INTRINSICALLY DISORDERED PROTEINS IN MOLECULAR RECOGNITION AND STRUCTURAL PROTEOMICS

Christopher John Oldfield

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the School of Informatics and Computing,

Indiana University

May 2014

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____

Sarath Chandra Janga, Ph.D., Chair

_____

A. Keith Dunker, Ph.D.

_____

Li Shen, Ph.D.

February 25, 2014

_____

Yuni Xia, Ph.D.

_____

Vladimir N. Uversky, Ph.D.

ACKNOWLEDGEMENTS

A. Keith Dunker, Ya-Yue Van Dunker, and Vladimir N. Uversky for the opportunities, advice, encouragement, and patience.

My family, John, Sharon, and Brian Oldfield, and Rachel Gilbert, for the encouragement and support.

Colleagues for the fruitful collaborations and interesting discussions, including Yugong Cheng, Marc S. Cortese, Wei-Lun Hsu, Fei Huang, Lilia M. Iakoucheva, Tanguy LeGall, Caron Morales, Vladimir Vacic, Maya Wagle, and Bin Xue.

Eli Lilly, NIH, IUPUI School of Informatics, and Molecular Kinetics, Inc. for financial support.

Christopher John Oldfield

INTRINSICALLY DISORDERED PROTEINS IN MOLECULAR RECOGNITION AND STRUCTURAL PROTEOMICS

Intrinsically disordered proteins (IDPs) are abundant in nature, being more prevalent in the proteomes of eukaryotes than those of bacteria or archaea. As introduced in Chapter I, these proteins, or portions of these proteins, lack stable equilibrium structures and instead have dynamic conformations that vary over time and population. Despite the lack of preformed structure, IDPs carry out many and varied molecular functions and participate in vital biological pathways. In particular, IDPs play important roles in cellular signaling that is, in part, enabled by the ability of IDPs to mediate molecular recognition. In Chapter II, the role of intrinsic disorder in molecular recognition is examined through two example IDPs: p53 and 14-3-3. The p53 protein uses intrinsically disordered regions at its N- and C-termini to interact with a large number of partners, often using the same residues. The 14-3-3 protein is a structured domain that uses the same binding site to recognize multiple intrinsically disordered partners. Examination of the structural details of these interactions highlights the importance of intrinsic disorder and induced fit in molecular recognition. More generally, many intrinsically disordered regions that mediate interactions share similar features that are identifiable from protein sequence. Chapter IV reviews several models of IDP mediated protein-protein interactions that use completely different parameterizations. Each model has its relative strengths in identifying novel interaction regions, and all suggest that IDP mediated interactions are common in nature. In addition to the biologic importance of IDPs, they are also practically important in the structural study of proteins. The presence of intrinsic disordered regions can inhibit crystallization and solution NMR studies of otherwise well-structured proteins. This problem is compounded in the context of high throughput structure determination. In Chapter III, the effect of IDPs on structure determination by X-ray crystallography is examined. It is found that protein crystals are intolerant of intrinsic

disorder by examining existing crystal structures from the PDB. A retrospective analysis of Protein Structure Initiative data indicates that prediction of intrinsic disorder may be useful in the prioritization and improvement of targets for structure determination.

Sarath Chandra Janga, Ph.D., Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

Intrinsically disordered proteins (IDPs)

The classic model of protein function is the sequence-structure-function paradigm; a protein's sequence determines the protein's three-dimensional structure (1), and the protein's structure determines the protein's function, via the lock-and-key mechanism (2). This view likely arose because enzymes were the primary focus of early functional and structural characterizations of proteins. Enzymes were convenient targets for study because their activities could be precisely quantified and related to physical characteristics of the enzyme, including protein structure. Many elegant studies have related the details of protein structure directly to function, e.g. the catalytic activity of α-chymotrypsin (3) and the allosteric mechanism of hemoglobin (4). Although contemporary structural biology appreciates the role of dynamics in protein function (5), these dynamics are relative to a stable equilibrium, and so the basic principle of sequence-structure-function has and still remains true for enzymes. However, not all proteins are enzymes.

Contrary to the classic model of protein function, intrinsically disordered proteins (IDPs) function without the prerequisite of a stable structure (6). The conformations of IDPs vary in time and over populations (7), and intrinsic disorder is encoded by the sequence of amino acids (8), similar to the encoding of specific protein structure in sequence. The extent of intrinsic disorder varies from protein to protein (here, the term IDP is used loosely to refer both to proteins that are completely intrinsically disordered and those that consist of a mixture of ordered and disordered residues): from completely disordered proteins, to mostly disordered proteins with local regions of order, to mostly ordered proteins with local regions of disorder, to completely ordered proteins (9). Several authors have investigated the functional repertoire of IDPs (7, 10, 11) , where one common functional indication is molecular recognition.

Physical characteristics

IDPs have been experimentally characterized by a variety of methods, each with their own strengths and weaknesses. The high resolution methods, X-ray crystallography and nuclear magnetic resonance (NMR), have the capability to characterize each residue of a protein as either disordered or ordered. Other lower resolution methods, such as circular dichroism and small angle X-ray diffraction, can characterize the relative content of disordered residues in a protein. Although lower resolution methods cannot provide a precise residue-by-residue characterization of a protein, they can provide high level characterizations that can help to understand the structural state of a protein. Further, low resolution methods can complement each other, e.g. combination of circular dichroism, limited proteolysis, and hydrophobic dye binding in characterizing the function of the molten globule clusterin (12), and be complemented by computational studies that can provide further insights, e.g. the combination of mass spectroscopy-resolved limited proteolysis and intrinsic disorder prediction to identify intrinsically disordered regions in XPA (13). Also, for proteins containing both order and disorder, use of both high and low resolution methods in concert can provide an improved overview of the spatial relationships between the ordered and disordered regions, e.g. combining NMR with small angle X-ray diffraction as used for HMGB1 (14) and 4EBP1 (15).

In the X-ray crystal structures of many proteins, some residues cannot be modeled due to the absence of corresponding electron density. In fact, the majority of known IDP regions have been identified from regions of missing density X-ray crystal structures (16). While rich in quantity, the quality of this characterization, in general, is poor. Missing density is not positive identification of disorder; rather, it only identifies regions that do not share the same lattice symmetry with the rest of the protein structure, where disorder is only one of many possible reasons for a lack of symmetry (other reasons are discussed in Chapter III). Some alternative methods for the solution of protein crystal structures have been developed that can give positive evidence of conformational disorder (17, 18), but these are not commonly used nor universally

applicable. Therefore, missing density should be viewed as correlated with IDP regions and not as definitive identification of IDP regions.

In contrast, several NMR experiments can not only give positive evidence for IDPs and IDP regions, they can also provide detailed information about the dynamics of each intrinsically disordered residue. In this regard, the ^{1}H-^{15}N hetero nuclear NOE experiment provides direct evidence for intrinsic disorder at the per-residue level (19). This experiment measures restriction of motion of backbone amide groups relative to the entire molecule, where a restrained backbone amide is strong evidence that a residue is ordered and an unrestrained backbone amide is strong evidence that a residue is disordered. Furthermore, a disordered region with conformational bias can be detected by this method, which is particularly compelling when combined with other evidence. For example, the IDP p27-KIP1 binds to the cyclin-CDK complex using two disordered regions separated by a third disordered linker region. This linker was shown to have significant structural bias toward helix through a combination of ^{1}H-^{15}N hetero nuclear NOE and chemical shift biases, where the latter gives an indication of secondary structure (20). Later work showed that this conformational bias is vital to the kinetics of the p27-cyclin/CDK interaction (21). While the ^{1}H-^{15}N hetero nuclear NOE experiment is very powerful at characterizing dynamics, it is a less common use of protein NMR than the solution of protein structures.

Solved NMR structures can also be used to identify intrinsically disordered residues in an analogous way to X-ray structures. Since NMR structures are underdetermined – fewer constraints are available than degrees of freedom in the protein – it is a common practice to report multiple models that fit the available NMR-derived constraints equally well (22). In these multiple structures, under- or un-constrained regions are often apparent through large variation in conformation. These regions have been shown to closely correspond to regions of missing density in crystal structures of the same protein (23). However, identification of disorder through conformational variability in NMR models has analogous drawbacks to identification of disorder by missing density in X-ray structures; several experimental and procedural complications can

give rise to lack of NMR constraints (22). Therefore, similar to missing density, an under-constrained region in a NMR structure is not a definitive indication of intrinsic disorder.

Several experimental methods can be used for a lower resolution characterization of intrinsic disorder. For example, limited proteolysis can identify whether protease cut sites reside in structured or unstructured regions. Proteases hydrolyze sequence specific sites, but the rate at which those sites are hydrolyzed is directly related to whether they are ordered or disordered, where the rate of hydrolysis of disordered sites is orders of magnitude faster than the rate of ordered sites (24). By resolving the proteolysis reaction over time and identification of proteolysis fragment via mass spectrometry, the relative cut rates of each site can be determined and thereby the order/disorder state of each cut site (13). The number of sequence-specific cut sites limits the resolution of this method but can be increased by using multiple proteases.

Another important experiment for the identification of IDPs is circular dichroism (CD). This experiment measures the differential absorption of left and right circularly polarized light, which is strongly dependent on the local environments of chiral centers in a molecule. Since each residue (excepting glycine) has a chiral center at the α-carbon, the far-UV CD spectrum of a protein is highly dependent on conformation of the protein backbone. Secondary structure types, α-helix and β-sheet, give distinct CD spectra, and lack of structure, so called random coil, also gives a distinct spectrum. Many IDPs have been identified by their distinct CD spectra (25). CD is an invaluable tool for initial, high level characterization of IDPs. However, some types of IDPs – molten globules – can exhibit high a high content of secondary structure, so other additional methods are required to characterize these proteins.

Finally, gel filtration and small angle X-ray scattering (SAXS) are experiments that can provide high level characterization of proteins as ordered or disordered. Gel filtration provides a relative measure of a protein's size in solution rather than its mass (26). Since IDPs are somewhat more extended in solution than ordered proteins, a protein's solution size, combined with its molecular weight, can be used as positive evidence that a protein is disordered (27). SAXS also

measures the solution dimensions of protein but in much more detail than gel filtration. In the case of ordered proteins, SAXS can be used to obtain a low resolution structure (28), but for highly dynamic IDPs, possible conformations are too degenerate and the shape of the conformational ensemble can only be described at a qualitative level (29). However, SAXS is very sensitive to residual structure in IDPs (29), and in the case of IDPs that contain partial structure or a biased conformational ensemble, a low resolution density map can be determined for constrained portions of the IDP (15).

Many more methods for physical characterization of IDPs have been developed (30, 31), including deuterium exchange (32) and spin labeling (33). All of these methods provide a unique perspective on the order-disorder state of a protein, probing different structural and chemical characteristics of protein structure. No single method gives a complete picture of the conformational characteristics of IDPs in the same way that an X-ray structure characterizes an ordered protein, but through application of multiple methods much can be learned (34). The physical characterization of IDPs is a field that is still evolving, and much of that evolution involves combination of both experimental and computational techniques, e.g. the combination of NMR and ensemble methods (35), and the combination of limited proteolysis and prediction of intrinsic disorder (13).

Sequence characteristics

Similar to the way a specific three dimension structure is encoded in the sequence of an ordered protein, the sequence also encodes a higher level, namely the ordered or disordered state of the protein. The relationship between intrinsic disorder and protein sequence was first noted in the case of nucleic acid associated proteins, which are rich in disorder and carry a relatively high net positive charge (36). Many subsequent studies have found that disordered proteins are also generally enriched in hydrophilic residues and depleted in hydrophobic residues, relative to structured proteins (37). In addition to these general biases, disordered regions are highly enriched in proline (8), likely because it disrupts regular secondary structure, and depleted in

cysteine (8), likely because of its ability to stabilize structure by forming crosslinks. Additionally, while it might be presumed that the compositions of IDPs would be very similar to the compositions of protein surfaces, there are some notable differences (38). One difference is that surfaces are generally enriched in all charged residues, but, compared to protein surfaces, IDPs are significantly less enriched in arginine and aspartic acid. Another difference is that IDPs are significantly depleted in all aromatic residues, but surfaces show little or no depletion of tryptophan and tyrosine and even a slight enrichment in histidine. The chemical and structural reasons for these asymmetries are currently not clear but likely relate to the hydrogen bonding potential of these three aromatics. Another notable feature of many disordered proteins is a low complexity sequence (8); many disordered proteins use a limited number of amino acids in their sequences. However, low complexity is not a requirement for a sequence to be disordered since many IDPs have a comparable complexity to ordered proteins (8).

The compositional biases of IDPs are an indication that sequence encodes for disorder, and additional evidence is provided by prediction of disorder from sequence. At the level of predicting disorder for an entire protein sequence or region, the mean absolute net charge and mean hydropathy separate ordered and disordered proteins very well (37), where disordered proteins tend to have a larger net charge and a lower hydrophobicity. The charge-hydropathy approach was also applied to per-residue prediction with good results (39), though more sophisticated approaches are more accurate. The first application of machine learning to prediction of intrinsic disorder from protein sequence was a neural network trained on several sequence attributes calculated on a small local window (40). A generalized extension of this predictor has good accuracy in predicting novel regions of intrinsic disorder (8). Subsequently, many predictors of intrinsic disorder from protein sequence have been developed using a variety of architectures (41), and prediction accuracy has reached a high level, as measured by objective evaluation (42). This high level prediction accuracy is due in no small part to the increasing number of well characterized IDPs and IDP regions, many of which have been collected into the

Database of Protein Disorder (DisProt) (16). The high level of prediction accuracy not only demonstrates the principle that intrinsic disorder is encoded by protein sequence, but provides a valuable research tool for the discovery of novel IDPs and investigating the role of IDPs in biology.

Functional characteristics

Application of disorder predictors to proteomes in all domains of life have helped provide insight into the biological roles of IDPs, and a significant amount of experimental data has been amassed on the novel and varied mechanisms by which IDPs carry out these roles (43). Correlations of functional annotations with disordered predictions (10, 11, 44) consistently indicate that intrinsic disorder plays a central role in cellular signaling. This idea is supported by the importance of disorder in the regulation of transcription (45) and some key signaling pathway and proteins (46–48). Also, eukaryotes are rich in intrinsic disorder relative to prokaryotes and archaea, which correlates with the increased complexity of signaling in eukaryotes (49). Signaling roles are fulfilled by IDPs' ability to mediate and regulate molecular recognition. Additionally, many IDPs have entropy-based functions; they have roles that rely directly on maintenance of their conformational heterogeneity. Signaling and entropy-based functions are not mutually exclusive, which will be illustrated by an example.

Many IDPs have been found to mediate interactions with other macromolecules, both nucleic acid and protein. Nucleic acid binding was one of the first functions associated with IDPs (36). Transcription factors (45), histones (50), and ribosomal proteins (51) are all rich in intrinsic disorder. Often the nucleic binding interface itself is disordered, e.g. Myc (52) and HMG-I(Y) (53), which is consistent with a high positive charge density to complement the high negative charge of nucleic acid backbones. In others, the nucleic acid binding domain is ordered, but disorder plays other important roles in other portions of these proteins (54), such as interactions with other proteins. Many examples of intrinsic disorder mediating interactions with partner proteins have been reported (7, 55). IDPs use fewer residues than ordered proteins to achieve

interfaces of comparable size (56) meaning that IDPs make efficient scaffolds (57) – scaffolds organize a large number of proteins to achieve higher level functions (58) and increase signaling efficiency (59). Additionally, the conformational flexibility of IDPs allows them to bind to multiple partners with distinct recognition sites (60). This ability to facilitate multiple specificity has been found in many IDPs (61) and is a focus of Chapter II.

The signaling functions of IDPs are regulated in several ways. Signaling via molecular recognition is often regulated via post-translational modifications, and the sites of many types of post translational modification are preferentially located in IDP regions (62–64). Post-translational modifications provided a direct mechanism for modulating signaling interactions mediated by IDPs, where modifications may directly participate in or disrupt an interaction (60), or induce or disrupt conformations required for an interaction (65). Also, mRNA encoding for IDP regions are preferential sites of alternative splicing (66), providing another mechanism for regulation of IDP proteins. Finally, other regulation mechanisms, such as protein degradation and translation regulation, have also been correlated with IDPs (67), although the mechanistic role of IDPs in these modes of regulation is not clear. The combined roles of IDPs in molecular recognition and regulation strongly suggest IDPs as a central feature of cellular signaling.

In addition to signaling functions, many IDPs perform functions that are quintessentially intrinsically disordered. These are functions that arise from the conformational heterogeneity of disordered proteins, which have been called entropy-based functions (7), and span a wide variety of mechanisms, which are illustrated here by a few examples. For example, the Shaker voltage gated potassium channel protein consists of a transmembrane channel subunit and an intrinsically disordered plug subunit (68). The duration of activation of the channel is determined by how long the plug subunit takes to rebind to the channel domain once it is displaced; the disordered plug domain acts as an entropic clock, where the duration of channel opening is related to the search space of the plug domain (69). For another example, the nuclear pore complex contains several proteins with long disordered tails consisting of FG repeats (70). These fill the wide

channel of the pore, which allows a free flow of ions but prevents larger molecules from freely passing through due to energy required to restrict the entropic freedom of the disordered tails (71). A final example is the neurofilament H protein that plays a role in maintaining axon structure. This IDP is highly phosphorylated and forms an extended bristle. Multiple bristles arising from multiple subunits form a brush around each neurofilament, preventing the axon from collapsing, with electrostatic repulsion enhancing the excluded volume effects of the brushes (72).

Molecular recognition, regulation, and entropic functions are not mutually exclusive, which is demonstrated by the example of the Sic1 protein. Sic1 is an intrinsically disordered protein (73) that inhibits cell cycle progression until Sic1 is inactivated by proteolytic degradation by the SCF complex (74). Recognition of Sic1 by SCF is mediated by Cdc4, which recognizes a phosphorylated motif. Sic1 contains 9 suboptimal copies of the Cdc4 recognition motif (73), and phosphorylation of multiple suboptimal motif instances is required for high affinity binding of Sic1 by Cdc4 (75). The requirement for multiple phosphorylations is remarkable because Cdc4 only has a single motif recognition site (75). Experiment (76) and modeling (77) results have indicated that tight binding is achieved through long range electrostatic interactions and transient interaction of multiple phosphorylated sites on Sic1 with the single Cdc4 motif binding interface, which is enabled by the intrinsic disorder of Sic1. Sic1 embodies many of the hallmark features of intrinsically disordered proteins: regulation by post-translational modification, molecular recognition, and function through conformational diversity.

There are also significant evolutionary implications of intrinsic disorder. In agreement with indications of the role of intrinsic disorder in cellular signaling, intrinsic disorder is over represented in eukaryotic organisms, relative to archaea or bacteria (49). Also, it has been suggested that the lack of requirement for order structure implies that IDPs should be more tolerant of random mutation due to genetic drift than ordered proteins, and for many IDPs this is the case (78, 79). Furthermore, a specific sequence is not required for function of some IDPs,

where the primary functional requirement appears to be amino acid composition (80). However, many other IDPs appear to be more conserved than expected (78, 79), possibly due to functional constraints on these proteins. Indeed, several well characterized domain families have been found to contain conserved disordered regions (81) or even to be entirely disordered (82), suggesting that intrinsic disorder is positively selected for in evolution.

Structural characteristics

Much recent work has focused on describing the conformational characteristics of intrinsically disordered proteins at various levels of resolution. At the coarsest level of resolution, the order-disorder continuum has been defined (83). This formalism quantifies the relative conformational diversity of a protein as a single number, which represents the number of structures required to represent the protein's conformational diversity, where a value of 1 indicates a single, stable equilibrium structure, and 0 represents a protein that requires an infinite number of structures to describe its diversity, which is not observed in practice. At finer levels of resolution, ensemble modeling methods have been developed (35, 84). These methods derive a set of static structures that in aggregate represent the solution structure of intrinsically disordered proteins. A set of structures, combined with their respective relative populations, can be used to model experimental data and predict protein properties. The finest resolution is provided by molecular dynamics simulations, which attempt to simulate the details of the behavior of proteins in solution (85). Disordered proteins are particularly difficult to address with these methods because the large scale motions they undergo are on a very large time scale, requiring an enormous amount of computational resources to simulate (86). Rather than direct simulation, approximation methods (87) have been very successful sampling the full conformational diversity of intrinsically disordered proteins.

The picture of the solution characteristics of IDPs has been largely determined in broad strokes, and is highly dependent on sequence properties and context. Sequence properties that affect solution conformations are primarily hydrophobicity and net charge; where hydrophobic

low charge sequences are more collapsed and hydrophilic, highly charged sequences are more expanded. In fact, the hydrodynamic radius of peptides can be modeled directly by these properties (88). The idea of collapsed IDPs composed of hydrophilic, uncharged residues is somewhat paradoxical, since one might expect that these residues should prefer to interact with water and form expanded structures. However, recent evidence has shown that water is a poor solvent for the protein backbone, which drives these IDPs to collapse (89), and only in the presence of significant charge-charge repulsion do sequences form expanded structures (90). Even the poly-glutamine sequence forms a collapsed rather than an extended form (89), likely from the just mentioned backbone effects and also because the amide groups prefer to hydrogen bond with each other rather than with water. Additionally, charge distribution plays a large part in the expansion of IDP sequences, where a highly asymmetrical charge distribution, e.g. acidic residues at the N-terminus and basic residues at the C-terminus, will cause a sequence to be less expanded than expected (91). These observations suggest that the solution behavior of IDPs can be fine-tuned by nature through coarse compositional changes.

The context dependence of intrinsic disorder is often functional. Many examples of IDPs undergoing a disorder-to-order transition on binding have been determined, and so intrinsic disorder is contextually dependent on the absence of these binding partners. This contextual dependence has inspired some to suggest that disorder is only as relevant to biology as protein folding, i.e. the bound state of these proteins is biologically active while the unbound state is not, and they prefer the nomenclature “proteins awaiting partners” to “IDPs” (92). However, this view simply ignores the many IDPs with entropic functions. Additionally, for IDPs that do bind to partners this view re-emphasizes the sequence-structure-function paradigm at the expense of consideration and understanding of the unbound state. The practical importance of the unbound state of IDPs is demonstrated by the development of small molecule inhibitors that directly target the unbound, disordered states of IDPs (93). Furthermore, maintained intrinsic disorder can play an integral role in binding, as in the case of Sic1. The more interesting question is not necessarily

the structures formed when IDPs bind to partners, but how IDPs recognize partners and the functional implications of the disordered state.

Disorder-to-order transitions (DOTs)

For IDPs, molecular recognition is often concomitant with folding, where the IDP undergoes a DOT (94). IDPs have been found to bind all types of biological molecules, including proteins, nucleic acids, small molecules, and various ions (7). As a consequence of folding and binding, the structures of IDPs with molecular interaction functions in complex with their partners can be determined experimentally via X-ray crystallography or NMR spectroscopy. For example, see (94) for examples of IDPs in complex with their partners. Also, covalent modifications, such as phosphorylation or acetylation, can also induce a DOT, a process that is thought to be important for signaling and regulation (6, 94). Finally, intrinsically disordered regions can become structured due to crystallization, particularly if they participate in crystal contacts (95), which are low affinity contacts that stabilize the crystal lattice.

Between biologically relevant interactions, covalent modifications, and artificial factors, there are several factors that can cause an IDP to appear as well ordered in structure determination experiments. This suggests that, at least potentially, there are a significant number of IDPs with structures in the PDB. This has significant implication for structural biology methods that interpret proteins as rigidly structured or moderately flexible, such as homology modeling and docking. This also has implications for bioinformatics, particularly prediction of intrinsic disorder from sequence. Currently, the presence of DOT regions is appreciated in the field. Unfortunately, since no reliable methods are available to identify these regions, this appreciation usually translates to exclusion of all ordered proteins that may contain DOT regions, namely all complex structures.

Computational analysis of DOTs

Several approaches have been used for examining DOT regions in experimentally determined protein structures. In comparative structural analysis methods (96–98), multiple X-

ray crystal structures of the same or similar protein are compared. Since regions of missing density in X-ray crystal structures are indicative of an intrinsically disordered region, DOT regions are inferred where a region is present in one crystal structure but absent in another. Structural attributes have also been used to examine DOT regions (99, 100), where static parameters of the protein structure are used to characterize these regions. Finally, a diverse set of sophisticated modeling methods has been brought to bear to investigate protein flexibility and intrinsic disorder (101–105).

Comparative analysis

Several groups have studied DOT regions in known protein structures by comparative structural analysis. Le Gall et al. (96) mapped protein sequences in PDB to reference sequences in the SwissProt database. DOTs were inferred from regions of defined density and regions of missing density that map to the same region of reference sequence structures. This procedure identified a large number of DOT residues, but the prevalence of these residues is difficult to assess because of how the data are reported. In a similar study, Zhang et al. (97) examined proteins with multiple structures in PDB and extracted regions that that were disordered in at least one structure but ordered in at least one other structure. These authors found 2,819 DOT regions (Zhang et al. call these regions “dual personality fragments”; the difference between this interpretation and the DOT interpretation is semantic) in 1,535 identical protein clusters, which corresponds to at least one DOT region in 45% of proteins with more than one structure in PDB. Zhang et al. demonstrated that DOT regions are fairly common in the PDB, but neither study attempted to link DOT regions directly with a causal factor, such as molecular interactions or crystal contacts.

A study by Fong et al. (98) focused on DOT regions in conserved binding modes (CBMs). CBMs are a set of protein domain-domain interfaces that have been filtered for artificial or otherwise spurious interactions by selecting only those interactions between domain families that have been observed in multiple protein structures (106). Structures with interactions

corresponding to CBMs were evaluated for DOTs by pairing these structures with structures of the monomeric form of the protein, an approach similar to that of Zhang et al. Unfortunately Fong et al. do not report the raw results of this analysis and instead report the results of a cautious statistical test for the bias of interactions' residues to be DOT residues. Despite this cautious approach, they found that 42-75% of interfaces show a significant bias toward DOT residues, depending on the dataset. However, many DOT regions were not accounted for by interactions, and many other DOT regions were found that are ordered in the monomeric structure and disordered in the complex structure, meaning that only a fraction of the DOT regions have causal indications in this study.

While comparative structural analysis is a useful method for identifying DOT regions in protein structures, it suffers from reliance on reference monomers. The fundamental difficulty is that reference monomers are not available for the majority of protein-protein, protein-nucleic acid, and protein-small molecule complexes. This suggests that there may be far more DOT regions involved in molecular interactions than comparative analysis suggests. The availability of reference monomers may even be biased against DOT containing proteins, since the presence of disorder is likely to be a confounding factor in protein crystallization. Furthermore, there are often many contextual differences between the complex structure and reference structure that complicate a direct interpretation of the differences between them. For instance, since the contents of the structures differ, there will nearly always be differences in crystallization conditions and crystal packing, both of which can induce DOTs. Additionally, analysis can become quite complex if all the constituents of crystal structure, e.g. additional domains, multiple binding partners, and small molecules are considered.

Structural analysis

A more direct approach to identifying DOT regions in protein structure is through structural analysis. The underlying hypothesis of this approach is that the structures of DOT regions and that of ordered regions differ in some significant respect that is detectable through

analysis of the protein's structure. Two groups have taken this approach to DOTs in protein structures.

Gunasekaran et al. studied the special case of an entire chain in a protein structure undergoing a DOT (100). From the literature, the authors constructed a set of complexes involving disordered proteins and complexes involving ordered proteins, and then they examined these two sets for structural attributes that distinguish ordered proteins from disordered proteins. The authors reported two primary attributes: normalized surface area of the protein, calculated disregarding the other constituents of the complex, and normalized interface area, calculated as the difference between the protein's surface area with and without other complex constituents. Both parameters are normalized by the number of residues in the protein. They found that the disordered proteins in their set were distinguished from the ordered proteins by both a significantly greater normalized surface area and a significantly greater normalized interface area.

Mohan et al. also studied the special case of entire protein chains undergoing a DOT (99). However, the focus of this work was Molecular Recognition Features (MoRFs), which are short DOT regions within a longer region of disorder that are responsible for molecular recognition. Because these are short regions within a longer region of intrinsic disorder, the structures of complexes involving MoRFs contain only this short region bound to a partner. Since there are only very few folded domains fewer than 70 residues in length, all proteins with a defined density shorter than this length and bound to a longer protein were selected from PDB and manually verified as lacking globular structure. This procedure found 372 DOT proteins. The DOT proteins found by this procedure were also consistent with the observations of Gunasekaran et al., with relatively large normalized surface and interface areas.

Both of these studies demonstrated that structural analysis methods are useful for identifying DOTs. However, both were limited to the case of entire proteins undergoing a DOT. For the more general case of DOTs that occur in the context of otherwise ordered proteins, such loops, termini, or inter-domain regions, the Mohan et al. method is inapplicable by definition and

the Gunasekaran et al. method fails because the normalized surface and interface areas are dominated by the ordered portion of the structure. The latter approach might be extended to handle the more general case of DOT regions, which is a topic for further research.

Model-based analysis

Several modeling methods have been applied to, or are applicable to, investigation of DOTs from protein structure. These methods are distinct from comparative and structural analysis in that they attempt to predict the physical behavior of proteins, to varying degrees of accuracy, based on an initial static structure. From a practical perspective, they are also distinguished by being relatively complex to implement and computationally intensive to apply.

Molecular dynamics has been applied by several groups in the study of IDPs in general and also to the study of DOTs specifically. In general, molecular dynamics involves simulating the atomic motions of a protein in explicit or implicit solvent. Molecular dynamics has successfully been used for discovering intrinsically disordered regions from proteins' structures, although relatively long simulations must be performed in order to access the large scale motions of IDPs (102, 104). While powerful and informative, molecular dynamics is difficult to apply on a wide scale, both because individual simulations require some fine tuning, and because long simulations are computationally intensive. Also, relatively extensive experience is required in order to perform a successful simulation, and so these methods are not generally accessible. As an alternative, less precise but more computationally tractable and more accessible methods have been developed for the study of protein flexibility and disorder in protein structures.

The approach of normal mode analysis has been developed for the study of protein flexibility (105). Both Gaussian network models (107) and elastic network models (108) are examples of formalisms for normal mode analysis. Normal mode analysis evaluates the characteristic frequency modes of a protein structure based on the force potentials between atoms, where Gaussian networks use realistic potentials and elastic networks use simplified spring potentials. This approach has been used to characterize stabilization induced through crystal

packing (95). Although this method has not been applied to the general study of IDP structures, it may be applicable in the sense of intrinsic disorder as an extreme case of flexibility. However, the model's assumption of static non-covalent contacts complicates the interpretation of frequency modes in terms of intrinsic disorder, i.e. low frequency intrinsic disorder modes may be trapped by transient contacts present in the protein structure.

Another modeling method developed for the study of protein flexibility is analysis of mechanical degrees of freedom, as implemented by the ProFlex algorithm (103). There are many degrees of freedom in the conformation of a protein, which can be represented by backbone Φ and Ψ angles and side chain χ angles. In a folded protein, these degrees of freedom are limited by the presence of stable non-covalent contacts between atoms. The ProFlex algorithm measures how much each of a protein's conformational degrees of freedom are limited based on the protein's structure. Over-constrained regions are considered stable and under-constrained regions are considered flexible. Similar to Gaussian network analysis, this method may be applicable to IDPs as an extreme case of flexibility. However, mechanical stability measurement does not make allowances for dynamic non-covalent contacts, so interpretation for intrinsically disordered regions is not straightforward.

In contrast to Gaussian network and mechanical stability analyses, a method developed for the study of protein flexibility, called the COREX algorithm (101), does not assume static non-covalent contacts. COREX is based on a two-state, i.e. folded and unfolded, model of thermodynamic stability. The algorithm enumerates all the possible states of the protein, where each region of the protein can either be folded or unfolded, and calculates the energy of each state relative to a reference state, in which every residue is in the folded state. The probability of a residue being unfolded is calculated as the sum of the Boltzmann weights of all states in which the residue is unfolded, which is the exponent of the state energy divided by the sum of the exponents of the energies of all of the states.

The COREX algorithm has been applied to several proteins with flexibility measured by various techniques, e.g. deuterium exchange (109), and shows good agreement with experimentally measured flexibilities. This method is appropriate for studying IDPs, since estimating flexibility as an unfolding process describes the DOT process. The disadvantages of this method are its computationally intensive nature and poor scaling properties. The COREX algorithm calculates the surface area of every protein state, and the number of states increases exponentially with protein length. Practically, the algorithm is computationally demanding for small proteins, around 180 residues, and quickly becomes intractable for average and long proteins.

Overview

Chapter II details published (60) work examining the ways in which disorder facilitates the regulatory functions of two protein hubs. This work shows that intrinsic disorder is crucial for function of these hub proteins by enabling multiple specificity protein-protein interactions. Chapter III is a published (110) examination of the relationship between intrinsic disorder and the determination of protein structures by X-ray crystallography (note that Figure 10 and accompanying text did not appear in the original publication). It was found that extensive intrinsic disorder can inhibit determination of a proteins structure and that protein crystals have an apparently limited tolerance for intrinsically disordered regions. Chapter IV is a review of Molecular Recognition Features (MoRFs) highlighting some of this authors contributions (56, 61, 99, 111, 112).

II. FLEXIBLE NETS: DISORDER AND INDUCED FIT IN THE ASSOCIATIONS OF P53 AND 14-3-3 WITH THEIR PARTNERS

Introduction

Protein-protein interaction (PPI) networks integrate various biological signals including those used for energy generation, cell division and growth to give a few notable examples. The architectures of the PPI networks indicate that they are nearly scale free (113–120). That is, a log-log plot of the number of nodes versus the number of links (or interactions) at each node gives a straight line with a negative slope. The negative slope means that these sets of interactions contain a few proteins (hubs) with many links and many proteins (non-hubs) with only a few links. The term 'hub protein' is relative to the other proteins in a given PPI network, with no agreed upon number of links separating hubs and non-hubs.

Several networks such as the internet, cellular phone systems, social interactions, author citations, and so on exhibit scale-free architecture. With regard to PPIs, scale-free network architecture is suggested to provide several biological advantages. For example, given the small fraction of hub proteins, random deleterious mutations will more likely occur in non-hub proteins. The elimination of the functions of such non-hub proteins typically have small effects and so, generally, are not serious. In contrast, a deleterious mutation of a hub protein is more likely to be lethal (116–121). Another advantage is that signals can traverse these networks in a small number of steps, so signal transduction efficiency is improved compared to that expected for random networks (119).

