HYDROGEN BOND ENERGY-BASED COMPARATIVE ANALYSIS OF PROTEIN-LIGAND INTERACTIONS AND SIMILARITY ASSESSMENT OF PROTEIN-DNA COMPLEX MODELS

by

Fareeha Kanwal Malik

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Bioinformatics and Computational Biology

Charlotte

2022

Approved by:

_____
Dr. Jun-tao Guo

_____
Dr. Way Sung

_____
Dr. Jennifer Weller

_____
Dr. Shan Yan

ABSTRACT

FAREEHA KANWAL MALIK. Hydrogen bond energy-based comparative analysis of protein-ligand interactions and similarity assessment of protein-DNA complex models. (Under the direction of DR. JUN-TAO GUO)

Hydrogen bonds play a vital role in protein-DNA interactions. In particular, side chain-base hydrogen bonds are crucial to the binding specificity between protein and DNA. Mutations effecting interface hydrogen bonds in protein-DNA complexes have been linked to changes in binding specificity and are implicated in various diseases. However, knowledge about the distribution of hydrogen bond energy (HBE) in protein-DNA complexes as compared to other important biomolecular complexes is unknown. Here, we performed a systematic comparative analysis of hydrogen bond energy (HBE) in three protein-ligand complexes; protein-DNA, protein-protein and protein-peptide. Our results show that while the hydrogen bonds in protein-protein and protein-peptide complexes are predominantly strong, a unique, almost equal distribution of strong and weak hydrogen bonds is observed in protein-DNA complexes. More importantly, more strong hydrogen bonds are observed in the minor grooves of highly specific protein-DNA complexes than multispecific complexes indicating the role of minor groove hydrogen bonds in protein-DNA binding specificity. The knowledge gained from these analyses was applied to develop a novel hydrogen bond energy-based method to assess the similarity between protein-DNA complex models and reference structures, an important step towards computational prediction of complex structures. We show that HBE based method provides more accurate assessment of similarity for models generated by both homology modeling and computational docking methods.

DEDICATION

For my late mother, Kalsum Akhtar, who was my best friend, my first teacher and my biggest inspiration.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Jun-tao Guo for his constant support and guidance throughout my Ph.D. His expertise, insightful feedbacks, patience yet extremely enthusiastic and encouraging mentorship has helped shape my skills as a scientist. I would also like to offer sincere gratitude to my committee members, Dr. Jennifer Weller, Dr. Way Sung and Dr. Shan Yan for their support, feedback and suggestions. I would also like to thank Dr. Donald Jacobs for training me in the beginning of my Ph.D. I am very thankful to all the current and past members of Guo lab, especially Dr. Alvin Farrel who introduced me to the lab and Dr. Maoxuan Lin for the productive discussions and team work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

HBE	Hydrogen bond energy
HBall	All hydrogen bonds
HBSP	Hydrogen bonds between specific amino acid sidechain and nucleotide base
iSA	Interface surface area
BB-BB	Backbone-backbone hydrogen bonds
SC-SC	Sidechain-sidechain hydrogen bonds
SC-BB/ BB-SC	Sidechain-backbone or Backbone-sidechain hydrogen bonds
HS	Highly specific protein-DNA complexes
MS	Multispecific protein-DNA complexes
Fnat	Fraction of native contacts
iRMSD	Interface root mean square deviation
lRMSD	Ligand root mean square deviation
MP2	Møller−Plesset theory (MP2)

MD	Molecular dynamics
PD	Protein-DNA complexes
PT	Protein-peptide complexes
PP	Protein-protein complexes
RDPD	Rigid docking protein-DNA complexes
PHDL	Protein homo/heterodimer library
RDPP	Rigid docking protein-protein complexes
PDnrall	Non-redundant protein-DNA complexes
PPnrall	Non-redundant protein-protein complexes
PTnrall	Non-redundant protein-peptide complexes
FIRST	Floppy inclusion and rigid substructure topography
CAPRI	Critical assessment of predicted interactions
PDB	Protein data bank
WHB	Weighted hydrogen bond

CHAPTER 1: INTRODUCTION

1.1. Background

Proteins bind to DNA with varying degrees of specificity ranging from highly specific, where proteins bind to specific DNA sequences, to non-specific where proteins can bind to variety of DNA sequences [1–6]. Previous studies have shown that structural or genetic mutations causing alterations in binding specificity of protein-DNA complexes can have serious medical consequences such as cancer [2,7–9]. In tumor suppressor transcription factor-DNA complexes, such as p53 and leucine zipper family, changes in binding specificity have been linked to problems with stem cell maintenance and differentiation and metastasis of tumor cells [3,10–12]. It is, therefore, crucial to consider binding specificity of protein-DNA complexes in various applications such as drug design or interface design of inhibitor-ligand complexes [11,13].

Binding specificity in protein-DNA complexes is achieved through two readout mechanisms; direct/base and indirect/shape [14–20]. Direct or base readout is achieved through several interactions such as hydrogen bonds, electrostatic interactions and hydrophobic interactions while indirect or shape readout of DNA by proteins is facilitated by changes in the shape and conformation of DNA [14–20]. Hydrogen bonds are weak electrostatic interactions which play a central role in protein-DNA binding specificity due to their directional nature [21–24]. While the role of hydrogen bonds in protein-DNA interactions has been extensively studied, knowledge about the energy distribution of hydrogen bonds and its application in structure prediction of protein-DNA complexes is under-explored [25–28]. In this dissertation, we provide an insight into the binding specificity of protein-DNA interactions based on a comparative analysis of energy distribution of hydrogen bonds in protein-ligand complexes. The knowledge gained from

this analysis is then applied to develop a novel approach to assess the similarity of computationally predicted models of protein-DNA complexes to their reference structure.

1.2. Hydrogen bonds

A hydrogen bond is a weak interaction in which a hydrogen atom is shared between two highly electronegative atoms, a donor atom and an acceptor atom. The donor atom carries a partial negative charge and is covalently bonded to a hydrogen atom whereas the acceptor atom has a lone pair of electrons and carries a partial negative charge. This electrostatic interaction creates a dipole, orienting hydrogen atoms toward the lone pair of acceptor atoms, providing directionality to hydrogen bonds [29–31]. In addition to electrostatic interactions, the directionality is also explained by partial covalent character of hydrogen bonds which results from the penetration of positively polarized hydrogen atom into the van der Waal's sphere of acceptor atom [32]. The orientation of hydrogen bonds is addressed through the angle and distance estimation between the interacting partners (Fig. 1.1.).

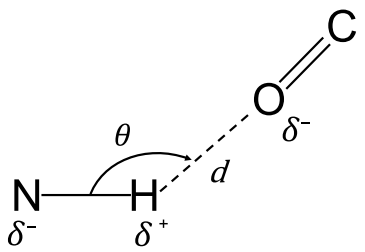

Figure 1.1. Hydrogen bond geometry representation between N and O atoms carrying partial negative charges. θ is the angle and d is the distance between interacting partners.

1.2.1. Energy of hydrogen bonds

The strength of hydrogen bond depends on geometry, nature and environment of the interacting atoms. Stronger hydrogen bonds form between neighboring atoms separated by smaller

interatomic distances. The optimal distance for a strong hydrogen bond in the biological systems is between 2.6 Å and 3.6 Å [33]. Quantum mechanics studies have shown that linear arrangement of donor, acceptor and hydrogen atom is ideal for a strong hydrogen bond. Hydrogen bonds within proteins are generally weaker because of the solvation effects. The energy of isolated hydrogen bonds in proteins is lower (~ -5-6 kcal/mol) as compared to that of proteins in solutions (-0.5 kcal/mol to -4.7 kcal/mol) because they do not have to compete with the solvation effects [34].

Estimating the energy of hydrogen bonds in a complex is a challenging task. Several quantum mechanical studies such as hybrid density functional theory and second-order Møller−Plesset theory (MP2) are used to estimate the energy of hydrogen bonds [35–44]. However, the application of these methods on larger biomolecular complexes such as protein-DNA complexes is complicated [38]. Several approximations, assumptions and simplifications are required for such purposes [45–47]. For example, a simplified version to estimate the hydrogen bond energy involves simply counting the number of donor and acceptor atoms in the complexes. One of the most widely used models to estimate hydrogen bond energy was proposed by Dahiyat et. al. which accounts for different hybridization states of the donor and acceptor atoms through the angle term [48] .

1.2.2. Hydrogen bonds in protein-ligand interactions

Hydrogen bonds play an important role in stability of monomeric proteins and nucleic acids structure, and the specificity of protein-ligand interactions [49–54]. In proteins, the folding of amino acid residues into secondary structures, such as alpha helices and beta sheets, is primarily facilitated by backbone hydrogen bonds [55–57]. The failure of a buried polar group to form hydrogen bonds results in destabilization of the structure [57]. In DNA, specific hydrogen bonds

between complementary nucleotides are central to the structure of double helix. Two hydrogen bonds are formed between adenine and thymine whereas three hydrogen bonds are formed between cytosine and guanine. Incorrect base pairing between nucleotides, such as cytosine-thymine instead of cytosine-guanine could cause a ten-fold difference in energy of DNA, resulting in destabilization[58]. In protein-ligand interactions, the directional nature of hydrogen bonds plays an important role to facilitate the ligand binding selectivity [25,59]. They are the second most frequent type of interaction observed in experimentally solved structures of protein-ligand complexes after hydrophobic interactions [60]. Hydrogen bonds are vital in enzyme catalysis and drug-target interactions. Introducing a hydrogen donating group in thrombin inhibitors, which act as anticoagulants, results in a remarkable increase in their binding affinity [61]. Mutations causing a decrease in the number of hydrogen bonds in several ribonuclease enzymes destabilize the structure of the complex [57,62–66]. Alterations in hydrogen bonding framework of protein-ligand complexes has shown to affect their function.

Hydrogen bonds mediate different types of protein-ligand interactions through differences in their geometry and packing at the interface. For instance, the binding interface in protein-protein complexes with stable monomeric partners in the unbound form exhibits dual features. It acts as protein surface in the unbound form and protein-interior in the bound form. This is thought to be achieved through a large number of less geometrically optimal hydrogen bonds at the binding interface than the protein interior [67–69]. On the other hand, small number of geometrically optimal hydrogen bonds are observed at the interface of permanent protein-protein complexes, where one of the partner is intrinsically unstructured in the unbound form [70]. The interfaces of protein-peptide complexes is also closely packed with hydrogen bonds to stabilize the otherwise unstructured peptide [34,71]. In protein-DNA complexes, In protein-DNA interactions, directional

nature of hydrogen bond plays a vital role in providing the shape complementarity for indirect readout [72]. They are major type of interaction involved in recognition, specificity and stability of the complex [73–76]. The hydrogen bonds formed between amino acid side chains and nucleotide base edges in particular have major role in binding specificity. Bidentate hydrogen bonds, where two hydrogen bonds are formed with a base or base pair, also play important roles in specific protein-DNA interactions [73]. Highly specific protein-DNA complexes contain more hydrogen bonds than multispecific and non-specific [5].

1.3. Computational modeling of protein-ligand complexes

Knowledge about the structure of protein-DNA complexes is important to gain insight into their function and is crucial in applications such as binding site prediction, drug-target interaction and rational drug design [76]. Despite recent advances in structural biology, the knowledge about the structure of protein-DNA complexes is incomplete [77–81]. Several technical challenges, such as solving phase problem in X-ray diffraction studies hinder successful structure determination of larger complexes involving nucleic acids [82–86]. Moreover, all the experimental methods are time consuming which limits their use in time-sensitive applications such as screening a large number of inhibitors in drug design. Therefore, *in-silico* methods for prediction of complex structures are being explored [87–92]. Two approaches of computational structure prediction are widely used to predict models of complexes; homology modeling and docking.

1.3.1. Homology modeling

Homology modeling of protein-DNA complexes, also referred to as comparative modeling or template-based modeling, employs the evolutionary principal; similar protein sequences share similar structures [93–95]. Conventional homology modeling approaches involve iteration of four

steps until the best model is predicted. First, an evolutionarily related homolog or set of homologs for the target sequence, called template(s), is identified by searching through an existing database, usually the PDB [96–99]. Second, an alignment of the query and template sequence is generated using any of the several different methods of sequence alignment [100–104]. Third, the three-dimensional atomic coordinates of the template and their corresponding alignments are used to generate a model of the query sequence [105–110]. Homology modeling of protein-DNA complexes requires some additional steps. Structural alignment of models predicted using aforementioned steps is performed to the native structure followed by complexing the model with the interacting DNA.

While the homology modeling methods have shown considerable success in drug-discovery among other important applications, it has several limitations [111–115]. Selection of an evolutionarily related template of high structural quality is crucial. A homology model is not considered reliable if no template with more than 25% sequence identity is found [116]. This is a major limitation because about 60% of DNA binding proteins do not have an experimentally solved homolog. Sequence alignment is another major bottleneck which effects the quality of predicted structure. A misplaced gap in the sequence alignment can put two otherwise adjacent residues 40 Å apart, resulting in an incorrect model [117]. Finally, two highly similar target-template complexes could potentially have different interaction interfaces [118]. New methods are being developed which use interaction interface as template to model novel protein-DNA interactions.

1.3.2. Docking

Docking methods predict structure of the complex by exploring the free energy landscape and identifying the lowest energy near-native binding state of the interacting partners [119]. The docking algorithms can be classified into two groups; rigid docking and flexible docking. Rigid docking algorithms sample the relative positions between protein and DNA while keeping the conformations of both protein and DNA molecules unchanged. Flexible docking algorithms, on the other hand, also consider the conformational changes of protein and DNA while sampling different positions between protein and DNA [120,121]. Semi-flexible docking methods explore the flexibility of the smaller, ligand molecule while keeping the larger receptor molecule rigid.

There are two main steps in all docking algorithms: sampling and scoring. The first step of docking involves exploring the entire energy landscape forming the conformational space of the complex and identifying the energy minima. This process is repeated until the identification of several "docked" poses of the complex. The second step involves evaluation of docked poses using a mathematical scoring function based on different parameters such as free energy of binding or a machine-learning based objective function based on different structural features [122].

Even though docking methods have shown success in practical applications, they are not devoid of limitations [123–136]. The selection of the best model is a daunting task because the energy minima of a complex explored by docking algorithm can account for a large number of possible solutions [137]. While a false negative model can result in missing the correct results whereas selection of a false positive model could lead to catastrophic results in drug development due to the underlying financial and liability cost. One way to filter false positives is to run a short molecular dynamics (MD) simulation. However, the MD simulations are very time consuming and computationally expensive specially in case of large number of macromolecules. Additionally, similar to homology models, this method is also dependent on the availability of experimentally

solved structures of the binding partners. An ideally designed docking algorithm explores all available degrees of freedom in both the unbound components of the complex to identify the best docked pose. This is computationally very expensive [123,138]. Another major limitation of docking algorithm is the scoring problem, the ability to distinguish between a near-native model and decoy. An accurate scoring function considers complex physical phenomenon such as entropy and electrostatic interaction. However, these calculations make scoring computationally expensive and time consuming. The existing scoring functions trade-off between computational cost and accuracy through over-simplification and assumptions of several physical phenomena.

1.4. Summary

Existing computational modeling methods of complexes require several improvements to accurately capture the structure of complexes. Modifications in existing methods or development of new methods is needed to tackle the challenges associated with predicting structures of complexes. In order to test the performance of novel structure prediction methods, the similarity between the structures of predicted complex models and reference structure needs to be accurately assessed. While several similarity assessment criteria exist for monomeric protein structures, no such metric exists for protein-ligand complexes which accurately considers the biologically relevant features to measure similarity. It is challenging to capture a variety of biologically relevant features unique to different types of protein-ligand interactions through existing metrics. Development of tailored methods for different types of protein-ligand complexes could potentially provide more accurate assessment of similarity. As discussed above, the role of hydrogen bonds is slightly different in different types of complexes and especially in protein-DNA complexes. The goal of this dissertation is to develop a hydrogen bond-based method for similarity assessment of protein-DNA complex models with their reference structure. However, instead of considering a

single distance or energy-based threshold to define hydrogen bonds, we aim to consider the overall energy distribution of hydrogen bonds at the interface of protein-DNA complexes. However, the role of energy of hydrogen bonds in protein-DNA complexes as compared to other types of protein-ligand interactions is not well understood. To this end, we have performed a systematic comparative analysis of energy of hydrogen bonds between protein-protein, protein-peptide and protein-DNA complexes. The knowledge gained by this analysis is implemented in the development of our novel similarity assessment method for protein-DNA complexes.

CHAPTER 2. INSIGHTS INTO PROTEIN-DNA INTERACTIONS FROM HYDROGEN BOND ENERGY-BASED COMPARATIVE PROTEIN-LIGAND ANALYSES

This project was originally published in Proteins: Structure, Function and Bioinformatics (https://onlinelibrary.wiley.com/doi/10.1002/prot.26313)

Hydrogen bonds play important roles in protein folding and protein-ligand interactions, particularly in specific protein-DNA recognition. However, the distributions of hydrogen bonds, especially hydrogen bond energy in different types of protein-ligand complexes, is unknown. Here we performed a comparative analysis of hydrogen bonds among three non-redundant datasets of protein-protein, protein-peptide and protein-DNA complexes. Besides comparing the number of hydrogen bonds in terms of types and locations, we investigated the distributions of hydrogen bond energy. Our results indicate that while there is no significant difference of hydrogen bonds within protein chains among the three types of complexes, interfacial hydrogen bonds are significantly more prevalent in protein-DNA complexes. More importantly, the interfacial hydrogen bonds in protein-DNA complexes displayed a unique energy distribution of strong and weak hydrogen bonds whereas majority of the interfacial hydrogen bonds in protein-protein and protein-peptide complexes are of predominantly high strength with low energy. Moreover, there is a significant difference in the energy distributions of minor groove hydrogen bonds between protein-DNA complexes with different binding specificity. Highly specific protein-DNA complexes contain more strong hydrogen bonds in the minor groove than multi-specific complexes, suggesting important role of minor groove in specific protein-DNA recognition. These results can help better understand protein-DNA interactions and have important implications in improving quality assessments of protein-DNA complex models.

2.1. Introduction

Proteins interact with DNA, peptides and other proteins to form macromolecular assemblies that carry out fundamental and essential biological functions [139]. Protein-DNA (PD) complexes, for example, play critical roles in regulation of gene expression, histone packaging, DNA replication, repair, modification and recombination [140]. The interactions between protein and DNA display different degrees of specificity that ranges from highly specific to non-specific [5]. Protein-peptide (PT) interactions account for up to 40% of cellular interactions and are involved in mediating signal transduction, regulating apoptotic pathways and immune responses [141–143]. Protein-protein (PP) interactions form essential complexes like hormone-receptor, antibody-antigen, and protease-inhibitor, which control cell signaling, electron transport, signal transduction, and cell metabolism [144]. Disruptions in these interactions can cause serious medical conditions such as cancer, cardiovascular and neurodegenerative disorders [2,7,9,144]. Knowledge of detailed interactions among these complexes at atomic resolution is therefore essential to understanding the underlying mechanisms that govern biochemical processes. It also has important implications in biomedical applications such as protein-ligand docking, *in-silico* design of inhibitors and interfaces, and virtual screening of drugs library in pharmaceutical industry.

Hydrogen bonds (HBs) play key roles in conferring binding specificity of macromolecular complexes [56,68,145,146]. An HB is generally considered as a weak, electrostatic interaction between a polar acceptor atom that carries a lone pair of electrons and a hydrogen atom that is covalently linked to a polar atom, oriented toward each other at an equilibrium distance. This orientation and distance dependent nature of hydrogen bonds is vital in providing the shape and chemical complementarity for selective recognition and binding of complexes [145]. In PD

complexes, for example, HBs play a key role in DNA base readout by proteins and act as the major contributor to binding specificity that is vital for the biomolecular function of protein-DNA complexes [147]. The recognition of DNA by proteins is guided by an innate hydrogen bonding pattern that generates an initial unstable non-specific, intermediate complex with high energy [3,148–150]. While most of this recognition is expected to occur through the signature hydrogen bonding pattern of major groove, many DNA binding proteins also bind to the minor groove through hydrogen bonding and shape readout [22,147]. Later, this complex transitions to a stable and highly specific low energy state through reversible structural deformations that are also guided by a specific HB pattern [145]. In PP complexes, HBs influence stability as well as binding specificity at the interface [68]. Interfacial hydrophilic side chains of a PP complex have a high charge density that is stabilized primarily through hydrogen bonding. Buried polar atoms at the interface not involved in hydrogen bonding may destabilize the complex [151–154]. Peptide binding, on the other hand, utilizes HBs to improve interface packing density as well as minimize the entropic cost of transitioning from a highly flexible, unstructured peptide to a well-defined rigid structure in a complex with protein [71]. On average, PT interface contains more HBs per 100 $Å^2$ interface area when compared to PP interface and PT interface HBs generally are more linearly oriented [71]. In addition to binding, HBs are the primary driving force in folding of protein chains into core secondary structures such as alpha helices and beta sheets and base pairing in nucleic acids [56]. HBs also bring flexibility to the structure, which is central to the dynamic nature of proteins and plays a key role in allosteric, catalytic, and binding activities [56,155].

The role of hydrogen bonds in binding and folding of complexes has previously been studied as individual cases as well as a group of cases [149,156–159]. Mandel-Gutfreund *et al.* studied different types of hydrogen bonds at the interface of 28 X-ray crystal structures of protein-

DNA complexes. The hydrogen bonds were classified according to the types of donor and acceptor atoms, such as backbone, sidechain or base edges [146]. Xu *et al*. performed a similar analysis on 319 protein-protein complexes [68]. London *et al*. compared the types of hydrogen bonds at the interface and within protein chains of 103 protein-peptide complexes. They further compared the types of hydrogen bonds in protein-peptide complexes to those in protein-protein complexes [71]. Rawat and Biswas in 2011 performed a comparison of HBs along with several other structural features to investigate the role of flexibility in protein-DNA, protein-RNA and protein-protein complexes [21]. Jiang *et. al*. demonstrated that in protein-protein complexes, the average energy contribution of a hydrogen bond is ~30% [160]. Zhou and Wang recently compared short hydrogen bonds, where donor-acceptor distance is less than 2.7Å, in1663 high quality protein, protein-ligand and protein-nucleic acid structures [161]. Itoh *et al*. showed that the interaction energy of even the weaker N^{+}-C-H···O hydrogen bonds is comparable to other protein-ligand interactions such as π/π interactions suggesting the importance of considering HB energy in drug design [28].

While analyses based on the number of hydrogen bonds with a single energy cutoff or a distance/angle cutoff can provide useful information about the role of hydrogen bonds in protein-ligand interaction, they have an intrinsic flaw since strong and weak hydrogen bonds are treated equally. Moreover, the distributions of interfacial hydrogen bonds in terms of HB strength or HB energy in protein-ligand complexes, and more importantly, the distributions of interfacial HB energy among different types of protein-ligand complexes remain unknown. To address these issues, in this study we performed a holistic statistical comparative analysis of hydrogen bonds across interfaces and within protein chains (intrachain) among PP, PT and PD complexes to get an insight into their roles in each type of complexes. In addition to comparing the types and locations of hydrogen bonds in each type of complexes, we investigated the HB energy distributions and

found significant differences among these three types of complexes, especially a unique pattern in protein-DNA complexes. To the best of our knowledge, an HB energy based large-scale comparison of macromolecular complexes has never been explored before.

2.2. Materials and methods

2.2.1 Datasets

Seven previously published and widely used datasets of protein-DNA, protein-peptide and protein-protein complexes were selected, including three datasets of protein-DNA complexes: highly specific (HS), multi-specific (MS) [5], and rigid docking protein-DNA (RDPD) complexes [162]; two protein-peptide complex datasets: LEADS-PEP [163] and InterPep [164]; and two datasets for protein-protein complexes: an updated M-TASSER dimer library [165] and the protein-protein Docking benchmark (RDPP, version 5) (Table 2.1.) [166] . Since the M-TASSER dimer library was published over 10 years ago, we generated an updated dataset, called protein homo/heterodimer library (PHDL) using some of the guidelines described in the original paper (Supplementary Table S1).

Each of the three datasets for PD represents a specific category of protein-DNA complexes. The HS dataset comprises 29 PD complexes with high binding specificity between protein and DNA whereas the MS dataset comprises 104 cases, in which proteins can bind to multiple conserved DNA sequences [5]. The RDPD dataset consists of 38 highly diverse non-redundant TF-DNA complexes that cover 11 structural folds, 15 super-families and 28 families [162].

Table 2.1. The protein-DNA, protein-peptide and protein-protein datasets

Types	Datasets	Number of complexes	Experimental method and selection criteria	Ligand	Average interface area
Protein-DNA	Highly Specific	28	X-ray (<= 3 Å) R-factor < 0.3	Double stranded DNA	~1100 Å^2
	Multi-specific	105	X-ray (<= 3 Å) R-factor < 0.3	Double stranded DNA	~700 Å^2
	Rigid docking	38	X-ray (<= 3 Å)	Double stranded DNA	~1100 Å^2
Protein-Peptide	InterPep	502	X-ray (<= 3 Å) or NMR	5-25 residues	~665 Å^2
	LEADS-PEP	53	X-ray < 2Å, R-factor < 0.3	3-12 residues	~512 Å^2
Protein-Protein	Protein homo/hetero dimer library	2608	X-ray (<= 3 Å)	>40 residues per protein chain	~1374 Å^2
	Docking Benchmark V5	230	X-ray (<=3.25 Å)	>= 30 residues per protein chain	~1847 Å^2

The two PT complex datasets differ mainly in the peptide chain lengths. InterPep comprises protein complexes with peptides ranging from 5 to 25 amino acids whereas peptides in LEADS-PEP are 3-12 amino acids long [163,164]. InterPep is a larger dataset with 502 X-ray and NMR structures, which was originally developed for testing a peptide-binding site prediction pipeline [164]. LEADS-PEP, on the other hand, is a much smaller dataset with 53 carefully curated and widely used complexes designed specifically for peptide-based therapeutics and peptide docking. It contains only X-ray crystal structures with a resolution better than 2Å [163].

The complexes in the PP datasets differ mainly in size and definition of interaction unit. The protein-protein docking benchmark (RDPP) has 230 complex structures that were experimentally solved with corresponding unbound components available [166]. The structures in the RDPP dataset represent a diverse combination of antigen-antibody, enzyme-substrate, enzyme-

regulatory complex, GPCR proteins and several other classes of proteins. The docking benchmark defines a true interaction as one that has functional significance as identified in the literature and agreed upon by the scientific community. The second PP dataset PHDL, a protein homo/hetero dimer library, determines the oligomeric state from PDB files [167]. PHDL contains non-redundant heterodimers (Supplementary Table S1A) and homodimers (Supplementary Table S1B), where no two chains share more than 30% sequence identity with each other and each interacting partner has at least 40 amino acids.

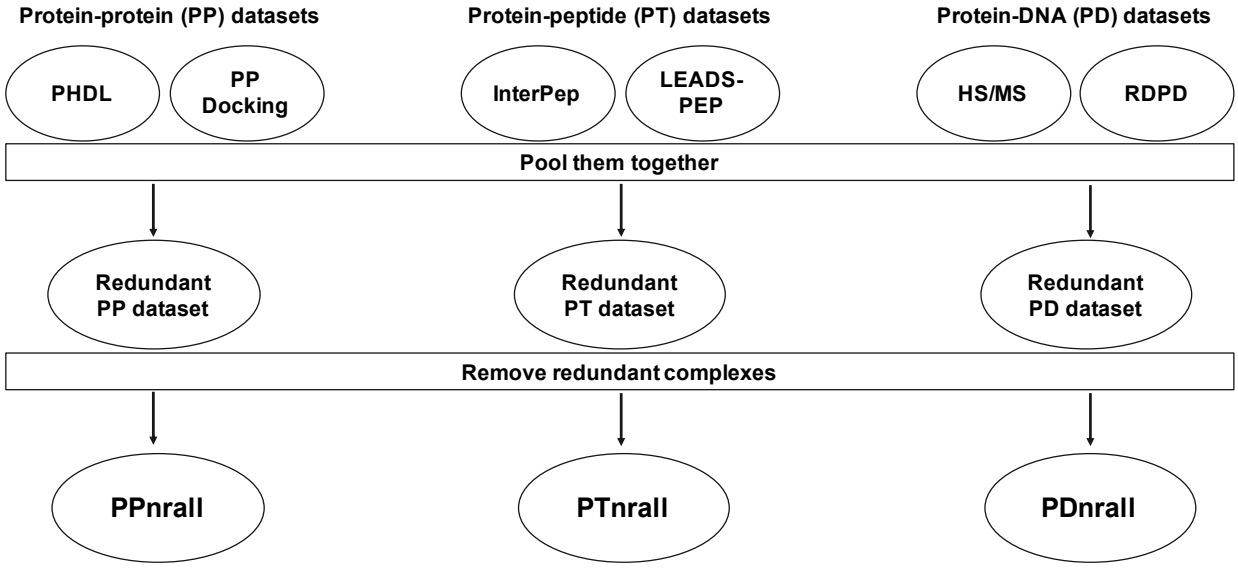

Figure 2.1. A flow chart for generating non-redundant datasets of protein-protein, protein-peptide and protein-DNA complexes.

In addition to these individual datasets, we pooled the datasets of the same type of complexes together and generated three larger, non-redundant and highly diverse datasets (Figure 2.1): (i) PDnrall, a protein-DNA dataset comprising HS, MS and RDPD; (ii) PTnrall, a protein-peptide dataset comprising LEADS-PEP and InterPep; and (iii) PPnrall, a protein-protein dataset comprising PHDL and RDPP. The redundancy after combining the respective datasets was removed with PISCES using a sequence identity cutoff of 30% [168], which resulted in 2724 non-

redundant protein-protein complexes (PPnrall), 346 non-redundant protein-peptide complexes (PTnrall) and 126 non-redundant protein-DNA complexes (PDnrall).

