

# INTRINSIC PROTEIN DISORDER AND PROTEIN-PROTEIN INTERACTIONS<sup>†</sup>

WEI-LUN HSU, CHRISTOPHER OLDFIELD, JINGWEI MENG, FEI HUANG,  
BIN XUE<sup>#</sup>, VLADIMIR N. UVERSKY<sup>#</sup>, PEDRO ROMERO, AND A. KEITH DUNKER<sup>\*</sup>

*Department of Biochemistry and Molecular Biology, Indiana University School of Medicine  
410 W 10<sup>th</sup> street, Suite 5000  
Indianapolis, IN 46202*

*Emails: {hsu20, cjoldfie, menj, huangfei, promero, or kedunker} @iupui.edu  
{binxue or vuversky} @health.usf.edu*

Intrinsically disordered proteins often bind to more than one partner. In this study, we focused on 11 sets of complexes in which the same disordered segment becomes bound to two or more distinct partners. For this collection of protein complexes, two or more partners of each disordered segment were selected to have less than 25% amino acid identity at structurally aligned positions. As it turned out that most of the examples so selected had similar 3D structure, the studied set was reduced to just these similar-fold cases. Based on the analyses of the interacting partners, the average sequence identity of the partners' binding regions showed substantially higher conservation as compared to the nonbinding regions: The residue identities, averaged over the 11 sets of partner proteins, were as follows: binding residues,  $42 \pm 6\%$ ; nonbinding residues  $20 \pm 3\%$ ; nonbinding buried residues  $26 \pm 5\%$ ; and nonbinding surface residues  $16 \pm 3\%$ . The higher sequence identity of the binding residues compared to the other sets of residues provides evidence that these observed interactions are likely to be meaningful biological interactions, not artifacts. Since many of the features of the various interactions indicate that the disordered binding segments were likely to have been disordered before binding, these results also add further weight to the existence and function of intrinsically disordered regions inside cells.

*Keywords: Molecular recognition, protein interaction, sequence conservation*

## 1. Introduction

Many proteins execute their biological functions through protein-protein interactions. By binding to interacting partners, proteins can deliver signals to other molecules. For example, hormone neurotransmitters and their receptors trigger various signal transduction pathways following their mutual interaction, antibody recognition of peptide antigens leads to B-cell activation, and the interaction between G-protein coupled receptors and G-proteins leads to the transduction of many signals.

---

<sup>†</sup> This work is supported by the grants R01 LM007688-01A1, GM071714-01A2 from the National Institute of Health and EF 0849803 from the National Science Foundation

<sup>\*</sup> Corresponding author

<sup>#</sup> Present Address: Department of Molecular Medicine, University of South Florida, Tampa, FL 33620

Mapping protein-protein interactions leads to networks that are far from random. A small number of highly connected proteins, called hubs, are observed, while most proteins have only a few connections [1, 2]. We proposed that hubs often use disordered regions to bind to multiple partners (one-to-many signaling), and we also proposed that structured hubs often bind to many partners via their partner's disordered regions (many-to-one signaling) [3]. The common involvement of disorder in hub proteins' interactions has been supported by several subsequent studies [4-6]. Our decision to test whether hub proteins depend on disorder was motivated by prior experiments showing that conformational disorder enabled one particular protein region to bind to multiple partners [7].

These short, partner-binding segments found within disordered regions have been studied under various names, such as eukaryotic linear motifs (ELMs) [8], short linear motifs (SLiMs) [9], ANCHOR-algorithm-indicated binding sites [10], and molecular recognition features (MoRFs) [11-13] among others. ELMs and SLiMs are based on sequence motifs, while ANCHOR-indicated sites and MoRFs both use features from disorder predictors. The motif-based and algorithmic approaches show significant overlap in their identification of their binding sites [14], suggesting that the different approaches associated with the different names are merely emphasizing different aspects of the same types of binding interactions.

The C-terminal region of p53 uses disorder to bind to more than 45 different proteins and to form a tetramer, but only six of these complexes and the tetramer have had their structures deposited in the Protein Data Bank (PDB) [15]. One particular p53 segment "SHLKSKKGQSTSRHKKLMFKTE" (residues 367-388), which is both an ELM and a MoRF and which is located at the C-terminus, morphs into an  $\alpha$ -helix when binding with S100 $\beta$ , into a  $\beta$ -sheet with sirtuin, into an irregular structure with CREB binding protein (CBP) and into another irregular structure with cyclin A2 as a partner [15].

Very different biological processes are transduced via these four different interactions involving the same segment of p53: The CDK2/cyclin A2 complex regulates progression of S phase of the eukaryote cell cycle by recognizing diverse but structurally constrained target sequences (KXL/RXL motif) from various substrates, including p53 [16]; deacetylase enzymes like the Sir 2 protein, which is a homologue of Sirtuin, can lead to down-regulation of p53-dependent transcription by binding to the acetylated p53 peptide on lysine 382 [17]; the recognition of acetylated lysine 382 in p53 by the conserved bromo-domain of transcriptional coactivator CBP is very specific, leading to the recruitment of p53 acetylation-dependent coactivator following DNA damage and to the activation of cyclin-dependent kinase inhibitor p21 [18]; dimeric S100 calcium binding protein B can sterically block the phosphorylation and acetylation sites of on p53 that are critical for the activation important transcription; finally, the peptide derived from the region of p53 was found to undergo a disorder-to-order conformational change while binding to Ca<sup>2+</sup> loaded S100 $\beta$  [19]. Thus, this same intrinsically disordered segment plays roles in a diverse set of signaling pathways.

