ELECTROSTATICS IN INTRINSICALLY DISORDERED PROTEINS

by

Eric Tsz Chung Wong

B.Sc., The University of British Columbia, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Biochemistry and Molecular Biology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

October 2012

Abstract

Protein-protein interactions are fundamental to many biological processes. A large proportion of proteins have been identified as partially or entirely disordered in their native state. Disordered protein regions appear to be especially adept for facilitating interactions and are often important for signaling and regulation. This is a relatively young field compared to the extensively studied structured proteins, which means the data available is not as abundant and well characterized. Consequently, the first goal of this thesis is to use a structure-based method to identify intrinsically disordered protein (IDP) complexes. We hypothesize that we can take advantage of the tendency of intrinsically disordered segments to adopt extended structures when bound to globular proteins to separate them from natively structured proteins. Radius of gyration is used to measure the extendedness of protein chains in protein data bank structures. The dataset of selected extended protein chains is then verified to be enriched in disorder through sequence based prediction. The second goal of the thesis is to characterize the identified IDP complexes. This involved an attempt to analyze the compaction of intrinsically disordered segments upon binding and analysis of the interface residue composition. Analysis of the interface residue composition revealed an interesting disparity between the residue compositions across the IDP complex interface. An enrichment of hydrophobic and depletion of charged residues on the IDP side of the interface seems to suggest that IDP interactions are relatively unspecific due to strong hydrophobic contributions. Their importance in signaling and regulation as well as studies that suggested highly specific interactions led us to investigate the polar interactions of IDP complexes. Computational alanine scanning and continuum electrostatic calculations on

IDP complexes reveal a class of protein-protein interaction that has high electrostatic complementarity in comparison to structured protein complexes. Charged residues on the IDP interface make greater contribution to binding compared to those of structured proteins. Furthermore, large desolvation penalties and Coulombic interactions balance out to contribute to highly specific yet low affinity interactions.

Preface

The majority of the research and results presented in this thesis are under preparation for publication. The title of the article is “Electrostatics in IDP complexes”. The conceptualization of the project is accredited to my supervisor Jörg Gsponer. I established all the methods and performed all the experiments under the guidance of my supervisor. Jörg Gsponer contributed a significant portion of writing to the manuscript. With the exception of the figures, only a minor portion of the writing from the manuscript is incorporated into this thesis.

Table of Contents

Abstract ii

Preface iv

Table of Contents v

List of Tables vii

List of Figures viii

List of Abbreviations ix

Acknowledgement x

Dedication xi

1 Introduction 1

1.1 Intrinsically Disordered Proteins 1

1.2 Protein-protein interactions 2

1.2.1 Solvation 3

1.2.2 Hydrophobic effect 4

1.2.3 Electrostatic Interactions 5

1.3 IDPs in Protein-Protein Interactions 8

1.3.1 Advantages of ID 9

1.3.2 Describing IDP Interactions 11

1.4 Research Overview 12

2 Methods 13

2.1 Dataset Creation 13

2.1.1 Testing Datasets 13

2.1.2 Sequence Alignments and Clustering 14

2.2 The Classifier 15

2.2.1 Radius of Gyration Over Protein Length 15

2.2.2 Testing Rg/N 16

2.2.3 Coiled Coils, Disulfide-Rich Domains, and Transmembrane Domains 17

2.3 Application of the Identification Filter on the PDB and Selection of Potential IDP Complexes 19

2.4 Secondary Structure Analysis 19

2.5 Validation 20
2.5.1 Disorder Prediction 20
2.5.2 Extended ID Segment Dataset 21
2.5.3 Solvent Accessible Surface Area of IDPs 22
2.6 Calculation of SASA 22
2.7 Radius of Gyration Ratio and Compaction of IDPs 23
2.8 Interface Residue Analysis 25
2.9 High Resolution Datasets 26
2.10 Hydrogen Bonds, Salt Bridges and Interface Area 27
2.11 Computational Alanine Scanning for Per Residue Contribution to Binding 28
2.12 Electrostatic Calculations Using CHARMM Parameters 30
2.12.1 CHARMM Structure Minimization 30
2.12.2 Electrostatic Calculations 31
3 Results & Discussion 38
3.1 Identification of ID segments in complex with partner proteins 38
3.2 Validation of identified proteins 42
3.2.1 Sequence disorder prediction 42
3.2.2 Amino acid composition 44
3.2.3 Solvent accessible surface area 46
3.3 Characterization of complexes 48
3.3.1 Predicting compaction in IDP binding 48
3.3.2 Secondary structure analysis 51
3.3.3 Interface residue analysis 53
3.4 Alanine scanning analysis 56
3.5 Role of electrostatics in ID segment-partner interaction 60
4 Conclusion 63
References 66

List of Tables

Table 1 List of datasets analyzed and the number of structures in each dataset............................38
Table 2 Hydrogen bonds, salt bridges and interface size. 56

List of Figures

Figure 1.1 Specificity conferred by electrostatic complementarity. 7
Figure 3.1 Scatter plot of IDP complexes and 3D complexes. 40
Figure 3.2 ROC curves and MCC curve evaluating the classifier. 41
Figure 3.3 Percentage predicted disordered. 44
Figure 3.4 Residue composition. 45
Figure 3.5 Ratio of SASA in complexed conformation and monomeric state. 47
Figure 3.6 ID segments Rg free in solution over Rg in complexed state. 50
Figure 3.7 Secondary structure composition of ID segments and predicted disorder. 52
Figure 3.8 Pie chart of secondary structure composition of 3D complex proteins. 52
Figure 3.9 Interface residue compositions. 54
Figure 3.10 Change in free energy of binding from alanine scan. 58
Figure 3.11 Polar desolvation, Coulombic, and total electrostatic energy of binding. 62

List of Abbreviations

Arel	Ratio of SASA in complexed conformation over monomeric conformation
AUC	Area under curve
DSSP	Define Secondary Structure of Proteins
FN	False negative
FP	False positive
FPR	False positive rate
ID	Intrinsic disorder
IDP	Intrinsically disordered protein
IQR	Interquartile range
MCC	Matthew's Correlation Coefficient
nrPDB	Non-redundant Protein Data Bank structures
PDB	Protein Data Bank
PBE	Poisson-Boltzmann Equation
rASA	Relative accessible surface area
Rg	Radius of gyration
Rg/N	Radius of gyration divided by protein chain length
Rh	Hydrodynamic radius
ROC	Receiver operating characteristic
SASA	Solvent accessible surface area
SMV	Support vector machine
TN	True negative
TP	True positive
TPR	True positive rate
VdW	Van der Waals

Acknowledgement

I am very thankful to Dr. Jörg Gsponer for the opportunities and support given. I arrived at his office not knowing exactly what to expect and he helped me get started by introducing me to the GSAT studentship program. He introduced me to computational biology and the study of intrinsically disordered proteins, which renewed and elevated my enthusiasm for research in biochemistry.

I also want to thank all the current and past lab members for all the help and inspirations they provided.

I thank Dr. Filip Van Petegem and Dr. Lawrence P. McIntosh for advice on my research and on the progress of my degree program. I also thank Dr. Suzana K. Straus for participating in the defence examination and for her advice for improving the thesis.

I thank GSAT for giving me the chance to get my first taste of research through their summer studentship program.

Importantly, I have to express my gratitude to Dr. Emmanuel D. Levy for providing perl scripts to help implement the definitions for core, rim and support interface regions.

Lastly, I thank my parents. I cannot adequately describe how grateful I am for all their support.

Dedication

To my parents

1 Introduction

1.1 Intrinsically Disordered Proteins

Ever since the first structure of myoglobin was resolved by Kendrew and colleagues more than half a century ago (Kendrew et al., 1958), the Protein Data Bank has grown to include more than 80 thousand structures. The immense progress in structural biology allows us to understand many cellular processes through making the connection between structure and function. This led to the view that proper protein function requires well-determined three-dimensional structures. In fact, many have viewed proteins as intricate machines that work in a well ordered manner in the cell (Alberts, 1998). This view slowly evolved with the discovery of unstructured proteins whose structure cannot be described by a single set of three-dimensional coordinates. NMR techniques have shown the presence of disorder in cellular proteins decades ago (Daniels et al., 1978), but the importance of these intrinsically disordered proteins (IDPs) has only been recognized more recently (Dyson and Wright, 2005). Consequently, the majority of previous research and methods cater towards structured proteins, which means that the study of IDPs is a relatively new field with many gaps in knowledge to fill.

Taking advantage of the availability of large quantities of protein sequences, computational programs were developed to identify intrinsically disordered regions from sequence (e.g. DISOPRED). A paradigm changing discovery was made when whole genome predictions of protein disorder revealed that greater than 30 percent of eukaryotic proteins have unstructured regions of over 30 residues long (Ward et al., 2004). Not only are IDPs prevalent in our proteome, they are also functional and are implicated in various important processes. One of the main motivations to study IDPs is the fact that they are often involved in signaling and regulation. Biophysical properties conferred by their unstructured nature makes ID regions suited for recognition of protein targets. Moreover, some IDPs are identified as hub proteins in protein interaction networks. Hub proteins play key roles in interaction networks through interacting with multiple partners and examples of hubs include p53, Mdm2 and p300 (Gsponer and Babu, 2009). The implication of these regulatory proteins in severe diseases, such as cancer, gives IDP research practical applications in medicine. In fact, because IDP binding sites on structured proteins tend to be deep grooves, IDP interactions show extra promise for drug development targeting protein-protein interactions (Meszaros et al., 2007).

1.2 Protein-protein interactions

No one would doubt the complexity of cellular systems because it is their functions that are allowing us to contemplate this very subject. These complex systems must have sophisticated mechanisms to regulate a myriad of metabolic processes and to respond to all sorts of external signals and stimuli. These mechanisms of signaling and regulation are primarily facilitated through protein-protein interactions. Consequently, a significant amount of research has been done to better understand the biophysical properties that control protein folding and interactions. Proper regulation of cellular processes generally requires specificity in protein-protein interactions and the means to regulate these interactions, such as post-translational modifications, alternate splicing or autoinhibition.

The chemical physics behind protein-protein interactions involves both the inter-molecular interactions (such as salt bridges) and the proteins' interaction with solvent.

1.2.1 Solvation

Protein folding is a subject of great interest in structural biology and its solution is often considered as the holy grail of structural biology. A protein is synthesized as a linear polypeptide and is driven to fold by the gradient in the free energy landscape (Baldwin, 2007). The search of the free energy minima results in the folded three-dimensional structure. It is well established that the most important driving force behind protein folding is the burial of non-polar (hydrophobic) surfaces from the polar environment of water (Baldwin, 2007). The hydrophobic effect (hydrophobic factor) is the same property that prevents oil from mixing well with water. In fact, free energy change of burying hydrophobic side chains in the protein interior was first modeled with dissolving hydrocarbons in water (Baldwin, 2007). A model suggested by Widom proposes that the hydrophobic effect comes from a balance between creating a cavity for the solute in the hydrogen-bonding structure of water and the new solute-solvent interactions (i.e. $\Delta G_{solvation} = \Delta G_{cavity} + E_{interaction}$) (Baldwin, 2007). Hence, the solvation energy refers to the change in free energy conferred by transferring protein from gas phase into the aqueous environment. The solvation of a hydrophobic surface would have a high solvation free energy because the cost of creating the cavity, which involves breaking the hydrogen bond network in bulk water, is not well compensated by solute-solvent interactions that primarily consist of van der Waals interactions. However, solvating polar surfaces would be more favorable due to the polar-polar interactions between the solute and water. Stable folded proteins selected by nature have interiors composed primarily of hydrophobic residues, with the exception of buried catalytic sites in enzymes. On the other hand, charged and polar residues are enriched on the surface of proteins where they are solvated.

Put in the context of protein-protein binding, solvation energy is the free energy for removing the water from the interface surfaces. This is generally unfavorable for the polar surfaces and favorable for hydrophobic surfaces. The balance of hydrophobicity and polarity varies between different types of protein interfaces with homo-dimers being the most hydrophobic and hetero-complex components that can exist individually being more polar (Jones and Thornton, 1997). The difference reflects the need for some proteins to be soluble on their own and also function in their oligomeric state.

1.2.2 Hydrophobic effect

Protein-protein interactions involve the same physics as folding so it is no surprise that the hydrophobic effect is a key thermodynamic driving force for protein binding. Protein-protein interfaces are enriched in hydrophobic residues, which is in contrast with the rest of the polar-residue-enriched and solvent-exposed protein surfaces. Hydrophobic surface patches are synonymous with protein-protein interactions. In fact, the drive to form non-polar interactions is so strong that hydrophobic patches sometimes promote nonspecific and detrimental interactions that lead to formation of protein aggregates. In one study, it was shown that predictors for aggregation are actually better at predicting binding interfaces on protein surfaces than hydrophobicity alone (Pechmann et al., 2009). Protein aggregates are implicated in serious neurodegenerative diseases such as Huntington's and Alzheimer's disease.

1.2.3 Electrostatic Interactions

Specificity of protein interactions is generally believed to rely on shape complementarity, hydrogen bonds and salt bridges. A protein has high specificity for a binding partner when their affinity is very high compared to the protein's affinities for all of its non-cognate binding targets. Shape complementarity is important for structured protein binding and has been successfully applied in protein-protein docking programs (Ritchie and Kemp, 2000). Since ID segments lack a unique three-dimensional structure, the focus of our attention is on electrostatic interactions. The electrostatic energy of protein interactions is an extensively studied subject and it encompasses much of the contributions from salt bridges and hydrogen bonds. Although electrostatics generally does not contribute greatly to the affinity of protein-protein binding, it has been revealed that polar interactions can sometimes contribute significantly to the binding free energy but more often confer specificity. Clackson and Wells showed that even though charged residues generally contribute less than hydrophobic residues, functionally important charged residues which make complementary contacts across the interface can be interaction "hot spots", whose replacement with alanine reduces binding affinity by more than 1 kcal/mol (Clackson and Wells, 1995).

Taking into account all the polar interactions, Sheinerman and Honig analyzed the interfaces of four protein-protein complexes and found that the total electrostatics contribution to binding is inversely correlated with non-polar solvent accessible surface area buried upon binding (Sheinerman and Honig, 2002). In other words, the electrostatic interactions are stabilizing in complexes with relatively weak non-polar contributions to binding while in others, the electrostatic contributions are strongly destabilizing but the non-polar contributions to binding are very high. Repeatedly, studies have shown that salt bridges in protein interfaces can contribute favorably in terms of electrostatic interaction to the free energy of binding but that the effect is often counterbalanced by very unfavorable polar solvation energy of binding, or desolvation (Elcock et al., 1999).

Despite the fact that the sum of electrostatic energy of binding is generally not a great contributor to affinity, polar interactions such as salt bridges are still important for specificity. In the simplified example in Figure 1.1, the desolvation cost of burying charged residues is compensated by making new specific interactions with residues of the opposing charge. When the charged residue is not making specific interactions, the desolvation cost is not compensated and the electrostatic component of the binding free energy becomes unfavorable. Consequently, even though electrostatic interactions are not the main contributor to affinity, they effectively penalize uncomplimentary protein binding to uphold specificity.

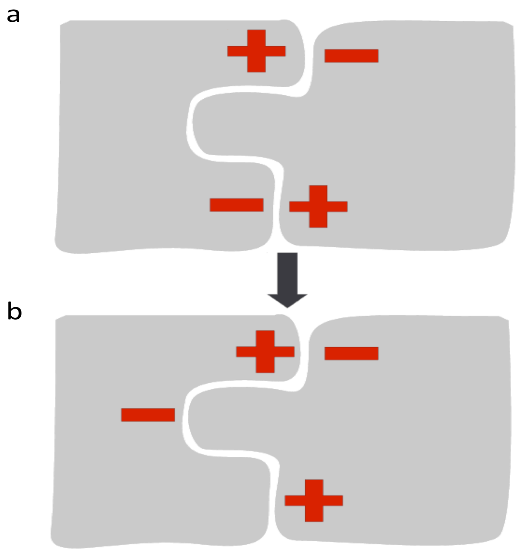

Figure 1.1 Specificity conferred by electrostatic complementarity.

