doi:10.1016/j.jmb.2006.07.087 *J. Mol. Biol.* (2006) **362**, 1043–1059

JMB

Available online at www.sciencedirect.com
ScienceDirect

ELSEVIER

Analysis of Molecular Recognition Features (MoRFs)

**Amrita Mohan[1], Christopher J. Oldfield[1,2], Predrag Radivojac[1]
Vladimir Vacic[2], Marc S. Cortese[2], A. Keith Dunker[2,3]*
and Vladimir N. Uversky[2,3,4]***

[1]*School of Informatics, Indiana University, Bloomington IN 47408, USA*

[2]*Center for Computational Biology and Bioinformatics Department of Biochemistry and Molecular Biology, Indiana University School of Medicine Indianapolis, IN 46202, USA*

[3]*Molecular Kinetics, Inc. Indianapolis, IN 46268, USA*

[4]*Institute for Biological Instrumentation, Russian Academy of Sciences 142290 Pushchino Moscow Region, Russia*

**Corresponding authors*

Several proteomic studies in the last decade revealed that many proteins are either completely disordered or possess long structurally flexible regions. Many such regions were shown to be of functional importance, often allowing a protein to interact with a large number of diverse partners. Parallel to these findings, during the last five years structural bioinformatics has produced an explosion of results regarding protein–protein interactions and their importance for cell signaling. We studied the occurrence of relatively short (10–70 residues), loosely structured protein regions within longer, largely disordered sequences that were characterized as bound to larger proteins. We call these regions molecular recognition features (MoRFs, also known as molecular recognition elements, MoREs). Interestingly, upon binding to their partner(s), MoRFs undergo disorder-to-order transitions. Thus, in our interpretation, MoRFs represent a class of disordered region that exhibits molecular recognition and binding functions. This work extends previous research showing the importance of flexibility and disorder for molecular recognition. We describe the development of a database of MoRFs derived from the RCSB Protein Data Bank and present preliminary results of bioinformatics analyses of these sequences. Based on the structure adopted upon binding, at least three basic types of MoRFs are found: α-MoRFs, β-MoRFs, and ι-MoRFs, which form α-helices, β-strands, and irregular secondary structure when bound, respectively. Our data suggest that functionally significant residual structure can exist in MoRF regions prior to the actual binding event. The contribution of intrinsic protein disorder to the nature and function of MoRFs has also been addressed. The results of this study will advance the understanding of protein–protein interactions and help towards the future development of useful protein–protein binding site predictors.

Keywords: intrinsic disorder; molecular recognition; PONDR; signaling; protein–protein interaction

Introduction

Traditional understanding of protein structure and function relationships relies on protein function being critically dependent on a well-defined three-dimensional protein structure. However, recent studies have revealed that the true functional state for many proteins and protein domains is intrinsically unstructured.[1–20] Intrinsically unstructured proteins, also known as intrinsically disordered proteins, lack a single, stable conformation in solution, where conformations fluctuate over time and over the population. Many proteins have been found to be entirely disordered while others contain both structured regions and disordered regions. The evidence that these intrinsically disordered proteins exist *in vitro* and *in vivo* is compelling and justifies considering them as a separate class within the protein universe.[5,11–13,21] A number of papers and reviews have reported and discussed advances in the

Present address: A. K. Dunker and V. N. Uversky, Center for Computational Biology and Bioinformatics, Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, 714 N. Senate Street, Suite 250, Indianapolis, IN 46202, USA.

Abbreviations used: MoRF, molecular recognition feature; OM, ordered monomers; OC, ordered complexes; PPII, poly-(L-proline) II.

E-mail addresses of the corresponding authors: kedunker@iupui.edu; vuversky@iupui.edu rapidly progressing study of intrinsically disordered proteins, with major efforts being directed towards gathering evidence for their unfolded nature and discussing the functional implications of their malleable structural states.[1–20]

Disordered proteins are common in various proteomes and their abundance increases with increasing organism complexity.[3,14,15,22–24] This increased prediction of disorder in eukaryotes compared with the prokaryotes or the archaea has been hypothesized to be a consequence of the increased need for cell signaling and regulation.[15,22,24,25] The functional importance of protein disorder is further emphasized by its role in various signal transduction processes, cell-cycle regulation, gene expression and molecular recognition.[2,4,5,18–20] The widespread prevalence and importance of these intrinsically disordered proteins has called for re-assessing the understanding of the classical protein structure–function paradigm.[1]

Among other functions, intrinsic disorder has been suggested to play an important role in molecular recognition.[4,10,19,20,26,27] Molecular recognition is defined as a process by which biological entities specifically interact with each other or with small molecules to form complexes. Complex formation is often a prerequisite for biological function, but also serves as a mechanism of functional modulation and signal transduction. Common features of intrinsic disorder-mediated molecular recognition are thought to be: (a) a combination of high specificity and low affinity; (b) binding diversity in which one region specifically recognizes different partners by structural accommodation; (c) binding commonality in which multiple, distinct sequences recognize a common binding site, where these sequences may assume different folds. These same features are thought to be crucial for interaction-mediated signaling and regulation, which suggests that intrinsic disorder may play a central role in signal transduction.[1,4,10,19,20,26]

Many examples of molecular recognition between partners that are wholly disordered prior to binding have been described.[9,19,20,27] These interactions have been shown to result in the formation of structured protein complexes and are said to undergo a disorder-to-order transition on binding. This mode of molecular recognition is clearly incompatible with the current view of structure–function relationships. Molecular complementarily, which is rooted in the century-old ''lock-and-key'' concept of Fischer,[28] is inappropriate for an interaction in which one partner has no predefined structure. The more recent concept of induced fit,[29] which takes into account that even structured proteins are flexible to some extent, does not describe the scale of conformational rearrangements observed for intrinsically disordered proteins. Clearly, the role of intrinsically disordered protein in molecular recognition requires a new model to describe the determinants and driving forces of this phenomenon.

Toward developing a model for intrinsic disorder-mediated interactions, the idea of molecular recognition features or elements (MoRFs) has previously been proposed.[26] The MoRF model describes regions of intrinsic disorder that undergo a disorder-to-order transition upon partner recognition, where the residues responsible for these interactions are typically linear in the protein sequence. These regions have been estimated to be common in proteomes, particularly eukaryotes, and may be enriched in proteins with regulatory and signal transduction functions.[26] This previous work focused on a small set of MoRFs, those that form α-helices when bound to partners (α-MoRFs).[26] However, the class of all MoRFs is thought to be much broader and include MoRFs that form β-strands (β-MoRFs), irregular structures (ι-MoRFs), and a combination of secondary structural forms (complex-MoRFs).

In this work, a representative set of all MoRF types was assembled and analyzed to reveal their common properties. There is currently a lack of information on the various features and characteristics of MoRFs and little is known about the mechanisms underlying the structural changes in MoRFs during their binding phase. The aim of our work was to begin to fill this knowledge gap. To this end, a dataset of MoRFs was collected from the RCSB PDB and evidence of their intrinsic disorder was collected through sequence and structure-based methods. These examples have also been characterized in terms of physiochemical properties, such as composition and charge. Another goal was to discover signs of inherent secondary structure preferences, if any, in MoRFs prior to binding, which could potentially influence their final structure in the ordered complex. Secondary structure propensity in MoRF sequences were assessed by a secondary structure predictor, PHD,[30,31] and compared to experimentally determined structures. The results of these analyses should help to further our understanding of the physicochemical and structural determinants of intrinsically disordered regions that serve as molecular recognition elements.

Results

MoRF and control dataset

An initial set of MoRFs was collected from the PDB by selecting protein chains of less than 70 residues bound to other protein chains greater than 100 residues. The choice for selecting protein chains with lengths less than 70 residues stemmed from the idea that such proteins would be less likely to form a stable structure prior to interaction with other proteins. In other words, such protein chains would less likely be able to develop significant buried surface area before participating in the molecular recognition event.

Using these criteria, a starting dataset consisting of 2512 protein chains was assembled, where these chains were reduced to give the final non-redundant MoRF dataset. Table 1 summarizes the major steps in the development of the MoRF dataset. The PDB files corresponding to the initial 2512 proteins

Table 1. Number of MoRFs after each data processing step

	Number of MoRFs	Residues
Initial MoRFs obtained using the PDB SEQRES dataset (July 2004)	2512	95,456
Filtering ambiguous data (X,Z), removal of sequences with less than 10 residues	1261	43,836
Sequence redundancy removal	372	10,434

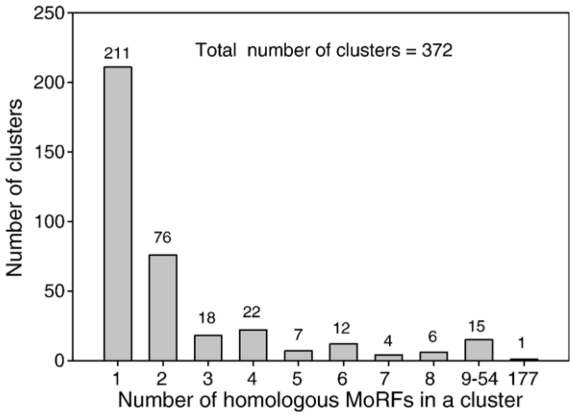

Figure 1. Frequency distribution of number of homologous MoRF sequences for 372 non-redundant MoRF dataset. The *x*-axis gives the cluster size and the *y*-axis shows the number of cluster with the given size.

were downloaded to obtain sequences, secondary structure, and information on Ramachandran's φ and ψ angles. The PDB SEQRES dataset contains all the protein sequences available in the PDB along with the residues observed in protein crystals or in solution. These sequences also included residues not present in the crystal model (e.g. disordered, lacking electron density, cloning artifacts, and His–tags).

The first step was to remove all chains with ambiguous sequence information (e.g. sequences containing X designations instead of standard amino acid designations). Protein chains ≤10 residues were also removed to facilitate mapping MoRF chains to their parent sequences. That is, many MoRF chains in the PDB are fragments of longer proteins and such short peptides may not be long enough to be specific to the parent protein sequence. At the end of this step, 1261 chains (approximately 44,000 residues with an average chain length of 34.9 residues) remained.

The next step was to remove sequence redundancy, which was done through application of length-dependent thresholds of sequence identity. This was necessary in order to overcome length variations and the overall short lengths of the MoRFs. It has been shown that pair-wise sequence identity alone is a poor definition of the twilight zone of sequence identity, which is the point where the ability to infer structural similarity from sequence becomes ambigous.[32] The use of length-dependent cut-offs to ascertain degrees of similarity within the MoRF dataset helps to correlate sequence alignments to actual structural similarity more strongly. Rost's formula[32] was used to dynamically calculate the sequence identity threshold based on each chain's length.