Besides p53 other MoRFs that bind to two or more partners and that have structures in PDB have not been systematically compared to understand how disorder can bind to multiple partners. To study more examples, we searched for other instances in PDB in which one disordered segment binds to more than one partner. We found 11 such sets for which the partners have both

superimposable 3D structures and pairwise sequence identities lower than 25%. As we show, the flexibility of the disordered regions is important for enabling the same protein segment to bind to alternative partners, thus extending our observations on p53 [15] to 11 other intrinsically disordered protein regions.

## 2. Results

### 2.1 Disordered hub dataset

Starting from 4,289 MoRF-containing complexes from PDB, various criteria were applied to extract those examples with the same MoRF sequences bound to collections of different globular partners, such that the partners to a particular MoRF exhibit less than 25% sequence identity. We identified 26 multi-partner sets having a total of 67 complexes. The point of the 25% sequence identity filter was to identify examples in which one MoRF binds to structurally dissimilar partners. Still, for 18 of these 26 examples, the MoRF partners were structurally superimposable even though their sequence identities were quite low and so, only those examples with similar fold were selected for our data set.

The RMSD values and the fraction of the total residues that gave good structural alignments were estimated for the sets of partners for the 18 MoRFs. The 11 sets that had both a relatively low RMSD (2.33 to 5.49 Å) and a relatively high fraction with good structural alignment (0.79 – 0.91), herein called coverage, are given in Table 1. The partners of the DNA repair protein have a reasonable RMSD (3.53) but a low coverage (0.36). This protein was not discarded because one of the partners had a large, non-alignable extra domain that was responsible for the low coverage.

Table 1. Disordered hub examples with structurally similar partners and low sequence identities

Disordered hub examples	Bound conformation					RMSD	Coverage
	Num	Helix	Sheet	Coil	Complex		
Nuclear receptor corepressor 2*	3	2	0	1	0	3.43	0.85
Thyroid receptor associated protein 220 (TRAP 220)*	3	3	0	0	0	3.05	0.91
Nuclear receptor coactivator 1*	2	2	0	0	0	5.49	0.85
Nuclear receptor 0B2 – near N-term*	2	2	0	0	0	3.74	0.86
Troponin I, cardiac muscles	2	0	0	1	1	3.01	0.79
Nuclear receptor 0B2 – near C-term*	2	1	0	1	0	3.88	0.80
Cell death protein GRIM	2	0	2	0	0	2.33	0.79
Beclin-1	2	2	0	0	0	4.10	0.84
BCL-2-like protein 11	2	2	0	0	0	2.72	0.90
Alzheimer’s disease amyloid A4 protein homolog	2	0	0	2	0	2.93	0.84
DNA repair protein	2	0	0	2	0	3.53	0.36

\*5 of 11 disordered hub examples belonged to the family of coregulatory proteins of nuclear receptor (NR), including thyroid receptor associated protein 220 (TRAP 220)

## ***2.2 Functional consequences of MoRF (or ELM) binding***

Interestingly, our dataset indicates that protein disorder is involved in coregulatory proteins of nuclear receptors (NR) such as coactivators and corepressors, making it possible for them to perform one-to-many signaling and to function as disordered hubs. The nuclear receptors (NRs) are a super-family of proteins, associated with other coregulatory proteins involved in the direct mediation and control of the expression of specific gene transcription in response to sensing the presence of hormones and other molecules. Recent data shows that, in addition to direct activation of the basal transcription machinery, nuclear receptors inhibit or enhance transcription by attracting an array of coactivator or corepressor proteins to the transcription complex.

NRs may be classified into two broad categories according to their mechanism of action and subcellular distribution in the absence of ligands. Ligands bind to type I NRs in the cytosol resulting in the dissociation of heat shock proteins, the formation of homo-dimers, translocation from cytoplasm into the cell nucleus, and binding to hormone response elements. Type I NRs include NR subfamily 3, which encompass androgen receptors, estrogen receptors, glucocorticoid receptors and progesterone receptors. Type II NRs, in contrast to type I NRs, are retained in the nucleus and heterodimerize upon binding to DNA in the absence of a ligand, when type II NRs are usually bound to a corepressor. Ligand binding to type II NRs triggers the dissociation of the corepressor, leading to the initiation of transcription by the coactivator. Type II NRs include NR subfamily 1, and receptor molecules such as retinoic acid receptor (RAR), retinoid x receptor (RXR), thyroid hormone receptor (TR) and vitamin D receptor (VDR).

