STRUCTURAL ANALYSIS OF PROTEIN-DNA BINDING SPECIFICITY AND ITS APPLICATION TO PROTEIN-DNA DOCKING ASSESSMENT

by

Rosario Ivetth Corona de la Fuente

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

2016

Approved by:

_____

Dr. Jun-tao Guo

_____

Dr. Xiuxia Du

_____

Dr. Dennis R. Livesay

_____

Dr. Jennifer W. Weller

_____

Dr. Richard Souvenir

ABSTRACT

ROSARIO IVETTH CORONA DE LA FUENTE. Structural analysis of protein-DNA binding specificity and its application to protein-DNA docking assessment. (Under the direction of DR. JUN-TAO GUO)

DNA-binding proteins are involved in essential biological processes including gene expression, DNA packaging and DNA repair. They bind to DNA target sequences with different degrees of binding specificity, ranging from highly specific to non-specific. Alterations of DNA-binding specificity, due to either genetic variation or somatic mutations, can lead to various diseases. In this study, a comparative analysis of protein-DNA complex structures was carried out to investigate the structural features for binding specificity. The analysis was done using three curated datasets of protein-DNA complexes with different degrees of DNA-binding specificity: highly specific (HS), multi-specific (MS), and non-specific (NS). We found a clear trend of structural features among these three classes, including amino acid binding propensities, simple and complex hydrogen bonds, major groove and base contacts, DNA shape, and conformational changes upon DNA-binding. These structural features were then applied to assess the accuracy of TF-DNA docking predictions. A binary classifier for evaluating the prediction accuracy was developed using a training dataset and the structural features as well as three binding affinity scores. The results on a test dataset show much improved prediction accuracy over previous methods.

DEDICATION

To my husband, my parents and all my family.

ACKNOWLEDGMENTS

I would like to acknowledge the amazing work of my advisor, Dr. Jun-tao Guo, who with his guidance helped me understand the qualities I needed to develop for the successful completion of this interdisciplinary doctoral program. I would also like to thank the members of the Guo lab for the discussions we had throughout the years.

I would like to thank all the members of my Ph.D. dissertation committee for their time and suggestions. Special thanks to Dr. Xiuxia Du who is a clear example of excellence in teaching, combining abstract and theoretical concepts with concrete applications. Also, thanks to Dr. Dennis R. Livesay, who gave me the possibility to explore new areas of research.

I would like to extend my appreciation to the University of North Carolina at Charlotte, Mexico's National Council of Science and Technology (CONACYT) and my advisor for providing financial support. Many thanks to the administrative staff of the International Students and Scholars Office and the Bioinformatics and Genomics department for their help.

Finally, I would like to acknowledge my husband because his strength has given me strength, and his humor and patience have made my life better. Thanks to all my family and friends, they are awesome!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Background

1.1.1 DNA-Binding Proteins

DNA-binding proteins are involved in many important biological processes in all living organisms. For example, DNA polymerase, DNA helicase, DNA topoisomerase, and DNA primase among others, play key roles in DNA replication, repair and recombination. Eukaryotes use histones, a family of basic DNA-binding proteins, to pack the DNA tightly in the nucleus of the cell. Another key function of DNA-binding proteins is to regulate gene expression, where DNA-binding proteins such as RNA polymerase and transcription factors (TFs) work together to either down-regulate or up-regulate gene expression. In bacteria, archea, and some viruses, restriction enzymes, an important group of DNA-binding proteins, are involved in protecting the organisms against foreign DNA, by identifying specific sequences in the invading DNA and cleaving at defined sites within the recognition sequence.

Transcription Factors

One of the largest and most diverse class of DNA-binding proteins are the transcription factors [82]. Transcription factors participate in regulating cell development, differentiation, and cell growth by binding specifically to short DNA sequences, known as transcription factor binding sites (TFBSs), and regulating gene expression. These

Table 1: Position weight matrix (PWM) of transcription factor Zif268 (Source: Jaspar [67]).

A	[	3	2	0	0	0	3	1	0	2	0	1]
C	[	4	1	13	0	0	0	0	0	10	0	0]
G	[	1	12	0	15	3	12	14	15	0	15	7]
T	[	7	0	2	0	12	0	0	0	3	0	7]

binding sites can be located in the promoter near the transcription start site, in an enhancer or other stretch of regulatory DNA many base pairs away from the promoter [84].

Transcription factors are key players in evolution. Changes affecting their function produce novel functions but they may also cause deleterious effects. Variations often occur in cis-regulatory elements [93]. The majority (≈93%) of disease- and trait-associated variants emerging from genome wide association studies and related strategies lie within noncoding sequence [70] that include transcription factor binding sites.

A transcription factor can recognize a collection of similar DNA-binding sites, which can be grouped together to define a DNA motif [95]. By assuming that each transcription factor-DNA base interaction is independent, the DNA-binding specificity of transcription factors can be expressed as a position weight matrix (PWM) (Table 1). PWMs describe the frequency of each nucleotide (A, C, G or T) at each position of a DNA-binding site [118], and can be visualized as motif logos (Figure 1). Recently, Yang *et al.* [118] demonstrated that augmenting existing motif databases with DNA shape features provides new insights into the mechanisms used by transcription factors to achieve DNA-binding specificity.

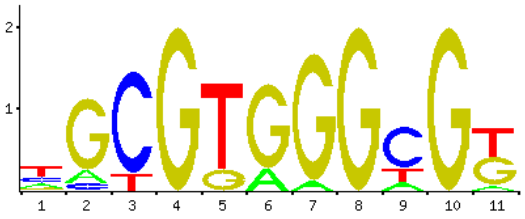


Figure 1: Motif logo of transcription factor Zif268 (Source: Jaspar [67]).