Complex (a) and (b) have the same shape complementarity but the first complex has higher electrostatic complementarity. In the first complex, the desolvation cost of the four surface charges cancels out the Coulombic energy from the two salt bridges. The second complex with approximately equal desolvation penalties will have unfavorable total electrostatic free energy of binding. In this way, electrostatic interactions can rule out the formation of the second complex despite the fact that there are no explicit same-charge interactions opposing binding.

Electrostatic interactions are also an important determinant for association kinetics. Simulations of enzyme-substrate association have demonstrated the role of electrostatics in the rate of association in diffusion-limited enzymes, for which the rate-limiting step in their reaction is association (Elcock et al., 1999). In one such case, McCammon and colleagues have shown that enzyme-substrate association rates for copper/zinc superoxide dismutase can be altered by mutating charged residues inside the interface and also outside to a lesser degree (Sines et al., 1990). The electrostatically enhanced association was also shown in protein-protein associations such as barnase with the inhibitor barstar (Gabdoulline and Wade, 1997). Even for cases where the electrostatic contributions to binding were found to be unfavorable in their bound state, electrostatic interactions could still contribute favorably when the two proteins are approaching (Elcock et al., 1999).

1.3 IDPs in Protein-Protein Interactions

Many functional advantages have been proposed for intrinsic disorder in proteins, all of which involve interactions with other proteins in some way. In contrast to structured protein complexes, much less is known about the energetics of interactions between ID segments and globular proteins. Notably, the role of electrostatics and specificity is not well characterized. Researchers tend to agree that IDP interactions are generally highly specific while often maintaining low affinity. Being regulatory or signaling elements requires these proteins to be able to selectively bind their targets while being able to quickly dissociate in response to signals, which is analogous to switches in the cellular pathways (Zhou, 2012). Studies have shown that the interface residues of ID segments are even more conserved than those of structured protein interfaces. This is unexpected because of the stark contrast to the rest of the ID regions, which have little conservation (Meszaros et al., 2007). Sequence conservation alludes to the importance of the residues to the interaction and also the importance of the interaction itself. But without rigid scaffolding behind the interface of ID segments to help enforce shape complementarity restrictions during binding, how do ID segments bind with high specificity? This is the one of the main focuses of this thesis.

1.3.1 Advantages of ID

Firstly, why do cells require ID segments? Some functional advantages of being unstructured proposed include (i) their ability to facilitate post-translational modifications (ii) their efficiency in recruiting and scaffolding proteins through the "fly-casting" mechanism and (iii) their ability to adapt different conformations, which allows them to bind to one or more partner with specificity (Gsponer and Babu, 2009). Unstructured segments of proteins have the advantage of being accessible by other proteins such as kinases, which are involved in post-translation modifications. For example, a study in yeast kinase-substrate network revealed that substrates of kinase are two times more enriched in disordered proteins than structured proteins (Gsponer et al., 2008). Post-translational modifications are crucial for signaling and regulation because they provide the cell with greater control through increasing the complexity of their interaction networks. P53 terminals are good examples of ID regions that are highly populated with sites of modifications. Phosphorylation, acetylation, ubiquitination, methylation and other post-translational modifications in different patterns or combinations can alter the affinity of the binding partners of the IDP and also regulate the protein's stability (Wright and Dyson, 2009). The fly-casting mechanism gives unstructured proteins a larger "capturing radius" than structured proteins. This mechanism gives ID segments greater association rates with binding partners by forming weaker binding intermediates while the rest of the two proteins are still separate (Levy et al., 2004; Shoemaker et al., 2000). Finally, ID segments can bind more than one partner with specificity; and HIF1α is a good example (Dyson and Wright, 2005). Hydroxylation of Asn830 of HIF1α is a mechanism that reduces its affinity for CPB/p300. However, binding to CPB/p300 requires a helical conformation while binding to the hydroxylase active site requires an extended conformation. The ability to alter affinity for different partners under different conditions/stimuli provides greater complexity for signaling and regulation.

Another advantage that ID regions provide is large surface area. By providing greater interaction area, ID regions can support interactions with several molecules to form large multimeric complexes (Gsponer and Babu, 2009). IDPs such as BRCA1 that act as a scaffold for complex assembly led to a study that revealed a correlation between protein complex size and the average predicted disorder (Hegyi et al., 2007). A similar proposed rationale for the presence of IDPs is that unstructured regions reduce the protein mass required to provide the needed interaction area (Gunasekaran et al., 2003). The authors of this study reasoned that if all the unstructured regions are replaced with structured domains, the complexes would be much larger in size and, consequently, so would the genome size and cell mass.

Clearly, intermolecular interaction is the key function of IDPs. However, much less is known about the characteristics of IDP complexes relative to the structured protein complexes which are traditionally studied. IDP complexes are defined as complexes where one protein subunit is disordered in isolation. Describing and uncovering more about the protein-protein interactions in IDP complexes are the focuses of the rest of the introduction and my research chapters.

1.3.2 Describing IDP Interactions

Two main models have been suggested for the binding of an ID segment to structured proteins: (i) induced folding upon binding and (ii) conformation selection. The first suggests that the unstructured ID segment folds into the bound state upon making intermolecular interactions with the structured interaction partner while the second suggests a pre-existing conformation is selected out of the population of sampled conformation during binding (Song et al., 2008; Sugase et al., 2007). A more recent interpretation attempts to unify the two models (Csermely et al., 2010). The role of entropy in binding differs from traditional structured protein binding because of the greater loss of conformational entropy upon binding. The magnitude of entropy loss is up for debate since it depends on the binding model and the degree of disorder in the complex. Not only is there a spectrum to the extent of disorder in native ID segments, but it has also been noted that many IDP complexes display varying degrees of disorder. Tompa and Fuxreiter coined the phenomenon "fuzziness", which may limit the entropic penalties of binding and further confer advantages in specificity and regulatory functions (Tompa and Fuxreiter, 2008). Essentially, quite a few binding models have been recognized through studying a relatively small fraction of IDPs. This highlights the great diversity in the mechanisms employed by IDP interactions.

Although IDP complexes are not as thoroughly studied as their structured counterparts, it is well established through previous studies that hydrophobic interactions are even more prevalent in IDP interfaces than structured protein interfaces. The enrichment of hydrophobic residues is one of the reasons why some researchers proposed that their interactions may be less specific than classical protein complexes (Vacic et al., 2007).

The large and hydrophobic interface of IDPs leads to high binding enthalpy. Nussinov proposed that tight interfaces provided by IDPs are required for effective signal propagation across proteins (Nussinov, 2012). The high binding enthalpy is counterbalanced by entropic cost of the transition from disorder to order state. This balance leads to low affinity. More specifically, Zhou pointed out that high dissociation rate is the reason why low affinity is an advantage (Zhou, 2012). He also proposed that the high dissociation rate comes from the flexibility of IDPs, which allows for smaller dissociation steps with lower energy barriers. The large interface area and precise fit is thought to be the source of specificity. However, is this alone enough to prevent non-specific interactions when IDPs can clearly adapt their conformation to the binding site?

1.4 Research Overview

The proposed lack of specificity in IDP interactions does not appear to agree with experimental studies that suggest high specificity in IDP interactions. Neither does it seem logical that cellular signaling and regulation, which require specificity for high fidelity, can rely on nonspecific binding interactions. This motivates our analysis of IDP interfaces for factors contributing to specificity, such as hydrogen bonding and polar interactions. Our hypothesis is that IDP interactions should generally be highly specific, and therefore exhibit strong polar interactions.

Firstly, we required a dataset of IDP complexes. An IDP complex dataset was created using the radius of gyration divided by protein length (Rg/N) as a classifier that selects for highly extended protein chains. The IDP complex dataset was then validated using Disopred2 for sequence-based disorder prediction. We also analyzed the residue composition and solvent accessible surface area of selected protein segments to check for agreement with previous studies. Because the method selected for highly extended proteins, we also compared the extendedness of the ID segments before and after binding.

The interface of the validated IDP complex dataset was analyzed and compared to a dataset enriched in structured proteins. Residue composition and alanine scanning analysis of the interface allowed us to roughly examine and compare the types of residues and their thermodynamic contributions to the interactions. Lastly, we provided a more accurate estimate of the electrostatic contributions to ID segment binding.

2 Methods

2.1 Dataset Creation

2.1.1 Testing Datasets

Initially, IDP complexes were collected from literature. The key words “intrinsic disorder”, “unstructured”, and “flexible” were used to search for studies involving IDPs. Many complexes were also found through the DisProt database, which is a curated database of IDPs (Sickmeier et al., 2007). We only kept protein chains longer than 20 residues to reduce the number of protein-peptide complexes. Long IDP segments were also taken from two articles by Dosztányi and coworkers (Meszaros et al., 2009;

Meszaros et al., 2007). Finally, a collection of 74 IDP chains were aligned and clustered (see below) to produce a dataset of 52 IDP chains in complexes.

In a second step, a dataset of structured proteins was assembled. This structured protein dataset was derived from the 3D Complex database (Levy et al., 2006). 3D Complex is a database of protein complexes classified by their known three-dimensional structure in a hierarchical way. The top level of the hierarchy, the Quaternary Structure Topologies, groups protein complexes by the number of chains and the pattern of contacts between them. Quaternary Structure Families is the second level of classification and further divides protein chains based on structural similarities (i.e. based on SCOP domain architecture). We selected dimers out of the Quaternary Structure Families of complexes and downloaded the structure coordinates from the Protein Data Bank (PDB) (Berman et al., 2000). The Quaternary Structure Families grouping has little to no relation to their sequence identity. Therefore, we could apply a sequence alignment and clustering procedure on the dataset (see below). The alignment returned 782 non-redundant structures. With protein chains that are shorter than 20 residues or had already been identified as IDPs in my literature search removed, a final dataset of 762 protein chains was created. I will refer to this dataset as the 3D complexes.

2.1.2 Sequence Alignments and Clustering

Sequence alignment followed by clustering removed redundant proteins from the protein datasets. EMBOSS Needle is a pairwise alignment tool that uses the Needleman-Wunsch global alignment algorithm (Needleman and Wunsch, 1970). We used the scoring matrix EBLOSUM62 and default gap penalties. Needle outputs sequence percentage similarity and identity for each chain pair. We defined redundant sequences using a sequence identity threshold determined by Rost's threshold for percentage identity (Rost, 1999):

$$p^i(n) = n + 480 \cdot L^{-0.32\cdot(1+e^{-L/1000})}$$

We choose $n = 3$ to create more conservative non-redundant datasets and L is the protein length.

Redundant structures were removed by grouping all the protein pairs that are above the identity threshold. One structure was selected out of each group to create a non-redundant dataset.

2.2 The Classifier

2.2.1 Radius of Gyration Over Protein Length

The radius of gyration (Rg) can be used as a measure of protein size. The Rg of a protein is the root mean square distance of the atoms from the center of mass:

$$R_g^2 = \sum_{i=0}^{N} m_i(r_i - R_C)^2/M$$

where the position of the atom i and the center of mass are r_i and R_C respectively. The mass of atom i is m_i and the total mass is M. In our case, only the alpha-carbon coordinates in the PDB files were included in the calculation, so $\sum_{i=0}^{N} m_i/M = 1/N$ where N is the number of alpha-carbons. By dividing the protein size by protein length, the resulting ratio (Rg/N) is a measure of the extendedness of the protein. Rg was calculated using the script rgyr.pl, which is part of the MMTSB toolset, with the '-caonly' option specified (Feig et al., 2004). The protein length was calculated by counting the alpha-carbons only.

2.2.2 Testing Rg/N

The performance of Rg/N for separating the IDP complexes and 3D complexes was first measured using Receiver Operating Characteristic (ROC) curves. They are often used in medical decision making, psychology and machine learning (Bradley, 1997; Swets et al., 2000). ROC curves are plots of a true positive rate on the y-axis and a false positive rate on the x-axis at varying thresholds or cutoff values. The true positive rate (TPR) is also known as the sensitivity and is described for a specific cutoff by this equation:

$$TPR = TP/(TP + FN)$$

where *TP* is the true positive, which is the number of positive cases (i.e. ID segments) correctly classified. *FN* is the false negative, which is the number of positive cases incorrectly classified as negative cases (i.e. structured protein). The false positive rate (FPR) is given by this equation:

$$FPR = FP/(FP + TN)$$

where the false positive (FP) is the number of negative cases classified as positives by the threshold while the true negative (TN) is the number of negative cases correctly classified.

The left side of the ROC curve represents very stringent thresholds where most positive and negative instances are classified as negative, and the threshold relaxes towards the right. The ROC curve can be used to compare how well classifiers perform in separating the positive and negative datasets. The area under the ROC curve (AUC) is a way of evaluating ROC curves. AUC is correlated to the accuracy of the classifier and has the advantage of being insensitive to class skew (Bradley, 1997).

However, the ROC curves do not visually show the optimal Rg/N threshold. A Matthew's Correlation Coefficient (MCC) curve is able to show that explicitly (Matthews, 1975). The MCC of each cutoff is a value between -1 and 1, where 1 is perfect prediction and zero indicates a random prediction. MCC can be calculated by:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FN) \cdot (TN + FP) \cdot (TP + FP) \cdot (TN + FN)}}$$

The ROC and MCC curves were plotted using the ROCR package (Sing et al., 2005).

2.2.3 Coiled Coils, Disulfide-Rich Domains, and Transmembrane Domains

The overlap between the Rg/N of the testing IDP complex dataset and the 3D complex dataset contained many coiled coils and disulfide rich domains. The removal of these structures from the 3D complex dataset increased the AUC slightly. Therefore coiled coils and disulfide-rich domains were removed as part of the process of filtering out structured protein chains.

Coiled coils are common structural motifs that consist of two to five alpha-helices that twist together to form a large coil (Burkhard et al., 2001). Some coiled coils are structured while others are disordered (Anurag et al., 2012). Because we cannot distinguish intrinsically structured from intrinsically unstructured coiled coils, we used Socket (Walshaw and Woolfson, 2001) to identify them and remove them from other ID segment candidates. Socket is a program that identifies coiled coils by recognizing the ‘knobs-into-holes’ packing patterns formed by interlocking alpha-helices in PDB file structures.

Disulfide-rich protein domains are small domains whose structures are stabilized by disulfide bonds. These protein domains tend to be shorter than 100 residues and have small hydrophobic cores. As a result, their structures are mainly stabilized by more than one disulfide bond (Cheek et al., 2006). Their small size and relatively irregular shape give them high Rg/N. Consequently, all protein chains shorter than 100 residues with two or more disulfide bonds were classified as structured proteins.

Another class of protein domains that are extended in structure are transmembrane domains. We used the HMMTOP server for the prediction of transmembrane helicies (Tusnady and Simon, 2001). It uses a hidden Markov model that was trained to recognize topological regions of experimentally determined transmembrane proteins based on sequence. For example, it makes use of the difference in residue composition between cytoplasmic and extracytoplasmic protein domains that flank the transmembrane region. The transmembrane domains are 17 to 25 helical residues long and stretch across membranes. There are beta-barrel transmembrane structures but they have only been recognized in bacterial membranes so far (Tusnady and Simon, 2001). Helical transmembrane domains are more common by far, so we removed only the structures identified by HMMTOP.

2.3 Application of the Identification Filter on the PDB and Selection of Potential IDP Complexes

The list of PDB files to be analyzed was from the nrtable (non-redundant protein dataset) of NCBI with nrlevel 0, which means sequences are grouped using BLAST *p* value of $10e^{-7}$ as a cutoff. Firstly, all the homomeric complexes and any protein chain from the list that consists of more than 50% unknown or non-canonical residues were removed. Then we selected protein chains at least 20 residues long to ensure that we do not end up with a large number of protein-peptide complexes. Because we wanted to analyze the secondary structure, the selected protein complex must also have a DSSP file deposited in the RCSB PDB (See Secondary Structure Analysis Below).