Peroxisome proliferator-activated receptor-binding protein (PBP), also known as thyroid hormone receptor-associated protein 220 (TRAP 220), is an anchor for multi-subunit mediator transcription complex. It functions as a transcription coactivator for nuclear receptors. These coactivator proteins often exhibit histone acetyltransferase (HAT) activity, which weakens the association of the histone to DNA, therefore promoting gene transcription. Three MoRFs in PBP have been found to be involved in multiple interactions with various type II nuclear receptors, such as vitamin D3 receptor (VDR), retinoic acid receptor-beta (RAR-beta) and retinoid x receptor-alpha (RXR-alpha) in our dataset. RAR or VDR, when forming a heterodimer with RXR, can bind to hormone response elements, forming a complex with corepressor protein in the absence of any ligands. When a ligand acting as agonist binds to RAR or VDR, it results in the dissociation of the corepressor and recruitment of coactivator which in turn, promotes transcription of downstream target gene.

When gene transcription is repressed by nuclear receptors, it is mediated by interactions with corepressor proteins. This reaction, in turn attracts histone deacetylases (HDACs) to the chromatin, triggering the strong binding of histone to DNA; thus repressing gene transcription. The antagonist further reinforces the binding of corepressor to the nuclear receptor. MoRF mechanisms are also involved in the down regulation of target gene expression when the nuclear receptor corepressor 2 binds to related nuclear receptors such as peroxisome proliferator activated receptor (PPAR), estrogen related receptor gamma (ERR-gamma) and progesterone receptor (PR) [20-22].

In the previous two examples, such coregulatory proteins can interact with various receptors with low sequence identity but high structure similarity using the same MoRF region. The configurations of secondary structures in those MoRFs bound to receptors are also comparable. Further experiments and analyses provided us more detailed and specific explanations regarding how disordered regions facilitate the binding diversity in different complex structures. The analyses of solvent surface area profiles from the examples with structurally different partners, shows different interfaces accommodate a variety of binding partners and those overlapping residues in interfaces bind to different molecules to varied extents. Otherwise, analogous binding profiles were observed within our 11 examples with similar partnerships.

### ***2.3 Binding to multiple partners, conservation at structure-matching sites***

In order to answer if having similar binding patterns mean MoRFs tend to bind a specific set of residues on partners, we analyzed the sequence identities of binding regions and nonbinding regions on partner side to determine whether the interacting residues are more selected during evolution (Table 2). Those interacting residues tend to be selected to form connection with same MoRF using similar binding patterns as we expected. The sequence conservation of the binding region is significantly higher than those in other parts of the protein.

Table 2. Sequence identities of binding regions and nonbinding regions in partners of 11 disordered hub examples based on MultiProt structural alignment

Disordered hub examples	B	NB	NB_B	NB_E	Overall
Nuclear receptor corepressor 2	0.36	0.17	0.22	0.15	0.21
Thyroid receptor associated protein 220	0.51	0.23	0.26	0.19	0.24
Nuclear receptor coactivator 1	0.62	0.16	0.22	0.08	0.16
Nuclear receptor 0B2 – near N-term	0.47	0.20	0.30	0.16	0.19
Troponin I, cardiac muscles	0.36	0.34	0.62	0.29	0.25
Nuclear receptor 0B2 – near C-term	0.27	0.16	0.13	0.25	0.12
Cell death protein GRIM	0.33	0.28	0.60	0.09	0.26
Beclin-1	0.45	0.17	0.26	0.06	0.17
BCL-2-like protein 11	0.44	0.16	0.26	0.11	0.21
Alzheimer’s disease amyloid A4 protein homolog	0.33	0.19	0.33	0.11	0.15
DNA repair protein	0.43	0.27	0.33	0.11	0.07
	0.42±0.06	0.20±0.03	0.26±0.05	0.16±0.03	0.19±0.04

The values in columns labeled B, NB, NB\_B, NB\_E, and Overall give the sequence identities of binding, nonbinding, nonbinding buried, nonbinding surface and all residues, respectively.

The overall mean identity and confidence interval for structurally aligned binding and non-binding residues are  $42 \pm 6\%$  and  $20 \pm 3\%$ , respectively, as shown in Table 2. These averages were taken over structurally matching residues. In comparison, for a collection of enzymes with  $\sim 25\%$  sequence identity, the active site residues exhibit sequence identities in the range of 43% to

70% [23]. Interestingly, the binding residues being discussed here for several of the proteins show sequence identities that overlap the lower part of the observed range for enzyme active site residues, which are known to have a high tendency to be conserved. Note also that the binding residues show a much higher conservation for the aligned residues as compared to the nonbinding surface residues (column B versus NB\_E) or to the nonbinding buried residues (column B versus NB\_B).

Three examples were chosen from our 11 sets in order to assess in more detail how a disordered region uses its conformational flexibility to form interactions with similar but not identical binding pockets. The three examples can be described as an  $\alpha$ -MoRF, a  $\beta$ -MoRF and an irregular-MoRF corresponding to the thyroid receptor associated protein 220 (TRAP220), the cell death protein GRIM and the Alzheimer's disease amyloid A4 protein homolog, respectively. Figure 1 shows the interacting residues and binding sites of the three selected cases.