Clusters were constructed using these dynamic thresholds and representatives of each cluster were selected as described in Materials and Methods, which resulted in a dataset of 372 MoRFs. These MoRFs, together with their major characteristics and binding partners, are listed in Table S1 (see Supplementary Data). The selected structures consisted of 336 X-ray structures, 23 NMR structures, and five cryo-electron microscopy structures. The average resolution of the X-ray structures was 2.41(±0.60) Å. Figure 1 shows the distribution of cluster members within the MoRF dataset. The minimum number of members per cluster was at least one and the maximum number of members was 177 (Thrombin, Alpha-Thrombin). Analysis of the lengths for all MoRFs showed that approximately two-thirds of the selected chains had lengths between 10 and 20 residues (Figure 2). Each selected MoRF was mapped to its parent sequence as described in Materials and Methods, where all but 53 were fragments of longer sequences.

For comparison, three datasets of ordered proteins, as described in Materials and Methods, were used, namely: ordered monomers (OM), ordered complexes (OC), and PDB select 25 (PDB_25). The OM set contained unique monomeric proteins from PDB X-ray structures with an average resolution of 2.04(±0.50)Å. The OC set represents chains from PDB X-ray structures that are ordered prior to complex formation, with an average resolution of 1.86(±0.43)Å. The PDB_25 set is a non-redundant dataset that is representative of all chains in the PDB, where no chain in this set has a resolution poorer than 3.5 Å.

Secondary structure analysis

The secondary structure assignments for each of the 372 MoRFs were determined by the DSSP program, which was designed to standardize

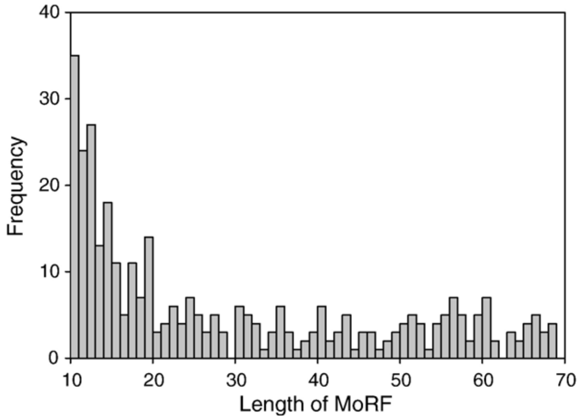

Figure 2. Length distribution of the MoRFs dataset.

protein secondary structure assignments.[30] DSSP accepts a single PDB file as input and assigns secondary structure types (*viz*. α-helices, β-strands and irregular) to each residue of the sequence. This analysis revealed that there are at least three basic types of MoRFs: those that form α-helical structures (α-MoRFs), β-strands (β-MoRFs), and irregular structures (ι-MoRFs) upon binding. Furthermore, several complex-MoRFs, i.e. containing different combinations of α-helical, β-structural and irregular elements, were also identified. Figure 3 represents several illustrative examples of MoRFs. This Figure is discussed in more detail below.

Secondary structure analysis revealed that 27% of the residues in the MoRF dataset had α-helical conformation, 12% were β-strands residues and approximately 48% were residues of irregular structure. The remaining 13% were residues with missing coordinates in the corresponding PDB files suggesting their highly flexible (disordered) nature. We compared these results with those from the OM dataset (Figure 4). The two distributions are significantly different by a χ^2 test ($p = 4 \times 10^{-80}$), but the relative χ^2 value (0.003) indicates that the difference, though significant, is relatively small. The content of irregular structure is the largest difference between the MoRF and OM datasets, with MoRFs having 6% more than OMs. Relative to OMs, MoRFs also have an increased content of residues with missing density and a corresponding decreased content of α-helix and β-strand residues. Over both sets, irregular structure is the most abundant secondary structure type and β-strand is the least.

The relative roles of local and non-local interactions in determining the secondary structures of MoRFs and OMs were studied. The role of non-local interaction, both inter-chain interactions and intra-chain interactions between residues distant in sequence, in the determination of local structure is somewhat controversial. Some authors have found local interactions to be dominant over non-local interactions in determining local structure,[33] while others have shown non-local interactions to have a direct effect on accuracy of predictions of local structure.[34] Here, we take the view that different proteins, and likely different regions in the same protein, vary in the relative contributions of local and non-local interaction to local structure, but we make no attempt to contribute to the debate over the magnitude of these contributions. The relative role of non-local interactions was investigated by comparing the secondary structure prediction accuracies, using the PHD algorithm,[30,31] for the MoRF and OM datasets. In the single sequence mode, the PHD secondary structure prediction algorithm uses a series of neural networks applied over the local sequence only. Since predictions do not consider the entire sequence, predicted secondary structure should reflect the secondary structural preferences of the local sequence, excluding influences from non-local interactions and bound partners. PHD is more commonly applied to sequence profiles generated from multiple alignments for better accuracy; however. the use of profiles is generally believed to provide a local encoding of information about long range interactions, which is not desired for the present analysis. In single sequence mode, differences between observed secondary structure and predicted secondary structure may reflect the extent to which interactions between residues distant in

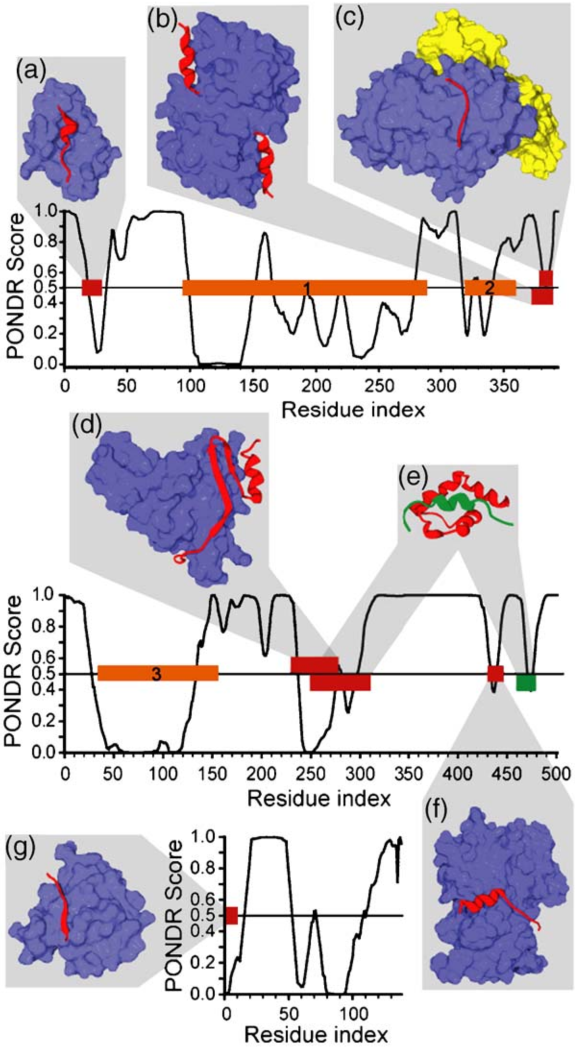

Figure 3. Examples of disorder predictions of MoRF-containing proteins and complexes between MoRFs and their binding partners. PONDR VL-XT predictions are shown for p53 (upper plot), WASP (center plot), and Grim (lower plot). The complexes shown are: (a) the α-MoRF of p53 bound to MDM2 (PDB code 1YCR); the ι/α -MoRF of p53 (b) bound to cyclin A2 (PDB code 1H26) and (c) bound to S100ββ (PDB code 1DT7); (d) the complex-MoRF of WASP bound to Cdc42 (PDB code 1CEE); (e) a complex between two MoRFs of WASP (PDB code 1EJ5); (f) the α-MoRF of WASP bound to actin (PDB code 1JD5); (g) the β-MoRF of Grim bound to DIAP1 (PDB code 2A3Z). The correspondence between sequence regions in the VL-XT plots and sequences represented in the structures is illustrated by the horizontal red and green boxes. Orange boxes represent regions with known or homology inferred structure (see the text for details).

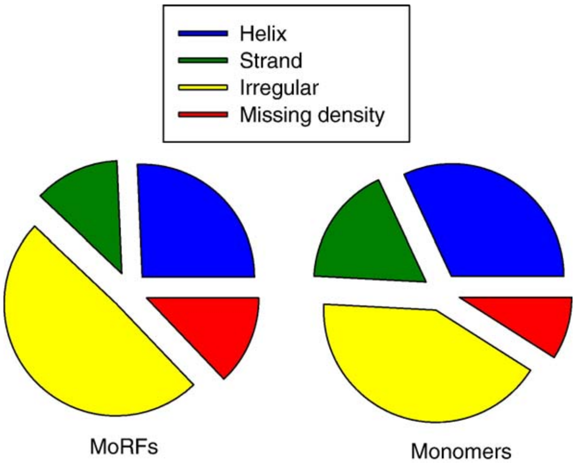

Figure 4. Secondary structure distribution of residues in the MoRF dataset and in the OM dataset.

sequence, in the case of monomers, or binding partners, in the case of MoRFs, influences the final protein conformation.

The overall prediction performance is consistent with the reported accuracy of PHD, with a single sequence prediction accuracy of 61% and a reduced accuracy of prediction for β-strands relative to α-helices and irregular structure (Table 2). Between MoRFs and OMs, the accuracy of secondary structure predictions for MoRFs is better than that for OMs by 5%. Furthermore, prediction accuracy is better for MoRFs for all defined secondary structure types, where much of this difference is due to the prediction accuracy for α-helices (9% better) rather than for β-strand (4% better) or irregular structure (3% better). These data suggest that the local secondary structural propensity of MoRFs is somewhat better preserved in their bound state, particularly for helical regions, than the local secondary structural propensity of OMs.

The secondary structure predictions for regions of missing density are also revealing; the missing density in MoRFs is predominantly predicted to be in an irregular conformation with much less of the missing density in OMs predicted to be irregular (31% difference). Missing density cannot be unequivocally related to intrinsic disorder, since missing density may correspond to mobile, structured domains. However, the lower content of predicted regular secondary structure in MoRFs, relative to OMs, may be an indication that the missing density in MoRFs is more likely to be disordered than the missing density in OMs. This provides further support to the idea that MoRFs occur in a disordered context, since the majority of missing density in these chains occurs in the N and C-terminal tails of the crystallized fragments.

Structural types were further analyzed in terms of contiguous structural regions. The MoRF set was broken into 1880 regions of sequence contiguous elements of secondary structure or missing density. Examination of the different structural types (Table 3) revealed that 269 regions were α-helical while 381 were β-strands. The larger number of β-strand regions than α-helix regions can be reconciled with the larger number of α-helix residues than β-strand residues (Figure 4) by observing that α-helical regions are on average longer than β-strand regions, with average lengths of 10±8 and 3±2 residues, respectively. More than half of the total regions (991) were found to have an irregular conformation. The remaining 239 regions were disordered. The average lengths of irregular regions (5±5 residues) and missing density regions (5±6 residues) are of intermediate length compared to α-helices and β-strands.