2.2.2. Dataset processing

The datasets were filtered rigorously for accurate analysis. In case of multiple models for one native structure as in the NMR entries, only the first model was selected. All the heteroatoms, including water molecules were removed since we do not consider solvation effects for the sake of simplicity and fair comparison. Proteins that have residues with insertion codes were renumbered accordingly. Since considering the alternate locations of a residue in an experimentally solved crystal structure may result in over counting the number of HBs, only the state with the highest occupancy for a given residue was included for analysis. The complexes with internal missing residues, i.e., residues that are not on the N or C terminal of the chain were discarded. Lastly, interactions between proteins and ligands were calculated based on interaction units for complexes composed of multiple chains of proteins and ligands. For example, 4FQI protein unit has two chains H, L and the ligand unit has six chains A, B, C, D, E, and F. For such cases, we only considered the inter-unit interaction between protein and ligand. In the case of 4FQI, H and L were identified as one unit while ABCDEF as another unit.

2.2.3. Identification of HBs

Two widely used hydrogen bond annotation programs, FIRST (Floppy Inclusion and Rigid Substructure Topography) and HBPLUS, were used to identify HBs with default parameters [33,169]. Reduce was used to add hydrogen atoms to pdb files for FIRST HB calculations while HBPLUS calculates the hydrogen atom positions within the program [170]. FIRST employs an energy-based approach and the HB energy is calculated as in Eq. 2.1 [33,48]

$$E_{HB} = V_0 \left\{ 5\left(\frac{d_0}{d}\right)^{12} - 6\left(\frac{d_0}{d}\right)^{10} \right\} F(\theta, \phi, \varphi) \qquad \textit{Eq. 2.1}$$

Where d is the donor-acceptor distance. d_0 (2.8 Å) and V_0 (8 kcal/mol) represent the equilibrium distance and well-depth respectively [171]. The angle term $F(\theta,\phi,\varphi)$ is calculated based on the hybridization state of the acceptor and donor atoms, where θ is the donor-hydrogen-acceptor angle, ϕ is the hydrogen-acceptor-base angle, and φ is the angle between the normals of the planes defined by the six atoms attached to the sp^2 center as described by Dahiyat *et al.*[48]. The FIRST program was used for both the number of hydrogen bonds annotations using a widely used HB energy cutoff of -0.6 kcal/mol as well as for HB energy-based analysis. HBPLUS identifies HB with a distance-angle approach and defines the optimal distance between the donor and acceptor as 2.5 Å or smaller and the optimal angle as 90 degrees or higher [169].

2.2.4. Interface analysis and comparison

Since the interface sizes are different among different types of complexes (Table 2.1), in order to accurately assess the roles of HB at the interface of PP, PT and PD complexes, the numbers of HBs were compared with respect to the interfacial surface area. The interfacial surface area (iSA) of a complex, was calculated using NACCESS v2.1.1 with default parameters as shown in Eq. 2 [162,172]:

$$iSA = \frac{SA_P + SA_L - SA_C}{2} \qquad \textit{Eq. 2.2}$$

where SA_L and SA_P represent the surface area of protein and ligand respectively, and SA_C is the surface area of the protein-ligand complex. For multichain components, SA_P is the surface area of the protein unit while SA_C is the surface area of the ligand unit.

The HB distributions were compared at three different aspects: HB types, HB locations, and HB energy ranges. The types of HB were grouped depending on the types of atoms involved in hydrogen bonding, sidechain (or base in DNA) or backbone. HB types include SC-SC (representing side chain-side chain in PP and PT or sidechain-base in PD), BB-BB (for backbone-backbone) and Mixed type (for SC-BB or BB-SC). A union of all three types encompasses all hydrogen bonds (HBall). The SC-SC hydrogen bonds, also termed here as HBSP, are generally considered more specific in molecular recognition and binding as the backbone atoms are the same for each type of molecules, protein or DNA. There are two different HB location types, interface (between proteins and ligands) and intrachain (within proteins).

Table 2.2. Hydrogen bond energy (HBE) categories based on energy ranges

CATEGORY	**HBE RANGE (KCAL/MOL)**
I	$-0.6 \leq HBE < -0.1$
II	$-1.0 \leq HBE < -0.6$
III	$-1.5 \leq HBE < -1.0$
IV	$HBE < -1.5$

We divided hydrogen bond energy (HBE) from the FIRST program into four categories based on different energy cutoffs used in previous studies [33,34,148] and personal communication with the FIRST program developer as shown in Table 2.2.

2.2.5. Statistical tests:

Wilcoxon rank sum test was employed to assess if there are significant differences between samples across datasets. Chi-squared goodness of fit test was used to test the categorical distributions of types and the energy of hydrogen bonds at interface and within intrachain.

2.3. Results

2.3.1. Hydrogen bonds at the interface of complexes

We first compared the number of HBall and HBSP in PDnrall, PPnrall and PTnrall datasets. Based on HB annotations from FIRST with the widely used energy cutoff of -0.6 kcal/mol [34], we found that the number of interface HBall and the number of interface HBSP in PD complexes are significantly higher than those in the PP and PT complexes (Figure 2.2 A&B). The number of HBall and HBSP in PT complexes are significantly less than those in PP complexes (Figure 2.2 A&B). Results from HBPLUS are consistent with the data from FIRST except that the number of HBSP in PP complexes is larger than that in PD complexes with HBPLUS (Figure S1 A&B). Interestingly, when the FIRST energy cutoff is set at -0.1 kcal/mol, the results are more similar to the HBPLUS data (Figure S2 A&B).

Since the interface areas among the three types of complexes are different with PP complexes having the largest average interfacial area and PT complexes having the smallest average interfacial area (Table 2.1), comparing the raw number of interface HBs might be biased towards the complexes with a larger contact surface. Therefore, we normalized the number of interface hydrogen bonds, HBall and HBSP, by the interfacial surface area (iSA). Figure 2.2C and 2.22D show that both HBall/iSA and HBSP/iSA ratios of PD complexes are significantly higher

than those in the PP complexes and PT complexes. There is a clear pattern for the iSA normalized HBSP, PD> PP> PT. When the analyses were carried out with HBPLUS, the results are consistent with the results from FIRST (Figure S1). Even though no significant difference of the ratio HBall/iSA from FIRST is found between PP and PT complexes for a two-tailed test (Figure 2C), one tailed test with a null hypothesis that HBall/iSA in PP is not smaller than HBall/iSA in PT results a p-value of 0.043, which is in line with the result from HBPLUS as well as that from FIRST with an energy cutoff at -0.1 kcal/mol: the ratio of HBall/iSA in PT complexes is significantly higher than PP complexes (Figure S1C & S2C). These results are also in agreement with a previous study that PT interface has more total HBs per 100 $Å^2$ interface area than that in PP [71]. However, the HBSP/iSA ratio is the opposite, suggesting relatively fewer interface HBSP in PT complexes when the interface area is taken into consideration.

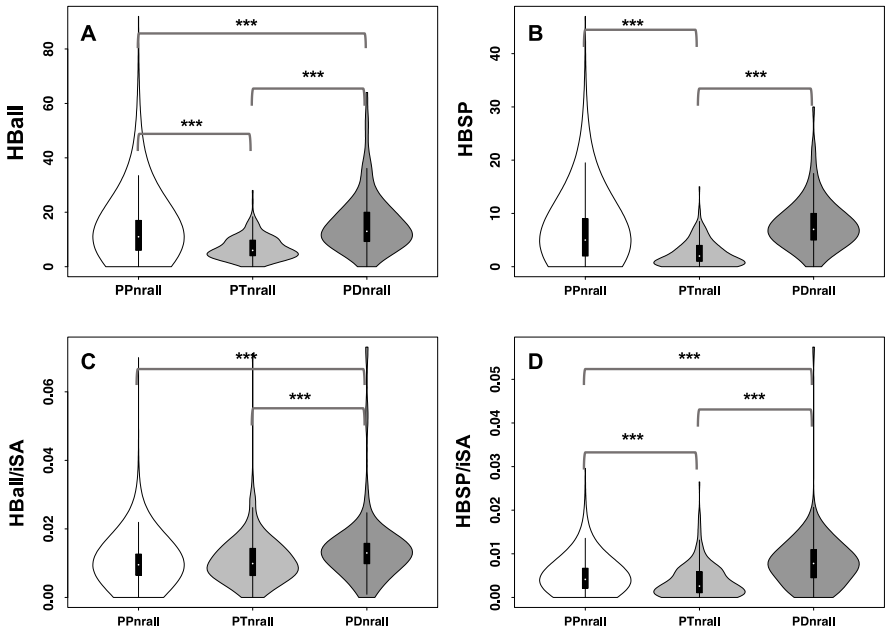

Figure 2.2. Comparison of interfacial hydrogen bonds based on FIRST with an energy cutoff of -0.6 kcal/mol: (A) the number of total hydrogen bonds (HBall); (B) the number of SC-SC or SC-Base hydrogen bonds (HBSP); (C) the ratio of HBall to interfacial surface area (iSA); and (D) the ratio of HBSP to iSA. *** = p-value ≤ 0.001, ** = p-value ≤ 0.01

2.3.2. Types of hydrogen bonds at interface and within intrachain

We compared the distributions of the HB types at complex interface or within protein (intrachain) in PP, PT and PD complexes and between individual complexes of the same type of complexes. Figure 2.3A and Table 2.3 show that there is no significant difference among the types of hydrogen bonds within proteins in all three types of complexes. BB-BB hydrogen bonds represent the largest number of overall hydrogen bonds within proteins (65-69%) followed by the Mixed (17-20%) and SC-SC (14-15%) hydrogen bonds respectively (Figure 2.3A). This is not surprising because the two major secondary structure types of the core protein structure, α-helices and β-sheets, are stabilized by backbone-backbone hydrogen bonds.

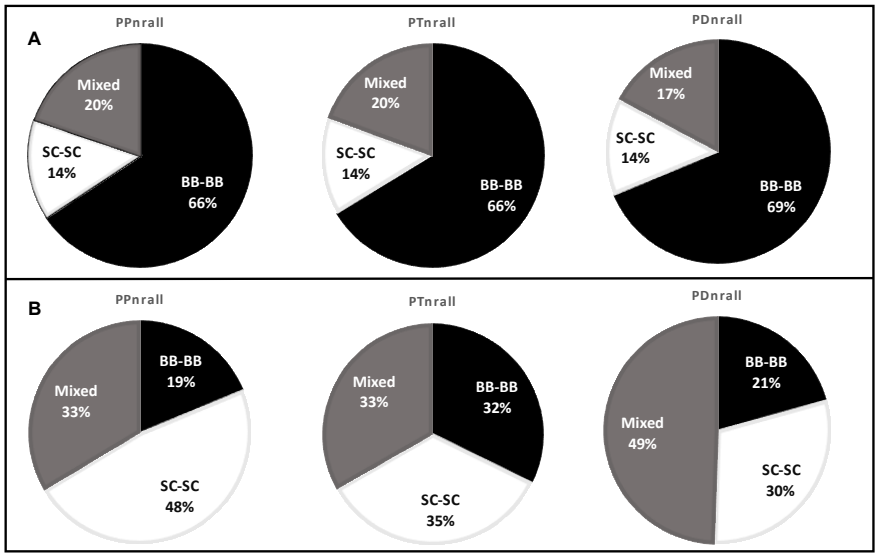

Figure 2.3. Comparisons of the distribution of different types of hydrogen bonds, backbone-backbone (BB-BB), sidechain-sidechain (SC-SC) and Mixed (BB-SC and SC-BB) for (A) intrachain within proteins and (B) at interface of PP, PT and PD complexes. The hydrogen bonds are annotated from the FIRST program with an energy cutoff of -0.6 kcal/mol.

The distributions of the hydrogen bond types at interface, however, are significantly different from the intrachain and among the three types of complexes (Figure 2.3B and Table 2.3).

The percentages of SC-SC hydrogen bonds at interface increase dramatically when compared with those within proteins while the BB-BB is the least type in all three complexes. The proportions of BB-BB hydrogen bonds at the interface are approximately one third of those from intrachain in PP and PD complexes and approximately half of that in PT complexes (Figure 2.3). The proportions of interface SC-SC HBs are at least twice more than those in intrachain in all three types of complexes. There is an increase of the Mixed HB type at interface when compared with intrachain. In PD complexes, the Mixed HB type consists of about half of all interfacial hydrogen bonds.

Table 2.3. p-values of chi-square tests between HB types from FIRST (-0.6 kcal/mol cutoff) and HBPLUS at interface and intrachain.

	Intrachain		**Interface**		**Interface/Intrachain**		
Dataset1/ Dataset2	**FIRST**	**HBPLUS**	**FIRST**	**HBPLUS**	**Dataset**	**FIRST**	**HBPLUS**
PPnrall, PDnrall	0.720	0.647	2.2e-16	0.025	PDnrall	<2.2e-16	<2.2e-16
PTnrall, PDnrall	0.874	0.945	0.002	0.0005	PPnrall	<2.2e-16	<2.2e-16
PTnrall, PPnrall	0.972	0.774	2.2e-16	<2.2e-16	PTnrall	8.904e-14	<2.2e-16

A previous study on protein-protein complexes indicated that the larger number of BB-BB hydrogen bonds within protein chains as compared to the interface is likely due to the differences in the degrees of freedom available to the corresponding atoms [68]. On both PP and PT interfaces, the highest proportion of HB types is SC-SC between interacting components while the percentage of BB-BB hydrogen bonds is the lowest. The percentage of interface BB-BB hydrogen bonds in PT complexes is higher than those in the PP and PD complexes. It has been suggested that a higher number of interface BB-BB hydrogen bonds in PT complexes is a result of bridging beta strands

at the interface between interacting peptides and protein molecules [71]. Once the interfacial beta-sheet containing complexes are removed from the dataset, BB-BB hydrogen bonds are comparable between PP and PT complexes [71]. Similar results were observed for the comparison of HB types annotated by HBPLUS and by FIRST with an energy cutoff of –0.1 kcal/mol (Table 2.3, Figure S3-S4 and Table S2) .

Besides comparisons among the three different types of non-redundant complexes, we also compared the distributions between individual datasets for each type of complexes (Figures S5-S6). For example, PHDL is composed of homodimers and heterodimers and the PD dataset has HS and MS complexes with different binding specificity. We found that there is no significant difference in the distribution of HB types for both intrachain and at interface between HS and MS (p-values of 0.3743 and 0.6685 respectively) as well as between homodimers and heterodimers (p-values of 0.9371 and 0.9746 respectively) from FIRST (Figure S5A). There is also no significant difference of HB type distributions for intrachain and at interface between PHDL and RDPP (p-values of 0.992 and 0.246 respectively). While there is no difference for the intrachain distributions between InterPep and LEADSPEP (p-value = 0.954), the interface distributions are different (p-value = 0.003) from FIRST HB annotations (Figure S6A). This might be a result of the relatively small LEADSPEP dataset with a small number of total hydrogen bonds (Figure 2.2). Similarly, no significant differences were found between any two of the above datasets of the same types of protein-ligand complexes based on HBPLUS annotations (Figures S5B and S6B).

2.3.3. Strength of hydrogen bonds at interface and within protein chain

We classified the strength of hydrogen bonds into four categories based on hydrogen bond energy from the FIRST program with different energy cutoffs used in previous studies as shown

in Table 2.2 [33,34,148]. For intrachain hydrogen bonds within proteins, no significant differences were found among the three types of complexes (Figure 2.4A and Table 2.4). Most of the hydrogen bonds (67-70%) are strong ones with lower than -1.5 kcal/mol energy (category IV) while very few of them are of intermediate energy (less than 15% when categories II and III are combined), suggesting that the hydrogen bonds in all types of proteins have similar energy distribution with predominantly strong hydrogen bonds.

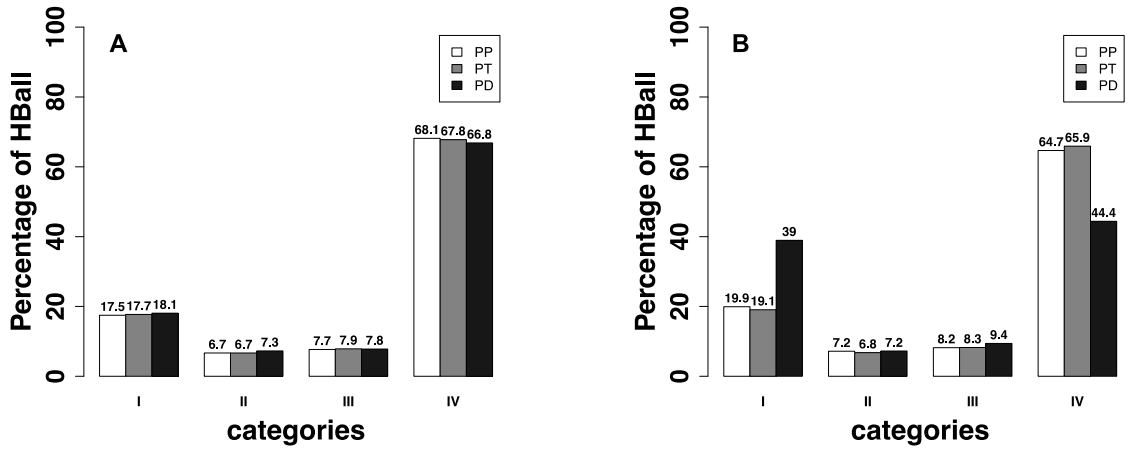

Figure 2.4. Comparisons of the distributions of hydrogen bond energy for (A) intrachain and (B) at interface.

To investigate if the energy categories are related to different HB types, we compared the distributions of each type of intrachain hydrogen bonds in each energy category (Figure 2.5A and Table S3). Similar trends for BB-BB, SC-SC and Mixed types were observed among the three types of complexes and there is no significant difference of intrachain hydrogen bond energy distribution for each HB type among the PP, PT, and PD complexes. There is a higher percentage of strong BB-BB hydrogen bonds in all complexes, but relatively fewer strong ones for the Mixed HBs, suggesting that the major secondary structure types patterned by the BB-BB hydrogen bonds are optimized in terms of both distance and angle and form strong hydrogen bonds. However, the

interface hydrogen bond energy distributions among different types of complexes are significantly different and exhibit a unique pattern for the PD complexes (Figure 2.4B and Table 2.4). There is a higher percentage of weak HB (category I) at PD complex interface when compared to those in PP and PT complexes as well as the intrachain HB energy in PD complexes. PD has the smallest percentage of strong HBs (category IV) among the three types of complexes. The difference between category I and IV HB percentage is much smaller in PD complexes (39% and 44.4%) than those in PP (18.9% and 66.2%) and PT (19.1% and 65.9%) complexes (Figure 2.4B). PP and PT complexes have similar distributions of interface HB energy categories. In addition, the interface and intrachain HB energy distributions in both PP and PT complexes are also similar (Table 2.4).

Table 2.4. p-values of chi-square tests between HBE categories at interface and within intrachain.

Dataset1/Dataset2	**intrachain**	**interface**	**Dataset**	**interface/intrachain**
PPnrall, PDnrall	0.919	2.2e-16	PDnrall	5.3e-07
PTnrall, PDnrall	0.994	3.73e-06	PPnrall	0.871
PTnrall, PPnrall	0.995	0.5247	PTnrall	0.979

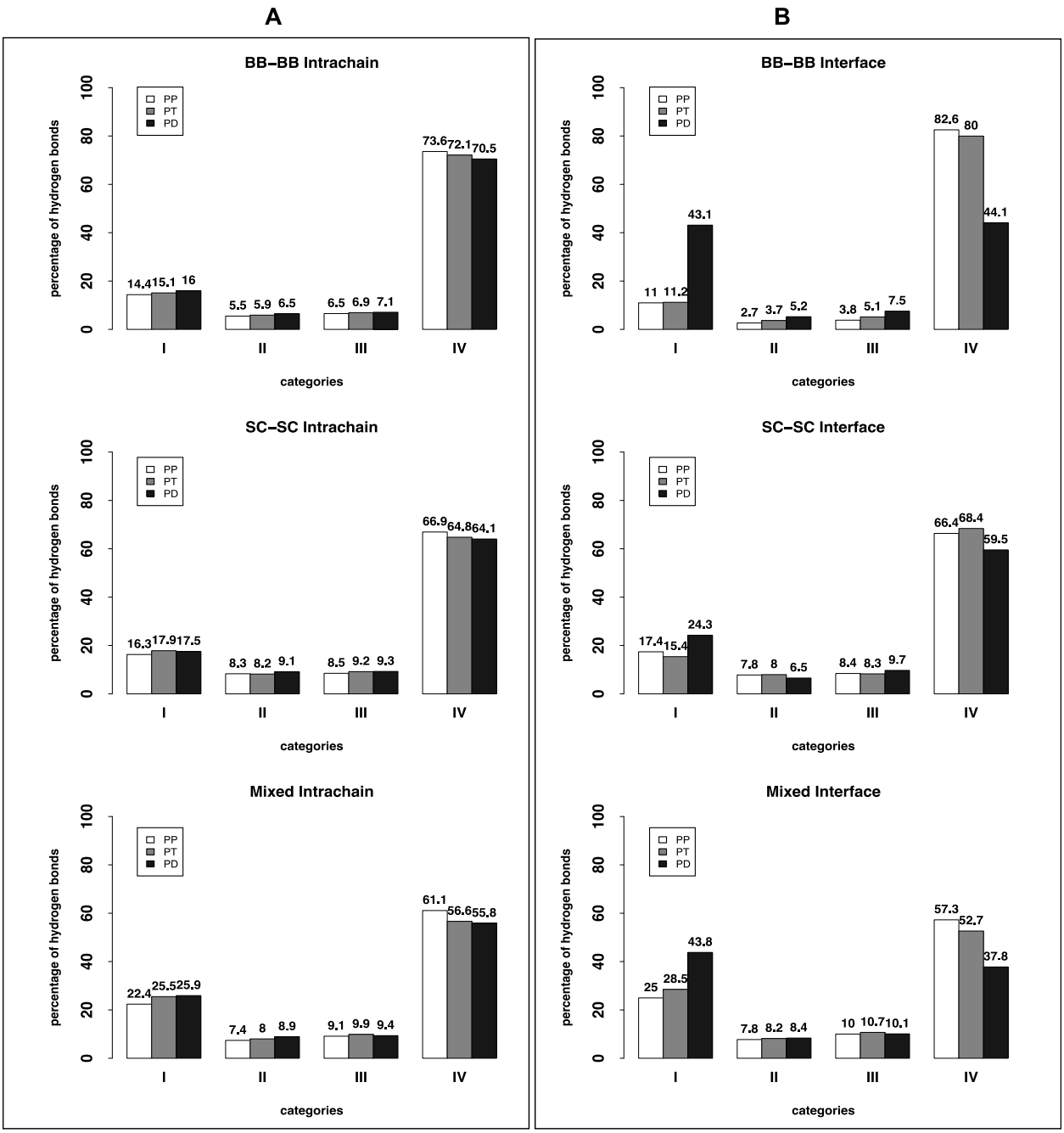

Figure 2.5. Comparison of (A) intrachain hydrogen bond energy and (B) interface hydrogen bond energy in different hydrogen bond types.

We also compared the energy distributions of each HB type across interfaces (Figure 2.5B). Similar to the pattern observed for all HBs in PD, energy distributions of different types of interface HB in PD complexes also differ significantly from PP and PT complexes while there is no significant difference between PP and PT complexes (Table S3). Interestingly, SC-SC HBs in PD complexes have a much larger percentage of strong, category IV HBs (59.5%) while the BB-BB and Mixed types in PD complexes have more weak, category I HBs (43.1% and 43.8% respectively) than the SC-SC HBs (24.3%), suggesting important functional applications of HBs in specific protein-DNA interactions.

2.3.4. Comparison of hydrogen bonds between HS and MS datasets

In our previous study, we demonstrated that highly specific HS protein-DNA complexes have more hydrogen bonds than the multi-specific MS protein-DNA complexes, including both total hydrogen bonds and sidechain-base hydrogen bonds [5]. It is intriguing to see whether there is any relationship between the HB strength and protein-DNA binding specificity. We first compared the HB types and energy categories within proteins as well as at the interface of HS and MS complexes. No significant differences between HS and MS complexes were found in terms of energy categories (Figures S7-S8) while there are significant differences between the intrachain and interface for both HS (p-value: 9.673e-07) and MS complexes (p-value: 6.413e-07). We did observe some statistically non-significant small differences. For example, the number of SC-SC interface HBs in HS (32%) is slightly higher than that in MS (28.2%) (Figure S5A). Both HS and MS complexes show similar interface HB energy distributions with an overall balance of strong and weak HBs, but HS complexes have a slightly higher percentage of HBs in category IV (Figure S7).

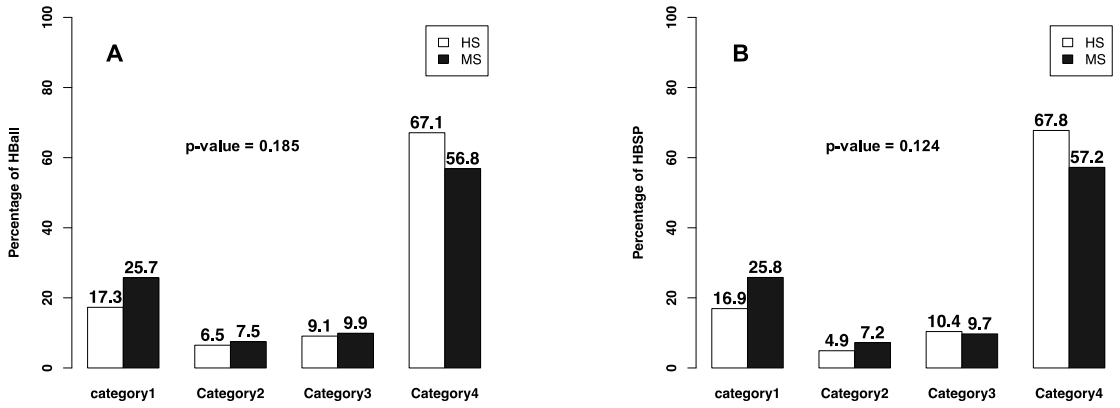

Figure 2.6. Comparison of major groove for (A) HBall and (B) HBSP energy distributions between HS and MS complexes.

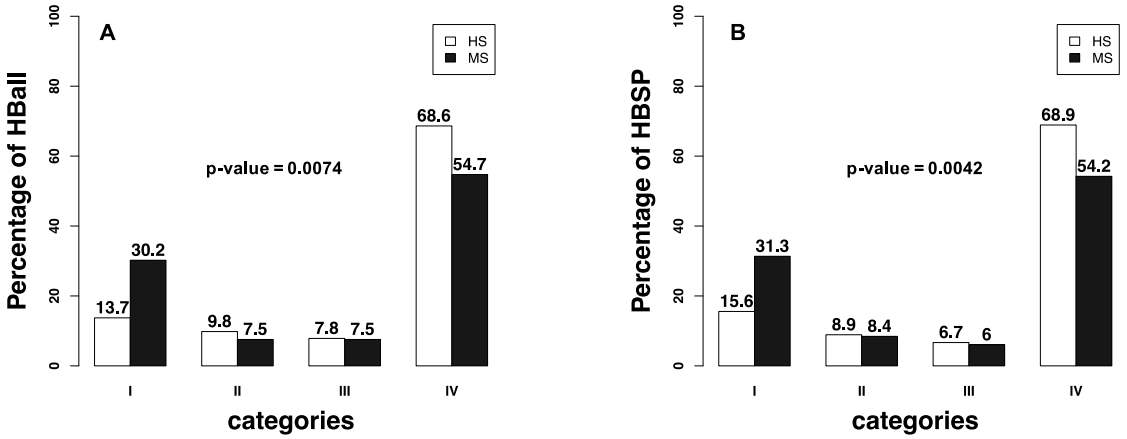

Figure 2.7. Comparison of minor groove for (A) HBall and (B) HBSP energy distributions between HS and MS complexes.

Since both major and minor grooves are known to play important roles in the base and shape readout mechanisms in specific protein-DNA recognition [5,22,147,173], we compared the energy distributions of total hydrogen bonds and sidechain-base hydrogen bonds in the major and minor grooves. Between major and minor grooves, there is no significant difference in terms of hydrogen bond energy distributions within each type of PD complexes, PDnrall, HS and MS with

high p-values (data not shown). For major groove HBs, while we observed more strong and fewer weak major groove HBs in HS complexes than those in the MS complexes, the differences in the energy distributions of HBall and HBSP in the major groove between HS and MS complexes are not statistically significant (Figure 2.6). However, we observed a significant difference in the energy distributions in the minor groove for both HBall and HBSP between HS and MS complexes (Figure 2.7). In general, HS complexes have more strong hydrogen bonds (category IV) and fewer weak hydrogen bonds (category I) than those in the MS complexes in the minor groove. The MS complexes have about double the percentage of weak hydrogen bonds in category I than that in HS complexes. These results suggest a clear and important role of HB energy of the minor groove in specific protein-DNA interaction.

2.4. Discussion

Despite the generally known importance of hydrogen bonds in protein-ligand interactions, the relative contribution of different types of hydrogen bonds, especially their energy in different types of complexes, is unknown. Previous studies mainly focused on analyses of the number of hydrogen bonds. Here we performed a systematic comparative analysis of hydrogen bonds and their energy at the interface and within protein chains among three non-redundant protein-ligand complexes, PP, PT and PD. To the best of our knowledge, this is the first study that compares the energy of hydrogen bonds in different types of complexes. In addition, our use of large non-redundant datasets not only maximizes diversity of the complexes but also avoids potential biases. Results between HBPLUS and FIRST are in high agreement even though they use different algorithms for identifying hydrogen bonds. We also showed similar results between individual datasets for each type of complexes suggesting the results are robust regardless of the datasets and the tools used for hydrogen bonds annotations.

Our analyses revealed several important findings. First, for intrachain hydrogen bonds, our analysis not only corroborates several previous findings [68,71], but also provide additional information by demonstrating no significant difference in the distributions of HB energy among different complexes. Second, at the interface, the hydrogen bond distributions of PD complexes differ from both PP and PT complexes significantly in three aspects: (a) the total number of hydrogen bonds, the number of sidechain-base hydrogen bonds, and the normalized numbers by interface area in PD complexes are significantly higher than those in both PP and PT complexes; (b) more importantly, PD complexes have significantly different distributions of HB types and energy than those of either PP or PT complexes. There is a unique balance between strong and weak hydrogen bonds in protein-DNA interfaces; and (c) there is a significant difference of the minor groove hydrogen bonds between HS and MS complexes with HS having more low energy strong HBs.