Figure 1.a shows four important residues (M5, L6, L9 and L10) on TRAP220 stretching into the clefts on the surfaces of receptor proteins with small structural variations. On the contrary, those residues on the non-buried side of the  $\alpha$ -MoRF have larger conformation fluctuations over the three complexes.

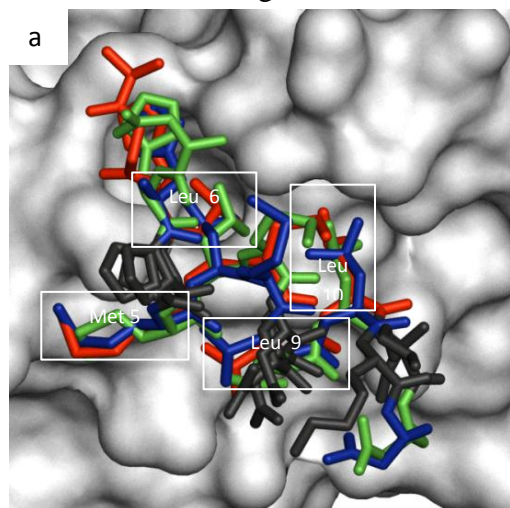
Alanine 1 and tyrosine 4 contribute most of the buried surface areas to the interaction of GRIM and IAP1 (Figure 1.b). The tyrosine side chain makes a huge rotation to fit distinct cavities of IAP1 while the backbone's  $\beta$ -sheet conformation and key interactions didn't change too much (Figure 1.b.1 and 1.b.2).

The MoRF in amyloid A4 protein is mostly coil but has local regions that could be classified as sheet and helix according the DSSP. Its central region (NPTY) adopts coiled conformation and maintains comparably similar structure (colored in black) in both complexes (Figure 1.c). The N-terminus (NGYE) of the green MoRF (Figure 1.c.1) stays in a coiled structure while the N-terminus of the red MoRF in (Figure 1.c.2) turns into  $\beta$ -strand to form an anti-parallel  $\beta$ -sheet with another strand on DAB1 protein. The spatial arrangement of a tyrosine 4 was observed to change substantially, suggesting that this change may facilitate the binding to two different surfaces by the same sequence.

In addition to gathering general evolutionary information for the whole interfaces that MoRFs associate with, some further calculations were carried out on the same three selected examples in order to explore more details at the residue level. We compared the partner residues with which each MoRF residue associates to determine if partner residue variability correlates with overall conformational variation of MoRF itself (Figure 2.a, 2.b and 2.c). Our hypothesis was: the more diverse the amino acids with which a MoRF residue associates, the greater the structural variability of the MoRF backbone. However, the correlation analyses between diversity of partner residues (not shown in Figure 2) and averaged root mean square standard deviation (RMSD) on C-alpha atom (line plots in Figure 2) did not show an obvious and strong correlation. While our particular hypothesis was not supported, Figure 2 nevertheless contains interesting summary information regarding the changes that are observed when one MoRF binds to multiple partners.

Note alternating burial and nonburial of residues (Figure 2.a); this pattern can be traced to the  $\alpha$ -helical structure (e.g. an  $\alpha$ -MoRF) of the thyroid receptor associated protein 220. Also notice that the buried residues are more hydrophobic, and, except for proline, the nonburied residues are more hydrophilic. While hydrophobic, proline is often found on the surfaces of proteins and, furthermore, frequently occupies positions near the amino-terminal ends of helices.

Several interesting points were found from the sequence and structure analyses of the two residues with the greatest contributions of solvent surface area in the GRIM-IAP1 interaction



(Figure 2.b). Although the side chain of alanine 1 is relatively small, a fairly large area becomes buried into the interface. More detailed analysis shows that, for this residue, not only the side chain participates the formation of interaction, but backbone atoms also play a significant role.

Big rotations on side chains also related to higher backbone structure variations, such as the cases on L9 in the  $\alpha$ -MoRF (Figure 2.a), Y4 in  $\beta$ -MoRF (Figure 2.b) and Y4 in irregular-MoRF (Figure 2.c).

The low RMSD of P7 in the irregular-MoRF example (Figure 2.c) may correlate with its capping function in the edge of helix in the X11 protein.