We wanted ID segments that are bound to globular proteins, so only protein chains that are in contact with at least one non-homologous protein chain that is more than or equal to 70 residues in length were selected. Two proteins were considered in contact when alpha-carbon atoms are 3.5Å to 9Å apart. The cutoff of 70 residues was chosen because that is the approximate size of some of the smallest, folded domains such as the SH3 and villin-type headpiece (D'Aquino and Ringe, 2003; Packer et al., 2011).

The key step in the selection process of potential ID segment is the application of the Rg/N threshold to select extended protein chains. We refer to the selected group of potential IDPs as IDP complexes or ID segments from this point forward.

2.4 Secondary Structure Analysis

The DSSP program (Define Secondary Structure of Proteins) was designed by Kabsch and Sander to assign secondary structures through hydrogen bonding patterns and the geometry of three-dimensional protein structures (Kabsch and Sander, 1983). It was used by many studies and bioinformatics programs and could be considered the gold standard for secondary structure assignments even today (Joosten et al., 2011). Furthermore, there is generally one DSSP file for every PDB file stored in the PDB. DSSP make assignments to alpha helices (H), beta bridge (B), extended strands in beta ladders (E), 3_{10} helix (G), pi helix (I), hydrogen bonded turn (T), bend (S), and irregular structures (Ir). For the purpose of this study, we combined H, G, and I into the helices group. Similarly, Ir, S, and B were all included in irregular structures.

2.5 Validation

2.5.1 Disorder Prediction

One of the reasons for choosing to use protein geometry to select ID segments was that we could use sequence analysis to validate the resulting dataset. Disopred2 is a knowledge-based prediction program that uses support vector machine (SMV) that was trained on sequences of high-resolution X-ray structures from the PDB (Ward et al., 2004). Disordered regions are identifiable in these PDB files because they would be present in the protein sequence but are missing in the coordinates due to their flexibility and, consequently, lack of electron density. For each residue in the input sequence, Disopred would output a probability of the residue being in a disordered region and identify residues that have higher disorder prediction than a default threshold producing a false positive rate of 5%. Disopred predictions were calculated for all protein chains in the nrPDB dataset, which included the ID segments. The full sequence of the protein chain was used as the input for disorder prediction because longer sequences would give better prediction. Furthermore, most protein chains came from the UniProt databank (Magrane and Consortium, 2011). For those that are linked to the UniProt databank, we tried to use the longer UniProt sequence instead. The percentage predicted disorder values of each protein chain were calculated with only residues that have alpha-carbon coordinates in the PDB file. The percentage predicted disorder is the number of residues predicted to be disordered divided by the total number of residues. IUPred (Dosztanyi et al., 2005), which is a disorder prediction program that uses a pair-wise inter-residue interaction scoring matrix, was also used to calculate the percentage predicted disordered and we arrived at similar results (not shown).

2.5.2 Extended ID Segment Dataset

Previous studies have shown higher hydrophobic residue composition at the interface of IDP complexes (Vacic et al., 2007). A way to check if the presence of order promoting sequences is due to an enrichment of interacting regions was through doing analysis with the addition of the flanking regions of the ID segments. An effort was made to extend each of the ID segments by 30 residues on each end by counting from the first and last residue with alpha-carbon coordinate in the PDB file. The sequences from UniProt were used preferentially because they are the full sequence of the native protein. The percentage predicted disorder was calculated with all the residues, regardless of whether the alpha-carbon coordinates were present.

2.5.3 Solvent Accessible Surface Area of IDPs

A feature of IDPs is high solvent accessible surface area (SASA) due to their lack of hydrophobic core (See below for SASA calculations). A recent study by Marsh and Teichmann compared the SASA of bound protein to that of their monomeric state (Marsh and Teichmann, 2011). The premise is that the SASA of monomeric proteins are highly correlated to mass and deviations of SASA in the bound state could indicate changes conformation upon binding. In the study by Marsh and Teichmann, they derived an equation that correlates the SASA of monomers ($A_s^{predicted}$) with their mass based on 4988 monomeric proteins:

$$A_s^{predicted} = 4.84M^{0.760}$$

The relative solvent accessible surface area (A_{rel}) was calculated by dividing the measured SASA by the predicted SASA of a monomer of the same mass:

$$A_{rel} = A_s^{observed}/A_s^{predicted}$$

The SASA was measured from the structure of the protein chain in isolation while in the bound conformation. The deviations of A_{rel} ratios from one provided an indication of the change in conformation upon binding.

2.6 Calculation of SASA

An atom is defined as accessible if a spherical probe with certain radius representing the solvent, such as water, could be brought into contact with the van der Waals (VdW) surface of the atom (Lee and Richards, 1971). Additionally, the solvent accessible surface area (SASA) is defined as the surface traced by the centre of a solvent sphere as it rolls across the van der Waals surface of the protein. Solvent accessible surface areas of protein structures in PDB files were calculated using Areaimol, which is a program that uses an algorithm by Shrake and Rupley (Shrake and Rupley, 1973). The solvent probe radius used was 1.4Å, which is the default radius used by Areaimol and by other studies (Levy, 2010; Miller et al., 1987). The point density used was 25Å^{-2} for relative high precision in exchange for higher computational resources. In addition to standard VdW radii used by Areaimol, the radii of zinc, calcium, and sodium ions were taken from CHARMM22 parameters. All crystal water molecules were ignored by default.

All complexes with DNA and RNA were removed from all analysis with SASA calculations because we were only interested in protein-protein interactions and the procedure was more consistent with the following molecular modeling analysis.

2.7 Radius of Gyration Ratio and Compaction of IDPs

Because the radius of gyration is a measure of protein size, comparison of the Rg of a protein before and after formation of a complex reveal its magnitude of compaction or expansion during the process. We do not know the size of the IDPs in isolation, but Marsh and Forman-Kay derived an equation to predict the hydrodynamic radius (Rh) of IDPs based on sequence (Marsh and Forman-Kay, 2010). They fitted Rh measurements from pulse field gradient (PFG) NMR and size-exclusion chromatography (SEC) experiments to an equation that accounted for net charge (Q), fraction of proline residues (P_{pro}), protein length (N), and presence of polyhistidine tag (S_{his}):

$$R_h = (1.24 \cdot P_{pro} + 0.904) \cdot (0.00759 \cdot |Q| + 0.963) \cdot (S_{his}) \cdot 2.49 \cdot N^{0.509}$$

S_{his} is equal to 0.901 with the presence of the polyhistidine tag and is equal to 1 when there is no tag.

Despite the fact that the compaction of our ID segments might depend on regions that are missing from the coordinates of the PDB files, our calculations of Rh only included residues with coordinates because these were the residues that we used to calculate Rg. The occurrence of polyhistidine tag was defined by the presence of at least five consecutive histidines in the protein sequence. The net charge of the protein chain was calculated at pH 7.2 using the Henderson-Hasselbalch Equation for each ionizable group. The pKa of the polypeptide side chains were taken from http://www.cem.msu.edu/~cem252/sp97/ch24/ch24aa.html with the exception of histidine, which we used the pKa of 6.8 to be consistent with Marsh and Forman-Kay. The pKa of the N and C-terminals were 8.2 and 3.65 respectively.

Hydrodynamic radius is the radius of a sphere that would have the same rate of diffusion as the molecule being measured (Marsh and Forman-Kay, 2010). The diffusion coefficient measured in dynamic light scattering experiments is often used to calculate Rh using the Stokes-Einstein equation. The diffusion of the protein is not only dependent on the size of the protein, but also the hydration of the protein. Therefore, the Rh value is dependent on the size of the protein with the hydration layers and the geometry of the protein. The radius of gyration is not directly comparable to Rh. For instance, the Rh of a globular protein is generally much larger than its radius of gyration. However, the Rh can be converted to Rg by using the following equation:

$$R_g = \rho R_h$$

The value ρ, which is the ratio of Rg over Rh, is dependent on the shape of the molecule, but we do not know the shape of the ID segments in solution (Wilkins et al., 1999). For this study, we used ρ values of $(3/5)^{1/2}$, 1.1, and 1.4 to show the possible outcomes of using different shape approximation to convert Rh to Rg. However, we suggest that ρ of 1.1 is a conservative average that could be used for IDPs (see results and discussion).

2.8 Interface Residue Analysis

The interface of a protein complex was defined as the region with a change in SASA upon binding. Similarly, we defined all residues with a ΔSASA upon binding as interface residues. For IDP complexes and 3D complexes, the SASA was calculated for the PDB structure of the whole complex, for the selected ID segment or the selected structured chain in isolation, and of the remaining protein chain(s) of the complex in isolation. The ΔSASA upon protein binding was calculated by subtracting the SASA in the complexed state by the SASA of the two subunits in isolation:

$$\Delta SASA = SASA_{complex(a:b)} - SASA_a - SASA_b$$

This static view of ΔSASA is a rough estimate because it assumes that the conformation of the complex subunits do not change upon binding. This definitely is not true, especially for IDPs. The study by Marsh and Teichmann on the A_{rel} showed that conformational changes that lead to change in total SASA is common in protein-protein interactions (Marsh and Teichmann, 2011). However, we do not know the structure of these protein chains in isolation, so it was an assumption that had to be made.

The protein-protein interfaces were further divided into different regions. A recent article by Levy proposed definitions for the core, rim and support regions of protein-protein interfaces based on SASA calculations (Levy, 2010). Relative ASA (rASA) was calculated by normalizing the residue X's SASA by their SASA in a Gly-X-Gly peptide. Residues with rASA greater than 0.25 in the complexed state were assigned to the rim. Residues with rASA less than 0.25 in the complexed state were assigned to the core if rASA was greater than 0.25 in the unbound state. The remaining residues, which were buried under the core of the interface even before binding, were assigned to the support. Importantly, there is generally no support region for ID segments because they bury very few residues beneath the surface of the interface. Therefore, we ignored the support residues in the analysis done in this study.

2.9 High Resolution Datasets

To avoid artifacts of poor resolution structures, high-resolution datasets for the IDP complexes and 3D complexes were created for the analysis of hydrogen bonds, salt bridges, computational alanine scanning, and continuum electrostatic calculations. The high-resolution datasets consisted of X-ray structures that are higher in resolution than 2.5Å. Consistent with the SASA analysis, all PDB complexes with DNA/RNA were excluded. We also excluded structures that have heterogens, which are marked with HETATM in the coordinate section of PDB files, because they could not be readily modeled using CHARMM (Chemistry at HARvard Macromolecular Mechanics) and other programs. An effort had been made to also include some of the IDP complex structures containing heterogens higher than the resolution of 2.0Å to maximize the size of the IDP dataset. Coordinate ions such as calcium and zinc could be easily included in CHARMM structures. Other small molecules that are not part of the native structure or do not interact with the IDP complex interface were ignored (e.g. glycerol from crystallization process).

2.10 Hydrogen Bonds, Salt Bridges and Interface Area

We used the program HBPlus to calculate potential hydrogen bonds from the high-resolution protein complex structures (McDonald and Thornton, 1994). The program predicts potential hydrogen bonding by calculating the positions of hydrogen and analyzing the structure for hydrogen bond donor and acceptor pairs that satisfy a hydrogen bonding criteria. We used the default criteria which included a distance threshold for donor and acceptor heavy atom distance of < 3.9Å and hydrogen and acceptor distance of < 2.5Å. All hydrogen bonds across protein complex interfaces were tabulated. The analysis did not include aromatic hydrogen bonds.

Any pair of nitrogen and oxygen atom from charged residues that are closer than 4.0Å apart was defined as a salt bridge (Barlow and Thornton, 1983). Terminal charged groups were also included in the analysis. But despite the large presence of terminal groups on relative short ID segments, the terminal groups contributed insignificantly (not shown). Distances between each charged interface residue and all other charged residues were calculated. The number of inter-chain and intra-chain interactions made by these interface residues were tabulated. Residues with no charge compensation within the distance threshold were also analyzed for the presence neighbouring crystal water because water molecules might be involved in water bridging.

2.11 Computational Alanine Scanning for Per Residue Contribution to Binding

Alanine-scanning mutagenesis is an important method often used to study the contribution of interface residues to protein-protein or protein-ligand binding. Residues are mutated to alanine to abolish the interactions made by all the side chain atoms aside from the β-carbon.

FoldX was chosen in this study due to its straightforward procedure for alanine scanning. The FoldX energy function for the calculation of the free energy of unfolding (ΔG; i.e. the change in stability of the protein) consists of van der Waals (ΔG_{vdw}), polar solvation (ΔG_{solvP}), apolar solvation (ΔG_{solvH}), hydrogen bonds (ΔG_{hbond}), electrostatic (ΔG_{el}), main chain entropy (ΔS_{mc}), and side chain entropy (ΔS_{sc}) contributions.

$$\Delta G = W_{vdw} \cdot \Delta G_{vdw} + W_{solvH} \cdot \Delta G_{solvH} + W_{solvP} \cdot \Delta G_{solvP} + \Delta G_{wb} + \Delta G_{hbond} + \Delta G_{el} + \Delta G_{Kon} + W_{mc} \cdot T \cdot \Delta S_{mc} + W_{sc} \cdot T \cdot \Delta S_{sc}$$

ΔG_{Kon} is related to the effect of electrostatic interaction on association rate and is not relevant to the method we used. ΔG_{wb} is the stabilization energy of bridging water molecules, which we ignored. These energy terms are from empirical data. Weights are applied to some energy terms through training of the energy function on the results of experimental mutation studies (Schymkowitz et al., 2005). The results of alanine scanning using FoldX had been compared to a Rosetta procedure and they yielded similar results, although FoldX's estimate tend to be slightly higher (London et al., 2010).

Before executing mutational analysis, it is recommended that a repair procedure is run on PDB structures to repair incorrect structures, such as removing van der Waals clashes. This procedure was done with the "VdWDesign" variable of two, which means that the maximum penalty was assigned for van der Waals clashes.

The mutational analysis was done using the “alascan” command. VdWDesign variable was set to zero to prevent large energetic penalties due to the mutations. The alascan procedure truncates every residue of the protein complex into alanine and does not move the neighboring residues. Residues are mutated to alanine one at a time and the output is a change in energy of folding ($\Delta\Delta G_{mut}$) for each residue. Alanine and glycine residues were not analyzed since their side chains cannot be truncated. We also chose not to include proline residues since their mutation involves changes to the backbone. The change in binding energy upon mutation ($\Delta\Delta\Delta G_{bind}$) for each of the interface residue defined by ΔSASA was calculated by subtracting the $\Delta\Delta G_{mut}$ of the residue in complexed state by the $\Delta\Delta G_{mut}$ of that residue in the unbound state. Mutations carried out without minimization are less accurate but the truncation of larger residues to alanine should not induce big changes in the structure of the protein. Furthermore, a comparable Rosetta mutation protocol with no rearrangement of neighboring residues and using a dampened Lennard-Jones repulsive energy was shown to perform relatively well for mutations of large to small residues (Kellogg et al., 2011).

All FoldX calculations were done at the temperature of 298K, pH 7, and ionic strength of 0.05M. Coordinated metal ions from the crystal structures were used while water molecules were ignored.