Our comparative analyses on energy categories are based on HB energy cutoffs (-0.1 kcal/mol, -0.6 kcal/mol, -1.0 kcal/mol, and -1.5kcal/mol) from previous studies (Table 2.2) [33,34,148]. To test if similar results can be observed with different HB energy discretization, the hydrogen bonds were grouped using a larger energy range separated by -0.1 kcal/mol, -0.7 kcal/mol, -1.3 kcal/mol, and -2.0 kcal/mol (Table S4). The results of HB energy distributions, shown in Figures S9-S12 and Table S5-S6, are in agreement with conclusions (Figures 2.4-2.7, Table 2.4 and Table S3) with energy ranges in Table 2.2, suggesting our key findings are not affected by different discretization of HB energy.

The above findings have important functional and practical implications. While omitting HB information in assessing predicted PP and PT complex models may have minimal effect, our results suggest consideration of hydrogen bonds is beneficial to quality assessment of protein-

DNA complexes models since both the raw number and the normalized number of interface HBs in PD complexes are much higher than those in PP and PT complexes. The use of conserved numbers of native hydrogen bonds in models was suggested to evaluate the quality of protein-peptide models [174]. We found that using the number of HBs can improve quality assessment of protein-DNA complex models [175]. However, due to the unique pattern of interface HB energy distributions in PD complexes and the dynamic nature of macromolecules, it could help model evaluations by considering the HB energy instead of using the raw number of HBs. We demonstrated in our previous study that the accuracy of structure-based prediction of transcription factor binding sites could be improved by adding an HB energy term [176,177].

Our data also provide an insight into the mechanism of binding specificity between protein and DNA. We observed an approximate balance of high and low energy interface hydrogen bonds in PD complexes, but not in the other two types of complexes (Figures 2.4B and 2.5B). One possibility of such difference lies in the geometry of interacting components as geometry is one of the key factors affecting hydrogen bond energy and strength [171]. While DNA is not a rigid molecule, the double helical nature restricts the atoms that can form optimal hydrogen bonds with protein sidechains while the peptide and protein surfaces have a relatively higher flexibility to position atoms for stronger HBs. Other than the unique structure of DNA double-helix that contributes to the pattern of energy distribution, it may also reflect the kinetics of protein-DNA recognition and binding, and the functions of many DNA binding proteins. For example, most of the DNA binding proteins are transcription factors, which bind to conserved DNA binding sequences while allowing variations at certain sites to regulate gene expression. Recent structural and dynamic analyses have shown that transcription factors typically bind to a preferred strand of the DNA double helix [150,178]. A fine balance of strong and weak HBs helps transcription factors

bind to conserved yet different sequences by allowing easier association and disengagement. This is further supported by the comparison between protein-DNA complexes of different binding specificity. Highly specific DNA binding proteins have more strong HBs than the MS group comprising transcription factors (Figures 2.6 and 2.7) [5].

The most interesting finding is from the DNA minor groove HB analysis. Both the energy of all hydrogen bonds and the sidechain-base hydrogen bonds of highly specific protein-DNA complexes are significantly different from that of multi-specific protein-DNA complexes (Figure 2.7). While it is generally thought that minor groove contacts play little role in conferring specific protein-DNA interactions, more studies have shown that this might not be the case. It has been reported that local sequence-dependent minor groove shape plays an important role in specific recognition between protein and DNA [22,147,179–181]. The number of contacts in minor grooves of HS complexes is more than that in MS complexes and the HS complexes contain wider minor grooves than MS [5], thus making it possible for optimal orientation of atoms to form stronger hydrogen bonds. Our results further demonstrate that the minor groove HBs play more critical roles in conferring binding specificity than previously thought.

Chapter 3. COMPAREPD: IMPROVING PROTEIN-DNA COMPLEX MODEL COMPARISON WITH HYDROGEN BOND ENERGY-BASED METRICS

Computational modeling of protein-DNA complexes is a cost-efficient alternative to experimental structure determination methods to fill the void in the protein-DNA complex structure landscape and has important implications in biomedical applications such as structure-based computer aided drug design. One of the key steps in developing methods for accurate modeling of protein-DNA complexes is the similarity assessment between models and reference complex structures. Existing methods assess complex structure similarities with distance-based methods such as interface root mean square deviation (iRMSD), ligand RMSD (lRMSD), and the fraction of the contacts in the native structure that is reproduced in the model (*Fnat*) or a combination of these terms (DockQ). However, one important interaction type between the complex units especially in protein-DNA interactions, hydrogen bonds, is generally not considered in scoring the complex similarity. Here, we present a new scoring function that considers interface hydrogen bond energy, ComparePD, for accurate similarity measure of protein-DNA complexes. ComparePD was tested on both homology and docking protein-DNA complex models. The results were compared with a modified version of DockQ, ModifiedDockQ (tailored for protein-DNA complexes) as well as the metrics employed in the community-wide experiment CAPRI (Critical Assessment of PRedicted Interactions). We demonstrated that ComparePD was able to accurately describe the similarity of protein-DNA complex models by capturing interface hydrogen bond interactions.

3.1. Introduction

Protein-DNA complexes play a vital role in the regulation of gene expression, cellular growth, differentiation and development [182,183]. Knowledge of protein-DNA complex structures is essential to gain insight into their function [184–188]. Structures of most of the protein-DNA complexes, however, remain unsolved due to the technical challenges in experimental methods to solve larger biomolecular assemblies and nucleic acids [189]. Recent data showed that protein-DNA complexes in Protein Data Bank (PDB) represent less than 5% of all the known structures [81,190]. To address this issue, *in-silico* prediction of protein-DNA complex structures has been explored using methods such as homology modeling and docking [191–193]. Both methods can predict a large number of models in a short time by exploring a wide conformational energy landscape [111,194,195]. While structure prediction of monomeric proteins is at an advanced stage, the methods for modeling complexes still has several challenges to overcome [111,196–199]. For example, sequence similarity, a limiting factor for the success of homology modeling methods, does not guarantee interface similarity. It has previously been shown that many homologous proteins can have different interaction modes [200]. New methods of structure prediction for complexes are increasingly being developed to overcome these challenges [201–203].

A key step in the development of computational modeling methods is to assess the structural similarity between predicted models and the experimentally solved reference structure. It is a challenging task to accurately capture the functionally meaningful interface structural similarity and no standard method has been widely accepted for complexes. Existing measures of similarity such as iRMSD, lRMSD and Fnat are all distance-based and have several limitations [204–214]. For example, RMSD treats each position equally and is therefore highly sensitive to conformational changes and can be misleading for complexes with larger loops [138,215]. Models

of similar quality can have largely different RMSDs because of larger flexibility in loops or termini of proteins [208]. Despite being widely used, these metrics alone do not accurately reflect the true overall quality of the model [205,208,216]. The community wide CAPRI (Critical Assessment of PRedicted Interactions) interaction prediction experiment group these metrics in four discrete classes (Table 3.1) [217]. Models are classified as high, medium, acceptable quality or incorrect. While this provides a reasonable initial assessment of the quality, classifying models into four main types is too restrictive for assessing true model quality. To this end, Basu *et. al.* proposed a continuous scoring function, DockQ which combines these criteria in a single score for assessment of complex similarity [205].

Table 3.1. Classification scheme of CAPRI based on Fnat, iRMSD and lRMSD ranges

Model quality	**Ranges**
High	$F_{nat} \geq 0.5\ and\ (lRMSD\ \leq 1.0$ Å $or\ iRMSD\ \leq 1.0$ Å
Medium	$(0.3 \leq F_{nat} < 0.5)\ and\ (lRMSD\ \leq 5.0$ Å $or\ iRMSD \leq 2.0$ Å$)$ $F_{nat} \geq 0.5\ and\ lRMSD > 1.0$ Å $and\ iRMSD\ > 1.0$ Å
Acceptable	$(0.1 \leq F_{nat} < 0.3)\ and\ (lRMSD\ \leq 10.0$ Å $or\ iRMSD \leq 4.0$ Å$)$ $F_{nat} \geq 0.3\ and\ lRMSD > 5.0$ Å $and\ iRMSD\ > 2.0$ Å
Incorrect	$F_{nat} < 0.1\ or\ (lRMSD > 10.0$ Å $and\ iRMSD > 4.0$ Å$)$

Despite being widely accepted, these methods do not represent a global measure of similarity for all complexes. For example, the distance cut-off in these metrics was modified for comparing protein-peptide complexes to reflect their smaller interface area [218,219]. A major limitation of these metrics is that they do not consider biological features of the complex relevant to its function which reflects a more meaningful comparison of complexes. For comparison of protein-DNA complexes, this can be achieved by incorporating hydrogen bonds which are central to their binding specificity and function [56,147–149,169,174,220–224]. Hydrogen bonds are

weak intermolecular interactions which are significantly more prevalent in protein-DNA complexes as compared to protein-protein and protein-peptide complexes [225]. Side chain-base hydrogen bonds between proteins and nucleic acids in particular play vital roles in binding specificity [149]. There is a significantly higher number of hydrogen bonds in highly specific protein-DNA complexes as compared to non-specific and multi-specific protein-DNA complexes [5]. Ferrell and Guo previously demonstrated that incorporating a hydrogen bonding term in a scoring function improves the prediction of transcription factor binding specificity [176]. A hydrogen bond-based comparison of protein-DNA complexes could result in the selection of more functionally meaningful models. A recent study by Marcu *et. al.* has suggested $fnat_{hb}$, the number of conserved hydrogen bonds in model, to compare the models with the native complex. An intrinsic limitation of the methods based on a single distance or energy cut-off is that hydrogen bonds of different strength are treated equally [56,171]. While this might be effective in complexes with mostly strong hydrogen bonds such as protein-protein and protein-peptide, it might not work well for protein-DNA complexes which present a unique, almost equal distribution of weak and strong hydrogen bonds [225]. Furthermore, conventional methods to identify hydrogen bonds are based on a single energy threshold. An energy threshold of -0.6 kcal/mol is generally suggested for protein-DNA complexes by the author of FIRST [33]. However, in terms of hydrogen bond comparison between protein-DNA complexes, a small difference in hydrogen bond energy between -0.595 kcal/mol and -0.605 kcal/mol would result in different conclusion: 0 vs. 1 hydrogen bond. Such discrete cut-off can be too rigid for protein-DNA complexes where hydrogen bonds of high energy are almost as prevalent as hydrogen bonds of low energy [225].

Here, we present a novel weighted hydrogen bond energy-based method to compare protein-DNA complexes. To the best of our knowledge, this is the first time an approach

considering different strengths of hydrogen bond has been employed to compare protein-DNA complexes. The results were compared with the standard CAPRI-based criteria of comparison (Table 3.1) and ModifiedDockQ to test the performance of our comparison score.

3.2. Materials and methods

3.2.1. Datasets

Two datasets of protein-DNA complex models generated through homology modeling and docking respectively were used [226,227]. The homology modeling dataset comprises 90 models of 5 non-redundant homeodomain complexes (Table 3.2). Several templates of high structural quality and varying sequence similarity ranging between 35% and 70% were selected for each target complex. Since the existing homology modeling methods do not efficiently model interfaces of larger biomolecular assemblies and nucleic acids, several additional steps were performed (Figure 3.1). First, the protein component of each template structure was used to generate 5 homology models of the target sequence. Next, each of these homology models was structurally aligned with the native complex. Finally, the aligned protein models complexed with native DNA were generated. Modeller was used to generate protein models and TM-align was used for structural alignment [228,229].

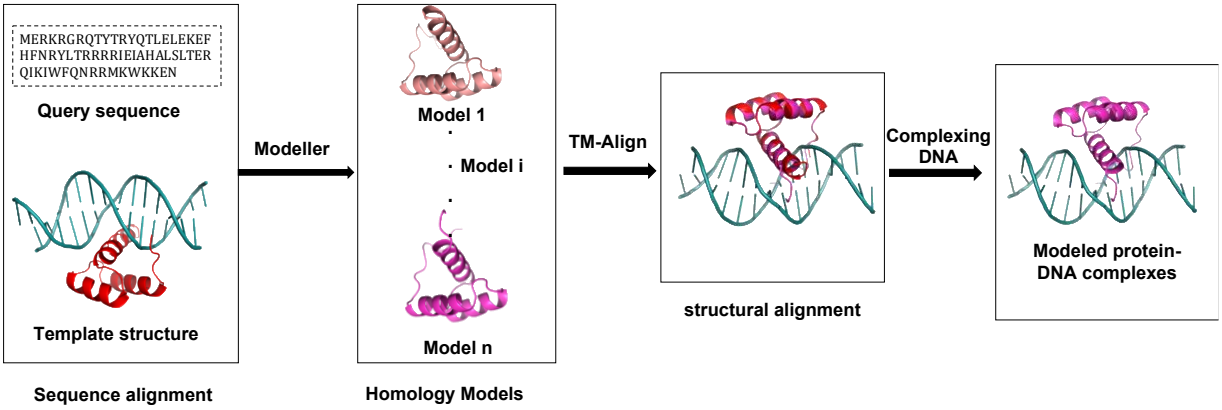

Figure 3.1. Homology modeling of 9antB protein-DNA complex using 1le8A as the template structure (n = 5).

Table 3.2. Homology models of homeodomain based on templates of varying sequence identity

Target	**Template (Sequence identity)**
1pufB	3cmyA (35%), 1akhA (46%) and 4xrmA (51%)
1zq3P	2xsdC (34%), 4rduA (45%) and 6m3dC (51%)
2me6A	1fjlA (40%), 1ig7A (59%), 6m3dC (48%)
3sjmA	3osfA (42%), 7c4pA (55%)
9antB	1le8A (33%), 4rduA (42%), 1jggA (52%), 1pufA (70%)

The docking dataset comprises 500 docked models of 25 protein-DNA complexes from previously published and widely used HADDOCK protein-DNA docking benchmark (Table 3.3) [226,227]. The models generated by Honorato *et. al.* using MARTINI force field for DNA in HADDOCK, were obtained from publicly available repository. The HADDOCK-MARTINI dataset originally comprises 43 protein-DNA complexes each comprising 1000 models ranked according to their HADDOCK score [225,230]. We filtered these complexes rigorously using methods described previously and eliminated 14 of these complexes [225]. Two more complexes with missing residues near the interface were removed after visual analysis of complexes. Of the resulting 27 complexes, FIRST was unable to annotate hydrogen bonds in two complexes resulting

in a final dataset of 25 complexes. Finally, the highest ranking 20 models for each complex according to HADDOCK score were selected.

Table 3.3. PDB ids of selected complexes from protein-DNA docking benchmark of HADDOCK

1azp	1f4k	1kc6	1rva	2fio
1a74	1fok	1mnn	1vas	2irf
1by4	1gi9z	1qrv	1z9c	3bam
1ddn	1h9t	1r4o	2oaa	3cro
1ea4	1h9c	1rpe	2fl3	7mht

In addition to the datasets of protein-DNA complex models, two previously published non-redundant datasets of native protein-protein complexes and two non-redundant datasets of native protein-DNA complexes were used to estimate the RMSD scaling factors and optimize weights in the ComparePD function [166,225]. Our previously developed dataset PPnrall was used to estimate the scaling factor for iRMSD [166,225]. It comprises high quality experimental structures of homodimers, heterodimers and protein-protein docking benchmark version 5.0 [166,225]. The protein-protein docking benchmark version 5.0 was used to estimate scaling factor of ligand RMSD [162,231]. For protein-DNA complexes, we used PDnrall which comprises non-redundant specific protein-DNA complexes and protein-DNA rigid docking benchmark [225]. In order to estimate weights for scoring function, we combined previously published highly specific and multi-specific datasets of protein-DNA complexes [5].

3.2.2. Hydrogen bond energy

FIRST (Floppy Inclusion and Rigid Substructure Topography) was used to annotate hydrogen bonds in the complex and calculate their energy using the equation 2.1 [33,171]. The

annotated hydrogen bonds were classified into three categories based on their energy (Table 3.4). Information from previous studies was used to define these energy bins [24,33,34,225]. Weights 0.5, 0.8 and 1.0 are arbitrarily assigned to reflect the strength of hydrogen bonds because we do not have sufficient number of protein-DNA complexes for training. In addition to FIRST, HBPLUS with default parameters was also used annotate hydrogen bonds in PPnrall, Docking benchmark version 5.0, PDnrall and pooled dataset of highly specific and multi-specific protein-DNA complexes [56].

Table 3.4. Energy bins for each category of hydrogen bonds energy (HBE) and their corresponding weights

Category	**HBE range (kcal/mol)**	**weights**
I	$-0.6 \leq HBE < -0.1$	0.5
II	$-1.0 \leq HBE < -0.6$	0.8
III	$HBE < -1.0$	1

3.2.3. ModifiedDockQ

DockQ is a continuous score between [0,1] based on Fnat, iRMSD, and lRMSD, originally developed for comparing protein-protein complexes [205]. Since protein-protein complexes and protein-DNA complexes have different interface [225], we implemented a modified version of DockQ for protein-DNA complexes using different scaling factors (Eq. 3.1).

$$ModifiedDockQ = \frac{F_{nat} + iRMSD_{scaled} + lRMSD_{scaled}}{3} \quad \text{Eq. 3.1}$$

A contact is defined between two heavy atoms if they are separated by a distance of 4.5 Å or less. Interface is defined as pairs of heavy atoms from the proteins and DNA within 10 Å of

each other. iRMSD is based on C_β atoms of proteins and N1 or N9 atoms of DNA in the binding interface ($iRMSD_{CB\text{-}N1N9}$). lRMSD is based on N1 and N9 interface atoms of DNA backbone ($lRMSD_{N1N9}$).

Estimation of scaling factors d_i and d_l

RMSDs in DockQ were scaled using inverse square scaling method to account for two problems. First, a near-native model has higher F_{nat} and lower RMSD values. Second, arbitrarily large RMSD values can be misleading. Inverse square scaling of iRMSD and lRMSD provides an efficient solution (Eq. 3.2). Basu et. al. have shown that scaled RMSD measures provide a more sensitive discrimination between the quality of protein-protein models [205]. The scaled RMSD values are calculated as in Eq. 3.2.

$$RMSD_{scaled_j} = \frac{1}{1 + \left(\frac{RMSD_j}{d_j}\right)^2} \quad \text{Eq. 3.2}$$

where j is for iRMSD or lRMSD respectively and d_j represents the scaling factor d_i for iRMSD and d_l for lRMSD respectively. In the original publication, the scaling factors d_i and d_l were optimized by performing a grid search on a dataset of 56,015 docked models of 118 protein-protein complexes. Due to the small number of experimentally solved protein-DNA complexes, a grid search optimization of scaling factors is not practical. We estimated the scaling factor protein-DNA complexes for iRMSD ($d_i \sim 1.2$) by comparing the average and median interface areas of PPnrall for protein-protein complexes and PDnrall for protein-DNA complexes [225]. Both the mean and median interface area of PPnrall are about 1.7 times those of PDnrall. In order to account for inverse-square scaling function, d_i for protein-protein complexes in the original publication

(1.5) was normalized by $\sqrt{1.7}$ to get an updated di of ~1.2. The scaling factor for lRMSD, d_l, was estimated similarly by comparing the number of backbone atoms ($d_l \sim 2$). Protein-DNA docking benchmark version 5 was used for protein-protein complexes to correspond better to the original publication since models of the native complexes in this benchmark were used to train d_l in the original publication.

3.2.4. ComparePD: a function for comparison of protein-DNA complexes

We developed a new linear continuous function for comparing protein-DNA complexes by combing the traditional features F_{nat}, iRMSD and lRMSD, and a novel weighted hydrogen bond energy-based score, $Composite_{HBE}$ (Eq. 3.3).

$$ComparePD = \frac{F_{nat} + iRMSD_{scaled} + lRMSD_{scaled} + Composite_{HBE}}{4} \quad \text{Eq. 3.3}$$

$Composite_{HBE}$: Weighted hydrogen bond energy algorithm

The key part of $Composite_{HBE}$ is a weighted hydrogen bond (WHB) based on the hydrogen bond energy between a protein-DNA complex model and a reference complex (Fig. 3.2). Hydrogen bonds in native and model complexes are first annotated with FIRST. Weights w_m and w_n are assigned to each hydrogen bond in the model and native complex, respectively (Table 4.4). The weight of each conserved hydrogen bonds in the model w_m is compared to the corresponding w_n in the reference structure. W_{HB} is then calculated to reflect the conservation of the hydrogen bond in the reference structure with a value between 0 and 1 (Eq. 3.4):

$$w_{HB} = \begin{cases} w_m / w_n , & w_m < w_n \\ w_n / w_m , & w_m \geq w_n \end{cases}$$ Eq. 3.4

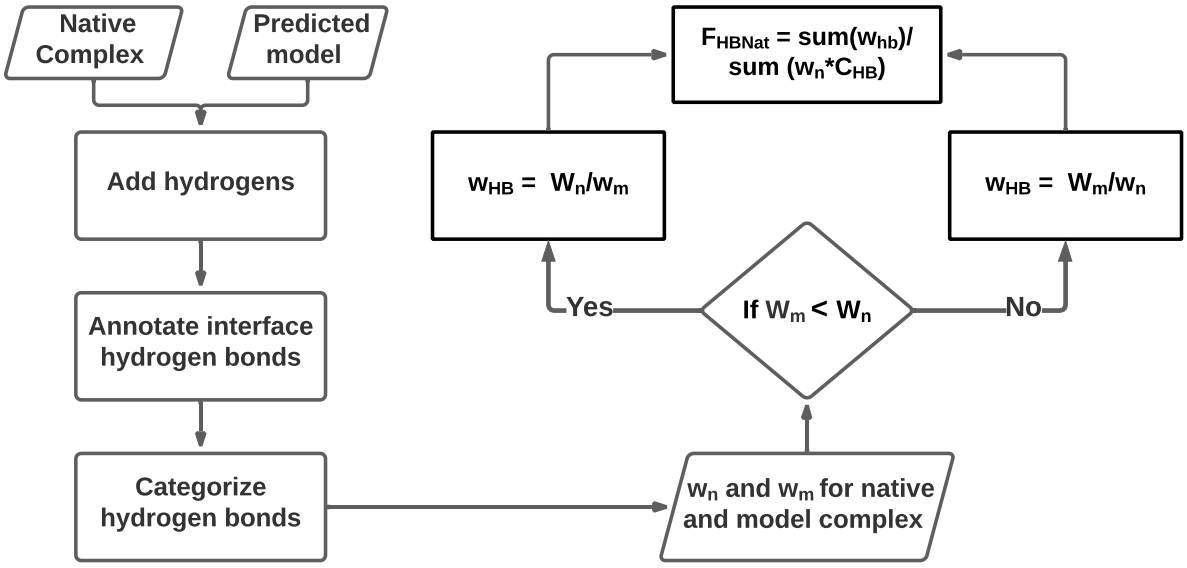

Figure 3.2. A flowchart of weighted hydrogen bond energy algorithm

Since side chain-base hydrogen bonds at the interface of protein-DNA complexes are a significant contributor to the binding specificity [22], we calculate them separately as w_{HBSP} using equation 3.5. w_{HB} and w_{HBSP} scores for all conserved hydrogen bonds at the interface of protein-DNA complex are then added and normalized by the weighted sum of total number of hydrogen bonds (Eq. 3.5, 3.6).

$$F_{HBNat} = \frac{\sum W_{HB}}{\sum_{i=1}^{3} W_i (C_{HB})_i}$$ Eq. 3.5

$$F_{HBSPNat} = \frac{\sum W_{HBSP}}{\sum_{i=1}^{3} W_i (C_{HBSP})_i}$$ Eq. 3.6

where F_{HBNat} and $F_{HBSPNat}$ are the scores that represent fraction of hydrogen bonds captured in the model. C_{HB} and C_{HBSP} are the total number of hydrogen bonds and sidechain-base hydrogen bonds respectively for the corresponding three energy categories (*i= 1,2,3)* of hydrogen bonds (Table 3.4). Finally, a composite score of F_{HBNat} and $F_{HBSPNat}$ is calculated to reflect overall similarity in the energy of hydrogen bonds between the native complex and the model (Eq. 3.7).

$$Composite_{HBE} = \frac{(w_1 C_{HB} F_{HBNat} + w_2 C_{HBSP} F_{HBSPNat})}{w_1 C_{HB} + w_2 C_{HBSP}} \quad \text{Eq. 3.7}$$

where $w_1 = 0.3$ and $w_2 = 1\text{-}w_1 = 0.7$. w_1 is estimated by comparing the average number of all interface hydrogen bonds to the sidechain-base hydrogen bonds in a pooled non-redundant dataset of highly specific and multi-specific protein-DNA complexes (Supplementary Figure S13). Higher weight is assigned to sidechain-base hydrogen bonds because of their important role in protein-DNA binding specificity. C_{HB} and C_{HBSP} represent the total number of hydrogen bonds and total number of sidechain-based hydrogen bonds in the reference protein-DNA complex respectively.

3.3. Results

ComparePD was used to rank MARTINI-based docking models of each of the 25 complexes from HADDOCK protein-DNA docking benchmark and the homology models of each of the 5 non-redundant homeodomain complexes. The results were compared with the standard CAPRI classification (Table 3.1) and ModifiedDockQ rankings (Eq. 3.2). The comparison of scoring methods was performed in two ways. First, the best model identified by ModifiedDockQ was compared to the corresponding top selection of ComparePD. Second, top three models ranked by both scores were compared to check if both methods agree on selection of at least one common model.

Table 3.5. CAPRI-based classification of models into 4 categories based on Table 3.1.

Category	Homology Models	Docked Models
High	1 (20%)	0
Medium	1 (20%)	19 (76%)
Acceptable	2 (40%)	6 (24%)
Incorrect	1 (20%)	0

Table 3.1. shows the best model category according to CAPRI classification in each case. Homology models of high quality were generated in only 1 out of 5 targets (Table 3.5). The MARTINI-based docking dataset did not generate any high-quality models. Most of the docked models are of medium (76%) or acceptable quality (24%).

ComparePD selects a different, potentially better-quality model in terms of hydrogen bonds for 52% of docked and 60% of homology modeled complexes as compared to ModifiedDockQ. When the top 3 models were compared, ComparePD selects at least one common model as ModifiedDockQ in 88% and orientation potential in 68% targets in docked models. In all homology models, at-least one of the top 3 models ranked by ComparePD is also captured by ModifiedDockQ (Tables 3.6).

Table 3.6A. Comparison of scoring methods for the selection of top model in docked dataset*

Same\Different	ComparePD (Top/Top3)	ModifiedDockQ (Top/Top3)
ComparePD		13 (52%)/3(12%)
ModifiedDockQ	12(48%)/22(88%)	

Table 3.6B. Comparison of scoring methods for the selection of top models in homology modeling dataset*

Same\Different	ComparePD	ModifiedDockQ

ComparePD		3 (60%)/0 (0%)
ModifiedDockQ	2(40%)/5(100%)	

* The lower half of the table presents similarities and the upper half shows differences between scores

3.3.1. Performance evaluation

The ability of ComparePD to accurately capture the distance-based measures was assessed through a correlation analysis (Fig. 3.3). Figure 3.3. shows correlations of ComparePD scores against ModifiedDockQ score, Fnat and RMSDs. Significant correlation of ComparePD with ModifiedDockQ and Fnat (correlation coefficient of 0.79 and 0.72) indicate its performance is comparable to these metrics. The RMSD values are also significantly correlated and similar pattern is observed for both homology modeling and docking datasets. lRMSD is only shown for docking models because native DNA is complexed with protein model in homology models (Fig. 3.3B).

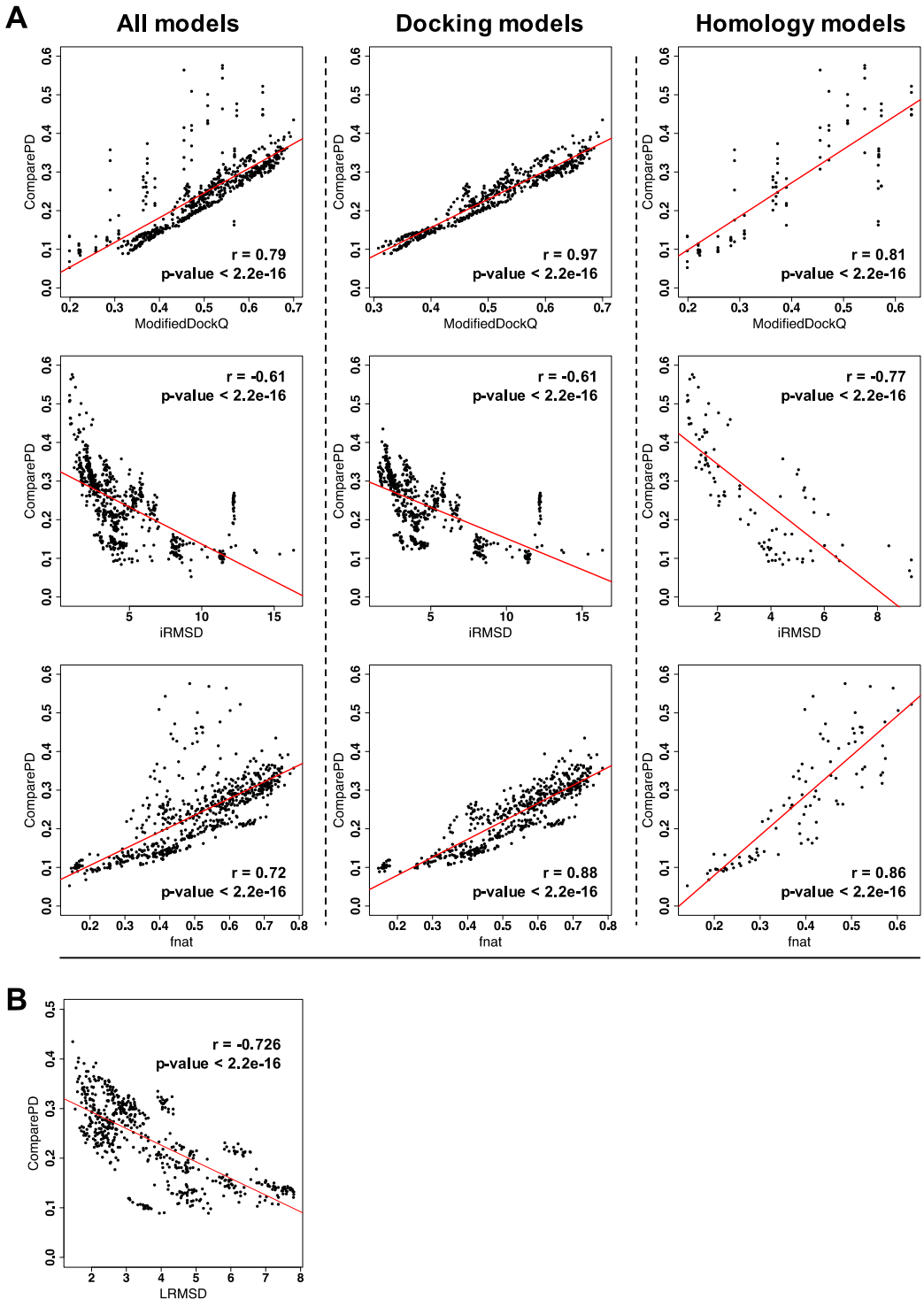

Figure 3.3. Correlation analysis of scores between (A) ComparePD, ModifiedDockQ, fnat and iRMSD for all models, docking models and homology models and (B) ComparePD against lRMSD for docking models.