TF-DNA binding specificity *in vivo* is very complex. For example, the human genome consists of around 700,000 possible response elements, but only 3,000 transcription factors [84]. Pan *et al.* [84] pointed out that by viewing the TF-DNA recognition problem from the perspective of the sequence variability of the response elements overlooks cellular effects. The cellular network also plays a role in selective binding, by controlling the expression and post-translational states of the transcription factors and its cofactors. They integrated observations on transcription factor binding and activation with concepts of dynamic conformational ensembles to classify the mechanisms of transcription factor selectivity into three groups in the order of transcription initiation events: (i) coregulator recruitment followed by response element binding; (ii) response element binding followed by coregulator recruitment; and (iii) enhanceosome-mediated response element binding. The mechanisms can be differentiated by the affinity of the transcription factors to its response elements, low affinity transcription factors use the first mechanism, while the high-affinity ones use the second mechanism. Assigning each known transcription factor to one of the mechanisms is not a simple task, because it requires the understanding of the transcription initiation events, and studying the transcription factor-DNA interaction in the context of the cell environment.

Jolma *et al.* [45] recently analyzed binding specificities of most human transcription factors using high-throughput SELEX. Comparison of 79 pairs of experiments for full-length transcription factors and their DNA-binding domains revealed that in general, the DNA-binding domains define the primary DNA-binding specificity, since position weight matrices obtained for full-length transcription factors and its corresponding DNA-binding domains were very similar. They investigated high-resolution DNA-binding specificity for a large fraction of human transcription factors and found that more than half of all binding models for transcription factors are more than 10 base pairs in length. They also compared ortholog transcription factors from human and mouse, and found no obvious changes in binding specificities. However, in paralog transcription factors, the dimer orientation and spacing preferences were divergent, suggesting that these features evolve faster than primary binding specificities. These features can give rise to the multi-specificity nature of transcription factors, *i.e.*, multi-specificity is due to the ability of transcription factors to bind to both a monomeric and a dimeric site, and/or multiple different dimeric configurations. Although binding specificity models such as position weight matrices, assume position independency, there are several cases where dependency is observed. In those cases, new models need to be developed that can take into account the interdependency of base positions. They developed two models to address the issues by using a first-order Markov chain and taking the spacing and orientation into consideration for dimeric sites.

The structures and DNA-binding properties of the transcription factors can help us understand how genetic information is utilized. Transcription factors are modular in structure, consisting of independently functional protein domains. Transcription

factors consist of domains involved: (i) in specific DNA recognition (DNA-binding), (ii) in formation of homodimeric or heterodimeric proteins (dimerization), and (iii) in transcription initiation signaling (activation). There is no consensus in how to classify DNA-binding domains, but a general grouping of the DNA-binding motifs in known transcription factor families include: helix-turn-helix (*e.g.*, homeodomain), zinc finger (*e.g.*, steroid and thyroid hormone receptor superfamily), leucine zipper (*e.g.*, C/EBP, c-Jun, and c-Fos), and helix-loop-helix (*e.g.*, MyoD and myogenin).

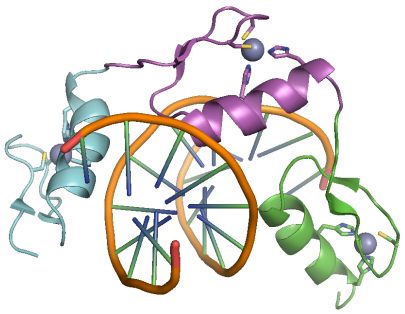
The zinc finger DNA-binding motif was first observed in transcription factor TFIIIA from the oocytes of the African clawed toads *Xenopus laevis*. A single zinc finger is approximately 30 residues in length, and may occur as monomers, dimers, or in sets of up to 30 zinc fingers [41]. Zif268 is the prototypic member of a family of immediate-early gene-encoded transcription factors that share highly similar Cys_2-His_2 zinc finger DNA-binding domains. The Cys_2-His_2 zinc finger motif is one of the most widely occurring eukaryotic DNA-binding domain structures. It folds into a compact globular domain that is composed of an antiparallel β-sheet followed by an α-helix and is stabilized by the coordination of a Zn^{2+} ion through two cysteine and two histidine residues (Figure 2a). [102]

GATA-binding proteins constitute a family of transcription factors that recognize a discrete target site, WGATAR (W=A or T, and R=G or A) [36]. Members of this family have been found in fungi, *Caenorhabditis elegans*, *Drosophila melanogaster*, birds, amphibians, and mammals [16]. DNA recognition is achieved through zinc fingers. In mammals, GATA-1, -2, -3, and -4, are expressed in distinct, yet often overlapping, cell types. The abilities of various members of the GATA family to

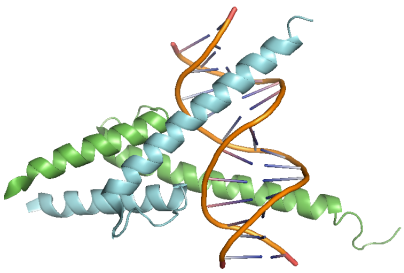
recognize closely related, but not identical, DNA sequence elements raises interesting possibilities as to how differential gene expression is accomplished in cells expressing more than one GATA protein. That is, differential regulation might be achieved by selective high-affinity binding of one, but no other, GATA family member to a target sequence because of subtle variations in their DNA-binding domains [72].

The basic helix-loop-helix (bHLH) transcription factors regulate gene expression by binding to specific DNA sequences. The basic domain of these proteins controls DNA binding to sites with the consensus sequence CANNTG (N=A or C or G or T), the E-box motif and is present in the regulatory regions of many tissue-specific genes. The various bHLH proteins can be divided into three groups: class A proteins (E12, E47, E2-2, and daughterless), the tissue-specific class B proteins (MyoD (Figure 2b), myogenin, MRF4, and achaete-scute), and class C proteins, which feature a tandem arrangement of bHLH and leucine zipper motifs (c-Myc, Max, upstream stimulatory factor [USF], AP4, TFE3, an TFEB).