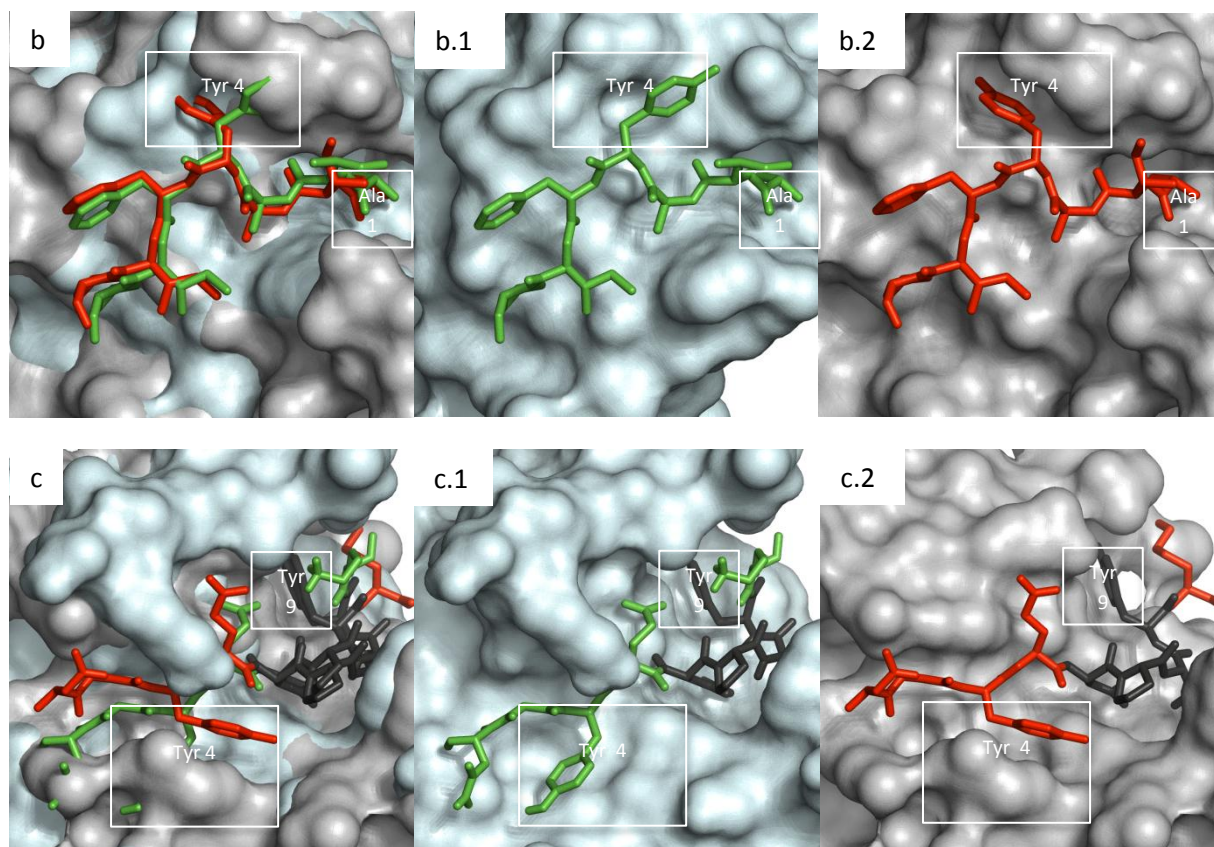


Fig.1. Conformational changes and variations of an identical MoRF binding to its structure-homologous partners from the three selected examples. Different MoRFs are shown in different colors. (a) A fragment from TRAP220 forming  $\alpha$ -MoRFs to associate with VDR, RAR-beta and RXR-alpha. Those residues on the exposed side of helices are colored in black. (b) The binding sites of a  $\beta$ -MoRF from GRIM and apoptosis 1 inhibitor (IAP1). (c) The irregular-MoRF in amyloid A4 protein adapts a highly flexible structure to accommodate the binding pockets of DAB1 and X11. Four structurally constrained residues with lowest RMSD are shown in black.