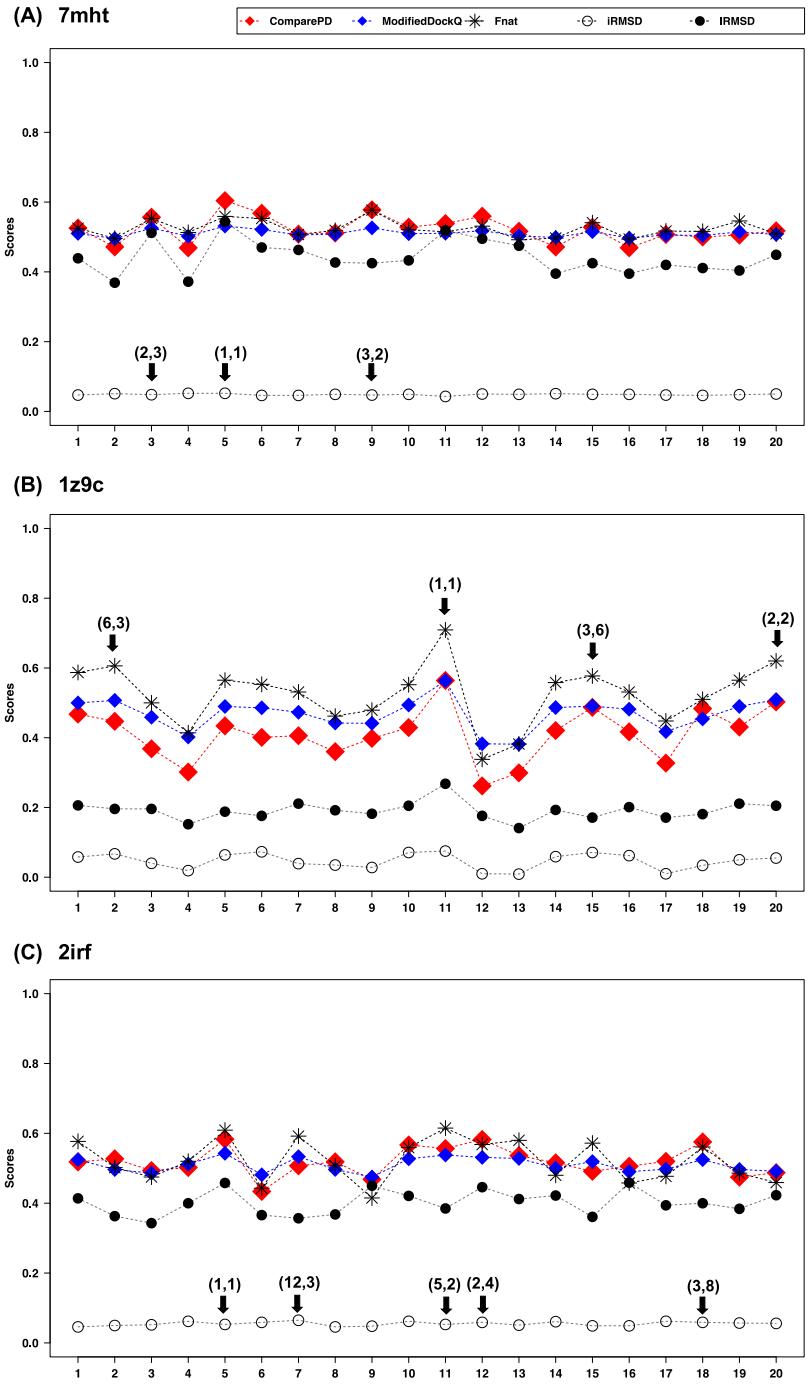

Figure 3.4. Comparison plots of models scored by ComparePD and ModifiedDockQ alongwith their individual Fnat, normalized iRMSD and lRMSD scores for (A) 7mht, (B) 1z9c and (C) 2irf. Top three models are highlighted, and corresponding ranks are reported as (ComparePD, ModifiedDockQ) for each complex.

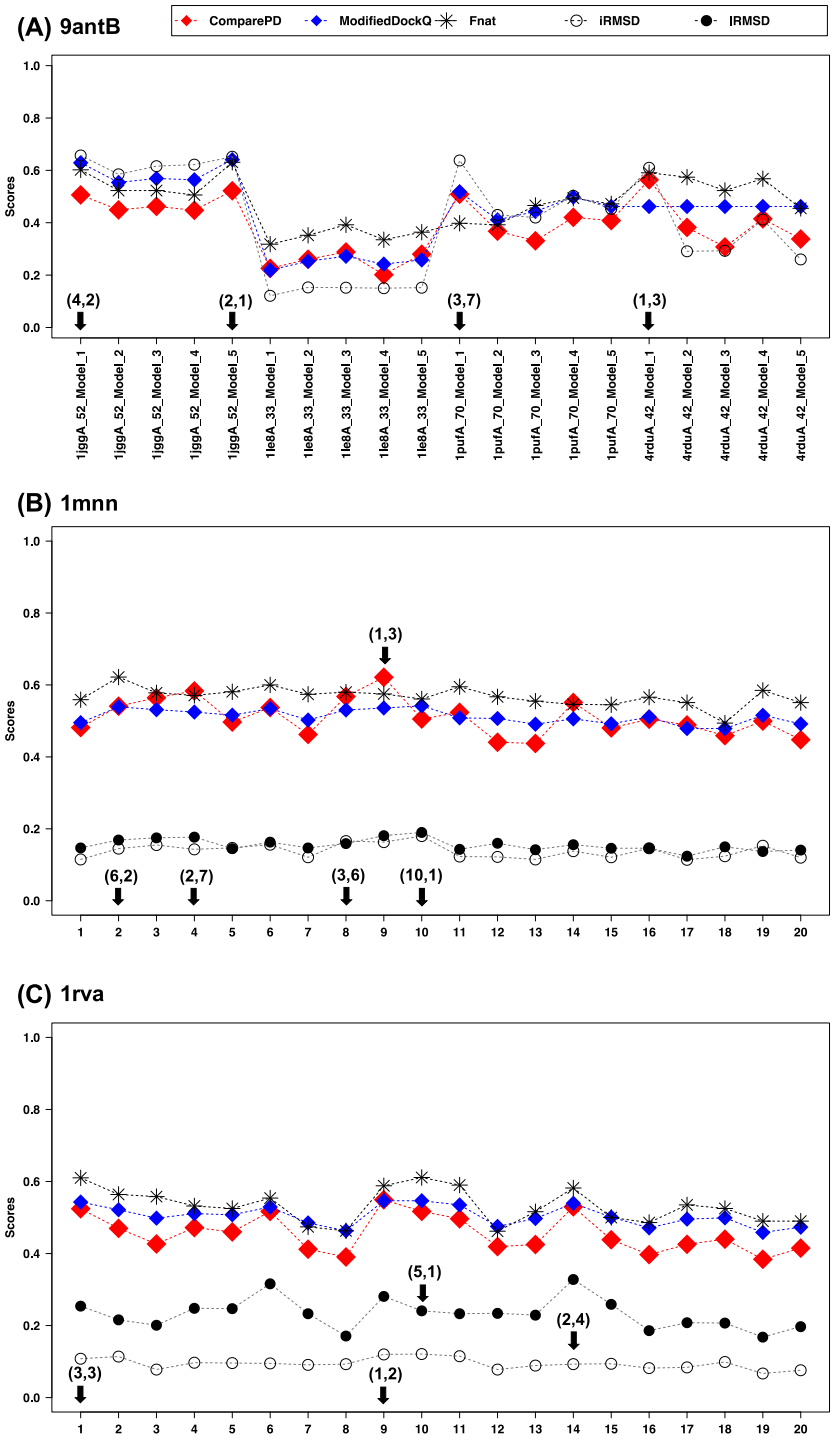

Figure 3.5. Comparison plots of models scored by ComparePD and ModifiedDockQ alongwith their individual Fnat, normalized iRMSD and lRMSD scores for (A) 9antB, (B) 1mnn and (C) 1rva. Top three models are highlighted, and corresponding ranks are reported as (ComparePD, ModifiedDockQ) for each complex.

In 48% cases of docking and 40% of homology modeling, ComparePD identifies same top model as ModifiedDockQ. Figure 3.4 shows three such examples. For 7mht, same top 3 models are selected by both scores although ranking of Model 9 and 3 is reversed in ComparePD (Fig. 3.4A). In case of 1z9c and 2irf, top two models selected by both methods are the same (Fig. 3.4B). ComparePD in these cases has been able to capture the top selection by ModifiedDockQ and it has provided further confidence in selection based on interface hydrogen bond energy. Thus, in addition to being structurally comparable, model predicted by ComparePD also has the nearest-native hydrogen bond energy distribution.

In 52% of the cases in docking dataset and 60% of cases in homology modeling, ComparePD identifies a different top model from ModifiedDockQ. Figure 3.5 shows three examples (1 homology modeling case 9antB and 2 docking cases 1mnn and 1rva) where ComparePD selects a better model in each case. A detailed analysis of hydrogen bonds and their energy in the model selected by both methods is given below.

3.3.1.1. 9antB: homeodomain homology models

We used ComparePD and ModifiedDockQ to score homology models of homeodomain protein-DNA complex 9ant built by using templates of varying sequence identity (Fig. 3.5A). 1jggA_52_Model_1 for example corresponds to the first model of 9antB using the experimental structure of 1jggA as template with a sequence identity of 52%. ComparePD ranks homology model 1 with template 4rduA (sequence identity 42%) as the top model, whereas ModifiedDockQ selects model 5 from template 1jggA (sequence identity 52%) as the top model. Even though the ModifiedDock scores for both models are similar (0.601 for top selection of ComparePD and 0.641 for top selection by ModifiedDockQ), an examination of hydrogen bonds in each of these models

indicate that ComparePD selects a better overall model (Fig. 3.6). The top model selected by ComparePD conserves 7 out of 8 native hydrogen bonds. Five of these conserved native hydrogen bonds also conserve the energy categories. These low energy, strong hydrogen bonds including: ARG55-C406, ARG33-A404, ARG7-T518 and ASN53-A520. The two hydrogen bonds whose energy category changes in the model selected by ComparePD (GLN8-A519 and ARG55-C406) are high energy, weaker hydrogen bonds. The energy of hydrogen bond between GLN8-A519 decreased in ComparePD selection whereas in ARG45-T521 it increases. This is also evident by the change in conformation of sidechains of the corresponding atoms. Sidechain confirmations for hydrogen bonds with conserved energy categories in the model selected by ComparePD is similar to near-native. The top model from ModifiedDockQ conserves only 5 hydrogen bonds. Four of the native hydrogen bonds (ARG33-A404, ARG45-T521 and two hydrogen bonds in ASN53-520) are of low energies. Despite the hydrogen bonds being conserved, the category of energy of 3 of these (ARG33-A404, ARG45-T521 and one hydrogen bond between ASN53-520) changes. The energy category of only one hydrogen bond (ARG45-T521) is conserved in this model. Even though Fnat and iRMSD of ModifiedDockQ selection (0.631 and 0.878 Å) is slightly better than ComparePD selection (0.591 and 0.959 Å), poor conservation of hydrogen bonds energy result in lower overall score for the latter.

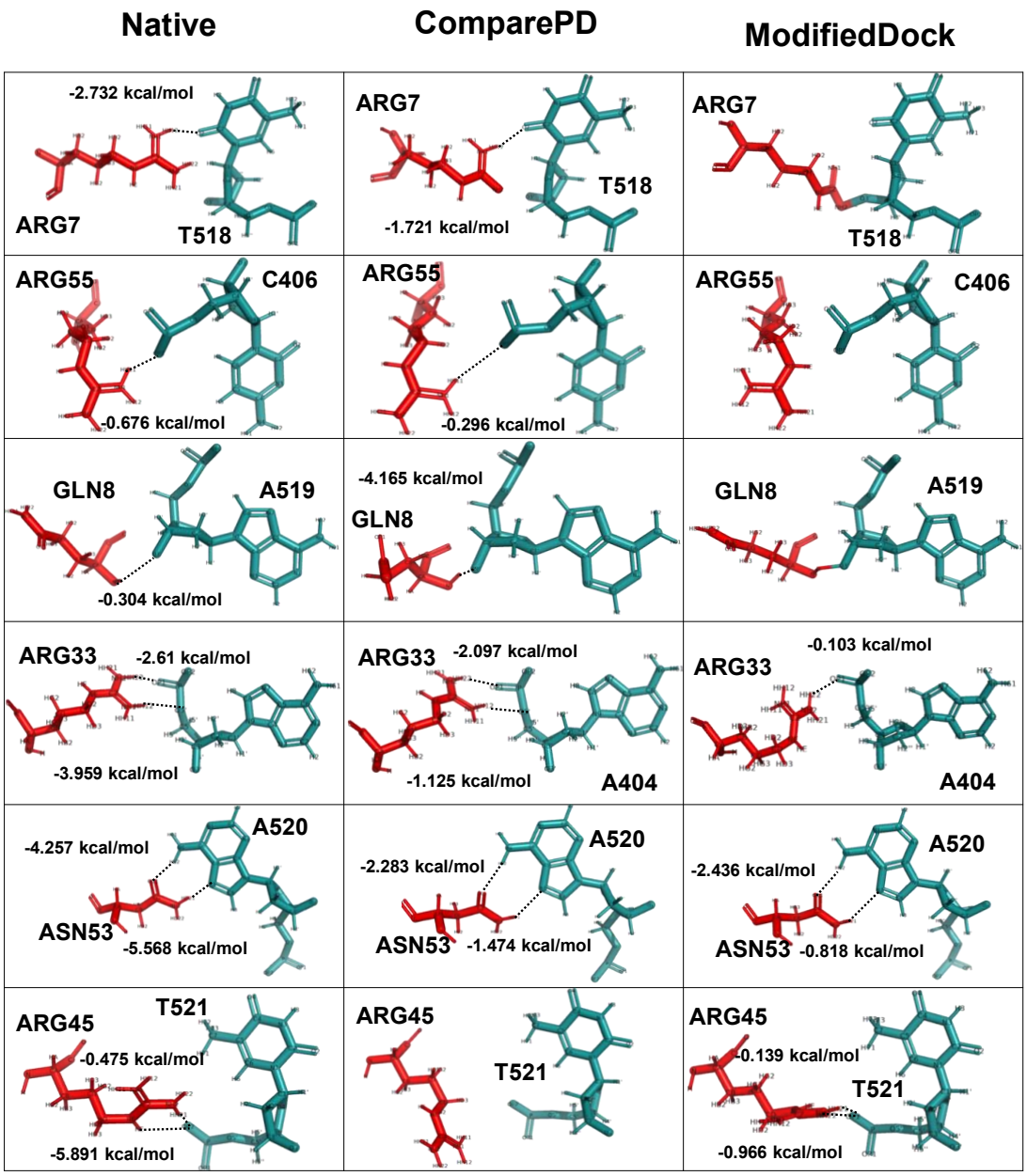

Figure 3.6. Interface hydrogen bonds and their energy in the native complex 9antB and top models predicted by ComparePD and ModifiedDockQ.

3.3.1.2. 1mnn: MARTINI-based docked models of HADDOCK protein-DNA benchmark

For 1mnn, ComparePD selects model 9 as the top model whereas ModifiedDockQ considers model 10 as the model most similar to the native complex (Fig. 3.5B). Despite having similar ModifiedDockQ score, an insight into the energetics of hydrogen bonds reveals that model

9 is a better model (Fig. 3.7). The model selected by ComparePD conserves 7 native hydrogen bonds, whereas the models selected by ModifiedDockQ only captures 2 hydrogen bonds. Four of the conserved native hydrogen bonds (ARG277-G4, ARG79-G25 and two hydrogen bonds in ARG65-G23) in model 9 also conserve the native energy category whereas neither of the hydrogen bonds conserved in model 10 conserve the hydrogen bond in the native energy category.

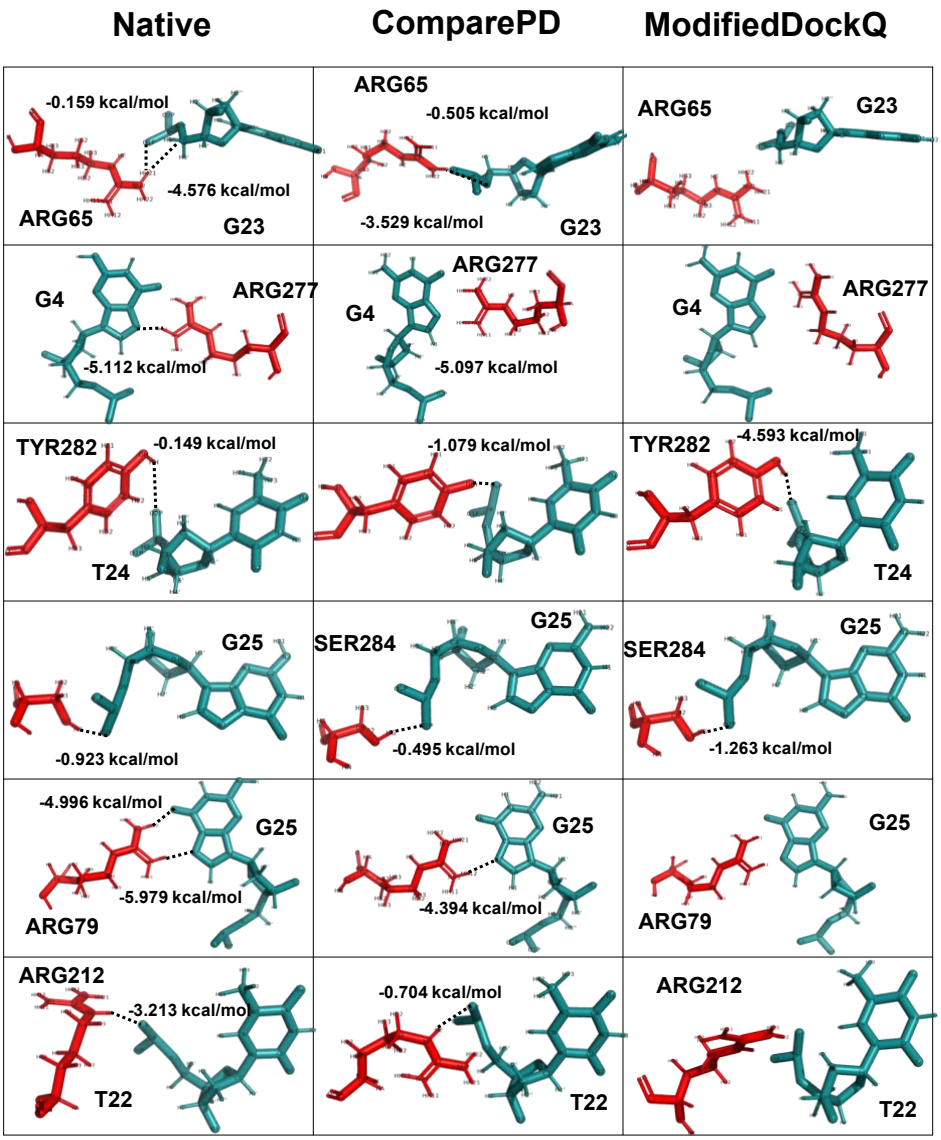

Figure 3.7. Interface hydrogen bonds and their energy in the native complex 1mnn and top models predicted by ComparePD and ModifiedDockQ.

3.3.1.3. 1rva: MARTINI-based docking models from the HADDOCK protein-DNA benchmark

The top selection by ComparePD is model 9 whereas that from ModifiedDockQ is model 10. (Fig. 3.5C). In this case the top ComparePD model is ranked the second best according to ModifiedDockQ. Figure 3.8 shows a comparison of hydrogen bonds and their energies in all these models to the native complex. Six low energy native hydrogen bonds are shown in the first column of the table. Model 10 has 5 conserved hydrogen bonds, whereas model 9 conserves 4 hydrogen bonds. Even though model 10 conserves more hydrogen bonds as compared to Model 9, the energy categories of two conserved hydrogen bonds (ASN184-GG16 and GLN312-C9) in model 10 is not conserved. Model 9, on the other hand, conserves energy categories of all the hydrogen bonds. Model 9 has better lRMSD but worse Fnat as compared to model 10 which is why it was ranked the second best by ModifiedDockQ.

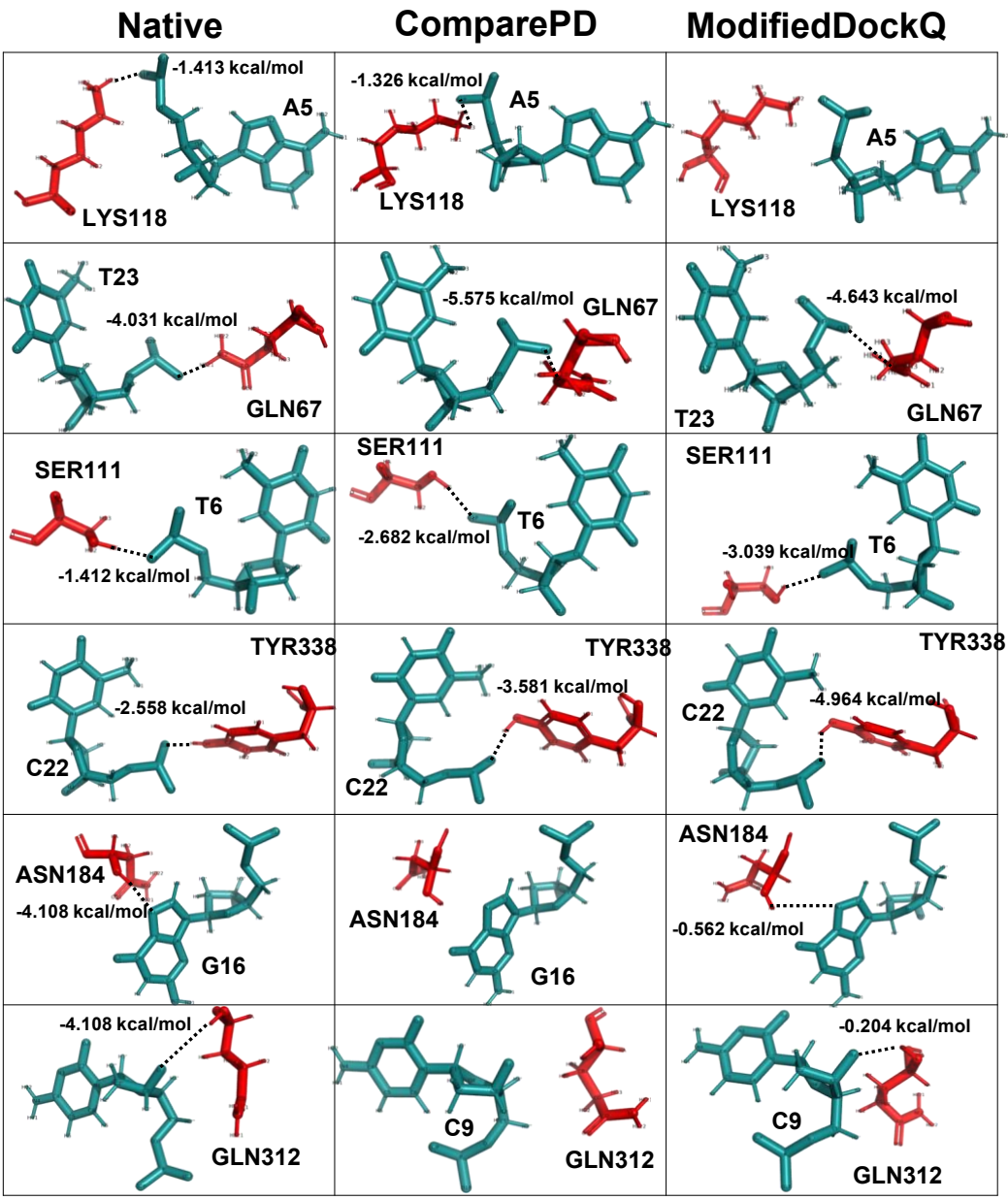

Figure 3.8. Interface hydrogen bonds and their energy in the native complex 1rva and top models predicted by ComparePD and ModifiedDockQ.

3.4. Discussion

Computational methods for predicting the structures of protein-DNA complexes not only fill the sequence-structure gap, they are also essential in time-sensitive applications such as

computer aided structure-based drug-design. Modeling the structure of complexes requires accurate mapping of interface features which is a challenging task and is not yet at a mature stage of development. New methods are being rapidly developed but they face several challenges. For example, unlike protein structure prediction where new methods can be assessed by several scores such as TM-score, GDT_TS and IS score, no standard criteria exist for comparing protein-DNA complexes [232–234]. Recent studies have stressed that incorporating biologically relevant measures in the development of in-silico structure prediction of complexes can help improve their performance [203]. We have presented here ComparePD, a novel scoring function to assess the similarity of protein-DNA complexes by incorporating energy of hydrogen bonds which are central to the binding specificity of protein-DNA complexes. A combination of conventional start-of-the-art criteria Fnat, iRMSD and lRMSD with the hydrogen bond energy in ComparePD could better reflect the true quality of a model. Proteins are dynamic in nature making it impossible to precisely conserve the exact energy of native hydrogen bonds. Instead of using discrete energy cut-offs, ComparePD treats native hydrogen bonds of varying strengths differently by assigning higher weights to stronger hydrogen bonds and lower weights to weaker. Small shifts in energy of hydrogen bonds within the same category are not penalized. A large shift in energy of conserved native hydrogen bond where the category of hydrogen bond has changed, is considered with a lower score (Table 3.4). While the use of hydrogen bond numbers for similarity assessment of complexes has previously been suggested, to the best of our knowledge, this is the first time a similarity assessment method based on different energy ranges has been explored to compare protein-DNA complexes.

ComparePD shows improvement in assessment of similarity over ModifiedDockQ by incorporating a single hydrogen bond energy-based term. It is significantly correlated to other

scoring metric indicating that while it focuses on hydrogen bonds, it does not neglect essential criteria of similarity assessment offered by other metrics. Overall, it outperforms ModifiedDockQ in more than 50% of the cases in our datasets by identifying a different, potentially better model. For instance, in homology models of 9antB, there is a clear difference between ModifiedDockQ and ComparePD scores for top models despite both having similar Fnat and iRMSD (Fig. 3.5A). A structural analysis of hydrogen bonds in both models reveals that the model selected by ComparePD is not only able to reproduce a larger number of native hydrogen bonds but also conserve the categories of their energy. In 1rva (Fig. 3.8) the model selected by ModifiedDockQ conserved more (5) hydrogen bonds as compared to ComparePD (4) but it does not conserve the energy of 3/5 hydrogen bonds. The selection by ComparePD conserves the energy of all four hydrogen bonds and also has better lRMSD as compared to both ModifiedDockQ selection (Fig. 3.5C).

ComparePD can serve as the standard criteria of comparison for protein-DNA complexes. It can facilitate the development of new methods for modeling and evaluating the structures of protein-DNA complexes. Since it is a continuous scoring function, it can be adapted in machine learning based methods to assess the quality of protein-DNA complexes. The performance of ComparePD can be further enhanced by training several weights used in this function.

Chapter 4. CONCLUSIONS AND FUTURE DIRECTIONS

4.1. Concluding remarks

Interface hydrogen bonds are pivotal to protein-DNA complexes and those formed between amino acid side chains and nucleotides base edges are central to the binding specificity. In this dissertation, we explored the importance of using hydrogen bond energy in similarity assessment of protein-DNA complexes, a key step in developing *in-silico* structure prediction methods. We first performed a comparative analysis of hydrogen bond energy distribution in protein-DNA complexes as compared to other biomolecular complexes [225]. The insight gained from these analyses was used to develop a novel hydrogen bond energy-based scoring function for structural similarity assessment of protein-DNA complexes. We showed that incorporating hydrogen bond energy leads to a clear improvement in accurate assessment of structural similarity in both homology and docking protein-DNA complex models. Our method selected different, potentially better models for more than 50% of targets which would otherwise be neglected using previous methods.

Computational structure prediction of protein-ligand complexes is a developing area of research and requires attention to the biologically relevant features of the type of complex. While the significance of hydrogen bonds in protein-DNA complexes is well studied, no comparative analysis of energy of hydrogen bonds exists. To this end, we first compared different features of hydrogen bonds, including their energy, in protein-DNA complexes, to two other major biomolecular complexes, protein-protein and protein-peptide complexes. The energy of hydrogen bonds was categorized into four bins corresponding to their strength. We showed that hydrogen bonds in protein-DNA complexes are significantly more prevalent than protein-protein and

protein-peptide complexes[225]. Moreover, a unique, almost equal distribution of weak and strong interface hydrogen bonds is observed in protein-DNA complexes. The hydrogen bonds at the interface of protein-protein and protein-peptide complexes are mostly strong. There is a significant difference in the energy of hydrogen bonds at the minor groove between highly specific and multi-specific complexes corroborating the importance of minor grooves in binding specificity of protein-DNA complexes reported by previous studies [22,147,179–181]. These results suggest that while a generic method of structural comparison can be used for protein-protein and protein-peptide complexes, special attention to hydrogen bonds is required in designing functions for protein-DNA complexes.

