proteins 61* (Suppl. 7), 176−182.
68. Anderson, C. W., and Appella, E. (2004) Signaling to the p53 tumor suppressor through pathways activated by genotoxic and nongenotoxic stress, in *Handbook of Cell Signaling* (Bradshaw, R. A., and Dennis, E. A., Eds.) pp 237−247, Academic Press, New York.
69. Hollstein, M., Sidransky, D., Vogelstein, B., and Harris, C. C. (1991) p53 mutations in human cancers, *Science 253*, 49−53.

70. Zhao, R., Gish, K., Murphy, M., Yin, Y., Notterman, D., Hoffman, W. H., Tom, E., Mack, D. H., and Levine, A. J. (2000) Analysis of p53-regulated gene expression patterns using oligonucleotide arrays, *Genes Dev. 14,* 981−993.

71. Kussie, P. H., Gorina, S., Marechal, V., Elenbaas, B., Moreau, J., Levine, A. J., and Pavletich, N. P. (1996) Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain, *Science 274*, 948−953.

72. Baudier, J., Delphin, C., Grunwald, D., Khochbin, S., and Lawrence, J. J. (1992) 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. 89*, 11627−11631.

73. Rustandi, R. R., Drohat, A. C., Baldisseri, D. M., Wilder, P. T., and Weber, D. J. (1998) The Ca(2+)-dependent interaction of S100B(beta beta) with a peptide derived from p53, *Biochemistry 37*, 1951−1960.

74. Rustandi, R. R., Baldisseri, D. M., and Weber, D. J. (2000) Structure of the negative regulatory domain of p53 bound to S100B(betabeta), *Nat. Struct. Biol. 7*, 570−574.

75. Ehretsmann, C. P., Carpousis, A. J., and Krisch, H. M. (1992) Specificity of Escherichia coli endoribonuclease RNase E: in vivo and in vitro analysis of mutants in a bacteriophage T4 mRNA processing site, *Genes Dev. 6*, 149−159.

76. Lee, K., Bernstein, J. A., and Cohen, S. N. (2002) RNase G complementation of rne null mutation identifies functional interrelationships with RNase E in Escherichia coli, *Mol. Microbiol. 43*, 1445−1456.

77. Chandran V., Luisi B. F. (2006) Recognition of enolase in the *Escherichia coli* RNA degradosome. *J. Mol Biol*. *358*, 8−15.

BI7012273