Amino acid composition, charge and aromaticity

It has been reported that local amino acid composition, flexibility, hydropathy, charge, coordination number and several other physiochemical properties of intrinsically disordered protein regions are significantly different from the same characteristics derived from ordered protein regions.[5,35–37] These properties have been examined for MoRFs, relative to ordered proteins, to investigate order/disorder propensity of MoRF regions. For this analysis, PDB_25 was used, since this set has been well characterized in terms of composition relative to intrinsically disordered proteins.[5,35–37]

The amino acid composition of intrinsically disordered proteins is characterized by depletion in order-promoting residues, such as C, V, L, I, M, Y, F, and W, and enrichment in disorder-promoting residues, such as Q, S, P, E, K, G, and A, relative to ordered proteins.[25,38] According to the MoRF

Table 2. PhD secondary structure prediction accuracies for MoRFs

		α-Helix	β-Strand	Irregular	Missing density
	DSSP (residues)	2469	1118	4359	1147
MoRFs and monomers	PHD (%)	**H: 74** B: 9 I: 17	H: 11 **B: 55** I: 34	H: 21 B: 15 **I: 64**	H: 18 B: 10 **I: 72**
	DSSP (residues)	35,938	19,363	47,029	10,189
Monomers	PHD (%)	**H: 65** B: 9 I: 32	H: 16 **B: 51** I: 32	H: 20 B: 18 **I: 61**	H: 31 B: 27 **I: 41**

Numbers indicated in bold represent the prevailing type of secondary structure.

Table 3. Region wise distribution in different structural types of MoRFs

Region length (in residues)	No. of missing density regions	No. of α-helical regions	No. of β-strand regions	No. of irregular regions
1–9	205	167	376	847
10–19	26	76	5	128
20–29	5	17	0	10
30–69	3	9	0	6
Total	239	269	381	991

hypothesis, MoRFs are disordered in the absence of binding partners, and consequently their amino acid compositional biases may be more similar to intrinsically disordered proteins than to ordered proteins. Alternatively, the binding propensities of these regions may give the compositions of these proteins a bias somewhere in between those of ordered and disordered proteins. To test the compositional bias of MoRFs, the fractional difference between MoRF compositions and PDB_25 compositions was calculated. The fractional difference was calculated as $(C_{\mathrm{MoRF}}–C_{\mathrm{order}})/C_{\mathrm{order}}$, where C_{MoRF} is the averaged amino acid composition of a MoRF dataset, and C_{order} is the averaged amino acid composition in PDB_25. The results are shown in Figure 5, where amino acids are arranged from the most rigid on the left to the most flexible on the right according to Vihinen.[39]

The comparison between amino acid compositions for MoRFs and PDB_25 (Figure 5(a)) shows that MoRFs are enriched in many of the disorder-promoting amino acids, arginine, glycine, serine, and proline, and depleted in many of the order promoting amino acids, tryptophan, isoleucine, tyrosine, valine, and leucine. These biases suggest that MoRFs are similar in composition to general intrinsically disordered proteins. However, several biases are inconsistent with this simple explanation. MoRFs are depleted or show similar composition to PDB_25 in charged residues other than arginine, which are generally disorder promoting. It is possible that the lower charge density of arginine relative to lysine makes arginine more amenable to a dual role in both ordered and disordered contexts. The general effect of this bias on the total charge of proteins is investigated in the next section.

Another bias inconsistency between MoRFs and intrinsically disordered proteins is the enrichment of MoRFs in cysteine and phenylalanine, which are generally order promoting. The effect of the phenylalanine composition on MoRF sequences is explored below. The bias of MoRFs toward cysteine is largely due to disulfide bonds. Of the 372 MoRFs, 36 contain at least one intra-chain disulfide bond, 18 contribute to at least one inter-chain disulfide bond, and four have at least one of each intra and inter-chain disulfide bonds. The cysteine residues involved in these bonds account for 73% of the cysteine residues in the MoRF dataset, which suggests that reduced cysteine is not a prevalent feature of MoRF regions. The presence of intra-chain disulfide bonds in MoRF sequences has clear implications for the hypotheses that these sequences are disordered in the absence of binding partners, since disulfides are well-known to stabilize protein structure.[40] As many as 11% of MoRFs in this dataset may be stabilized by disulfide bonds in the absence of their binding partners. This violates the MoRF hypothesis, and so we classify these chains as pseudo-MoRFs.

Compositional biases were also examined in terms of the compositions of different secondary structure elements (Figure 5(b)), where the fractional difference between MoRFs residues in helices, strands, and irregular conformations and PDB_25 residues in the same conformations were calculated. Generally, the compositions show the same biases as the bulk residues, or insignificant differences, due in part to the smaller sizes of the sets. Notable compositional biases of MoRFs include: the large bias of strand, and to a lesser extent, irregular, residues toward cysteine; the bias of strands and irregular residues toward tyrosine; the bias of strands and irregular residues toward lysine; and the bias of helical residues against proline. The latter bias is

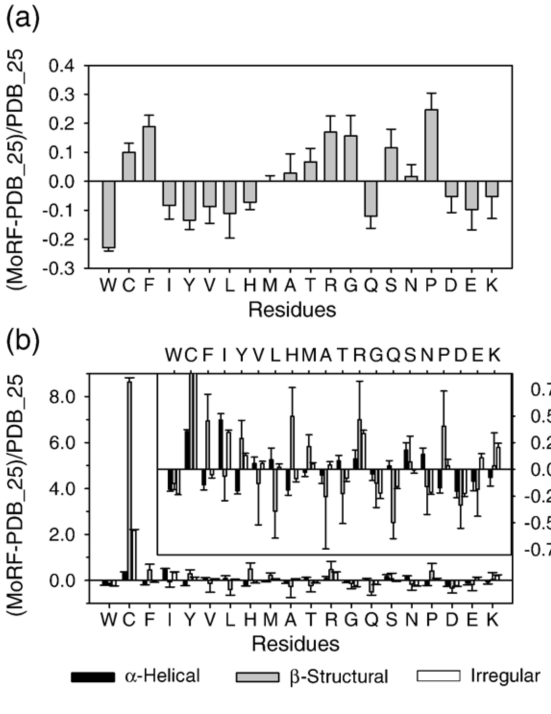

Figure 5. (a) Relative amino acid composition of MoRFs with respect to PDB_25. The fractional difference of the amino acid compositions of MoRFs and ordered monomers is calculated as (MoRFs–PDB_25)/PDB_25. (b) Relative amino acid composition of different structural types (α-helical, β-structural, and irregular) of MoRFs with respect to the same structural types in PDB_ 25. The inset represents the same graph with a reduced relative frequency range. The amino acid residues on the *x*-axis are arranged from the most rigid to the most flexible according to Vihinen *et al.*[101] Error bars represent one standard error, calculated from 200 bootstrap iterations.

particularly interesting, since MoRFs in general are enriched in proline and proline is generally regarded as a helix-breaking residue. This suggests that the helices in MoRFs are more sensitive to the conformational restrictions of proline than are helices in proteins in general.

A comparison of the total charge (K+R+D+E), net charge (K+R–D–E), proline, and aromatic content (F+W+Y) of MoRF proteins and PDB_25 proteins is shown in Figure 6. Despite being depleted in lysine, aspartic acid, and glutamic acid, MoRFs demonstrate a higher net charge than the PDB_25 proteins. The enrichment in arginine in MoRFs is apparent from the positive net charge of MoRFs, compared to the negative net charge in PDB_25. This is similar to a previous description of intrinsically disordered proteins.[2] MoRFs also show lower proportions of aromatic amino acid residues in comparison with PDB_25 proteins, despite being enriched in phenylalanine. However, the vast majority of MoRF regions contained at least one aromatic residue, often phenylalanine. This is consistent with the molecular recognition function of MoRFs, since the side-chains of aromatic amino acids tend to make strong and specific interactions.[41]

Finally, the proline content observed in MoRFs exceeds that found in PDB_25 proteins by nearly 50%. This high concentration of proline was the motivation for examining the prevalence of polyproline II helices (PPII helices) in MoRFs.

PPII helices

The poly-(L-proline) II helix, (PPII) is a left-handed helix with an axial translation of 3.20 Å, a rise of three residues per turn, and ideal backbone angles of (φ, ψ) = (−75°, +145°). The range of backbone φ- and ψ-angles is illustrated in Figure 7, where the φ and ψ-angles of the MoRF dataset are also plotted. The PPII helix is often observed in the context of proline-rich sequences,[42] but sequences that are not enriched in proline can adopt this structure.[43,44] PPII helices have even been hypothesized to be a major, though transient, conformation of protein denatured states.[45–50] For example, ROA spectra of α-synuclein, caseins and tau-protein suggest that these proteins may contain some PPII conformation.[50]

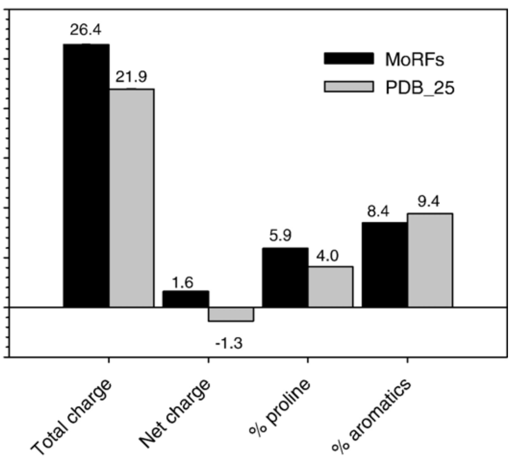

Figure 6. Total and net charge (calculated as charge per 100 residues) and the proportion of proline and aromatic amino acid residues in MoRFs and PDB_25. Error bars representing one standard error, calculated from 200 bootstrap iterations, are plotted but are narrower than the width of the bar boarders.

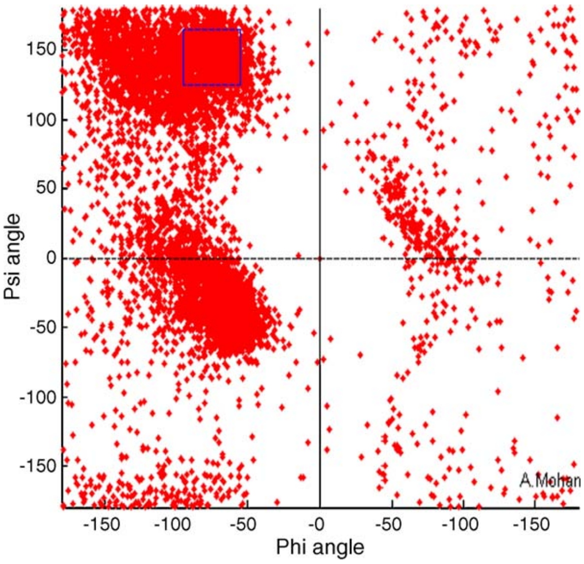

Figure 7. Ramachandran plot for MoRFs. Psi and phi-angles that correspond to the PPII residues in a PPII conformation are indicated (black box).