A key step in developing the computational methods for structure prediction of protein-DNA complexes is the comparison of predicted models to experimentally determined native structures. Accurate assessment of structural similarity is an important measure to evaluate and compare the performance of modeling methods. While generic criteria of evaluation have shown considerable success in the similarity assessment of monomeric proteins, special attention to the biologically relevant features is required in designing methods for complexes. For example, CAPRI based criteria of comparison use different distance cut-offs in protein-peptide complexes to reflect their compact interface as compared to protein-protein complexes. The existing methods of comparison are distance-based and do not consider the relevant energetic contribution of interactions. We developed a scoring function which combines the existing criteria of comparison to a novel hydrogen bond-based scoring term to compare the structures of protein-DNA complexes. We showed that assessing the similarity of complexes based on hydrogen bonds, which are critical to the function and binding specificity of protein-DNA complexes, results in marked improvements over existing methods.

4.2. Future directions: model quality assessment of protein-DNA complexes

The ultimate goal is to develop an accurate protein-DNA complex model assessment by applying the knowledge gained by comparative analysis of hydrogen bonds towards development of new interface potentials for quality assessment of protein-DNA complex models. We performed preliminary analyses by comparing the performance of three previously developed interface potentials to select the best model against our hydrogen bond energy-based similarity assessment method. Two knowledge-based potentials, orientation potential and multibody potential, and one physics based potential, HADDOCK score were used. Multibody potential is a distance-based, residue level energy function which assesses the interactions between amino acids and interaction units of DNA. DNA tri-nucleotides, referred to as triplets by multibody potential are used as interaction units to calculate two body, three body and four body interactions. Short distance hydrogen bonds and van der Waals interactions are considered in multibody potential [235]. Orientation potential is a knowledge-based, residue level energy function previously developed by our lab to discriminate near-native models from decoys [236]. The interaction energy in this method is calculated through statistical thermodynamic calculations of distance-based contact energies obtained for all residue-base interactions. Each interaction is assessed based on its geometrical properties, distance and angle. Due to its orientation dependent nature, it implicitly captures hydrogen bonds. HADDOCK score is a physics-based function of weighted energy terms such as van der Waals, electrostatic, desolvation, buried surface area and restraint violation energies [230,237]. The scoring function has previously been used to distinguish between near-native, acceptable and incorrect decoys [230,238]. Our analysis shows that no energy potential was able to select the model with the best hydrogen bond energy for more than 80% of targets in both docking and homology modeling dataset (Table 4.1). While both knowledge-based potentials

do not agree over the selection of top models in most cases (92% targets in docking dataset and 4 out of 5 targets in homology models), the orientation potential performs slightly better (Table 4.1). However, implicitly including hydrogen bonds through orientation potential results in selection of only 12% models with better hydrogen bond energy. While this is an improvement over multibody potential (4% same as ComparePD in docking and 1 out of 5 in homology modeling), it is not sufficient for accurate model quality assessment. Even though ComparePD and HADDOCK score do not agree over selection of top model for any complexes, both identify at least common model ranked among top three best models in 48% cases (Table 4.2).

Table 4.1A. Comparison of scoring methods for the top model in the docking dataset*

Same\Different	**ComparePD**	**Orientation potential**	**Multibody potential**	**HADDOCK score**
ComparePD		22 (88%)	24(96%)	25(100%)
Orientation potential	12(48%)		23(92%)	25(100%)
Multibody potential	1 (4%)	2(8%)		23(92%)
HADDOCK score	0 (0%)	0(0%)	2(8%)	

Table 4.1B. Comparison of scoring methods for the top model in the homology modeling dataset*

Same\Different	**ComparePD**	**Orientation potential**	**Multibody potential**
ComparePD		4 (80%)	5(100%)
Orientation potential	1 (20%)		4(80%)
Multibody potential	0 (0%)	1(20%)	

Table 4.2A. Comparison of scoring methods for the top three models in the docking dataset*

Same\Different	**ComparePD**	**Orientation potential**	**Multibody potential**	**HADDOCK score**

ComparePD		8 (32%)	11(44%)	25(100%)
Orientation potential	17(68%)		11(44%)	15(60%)
Multibody potential	14 (56%)	14(56%)		14(56%)
HADDOCK score	12(48%)	10(40%)	11(44%)	

Table 4.2B. Comparison of scoring methods for the top three models in the homology modeling dataset*

Same\Different	**ComparePD**	**Orientation potential**	**Multibody potential**
ComparePD		0(0%)	4(80%)
Orientation potential	5(100%)		3(60%)
Multibody potential	1(20%)	2(40%)	

* The lower half of the table presents similarities and the upper half shows differences between scores

Figure 4.1 shows three cases where these energy functions were compared against our method. Overall, orientation potential performed better than both HADDOCK and multibody potential. For example, in 7mht, only orientation potential predicted the near-native model with better hydrogen bond energy distribution (Fig. 4.1B). In the quality assessment of homology models of homeodomain-DNA complex 9antB, orientation potential also performed better than multibody and HADDOCK score. Yet, it still does not identify the best model. The model with best hydrogen bond energy, as identified by our method, ComparePD, is not even in the top 3 rankings by either of the energy function (Fig. 4.1A). The ideal candidate in terms of hydrogen bond energy should be model 1 predicted by using 4rduA as template with sequence identity of 42%. It conserves 7 out of 8 native hydrogen bonds and also conserves the energy of 5 strong hydrogen bonds (Fig. 4.2). All of these are low energy, strong hydrogen bonds including: ARG55-C406, ARG33-A404, ARG7-T518 and ASN53-A520. The two hydrogen bonds whose energy category changes in the model selected by ComparePD (GLN8-A519 and ARG55-C406) are high

energy, weaker hydrogen bonds. The energy of hydrogen bond between GLN8-A519 decreased in ComparePD selection whereas in ARG45-T521 it increases. This is also evident by the change in conformation of sidechains of the corresponding atoms. Sidechain conformations for hydrogen bonds with conserved energy categories in the model selected by ComparePD are similar to the near-native. The model selected by orientation potential is better than multibody potential in that it conserves 5 hydrogen bonds (ARG33-A404, ARG45-T521 and two hydrogen bonds in ASN53-520) with a change in category of 3 conserved hydrogen bonds (ARG33-A404, ARG45-T521 and one hydrogen bond between ASN53-520). The model selected by multibody potential reproduced only two of these hydrogen bonds (ASN53-520) and also conserves the category of their energy. A comparison of scores in the only case where multibody identified the model with near-native hydrogen bond energy distribution is shown (Fig. 4.1C).

Our analyses show that the existing methods to assess the model quality assessment of protein-DNA complexes do not capture the hydrogen bond energy framework of native complex. New methods are needed which explicitly incorporate the energy of hydrogen bonds to assess the quality of computationally predicted models. The knowledge gained in this study can be expanded towards the development of such functions.

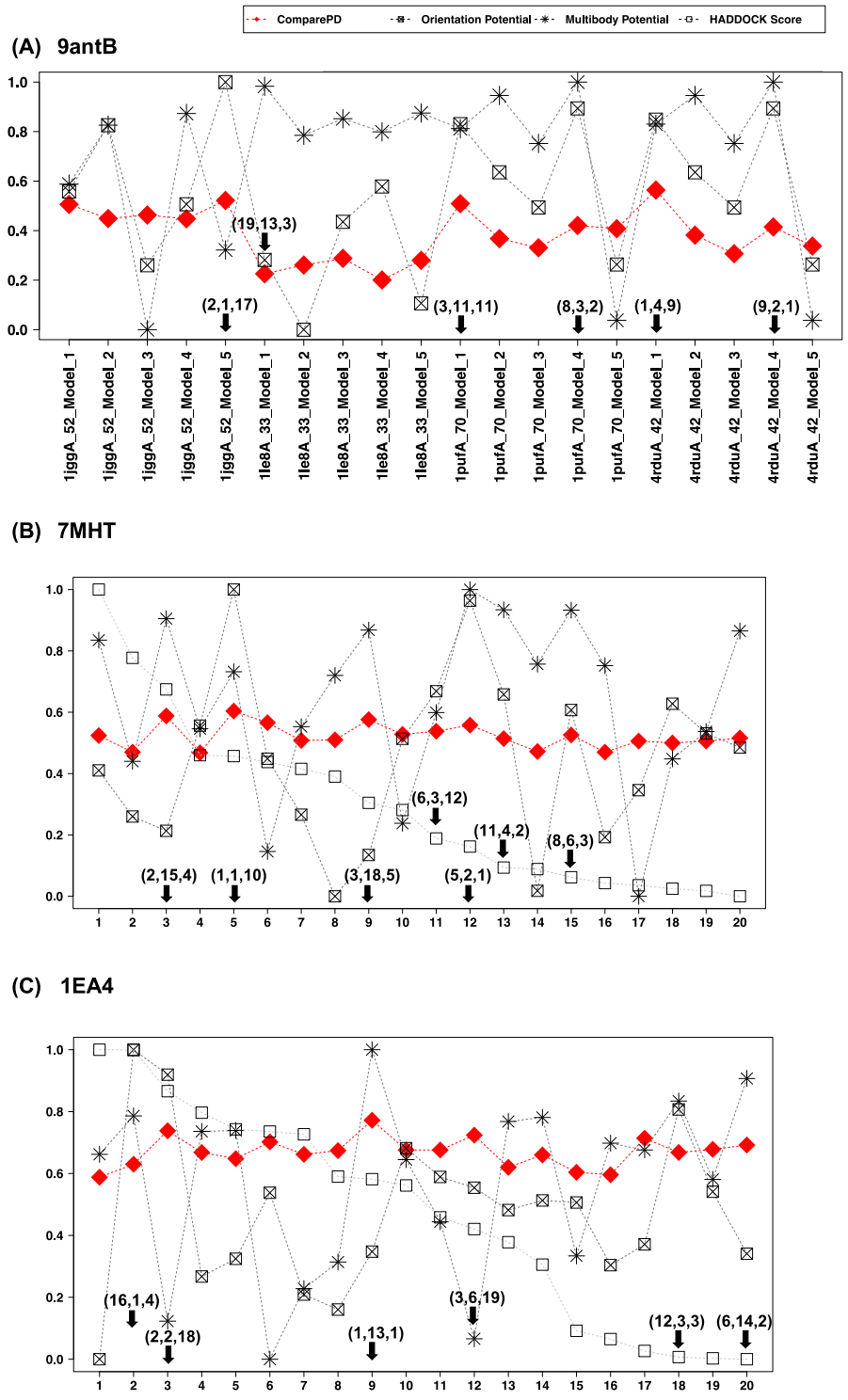

Figure 4.1. Comparison plots of the models scored by ComparePD along with the corresponding orientation potential, multibody potential and HADDOCK score for (A) 9antB, (B) 7mht and (C) 1ea4. The top three models selected by each method are highlighted and corresponding ranks are reported as (ComparePD, Orientation potential, multibody potential) for each complex.

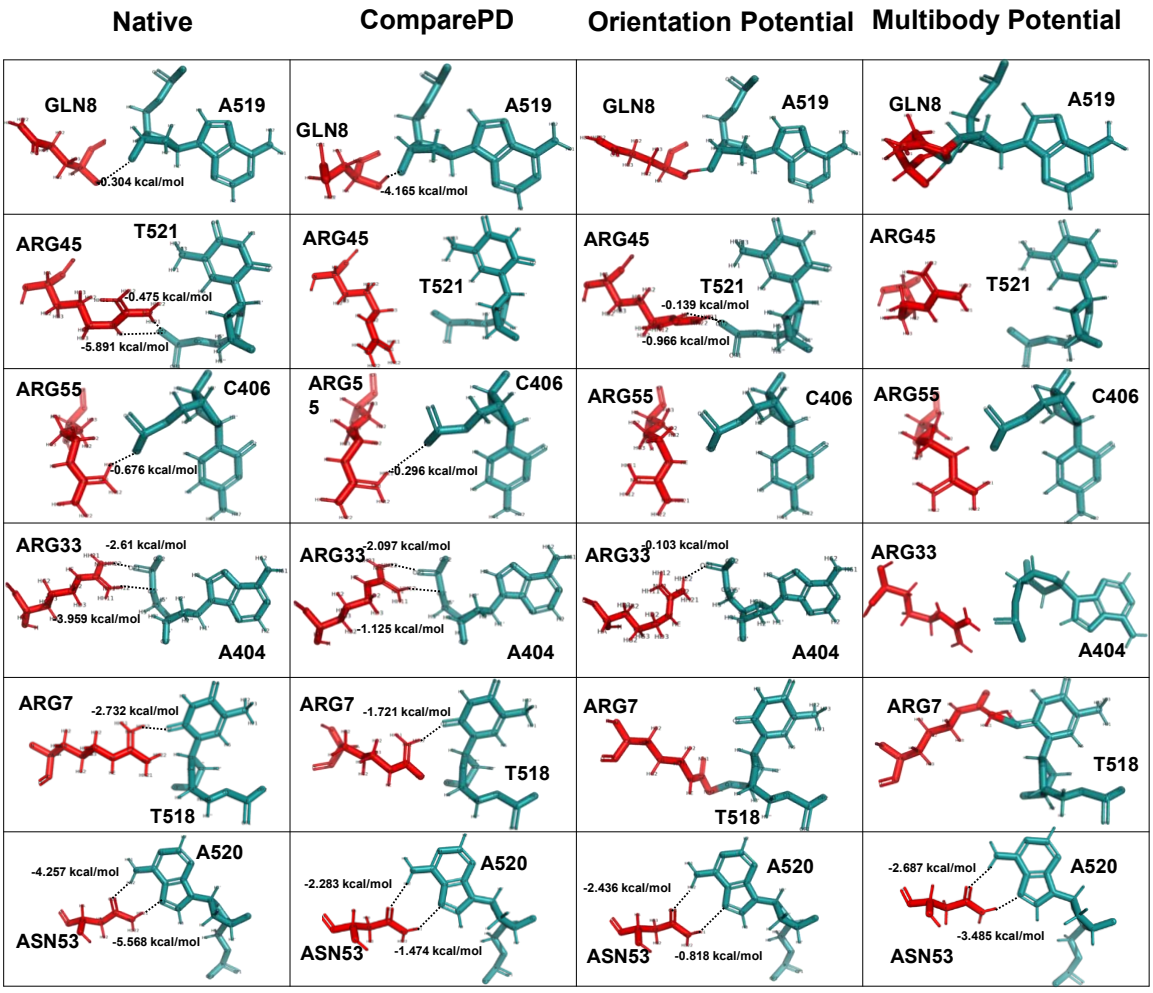

Figure 4.2. Hydrogen bond comparison in different homology models of the homeodomain-DNA complex 9antB selected by ComparePD, orientation potential and multibody potential.

REFERENCES

1. Schott J-J, Benson DW, Basson CT, et al. Congenital Heart Disease Caused by Mutations in the Transcription Factor NKX2-5. Science (80-.). 1998; 281:108–111

2. Filippova GN, Qi C, Ulmer JE, et al. Advances in Brief Tumor-associated Zinc Finger Mutations in the CTCF Transcription Factor Selectively Alter Its DNA-binding Specificity 1. 2002; 48–52

3. Luscombe NM, Thornton JM. Protein–DNA Interactions: Amino Acid Conservation and the Effects of Mutations on Binding Specificity. J. Mol. Biol. 2002; 320:991–1009

4. Latchman DS. Transcription-Factor Mutations and Disease. N. Engl. J. Med. 1996; 334:28–33

5. Corona RI, Guo J. Statistical analysis of structural determinants for protein–DNA-binding specificity. Proteins Struct. Funct. Bioinforma. 2016; 84:1147–1161

6. Agback P, Baumann H, Knapp S, et al. Architecture of nonspecific protein–DNA interactions in the Sso7d–DNA complex. Nat. Struct. Biol. 1998; 5:579–584

7. Göhler T, Jäger S, Warnecke G, et al. Mutant p53 proteins bind DNA in a DNA structure-selective mode. Nucleic Acids Res. 2005; 33:1087–1100

8. Thukral SK, Lu Y, Blain GC, et al. Discrimination of DNA binding sites by mutant p53 proteins. Mol. Cell. Biol. 1995; 15:5196–5202

9. Chène P. Mutations at Position 277 Modify the DNA-Binding Specificity of Human p53 in Vitro. Biochem. Biophys. Res. Commun. 1999; 263:1–5

10. Shiroma Y, Takahashi R-U, Yamamoto Y, et al. Targeting DNA binding proteins for cancer therapy. Cancer Sci. 2020; 111:1058–1064

11. Radaeva M, Ton A-T, Hsing M, et al. Drugging the ‘undruggable’. Therapeutic targeting of protein–DNA interactions with the use of computer-aided drug discovery methods. Drug Discov. Today 2021; 26:2660–2679

12. Kim S, Kang M, Ko J. Small leucine zipper protein promotes the metastasis of castration-resistant prostate cancer through transcriptional regulation of matrix metalloproteinase-13. Carcinogenesis 2021; 42:1089–1099

13. Liang S, Thomas SE, Chaplin AK, et al. Structural insights into inhibitor regulation of the DNA repair protein DNA-PKcs. Nature 2022; 601:643–648

14. Siggers T, Gordân R. Protein–DNA binding: complexities and multi-protein codes. Nucleic Acids Res. 2014; 42:2099–2111

15. Sarai A, Kono H. Protein-DNA Recognition Patterns and Predictions. Annu. Rev. Biophys. Biomol. Struct. 2005; 34:379–398

16. K. OW, A. GA, Xiang-Jun L, et al. DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. Proc. Natl. Acad. Sci. 1998; 95:11163–11168

17. Drew HR, Travers AA. Structural junctions in DNA: the influence of flanking sequence on nuclease digestion specificities. Nucleic Acids Res. 1985; 13:4445–4467

18. Michael Gromiha M, Siebers JG, Selvaraj S, et al. Intermolecular and Intramolecular Readout Mechanisms in Protein–DNA Recognition. J. Mol. Biol. 2004; 337:285–294

19. Sarai A, Mazur J, Nussinov R, et al. Sequence dependence of DNA conformational flexibility. Biochemistry 1989; 28:7842–7849

20. Hogan ME, Austin RH. Importance of DNA stiffness in protein–DNA binding specificity. Nature 1987; 329:263–266

21. Rawat N, Biswas P. Shape, flexibility and packing of proteins and nucleic acids in complexes. Phys. Chem. Chem. Phys. 2011; 13:9632–9643

22. Rohs R, West SM, Sosinsky A, et al. The role of DNA shape in protein-DNA recognition. Nature 2009; 461:1248–1253

23. Fersht AR. The hydrogen bond in molecular recognition. Trends Biochem. Sci. 1987; 12:301–304

24. Dixit SB, Arora N, Jayaram B. How do hydrogen bonds contribute to protein-DNA recognition? J. Biomol. Struct. Dyn. 2000; 17:109–112

25. Nittinger E, Inhester T, Bietz S, et al. Large-Scale Analysis of Hydrogen Bond Interaction Patterns in Protein–Ligand Interfaces. J. Med. Chem. 2017; 60:4245–4257

26. Panigrahi SK, Desiraju GR. Strong and weak hydrogen bonds in the protein–ligand interface. Proteins Struct. Funct. Bioinforma. 2007; 67:128–141

27. Sarkhel S, Desiraju GR. N□H…O, O□H…O, and C□H…O hydrogen bonds in protein–ligand complexes: Strong and weak interactions in molecular recognition. Proteins Struct. Funct. Bioinforma. 2004; 54:247–259

28. Itoh Y, Nakashima Y, Tsukamoto S, et al. N(+)-C-H···O Hydrogen bonds in protein-ligand

complexes. Sci. Rep. 2019; 9:767

29. Shahi A, Arunan E. Why are Hydrogen Bonds Directional? J. Chem. Sci. 2016; 128:

30. Freddolino PL, Harrison CB, Liu Y, et al. Challenges in protein-folding simulations. Nat. Phys. 2010; 6:751–758

31. Morozov A V, Kortemme T, Tsemekhman K, et al. Close agreement between the orientation dependence of hydrogen bonds observed in protein structures and quantum mechanical calculations. Proc. Natl. Acad. Sci. 2004; 101:6946–6951

32. Kortemme T, Morozov A V, Baker D. An Orientation-dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein–Protein Complexes. J. Mol. Biol. 2003; 326:1239–1259

33. Jacobs DJ, Rader AJ, Kuhn LA, et al. Protein flexibility predictions using graph theory. Proteins Struct. Funct. Bioinforma. 2001; 44:150–165

34. Sheu S-Y, Yang D-Y, Selzle HL, et al. Energetics of hydrogen bonds in peptides. Proc. Natl. Acad. Sci. 2003; 100:12683 LP – 12687

35. Huang N, MacKerell AD. An ab Initio Quantum Mechanical Study of Hydrogen-Bonded Complexes of Biological Interest. J. Phys. Chem. A 2002; 106:7820–7827

36. Lozynski M, Rusinska-Roszak D, Mack H-G. MP2 and Density Functional Studies of Hydrogen Bonding in Model Trioses: d-(+)-Glyceraldehyde and Dihydroxyacetone. J. Phys. Chem. A 1997; 101:1542–1548

37. Rablen PR, Lockman JW, Jorgensen WL. Ab Initio Study of Hydrogen-Bonded Complexes

of Small Organic Molecules with Water. J. Phys. Chem. A 1998; 102:3782–3797

38. Bondesson L, Mikkelsen K V, Luo Y, et al. Density functional theory calculations of hydrogen bonding energies of drug molecules. J. Mol. Struct. THEOCHEM 2006; 776:61–68

39. Elstner M, Hobza P, Frauenheim T, et al. Hydrogen bonding and stacking interactions of nucleic acid base pairs: A density-functional-theory based treatment. J. Chem. Phys. 2001; 114:5149–5155

40. Mmereki BT, Donaldson DJ. Ab Initio and Density Functional Study of Complexes between the Methylamines and Water. J. Phys. Chem. A 2002; 106:3185–3190

41. González L, Mó O, Yáñez M. High-level ab initio versus DFT calculations on (H2O2)2 and H2O2–H2O complexes as prototypes of multiple hydrogen bond systems. J. Comput. Chem. 1997; 18:1124–1135

42. Sim F, St. Amant A, Papai I, et al. Gaussian density functional calculations on hydrogen-bonded systems. J. Am. Chem. Soc. 1992; 114:4391–4400

43. Tuma C, Daniel Boese A, C. Handy N. Predicting the binding energies of H-bonded complexes: A comparative DFT study. Phys. Chem. Chem. Phys. 1999; 1:3939–3947

44. Latajka Z, Bouteiller Y. Application of density functional methods for the study of hydrogen-bonded systems: The hydrogen fluoride dimer. J. Chem. Phys. 1994; 101:9793–9799

45. Laffort P. A Slightly Modified Expression of the Polar Surface Area Applied to an Olfactory Study. Open J. Phys. Chem. 2013; 03:150–156

46. Kaur D, Khanna S. Intermolecular hydrogen bonding interactions of furan, isoxazole and

oxazole with water. Comput. Theor. Chem. 2011; 963:71–75

47. Gancia E, Montana J, Manallack D. Theoretical hydrogen bonding parameters for drug design. J. Mol. Graph. Model. 2001; 19:349–362

48. Dahiyat BI, Gordon DB, Mayo SL. Automated design of the surface positions of protein helices. Protein Sci. 1997; 6:1333–1337

49. Pauling L, Corey RB. Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. Proc. Natl. Acad. Sci. U. S. A. 1951; 37:729

50. Pauling L, Corey RB, Branson HR. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. Proc. Natl. Acad. Sci. 1951; 37:205–211

51. Pace CN, Fu H, Lee Fryar K, et al. Contribution of hydrogen bonds to protein stability. Protein Sci. 2014; 23:652–661

52. Gao J, Bosco DA, Powers ET, et al. Localized thermodynamic coupling between hydrogen bonding and microenvironment polarity substantially stabilizes proteins. Nat. Struct. Mol. Biol. 2009; 16:684–690

53. Sawada T, Fedorov DG, Kitaura K. Role of the key mutation in the selective binding of avian and human influenza hemagglutinin to sialosides revealed by quantum-mechanical calculations. J. Am. Chem. Soc. 2010; 132:16862–16872

54. Salentin S, Haupt VJ, Daminelli S, et al. Polypharmacology rescored: Protein–ligand interaction profiles for remote binding site similarity assessment. Prog. Biophys. Mol. Biol. 2014; 116:174–186

55. Rehman I, Farooq M, Botelho S. Biochemistry, secondary protein structure. StatPearls [Internet] 2021;

56. Hubbard RE, Kamran Haider M. Hydrogen Bonds in Proteins: Role and Strength. eLS 2010;

57. Takano K, Scholtz JM, Sacchettini JC, et al. The contribution of polar group burial to protein stability is strongly context-dependent. J. Biol. Chem. 2003; 278:31790–31795

58. Milo R, Phillips R. Cell biology by the numbers. 2015;

59. Steiner T. The hydrogen bond in the solid state. Angew. Chemie Int. Ed. 2002; 41:48–76

60. Ferreira de Freitas R, Schapira M. A systematic analysis of atomic protein–ligand interactions in the PDB. Medchemcomm 2017; 8:1970–1981

61. Muley L, Baum B, Smolinski M, et al. Enhancement of Hydrophobic Interactions and Hydrogen Bond Strength by Cooperativity: Synthesis, Modeling, and Molecular Dynamics Simulations of a Congeneric Series of Thrombin Inhibitors. J. Med. Chem. 2010; 53:2126–2135

62. Shirley BA, Stanssens P, Hahn U, et al. Contribution of hydrogen bonding to the conformational stability of ribonuclease T1. Biochemistry 1992; 31:725–732

63. Giletto A, Pace CN. Buried, charged, non-ion-paired aspartic acid 76 contributes favorably to the conformational stability of ribonuclease T1. Biochemistry 1999; 38:13379–13384

64. Serrano L, Kellis Jr JT, Cann P, et al. The folding of an enzyme: II. Substructure of barnase and the contribution of different interactions to protein stability. J. Mol. Biol. 1992; 224:783–804

65. Hebert EJ, Giletto A, Sevcik J, et al. Contribution of a conserved asparagine to the conformational stability of ribonucleases Sa, Ba, and T1. Biochemistry 1998; 37:16192–16200

66. Pace CN, Horn G, Hebert EJ, et al. Tyrosine hydrogen bonds make a large contribution to protein stability. J. Mol. Biol. 2001; 312:393–404

67. Tsai C, Xu D, Nussinov R. Structural motifs at protein-protein interfaces: Protein cores versus two-state and three-state model complexes. Protein Sci. 1997; 6:1793–1805

68. Xu D, Tsai CJ, Nussinov R. Hydrogen bonds and salt bridges across protein-protein interfaces. Protein Eng. 1997; 10:999–1012

69. Erijman A, Rosenthal E, Shifman JM. How structure defines affinity in protein-protein interactions. PLoS One 2014; 9:e110085–e110085

70. Jones S, Thornton JM. Principles of protein-protein interactions. Proc. Natl. Acad. Sci. 1996; 93:13–20

71. London N, Movshovitz-Attias D, Schueler-Furman O. The Structural Basis of Peptide-Protein Binding Strategies. Structure 2010; 18:188–199

72. Yeh C-S, Chen F-M, Wang J-Y, et al. Directional shape complementarity at the protein–DNA interface. J. Mol. Recognit. 2003; 16:213–222

73. Luscombe NM, Laskowski RA, Thornton JM. Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. Nucleic Acids Res. 2001; 29:2860–2874

74. Baker CM, Grant GH. Role of aromatic amino acids in protein–nucleic acid recognition.

Biopolym. Orig. Res. Biomol. 2007; 85:456–470

75. Ashworth J, Havranek JJ, Duarte CM, et al. Computational redesign of endonuclease DNA binding and cleavage specificity. Nature 2006; 441:656–659

76. Angarica VE, Pérez AG, Vasconcelos AT, et al. Prediction of TF target sites based on atomistic models of protein-DNA complexes. BMC Bioinformatics 2008; 9:436

77. Sheehan D, O'Sullivan S. Online homology modelling as a means of bridging the sequence-structure gap. Bioeng. Bugs 2011; 2:299–305

78. Goodsell DS. The Protein Data Bank: exploring biomolecular structure. Nat. Educ. 2010; 3:39

79. Campagne S, Gervais V, Milon A. Nuclear magnetic resonance analysis of protein-DNA interactions. J. R. Soc. Interface 2011; 8:1065–1078

80. Carey MF, Peterson CL, Smale ST. Experimental Strategies for the Identification of DNA-Binding Proteins. 2012; 18–34

81. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. Nucleic Acids Res. 2000; 28:235–242

82. Gawas UB, Mandrekar VK, Majik MS. Chapter 5 - Structural analysis of proteins using X-ray diffraction technique. 2019; 69–84

83. Navaza J. AMoRe: an automated package for molecular replacement. Acta Crystallogr. Sect. A Found. Crystallogr. 1994; 50:157–163

84. Wüthrich K. Protein structure determination in solution by NMR spectroscopy. J. Biol. Chem. 1990; 265:22059–22062

85. Banci L, Bertini I, Luchinat C, et al. NMR in structural proteomics and beyond. Prog. Nucl. Magn. Reson. Spectrosc. 2010; 56:247–266

86. Rehm T, Huber R, Holak TA. Application of NMR in structural proteomics: screening for proteins amenable to structural analysis. Structure 2002; 10:1613–1618

87. Schwede T. Protein modeling: what happened to the 'protein structure gap'? Structure 2013; 21:1531–1540

88. Schlick T, Collepardo-Guevara R, Halvorsen LA, et al. Biomolecular modeling and simulation: a field coming of age. Q. Rev. Biophys. 2011; 44:191–228

89. Gao M, Zhou H, Skolnick J. DESTINI: A deep-learning approach to contact-driven protein structure prediction. Sci. Rep. 2019; 9:

90. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. Nature 2020; 577:706–710

91. Zheng W, Li Y, Zhang C, et al. Deep-learning contact-map guided protein structure prediction in CASP13. Proteins Struct. Funct. Bioinforma. 2019; 87:1149–1164