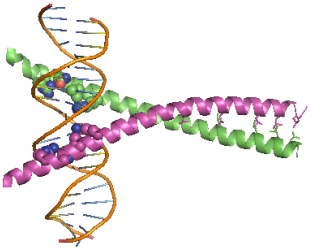
The leucine zipper (bZip) proteins possess a distinctive structural motif that consists of two sub-domains: a region of basic amino acids, which directly contacts DNA, adjacent to a hydrophobic heptad repeat, and a leucine zipper dimerization domain. CCAAT/enhancer-binding protein (C/EBP) family members (C/EBPα (Figure 2c), C/EBPβ, C/EBPγ, C/EBPδ, C/EBPϵ, and CHOP 10) are among the basic leucine zipper transcription factors, and they bind to specific DNA sequences as dimers. C/EBP family members show similar sequence preferences, and the consensus sequence is RTTGCGYAAY (R=G or A, and Y=C or T). The specificity of C/EBP family members may be derived from the characteristics of each factor, including the



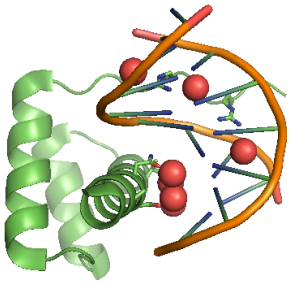
(a) Transcription factor Zif268 with three zinc finger motifs (d1aaya1 in cyan, d1aaya2 in magenta, and d1aay3 in green) in complex with DNA. Zn^{2+} ion is represented as a blue sphere. PDB identifier: 1aay



(b) Transcription factor MyoD with two bHLH motifs (d1mdya_ in green and d1mdyb_ in cyan) in complex with DNA. PDB identifier: 1mdy



(c) Transcription factor C/EBPα in complex with cognate DNA showing a leucine zipper (bZip) motif. DNA-binding residues are represented as "spheres" and the leucine residues that are part of the "zipper" are shown in "stick representation". PDB identifier: 1nwq



(d) Transcription factor PAX3 homeodomain in complex with DNA (HTH motif). DNA-binding residues are shown in "stick representation" and water molecules are represented as "spheres". PDB identifier: 3cmy

Figure 2: DNA-binding motifs of transcription factors.

expression profiles, the DNA binding affinities, the cofactors, and so on, in addition to the DNA-binding specificities. [80]

Homeodomain is a highly conserved DNA-binding domain found in many tran-

scription factors. The regulatory function of a homeodomain protein derives from the specificity of its interactions with DNA and with other proteins such as RNA polymerase or accessory transcription factors. Homeodomains utilize a helix-turn-helix (HTH) fold to contact DNA in the major groove, and utilize an N-terminal arm to contact DNA in the minor groove (Figure 2d). This contact is sequence-specific and contributes to the high affinity of homeodomains for DNA. [55]

Non-Specific DNA-Binding Proteins

All DNA-binding proteins show non-specific protein-DNA interactions [79]. However, some proteins, even though they interact with the DNA bases, are known to bind indiscriminately to any DNA sequence. To define non-specific DNA-binding proteins first we have to define what specificity is. Specificity involves binding one or several DNA sequences with higher affinity than the other DNA sequences. Therefore, non-specificity describes binding to any DNA sequence with practically the same affinity [100].

Non-specific DNA-binding proteins are important in many biological processes, such as DNA replication (Figure 3a), repair, and recombination (DNA polymerases, DNA helicases, DNA topoisomerases, and DNA primases), gene regulation (RNA polymerases), and cellular organization and metabolism (histones).

DNA polymerases (pols) α, β, γ, δ, and ϵ are the key enzymes required to maintain the integrity of the genome. DNA polymerases synthesize DNA efficiently and accurately, which is crucial to ensure the faithful transmission of genetic information from parents to offspring. All free-living organisms encode several DNA polymerases,

but there is a rich variety within the DNA polymerase family. The function of the core polymerase activity is to add deoxynucleotides onto the growing end of a DNA primer strand, although another important attribute of enzymes of this type, imply that many of the physicochemical mechanisms used to discriminate between correct and incorrect base pairs have been preserved throughout this family of enzymes. [47]

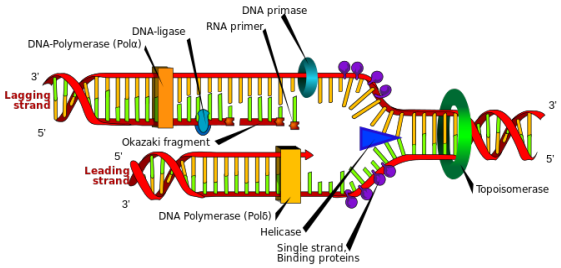
DNA helicases are enzymes that facilitate the unwinding of duplex DNA, which is a prerequisite for DNA replication and repair, and provides the single-stranded DNA template for DNA polymerase to copy. DNA helicases disrupt the hydrogen bonds that hold the two strands of duplex DNA together [68]. DNA topoisomerases are enzymes that also disentangle DNA strands or duplexes in a cell. They play an important role in replication, transcription, chromosome condensation, and maintenance of genome stability. They function differently from DNA helicases, since DNA topoisomerases alter the linking number of the duplex DNA molecule through phosphodiester bond breakage and reunion.

DNA primases are enzymes involved in DNA replication. Most DNA primases can be divided into two classes. The first class contains bacterial and bacteriophage enzymes found to be associated with replicative DNA helicases. These prokaryotic primases contain three distinct domains: an amino terminal domain with a zinc ribbon motif involved in binding template DNA, a middle RNA polymerase domain, and a carboxyl-terminal region that either is a DNA helicase or interacts with a DNA helicase. The second major primase class comprises heterodimeric eukaryotic primases that form a complex with DNA polymerase alpha and its accessory B subunit. The small eukaryotic primase subunit contains the active site for DNA synthesis, and its

activity correlates with DNA replication during cell cycle.

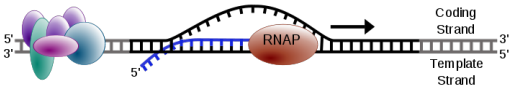
DNA-dependent RNA polymerases (Pol I, Pol II, and Pol III) are responsible for the synthesis of all cellular RNA and play a central role in gene expression (Figure 3b). Pol I produces ribosomal RNA, Pol II synthesizes messenger RNAs and small nuclear RNAs, and Pol III produces transfer RNAs and other RNAs. RNA polymerases are large and complex enzymes composed of several polypeptide chain subunits. Pol I, II and III comprise 14, 12, and 17 subunits, respectively. Ten subunits form a structurally conserved core, and additional subunits are located in the periphery. The complexity and large size of multisubunit RNA polymerases have prevented elucidation of their structure for a long time, but even with the structural information available, many aspects of RNA polymerases remain unresolved. Among these open issues are how these enzymes are regulated by coregulatory assemblies, *e.g.*, the mechanisms involved in the interactions with other molecules, including DNA. [19]

In eukaryotes, chromosomal DNA is complexed with many DNA-binding proteins such as histones [10] that function as building blocks to package eukaryotic DNA into repeating nucleosomal units that are folded into higher-order structures (Figure 3c). Histones are small basic proteins consisting of a globular domain and a more flexible and charged NH_2-terminus (histone tail) that protrudes from the nucleosome [44]. Once thought of as static, non-participating structural elements, it is now clear that histones are integral and dynamic components of the machinery responsible for regulating gene transcription [101].