Understanding PPI network evolution across different species is an important problem (122–125). From this body of work, hub proteins appear to evolve more slowly than non-hub proteins, an observation that is consistent with Fisher's classic proposal that pleiotropy constrains evolution (126, 127). Some proteins have multiple, simultaneous interactions ("party hubs") (128) while others have multiple, sequential interactions ("date hubs") (128). Date hubs appear to

connect biological modules to each other (129) while party hubs evidently form scaffolds that assemble functional modules (128).

The idea that PPI networks use scale-free network topology is receiving considerable attention, but some caution is in order. Currently constructed networks are noisy, with both false positive and false negative interactions (120, 130–132). Also, network coverage to date (133–136) is not sufficient to prove scale-free architecture (137). Whether PPI networks are truly scale-free or only approximately so, it nevertheless appears to be true that a relatively small number of proteins interact with many partners, either as date hubs or party hubs, while many proteins interact with just a few partners.

The ability of a protein to bind to multiple partners was suggested to involve new principles (138). Indeed, neither the lock-and-key (2) nor the original induced-fit (139) readily explains how one protein can bind to multiple partners. Note that the original induced fit mechanism was defined as changes in a structured binding site upon binding to the partner (139), changes that are analogous to a glove altering its shape to fit a hand. On the other hand, both theoretical and experimental studies over many years suggested that natively unstructured or intrinsically disordered proteins form multi-structure ensembles that present different structures for binding to different partners (140–146). Based on these prior studies, we proposed that molecular recognition via DOTs provides a mechanism for hub proteins to specifically recognize multiple partners (47). We pointed out earlier that intrinsic disorder could enable one protein to associate with multiple partners (one-to-many signaling) and could also enable multiple partners to associate with one protein (many-to-one signaling) (146).

Recent bioinformatics studies support the importance of protein disorder for hubs (147–151). While disorder appears to be more clearly associated with date hubs (149, 151) than with party hubs, some protein complexes clearly use long regions of disorder as a scaffold for assembling an interacting group of proteins (152–160). Thus, the importance of disorder for party hubs needs to be examined further. Additional evidence for the importance of disorder for

highly connected hub proteins comes from a structure-based study of the yeast protein interaction network (161). The authors considered only interactions that could be mediated by domains with known structures and found that the degree distribution of the resulting network contained no proteins with more than 14 interactions, which is more than an order of magnitude less than is observed in one unfiltered, high confidence dataset (Jake Chen, personal communication). This result indicates that a structure-based view of hub proteins is insufficient to explain the multitude of partners that interact with hub proteins.

To improve understanding of the use of disorder for binding diversity, we studied two prototypical examples: p53 and 14-3-3. Both are hubs that are clearly involved in crucial biological functions. For example, p53 is a key player in a large signaling network involving the expression of genes carrying out such processes as cell cycle progression, apoptosis induction, DNA repair, response to cellular stress, etc. (162). Loss of p53 function, either directly through mutation or indirectly through several other mechanisms, is often accompanied by cancerous transformation (163). Cancers with mutations in p53 occur in colon, lung, esophagus, breast, liver, brain, reticuloendothelial tissues, and hemopoietic tissues (163). The p53 protein induces or inhibits over 150 genes, including p21, GADD45, MDM2, IGFBP3, and BAX (164).

The four regions or (not necessarily structured) domains in p53 are the N-terminal transcription activation domain, the central DNA binding domain, the C-terminal tetramerization domain, and the C-terminal regulatory domain. The last two could be considered to be a single C-terminal domain with two subregions. The transactivation region interacts with TFIID, TFIIH, Mdm2, RPA, CBP/p300 and CSN5/Jab1 among many other proteins (162). The C-terminal domain interacts with GSK3β, PARP-1, TAF1, TRRAP, hGcn5, TAF, 14-3-3, S100B(ββ), and many other proteins (162).

As for 14-3-3 proteins, they contribute to a wide range of crucial regulatory processes including signal transduction, apoptosis, cell cycle progression, DNA replication, and cell malignant transformation (165). These activities involve 14-3-3 interactions with various

proteins in a phosphorylation-dependent manner. More than 200 proteins have been shown to interact with members of 14-3-3 family (166–168), with these 14-3-3-interacting proteins amounting to approximately 0.6% of the human proteome (168). One proposed functional model is that 14-3-3 binds to the specific target as a molecular anvil causing conformational changes in the partner. In their turn, these changes can affect enzymatic (biological) activity of a target protein or mask or reveal specific motifs that regulate its localization, activity, phosphorylation state, and/or stability (169).

The 14-3-3 protein has at least nine sequence isomers, called α, β, γ, δ, ε, η, σ, τ, and ζ (170). All isomers are structured dimers with grooves that bind to more than 200 different partners, and the different partners have different sequences for their binding regions. Screening experiments have identified individual peptides that bind to all the different isomers, suggesting that the binding grooves in the different isomers have some common features (171). A recent bioinformatics study suggests that the partners of 14-3-3 utilize intrinsic disorder for binding (172).

The interactions of p53 and 14-3-3 with their partners as reported previously (170, 173–188) are examined herein but from an order-disorder point of view. In the case of p53, different regions in the disordered tails enable this protein to bind to multiple partners at the same time. In addition, one single region of disorder adopts clearly different secondary structures and uses the same amino acids to different extents in different binding interactions. For this case the plasticity of the disordered region clearly enables the binding to multiple partners. In the case of 14-3-3, the different partners have distinct sequences. Their interactions with 14-3-3 show characteristics, such as hydrogen bonds between side chains of 14-3-3 and the backbone of the partners and such as hydrogen bonds between the backbone of the partners and water, indicating that the two partners were very likely unfolded in water just prior to association with 14-3-3. The distinct sequences of the partners do not adopt identical backbone structures, and the various side chain interactions between 14-3-3 and the two different partners involve induced-fit adjustments

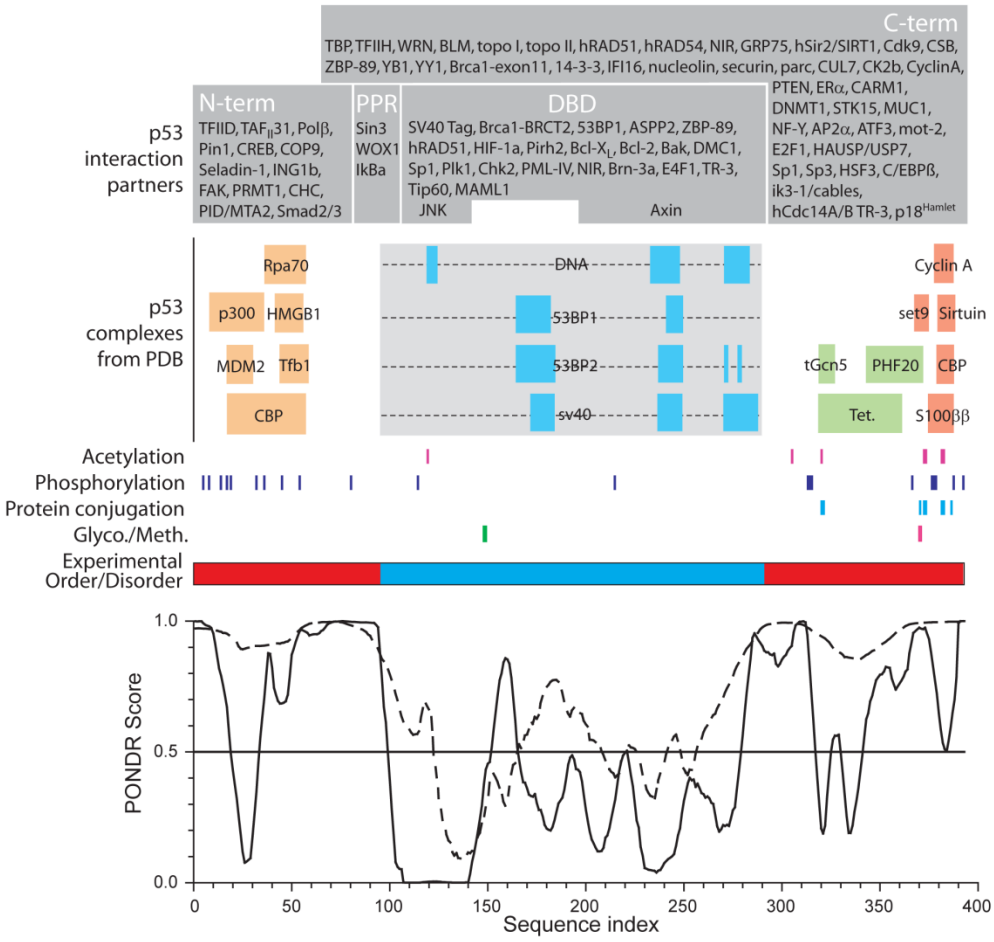
of the 14-3-3 structure. Overall, these studies show how the plasticity of disordered proteins is used to enable the binding diversity of hub proteins, both for a single disordered region binding to multiple partners and for multiple disordered regions binding to the same partner.

Results

Intrinsic disorder and the molecular interactions of p53

The p53 molecule interacts with many other proteins in order to carry out its signal transduction function. A number of these are downstream targets, such as transcription factors, and others are activators or inhibitors of p53's transactivation function. Many of these interactions have been mapped to regions of the p53 sequence (Figure 1, gray boxes): the N-terminal domain (i.e., the transactivation domain), the C-terminal domain (i.e., the regulatory domain), and the DNA binding domain (DBD). These domains have also been characterized in terms of their structure or lack thereof (Figure 1, red (disordered) and blue (structured) segments), where the DNA binding domain is intrinsically structured and the terminal domains are intrinsically disordered (189, 190). While the tetramerization domain is structured (191), the structure is acquired upon the formation of the complex. Additionally, multiple different posttranslational modifications have been identified in p53 (Figure 1, vertical ticks). These modifications are relevant here because they are a common method for altering protein interactions.

Figure 1. Summary of p53 interactions and structure. Dark gray boxes indicate the approximate binding regions of p53's known binding partners. The regions of p53 represented in structure complexes in PDB are represented by horizontal bars, labeled with the name of the binding partner. For the DBD, the extent of the globular domain is indicated by the light gray box, where the internal horizontal bars indicate regions involved in binding to a particular partner. Post translational modifications sites are represented by vertical ticks. Experimentally characterized regions of disorder (red) and order (blue) are indicated by the horizontal bar. Finally, predictions of disorder (scores > 0.5) and order (scores < 0.5) are shown for two PONDR predictors: VLXT (solid line) and VSL2P (dashed line). All features are presented to scale, as indicated by the horizontal axis. The p53 interaction partners and post translational modification sites have been adapted from Anderson & Appella (162).



Comparing the regions of order and disorder reveals a strong bias towards the localization of the interactions within the intrinsically disordered regions. Overall, 60/84 = 71% of the interactions are mediated by intrinsically disordered regions in p53. A bias toward intrinsically disordered regions is even more pronounced in the sites of posttranslational modifications, with 86%, 90%, and 100% of observed acetylation, phosphorylation, and protein conjugation sites, respectively, found in the disordered regions. This is consistent with previous observation of a strong bias for post translational modifications toward intrinsically disordered regions (62). This concentration of functional elements within intrinsically disordered regions compares to just 29% of the residues being disordered (47). Clearly, p53 exhibits a highly biased use of disordered regions for mediating and modulating interactions with other proteins.

In addition to experimentally characterized disorder, predictions of intrinsic disorder for p53 using both PONDR VL-XT (8) and VSL2 predictions (192) were carried out (Figure 1, graph). The latter is one of the highest accuracy prediction algorithms available (193), whereas the former has been observed to be especially useful in identifying binding regions within longer regions of disorder (111, 194, 195) and to be much better at identifying such sites as compared to a number of different disorder predictors (196). Both predictors give good agreement with the experimental determination of intrinsic disorder (7, 13, 44, 45, 62, 197–211), and in the case of p53 both of their predictions agree well with experimental characterization.

Analysis of associations involving p53 using 3D structures

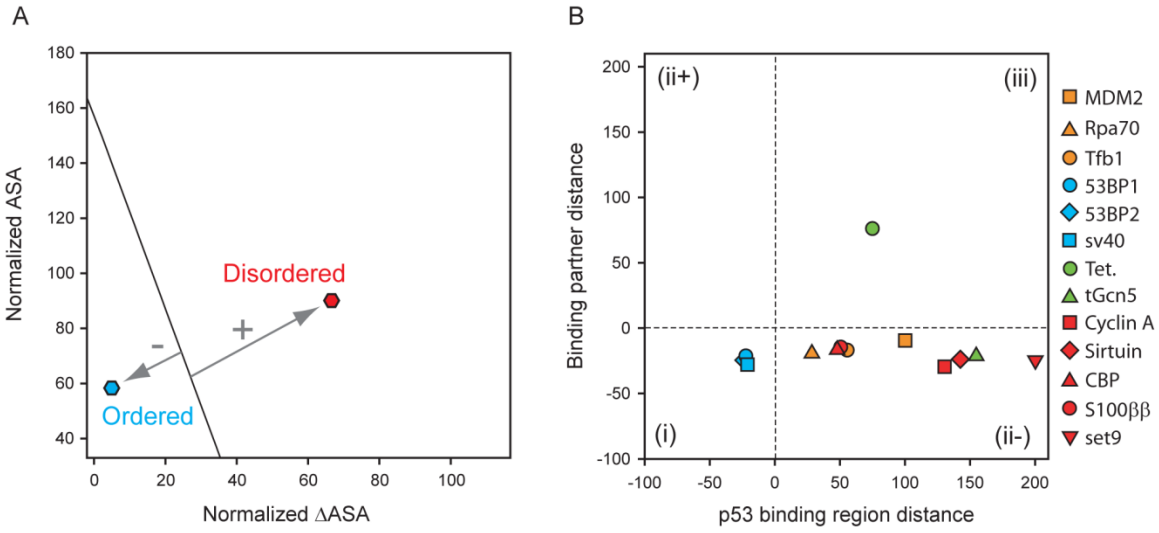
The structures of 14 complexes between various regions of p53 and unique binding partners have been determined (Figure 1, horizontal bars). For 10 of these partners, the interactions are mediated by regions experimentally characterized as intrinsically disordered, where PONDR VL-XT detects the majority of these binding regions as short predictions of order within a longer prediction of disorder. These structures are complexes between p53 and endogenous partners: cyclin A (173), sirtuin (174), CBP (175), S100ββ (176), set9 (177), tGcn5 (178), Rpa70 (179), MDM2 (180), Tfb1 (181), and itself (182). The remaining four interactions are mediated by the

structured DBD, namely between p53 and three endogenous partners – DNA (183), 53BP1 (184), and 53BP2 (185) – and one exogenous partner – the large-T antigen (LTag) from simian virus 40 (188).

Protein complexes can be formed from the association of structured proteins by the folding of one disordered protein onto the surface of a structured partner or by the coupled folding and binding of intrinsically disordered proteins (100, 212–218). Nussinov and collaborators (100) showed that a plot of normalized monomer area (NMA) versus normalized interface area (NIA) nicely separates complexes formed from structured proteins as compared to complexes formed from unfolded proteins by coupled binding and folding. That is, associations of structured proteins exhibit small NMAs and NIAs and so lie near the origin of the NMA-NIA plot. Conversely, complexes formed by coupled binding and folding have much larger NMAs and NIAs and so are spread out and lie far from the origin of the NMA versus NIA plot. Indeed, a linear boundary separates the two groups (100). It should be emphasized that the NMA-NIA plot approach is a global measure of a protein's order-disorder monomeric state and has not been characterized on local order-disorder transitions (e.g. disordered binding loops in an otherwise well-ordered protein).

As described in more detail in the implementation, by developing two separate NMA-NIA plots, one for each partner of a complex (Figure 2A), and then by determining the distance to the linear boundary in each plot, a double NMA-NIA plot (Figure 2B) can be produced. Interacting pairs can be divided into the three groups given above, namely: (1) both partners are structured (region (i) of Figure 2B), i.e. both distances are negative; (2) one partner is structured and the second partner is disordered, i.e. the ordered partner has a negative distance and the disordered partner has a positive distance (regions (ii+ and ii-) of Figure 2B); and (3) both partners are intrinsically disordered, i.e. both distances are positive (region (iii) of Figure 2B).

Figure 2. Double NMA-NIA plot for p53 complexes. (A) The definition of boundary distance used in the double NMA-NIA plot, where ordered structures have a negative boundary distance and disordered structures have a positive boundary distance. (B) The double NMA-NIA plot for the p53 structures shown in Figure 1, with the exception of DNA-bound p53.



A double NMA-NIA plot was calculated for 13 of the p53 complex structures (Figure 2B). The p53-DNA complex was excluded since the NMA-NIA analysis is not relevant for nucleic acids. In the general case, the distinction between the distances of the two partners is arbitrary, so that the double NMA-NIA plot is symmetric about the diagonal. However, here we restrict the p53 distance to one axis, so that group (2) is split into two sub groups (regions (ii+ and ii-)): the p53 segment is disordered and the partner is ordered (region (ii-)), and the p53 region is ordered and the partner is disordered (region (ii+)). One interaction, the formation of the p53 tetramer, is in the third group (region (iii)) and so therefore likely involves an association between two disordered partners. This is consistent with experimental data (189). At the opposite side of the spectrum, the three protein-protein complexes involving the p53 DBD domain are in group 1 (region (iii)), indicating that all three are ordered prior to binding, which is consistent with the solution of structures for identical or homologous monomeric domains (e.g. p53 DBD (219), 53BP1 BRCT domain (220), 53BP2 SH3 domain (221), and LTag (188)). The other nine p53 complexes found so far in the PDB are all in the group 2 quadrant (that is, in region (ii-), and so all likely involve a disordered region of p53 associating with a structured partner. These results are likewise consistent with experimental data. That is, these p53 regions are disordered in the unbound state (189, 190), and the isolated partners appear to be structured: MDM2 (222), Rpa70 (179), Tfb1 (181), tGCN5 (223), Cyclin A/CDK2 (224), sirtuin (225), CBP (226), S100ββ (227), and set9 (228)).

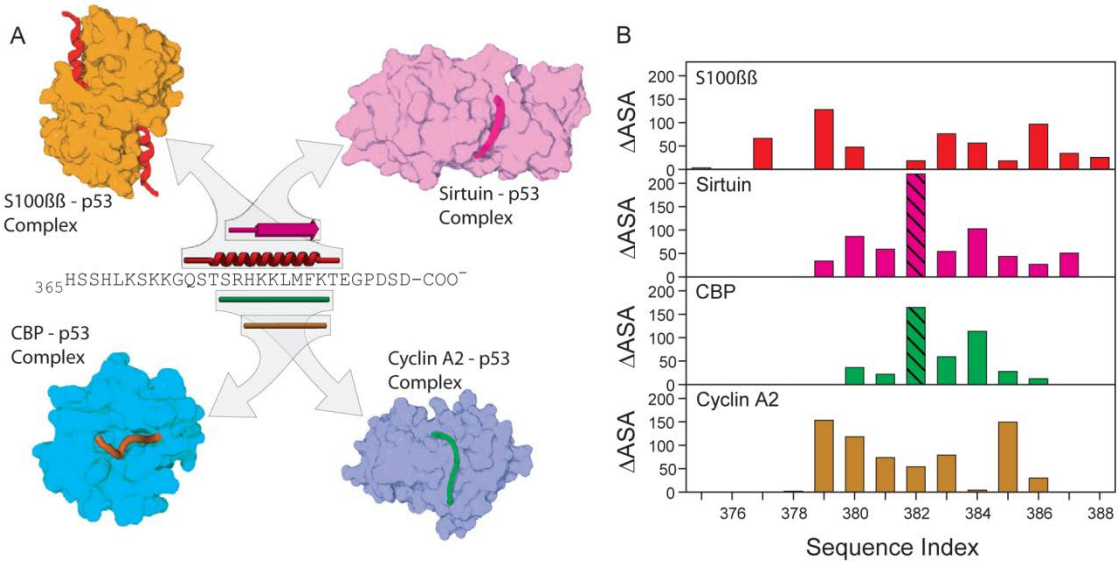
In summary, these data point out the importance of DOTs for many of the structurally characterized interactions involving the p53 hub protein. While many previous studies discuss these same interactions, to our knowledge the importance of disorder has not been emphasized in those previous studies.

Analysis of multiple specificities in the p53 C-terminus

So far, complexes involving one region of the p53 sequence bound to four different partners have been determined and deposited in the PDB. This region is from residue 374 to 388

in the p53 sequence bound to one of the following: cyclin A (173), sirtuin (174), CBP (175), or S100ββ (176). The regions that mediate these interactions and their respective secondary structures were mapped precisely to the p53 sequence (Figure 3A). Although slightly different residues of the p53 sequence are used in each interaction, there is a very high degree of overlap, with a span of 7 core residues being the same (Figure 3A). Interestingly, the four complexes display all three major secondary structure types. The core span becomes a helix when binding to S100ββ, a sheet when binding to sirtuin, and a coil with two distinct backbone trajectories when binding to CBP and cyclin A2 (Figure 3A).

Figure 3. Sequence and structure comparison for the four overlapping complexes in the C-terminus of p53. (A) Primary, secondary, and quaternary structure of p53 complexes. (B) The ΔASA for rigid association between the components of complexes for each residue in the relevant sequence region of p53. The two hatched bars indicate acetylated lysine residues.

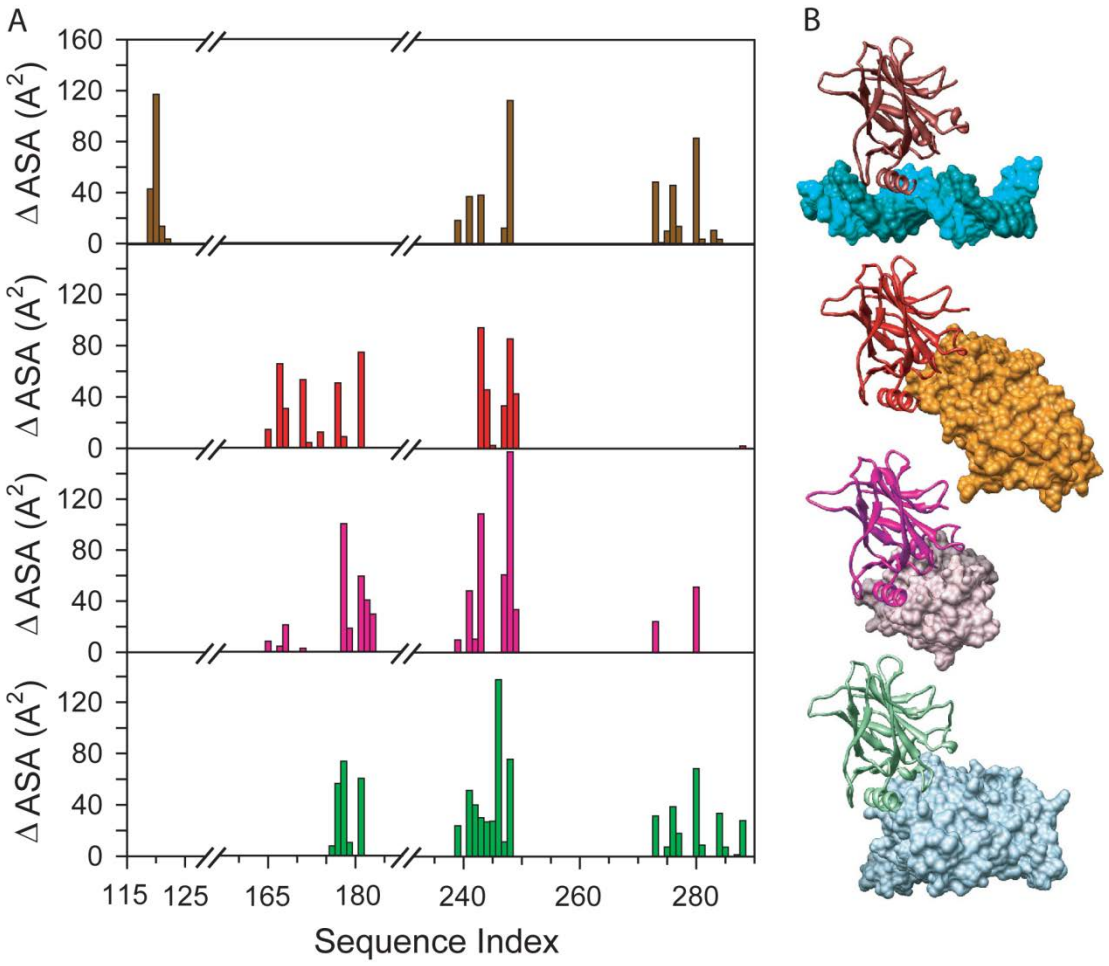


Because the secondary structures are distinct, it seems likely that p53 utilizes different residues for the interactions with these four different partners. To examine this, the buried surface area for each residue in each interaction was quantified by calculating the ΔASA (Figure 3B). Different amino acid interaction profiles are seen for each of the interactions, showing that the same residues are used to different extents in the four interfaces. The particularly large ΔASA peaks for K382 in complexes with CBP and sirtiun (indicated by the hatched bar) are due to extra buried areas arising from the acetylation of this residue. This highlights the importance of posttranslational modification for altering PPI networks.

Analysis of multiple specificities of the p53 DBD

The p53 molecule contains another set of overlapping interactions that contrasts with those at the C-terminus. These interactions are mediated by the DNA binding domain and include interactions with DNA (183), the BRCT domain of 53BP1 (184), the SH3 domain of 53BP2 (185), and the large T-antigen (LTag) of Simian Virus 40 (188). Here we compare these four interactions using the methods described in Figure 4.

Figure 4. p53 DBD interaction with different binding partners. The interaction profiles (A) and rendered structures (B) for the four unique complexes of the p53 DBD. Rendered structures depict p53 as a ribbon and each interaction partner as a molecular surface. The interaction profile-structure pairs are (from top to bottom): p53-DNA, p53-53BP1, p53-53BP2, and p53-sv40.

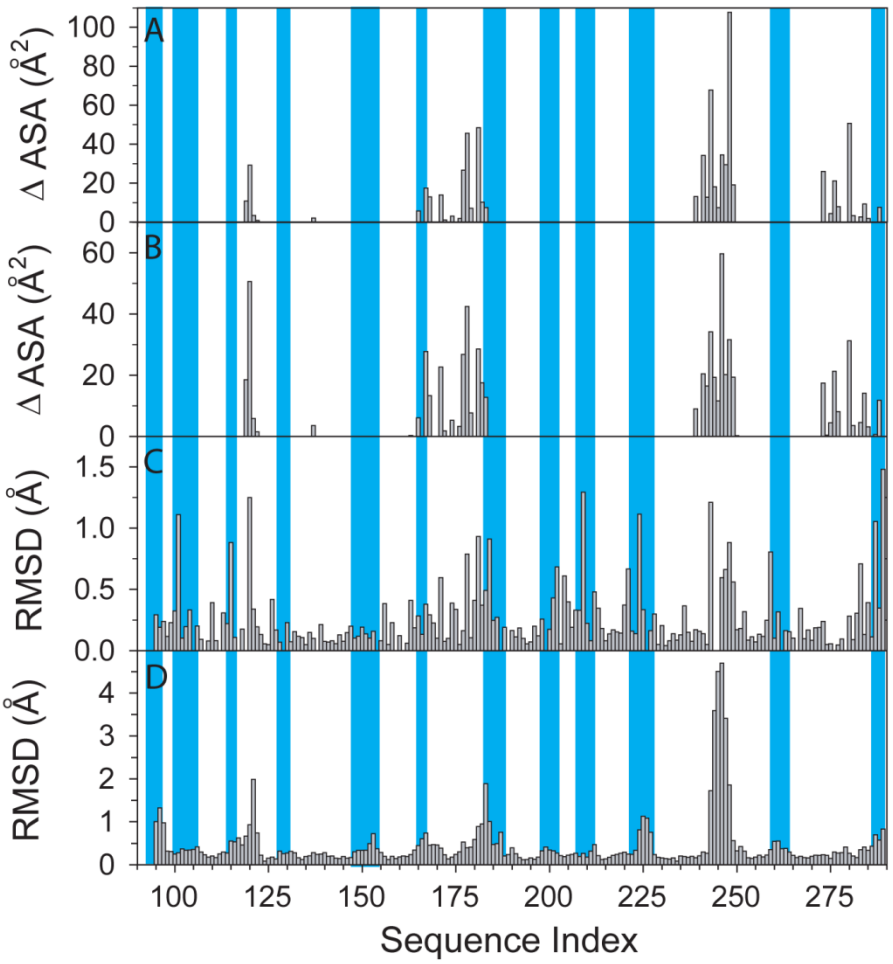


The structures of the p53-DNA, the p53-53BP1, the p53-53BP2, and the p53-LTag complexes are shown (Figure 4B). While all of the ligands are different, they all bind to basically the same region of p53.

Comparison of the interface profiles of the four complexes (Figure 4A) shows a large difference in the pattern of interface residues used by p53. For instance, there are several residues at the N-terminal end of the DBD which are only found in interaction with DNA. Similarly, interface residues near the C-terminal end participate in binding to different extents in three interactions but not at all in the p53-53BP1 interaction. The differing usage of residues in each interaction is the most prevalent feature of this data. However, there are also several residues contributing an exceptionally large amount of surface area in each complex (*e.g.*, M243 and R248).

While the focus of this paper is on the roles of disorder in the interactions involving two different hub proteins, the DNA binding domain of p53 presents the opportunity to study structural changes involving one structured region binding to several different structured partners. For this purpose, we compiled a 4-panel set of plots for characterizing the induced fit as one protein binds to different partners (Figure 5). These panels show the average interface area (Figure 5A), the standard deviation of the interface area (Figure 5B), the differences in side chain conformation (Figure 5C), and differences in backbone conformation (Figure 5D). Furthermore, regions that are highly exposed to solvent are also indicated (Figure 5, blue shading), so that structural differences due to interactions can be distinguished from those due from intrinsic flexibility – disordered loops – or crystallization artifacts.

Figure 5. Comparison of residue interactions with structural differences for bound p53 DBD. The average (A) and standard deviations (B) were calculated over the four interaction profiles of the p53 DBD shown in Figure 4. These are shown aligned with the side chain RMSF (C) and the backbone RMSF (D) calculated from the four structures of bound p53 DBD. Regions of residues that are highly exposed to solvent in all complex structures are indicated by the blue-shaded regions.



These induced-fit profiles exhibit a number of interesting features (Figure 5). The most striking of these is the region from residue 240 to residue 250. This region shows a large and variable interaction interface, which is associated with large side chain and backbone conformational differences. This is true also of a smaller region around residue 120. Other interaction regions show only side chain conformational differences associated with variable interface areas. Other conformational differences observed are limited regions of high solvent exposure, which suggests that these changes are due the details of the crystallization conditions more than interaction with a particular binding partner.

Together, these results suggest that multiple partners of p53 are accommodated by reusing similar binding interfaces. This is facilitated by small scale or large scale structural differences, which range from differences in side chain conformation to backbone rearrangements. It should be noted that this differs from our finding in a more limited analysis on only the p53-53BP1 and -53BP complexes (229).

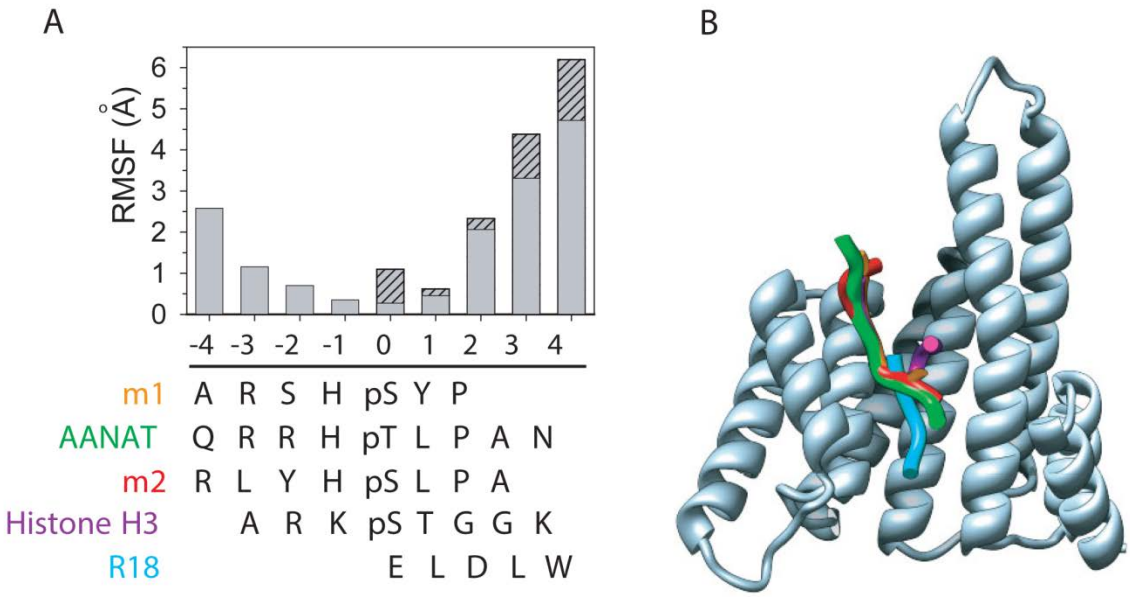
Analysis of the multiple specificities of 14-3-3

Five different 3-D structures of the 14-3-3ζ protein bound to distinct partners were found in PDB. These partners include a peptide from the tail of histone H3 (230), serotonin N-acetyltransferase (AANAT) (186), a phage display-derived peptide (R18) (187), and motif 1 and 2 peptides (m1 and m2, respectively) (170). For AANAT, only the region within the canonical 14-3-3 binding site is included in our analysis with the globular region being deleted. Two additional structures were not included because they were either unsuitable for structural analysis or were highly redundant with another structure. All peptides are phosphorylated in their respective structures except R18, which contains a glutamate in place of the phosphoserine.

The five bound peptides' sequences were aligned structurally as described in the methods. Likewise, the 14-3-3 domain structures were independently aligned, without considering the bound peptides. Next the 14-3-3 alignment was anchored manually by the observed correspondence of the bound peptide C_α atoms at the 0 and -1 positions and by extending the

alignment without gaps from the anchor positions, thereby giving the final structural alignment (Figure 6A). In terms of sequence, the R18 sequence has no identical positions to any other peptide. The number of identities between the other peptides range from 1 to 4.