The enrichment of MoRFs in proline and the possible enrichment of PPII helix in intrinsically disordered proteins in general motivated the investigation of the prevalence of PPII helix in the bound structures of MoRFs. That is, the possible conformational preference of these proteins for the PPII helix may be reflected in a higher content of PPII helices in bound structures. Using the algorithm from Sreerama *et al*.[51] and Stapley & Creamer[52] to calculate the presence of PPII helices, 53 PPII regions with lengths between 4 and 12 residues were identified in the MoRF dataset. These regions included 2.6% of the residues in the MoRF dataset. For comparison, a previous study identified PPII helices of at least four residues in length accounted for 2% of all residues in known protein structures.[53] For a comparable set of 101 PDB_25 proteins that were randomly selected until ~9000 residues were in the set, only 17 unique PPII were found with no region greater than four residues in length. These data suggest that MoRFs are slightly enriched in PPII helices as compared to bulk protein structures.

Order/disorder predictions

Computational structure and sequence-based evaluations of ordered and disorder were performed to provide support for the idea that MoRFs are disordered in isolation and undergo a disorder-to-order transition upon binding. Structure-based evaluations of disorder were performed using the criteria of Gunasekaran *et al*.,[54] who showed that the complexes of intrinsically disordered proteins have much larger interface and surface areas than those of ordered proteins. Sequence-based evaluations used prediction of disorder from sequence using both PONDR® VL-XT[36,38,55] and VL3.[35] The behavior of PONDR VL-XT on MoRF containing proteins has been characterized on a small set of validated MoRFs,[26] whereas behavior of VL3 has not been characterized in this respect.

Gunasekaran *et al*.[54] have demonstrated that intrinsic disorder in the unbound state is reflected in structures of the bound state through relatively large surface and interface areas. A structural analysis of the bound structures of MoRFs in this dataset was carried out, using the previously characterized[54] OC dataset as a negative control (Figure 8). Almost all MoRFs in the dataset were above the order-disorder boundary suggested by Gunasekaran *et al*., which indicates that these regions are likely to be disordered in isolation, while all structured proteins were below this boundary, which indicates these proteins are probably ordered in isolation. Only two of the β-MoRFs and one of the ι-MoRFs falls below the suggested boundary. This analysis should be viewed with some caution, since the dataset used to derive the boundary was relatively small. Indeed, only a very slight shift in the boundary would put all of the MoRFs above it. Thus, the boundary provides a strong indication that the MoRFs in this dataset are indeed disordered in the absence of their binding partners and undergo a disorder-to-order transition upon complex formation. It should also be noted that disulfide bonds are not considered in this analysis, and so the indication that oxidized pseudo-MoRFs are disordered in the absence of their binding partners is likely to be in error. However, this analysis suggests that pseudo-MoRFs would probably be disordered in the absence of their binding partners and in the reduced state.

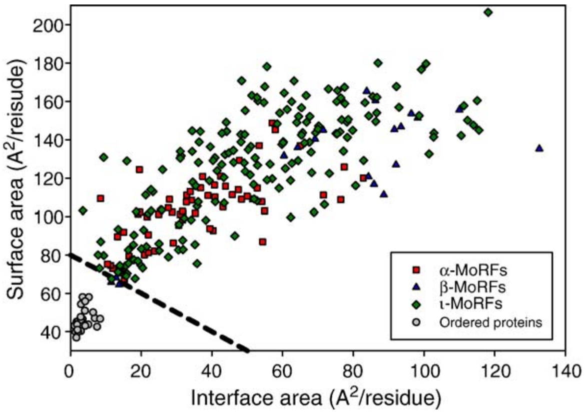

Figure 8. Surface and interface area normalized by the number of residues in each chain for the MoRF and the OC datasets.

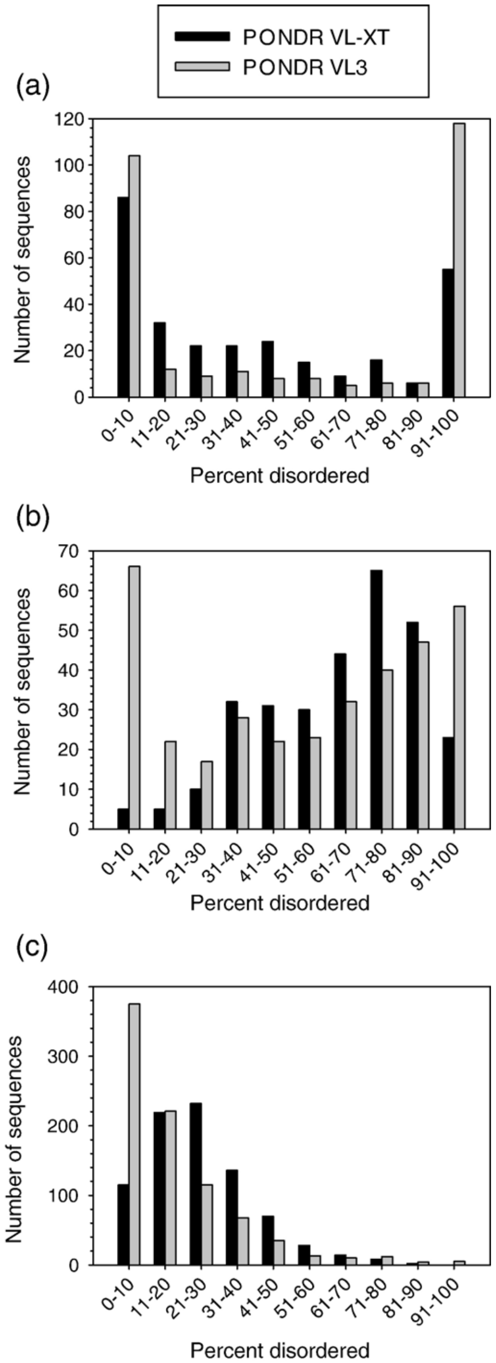

Figure 9. Disorder distribution in (a) MoRFs and (b) MoRF containing proteins and (c) OM proteins estimated by PONDR® VL-XT and VL3 predictors.

Sequence based predictions of order/disorder, made with both the PONDR® VL-XT[36,38,55] and VL3[35] predictors, seem to contradict the structure-based results. Specifically, predictions of disorder in MoRF regions (Figure 9(a)) suggest that, while many MoRFs are highly disordered, many MoRFs may be ordered. This is in part due to the large content of cysteine in these sequences, which is strongly correlated with prediction of order.[38] Also, it has been previously observed that disorder-to-order binding regions within larger disordered regions are often predicted to be ordered,[26,56] and these predictions likely reflect these previous observations.

Examination of predicted disorder with respect to secondary structure (Table 4) of MoRFs reveals some bias toward prediction of disorder in irregular residues and against prediction of disorder in β-strand residues, with α-residues showing an intermediate bias. Residues with missing density are the most likely to be predicted to be disordered, which is expected since these regions are very often disordered in solution. In comparison to OM dataset, a higher proportion of helix and irregular residues in MoRFs are predicted to be disordered. A lower proportion of MoRF strand residues are predicted to be disordered compared to OM proteins, although strand residues are the least likely to be predicted to be disordered in both datasets. The higher proportion of residues with missing density predicted to be disordered in MoRFs relative to OMs (20% higher) agrees with the secondary structure prediction accuracy analysis, which provides additional support to the idea that regions of missing density in MoRFs are likely to be disordered and consequently that MoRFs occur in a disordered context.

The previously observed bias of disorder-to-order transition regions to be predicted to be ordered[26,56] gives a false indication of intrinsic order in many MoRF sequences. This bias is evident by the extreme behavior of disordered predictions for MoRFs (Figure 9(a)), where most MoRFs are predicted to be either highly disorder or highly ordered. Therefore, disorder predictions were also examined for the entire sequences of proteins containing MoRFs and the sequence regions to the N and C sides of MoRFs in these sequences, in order to provide support for the idea that these regions occur in longer region of disorder. Disorder predictions for the full-length proteins that contain MoRFs (Figure 9(b)), relative to OM proteins (Figure 9(c)), suggest that many MoRF containing proteins are highly disordered. For the calculation of disorder in regions surrounding MoRFs, the fraction of residues predicted to be disordered was calculated over two windows of residues in the parent sequence of the MoRF, one on the C side and one on the N side of the MoRF. For ordered proteins, random sequence windows of equal size were taken from the OM set. Similar to the entire sequence of proteins containing MoRFs, the sequence regions immediately surrounding MoRFs show a high content of predicted disordered residues, relative to OM proteins (Figure 10). This suggests that these MoRFs frequently occur in longer regions of predicted disorder.

Table 4. PONDR VL-XT predictions of order/disorder for different classes of MoRFs

Dataset	Residue type	Predicted disordered (%)	Predicted ordered (%)
MoRFs	α–residues	9	18
	β–residues	2	10
	ι–residues	18	30
	Missing density	7	7
Monomers	α–residues	8	25
	β–residues	4	14
	ι–residues	7	32
	Missing density	3	7

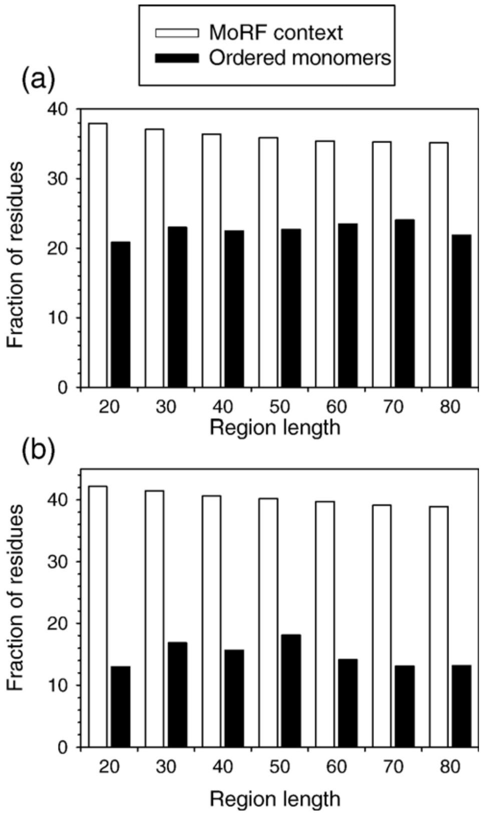

Figure 10. Fraction of residues predicted to be disordered for regions surrounding MoRFs and regions taken from ordered monomers using PONDR (a) VL-XT and (b) VL3.