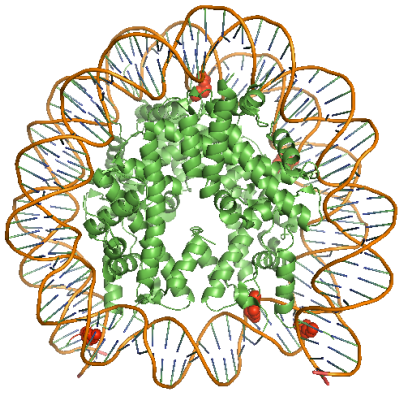
(a) Many non-specific DNA-binding proteins work together in the DNA replication fork. [a]

[a]From Wikipedia: DNA replication



(b) In the transcription process, RNA polymerase (RNAP) uses DNA (black) as a template to produce RNA (blue) [a].

[a]From Wikipedia: Gene expression



(c) Histone octamer forming a human nucleosome core particle. Red spheres show DNA base binding residues, that in this case are all arginines. PDB identifier: 3WKJ.

Figure 3: Non-specific DNA-binding proteins.

Type II Restriction Enzymes

Restriction-modification systems [111] comprise pairs of opposing intracellular enzyme activities: an endodeoxyribonuclease (ENase) and a DNA-methyltransferase (MTase). The enzymes interact with specific sequences of nucleotides in DNA and recognize double-stranded DNA; a few also recognize single-stranded DNA. ENases and MTases from the same system recognize the same sequences. In some restriction-modification systems, the two activities are combined in a single, multi-subunit enzyme, but in most systems they are separate.

Restriction endonucleases catalyze double-strand cleavage of DNA. Cleavage occurs once for each occurrence of the recognition sequence, and is accomplished by hydrolysis of one phosphate-deoxyribose bond in the backbone of each DNA strand. In many systems, cleavage occurs at a fixed position with respect to the recognition sequence, either within the sequence or a few bases to one side of it. In others, hydrolysis takes place at an indefinite distance from the recognition sequence.

Type II systems are the simplest and the most numerous. Type II endonucleases and methyltransferases act independently and have simple requirements: the endonucleases require Mg^{2+}, the methyltransferases require AdoMet.

Type II recognition sequences are mainly symmetric (Table 2). Some sequences are continuous (*e.g.*, BamHI, BglII, and EcoRI) while others are interrupted. The interruptions can be short (*e.g.*, BcnI, BsoBI, and EcoRII) or relatively long (*e.g.*, BglI and SfiI). The sequences comprise four to eight specific nucleotides, and they vary in base composition. Symmetric sequences are economical sequences; one protein can react with both strands of duplex since it appears the same regardless of orientation. Type II endonucleases generally act as homodimers, an association that facilitates the coordinated cleavage of both strands. Cleavage by type II endonucleases occurs symmetrically within the recognition sequences. Some endonucleases cleave on the 5' side of the dyad axis (*e.g.*, BamHI, BcnI, and BglII), producing fragments with 5' single-stranded termini of various lengths; others cleave in the center (*e.g.*, EcoRV, HincII, and NaeI), producing flush termini; yet others cleave on the 3' site (*e.g.*, BglI, Hpy188I, and PacI), producing 3' single-stranded termini.

Due to its simplicity and its highly-specific DNA-binding nature, type II restriction

Table 2: Examples of recognition sequences of type II restriction enzymes.

Name	**Organism**	**Recognition Sequence**
BamHI	*Bacillus amyloliquefaciens* H	GˆGATCC
BcnI	*Bacillus centrosporus* RFL1	CCˆSGG
BglI	*Bacillus globigii*	GCCNNNNˆNGGC
BglII	*Bacillus globigii*	AˆGATCT
BpuJI	*Bacillus pumilus* RFL1458	CCCGT
BsoBI	*Bacillus stearothermophilus* JN2091	CˆYCGRG
BstYI	*Bacillus stearothermophilus* Y406	RˆGATCY
Ecl18kI	*Enterobacter cloaceae* 18k	ˆCCNGG
EcoO109I	*Escherichia coli* H709c	RGˆGNCCY
EcoRI	*Escherichia coli* RY13	GˆAATTC
EcoRII	*Escherichia coli* R245	ˆCCWGG
EcoRV	*Escherichia coli* J62 pLG74	GATˆATC
FokI	*Flavobacterium okeanokoites*	GGATG (9/13)
HincII	*Haemophilus influenzae* Rc	GTYˆRAC
HindIII	*Haemophilus influenzae* Rd	AˆAGCTT
HinP1I	*Haemophilus influenzae* P1	GˆCGC
Hpy188I	*Helicobacter pylori* J188	TCNˆGA
Hpy99I	*Helicobacter pylori* J99	CGWCGˆ
MspI	*Moraxella species*	CˆCGG
MvaI	*Micrococcus varians* RFL19	CCˆWGG
NaeI	*Nocardia aerocolonigenes*	GCCˆGGC
NgoMIV	*Neisseria gonorrhoeae* MS11	GˆCCGGC
NotI	*Nocardia otitidis-caviarum*	GCˆGGCCGC
PacI	*Pseudomonas alcaligenes*	TTAATˆTAA
PspGI	*Pyrococcus species* G1H	ˆCCWGG
PvuII	*Proteus vulgaris*	CAGˆCTG
SfiI	*Streptomyces fimbriatus*	GGCCNNNNˆNGGCC
SgrAI	*Streptomyces griseus*	CRˆCCGGYG
ThaI	*Thermoplasma acidophilum*	CGˆCG

enzymes are a good model to study the mechanisms of DNA-binding specificity.