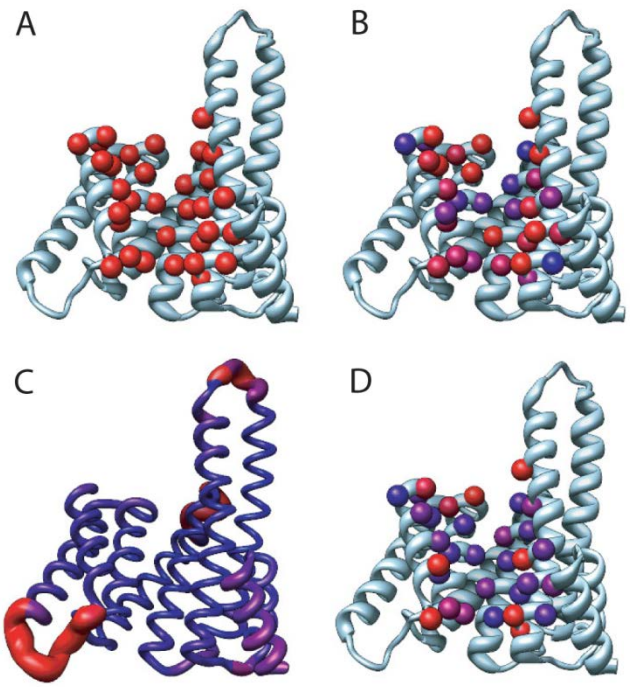
Figure 6. Sequence and structure for five peptides bound to 14-3-3ζ. (A) Sequence alignment of the bound peptides and the RMSF of their conformations. Solid gray bars give the RMSF for four peptides – excluding R18 – and the hatched bars give the RMSF for all five peptides. (B) Aligned ribbon representations of the structures of the five peptides, which were aligned through multiple alignment of their respectively bound 14-3-3 domains, shown along with a representative ribbon representation of a 14-3-3 domain.



There is high overlap in the backbone trajectories of the 5 peptides from position -3 to 1 but large divergences at either end of the structural alignment (Figure 6A). This divergence at the ends is apparent qualitatively in the superimposed structures of the five peptides (Figure 6B). Structural divergence and sequence variability are loosely correlated, where positions with three identical residues have a lower divergence than those with no identical residues. This suggests that 14-3-3 may use different binding pocket residues to interact with different peptide residues. The R18 sequence, which is divergent from the others, makes a large contribution to the estimated RMSF values (indicated by the cross-hatched bars, Figure 6A).

The factors contributing to the ability of 14-3-3 to bind to distinct peptides were estimated by a detailed structural analysis. The peptide binding residues of 14-3-3 are located primarily in a central cleft, made up of four helices (Figure 7A), which has been noted previously by several researchers. The standard deviation of ΔASA for the peptide binding residues (Figure 7B) show that the residues with the most binding variability are located at either end of the central cleft, which is consistent with the variation of peptide backbone trajectories in these regions. Backbone variability in bound 14-3-3 structures (Figure 7C) is restricted to the ends of most of the binding cleft helices. These observations suggest that large a conformational change in 14-3-3 is not necessary for multiple specificities, although some small adjustments at the ends of binding helices may be necessary.

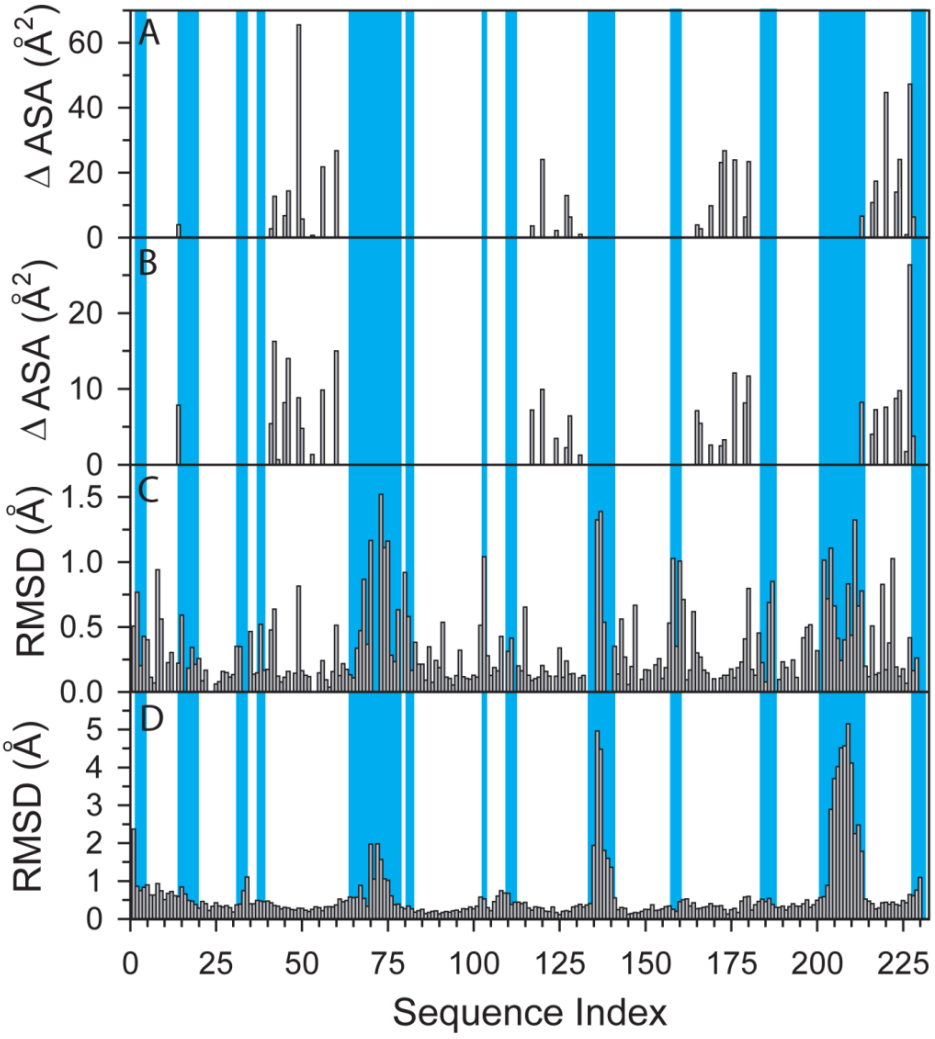
Figure 7. Peptide binding residues of 14-3-3ζ. (A) The C_β atoms of all residues involved in binding in any of the five peptide bound structures are shown (red) along with the rest of the backbone (light blue ribbon). (B) The standard deviation in the area bound on complex formation is displayed by coloring the C_β atoms of peptide binding residues on a gradient, from a standard deviation of 0Å^2 (blue) to 10Å^2 and greater (red). (C) The backbone RMSF of the 14-3-3 domain calculated over C_α atoms displayed as a color and radius gradient, from an RMSF of 0Å (blue, 0.25Å) to an RMSF of 2.0Å and greater (red, 2.0Å). (D) The side chain RMSF is displayed by coloring the C_β atoms of peptide binding residues on a gradient, from a RMSF of 0Å (blue) to an RMSF of 0.50Å and greater (red). All parameters were calculated using all five of the peptide-14-3-3 complexes.



To assess the role of side chain conformational changes in peptide binding, the RMSF of side chain atoms was calculated (Figure 7D). The side chain RMSF and standard deviation of ΔASAs give similar indications for many binding site residues, where residues used inconsistently across multiple complexes are the most likely to undergo conformational rearrangement. These are the same residues that are located at the broadest parts of the binding site. However, a few residues deep in the binding grove show both consistent participation in the binding interface and variable side chain conformation. These observations suggest that the primary, high level mechanisms of 14-3-3 multiple specificity are a broad binding site that allows multiple trajectories (and therefore interaction with different residues) and side chain rearrangement to accommodate different peptide sequences.

To further analyze the conformational changes in 14-3-3 upon binding to its multiple partners, we show the 4-panel induced-fit profile described above (Figure 8). Contrary to the results seen for the p53 DBD, 14-3-3 is much more static in its multiple interactions. All regions displaying large conformational differences across bound complexes are also highly exposed to solvent and play no direct role in mediating binding to any peptide. The plots do show several small scale structural differences – side chain rearrangements – associated with variable participation in peptide binding, particularly in the regions 40-60 and 215-230.

Figure 8. Comparison of residue interactions with structural differences for bound 14-3-3ζ. The average (A) and standard deviations (B) were calculated over the five 14-3-3ζ-peptide interaction profiles. These are shown aligned with the side chain RMSF (C) and the backbone RMSF (D) calculated from the five structures of bound 14-3-3ζ. Regions of residues that are highly exposed to solvent in all complex structures are indicated by the blue-shaded regions.

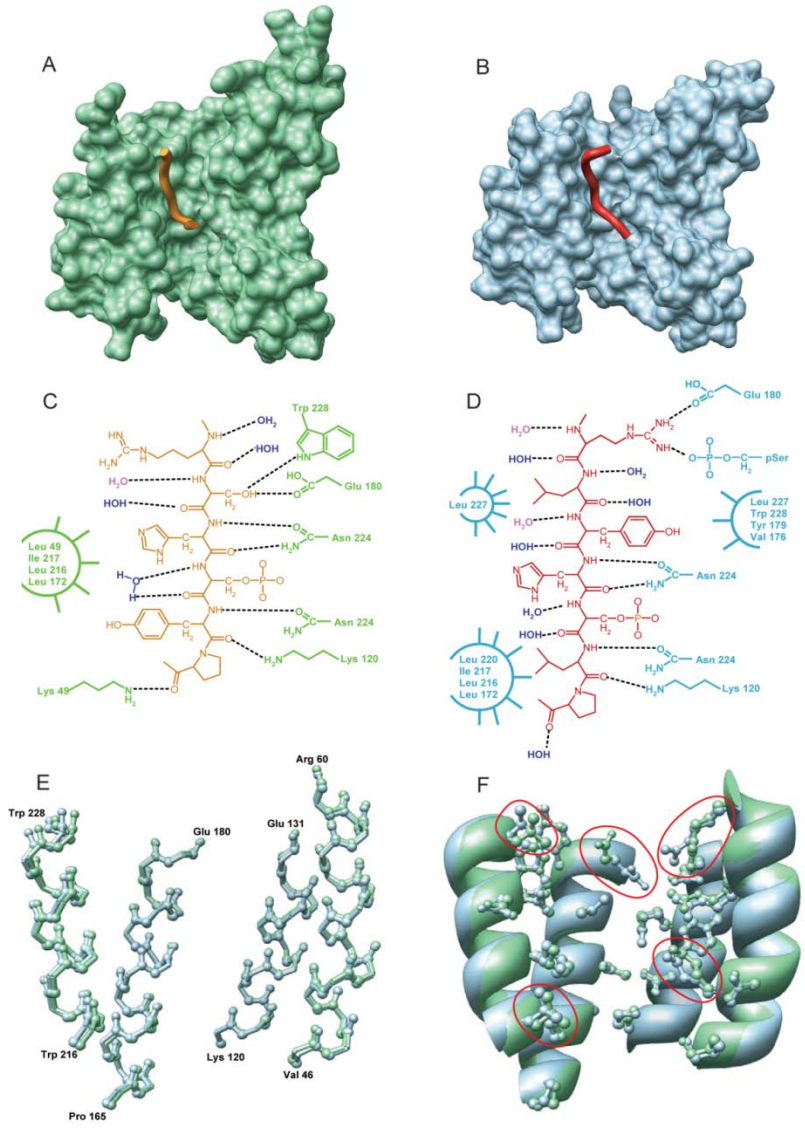


14-3-3 binding to two different partners

To gain further insight into 14-3-3 binding to different partners, we compared a pair of 14-3-3 binding peptides in detail. These two peptides, m1 and m2, were derived from two motifs, identified through the screening of peptide libraries for sequences that bound to all 14-3-3 isoforms (171). These two peptide structures have been compared previously (170), but here we reanalyze these structural data from the order-disorder point of view.

As noted previously (170), the backbone traces of the two peptides are noticeably different even though the m1 and m2 peptides bind to essentially the same region of 14-3-3ζ (Figure 9A and B, respectively). Examining the side chain interactions of these peptides with specific 14-3-3 residues (Figure 9C and D) shows that there is difference in the location and identity of the residues involved, which is consistent with the aggregate findings (Figure 7 and Figure 8). Similarly, distinctive hydrogen bonding patterns are exhibited between the two peptides and 14-3-3ζ and between the two peptides and bound water (Figure 9C and D). Since a cardinal feature of a structured protein is internal satisfaction of hydrogen bond donors and acceptors, these data are both consistent with the peptides being from unstructured regions of protein before binding.

Figure 9. Detailed analysis of 14-3-3ζ peptide binding. The m1 peptide (A, orange ribbon) and m2 peptide (B, red ribbon) bound to 14-3-3 (A and B, shown by the green and blue surface, respectively). Details of 14-3-3 peptide binding are shown by a chemical schematic for the m1 peptide (C) and the m2 peptide (D), where both crystallographic waters (blue) and implicit waters (red) are shown. (E) Superposition of the backbone atoms from the 4 helices with the primary peptide binding residues for m1 (green) and m2 (blue) bound 14-3-3. (F) Superposition of ribbons of the same 4 helices showing the side chains of the residues that participate in m1 (green) and/or m2 (blue) binding.



The above data on the complexes suggest that 14-3-3ζ has distinct conformations when bound to the two different peptides. Overlaying the backbone structures of the four binding helices from both complexes – based on a pair-wise alignment of the complete domains – shows only minor variability in conformation, with the most occurring at the helix spanning residues 216 to 228 (Figure 9E). Finally, comparison of side chain conformation in the two complexes shows significant differences in several of 14-3-3ζ side chains (Figure 9F, residues outlined in red show significant movement) and several other minor differences. Overall, these data suggest that a difference in the conformations of some side chains with rather less difference in backbone conformations is sufficient to accommodate the binding of two different phosphopeptides by the 14-3-3ζ molecule.

Discussion

Use of disordered regions for binding

The large majority of the binding sites on the p53 sequence map to the disordered regions of this protein (Figure 1), indicating that intrinsic disorder commonly provides the binding sites for the various partners that associate with p53. Recent bioinformatics investigations suggest that the majority use of disorder for binding to multiple partners is quite likely to be a general result (147–151).

The p53 binding sites are often indicated on the order-disorder predictions as dips, in other words as short segments with structure tendency flanked by regions of disorder tendency on both sides. Starting from this observation, we previously developed a predictor of such regions, which we called molecular recognition features, or MoRFs, because such regions "morph" from disorder to order upon binding (111, 194). Others have used the PONDR VL-XT order/disorder plots or MoRF predictors to identify potential binding sites that were subsequently verified by laboratory experiments (195, 231). For some of these predicted examples, the regions did indeed form helices upon binding to their partners (232, 233). By greatly enlarging the training set, we recently improved the MoRF predictor. Interestingly, when tested against several order-disorder

predictors including ones from other laboratories, PONDR VL-XT gave the clearest indication of binding sites within disordered regions (196).

Others developed a sequence-based approach to identify short, conserved recognition sites, called eukaryotic linear motifs (ELMs) (234–236). While MoRFs are identified by general order/disorder tendencies and while ELMs are identified by motif discovery from sequence analysis, the resulting binding sites identified by both methods share several features (237). The use of different residues in the same disordered fragment for one-to-many signaling leads to a potential problem with the ELM model. That is, the concept behind ELMs is that each ELM uses a common set of amino acids for binding to different partners. These common amino acids therefore show up as an over-represented pattern leading to a "linear motif". What if a region used to bind to multiple partners uses different secondary structures and different amino acids? In such a case, the residues in the "linear motif" would not necessarily be over-represented. It will be interesting over time to determine whether ELMs having stronger signals use a reduced set of structures for their interactions.

While the observed binding sites in the disordered regions of p53 have a localized tendency for ordered structure, not all disorder-associated binding sites exhibit such features. We have found many binding sites that are associated with high disorder prediction values across the entire spans of the binding sites, one example of which was recently published (238). Many of these dipless MoRFs form irregular structures upon binding with their partners, and often such binding regions are rich in proline. Our recent study of the complexes that form when various disordered segments bind to ordered partners indicates that the disorder-associated binding regions have distinct sequence features, even when the bound structure is irregular or sheet instead of helix, and so it should be feasible to develop a specific predictor for each of the different types of MoRFs (56).

One-to-many signaling

Date hubs bind to different proteins at different times. Figure 3 shows how a single region of p53 binds to four different partners. The amino acids involved in each interaction show a significant overlap, and no two of these interactions could exist simultaneously. Furthermore, the same residues adopt helix, sheet, and two different irregular structures when associated with the different partners. Finally, the same amino acids are buried to very different extents in each of the molecular associations. These results show very clearly how one segment of disordered protein can bind to multiple partners via the ability to adopt distinct conformations.

The idea that one segment of protein can adopt different secondary structures depending on the context is not new. Many unrelated proteins have identical subsequences of length six, and sometimes even up to length eight, with the same sequences often adopting different secondary structures in different contexts (239–241). Such sequences have been called chameleons for their ability to adopt different structures in different environments (240–246). Chameleon behavior could be an important feature that enables one disordered region to bind to multiple partners. With different secondary structures and with different side chain participation in the different complexes, it is as if one sequence can be "read" in multiple ways by the various binding partners.

Chameleon behavior occurs for short peptides (octamers), for longer protein fragments, and even for entire proteins. For example, the 17 residues-long arginine-rich RNA binding domain (residues 65–81) of the Jembrana disease virus (JDV) Tat protein recognizes two different transactivating response element (TAR) RNA sites, from human and bovine immunodeficiency viruses (HIV and BIV, respectively). The JDV segment adopts different conformations in the two RNA contexts and uses different amino acids for recognition (243). In addition to the above conformational differences, the JDV domain requires the cyclin T1 protein for high-affinity binding to HIV TAR but not to BIV TAR (243). Another protein with chameleon properties is human α-synuclein, which is implicated in Parkinson's disease and in a

number of other neurodegenerative disorders known as synucleinopathies. This protein may remain substantially unfolded, or it may adopt an amyloidogenic partially folded structure, or it may fold into α-helical or β-strand species, including both monomeric and oligomeric species. In addition, this protein can form several morphologically different types of aggregates, including oligomers (spheres or doughnuts), amorphous aggregates, and amyloid-like fibrils (145).

Such chameleon sequences likely underlie the multiple specificity binding sites common in p53. For a quick calculation of the implied degree of interface overlap, assume that each residue in a region has equal probability to interact with a partner and consider the C-terminus of p53. The disordered C-terminus (~100 residues) associates with at least 44 distinct partners. The average length of a binding site in this region is ~14 residues, which means that on average only 100/14=7 partners bind at any given residue in the C-terminus. This simple back-of-the-envelope calculation suggests that multiple specificity sequences may be the rule for p53 interactions, rather than a curiosity of a single region. However, available data suggest that interactions do not overlap in a random fashion, but rather interactions are localized to specific regions. For example, consider that the majority of the structures available for the C-terminus of p53 involved the same region of sequence. Therefore, the back-of-the-envelope calculation provides an approximate minimum degree of overlap, where the actual degree of overlap is likely much higher. This idea, which is an extension of a previous proposal (100), further suggests a general mechanism by which hub proteins could bind to such a large multitude of partners, which cannot be explained from the viewpoint of interaction between two structured proteins (161).

Finally, the p53 DBD offers a counter example to the disorder-based view of date hubs. That is, it uses the same or similar face of its globular structure to bind to multiple partners. While the p53 DBD is a folded protein, it does exhibit some remarkable structural differences when bound to different partners. It seems unlikely that these *local* regions of the p53 DBD structure are well folded in isolation, otherwise the association rate of some or all of these complexes would be relatively low. This idea is supported by the finding that the p53 DBD is

only marginally stable at physiological temperature (247). Therefore, it is plausible that these regions of the monomeric DBD are only transiently folded in solution, where crystallization conditions cause a shift toward the folded state in monomeric crystal structures. The double NIA-NMA plot data (Figure 2B) does not contradict this idea, since it is limited to global analysis and this idea only applies to local regions of the DBD. This idea is conjecture and further experimental or simulation evidence is needed to test this idea. In any event, however, the p53 DBD demonstrates that even in proteins generally thought to be well folded, structural changes can still occur in association with multiple specificity.

Many-to-one signaling

In 14-3-3, a common binding groove in a structured dimeric protein can be fitted by multiple, distinct sequences provided by many different binding partners. A recent bioinformatics study (172) found that14-3-3 proteins, and the 14-3-3 binding regions in particular, are predicted to be highly disordered by multiple disorder prediction methods. The authors proposed that 14-3-3 recognition generally involved coupled binding and folding of the recognition region. Our results support this conclusion because the backbone of m1 and m2 peptides are highly hydrated in the bound state (Figure 9C and D), indicating that the binding peptide is likely to be unstructured prior to binding (62).

One idea is that 14-3-3 holds its bound partner in a non-active state (172). Even though 14-3-3 likely binds to disordered regions in its partners (data herein and (172)), this idea of blocking the active structure could still be true. For example, the productive state of 14-3-3's partner might involve the binding of the partner to a second partner via the same disordered region that binds to 14-3-3, in which case 14-3-3 binding would prevent the formation of the productive complex. Another possibility is that the disordered region exhibits an equilibrium between a bound state that activates the protein and an unbound state that inactivates the protein. The association of the unbound disordered region with 14-3-3 would then hold 14-3-3's partner in the non-productive state as proposed previously.

We previously suggested that disordered segments with different sequences could use their flexibility to bind to a common binding site, thereby facilitating many-to-one signaling (146). The multiple recognition of 14-3-3 depends on this mechanism to a considerable degree, with the different peptides taking different paths through the binding cleft and interacting with binding site residues in distinct ways (Figure 6B).

In addition, structured proteins also have a degree of flexibility, and so the binding site backbone and side chain residues can undergo shifts (induced-fit mechanisms) to help accommodate interactions with distinct sequences (Figure 6 and Figure 8). Thus, induced-fit mechanisms are important for structured protein interactions with different partners whether the partners are structured or intrinsically disordered.

The induced-fit mechanisms observed for 14-3-3 and the DNA binding domain of p53 are commonly observed in other situations. For example, tethering, in which a peptide is covalently linked to its protein target to allow detection of low affinity interactions, often results large-scale side chain movements concomitant with peptide binding (248). Also, when many different MoRFs and their binding partners are examined, induced-fit movements in the structured partners are very commonly observed (56). Similarly, small backbone shifts and side chain conformational changes are both important for 14-3-3's ability to bind multiple partners. For all of these examples, the associations involve coupled binding and folding for the disordered peptide partner coupled with a near universal classical induced fit for the structured side of the partnership.

One-to-many signaling vs. many-to-one signaling

The p53 C-terminus and 14-3-3 use intrinsic disorder differently with regard to enabling multiple binding specificities. In p53, drastic conformational changes enable distinct surfaces to be exposed to binding partners. In 14-3-3, subtle differences in 14-3-3 conformation and peptide binding locations enable multiple specificities. Why would nature use one mechanism rather than the other for a particular biological role? The interactions of p53 serve to activate or inhibit its

primary role as a transcription regulator, while 14-3-3 alters the functions or subcellular localization of many proteins. From this, one can make some highly speculative proposals: (1) disorder binding regions play a passive role in regulation by providing a specific binding site – i.e. the disordered regions are the identification sites of the protein to be regulated (9) – and (2) ordered proteins play the active role – i.e. altering the activity of the proteins they bind to – where recognition of disordered regions allows for a generalized specificity so that a single protein can alter the activity of many others. Validation of the accuracy and generality of these ideas requires further study.

Conclusion

Here we have examined the mechanisms of multiple specificities in two date hub-like hub proteins. Evidence here and elsewhere (147–151) suggests that disordered regions may be an extremely common mechanism by which hub proteins bind to their multitude of partners. The specific examples of p53 and 14-3-3 contrast the mechanisms by which disorder facilitates multiple recognition. Two regions of p53 are highlighted: the C-terminus, which is intrinsically disordered, and the central DNA binding domain, which is structured. The C-terminus binds four different partners with the same residues, forming distinct structures and using distinct residues in their respective complexes. The structural heterogeneity of the bound complexes highlights the importance of intrinsic disordered in the unbound state. The DNA binding domain also binds to four different partners, where these complexes use some distinct residues and some common residues. Residues that participate in multiple complexes show small to large scale structural rearrangements, which indicate either induced fit and/or structural instability in these regions. Both the C-terminal and DBD examples demonstrate how proteins can accommodate multiple partners through distinct conformations. A contrasting example is provided by 14-3-3, which binds to multiple partners without large scale conformational changes. 14-3-3 accommodates multiple partners with distinct sequences via broad binding surfaces surrounding the common charge clamp. Individual partners take different path over the binding surface with

accompanying changes in side chain conformation. The variably in conformations of bound partners highlights how disordered fascilitates multiple recognition of 14-3-3.

These examples suggest two distinct models of disorder mediated multiple recognition: one-to-many, where one disordered region recognizes multiple partners though assuming distinct conformations; and many-to-one, where many disordered regions recognize a single binding site on a structured partner. Work subsequent to that presented here has found dozens of more examples of one-to-many (61) and many-to-one (Hsu et al., manuscript in preparation) binding in PDB. For one-to-many binding, examples mirror binding observed for p53 C-terminus, where the region adopts unique conformations to accommodate different partners. For many-to-one binding, binding sites showed widely varying amounts of overlap on the binding surface, which is consistent with the findings for 14-3-3. This work indicates that these binding models may be common in nature.

A next step in this research is to distinguish single target disordered regions from one-to-many binding. Specifically, several methods for predicting disordered regions have been developed (see Chapter IV), and could be extended for this purpose. One possible avenue for extension is examination of relative conservation of single target and one-to-many disordered binding regions. The hypothesis is that sequences of one-to-many binding regions will be more highly constrained between species than single target binding regions. That is, residues in one-to-many regions will be more resistant to mutation due to their participation in multiple interfaces, versus single target regions that can undergo correlated mutation with their partners without effecting other interactions. Distinguishing single target and one-to-many regions is useful generally to find the relative prevalence of these regions, and also to help determine possible functions of novel binding regions.

Methods

PONDRs VL-XT and VSL1

Predictions of intrinsic disorder in HPV proteins were performed using a set of PONDR (Predictor Of Natural Disordered Regions) predictors, VL-XT and VSL2. PONDR VL-XT integrates three feed forward neural networks: the Variously characterized Long, version 1 (VL1) predictor from Romero *et al*. 2001 (8), which predicts non-terminal residues, and the X-ray characterized N- and C- terminal predictors (XT) from Li *et al.* 1999 (249), which predicts terminal residues. Output for the VL1 predictor starts and ends 11 amino acids from the termini. The XT predictor's 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 9 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.

The recently developed Various Short-Long, version 1 (PONDR-VSL1) algorithm is an ensemble of logistic regression models that predict per-residue order-disorder (192, 250). Two models predict either long (>30 residues) or short (<15 residues) disordered regions based on features similar to those used by VL-XT. The algorithm calculates a weighted average of these predictions, where the weights are determined by a meta-predictor that approximates the likelihood of a long disordered region within its 61-residue window. Predictor inputs include PSI-blast profiles (251), and PHD (252), and PSI-pred (253) secondary structure predictions.

Structure surface and complex interface analysis

Solvent accessible surface area (ASA) was calculated from atomic protein structure numerically using the double cubic lattice method (254) as implemented in the Biochemical Algorithms Library (255). Using this algorithm, ASA of residues and entire chains can be calculated.

To determine interface areas, for example between two chains, the ASA of each individual chain is calculated, as well as the ASA of the complex. The interface area is then

calculated as the change in ASA (ΔASA), i.e. the sum of the individual chain ASA minus the complex ASA. Residues directly involved in interactions were identified from molecular structures as residues with a ΔASA greater than 1 $Å^2$ (214, 215). All calculations used a probe radius of 1.4 Å, which roughly corresponds to the size of a water molecule.

Order-disorder evaluation from known structure

The work of Gunasekaran et al. has previously shown that, in many cases, the order-disorder state of a protein prior to complex formation is reflected in the complex structure (100). Specifically, a plot of the normalized monomer area (NMA) – ASA divided by the number of monomer residues – versus the normalized interface area (NIA) – ΔASA divided by the number of monomer residues – effectively distinguishes between ordered and disordered monomers using a linear boundary. This effectiveness of this NMA-NIA plot has been validated on an expanded dataset and an optimal linear boundary has been estimated and evaluated. The equation for the novel boundary is:

$$\langle NMA \rangle = 157.43 - 3.51 \langle NIA \rangle$$

Since the NMA-NIA plot can only represent one partner of a complex, the double NMA-NIA plot was developed to simultaneously represent both monomers of a binary complex – or complexes that can be treated as binary, such as two monomers bound to a dimer. Rather than plotting the NMA and NIA directly, the Euclidean distance to the order-disorder boundary is calculated, where disordered monomers have a positive distance and ordered monomers have a negative distance. Then the boundary distances of each monomer in a binary complex can be plotted against each other to give an overall order-disorder prediction for the complex.

Other structure calculations

The root mean squared fluctuation (RMSF) is a commonly used measure of variability across multiple structure alignments. Here, RMSF of the protein backbone is approximated as the RMSF of the C_α atoms. The equation used is

$$RMSF_i = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(C_{\alpha_{j,i}} - \overline{C_{\alpha_i}}\right)^2}$$

where $C_{\alpha_{j,i}}$ is the position vector of the ith C_α atom of the jth complex and $\overline{C_{\alpha_i}}$ is the averaged position for the ith amino acid from the multiple sequence alignment of N structures. The program MultiProt (256) was used to generate the multiple sequence alignments for RMSF calculation and structure rendering.

To estimate side chain conformation variability among multiple protein structures, the RMSF of side chain residues was calculated. In this calculation, the residue atoms C_α,C_β, backbone carbonyl carbon, and backbone nitrogen were used to align a residue to a selected reference residue of the same type. Thus aligned, the RMSF was calculated over side chain carbons beyond the C_β. Consequently, no side chain RMSF was calculated for Glycine or Alanine residues. The RMSF was also corrected for the number of atoms in the side chain beyond the C_β.

The solvent accessibility of individual residues was calculated relative to an extended Gly-X-Gly model peptide (257), which gives a conservative estimate of relative solvent exposure, i.e. underestimates relative solvent exposure. Residues exposed to solvent were defined as those with an accessible surface area at least 40% of that of the reference area for that residue type. This cutoff is arbitrary, but cutoffs for solvent exposed residues as low as 20% have been used by others, e.g. (258). Solvent exposures were calculated in the context of binary complexes, which is valid for p53 complexes. In 14-3-3 complexes, 14-3-3 forms homotypic dimmers in addition to binding to phosphopeptides, so residues found to be highly solvent exposed are either actually exposed to solvent or involved in the homodimer interface.

III. UTILIZATION OF PROTEIN INTRINSIC DISORDER KNOWLEDGE IN STRUCTURAL PROTEOMICS

Introduction

Structure determination, historically, has been attempted on a protein-by-protein basis, typically after an accumulation of years or decades of study on each particular protein. Information regarding solubility, stability, pH range, and temperature sensitivity was therefore generally well known. Consequently, due to lack of suitable samples, structure determination was not usually attempted on proteins that were ill behaved in solution. The Protein Structure Initiative (PSI) or Structural Genomics Initiative (SGI) (259–262) has turned the *status quo* on its head by attempting to determine structures without prior knowledge of a protein's behavior and to do this rapidly on a large scale. Despite the obvious success of the PSI Centers in decreasing the overall cost of determining novel structures (263), ill-behaved proteins continue to represent major challenges that hamper the efficiency of structure determination. One source of ill-behaved proteins is intrinsic protein disorder.

Prediction techniques and data mining have shown that intrinsically disordered proteins (IDPs) and proteins with regions of intrinsic disorder are likely to be quite common (7, 8, 10, 49, 264–266). Disordered proteins or regions are defined herein as entire proteins or regions of proteins that lack a fixed tertiary structure. A given region of intrinsic disorder might be ordered or disordered, depending on the physiological or experimental conditions. For instance, a protein may undergo a DOT upon binding to a cofactor, DNA, or protein partner. On the other hand, proteins that lose rigid structure in the presence of denaturants are not considered to be disordered proteins. Similarly, proteins that fold into specific 3-D structure only under condition of extreme molecular crowding are not considered ordered. IDPs and intrinsically disordered regions are noticeably different from structured globular proteins and their domains at several levels, including amino acid composition, sequence complexity, hydrophobicity, charge, and flexibility.

Many of these differences were used in the development of several disorder predictors (reviewed (41, 267–271)).

Disordered protein can impact many of the processes in the structure determination pipeline, including expression and stability (272, 273), solubility (201, 274), and crystallization (12, 13, 201, 275–277). Therefore, it is advantageous to filter highly disordered proteins from the target list. Also, many structured proteins contain isolated regions of intrinsic disorder, which can inhibit crystallization. Fragments remaining after removal of disordered regions may crystallize when the whole protein did not (201), and so a method for identification of disordered regions to allow for intelligent target improvement would be of great utility.

Here we examine the impact of intrinsic disorder on the various stages of the structure determination pipeline and outline several applications of Predictors of Natural Disordered Regions (PONDRs) that can improve the efficiency of structure determination efforts. To evaluate the tolerance of intrinsic disorder in protein crystals, missing density in the Protein Data Bank (PDB) was evaluated in light of the various sources of missing density in addition to intrinsic disorder. This analysis indicates that extensive disorder in PDB is relatively rare and highlights the benefit of filtering crystallization targets for intrinsic disorder. Several disorder-based target prioritization criteria are evaluated and retrospectively applied to protein targets in the TargetTrack database, the progress tracking module of the PSI Structural Genomics Knowledgebase. This evaluation suggests that disorder prediction provides an effective means for prioritizing targets for structure determination pipelines. Finally, the use of disorder prediction for tailoring proteins for structure determination is examined.

Materials and Methods

Protein datasets

For analysis of missing density, protein structures without nucleic acid from the July 2012 version of PDB (278) were used in conjunction with the S2C database (http://dunbrack.fccc.edu/Guoli/s2c/index.php). The latter provides alignments between residues

with defined density and the reported sequence (SEQRES), greatly simplifying identification of apparent missing density. The number of residues and proportion of apparent missing density were calculated with respect to the contents of the asymmetric unit. Several structures were excluded from the list of cases of extreme missing density due to idiosyncratic issues with these structures: 1EFM because a large portion of defined density was not fit by the author (279), 1YD7 because no supporting reference was available and neither solvent content nor Mathew's coefficient was reported, and 3JQH and 4HB1 because these structures contain chains with exact repeats and the asymmetric unit contains less than one copy of the chain in the asymmetric unit. Three idealized sets of proteins were assembled for these studies. The "ordered proteins" set consists of the entire sequences of all the X-ray crystallography-determined structures from PDB Select 25 proteins (280). This set contains 929 proteins with 230,205 residues. The "disordered proteins" set consists of proteins compiled from literature searches where the known disordered regions comprise $\geq 80\%$ of the entire sequences. This set contains 48 proteins with 9,219 residues. The "X-ray proteins with disorder" set consists of crystallized proteins with disordered regions ≥ 30 residues. This set contains 53 proteins with 17,191 residues.