92. Narykov O, Srinivasan S, Korkin D. Computational protein modeling and the next viral pandemic. Nat. Methods 2021; 18:444–445

93. Gu J, Bourne PE. Structural bioinformatics. 2009; 44:

94. Kaczanowski S, Zielenkiewicz P. Why similar protein sequences encode similar three-dimensional structures? Theor. Chem. Acc. 2010; 125:643–650

95. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J. 1986; 5:823–826

96. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402

97. Apweiler R, Martin MJ, O'Donovan C, et al. The universal protein resource (UniProt) in 2010. Nucleic Acids Res. 2010; 38:D142–D148

98. Wang G, Dunbrack Jr RL. Scoring profile-to-profile sequence alignments. Protein Sci. 2004; 13:1612–1626

99. Söding J. Protein homology detection by HMM–HMM comparison. Bioinformatics 2005; 21:951–960

100. Dalton JAR, Jackson RM. An evaluation of automated homology modelling methods at low target–template sequence similarity. Bioinformatics 2007; 23:1901–1908

101. Larkin MA, Blackshields G, Brown NP, et al. Clustal W and Clustal X version 2.0. bioinformatics 2007; 23:2947–2948

102. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 2000; 302:205–217

103. O'Sullivan O, Suhre K, Abergel C, et al. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. J. Mol. Biol. 2004; 340:385–395

104. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32:1792–1797

105. Katebi AR, Kloczkowski A, Jernigan RL. Structural interpretation of protein-protein interaction network. BMC Struct. Biol. 2010; 10:1–11

106. Orengo CA, Sillitoe I, Reeves G, et al. What can structural classifications reveal about protein evolution? J. Struct. Biol. 2001; 134:145–165

107. Kim T, Oh S, Yang JS, et al. A simplified homology-model builder toward highly protein-like structures: An inspection of restraining potentials. J. Comput. Chem. 2012; 33:1927–1935

108. Arnold K, Bordoli L, Kopp J, et al. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 2006; 22:195–201

109. Levitt M. Accurate modeling of protein conformation by automatic segment matching. J. Mol. Biol. 1992; 226:507–533

110. Šali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 1993; 234:779–815

111. Muhammed MT, Aki-Yalcin E. Homology modeling in drug discovery: Overview, current applications, and future perspectives. Chem. Biol. Drug Des. 2019; 93:12–20

112. Wang J, Li W, Wang B, et al. In silicon approach for discovery of chemopreventive agents. Curr. Pharmacol. Reports 2017; 3:184–195

113. Balmith M, Faya M, Soliman MES. Ebola virus: A gap in drug design and discovery-experimental and computational perspective. Chem. Biol. Drug Des. 2017; 89:297–308

114. Ramharack P, Soliman MES. Zika virus NS5 protein potential inhibitors: an enhanced in silico approach in drug discovery. J. Biomol. Struct. Dyn. 2018; 36:1118–1133

115. Koseki Y, Aoki S. Computational medicinal chemistry for rational drug design: Identification of novel chemical structures with potential anti-tuberculosis activity. Curr. Top. Med. Chem. 2014; 14:176–188

116. Rost B, Sander C. Bridging the protein sequence-structure gap by structure predictions. Annu. Rev. Biophys. Biomol. Struct. 1996; 25:113–136

117. Wallner B, Elofsson A. All are not equal: A benchmark of different homology modeling programs. Protein Sci. 2005; 14:1315–1327

118. N. R, S.M. M, C. N, et al. Interaction interfaces of protein domains are not topologically equivalent across families within superfamilies: Implications for metabolic and signaling pathways. Proteins Struct. Funct. Bioinforma. 2004; 58:339–353

119. Wang Z, Sun H, Yao X, et al. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. Phys. Chem. Chem. Phys. 2016; 18:12964–12975

120. Andrusier N, Mashiach E, Nussinov R, et al. Principles of flexible protein-protein docking. Proteins 2008; 73:271–289

121. Huang S-Y. Comprehensive assessment of flexible-ligand docking algorithms: current effectiveness and challenges. Brief. Bioinform. 2018; 19:982–994

122. Khamis MA, Gomaa W, Ahmed WF. Machine learning in computational docking. Artif.

Intell. Med. 2015; 63:135–152

123. Forli S, Huey R, Pique ME, et al. Computational protein–ligand docking and virtual drug screening with the AutoDock suite. Nat. Protoc. 2016; 11:905–919

124. Meng X-Y, Zhang H-X, Mezei M, et al. Molecular docking: a powerful approach for structure-based drug discovery. Curr. Comput. Aided. Drug Des. 2011; 7:146–157

125. Goodsell DS, Morris GM, Olson AJ. Automated docking of flexible ligands: applications of AutoDock. J. Mol. Recognit. 1996; 9:1–5

126. Soares TA, Goodsell DS, Briggs JM, et al. Docking of 4-oxalocrotonate tautomerase substrates: Implications for the catalytic mechanism. Biopolym. Orig. Res. Biomol. 1999; 50:319–328

127. Rosenfeld RJ, Goodsell DS, Musah RA, et al. Automated docking of ligands to an artificial active site: augmenting crystallographic analysis with computer modeling. J. Comput. Aided. Mol. Des. 2003; 17:525–536

128. Leonard SE, Garcia FJ, Goodsell DS, et al. Redox-based probes for protein tyrosine phosphatases. Angew. Chemie Int. Ed. 2011; 50:4423–4427

129. Brik A, Muldoon J, Lin Y, et al. Rapid diversity-oriented synthesis in microtiter plates for in situ screening of HIV protease inhibitors. ChemBioChem 2003; 4:1246–1248

130. Brik A, Alexandratos J, Lin Y, et al. 1, 2, 3-Triazole as a peptide surrogate in the rapid synthesis of HIV-1 protease inhibitors. ChemBioChem 2005; 6:1167–1169

131. Perryman AL, Zhang Q, Soutter HH, et al. Fragment-based screen against HIV protease. Chem. Biol. Drug Des. 2010; 75:257–268

132. Cosconati S, Hong JA, Novellino E, et al. Structure-based virtual screening and biological evaluation of Mycobacterium tuberculosis adenosine 5′-phosphosulfate reductase inhibitors. J. Med. Chem. 2008; 51:6627–6630

133. Perryman AL, Yu W, Wang X, et al. A virtual screen discovers novel, fragment-sized inhibitors of Mycobacterium tuberculosis InhA. J. Chem. Inf. Model. 2015; 55:645–659

134. Cosconati S, Huey R, Marinelli L, et al. Identification of novel β-secretase inhibitors through the inclusion of protein flexibility in virtual screening calculations. 2008;

135. Cosconati S, Marinelli L, Di Leva FS, et al. Protein flexibility in virtual screening: the BACE-1 case study. J. Chem. Inf. Model. 2012; 52:2697–2704

136. Goodsell DS, Lauble H, Stout CD, et al. Automated docking in crystallography: analysis of the substrates of aconitase. Proteins Struct. Funct. Bioinforma. 1993; 17:1–10

137. Li J, Fu A, Zhang L. An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking. Interdiscip. Sci. Comput. Life Sci. 2019; 11:320–328

138. Yuriev E, Agostino M, Ramsland PA. Challenges and advances in computational docking: 2009 in review. J. Mol. Recognit. 2011; 24:149–164

139. Du X, Li Y, Xia Y-L, et al. Insights into Protein-Ligand Interactions: Mechanisms, Models, and Methods. Int. J. Mol. Sci. 2016; 17:144

140. Pandey P, Hasnain S, Ahmad S. Encyclopedia of Bioinformatics and Computational

Biology. 2019; 142–154

141. Stanfield RL, Wilson IA. Protein-peptide interactions. Curr. Opin. Struct. Biol. 1995; 5:103–113

142. Mendoza F, Espino P, Cann K, et al. Anti-tumor chemotherapy utilizing peptide-based approaches - Apoptotic pathways, kinases and proteasome as targets. Arch. Immunol. Ther. Exp. (Warsz). 2005; 53:47–60

143. Trellet M, Melquiond ASJ, Bonvin AMJJ. A Unified Conformational Selection and Induced Fit Approach to Protein-Peptide Docking. PLoS One 2013; 8:e58769

144. Hardcastle IR. 5.06 - Protein–Protein Interaction Inhibitors in Cancer. 2017; 154–201

145. Coulocheri SA, Pigis DG, Papavassiliou KA, et al. Hydrogen bonds in protein-DNA complexes: Where geometry meets plasticity. Biochimie 2007; 89:1291–1303

146. Mandel-Gutfreund Y, Schueler O, Margalit H. Comprehensive Analysis of Hydrogen Bonds in Regulatory Protein DNA-Complexes: In Search of Common Principles. J. Mol. Biol. 1995; 253:370–382

147. Rohs R, Jin X, West SM, et al. Origins of Specificity in Protein-DNA Recognition. Annu. Rev. Biochem. 2010; 79:233–269

148. Dixit SB, Arora N, Jayaram B. How Do Hydrogen Bonds Contribute to Protein-DNA Recognition? J. Biomol. Struct. Dyn. 2000; 17:109–112

149. Mukherjee S, Majumdar S, Bhattacharyya D. Role of Hydrogen Bonds in Protein−DNA Recognition: Effect of Nonplanar Amino Groups. J. Phys. Chem. B 2005; 109:10484–10492

150. Dai L, Xu Y, Du Z, et al. Revealing atomic-scale molecular diffusion of a plant-transcription factor WRKY domain protein along DNA. Proc. Natl. Acad. Sci. U. S. A. 2021; 118:1–10

151. Lu H, Lu L, Skolnick J. Development of unified statistical potentials describing protein-protein interactions. Biophys. J. 2003; 84:1895–1901

152. Levy ED. A Simple Definition of Structural Regions in Proteins and Its Use in Analyzing Interface Evolution. J. Mol. Biol. 2010; 403:660–670

153. Worth CL, Blundell TL. Satisfaction of hydrogen-bonding potential influences the conservation of polar sidechains. Proteins Struct. Funct. Bioinforma. 2009; 75:413–429

154. Kota P, Ding F, Ramachandran S, et al. Gaia: automated quality assessment of protein structure models. Bioinformatics 2011; 27:2209–2215

155. Song W, Guo J-T. Investigation of arc repressor DNA-binding specificity by comparative molecular dynamics simulations. J. Biomol. Struct. Dyn. 2015; 33:2083–2093

156. Laederach A, Reilly PJ. Specific empirical free energy function for automated docking of carbohydrates to proteins. J. Comput. Chem. 2003; 24:1748–1757

157. Morozov A V, Havranek JJ, Baker D, et al. Protein-DNA binding specificity predictions with structural models. Nucleic Acids Res. 2005; 33:5781–5798

158. Eildal JNN, Hultqvist G, Balle T, et al. Probing the Role of Backbone Hydrogen Bonds in Protein–Peptide Interactions by Amide-to-Ester Mutations. J. Am. Chem. Soc. 2013; 135:12998–13007

159. Stranges PB, Kuhlman B. A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds. Protein Sci. 2013; 22:74–82

160. Jiang L, Lai L. CH...O hydrogen bonds at protein-protein interfaces. J. Biol. Chem. 2002; 277:37732–37740

161. Zhou S, Wang L. Unraveling the structural and chemical features of biological short hydrogen bonds. Chem. Sci. 2019; 10:7734–7745

162. Kim R, Corona RI, Hong B, et al. Benchmarks for flexible and rigid transcription factor-DNA docking. BMC Struct. Biol. 2011; 11:45

163. Hauser AS, Windshügel B. LEADS-PEP: A Benchmark Data Set for Assessment of Peptide Docking Performance. J. Chem. Inf. Model. 2016; 56:188–200

164. Johansson-Åkhe I, Mirabello C, Wallner B. Predicting protein-peptide interaction sites using distant protein complexes as structural templates. Sci. Rep. 2019; 9:4267

165. Chen H, Skolnick J. M-TASSER: an algorithm for protein quaternary structure prediction. Biophys. J. 2008; 94:918–928

166. Vreven T, Moal IH, Vangone A, et al. Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. J. Mol. Biol. 2015; 427:3031–3041

167. Berman HM, Battistuz T, Bhat TN, et al. The Protein Data Bank. Acta Crystallogr. Sect. D 2002; 58:899–907

168. Wang G, Dunbrack Jr RL. PISCES: a protein sequence culling server. Bioinformatics

2003; 19:1589–1591

169. McDonald IK, Thornton JM. Satisfying Hydrogen Bonding Potential in Proteins. J. Mol. Biol. 1994; 238:777–793

170. Word JM, Lovell SC, Richardson JS, et al. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J. Mol. Biol. 1999; 285:1735–1747

171. Dahiyat BI, Mayo SL. Probing the role of packing specificity in protein design. Proc. Natl. Acad. Sci. 1997; 94:10172 LP – 10177

172. Hubbard S, Thornton J. NACCESS: Department of Biochemistry and Molecular Biology, University College London. 1993;

173. Seeman NC, Rosenberg JM, Rich A. Sequence-specific recognition of double helical nucleic acids by proteins. Proc. Natl. Acad. Sci. U. S. A. 1976; 73:804–808

174. Marcu O, Dodson E-J, Alam N, et al. FlexPepDock lessons from CAPRI peptide–protein rounds and suggested new criteria for assessment of model quality and utility. Proteins Struct. Funct. Bioinforma. 2017; 85:445–462

175. Corona RI, Sudarshan S, Aluru S, et al. An SVM-based method for assessment of transcription factor-DNA complex models. BMC Bioinformatics 2018; 19:506

176. Farrel A, Murphy J, Guo J. Structure-based prediction of transcription factor binding specificity using an integrative energy function. Bioinformatics 2016; 32:i306–i313

177. Farrel A, Guo J. An efficient algorithm for improving structure-based prediction of transcription factor binding sites. BMC Bioinformatics 2017; 18:342

178. Lin M, Guo JT. New insights into protein-DNA binding specificity from hydrogen bond based comparative study. Nucleic Acids Res. 2019; 47:11103–11113

179. Slattery M, Zhou T, Yang L, et al. Absence of a simple code: how transcription factors read the genome. Trends Biochem. Sci. 2014; 39:381–399

180. Chiu T-P, Rao S, Mann RS, et al. Genome-wide prediction of minor-groove electrostatic potential enables biophysical modeling of protein–DNA binding. Nucleic Acids Res. 2017; 45:12565–12576

181. Dantas Machado AC, Cooper BH, Lei X, et al. Landscape of DNA binding signatures of myocyte enhancer factor-2B reveals a unique interplay of base and shape readout. Nucleic Acids Res. 2020; 48:8529–8544

182. Hudson WH, Ortlund EA. The structure, function and evolution of proteins that bind DNA and RNA. Nat. Rev. Mol. cell Biol. 2014; 15:749–760

183. Wang L, Brown SJ. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res. 2006; 34:W243–W248

184. Tan S, Davey CA. Nucleosome structural studies. Curr. Opin. Struct. Biol. 2011; 21:128–136

185. Lilley DMJ. The interaction of four-way DNA junctions with resolving enzymes. Biochem. Soc. Trans. 2010; 38:399–403

186. Budayeva HG, Cristea IM. A mass spectrometry view of stable and transient protein interactions. Adv. Mass Spectrom. Biomed. Res. 2014; 263–282

187. Iacobucci I, Monaco V, Cozzolino F, et al. From classical to new generation approaches: An excursus of-omics methods for investigation of protein-protein interaction networks. J. Proteomics 2021; 230:103990

188. Stoddard BL. Homing endonucleases: from microbial genetic invaders to reagents for targeted DNA modification. Structure 2011; 19:7–15

189. Pröpper K, Meindl K, Sammito M, et al. Structure solution of DNA-binding proteins and complexes with ARCIMBOLDO libraries. Acta Crystallogr. D. Biol. Crystallogr. 2014; 70:1743–1757

190. Berman HM, Bhat TN, Bourne PE, et al. The Protein Data Bank and the challenge of structural genomics. Nat. Struct. Biol. 2000; 7:957–959

191. Hameduh T, Haddad Y, Adam V, et al. Homology modeling in the time of collective and artificial intelligence. Comput. Struct. Biotechnol. J. 2020; 18:3494–3506

192. Fan J, Fu A, Zhang L. Progress in molecular docking. Quant. Biol. 2019; 7:83–89

193. De P, Roy K. Computational Modeling of ACE2-Mediated Cell Entry Inhibitors for the Development of Drugs Against Coronaviruses. Silico Model. Drugs Against Coronaviruses 2021; 495–539

194. Chen Y-C. Beware of docking! Trends Pharmacol. Sci. 2015; 36:78–95

195. Haddad Y, Adam V, Heger Z. Ten quick tips for homology modeling of high-resolution protein 3D structures. PLOS Comput. Biol. 2020; 16:e1007449

196. Kairys V, Gilson MK, Fernandes MX. Using protein homology models for structure-based

studies: approaches to model refinement. ScientificWorldJournal. 2006; 6:1542–1554

197. Leman JK, Weitzner BD, Lewis SM, et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. Nat. Methods 2020; 17:665–680

198. Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018; 46:W296–W303

199. Launay G, Simonson T. Homology modelling of protein-protein complexes: a simple method and its possibilities and limitations. BMC Bioinformatics 2008; 9:427

200. Aloy P, Ceulemans H, Stark A, et al. The Relationship Between Sequence and Interaction Divergence in Proteins. J. Mol. Biol. 2003; 332:989–998

201. Quadir F, Roy RS, Soltanikazemi E, et al. DeepComplex: A Web Server of Predicting Protein Complex Structures by Deep Learning Inter-chain Contact Prediction and Distance-Based Modelling . Front. Mol. Biosci. 2021; 8:

202. Susanty M, Rajab TE, Hertadi R. A Review of Protein Structure Prediction using Deep Learning. BIO Web Conf. 2021; 41:

203. Zahiri J, Emamjomeh A, Bagheri S, et al. Protein complex prediction: A survey. Genomics 2020; 112:174–183

204. Gao M, Skolnick J. New benchmark metrics for protein-protein docking methods. Proteins Struct. Funct. Bioinforma. 2011; 79:1623–1634

205. Basu S, Wallner B. DockQ: A quality measure for protein-protein docking models. PLoS One 2016; 11:

206. Janin J, Henrick K, Moult J, et al. CAPRI: a critical assessment of predicted interactions. Proteins Struct. Funct. Bioinforma. 2003; 52:2–9

207. Xue LC, Jordan RA, El-Manzalawy Y, et al. DockRank: ranking docked conformations using partner-specific sequence homology-based protein interface prediction. Proteins 2014; 82:250–267

208. Kufareva I, Abagyan R. Methods of protein structure comparison. Methods Mol. Biol. 2012; 857:231–257

209. Velázquez-Libera JL, Durán-Verdugo F, Valdés-Jiménez A, et al. LigRMSD: a web server for automatic structure matching and RMSD calculations among identical and similar compounds in protein-ligand docking. Bioinformatics 2020; 36:2912–2914

210. Guo F, Zou Q, Yang G, et al. Identifying protein-protein interface via a novel multi-scale local sequence and structural representation. BMC Bioinformatics 2019; 20:1–11

211. Das S, Chakrabarti S. Classification and prediction of protein–protein interaction interface using machine learning algorithm. Sci. Rep. 2021; 11:1–12

212. Pérez-Cano L, Solernou A, Pons C, et al. Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical potentials. Biocomput. 2010 2010; 293–301

213. Pozatti G, Kundrotas P, Elofsson A. Improved protein docking by predicted interface residues. bioRxiv 2021;

214. Jain AN. Scoring functions for protein-ligand docking. Curr. Protein Pept. Sci. 2006;

7:407–420

215. Zemla A. LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res. 2003; 31:3370–3374

216. Jandova Z, Vargiu AV, Bonvin AMJJ. Native or Non-Native Protein–Protein Docking Models? Molecular Dynamics to the Rescue. J. Chem. Theory Comput. 2021; 17:5944–5954

217. Lensink MF, Méndez R, Wodak SJ. Docking and scoring protein complexes: CAPRI 3rd Edition. Proteins Struct. Funct. Bioinforma. 2007; 69:704–718

218. Lensink MF, Velankar S, Wodak SJ. Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition. Proteins Struct. Funct. Bioinforma. 2017; 85:359–377

219. Parisien M, Freed KF, Sosnick TR. On docking, scoring and assessing protein-DNA complexes in a rigid-body framework. PLoS One 2012; 7:

220. Yamagata Y, Kubota M, Sumikawa Y, et al. Contribution of hydrogen bonds to the conformational stability of human lysozyme: calorimetry and X-ray analysis of six tyrosine→ phenylalanine mutants. Biochemistry 1998; 37:9355–9362

221. Lin M, Guo J. New insights into protein–DNA binding specificity from hydrogen bond based comparative study. Nucleic Acids Res. 2019; 47:11103–11113

222. Zhang J, Fan J-S, Li S, et al. Structural basis of DNA binding to human YB-1 cold shock domain regulated by phosphorylation. Nucleic Acids Res. 2020; 48:9361–9371

223. Iftode C, Daniely Y, Borowiec JA. Replication protein A (RPA): the eukaryotic SSB. Crit. Rev. Biochem. Mol. Biol. 1999; 34:141–180

224. Déjardin J, Kingston RE. Purification of proteins associated with specific genomic Loci. Cell 2009; 136:175–186

225. Malik FK, Guo J. Insights into protein–DNA interactions from hydrogen bond energy-based comparative protein–ligand analyses. Proteins Struct. Funct. Bioinforma. 2022; n/a:

226. van Dijk M, Bonvin AMJJ. A protein-DNA docking benchmark. Nucleic Acids Res. 2008; 36:

227. Honorato R V, Roel-Touris J, Bonvin AMJJ. MARTINI-Based Protein-DNA Coarse-Grained HADDOCKing . Front. Mol. Biosci. 2019; 6:102

228. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 2005; 33:2302–2309

229. Fiser A, Šali ABT-M in E. Modeller: Generation and Refinement of Homology-Based Protein Structure Models. Macromol. Crystallogr. Part D 2003; 374:461–491

230. Vangone A, Rodrigues JPGLM, Xue LC, et al. Sense and simplicity in HADDOCK scoring: Lessons from CASP-CAPRI round 1. Proteins 2017; 85:417–423

231. Kanwal F, Okwei E, Troutman J, et al. Understanding Stability and Flexibility Relationships across Undecaprenyl Diphosphate Synthase Superfamily QSFR Within the Cis-IPPS Superfamily Methods. 2005; 101570

232. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins Struct. Funct. Bioinforma. 2004; 57:702–710

233. Zemla A, Venclovas Č, Moult J, et al. Processing and analysis of CASP3 protein structure

predictions. Proteins Struct. Funct. Bioinforma. 1999; 37:22–29

234. Moult J, Pedersen JT, Judson R, et al. A large-scale experiment to assess protein structure prediction methods. Proteins Struct. Funct. Bioinforma. 1995; 23:ii–iv

235. Liu Z, Mao F, Guo J, et al. Quantitative evaluation of protein-DNA interactions using an optimized knowledge-based potential. Nucleic Acids Res. 2005; 33:546–558

236. Takeda T, Corona RI, Guo J. A knowledge-based orientation potential for transcription factor-DNA docking. Bioinformatics 2012; 29:322–330

237. Fernandez-Recio J, Totrov M, Abagyan R. Identification of protein–protein interaction sites from docking energy landscapes. J. Mol. Biol. 2004; 335:843–865

238. Spiliotopoulos D, Kastritis PL, Melquiond ASJ, et al. dMM-PBSA: A New HADDOCK Scoring Function for Protein-Peptide Docking . Front. Mol. Biosci. 2016; 3:

APPENDIX A: SUPPLEMENTARY TABLES

Table S1. PDB ids in the protein homo/heterodimer library (PHDL)

(A) PDB ids of the heterodimers in PHDL

1AY7	1BDJ	1BH9	1BVN	1CXZ	1D0D	1D4T	1DJ7	1DOW	1DS6	1E44	1E96	1EUV	1F3V	1FM2
1FXW	1GK9	1J2J	1JIW	1JQL	1KTP	1M2T	1MK2	1MTP	1MZW	1NME	1NPE	1OO0	1ORY	1PDK
1QGE	1QTX	1R0R	1R8S	1SPP	1SVD	1T0P	1TA3	1TMQ	1U0S	1UGH	1V5I	1V74	1W98	1WMH
1WQJ	1WRD	1WYW	1XG2	1XTG	1Y43	1Z0J	1Z3E	1Z5Y	2A5D	2A9K	2AQ2	2BCG	2BKR	2C1M
2C7M	2D5R	2DVW	2EHB	2F4M	2FCW	2FHZ	2FTX	2GSK	2H7Z	2H9A	2HRK	2HTH	2IE4	2O3B
2OOB	2OZN	2P45	2P8Q	2PA8	2PTT	2QWO	2R25	2UUY	2V3B	2V8S	2V9T	2VPB	2WBW	2WWX
2WY8	2XJY	2XN6	2XPP	2YGG	2Z30	3A2F	3A8G	3AA7	3AON	3AQF	3B0C	3BH7	3BS5	3BY4
3CF4	3CKI	3CNQ	3D3B	3D6N	3DAW	3DGP	3EGV	3F1P	3F75	3FJU	3FMO	3FPU	3FXE	3GJ3
3GOV	3K1R	3KNB	3L51	3LQC	3MCB	3ME0	3MKR	3MXN	3N1M	3NCE	3NVN	3NY7	3O2P	3O3O
3ONA	3OQ3	3P71	3P73	3PLV	3PNL	3PT8	3QDR	3QHY	3QQ8	3SDE	3SHG	3TBI	3TJ5	3UB5
3V61	3VF0	3VRD	3VYR	3VZ9	3WDG	3X37	3YGS	3ZG9	4APX	4BVX	4C2A	4C4P	4CMM	4CRW
4CSR	4DBG	4DHI	4DRI	4F7G	4FBJ	4G01	4G6T	4GN4	4H5S	4H6J	4HST	4HT3	4IU3	4IUM
4J38	4JE3	4JS0	4K12	4K5A	4KAX	4L2I	4LJO	4LLD	4LZX	4M0W	4MRT	4NBX	4NTQ	4NUT
4OB0	4PAS	4PZ5	4QJF	4QLP	4QO1	4R1D	4RCA	4RHZ	4RLJ	4U9H	4UAF	4UHZ	4UN2	4UQZ
4UYQ	4UZZ	4W8P	4WKS	4X86	4X8K	4XAX	4XYD	4YH8	4YI0	4YYP	4ZGM	4ZHY	4ZQU	5AQV
5B64	5B78	5BY8	5BZ0	5C50	5CEC	5CHL	5D6J	5DYN	5EU0	5EUI	5F22	5FOY	5G1X	5GNA
5GXW	5GZT	5H3J	5HE9	5HKQ	5HKY	5I4H	5INB	5IVA	5JCA	5JP1	5JW9	5KYC	5L0R	5L0V
5L3D	5L9Z	5LSI	5LXR	5M0Y	5M2O	5M72	5MAW	5ML9	5MS2	5MU7	5NCW	5NRM	5O33	5OOV
5OW0	5OXZ	5OYL	5SVH	5T51	5T86	5TUU	5TVQ	5TZP	5UIW	5UN7	5UNI	5UUK	5V7P	5VGB
5VKO	5VMO	5WUJ	5WXK	5XA5	5XEC	5XLU	5Y27	5Y38	5YCA	5YR0	5YWR	5Z51	5ZNG	5ZWL
5ZZA	6APP	6AU8	6BN1	6BSC	6BW9	6DLM	6DRE	6DXZ	6EH4	6EM7	6ES1	6F2G	6F6R	6FDK
6FFA	6FUD	6GHO	6GR8	6H02	6H9U	6HM3	6HUL	6IUA	6J4P	6JLE	6JXH	6K06	6K3B	6KGC
6KHS	6KMJ	6KXD	6L4P	6L8G	6LBX	6LKI	6LPH	6M0J	6MBB	6MGN	6MIB	6MS4	6NE2	6NVX
6ODD	6OP8	6OQ7	6OVM	6OX6	6Q00	6QBA	6QUP	6R6M	6RCX	6RM9	6RTW	6S07	6S3F	6S8Q
6SWT	6U3B	6U54	6UUI	6V7M	6VE5	6VJJ	6W0V	6W9S	6WCW	6WG4	6WH1	6WJC	6WUD	6XRU
6XZU	6YX5	6YZ5	6ZXW	7A48	7BQV	7BZK	7C96	7CE4	7CN7	7CQ3	7EDP	7JTU	7MC5	