Functional analysis of MoRFs

The functions of MoRF containing proteins were investigated by examining the keywords associated with those MoRFs with sequences in Swiss-Prot. Of MoRFs in the current set, 227 MoRFs were found in 201 Swiss-Prot sequences, due to non-overlapping MoRFs from the same parent sequence, and a summary of the keywords associated with these MoRFs is shown in Table 5. The most frequent keyword found for MoRF sequences, 3D structure, indicates only that these sequences have structures in PDB, which contains no information for the present analysis and is excluded from Table 5. The high number of hits for such keywords as "signal" and "alternative splicing" corresponding to the MoRF dataset suggests that sequences containing MoRFs are likely to be involved in signaling processes or being alternatively spliced. The last

Table 5. Top 8 Swiss Prot functional classes returned for MoRFs

SW keyword	Frequency
Signal	57
Glycoprotein	41
Transmembrane	37
Alternative splicing	35
Hydrolase	25
DNA binding	24
Transcription regulation	23
Serine protease inhibitor	21

observation is of crucial importance, as our recent study revealed the existence of strong correlation between intrinsic disorder and alternative splicing.[57] It has been emphasized that associating alternative splicing with protein disorder enables temporal and tissue-specific modulation of protein function needed for cell differentiation and for the evolution of multi-cellular organisms. These data suggest that MoRFs are associated with alternative spliced regions, which in turn suggests a functional role for alternatively spliced intrinsically disordered regions. That is, MoRFs provide a mechanism by which the functional profile of a protein may be modulated by alternative splicing without the restriction of structural constraints that would be present for structured proteins.

Phosphorylation modulates the activity of numerous proteins involved in signal transduction and may generally function through alteration of binding affinities.[58] Since the primary role of MoRFs is hypothesized to be molecular recognition, we estimated the extent of phosphorylation in MoRFs using the DISorder-enhanced PHOSphorylation predictor (DisPhos).[59] Briefly, DisPhos leverages both sequence profiles and disordered predictions for improved prediction of phosphorylation sites in protein sequences. The application of DisPhos to the MoRF-containing proteins revealed that in the 305 MoRFs with lengths ≥12 residues, 159 had potential phosphorylation sites. Of the 1082 phosphorylation sites predicted by DisPhos for these MoRFs, 45%, 36%, and 19% were serine, threonine, and tyrosine phosphorylation sites, respectively. A parallel PROSITE[60] search also showed that a third of MoRFs potentially contain phosphorylation sites. The DisPhos and PROSITE results suggest that phosphorylation may be a very common mechanism for the regulation of the binding affinities of MoRFs.

Examination of MoRF examples

The structures of a few examples of MoRFs are illustrated in the context of their PONDR VL-XT predictions in Figure 3. This provides an example of each of α, β, ι, and complex-MoRFs (e.g. Figure 3(a), (g), (c), and (d), respectively), and also provides examples of structural polymorphism in a MoRF bound to two different partners (Figure 3(b) and (c), (d) and (e)).

Tumor suppressor p53 plays a vital role in the regulation of cellular division in response to DNA damage and mutations of this gene are estimated to be present in ~50% of human cancer cases.[61] The four domains crucial to p53 function are shown in context of the PONDR VL-XT prediction for p53 (Figure 3, upper plot): an N-terminal MoRF, the DNA binding domain (Figure 3, box 1), the tetramerization domain (Figure 3, box 2), and a C-terminal MoRF (Figure 3, both overlapping red boxes). Both the N and C-terminal MoRFs have been verified to be disordered in the absence of binding partners.[62,63] The N-terminal MoRF is an example of an α-MoRF and corresponds to the transactivation domain of p53 bound to MDM2 (Figure 3(a)),[64] where this interaction inhibits p53′s transactivation activity and subsequent cell cycle arrest.[65] The C-terminal MoRF is an example of an ι-MoRFs and is shown interacting with the CDK2/cyclin A complex (Figure 3(b)).[66] This interaction facilitates phosphorylation and thereby activation of p53.[67] An overlapping region of p53 also interacts with S100ββ, an interaction that blocks oligomerization[68] and phoshorylation,[69] thereby blocking activation, of p53, but forms an α-helix when bound (Figure 3(c)).[62] The C-terminal region of p53 represents a single MoRF that interacts with multiple partners; it is an example of the richness of function possible under the MoRF model.

Wiskott–Aldrich syndrome protein (WASP) plays an important role in Arp2/3-mediated regulation of the actin cytoskeleton.[70] Four domains important for WASP function are shown in the context of the WASP PONDR VL-XT prediction (Figure 3, center plot), which are, from the N to C termini: the N-terminal WH1 domain (Figure 3, box 3), a complex-MoRF that corresponds to the GTPase binding domain (GBD; Figure 3, both overlapping red boxes), an α-MoRF corresponding to the WH2 domain, and the C-terminal VCA region (Figure 3, green box). Note that the GDB MoRF is the only MoRF in the current dataset; the WH2-actin complex structure (Figure 3(f))[71] was released after construction of the dataset and the VCA-GDB complex (Figure 3(e))[72] is actually a single chimerical chain, which was discarded by the MoRF selection criteria. However, both the VCA and WH2 domain are consistent with MoRF criteria and are considered such here.

The VCA domain interacts directly with the Arp2/3 complex and, together with the actin binding activity of the WH2 domain (Figure 3(f)),[71] stimulates polymer nucleation.[70] Interestingly, the Arp2/3 binding activity of the VCA MoRF is autoinhibited by the GDB MoRF (Figure 3(e)).[72] This auto-inhibitory interaction is interrupted by binding of the GDB MoRF to activated Cdc42 (Figure 3(d)),[73] which releases the VCA MoRF to interact with Arp2/3.[74] The two GDB MoRF complexes (Figure 3(d) and (e)) show radically different structures, which is an extreme example of multiple binding affinities through bound structure conformational heterogeneity.

Grim protein bound to the DIAP1 apoptosis inhibitor is an example of an extended β-MoRF (Figure 3(g)).[75] In *Drosophila*, DIAP1 inhibits apoptosis through interaction with and subsequent down-regulation of caspase activity.[76] Grim prevents the interaction of DIAP1 with caspases by competitively binding to DIAP1 with a short stretch of residues at its N terminus.[77] This mechanism of inhibiting DIAP1 activity is not unique to Grim, but is shared by the Hid and Reaper proteins,[77] which contain sequences homologous to Grim over their N-terminal 14 residues. That three otherwise non-homologous proteins share a similar mechanism for inhibiting DIAP1 demonstrates one very important aspect of MoRF function; under the MoRF model, gain of function may be obtained through mutation or duplication of relatively few residues without the restriction of maintenance of globular structure.

The examples illustrated here of the bound morphological heterogeneity of MoRFs (Figure 3(b)–(e)) may represent a more general phenomenon of molecular recognition by MoRFs. In a protein that is ordered prior to binding its partners, overlapping interaction sites on its surface are necessarily the same, which imposes severe constraints on the interface properties of its partners. For MoRFs, which are disordered prior to binding by definition, the residues used for partner recognition may be very different. This transfers the burden of specificity determination from the partner to the MoRF, relative to ordered proteins. The number of partners that can be encoded by a given sequence is likely limited, but further research is required to determine this limitation.

Finally, a clear relationship between PONDR VL-XT predictions and the sequence location of MoRFs is apparent from Figure 3; MoRFs are often predicted to occur in ordered regions flanked by long predictions of disorder. This general feature has been noted in the context of other proteins by two groups, called indications of binding regions[56] or regions of intrinsic structural preference,[78] and form the basis for the α-MoRF predictor.[26] Such examples from the MoRF dataset indicate the possibility of discovering novel binding regions in other proteins containing MoRFs. However, this heuristic is clearly not generally applicable; Figure 9(a) shows that many MoRF regions are predicted to be highly disordered. Further work is required to determine if these two types of MoRFs, predicted ordered and predicted disordered, have real physical differences, or if they can be somehow grouped into a single, novel prediction scheme.

Discussion

Intrinsic disorder and conformational bias

The express purpose of the selection process and manual inspection of MoRFs was to gather examples of proteins or protein fragments that appeared to envelop their respective partners and participate in molecular recognition mediated by a disorder-to-order transition. The fact that these MoRFs are disordered in the unbound state is supported by several lines of evidence presented here. (1) Compositional analysis shows that these MoRFs have compositions generally more similar to intrinsically disordered proteins than to ordered proteins. (2) Structure-based examination of order-disorder indicates that the structures of most of these MoRFs have surface and interface areas similar to disordered proteins. (3) Sequence-based predictions indicate that many MoRF sequences are likely to be disordered in isolation. In consideration of the previously described bias of PONDR predictors to indicate order in regions that undergo a disorder-to-order transition upon binding,[26,56] disorder predictions of MoRF regions should be considered overly conservative. (4) To compensate for this predictor bias, sequence based predictions were examined for sequences containing MoRFs and the sequence regions immediately surrounding MoRF regions. These predictions indicate that many MoRF regions are likely to occur in a disordered context. This evidence suggests that the MoRF examples examined here conform to the MoRF hypothesis.

The general molecular recognition function of MoRFs involves binding to specific partners through a disorder-to-order transition.[26] This binding process can be considered as a special type of protein folding. In protein folding, formation of tertiary structure stabilizes secondary structural elements. Similarly in disorder-to-order transitions, formation of intermolecular contacts between the MoRF and its binding partner stabilizes secondary structure elements in the MoRF. By this analogy with globular protein folding, two mechanisms of the formation of structure in MoRFs can be suggested. The first mechanism,the inherent-structure mechanism, involves the predominance of a particular local secondary structure among the highly fluctuating conformations sampled by the unbound MoRF.[79] In this case, the structure of the MoRF is not entirely random and shows some features that will later be stabilized in the bound conformation. The second mechanism, the induced-structure mechanism, is that the MoRF is in a completely disordered state prior to binding and makes initial intra- and interchain contacts randomly. These contact points serve as nucleation sites for the subsequent folding and formation of secondary structure under the influence of successive contacts with the partner. In such a mechanism, the inherent conformational preferences of the intrinsically disordered protein itself may be overridden by interactions with the partner, resulting in significantly different secondary structure elements in its uncomplexed and bound state.