We also analyzed all the proteins currently available at the PSI Structural Biology Knowledgebase (http://sbkb.org), namely at its target progress tracking module (TargetTrack: http://sbkb.org/tt/), which provides status and tracking information on the production protein targets and their structure solution (281). Targets' sequence and status were taken from the XML distribution of the database, as of July 2012. Status tags for each trial were normalized by adding tags from previous pipeline steps that were missing. For targets with multiple trials, the trial showing the furthest pipeline progression was selected as representative of the target. From 302,311 selected targets deposited by worldwide Contributing Centers in to TargetTrack, 201,825 were cloned, of which 120,802 were expressed. Of the expressed proteins, 51,303 were purified, of which 8,168 were crystallized, of which 5,125 produced diffraction-quality crystals. So far,

according to the July 2012 TargetTrack, 4,203 crystal and 1,600 NMR structures have been solved.

Intrinsic disorder prediction

PONDR algorithms predict per-residue, two-state order/disorder from amino acid sequence (40, 249, 282). Although several PONDRs have been developed to date (8, 40, 193, 249, 250, 282–286), only one of these predictors, VL-XT, is used here. This choice of this predictor was made because PONDR VL-XT is one of the best characterized members of the set of PONDR predictors.

PONDR predictions are real numbers between 0 and 1, where values of 0.5 or more are predictions of disorder. Raw PONDR prediction output is commonly plotted, as distribution of disorder scores over the protein sequence. An alternative representation is to map predictions to a bar, where black represents a region predicted to be disordered and white represents a region predicted to be ordered. Predictions can also be quantified per-protein, using measures such as percent of residues predicted to be disordered or the longest predicted disordered region. These measures are particularly useful because of the dependency of false positive rates on the length of the disorder prediction.

PONDR VL-XT was constructed from a neural network trained on disordered proteins collected from the literature. The VL-XT predictor integrates three feed-forward neural network predictors: the various long predictor version 1.0 (VL1) (284), the N-terminus predictor (XN), and the C-terminus predictor (XC) (249), and achieves good accuracy both on held-out training set sequences and on out-of-sample sequences (8). Prediction of VL-XT on O_PDB_S25 (8), which is a set of defined density regions from sequence unique structures from PDB, gave an accuracy of 80%. Prediction on the disordered data gave an accuracy of 60%, suggesting that VL-XT does not generalize well for prediction of disorder (8).

PONDR is available on the web at http://www.pondr.com. The web interface does not support batch predictions. Arrangements for large numbers of predictions can be made by

contacting the corresponding authors. It is important to note that although there are many predictors of protein intrinsic disorder (41, 267–271), we anticipate that similar results would be found with other predictors.

Model fitting

To evaluate combinations of disorder prediction based criteria, logistic regression models were fit using proteins that reached a given status in the structure determination pipeline as a positive set and remaining proteins as a negative set. The purpose of fitting these models was not to create binary predictors of status progression, but rather to determine an appropriate linear combination of parameters. The statistics package R (http://www.r-project.org/) was used to fit models, using the standard generalized linear model function to fit models and the package ROCR (287) to calculate the area under the curve (AUC) measure of predictor performance. The AUC was selected to compare models because it emphasizes the separation of the positive and negative dataset, where 1.0 is prefect separation and 0.5 is completely interspersed, and is independent of any kind of prediction threshold.

Results and Discussion

Missing density and intrinsic disorder in the PDB

To evaluate the tolerance of crystal structures for intrinsic disorder, we examined the -extent of missing density in existing crystal structures. Although missing density regions have often been equated to intrinsic disorder, apparent missing density in crystal structures may arise from several sources, including disordered regions, mobile domains, and proteolysis. Disordered regions may be present in the crystal but fail to diffract X-rays due to heterogeneous or dynamic structure. For multi-domain proteins, mobile domains may be present; a domain may have a heterogeneous or dynamic orientation and position in the crystal, resulting in missing density in the solved crystal structure. Although not intrinsically disordered themselves, the mobility of domains may be facilitated by intrinsically disordered linkers. Finally, rational or fortuitous proteolysis may occur prior to or during crystallization, but these altered sequences are often not

reflected in the PDB record. While several authors have noted that removal of intrinsically disordered regions can facilitate crystallization and diffraction, e.g. Ref. (201), the use of proteolysis is not consistently reported and the integrity of crystallized chains not routinely verified.

The most extreme cases of apparent missing density in terms of longest region of missing density and largest fraction of residues in missing density regions were collected from PDB (Table 1). All of these structures have over half of their residues in apparent missing density regions or at least one apparent missing density region longer than 200 residues. Much of the missing density found was due to the anticipated sources, (likely) proteolysis, (likely) mobile domains and intrinsic disorder, but much apparent missing density was due to misannotation. Misannotations are less frequent than other sources of missing density and are limited to cases where only a portion of a sequence was included in experiments, but a longer sequence was reported in the PDB record, e.g. PDB ID: 4ACO. For the purposes of this study, apparent missing density due to misannotations is functionally equivalent to proteolysis. Discounting misannotation and proteolysis, and PDB record limitations, missing density regions attributable to residues physically present in protein crystals are limited to about 50% of residues when mobile domains (around 40% of residues for confirmed cases of mobile domains) are present and about 30% of residues for intrinsically disordered regions.

Table 1. Extreme cases of missing density in the PDB. The most extreme cases of apparent missing density in the PDB in terms of the longest missing density region (LMDR) and the fraction of residues in missing density regions (FMDR).

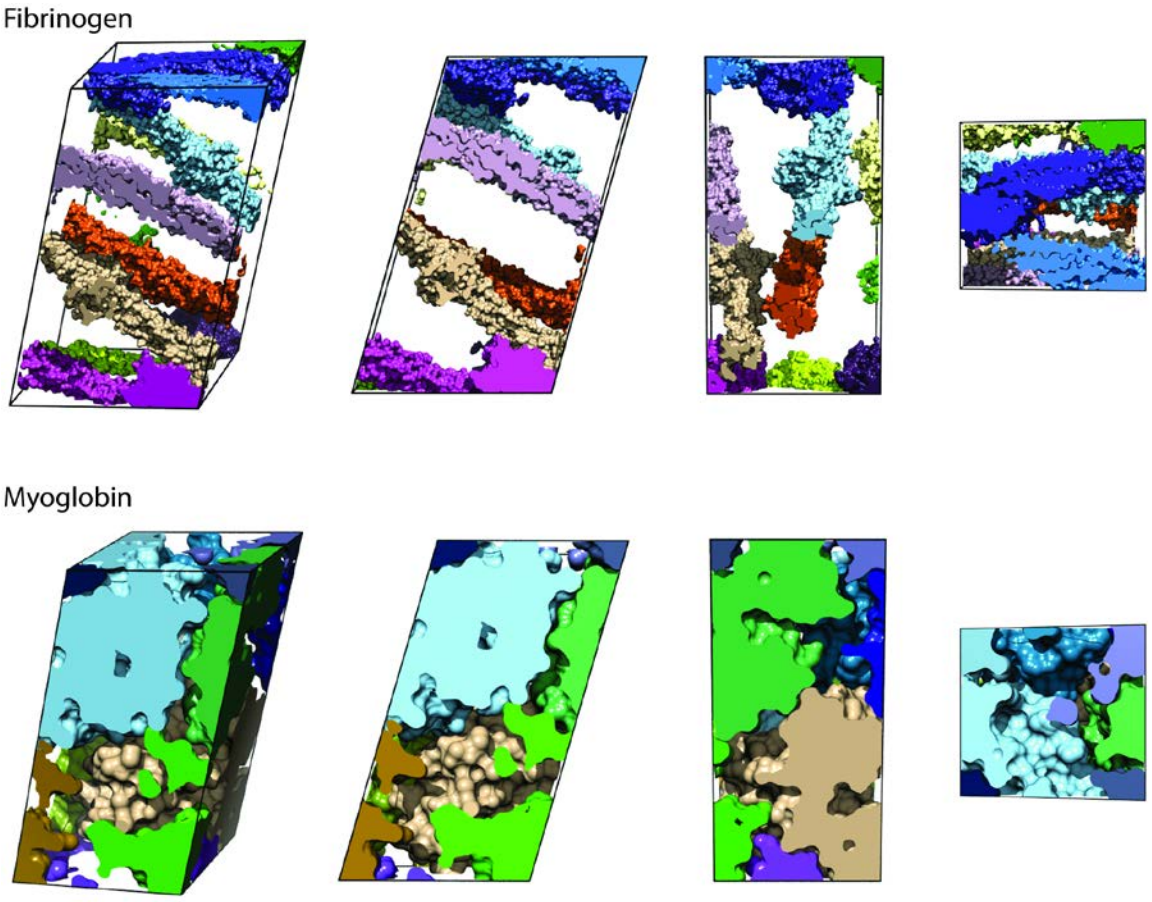
Structure	Residues	LMDR	FMDR	Reason for extreme missing density	Reference
1M1J	2744	273	0.28	Disordered	(288)
1SCF	1092	122	0.59	Misannotation	(289)
1SRQ	1364	208	0.25	Mobile domain	(290)
2HQX	492	116	0.63	Likely proteolysis	(291)
2HZ7	851	220	0.35	Mobile domain	(292)
2IG2	671	217	0.32	Likely mobile domain[1]	
2J4W	883	378	0.47	Likely mobile domain	(293)
2J5L	1019	446	0.54	Likely mobile domain	(293)
2JCH	720	217	0.36	Proteolysis	(294, 295)
2JE5	1440	222	0.36	Proteolysis	(295)
2JKU	94	36	0.63	Proteolysis	(296)
2VF4/2VF5	608	242	0.41	Mobile domain	(297)
2WZL	303	196	0.65	Proteolysis	(298)
3BJC	878	537	0.65	Misannotation	(299)
3DUF/3DVA/ 3DV0	3632	263	0.22	Misannotation	(300)
3EVY	478	147	0.65	Proteolysis[2]	
3FG7	796	277	0.73	Proteolysis	(301)
3GHG	5768	362	0.33	Proteolysis and Disordered	(302)
3JZY	510	385	0.76	Proteolysis[2]	
3KIC/3KIE	14720	216	0.29	Likely disordered[1]	(303)
3L7L/3L7M/ 3L7J/3L7I	2916	313	0.44	Misannotation	(304)
3NG9	7360	216	0.29	Likely disordered[1]	
3ORE	988	220	0.25	Likely mobile domain	(305)
3R05/3QCW	2048	230	0.23	Likely mobile domain	(306)
4ACO	956	419	0.53	Misannotation	(307)
4AQ1	2044	264	0.22	Likely mobile domain	(308)

[1] Classification based in part on disorder prediction

[2] Based on authors note in the PDB record

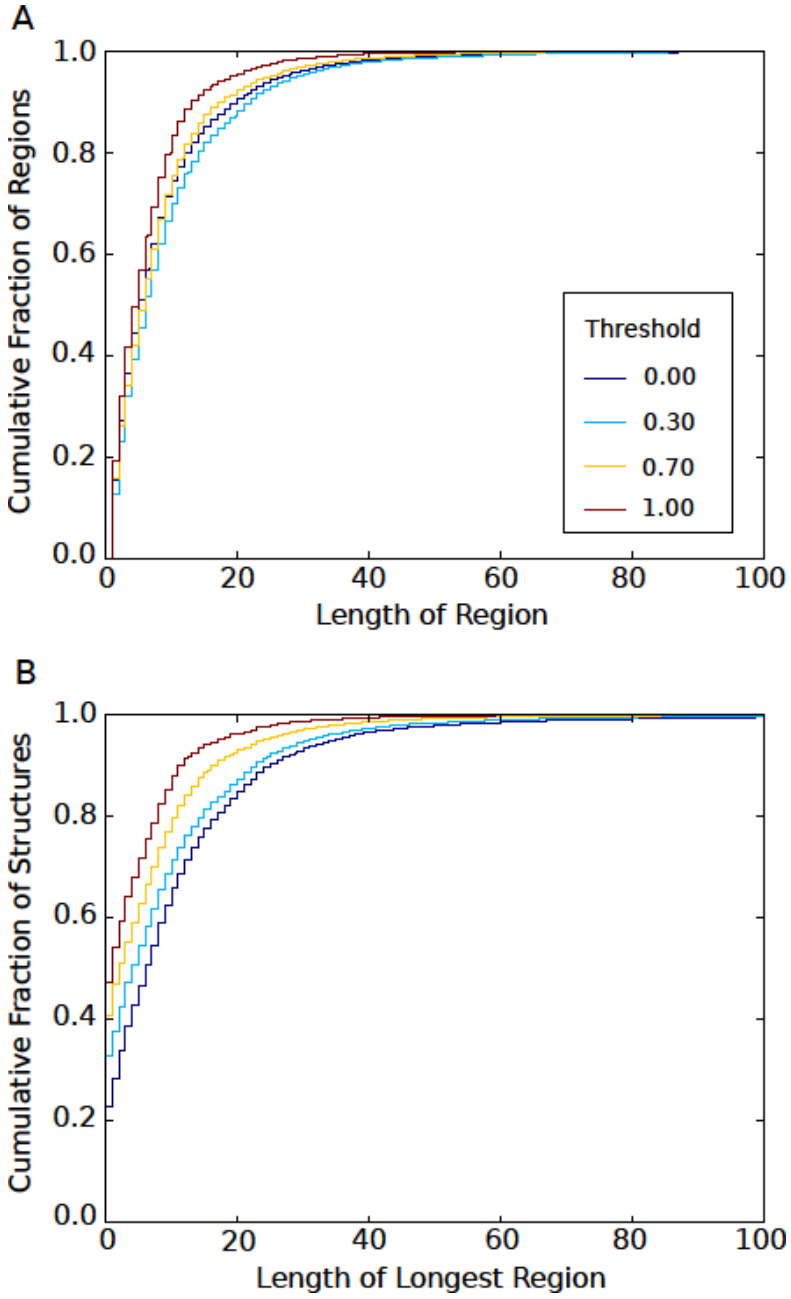
To illustrate the implications of a large amount of intrinsic disorder on a protein crystal, the crystal lattice of the protein with the longest disordered region – fibrinogen (Figure 10, upper panel) – was rendered and compared to that of a well ordered protein – myoglobin (Figure 10, lower panel). The crystal packing for fibrinogen clearly displays large gaps relative to the tight crystal packing of myoglobin. However, the reported solvent content of each of these structures is comparable, 65.3% and 64.4% for fibrinogen and myoglobin, respectively. This implies that the large gaps in the crystal lattice are packed with relatively compact intrinsically disordered residues. One possible explanation for the high tolerance for intrinsic disorder in fibrinogen crystals is that the elongated structure of fibrinogen allows for a lattice with more gaps that would not be tolerated by packing of a globular protein.

Figure 10. Examples of the crystal packing. Packing of the unit cells of two different crystal structures are shown: fibrinogen (1M1J) and myoglobin (1MDN). The surfaces of each copy of the asymmetric unit are rendered in different colors and truncated by the faces of the unit cell.



In addition to extreme cases of missing density in PDB, missing density across crystal structures in PDB was evaluated (Figure 11). To account for the presence of mobile domains, we applied disorder prediction to the sequences of all PDB chains and filtered missing density regions by the fraction of residues predicted to be disordered. However, we were unable to account for proteolysis or misannotation in this analysis, so these results should be considered an upper bound; there is less missing density due to residues physically present in the protein crystal than indicated by these results. The cumulative fraction of missing density regions by length (Figure 11A) shows that missing density regions are often short, with between 85% and 95% of regions being 20 residues long or shorter depending on the predicted disorder cutoff. Also, missing density regions with a moderate fraction of predicted disorder (0.3) tend to be longer than all missing density regions and regions with a larger proportion of predicted disorder, which suggests that many of these intermediately disordered regions may include mobile domains. The cumulative fraction of structures by length of its longest missing density region (Figure 11B) shows that between 23% and 47% of structures have no missing density residues and that the vast majority of structures, between 85% and 95%, contain no missing density regions longer than 20 residues.

Figure 11. Cumulative histograms of missing density in the PDB. Cumulative histograms of (A) missing density regions by length and (B) structures by length of the longest missing density region in any protein chain in the structure. For each, missing density regions were filtered by the fraction of residues predicted to be disordered: no filtering (threshold of 0.0), at least partially disordered (threshold of 0.3), mostly disordered (threshold of 0.7), and completely disordered (threshold of 1.0).



The overall picture of missing density in the PDB provided by this analysis is that missing density is infrequent in PDB structures, and when present tends to be limited to relatively short regions. When long regions of disorder are present in the crystal (Table 1, structures 1M1J, 3KIC, and 3NG9) the total proportion of disordered residues in the structure is below 30%. Disorder prediction suggests that much of the missing density in PDB does arise from intrinsically disordered regions, while some arises from mobile domains and an unknown amount from proteolysis and misannotation. The limited amount of intrinsic disorder present as missing density regions agrees with the idea that intrinsically disordered regions, particularly long disordered regions, inhibit successful determination of crystal structures, and it suggests that avoiding or tailoring disordered proteins may aid in the determination of crystal structures.

Target prioritization with disorder prediction

Per-protein assessments of disorder by PONDR can contribute to a target prioritization scheme in a straightforward manner. For example, a cut-off value for a given quantity establishes a filter, while weights determined from the given measure provide a means for prioritization. Using different scoring schemes, specific cut-off values are suggested below, but it is suggested that alternative values be considered for each particular application in order to balance the rates of false prediction of disorder with tolerable disorder content. Likewise, the relative weights calculated from the different disorder schemes could be used to prioritize the selected targets with the proteins having less predicted disorder receiving the higher ranks.

The genome-wide measures of predicted disorder will be effective in prioritization, provided proper attention is paid to the balance between false prediction of disorder and false prediction of order. As illustrated in Table 2, the rate of false prediction of disorder drops drastically with increasing number of contiguous predictions of disorder. However, there is a corresponding increase in false prediction of order. There is no clear indication of the extent of this increase, due to the limited number of proteins in the current disorder protein dataset and uncertainty in the boundaries of disordered regions. Prediction sensitivity will be most directly

affected by limiting the minimum size of disordered regions that prediction will detect. Because many procedures in molecular biology, including protein crystallization, can be relatively insensitive to small regions of disorder, ignoring contiguous disorder predictions of modest length (e.g., less than 40 residues) should not adversely affect the practical use of PONDR predictions. The thresholds for any of the weighting schemes presented should be selected, in reference to Table 2, so that the overall rate of false prediction of disorder is in balance with the desired preference for the sensitivity of the predictions. Finally, it should be noted that the rate of false prediction of disorder may be somewhat overestimated in Table 2. The O_PDB_S25 contains chains from complex structures, which may contain DOT regions involved in binding (7) that are correctly predicted to be disordered.

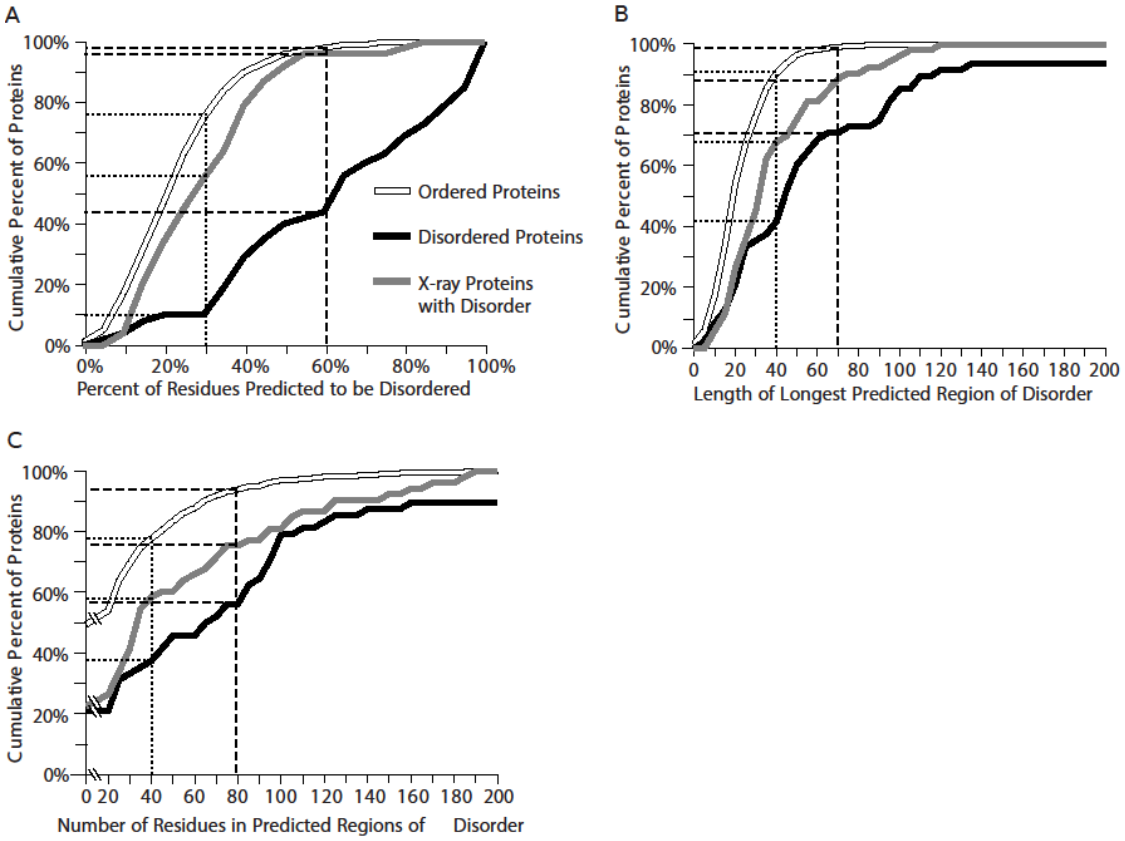
Table 2. PONDR VL-XT error rates as a function of number of consecutive prediction of disorder. Pre-residue error rates are calculated as the number of amino acids in predicted disordered region of the given length or longer, divided by the number of residues in the set of ordered proteins. Per-protein error rates are calculated as the number of proteins containing at least one predicted disordered region of the indicated length or longer, divided by the number of proteins in the set of ordered proteins. These calculations were made using only residues with defined coordinates from the "ordered proteins" set.

Minimum Number of Consecutive Predictions of Disorder	**Per-Residue Error Rate**	**Per-Protein Error Rate**
1	20%	98%
10	15%	82%
20	8%	46%
30	4%	29%
40	2%	8%
50	0.5%	2%
60	0.1%	0.5%

Target prioritization using percent predicted protein disorder

Shown in Figure 12A are the cumulative histograms for the ordered, disordered, and X-ray with disorder datasets as a function of the percent of total sequence predicted to be disordered. This measure gives good separation between the sets of structured proteins and disordered proteins and could serve to indicate proteins that are mostly disordered. The weakness of this approach is that, while proteins that are highly disordered can be filtered, mostly ordered sequences containing enough disorder to hinder high throughput protocols may be overlooked by this measure. However, if longer proteins receive a lower priority, a simple percent of residues predicted to be disordered cutoff may be useful without additional criteria.

Figure 12. Cumulative histograms of predicted disorder in the ordered, disordered, and X-ray with disorder datasets. Disorder is summarized for each protein in the set by three different methods: (A) the percentage of residues per-protein predicted to be disordered, (B) the longest consecutive disorder prediction, and (C) the number of residues within predicted regions of disorder of 20 residues or longer. The dotted and dashed lines in each plot represent aggressive and conservative cutoffs, respectively, and are discussed in the text.



The percent of proteins with proportions of predicted disorder less than that indicated on the *X*-axis rises much more sharply for the ordered proteins and the X-ray proteins with disorder sets than the disordered protein set (see Figure 12A) . Although such a relationship is expected for an order-disorder predictor of reasonable accuracy, the position of the X-ray disorder set is particularly important. A large fraction of proteins in many genomes likely contain relatively long regions of disorder (10, 49, 264, 265), and such proteins should be removed from the structure determination target list. However, the presence of relatively short regions of disorder does not imply that a protein would not be amenable to structure determination. Since a high proportion of X-ray disorder proteins have relatively low predicted disorder content, a prioritization scheme using the percent of residues predicted to be disordered would be useful.

Two example cut-off values are indicated in the cumulative histograms of Figure 12A. The dotted lines in the figure illustrate an aggressive filtering and weighting scheme using 30% or less of residues predicted to be disordered as the cutoff. Based on these example datasets, this would eliminate 90% of the disordered proteins while retaining 75% of the ordered proteins and 55% of the crystallizable proteins having disordered regions. A more conservative cutoff is illustrated with the dashed lines in Figure 12A. Using this threshold of 60% or less of each sequence predicted to be disordered, nearly all of the ordered proteins and the X-ray disordered proteins (98% and 96%, respectively) can be retained while eliminating over half (56%) of the disordered proteins.

Target prioritization using total predicted protein disorder

For larger proteins, a small percentage of predicted disorder could correspond to a large amount of total disorder, which could inhibit the structure determination process. Two alternative measures of total predicted disorder are suggested as possibilities to overcome this problem (see Figure 12B and C). Proteins removed from a target list by these two measures could be separated into their ordered regions and their disordered regions for further study.

The two suggested measures of total predicted disorder are analyzed in Figure 12B and C. The first measure, Figure 12B, considers the longest disorder prediction in each protein, where the length is given by the *x*-axis value. The second measure, Figure 12C, presents the number of residues in predicted regions of disorder of 20 residues or longer in each protein, where the number of residues is given by the *x*-axis value.

The longest predicted disordered region could be used for prioritization by eliminating proteins with predicted disordered regions of 40 residues or longer. Such an aggressive scheme would retain 91% of ordered proteins and 68% of X-ray disorder proteins as targets, and eliminate 58% of disordered proteins from consideration, as indicated by the dotted line in Figure 12B. A longest predicted disordered region cutoff of 70 residues, represented by the dashed line in the same figure, represents a more conservative filtering scheme. This would retain 99% of ordered proteins and 89% of X-ray disordered proteins but eliminate only 29% of completely disordered proteins.

Alternatively, the total number of residues in predicted disordered regions of 20 or longer could be used as a prioritization attribute, with a cutoff of 40 residues as indicated by the dashed line in Figure 12C. Approximately 62% of disordered proteins could be filtered while retaining 78% of ordered proteins and 59% of X-ray disorder proteins. With a conservative threshold of 80 residues, 94% of ordered proteins and 76% of X-ray disorder proteins are retained while 44% of the disordered proteins would be eliminated.

Retrospective evaluation of predicted disorder based filtering

The effectiveness of the proposed methods of disorder-based filtering for crystallization targets was examined through a retrospective evaluation. For this, disorder predictions were made and evaluated for all proteins in the TargetTrack database. Since the use of disorder-based filtering in selected proteins would confound retrospective evaluation, we examined the TargetTrack database for biases in the content of predicted disorder. Based on the idea that targets were not filtered for disorder in the initial years of the PSI, targets were grouped by the

year they were first selected and the values of the filtering criteria were compared. As can be seen by the plot of mean percent predicted disorder by year (Figure 13), targets selected in early years of the project have significantly more disorder than those selected in the latter years of the project. Note that the high proportion of predicted disordered in targets selected in 2004 and prior relative to those selected in 2007 and later is not reflected in analogous plots for total disorder (Figure 14 and Figure 15). The latter plots show no pattern in the bias of disorder content based on the selection year of targets. These results suggest that disorder prediction has been used for target filtering by contributing centers using a criterion similar to percent predicted disorder. Based on this result, targets selected in 2004 or before were taken as a set of unfiltered structure targets, and targets selected in 2007 and before were taken as a set of filtered targets.

Figure 13. Predicted disorder in the TargetTrack database by year. Means of the percentage of predicted disorder for targets in the TargetTrack database. Targets are grouped by the year targets were first selected. Error bars give the 95% confidence interval of the mean.

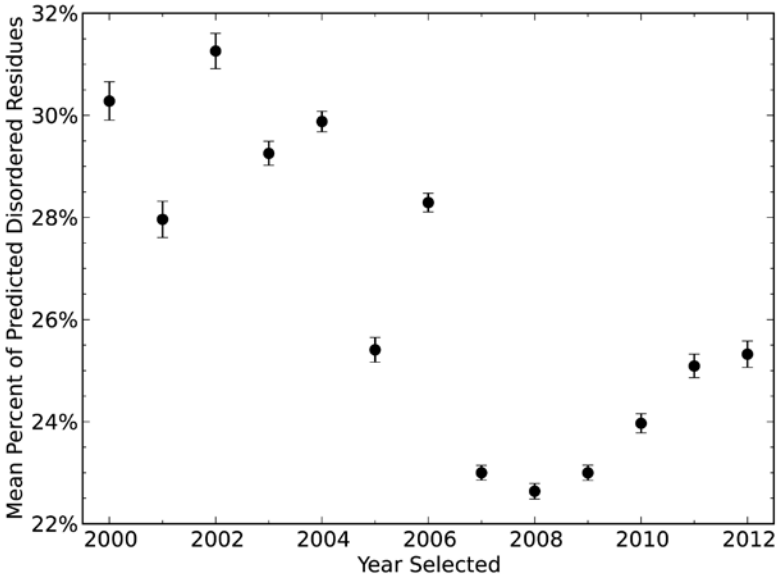


Figure 14. Means of the longest disordered region of targets in the TargetTrack database. Targets are grouped by the year targets were first selected. Error bars give the 95% confidence interval of the mean.

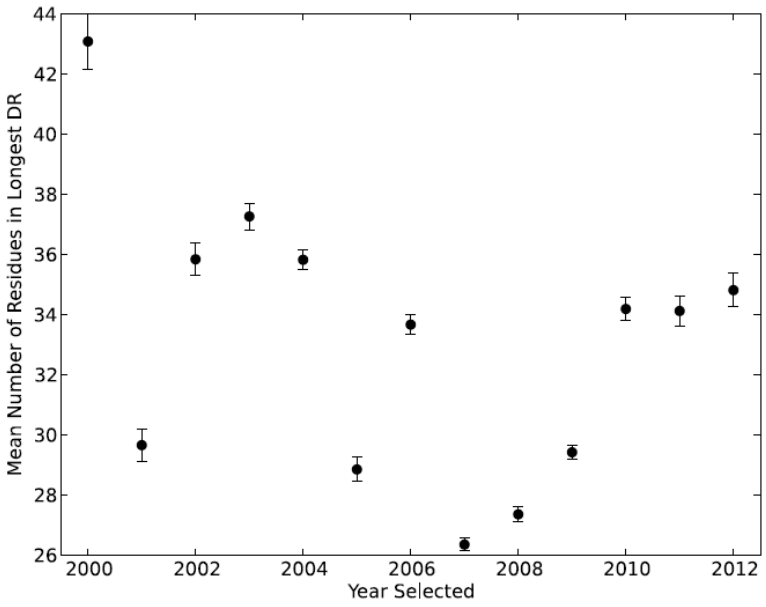
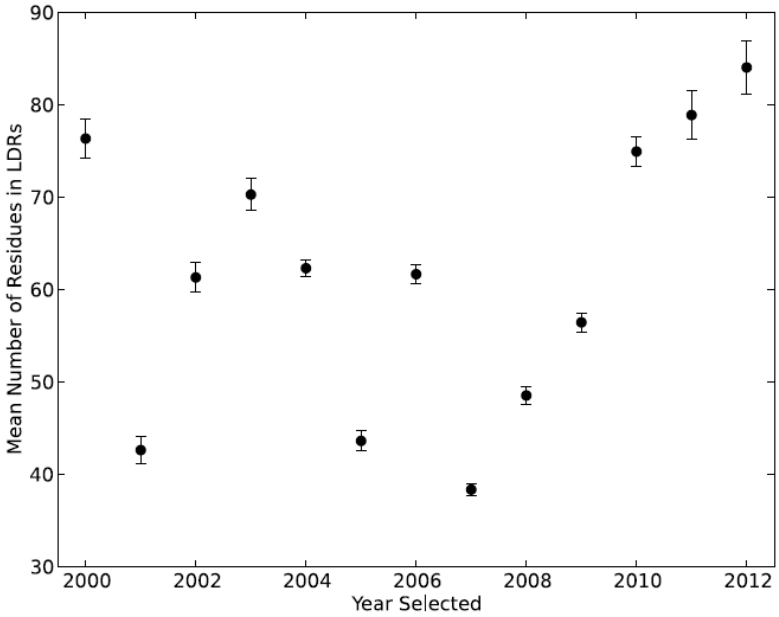


Figure 15. Means of the number of residues within predicted regions of disorder of 20 residues or longer of targets in the TargetTrack database. Targets are grouped by the year targets were first selected. Error bars give the 95% confidence interval of the mean.



Disorder prediction based filtering was evaluated by CDF plots for unfiltered targets using the proportion predicted disorder (Figure 16A), number of residues in disordered regions of 20 residues or longer (Figure 16B), and longest disordered region (Figure 17) criteria. These plots have one curve for each target status in the crystal structure pipeline, which progresses in the following order: selection, cloning, expression, purification, crystallization, diffraction, and structure determination. Each criterion shows a progressive loss of disordered proteins in subsequent steps of the structure determination pipeline. Also, the curves for targets with determined structures (black line in Figure 16A, B, and Figure 17) closely follow the curves for ordered proteins (empty line in Figure 12A, B, and C), which indicates that ordered proteins are an appropriate model for successful structure targets. Comparison of the curves for selected targets with curves for targets with determined structures suggests that disorder based filtering provides an effective method for increasing pipeline efficiency; depending on criteria and threshold selection, 5% to 20% of selected proteins could be filtered at the suggested thresholds while losing relatively few successful structure targets.

Figure 16. Cumulative histograms for targets in the TargetTrack database for (A and C) percentage of predicted disorder and (B and D) number of residues within predicted regions of disorder of 20 residues or longer. Plots for two selection date thresholds are shown: (A and B) 2004 or earlier and (C and D) 2007 or later.