2.12 Electrostatic Calculations Using CHARMM Parameters

2.12.1 CHARMM Structure Minimization

Before calculating the electrostatic free energies for the high-resolution datasets, the structures were first processed by CHARMM (Brooks et al., 1983). CHARMM22 parameters were chosen for the electrostatic calculations. The script named convpdb.pl from the MMTSB toolset was used to convert PDB files into CHARMM22 format (Feig et al., 2004). During the conversion process, all histidine residues were converted from HIS to HSD. HSD is a neutral histidine residue with the proton on the ND1 atom. Determining the protonation state of each histidine residue would be difficult and approximating all histidine as the protonated state (HSP) did not have a big effect on our comparison between IDP and 3D complexes. After we generated each of the protein chains in CHARMM, we added all the hydrogen using the ‘hbuild’ command. During this process, missing atoms and residues were also generated in their extended conformation by CHARMM. Missing segments in the middle of a protein chain of less than or equal to six residues were allowed to be built. However, PDB files with longer missing mid-chain segments were removed because it is less likely that the residues’ position would be predicted correctly. For the same reason, a maximum of two missing residues were generated at the ends of the protein chains. The structures were then minimized using a procedure similar to the one used in a Sheinerman and Honig study (Sheinerman and Honig, 2002). In detail, each complex was subjected to 20 steps of steepest descent minimization at 0.05kcal/mol gradient tolerance. A low number of steps was used due to the fact that these are high-resolution structures. Steepest descent is less precise but can be used to quickly remove obviously bad clashes in the structure. Next, harmonic constraints with the force of 50kcal/mol/Å^2 were applied to the heavy atoms with coordinates in the original PDB file before minimizing with 5000 conjugate gradient steps at 0.02kcal/mol gradient tolerance. This would allow the originally missing atoms that were built by CHARMM to be minimized while the rest of the atoms in the complex were held near their original positions. The resulting protein coordinates were used for electrostatic calculations with DelPhi.

Both minimization steps were done with FACTS as continuum solvent (Haberthur and Caflisch, 2008). FACTS is a generalized Born implicit solvent model. It accounts for the effects of solvation on the protein structure implicitly, which means no explicit water molecule is added to the system. Implicit solvent model was used because minimization in vacuum may produce non-physical conformations.

2.12.2 Electrostatic Calculations

A procedure for using the program DelPhi for calculating the electrostatic free energy of binding for protein-protein complexes is described by Sheinerman and Honig (Sheinerman and Honig, 2002). As discussed in that article, this is a simplified approximation that assumes the components of the complex do not undergo significant conformational changes upon binding. They described the free energy change of binding ($\Delta\Delta G_{bind}$) as a sum of the free energy change due to changes in conformation during binding ($\Delta\Delta G_{strain}$) and free energy change of rigid binding ($\Delta\Delta G_{rigid}$). $\Delta\Delta G_{strain}$ consists of enthalpic and entropic changes upon binding and is always positive in value, which is unfavorable for binding (Sheinerman et al., 2000). Similar to their study, the electrostatic calculations here also consisted only of $\Delta\Delta G_{rigid}$.

DelPhi is a program for calculating electrostatic potential using the Poisson-Boltzmann Equation (PBE) (Honig and Nicholls, 1995). The PBE provides a way to describe the electrostatic interaction of molecules in ionic solution. DelPhi presents a method for calculating the numerical solution to the PBE, which could not be solved analytically for complex systems such as proteins. The Poisson's equation is a partial differential equation that can be applied to electrostatics *in vacuo* to give the following equation:

$$-\nabla \cdot \nabla \varphi(r) = {}^{\rho(r)}/_{\varepsilon_o}$$

where $\varphi(r)$ and $\rho(r)$ are the position specific electrostatic potential and charge density respectively. $\epsilon_o = 8.85\times10^{-12} F/m$ is the vacuum permittivity and permittivity (ε) (units in farads per meter) is a measure of a medium's resistance to the formation of a electric field.

The permittivity is proportional to the dielectric constant:

$$\varepsilon_r = \varepsilon/\varepsilon_o$$

Therefore, the dielectric constant (ε_r) for a vacuum environment is one. The dielectric constant increases for media that are more polarizable. An electric field can induce dipole moments in atoms and molecules through electronic polarization and orientational polarization. Electronic polarization involves the shifting of electrons, which moves the negative charge relative to the positive charge in an atom and creates a dipole. Water molecules and polar protein side chains have permanent dipoles, which can participate in orientational polarization. Molecules with permanent dipole moments can reorient or change their conformation to align with the electric field and permanent dipoles are generally much stronger than induced dipoles. A dipole moment induced or oriented by an electric field creates its own electric field (reaction field) that opposes and reduces the strength of the inducing field. Consequently, highly polarizable media with high dielectric constants reduce or screen electric fields more effectively. Media such as the protein interior, which has limited allowance for reorientation due to constraints from hydrogen bonding networks and tight packing, generally can only polarize electronically and have low ε_r of around two. In contrast, water can freely participate in orientational polarization, so water has a high ε_r of 80. Importantly, calculation of electrostatic potential of systems with water and protein requires two dielectric constants. The Poisson equation for these systems is as follows:

$$-\nabla \cdot [D(r)\nabla\varphi(r)] = {}^{\rho(r)}/_{\varepsilon_o}$$

where $D(r)$ is the position specific dielectric constant. Regions accessible by water, as defined by solvent accessible surface using Lee and Richard's definition (Lee and Richards, 1971), are assigned the high external dielectric constant 80. The regions not accessible by solvent are assigned the low interior dielectric.

Lastly, salt concentration in the solvent also affects the electrostatic potential because they are sources of charge. Ions are free to position themselves closer and away from molecules or surfaces with the opposite charge and like charge respectively. This has a similar effect as polarization and reduces the electrostatic potential of the screened source charge. The distribution of ions free in solution is described as a Boltzmann distribution. The Debye-Huckel theory uses the Boltzmann factor to describe the local concentration of ions relative to the bulk concentration in the solvent ($c_{i,bulk}$). The

Boltzmann factor is $e^{-\beta U}$ where *U* is the electrostatic energy of the ion, which is equal to the product of the electrostatic potential and the charge of the ion ($\varphi(r)q$). Therefore, the concentration of ion species *i* at position *r* is:

$$c_{i,bulk}e^{-\beta\varphi(r)q_i}$$

The full nonlinear PBE that takes into account the salt concentration's effect on the electrostatic potential is:

$$-\varepsilon_o\nabla\cdot[D(r)\nabla\varphi(r)] = \rho(r) + \sum_{i=1}^{N} q_i c_{i,bulk}e^{-\beta\varphi(r)q_i}$$

where $\rho(r)$ is the density of the source charge (i.e. charges in the protein). The contribution of salt to the electrostatic potential is all in the second term. Finally, an approximation can be made to produce the linear PBE:

$$-\varepsilon_o\nabla\cdot[D(r)\nabla\varphi(r)] = \rho(r) + \varepsilon_o D(r)\kappa^2(r)\varphi(r)$$

The term κ^2, which is equal to zero in the protein due to absence of salt, in solvent is equal to:

$$\kappa^2 = \frac{\beta}{D\varepsilon_o}\sum_{i=1}^{N} q_i^2 c_{i,bulk}$$

Solving the PBE gives the electrostatic potential in every point in space. DelPhi provides the approximate solution for PBE using finite difference methods on a three dimensional cubic grid. All the charges in the PDB coordinates are mapped onto the grid points. For example, a charge from the coordinate will be mapped onto the 8 grid points surrounding it. The dielectric will be defined for each grid line joining the grid points. The electrostatic potential defined by PBE can be calculated for a grid point using a finite difference equation. The electrostatic potential at a grid point can be calculated from its charge density, its κ, the electrostatic potential of the 6 neighboring grid points, and the dielectric constants on the 6 adjoining grid lines. Consequently, the electrostatic potential of each grid point can be calculated with the exception of the points on the outer boundaries, which are all missing at least one neighboring grid point. The electrostatic potential at these boundary points must be defined by one of several approximations, which are called boundary conditions by DelPhi.

With the electrostatic potential calculated, the electrostatic free energy can be calculated. The electrostatic potential energy of a charged particle in a system with only two charges is:

$$U = q_1 \cdot \varphi_2(r_1)$$

where q_1 is the charge of the first particle and $\varphi_2(r_1)$ is the electrostatic potential at the position of the first charge generated by the second charged particle. The total electrostatic free energy of an entire protein can be calculated as the sum of energy of all the charges in the potential field (Sheinerman and Honig, 2002).

$$\Delta G_{elec} = \frac{1}{2}\sum_{i=1}^{N} q_i \cdot \varphi(r_i)$$

The total electrostatic free energy is partitioned into (i) the Coulombic energy and (ii) the solvation energy by DelPhi. (i) The Coulombic energy is the sum of the energy of each fixed charge in the protein multiplied by the electrostatic potential generated from all the other fixed charges in the protein (φ_{coul}). The φ_{coul} at position j is given by this equation (Rocchia et al., 2001):

$$\varphi_{coul}(r_j) = \sum_{i \neq j} \frac{q_i}{4\pi\varepsilon_o\varepsilon_i r_{ij}}$$

where the ε_i is equal to the internal dielectric constant and r_{ij} is the distance from charge i.

(ii) The solvation energy is equivalent to the energy of moving the protein from vacuum (i.e. dielectric constant of 1) to the solvent (i.e. external dielectric constant of 80). The source of this energy is due to the polarizing effect of the electric field generated by the fix charges. The electric field induces surface charges on the boundary between two different dielectric media and the resulting induced field is called the reaction field. In the case of a protein, this boundary is between protein and water as defined by Lee and Richard (Lee and Richards, 1971). The reaction field energy is the energy from the product of all the fixed charges and the potential generated by the induced surface charges (φ_{react}). The equation for φ_{react} is:

$$\varphi_{react}(r_j) = \sum_{p} \frac{\delta_p}{4\pi\varepsilon_o r_{pj}}$$

where δ_p is the surface charge at position p (Rocchia et al., 2001).

Electrostatic calculations with DelPhi were done using charge and radii from CHARMM22 parameters. We used a 401*401*401 point cubic grid with a scale of 2 grids/Å. This is a very big grid chosen in order to fit all of our complexes with the longest dimension of the largest complexes filling less than 80% of the grid. The accuracy of the calculation is dependent on the grid spacing, which affects the resolution of the molecule for the calculation of the electrostatic potential map. A 2 grids/Å or greater spacing is recommended. However, the number of grid points possible is limited by the computational resources available. This grid was centered on the geometrical center of all the atoms of the protein complex. The calculations on the subunits of the complex were done on the same grid position that was used for the whole complex. The probe radius for determining the solvent accessible surface area was 1.4Å. An ion exclusion (Stern) layer of 2.0Å surrounded the protein where the ion concentration was zero. The boundary potential at the edge of the grid was calculated as the Debye-Huckel potential based on the charge distribution of the system. Each calculation consisted of 2000 iterations of the linear PBE. Generally, 1000 iterations were more than enough for our systems to converge. The linear PBE is not as accurate as the nonlinear PBE, but the linear PBE converges more quickly and is suitable for molecules that have lower net charges, such as proteins (Rocchia et al., 2001).

The interior (protein) and exterior (solvent) dielectric constants were 2 and 80 respectively. The choice of dielectric constant for protein varies for different experiments. The dielectric constant of 1 is used for a medium that is completely non-polarizable, which is also the dielectric of a vacuum environment. Sheinerman and Honig used the dielectric constant of 2 for rigid binding electrostatic calculations (Sheinerman and Honig, 2002). The dielectric constant of 2 means that the protein structure does not undergo conformational changes but it does undergo electronic relaxation. The dielectric constant of 4 implicitly accounts for some changes in the protein structure. Interestingly, dielectric constants of 8 or higher are used for calculation of pKa values of titratable side chains (Nielsen and McCammon, 2003). In accordance to the method used by Sheinerman and Honig, we used the dielectric constant of 2. We did redo the calculations with the dielectric constant of 4. We got similar results when we compared the electrostatic energies of binding of the IDP complexes to the 3D complexes, but their magnitudes are much smaller (not shown).

Similar to the alanine scanning calculations, the electrostatic free energy of binding was calculated as the change in electrostatic free energy between the complex and the two subunits in their unbound states. Wang and Kollman showed that the total electrostatic contributions to the free energy of binding of proteins in water could be calculated as follows (Wang and Kollman, 2000):

$$\Delta\Delta G_{elec} = \Delta\Delta G_{int}^{ele} + (\Delta G_{RFE\ a:b}^{2-80} - \Delta G_{RFE\ a}^{2-80} - \Delta G_{RFE\ b}^{2-80})$$

where $\Delta\Delta G_{int}^{ele}$ is the electrostatic energy of interaction from binding (i.e. Coulombic energy). $\Delta G_{RFE\ a:b}^{2-80}$, $\Delta G_{RFE\ a}^{2-80}$, and $\Delta G_{RFE\ b}^{2-80}$ are the reaction field energies of the whole complex, protein *a* in isolation, and protein *b* in isolation calculated by DelPhi with interior and exterior dielectric constants of 2 and 80 respectively. The change in free energy of solvation from binding is equal to the reaction field energy of the complex minus the reaction field energies of the two protein complex subunits in isolation.

3 Results & Discussion

Table 1 List of datasets analyzed and the number of structures in each dataset.

The numbers in brackets are the number of structures used in the electrostatics calculations (see methods in CHARMM structure minimization).

Dataset Name	Number of structures
Non-redundant PDB	6918
IDP Complexes	368
High-resolution IDP Complexes	91(71)
3D Complexes	762
High-resolution 3D Complexes	140(109)

3.1 Identification of ID segments in complex with partner proteins

It has been shown that ID segments burry more solvent accessible surface area relative to their length than do structured proteins when they interact with other macromolecules (Gunasekaran et al., 2004). In addition, several examples have been reported in which ID segments wrap around binding partners [e.g. p27Kip1 (Russo et al., 1996)]. The reason why ID segments are unstructured in isolation is that they have little to no hydrophobic core, and consequently, are not globular in shape. When they form a complex, they are stabilized by the interactions with the surface of the binding partner and tend to wrap around their structured partner. This motivated us to hypothesize that ID segments could be identified based on their geometry when they are bound to structured proteins. Specifically, the radius of gyration (Rg) of a protein is a measure of its size and will reveal the extendedness of the protein chain when divided by chain length (N). We tested this hypothesis on a set of 52 experimentally verified long ID segments for which we know the structure when in complex with their partner chains(s), which is generally a globular protein. As a negative set, we selected 762 complexes from the 3D Complex database, which is a database of proteins classified based on sequence, structure and topology. This should give a negative set that is enriched in structured proteins that are structurally diverse. We will refer to this dataset as the 3D complexes (See Table 1).

Consistent with our hypothesis and previous observations (Meszaros et al., 2007), we find that ID segments in complex with structured proteins tend to have larger Rg values for a given protein length than do structured proteins when interacting with partner(s) (Figure 3.1). In order to see whether Rg/N can be used as an effective classifier of these structures, we used receiver operating characteristic (ROC) curves. The ROC curve constructed from calculating the Rg/N of the positive ID segment and the negative structured proteins datasets has an AUC of 0.986 (Figure 3.2a). To find an optimal threshold, we then plotted the Matthews correlation coefficient (MCC) curve (Figure 3.2b), which allows identifying thresholds with the highest performance for the classifier (Matthews, 1975). At the Rg/N of 0.22Å, the MCC is maximized to a value of 0.87±0.06.

The chosen IDP dataset is biased to long and extended ID segments, so measures of performance should be interpreted accordingly. Overall, these parameters clearly demonstrate that Rg/N is a very effective classifier to distinguish extended intrinsically disordered segments from structured proteins when in complex with partner molecules.