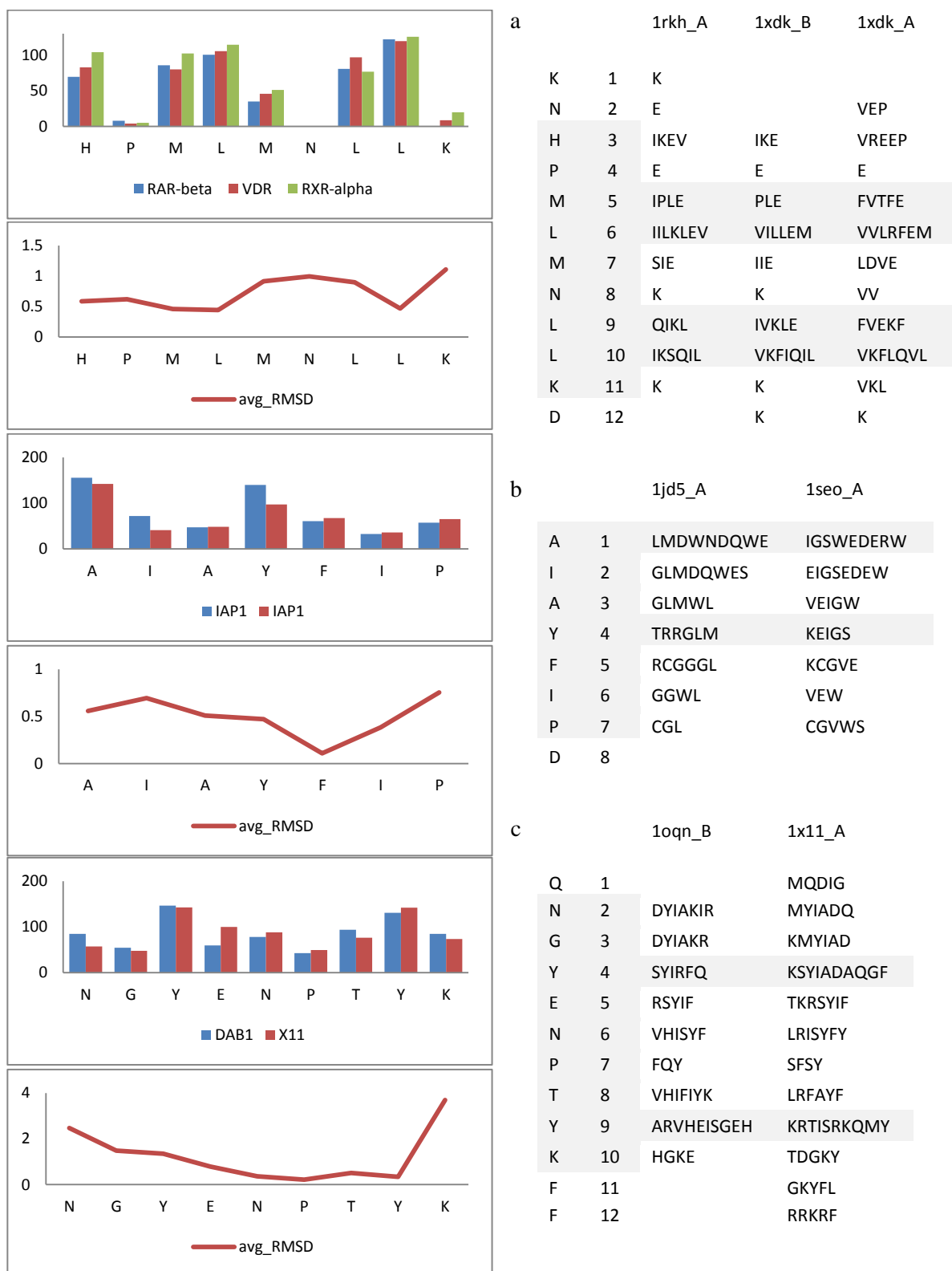


Fig.2. Solvent surface area profiles (bar graphs), averaged RMSD (line plots), and interacting residues on various globular partners (right) calculated on each residue from the three selected MoRF sets: (a)  $\alpha$ -MoRF: thyroid receptor associated protein 220, (b)  $\beta$ - MoRF: Cell death protein GRIM and (c) irregular-MoRF: Alzheimer's disease amyloid A4 protein homolog. Notice in group (c) the specific interactions between 2 aromatic groups that oriented themselves orthogonally.



### 3. Discussion

When the idea of hub-based “scale free” protein-protein interaction networks was first proposed [1], a News and Views article pointed out that such multiple interactions were unlike what had been studied up to that time and that an understanding of these multiple interactions would likely require the discovery of new concepts [24]. We immediately tried to suggest that the new principle was likely the use of disordered proteins by means of coupled binding and folding, which had been previously suggested for protein-DNA interactions [25] as well as for one protein binding to several partners [26], but publication of these ideas for hub proteins was delayed somewhat [3]. By now there is strong evidence that, at least for many if not all hub-partner interactions, disorder plays an important role in enabling one protein to bind to multiple partners [3-6].

As we have shown here, one MoRF can bind to multiple partners and, through these sets of interactions, a small region of one protein can play a role in multiple signaling events, thereby affecting several different cellular functions. Limitations on the availability of multiple interactions in PDB led to a small dataset, which, nevertheless, showed consistent results in terms of residue conservation in the binding partners. By restricting sequence identity of binding partners to 25% we were hoping to find MoRFs binding to structurally diverse partners, but it turned out that in most cases the partner’s folds were very much conserved, despite the sequence differences. Aromatic residue side chain re-orientation was shown to be important in two and three interactions when the backbone conformation remained constant. Overall, these results paint a more detailed picture of multiple interactions than what has been available to date, and support the notion of intrinsic disorder or flexibility as an important factor in the development of non-random protein interaction networks characterized by multiple binding partners.

Despite the biological and structure-function importance of these disorder-based multiple protein interactions, there have been surprisingly few studies indicating in detail how such multiple interactions are brought about. Investigating more examples in detail is needed to provide a clearer and more general picture of how the one MoRF sequence can bind to two or more different partners.

To broaden our understanding, it makes sense to study different collections of single proteins binding to multiple partners and to study multiple proteins binding to the same partner. Previously we investigated one segment binding to completely unrelated partners (e.g., one-to-many signaling) and a collection of unrelated disordered partners binding to a single binding site on one structured protein (e.g. many-to-one signaling) [15]. Rather than focusing on individual examples as before [15], here we studied a collection of MoRFs involved in binding to more than one partner.

A distinctive feature of this study is the partners to a given MoRF had a sequence identity less than 25% yet displayed the same overall fold. In general, the mutations on one partner are frequently linked to mutations on the other partner, indicating structural compensation or coadaptation across the binding interface [26]. However, the interacting protein pairs we collected here are a special set in that only the partners show amino acid substitutions in their sequences, whereas the MoRFs’ sequences are unchanged.