1.1.2 Protein-DNA Binding Specificity

Proteins that bind to specific recognition sequences on DNA do so against a background of a large number of more or less similar non-specific sequences in the genome. To appreciate the functional specificity of a particular binding site, one must know not only its specific binding affinity for the regulatory protein, but also the distribution of

binding affinities for all possible competitive sites. The binding affinity of the protein P for the sequence S can be defined by the dissociation constant K_d, which is the ratio of the off-rate k_{off} to the on-rate k_{on} that governs the binding process of P and S. Stormo and Zhao [100] define the term specificity as to "how well a protein can distinguish between different sequences." They suggest that the complete specificity of a protein can be defined by the list of K_ds to all possible binding sites. Therefore, the terms affinity and specificity are not independent, the binding specificity of a protein requires, ideally, the characterization of the binding affinities of the protein against all possible DNA sequences.

The field of drug design is in need of understanding the forces that drive protein-DNA interactions. The ability to design molecules that bind to specific DNA has many potential applications including the directed control of gene expressions, for example, the inhibition of the c-Myc transcription factor, which is over-expressed in most human cancers [23].

DNA recognition is a key step of biological regulatory processes. The increase of protein-DNA complexes in the Protein Data Bank (PDB) [9] has provided an insight on how proteins interact with DNA. Protein-DNA binding specificity often involve the formation of hydrogen bonds between protein side chains and DNA bases. It is known that every DNA base pair has a unique hydrogen-bonding signature in the major groove, but not in the minor groove. Thus, the recognition of specific DNA sequences would be expected to take place primarily in the major groove by the formation of a series of amino-acid- and base-specific hydrogen bonds [91]. This "base readout" mechanism can explain most of the binding speificity, but it is not

the entire story.

It has been suggested that the binding specificity is contributed by two mechanisms. The first one, known as direct or base readout, invokes contact of protein moieties with base-specific functional groups on the nucleic acid. Base readout interactions involve the relative three-dimensional orientation of various contact points in a given sequence. These interactions of amino acids side chains and the array of hydrogen-binding and van der Waals contacts available on DNA comprise combinations of charge and shape complementarity [29].

The second, the indirect or shape readout mechanism, has been proposed to explain cases in which the specificity observed in biochemical experiments cannot be accounted for by direct hydrogen bonding interactions between the macromolecules. In indirect or shape readout, the sequence-dependent conformation of nucleic acid structure is recognized instead, via protein contacts with the sugar-phosphate backbone and/or with nonspecific portions of the base [37]. Shape recognition of nucleic acid is being increasingly recognized as playing an equally important role in DNA recognition [116]. In many complexes, the DNA assumes conformations that deviate from the structure of an ideal B-form double helix, sometimes bending in such a way to optimize the protein-DNA interaction, and in some cases undergoing large conformational changes as in the opening of the minor groove in the complex formed between TBP and the TATA box. The term "indirect readout" was first coined to describe such recognition mechanisms that depend on the propensity of a given sequence to assume a conformation that facilitates its binding to a particular protein. The bases involved in such mechanisms need not be in contact with the protein

and, for example, can be found in linker sequences that connect two half-sites that are themselves bound by individual protein subunits [91]. Proteins are able to perform DNA recognition by using both base readout and shape readout mechanisms, the combination of which allows the protein great subtlety in sequence recognition through an ensemble of non-covalent contacts.

Molecular interactions have been studied for several decades. Protein-protein interactions have been studied in more depth than protein-nucleic acid interactions. Nadassy *et al.* [78] showed that protein-protein interactions are different from protein-DNA interactions, therefore, specialized protein-DNA models need to be developed, in order to understand these complex interactions and ultimately being able to predict the effect of mutations in DNA-binding proteins.

A simple protein-DNA recognition code does not exist, however, some proteins present simple recognition mechanisms and can be modeled more easily. Mandel-Gutfreund and Margalit [65] described a quantitative measure of base-amino acid interactions obtained by computing the log odds of the observed pair frequencies and those expected at random. The results reflect a correspondence between the computed scores and results of binding experiments of the protein Zif268. The drawback of the study is that position independence needs to be assumed, and that the correlation with the Zif268 protein might be due to the simple binding mechanism of this particular protein. With the increasing number of protein-DNA complexes in the PDB, quantitation of the different parameters, like position-dependent effects and coupled interactions, as well as predictions of the DNA structure in the binding site can be obtained. Still, the effects of each mechanism (direct or indirect readout)

needs to be quantified as well.

Contreras-Moreira *et al.* [18] systematically explored the conservation of structural features of binding interfaces, centering the study both at the protein and DNA sides of docked complexes. They estimated that the average contribution of indirect readout to specific binding is approximately of one every five DNA bases, with the notable exception of restriction enzymes, which doubles its contribution. With respect to direct readout, hydrogen bonds dominate DNA recognition, with a minor fraction of hydrophobic interactions. Luscombe *et al.* [61] studied protein-DNA interactions at an atomic level, and concluded that van der Waals contacts are mostly used to stabilize the complex, water-mediated bonds are mostly used as gap fillers in the protein-DNA interface, and complex interactions are expected to play an important role in providing specificity.

To describe the effect of each protein-DNA interaction as specific or non-specific, a study performed by Ashworth and Baker [4] utilize the atomic model developed by Havranek *et al.* [38] of the energetics of amino acid-nucleotide interactions to estimate the extent to which amino acids are optimal for affinity or specificity. The correspondence with experimental results suggested the usefulness of the method for rapidly formulating hypotheses about the roles of amino acids at protein-DNA interfaces, given a high-resolution structure of the protein-DNA complex. But still, the method significantly underestimates the optimization of native amino acid sequence for specificity in complexes in which sequence recognition is dominated by indirect readout mechanisms.