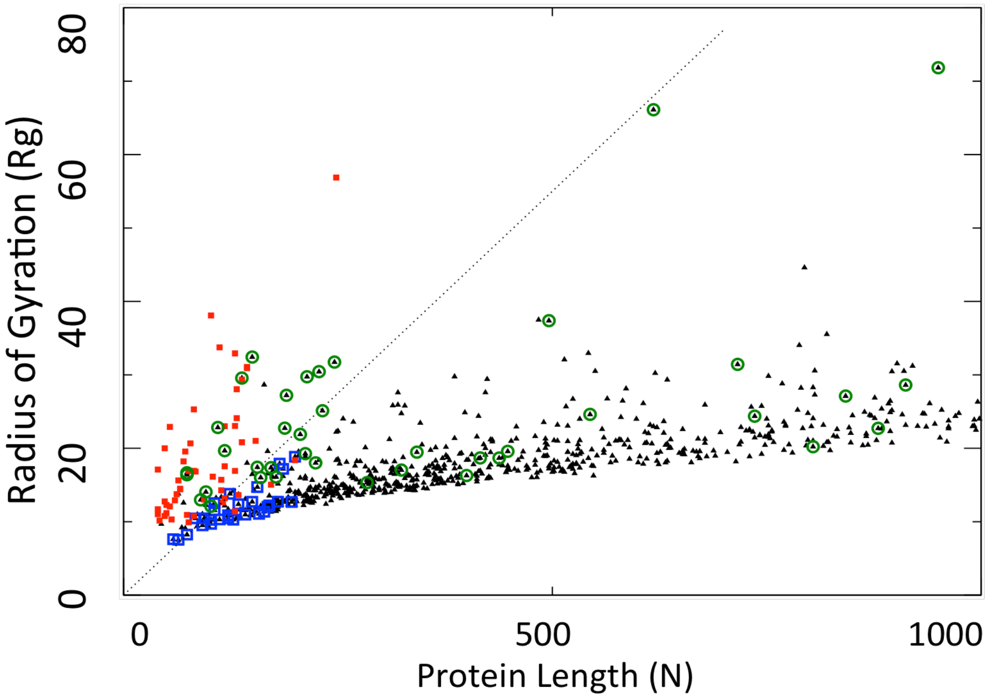

Figure 3.1 Scatter plot of IDP complexes and 3D complexes.

There are 52 ID segments (red squares) and 762 3D complexes (black triangles). The Rg/N threshold of 0.22Å is represented by the dotted line. Disulfide-rich proteins and coiled coils were identified from the 3D complex dataset, which comprises most of the structures that overlap with Rg/N of IDP structures. There are 29 disulfide rich structures and 42 coiled coils represented by blue boxes and green circles respectively.

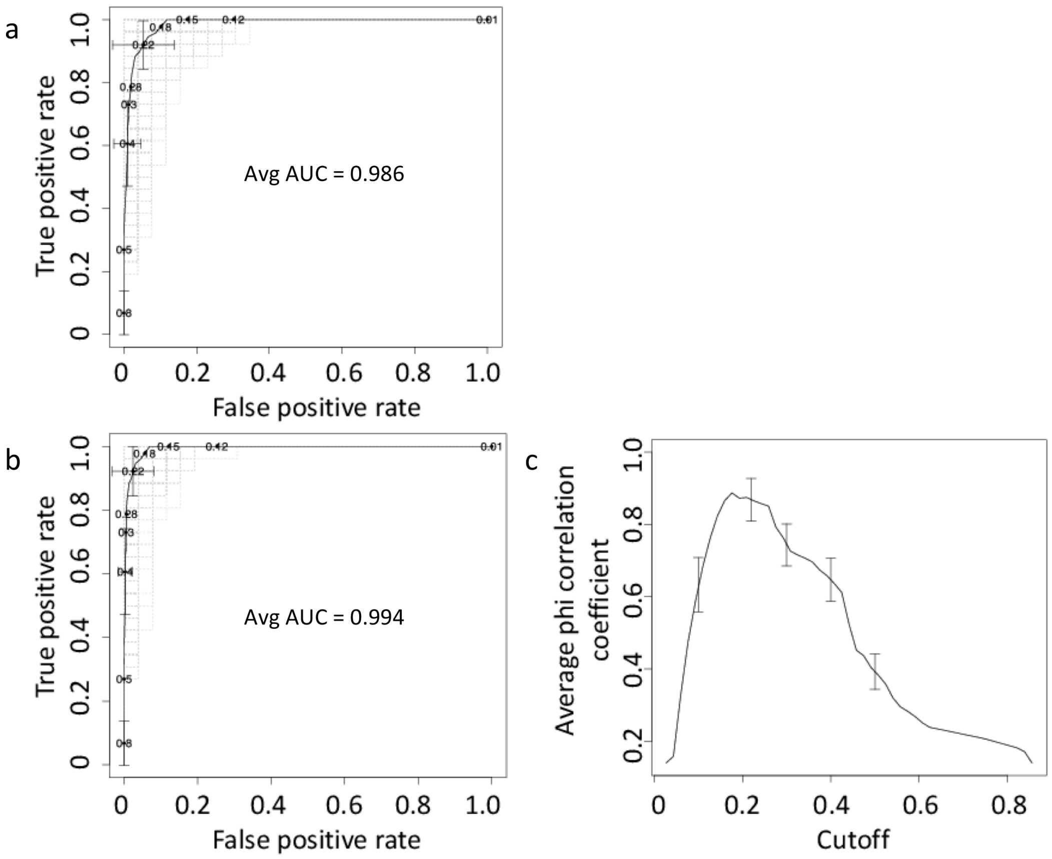

Figure 3.2 ROC curves and MCC curve evaluating the classifier.

(a) ROC and (b) MCC curves were plotted from varying the Rg/N threshold. ID segments are the positive set and 3D complex proteins are the negative set. Figure (c) is a ROC curve of the positive ID segments and the negative structured proteins with disulfide-rich domains and coiled coils taken out. All curves were plotted by randomly sampling 26 structures from the positive and negative sets and averaging over 1000 repetitions.

Despite the excellent performance, we were interested in determining which proteins were falsely identified as intrinsically disordered. A close inspection of Figure 3.1 reveals that most of the false positives are from smaller proteins chains. Rg/N as a classifier appears to have difficulty discerning short ID segments from small, folded structures. Folded proteins shorter than 100 residues are often stabilized by disulfide bonds (Cheek et al., 2006), which allow them to adopt relatively expanded conformations.

Indeed, proteins with disulfide bonds are enriched among the false positives that we identified with the Rg/N classifier. Therefore, we removed proteins that are rich in disulfide bridges from the negative set. In addition, coiled coils often form long stretches of helical structure. Although some coiled coils are known to be intrinsically disordered (Brunger, 2005), this is not the case for all coiled coils. Consequently, coiled coils were also removed from the negative set. As a result of the exclusion of disulfide-rich complexes and coiled coils, the AUC rises slightly (Figure 3.2c).

Next, we applied the method on a non-redundant set of PDB files. The result of the procedure is 368 potential ID segments from 6379 nrPDB chains. IDP complexes or ID segments will be referring to this group of potential IDPs from this point forward.

3.2 Validation of identified proteins

3.2.1 Sequence disorder prediction

The motivation for using a measure of geometry to select interacting ID segments of proteins is to have a method that does not rely on sequence information, which avoids biases in the analyses presented below. Sequence-based disorder predictions can instead be used to cross-validate the selected structures. First, we predicted the disorder for the selected polypeptide chains, i.e. only those residues with known coordinates (Figure 3.3). The selected chains have significantly higher predicted disorder content than a control set of non-redundant structures from the PDB (*p* value < $2.2e^{-16}$; Wilcoxon test). Moreover, the selected polypeptide chains are also predicted to be significantly more disordered when compared to chains that have Rg/N < 0.22Å (*p* value < $2.2e^{-16}$; Wilcoxon test).

Nevertheless, there are still sequences in the identified dataset that have low intrinsic disorder predicted based on their primary structure. The lower than expected percentage disorder for some sequences may be explainable by the enrichment of ID interaction regions in the dataset. These interacting segments are often called Molecular Recognition Features (MoRFs) (Vacic et al., 2007). Interface regions within ID segments are known to be highly hydrophobic and have residue compositions more similar to the buried regions of ordered proteins than to the rest of the ID segments (Vacic et al., 2007). To test this reasoning, we added 30 residues from each of the two flanking regions of the identified ID segments and repeated the analysis. Indeed, the distribution of percentage disorder increases significantly when the ID segment sequences are extended (Figure 3.3, *p* value = 0.002; Wilcoxon test), which suggests that the flanking regions of these ID segments are often more disordered than the interacting regions. A considerable number of ID segments in our set do not have both flanking regions because they are located at the protein termini or the extended sequences were not readily found. Consequently, we expect the difference to be greater still if we could include more flanking regions or exclude the interacting residues.

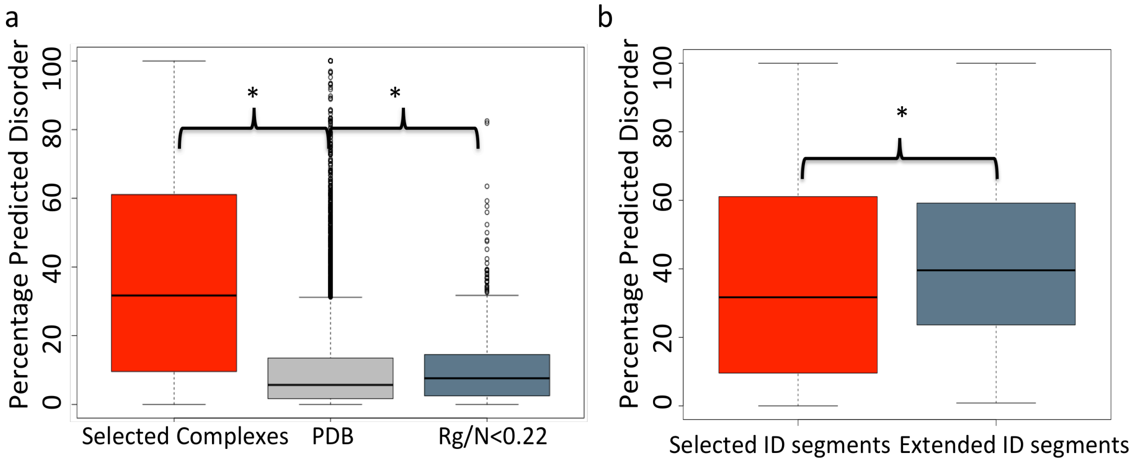

Figure 3.3 Percentage predicted disordered.

The percentage predicted disorder of protein chains were calculated using Disopred2 and their distributions are represented by these box-plots. The horizontal lines of a box-plot listed starting from the bottom are the minimum, first quartile, median, third quartile, and maximum of the distribution respectively. The interquartile range (IQR), which is represented by the coloured region, is the difference between third and first quartiles. Circles above the maximum or below the minimum are data points greater than or less than 1.5*IQR, which are considered to be outliers. (a) The percentage disorder of the selected ID segments, the whole nrPDB dataset, and the low Rg/N protein chains were plotted in red, light grey, and dark grey respectively. (b) The percentage predicted disorder of the selected ID segments and the extended ID segments were plotted in red and dark grey respectively. Distributions with *p* values below 0.05 are marked by asterisk.

3.2.2 Amino acid composition

The amino acid composition of the ID segments is in agreement with the disorder prediction. In comparison to the 3D complex proteins, the composition of the identified protein segments shown in Figure 3.4 is enriched in disorder promoting residues (Romero et al., 2001), especially charged residues such as R and K (R p value = 1.14e^{-10}, K p value = 2.69e^{-5}; Wilcoxon Test). The order-promoting residues, which include most of the hydrophobic amino acids (W p value = 1.29e^{-13}, C p value < 2.2e^{-16}, F p value =

4.18e^{-10}, I *p* value = 0.00080, V *p* value < 2.2e^{-16}, Y *p* value = 2.10e^{-12}, L *p* value = 0.269; Wilcoxon Test), are depleted in our dataset. Surprisingly, some disorder promoting residues are not significantly enriched, or are even depleted, in the ID segments, such as D and E (D *p* value 8.66e^{-6}, E *p* value 0.653; Wilocoxon Test). The overall residue compositions of ID segments and 3D complex proteins are similar to some extant despite the great difference in predicted percentage disorder, which takes into account the importance of other properties like sequence complexity. Since we only counted residues with alpha-carbon coordinates, which are usually residues near the binding site that are stabilized through inter-molecular interactions, we could expect more disorder promoting residues when including the flanking regions. When we extended the number of residues analysed by 30 on each end of the selected polypeptide chains, there is a modest decrease in the percentage of order promoting residues (Figure 3.4). In summary, these results show that our method is able to select for ID proteins, and respectively, for ID protein segments that are in complex with other macromolecules.

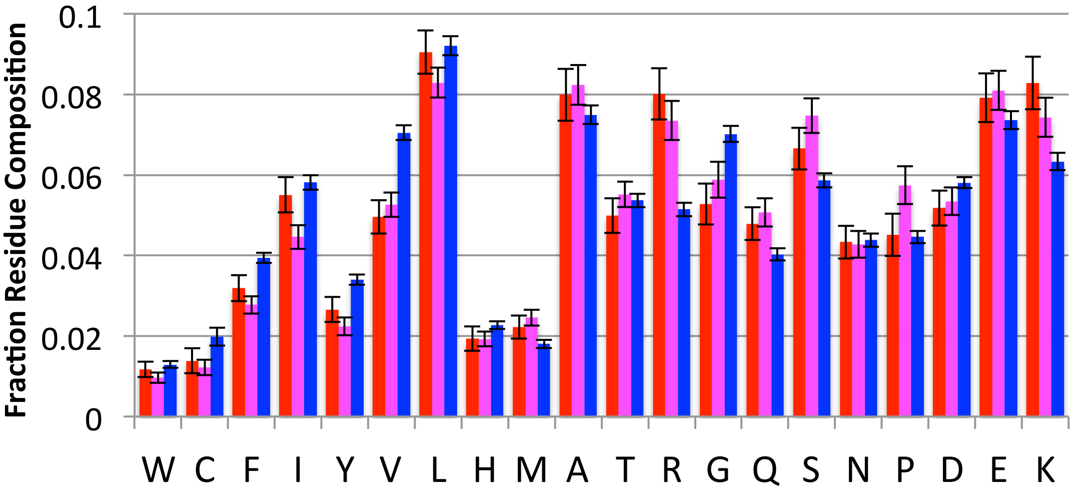

Figure 3.4 Residue composition.

Residue composition of ID segment, extended ID segment, and 3D complex protein datasets in red, magenta, and blue respectively.

3.2.3 Solvent accessible surface area

The increase in SASA of IDPs in bound conformation in correlation with the increase in chain length is much steeper compared to structured proteins, which makes large SASA one of the defining features of IDPs (Gunasekaran et al., 2004). As discussed in the introduction, one of the advantageous features of IDPs is the large surface area they provide. The lack of hydrophobic core in their native state leads to large surface areas that can facilitate protein-protein interactions. Marsh and Teichmann defined relative accessible surface area (A_{rel}) to predict conformational changes upon binding based on the bound protein SASA's deviation from the predicted SASA in the monomeric state (Marsh and Teichmann, 2011). A_{rel} is computed by taking the ratio of the calculated SASA over the SASA predicted by mass. An $A_{rel} > 1$ indicates the bound protein conformation has higher SASA relative to its monomeric state, which means it has undergone conformational changes. Higher A_{rel} values suggest larger conformational changes.

Based on our selection procedure, we expect the ID segments to have A_{rel} significantly larger than one. Indeed, the A_{rel} is greater than one for nearly all the selected ID segments. The A_{rel} of our ID segments are much larger than those of 3D complex proteins as well (*p* value $< 2.2e^{-16}$; Wilcoxon test; Figure 3.5). Interestingly, comparing our results for 3D and IDP datasets to their corresponding structured and disordered datasets in the Marsh and Teichmann's study, both of our datasets have higher A_{rel} distributions. The higher A_{rel} distribution in our 3D dataset suggests that it is relatively heterogeneous and is not void of proteins with flexible structures. Importantly, the A_{rel} in our IDP dataset is generally much higher than the disordered proteins in their study. This probably reflects our selection process, which is based on the extendedness of the protein. Proteins with higher Rg/N are expected to have larger SASA due to their extended structure. In contrast, their dataset is defined using a sequence prediction method.