The binding sites on the structured partners may also bind additional disordered sequences that have amino acid substitutions (e.g. many-to-one signaling). If such complexes exist, it would be interesting to determine whether the amino acid changes in the MoRFs compensate for the already observed amino acid changes in the structured binding partners. We have a collection of many-to-one examples (manuscript in preparation), so we can search this set of interactions to determine if any of the structured proteins in the many-to-one collection match any of the structured partners

discussed herein. Such a finding would not only provide information about possible mutation compensation across protein-protein interaction interfaces involving disordered protein regions, but would also suggest new concepts with regard to the structural basis of protein-protein interaction networks.

The indispensability of hub proteins is apparent, as they appear to evolve more slowly and are more likely to be vital for survival [1]. MoRF-protein interactions likely have the combination of high specificity coupled with low affinity [12]. The latter property may facilitate the discovery of small drug molecules that block the interactions. This study and others like it have the potential for a new strategy for drug discovery, namely to search for molecules that selectively block certain protein-protein interactions involving a given protein but not others, by taking advantage of different conformations in the different interactions. This would allow the development of drugs that target specific pathways or even particular pathways in particular tissue types.

Observations have been made that the residues in enzyme active sites tend to evolve more slowly than other parts of the same proteins [23]. We wondered whether the same trend would also be found in the binding sites of the structured partners. By analyzing the 11 sets of interactions that we collected, we found that, like the active-site residues of enzymes, the binding residues of the structured partners exhibited a higher conservation as compared to the non-binding residues.

A recent study showed that protein-protein interactions in which a disordered region binds to a structured partner often involves interactions between two aromatic groups, for which the aromatic residues are frequently not stacked but rather oriented in such a way that a hydrogen of one aromatic ring points towards the centers of the conjugated electron rings of the other [27]. In agreement with this study, many of the protein-protein interactions investigated herein do indeed involve interacting aromatic residues (Figure 2.c), but specific examples without such aromatic-aromatic interactions were also found (Figure 2.a).

Interactions between globular proteins and MoRFs often contain disordered residues as part of the MoRFs [11]. Others have shown that, even though such local regions remain unstructured, they can still affect the overall binding affinity. Such “fuzzy complexes” thus bind to their partners without undergoing complete conversion to structure with the regions that remain disordered still contributing to the energetics of the interaction [28]. From what we have shown here, a search for fuzzy complexes involved in one-to-many and many-to-one signaling could shed new light on these novel and interesting protein-protein “flexible nets”.

## **4. Methods**

### ***4.1 Disordered hub dataset***

Our disordered hub dataset was extracted from Protein Data Bank (PDB) released on March 28, 2008 by analyzing the complex structures that have short protein fragments bound to globular structured partners. The lengths of MoRFs and their globular structured partners were restricted as 5 to 25 and greater than 70, respectively. Only those interactions whose changes of solvent surface area ( $\Delta$ ASA) upon binding are greater than  $400\text{\AA}^2$  were considered as biological interactions. Since our goal is to gather all the complexes with MoRF mechanisms which use a specific region to bind multiple structured partners like p53, Fasta program was used to align each MoRF sequence to UniProt database in order to pinpoint its position in its parent sequence (Expectation value was set at 1000). Following that, we only kept those MoRFs which had overlapping regions

in their parent sequence mapping and used a cluster algorithm (wherein at least one residue overlapped with the rest of the molecules in the same cluster).

As our research is focused upon those MoRFs from the same disordered region which bind to structurally different partners, we used a blast-cluster program to remove any redundant structured partners in our dataset based on 100% and 25% sequence identity. That meant that those specific MoRFs are in one disordered region, but they may use distinct residues to form bonding with different structured partners. After examination of the entire dataset, several unexpected cases were removed, such as those cases involving one MoRF interacting with more than one partner in a single PDB entry or a partner molecule which may be a part of a subset of the same cluster.

#### 4.2 Sequence and Structure analysis

We previously differentiated the MoRFs into 4 different types ( $\alpha$ ,  $\beta$ ,  $\iota$  and complex) based on their secondary structure type which has the largest percentage value of the four types mentioned above [11]. If there is no clear preponderance of any one secondary type (which is at least 1% greater than the other 2 types), we called it “a complex-MoRF”. Only the residues on the interface were counted. The secondary structure assignment program used here was the DSSP program. The root mean square deviation (RMSD) of pairwise proteins was calculated by CEalign [29]. The coverage of alignable region is calculated by length of aligned regions dividing by average length of all sequences. The transposed coordinates and multiple structure alignments were generated by MultiProt algorithm [30] using the complex structures including both MoRF and partner. Sequence identity calculations are based on the structure alignments.

Both residues in each aligned pair were compared to see if they are both in the binding or nonbinding region. The alignment will be considered identical only when the position in both proteins is assigned in the same class: either binding or nonbinding. For the case with more than 2 partners, we averaged all the identities together. Those aligned residues not consistent with their binding/nonbinding status (one is on binding region, but the other one is not) will be classified into another category that didn't show on table 2. Here, those residues with higher solvent surface changes (greater than  $1 \text{ \AA}^2$ ) will be considered as interacting residues. Error bars that represent the 95% confidence interval (CI) of a mean are approximated from 3000 random samplings with replacement generated by the bootstrapping method. The molecular images in Figure 1 were generated by PyMol software.