Another interesting feature of DNA-binding proteins is the level of specificity. In

that extent, Luscombe and Thorton [64] compared the conservation of amino acid residue sequences in DNA-binding protein families with different levels of specificity. The protein families were classified into one of three classes on the basis of their DNA-binding specificities: (i) non-specific families, proteins that bind promiscuously and have no requirement for any specific base sequence; (ii) highly specific families, proteins that bind DNA specifically, and all members target a common base sequence; and (iii) multi-specific families, proteins that bind specifically, but different members bind distinct and different targets. The study shows a clear difference between DNA-binding proteins with different levels of specificity in terms of residue conservation patterns. However, we believe that DNA-binding proteins are different in other aspects besides residue conservation, like the combination of base and shape readout mechanisms they use to perform DNA recognition.

In conclusion, very interesting results have been found so far by studying protein-DNA complexes using a structural approach. A more detailed analysis that unveils the protein-DNA interactions and structural features used by the proteins to achieve DNA-binding specificity will be useful to out understanding of protein-DNA recognition.

1.1.3 Protein Flexibility and Intrinsic Disorder of DNA-Binding Proteins

It has been recently believed that numerous proteins lack intrinsic globular structure or contain long disordered segments and that disorder is their normal, functional state [24]. Disordered segments appear to be common in proteins encoded by higher eukaryote genomes [24]. The intrinsic lack of structure can confer functional

advantages, including the ability to bind to several different targets [115]. Many transcriptional activation domains are either unstructured or partly structured, and their interactions with their targets involve coupled folding and binding events. Well-characterized examples include the transactivation domain of p53, which undergoes a coil-to-helix folding transition on binding to the cellular oncoprotein MDM2 [24].

Wild-type p53 protein is commonly described as a tumor suppressor or an antioncogene product. Alteration or loss of p53 function is associated with a wide variety of human tumor cells. Mutations in the p53 gene are the most frequently observed genetic lesions in spontaneous human cancers. p53 functions as a node in numerous signaling pathways such that it regulates many important biological activities, from fertility and development to maintaining genomic stability and cell death. As the diversity of p53-dependent activities widens to include key roles in metabolism and development, more questions arise, but it is clear that p53 is therapeutically important and numerous approaches are being employed to reconstitute its expression in tumors.

Fong *et al.* [27] used missing residues in PDB structures to define disordered regions. Their analysis reveals a variety of categories where intrinsic disorder can play an important functional role, the most frequent of them being nucleic acid binding proteins, enzymes, ATP binding proteins, receptor binding proteins, and other ligand binding proteins.

Since intrinsic disorder is important to out understanding of the mechanisms involved in molecular interactions, it is desirable to predict disordered regions. A number of predictions have been developed based on the characteristics of disordered

fragments of proteins. Linding *et al.* [57] developed a sequence-based tool to predict the propensity of protein regions to be ordered or disordered. They compared their results with Cα B-factor values of a set of PDB structures, and achieved a specificity of 88%. Radivojac *et al.* [88] concluded that high B-factor ordered regions are more similar to disordered regions than to low B-factor ordered regions. This means that high-B factor ordered regions and missing residues can be used to define disordered regions in protein structures.

Liu *et al.* [58] examine the linkage between disorder and protein function from a thermodynamics point of view. The results show that eukaryotic genomes have more disordered residues than prokaryotic genomes. They also concluded that the distribution of the amount of disorder depends strongly on protein function, *e.g.*, proteins with "protein binding" function present a large range of disorder whereas proteins involved in "catalytic activity" have a strong preference for a stable folded state. A similar analysis can be performed to explore the propensity of disorder in DNA-binding proteins, and look for a relationship of DNA-binding specificity and intrinsic disorder.

Günter *et al.* [35] analyzed conformational diversity within seven DNA-binding proteins that have frequently been crystallized in DNA-complexed and free states. The local structure of the DNA-binding sites of all seven proteins is influenced by DNA. This constitutes a problem for protein-DNA docking prediction models, where conformational space increases enormously when considering protein and DNA flexibility. A more promising way of predicting protein-DNA interactions is to combine geometric criteria with additional physical parameters to narrow down the conforma-

tional space by several orders of magnitude.

Vuzman and Levy [110] studied the effect of disordered regions on protein-DNA interactions. They showed that disordered tails have higher occurrence in DNA-binding proteins than in non-DNA-binding proteins. In conclusion, they mentioned that the composition and distribution of charges within intrinsically disordered regions regulates the strength of protein-DNA interactions. Dunker and Uversky [23] found that protein clouds (dynamic ensembles of intrinsically disordered regions) are druggable, which is a desirable feature, since the transcription factors might contain significant amounts of intrinsic disorder, according to computational analysis. Transcription factors present a higher degree of disorder in the activation domains than in the DNA-binding domains. However, Guo *et al.* [33] showed that the flanking regions of DNA-binding domains in human transcription factors generally exhibit significant disorder.

In summary, disorder has been studied extensively in the last decade, but its contribution to DNA-binding specificity is still unknown. By comparing the level of disorder in DNA-binding proteins we can measure if disorder or flexibility is directly involved in DNA-binding specificity.

As described above, though previous studies have revealed many important characteristics in protein-DNA recognition, it still not clear how the protein-DNA binding specificity is determined. In my dissertation research, I carried out a statistical analysis on DNA-binding protein structures, and compare static and dynamic structural features to identify major structural determinants of DNA-binding specificity. I addition, we applied these features for protein-DNA docking assessment and showed a

major improvement respect to the previous methods.

CHAPTER 2: STATISTICAL ANALYSIS OF STRUCTURAL DETERMINANTS FOR PROTEIN-DNA BINDING SPECIFICITY

2.1 Introduction

Specific interactions between proteins and their DNA target sequences are essential in many fundamental biological processes and aberrant changes in binding specificity can cause serious consequences [94, 26, 63, 54]. It has been demonstrated that altered binding specificity between mutated transcription factors and their DNA target sequences plays a role in a broad variety of cancers [26, 34, 104, 17]. On the other side of the specificity spectrum, many DNA-binding proteins can bind to a wide range of DNA sequences. These non-specific DNA-binding proteins are also critical for fundamental cellular functions, including processing and packaging of DNA [1].