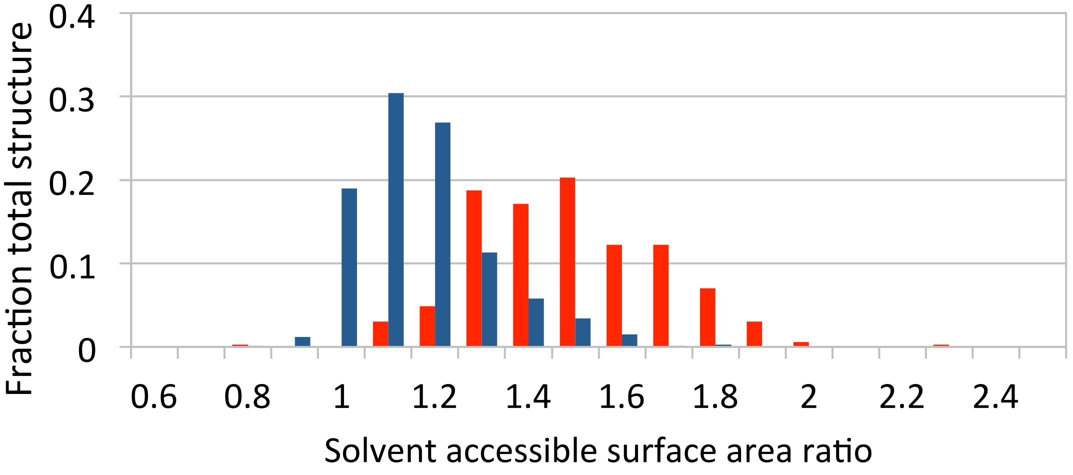

Figure 3.5 Ratio of SASA in complexed conformation and monomeric state. The SASA ratio of ID segments and 3D complex proteins are the red and blue columns respectively.

If we make the assumption that unbound conformations of IDPs are similar to structured monomeric proteins in shape, the results may indicate that our ID segments have extreme changes in conformation upon binding. It is likely true that ID segments do have large conformational changes during binding. However, it is difficult to imagine that ID segments have similar SASA to folded proteins when they are isolated in solution. This analysis is more suited for general protein complexes. Therefore, we can only conclude that our ID segment dataset have high SASA in their protein-bound conformation, which has been recognized in other studies as a discriminating feature of IDPs.

3.3 Characterization of complexes

3.3.1 Predicting compaction in IDP binding

How expanded are the ID segments in their bound states compared to their free states? Marsh and Forman-Kay recently derived a formula from experimental results of SEC and PFG NMR studies to determine the hydrodynamic radius (Rh) that ID proteins have when free in solution (Marsh and Forman-Kay, 2010). To compare the expansion of ID segments before and after binding, we used their equation to predict the Rh that the ID segments have in solution and then transformed it to the Rg using a shape approximation (Wilkins et al., 1999). The conversion ratio ρ of Rg/Rh is dependent on the shape of the molecule.

The ID segments could behave like random coils, so an extreme option would be to use ρ around 1.5 determined for linear chain polymers (Burchard et al., 1980). The structure of IDPs in their native state is different from chemically denatured proteins due to residual structural properties of IDPs that tend to make them more compact (Marsh and Forman-Kay, 2010). It has been suggested that unfolded proteins with residual structures can have a ρ value anywhere between $(3/5)^{1/2}$ and approximately 1.4, which are the ρ values for compact spherical molecules and denatured proteins respectively (Receveur-Brechot and Durand, 2012). Wilkins et al. also calculated the ρ of 1.06 for highly denatured protein chains (Wilkins et al. 1999). On the other hand, a recent study on the effects of charge interactions on the size of IDPs in solution showed that IDPs with high net charge can be much more expanded than a neutral denatured protein due to charge repulsion (Muller-Spath et al., 2010). Förster resonance energy transfer (FRET) was used in that study to analyze the conformation of IDPs and they calculated the Rg of three ID segments. Using their Rg and using Marsh and Forman-Kay's equation to estimate the Rh based on sequence, I derived ρ values of approximately 1.3, 1.7 and 1.8 for HIV-1 integrase, human prothymosin α variant N, and variant C respectively. Human prothymosin α is an extreme case due to the large net charge. Nonetheless, IDPs tend to be enriched in charged residues, so the expansion of IDPs due to net charge should be applicable to many cases. In Marsh and Forman-Kay's study, they compared Rh measurements of relatively large numbers of disordered, denatured, and structured proteins in solution (Marsh and Forman-Kay, 2010). Their study clearly showed that the Rh values of IDPs are much closer to denatured proteins than structured proteins of equal length.

Because we do not know the shape of ID segments in their native state, three different conversion ratios of Rg/Rh (ρ) are analyzed. Aside from the ρ values of $(3/5)^{(1/2)}$ and 1.4, we also used an intermediate ρ of 1.1. The ratio between the Rg of free ID segments (i.e. converted from predicted Rh) and the Rg of the bound conformation is calculated using the three different shape assumptions (Figure 3.6). A ratio higher than one suggests that the ID segment is more extended when free in solution and that implies compaction during binding. Marsh and Forman-Kay also used the ratio between measured Rh over predicted Rh as a measure of compaction. They used this ratio to study the correlation of different sequence features to IDP compaction in solution.

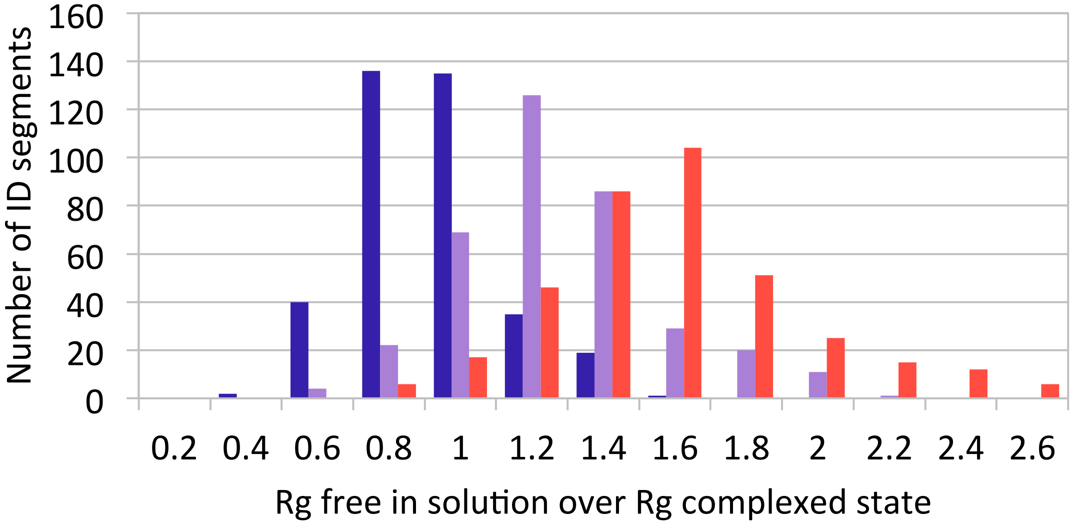

Figure 3.6 ID segments Rg free in solution over Rg in complexed state.

The distributions of the ratios of predicted Rg of the ID segment free in solution over Rg of the ID segment in bound conformation is shown. The blue, purple, and red columns correspond to the ρ values of $(3/5)^{(1/2)}$, 1.1, and 1.4 respectively.

Figure 3.6 reveals that the Rg values of the unbound ID segments are generally higher than the bound Rg for both ρ of 1.1 and 1.4. The Rg ratios from the spherical approximation are predominantly below 1. However, the ρ of $(3/5)^{(1/2)}$ is used as an extreme example that is unlikely to be applicable to ID segments because it assumes the protein chain is densely packed in solution. The average ρ of ID segments is also likely lower than 1.4 due to the presence of residual structures. As discussed above, the extendedness of an ID segment is largely dependent on net charge. An averaged absolute net charge of 3.8±3.7 is calculated for our ID segments. This shows a great variability in net charge but many ID segments have net charge of 1 to 2, which suggests relatively low extendedness. Consequently, we suggest a low *ρ* of 1.1 as a reasonable average approximation. For the *ρ* of 1.1, some structures have ratio below 1, which may indicate that some of the longer ID segments that wrap around large binding partners are in conformations more extended than they normally are in the native state, being held open by their binding partners. However, the general distribution suggests that the majority of ID segments are less extended in their bound state in spite of the fact that our ID segments are biased toward extended bound structures. This is in agreement with the model of folding up upon binding where the ID segment contracts as secondary structures are formed with the help of inter-molecular interactions. Although, we must reiterate that the real *ρ* varies greatly between individual ID segments with different residual structures and net charges.

3.3.2 Secondary structure analysis

Since we are not aware of a dataset of IDP complexes of this size, we thought it useful to tabulate the secondary structures involved. Figure 3.7 shows that helix is by far the most common secondary structure found. Loops or irregular structures involve the second largest group of residues. Both these categories are overrepresented compared to structured proteins (Figure 3.8). Beta structures are much less common in these ID segments.

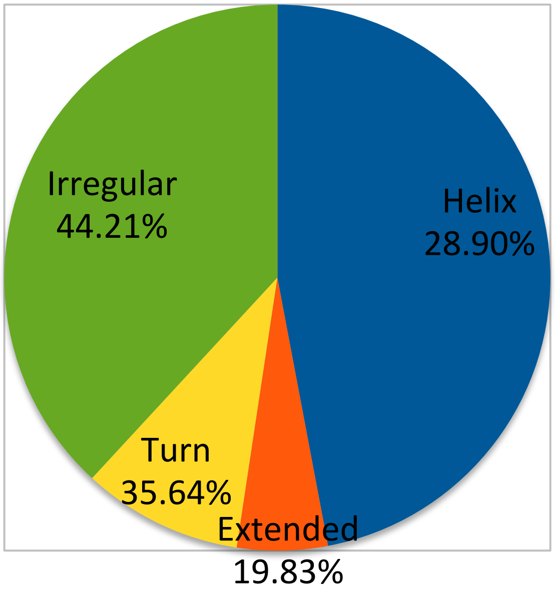

Figure 3.7 Secondary structure composition of ID segments and predicted disorder. The secondary structure composition of the entire ID segment dataset was tabulated. The total percentage of disordered residues involved in each type of secondary structure is the number below each label.

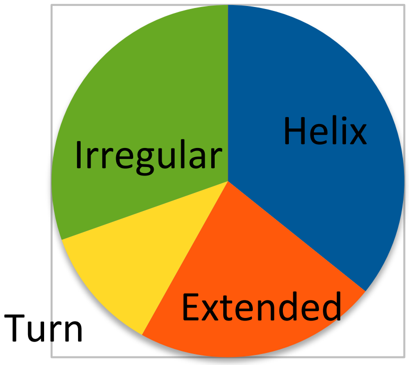

Figure 3.8 Pie chart of secondary structure composition of 3D complex proteins.

The secondary structure of the ID segments are also aligned and compared with disorder prediction in Figure 3.7. The percentage disorder for each secondary structure category is the sum of all corresponding residues predicted to be disordered divided by the total number of residues in that category. The helices and beta strands have lower percentage predicted disorder than residues in irregular structures, indicating that there may be some residual secondary structure in their native state. In fact, only around 30% of the helices are predicted to be greater than 50% disordered in their native state. Irregular structures have the highest predicted disorder, which is expected since they are less likely to have sequence propensity for forming secondary structures or folding.

3.3.3 Interface residue analysis

The main focus of our study is on the interaction between IDPs and their structured binding targets, so we analysed the interface residue composition. Many studies noted the differences in residue composition in the core of the interface compared to the rest of the interface (Bahadur et al., 2003). The cores of structured protein interfaces are the most buried regions and their residue compositions are enriched in hydropobic residues like the interior of globular proteins. On the other hand, the rim region surrounds the core and is enriched with charged and polar residues that are partially solvated even in the complexed state.

We observed the expected enrichment of hydrophobic residues at the cores of complex interfaces (Figure 3.9a)(e.g. IDP core vs. IDP rim: I p value = $1.459e^{-15}$). Importantly, the core regions of the ID segments are much more enriched in hydrophobic residues than the core regions of 3D complexes (e.g. IDP vs. 3D C p value = $1.891e^{-6}$, F p value = $2.425e^{-6}$ (W insignificantly higher p value = 0.1529)), which is in agreement with previous studies (Vacic et al., 2007). Moreover, comparison of Figure 3.9a and 3.9b demonstrates that the rim contains more charged and polar residues for both IDP and 3D complexes. Therefore, the core and rim regions of IDP complexes have the same general properties as those of structured complexes. Furthermore, IDP complex interfaces are definitely more enriched in hydrophobic residues, but most hydrophobic residues are found in the core regions only.

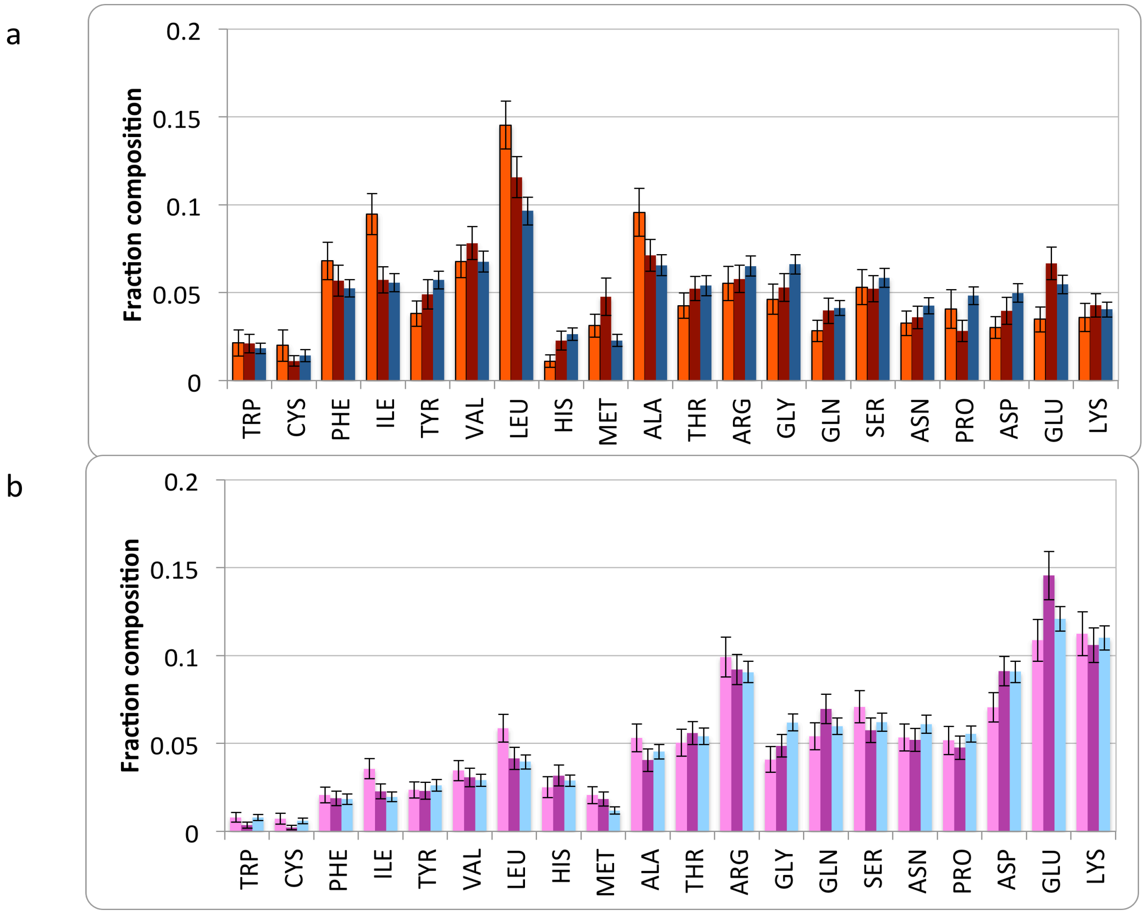

Figure 3.9 Interface residue compositions.

