Coupled Folding and Binding with α-Helix-Forming Molecular Recognition Elements†

Christopher J. Oldfield,[‡,§] Yugong Cheng,[‡] Marc S. Cortese,[||] Pedro Romero,[||,⊥] Vladimir N. Uversky,[‡,||,#] and A. Keith Dunker*[,‡,||]

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

Received April 21, 2005; Revised Manuscript Received June 28, 2005

ABSTRACT: Many protein−protein and protein−nucleic acid interactions involve coupled folding and binding of at least one of the partners. Here, we propose a protein structural element or feature that mediates the binding events of initially disordered regions. This element consists of a short region that undergoes coupled binding and folding within a longer region of disorder. We call these features "molecular recognition elements" (MoREs). Examples of MoREs bound to their partners can be found in the α-helix, β-strand, polyproline II helix, or irregular secondary structure conformations, and in various mixtures of the four structural forms. Here we describe an algorithm that identifies regions having propensities to become α-helix-forming molecular recognition elements (α-MoREs) based on a discriminant function that indicates such regions while giving a low false-positive error rate on a large collection of structured proteins. Application of this algorithm to databases of genomics and functionally annotated proteins indicates that α-MoREs are likely to play important roles protein−protein interactions involved in signaling events.

Protein−protein and protein−nucleic acid interactions are central to many processes in molecular biology. Through such interactions, translation is initiated (*1*) or terminated (*2*), apoptotic signals are stimulated (*3*) or inhibited (*4*), transcription is activated (*5*) or repressed (*6*), and a whole host of other cellular processes relying on recognition, regulation, and signaling are performed. Thus, understanding protein−protein and protein−nucleic acid interactions is critical for gaining insight into signaling and regulation within biological systems. Knowledge of these interactions might enable the development of small molecule therapies that could target a multitude of human diseases (*7*, *8*), thus highlighting the practical importance of understanding such interactions.

The ability to predict protein−protein interactions from sequence and structure would be very useful for guiding laboratory experiments. Both binding regions (*9*, *10*) and binding partners (*11*) can be predicted with some success from known protein structure, especially when structural knowledge is combined with evolutionary information (*9*−*11*).

Several well-characterized protein−nucleic acid and protein−protein interactions involve disorder-to-order transitions upon binding, which is also called coupled binding and folding (*12*−*23*). One or even both protein partners can be disordered prior to the interaction (reviewed in ref *24*). Naturally disordered proteins have also been called natively unfolded (*25*), intrinsically unstructured (*13*), and natively disordered (*26*). When a protein−protein interaction involves a natively unfolded partner, the methods developed for predicting protein−protein interactions based on known structures are simply inapplicable. For intrinsically disordered proteins, new methods and new approaches are needed.

The importance of predicting regions of disordered proteins that bind to partners of course depends on the commonness of such proteins. Several computational experiments indicate that natively unfolded or intrinsically disordered regions are a common phenomenon (*14*, *27*−*30*). For example, more than 15 000 out of 91 000 proteins in the then-current Swiss Protein database were identified as having long regions of intrinsic disorder (*27*) using PONDR[1] (*14*, *31*−*34*). The analogous conclusion has been made based on the results of disorder prediction for 31 genomes that span the three kingdoms. Using predictive methods it was shown that eukaryotes contain more disorder than either the prokaryotes or the archaea, with *Caenorhabditis elegans*, *Arabidopsis*

† This work was supported in part by NIH Grants R01 LM07688 (A.K.D.) and R43 GM06412 (V.N.U., Y.C.) and also in part by the Lilly Endowment via the Indiana Genomics Initiative.

* To whom correspondence should be addressed at Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, School of Medicine, Indiana University−Purdue University at Indianapolis, 714 N. Senate St., Suite 250, Indianapolis, IN 46202. Phone: 317-278-9650. Fax: 317-278-9217. E-mail: kedunker@iupui.edu.

‡ Molecular Kinetics Inc.

§ Present address: Biophysics Department, Institute for Molecular Virology 1525 Linden Drive, Madison, WI 53706.

|| School of Medicine, Indiana University−Purdue University at Indianapolis.

⊥ School of Informatics, Indiana University−Purdue University at Indianapolis.

Russian Academy of Sciences.

10.1021/bi050736e CCC: $30.25
Published on Web 08/24/2005

thaliana, *Saccharomyces cerevisiae*, and *Drosophila melanogaster* having 52−67% of their proteins with segments predicted to have ≥40 consecutive disordered residues, whereas bacteria and archaea were indicated to have 16−45% and 26−51% of their proteins with such long regions of predicted disorder, respectively (*29*). Finally, the application of a predictor combining two distinct approaches to elucidate proteins that are likely to be wholly disordered [the net charge−hydropathy distribution (*12*) and the distribution of disorder prediction score derived from PONDR VL-XT (*14*, *31*−*34*)] provided additional support for these observations (*30*). Using this consensus method it was shown that approximately 4.5% of *Yersinia pestis*, 5% of *Escherichia coli* K12, 6% of *Archaeoglobus fulgidus*, 8% of *Methanobacterium thermoautotrophicum*, 23% of *Arabidopsis thaliana*, and 28% of *Mus musculus* proteins are likely to be highly disordered in their entireties (*30*).

The increased prediction of disorder in eukaryotes compared with the other kingdoms has been suggested to be a consequence of the increased need for cell signaling and regulation (*14*, *29*, *35*−*37*). Laboratory experiments support this suggestion. In a collection of more than 100 experimentally characterized regions of disorder from a diverse set of proteins selected only because they contained regions of intrinsic disorder, a large majority of these examples were found by literature mining to participate in cell-signaling or regulation via noncatalytic interactions with DNA, RNA, or other proteins (*15*). The number of intrinsically disordered proteins or protein domains identified to be involved in cell-signaling, recognition, and regulation is growing rapidly. In addition, computational experiments have revealed that 79% ± 5% and 66% ± 6% of human cancer-associated proteins and signaling proteins, respectively, contain regions in which ≥30 consecutive residues were predicted to be disordered, whereas just 47% ± 4% and 13% ± 4% of general eukaryotic proteins and proteins with a known 3-D structure, respectively, contain regions with such long predictions of disorder (*36*). This study strongly implicated intrinsic disorder in cancer-associated and signaling proteins as compared to the two control sets. In the same study it was shown that proteins involved in metabolism, biosynthesis, and degradation together with kinases, inhibitors, transporters, G-protein coupled receptors, and membrane proteins were predicted to possess at least 2-fold less disorder than regulatory, cancer-associated, and cytoskeletal proteins (*36*). Based on these observations it was further suggested that intrinsically disordered proteins played key roles in cell-signaling, regulation, recognition, and cancer, where coupled folding and binding is a common mechanism (*36*).

Unlike well-structured binding proteins, intrinsically disordered binding proteins are inherently intractable to structure determination methods unless their binding partners are first identified and isolated. Once isolated, the structures of the disordered binding regions can be determined in combination with their structured partners using standard methods. For some disordered proteins, identification of the binding partners may simply be a matter of subjecting the protein to a binding screen against cellular extracts, but for many proteins the situation may not be that simple. Given the modular nature of proteins (*38*, *39*), the results of such screens may be complex. Also, if modules are connected by flexible linkers, the structure of the entire protein will still be difficult to obtain until the specific interaction regions of the binding partners are identified by deletion mutagenesis or by proteolytic digestion studies (*40*). In this sense, the in silico identification of binding regions within disordered proteins would be useful in accelerating the process of structure determination for complexes that contain disordered proteins.

Based on previous observations (*41*) and the results of this work, a specific structural element is proposed that mediates many of the binding events of disordered regions. This element consists of a short region (on the order of 20 residues) that undergoes a disorder-to-order transition that is stabilized by binding to its partner; this short region is within a segment of disorder. These molecular recognition elements, MoREs, are proposed to function in the recognition of protein or nucleic acid partners. The term "element" is applied to these regions to distinguish them from globular domains and from sequence-specific motifs. The proposed structural element is not constrained concerning the secondary structure in the bound state. Bound MoREs can adopt α-helix, β-structure, nonregular secondary structure, and even polyproline II helix conformations (Mohan, Romero, Uversky, and Dunker, manuscript in preparation). Here we focus on the α-helix-forming subset of MoREs, or α-MoREs. The reasons for this are practical: α-MoREs represent a sizable fraction of the MoREs studied to date (*42*), and concepts enabling their identification from sequence have been the most fully developed.

Based on examples derived from database and literature sources, an indicator of α-MoRE-containing regions is described herein. Results derived with this algorithm were already presented in one published study (*43*) with the identified α-MoRE being later confirmed by structure determination (*44*). We describe here the procedures employed for the generation of a data set of observed α-MoREs and the construction of an algorithm for indicating α-MoRE regions. The paucity of examples prevented the development of a formally trained predictor. Instead, our goal was to develop an algorithm that correctly identified the few known examples while yielding a low false positive error rate on a large collection of structured proteins. Next, we show the application of the algorithm to genomes and databases of functionally annotated proteins. Some interesting examples of α-MoRE indications are compared with biological data. The biological importance and implications of the MoRE concept are discussed.

MATERIALS AND METHODS

Data Sets

Examples of α-MoREs were derived from structures in the PDB (*45*) as of Nov 10, 2002. PDB Select 25 (*46*), which is a sequence nonredundant (≤25% identity) set of chains from the PDB, was used as a negative control in this work:

[1] Abbreviations: PONDR, predictor of natural disordered regions (PONDR is a registered trademark of Molecular Kinetics, Inc.); VL-XT, variously characterized long disordered regions in the internal regions training set, X-ray characterized terminal disordered regions in the end regions training set; MoRE, molecular recognition element; α-MoRE, α-helix-forming molecular recognition element; PDB, protein data bank; HCA, hydrophobic cluster analysis; MV, measles virus; GO, gene ontology.

By selecting chains longer than 30 residues, potential MoRE regions were explicitly eliminated from the data set. Therefore, any MoRE pattern in this set of structured proteins is false. PDB Select 25 sequences longer than 30 residues were used, which includes 1117 chains containing 220 668 residues.

Genomic data sets were obtained from the Entrez database at the National Center for Biotechnology (NCBI) web site (http://www.ncbi.nlm.nih.gov/entrez/). Functional data sets were derived from the SWISS-PROT database (*47*) as described previously (*36*). The data sets were composed of 9 eukaryotic genomes with 123 845 sequences and 51 799 104 total residues, 16 archaeal genomes with 37 128 sequences and 10 704 340 total residues, and 57 bacterial genomes with 162 692 sequences and 50 989 849 total residues. Additional details on the genomic data sets are available (Supporting Information). Correlations between MoRE patterns and protein function and cellular localization were evaluated using gene ontology (GO) annotations for budding yeast (*48*, *49*) following methods described previously (*50*).

Examples of α*-MoREs*

α-MoREs were extracted from PDB entries using the following procedure:

(1) Select chains in the PDB of 30 residues or shorter for which a valid reference sequence identification (i.e., Swiss-Prot or PIR ID) is given in the PDB entry, and which are bound to another protein chain longer than 30 residues.

(2) Retain all chains with helical content, as indicated by DSSP (*51*) classification. Discard all other chains.

(3) Discard all chains for which another chain longer than 30 residues in the structure is composed of a different region of the same reference sequence. That is, keep only chains bound to a *different molecule*.

(4) Verify remaining chains individually to make sure they are not fragments of a known globular domain. This is done by comparing the chain's reference sequence to a nonredundant set of protein sequences in PDB.

All chains selected through this procedure were examined in reference to the PDB file remarks, the coordinates of the structure, BLAST database searches, the reference publication, and additional literature. This detailed examination verified that each selected protein chain was bound to another protein and that each chain is consistent with the α-MoRE model; i.e., disordered in isolation. Due to the highly restrictive nature of this search method, these examples should not be viewed as an estimate of the number of α-MoREs in the PDB. These are only a set of examples of α-MoREs.