DNA-binding specificity generally refers to two interrelated terms: "sequence specificity" and "degree of specificity" [98]. For example, type II restriction endonucleases EcoRI and BamHI specifically recognize their DNA target sequences GAATTC and GGATCC, respectively. Both enzymes show very high degrees of specificity towards different DNA sequences. Some transcription factors, such as homeodomains Ubx (from *Drosophila melanogaster*) and Nkx3-1 (from *Homo sapiens*), bind to different DNA sequence patterns, but with similar, high sequence conservation [66]. On the other hand, homeodomain Dbx1 (from *Mus musculus*) has a similar binding sequence pattern to Ubx, but most positions allow more variations and are less conserved

[98, 66]. Most experimental and computational studies have focused on identifying sequence specificity or sequence patterns. No simple recognition rules between particular amino acids and specific DNA bases have been found, although some preferred pairings were observed [69, 81, 62, 59, 103, 117]. In this study, we focus on analysing structural determinants for different degrees of protein-DNA binding specificity.

Current structural studies range from individual cases to comparative analyses. Homing endonucleases [6, 77, 108] and zinc fingers [83, 49, 96] are two widely studied family proteins. Ashworth *et al.* developed a computational model and applied it to redesign the specificity of a homing endonuclease, I-MsoI [5]. In their model, the specificity is described by packing, hydrogen bonding, solvation and electrostatic interactions. Several comparative studies have also been conducted to examine DNA-binding specificity. Luscombe and Thornton investigated the effects of individual mutations on binding specificity using small datasets, due to limited availability of protein-DNA complex structures at that time. They carried out a comparative analysis on two groups of transcription factors (including highly specific and multi-specific) and non-specific DNA-binding proteins [63]. Ashworth *et al.* predicted the contribution of each interface residue to the binding affinity and binding specificity of four types of DNA-binding proteins: a) helical-motif transcription factors, b) restriction endonucleases, c) homing endonucleases, and d) non-specific DNA-binding enzymes [4]. Another comparative analysis was performed on nine SCOP superfamilies, including homing nucleases, ribbon-helix-helix, glucocorticoid receptor-like, zinc fingers, homeodomain-like, winged helix, P53-like, lambda repressor-like, and restriction endonuclease-like [18]. By comparing the ratio of indirect/direct readout and

the frequency of atomic interactions, Contreras-Moreira *et al.* concluded that these specificity features are generally conserved and superfamily-specific [18].

Two readout mechanisms are considered to contribute to the binding specificity between proteins and DNA, base readout and shape readout (also called direct and indirect readout, respectively) [63, 75, 91, 89, 121]. The base readout describes contributions from direct interaction of protein side-chains with DNA bases. The shape readout, on the other hand, describes the role of DNA shape and indirect contacts between proteins and DNA [91, 89, 90]. The combination of base and shape readouts provides a general picture for specific protein-DNA interactions. However, what controls the degree of binding specificity, or why some proteins are highly selective on binding sequences while others are less stringent, is still not clear.

Protein-DNA recognition is by nature a dynamic process that involves delicate structural fitting between proteins and DNA [30, 42]. However, the exact role of flexibility and intrinsic disorder to the binding specificity is not well understood. As the specific interactions are mainly contributed by hydrogen bonding between proteins and DNA, high specificity between proteins and their cognate binding sequences is considered an optimized result of shape fit and binding thermodynamics. We have demonstrated previously that a point mutation F10V in P22 Arc repressor, which does not make direct DNA base contact, affects the degree of binding specificity by altering the flexibility of residues involved in direct base contacts [98]. Therefore, more complete description in terms of both static and dynamic features is needed to fully understand the specificity in protein-DNA recognition. With the advancement of structure determination techniques, the number of protein-DNA complex structures

in Protein Data Bank (PDB) is increasing at a higher rate [9]. Currently there are over 3000 protein-DNA complex structures in PDB. The availability of a large number of protein-DNA complexes and their corresponding unbound protein structures makes it feasible to conduct a more comprehensive study of protein-DNA binding specificity. In this paper, we carried out a comparative analysis to investigate the static and dynamic structural features for protein-DNA binding specificity.

We first constructed datasets of protein-DNA complex structures and group these DNA-binding proteins into three general classes based on decreasing degrees of DNA-binding specificity: type II restriction enzymes (highly specific, HS), transcription factors (multi-specific, MS), and non-specific (NS) DNA-binding proteins. It should be noted that there are no distinct groups with respect to DNA-binding specificity; rather, we consider that DNA-binding proteins run a gamut of specificities from very specific (recognize exact sequences) to non-specific. For example, type II restriction enzyme MvaI recognizes CCWGG (W can be either A or T). On the other hand, some transcription factors, such as some nuclear receptors, exhibit high specificity [45, 32, 31]. Nevertheless, type II restriction enzymes, in general, have higher binding specificity than transcription factors. In this study, type II restriction enzymes with lower binding specificity, such as BglI (recognition sequence GCCNNNNN^NGCC, where N represents any base), are not included in the HS dataset to minimize the potential specificity overlap between the HS and MS groups. In addition to the three-class design, we used bound-unbound (or holo-apo) pairs for identifying dynamic structural features that contribute to binding specificity, such as the range of conformational change upon DNA-binding [42]. Furthermore, to assess the relation-

ship between protein flexibility and binding specificity, we compared the structural diversity of DNA-binding proteins, by comparing multiple apo and holo structures of the same DNA-binding protein.

Our results demonstrated a trend in several static structural features: amino acid propensities, interface size, number of residue-base contacts, backbone to base contact ratios, major to minor groove contact ratios, number of protein-DNA hydrogen bonds, and DNA shape parameters, among the three groups. We found that negatively charged aspartate is highly enriched in base interactions in highly specific DNA-binding proteins while it is depleted in multi-specific and non-specific DNA-binding proteins. Our data revealed a tight connection between aspartate and the cytosine base. We also showed the importance of two aromatic residues, tyrosine and histidine, in conferring specific protein-DNA binding. To our knowledge, this is the first large-scale comparative study to demonstrate the critical role of aspartate, tyrosine and histidine in specific protein-DNA recognition. In terms of dynamic features, we analyzed the protein conformational changes upon DNA-binding and their structural variations in both free form and bound state. We found that highly specific DNA-binding proteins show larger conformational changes upon DNA-binding while the non-specific DNA-binding proteins have smaller structural variations and conformational changes.