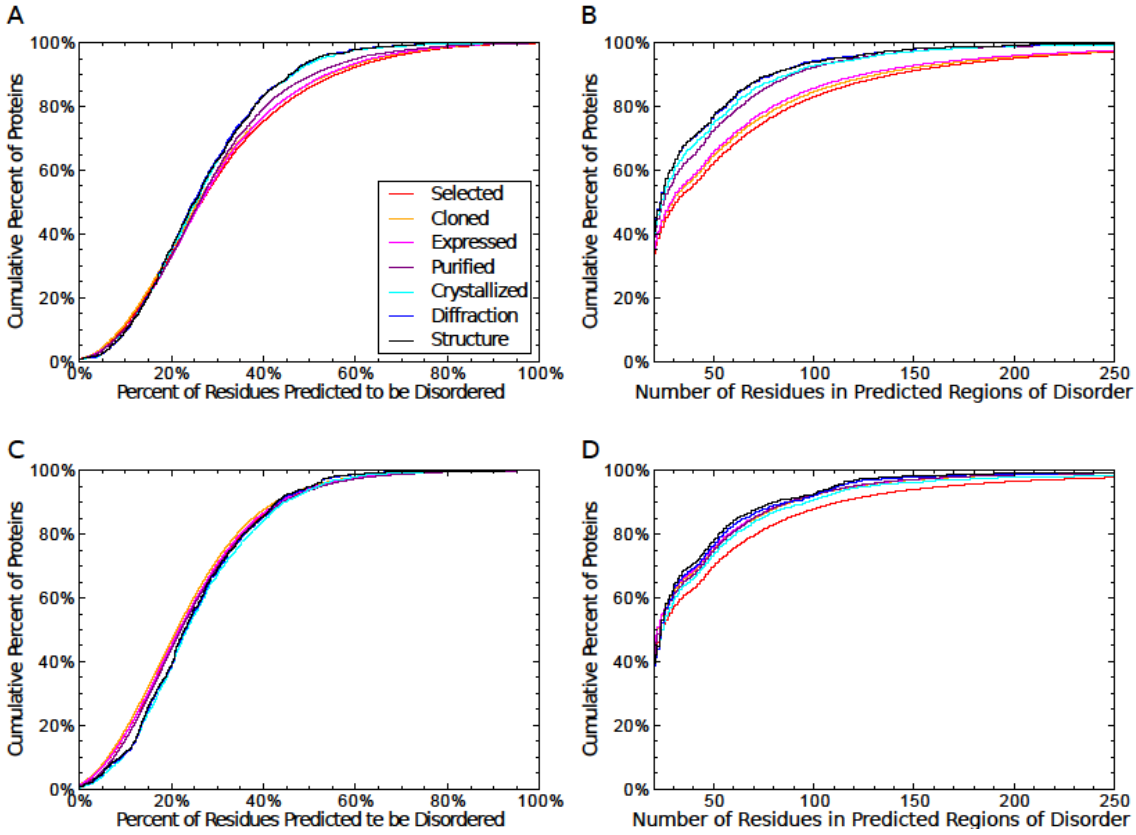
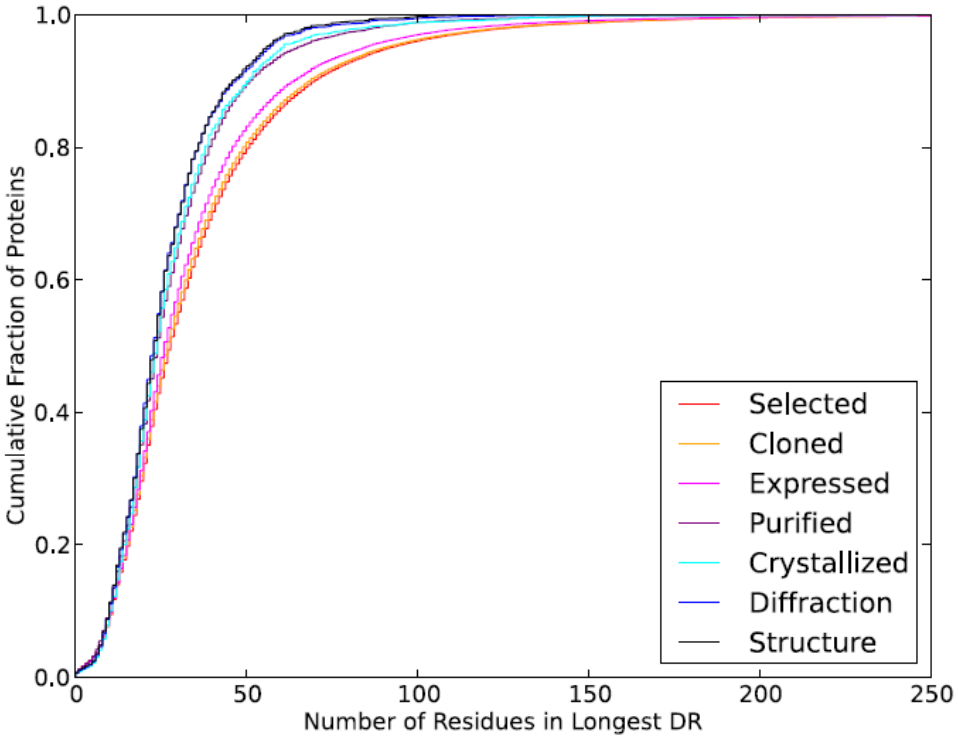


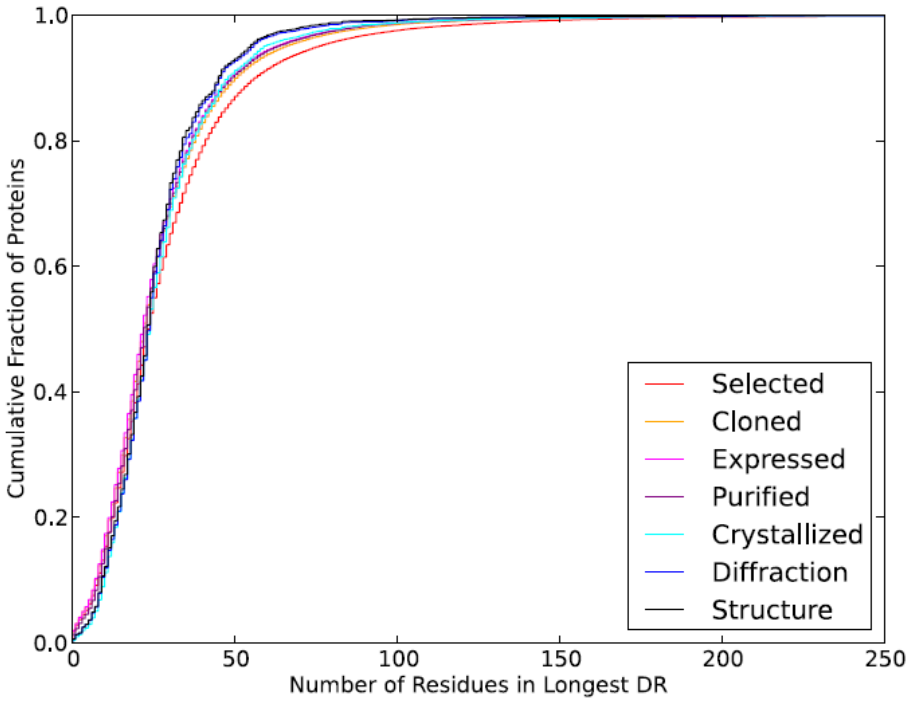
Figure 17. Cumulative histogram for the longest consecutive disorder prediction for targets in the TargetTrack database with selection dates in 2004 or earlier.



The impact of disorder based filtering was also investigated by examining filtering criteria for the set of filtered targets, i.e. those targets selected in 2007 and later. As expected, the plots of proportion of predicted disorder (Figure 16C) show little to distinguish selected targets from successful structure targets. Therefore, little efficiency could be gained from further filters based on this criterion. However, both total disorder criteria, number of residues in LDRs (Figure 16D) and longest disordered region (Figure 18), show a large difference in predicted disorder between selected and successful structure targets, although this difference is somewhat reduced compared to unfiltered targets (Figure 16 and Figure 18, respectfully). This suggests that not only is a low proportion of intrinsic disorder important for successful structure determination, but also that there is a limit to the absolute number of disordered residues a well-diffracting crystal can tolerate. Therefore, further gains in pipeline efficiency could possibly be gained by the use of multiple criteria.

Combinations of disorder criteria were examined by fitting logistic regression models and evaluating these models with the AUC performance measure. Single parameter models were also evaluated for comparison. These models were fit on proteins selected in 2004 and prior from the TargetTrack database, using proteins yielding crystal structures as one class and all other proteins as the other class. The goal in fitting these models was not to develop binary predictors of whether or not a protein will yield a crystal structure, but rather to determine an appropriate linear combination of parameters. AUC is used as a measure of separation between the two classes, which is appropriate since AUC is independent of threshold and the logistic function transform.

Figure 18. Cumulative histogram for the longest consecutive disorder prediction for targets in the TargetTrack database with selection dates in 2007 or later.



Comparison of AUCs for combinations of parameters (Table 3) shows that addition of parameters improves separation performance. Combination of all three disorder prediction parameters provides the best performance. However, all three parameters is only marginally better than the next best combination, fraction of disordered residues (FDR) and number of residues in long disordered regions (NDR), and the latter is examined further. The cumulative histogram of the linear combination of FDR and NDR is shown in Figure 19. Note that the linear combination is 0.0092 NDR–0.72 FDR, where the signs indicate that NDR has a negative impact on crystallization but that FDR has a positive impact on crystallization. The apparent positive influence of FDR on pipeline progress, which contradicts the current working hypothesis, was investigated further by fitting FDR-NDR logistic regression models for each combination of subsequent target status tags (Table 4). The positive relationship between FDR and structure pipeline progression persists except for progression of purified targets, for which both FDR and NDR are detrimental.

Table 3. Logistic regression classification performance for combinations of disorder prediction-based parameters, longest predicted disordered region (LDR), number of residues in predicted disordered regions longer than 20 residues, and fraction of residues predicted to be disordered, measured by the area under the curve (AUC).

Disorder Parameter			
LDR	**NDR**	**FDR**	**AUC**
+	-	-	0.578 ± 0.020
-	+	-	0.584 ± 0.013
-	-	+	0.533 ± 0.023
+	+	-	0.583 ± 0.020
+	-	+	0.589 ± 0.017
-	+	+	0.595 ± 0.020
+	+	+	0.600 ± 0.020

Figure 19. Cumulative histogram for targets in the TargetTrack database selected in 2004 or earlier by the linear combination of percentage of predicted disorder and number of residues within predicted regions of disorder of 20 residues or longer (equation given in text).

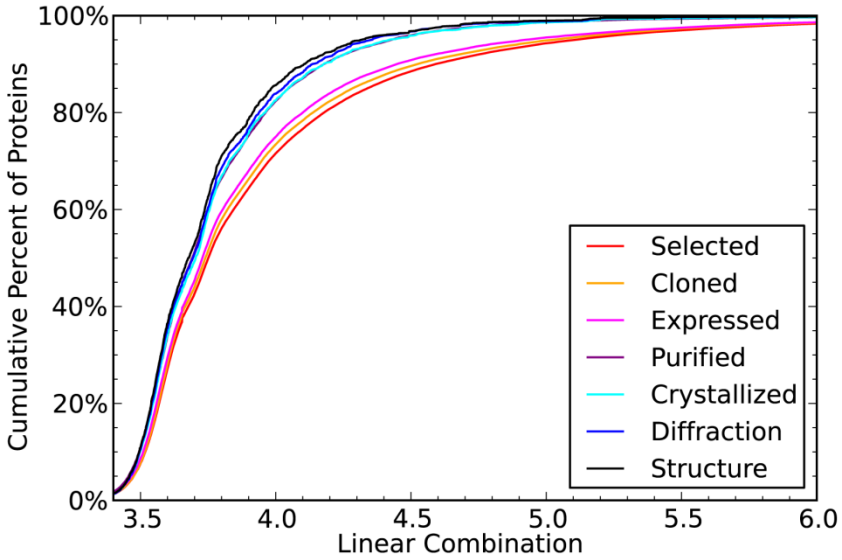


Table 4. Performance and parameter ranges of logistic regression models from cross validation for targets selected in 2004 or earlier. Positive (negative) parameter values indicate that greater (lesser) values inhibit (enhance) progression from the start status to the end status.

Status Tags			**Parameter Range**	
Start	**End**	**AUC**	**FDR**	**NDR**
selected	cloned	0.547	[0.595, 0.684]	[0.00060, 0.00082]
selected	expressed	0.543	[-0.245, -0.171]	[0.00182, 0.00202]
selected	purified	0.597	[-1.180, -1.114]	[0.00834, 0.00900]
selected	crystallized	0.573	[-0.319, -0.248]	[0.00592, 0.00638]
selected	diffraction	0.583	[-0.513, -0.328]	[0.00695, 0.00828]
selected	crystal structure	0.594	[-0.887, -0.627]	[0.00843, 0.00954]
cloned	expressed	0.551	[-0.769, -0.622]	[0.00167, 0.00203]
cloned	purified	0.599	[-1.508, -1.434]	[0.00848, 0.00878]
cloned	crystallized	0.562	[-0.566, -0.312]	[0.00514, 0.00648]
cloned	diffraction	0.573	[-0.618, -0.493]	[0.00643, 0.00739]
cloned	crystal structure	0.588	[-0.995, -0.766]	[0.00826, 0.00915]
expressed	purified	0.595	[-1.231, -1.108]	[0.00809, 0.00860]
expressed	crystallized	0.544	[-0.073, 0.219]	[0.00403, 0.00459]
expressed	diffraction	0.556	[-0.194, 0.030]	[0.00520, 0.00603]
expressed	crystal structure	0.572	[-0.488, -0.291]	[0.00642, 0.00752]
purified	crystallized	0.519	[0.673, 1.073]	[-0.00174, -0.00078]
purified	diffraction	0.521	[0.663, 0.907]	[-0.00029, 0.00022]
purified	crystal structure	0.520	[0.344, 0.502]	[0.00112, 0.00210]
crystallized	diffraction	0.531	[-0.333, 0.049]	[0.00276, 0.00408]
crystallized	crystal structure	0.565	[-1.113, -0.775]	[0.00495, 0.00656]
diffraction	crystal structure	0.559	[-1.949, -0.741]	[0.00538, 0.00642]

The relationship between predicted disorder and structure pipeline progression suggested by these results is 3-fold: (1) NDR generally inhibits pipeline progression, while (2) FDR (in conjunction with NDR) benefits pipeline progression from expression through purification but (3) FDR (in conjunction with NDR) inhibits pipeline progression from purification through to crystallization. Relationship (2) is somewhat contrary to the current working hypothesis, but can be reconciled in the following way. Successful protein purification is dependent on protein solubility, so the current observation is consistent with recent results that show that intrinsic disorder can greatly increase protein solubility (309). Relationships (1) and (3) agree with expectation since disorder has been observed to complicate each step in the structure determination process, from expression to solubility to crystallization.

Expression levels can suffer from the reduced stability of intrinsically disordered proteins, due to their inherent proteolytic sensitivity; intrinsically disordered proteins are digested much faster than ordered proteins (24, 310–313). Solubility of some proteins with disordered regions can be very low under conditions promoting the native state. For example, the aggregation of wild type Bcl-2 was remedied by replacing its disordered region of Bcl-2 with a portion of the disordered region of Bcl-xL, reducing aggregation and allowing structure determination of the chimera (274). Crystallization is unlikely for completely disordered proteins or proteins with too much intrinsic disorder because of their flexible and very dynamic nature. Myelin basic protein (MBP) exemplifies these troublemakers (314). One exhaustive series of attempts to crystallize MBP for X-ray diffraction has been reported, where the authors tried 4,600 different crystallization conditions but were unable to induce crystallization of MBP (275). Based on these observations the myelin basic protein has been suggested to belong to the category on “uncrystallizable” proteins. It can be safely assumed that many other unsuccessful crystallization attempts for numerous other proteins have not been reported, since negative results are generally assumed to be unsuitable for publication. In the case of MBP, several additional studies suggest that this protein lacks fixed 3D structure, existing instead as in intrinsically

disordered ensemble, which in turn have been suggested to provide the basis for its multifunctionality (315).

Target improvement with disorder prediction

Many protein structures in PDB are fragments of full-length gene products. These structures often represent independently folding domains and are invaluable for understanding the biological functions of these proteins. Ordered regions need to be identified and isolated, preferably without the use of extensive (and costly) experimentation. This section illustrates some ways in which expert analysis of PONDR predictions can provide a starting point for the isolation of ordered regions and disordered regions from their larger host proteins.

Disordered termini detection

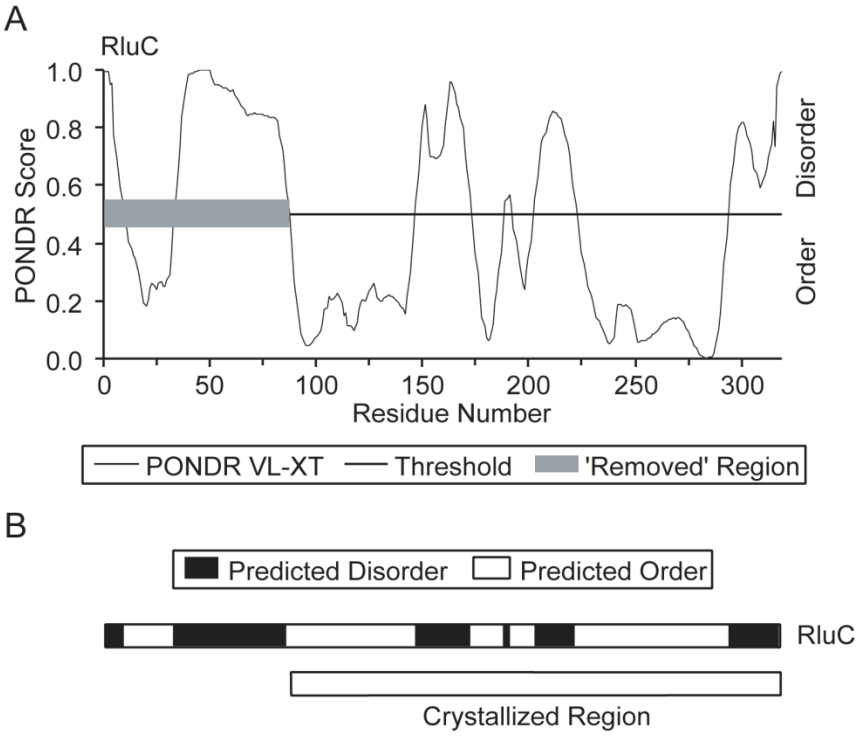
Many proteins contain long regions of disorder at either the amino or carboxyl terminus or at both termini. Removing these terminal disordered regions can often improve protein handling and crystallizability. Crystallization trials for three proteins, NEIL1 (201), RluC, and β-recombinase, exemplify such a target improvement scheme. The original crystallization attempts for these proteins began with the full-length sequences.

NEIL1, a human homolog of *Escherichia coli* DNA glycosylase endonuclease VIII, represents an illustration of the non-trivial correlation between disorder, solubility, and crystallizability. Crystallization trials for full-length C-terminally His-tagged NEIL1 failed to yield any crystals. This inability to grow crystals was corroborated by the fact that the protein was polydisperse based on the dynamic light-scattering analysis, and this polydisperse feature persisted regardless of the temperature or buffer conditions used (201). To overcome this obstacle, PONDR VL-XT was used to detect disordered region(s) in NEIL1 that may hinder crystallization. This analysis predicted that the protein has a disordered C-terminal region (201). Based on this observation, a construct missing the entire disordered C-terminal region (last 106 amino acids) in NEIL1 was designed. Little or no soluble protein expression was observed with this construct. A series of shorter C-terminal deletion constructs were cloned and checked for

expression. These studies showed that deletions of > 100 amino acids did not yield any protein expression. Therefore, a NEIL1 construct missing the C-terminal 100 amino acids (NEIL1CΔ100) was tested in crystallization studies. This active construct was successfully overexpressed in *E. coli*, purified and crystallized (201). Thus, this example clearly illustrates the potential utility of disorder prediction to increase chances for successful protein crystallization.

Crystals were obtained for RluC after a 6-week crystallization period and the protein in the crystal was found to lack 88 residues from the amino terminus. Further analysis showed that trace protease activity was likely responsible for the hydrolysis of the backbone. Further crystallization trials using truncated RluC succeeded in 2–3 days (316). PONDR predicts the removed region to be highly disordered, with a disorder–order boundary within 2 residues of the observed cleavage site. Figure 20 shows this truncation from two perspectives: Figure 20A is a plot of the PONDR prediction with the removed region indicated and Figure 20B gives the order–disorder prediction in the form of a bar with the crystallized region as indicated. This bar-type representation is indicated here for later use.

Figure 20. PONDR VL-XT predictions on the protein RluC. PONDR VL-XT prediction output values plotted versus residue number, where the order-disorder threshold is indicated by the thin horizontal line at a score of 0.5 (A). The region of this protein that was removed by proteolysis during crystallization trials is indicated by the thick gray bar. In an alternative representation, PONDR VL-XT output scores are indicated in the upper bar of (B) in binary form where black bars (score ≥ 0.5) indicate disorder and white bars (score < 0.5) indicate order, with the sequence location indicated by the distance from the left end of the bar. The region of the protein that was crystallized is shown by the lower bar of (B), coded white to indicate the region is ordered.



Crystallization of full-length β-recombinase required 6 months to produce small crystals. Analysis of the protein in the crystal revealed that 82 residues at the carboxyl terminal had been cleaved (317). The removed region is predicted to be highly disordered by PONDR, with an order-disorder boundary within 10 residues of the cleavage site (data not shown).

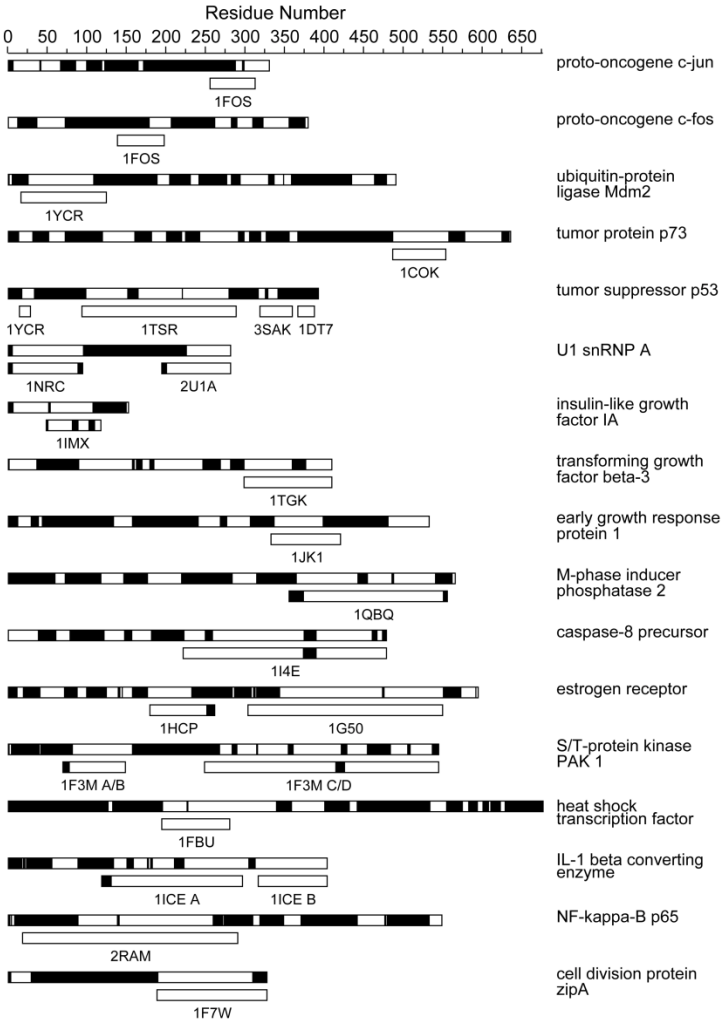
The removed regions of RluC and β-recombinase have not been convincingly characterized as completely disordered. However, the extreme protease sensitivity of disordered regions suggests that these two proteins contain disorder, at least in the regions surrounding their cleavage sites. Both of the cleaved loci are predicted to be disordered by PONDR VL-XT. Although neither of the termini of either protein is predicted to be completely disordered, removal of the bulk of the predicted disorder led to crystallization of both proteins.

In practice, the fraction of residues predicted to be disordered at each terminus should be considered, not just long predictions of disordered at the termini. If the terminal disorder represents a large portion of the predicted disorder in a protein, it is likely that removing such a region will improve a protein's crystallizability.

Detection of ordered fragments

Disorder prediction can be useful to identify the specific regions of proteins that may form rigid structures. Figure 21 provides examples showing the relationship between PONDR predictions (upper bar) and the locations of structured fragments (lower bar) for a representative set of proteins.

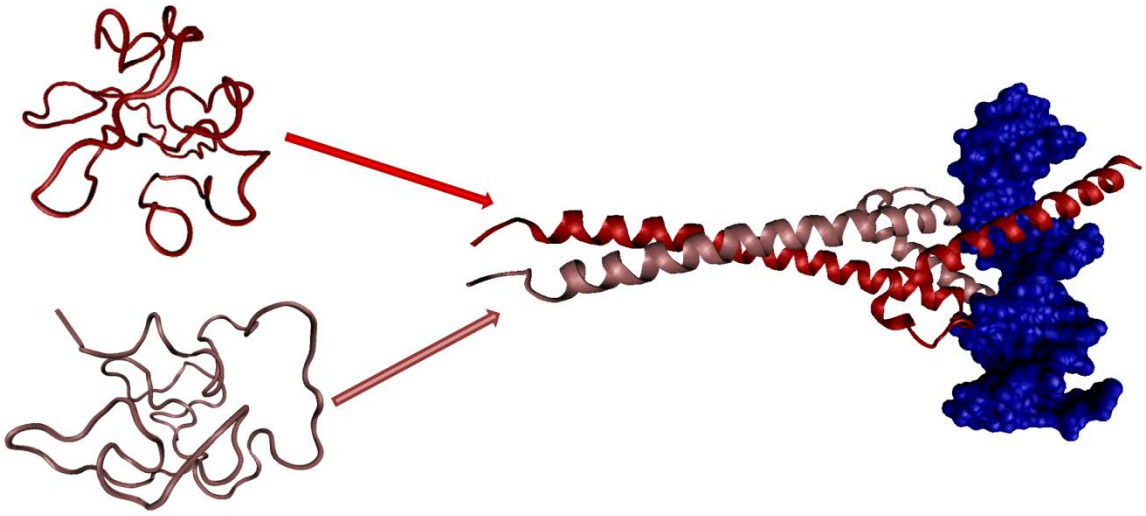
Figure 21. Comparison of PONDR predictions and protein fragments having 3-D structure. Proteins were chosen as representative of accurate and inaccurate PONDR VL-XT predictions. The PONDR VL-XT prediction for each protein is represented by the upper bar, where black disorder and white indicates order. The lower bars represent determined structures with PDB accessions as labeled, where white indicates defined coordinates and black represents missing coordinates or divergent NMR structures. Some examples are for proteins in isolation (1COK, 1TSR, 1NRC, 2U1A, 1IMX, 1HCP, 1F3M A/B, 1F3M C/D, 1F7W), other structures are for long protein fragments in complexes with other proteins (mdm2 in 1YCR, 1I4E, 1QBQ) or nucleic acid (1JK1, 2RAM), some are for homooligomers (1TGK, 1G50, 1FBU, 1ICE), and some are for short protein fragments in complexes with other proteins (1FOS, p53 in 1YCR, 3SAK, 1DT7).



The predictions are obviously not perfect. Short regions of predicted disorder are contained within ordered regions, and the fragment boundaries are often over- or under-predicted. Such inaccuracies can be partially attributed to disordered regions that are misclassified as ordered regions, as exemplified the c-fos and c-jun entries.

Many ordered protein regions in PDB are involved in protein-protein or protein-nucleic acid complexes, yet these same regions are disordered in the absence of their partners. By our definition, such regions should be classified as disordered, but such regions are typically misclassified as ordered in current data sets. This adds noise to the data, which contributes to the observed error rates. For example, DNA binding regions of proteins are often predicted to be disordered (264). For some of these proteins experimental evidence has shown that such regions are disordered in the absence of DNA. An example of this type is the c-fos/c-max leucine zipper DNA binding region. Leucine zippers have been shown to be disordered in isolation from their zipper partner (318); therefore these regions of c-fos and c-jun are classified as disordered. The 1FOS crystal structure contains the leucine zipper dimerization regions of c-fos and c-jun, which are complexed to DNA. The portions of these helices that make contact with the DNA are isolated from any other protein contacts, which suggest that these regions would be highly unstable in the absence of DNA, and indeed, these regions were predicted to be disordered. Similar behavior was also found for the c-Myc/Max leucine zipper DNA binding region. Furthermore, although c-Myc and Max were characterized by the well-defined, ordered structure in their tertiary complex with DNA (1NKP), the detailed structural characterization of unbound Max and c-Myc revealed that both proteins were highly disordered in their free forms and underwent mutual coupled binding and folding when their zipper domains interacted to form a helical coiled coil (Figure 22) (93, 319). These examples show that comprehensive study of disorder predictions for regions involved in protein-ligand interactions may provide useful information for future attempts at identifying partners of intrinsically disordered protein regions.

Figure 22. Intrinsic disorder-based interaction between c-Myc, Max and DNA. Both c-Myc and Max are intrinsically disordered in their unbound states (left side), but fold into coiled-coil dimer upon interaction with DNA (right side) (PDB ID: 1NKP). c-Myc, Max, and DNA are shown as red ribbon, pink ribbon, and blue surface, respectively).



These results point to the importance of specific investigations of order–disorder boundaries to determine whether their prediction can be improved. However, even without further refinement, disorder prediction can at least provide a useful starting point for structure studies by indicating possible locations of the ordered regions of proteins.

Inclusion of binding partners

While this paper has focused on the inhibition of protein structure determination by excessive amounts of intrinsic disorder, there are counter examples that do not fit this general trend, including structured proteins that do not form suitable crystals and disorder-containing proteins that do form suitable crystals. Both of these features are discussed below in terms of a specific example.

When a protein fails to crystallize or gives poor quality crystals, a common approach is to attempt to crystallize a homologue, often leading to a successful end result. An interesting example is provided by attempts to determine the structure of protein phosphatase 1 (PP1) complexed with inhibitor 2 (I2). When rabbit muscle PP1 was complexed with the rabbit I2 protein, crystals that diffracted to ~ 4 Å were obtained (Tom Hurley, personal communication). However, when experiments were carried out using rat PP1 complexed with mouse I2, crystals were also obtained, but these diffracted to 2.5 Å, which was sufficient for 3D structure determination (320).

The PP1:I2 complex also illustrates that proteins with substantial amounts of disorder can sometimes form crystals of sufficient quality for structure determination. When not bound to PPI, the I2 protein is unstructured (320, 321). Upon complex formation with PP1, three separate segments of I2 become structured, giving a total of 59 structured residues but with 154 residues that remain unobserved in the complex. An NMR study of PP1 complexed with a large I2 fragment, from residue 9 to 169, shows that 100 residues corresponding to two unobserved segments in the crystal indeed remain as a random coil in the complex. Furthermore, large holes of sufficient size to accommodate models constructed for the IDP ensemble are observed for the

PP1:I2 crystal lattice (322). This crystal structure is similar to those presented in Table 1, where a large disordered region can be accommodated by the crystal lattice if a sufficient number of ordered residues are present to support the lattice.

Integration of disorder prediction with high throughput structure determination

One use of PONDR in structural genomics would be to divide the target list into two separate pipelines: a structure pipeline, for whole proteins or protein domains for which standard structure determination would be appropriate (i.e., for highly ordered protein); and a disorder pipeline, for proteins or protein domains for which other methods of study would be appropriate (i.e., for highly disordered proteins). Although a structure pipeline represents the current Structure Genomics methodologies, a disorder pipeline has not been implemented as of yet.

The underlying rational for the PSI is that protein function can be inferred through knowledge of protein structure (259, 260). Disordered proteins, however, are not encompassed by this view of protein function, including traditional ideas regarding protein flexibility, e.g. allostery and induced fit. While much information can be gathered from disordered proteins through biophysical and biochemical methods, new models are required to translate this information into functional mechanisms. The first step in the development of these new models is the accumulation of biophysical parameters from a broad range of disordered proteins. Through a centralized dataset of such information (e.g., DisProt database (16) or similar resources), generalities may emerge that will enable functional inferences, as is now possible with ordered proteins. The lack of this accumulated information on disordered protein is a major block to understanding the mechanisms of disordered protein function.

Disordered proteins can be analyzed by traditional biophysical methods including circular dichroism, infrared spectroscopy, fluorescence, various hydrodynamic studies, small-angle X-ray or neutron scattering, etc. Studies using these methods provide a picture of the average conformational range of proteins (i.e., such methods when used in combination differentiate extended and collapsed disorder) (30, 31, 34, 313, 323). Mapping protein disorder by the means

of limited proteolysis resolved through both SDS–PAGE and mass spectroscopy, which were successfully used to map disordered regions in XPA (13) and clusterin (12), would also be a generally applicable methodology. These mapping studies localize disordered regions in terms of protein primary structure and can indicate the presence of structured domains, as in XPA (13), or indicate the presence of regions of secondary structure, as in clusterin (12). A more complete picture of the conformational dynamics of disordered proteins could be obtained through the use of small molecule probes and cross-linkers that would indicate the time averaged spatial relationships of individual residues, although these techniques have yet to be explored in this context.

A survey of disorder–function relationships for more than 90 proteins with experimentally determined regions of disorder of length $\geq$ 30 residues ascribed more than 25 separate functions to the regions of disorder, with these functions falling into four broad categories: (i) molecular recognition, (ii) molecular assembly, (iii) protein modification, and (iv) entropic chains (7). The molecular recognition category is especially interesting. Molecular recognition by ordered proteins typically involves catalysis that can be described by the lock-and-key paradigm, whereas molecular recognition by disordered proteins typically involves signaling with a DOT playing a key role. Given that disordered regions have been implicated as the primary interaction regions in signaling events (9, 44, 47, 94, 264, 324, 325), these regions of proteins are prime candidates for use in proteomic binding assays. Using protein chip technology (326), display of intrinsically disordered protein segments for screening against a host of gene products could prove to be advantageous in elucidating the web that underlies cell signaling. Thus, armed with biophysical and functional evidence, large advances in the understanding of disordered protein function could be made.

Disorder prediction can be used to differentiate proteins that are likely to crystallize from proteins that are unlikely to crystallize or that are almost certain to not crystallize. Proteins predicted to contain both ordered and disordered regions can be studied in both pipelines, using

disorder prediction as a starting point for any modification necessary to facilitate analyses (e.g., removal of disorder from structured domains and isolation of disordered regions). Disorder prediction would also provide the basis for helping to identify structured and unstructured domains within multidomain proteins; such identified domains could then be recovered by genetic engineering and sent to their respective pipelines.