Construction of an Algorithm That Indicates α*-MoRE Propensities*

The algorithm indicating α-MoRE propensities was designed using a stacked algorithm technique, meaning that distinct algorithms are applied to an input in serial to obtain a final estimation. Here, three algorithms are used. The first algorithm is PONDR VL-XT, which provides a prediction of order/disorder for each residue. The second algorithm defines the location of a potential α-MoRE within a given protein sequence through identification of a particular pattern in the PONDR predictions, herein called a PONDR pattern. The third algorithm discriminates between locations that are similar to the example α-MoREs and locations that are similar to a collection of spurious patterns.

A generalized heuristic was developed to identify the locations of potential α-MoREs, such that PONDR prediction patterns were associated with all α-MoRE regions within the set of example α-MoREs. Generally, the PONDR pattern was a short predicted ordered region between two predicted disordered regions (see Results for details). Note that unless specified otherwise, PONDR disorder predictions are made at a threshold of 0.5, where disorder is indicated by a score greater than the threshold and structure is indicated by a score less than the threshold. For the PONDR pattern definition, however, the threshold was adjusted to improve the fit to the known examples. Since these PONDR prediction patterns are relatively common for many protein sequences, the third algorithm was required to discriminate between patterns associated with α-MoREs (true patterns) and spurious patterns (false patterns).

To discriminate between true and false patterns, parametric discriminant models were used. Specifically, sequence attributes associated with the patterns were used to fit quadratic multivatriate discrimination models. The various attributes were calculated from the sequence of each protein, which included estimated physiochemical properties, predicted disorder, and predicted secondary structure. Net charge, total charge, Kyte−Doolittle hydropathy (*52*), Vihinen flexibility (*53*), sequence conservation measured as Shannon's entropy (*54*), and Eisenberg's hydrophobic moment (*55*) were the protein attributes considered in the analysis. The disorder prediction parameters considered were PONDR VL-XT (*32*, *33*), VL-2, VL-2C, VL-2V, and VL-2S (*37*). For secondary structure prediction both PSIPRED (*56*) and GOR-I (*57*) were examined. Finally, hydrophobic cluster analysis, HCA (*58*), was implemented and was quantified by calculating the fraction of residues in a given region that were included in a hydrophobic cluster. Each of these quantities was averaged over three regions relative to the α-MoRE pattern: (a) the α-MoRE region; (b) flanking regions; and (c) both the α-MoRE region and the flanking regions. This provided a total of 66 attributes. A course window optimization was carried out for the definition of the flanking regions, where 30 residues from either side of the indicated α-MoRE region gave the best results.

The attributes were calculated for the set of true patterns from the example α-MoREs and the set of PONDR patterns from PDB Select 25 sequences longer than 30 residues. The attributes of true and false patterns were fit to separate quadratic discriminant models, which take the following form (*59*):

$$f_{m,x}(x;\mu_m,\Sigma_m) = \frac{\exp^{1}/_{2}(x-\mu_m)'\Sigma_m^{-1}(x-\mu_m)}{(2\pi)^{P/2}|\Sigma|^{1/2}}$$

where x is a vector of parameter values, $\boldsymbol{\mu}$ is a vector of parameter means, and $\boldsymbol{\Sigma}$ is the covariance matrix, where m denotes either the class of true or false examples. True and false MoRE patterns were used to estimate $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for both classes. A novel example was assigned a class based on which of the models, true or false, gave the largest value of the likelihood function.

Throughout this work, model accuracy was estimated by jack-knifing analysis. For this method of model assessment, a single example is held out, the model is calculated using the remaining data, and the class of the left-out example is tested. This procedure is repeated, holding out each of the examples once, and the estimated accuracy is taken as the average accuracy over each left-out example. This procedure gives a relatively unbiased estimation of the accuracy of a model (*59*). Overall accuracy was taken as the average of the sensitivity, i.e., the proportion of true patterns indicated correctly, and specificity, i.e., the proportion of false patterns indicated correctly.

Attributes for the final algorithm were selected incrementally using a greedy algorithm, despite that a greedy approach is not guaranteed to find an optimal solution for a quadratic discrimination model. However, the alternative of assessing all possible subsets of attributes (2^{66} possibilities) was computationally intractable. A suboptimal model was found by using a 5-fold greedy approach, which involves selecting the five best models at each round of attribute selection. For example, in the first round of selection, the five attributes that alone gave the best overall discrimination were carried into the next round, where all pairwise combinations of these five attributes with all other attributes were assessed. In the third round, all combinations of three attributes including the five best pairwise attributes were assessed, and so on. This selection process was carried out for at least nine rounds.

Correlations between High Pattern Propensities and Protein Function

To investigate the correlation between regions with α-MoRE propensities and protein functions, outputs were compared to GO annotations for budding yeast (*48*, *49*), using a procedure similar to that of Ward et al. (*50*). From the regions with α-MoRE propensities in budding yeast, the number of times each annotation was associated with an α-MoRE pattern was counted. This observed count was compared to a null model, which held that annotations are randomly associated with α-MoRE patterns. To generate the null model distribution, the same numbers of α-MoRE patterns were assigned randomly across all yeast proteins, where every valid sequence position in the data set had an equal probability of being assigned as an α-MoRE. The correspondence between random α-MoRE assignments and annotations was counted, and the procedure was repeated 10 000 times. The distribution of the null model for each annotation was modeled as a normal distribution and used to calculate a *z*-score for the observed counts. The *z*-score for an annotation was calculated as the observed count minus the mean of the null model, divided by the standard deviation of the null model. The same procedure was carried out for disorder predictions, except that each predicted disordered region, rather than each residue, was randomly distributed in the null model.

The architecture of GO annotation terms was used to remove dependencies in the results. Annotation terms are related in a directed acyclic graph (DAG), where child nodes become more specific further down the DAG. Furthermore, annotation terms follow the true path rule, which requires that all ancestors of an annotation must be consistent with that annotation. In terms of examining correlations with functional annotations, this structure provides a convenient way of removing dependent annotations, which is important since failure to consider data dependence can result in large reduction in the power of significance tests. Beginning with the annotation term with the largest absolute *z*-score, all ancestors and descendents of that term were removed from the results. This process was iterated, using the remaining annotation term with next highest absolute *z*-score, until the smallest remaining absolute *z*-score was reached. The significance of annotation *z*-scores was assessed using a step-down procedure, where a significance level of 5% was Bonferroni-adjusted in terms of the number of remaining annotation terms (*60*). Terms that failed this significance test were discarded.

RESULTS

PONDR VL-XT Indications of Binding Regions

Previous observations have shown that PONDR VL-XT gives indications of short binding regions within longer regions of intrinsic disorder (*15*, *41*). Based on two examples (*61*, *62*), a general subset of these indications appears to be short, α-helical regions involved in protein−protein interactions. One of these examples, 4E binding protein 1 (4E-BP1), was shown by NMR to be entirely disordered in solution (*2*, *63*), but a short, central region undergoes a disorder-to-order transition upon binding to eukaryotic translation initiation factor 4E (*62*). The other example is the autoinhibitory helix of calcineurin. Calcineurin is a calcium and calmodulin dependent protein phosphatase that is important in multiple signaling pathways. In terms of the sequence of the A subunit, the catalytic domain is followed successively by a disordered region of 95 residues, an autoinhibitory helix of 18 residues, and an uncharacterized region of 35 residues (*61*). The conformational preference of the isolated autoinhibitory helix is not known, but our hypothesis is that this region undergoes a disorder-to-order transition upon binding of the active site.

PONDR VL-XT predicts the entirety of 4E-BP1 and the calcineurin C-terminus to be completely disordered, except for short predictions of order corresponding to the experimentally determined disorder-to-order transition regions. For 4E-BP1 the predicted region of order corresponds very closely to the binding region, whereas for calcineurin the predicted ordered region is shifted slightly toward the amino terminus. On the basis of these two protein regions, more examples of short α-helical binding regions were mined from PDB.

α-*MoREs*

The selection process, described in Materials and Methods, resulted in a set of 13 proteins containing 15 potential α-MoREs. Two of these chains were removed upon manual inspection. The first was the antigenic region from an antibody−antigen complex, PDB code 2AP2 (*64*). This structure represents a type of interaction that is distinct from canonical α-MoRE mediated interactions. The second removed chain corresponded to a central portion of the human BAK protein (*65*) complexed with Bcl-xL, PDB code 1BXL (*66*). The authors of a paper describing this structure suggest that it may represent a conformational rearrangement of the full length structure of BAK (*66*), rather than a structurally

Table 1: Fourteen Examples of α-Helical Molecular Recognition Elements (α-MoREs) in Twelve Proteins

accession code	name	general function	MoRE Region start	MoRE Region end	PDB example	binding partner	function of binding example
APC_HUMAN	APC	tumor suppressor	2034	2049	1emu	axin	β-catenin phosphorylation complex formation
BAD_MOUSE	Bad	cell death promoter	140	164	1g5j	Bcl-x_L	inhibition of the anti-apototic activity of Bcl-x_L
FTSZ_ECOLI	FtsZ	cellular division	367	383	1f47	ZipA	stabilization of the septal ring
FLIM_ECOLI	FliM	flagellar motor switch component	1	16	1f4v	CheY	reversal of the direction of flagellar rotation
IPKA_MOUSE	PKI-α	inhibitor of cAMP-dependent protein kinase	5	24	1jbp	cAMP-dependent protein kinase	competitive inhibition of cAMP-dependent protein kinase
U2AF_HUMAN	$U2AF^{65}$	spliceosome assembly	85	112	1jmt	$U2AF^{35}$	formation of the U2AF recruitment complex
NCO2_MOUSE	NcoA-2	nuclear receptor coactivation	686	698	1l2i	estrogen receptor α	estrogen receptor α coactivation
P53_HUMAN	p53	tumor suppressor	13 367	29 388	1ycr 1dt7	Mdm2 S100B(ββ)	inhibition of p53 transactivation blocks activation of p53
MAD_HUMAN	Mad	transcription repression	8	20	1e91	SINB3	recruitment of the Sin3-histone deacetylase corepressor complex
gi|4758258	4E-BP1	inhibition of eIF4E	51	64	1ej4	eIF4E	blocks initiation of translation
gi|1906028	SRC-1	nuclear receptor coactivation	631 688	637 695	2prg	PPAR-γ	receptor coactivation
gi|2781188	calcineurin	protein phosphatase	469	486	1aui	calcineurin catalytic domain	competitive autoinhibition of the phosphatase's activity

independent interaction region. This view is supported by the BH1, BH2, BH3, and pseudo-BH4 domain homology of full length BAK (*67*), which suggests that this protein has an overall structure similar to Bcl-2. As a consequence of the selection criteria, many autoinhibitory regions were likely discarded. However, some autoinhibitory regions may satisfy the α-MoRE criteria. To include a representative of these interactions, the autoinhibitory helix from human calcineurin was added to the data set after selection.

Thus, the final set of α-MoRE examples contained 14 regions from 12 proteins (Table 1). Only a few of these examples have been shown to be disordered or occur in a disordered context in the absence of their binding partners [e.g., 4E-BP1 (*2*), calcineurin (*61*), and a region of p53 near the amino terminus (*68*)]. This means that some considered fragments may represent isolated helices from globular domains. However, all regions are consistent with the model of α-MoRE regions: none are known to be part of a structured domain, and all mediate cell signaling events. Support for this working assumption is given by the fact that the unverified regions have all been shown to be minimal binding domains [p53 (*69*, *70*), FliM (*71*), FtsZ (*72*), Mad1 (*6*), Bad (*4*), U2AF65 (*73*), PKI (*74*), APC (*75*), SRC-1 and NcoA-1 (*76*)]. This demonstrates that these regions can exhibit function without globular structure.

Functions carried out by these α-MoREs are diverse, including nuclear receptor activation, complex stabilization, complex formation, inhibition of complex formation, and enzyme inhibition. The wide range of functional and sequence diversity of the α-MoREs listed in Table 1 precluded traditional motif-prediction approaches, which rely primarily on sequence similarity for detection. To test the supposition that α-MoREs represent a distinct structural element or feature, a preliminary algorithm for indicating α-MoRE patterns was developed based on a distinctive signature observable in PONDR VL-XT predictions indicative of binding regions.