(B) PDB ids of the homodimers in PHDL

1A8U	1AA7	1AOC	1B2P	1B43	1B5E	1B6Z	1BD9	1BDY	1BO4	1BYF	1C77	1C8U	1CI9	1CKM
1CKU	1CQX	1D0C	1D0Q	1D1G	1D2C	1D2O	1D7F	1D9C	1DEB	1DJ0	1DL5	1DOK	1DPG	1DQE
1DQP	1DQZ	1DU5	1DYS	1E0B	1E19	1E5L	1E7L	1E8U	1EAJ	1EBF	1EE8	1EEJ	1EF0	1EJF
1ELU	1EVX	1EXT	1EYQ	1EZG	1F08	1F0K	1F46	1F5V	1F86	1FBQ	1FBT	1FLM	1FN9	1G29
1G64	1G8E	1G8L	1G8M	1GDE	1GE7	1GQI	1GT1	1GU7	1GVJ	1GXJ	1GYO	1GYX	1H18	1HKQ
1HQS	1HRU	1HYO	1HZ5	1I07	1I0R	1I4S	1I4U	1I6W	1I78	1I7N	1IAZ	1IG0	1IGU	1II7
1IPS	1IQ8	1IRQ	1ITU	1ITV	1IUJ	1IX2	1IX9	1IYB	1IZY	1J0H	1J3M	1J49	1JAD	1JAY
1JK6	1JLY	1JNP	1JR8	1JYA	1JZT	1K3Y	1K4Z	1K66	1KAE	1KDG	1KFI	1KJN	1KKO	1KNQ
1KPT	1KQL	1KQP	1KTJ	1KV0	1KZQ	1LGQ	1LJM	1LN0	1LQ9	1LQA	1M2D	1M4J	1M76	1MBY
1MK4	1MKF	1MKK	1MKZ	1MO9	1MXR	1MY7	1MZG	1N1E	1N2Z	1NBC	1ND4	1NKI	1NNW	1NO5
1NS5	1NSZ	1NU4	1NV7	1NWP	1NWW	1NXM	1NXU	1O5X	1O63	1O6A	1OC2	1OC9	1OCK	1OFZ
1OH0	1OI6	1OKI	1ON2	1OOE	1OR4	1ORD	1ORU	1OSY	1OTK	1OVN	1OX8	1P1C	1P4O	1P5T
1P65	1P6O	1PC6	1PIW	1PIX	1PKV	1PL5	1PPV	1PSR	1Q6O	1QAH	1QC5	1QFH	1QH5	1QI9
1QKS	1QL0	1QLW	1QMH	1QO8	1QQ5	1QSD	1QUP	1QVE	1QVZ	1QXR	1R11	1R12	1R1D	1R61
1R7A	1RDO	1REG	1RFY	1RKT	1RKU	1RW0	1S0P	1S4K	1S9R	1SBY	1SD4	1SEI	1SFN	1SGM
1SJ1	1SMO	1SNN	1SQS	1SU2	1SXH	1SXR	1SZQ	1T06	1T1V	1T3C	1T6S	1T6T	1T7S	1T92
1TBX	1TE2	1TE5	1TEJ	1TJ7	1TLJ	1TU1	1TV8	1TVN	1TXG	1U07	1U5U	1U6R	1U6Z	1UCR
1UDV	1UIX	1UKK	1USC	1USO	1UWK	1UZ3	1V4E	1V58	1V5V	1V5X	1V6P	1V6Z	1V7L	1V7O
1V8H	1VB5	1VC4	1VH5	1VHD	1VHQ	1VHZ	1VJH	1VJQ	1VL7	1VSC	1W5R	1W9C	1WKV	1WLG
1WMX	1WPN	1WR8	1WRA	1WTJ	1WWA	1WWP	1WWZ	1WY2	1WY5	1WZ3	1WZD	1X2I	1X7D	1X9I
1X9Z	1XEQ	1XGS	1XHK	1XJ4	1XNF	1XNG	1XRK	1XRU	1XSV	1XTA	1XVI	1XVS	1Y0H	1Y0U
1Y2O	1Y7M	1Y7R	1Y89	1Y9B	1Y9W	1YDY	1YGA	1YLM	1YLQ	1YLR	1YLX	1YOC	1YRB	1Z41
1Z4E	1Z5B	1ZBO	1ZBR	1ZC6	1ZK8	1ZKI	1ZO2	1ZQ9	1ZUO	1ZV1	1ZVF	1ZZG	2A0U	2A2J
2A4N	2A9U	2AIB	2AKZ	2ANX	2AQ6	2AQP	2AQX	2ARC	2ASK	2AUW	2AXW	2AYT	2B4H	2B6C
2B9D	2BDR	2BJI	2C0D	2C1L	2C2I	2C49	2C5A	2CAR	2CB5	2CC0	2CDU	2CH7	2CMG	2CO3

2CO5	2CTZ	2CU6	2CUN	2CVI	2CWK	2CXN	2D4G	2D73	2D7V	2D8D	2DBS	2DC0	2DC1	2DC3
2DC4	2DCT	2DFJ	2DJ5	2DKJ	2DLB	2DM9	2DOU	2DQL	2DR1	2DS5	2DSJ	2DSK	2DST	2DTC
2DXQ	2E2N	2E2X	2E5F	2E5Y	2E85	2EBE	2ECS	2ECU	2EG4	2EGD	2EGV	2EIX	2EJN	2EK0
2ERB	2ESR	2ETX	2EV1	2F02	2F07	2F1F	2F22	2F2E	2F48	2F5G	2F62	2F96	2F9H	2FA1
2FAE	2FBN	2FCA	2FFG	2FG0	2FHQ	2FIU	2FJR	2FM6	2FNU	2FP1	2FRE	2FSW	2FTR	2FUR
2FXV	2FZF	2FZT	2G3W	2G84	2GA1	2GAN	2GAX	2GBO	2GEC	2GEX	2GFF	2GIY	2GJ3	2GJA
2GKM	2GLZ	2GOM	2GSV	2GU9	2GUD	2GV8	2H1T	2H28	2H2N	2H2R	2H8G	2H98	2HA8	2HBV
2HDW	2HHJ	2HIN	2HIQ	2HO1	2HQ7	2HQY	2HS1	2HXR	2HZG	2I2O	2I5E	2I5G	2I7R	2I8D
2I9U	2IAB	2IB0	2IG3	2IGI	2IPR	2IUT	2IYC	2J05	2J85	2J8W	2J98	2JD3	2JDJ	2JHF
2NLV	2NNH	2NOG	2NQL	2NQT	2NS9	2NTT	2NX9	2NXV	2NYS	2NZ5	2NZ7	2O4C	2O6P	2O7M
2OB3	2OD0	2OD4	2ODA	2OEM	2OFC	2OGB	2OGI	2OHC	2OKU	2OKX	2OM6	2OMD	2OND	2ONF
2OPI	2OPL	2OQB	2OR2	2ORI	2OTA	2OU3	2OU5	2OXL	2OY9	2P08	2P0M	2P12	2P1A	2P1J
2P23	2P2S	2P3Y	2P4P	2P4R	2P62	2P64	2P8U	2P97	2P9H	2PA7	2PEB	2PFW	2PH0	2PIH
2PJS	2PL7	2PO3	2PR8	2PRV	2PS1	2PS5	2PUZ	2Q03	2Q0X	2Q24	2Q3V	2Q5C	2Q6Q	2Q7A
2Q8O	2Q8V	2QBU	2QE8	2QFR	2QGY	2QH9	2QHQ	2QIW	2QJD	2QJF	2QL8	2QLX	2QMX	2QND
2QQZ	2QRR	2QSI	2QU7	2QV0	2QVH	2QXY	2QYC	2QZZ	2R15	2R1F	2R1I	2R5O	2R74	2R8Q
2R8W	2RAS	2RB7	2RBB	2RBG	2RC8	2RCZ	2RDC	2RDE	2RGM	2RK0	2UUZ	2UW1	2V27	2V6K
2V9B	2VD8	2VGX	2VH3	2VKJ	2VOK	2VQ3	2VSW	2VVW	2W1T	2W1V	2W2K	2W31	2W3G	2W43
2W6A	2W8X	2WCR	2WCU	2WD6	2WK4	2WLV	2WMM	2WNS	2WNW	2WU9	2WUF	2WVF	2WW4	2WZV
2X2W	2X65	2X7X	2XDG	2XFN	2XFV	2XGG	2XHF	2XJ3	2XMJ	2XOL	2XR4	2XT2	2XUA	2XW7
2XWL	2XZ8	2XZ9	2Y27	2Y43	2YA8	2YEQ	2YFA	2YIO	2YMA	2YMQ	2YMY	2YR2	2YVE	2YVS
2YW2	2YWL	2YWW	2YXH	2YYB	2YYV	2YYY	2YZI	2Z0U	2Z5E	2Z6R	2Z73	2Z76	2Z8R	2ZB9
2ZDP	2ZEW	2ZGL	2ZGY	2ZO9	2ZOG	2ZVX	2ZVY	2ZW2	2ZW5	2ZX2	2ZYQ	2ZZV	3A1D	3A3D
3A8R	3AAB	3ABH	3AIA	3AJ6	3ALY	3AMI	3AOW	3ATJ	3B0F	3B42	3B4U	3B73	3B8X	3BA3
3BBD	3BBZ	3BCW	3BED	3BGA	3BHQ	3BJE	3BKX	3BL4	3BMZ	3BNK	3BOS	3BRC	3BRU	3BS9
3BWS	3BYP	3BZY	3C1Q	3C3Y	3C8C	3CCD	3CGU	3CJL	3CJP	3CKA	3CNK	3CP7	3CQR	3CRN
3CRY	3CSX	3CT6	3CTP	3CU2	3CW9	3CZ1	3CZ6	3CZZ	3D0F	3D34	3D3I	3D5P	3D7A	3DA5
3DFU	3DMC	3DME	3DN7	3DNF	3DP7	3DS2	3DSB	3DUP	3DUW	3DXO	3E1W	3E2C	3E2D	3E48
3E4V	3E7Q	3E8O	3E96	3EDE	3EDN	3EFY	3EGO	3EIK	3EKG	3ENT	3EO6	3EOF	3EOQ	3EPY
3EQZ	3ER7	3ERX	3ES4	3EUU	3EVI	3EWW	3EY8	3EZH	3F08	3F1L	3F3S	3F5H	3F6G	3F6O
3F6T	3F7E	3F84	3F9S	3F9T	3F9U	3FA5	3FCH	3FD4	3FD7	3FF9	3FGV	3FGY	3FH3	3FHU
3FIL	3FJ4	3FK9	3FKR	3FLD	3FOU	3FPF	3FPK	3FQM	3FR7	3FRQ	3FV6	3FVV	3FX7	3FYB
3G0T	3G16	3G1P	3G3Q	3G3S	3G3Z	3G46	3G4E	3G67	3G8K	3G8R	3GAE	3GAZ	3GB3	3GBY
3GDW	3GE6	3GFA	3GKX	3GLV	3GMG	3GMX	3GO6	3GOC	3GPV	3GR3	3GRD	3GRN	3GRO	3GU3
3GVE	3GW4	3GWK	3GWL	3GWN	3GWO	3GWR	3GZR	3H2B	3H3N	3H6R	3H8L	3H8U	3HA2	3HCN
3HDO	3HEB	3HG9	3HIM	3HIN	3HJ9	3HJG	3HL4	3HLU	3HLX	3HM4	3HMT	3HN0	3HNW	3HO7
3HOA	3HPE	3HPF	3HR0	3HS3	3HU5	3HUP	3HV2	3I0Z	3I2Z	3I3W	3I5Q	3I9F	3IA1	3IAV
3IBS	3IBW	3ICY	3IGR	3IJM	3IKK	3ILW	3IN6	3IPO	3ITF	3IUO	3IUP	3IUW	3IX1	3IX3
3IX7	3JSL	3JU7	3JX9	3JXO	3K0Z	3K2N	3K67	3K86	3K8R	3K9U	3K9V	3KBY	3KD4	3KD6
3KE7	3KEA	3KF3	3KGZ	3KHF	3KIZ	3KKB	3KKZ	3KMA	3KPH	3KUV	3KUZ	3KWR	3KWS	3KZP
3KZT	3L0Q	3L32	3L46	3L5Z	3L6I	3L6U	3LAG	3LAS	3LF5	3LF6	3LFI	3LGD	3LHN	3LHR
3LIA	3LID	3LJD	3LM2	3LMB	3LQ6	3LQS	3LR2	3LRT	3LS9	3LV4	3LVC	3LYD	3LYN	3LYY
3LZX	3LZZ	3M33	3M8J	3MAB	3MAD	3MBK	3MC1	3MCW	3MCZ	3MEX	3MGD	3MGG	3MGJ	3MGK
3MIL	3MIZ	3MJQ	3MMH	3MOZ	3MQM	3MQQ	3MTR	3MUJ	3MUQ	3MUX	3MVE	3MVG	3MWJ	3MZ2
3N08	3N10	3N1E	3N8B	3NAU	3NAW	3NDO	3NEK	3NI0	3NI6	3NJ2	3NO7	3NOI	3NPF	3NPI
3NPP	3NQB	3NQW	3NRL	3NS6	3NTL	3NTV	3NUF	3NVA	3NX3	3NYD	3O0L	3O4W	3O5Y	3O6V
3O7O	3OAJ	3OCP	3OFG	3OHE	3OMT	3OMY	3ONX	3OOO	3OPC	3OQ2	3OQP	3OT2	3OTN	3OVP
3OY2	3OZI	3OZY	3P1X	3P2C	3P6B	3P6K	3P7J	3P8T	3P9V	3PA8	3PC7	3PDY	3PET	3PFO
3PIJ	3PJT	3PJV	3PJY	3PMC	3PMR	3PN3	3PPB	3PPL	3PPM	3PSM	3PU9	3PUB	3PUH	3PX2
3Q18	3Q20	3Q31	3Q4N	3QBM	3QGU	3QHA	3QKC	3QTA	3QWU	3QYF	3R3P	3R41	3R5G	3R89
3RA5	3RAU	3RBY	3RKC	3ROT	3RQ9	3RQB	3RRI	3RRS	3S06	3S18	3S84	3SBU	3SG8	3SK2
3SLZ	3SON	3SY6	3T2Z	3T6S	3T7Y	3T8K	3TAK	3TB6	3TC9	3TDQ	3TE8	3TFJ	3THF	3TJ8
3TP9	3TRI	3TY2	3TYY	3U1Y	3U4Z	3U6G	3U7R	3U96	3UB6	3UBU	3UEJ	3UEP	3UF6	3UFE
3UHA	3UMO	3UMZ	3UPL	3URY	3USS	3UT4	3UUN	3UV0	3UV1	3UX3	3V1E	3V4K	3V4M	3V67
3V6G	3VAY	3VB8	3VCC	3VEJ	3VK5	3VM9	3VRC	3VTX	3VW9	3VZX	3W08	3W0E	3W1O	3W36
3W77	3WAE	3WGT	3WHA	3WJE	3WRB	3WSC	3WV8	3WX7	3X3Y	3ZFI	3ZIG	3ZIT	3ZJL	3ZRP
3ZRX	3ZTB	3ZTH	3ZX4	3ZYL	3ZYY	4A7U	4AB5	4AE4	4AG0	4AG7	4AML	4AUU	4AVR	4AXO
4AYN	4B0N	4B0Z	4B54	4BE3	4BE9	4BF5	4BG7	4BG8	4BI3	4BK0	4BLG	4BND	4BOL	4BRC
4BWO	4BWV	4BX2	4C0R	4C86	4CHI	4CI8	4CJN	4CL3	4COB	4CWC	4D3D	4D3Q	4DCZ	4DJN
4DMG	4DNN	4DNX	4DO2	4DT5	4DZZ	4E0A	4E0U	4EBG	4EF0	4EGU	4EHS	4EHU	4EI0	4EIB
4EIR	4EJR	4EP4	4EPU	4EQ7	4EQQ	4EQS	4ESW	4ETK	4EU9	4EVX	4EW5	4EZG	4FBM	4FDI
4FKB	4FKZ	4FRY	4FU3	4FVF	4FYP	4FZL	4G06	4G3V	4G5A	4GEK	4GHO	4GI2	4GIT	4GKM
4GOF	4GP7	4GR6	4GXO	4GYT	4H5B	4H7L	4H8A	4HAH	4HBE	4HBQ	4HCE	4HCF	4HEH	4HEI
4HEQ	4HFQ	4HFS	4HHV	4HI7	4HIA	4HL2	4HMS	4HU7	4HW5	4HWV	4HYL	4I1Q	4I4K	4I4O
4I6R	4I6Y	4IBG	4IC3	4ICS	4ID0	4IGU	4IHU	4IJ5	4IJ7	4IJZ	4IKB	4IP5	4IQD	4IQI
4ITB	4IV9	4IX3	4IXN	4IY4	4IYJ	4J0N	4J3Y	4J42	4J5R	4J6C	4J7R	4J8C	4J8E	4J8Z
4JAW	4JEM	4JG9	4JGP	4JLE	4JN9	4JOQ	4JTM	4JXR	4JYS	4K0U	4K26	4K28	4K6H	4KEM
4KR5	4KTP	4KTW	4KV2	4L1J	4L3K	4L3R	4L57	4L7A	4L9C	4LAN	4LIR	4LJ3	4LJI	4LJL
4LM4	4LMY	4LS9	4LSM	4LTB	4LXQ	4M0Q	4M0S	4M73	4M7Y	4MAC	4MAE	4MAK	4MAM	4MDU
4MEB	4MGE	4MIS	4MJD	4MN7	4MPM	4MT8	4MUV	4MYP	4N04	4N06	4N0R	4N0V	4N6J	4N7W
4N8O	4NAD	4NC7	4NDS	4NEX	4NK2	4NLH	4NOG	4NPR	4NQ8	4NQF	4NRN	4NSV	4NTC	4NU3

4O6I	4O6Y	4O7J	4O9K	4OH9	4OHJ	4OK4	4OKE	4OKI	4OM8	4OO4	4OPM	4OQQ	4OS3	4OTN
4OYU	4OZ0	4P33	4P5N	4P7C	4P93	4P94	4PAG	4PE0	4PHJ	4PIC	4PRS	4PUH	4PVC	4PXE
4PYQ	4PZK	4Q04	4Q1V	4Q25	4Q51	4Q69	4Q6Z	4Q7O	4Q9A	4Q9V	4QBN	4QE0	4QGB	4QGE
4QGX	4QHJ	4QI3	4QJY	4QNC	4QR8	4QUS	4R16	4R27	4R3N	4R60	4R8D	4R8O	4R8Z	4R9X
4RAY	4RBR	4RDZ	4RE5	4RGB	4RGD	4RGP	4RLZ	4RO3	4RP3	4RPT	4RRQ	4RSW	4RT5	4RUN
4RVS	4RZ3	4RZB	4S1H	4S23	4S26	4S3I	4S3P	4TLJ	4TMT	4TN5	4TPV	4TQJ	4TR6	4TRH
4TRT	4TSD	4TT0	4TTY	4TVI	4TWL	4TX5	4U13	4U5G	4U9N	4UAB	4UAI	4UC2	4UEJ	4UG1
4UIQ	4UNU	4UOP	4UP3	4UR6	4USK	4UTU	4UU3	4UUL	4UX7	4UXU	4UZ8	4V15	4V17	4V29
4W7Y	4W9R	4WBP	4WF5	4WH5	4WJT	4WPM	4WWF	4WX0	4WZN	4X08	4X3L	4X51	4X6X	4X8Y
4XFW	4XIN	4XO6	4XQ4	4XQC	4XVV	4XWT	4XZZ	4Y1R	4Y7D	4YEA	4YEP	4YMG	4YNX	4YPO
4YSL	4YT2	4YTD	4YTO	4YX1	4YY5	4YZG	4YZZ	4Z24	4Z27	4Z39	4Z4A	4ZBD	4ZBW	4ZCE
4ZDS	4ZFV	4ZKY	4ZO2	4ZSI	4ZUR	4ZV5	4ZVA	4ZVC	5A3V	5A48	5A9D	5ACS	5AIF	5AL7
5AMT	5AQ0	5AVN	5AWI	5AXG	5AYV	5AZW	5B08	5B0H	5B0P	5B1Q	5B4N	5B5I	5B7G	5BIR
5BJX	5BNC	5BR4	5BTU	5BU6	5BWI	5C04	5C1F	5C40	5C5Z	5C7Q	5C8Z	5CES	5CL2	5CQG
5CR4	5CRB	5CRH	5CUO	5CX8	5CXO	5CYJ	5D1P	5D1R	5D1V	5D3A	5DCL	5DY1	5E2C	5ECC
5EDX	5EIU	5EK5	5EQ2	5ER9	5EUV	5F29	5F2K	5F46	5F5N	5F6R	5FAV	5FCN	5FFP	5FFQ
5FI3	5FID	5FIS	5FLH	5FVJ	5FXD	5FZP	5G4I	5GGY	5GPK	5GSM	5GT5	5GUK	5GVY	5GX8
5GXE	5GXX	5GY7	5H1N	5H34	5H3Z	5H78	5HB6	5HCB	5HDM	5HEE	5HHJ	5HI8	5HIF	5HJL
5HOP	5HRA	5HS7	5HTL	5HWV	5HX0	5I0Y	5I5M	5I90	5I96	5IDB	5IN1	5IOJ	5IPY	5IRB
5IT3	5ITJ	5IW9	5IXV	5IZ3	5J0A	5J41	5J4I	5J7M	5J90	5JAZ	5JBR	5JE6	5JEL	5JHX
5JIP	5JKJ	5JNP	5JNU	5JSI	5JTD	5JWC	5K3X	5K4W	5KAY	5KEF	5KHD	5KO4	5KX4	5L0L
5L73	5LLJ	5LTL	5LVS	5LWK	5LZK	5M7C	5M97	5M99	5MOZ	5MQ8	5MUY	5MWX	5N6X	5NCK
5NCR	5NEG	5NL6	5NLZ	5NO5	5NZO	5O10	5O2Z	5OI7	5OLY	5OO7	5ORG	5OVY	5SY4	5T3E
5T3U	5TD6	5TFP	5TJJ	5TO5	5TTA	5TXC	5U35	5U4H	5U5N	5U85	5UCT	5UE1	5UE7	5UEJ
5UF5	5UFN	5UH7	5UI9	5UJD	5UKV	5UQS	5UUO	5UZX	5V01	5V4A	5V4P	5V4R	5V5U	5V6I
5VAZ	5VDN	5VHT	5VJ4	5VM2	5VSJ	5VT2	5VX1	5W4Z	5W8Q	5WEC	5WFX	5WHX	5WI2	5WPP
5WUT	5WWD	5X03	5X56	5X9I	5XAQ	5XGT	5XKT	5XNA	5XNE	5XOM	5XP0	5XPV	5XSP	5XUN
5XVJ	5XXA	5Y78	5Y8L	5Y9Q	5Y9Z	5YA6	5YAD	5YAT	5YDD	5YET	5YGE	5YGH	5YHR	5YJC
5YKR	5YKZ	5YN4	5YNX	5YRH	5YZ1	5Z11	5Z16	5Z28	5Z2G	5Z2H	5Z2V	5Z49	5Z50	5Z8O
5ZFK	5ZI1	5ZI2	5ZKT	5ZQJ	5ZUM	5ZVV	5ZXN	6A51	6A55	6A5F	6A6F	6A71	6A80	6AE9
6AEF	6AEP	6ALL	6AMG	6AQE	6AR4	6AT3	6AWL	6AWR	6B7C	6B9F	6BHY	6BIE	6BND	6BSU
6BSY	6C0G	6C3C	6C5B	6C6N	6C8R	6CDB	6CKK	6CMK	6COF	6CPB	6CPD	6CQP	6CS9	6CW0
6CWW	6D2W	6D3V	6D41	6DAO	6DB1	6DBP	6DEB	6DGK	6DGM	6DJC	6DKK	6DQP	6DT3	6DVR
6E28	6EDQ	6EID	6EJT	6EL2	6ENI	6EP6	6ES9	6EW7	6EWM	6EY5	6F1J	6F43	6F5C	6FDC
6FF2	6FHG	6FIY	6FP5	6FU3	6G6U	6G96	6GDJ	6GF6	6GFB	6GHU	6GU1	6GYG	6GZA	6H1W
6H31	6H59	6H6O	6H86	6H8F	6HAT	6HAZ	6HBV	6HIU	6HJO	6HK8	6HNM	6HPQ	6HQ2	6HQZ
6HTJ	6HZY	6I1A	6I5B	6I6S	6IAU	6IFQ	6ILS	6IME	6IOW	6IPT	6IRP	6J1O	6J25	6J3E
6J4K	6J5X	6J66	6J6A	6J6L	6J8L	6J94	6JDH	6JHV	6JIE	6JNJ	6JQW	6JSX	6K2F	6K2Y
6K62	6K7C	6K8V	6KEW	6KFM	6KGJ	6KHL	6KI2	6KLK	6KNL	6L2U	6L3Q	6L5H	6L6G	6L85
6LAC	6LCQ	6LEB	6LF1	6LGI	6LH6	6LIY	6LPN	6LZH	6M2O	6M31	6M4B	6M9G	6MB8	6MRV
6MTW	6MX1	6MXV	6N7O	6N91	6N9Q	6NAL	6NDI	6NIM	6NJC	6NK3	6NL2	6NNH	6NNR	6NNW
6NP6	6NQY	6NRX	6O0B	6O14	6O5K	6O6Y	6O8N	6OH8	6OIB	6OJF	6OMP	6ON4	6OVP	6OZU
6P1E	6P2I	6P58	6P73	6PCE	6PNR	6PT4	6PT8	6Q2C	6QJ6	6QLA	6QSI	6QUW	6QWO	6R5J
6R6U	6RCH	6RIV	6RJB	6RK0	6RK1	6RS4	6RWD	6RYK	6S2R	6S33	6S6F	6S7X	6S95	6SAN
6SCB	6SCQ	6SEK	6SF4	6SFH	6SI6	6SIZ	6SJ8	6SRB	6SSG	6SU3	6SW4	6T7O	6TCB	6TEK
6TJ8	6TJR	6TL7	6TVV	6TY0	6TY2	6TYK	6U2U	6U60	6UBL	6UBO	6UD6	6UH8	6UN8	6URE
6USS	6UXU	6V1B	6V3Z	6V42	6VD8	6VH6	6VJC	6VJU	6VPE	6VTV	6VUD	6VZ0	6W40	6W6X
6WE8	6WJA	6WN2	6WU7	6WXW	6XB6	6XNO	6XPH	6Y04	6Y1W	6Y1Y	6Y7F	6YF6	6YIZ	6YJ9
6YKB	6Z68	6ZA0	6ZII	6ZK8	6ZMB	6ZN7	6ZT4	7A1F	7A5C	7AED	7AG6	7AO3	7APP	7ASV
7B5J	7B67	7BB3	7BIO	7BJN	7BM8	7BR1	7BRA	7BU2	7C02	7C23	7C38	7C4A	7C5Y	7C64
7C8G	7C8P	7CBI	7CCB	7CDV	7CIK	7CJ3	7CJ7	7CKH	7CMA	7CSV	7CWQ	7EV1	7JJV	7JKV
7JW2	7KB9	7KL8	7KPZ	7KQA	7KSB	7KWD	7LZG	7MBK	7NBI	7NET	7NUU	7O39	12AS	

Table S2. p-values of chi-square tests between hydrogen bond types from FIRST with an energy cutoff of -0.1 kcal/mol.

	Intrachain	**Interface**	**Interface/Intrachain**	
Dataset1/Dataset2	**p-values**	**p-values**	**Dataset**	**p-values**
PPnrall, PDnrall	0.858	0.0047	PDnrall	6.941e-10

PTnrall, PDnrall	0.845	0.0043	PPnrall	3.831e-10
PTnrall, PPnrall	0.963	0.0137	PTnrall	3.369e-06

Table S3. p-values of chi squared tests comparing proportions of different types of HB energy categories based on Table 2.

Intrachain		**Interface**	
Dataset1/Dataset2	**p-values**	**Dataset1/Dataset2**	**p-values**
BB-BB: PDnrall, PPnrall	0.924	*BB-BB: PDnrall, PPnrall	2.2e-16
BB-BB: PDnrall, PTnrall	0.986	*BB-BB: PDnrall, PTnrall	2.2e-16
BB-BB: PPnrall, PTnrall	0.991	*BB-BB: PPnrall, PTnrall	0.703
SC-SC: PDnrall, PPnrall	0.948	SC-SC: PDnrall, PPnrall	4.031e-06
SC-SC: PDnrall, PTnrall	0.989	SC-SC: PDnrall, PTnrall	1.036e-10
SC-SC: PPnrall, PTnrall	0.994	SC-SC: PPnrall, PTnrall	0.399
Mixed: PDnrall, PPnrall	0.741	Mixed: PDnrall, PPnrall	2.2e-16
Mixed: PDnrall, PTnrall	0.987	Mixed: PDnrall, PTnrall	2.2e-16
Mixed: PPnrall, PTnrall	0.839	Mixed: PPnrall, PTnrall	0.816

* Since the numbers of HBs of the interface BB-BB types for category II and III are small, the chi-square statistical analysis was performed by combining the numbers in category II and III.

Table S4. HB energy (HBE) categories based on different energy ranges

Category	**HBE range (kcal/mol)**
I	$-0.7 \leq HBE < -0.1$
II	$-1.3 \leq HBE < -0.7$
III	$-2.0 \leq HBE < -1.3$
IV	$HBE < -2.0$

Table S5. p-values of chi-square tests between hydrogen bond energy categories (based on the discretization in Table S4) at interface and within intrachain.

Dataset1/Dataset2	**p-values (intrachain)**	**p-values (interface)**	**Dataset**	**p-values (interface/intrachain)**
PPnrall, PDnrall	0.959	0.007	PDnrall	0.005

PTnrall, PDnrall	0.999	0.009	PPnrall	0.944
PTnrall, PPnrall	0.980	0.999	PTnrall	0.99

Table S6. p-values of chi squared tests comparing proportions of different types of HB energy categories based on the discretization in Table S4.

Intrachain		**Interface**	
Dataset1/Dataset2	**p-values**	**Dataset1/Dataset2**	**p-values**
BB-BB: PDnrall, PPnrall	0.859	*BB-BB: PDnrall, PPnrall	2.2e-16
BB-BB: PDnrall, PTnrall	0.986	BB-BB: PDnrall, PTnrall	2.2e-16
BB-BB: PPnrall, PTnrall	0.973	*BB-BB: PPnrall, PTnrall	0.726
SC-SC: PDnrall, PPnrall	0.926	SC-SC: PDnrall, PPnrall	6.935e-15
SC-SC: PDnrall, PTnrall	0.948	SC-SC: PDnrall, PTnrall	4.83e-12
SC-SC: PPnrall, PTnrall	0.946	SC-SC: PPnrall, PTnrall	0.968
Mixed: PDnrall, PPnrall	0.926	Mixed: PDnrall, PPnrall	9.802e-05
Mixed: PDnrall, PTnrall	0.948	Mixed: PDnrall, PTnrall	0.009
Mixed: PPnrall, PTnrall	0.946	Mixed: PPnrall, PTnrall	0.756

* Since the numbers of HBs of the interface BB-BB types for category II and III are small, the chi-square statistical analysis was performed by combining the numbers in category II and III.