The structure pipeline and disorder pipeline would collect ordered structure information and disorder structural information in complementary databases that would allow for concurrent study of some proteins in both contexts for the greatest possible efficiency and expediency. Regardless of whether an approach such as this is used, disorder prediction could aid selection and improvement of protein targets for the structure-determination pipeline through the filtering of disordered protein targets and isolation of ordered domains.

Conclusion

The impact of intrinsic disorder on the structure determination pipeline was studied retrospectively by examining missing density in the crystal structures and target progression in structural genomics relative to predicted intrinsic disorder. Intrinsic disorder is tolerated to some extent in crystal structures, although extensive regions of intrinsic disorder are infrequent and are likely to require a sufficient relative proportion of ordered residues to support the crystal lattice. In fact, as shown in Table 1 and the associated references, large regions of apparently missing electron density in X-ray-determined protein structures in the PDB can arise from true intrinsic disorder, from mobile domains, from misannotation, or from unreported proteolysis, with intrinsic disorder accounting for about 30% of the mass of the protein in crystals containing the largest amounts of disorder. Given the high solvent content (~ 50%) of typical protein crystals (Tom Hurley, personal communication), the capacity of protein crystals to accommodate such large amounts of disorder is not surprising. This accommodation of disorder and the inaccuracy of disorder prediction (Table 2) combine to limit the ease of use of disorder prediction for filtering structural proteomic target lists.

Despite these limitations, the use of intrinsic disorder prediction for filtering targets is a viable method to improve the proportion of ordered targets in the structure determination pipeline. Retrospective application of these methods additionally demonstrates that the effect of intrinsic disorder on structure determination is not only detrimental; increased success in purification is correlated with a higher proportion of predicted disordered residues, which agrees with recent observations (309). For structure determination of individual proteins, prediction of intrinsic disorder can be an effective tool for tailoring proteins for structure determination. Though not developed to the extent to which automated analysis would be possible, consideration of prediction results on a protein-by-protein basis is certainly feasible. For the time being, these methods will likely prove useful in a supplemental capacity. That is, when difficult proteins are encountered, their PONDR predictions can be consulted to determine if they suggest an ordered region that might crystallize or a disordered terminus that could be easily removed. In this way, PONDR predictions could be used to salvage regions of proteins for the structure pipeline and provide input for disorder analysis. Finally, while this work focuses on the use of disorder prediction to facilitate structure determination, complementary approaches can be used to study those proteins and regions that fall outside the purview of direct structure determination methods.

Several avenues are available to extend the current work. One major challenge in studying intrinsic disorder from crystal structures is uncertainty in whether missing density is actually present in the crystal structure, or instead due to proteolysis or misannotation. But, it may be possible to determine the presence or absence of missing directly from crystal structure data. Specifically, protein crystal structures have a relatively narrow range of observed solvent contents. Back calculation of the solvent content from the contents of the unit cell with and without inclusion of missing density regions may show whether presence or absence of missing density regions is more consistent with the allowable range of solvent contents. While this method will not be sensitive enough for short missing density regions, it may be sensitive enough for large missing density regions. Another avenue of further research is in structure pipeline

progress prediction, specifically development of stage specific predictors. Model fitting results suggest that various stages of structure determination may have different, perhaps conflicting, attributes contributing to pipeline progression. Stage focused models with a broader range of attributes will likely show better detection of targets that will progress through the structure determination pipeline.

IV. MOLECULAR RECOGNITION FEATURES

The generality of IDP-mediated interactions is suggested by the large number of characterized DOT binding regions and the correlation between intrinsic disorder and regulatory functions. Toward the end of estimating the prevalence of DOT regions and providing tools for helping to identify and characterize novel examples, several groups have developed predictors of DOT regions and related regions (111, 196, 234, 327–329). Because these regions lack structure in the unbound state, all predictors developed to date are based on prediction from protein sequence. Also, all predictors focus on regions that mediate interactions with proteins, thus so far ignoring the likely possibility that particular IDP subregions have evolved to bind DNA or RNA.

Predictors of intrinsically disordered interaction regions

A variety of methods have been developed to predict intrinsically disordered binding regions from sequence. All of these methods have focused on regions shorter than typical folded domains. This emphasis on short regions is based on the particular type region being modeled and complications in defining the interface of long DOT regions. These predictors are based on two different views of these regions.

In one view, these regions are modeled as short (on the order of 20 residues), DOT regions occurring within longer regions of intrinsic disorder, called Molecular Recognition Features (MoRF) (99, 111). The inception of the MoRF model was the observation that some long regions of intrinsic disorder contained short predictions of order that correspond to regions responsible for molecular recognition (194). In other words, while these regions had been demonstrated to be disordered when isolated from their binding partners, the sequence itself contains an intrinsic propensity to form structure, a propensity that is realized when binding to its partner. Initial studies suggested that intrinsically disordered regions that formed α-helices in the bound state had particularly strong propensity to form structure, a property that was used to develop predictors of α-helix forming MoRFs (α-MoRFs) (111, 196). Subsequently, a

generalized algorithm was developed for prediction of MoRFs regardless of the conformation of the bound state, MoRFpred (112). Also, the independently developed ANCHOR predictor (327) is based on a definition of short, DOT transitions regions similar to the MoRF definition.

The other view of short intrinsically disordered binding regions is based on the idea of conserved, short (from 3 to no more than 15 residues), functional motifs, called Short Linear Motifs (SLiMs) (236). The SLiM model is not directly dependent on intrinsic disorder in the unbound state. However, instances of SLiMs have been shown to correspond closely to the MoRF model (237). Generally, SLiM regions occur within longer regions of intrinsic disorder, but SLiMs show a propensity to be structured. Regions flanking SLiMs have a propensity to be predicted to be disordered and flanking regions have compositions similar to intrinsically disordered proteins. SLiMs, relative to their flanking regions, show an increased propensity to be predicted to be ordered and have compositions closer to ordered proteins. Although SLiMs are in general correlated with regions of intrinsic disorder, it should be noted that some specific types and instances of SLiMs occur in an ordered context. Several prediction methods are based on the SLiM model, including the ELM Server (234) and the Minimotif Miner (330). Another prediction method, SLiMPred (328), while based on a SLiM dataset, takes a non-motif based approach to prediction of SLiMs.

Datasets

Ideally, development of a predictor of DOT binding regions would require only a positive dataset of known DOT binding regions and a negative dataset of other types of regions: folded domains and non-binding disordered regions. Since the absence of known binding cannot be extrapolated to mean that a region cannot bind to any partner, it is difficult to assemble a negative dataset of non-binding disordered regions. Different research groups have taken different approaches to addressing the problems associated with these types of regions.

α-MoRFs. According to the MoRF model, MoRFs are short DOT regions. Therefore, datasets for the various MoRF predictors were gathered from complexes in the PDB (278). For the α-MoRF I dataset (14 α-MoRFs), helical fragments bound to longer, globular proteins were selected and verified for literature evidence of intrinsic disorder in the unbound state (111). This approach was repeated for the α-MoRF II dataset, resulting in a lager set of MoRFs (196). Additionally, to further expand the dataset, homologous sequences to each MoRF-containing protein from UniProt were aligned to each MoRF region, and additional MoRF examples were selected at random from among those that align well to the MoRF region (54 MoRF + 48 homologous MoRFs). Control datasets were constructed from PDB monomers, i.e. sequences very unlikely to contain MoRF regions. Development of the α-MoRFs predictors did not address false prediction of MoRFs within intrinsically disordered regions. Again the problem is that if a given region predicted to be an α-MoRF is not known to bind to any partner, that does not mean that it does not bind to some yet to be discovered partner, so such false positives would be difficult to identify.

MoRFpred. A much larger dataset (840 MoRFs, 427 sequence unique clusters) was constructed by selecting bound fragments from PDB regardless of secondary structure and relaxing literature verification of intrinsic disorder in the unbound state (112). For verification, fragments were analyzed by examining the relative bound and exposed surface areas (100), where this analysis indicated that fragment structures were consistent with disorder in the unbound state. Similar to α-MoRF predictors, a control dataset of monomers from PDB was used, but additional control sequences were taken from regions flanking MoRF regions. The latter set of sequences is intrinsically disordered but assumed to be unlikely to contain MoRF regions. Although there are some known exceptions to this assumption, e.g. the N-terminal MDM2 binding region of p53 is flanked by regions that interact with several other proteins (60), these types of exceptions were assumed to be rare.

ANCHOR. Similar to MoRFs, the ANCHOR dataset is composed of short DOT binding regions (327). All regions were verified to have experimental evidence of disorder in the unbound state in the literature. The dataset constructed for ANCHOR optimization included an independent set of MoRF regions containing 46 examples, monomeric, ordered proteins, and bulk intrinsically disordered regions from the DisProt database (16). Additionally, two validation sets were used: long DOT binding regions containing 28 examples and ordered protein that participate in complexes with other ordered proteins with 72 complexes in all.

SLiMs. Unlike the MoRF predictors and ANCHOR, the SLiMPred, ELM Server, and Minimotif Miner datasets are based on functional motifs rather than structural or functional propensities. Consequently, the datasets used for these predictors do not rely on experimental structures. For SLiMPred (328) and ELM server (234), the ELM database (331) is the primary dataset, which is a manually annotated set of functional motifs. Minimotif Miner draws from several databases, as well as manually annotated motifs (330). The control set for SLiMPred consisted of domains (as determined by a SMART (332) search) and non-SLiM residues of SLiM containing proteins (300 sequence unique SLiM containing proteins). ELM server and Minimotif Miner do not use a control set per se, but rather use contextual filters to remove false positives after scanning for motifs.

The datasets used in training and evaluation of predictors of DOT binding regions vary widely in terms of definition and experimental verification. The α-MoRF predictors and ANCHOR used smaller, literature validated positive datasets, while the MoRFPred and SliMPred used much larger, but less well validated positive datasets. Negative datasets vary even more than positive datasets, particularly with respect to treatment of disordered proteins, which are avoided in α-MoRF predictors, used in bulk in ANCHOR, and filtered for binding regions in MoRFPred and SLiMPred. Ordered regions sets are similar across all predictors, consisting of monomeric, globular domains from PDB, except for MoRFPred, which additionally filtered out

structures with any missing residues that resulted in a significantly smaller ordered region dataset. Finally, ANCHOR also used additional evaluation sets: interfaces of structured proteins and long DOT binding regions.

Architecture

α-*MoRFs.* The α-MoRF I (111) and II (196) predictors use similar architectures. Predictors are stacked with two stages: first, identification of potential MoRF regions, and second, classification of potential regions as MoRFs or non-MoRFs. The first stage of the predictor identifies patterns in PONDR VLXT predictions that indicate the locations of regions of order propensity within longer regions of disorder. The second stage distinguishes between DOT binding regions and other causes of the disorder pattern, such as loops flanking a structured segment or false disorder predictions. This predictor uses several sequence properties, including disorder prediction, secondary structure prediction, and hydropathy, which are averaged over the potential MoRF region, flanking regions, or both. These attributes were used to train a quadratic discriminator and a neural network for α-MoRF I and II, respectively. In both cases, averaging sequence properties over regions relative to the potential MoRF reduces the feature space sufficiently (6 features in both cases) to compensate for the limited number of examples.

MoRFpred. Unlike the α-MoRF prediction methods, which predict MoRFs per-region, the MoRFpred algorithm predicts MoRFs per-residue (112). MoRFpred uses a support vector machine (SVM) with sequence and predicted features. In addition to features considered in α-MoRF predictors, the outputs of other prediction methods, such as B-factor and surface exposure prediction, and PSI-BLAST derived profiles were considered. Features were aggregated over a series of sliding windows relative to the residue to be predicted, similar to region averaging in α-MoRF predictors, and additionally features were calculated relative to a larger sequence window. Multiple sequence selection methods were used to select 24 features for the final SVM model.

SVM predictions were also augmented with alignments to MoRFs in the training set, which gave a small improvement to the true positive rate.

ANCHOR. The ANCHOR prediction method models DOT binding regions literally as disordered residues with a propensity to fold in a globular context (327). The prediction score is a linear combination of the IUPred (333) disorder prediction score, the self-interaction energy, and the interaction energy gained in a globular context. Energies are estimated as in IUPred, which are the averages of pairwise interaction energies between the residue of interest and a particular environment. Model parameters were fit on short DOT binding regions, using globular proteins as a control set. Concurrently, the rate of DOT binding residues in bulk intrinsic disorder was also coarsely minimized, which addresses the important objective of minimizing false prediction in intrinsically disordered regions in the absence of a disorder control set. The set of long DOT binding regions was not used in training because the others observed that binding regions seem to be limited to discrete regions. Rather than defining the discrete binding region within long DOT binding regions, they examined the behavior of the final predictor relative to the participation in the interface of each residue.

SLiMPred. MoRF predictors consider prediction context explicitly using region averaging. Prediction context in the SLiMPred algorithm (328) is established through use of a bidirectional recurrent neural network (BRNN). Connections in BRNNs link the internal state of an arbitrary number of previous and following residues to the prediction of the current residue. Inputs to the BRNN were protein sequence and predicted secondary structure, solvent accessibility, and structural motif (i.e. local conformation).

Motif. ELM Server (234) and Minimotif Miner (330) are based on matching motif patterns to regular expressions. SLiM motifs are short (15 residues or less) and consequently many matches are expected to occur at random. To reduce identification of false positive motif predictions or motif predictions that are irrelevant for a particular query, both methods apply

filters. Two general types of filters are used: query sequence-based filters and database-based filters. Query filters retain or remove predicted motifs based on predicted properties of a query sequence that favor or disfavor motif function, respectively. For example, a motif found to be buried within a globular domain is likely to be a false positive motif prediction. Other examples of query-based filters are domain prediction (234) and accessibility prediction (329). Database-based filters do not reduce false positives per se (330), but rather limit motif prediction to motifs that fit a certain set of criteria relevant to a given protein or query; these filters allow a subset of motifs to be selected based on information about the protein of interest. For example, motifs that have been observed in a given taxonomic group or observed to be involved in a given biological process or cellular compartment can be selected (334). Additionally, frequency statistics are used to rank matched motifs in reverse order of frequency of expected random matches (335).

The architectures used for DOT datasets vary widely and the complexity of the approach, in terms of the number of parameters, is correlated with data set size. The α-MoRF predictors have the smallest datasets and are relatively simple predictors, with a small number of parameters used in simple pattern recognition and a small statistical or neural network model. The ANCHOR predictor has a comparable size DOT dataset as compared to the MoRF predictors, but uses a large number of interaction potentials. However, potentials are derived primarily from the much larger ordered dataset, leaving only three free parameters trained on binding regions, making ANCHOR arguably the simplest DOT binding region predictor. MoRFPred and SLiMPred have relatively complex architectures with a large number of parameters, but the dataset used to train these parameters is much larger than the other predictors. In terms of parameters, motif-based predictors are among the most complex since they require a pattern for each type of motif, in addition to instance sequences.

Also, all prediction architectures consider the larger sequence context when predicting DOT binding regions, though each in a different way. The α-MoRF predictor and MoRFPred

consider context explicitly by considering features averaged over windows of the potential binding region and in the flanking sequence. ANCHOR uses an extended window to measure the disordered context and intra-sequence interaction propensity of a target residue. Context is an intrinsic feature of the bidirectional recurrent neural network (BRNN) architecture used by SLiMPred. Finally, the sequence filters used in motif-based prediction establish the, for example, order-disorder context of predicted motifs through domain prediction.

Evaluation

Prediction methods. The common evaluation for all prediction methods was the true positive rate for DOT binding regions and the false positive rate for monomeric globular proteins. However, this evaluation ignores some important cases, namely intrinsically disordered regions that do not contain binding sites and interfaces of ordered proteins. The former is a difficult problem common to many bioinformatics prediction tasks due to lack of annotation not being equivalent to a negative annotation, and is dealt with in different ways by MoRFPred and ANCHOR authors. ANCHOR authors also addressed the ordered interface control, and additionally they evaluated performance on long DOT binding regions.

The primary evaluation of prediction methods can be grouped into two types of approaches relative to the positive and negative datasets: (i) a small, high-confidence positive set of DOT binding regions and a negative set of ordered regions and (ii) a large positive set of DOT binding regions and a negative set of disordered regions. Approach (i) is in some senses an easier problem than approach (ii) for several reasons. The larger positive datasets are likely to be noisier than the smaller, literature verified datasets. Also, disordered negative datasets are noisier than ordered negative datasets; many unannotated binding regions may occur in disordered regions but not in ordered regions. Finally, several of the prediction methods make use of disorder prediction directly as a base attribute, meaning that the there is a greater contrast between positive and negative datasets when using a ordered negative dataset as opposed to a disordered negative dataset.

The α-MoRF I and II predictors and ANCHOR use approach (i) to determine accuracy and both demonstrate high accuracy on their respective datasets. MoRFPred and SLiMPred used approach (ii) to determine accuracy. For a negative set, MoRFPred used non-MoRF residues from MoRF containing proteins, which contain both ordered and disordered residues, and residues flanking MoRF regions, which are likely to be disordered. Similarly, SLiMPred used non-SLiM/non-domain residues for a negative set. The demonstrated accuracies of these algorithms on these sets are markedly lower than in the cases of α-MoRF I and II predictors and ANCHOR. This difference in evaluated accuracy is due to differences in the evaluation sets and does not reflect relative predictor accuracy.

Additionally, both MoRFPred and SLiMPred authors compared their respective performances to other MoRF predictors, which also showed the same reduced accuracy on these datasets relative to their own datasets. In the case of the SLiMPred dataset, ANCHOR showed a very similar accuracy to SLiMPred. In the case of the MoRFPred dataset, α-MoRF predictors had a very low false positive rate, but also a low true positive rate. The latter is not unexpected given the limited scope of these predictors. ANCHOR showed a good true positive rate on this dataset, but also a very large false positive rate.

Finally, ANCHOR was evaluated on additional datasets. The fraction of bulk disorder residues predicted to be binding regions was reported to be about 45%. Due to the ambiguity noted above, it is not known what the real target quantity should be, and it is constrained to be less than 50% during training. ANCHOR was also evaluated on long DOT binding regions. Rather than predicting these long regions to be a single binding region, ANCHOR predicts them to be composed of several binding regions. This segmentation of prediction suggests that long disordered binding regions are composed of several independent short binding regions. This idea is supported by data for the p27-Cyclin/CDK complex, which shows that the prediction score is correlated with the number of atomic contacts per residue (327).

Motif methods. It is generally believed that motif methods have a high false positive rate due to the low complexity of short motif patterns may frequently find matches at random. However, accuracy evaluations of the kind performed for prediction methods have not been reported for motif based methods, either with or without filtering. This is at least in part due to different nature of the two methods; supervised predictors require a positive examples and negative examples for training that can also be used for evaluation, but motif-based methods require only positive examples. Filter approaches partially mitigate the high false positive rate of motif methods. Sequence filters remove hits that might occur within a structured protein and motif filters remove hits that arise from irrelevant motifs, in terms of cellular or functional context.

An indirect evaluation of unfiltered motifs predictions confirms that the false positive rate is indeed quite high (336): (i) eukaryotic-specific domains are predicted at the same rate across eukaryotes, bacteria, and archaea, and (ii) motifs with narrow roles are predicted to occur frequently. Motif prediction filter methods adequately address the source of error (i), since source organisms are generally well annotated. Performance of filter methods for the source of error (ii) depends greatly on the completeness of the relevant functional annotations; filters will function properly for well annotated proteins, but will return false positives and negatives for missing and incorrect annotations, respectively.

Comparison

The SLiMPred authors compared their algorithm to ANCHOR and showed that the two algorithms have similar prediction accuracy – with nearly identical ROC curves – to ANCHOR across all examined datasets including long disordered regions (328) with two exceptions. When applied to secondary structure sub-sets, the two prediction methods have similar accuracies for all subsets except for polyproline II helices, where SLiMPred shows a markedly higher accuracy. Also for prediction of ordered SLiMs and disordered SLiMs, i.e. SLiMs predicted to be ordered or disordered respectively, the algorithms showed similar accuracy on disordered SLiMs, but

SLiMPred demonstrated much higher prediction accuracy on ordered SLiMs. It should be noted that IUPred was used to classify ordered and disordered SLiMs, and that the IUPred prediction score is one of ANCHORs three primary attributes. Therefore this test is inherently biased against ANCHOR prediction, however it does beg for an evaluation of experimentally verified ordered and disordered SLiMs by ANCHOR and SLiMPred.

In terms of best accuracy, there is no direct comparison of all methods, so no definitive statements can be made. Inferring from similar evaluation accuracy of SLiMPred and ANCHOR (328) and the comparison of MoRFPred and ANCHOR (112), it is likely that MoRFPred has the highest prediction accuracy, with a slightly lower true positive rate but a much lower false positive rate than either ANCHOR or SLiMPred. The α-MoRF predictors do not compare to the other predictors in terms of their true positive rate, due to the limited scope of these algorithms. However, they do compare well in terms of their false negative rates, which is somewhat surprising given that intrinsic disorder was not included in their negative training sets.

An extensive comparison of DOT binding region prediction and motif prediction demonstrates a useful synergy between these approaches (336). For many type of ligand binding motifs from the ELM database, two thirds of motif instances overlap with ANCHOR-predicted DOT binding motifs, but less than a fifth of predicted ligand binding motifs ANCHOR predictions. This result suggest that DOT binding region may make an effective sequence filter of motif predictions by greatly increasing specificity with only a modest reduction in recall. The observed overlap is highly dependent of the type of motif. Motif instances occur in predicted ordered regions, disordered regions, or mixed environments. Motif instances occurring in disordered regions generally show a high overlap with ANCHOR predictions, and motif instances in mixed order-disorder regions show a good but reduced overlap with ANCHOR predictions. Those few motif instances in ordered regions show very little overlap with ANCHOR predictions, which suggests that predictions of ordered regions would be an appropriate filter for these motif

types. Also, secondary structure of the bound motif instance effects filtering efficiency; overlap with ANCHOR predictions is increased for motifs that form helix and somewhat decreased for motifs that form extended structures. Finally, DOT binding region prediction can enrich predicted motifs in biologically relevant predictions, as demonstrated by overlap between predicted nuclear receptor box motifs and ANCHOR; ANCHOR filters nearly four fifths of the predicted nuclear receptor box motifs in the human proteome and the remaining predicted motifs are enriched in relevant Gene Ontology annotations: nuclear localization, regulation of transcription, transcription cofactor, and protein binding annotations.

There are commonalities between the reviewed methods, but also important differences. The differences are particularly important for datasets and evaluation methods. Some aspects of these differences may be leveraged in order to create improved predictors and extend biological knowledge.

Conclusions

Current DOT prediction methods are varied in every aspect: from the underlying model to the construction of datasets to methods of evaluation. The similarity between the MoRF, SLIM, and motif models has been demonstrated. But these models, which all address short DOT regions, have yet to be reconciled with long DOT regions, and a combined model of short and long regions would result in expanding available data for predictor development. Also, control datasets need to be refined and expanded, particularly for non-interacting IDP control sequences. Finally, the complexity observed for some IDPs with respect to multiple binding suggests that prediction of binding sites from single sequences may be too limited.

Expanding DOT Datasets

Several surveys of DOT binding regions have shown that these regions come in various lengths: from short segments of a few residues to long regions of a size comparable to small folded domains. Given the length heterogeneity of DOT binding regions, the current generation of predictors would seem to address only a portion of the prediction problem, which is the

prediction of short DOT regions. Indeed, coverage of long disorder-to-order binding regions by current predictors has been shown to be quite poor (327, 328). However, it may be the case that long DOT binding regions are composed of multiple, short DOT binding regions, which is suggested by one analysis (327). Additional experimental evidence is required to verify if decomposition of long DOT binding regions to a combination of short regions is a generally valid approach. However, this approach is consistent with the observation that long DOT regions have a low amount of intra-protein buried surface area in bound structures (100, 337), with much of the buried surface occurring between the disordered protein and its partner. These observations suggest that the use of short regions for predictor development may be an adequate, in the absence of additional experimental evidence.

Many additional DOT examples may be available from the PDB, but to date, there has not been a comprehensive study of the DOTs present in protein structures in the PDB. Reported surveys, which rely on comparative analysis, have been limited by availability and reliability of reference monomer structures. Though limited, these surveys suggest that DOTs may be a common feature in protein complex structures. Therefore, a survey of DOTs independent of reference monomers could potentially reveal many more novel DOT regions. Model-based analyses, particularly molecular dynamics and thermodynamic stability analysis (i.e. COREX), are theoretically capable of performing a survey of DOTs in the PDB. However, the computational demands and practical limitations make such a survey intractable. Alternatively, structural analysis has been shown to be a promising method for analyzing complex structures for the presence of DOTs (e.g. surface area analysis (100)). However, current methods are only applicable to entire proteins, and are applicable to protein containing both DOT and structured regions. At least one alternative method designed for the classification of DOT and structured regions has been developed (51), but not well characterized. Also, other methods developed for flexibility analysis, such as normal mode (105) or mechanical (103) analyses, may be repurposed

for identification of DOT regions. Development and characterization of new analyses will help to greatly expand DOT region datasets.

Another avenue for expanding DOT datasets is to combine the datasets of the current methods. All methods reviewed here relied on short binding regions as a positive dataset for training. Generally, complex structures were used to determine the bounds of the DOT binding region and disorder in the unbound state was verified to differing extents, where larger datasets precluded strict validation used for smaller datasets. For motif based methods, short motif patterns and instances were collected from the literature and are consequently well verified. Although there exists some overlap between these sources of short DOT binding regions, a union of these two sources would significantly increase the size of available datasets. Such an approach should consider that some minority of motifs are known to be or likely to be in a structured context or themselves structured prior to binding.

It is a general concern that some ordered regions might be included in constructing datasets of DOT regions. Some smaller datasets require literature evidence of intrinsic disorder in the unbound state (111, 327), but this approach is prohibitive for construction of larger datasets. Although structural analysis can give an indication of the order-disorder state of the isolated monomer (100), this analysis assumes a structure is available and only applies to the fragment present in the structure. The more important question is if the fragment is structured or disordered in the context of the parent protein. Several ELM motifs types seem to occur within a disordered context (237), but the binding regions themselves are not necessarily disordered. In other words, a DOT region may occur as an unstructured loop or tail of a structured domain, e.g. 14-3-3 binding region of serotonin N-acetyltransferase (186). Such regions lack the disordered context of typical MoRFs and are difficult for predictors that rely on this context for prediction. It is clear that these types of DOTs do not fit neatly into the MoRF model and may require treatment as a special case. However, the increased performance of SLiMPred on these regions,

relative to ANCHOR, suggests that special treatment may not be necessary given the appropriate attributes and prediction architecture. Similar difficulties arise when constructing an appropriate control set.

Control Datasets

A control set is necessary for training prediction methods and for estimation of the rate of false prediction. In the motif-based formulation of the DOT binding region prediction problem, sets of motif instances that do not bind to associated domains are very difficult, if not infeasible, to construct. For this reason, there has been no direct quantification of the false positive rate of motif-based methods. Prediction methods however require a control set for training, and generally two types of protein regions are used, ordered and disordered.

For ordered sequences, generally only sequences with monomeric structures are used. Generally complexes are avoided due to the possibility that one or more of the partners is disordered in the unbound state, except where additional evidence is available to indicate the complex is formed by ordered monomers.

Intrinsically disordered regions are problematic for the same reason that motif control datasets are problematic; it is generally not known whether a given disordered region binds to any partners. One approach to this problem is to assume that interaction regions are sparse and that residues outside of known binding domains do not interact with partners, which is the approach taken by MoRFPred and SLiMPred. However, the interaction map of p53 (60), one of the most intensely studied proteins in the human genome, shows that its interactions are tiled over the disordered termini. While p53 may not be a prototypical protein, this observation does call into question the assumption of sparsity of interaction regions within disordered segments.

Improving Prediction Model

The example of p53 raises two complications with prediction of intrinsically disordered binding regions. Under the current model of DOT binding region prediction, a residue is predicted to be a binding residue if it interacts with any partner. This implies that the correct

prediction for the N-terminus of p53 – according to the interaction map which shows that the N-terminus is tiled with interaction sites, would have nearly the entirety of the N-terminus predicted as binding residues. While correct under the current model, such a prediction is not practically very useful for novel proteins and may be misleading. The other complication is one-to-many binding regions, in which the same region or overlapping regions bind to multiple partners. That is, it is tempting to interpret predictions of binding sites as binding to a single partner, but multiple partners may be involved. Current predictors do not address this complication. Motif-based methods do not suffer these complications, but do not predict novel binding sites.

A more useful and transparent prediction model would decompose a protein into non-binding sites and competent, perhaps overlapping, binding sites. However, direct application of a single sequence approach to decomposition represents a large increase in algorithm sophistication and complexity, and would not address the case of one-to-many binding. An alternative approach that may be more effective would be to combine binding site prediction with knowledge of a particular partner. Such an approach has been applied in the analysis of peptide binding domains, such as SH3 (338) domains. This approach considers the peptide binding site of the peptide binding domain to predict peptide sequences compatible with binding to a particular instance of a peptide binding domain. In a sense, this method constructs a pseudo motif, which is dependent on the details of the binding partner. Another approach is to consider coevolving residues between two binding partners in the form of mutual information (339). Generalization of these types of methods, in combination with consideration of context provided by the current generation of binding site predictors, may be the way forward to the next generation of DOT binding region prediction.

REFERENCES

1. Anfinsen CB. 1973. Principles that govern the folding of protein chains. *Science*. 181(96):223–30
2. Fischer E. 1894. Einfluss der configuration auf die wirkung der enzyme. *Berichte der deutschen chemischen Gesellschaft*. 27(3):2993, 2985
3. Blow DM. 1997. The tortuous story of asp…his…ser: structural analysis of α-chymotrypsin. *Trends in Biochemical Sciences*. 22(10):405–8
4. Perutz MF, Wilkinson AJ, Paoli M, Dodson GG. 1998. The stereochemical mechanism of the cooperative effects in hemoglobin revisited. *Annual Review of Biophysics and Biomolecular Structure*. 27(1):1–34
5. Daniel RM, Dunn RV, Finney JL, Smith JC. 2003. The role of dynamics in enzyme activity. *Annual Review of Biophysics and Biomolecular Structure*. 32(1):69–92
6. Dunker AK, Obradovic Z. 2001. The protein trinity: linking function and disorder. *Nat Biotech*. 19(9):805–6
7. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradović Z. 2002. Intrinsic disorder and protein function. *Biochemistry*. 41(21):6573–82
8. Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. 2001. Sequence complexity of disordered protein. *Proteins*. 42(1):38–48
9. Uversky VN, Oldfield CJ, Dunker AK. 2005. Showing your id: intrinsic disorder as an id for recognition, regulation and cell signaling. *J. Mol. Recognit*. 18(5):343–84
10. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol*. 337(3):635–45
11. Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, et al. 2007. Functional anthology of intrinsic disorder. 1. biological processes and functions of proteins with long disordered regions. *J Proteome Res*. 6(5):1882–98
12. Bailey RW, Dunker AK, Brown CJ, Garner EC, Griswold MD. 2001. Clusterin, a binding protein with a molten globule-like region. *Biochemistry*. 40(39):11828–40
13. Iakoucheva LM, Kimzey AL, Masselon CD, Bruce JE, Garner EC, et al. 2001. Identification of intrinsic order and disorder in the dna repair protein xpa. *Protein Science*. 10(3):560–71
14. Stott K, Watson M, Howe FS, Grossmann JG, Thomas JO. 2010. Tail-mediated collapse of hmgb1 is dynamic and occurs via differential binding of the acidic tail to the a and b domains. *Journal of Molecular Biology*. 403(5):706–22
15. Gosselin P, Oulhen N, Jam M, Ronzca J, Cormier P, et al. 2011. The translational repressor 4e-bp called to order by eif4e: new structural insights by saxs. *Nucleic Acids Res*. 39(8):3496–3503
16. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, et al. 2007. Disprot: the database of disordered proteins. *Nucleic Acids Res*. 35(Database issue):D786–93
17. Schomaker V, Trueblood KN. 1968. On the rigid-body motion of molecules in crystals. *Acta Crystallographica Section B Structural Crystallography and Crystal Chemistry*. 24(1):63–76
18. Levin EJ, Kondrashov DA, Wesenberg GE, Phillips Jr. GN. 2007. Ensemble refinement of protein crystal structures: validation and application. *Structure*. 15(9):1040–52
19. Kay LE. 1998. Protein dynamics from nmr. *Nat Struct Mol Biol*. 5:513–17
20. Lacy ER, Filippov I, Lewis WS, Otieno S, Xiao L, et al. 2004. P27 binds cyclin–cdk complexes through a sequential mechanism involving binding-induced protein folding. *Nat Struct Mol Biol*. 11(4):358–64