Support for the inherent-structure mechanism is provided by comparison of experimental and predicted secondary structure. This comparison suggests that the conformation of the bound form of MoRFs is more dependent on local sequence, relative to monomers, and not strictly determined by the binding partner. In other words, the conformational space of unbound MoRFs may be limited by their conformational preferences and may restrict the set of possible structures in the bound state. This idea agrees with previously reported observations that MoRFs display signs of residual structure, for example p27(Kip1),[80] p53[80] and GCN4.[81] A restricted choice of available conformational states would serve to reduce the entropic cost of binding, thereby increasing affinity, provided that the predominant conformations resemble the bound conformation. The secondary structure accuracy rates of MoRF structures suggest that this is the case; the intrinsic structural propensity of the MoRF sequence is reflected in the bound state.

Support for the induced-structure mechanism is also provided by the secondary structure predictions; the accuracy of secondary structure predictions for MoRFs is only marginally better than the accuracy of secondary structure predictions for monomers. This suggests that not all of MoRF residues have strong, local structural preferences, relative to monomers, or at least that these structural preferences are not always satisfied in the bound state. The examples of structural polymorphism in MoRFs from p53 and WASP (Figure 3) demonstrate that, at least for these examples, that the intrinsically structural preferences of MoRFs cannot always be satisfied. That is, the structural preferences of the MoRF cannot be an overriding factor in determining bound MoRF conformation if MoRFs can adopt multiple conformations when bound. Another well characterized protein that exhibits this flexible binding mode is the Cdk inhibitor p21Cip1, which can interact with CycA-Cdk2, CycE-Cdk2, CycD-Cdk4 complexes[80] and apoptosis signal-regulating kinase 1[82] under different conditions. In fact, it has been shown for at least one MoRF, p27(Kip1),[80] that over-stabilization of secondary structural elements can decrease the rate constant of complex formation, which can be interpreted as residual structure interfering with the MoRF reaching the correct bound conformation.

It seems unlikely that either of the inherent or induced-structure mechanisms is completely correct, and it is more likely that both of these mechanisms are at play in MoRF mediated interactions. MoRFs can be regarded as "mixtures" of segments with strong or weak secondary structure preferences, where the strength of these preferences may serve to modulate affinity for their binding partner.

PPII helices in MoRFs

The existence of PPII peptides in our MoRF dataset suggests that the extended and rather stiff PPII helix conformation in MoRFs might be important for protein–protein interactions. The PPII conformation may be advantageous in protein interactions for several reasons, for instance the backbone atoms of peptide can form hydrogen bonds with the protein receptor at the interface of the peptide–protein complex.[83] An earlier study revealed that a great number of linear peptides, whose extended structure was determined by X-ray or NMR studies, are involved in molecular recognition processes.[84]

The PPII left-handed helical structure was almost unknown until recently, being often confused with unordered, disordered, irregular, unstructured, extended, or random coil conformations because it is neither α-helical nor β-turn nor β-strands; i.e. a classical structure.[83,84] The overall importance of this conformation has recently become apparent,[48,84] as it has been recognized that PPII may play a central role in numerous processes including signal transduction, transcription, cell motility, and the immune response.[48] Furthermore, the results of recent studies on Raman optical activity (ROA) spectra and NMR analysis provide good evidence that proteins previously thought to be in a statistical coil state may in fact be flickering in and out of a metastable PPII helical conformation.[43,85,86] It has been also hypothesized that PPII, being transiently populated by a polypeptide chain in a major amyloidogenic conformation, pre-molten globule state[87] may play a crucial role in the protein fibrillogenesis.[88] For example, it has recently been shown that the hydrated α-helix in human lysozyme readily undergoes a conformational change to PPII structure on heating, i.e. under conditions favoring fibrillation.[88] It has been assumed that this conformational change may be a key step in the conversion of α-helix into β-strand associated with the formation of amyloid fibrils in this protein. Furthermore, since the PPII helix is extended, flexible, lacks intra-chain hydrogen bonds and is fully hydrated in aqueous solution, it has the appropriate characteristics to be implicated as a critical conformational element in conformational diseases.[88]

Molecular function and MoRFs

The functional analysis of MoRFs here agrees with previous functional analyses of intrinsic disorder in general. Specifically, it was found that many MoRFs are associated with signaling and alternative splicing and that phosphorylation may be a general mechanism for regulating MoRF binding functions. For intrinsic disorder in general, previous studies[4,5,11,16,17] have found that intrinsic disorder is strongly associated with signal transduction, cell-cycle regulation and gene expression and thus may often be implicated in various cancer types.[20] Disorder is also strongly correlated with the sites of post-translational modification, such as phosphorylation, acetylation, ubiquitination, hydroxylation, and proteolysis.[5,9–12,16,19] Protein phosphorylation represents an important regulatory mechanism in eukaryotic cells, where at least one-third of proteins undergo reversible phosphorylation.[89]

Another prominent feature of intrinsic disorder is that its extreme proteolytic sensitivity, in principle, allows for effective and temporally responsive control *via* rapid turnover. Disordered proteins, some of which are known to be short-lived and subject to rapid turnover, are prevalent among signaling, regulatory and cancer-associated proteins.[3,15] Furthermore, disorder itself constitutes an integral part of the proteasomal destruction signal in two distinct ways. On one hand, non-ubiquitinated intrinsic disorder may be directly degraded by the 20 S proteasome, as shown for the disordered proteins p21Cip1[90] and tau.[91] On the other hand, this mechanism may also play a more subtle regulatory role; disordered segments in multidomain proteins may be processed, thereby releasing the flanking, constitutively activated globular domains due to the endoproteolytic activity of the proteasome.[92] Disorder may also constitute part of the signal to the ubiquitination system itself. For example the regions of securin and cyclin B recognized by the ubiquitination machinery have been shown recently to be natively unfolded.[93] We hypothesize that many of the properties of intrinsic disorder also apply to MoRFs, given that they are disordered in the absence of binding partners. Conversely, MoRFs are primarily responsible for a subset of the properties of intrinsic disorder in general.

Functional disorder was noted to be associated with molecular recognition that involves protein binding to RNA, DNA, other proteins, and small molecule ligands.[2–4,7–12,15,16,19,24,25] For protein–protein interactions, MoRFs may provide a general mechanism by which intrinsic disorder mediates these interactions. MoRFs, and intrinsic disorder in general, provide several desirable properties associated with protein–protein interactions that mediate signaling events. The relatively large exposed surface area of unbound MoRFs enables them to contact their partner(s) over a large binding surface for a protein of the given size. This allows an interaction potential to be realized by shorter proteins, thus facilitating more economical encoding, transcription, translation and spatial requirements for a given recognition function.[94] In addition to these advantages, the flexibility itself is instrumental to the assembly process, as certain complexes may not be assembled successfully from rigid components. The open, extended structure of MoRFs may enable an increased speed of interaction; macromolecular association rates are thought to be substantially improved by an initial, relatively non-specific association enabled by flexible (disordered) recognition segments, mechanistically describe as the ‘‘fly-casting’’ model.[95] Another unique consequence of the structural flexibility of MoRFs is their capability to adapt to the structure of different partners, which allows increased plasticity in signaling interactions. Such a molecular recognition mechanism, which is coupled to the folding process, has been noted to confer exceptional specificity and versatility.[5,16,27,94] All these features help explain the prevalence of structural disorder in signaling and regulatory proteins.[16,27] The interaction of intrinsically disordered proteins with their partners highlights the need and importance of understanding the mechanism of the induced folding process. Since effective functioning of intrinsically disordered proteins requires fast formation of the folded state,[95] their template-induced folding represents a special and interesting case of protein folding.

Materials and Methods

Assembling the MoRF dataset

Protein segments shorter than 70 residues, which were observed to be bound to other proteins longer than 100 residues, were collected from the Protein Data Bank (PDB)[96] using the provided SEQRES data. Then, all sequences containing non-standard residues (X or Z annotations) and all protein chains with lengths of ten residues and shorter were removed and the resulting dataset was subjected to redundancy analysis. Sequence redundancy was eliminated by applying a dynamic sequence identity threshold (described in Results) to each sequence pair and clustering all sequences using the “more similar than identical” rule described by Rost.[32] If the sequence identity between any two MoRFs was higher than a threshold for similarity, they were considered to be structurally similar and hence part of the same cluster. If sequence identity was lower than the threshold then sequence was assigned to a new cluster. Our interest in this formula was largely due to the length restrictions of MoRFs (i.e. ≤70 and wherein using the standard 25% identity cut off would produce many false positives) as this method allowed us to use a "length dependent" threshold calculation method. A representative of each cluster was selected by these criteria. The selected structures were examined using the Swiss PDB Viewer. All structures that appeared to be globular were eliminated, where only four such examples were found. The selection process resulted in 372 chains that make up the MoRF dataset.

Using other database references (Swiss-Prot,[97] PIR,[98] and NCBI referenced in the respective PDB files for each of the MoRF sequences) we were able to extract 301 sequences containing these 372 MoRF chains. All but 53 of the total MoRFs were found to be fragments of larger sequences. The final task after collecting and processing these MoRFs was to design a database for storing the MoRF data. For this, we used MySQL as the backend and Perl scripts to load information about each MoRF such as sequence, secondary structure, binding partner, disorder predictions, etc.

Ordered protein control sets

Three sets of ordered proteins were used as controls in this work. Two of these sets have been described, PDB select 25 (PDB_25)[99] and the ordered protein complex set (OC).[54] PDB_25 is a representative set of chains from the PDB where no two chains have greater that 25% sequence identity. Furthermore, structures were selected based on quality, where no structure in the set has a resolution poorer than 3.5 Å or an *R*-factor greater than 30. The version of PDB_25 used was released in March 2005 and contained 1765 sequences. The OC set is a collection of protein–protein complexes, including both complexes present in solution and artificial complexes due to crystal packing, where the partners have been shown to be ordered prior to complex formation. This set included 26 structures.

The third set, ordered monomers (OM), was derived from the protein quaternary structure (PQS) server.[100] This resource infers the solution oligomeric state of a protein structure based on the surface area buried between individual subunits in the asymmetric unit and symmetry related molecules. Monomeric proteins from X-ray crystal structures were selected and further filtered for structures that contain only a single chain in the asymmetric unit. The SEQRES records were used to cluster these chains for sequence similarity at threshold of 25% sequence identity and 60% coverage using the blastclust program provided by NCBI. The resulting set contained 848 proteins chains.

Compositional profiling

The analysis of amino acid composition in the MoRF dataset was based on the approach recently developed for intrinsically disordered proteins.[5] Briefly, this consists of calculating the fractional difference in composition between the set of proteins being studied and a set of ordered proteins for each amino acid residue. The fractional difference is calculated as $(C_X\text{-}C_{\text{ordered}})/C_{\text{ordered}}$, where C_X is the content of a given amino acid in the set of proteins being studies, and C_{ordered} is the corresponding content in a set of ordered proteins. Subsequently, plots of $(C_X\text{-}C_{\text{ordered}})/C_{\text{ordered}}$ are constructed with the amino acids arrayed from the most rigid to the most flexible according to averaged backbone atom *B*-factor values as determined by Vihinen *et al.*[101] Standard errors were calculated from 200 bootstrap iterations.