The interface area was divided into the core and rim regions. The core region (a) interface residue compositions of ID segments, IDP binding partner, and the 3D complex proteins correspond to red, maroon, and dark blue columns respectively. The rim region (b) interface residue composition of ID segments, IDP binding partners, and the 3D complex proteins correspond to magenta, purple, and light blue columns respectively. The error bars are the 95% confidence intervals calculated from the mean composition from individual interfaces. The interface regions of IDP binding partners from the same complex were counted as a single interface.

Interestingly, both Figure 3.9a and 3.9b show that charged and polar residues are generally depleted on the IDP side of the interface compared to their structured binding partners. Somewhat unexpected is the evidence that the rim of the IDP interface is not heavily enriched in polar residues. Because the core is so depleted of polar residues (e.g. IDP core vs. ID partner core: K *p* value = 0.03157, E *p* value = 3.846e^{-07}), we thought there would be more in the rim to compensate. Instead, the lack of extra charged groups on the ID segment interface appears to support the idea of less specific interactions. In fact, there are slightly greater percentage of charged residue compositions on the ID partner rim than the IDP's side (E *p* value = 0.0001178, D *p* value = 0.0001038, H *p* value = 0.01886 (insignificantly lower K *p* value = 0.8263, R *p* value = 0.8095); Wilcoxon test). This could suggest that there are charged residues on the structured partner that are buried but not compensated with opposing charges upon binding. This would lead to an overall unfavourable electrostatic energy for binding and, consequently, low specificity. Alternatively, charged residues on the ID segments may be placed more strategically for greater electrostatic complementarity. To test this hypothesis, we evaluated the salt bridge and hydrogen bond interactions across complex interfaces.

In agreement with our hypothesis, there are a greater number of salt bridges found in the interfaces of IDP complex compared to the 3D complex dataset. In Table 2, there are on average 6.9 possible salt bridge interactions found in IDP complex interfaces while 4.7 are found in structured-structured complexes. The number of salt bridges on the IDP complexes is not significantly higher after normalization by the average interface area, but the interface areas of IDP complexes are generally much larger. On average, there are only slightly more hydrogen bonds in IDP complex interfaces (*p* value =

0.05324; Wilcoxon test). The numbers of hydrogen bonds in IDP complexes is comparable, or even slightly lower, when we normalize by interface area. IDP complex interfaces do not appear to be significantly enriched in hydrogen bonds, but it appears that a greater percentage of charged residues are making complementary bridging interactions. Although the percentages of charged residues are smaller on the ID segment interface regions, they may be able to make greater contributions to binding.

Table 2 Hydrogen bonds, salt bridges and interface size.

	IDP Complexes	3D Complexes	*p* values
Hydrogen bonds	12.44	10.64	0.05324
Per 100 Å^2	0.85	0.92	0.3488
Salt-bridges	6.95	4.51	0.00357
Per 100 Å^2	0.49	0.44	0.1308
Unpaired charges	11.48	10.93	0.9389
Per 100 Å^2	0.8	1.05	$1.914e^{-5}$
Unpaired with intra-chain bridges	3.37	2.93	0.2034
Per 100 Å^2	0.22	0.28	0.2742
Interface SASA(Total/2)	1443.92	1148.43	$6.507e^{-5}$

3.4 Alanine scanning analysis

Consistent with previous analysis of ID segments in complex with partner proteins, we find that they have enrichment for hydrophobic residues at the interface. Previous studies proposed, based on these findings, that ID segment-partner interactions may be less specific than other protein-protein interactions. This interpretation seems to be at odds with various experimental findings of high specificity in ID-segment mediated interactions. Therefore, we calculated the energetic contributions of interface residues to binding by carrying out computational alanine scans. It is well understood that, in general, only a small number of interface residues are making the essential contributions to binding and they are called “hot spots”. We defined hot spot residues as those which have a $\Delta\Delta\Delta G_{bind}$ of >1.5kcal/mol. In order to avoid artefacts from low-resolution data, we first assembled high-resolution datasets consisting of crystal structures < 2.5Å in resolution. We also analysed our results through comparison with the 3D complexes to minimize any bias in the methods of calculation. Analysis of our high-resolution datasets reveals a greater percentage of interface residues qualifying as hot spots in IDP complexes than in 3D complex proteins at 28% and 20% respectively (*p* value = $3.18e^{-8}$; Wilcoxon test).

Next, we analysed the contribution of different groups of amino acids to the interface energy. Figure 3.10a shows the distribution of changes in binding free energy ($\Delta\Delta\Delta G_{bind}$) for hydrophobic residues (V, L, I, F, M, Y, W) in IDP complexes and 3D complexes. Interestingly, the distribution of $\Delta\Delta\Delta G_{bind}$ in the ID segments and their binding partners' side of the interface is comparable (*p* value = 0.81; Wilcoxon test), but the hydrophobic residues in IDP complexes generally contribute significantly more to the interface energy than hydrophobic residues in structured complexes (IDP partner *p* value = $3.6e^{-14}$, ID segment *p* value < $2.2e^{-16}$; Wilcoxon test). Greater interface percentage composition of hydrophobic residues with larger surface area, such as phenylalanine, isoleucine and leucine (e.g. Figure 3.9 ID segment and 3D complex core F *p* value = $2.425e^{-6}$;Wilcoxon test), would contribute to greater mean $\Delta\Delta\Delta G_{bind}$. Another possible explanation is that the packing of the side chains in the IDP interface is more optimal compared to the rigid binding structured proteins. Better packing means greater enthalpy from van der Waals interactions, which can translate to higher specificity. A study showed that IDP interface residues make more contacts compared to structured proteins, which could give rise to greater interaction energy for polar and apolar interface residues (Meszaros et al., 2007).

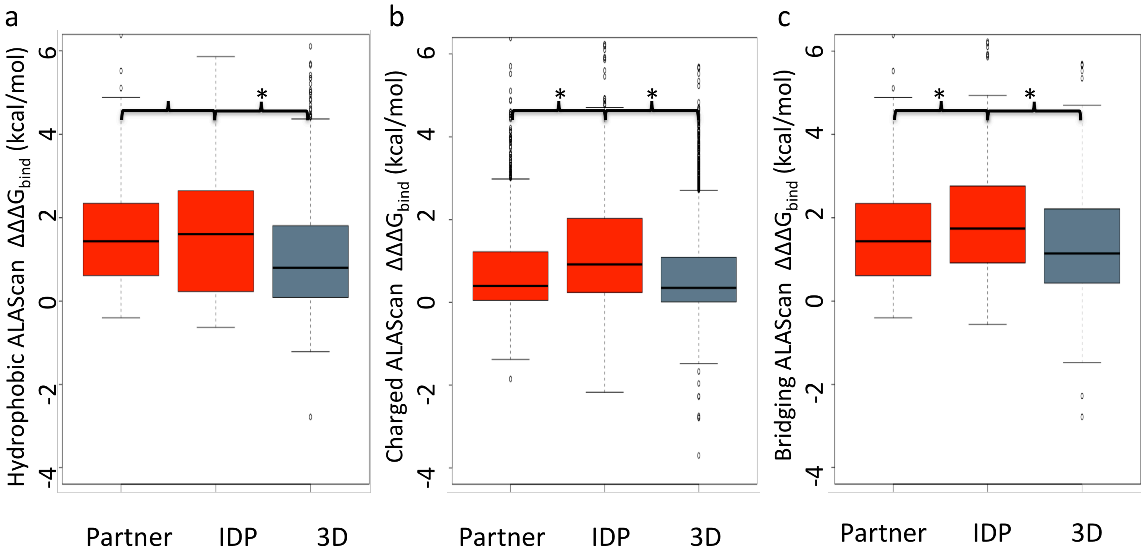

Figure 3.10 Change in free energy of binding from alanine scan.
The box plots show the $\Delta\Delta\Delta G_{bind}$ from alanine scanning results. The residues from the IDP binding partner interface and the ID segment interface were plotted in red. Both sides of the 3D complex interface were combined and plotted in grey. Figure (a) includes all hydrophobic residues, figure (b) includes all charged residues, and figure (c) includes only charged residues that are forming salt bridge interactions. Asterisks identify distributions that have *p* values below 0.05.

Figures 3.10b and 3.10c reveal the distribution of the charged residues (E, D, R, K, H) and the residues that can form salt bridges respectively. As expected, the ion-pairing residues have higher $\Delta\Delta\Delta G_{bind}$ due to the favourable electrostatic interactions. For all charged interface residues (Figure 3.10b), we again find higher per-residue contributions to binding on the IDP complexes relative to 3D complexes (ID segment *p* value < $2.2e^{-16}$, IDP partner *p* value = 0.01057; Wilcoxon Test). Intriguingly, the charged interface residues on the IDP side of the interface have much higher average contribution than those on the structured-partner's side (*p* value < 2.2e^{-16}; Wilcoxon Test). We propose that a greater proportion of the charged residues on the ID segment's interface are making specific interactions, such as salt bridges, than the charged residues on the partner protein. Indeed, this disparity in $\Delta\Delta\Delta G_{bind}$ is smaller when we analysed only the salt bridge forming residue. However, there is still a difference (*p* value 0.009444; Wilcoxon Test), which could arise from cases where a charged residue on the ID segment make multiple salt bridge interactions with the structured partner. The interface residue composition analysis above, which showed a smaller population of charged residues on the IDP, is in agreement with this explanation. A study involving experimental and computational double or higher order mutations demonstrated that polar/charged interactions can form cooperative networks that contribute favourably to structured protein binding (Albeck et al., 2000). Furthermore, the $\Delta\Delta\Delta G_{bind}$ distribution of salt bridging residues in IDP complexes is even higher compared to that of the 3D complexes (ID segment *p* value = 2.966e^{-7}, IDP partner *p* value = 0.01818; Wilcoxon Test). This is more evidence that points to high specificity of IDP interaction. Each residue in the IDP complex interface, especially those in the ID segment, appears to have greater role in binding compared to structured protein-protein complexes. The greater residue conservation observed on the ID segment's interface is also consistent with our explanation (Meszaros et al., 2007). This paints a very different picture for IDP interactions than more promiscuous hydrophobic interface patches that some studies have suggested.

Furthermore, comparing the percentage of interface residue qualifying as hot spot residues on the IDP against those on their partner protein reveals that there is a greater percentage on the IDP side. The percentage of interface residues qualifying as hot spots averages at 41% and 21% (*p* value < $2.2e^{-16}$;Wilcoxon test) for the ID segment and their partner respectively. However, there are greater numbers of residues on the structured partners' interface that may dilute the percentage. The structured protein's side of the interface includes more residues because some residues are partially buried beneath the interface and, in many cases, IDP binding sites are deep grooves with large surface areas. Put it another way, the average IDP interface contain roughly half the residues but have the same number of hot spots as their structured partner. However, this does not preclude the possibility that residues on the ID segment may be taking up more optimal positions to contribute more to binding energy. Studies have shown that structured protein involved in relatively rigid binding have interface residues that are off their minima due to structural constraints. These constraints are relaxed in IDP binding because they can reorganize their conformation upon binding to make more optimal interactions, such as hydrogen bonding angles (Xu et al., 1997).

3.5 Role of electrostatics in ID segment-partner interaction

In order to get more quantitative insights into the contribution of electrostatics to the binding of ID segments to their partner proteins, we used the same approach that has previously been used to study the binding of folded proteins (Sheinerman and Honig, 2002). Results from a comparable procedure using DelPhi for electrostatic calculation have been shown to have good correlation to experimental mutation data for salt bridges (Albeck et al., 2000). Predicted association rates of proteins at varying salt concentration have also been shown to correlate well with experimental results (Selzer and Schreiber, 1999). In the same article, Sheinerman and Honig also revealed a trend of larger surface area corresponding to greater hydrophobic contributions (Sheinerman and Honig, 2002). Since IDP complexes have large interface areas, we were interested to see if they defy this trend.

We compared the electrostatic contributions to the interface stability in classical protein-protein complexes to complexes where one partner is ID in isolation. The distribution of polar desolvation energy of binding is shown in Figure 3.11a. As expected, the electrostatic desolvation energy of binding is much higher for IDP complexes (*p* value = $7.997e^{-11}$). The high electrostatic desolvation energy reflects the cost of removing the larger polar interfaces from the high dielectric environment of the polar solvent and burying them within the low dielectric protein environment. Despite the slightly lower charged residue percentage composition involved in the interaction of IDP complexes (Figure 3.9), the desolvation energy is still much higher than the structured protein complexes. While most IDP complexes have very unfavorable desolvation energy of binding, it is favorable for more than a quarter of the structured complexes. However, the unfavorable polar desolvation may be compensated by inter-chain electrostatic interactions, such as salt-bridges.

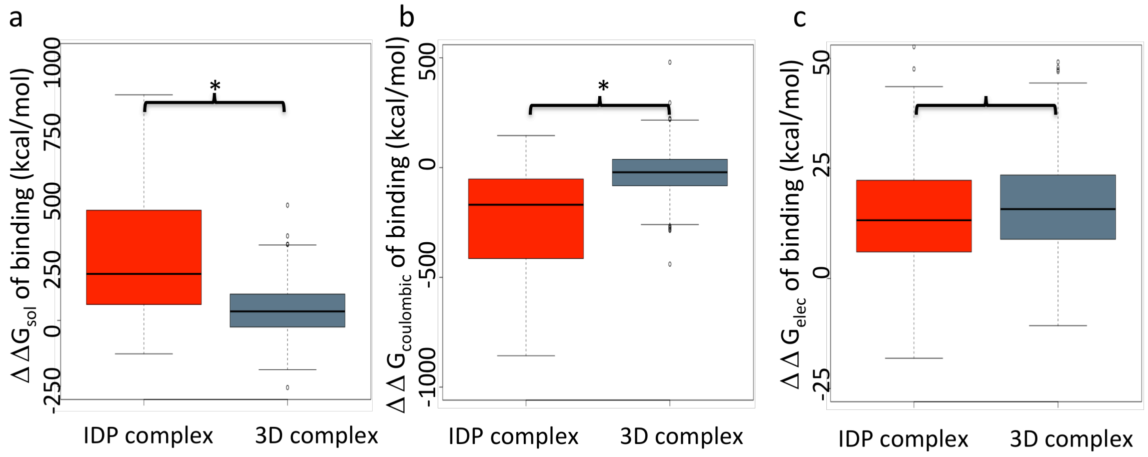

Figure 3.11 Polar desolvation, Coulombic, and total electrostatic energy of binding. Electrostatic contributions to binding were plotted for IDP complexes in red and 3D complexes in grey. Figure (a) shows the desolvation energy of binding. Figure (b) shows the Coulombic interaction energy of binding. Figure (c) shows the total electrostatic energy of binding.

Initially, the residue composition at the interface suggests that the total electrostatic energy of binding for IDP complexes will be somewhat destabilizing because of the uncompensated charges on the partner protein. Surprisingly, the electrostatic energy of binding for IDP complexes is slightly, but not significantly, more favorable than for structured protein complexes (*p* value = 0.0988; Wilcoxon test) (Figure 3.11c). Importantly, the total electrostatic energy of binding is not a very good predictor of affinity (Sheinerman and Honig, 2002). What is important for specificity is the balance between the polar-desolvation energy and the Coulombic interaction energy, which comes from charge/polar protein-protein interactions.

Regardless of whether the total electrostatic energy contributes positively to binding, we can already see that the contribution of Coulombic energy in IDP complexes is high enough to overcome most of the highly unfavorable polar-desolvation energy.

Figure 3.11b shows that Coulombic contributions to binding are much greater in IDP than conventional complexes (*p* value $4.44e^{-11}$; Wilcoxon test). The large differences between the total electrostatic energies of binding and the polar desolvation penalties observed in IDP complexes leaves a great deal of room for the modulation of specificity. Uncomplimentary charged/polar groups will cause protein association to become unfavorable by not compensating for the penalties for burying themselves in a low dielectric environment. This demonstrates the importance of charged and polar interactions in the IDP complexes. With the shape complementarity likely playing a lesser role in IDP complexes due to the lack of scaffolding in IDPs, the role of electrostatic complementarity is essential to specificity. The difference in Coulombic free energy of binding between IDP and conventional protein complexes shows that IDP complexes are clearly in a different category in regards to protein-protein association.