Supplementary Table S7. Mean and median of interface surface area and backbone atoms at interface

Statistic\Dataset	**PDnrall**	**PPnrall**	**RDPP**
mean(interface surface area)	1404.46	2926.04	
median(interface surface area)	798.87	1703.67	
mean(number of backbone atoms at interface)	35.021		722.139
median(number of backbone atoms at interface)	32		560

APPENDIX B: SUPPLEMENTARY FIGURES

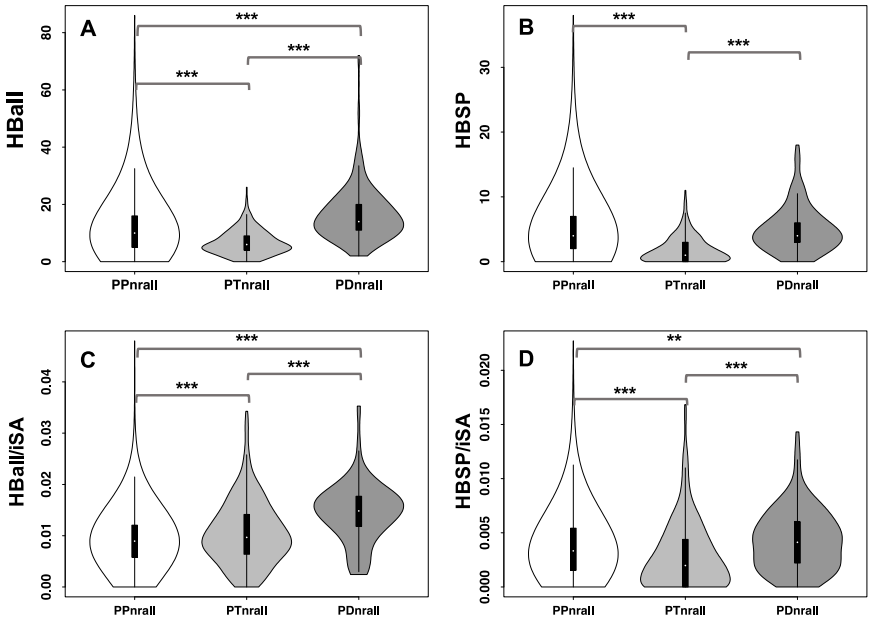

Figure S1. Comparison of interfacial hydrogen bonds based on HBPLUS with default parameters: (A) the number of total hydrogen bonds (HBall); (B) the number of SC-SC or SC-base hydrogen bonds (HBSP); (C) the ratio HBall to interfacial surface area (iSA); and (D) the ratio of HBSP to iSA.

*** = p-value $\leq$ 0.001; ** = p-value $\leq$ 0.01

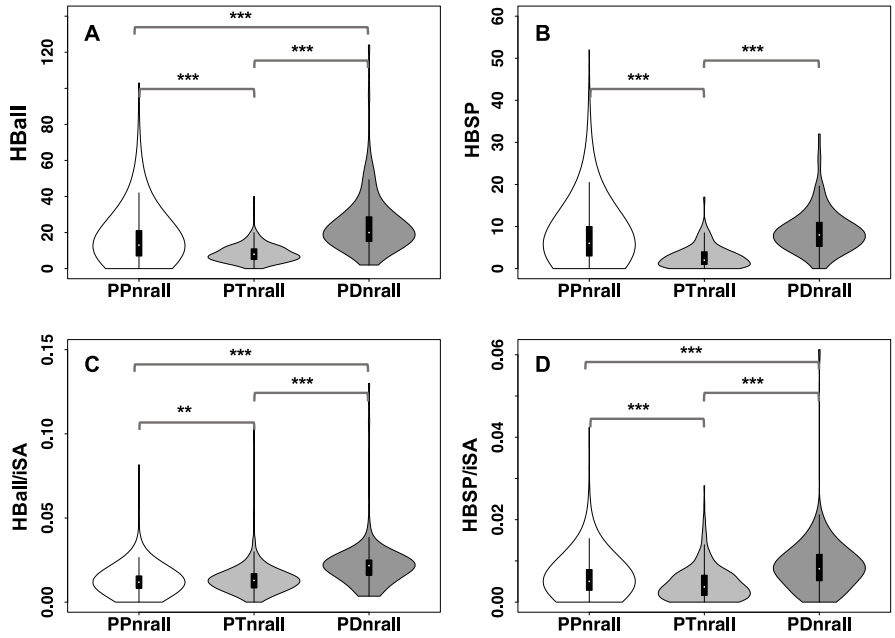

Figure S2. Comparison of interfacial hydrogen bonds based on FIRST with an energy cutoff of -0.1 kcal/mol: (A) the number of total hydrogen bonds (HBall); (B) the number of SC-SC or SC-Base hydrogen bonds (HBSP); (C) the ratio of HBall to interfacial surface area (iSA); and (D) the ratio of HBSP to iSA.

*** = p-value ≤ 0.001, ** = p-value ≤ 0.01

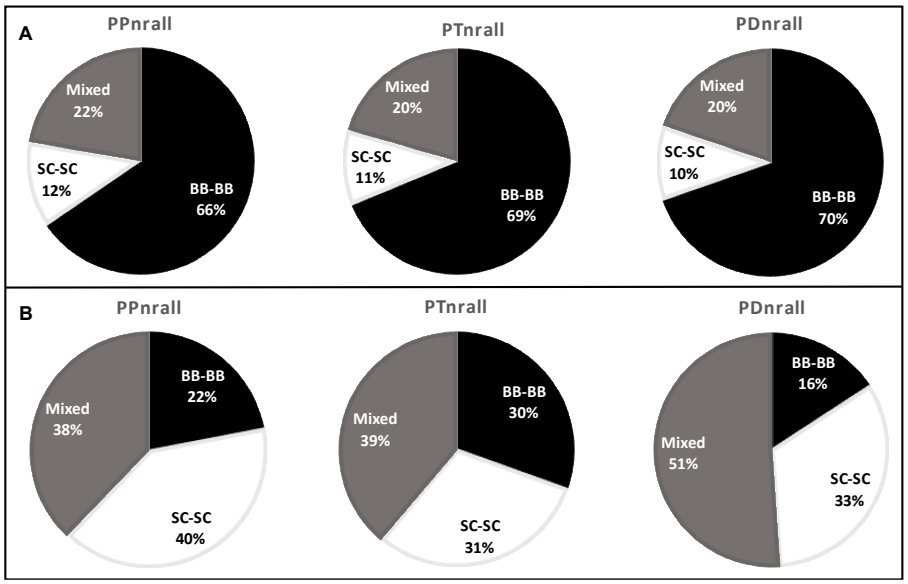

Figure S3. Comparison of the distributions of hydrogen bond types with HBPLUS: backbone-backbone (BB-BB), sidechain-sidechain (SC-SC) and mixed (BB-SC and SC-BB) at (A) intrachain and (B) interface of PP, PT and PD complexes. (See p-values in Table 3)

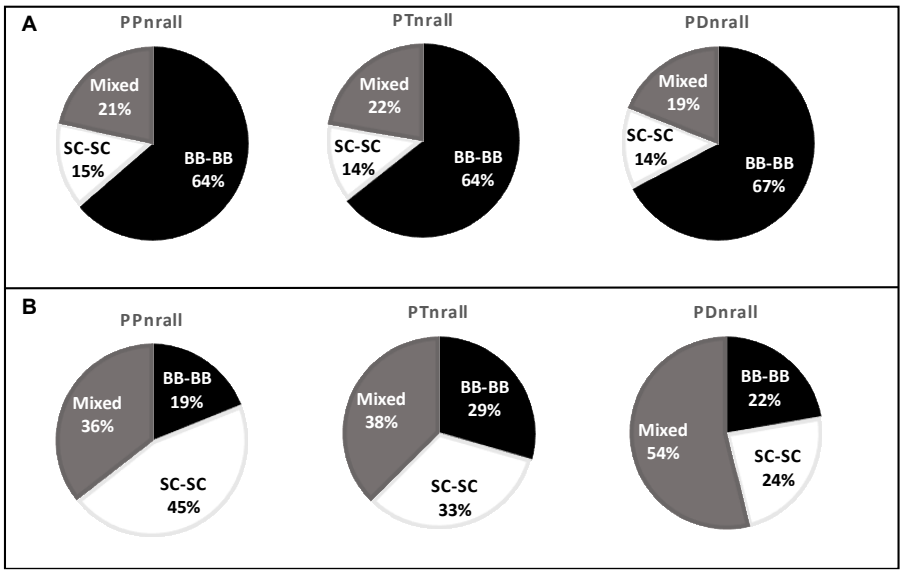

Figure S4. Comparisons of the distribution of different types of hydrogen bonds, backbone-backbone (BB-BB), sidechain-sidechain (SC-SC) and Mixed (BB-SC and SC-BB) for (A) intrachain within proteins and (B) at interface of PP, PT and PD complexes. The hydrogen bonds are annotated from the FIRST program with an energy cutoff of -0.1 kcal/mol. (See p-values in Table S2)

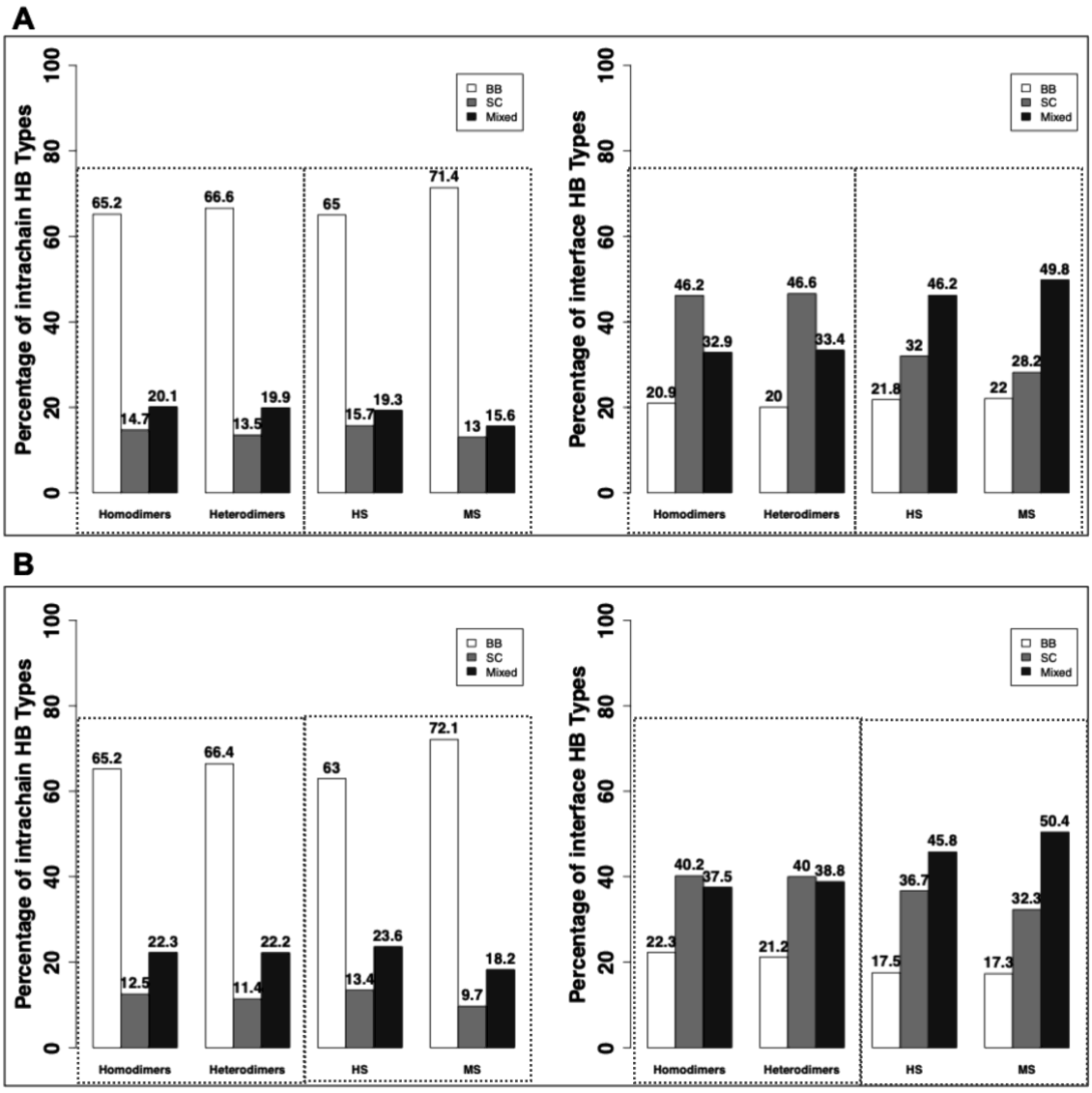

Figure S5. Comparison of the percentages of HB types: backbone-backbone (BB-BB), sidechain-sidechain (SC-SC) and mixed (BB-SC and SC-BB) in intrachain and interface of homodimers, heterodimers, highly specific and multi-specific protein-DNA complexes. (A)The hydrogen bonds are annotated by FIRST with an energy cutoff of -0.6 kcal/mol. (B) The hydrogen bonds are annotated by HBPLUS.

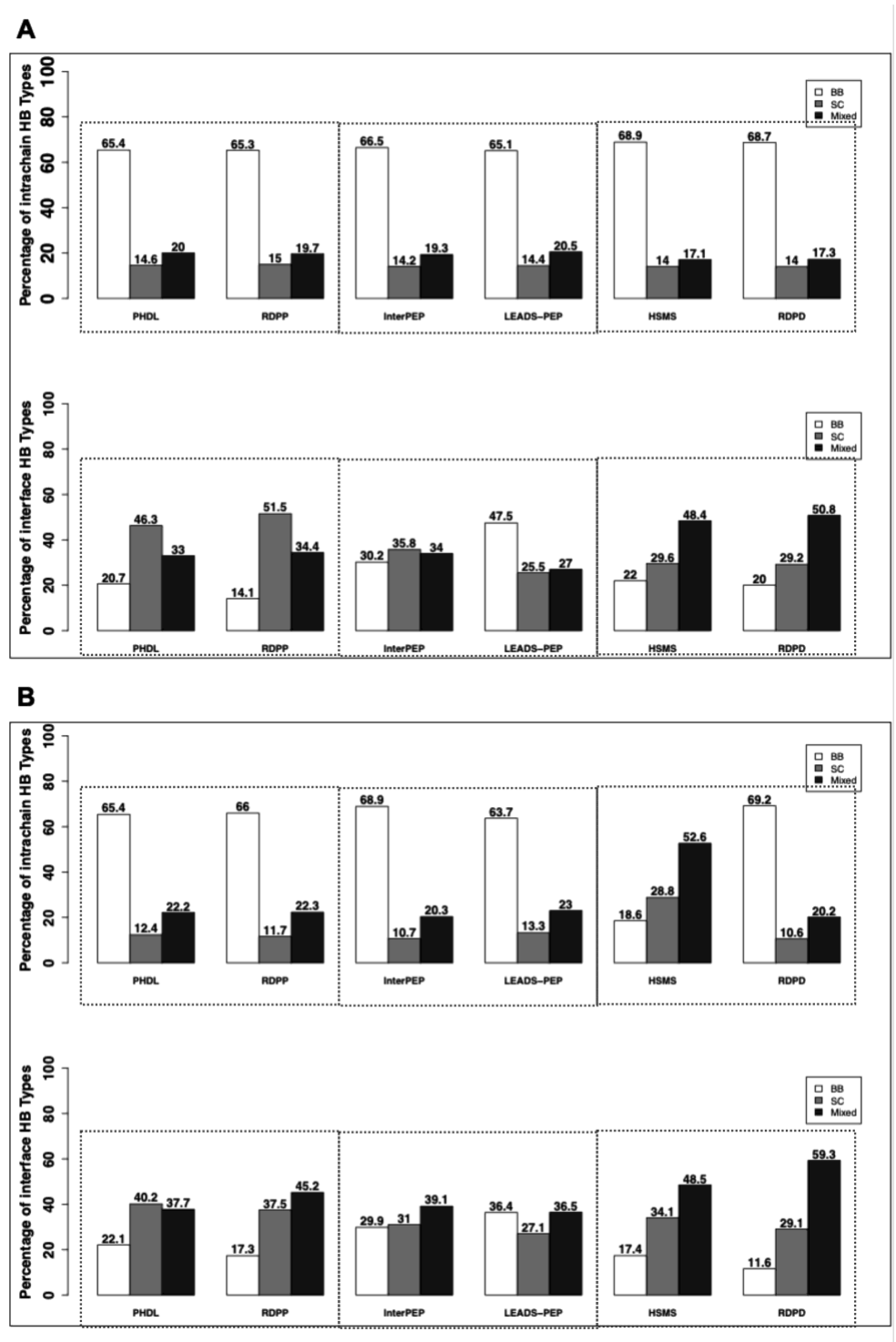

Figure S6. Comparison of the percentages of HB types: backbone-backbone (BB-BB), sidechain-sidechain (SC-SC) and mixed (BB-SC and SC-BB) for intrachain and interface of individual PP, PT and PD complexes. (A) The hydrogen bonds are annotated by FIRST with an energy cutoff of -0.6 kcal/mol. (B) The hydrogen bonds are annotated by HBPLUS.

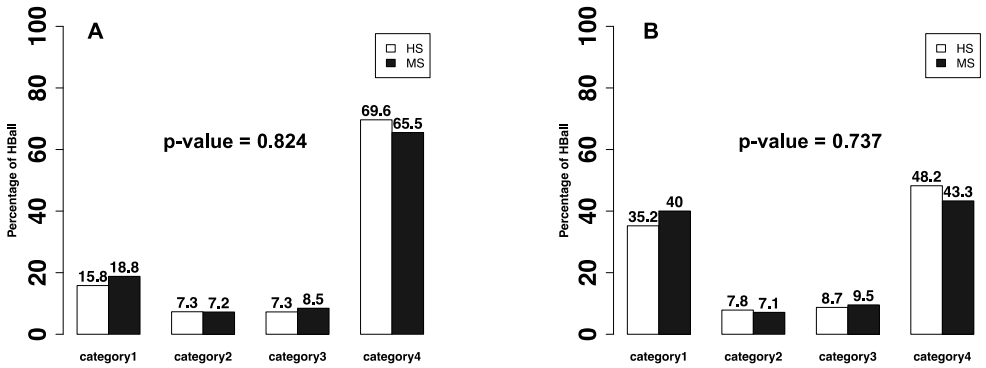

Figure S7. Comparison of the categories of hydrogen bond energy (based on Table 2) between HS and MS complexes. (A) intrachain; (B) interface.

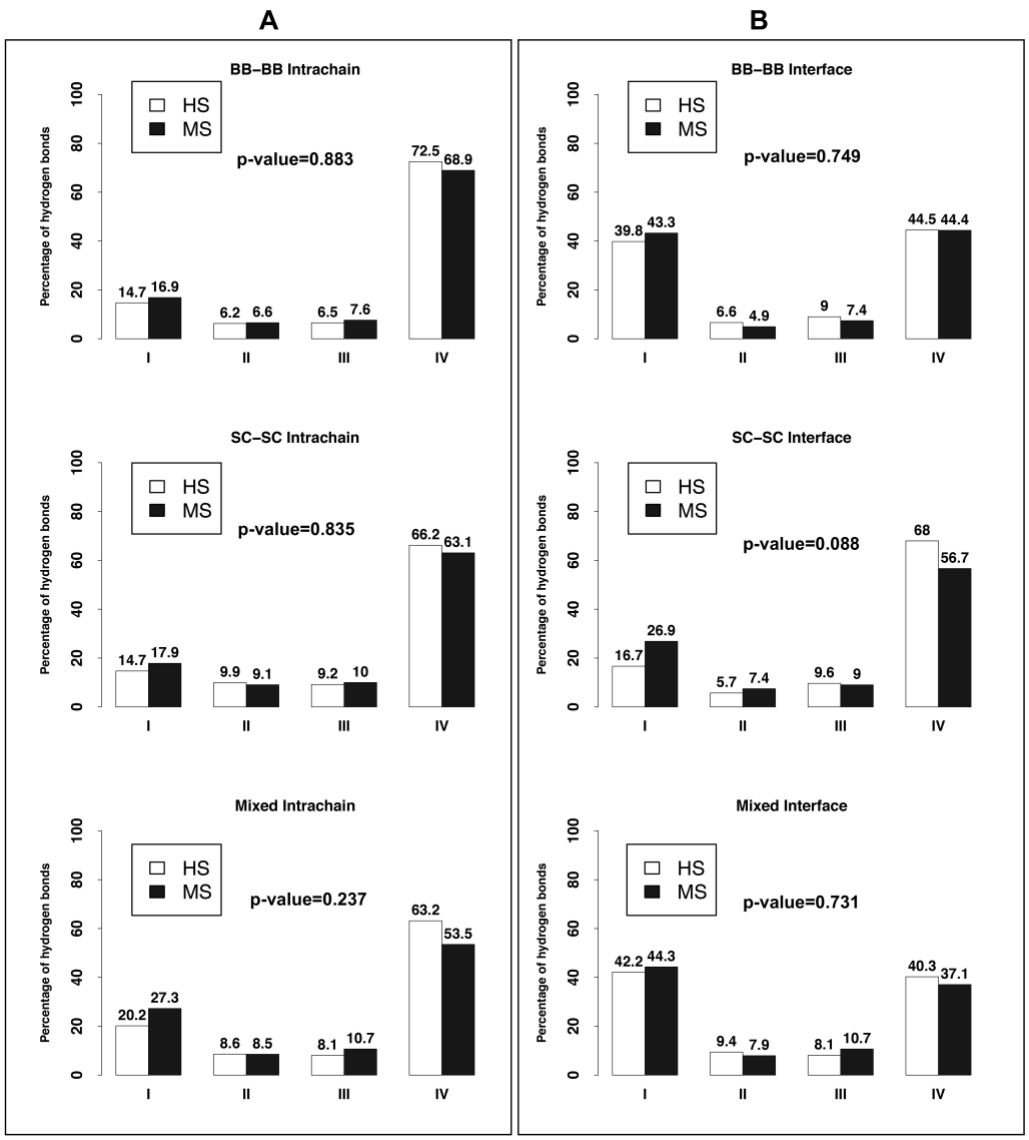

Figure S8. Comparison of hydrogen bond energy categories (based on Table 2) in different hydrogen bond types between HS and MS complexes. (A) intrachain; (B) interface.

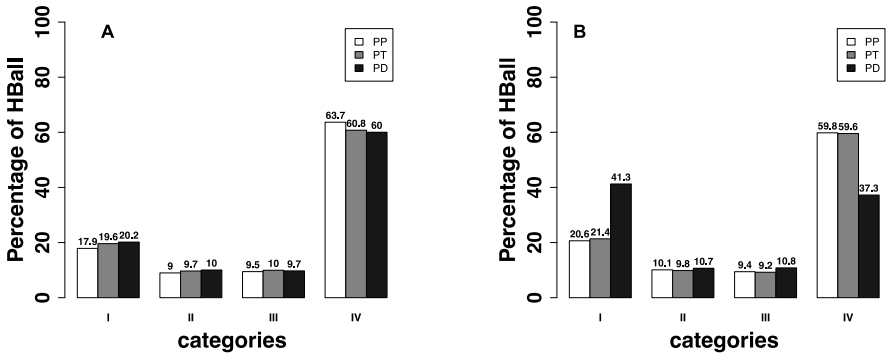

Figure S9. Comparisons of the distributions of hydrogen bond energy based on the discretization in Table S4 for (A) intrachain and (B) at interface. (See Table S5 for p-values).

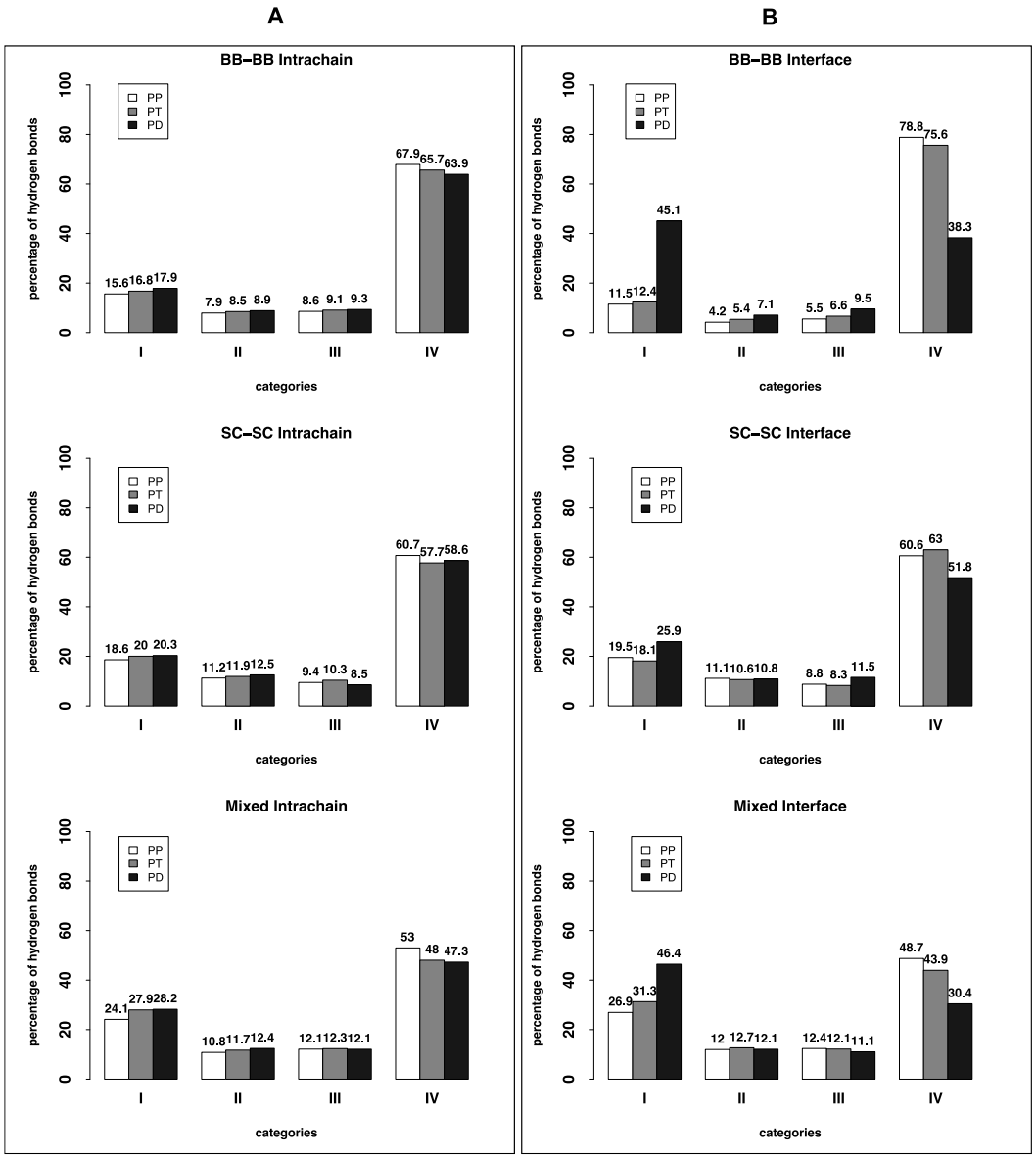

Figure S10. Comparison of (A) intrachain hydrogen bond energy and (B*)* interface hydrogen bond energy (based on the discretization in Table S4) in different hydrogen bond types (See Table S6 for p-values).

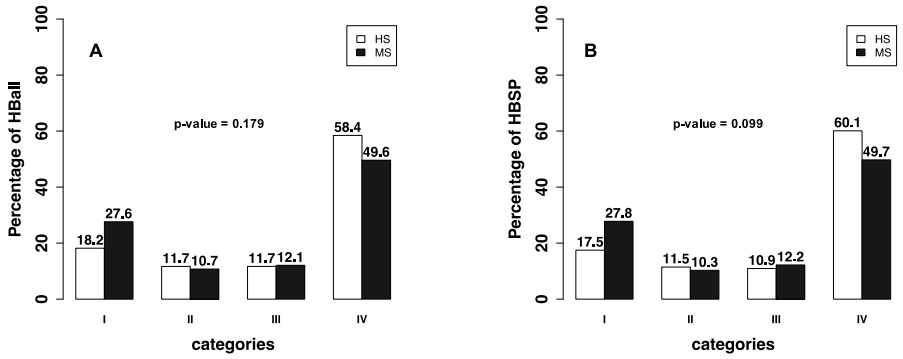

Figure S11. Comparison of major groove for (A) HBall and (B) HBSP energy distributions (based on the discretization in Table S4) between HS and MS complexes.

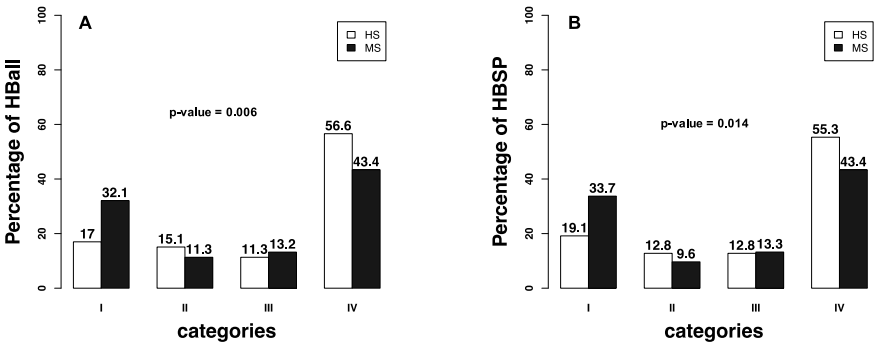

Figure S12. Comparison of minor groove for (A) HBall and (B) HBSP energy distributions (based on the discretization in Table S4) between HS and MS complexes.

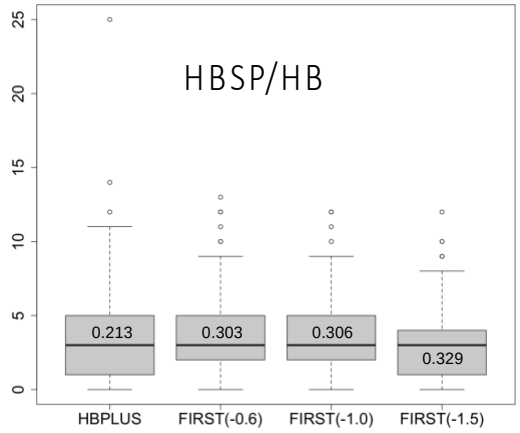

Figure S13. Comparison of side chain-base hydrogen bonds to all hydrogen bonds at the interface of protein-DNA complexes from HBPLUS and FIRST with three different energy thresholds.