PONDR Pattern

The set of α-MoRE examples listed in Table 1 represents a diverse set of functions mediated through a common structural feature. Thus, finding regions with α-MoRE patterns was approached as a problem of structural classification, rather than a sequence-based motif. The attributes and methods examined for the identification of α-MoRE patterns from amino acid sequence have the effect of abstracting the underlying sequence, thereby making it possible to bring these diverse sequences into a common classification.

The basis of the α-MoRE pattern identification algorithm is an elaboration of the previously reported PONDR VL-XT binding region prediction signature (*41*). Based on examination of the PONDR VL-XT predictions of the 14 α-MoRE examples, the following generalization was established as the α-MoRE PONDR pattern: (1) A predicted ordered region less than 60 residues in length, at a threshold of 0.6 PONDR score. (2) This short prediction of order is flanked by two predicted disordered regions or by one predicted disordered region and the amino or carboxy terminus. (3) Subsequently, a modest requirement was added that necessitated at least one flanking disordered region be at least 20 residues in length, which reduced significantly the number of false patterns found in PDB Select 25 (data not shown). Based on the mean length of the α-MoRE examples, the α-MoRE region was assigned to a 19 residue window, centered on the middle residue of the predicted ordered region. This pattern gave indicated α-MoRE regions that agreed well with the examples (see Table 2), although none precisely matched the experimentally determined boundaries. For the purposes of this study, α-MoRE patterms were considered correct if an α-MoRE pattern overlapped a known α-MoRE region and incorrect if the α-MoRE pattern occurred in a region that has been structurally characterized as a nonhelical MoRE region.

Table 2: Comparison of Defined α-MoRE Positions and Indicated α-MoRE Patterns

name	α-MoRE region	indicated α-MoRE pattern
APC	2034−2049	2031−2049
Bad	140−164	144−162
FtsZ	367−383	353−371
FliM	1−16	1−19
PKI-α	5−24	14−32
U2AF^{65}	85−112	92−110
NcoA-2	686−698	681−699
p53	13−29	17−35
	367−388	374−392
Mad	8−20	1−19
4E-BP1	51−64	45−63
SRC-1	628−640	627−645
	685−703	681−699
calcineurin	469−486	456−474

PONDR patterns found in PDB Select 25 were classified as false α-MoRE patterns. The pattern occurred 863 times in this data set, at a rate of 3.69×10^{-3} patterns per residue. The true α-MoRE PONDR patterns, corresponding to example α-MoRE regions, and the 863 false α-MoRE patterns were compared based on the set of attributes described in the Materials and Methods section. Of the 66 parameters examined, 46 parameters showed significantly different means that could potentially discriminate between the true and false α-MoRE patterns at a 95% confidence interval (data not shown), suggesting that discriminant analysis was an appropriate technique for this classification problem.

The differences observed were also consistent with our model of α-MoRE regions. For example, the disordered regions that flank the true α-MoRE binding regions had higher average disorder predictions, lower fractions of predicted α-helix and β-sheet, and a lower HCA fraction than the flanking regions of the false α-MoRE patterns derived from structured proteins. For the internal predicted-to-be-ordered binding regions of the true and false α-MoRE patterns, most parameters showed no significant differences for most of the sequence features. Three of the five disorder predictors gave higher values for the true binding regions, and the true binding regions also had a larger flexibility index, slightly larger loop content, and less hydrophobicity than the false α-MoRE patterns. Differences in whole α-MoRE regions were qualitatively the same as flanking region differences, due to the large contributions from the flanking regions.

Algorithm for Identifying Regions of High α-MoRE Propensity

Discriminant analysis was used to model the difference between true α-MoRE PONDR patterns and false α-MoRE PONDR patterns. Attribute selection was performed as described in Materials and Methods to find the best performing set of attributes. Though attribute combinations including PSIPRED produced slightly better results, the margin of this improvement for peak accuracy over attribute combinations without PSIPRED predictions, ~0.5% (data not shown), did not justify the additional computational intensity of this algorithm, relative to GOR-I. Table 3 summarizes the attribute selection process for the final model. Only two parameters were necessary to differentiate all true patterns from false patterns, but 74, or 8% of false patterns, were still classified as true. Addition of subsequent parameters reduced the misclassification of false patterns. The peak accuracy was observed at six parameters, which classified all true patterns correctly and misclassified <2% of false patterns. Model accuracy degraded after addition of a seventh parameter; therefore the first through the sixth parameters were used in the final model. Figure 1 shows the distribution of values for these parameters for the true and false patterns, relative to the overall balanced mean of the two groups. The distributions suggest that the assumption of normality of parameters is roughly satisfied. The accuracy of the final model was 99%, with 100% correct classification of true patterns (i.e., sensitivity) and 98% correct classification of

Table 3: Summary of the Accuracy of the Discriminant Model versus the Number and Identity of Attributes

algorithm	true α-MoRE patterns	false α-MoRE patterns	false α-MoRE patterns/residue ($\times 10^{-3}$)
PONDR pattern	14	863	3.689
discriminant analysis			
1st parameter: flanking VL-2 value	13	113	0.483
2nd parameter: flanking fraction of predicted turn	14	74	0.316
3rd parameter: MoRE fraction of predicted loop	14	40	0.171
4th parameter: flanking HC fraction	14	29	0.124
5th parameter: whole hydrophobic moment	14	20	0.086
6th parameter: flanking VL-XT value	14	14	0.060
7th parameter: whole fraction of predicted sheet	13	15	0.064
8th parameter: flanking VL-2S value	11	9	0.039
9th parameter: MoRE sequence entropy	8	5	0.021

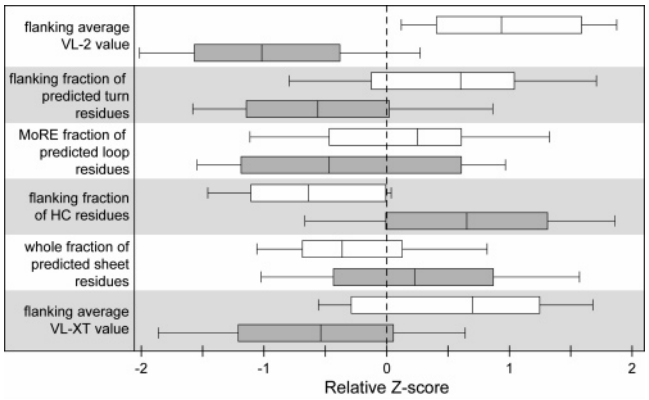

FIGURE 1: Bar plot of attributes used in the final α-MoRE model. The range of *z*-scores for each attribute and for both pattern types—true patterns (white bars) and false patterns (gray bars)—is shown for each example. *z*-Scores are translated so that 0 corresponds to the mean of the positive pattern mean and negative pattern mean. The median (center line), 25% and 75% quantiles (thick bar), and 10% and 90% quantiles (whiskers) of *z*-scores are shown.

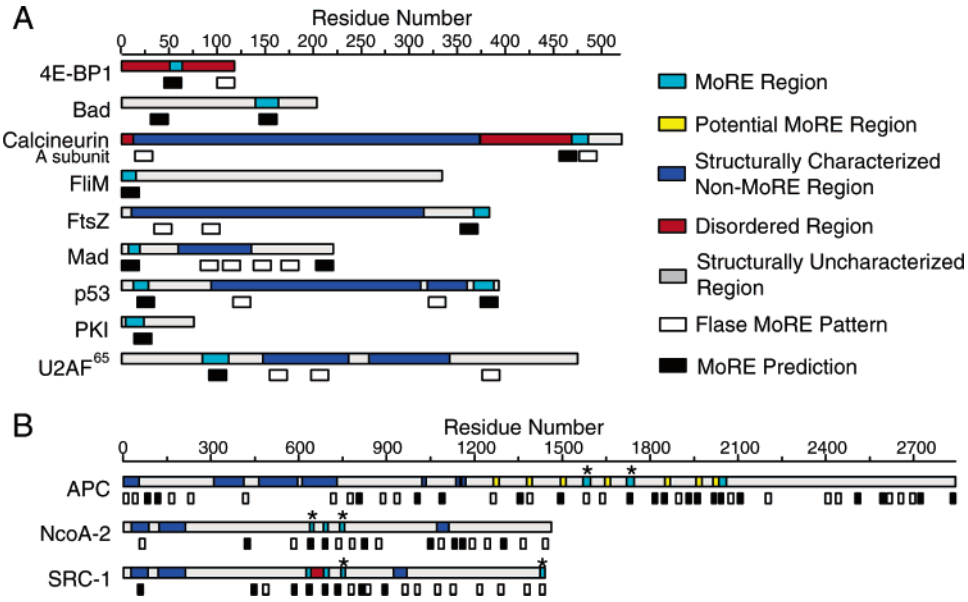

FIGURE 2: Bar representation of the 12 example proteins that contain α-MoREs. Two rows of bars are displayed for each protein, where the name of the protein is given to the left of each pair of rows. The upper bar represents characterized regions of each sequence: MoRE regions, disordered regions, and structurally characterized non-MoRE regions. The lower bar represents PONDR MoRE patterns indicated to be false by discriminant analysis and patterns indicated to be true by discriminant analysis. Proteins are divided into those with (A) sequences shorter than 600 residues and (B) sequences longer than 600 residues, for better clarity of the former.

false patterns (i.e., specificity). This high accuracy is probably due in significant degree to the small number of α-MoRE examples that are currently known; as the set of characterized α-MoREs increases, the accuracy of our algorithm is very likely to fall dramatically.

α-MoRE propensities for the 12 example proteins examined are illustrated in the contexts of the entire proteins in Figure 2. Most verified α-MoREs regions are predicted with good region agreement, and all regions were predicted by a true pattern that overlaps the experimentally derived region to some extent, as shown in Table 3. By the criteria used here, the predictions in Figure 2A show perfect agreement with the characterized regions of these proteins. In fact, all α-MoRE regions were correctly identified and no α-MoREs were indicated in structurally characterized non-MoRE regions. Two α-MoREs were suggested in uncharacterized regions of two proteins: residues 31−49 of Bad and residues 203−221 of Mad.

The α-MoREs in Figure 2B are also all correctly indicated. The six α-MoREs marked by an asterisk (*) in Figure 2B are regions that have not been verified to be α-MoREs; however, these regions contain a repeated sequence that is also present in the verified α-MoRE example region of that protein. For NcoA-2 and SRC-1, these regions have also been shown to bind to the same partner as the verified α-MoRE (*76*). All six of these putative α-MoREs are associated with α-MoRE patterns, but only three of them were indicated to contain such regions. The only verifiably false pattern for all 12 of the α-MoRE containing proteins is in the amino terminal helix−loop−helix domain of SRC-1. APC, NcoA-2, and SCR-1 contain additional regions with high α-MoRE propensities in uncharacterized regions of the sequences.

The seven potential α-MoREs in APC, shown in Figure 2B, are 20-residue repeats of a consensus sequence. The consensus sequence alone is capable of binding a partner, *β*-catenin (*77*). The structure of a distinct 15-residue repeat sequence, of which there are three in the APC sequence, has been structurally characterized to bind in an extended conformation (*78*). By the definition of MoREs, these regions qualify as extended MoREs, but are not identified to have high α-MoRE propensities. The 15-residue repeats lack three conserved serines, critical for the binding activity of the 20-residue repeats (*77*). This suggests that the 20-residue repeat sequence binds *β*-catenin by alternative means or at reduced affinity or specificity. All seven of these repeats are associated with α-MoRE patterns and four are indicated to be α-MoREs, which suggests that these regions may be α-MoREs.