21. Bienkiewicz EA, Adkins JN, Lumb KJ. 2002. Functional consequences of preorganized helical structure in the intrinsically disordered cell-cycle inhibitor p27kip1†. *Biochemistry*. 41(3):752–59
22. Markwick PRL, Malliavin T, Nilges M. 2008. Structural biology by nmr: structure, dynamics, and interactions. *PLoS Comput Biol*. 4(9):e1000168
23. Ota M, Koike R, Amemiya T, Tenno T, Romero PR, et al. 2013. An assignment of intrinsically disordered regions of proteins based on nmr structures. *Journal of Structural Biology*. 181(1):29–36
24. Fontana A, Polverino de Laureto P, De Filippis V, Scaramella E, Zambonin M. 1997. Probing the partly folded states of proteins by limited proteolysis. *Fold Des*. 2(2):R17–26
25. Chemes LB, Alonso LG, Noval MG, Prat-Gay G de. 2012. Circular dichroism techniques for the analysis of intrinsically disordered proteins and domains. In *Intrinsically Disordered Protein Analysis*, ed VN Uversky, AK Dunker, pp. 387–404. Humana Press
26. Wang Y, Teraoka I, Hansen FY, Peters GH, Hassager O. 2010. A theoretical study of the separation principle in size exclusion chromatography. *Macromolecules*. 43(3):1651–59
27. Uversky VN. 2002. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci*. 11(4):739–56
28. Lipfert J, Doniach S. 2007. Small-angle x-ray scattering from rna, proteins, and protein complexes. *Annual Review of Biophysics and Biomolecular Structure*. 36(1):307–27
29. Receveur-Brechot V, Durand D. 2012. How random are intrinsically disordered proteins? a small angle scattering perspective. *Current Protein & Peptide Science*. 13(1):55–75
30. Uversky VN, Dunker AK, eds. 2012. *Intrinsically Disordered Protein Analysis - Volume 1, Methods and Experimental Tools*, Vol. 895. Totowa, NJ, USA: Humana Press. 511 pp.
31. Uversky VN, Dunker AK, eds. 2012. *Intrinsically Disordered Protein Analysis - Volume 2, Methods and Experimental Tools*, Vol. 896. Totowa, NJ, USA: Humana Press. 454 pp.
32. Balasubramaniam D, Komives EA. 2013. Hydrogen-exchange mass spectrometry for the study of intrinsic disorder in proteins. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*. 1834(6):1202–9
33. Dedmon MM, Lindorff-Larsen K, Christodoulou J, Vendruscolo M, Dobson CM. 2005. Mapping long-range interactions in α-synuclein using spin-label nmr and ensemble molecular dynamics simulations. *J. Am. Chem. Soc.* 127(2):476–77
34. Uversky VN, Dunker AK. 2012. Multiparametric analysis of intrinsically disordered proteins: looking at intrinsic disorder through compound eyes. *Anal. Chem.* 84(5):2096–2104
35. Marsh JA, Forman-Kay JD. 2009. Structure and disorder in an unfolded state under nondenaturing conditions from ensemble models consistent with a large number of experimental restraints. *Journal of Molecular Biology*. 391(2):359–74
36. Sigler PB. 1988. Acid blobs and negative noodles. *Nature*. 333(6170):210–12
37. Uversky VN, Gillespie JR, Fink AL. 2000. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins: Structure, Function, and Bioinformatics*. 41(3):415–27
38. Vacic V, Uversky VN, Dunker AK, Lonardi S. 2007. Composition profiler: a tool for discovery and visualization of amino acid composition differences. *BMC Bioinformatics*. 8(1):211
39. Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, et al. 2005. Foldindex©: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics*. 21(16):3435–38
40. Romero, Obradovic, Dunker. 1997. Sequence data analysis for long disordered regions prediction in the calcineurin family. *Genome Inform Ser Workshop Genome Inform*. 8:110–24

41. Ferron F, Longhi S, Canard B, Karlin D. 2006. A practical overview of protein disorder prediction methods. *Proteins: Structure, Function, and Bioinformatics*. 65(1):1–14
42. Monastyrskyy B, Fidelis K, Moult J, Tramontano A, Kryshtafovych A. 2011. Evaluation of disorder predictions in casp9. *Proteins: Structure, Function, and Bioinformatics*. 79(S10):107–18
43. Dunker AK, Oldfield CJ, Meng J, Romero P, Yang JY, et al. 2008. The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics*. 9(Suppl 2):S1
44. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK. 2002. Intrinsic disorder in cell-signaling and cancer-associated proteins. *Journal of Molecular Biology*. 323(3):573–84
45. Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN, Dunker AK. 2006. Intrinsic disorder in transcription factors. *Biochemistry*. 45(22):6873–88
46. Xue B, Dunker AK, Uversky VN. 2012. The roles of intrinsic disorder in orchestrating the wnt-pathway. *Journal of Biomolecular Structure and Dynamics*. 29(5):843–61
47. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. 2005. Flexible nets: the roles of intrinsic disorder in protein interaction networks. *FEBS Journal*. 272(20):5129–48
48. Cortese MS, Uversky VN, Keith Dunker A. 2008. Intrinsic disorder in scaffold proteins: getting more from less. *Progress in Biophysics and Molecular Biology*. 98(1):85–106
49. Xue B, Dunker AK, Uversky VN. 2012. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *Journal of Biomolecular Structure and Dynamics*. 30(2):137–49
50. Peng Z, Mizianty MJ, Xue B, Kurgan L, Uversky VN. 2012. More than just tails: intrinsic disorder in histone proteins. *Mol. BioSyst.* 8(7):1886–1901
51. Peng Z, Oldfield CJ, Xue B, Mizianty MJ, Dunker AK, et al. A creature with a hundred waggly tails: intrinsically disordered proteins in the ribosome. *Cell. Mol. Life Sci.*, pp. 1–28
52. Fladvad M, Zhou K, Moshref A, Pursglove S, Säfsten P, Sunnerhagen M. 2005. N and c-terminal sub-regions in the c-myc transactivation region and their joint role in creating versatility in folding and binding. *Journal of Molecular Biology*. 346(1):175–89
53. Huth JR, Bewley CA, Nissen MS, Evans JNS, Reeves R, et al. 1997. The solution structure of an hmg-i(y)–dna complex defines a new architectural minor groove binding motif. *Nat Struct Mol Biol*. 4(8):657–65
54. Guo X, Bulyk ML, Hartemink AJ. 2012. Intrinsic disorder within and flanking the dna-binding domains of human transcription factors
55. Dyson HJ, Wright PE. 2002. Coupling of folding and binding for unstructured proteins. *Current Opinion in Structural Biology*. 12(1):54–60
56. Vacic V, Oldfield CJ, Mohan A, Radivojac P, Cortese MS, et al. 2007. Characterization of molecular recognition features, morfs, and their binding partners. *J Proteome Res*. 6(6):2351–66
57. Gunasekaran K, Tsai C-J, Kumar S, Zanuy D, Nussinov R. 2003. Extended disordered proteins: targeting function with less scaffold. *Trends in Biochemical Sciences*. 28(2):81–85
58. Good MC, Zalatan JG, Lim WA. 2011. Scaffold proteins: hubs for controlling the flow of cellular information. *Science*. 332(6030):680–86
59. Ferrell JE Jr. 2000. What do scaffold proteins really do? *Sci. STKE*. 2000(52):pe1
60. Oldfield CJ, Meng J, Yang JY, Yang MQ, Uversky VN, Dunker AK. 2008. Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. *BMC Genomics.* 9(Suppl 1):S1

61. Hsu W-L, Oldfield CJ, Xue B, Meng J, Huang F, et al. 2013. Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding. *Protein Science*. 22(3):258–73
62. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, et al. 2004. The importance of intrinsic disorder for protein phosphorylation. *Nucl. Acids Res.* 32(3):1037–49
63. Daily KM, Radivojac P, Dunker AK, Methylation AP. 2005. Intrinsic disorder and protein modifications: building an svm predictor for methylation
64. Gao J, Xu D. 2012. Correlation between posttranslational modification and intrinsic disorder in protein
65. Mitrea DM, Kriwacki RW. 2012. Cryptic disorder: an order-disorder transformation regulates the function of nucleophosmin. *Pac Symp Biocomput*, pp. 152–63
66. Romero PR, Zaidi S, Fang YY, Uversky VN, Radivojac P, et al. 2006. Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *PNAS*. 103(22):8390–95
67. Babu MM, van der Lee R, de Groot NS, Gsponer J. 2011. Intrinsically disordered proteins: regulation and disease. *Current Opinion in Structural Biology*. 21(3):432–40
68. Magidovich E, Orr I, Fass D, Abdu U, Yifrach O. 2007. Intrinsic disorder in the c-terminal domain of the shaker voltage-activated k+ channel modulates its interaction with scaffold proteins. *PNAS*. 104(32):13022–27
69. Zhou M, Morais-Cabral JH, Mann S, MacKinnon R. 2001. Potassium channel receptor site for the inactivation gate and quaternary amine inhibitors. *Nature*. 411(6838):657–61
70. Denning DP, Patel SS, Uversky V, Fink AL, Rexach M. 2003. Disorder in the nuclear pore complex: the fg repeat regions of nucleoporins are natively unfolded. *Proc Natl Acad Sci U S A*. 100(5):2450–55
71. Ma J, Goryaynov A, Sarma A, Yang W. 2012. Self-regulated viscous channel in the nuclear pore complex. *PNAS*
72. Brown HG, Hoh JH. 1997. Entropic exclusion by neurofilament sidearms: a mechanism for maintaining interfilament spacing†. *Biochemistry*. 36(49):15035–40
73. Mittag T, Orlicky S, Choy W-Y, Tang X, Lin H, et al. 2008. Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proceedings of the National Academy of Sciences*. 105(46):17772 –17777
74. Verma R, Annan RS, Huddleston MJ, Carr SA, Reynard G, Deshaies RJ. 1997. Phosphorylation of sic1p by g1 cdk required for its degradation and entry into s phase. *Science*. 278(5337):455–60
75. Nash P, Tang X, Orlicky S, Chen Q, Gertler FB, et al. 2001. Multisite phosphorylation of a cdk inhibitor sets a threshold for the onset of dna replication. *Nature*. 414(6863):514–21
76. Mittag T, Marsh J, Grishaev A, Orlicky S, Lin H, et al. 2010. Structure/function implications in a dynamic complex of the intrinsically disordered sic1 with the cdc4 subunit of an scf ubiquitin ligase. *Structure*. 18(4):494–506
77. Borg M, Mittag T, Pawson T, Tyers M, Forman-Kay JD, Chan HS. 2007. Polyelectrostatic interactions of disordered ligands suggest a physical basis for ultrasensitivity. *Proceedings of the National Academy of Sciences*. 104(23):9650 –9655
78. Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, et al. 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. *J. Mol. Evol.* 55(1):104–10
79. Brown CJ, Johnson AK, Dunker AK, Daughdrill GW. 2011. Evolution and disorder. *Curr. Opin. Struct. Biol.* 21(3):441–46
80. Lu X, Hamkalo B, Parseghian MH, Hansen JC. 2009. Chromatin condensing functions of the linker histone c-terminal domain are mediated by specific amino acid composition and intrinsic protein disorder†. *Biochemistry*. 48(1):164–72

81. Chen JW, Romero P, Uversky VN, Dunker AK. 2006. Conservation of intrinsic disorder in protein domains and families: i. a database of conserved predicted disordered regions. *Journal of Proteome Research*. 5(4):879–87
82. Williams RW, Xue B, Uversky VN, Dunker AK. 2013. Distribution and cluster analysis of predicted intrinsically disordered protein pfam domains. *Intrinsically Disordered Proteins*. 1(1):82–107
83. Fisher CK, Stultz CM. 2011. Protein structure along the order–disorder continuum. *J. Am. Chem. Soc.* 133(26):10022–25
84. Fisher CK, Huang A, Stultz CM. 2010. Modeling intrinsically disordered proteins with bayesian statistics. *J. Am. Chem. Soc.* 132(42):14919–27
85. Karplus M, McCammon JA. 2002. Molecular dynamics simulations of biomolecules. *Nat Struct Mol Biol*. 9(9):646–52
86. Lindorff-Larsen K, Trbovic N, Maragakis P, Piana S, Shaw DE. 2012. Structure and dynamics of an unfolded protein examined by molecular dynamics simulation. *J. Am. Chem. Soc.* 134(8):3787–91
87. Vitalis A, Pappu RV. 2009. Methods for monte carlo simulations of biomacromolecules. *Annu Rep Comput Chem*. 5:49–76
88. Marsh JA, Forman-Kay JD. 2010. Sequence determinants of compaction in intrinsically disordered proteins. *Biophysical Journal*. 98(10):2383–90
89. Vitalis A, Wang X, Pappu RV. 2007. Quantitative characterization of intrinsic disorder in polyglutamine: insights from analysis based on polymer theories. *Biophysical Journal*. 93(6):1923–37
90. Mao AH, Crick SL, Vitalis A, Chicoine CL, Pappu RV. 2010. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* 107(18):8183–88
91. Das RK, Pappu RV. 2013. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *PNAS*, p. 201304749
92. Janin J, Sternberg MJE. 2013. Protein flexibility, not disorder, is intrinsic to molecular recognition. *F1000 Biol Rep*. 5:2
93. Metallo SJ. 2010. Intrinsically disordered proteins are potential drug targets. *Curr Opin Chem Biol*. 14(4):481–88
94. Wright PE, Dyson HJ. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol*. 293(2):321–31
95. Kundu S, Melton JS, Sorensen DC, Phillips GN. 2002. Dynamics of proteins in crystals: comparison of experiment with simple models. *Biophys J*. 83(2):723–32
96. Le Gall T, Romero PR, Cortese MS, Uversky VN, Dunker AK. 2007. Intrinsic disorder in the protein data bank. *J. Biomol. Struct. Dyn*. 24(4):325–42
97. Zhang Y, Stec B, Godzik A. 2007. Between order and disorder in protein structures: analysis of “dual personality” fragments in proteins. *Structure*. 15(9):1141–47
98. Fong JH, Shoemaker BA, Garbuzynskiy SO, Lobanov MY, Galzitskaya OV, Panchenko AR. 2009. Intrinsic disorder in protein interactions: insights from a comprehensive structural analysis. *PLoS Comput Biol*. 5(3):e1000316
99. Mohan A, Oldfield CJ, Radivojac P, Vacic V, Cortese MS, et al. 2006. Analysis of molecular recognition features (morfs). *J Mol Biol*. 362(5):1043–59
100. 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(5):1327–41
101. Hilser VJ, García-Moreno E B, Oas TG, Kapp G, Whitten ST. 2006. A statistical thermodynamic model of the protein ensemble. *Chem. Rev*. 106(5):1545–58

102. Li L, Uversky VN, Dunker AK, Meroueh SO. 2007. A computational investigation of allostery in the catabolite activator protein. *J. Am. Chem. Soc*. 129(50):15668–76
103. Jacobs DJ, Rader AJ, Kuhn LA, Thorpe MF. 2001. Protein flexibility predictions using graph theory. *Proteins*. 44(2):150–65
104. Liang S, Li L, Hsu W-L, Pilcher MN, Uversky V, et al. 2009. Exploring the molecular design of protein interaction sites with molecular dynamics simulations and free energy calculations. *Biochemistry*. 48(2):399–414
105. Bahar I, Rader A. 2005. Coarse-grained normal mode analysis in structural biology. *Curr Opin Struct Biol*. 15(5):586–92
106. Shoemaker BA, Panchenko AR, Bryant SH. 2006. Finding biologically relevant protein domain interactions: conserved binding mode analysis. *Protein Sci*. 15(2):352–61
107. Haliloglu T, Bahar I, Erman B. 1997. Gaussian dynamics of folded proteins. *Phys. Rev. Lett.* 79(16):3090
108. Tirion. 1996. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett*. 77(9):1905–8
109. Hilser VJ, Freire E. 1996. Structure-based calculation of the equilibrium folding pathway of proteins. correlation with hydrogen exchange protection factors. *J. Mol. Biol*. 262(5):756–72
110. Oldfield CJ, Xue B, Van Y-Y, Ulrich EL, Markley JL, et al. 2013. Utilization of protein intrinsic disorder knowledge in structural proteomics. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*. 1834(2):487–98
111. Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, Dunker AK. 2005. Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry*. 44(37):12454–70
112. Disfani FM, Hsu W-L, Mizianty MJ, Oldfield CJ, Xue B, et al. 2012. Morfpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics*. 28(12):i75–i83
113. Goh K-I, Oh E, Jeong H, Kahng B, Kim D. 2002. Classification of scale-free networks. *PNAS*. 99(20):12583–88
114. Watts DJ, Strogatz SH. 1998. Collective dynamics of "small-world" networks. *Nature*. 393(6684):440–42
115. Erdős P, Rényi A. 1960. On the evolution of random graphs
116. Barabási A-L, Bonabeau E. 2003. Scale-free networks. *Scientific American*, pp. 60–69
117. Albert R, Jeong H, Barabási A-L. 2000. Error and attack tolerance of complex networks : article : nature. *Nature*. 406(6794):378–82
118. Jeong H, Mason SP, Barabási A-L, Oltvai ZN. 2001. Lethality and centrality in protein networks. *Nature*. 411(6833):41–42
119. Milgram S. 1967. The small world problem. *Psychology Today*. 2:60–67
120. Bork P, Jensen LJ, von Mering C, Ramani AK, Lee I, Marcotte EM. 2004. Protein interaction networks from yeast to human. *Current Opinion in Structural Biology*. 14(3):292–99
121. Barabási A-L, Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. *Nat Rev Genet*. 5(2):101–13
122. Wu CH, Huang H, Nikolskaya A, Hu Z, Barker WC. 2004. The iproclass integrated database for protein functional analysis. *Comput Biol Chem*. 28(1):87–96
123. Huang T-W, Tien A-C, Huang W-S, Lee Y-CG, Peng C-L, et al. 2004. Point: a database for the prediction of protein–protein interactions based on the orthologous interactome. *Bioinformatics*. 20(17):3273–76
124. Kelley BP, Yuan B, Lewitter F, Sharan R, Stockwell BR, Ideker T. 2004. Pathblast: a tool for alignment of protein interaction networks. *Nucl. Acids Res.* 32(suppl 2):W83–W88

125. Mering C von, Jensen LJ, Snel B, Hooper SD, Krupp M, et al. 2005. String: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucl. Acids Res.* 33(suppl 1):D433–D437
126. Hahn MW, Kern AD. 2005. Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol*. 22(4):803–6
127. Huang S. 2004. Back to the biology in systems biology: what can we learn from biomolecular networks? *Briefings in Functional Genomics and Proteomics*. 2(4):279–97
128. Han J-DJ, Bertin N, Hao T, Goldberg DS, Berriz GF, et al. 2004. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*. 430(6995):88–93
129. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. 1999. From molecular to modular cell biology. *Nature*. 402:C47–C52
130. Cesareni G, Ceol A, Gavrila C, Palazzi LM, Persico M, Schneider MV. 2005. Comparative interactomics. *FEBS Letters*. 579(8):1828–33
131. Von Mering C, Krause R, Snel B, Cornell M, Oliver SG, et al. 2002. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*. 417(6887):399–403
132. Bader GD, Hogue CWV. 2002. Analyzing yeast protein-protein interaction data obtained from different sources. *Nat. Biotechnol.* 20(10):991–97
133. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, Sakaki Y. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *PNAS*. 98(8):4569–74
134. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, et al. 2000. A comprehensive analysis of protein–protein interactions in saccharomyces cerevisiae. *Nature*. 403(6770):623–27
135. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, et al. 2004. A map of the interactome network of the metazoan c. elegans. *Science*. 303(5657):540–43
136. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, et al. 2003. A protein interaction map of drosophila melanogaster. *Science*. 302(5651):1727–36
137. Han J-DJ, Dupuy D, Bertin N, Cusick ME, Vidal M. 2005. Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotech*. 23(7):839–44
138. Hasty J, Collins JJ. 2001. Protein interactions: unspinning the web. *Nature*. 411(6833):30–31
139. KOSHLAND DE Jr, RAY WJ Jr, ERWIN MJ. 1958. Protein structure and enzyme action. *Fed. Proc.* 17(4):1145–50
140. Landsteiner K. 1936. *The Specificity of Serological Reactions*. Mineola, New York: Courier Dover Publications
141. Pauling L. 1940. A theory of the structure and process of formation of antibodies*. *J. Am. Chem. Soc.* 62(10):2643–57
142. Karush F. 1950. Heterogeneity of the binding sites of bovine serum albumin1. *J. Am. Chem. Soc.* 72(6):2705–13
143. Meador WE, Means AR, Quiocho FA. 1993. Modulation of calmodulin plasticity in molecular recognition on the basis of x-ray structures. *Science*. 262(5140):1718–21
144. Kriwacki RW, Hengst L, Tennant L, Reed SI, Wright PE. 1996. Structural studies of p21waf1/cip1/sdi1 in the free and cdk2-bound state: conformational disorder mediates binding diversity. *PNAS*. 93(21):11504–9
145. Uversky VN. 2003. A protein-chameleon: conformational plasticity of alpha-synuclein, a disordered protein involved in neurodegenerative disorders. *J. Biomol. Struct. Dyn.* 21(2):211–34
146. Dunker AK, Garner E, Guilliot S, Romero P, Albrecht K, et al. 1998. Protein disorder and the evolution of molecular recognition: theory, predictions and observations. *Pac Symp Biocomput*, pp. 473–84

147. Patil A, Nakamura H. 2006. Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks. *FEBS Letters*. 580(8):2041–45
148. Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME, et al. 2006. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. *PLoS Comput Biol*. 2(8):e100
149. Ekman D, Light S, Björklund ÅK, Elofsson A. 2006. What properties characterize the hub proteins of the protein-protein interaction network of saccharomyces cerevisiae? *Genome Biology*. 7(6):R45
150. Dosztányi Z, Chen J, Dunker AK, Simon I, Tompa P. 2006. Disorder and sequence repeats in hub proteins and their implications for network evolution. *J. Proteome Res.* 5(11):2985–95
151. Singh GP, Ganapathi M, Sandhu KS, Dash D. 2006. Intrinsic unstructuredness and abundance of pest motifs in eukaryotic proteomes. *Proteins: Structure, Function, and Bioinformatics*. 62(2):309–15
152. Marinissen MJ, Gutkind JS. 2005. Scaffold proteins dictate rho gtpase-signaling specificity. *Trends in Biochemical Sciences*. 30(8):423–26
153. Jaffe AB, Aspenström P, Hall A. 2004. Human cnk1 acts as a scaffold protein, linking rho and ras signal transduction pathways. *Mol. Cell. Biol.* 24(4):1736–46
154. Jaffe AB, Hall A. 2005. Rho gtpases: biochemistry and biology. *Annual Review of Cell and Developmental Biology*. 21(1):247–69
155. Hohenstein P, Giles RH. 2003. Brca1: a scaffold for p53 response? *Trends in Genetics*. 19(9):489–94
156. Luo W, Lin S-C. 2004. Axin: a master scaffold for multiple signaling pathways. *Neurosignals*. 13(3):99–113
157. Rui Y, Xu Z, Lin S, Li Q, Rui H, et al. 2004. Axin stimulates p53 functions by activation of hipk2 kinase through multimeric complex formation. *EMBO J*. 23(23):4583–94
158. Salahshor S, Woodgett JR. 2005. The links between axin and carcinogenesis. *J Clin Pathol*. 58(3):225–36
159. Wong W, Scott JD. 2004. Akap signalling complexes: focal points in space and time. *Nat Rev Mol Cell Biol*. 5(12):959–70
160. Carpousis AJ. 2007. The rna degradosome of escherichia coli: an mrna-degrading machine assembled on rnase e. *Annual Review of Microbiology*. 61(1):71–87
161. Kim PM, Lu LJ, Xia Y, Gerstein MB. 2006. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*. 314(5807):1938 –1941
162. Anderson C, Appella E. 2003. Signaling to the p53 tumor suppressor through pathways activated by genotoxic and nongenotoxic stress. In *Handbook of Cell Signaling*, ed R Bradshaw, pp. 237–47. New York: Academic Press
163. Hollstein M, Sidransky D, Vogelstein B, Harris CC. 1991. P53 mutations in human cancers. *Science*. 253(5015):49–53
164. Zhao R, Gish K, Murphy M, Yin Y, Notterman D, et al. 2000. Analysis of p53-regulated gene expression patterns using oligonucleotide arrays. *Genes Dev*. 14(8):981–93
165. Dougherty MK, Morrison DK. 2004. Unlocking the code of 14-3-3. *J Cell Sci*. 117(10):1875–84
166. Rubio MP, Geraghty KM, Wong BHC, Wood NT, Campbell DG, et al. 2004. 14-3-3-affinity purification of over 200 human phosphoproteins reveals new links to regulation of cellular metabolism, proliferation and trafficking. *Biochemical Journal*. 379(2):395
167. Meek SEM, Lane WS, Piwnica-Worms H. 2004. Comprehensive proteomic analysis of interphase and mitotic 14-3-3-binding proteins. *J. Biol. Chem.* 279(31):32046–54

168. Jin J, Smith FD, Stark C, Wells CD, Fawcett JP, et al. 2004. Proteomic, functional, and domain-based analysis of in vivo 14-3-3 binding proteins involved in cytoskeletal regulation and cellular organization. *Current Biology*. 14(16):1436–50
169. Yaffe MB. 2002. How do 14-3-3 proteins work? – gatekeeper phosphorylation and the molecular anvil hypothesis. *FEBS Letters*. 513(1):53–57
170. Rittinger K, Budman J, Xu J, Volinia S, Cantley LC, et al. 1999. Structural analysis of 14-3-3 phosphopeptide complexes identifies a dual role for the nuclear export signal of 14-3-3 in ligand binding. *Molecular Cell*. 4(2):153–66
171. Yaffe MB, Rittinger K, Volinia S, Caron PR, Aitken A, et al. 1997. The structural basis for 14-3-3:phosphopeptide binding specificity. *Cell*. 91(7):961–71
172. Bustos DM, Iglesias AA. 2006. Intrinsic disorder is a key characteristic in partners that bind 14-3-3 proteins. *Proteins: Structure, Function, and Bioinformatics*. 63(1):35–42
173. Lowe ED, Tews I, Cheng KY, Brown NR, Gul S, et al. 2002. Specificity determinants of recruitment peptides bound to phospho-cdk2/cyclin a. *Biochemistry*. 41(52):15625–34
174. Avalos JL, Celic I, Muhammad S, Cosgrove MS, Boeke JD, Wolberger C. 2002. Structure of a sir2 enzyme bound to an acetylated p53 peptide. *Molecular Cell*. 10(3):523–35
175. Mujtaba S, He Y, Zeng L, Yan S, Plotnikova O, et al. 2004. Structural mechanism of the bromodomain of the coactivator cbp in p53 transcriptional activation. *Molecular Cell*. 13(2):251–63
176. Rust RR, Baldisseri DM, Weber DJ. 2000. Structure of the negative regulatory domain of p53 bound to s100b(ββ). *Nat Struct Mol Biol*. 7(7):570–74
177. Chuikov S, Kurash JK, Wilson JR, Xiao B, Justin N, et al. 2004. Regulation of p53 activity through lysine methylation. *Nature*. 432(7015):353–60
178. Poux AN, Marmorstein R. 2003. Molecular basis for gcn5/pcaf histone acetyltransferase selectivity for histone and nonhistone substrates†,‡. *Biochemistry*. 42(49):14366–74
179. Bochkareva E, Kaustov L, Ayed A, Yi G-S, Lu Y, et al. 2005. Single-stranded dna mimicry in the p53 transactivation domain interaction with replication protein a. *PNAS*. 102(43):15412–17
180. Kussie PH, Gorina S, Marechal V, Elenbaas B, Moreau J, et al. 1996. Structure of the mdm2 oncoprotein bound to the p53 tumor suppressor transactivation domain. *Science*. 274(5289):948–53
181. Di Lello P, Jenkins LMM, Jones TN, Nguyen BD, Hara T, et al. 2006. Structure of the tfb1/p53 complex: insights into the interaction between the p62/tfb1 subunit of tfiih and the activation domain of p53. *Molecular Cell*. 22(6):731–40
182. Kuszewski J, Gronenborn A, Clore M. 1999. Improving the packing and accuracy of nmr structures with a pseudopotential for the radius of gyration. *Journal of the American Chemical Society*. 121(10):2337–38
183. Cho Y, Gorina S, Jeffrey PD, Pavletich NP. 1994. Crystal structure of a p53 tumor suppressor-dna complex: understanding tumorigenic mutations. *Science*. 265(5170):346–55
184. Joo WS, Jeffrey PD, Cantor SB, Finnin MS, Livingston DM, Pavletich NP. 2002. Structure of the 53bp1 brct region bound to p53 and its comparison to the brca1 brct structure. *Genes Dev*. 16(5):583–93
185. Gorina S, Pavletich NP. 1996. Structure of the p53 tumor suppressor bound to the ankyrin and sh3 domains of 53bp2. *Science*. 274(5289):1001–5
186. Obsil T, Ghirlando R, Klein DC, Ganguly S, Dyda F. 2001. Crystal structure of the 14-3-3ζ:serotonin n-acetyltransferase complex: a role for scaffolding in enzyme regulation. *Cell*. 105(2):257–67
187. Petosa C, Masters SC, Bankston LA, Pohl J, Wang B, et al. 1998. 14-3-3ζ binds a phosphorylated raf peptide and an unphosphorylated peptide via its conserved amphipathic groove. *J. Biol. Chem.* 273(26):16305–10

188. Lilyestrom W, Klein MG, Zhang R, Joachimiak A, Chen XS. 2006. Crystal structure of sv40 large t-antigen bound to p53: interplay between a viral oncoprotein and a cellular tumor suppressor. *Genes Dev.* 20(17):2373–82
189. Dawson R, Müller L, Dehner A, Klein C, Kessler H, Buchner J. 2003. The n-terminal domain of p53 is natively unfolded. *Journal of Molecular Biology*. 332(5):1131–41
190. Lee H, Mok KH, Muhandiram R, Park K-H, Suk J-E, et al. 2000. Local structural elements in the mostly unstructured transcriptional activation domain of human p53. *J. Biol. Chem.* 275(38):29426–32
191. Jeffrey P, Gorina S, Pavletich N. 1995. Crystal structure of the tetramerization domain of the p53 tumor suppressor at 1.7 angstroms. *Science*. 267(5203):1498–1502
192. Obradovic Z, Peng K, Vucetic S, Radivojac P, Dunker AK. 2005. Exploiting heterogeneous sequence properties improves prediction of protein disorder. *Proteins: Structure, Function, and Bioinformatics*. 61(S7):176–82
193. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. 2006. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*. 7(1):208
194. Garner, Romero, Dunker, Brown, Obradovic. 1999. Predicting binding regions within disordered proteins. *Genome Inform Ser Workshop Genome Inform*. 10:41–50
195. Callaghan AJ, Aurikko JP, Ilag LL, Günter Grossmann J, Chandran V, et al. 2004. Studies of the rna degradosome-organizing domain of the escherichia coli ribonuclease rnase e. *Journal of Molecular Biology*. 340(5):965–79
196. Cheng Y, Oldfield CJ, Meng J, Romero P, Uversky VN, Dunker AK. 2007. Mining α-helix-forming molecular recognition features (α-morfs) with cross species sequence alignments. *Biochemistry*. 46(47):13468–77
197. Adkins JN, Lumb KJ. 2002. Intrinsic structural disorder and sequence features of the cell cycle inhibitor p57kip2. *Proteins: Structure, Function, and Bioinformatics*. 46(1):1–7
198. Longhi S, Receveur-Bréchot V, Karlin D, Johansson K, Darbon H, et al. 2003. The c-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the c-terminal moiety of the phosphoprotein. *J. Biol. Chem.* 278(20):18638–48
199. Karlin D, Ferron F, Canard B, Longhi S. 2003. Structural disorder and modular organization in paramyxovirinae n and p. *J Gen Virol*. 84(12):3239–52
200. Munishkina LA, Fink AL, Uversky VN. 2004. Conformational prerequisites for formation of amyloid fibrils from histones. *Journal of Molecular Biology*. 342(4):1305–24
201. Bandaru V, Cooper W, Wallace SS, Doublié S. 2004. Overproduction, crystallization and preliminary crystallographic analysis of a novel human dna-repair enzyme that recognizes oxidative dna damage. *Acta Crystallographica Section D Biological Crystallography*. 60(6):1142–44
202. Oldfield CJ, Ulrich EL, Cheng Y, Dunker AK, Markley JL. 2005. Addressing the intrinsic disorder bottleneck in structural proteomics. *Proteins: Structure, Function, and Bioinformatics*. 59(3):444–53
203. Hansen JC, Lu X, Ross ED, Woody RW. 2006. Intrinsic protein disorder, amino acid composition, and histone terminal domains. *J. Biol. Chem.* 281(4):1853–56
204. Haag Breese E, Uversky VN, Georgiadis MM, Harrington MA. 2006. The disordered amino-terminus of simpl interacts with members of the 70-kda heat-shock protein family. *DNA Cell Biol.* 25(12):704–14
205. Radivojac P, Vucetic S, O'Connor TR, Uversky VN, Obradovic Z, Dunker AK. 2006. Calmodulin signaling: analysis and prediction of a disorder-dependent molecular recognition. *Proteins: Structure, Function, and Bioinformatics*. 63(2):398–410
206. Uversky VN, Roman A, Oldfield CJ, Dunker AK. 2006. Protein intrinsic disorder and human papillomaviruses: increased amount of disorder in e6 and e7 oncoproteins from high risk hpvs. *J. Proteome Res.* 5(8):1829–42

207. Cheng Y, LeGall T, Oldfield CJ, Dunker AK, Uversky VN. 2006. Abundance of intrinsic disorder in protein associated with cardiovascular disease†. *Biochemistry*. 45(35):10448–60
208. Sigalov AB, Aivazian DA, Uversky VN, Stern LJ. 2006. Lipid-binding activity of intrinsically unstructured cytoplasmic domains of multichain immune recognition receptor signaling subunits†. *Biochemistry*. 45(51):15731–39
209. Singh VK, Zhou Y, Marsh JA, Uversky VN, Forman-Kay JD, et al. 2007. Synuclein-γ targeting peptide inhibitor that enhances sensitivity of breast cancer cells to antimicrotubule drugs. *Cancer Res*. 67(2):626–33
210. Ng KP, Potikyan G, Savene ROV, Denny CT, Uversky VN, Lee KAW. 2007. Multiple aromatic side chains within a disordered structure are critical for transcription and transforming activity of ews family oncoproteins. *PNAS*. 104(2):479–84
211. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker AK. 2007. Intrinsic disorder and functional proteomics. *Biophysical Journal*. 92(5):1439–56
212. Jones S, Thornton JM. 1995. Protein-protein interactions: a review of protein dimer structures. *Progress in Biophysics and Molecular Biology*. 63(1):31–59
213. Jones S, Thornton JM. 1996. Principles of protein-protein interactions. *Proc Natl Acad Sci U S A*. 93(1):13–20
214. Jones S, Thornton JM. 1997. Analysis of protein-protein interaction sites using surface patches. *Journal of Molecular Biology*. 272(1):121–32
215. Jones S, Thornton JM. 1997. Prediction of protein-protein interaction sites using patch analysis. *Journal of Molecular Biology*. 272(1):133–43
216. Larsen TA, Olson AJ, Goodsell DS. 1998. Morphology of protein–protein interfaces. *Structure*. 6(4):421–27
217. Lo Conte L, Chothia C, Janin J. 1999. The atomic structure of protein-protein recognition sites. *Journal of Molecular Biology*. 285(5):2177–98
218. Smith GR, Sternberg MJE, Bates PA. 2005. The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. *Journal of Molecular Biology*. 347(5):1077–1101
219. Ho WC, Luo C, Zhao K, Chai X, Fitzgerald MX, Marmorstein R. 2006. High-resolution structure of the p53 core domain: implications for binding small-molecule stabilizing compounds. *Acta Crystallographica Section D Biological Crystallography*. 62(12):1484–93
220. Williams RS, Green R, Glover JNM. 2001. Crystal structure of the brct repeat region from the breast cancer-associated protein brca1. *Nat Struct Mol Biol*. 8(10):838–42
221. Ferreon JC, Volk DE, Luxon BA, Gorenstein DG, Hilser VJ. 2003. Solution structure, dynamics, and thermodynamics of the native state ensemble of the sem-5 c-terminal sh3 domain†. *Biochemistry*. 42(19):5582–91
222. Uhrinova S, Uhrin D, Powers H, Watt K, Zheleva D, et al. 2005. Structure of free mdm2 n-terminal domain reveals conformational adjustments that accompany p53-binding. *Journal of Molecular Biology*. 350(3):587–98
223. Rojas JR, Trievel RC, Zhou J, Mo Y, Li X, et al. 1999. Structure of tetrahymena gcn5 bound to coenzyme a and a histone h3 peptide. *Nature*. 401(6748):93–98
224. Jeffrey PD, Russo AA, Polyak K, Gibbs E, Hurwitz J, et al. 1995. Mechanism of cdk activation revealed by the structure of a cyclina-cdk2 complex. *Nature*. 376(6538):313–20
225. Avalos JL, Boeke JD, Wolberger C. 2004. Structural basis for the mechanism and regulation of sir2 enzymes. *Molecular Cell*. 13(5):639–48
226. Sachchidanand, Resnick-Silverman L, Yan S, Mutjaba S, Liu W, et al. 2006. Target structure-based discovery of small molecules that block human p53 and creb binding protein association. *Chemistry & Biology*. 13(1):81–90