Disorder prediction

PONDR® (predictor of natural disordered regions) is a set of neural network predictors of disordered regions based on local amino acid composition, flexibility, hydropathy, coordination number and other factors. These predictors classify each residue within a sequence as either ordered or disordered. PONDR® VL-XT integrates three feed forward neural networks: the variously characterized long, version 1 (VL1) predictor from Romero *et al.* 2001,[38] which predicts non-terminal residues, and the X-ray characterized N and C-terminal predictors (XT) from Li *et al.* 1999,[55] which predicts terminal residues. Output for the VL1 predictor starts and ends 11 amino acids from the termini. The XT predictors output provides predictions up to 14 amino acids from their respective ends. A simple average is taken for the overlapping predictions; and a sliding window of nine amino acids is used to smooth the prediction values along the length of the sequence. Unsmoothed prediction values from the XT predictors are used for the first and last four sequence positions.

PONDR® VL3 combines the predictions of 30 neural networks for the entire protein sequence and was trained using disordered regions from more than 150 proteins characterized by the methods mentioned above plus circular dichroism, limited proteolysis and other physical approaches.[102]

PONDR® VL-XT and VL3 predictions were performed on all of the protein sequences in the database. The resulting disorder score for each amino acid position was stored for later use.

Secondary structure analysis

The predisposition of each MoRF sequence to form secondary structure was assessed by the secondary structure predictor PHD.[30,31] Additionally, information about the secondary structure of the MoRFs in the bound state was extracted from the corresponding PDB files using the DSSP program.[30]

Identification of polyproline type II helices

Using approaches developed earlier by Sreerama *et al.* 1994[51] and Stapley & Creamer[52] we identified polyproline type II (PPII) helices as a stretch containing minimum of four contiguous residues having φ and ψ angles within the regions from 125° to 165° and from −95° to −55°, respectively. The natural restriction of the φ angle of a proline side-chain in any polypeptide within the range from −48° to −78° forms the basis of these φ and ψ ranges.

Identification of post-translational modification sites

Post-translational modification sites in the members of the MoRF dataset were identified by the searching the PROSITE database.[60] In addition, an intrinsic disorder-based algorithm for the prediction of phosphorylation sites, DisPhos,[59] was applied to 305 MoRFs whose lengths were ≥12 residues.

Acknowledgements

The Indiana Genomics Initiative (INGEN) and NIH grant 1 R01 LM007688-0A1 provided support for A.K.D. This work was supported in part by the Programs of the Russian Academy of Sciences for the Molecular and Cellular Biology and Fundamental Science for Medicine (V.N.U.).

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.07.087

References

1. Wright, P. E. & Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**, 321–331.
2. Uversky, V. N., Gillespie, J. R. & Fink, A. L. (2000). Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins: Struct . Funct. Genet.* **41**, 415–427.
3. Dunker, A. K., Obradovic, Z., Romero, P., Garner, E. C. & Brown, C. J. (2000). Intrinsic protein disorder in complete genomes. *Genome Inform. Ser. Workshop Genome Inform.* **11**, 161–171.
4. Dunker, A. K. & Obradovic, Z. (2001). The protein trinity–linking function and disorder. *Nature Biotechnol.* **19**, 805–806.

5. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S. *et al.* (2001). Intrinsically disordered protein. *J. Mol. Graph. Model.* **19**, 26–59.
6. Demchenko, A. P. (2001). Recognition between flexible protein molecules: induced and assisted folding. *J. Mol. Recognit.* **14**, 42–61.
7. Namba, K. (2001). Roles of partly unfolded conformations in macromolecular self-assembly. *Genes Cells*, **6**, 1–12.
8. Dunker, A. K., Brown, C. J. & Obradovic, Z. (2002). Identification and functions of usefully disordered proteins. *Adv. Protein Chem.* **62**, 25–49.
9. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M. & Obradovic, Z. (2002). Intrinsic disorder and protein function. *Biochemistry*, **41**, 6573–6582.
10. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradovic, Z. & Dunker, A. K. (2002). Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* **323**, 573–584.
11. Tompa, P. (2002). Intrinsically unstructured proteins. *Trends Biochem. Sci.* **27**, 527–533.
12. Uversky, V. N. (2002). Natively unfolded proteins: a point where biology waits for physics. *Protein Sci.* **11**, 739–756.
13. Uversky, V. N. (2002). What does it mean to be natively unfolded? *Eur. J. Biochem.* **269**, 2–12.
14. 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.
15. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F. & 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.
16. Dyson, H. J. & Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nature Rev. Mol. Cell Biol.* **6**, 197–208.
17. Fink, A. L. (2005). Natively unfolded proteins. *Curr. Opin. Struct. Biol.* **15**, 35–41.
18. Daughdrill, G. W., Pielak, G. J., Uversky, V. N., Cortese, M. S. & Dunker, A. K. (2005). Natively disordered proteins. In *Handbook of Protein Folding* (Buchner, J. & Kiefhaber, T., eds), pp. 271–353, VCH Wiley-Verlag GmbH and Co. KGaA, Weinheim, Germany.
19. Uversky, V. N., Oldfield, C. J. & Dunker, A. K. (2005). Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. *J. Mol. Recognit.* **18**, 343–384.
20. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M. & Uversky, V. N. (2005). Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS J.* **272**, 5129–5148.
21. Dedmon, M. M., Patel, C. N., Young, G. B. & Pielak, G. J. (2002). FlgM gains structure in living cells. *Proc. Natl Acad. Sci. USA*, **99**, 12681–12684.
22. Oldfield, C. J., Cheng, Y., Cortese, M. S., Brown, C. J., Uversky, V. N. & Dunker, A. K. (2005). Comparing and combining predictors of mostly disordered proteins. *Biochemistry*, **44**, 1989–2000.
23. Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J. & Russell, R. B. (2003). Protein disorder prediction: implications for structural proteomics. *Structure (Camb)*, **11**, 1453–1459.
24. Liu, J. & Rost, B. (2001). Comparing function and structure between entire proteomes. *Protein Sci.* **10**, 1970–1979.
25. Vucetic, S., Brown, C. J., Dunker, A. K. & Obradovic, Z. (2003). Flavors of protein disorder. *Proteins: Struct. Funct. Genet.* **52**, 573–584.
26. Oldfield, C. J., Cheng, Y., Cortese, M. S., Romero, P., Uversky, V. N. & Dunker, A. K. (2005). Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry*, **44**, 12454–12470.
27. Dyson, H. J. & Wright, P. E. (2002). Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.* **12**, 54–60.
28. Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. *Ber. Dt. Chem. Ges.* **27**, 2985–2993.
29. Koshland, D. E. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proc. Natl Acad. Sci. USA*, **44**, 98–104.
30. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
31. Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584–599.
32. Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Eng.* **12**, 85–94.
33. Fiser, A., Dosztanyi, Z. & Simon, I. (1997). The role of long-range interactions in defining the secondary structure of proteins is overestimated. *Comput. Appl. Biosci.* **13**, 297–301.
34. Kihara, D. (2005). The effect of long-range interactions on the secondary structure formation of proteins. *Protein Sci.* **14**, 1955–1963.
35. Romero, P., Obradovic, Z. & Dunker, A. K. (1997). Sequence data analysis for long disordered regions prediction in the calcineurin family. *Genome Informat.* **8**, 110–124.
36. Romero, P., Obradovic, Z., Kissinger, C., Villafranca, J. E. & Dunker, A. K. (1997). Identifying disordered regions in proteins from amino acid sequence. *1997 Proc. Internat. Confer. Neural Networks*, vol. 1, pp. 90–95, Wiley-IEEE Press, Houston, TX.
37. Dunker, A. K., Garner, E., Guilliot, S., Romero, P., Albrecht, K., Hart, J. *et al.* (1998). Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac. Symp. Biocomput.* **3**, 473–484.
38. Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J. & Dunker, A. K. (2001). Sequence complexity of disordered protein. *Proteins: Struct. Funct. Genet.* **42**, 38–48.
39. Vihinen, M. (1987). Relationship of protein flexibility to thermostability. *Protein Eng.* **1**, 477–480.
40. Thornton, J. M. (1981). Disulphide bridges in globular proteins. *J. Mol. Biol.* **151**, 261–287.
41. Burley, S. K. & Petsko, G. A. (1985). Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science*, **229**, 23–28.
42. Williamson, M. P. (1994). The structure and function of proline-rich regions in proteins. *Biochem. J.* **297**, 249–260.
43. Shi, Z., Olson, C. A., Rose, G. D., Baldwin, R. L. & Kallenbach, N. R. (2002). Polyproline II structure in a sequence of seven alanine residues. *Proc. Natl Acad. Sci. USA*, **99**, 9190–9195.
44. Pappu, R. V. & Rose, G. D. (2002). A simple model for polyproline II structure in unfolded states of alanine-based peptides. *Protein Sci.* **11**, 2437–2455.
45. Dukor, R. K. & Keiderling, T. A. (1991). Reassessment of the random coil conformation: vibrational CD study of proline oligopeptides and related polypeptides. *Biopolymers*, **31**, 1747–1761.