From a PDB-derived data set of 1117 chains and 220 668 residues, the current algorithm gave only 14 false positive α-MoREs. These putative errors were from 12 different chains from nine separate structures. Half of these false patterns occurred in nucleic acid bound proteins, five of which are from the same ribosome structure, 1FFK. Nucleic acid binding proteins can be disordered in the absence of nucleic acid (*13*), which may explain the presence of α-MoRE patterns in these chains. Three of the remaining false patterns occurred near the amino terminus with varying amounts of missing density between the α-MoRE pattern and the true chain terminus: 1MAI with a MoRE pattern at residues 5−23 is missing residues 1−11, 1AGQ D with an α-MoRE pattern at residues 52−70 is missing residues 1−40, and 1D8E A with an α-MoRE pattern at residues 57−75 is missing residues 1−54. These missing density regions are likely disordered regions, and the patterns are likely misclassified as true patterns because the amino terminal regions resemble those of true α-MoRE patterns. The remaining three patterns occur in a surface exposed helix (residues 1−19 of 1QH8 A), in a surface exposed loop (residues 135−153 of 1DO0 A), and in a region involved in subunit interactions (residues 63−81 of 1NPO C). No explanation of these errors is apparent from the respective structures; they are likely due to uncertainty in model parameter estimation.

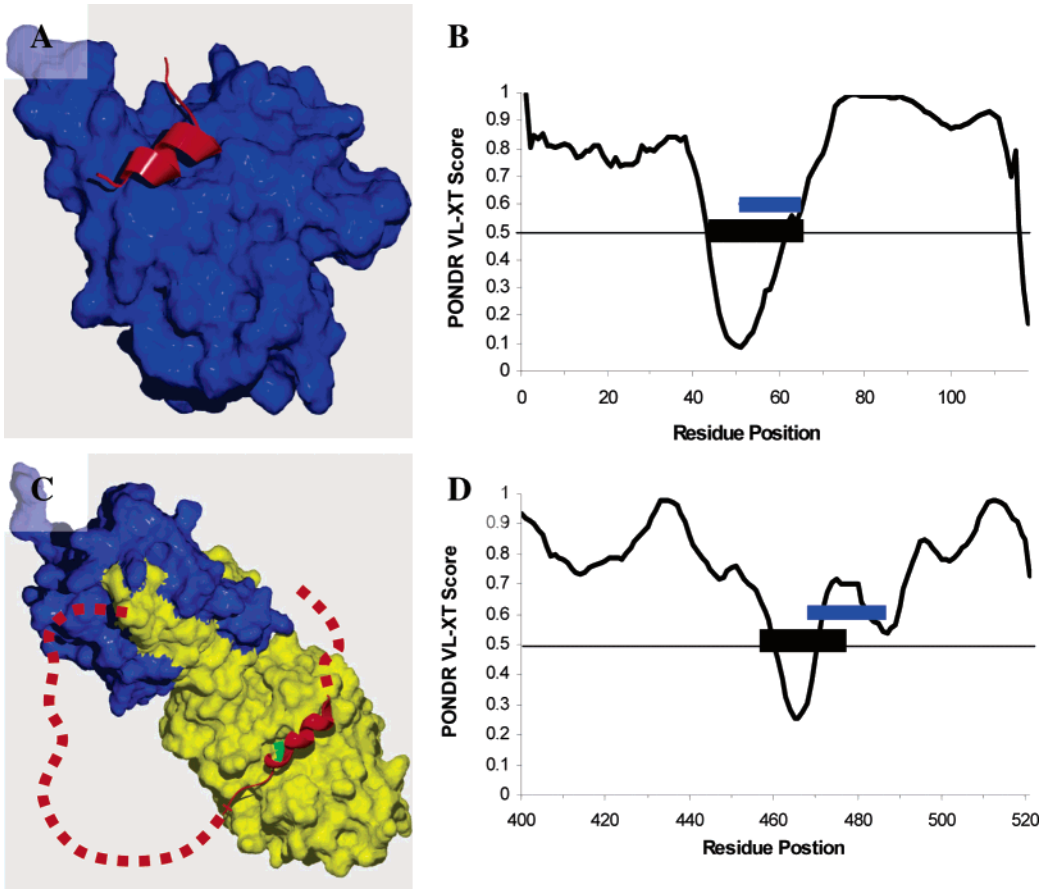

FIGURE 3: Examples of binding regions and their positions relative to the regions of predicted order (PONDR VL-XT score) and α-MoREs. (A) Eukaryotic initiation factor (blue) and the binding region of 4E-BP1 (red). (B) The PONDR VL-XT prediction for 4E-BP1 with the binding region (blue bar) and the α-MoRE pattern region (black bar) shown. (C) The B (blue) and A (yellow) subunits of calcineurin and the autoinhibitory region of the A subunit (red helix) in the midst of observed disordered sequence (red dashes). (D) The PONDR VL-XT prediction for the last 121 amino acid residues of the A subunit with the autoinhibitory region (blue bar) and the α-MoRE pattern region (black bar) indicated.

α-*MoREs: Illustrative Examples*

It is helpful to observe the relationships among the PONDR VL-XT predictions, the α-MoRE patterns, and the resulting three-dimensional structures. Four illustrative examples are shown in Figures 3−5. Each of these examples will be discussed briefly.

4E Binding Protein 1. As it has been already pointed out, 4E binding protein 1 (4E-BP1), being entirely disordered in solution (*2*, *63*), contains a short region that is able to undergo a disorder-to-order transition upon interaction with the binding partner, eukaryotic translation initiation factor 4E (*62*). The structure of this complex is shown in Figure 3A, with the results of PONDR VL-XT and the indicated α-MoRE patterns shown in Figure 3B. Note that there is a sharp dip in the PONDR score in an area overlapping with the experimentally established binding region. This dip is flanked by extended fragments of predicted disorder and corresponds to the disorder-to-order transition region. It is this pattern that caught our attention and was used as a basis for the development of our algorithm for identifying α-MoRE patterns.

Autoinhibitory Helix of Calcineurin. The other example of an α-MoRE is the autoinhibitory helix of calcineurin, which is a Ca^{2+} and calmodulin dependent protein phosphatase and which plays crucial roles in several signaling pathways. A representation of the crystal structure of this complex is shown in Figure 3C. The A subunit's catalytic domain is followed by a stretch of 95 residues invisible in

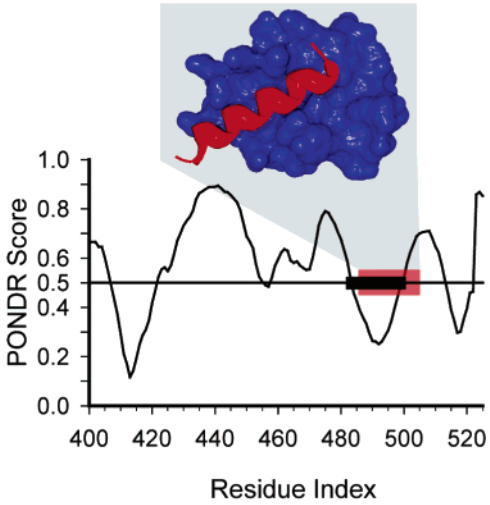

FIGURE 4: PONDRing measles virus protein N: The order/disorder tendencies of N as revealed by PONDR VL-XT. The C-terminal end of the protein is predicted to be disordered except for a small ordered region shown by a black bar. The identified α-MoRE pattern is shown by a red bar (*43*). Experimentally established fragment undergoing disorder-to-α-helix transition (*44*) upon protein P binding is also shown by the black bar. Note, the binding partner of this α-MoRE is known to be the measles virus protein P, which is shown in blue.

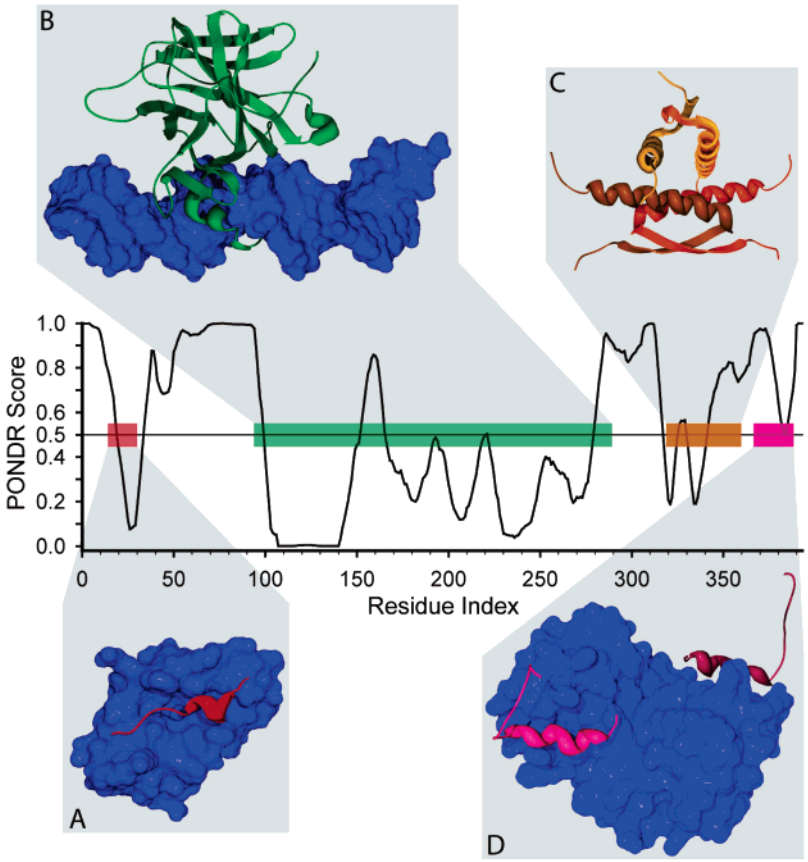

FIGURE 5: PONDRing p53: The order/disorder tendencies of p53 as revealed by PONDR VL-XT show interesting correlations with the well-understood domains of this molecule. These correlations are as indicated in order from amino terminal to carboxyl terminal along the sequence correlating with a clockwise arrangement for the molecular structures: (A) a downward spike in the plot roughly matches the transcription activation domain that binds to Mdm2; (B) a long prediction of mostly ordered structure matches the domain that binds to DNA; (C) a sharp downward spike doublet overlaps the tetramerization domain; and (D) a sharp downward spike matches the negative regulatory domain that binds to the S100B($\beta\beta$) dimer. The binding partners (Mdm2, DNA, and S100B($\beta\beta$) dimer) are given in blue, and the structured domains in p53 are color-coded to bars that indicate their locations along the sequence.

the X-ray crystal structure, a short (18 residue) helix spanning the autoinhibitory domain, and another stretch of 35 residues missing from the X-ray crystal (*61*). The PONDR VL-XT prediction of most of this region is shown in Figure 3D. Importantly, in this case the region of predicted order also overlaps with the autoinhibitory helix. Overall, Figure 3 shows that PONDR VL-XT predicts the entirety of 4E-BP1 and the C-terminus of calcineurin to be completely disordered, except for short predictions of order, which correspond to the disorder-to-order transition regions. Furthermore, for 4E-BP1 the predicted region of order corresponds very closely to the binding region, whereas for calcineurin the predicted ordered region is shifted slightly toward the amino terminus.

Measles Virus Nucleoprotein N. As it has been pointed out in the introduction, the prerelease version of the α-MoRE pattern-finding algorithm had already been successfully applied to elucidate the molecular mechanism underlying the binding of protein N to its partner, protein P in measles virus (*43*). The application of our algorithm indicated an α-MoRE region located at residues 488−499 of protein N (Figure 4). The involvement of the suggested α-MoRE in interaction with the protein P was confirmed experimentally, as a truncated form of N_{TAIL} (amino acid residues 401−488) lacking the suggested α-MoRE region was shown to lose ability to bind to protein P (*43*).

More recent NMR and X-ray crystallography data are in excellent agreement with the indicated α-MoRE pattern, as they show that $N_{487-503}$ binds as a helix to the surface created by the second (α2) and third (α3) helices of $P_{457-507}$, in an orientation parallel to helix α3, thereby creating a four-helix bundle (*44*). Furthermore, it was established that the binding interface is tightly packed and dominated by hydrophobic amino acids, confirming that binding and folding of $N_{487-503}$ are coupled (*44*). Importantly, these direct experimental data are in excellent agreement with the results of our computational analysis of the N_{TAIL} sequence. In fact, the NMR and X-ray results considered above show that the 487−503 fragment of protein N, being substantially disordered when free in solution, binds as a helix to the surface of $P_{457-507}$. This foldable $N_{487-503}$ fragment almost exactly coincides with the 488−499 region of N indicated to have an α-MoRE pattern (see Figure 4). Thus, these studies illustrate the usefulness of the application of the disorder prediction for the functional analysis of intrinsically disordered proteins.