227. Smith SP, Shaw GS. 1998. A novel calcium-sensitive switch revealed by the structure of human s100b in the calcium-bound form. *Structure*. 6(2):211–22
228. Kwon T, Chang JH, Kwak E, Lee CW, Joachimiak A, et al. 2003. Mechanism of histone lysine methyl transfer revealed by the structure of set7/9–adomet. *EMBO J*. 22(2):292–303
229. Christopher J. Oldfield JM. 2007. Intrinsic disorder in protein-protein interaction networks: case studies of complexes involving p53 and 14-3-3. , pp. 553–66
230. Macdonald N, Welburn JPI, Noble MEM, Nguyen A, Yaffe MB, et al. 2005. Molecular basis for the recognition of phosphorylated and phosphoacetylated histone h3 by 14-3-3. *Molecular Cell*. 20(2):199–211
231. Bourhis J-M, Johansson K, Receveur-Bréchot V, Oldfield CJ, Dunker KA, et al. 2004. The c-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner. *Virus Research*. 99(2):157–67
232. Kingston RL, Hamel DJ, Gay LS, Dahlquist FW, Matthews BW. 2004. Structural basis for the attachment of a paramyxoviral polymerase to its template. *PNAS*. 101(22):8301–6
233. Chandran V, Luisi BF. 2006. Recognition of enolase in the escherichia coli rna degradosome. *Journal of Molecular Biology*. 358(1):8–15
234. Puntervoll P, Linding R, Gemünd C, Chabanis-Davidson S, Mattingsdal M, et al. 2003. Elm server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucl. Acids Res*. 31(13):3625–30
235. Neduva V, Linding R, Su-Angrand I, Stark A, Masi F de, et al. 2005. Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol*. 3(12):e405
236. Neduva V, Russell RB. 2005. Linear motifs: evolutionary interaction switches. *FEBS Letters*. 579(15):3342–45
237. Fuxreiter M, Tompa P, Simon I. 2007. Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*. 23(8):950 –956
238. Dunker AK. 2007. Another window into disordered protein function. *Structure*. 15(9):1026–28
239. Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 22(12):2577–2637
240. Minor DL, Kim PS. 1996. Context-dependent secondary structure formation of a designed protein sequence. *Nature*. 380(6576):730–34
241. Jacoboni I, Martelli PL, Fariselli P, Compiani M, Casadio R. 2000. Predictions of protein segments with the same aminoacid sequence and different secondary structure: a benchmark for predictive methods. *Proteins: Structure, Function, and Bioinformatics*. 41(4):535–44
242. Mezei M. 1998. Chameleon sequences in the pdb. *Protein Eng*. 11(6):411–14
243. Smith CA, Calabro V, Frankel AD. 2000. An rna-binding chameleon. *Molecular Cell*. 6(5):1067–76
244. Yoon S, Jung H. 2006. Analysis of chameleon sequences by energy decomposition on a pairwise per-residue basis. *Protein J*. 25(5):361–68
245. Guo J-T, Jaromczyk JW, Xu Y. 2007. Analysis of chameleon sequences and their implications in biological processes. *Proteins: Structure, Function, and Bioinformatics*. 67(3):548–58
246. Takano K, Katagiri Y, Mukaiyama A, Chon H, Matsumura H, et al. 2007. Conformational contagion in a protein: structural properties of a chameleon sequence. *Proteins: Structure, Function, and Bioinformatics*. 68(3):617–25
247. Bullock AN, Henckel J, Fersht AR. 2000. Quantitative analysis of residual folding and dna binding in mutant p53 core domain: definition of mutant states for rescue in cancer therapy. *Oncogene*. 19(10):1245–56

248. Erlanson DA, Wells JA, Braisted AC. 2004. Tethering: fragment-based drug discovery. *Annual Review of Biophysics and Biomolecular Structure*. 33(1):199–223
249. Li, Romero, Rani, Dunker, Obradovic. 1999. Predicting protein disorder for n-, c-, and internal regions. *Genome Inform Ser Workshop Genome Inform*. 10:30–40
250. Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK, Obradovic Z. 2005. Optimizing long intrinsic disorder predictors with protein evolutionary information. *J Bioinform Comput Biol*. 3(1):35–60
251. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. 1997. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucl. Acids Res.* 25(17):3389–3402
252. Rost B, Sander C, Schneider R. 1994. Phd--an automatic mail server for protein secondary structure prediction. *Comput. Appl. Biosci.* 10(1):53–60
253. Jones DT, Ward JJ. 2003. Prediction of disordered regions in proteins from position specific score matrices. *Proteins: Structure, Function, and Bioinformatics*. 53(S6):573–78
254. Eisenhaber F, Lijnzaad P, Argos P, Sander C, Scharf M. 1995. The double cubic lattice method: efficient approaches to numerical integration of surface area and volume and to dot surface contouring of molecular assemblies. *J. Comput. Chem.* 16(3):273–84
255. Kohlbacher O, Lenhof H-P. 2000. Ball—rapid software prototyping in computational molecular biology. *Bioinformatics*. 16(9):815–24
256. Shatsky M, Nussinov R, Wolfson HJ. 2004. A method for simultaneous alignment of multiple protein structures. *Proteins: Structure, Function, and Bioinformatics*. 56(1):143–56
257. Chothia C. 1975. Structural invariants in protein folding. *Nature*. 254(5498):304–8
258. Ma B, Elkayam T, Wolfson H, Nussinov R. 2003. Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *PNAS*. 100(10):5772–77
259. Burley SK. 2000. An overview of structural genomics. *Nature Structural Biology*. 7 Suppl 1(11):932–34
260. Chance MR, Bresnick AR, Burley SK, Jiang J-S, Lima CD, et al. 2002. Structural genomics: a pipeline for providing structures for the biologist. *Protein Sci*. 11(4):723–38
261. Fox BG, Goulding C, Malkowski MG, Stewart L, Deacon A. 2008. Structural genomics: from genes to structures with valuable materials and many questions in between. *Nature Methods*. 5(2):129–32
262. Dessailly BH, Nair R, Jaroszewski L, Fajardo JE, Kouranov A, et al. 2009. Psi-2: structural genomics to cover protein domain family space. *Structure*. 17(6):869–81
263. Chandonia J-M, Brenner SE. 2006. The impact of structural genomics: expectations and outcomes. *Science*. 311(5759):347–51
264. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, et al. 2001. Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling*. 19(1):26–59
265. Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. 2000. Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform*. 11:161–71
266. Uversky VN. 2010. The mysterious unfoldome: structureless, underappreciated, yet vital part of any given proteome. *Journal of Biomedicine and Biotechnology*. 2010:1–15
267. Esnouf RM, Hamer R, Sussman JL, Silman I, Trudgian D, et al. 2006. Honing the in silico toolkit for detecting protein disorder. *Acta Crystallogr D Biol Crystallogr*. 62(Pt 10):1260–66
268. Bourhis J, Canard B, Longhi S. 2007. Predicting protein disorder and induced folding: from theoretical principles to practical applications. *Current Protein & Peptide Science*. 8(2):135–49
269. Dosztanyi Z, Sandor M, Tompa P, Simon I. 2007. Prediction of protein disorder at the domain level. *Current Protein & Peptide Science*. 8(2):161–71

270. Dosztányi Z, Tompa P. 2008. Prediction of protein disorder. In *Structural Proteomics*, ed B Kobe, M Guss, T Huber, pp. 103–15. Humana Press
271. He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK. 2009. Predicting intrinsic disorder in proteins: an overview. *Cell Research*. 19(8):929–49
272. Reeves R, Nissen MS. 1999. Purification and assays for high mobility group hmg-i(y) protein function. In *Methods in Enzymology*, ed APW Paul M. Wassarman. Volume 304:155–88. Academic Press
273. Stewart AA, Ingebritsen TS, Cohen P. 1983. The protein phosphatases involved in cellular regulation. *European Journal of Biochemistry*. 132(2):289–95
274. Petros AM, Medek A, Nettesheim DG, Kim DH, Yoon HS, et al. 2001. Solution structure of the antiapoptotic protein bcl-2. *PNAS*. 98(6):3012–17
275. Harauz G, Ishiyama N, Hill CM., Bates IR, Libich DS, Farès C. 2004. Myelin basic protein—diverse conformational states of an intrinsically unstructured protein and its roles in myelin assembly and multiple sclerosis. *Micron*. 35(7):503–42
276. Cary PD, King DS, Crane-Robinson C, Bradbury EM, Rabbani A, et al. 1980. Structural studies on two high-mobility-group proteins from calf thymus, hmg-14 and hmg-20 (ubiquitin), and their interaction with dna. *European Journal of Biochemistry*. 112(3):577–80
277. Daughdrill GW, Chadsey MS, Karlinsey JE, Hughes KT, Dahlquist FW. 1997. The c-terminal half of the anti-sigma factor, flgm, becomes structured when bound to its target, σ28. *Nature Structural & Molecular Biology*. 4(4):285–91
278. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. 2000. The protein data bank. *Nucleic Acids Res*. 28(1):235–42
279. Jurnak F. 1985. Structure of the gdp domain of ef-tu and location of the amino acids homologous to ras oncogene proteins. *Science*. 230(4721):32–36
280. Hobohm U, Scharf M, Schneider R, Sander C. 1992. Selection of representative protein data sets. *Protein Sci*. 1(3):409–17
281. Chen L, Oughtred R, Berman HM, Westbrook J. 2004. Targetdb: a target registration database for structural genomics projects. *Bioinformatics*. 20(16):2860–62
282. Li X, Obradovic Z, Brown CJ, Garner EC, Dunker AK. 2000. Comparing predictors of disordered protein. *Genome Inform Ser Workshop Genome Inform*. 11:172–84
283. Garner, Cannon, Romero, Obradovic, Dunker. 1998. Predicting disordered regions from amino acid sequence: common themes despite differing structural characterization. *Genome Inform Ser Workshop Genome Inform*. 9:201–13
284. Romero P, Obradovic Z, Kissinger C, Villafranca JE, Dunker AK. 1997. Identifying disordered regions in proteins from amino acid sequence
285. Vucetic S, Radivojac P, Obradovic Z, Brown CJ, Dunker AK. 2001. Methods for improving protein disorder prediction
286. Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN. 2010. Pondr-fit: a meta-predictor of intrinsically disordered amino acids. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*. 1804(4):996–1010
287. Sing T, Sander O, Beerenwinkel N, Lengauer T. 2005. Rocr: visualizing classifier performance in r. *Bioinformatics*. 21(20):3940–41
288. Yang Z, Kollman JM, Pandi L, Doolittle RF. 2001. Crystal structure of native chicken fibrinogen at 2.7 å resolution†,‡. *Biochemistry*. 40(42):12515–23
289. Jiang X, Gurel O, Mendiaz EA, Stearns GW, Clogston CL, et al. 2000. Structure of the active core of human stem cell factor and analysis of binding to its receptor kit. *EMBO J*. 19(13):3192–3203
290. Daumke O, Weyand M, Chakrabarti PP, Vetter IR, Wittinghofer A. 2004. The gtpase-activating protein rap1gap uses a catalytic asparagine. *Nature*. 429(6988):197–201

291. Shaw N, Zhao M, Cheng C, Xu H, Saarikettu J, et al. 2007. The multifunctional human p100 protein “hooks” methylated ligands. *Nat Struct Mol Biol*. 14(8):779–84
292. Deniziak M, Sauter C, Becker HD, Paulus CA, Giegé R, Kern D. 2007. Deinococcus glutaminyl-trna synthetase is a chimer between proteins from an ancient and the modern pathways of aminoacyl-trna formation. *Nucl. Acids Res.* 35(5):1421–31
293. Igonet S, Vulliez-Le Normand B, Faure G, Riottot M-M, Kocken CHM, et al. 2007. Cross-reactivity studies of an anti-plasmodium vivax apical membrane antigen 1 monoclonal antibody: binding and structural characterisation. *Journal of Molecular Biology*. 366(5):1523–37
294. Macheboeuf P, Fischer DS, Brown T, Zervosen A, Luxen A, et al. 2007. Structural and mechanistic basis of penicillin-binding protein inhibition by lactivicins. *Nat Chem Biol*. 3(9):565–69
295. Macheboeuf P, Guilmi AMD, Job V, Vernet T, Dideberg O, Dessen A. 2005. Active site restructuring regulates ligand recognition in class a penicillin-binding proteins. *PNAS*. 102(3):577–82
296. Healy S, McDonald MK, Wu X, Yue WW, Kochan G, et al. 2010. Structural impact of human and escherichia coli biotin carboxyl carrier proteins on biotin attachment. *Biochemistry*. 49(22):4687–94
297. Mouilleron S, Badet-Denisot M-A, Golinelli-Pimpaneau B. 2008. Ordering of c-terminal loop and glutaminase domains of glucosamine-6-phosphate synthase promotes sugar ring opening and formation of the ammonia channel. *Journal of Molecular Biology*. 377(4):1174–85
298. Assenberg R, Delmas O, Ren J, Vidalain P-O, Verma A, et al. 2010. Structure of the nucleoprotein binding domain of mokola virus phosphoprotein. *J. Virol.* 84(2):1089–96
299. Chen G, Wang H, Robinson H, Cai J, Wan Y, Ke H. 2008. An insight into the pharmacophores of phosphodiesterase-5 inhibitors from synthetic and crystal structural studies. *Biochemical Pharmacology*. 75(9):1717–28
300. Pei XY, Titman CM, Frank RAW, Leeper FJ, Luisi BF. 2008. Snapshots of catalysis in the e1 subunit of the pyruvate dehydrogenase multienzyme complex. *Structure*. 16(12):1860–72
301. Wang H, Chumnarnsilpa S, Loonchanta A, Li Q, Kuan Y-M, et al. 2009. Helix straightening as an activation mechanism in the gelsolin superfamily of actin regulatory proteins. *J. Biol. Chem.* 284(32):21265–69
302. Kollman JM, Pandi L, Sawaya MR, Riley M, Doolittle RF. 2009. Crystal structure of human fibrinogen. *Biochemistry*. 48(18):3877–86
303. Lerch TF, Xie Q, Chapman MS. 2010. The structure of adeno-associated virus serotype 3b (aav-3b): insights into receptor binding and immune evasion. *Virology*. 403(1):26–36
304. Lovering AL, Lin LY-C, Sewell EW, Spreter T, Brown ED, Strynadka NCJ. 2010. Structure of the bacterial teichoic acid polymerase tagf provides insights into membrane association and catalysis. *Nat Struct Mol Biol*. 17(5):582–89
305. Jacques DA, Langley DB, Kuramitsu S, Yokoyama S, Trewhella J, Guss JM. 2011. The structure of ttha0988 from thermus thermophilus , a kipi–kipa homologue incorrectly annotated as an allophanate hydrolase. *Acta Crystallographica Section D Biological Crystallography*. 67(2):105–11
306. Chen F, Venugopal V, Murray B, Rudenko G. 2011. The structure of neurexin 1α reveals features promoting a role as synaptic organizer. *Structure*. 19(6):779–89
307. Perriches T, Singleton MR. 2012. Structure of yeast kinetochore ndc10 dna-binding domain reveals unexpected evolutionary relationship to tyrosine recombinases. *J. Biol. Chem.* 287(7):5173–79

308. Baranova E, Fronzes R, Garcia-Pino A, Van Gerven N, Papapostolou D, et al. 2012. Sbsb structure and lattice reconstruction unveil ca2+ triggered s-layer assembly. *Nature*. advance online publication:
309. Santner AA, Croy CH, Vasanwala FH, Uversky VN, Van Y-YJ, Dunker AK. 2012. Sweeping away protein aggregation with entropic bristles: intrinsically disordered protein fusions enhance soluble expression. *Biochemistry*. 51(37):7250–62
310. Fontana A, Fassina G, Vita C, Dalzoppo D, Zamai M, Zambonin M. 1986. Correlation between sites of limited proteolysis and segmental mobility in thermolysin. *Biochemistry*. 25(8):1847–51
311. Fontana A, Zambonin M, Polverino de Laureto P, De Filippis V, Clementi A, Scaramella E. 1997. Probing the conformational state of apomyoglobin by limited proteolysis. *Journal of Molecular Biology*. 266(2):223–30
312. Spolaore B, Bermejo R, Zambonin M, Fontana A. 2001. Protein interactions leading to conformational changes monitored by limited proteolysis: apo form and fragments of horse cytochrome c. *Biochemistry*. 40(32):9460–68
313. Receveur-Bréchot V, Bourhis J-M, Uversky VN, Canard B, Longhi S. 2006. Assessing protein disorder and induced folding. *Proteins: Structure, Function, and Bioinformatics*. 62(1):24–45
314. Sedzik J, Kirschner DA. 1992. Is myelin basic protein crystallizable? *Neurochem. Res.* 17(2):157–66
315. Harauz G, Ladizhansky V, Boggs JM. 2009. Structural polymorphism and multifunctionality of myelin basic protein. *Biochemistry*. 48(34):8094–8104
316. Corollo D, Blair-Johnson M, Conrad J, Fiedler T, Sun D, et al. 1999. Crystallization and characterization of a fragment of pseudouridine synthase rluc from escherichia coli. *Acta Crystallographica Section D Biological Crystallography*. 55(1):302–4
317. Orth P, Jekow P, Alonso JC, Hinrichs W. 1999. Proteolytic cleavage of gram-positive β recombinase is required for crystallization. *Protein Eng.* 12(5):371–73
318. Bracken C. 2001. Nmr spin relaxation methods for characterization of disorder and folding in proteins. *Journal of Molecular Graphics and Modelling*. 19(1):3–12
319. Hammoudeh DI, Follis AV, Prochownik EV, Metallo SJ. 2009. Multiple independent binding sites for small-molecule inhibitors on the oncoprotein c-myc. *J. Am. Chem. Soc.* 131(21):7390–7401
320. Hurley TD, Yang J, Zhang L, Goodwin KD, Zou Q, et al. 2007. Structural basis for regulation of protein phosphatase 1 by inhibitor-2. *J. Biol. Chem.* 282(39):28874–83
321. Huang HB, Chen YC, Tsai LH, Wang H, Lin FM, et al. 2000. Backbone 1h, 15n, and 13c resonance assignments of inhibitor-2 -- a protein inhibitor of protein phosphatase-1. *J. Biomol. NMR*. 17(4):359–60
322. Marsh JA, Dancheck B, Ragusa MJ, Allaire M, Forman-Kay JD, Peti W. 2010. Structural diversity in free and bound states of intrinsically disordered protein phosphatase 1 regulators. *Structure*. 18(9):1094–1103
323. Uversky VN, Longhi S, eds. 2010. *Instrumental Analysis of Intrinsically Disordered Proteins: Assessing Structure And Conformation*
324. Dyson HJ, Wright PE. 2005. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*. 6(3):197–208
325. Tompa P. 2002. Intrinsically unstructured proteins. *Trends in Biochemical Sciences.* 27(10):527–33
326. Zhu H, Snyder M. 2001. Protein arrays and microarrays. *Current Opinion in Chemical Biology*. 5(1):40–45
327. Mészáros B, Simon I, Dosztányi Z. 2009. Prediction of protein binding regions in disordered proteins. *PLoS Comput Biol*. 5(5):e1000376

328. Mooney C, Pollastri G, Shields DC, Haslam NJ. 2012. Prediction of short linear protein binding regions. *Journal of Molecular Biology*. 415(1):193–204
329. Balla S, Thapar V, Verma S, Luong T, Faghri T, et al. 2006. Minimotif miner: a tool for investigating protein function. *Nat Meth*. 3(3):175–77
330. Mi T, Merlin JC, Deverasetty S, Gryk MR, Bill TJ, et al. 2012. Minimotif miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences. *Nucleic Acids Res*. 40(D1):D252–D260
331. Dinkel H, Michael S, Weatheritt RJ, Davey NE, Roey KV, et al. 2011. Elm—the database of eukaryotic linear motifs. *Nucl. Acids Res.*
332. Letunic I, Doerks T, Bork P. 2011. Smart 7: recent updates to the protein domain annotation resource. *Nucleic Acids Research*. 40(D1):D302–D305
333. Dosztányi Z, Csizmók V, Tompa P, Simon I. 2005. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol*. 347(4):827–39
334. Rajasekaran S, Mi T, Merlin JC, Oommen A, Gradie P, Schiller MR. 2010. Partitioning of minimotifs based on function with improved prediction accuracy. *PLoS One*. 5(8):
335. Rajasekaran S, Merlin JC, Kundeti V, Mi T, Oommen A, et al. 2011. A computational tool for identifying minimotifs in protein–protein interactions and improving the accuracy of minimotif predictions. *Proteins: Structure, Function, and Bioinformatics*. 79(1):153–64
336. Mészáros B, Dosztányi Z, Simon I. 2012. Disordered binding regions and linear motifs—bridging the gap between two models of molecular recognition. *PLoS ONE*. 7(10):e46829
337. Mészáros B, Tompa P, Simon I, Dosztányi Z. 2007. Molecular principles of the interactions of disordered proteins. *Journal of Molecular Biology*. 372(2):549–61
338. Brannetti B, Via A, Cestra G, Cesareni G, Citterich MH. 2000. Sh3-spot: an algorithm to predict preferred ligands to different members of the sh3 gene family. *Journal of Molecular Biology*. 298(2):313–28
339. Yip KY, Utz L, Sitwell S, Hu X, Sidhu SS, et al. 2011. Identification of specificity determining residues in peptide recognition domains using an information theoretic approach applied to large-scale binding maps. *BMC Biol*. 9:53

CURRICULUM VITAE

Christopher John Oldfield

Education

PhD in Bioinformatics, Indiana University, Indianapolis, 2005-2014
Graduate coursework in Biophysics, University of Wisconsin, Madison, 2003-2005
BS in Biochemistry, Washington State University, Pullman, 1997-2003

Experience

Research Assistant, Indiana University, Indianapolis, 2005-2012
Research Assistant, University of Wisconsin, Madison, 2003-2005
Researcher, Molecular Kinetics, Inc., Pullman, WA, 2001-2003
Undergraduate Research Assistant, Pullman, WA, 1999-2001

Scholarships and Conferences

Scholarships: Computation and Informatics in Biology and Medicine Predoctoral Trainee (NLM funded, September 2003 to May 2005), Eli Lilly IUPUI Informatics Fellowship (September 2008 to May 2010)
Presentations: Biophysical Society (2009), GRC Intrinsically Disordered Proteins (2010)
Posters: Biophysical Society (2008), Pacific Symposium on Biocomputing (2009)

Primary Publications

1. **Oldfield, C.J.**, B. Xue, Y.Y. Van, E.L. Ulrich, J.L. Markley, A.K. Dunker, V.N. Uversky. (2013) *Utilization of protein intrinsic disorder knowledge in structural proteomics.* Biochim Biophys Acta. 1834: 487-98.
2. **Oldfield, C.J**., J. Meng, J.Y. Yang, M.Q. Yang, V.N. Uversky, and A.K. Dunker. (2008) *Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners.* BMC Genomics 9 Suppl 1: S1.
3. **Oldfield, C.J.**, E.L. Ulrich, Y. Cheng, A.K. Dunker, and J.L. Markley. (2005) *Addressing the intrinsic disorder bottleneck in structural proteomics.* Proteins 59: 444-53.
4. **Oldfield, C.J.**, Y. Cheng, M.S. Cortese, P. Romero, V.N. Uversky, and A.K. Dunker. (2005) *Coupled folding and binding with alpha-helix-forming molecular recognition elements.* Biochemistry 44: 12454-70.
5. **Oldfield, C.J.**, Y. Cheng, M.S. Cortese, C.J. Brown, V.N. Uversky, and A.K. Dunker. (2005) *Comparing and combining predictors of mostly disordered proteins.* Biochemistry 44: 1989-2000.

Contributing Publications

6. Hsu, W.L., **C.J. Oldfield**, B. Xue, J. Meng, F. Huang, P. Romero, V.N. Uversky, A.K. Dunker. (2013) *Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding.* Protein Sci. 22: 258-73.
7. Xue, B., **C.J. Oldfield**, Y.Y. Van, A.K. Dunker, V.N. Uversky. (2012) *Protein intrinsic disorder and induced pluripotent stem cells.* Mol Biosyst. 8: 134-50.

8. Hsu, W.L., **C.J. Oldfield**, J. Meng, F. Huang, B. Xue, V.N. Uversky, P. Romero, A.K. Dunker. (2012) *Intrinsic protein disorder and protein-protein interactions.* Pac Symp Biocomput. 2012: 116-27.
9. Huang, F., **C.J. Oldfield**, J. Meng, W.L. Hsu, B. Xue, V.N. Uversky, P. Romero, A.K. Dunker. (2012) *Subclassifying disordered proteins by the ch-cdf plot method.* Pac Symp Biocomput. 2012: 128-39.
10. Johnson, D.E., B. Xue, M.D. Sickmeier, J. Meng, M.S. Cortese, **C.J. Oldfield**, T. Le Gall, A.K. Dunker, V.N. Uversky. (2012) *High-throughput characterization of intrinsic disorder in proteins from the Protein Structure Initiative*. J Struct Biol. 180: 201-15.
11. Disfani, F.M., W.L. Hsu, M.J. Mizianty, **C.J. Oldfield**, B. Xue, A.K. Dunker, V.N. Uversky, L. Kurgan. (2012) *MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins.* Bioinformatics. 28:i75-83.
12. Vacic, V., P.R. Markwick, **C.J. Oldfield**, X. Zhao, C. Haynes, V.N. Uversky, L.M. Iakoucheva. (2012) *Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder*. PLoS Comput Biol. 8:e1002709
13. Sun X., W.T. Jones, D. Harvey, P.J. Edwards, S.M. Pascal, C. Kirk, T. Considine, D.J. Sheerin, J. Rakonjac, **C.J. Oldfield**, B. Xue, A.K. Dunker, V.N. Uversky. (2010) *N-terminal domains of DELLA proteins are intrinsically unstructured in the absence of interaction with GID1/gibberellic acid receptors*. J Biol Chem. 285: 11557-71.
14. Xue B., R.W. Williams, **C.J. Oldfield**, G.K. Goh, A.K. Dunker, V.N. Uversky. (2010) *Viral disorder or disordered viruses: do viral proteins possess unique features?* Protein Pept Lett. 17: 932-51.
15. Xue B., R.W. Williams, **C.J. Oldfield**, A.K. Dunker, V.N. Uversky. (2010) *Archaic chaos: intrinsically disordered proteins in Archaea*. BMC Syst Biol. 4 Suppl 1: S1.
16. Midic U., **C.J. Oldfield**, A.K. Dunker, Z. Obradovic, V.N. Uversky. (2009) *Unfoldomics of human genetic diseases: illustrative examples of ordered and intrinsically disordered members of the human diseasome*. Protein Pept Lett. 16: 1533-47.
17. Uversky V.N., **C.J. Oldfield**, U. Midic, H. Xie, B. Xue, S. Vucetic, L.M. Iakoucheva, Z. Obradovic, A.K. Dunker. (2009) *Unfoldomics of human diseases: linking protein intrinsic disorder with diseases*. BMC Genomics. 10 Suppl 1: S7.
18. Midic U., C.J. Oldfield, A.K. Dunker, Z. Obradovic, V.N. Uversky. (2009) *Protein disorder in the human diseasome: unfoldomics of human genetic diseases.* BMC Genomics. 10 Suppl 1: S12.
19. Xue B., **C.J. Oldfield**, A.K. Dunker, V.N. Uversky. (2009) *CDF it all: consensus prediction of intrinsically disordered proteins based on various cumulative distribution functions.* FEBS Lett. 583: 1469-74.
20. Tompa P., M. Fuxreiter, **C.J. Oldfield**, I. Simon, A.K. Dunker, V.N. Uversky. (2009) *Close encounters of the third kind: disordered domains and the interactions of proteins.* Bioessays. 31: 328-35.
21. Tóth-Petróczy A., **C.J. Oldfield**, I. Simon, Y. Takagi, A.K. Dunker, V.N. Uversky, M. Fuxreiter. (2008) *Malleable machines in transcription regulation: the mediator complex.* PLoS Comput Biol. 4: e1000243.
22. Dunker A.K., **C.J. Oldfield**, J. Meng, P. Romero, J.Y. Yang, J.W. Chen, V. Vacic, Z. Obradovic, V.N. Uversky. (2008) *The unfoldomics decade: an update on intrinsically disordered proteins.* BMC Genomics. 9 Suppl 2: S1.
23. Uversky V.N., **C.J. Oldfield**, A.K. Dunker. (2008) *Intrinsically disordered proteins in human diseases: introducing the D2 concept.* Annu Rev Biophys. 37: 215-46.
24. Xie, H., S. Vucetic, L.M. Iakoucheva, **C.J. Oldfield**, A.K. Dunker, V.N. Uversky, and Z. Obradovic. (2007) *Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions.* J Proteome Res. 6: 1882-98.

25. Vucetic, S., H. Xie, L.M. Iakoucheva, **C.J. Oldfield**, A.K. Dunker, Z. Obradovic, and V.N. Uversky. (2007) *Functional anthology of intrinsic disorder. 2. Cellular components, domains, technical terms, developmental processes, and coding sequence diversities correlated with long disordered regions.* J Proteome Res. 6: 1899-916.
26. Xie, H., S. Vucetic, L.M. Iakoucheva, **C.J. Oldfield**, A.K. Dunker, Z. Obradovic, and V.N. Uversky. (2007) *Functional anthology of intrinsic disorder. 3. Ligands, post-translational modifications, and diseases associated with intrinsically disordered proteins.* J Proteome Res. 6: 1917-32.
27. Vacic, V., **C.J. Oldfield**, A. Mohan, P. Radivojac, M.S. Cortese, V.N. Uversky, and A.K. Dunker. (2007) *Characterization of molecular recognition features, MoRFs, and their binding partners.* J Proteome Res. 6: 2351-66.
28. Radivojac, P., L.M. Iakoucheva, **C.J. Oldfield**, Z. Obradovic, V.N. Uversky, and A.K. Dunker. (2007) *Intrinsic disorder and functional proteomics.* Biophys J. 92: 1439-56.
29. Cheng, Y., **C.J. Oldfield**, J. Meng, P. Romero, V.N. Uversky, and A.K. Dunker. (2007) *Mining alpha-helix-forming molecular recognition features with cross species sequence alignments.* Biochemistry 46: 13468-77.
30. Haynes, C., **C.J. Oldfield**, F. Ji, N. Klitgord, M.E. Cusick, P. Radivojac, V.N. Uversky, M. Vidal, L.M. Iakoucheva. *Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes.* (2006) PLoS Comput Biol. 2: e100.
31. Uversky, V.N., A. Roman, **C.J. Oldfield**, and A.K. Dunker. (2006) *Protein intrinsic disorder and human papillomaviruses: increased amount of disorder in E6 and E7 oncoproteins from high risk HPVs.* J Proteome Res. 5: 1829-42.
32. Romero, P.R., S. Zaidi, Y.Y. Fang, V.N. Uversky, P. Radivojac, **C.J. Oldfield**, M.S. Cortese, M. Sickmeier, T. LeGall, Z. Obradovic, and A.K. Dunker. (2006) *Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms.* PNAS 103: 8390-5.
33. Mohan, A., **C.J. Oldfield**, P. Radivojac, V. Vacic, M.S. Cortese, A.K. Dunker, and V.N. Uversky. (2006) *Analysis of molecular recognition features.* J Mol Biol. 362: 1043-59.
34. Liu, J., N.B. Perumal, **C.J. Oldfield**, E.W. Su, V.N. Uversky, and A.K. Dunker. (2006) *Intrinsic disorder in transcription factors.* Biochemistry 45: 6873-88.
35. Cheng, Y., T. LeGall, **C.J. Oldfield**, J.P. Mueller, Y.Y. Van, P. Romero, M.S. Cortese, V.N. Uversky, and A.K. Dunker. (2006) *Rational drug design via intrinsically disordered protein.* Trends Biotechnol. 24: 435-42.
36. Cheng, Y., T. LeGall, **C.J. Oldfield**, A.K. Dunker, and V.N. Uversky. (2006) *Abundance of intrinsic disorder in protein associated with cardiovascular disease.* Biochemistry 45: 10448-60.
37. Uversky, V.N., **C.J. Oldfield**, and A.K. Dunker. (2005) *Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling.* J Mol Recognit. 18: 343-84.
38. Bourhis, J.M., K. Johansson, V. Receveur-Brechot, **C.J. Oldfield**, K.A. Dunker, B. Canard, and S. Longhi. (2004) *The C-terminal domain of measles virus nucleoprotein belongs to the class of intrinsically disordered proteins that fold upon binding to their physiological partner.* Virus Res. 99: 157-67.
39. Brown, C.J., S. Takayama, A.M. Campen, P. Vise, T.W. Marshall, **C.J. Oldfield**, C.J. Williams, and A.K. Dunker. (2002) *Evolutionary rate heterogeneity in proteins with long disordered regions.* J Mol Evol. 55: 104-10.
40. Dunker, A.K., J.D. Lawson, C.J. Brown, R.M. Williams, P. Romero, J.S. Oh, **C.J. Oldfield**, A.M. Campen, C.M. Ratliff, K.W. Hipps, J. Ausio, M.S. Nissen, R. Reeves, C. Kang, C.R. Kissinger, R.W. Bailey, M.D. Griswold, W. Chiu, E.C. Garner, and Z. Obradovic. (2001) *Intrinsically disordered protein.* J Mol Graph Model. 19: 26-59.