46. Wilson, G., Hecht, L. & Barron, L. D. (1996). Residual structure in unfolded proteins revealed by Raman optical activity. *Biochemistry*, **35**, 12518–12525.
47. Park, S. H., Shalongo, W. & Stellwagen, E. (1997). The role of PII conformations in the calculation of peptide fractional helix content. *Protein Sci.* **6**, 1694–1700.
48. Kelly, M. A., Chellgren, B. W., Rucker, A. L., Troutman, J. M., Fried, M. G., Miller, A. F. & Creamer, T. P. (2001). Host-guest study of left-handed polyproline II helix formation. *Biochemistry*, **40**, 14376–14383.
49. Barron, L. D., Blanch, E. W. & Hecht, L. (2002). Unfolded proteins studied by Raman optical activity. *Adv. Protein Chem.* **62**, 51–90.
50. Syme, C. D., Blanch, E. W., Holt, C., Jakes, R., Goedert, M., Hecht, L. & Barron, L. D. (2002). A Raman optical activity study of rheomorphism in caseins, synucleins and tau. New insight into the structure and behaviour of natively unfolded proteins. *Eur. J. Biochem.* **269**, 148–156.
51. Sreerama, N. & Woody, R. W. (1994). Poly(pro)II helices in globular proteins: identification and circular dichroic analysis. *Biochemistry*, **33**, 10022–10025.
52. Stapley, B. J. & Creamer, T. P. (1999). A survey of left-handed polyproline II helices. *Protein Sci.* **8**, 587–595.
53. Adzhubei, A. A. & Sternberg, M. J. (1993). Left-handed polyproline II helices commonly occur in globular proteins. *J. Mol. Biol.* **229**, 472–493.
54. Gunasekaran, K., Tsai, C. J. & Nussinov, R. (2004). Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. *J. Mol. Biol.* **341**, 1327–1341.
55. Li, X., Romero, P., Rani, M., Dunker, A. K. & Obradovic, Z. (1999). Predicting Protein Disorder for N-, C-, and Internal Regions. *Genome Inform. Ser. Workshop Genome Inform.* **10**, 30–40.
56. Garner, E., Romero, P., Dunker, A. K., Brown, C. & Obradovic, Z. (1999). Predicting Binding Regions within Disordered Proteins. *Genome Inform. Ser. Workshop Genome Inform.* **10**, 41–50.
57. Romero, P. R., Zaidi, S., Fang, Y. Y., Uversky, V. N., Radivojac, P., Oldfield, C. J. *et al.* (2006). Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proc. Natl Acad. Sci. USA*, **103**, 8390–8395.
58. Zor, T., Mayr, B. M., Dyson, H. J., Montminy, M. R. & Wright, P. E. (2002). Roles of phosphorylation and helix propensity in the binding of the KIX domain of CREB-binding protein by constitutive (c-Myb) and inducible (CREB) activators. *J. Biol. Chem.* **277**, 42241–42248.
59. Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z. & Dunker, A. K. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucl. Acids Res.* **32**, 1037–1049.
60. Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K. & Bairoch, A. (2002). The PROSITE database, its status in 2002. *Nucl. Acids Res.* **30**, 235–238.
61. Prives, C. & Hall, P. A. (1999). The p53 pathway. *J. Pathol.* **187**, 112–126.
62. Rustandi, R. R., Baldisseri, D. M. & Weber, D. J. (2000). Structure of the negative regulatory domain of p53 bound to S100B(betabeta). *Nature Struct. Biol.* **7**, 570–574.
63. Lee, H., Mok, K. H., Muhandiram, R., Park, K. H., Suk, J. E., Kim, D. H. *et al.* (2000). Local structural elements in the mostly unstructured transcriptional activation domain of human p53. *J. Biol. Chem.* **275**, 29426–29432.
64. Kussie, P. H., Gorina, S., Marechal, V., Elenbaas, B., Moreau, J., Levine, A. J. & Pavletich, N. P. (1996). Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science*, **274**, 948–953.
65. Oliner, J. D., Pietenpol, J. A., Thiagalingam, S., Gyuris, J., Kinzler, K. W. & Vogelstein, B. (1993). Oncoprotein MDM2 conceals the activation domain of tumour suppressor p53. *Nature*, **362**, 857–860.
66. Lowe, E. D., Tews, I., Cheng, K. Y., Brown, N. R., Gul, S., Noble, M. E. *et al.* (2002). Specificity determinants of recruitment peptides bound to phospho-CDK2/cyclin A. *Biochemistry*, **41**, 15625–15634.
67. Schulman, B. A., Lindstrom, D. L. & Harlow, E. (1998). Substrate recruitment to cyclin-dependent kinase 2 by a multipurpose docking site on cyclin A. *Proc. Natl Acad. Sci. USA*, **95**, 10453–10458.
68. Baudier, J., Delphin, C., Grunwald, D., Khochbin, S. & Lawrence, J. J. (1992). Characterization of the tumor suppressor protein p53 as a protein kinase C substrate and a S100b-binding protein. *Proc. Natl Acad. Sci. USA*, **89**, 11627–11631.
69. Wilder, P. T., Rustandi, R. R., Drohat, A. C. & Weber, D. J. (1998). S100B(betabeta) inhibits the protein kinase C-dependent phosphorylation of a peptide derived from p53 in a Ca^{2+}-dependent manner. *Protein Sci.* **7**, 794–798.
70. Machesky, L. M. & Insall, R. H. (1999). Signaling to actin dynamics. *J. Cell Biol.* **146**, 267–272.
71. Chereau, D., Kerff, F., Graceffa, P., Grabarek, Z., Langsetmo, K. & Dominguez, R. (2005). Actin-bound structures of Wiskott-Aldrich syndrome protein (WASP)-homology domain 2 and the implications for filament assembly. *Proc. Natl Acad. Sci. USA*, **102**, 16644–16649.
72. Kim, A. S., Kakalis, L. T., Abdul-Manan, N., Liu, G. A. & Rosen, M. K. (2000). Autoinhibition and activation mechanisms of the Wiskott-Aldrich syndrome protein. *Nature*, **404**, 151–158.
73. Abdul-Manan, N., Aghazadeh, B., Liu, G. A., Majumdar, A., Ouerfelli, O., Siminovitch, K. A. & Rosen, M. K. (1999). Structure of Cdc42 in complex with the GTPase-binding domain of the 'Wiskott-Aldrich syndrome' protein. *Nature*, **399**, 379–383.
74. Rohatgi, R., Ma, L., Miki, H., Lopez, M., Kirchhausen, T., Takenawa, T. & Kirschner, M. W. (1999). The interaction between N-WASP and the Arp2/3 complex links Cdc42-dependent signals to actin assembly. *Cell*, **97**, 221–231.
75. Wu, J. W., Cocina, A. E., Chai, J., Hay, B. A. & Shi, Y. (2001). Structural analysis of a functional DIAP1 fragment bound to grim and hid peptides. *Mol. Cell*, **8**, 95–104.
76. Hawkins, C. J., Wang, S. L. & Hay, B. A. (1999). A cloning method to identify caspases and their regulators in yeast: identification of *Drosophila* IAP1 as an inhibitor of the *Drosophila* caspase DCP-1. *Proc. Natl Acad. Sci. USA*, **96**, 2885–2890.
77. Goyal, L., McCall, K., Agapite, J., Hartwieg, E. & Steller, H. (2000). Induction of apoptosis by Drosophila reaper, hid and grim through inhibition of IAP function. *EMBO J.* **19**, 589–597.
78. Callaghan, A. J., Aurikko, J. P., Ilag, L. L., Gunter Grossmann, J., Chandran, V., Kuhnel, K. *et al.* (2004). Studies of the RNA degradosome-organizing domain

of the *Escherichia coli* ribonuclease RNase E. *J. Mol. Biol.* **340**, 965–979.

79. Kim, T. D., Ryu, H. J., Cho, H. I., Yang, C. H. & Kim, J. (2000). Thermal behavior of proteins: heat-resistant proteins and their heat-induced secondary structural changes. *Biochemistry*, **39**, 14839–14846.
80. Sherr, C. J. & Roberts, J. M. (1995). Inhibitors of mammalian G1 cyclin-dependent kinases. *Genes Dev.* **9**, 1149–1163.
81. Hinnebusch, A. G. & Fink, G. R. (1983). Positive regulation in the general amino acid control of Saccharomyces cerevisiae. *Proc. Natl Acad. Sci. USA*, **80**, 5374–5378.
82. Kanamoto, T., Mota, M., Takeda, K., Rubin, L. L., Miyazono, K., Ichijo, H. & Bazenet, C. E. (2000). Role of apoptosis signal-regulating kinase in regulation of the c-Jun N-terminal kinase pathway and apoptosis in sympathetic neurons. *Mol. Cell. Biol.* **20**, 196–204.
83. Siligardi, G. & Drake, A. F. (1995). The importance of extended conformations and, in particular, the PII conformation for the molecular recognition of peptides. *Biopolymers*, **37**, 281–292.
84. Bochicchio, B. & Tamburro, A. M. (2002). Polyproline II structure in proteins: identification by chiroptical spectroscopies, stability, and functions. *Chirality*, **14**, 782–792.
85. Shi, Z., Woody, R. W. & Kallenbach, N. R. (2002). Is polyproline II a major backbone conformation in unfolded proteins? *Adv. Protein Chem.* **62**, 163–240.
86. Avbelj, F. & Baldwin, R. L. (2003). Role of backbone solvation and electrostatics in generating preferred peptide backbone conformations: distributions of phi. *Proc. Natl Acad. Sci. USA*, **100**, 5742–5747.
87. Uversky, V. N. & Fink, A. L. (2004). Conformational constraints for amyloid fibrillation: the importance of being unfolded. *Biochim. Biophys. Acta*, **1698**, 131–153.
88. Blanch, E. W., Morozova-Roche, L. A., Cochran, D. A., Doig, A. J., Hecht, L. & Barron, L. D. (2000). Is polyproline II helix the killer conformation? A Raman optical activity study of the amyloidogenic prefibrillar intermediate of human lysozyme. *J. Mol. Biol.* **301**, 553–563.
89. Marks, F. (1996). *Protein Phosphorylation.* VCH Weinheim, New York.
90. Sheaff, R. J., Singer, J. D., Swanger, J., Smitherman, M., Roberts, J. M. & Clurman, B. E. (2000). Proteasomal turnover of p21Cip1 does not require p21Cip1 ubiquitination. *Mol. Cell*, **5**, 403–410.
91. David, D. C., Layfield, R., Serpell, L., Narain, Y., Goedert, M. & Spillantini, M. G. (2002). Proteasomal degradation of tau protein. *J. Neurochem.* **83**, 176–185.
92. Liu, C. W., Corboy, M. J., DeMartino, G. N. & Thomas, P. J. (2003). Endoproteolytic activity of the proteasome. *Science*, **299**, 408–411.
93. Cox, C. J., Dutta, K., Petri, E. T., Hwang, W. C., Lin, Y., Pascal, S. M. & Basavappa, R. (2002). The regions of securin and cyclin B proteins recognized by the ubiquitination machinery are natively unfolded. *FEBS Lett.* **527**, 303–308.
94. Gunasekaran, K., Tsai, C. J., Kumar, S., Zanuy, D. & Nussinov, R. (2003). Extended disordered proteins: targeting function with less scaffold. *Trends Biochem. Sci.* **28**, 81–85.
95. Shoemaker, B. A., Portman, J. J. & Wolynes, P. G. (2000). Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc. Natl Acad. Sci. USA*, **97**, 8868–8873.
96. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
97. Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45–48.
98. Wu, C. H., Yeh, L. S., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y. *et al.* (2003). The Protein Information Resource. *Nucl. Acids Res.* **31**, 345–347.
99. Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522–524.
100. Henrick, K. & Thornton, J. M. (1998). PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* **23**, 358–361.
101. Vihinen, M., Torkkila, E. & Riikonen, P. (1994). Accuracy of protein flexibility predictions. *Proteins: Struct. Funct. Genet.* **19**, 141–149.
102. Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C. J. & Dunker, A. K. (2003). Predicting intrinsic disorder from amino acid sequence. *Proteins: Struct. Funct. Genet.* **53** (Suppl. 6), 566–572.

(Received 16 February 2006; received in revised form 8 July 2006; accepted 31 July 2006)
Available online 4 August 2006