The Tumor Suppressor p53. The p53 tumor suppressor is one of the most extensively studied signaling proteins. This protein is involved in binding more than 40 protein partners as well as DNA (*79*). Figure 5 presents the results of the PONDR VL-XT order/disorder prediction for p53 together with two regions with α-MoRE patterns visualized by red and magenta bars. Figure 5 shows that these α-MoRE

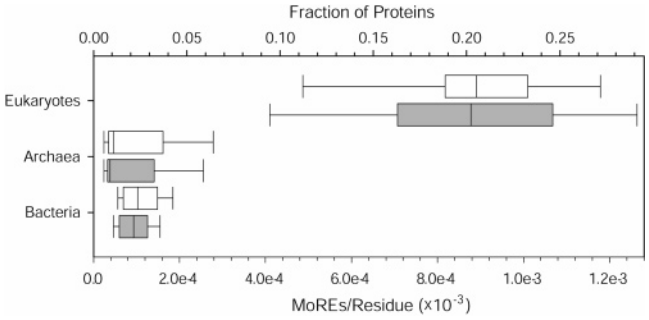

FIGURE 6: Bar plot of helical MoRE patterns across the three domains of life. Both the fraction of proteins found to contain α-MoRE patterns and the rate of α-MoRE patterns per unit length of sequence are plotted. Bars represent the range of results obtained for each of 9 eukaryotic, 16 archaeal, and 57 bacterial genomes, where the median (center line), 25% and 75% quantiles (thick bar), and 10% and 90% quantiles (whiskers) are shown.

patterns coincide with downward spikes (indicating order within a disordered region) in the PONDR VL-XT plot. Importantly, the first α-MoRE pattern is located at residues 17−35, which strongly overlaps with the fragment of p53 known to be involved in Mdm2 binding and which undergoes a disorder-to-α-helix transition coincident with this interaction (residues 13−29) (*70*). The second α-MoRE pattern (residues 374−392) is located in the C-terminal negative regulatory domain. This α-MoRE pattern overlaps with the region 367−388, which is known to fold upon binding to S100B(*ββ*) (*80*). Both of the α-MoRE regions in p53 were included in the training set of the algorithm for finding α-MoRE patterns and so do not represent novel findings. However, these regions are exemplary of the functional diversity that can be mediated by α-MoREs.

α-*MoRE Propensities Across Genomes*

The α-MoRE propensity algorithm was applied to known and putative proteins from 82 sequenced genomes to estimate the prevalence of regions having α-MoRE propensities. A summary of α-MoRE propensities across each domain of life is given in Figure 6. To control for the much longer average length of eukaryotic proteins compared to bacterial or archaeal proteins, both the proportion of proteins with indicated α-MoRE patterns and the rate of α-MoRE patterns per length of sequence are shown for each of the three domains (see Supporting Information for full results). The algorithm estimates that the median eukaryotic genome has a greater than 18-fold higher fraction of proteins with α-MoRE propensities than the median archaeal genome and a greater than 8-fold higher fraction of proteins with α-MoREs than the median bacterial genome. In terms of the rate of α-MoRE indications, α-MoREs are suggested to occur with greater than 9-fold higher frequency in the median eukaryote than in the median bacteria or archaea. All eukaryotic genomes show higher frequencies of α-MoRE indications than all bacterial and archaeal genomes (Supporting Information).

Table 4 was compiled to place these propensities in the context of the error rate of the algorithm and examine which tier of the algorithm, the PONDR patterns (that is, dips in the PONDR VL-XT plots flanked by predictions of disorder) or the α-MoRE patterns filtered by the discriminant analysis model, is responsible for differences in α-MoRE pattern frequencies among different organisms. α-MoRE patterns occurred 1.6 to 1.2 times more frequently in the genomic data sets than in PDB Select 25, with α-MoRE patterns most frequent in eukaryotic proteins. Eukaryotes also have the largest proportion of PONDR patterns that are suggested to be α-MoREs, 16.6%, which is 10-, 9-, and 7- fold greater than the proportion of PONDR patterns indicated to be α-MoREs in PDB Select 25, bacteria, and archaea, respectively. The higher rate of occurrence of regions with α-MoRE propensities in eukaryotes is due primarily to a higher proportion of putatively true α-MoRE patterns as judged by the discriminant analysis model, and to a lesser extent, due to an increase in the frequency of PONDR patterns.

Table 4: Comparison of the PONDR Pattern Rate and α-MoRE Pattern Rate in the Three Kingdoms and PDB

data set	PONDR pattern/residue ($\times 10^{-3}$)	α-MoRE pattern/residue ($\times 10^{-3}$)
PDB Select 25	3.698	0.060
eukaryotes	5.834	0.971
archaea	5.067	0.098
bacteria	4.479	0.107

The frequency of α-MoRE patterns in bacteria and archaea is only 1.6- and 1.8-fold greater than the frequency of PONDR patterns. Although this indicates that α-MoRE indications in these kingdoms are likely to contain a higher proportion of false patterns, it does not necessarily mean that the majority of α-MoRE patterns are false. The error rate estimate given by the PDB Select 25 α-MoRE pattern rate is conservative. That is, this data set consists exclusively of structured proteins. The complement of all proteins encoded by the genomes of individual organisms, however, is likely to contain a significant proportion of disordered residues (*29*). Therefore, this error rate should be viewed as a worst case—if all proteins in a genome are completely structured.

Correlations of α-*MoRE Patterns with Protein Function and Cellular Organization*

The role of α-MoREs in biological function was investigated by finding GO annotations for budding yeast that correspond to indicated α-MoREs. Patterns were compared to each of the three branches of GO annotations: molecular function (Figure 7), biological process (Figure 8), and cellular component (Figure 9). The correlations were quantified by counting the number of times a region of high α-MoRE propensity occurred in a protein with a given annotation. To test for significant correlations between patterns and annotation, a null model was used that assumed a random association of patterms and annotations. Under the null model, the α-MoRE pattern frequencies in budding yeast were redistributed at random, a process that was repeated to obtain a mean and standard deviation for the number of associations that occur by chance under the null model. The observed associations are given in terms of the number of standard deviations (*z*-score) the observed value was above (positive) or below (negative) the mean of the number of associations observed by chance. Also, the results from a similar analysis for PONDR VL-XT alone are also given for comparison.

These correlations do not suggest functions for the suggested α-MoREs, but rather they indicate the functions of proteins that contain regions of α-MoRE propensity. For

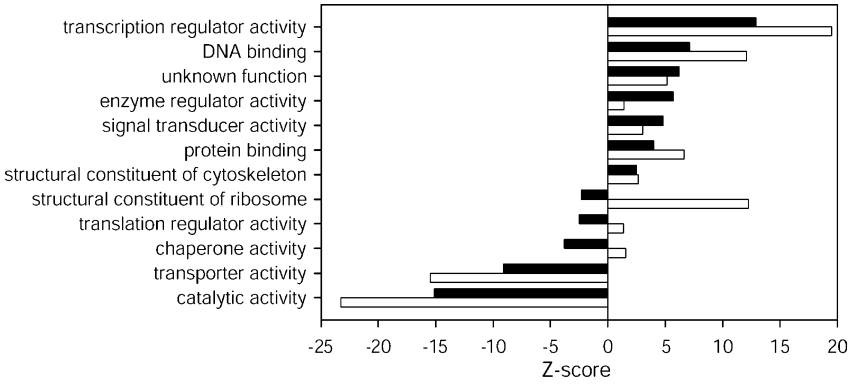

FIGURE 7: Correlations of α-MoRE patterns (black bars) and PONDR VL-XT predictions (open bars) with the molecular function ontology. Ontology entries with more or fewer α-MoRE patterns than expected by chance have a positive or negative *z*-score, respectively. See Materials and Methods for details.

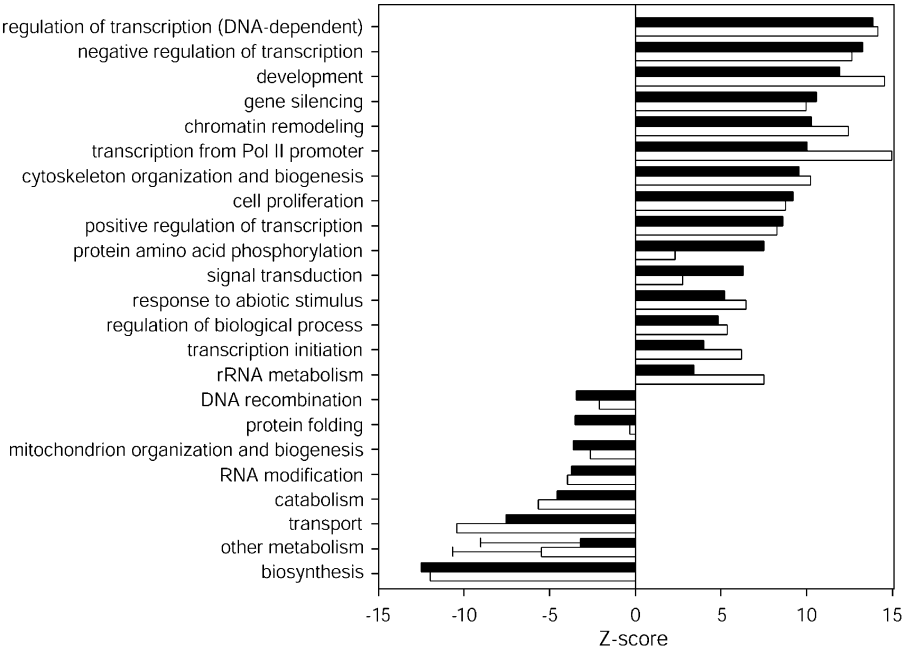

FIGURE 8: Correlations of α-MoRE patterns (black bars) and PONDR VL-XT predictions (open bars) with the biological process ontology. Ontology entries with more or fewer α-MoRE patterns than expected by chance have a positive or negative *z*-score, respectively. See Materials and Methods for details. The "other metabolism" figure entry contains all specific metabolism-associated proteins that are negatively correlated with α-MoRE patterns and with PONDR VL-XT disorder scores, where the range of the *z*-scores is given by the error bar. The full figure is given in the Supporting Information.

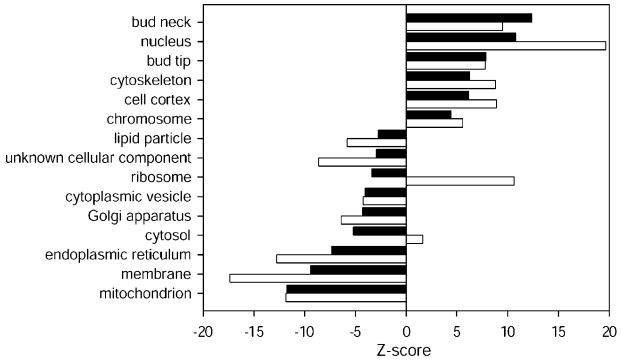

FIGURE 9: Correlations of α-MoRE patterns (black bars) and PONDR VL-XT predictions (open bars) with the cellular component ontology. Ontology entries with more or fewer α-MoRE patterns than expected by chance have a positive or negative *z*-score, respectively. See Materials and Methods for details.

example, the strong positive correlation between α-MoRE patterns and DNA binding proteins (Figure 7) does not necessarily implicate α-MoREs in mediating interactions with DNA, although such is possible. The correlation indicates only that proteins that bind DNA are likely to contain α-MoREs. The protein p53, discussed above, is an example of a protein that contains both a DNA binding domain and α-MoREs, but the α-MoREs are involved in protein−protein interactions, not protein−DNA interactions.

The molecular function ontology contains annotations detailing the physical activities of genes on the molecular level. The correlations between indicated α-MoREs and molecular function annotations (Figure 7) suggest many functional associations that are consistent with the α-MoRE hypothesis. Most directly, α-MoREs are indicated in proteins with signaling functions, such as protein binding, signal transduction, enzyme regulation, transcription regulation, and DNA binding, at a much higher rate then expected at random. Also, regions with high α-MoRE propensities occur infrequently in proteins with catalytic activity, which is consistent with but not required by the proposed model for α-MoREs.