#### References

1. Jeong, H., et al., *Lethality and centrality in protein networks*. Nature, 2001. **411**(6833): p. 41-2.
2. Barabasi, A.L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization*. Nat Rev Genet, 2004. **5**(2): p. 101-13.
3. Dunker, A.K., et al., *Flexible nets. The roles of intrinsic disorder in protein interaction networks*. FEBS J, 2005. **272**(20): p. 5129-48.
4. Patil, A., K. Kinoshita, and H. Nakamura, *Hub promiscuity in protein-protein interaction networks*. Int J Mol Sci, 2010. **11**(4): p. 1930-43.
5. Kim, P.M., et al., *The role of disorder in interaction networks: a structural analysis*. Mol Syst Biol, 2008. **4**: p. 179.
6. Haynes, C., et al., *Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes*. PLoS Comput Biol, 2006. **2**(8): p. e100.
7. Kriwacki, R.W., et al., *Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity*. Proc Natl Acad Sci U S A, 1996. **93**(21): p. 11504-9.

8. Gould, C.M., et al., *ELM: the status of the 2010 eukaryotic linear motif resource*. Nucleic Acids Res, 2010. **38**(Database issue): p. D167-80.
9. Edwards, R.J., N.E. Davey, and D.C. Shields, *SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins*. PLoS One, 2007. **2**(10): p. e967.
10. Dosztanyi, Z., B. Meszaros, and I. Simon, *ANCHOR: web server for predicting protein binding regions in disordered proteins*. Bioinformatics, 2009. **25**(20): p. 2745-6.
11. Mohan, A., et al., *Analysis of molecular recognition features (MoRFs)*. J Mol Biol, 2006. **362**(5): p. 1043-59.
12. Oldfield, C.J., et al., *Coupled folding and binding with alpha-helix-forming molecular recognition elements*. Biochemistry, 2005. **44**(37): p. 12454-70.
13. Cheng, Y., et al., *Mining alpha-helix-forming molecular recognition features with cross species sequence alignments*. Biochemistry, 2007. **46**(47): p. 13468-77.
14. Fuxreiter, M., P. Tompa, and I. Simon, *Local structural disorder imparts plasticity on linear motifs*. Bioinformatics, 2007. **23**(8): p. 950-6.
15. Oldfield, C.J., et al., *Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners*. BMC Genomics, 2008. **9** **Suppl 1**: p. S1.
16. Lowe, E.D., et al., *Specificity determinants of recruitment peptides bound to phospho-CDK2/cyclin A*. Biochemistry, 2002. **41**(52): p. 15625-34.
17. Avalos, J.L., et al., *Structure of a Sir2 enzyme bound to an acetylated p53 peptide*. Mol Cell, 2002. **10**(3): p. 523-35.
18. Mujtaba, S., et al., *Structural mechanism of the bromodomain of the coactivator CBP in p53 transcriptional activation*. Mol Cell, 2004. **13**(2): p. 251-63.
19. Rustandi, R.R., D.M. Baldisseri, and D.J. Weber, *Structure of the negative regulatory domain of p53 bound to S100B(beta-beta)*. Nat Struct Biol, 2000. **7**(7): p. 570-4.
20. Wang, L., et al., *X-ray crystal structures of the estrogen-related receptor-gamma ligand binding domain in three functional states reveal the molecular basis of small molecule regulation*. J Biol Chem, 2006. **281**(49): p. 37773-81.
21. Madauss, K.P., et al., *A structural and in vitro characterization of asoprisnil: A selective progesterone receptor modulator*. Molecular Endocrinology, 2007. **21**(5): p. 1066-1081.
22. Xu, H.E., et al., *Structural basis for antagonist-mediated recruitment of nuclear co-repressors by PPAR alpha*. Nature, 2002. **415**(6873): p. 813-817.
23. Grishin, N.V. and M.A. Phillips, *The subunit interfaces of oligomeric enzymes are conserved to a similar extent to the overall protein sequences*. Protein Sci, 1994. **3**(12): p. 2455-8.
24. Hasty, J. and J.J. Collins, *Protein interactions. Unspinning the web*. Nature, 2001. **411**(6833): p. 30-1.
25. Spolar, R.S. and M.T. Record, Jr., *Coupling of local folding to site-specific binding of proteins to DNA*. Science, 1994. **263**(5148): p. 777-84.
26. Fares, M.A., M.X. Ruiz-Gonzalez, and J.P. Labrador, *Protein coadaptation and the design of novel approaches to identify protein-protein interactions*. IUBMB Life, 2011. **63**(4): p. 264-71.
27. Espinoza-Fonseca, L.M., *Aromatic residues link binding and function of intrinsically disordered proteins*. Molecular BioSystems, 2011.
28. Tompa, P. and M. Fuxreiter, *Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions*. Trends Biochem Sci, 2008. **33**(1): p. 2-8.
29. Shindyalov, I.N. and P.E. Bourne, *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path*. Protein Eng, 1998. **11**(9): p. 739-47.
30. Shatsky, M., R. Nussinov, and H.J. Wolfson, *A method for simultaneous alignment of multiple protein structures*. Proteins, 2004. **56**(1): p. 143-56.