4 Conclusion

The abundance of IDPs in our proteome and their involvement in morbid diseases make the understanding of IDPs very important. Their discovery changed our understanding of protein-protein interactions from a more static and rigid view to a much more dynamic one. Their ability to associate with multiple proteins and propensity for post-translational modification provide greater complexity to cellular signaling. Not only is there a varying degree of disorder in their native states, but some complexes are also shown to be very dynamic and to lack a single well-defined binding conformation. Undoubtedly, IDPs add a new level of complexity to our understanding of protein-protein interactions and cellular functions.

This field has grown rapidly and researchers have proposed many functional reasons for why IDPs are employed by cells and have studied their structural properties in complexes. Repeatedly, hydrophobic residues were noted to be enriched in IDP interactions and hydrophobic interactions are associated to non-specific interactions.

In this thesis, the goal is to create a large dataset of IDP complexes and characterize them. We observed many of the same properties found in other studies of IDPs in our selected complexes, which include high accessible surface area and enrichment of hydrophobic residues in the interface. Indeed, the hydrophobic effect is a strong driving force for binding and IDPs require high enthalpy to compensate for the loss of entropy. However, our study of these IDP complexes revealed the significance of electrostatic interactions. Previous studies often focused more on the core of the IDP complex interfaces and neglected the surrounding charged residues. Moreover, the direct analysis of hydrogen bonds and salt bridges often fails to capture the full electrostatic complementarity in protein-protein interactions. The continuum electrostatic approach used here more accurately accounts for all the electrostatic contributions to binding.

We showed that the magnitude of Coulombic and desolvation energy in IDP binding are both very large. This could be another explanation for the high specificity and low affinity proposed by other researchers. The current opinion on the subject often relates the high specificity to the enthalpy conferred by the large interface and low affinity to the entropic cost of stabilizing the IDP. We proposed that electrostatic complementarity is an important contributor to these attributes because the large Coulombic interaction energy is often completely canceled by the equally large desolvation penalties. We also showed that the fewer charged residues on ID segment interfaces are able to contribute more towards binding than those on structured proteins or even the binding partners of ID segments. These properties could be another defining feature of IDPs.

The higher energy contributions to binding of charged residues on the ID segments may be due to the cooperative charged residue networks on the opposing interface. This could be a worthwhile subject for further research. Defining features like this could be used for prediction of IDP binding sites, which is one of the long-term goals in the Gsponer lab. This is an ambitious goal with the potential for great rewards. Not only would it allow the prediction of novel interactions but it could also help identify new drug targets. Another tool that would be useful to the field is a classification system for IDP complexes. The diverse folding patterns of IDPs in their complexed states may make classification using structural features more difficult than for structured protein complexes, which can rely on well-characterized folding domains. It may be more practical to classify IDP complexes based on biophysical properties such as electrostatics because they may be more relevant to the function of these interactions. For instance, the balance in polar and apolar contributions would likely be different for ID segments involved in signaling as opposed to scaffolding. A classification system that is reflective of their functions could be useful.

In summary, we have shown that strong electrostatic interactions are prominent features of IDP complexes that likely play an important role in modulating their specificity. Furthermore, with a large dataset of IDP complex structures, there are many other experiments and possibilities to explore.

References

Albeck, S., Unger, R., and Schreiber, G. (2000). Evaluation of direct and cooperative contributions towards the strength of buried hydrogen bonds and salt bridges. Journal of molecular biology *298*, 503-520.

Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell *92*, 291-294.

Anurag, M., Singh, G.P., and Dash, D. (2012). Location of disorder in coiled coil proteins is influenced by its biological role and subcellular localization: a GO-based study on human proteome. Molecular bioSystems *8*, 346-352.

Bahadur, R.P., Chakrabarti, P., Rodier, F., and Janin, J. (2003). Dissecting subunit interfaces in homodimeric proteins. Proteins *53*, 708-719.

Baldwin, R.L. (2007). Energetics of protein folding. Journal of molecular biology *371*, 283-301.

Barlow, D.J., and Thornton, J.M. (1983). Ion-pairs in proteins. Journal of molecular biology *168*, 867-885.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic acids research *28*, 235-242.

Bradley, A.P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn *30*, 1145-1159.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S., and Karplus, M. (1983). Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. Journal of computational chemistry *4*, 187-217.

Brunger, A.T. (2005). Structure and function of SNARE and SNARE-interacting proteins. Quarterly reviews of biophysics *38*, 1-47.

Burchard, W., Schmidt, M., and Stockmayer, W.H. (1980). Information on Polydispersity and Branching from Combined Quasi-Elastic and Intergrated Scattering. Macromolecules *13*, 1265-1272.

Burkhard, P., Stetefeld, J., and Strelkov, S.V. (2001). Coiled coils: a highly versatile protein folding motif. Trends in cell biology *11*, 82-88.

Cheek, S., Krishna, S.S., and Grishin, N.V. (2006). Structural classification of small, disulfide-rich protein domains. Journal of molecular biology *359*, 215-237.

Clackson, T., and Wells, J.A. (1995). A hot spot of binding energy in a hormone-receptor interface. Science *267*, 383-386.

Csermely, P., Palotai, R., and Nussinov, R. (2010). Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. Trends in biochemical sciences *35*, 539-546.

D'Aquino, J.A., and Ringe, D. (2003). Determinants of the SRC homology domain 3-like fold. Journal of bacteriology *185*, 4081-4086.

Daniels, A.J., Williams, R.J., and Wright, P.E. (1978). The character of the stored molecules in chromaffin granules of the adrenal medulla: a nuclear magnetic resonance study. Neuroscience *3*, 573-585.

Dosztanyi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics *21*, 3433-3434.

Dyson, H.J., and Wright, P.E. (2005). Intrinsically unstructured proteins and their functions. Nature reviews Molecular cell biology *6*, 197-208.

Elcock, A.H., Gabdoulline, R.R., Wade, R.C., and McCammon, J.A. (1999). Computer simulation of protein-protein association kinetics: acetylcholinesterase-fasciculin. Journal of molecular biology *291*, 149-162.

Feig, M., Karanicolas, J., and Brooks, C.L., 3rd (2004). MMTSB Tool Set: enhanced sampling and multiscale modeling methods for applications in structural biology. Journal of molecular graphics & modelling *22*, 377-395.

Gabdoulline, R.R., and Wade, R.C. (1997). Simulation of the diffusional association of barnase and barstar. Biophysical journal *72*, 1917-1929.

Gsponer, J., and Babu, M.M. (2009). The rules of disorder or why disorder rules. Progress in biophysics and molecular biology *99*, 94-103.

Gsponer, J., Futschik, M.E., Teichmann, S.A., and Babu, M.M. (2008). Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. Science *322*, 1365-1368.

Gunasekaran, K., Tsai, C.J., Kumar, S., Zanuy, D., and Nussinov, R. (2003). Extended disordered proteins: targeting function with less scaffold. Trends in biochemical sciences *28*, 81-85.

Gunasekaran, K., Tsai, C.J., and Nussinov, R. (2004). Analysis of ordered and disordered protein complexes reveals structural features discriminating between stable and unstable monomers. Journal of molecular biology *341*, 1327-1341.

Haberthur, U., and Caflisch, A. (2008). FACTS: Fast analytical continuum treatment of solvation. Journal of computational chemistry *29*, 701-715.

Hegyi, H., Schad, E., and Tompa, P. (2007). Structural disorder promotes assembly of protein complexes. BMC structural biology *7*, 65.

Honig, B., and Nicholls, A. (1995). Classical electrostatics in biology and chemistry. Science *268*, 1144-1149.

Jones, S., and Thornton, J.M. (1997). Prediction of protein-protein interaction sites using patch analysis. Journal of molecular biology *272*, 133-143.

Joosten, R.P., te Beek, T.A., Krieger, E., Hekkelman, M.L., Hooft, R.W., Schneider, R., Sander, C., and Vriend, G. (2011). A series of PDB related databases for everyday needs. Nucleic acids research *39*, D411-419.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577-2637.

Kellogg, E.H., Leaver-Fay, A., and Baker, D. (2011). Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins *79*, 830-838.

Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H., and Phillips, D.C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature *181*, 662-666.

Lee, B., and Richards, F.M. (1971). The interpretation of protein structures: estimation of static accessibility. Journal of molecular biology *55*, 379-400.

Levy, E.D. (2010). A simple definition of structural regions in proteins and its use in analyzing interface evolution. Journal of molecular biology *403*, 660-670.

Levy, E.D., Pereira-Leal, J.B., Chothia, C., and Teichmann, S.A. (2006). 3D complex: a structural classification of protein complexes. PLoS computational biology *2*, e155.

Levy, Y., Wolynes, P.G., and Onuchic, J.N. (2004). Protein topology determines binding mechanism. Proceedings of the National Academy of Sciences of the United States of America *101*, 511-516.

London, N., Movshovitz-Attias, D., and Schueler-Furman, O. (2010). The structural basis of peptide-protein binding strategies. Structure *18*, 188-199.

Magrane, M., and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. Database : the journal of biological databases and curation *2011*, bar009.

Marsh, J.A., and Forman-Kay, J.D. (2010). Sequence determinants of compaction in intrinsically disordered proteins. Biophysical journal *98*, 2383-2390.

Marsh, J.A., and Teichmann, S.A. (2011). Relative solvent accessible surface area predicts protein conformational changes upon binding. Structure *19*, 859-867.

Matthews, B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et biophysica acta *405*, 442-451.

McDonald, I.K., and Thornton, J.M. (1994). Satisfying hydrogen bonding potential in proteins. Journal of molecular biology *238*, 777-793.

Meszaros, B., Simon, I., and Dosztanyi, Z. (2009). Prediction of protein binding regions in disordered proteins. PLoS computational biology *5*, e1000376.

Meszaros, B., Tompa, P., Simon, I., and Dosztanyi, Z. (2007). Molecular principles of the interactions of disordered proteins. Journal of molecular biology *372*, 549-561.

Miller, S., Janin, J., Lesk, A.M., and Chothia, C. (1987). Interior and surface of monomeric proteins. Journal of molecular biology *196*, 641-656.

Muller-Spath, S., Soranno, A., Hirschfeld, V., Hofmann, H., Ruegger, S., Reymond, L., Nettels, D., and Schuler, B. (2010). From the Cover: Charge interactions can dominate the dimensions of intrinsically disordered proteins. Proceedings of the National Academy of Sciences of the United States of America *107*, 14609-14614.

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology *48*, 443-453.

Nielsen, J.E., and McCammon, J.A. (2003). Calculating pKa values in enzyme active sites. Protein science : a publication of the Protein Society *12*, 1894-1901.

Nussinov, R. (2012). How do dynamic cellular signals travel long distances? Molecular bioSystems *8*, 22-26.

Packer, L.E., Song, B., Raleigh, D.P., and McKnight, C.J. (2011). Competition between intradomain and interdomain interactions: a buried salt bridge is essential for villin headpiece folding and actin binding. Biochemistry *50*, 3706-3712.

Pechmann, S., Levy, E.D., Tartaglia, G.G., and Vendruscolo, M. (2009). Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. Proceedings of the National Academy of Sciences of the United States of America *106*, 10159-10164.

Receveur-Brechot, V., and Durand, D. (2012). How random are intrinsically disordered proteins? A small angle scattering perspective. Current protein & peptide science *13*, 55-75.

Ritchie, D.W., and Kemp, G.J. (2000). Protein docking using spherical polar Fourier correlations. Proteins *39*, 178-194.

Rocchia, W., Alexov, E., and Honig, B. (2001). Extending the Applicability of the Nonlinear Poisson−Boltzmann Equation: Multiple Dielectric Constants and Multivalent Ions†. The Journal of Physical Chemistry B *105*, 6507-6514.

Romero, P., Obradovic, Z., Li, X., Garner, E.C., Brown, C.J., and Dunker, A.K. (2001). Sequence complexity of disordered protein. Proteins *42*, 38-48.

Rost, B. (1999). Twilight zone of protein sequence alignments. Protein engineering *12*, 85-94.

Russo, A.A., Jeffrey, P.D., Patten, A.K., Massague, J., and Pavletich, N.P. (1996). Crystal structure of the p27Kip1 cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex. Nature *382*, 325-331.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., and Serrano, L. (2005). The FoldX web server: an online force field. Nucleic acids research *33*, W382-388.

Selzer, T., and Schreiber, G. (1999). Predicting the rate enhancement of protein complex formation from the electrostatic energy of interaction. Journal of molecular biology *287*, 409-419.

Sheinerman, F.B., and Honig, B. (2002). On the role of electrostatic interactions in the design of protein-protein interfaces. Journal of molecular biology *318*, 161-177.

Sheinerman, F.B., Norel, R., and Honig, B. (2000). Electrostatic aspects of protein-protein interactions. Current opinion in structural biology *10*, 153-159.

Shoemaker, B.A., Portman, J.J., and Wolynes, P.G. (2000). Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. Proceedings of the National Academy of Sciences of the United States of America *97*, 8868-8873.

Shrake, A., and Rupley, J.A. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. Journal of molecular biology *79*, 351-371.

Sickmeier, M., Hamilton, J.A., LeGall, T., Vacic, V., Cortese, M.S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V.N.*, et al.* (2007). DisProt: the Database of Disordered Proteins. Nucleic acids research *35*, D786-793.

Sines, J.J., Allison, S.A., and McCammon, J.A. (1990). Point charge distributions and electrostatic steering in enzyme/substrate encounter: Brownian dynamics of modified copper/zinc superoxide dismutases. Biochemistry *29*, 9403-9412.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. Bioinformatics *21*, 3940-3941.

Song, J., Guo, L.W., Muradov, H., Artemyev, N.O., Ruoho, A.E., and Markley, J.L. (2008). Intrinsically disordered gamma-subunit of cGMP phosphodiesterase encodes functionally relevant transient secondary and tertiary structure. Proceedings of the National Academy of Sciences of the United States of America *105*, 1505-1510.

Sugase, K., Dyson, H.J., and Wright, P.E. (2007). Mechanism of coupled folding and binding of an intrinsically disordered protein. Nature *447*, 1021-1025.

Swets, J.A., Dawes, R.M., and Monahan, J. (2000). Psychological Science Can Improve Diagnostic Decisions. Psychological Science in the Public Interest *1*, 1-26.

Tompa, P., and Fuxreiter, M. (2008). Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. Trends in biochemical sciences *33*, 2-8.

Tusnady, G.E., and Simon, I. (2001). The HMMTOP transmembrane topology prediction server. Bioinformatics *17*, 849-850.

Vacic, V., Oldfield, C.J., Mohan, A., Radivojac, P., Cortese, M.S., Uversky, V.N., and Dunker, A.K. (2007). Characterization of molecular recognition features, MoRFs, and their binding partners. Journal of proteome research *6*, 2351-2366.

Walshaw, J., and Woolfson, D.N. (2001). Socket: a program for identifying and analysing coiled-coil motifs within protein structures. Journal of molecular biology *307*, 1427-1450.

Wang, W., and Kollman, P.A. (2000). Free energy calculations on dimer stability of the HIV protease using molecular dynamics and a continuum solvent model. Journal of molecular biology *303*, 567-582.

Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. Journal of molecular biology *337*, 635-645.

Wilkins, D.K., Grimshaw, S.B., Receveur, V., Dobson, C.M., Jones, J.A., and Smith, L.J. (1999). Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques. Biochemistry *38*, 16424-16431.

Wright, P.E., and Dyson, H.J. (2009). Linking folding and binding. Current opinion in structural biology *19*, 31-38.

Xu, D., Tsai, C.J., and Nussinov, R. (1997). Hydrogen bonds and salt bridges across protein-protein interfaces. Protein engineering *10*, 999-1012.

Zhou, H.X. (2012). Intrinsic disorder: signaling via highly specific but short-lived association. Trends in biochemical sciences *37*, 43-48.