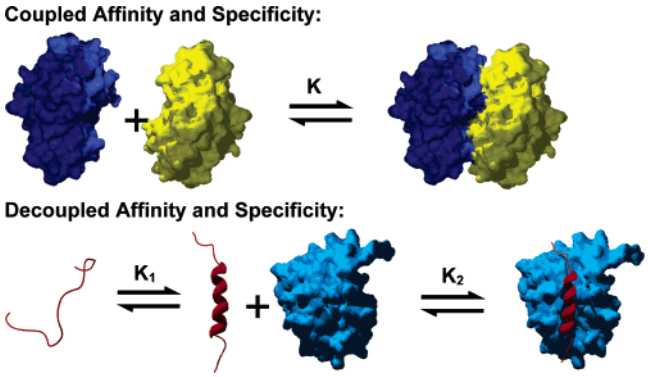

FIGURE 10: Conceptual diagram of coupled and decoupled specificity and affinity. (Upper) For the binding of two ordered domains, affinity and specificity are expected to be proportional. (Lower) For proteins that undergo disorder-to-order transitions when binding, such as α-MoREs, specificity and affinity are relatively uncoupled by the required ordering that is coupled to binding. The structure on the bottom right is the structure of the Bad α-MoRE bound to Bcl-xL.

Pathways and other high-order processes are annotated with the biological process ontology, and the correlations of these annotations to α-MoRE patterns (Figure 8) are consistent with correlations found with the molecular function ontology. In particular, α-MoRE patterns are frequently indicated in regulatory and signal transduction proteins, as well many vital, high-level systems, including development and proliferation. Also, α-MoREs are indicated infrequently in proteins involved in catabolic and most metabolic pathways.

From the cellular component ontology analysis (Figure 9) it is apparent that proteins likely to contain α-MoREs in yeast are largely localized to intracellular structural elements, the nucleus, and reproductive structures. α-MoREs are underrepresented in other general areas of the cell, such as the cytosol and cellular membranes.

In general, PONDR VL-XT predictions show the same correlations with GO annotations as indications of α-MoRE patterns, which is expected since the α-MoRE algorithm is based on PONDR VL-XT. Notable exceptions are correlations with ribosomal proteins, for which PONDR VL-XT indicates that many are disordered while very few are indicated to contain α-MoREs (Figures 7 and 9). Many ribosomal proteins are predicted to be disordered by PONDR VL-XT and have been shown to be disordered in isolation (*81*, *82*). However, it seems unlikely that α-MoREs would be present in these proteins, since they interact primarily with RNA. The α-MoRE patterns support this idea and demonstrate that the existence of α-MoRE patterns is not always directly related to predictions of disorder by PONDR VL-XT.

Advantages of α-MoREs for Cell Signaling

The frequency of α-MoREs in various types of proteins is highest in those associated with signaling and lowest in the metabolic enzymes. Evidently, these elements have advantages for cell signaling. The advantages of such a simple element in signaling are, among others, the decoupling of specificity/affinity, as illustrated in Figure 10. For the interaction of two globular domains, specificity and affinity are typically coupled, because as binding specificity is increased by additional contacts, the free energy of binding increases as the sum total of these contacts (*83*). This type of interaction can be viewed as a single equilibrium process (K). However, for an isolated helix, as for other disorder-to-order transition regions, entropy is a significant factor in the free energy of binding. In reference to Figure 7 the α-helical propensity (i.e., the equilibrium constant of the coil−helix transition, K_1) of the α-MoRE region determines the magnitude of the unfavorable free energy in helix formation, but this equilibrium does not affect binding specificity. After helical ordering, binding is analogous to the interaction of two globular domains, with coupled affinity and specificity (equilibrium constants K and K_2). This interaction can be viewed as a double equilibrium process, where the first transition affects only interaction affinity and the second reaction affects both specificity and affinity. These separate processes allow for specificity and affinity to be largely decoupled. In short, the decoupling of specificity and affinity provides a mechanism by which the strength and duration of signaling events can evolve separately.

DISCUSSION

The High Percentage of α-MoRE Patterns

Prediction methods rely on the extent to which training examples are representative of the problem domain. The small number of α-MoRE examples used here argues against the representative nature of this set due to its insufficient size. Because of this limitation, we have not attempted to develop a formal predictor, but rather our goal has been to develop an algorithm that identifies sequence patterns similar to those of the small set while attempting to reduce the false positive identification rate. Furthermore, all of the training examples are correctly identified, suggesting the strong likelihood of overfitting. If a true α-MoRE has properties that differ even slightly from those of the training set used here, then such a binding element would likely be missed by the current algorithm. One example in which our algorithm misses an α-MoRE is the enolase binding segment of the organizing domain of RNase E (*84*). This segment forms an α-helix upon binding to enolase (Luisi, personal communication) and so is a true α-MoRE. This segment is missed by our algorithm despite a strong dip in the PONDR VL-XT plot flanked by long predictions of disorder (*84*). To further test the possibility that true α-MoREs are frequently missed, we constructed sequence homologue families with high sequence identity to known α-MoREs and applied our algorithm to these homologues. In many cases the α-MoRE patterns were not indicated by our algorithm despite similar-appearing dips in the PONDR VL-XT plots (work in progress). The presence and absence of α-MoRE patterns in very similar sequences indicates that the current algorithm is indeed overly sensitive (i.e., overfitted to the training data).

The criteria for the indication of an α-MoRE was deliberately set to be highly restrictive, in order to give low false positive error rates. As discussed in the results, just 14 false positive pattern identifications were made on a nonredundant set of 1117 chains and 220 668 residues. This false positive error rate is estimated from ordered protein. Currently, there is simply no information regarding the false positive error rate for α-MoRE pattern indications on regions

of disorder. An α-MoRE pattern within a disordered region might or might not be the binding target for a protein or nucleic acid partner.

Despite all the limitations discussed above, application of this first-generation algorithm for identifying α-MoRE patterns indicates that associations like the very well studied example involving p53 and Mdm2 might be an extremely common signaling mechanism in eukaryotic cells. As much as 20% of all eukaryotic proteins (Figure 6) and nearly 50% of signaling and regulatory proteins (data not shown) might bind to partners via a p53/Mdm2-like mechanism.

The estimated commonness of the α-MoRE seems to be too high to be true, especially since the current estimate likely misses many examples because the algorithm is probably overfitted to the training set. A simple explanation would be that most of the indicated α-MoRE patterns are simply false positives. On the other hand, a study of these examples indicates that each α-MoRE typically provides five or more residues that are critical for the interaction with its partner. Assuming that just 10 of the 20 amino acids can contribute to the critical binding surface on an α-MoRE, such a feature would have a minimal diversity of about 10^5 distinct binding surfaces, which is comparable to the number of proteins in the human genome even if alternative splicing were considered. Thus, the structurally simple α-MoRE does indeed contain sufficient diversity to be as common as suggested by the number of patterns found by our algorithm. We therefore suggest that it would be useful to test for the commonness of this feature by finding α-MoRE patterns in various proteins and then carrying out laboratory experiments to determine what fraction of the patterns are true positives and what fraction are false positives.

α-*MoREs and Cell Signaling*

All of the examples of α-MoREs found here are related to cell signaling of one form or another. Some of the α-MoREs including the calcineurin inhibitory helix and IPKA are involved in inhibition of the active sites of enzyme partners. These two examples inhibit a phosphatase and a kinase, respectively, and may play signal-gating roles in phosphorylation-dependent signaling events. Another inhibiting α-MoRE, 4E-BP1, blocks protein−protein interactions by steric interference in the binding interface between eIF4E and eIF4G by functioning as a molecular mimic of eIF4G (*85*). The α-MoRE region in Bad may very well function in an analogous manner to 4E-BP-1 (i.e., blocking Bcl-xL interactions with other molecules). Other α-MoRE examples are involved in complex formation and/or recruitment of other molecules, such as in APC, FtsZ, and U2AF65. The binding mechanism of the FliM α-MoRE to CheY is not clear. However, this binding is clearly a signaling event, which results in the reversal of direction of flagellar rotation.

Finally, four of the α-MoREs described here play a role in transcription regulation and control. NcoA-2 and SRC-1 are paralogous nuclear receptor coactivators, where multiple α-MoRE regions, scattered through their sequences, are the regions primarily responsible for receptor coactivation (*76*). Mad binds DNA, and the α-MoRE region recruits the Sin3-histone deacetylase repressor complex (*6*). The two α-MoREs examples in p53 function in regulation of its transactivation activity (*70, 80*).

Clearly, the diversity of functions represented by this single, simple structural feature can only be understood in terms of cell signaling. This commonness and functional diversity also raises the possibility that the high number of α-MoREs suggested herein is not a mistake, but rather an indication of a very widely used mechanism in eukaryotic signaling.

Illustrative Examples

The nucleoprotein N from measles virus (MV) and the p53 tumor suppressor are very different proteins that both use α-MoREs to carry out function. A further examination of these proteins provides additional insight into the α-MoRE concept.

MV Nucleoprotein N. MV is an enveloped virus whose RNA genome is encapsulated within a helical nucleocapsid by the nucleoprotein, N. This N-RNA template is transcribed and replicated by a viral RNA-dependent RNA polymerase complex, consisting of the phosphoprotein P, and the large protein L (*87*). The polymerase catalytic activity of this RNA polymerase resides within L, with P being responsible for numerous other activities, including anchoring of the polymerase to the nucleocapsid (*88*). Protein N can be divided on two functional regions, N_{CORE}, which is a well-conserved and proteolysis-resistant N-terminal domain (residues 1−400), and the C-terminal fragment (residues 401−525) N_{TAIL}, which is hypervariable and hypersensitive to proteolysis. Interestingly, N_{CORE} includes all the regions necessary for self-assembly and RNA binding, whereas N_{TAIL} binds P and is required for N-RNA to act as a template for viral RNA synthesis (*89*−*93*). Computational analysis of the protein N sequence using PONDR VL-XT (*14*, *31*−*34*) and charge−hydropathy plot (*12*) predicted that N_{CORE} was ordered, whereas N_{TAIL} was predicted to be disordered (*43*).

These predictions for the intrinsically disordered nature of N_{TAIL} were confirmed experimentally using a wide range of biophysical techniques (*43*, *94*). Furthermore, binding of N_{TAIL} to the C-terminal domain of P protein was shown to result in a strong disorder-to-order transition, as indicated by the considerable increase in α-helical structure (*94*). This example shows how the prediction of disorder and the indication of an α-MoRE pattern within the predicted region of disorder can be used as a guide for experiments.

The Tumor Suppressor p53. Inactivation of p53 is one of the most common events in neoplastic transformation. In approximately 50% of all cancer cases so far studied, p53 is inactivated by mutations and related genomic alterations. In many of the remaining cases, p53 becomes functionally inactivated by elevated levels of the Mdm2 oncoprotein. Mdm2 acts by binding to the transactivation domain of p53, leading in turn to p53 degradation via the ubiquitin−ubiquitin lyase−proteasome pathway (*70*, *79*).

The structure of the complex between Mdm2 and p53 transcription activation domain has been determined by X-ray crystallography (*70*). Mdm2 in this complex exhibits a concave surface reminiscent of a twisted trough, with the cleft being lined with hydrophobic amino acid residues. The p53 peptide forms an amphipathic helix of ∼8 residues (∼2.5 turns), which nestles into the Mdm2 cleft (*70*). Importantly, in the absence of Mdm2 or other binding partners, this helix is rapidly digested by proteases, reflecting high flexibility

of the peptide in the nonbound state (*70*). Recent NMR studies confirmed the disordered nature of the isolated transcription activation domain, revealing however that the functional helix might be present as a measurable fraction of the overall structural ensemble (*95*).

Similarly, a peptide derived from the C-terminal region of p53 (residues 367−388) was found to have no regular structure in its native form by NMR spectroscopy, but became α-helical when bound to Ca^{2+}-loaded S100B(*ββ*) (*80*). The 3-D structure of this complex revealed several favorable hydrophobic and electrostatic interactions between S100B(*ββ*) and the p53 peptide, with the binding of S100B(*ββ*) leading to the blockage of sites for phosphorylation and acetylation on p53. These chemical modifications are important for subsequent transcription activation (*80*).

Figure 5 also illustrates two additional interesting features in the PONDR VL-XT prediction for p53: (a) a rather long central region with a high level of predicted order (residues 100−290) that matches almost exactly the region identified as the DNA binding domain (*96*); and (b) a downward spike doublet region (residues 320−360), which gives an excellent match to the tetramerization domain (*97*). The formation of the tetramer domain starts with a *β*-strand and an α-helix which associate with a second molecule across an antiparallel *β*-sheet and an antiparallel helix−helix interface to form a dimer. Two of these dimers interact across a second and distinct parallel helix−helix interface to form the tetramer. Notice that this tetramer contains very little intramolecularly buried surface area, so tetramer formation must involve large burial of accessible surface area and water release as has been observed for several coupled binding and folding interactions. From our viewpoint, the tetramerization domain is a complex α/*β*-MoRE that involves both α- and *β*-structures and undergoes mutually induced folding by self-association.

Unlike the studies on the N protein from measles virus, predictions of disorder and indications of α-MoRE patterns played no role in guiding the experiments on p53 and its self-association or its association with binding partners. However, p53 is one of the most thoroughly studied proteins that serves as a prototype for signaling and regulation. According to the findings presented herein, the use of α-MoRE patterns by p53 could represent a much more general phenomenon than previously realized.

Potential Biological Roles of α*-MoREs*

The correlation of α-MoRE patterns with functional and cellular localization annotations suggests some general biological roles for α-MoREs. Many of these correlations support our model of α-MoREs as a key component for mediating protein−protein interactions. The correlations also suggest a role for MoREs in many signaling events and pathways, presumably also involving protein interactions. One particular set of correlations suggests a central role for α-MoREs in mediating protein interactions involved is yeast replication. Specifically, many proteins localized to bud structures, as well as proteins involved in replication, are indicated to contain MoRE patterns.

One unexpected result of this analysis was the high frequency of α-MoRE patterns in proteins localized to cellular cortex and cytoskeleton. These correlations do not indicate the role of α-MoREs in this context, but it seems reasonable to suggest that they play a role in recruiting proteins to the particular sites on the cytoskeleton. Such a role for α-MoREs would be advantageous, since the surrounding disordered region would give the α-MoRE a larger diffusion search space, relative to an ordered domain anchored to the cytoskeleton. Also, the disorder inherent in unbound α-MoREs may enhance binding rates through the proposed fly-casting model (*98*), which contains ideas similar to much earlier proposals regarding the importance of disorder in protein associations (*99*).

Advantages of α*-MoREs for Cell Signaling*

The frequency of α-MoREs in various types of proteins is highest in those associated with signaling and lowest in the metabolic enzymes. Evidently, these elements have advantages for cell signaling. The advantages of such a simple element in signaling are, among others, the decoupling of specificity/affinity. This idea is supported by the work of Petros et al. (*86*), but is discussed explicitly here. For the interaction of two globular domains, specificity and affinity are often coupled, because as binding specificity is increased by additional contacts, the free energy of binding increases as the sum total of these contacts (*83*). This type of interaction can be viewed as a single equilibrium process. However, for an isolated helix, as for other disorder-to-order transition regions, entropy is a significant factor in the free energy of binding. The α-helical propensity (i.e., the equilibrium constant of the coil−helix transition) of the α-MoRE region determines the magnitude of the unfavorable free energy in helix formation, but this equilibrium does not affect binding specificity. After helical ordering, binding is analogous to the interaction of two globular domains, with coupled affinity and specificity. This interaction can be viewed as a double equilibrium process, where the first transition affects only interaction affinity and the second reaction affects both specificity and affinity. These separate processes allow for specificity and affinity to be largely decoupled. This decoupling allows the strength and duration of signaling events to respond separately to distinctive evolutionary selection pressures. This decoupling should occur whether folding and binding are entirely separate as shown in Figure 10 for simplicity or whether folding and binding occur concomitantly in a coordinated manner. We believe that the decoupling arises from the energetics of the overall process, not from the detailed steps in the coupled folding and binding reaction.

In an especially interesting example, Petros et al. (*86*) found the minimal region of Bad (16 residues) that was able to weakly bind to Bcl-xL. They constructed a series of mutant peptides with increased helical propensity by altering only residues that are not involved in the Bad−Bcl-xL interaction. These modified fragments were shown to bind Bcl-xL with increased affinity, where the helical propensities, across multiple peptide mutants, were directly proportional to binding affinity. Although this is an artificial example, it demonstrates the biophysical underpinnings of this important feature of α-MoRE mediated signaling events.

Implications for Structural and Functional Genomics

The identification of α-MoRE patterns may provide structural genomics centers with a powerful tool for the

discovery of protein−protein interaction regions that could contribute significantly to the solution of protein complex structures. Putative binding regions based on α-MoRE patterns can be isolated by cloning (*40*) or synthesized (*100*) and used in panning experiments. Bound proteins can then be identified through proteolysis and mass spectroscopy (*101*). The corresponding sequences can then be expressed and purified. The complex structure can then be determined by coanalyzing the MoRE region and its structured partner. In some cases, it is likely that additional deletion or proteolysis for the binding partners of α-MoREs will be required in order to isolate a highly structured complex. However, the complexity of such experiments is greatly reduced by prior knowledge of the binding region of one of the partners. We believe that α-MoRE identification algorithm developed here has the potential to accelerate structure determination of complexes involving disordered regions of proteins.

ACKNOWLEDGMENT

Ya-Yue Van of Molecular Kinetics, Inc. and Zoran Obradovic of Temple University are thanked for their continuing support of our work on the roles of intrinsic disorder in protein function.

SUPPORTING INFORMATION AVAILABLE

Summary information and α-MoRE predictions for 82 genomes (Table S1), proportion of proteins with indicated α-MoRE patterns and the rate of occurrence of α-MoRE patterns in each of the 82 genomes studied (Figure S1), and the correlation of α-MoRE patterns and PONDR VL-XT predictions with the biological process ontology (Figure S2). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES

1. Hershey, P. E., McWhirter, S. M., Gross, J. D., Wagner, G., Alber, T., and Sachs, A. B. (1999) The Cap-binding protein eIF4E promotes folding of a functional domain of yeast translation initiation factor eIF4G1, *J. Biol. Chem. 274*, 21297−21304.
2. 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.
3. Pan, G., O'Rourke, K., and Dixit, V. M. (1998) Caspase-9, Bcl-XL, and Apaf-1 form a ternary complex, *J. Biol. Chem. 273*, 5841−5845.
4. Kelekar, A., Chang, B. S., Harlan, J. E., Fesik, S. W., and Thompson, C. B. (1997) Bad is a BH3 domain-containing protein that forms an inactivating dimer with Bcl-XL, *Mol. Cell. Biol. 17*, 7040−7046.
5. Nolte, R. T., Wisely, G. B., Westin, S., Cobb, J. E., Lambert, M. H., Kurokawa, R., Rosenfeld, M. G., Willson, T. M., Glass, C. K., and Milburn, M. V. (1998) Ligand binding and co-activator assembly of the peroxisome proliferator-activated receptor-gamma, *Nature 395*, 137−143.
6. Spronk, C. A., Tessari, M., Kaan, A. M., Jansen, J. F., Vermeulen, M., Stunnenberg, H. G., and Vuister, G. W. (2000) The Mad1-Sin3B interaction involves a novel helical fold, *Nat. Struct. Biol. 7*, 1100−1104.
7. Cochran, A. G. (2000) Antagonists of protein-protein interactions, *Chem. Biol. 7*, R85−R94.
8. Sharma, S. K., Ramsey, T. M., and Bair, K. W. (2002) Protein-protein interactions: lessons learned, *Curr. Med. Chem.: Anti-Cancer Agents 2*, 311−330.
9. Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996) An evolutionary trace method defines binding surfaces common to protein families, *J. Mol. Biol. 257*, 342−358.
10. Fariselli, P., Pazos, F., Valencia, A., and Casadio, R. (2002) Prediction of protein-protein interaction sites in heterocomplexes with neural networks, *Eur. J. Biochem. 269*, 1356−1361.
11. Valencia, A., and Pazos, F. (2002) Computational methods for the prediction of protein interactions, *Curr. Opin. Struct. Biol. 12*, 368−373.
12. Uversky, V. N., Gillespie, J. R., and Fink, A. L. (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions?, *Proteins 41*, 415−427.
13. Wright, P. E., and Dyson, H. J. (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm, *J. Mol. Biol. 293*, 321−331.
14. 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.
15. 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.
16. Uversky, V. N. (2002) Natively unfolded proteins: a point where biology waits for physics, *Protein Sci. 11*, 739−756.
17. Uversky, V. N. (2002) What does it mean to be natively unfolded?, *Eur. J. Biochem. 269*, 2−12.
18. Dyson, H. J., and Wright, P. E. (2002) Coupling of folding and binding for unstructured proteins, *Curr. Opin. Struct. Biol. 12*, 54−60.
19. Tompa, P. (2002) Intrinsically unstructured proteins, *Trends Biochem. Sci. 27*, 527−533.
20. Demchenko, A. P. (2001) Recognition between flexible protein molecules: induced and assisted folding, *J. Mol. Recognit. 14*, 42−61.
21. Gunasekaran, K., Tsai, C. J., Kumar, S., Zanuy, D., and Nussinov, R. (2003) Extended disordered proteins: targeting function with less scaffold, *Trends Biochem. Sci. 28*, 81−85.
22. Namba, K. (2001) Roles of partly unfolded conformations in macromolecular self-assembly, *Genes Cells 6*, 1−12.
23. 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.
24. Dyson, H. J., and Wright, P. E. (2005) Intrinsically unstructured proteins and their functions, *Nat. Rev. Mol. Cell Biol. 6*, 197−208.
25. Weinreb, P. H., Zhen, W., Poon, A. W., Conway, K. A., and Lansbury, P. T., Jr. (1996) NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded, *Biochemistry 35*, 13709−13715.
26. Daughdrill, G. W., Pielak, G. J., Uversky, V. N., Cortese, M. S., and Dunker, A. K. (2005) Natively disordered protein, in *Protein Folding Handbook* (Buchner, J., and Kiefhaber, T., Eds.) pp 275−357, Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim.
27. Romero, P., Obradovic, Z., Kissinger, C. R., Villafranca, J. E., Garner, E., Guilliot, S., and Dunker, A. K. (1998) Thousands of proteins likely to have long disordered regions, *Pac. Symp. Biocomput.* 437−448.
28. Dunker, A. K., Garner, E., Guilliot, S., Romero, P., Albrecht, K., Hart, J., Obradovic, Z., Kissinger, C., and Villafranca, J. E. (1998) Protein disorder and the evolution of molecular recognition: theory, predictions and observations, *Pac. Symp. Biocomput.* 473−484.
29. Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C., and Brown, C. J. (2000) Intrinsic protein disorder in complete genomes, *Genome Inf. Ser.* No. 11, 161−171.
30. 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.
31. Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J. E., and Dunker, A. K. (1997) Identifying disordered regions in proteins from amino acid sequence, *Proc. IEEE Int. Conf. Neuronal Networks 1*, 90−95.
32. Romero, P., Obradovic, Z., and Dunker, A. K. (1997) Sequence data analysis for long disordered regions prediction in the calcineurin family, *Genome Inf. Ser.* No. 8, 110−124.
33. Li, X., Romero, P., Rani, M., Dunker, A. K., and Obradovic, Z. (1999) Predicting Protein Disorder for N-, C-, and Internal Regions, *Genome Inform. Ser.* No. 10, 30−40.

34. Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., and Dunker, A. K. (2001) Sequence complexity of disordered protein, *Proteins 42*, 38−48.
35. Liu, J., and Rost, B. (2001) Comparing function and structure between entire proteomes, *Protein Sci. 10*, 1970−1979.
36. 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.
37. Vucetic, S., Brown, C. J., Dunker, A. K., and Obradovic, Z. (2003) Flavors of protein disorder, *Proteins 52*, 573−584.
38. Baron, M., Norman, D. G., and Campbell, I. D. (1991) Protein modules, *Trends Biochem. Sci. 16*, 13−17.
39. Pawson, T. (1995) Protein modules and signalling networks, *Nature 373*, 573−580.
40. Campbell, I. D., and Baron, M. (1991) The structure and function of protein modules, *Philos. Trans. R. Soc. London, B 332*, 165−170.
41. Garner, E., Romero, P., Dunker, A. K., Brown, C., and Obradovic, Z. (1999) Predicting Binding Regions within Disordered Proteins, *Genome Inf. Ser*. No. 10, 41−50.
42. 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.
43. Bourhis, J. M., Johansson, K., Receveur-Brechot, V., Oldfield, C. J., Dunker, K. A., 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.
44. 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.
45. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000) The Protein Data Bank, *Nucleic Acids Res. 28*, 235−242.
46. Hobohm, U., and Sander, C. (1994) Enlarged representative set of protein structures, *Protein Sci. 3*, 522−524.
47. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, *Nucleic Acids Res. 31*, 365−370.
48. The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation, *Genome Res. 11*, 1425−1433.
49. Dwight, S. S., Harris, M. A., Dolinski, K., Ball, C. A., Binkley, G., Christie, K. R., Fisk, D. G., Issel-Tarver, L., Schroeder, M., Sherlock, G., Sethuraman, A., Weng, S., Botstein, D., and Cherry, J. M. (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO), *Nucleic Acids Res. 30*, 69−72.
50. 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.
51. Kabsch, W., and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers 22*, 2577−2637.
52. Kyte, J., and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol. 157*, 105−132.
53. Vihinen, M., Torkkila, E., and Riikonen, P. (1994) Accuracy of protein flexibility predictions, *Proteins 19*, 141−149.
54. Shannon, C. E. (1948) A mathematical theory of communication, *Bell Syst. Tech. J.* 379−423.
55. Eisenberg, D., Weiss, R. M., and Terwilliger, T. C. (1982) The helical hydrophobic moment: a measure of the amphiphilicity of a helix, *Nature 299*, 371−374.
56. Jones, D. T. (1999) Protein secondary structure prediction based on position-specific scoring matrices, *J. Mol. Biol. 292*, 195−202.
57. Garnier, J., Osguthorpe, D. J., and Robson, B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, *J. Mol. Biol. 120*, 97−120.
58. Callebaut, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J., Henrissat, B., and Mornon, J. P. (1997) Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives, *Cell. Mol. Life Sci. 53*, 621−645.
59. Sprent, P. (1993) *Applied Nonparametric Statistical Methods*, Chapman and Hall, London.
60. Ewens, W. J., and Grant, G. G. (2001) *Statistical Methods in Bioinformatics*, Springer, New York.
61. Kissinger, C. R., Parge, H. E., Knighton, D. R., Lewis, C. T., Pelletier, L. A., Tempczyk, A., Kalish, V. J., Tucker, K. D., Showalter, R. E., Moomaw, E. W., Gastinel, L. N., Habuka, N., Chen, X. H., Maldonado, F., Barker, J. E., Bacquet, R., and Villafranca, J. E. (1995) Crystal structures of human calcineurin and the human FKBP12-FK506-calcineurin complex, *Nature 378*, 641−644.
62. 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.
63. 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.
64. Hoedemaeker, F. J., Signorelli, T., Johns, K., Kuntz, D. A., and Rose, D. R. (1997) A single chain Fv fragment of P-glycoprotein-specific monoclonal antibody C219. Design, expression, and crystal structure at 2.4 Å resolution, *J. Biol. Chem. 272*, 29784−29789.
65. Kiefer, M. C., Brauer, M. J., Powers, V. C., Wu, J. J., Umansky, S. R., Tomei, L. D., and Barr, P. J. (1995) Modulation of apoptosis by the widely distributed Bcl-2 homologue Bak, *Nature 374*, 736−739.
66. Sattler, M., Liang, H., Nettesheim, D., Meadows, R. P., Harlan, J. E., Eberstadt, M., Yoon, H. S., Shuker, S. B., Chang, B. S., Minn, A. J., Thompson, C. B., and Fesik, S. W. (1997) Structure of Bcl-xL-Bak peptide complex: recognition between regulators of apoptosis, *Science 275*, 983−986.
67. Kelekar, A., and Thompson, C. B. (1998) Bcl-2-family proteins: the role of the BH3 domain in apoptosis, *Trends Cell Biol. 8*, 324−330.
68. Dawson, R., Muller, L., Dehner, A., Klein, C., Kessler, H., and Buchner, J. (2003) The N-terminal domain of p53 is natively unfolded, *J. Mol. Biol. 332*, 1131−1141.
69. Willenbrock, F., Crabbe, T., Slocombe, P. M., Sutton, C. W., Docherty, A. J., Cockett, M. I., O'Shea, M., Brocklehurst, K., Phillips, I. R., and Murphy, G. (1993) The activity of the tissue inhibitors of metalloproteinases is regulated by C-terminal domain interactions: a kinetic analysis of the inhibition of gelatinase A, *Biochemistry 32*, 4330−4337.
70. 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.
71. Bren, A., and Eisenbach, M. (1998) The N terminus of the flagellar switch protein, FliM, is the binding domain for the chemotactic response regulator, CheY, *J. Mol. Biol. 278*, 507−514.
72. Mosyak, L., Zhang, Y., Glasfeld, E., Haney, S., Stahl, M., Seehra, J., and Somers, W. S. (2000) The bacterial cell-division protein ZipA and its interaction with an FtsZ fragment revealed by X-ray crystallography, *EMBO J. 19*, 3179−3191.
73. Rudner, D. Z., Kanaar, R., Breger, K. S., and Rio, D. C. (1998) Interaction between subunits of heterodimeric splicing factor U2AF is essential in vivo, *Mol. Cell. Biol. 18*, 1765−1773.
74. Cheng, H. C., Kemp, B. E., Pearson, R. B., Smith, A. J., Misconi, L., Van Patten, S. M., and Walsh, D. A. (1986) A potent synthetic peptide inhibitor of the cAMP-dependent protein kinase, *J. Biol. Chem. 261*, 989−992.
75. Spink, K. E., Polakis, P., and Weis, W. I. (2000) Structural basis of the Axin-adenomatous polyposis coli interaction, *EMBO J. 19*, 2270−2279.
76. Heery, D. M., Kalkhoven, E., Hoare, S., and Parker, M. G. (1997) A signature motif in transcriptional co-activators mediates binding to nuclear receptors, *Nature 387*, 733−736.
77. Rubinfeld, B., Albert, I., Porfiri, E., Munemitsu, S., and Polakis, P. (1997) Loss of beta-catenin regulation by the APC tumor suppressor protein correlates with loss of structure due to common somatic mutations of the gene, *Cancer Res. 57*, 4624−4630.

78. Eklof, S. K., Fridman, S. G., and Weis, W. I. (2001) Molecular mechanisms of beta-catenin recognition by adenomatous polyposis coli revealed by the structure of an APC-beta-catenin complex, *EMBO J. 20*, 6203−6212.
79. Prives, C., and Hall, P. A. (1999) The p53 pathway, *J. Pathol. 187*, 112−126.
80. 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.
81. Venyaminov, S. Y., Gudkov, A. T., Gogia, Z. V., and Tumanova, L. G. (2005) *Absorption and Circular Dichroism Spectra of Individual Proteins from Escherichia Coli Ribosome*, Biological Research Center, Institute of Protein Research, AS USSR, Pushchino.
82. Ban, N., Nissen, P., Hansen, J., Capel, M., Moore, P. B., and Steitz, T. A. (1999) Placement of protein and RNA structures into a 5 Å-resolution map of the 50S ribosomal subunit, *Nature 400*, 841−847.
83. 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.
84. Callaghan, A. J., Aurikko, J. P., Ilag, L. L., Gunter, G. 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.
85. Marcotrigiano, J., Gingras, A. C., Sonenberg, N., and Burley, S. K. (1999) Cap-dependent translation initiation in eukaryotes is regulated by a molecular mimic of eIF4G, *Mol. Cell 3*, 707−716.
86. Petros, A. M., Nettesheim, D. G., Wang, Y., Olejniczak, E. T., Meadows, R. P., Mack, J., Swift, K., Matayoshi, E. D., Zhang, H., Thompson, C. B., and Fesik, S. W. (2000) Rationale for Bcl-xL/Bad peptide complex formation from structure, mutagenesis, and biophysical studies, *Protein Sci. 9*, 2528−2534.
87. Lamb, R. A., and Kolakofsky, D. (2001) *Paramyxoviridae*: the viruses and their replication, in *Fields Virology* (Fields, B. N., Knipe, D. M., and Howley, P. M., Eds.) 4th ed., pp 1305−1340, Lippincott Williams & Wilkins, Philadelphia.
88. Portner, A., Murti, K. G., Morgan, E. M., and Kingsbury, D. W. (1988) Antibodies against Sendai virus L protein: distribution of the protein in nucleocapsids revealed by immunoelectron microscopy, *Virology 163*, 236−239.
89. Bankamp, B., Horikami, S. M., Thompson, P. D., Huber, M., Billeter, M., and Moyer, S. A. (1996) Domains of the measles virus N protein required for binding to P protein and self-assembly, *Virology 216*, 272−277.
90. Buchholz, C. J., Retzler, C., Homann, H. E., and Neubert, W. J. (1994) The carboxy-terminal domain of Sendai virus nucleocapsid protein is involved in complex formation between phosphoprotein and nucleocapsid-like particles, *Virology 204*, 770−776.
91. Curran, J., Homann, H., Buchholz, C., Rochat, S., Neubert, W., and Kolakofsky, D. (1993) The hypervariable C-terminal tail of the Sendai paramyxovirus nucleocapsid protein is required for template function but not for RNA encapsidation, *J. Virol. 67*, 4358−4364.
92. Harty, R. N., and Palese, P. (1995) Measles virus phosphoprotein (P) requires the NH2- and COOH-terminal domains for interactions with the nucleoprotein (N) but only the COOH terminus for interactions with itself, *J. Gen. Virol. 76* (Part 11), 2863−2867.
93. Nishio, M., Tsurudome, M., Ito, M., Kawano, M., Kusagawa, S., Komada, H., and Ito, Y. (1999) Mapping of domains on the human parainfluenza virus type 2 nucleocapsid protein (NP) required for NP-phosphoprotein or NP-NP interaction, *J. Gen. Virol. 80* (Part 8), 2017−2022.
94. Longhi, S., Receveur-Brechot, V., Karlin, D., Johansson, K., Darbon, H., Bhella, D., Yeo, R., Finet, S., and Canard, B. (2003) The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein, *J. Biol. Chem. 278*, 18638−18648.
95. 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.
96. Cho, Y., Gorina, S., Jeffrey, P. D., and Pavletich, N. P. (1994) Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations, *Science 265*, 346−355.
97. Jeffrey, P. D., Gorina, S., and Pavletich, N. P. (1995) Crystal structure of the tetramerization domain of the p53 tumor suppressor at 1.7 angstroms, *Science 267*, 1498−1502.
98. Shoemaker, B. A., Portman, J. J., and Wolynes, P. G. (2000) Speeding molecular recognition by using the folding funnel: the fly-casting mechanism, *Proc. Natl. Acad. Sci. U.S.A. 97*, 8868−8873.
99. Pontius, B. W. (1993) Close encounters: why unstructured, polymeric domains can increase rates of specific macromolecular association, *Trends Biochem. Sci. 18*, 181−186.
100. Soellner, M. B., Nilsson, B. L., and Raines, R. T. (2002) Staudinger ligation of alpha-azido acids retains stereochemistry, *J. Org. Chem. 67*, 4993−4996.
101. Lahm, H. W., and Langen, H. (2000) Mass spectrometry: a tool for the identification of proteins separated by gels, *Electrophoresis 21*, 2105−2114.

BI050736E