2.2 Materials and Methods

2.2.1 Datasets

Three different datasets were generated in this study for different comparative analyses: (i) pdNR30, a non-redundant protein-DNA complex dataset, for investigation of static structural features related to protein-DNA interactions; (ii) pairNR30, a non-redundant bound-unbound pairs of DNA-binding domains, for comparing conformational changes upon DNA-binding; and (iii) svSet, a dataset for comparison of structural variations of DNA-binding domains.

A total of 3,098 protein-DNA complexes were selected from the PDB [9]. Of these complexes, some contain only DNA-binding domains while others represent full-length DNA-binding proteins, including signal-sensing domains or trans-activating domains besides DNA-binding domains. In this work, we used DNA-binding domains in protein-DNA complexes as comparison units to maintain consistency. For structural domain annotation, we combined the two most widely used structural classification databases, CATH [97] and SCOPe [28], with manual inspection if an annotation is not available in either database (Figure 4). A DNA-binding domain was selected if there are at least 4 protein-DNA contacts with a distance cutoff of 3.9Å, and the domain has 40 or more amino acids.

Figure 5 shows how pdNR30 was generated. First, all the X-ray crystal structures of protein-DNA complexes were selected from PDB. A series of quality filtering steps were then carried out. X-ray structures with resolution higher than 3Å and R-factor more than 0.3 were removed. Protein-DNA complexes with single-stranded DNA

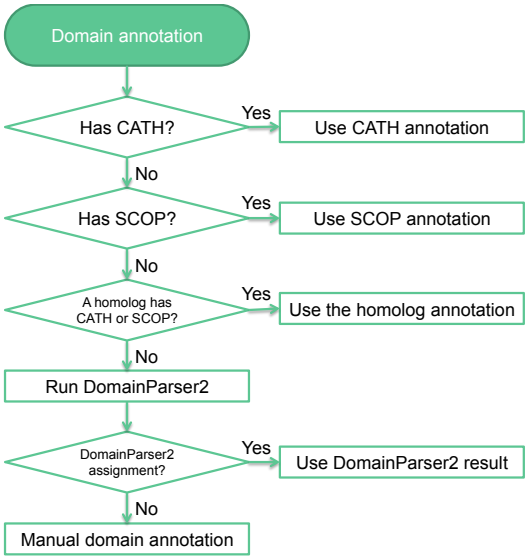


Figure 4: Flowchart of DNA-binding domain annotations.

(ssDNA) were also filtered out. For the false ssDNA complexes, in which coordinates are provided for only one DNA chain of a double-stranded DNA, we used our in-house program PDA (Protein-DNA complex structure Analyzer) to reconstruct these protein-DNA complexes by calculating the positions of the missing complementary DNA chain [51]. Since the main goal of this analysis is to study the structural features that contribute to the degree of protein-DNA binding specificity, removing mutant protein structures and non-cognate protein-DNA complexes is essential as it would add noise to our analysis. For example, researchers often use protein and/or DNA mutants to study the effects of mutations on protein-DNA binding specificity [92].

The DNA-binding domains that interact with double-stranded DNA in the complex structures were then annotated as HS (highly specific), MS (multi-specific), or NS (non-specific) DNA-binding proteins [63] based on their DNA-binding specificity and

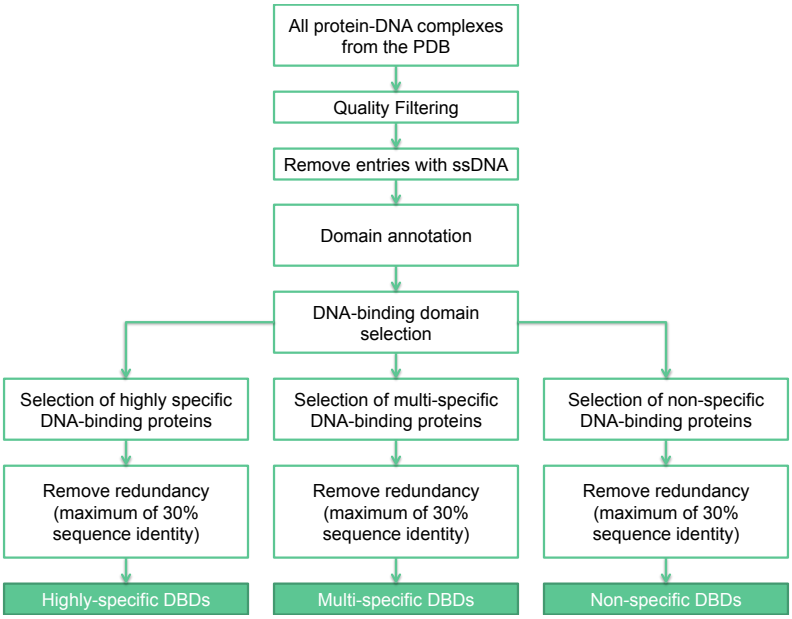


Figure 5: Flowchart for compiling the non-redundant datasets of DNA-binding domains.

function. Type II restriction enzymes generally belong to the highly specific group and were selected based on enzyme classification number 3.1.21.4 and keywords in the PDB, combined with manual inspection of the recognition sequences to assure that the bindings are highly specific. Transcription factors belong to the multi-specific group, since they generally recognize multiple conserved sequences. Transcription factors were selected using TFinDit, a data repository for known transcription factor-DNA complex structures [105]. Except for histones, DNA polymerases and RNA polymerases, the annotation of other non-specific DNA-binding proteins is not trivial, which was done based on manual inspection of the PDB entry and related references. After clustering with a sequence identity of 30% using CD-HIT [56], the non-redundant set pdNR30, was generated by selecting one representative from each cluster, based on resolution and the number of missing residues. The pdNR30 dataset

has 28 HS, 115 MS and 52 NS DNA-binding domains in complex with DNA (Table S1).

The second dataset, pairNR30, was generated in a similar way except that we started with a list of DNA-binding domains with both bound and unbound structures in PDB. The DNA-binding domains in free, unbound state were selected if they have 100% sequence identity and at least 80% coverage with their corresponding structure in the dataset of bound structures (Figure 6). The pairNR30 dataset consists of 11 HS, 41 MS and 16 NS bound-unbound DNA-binding domain pairs (Table S2).

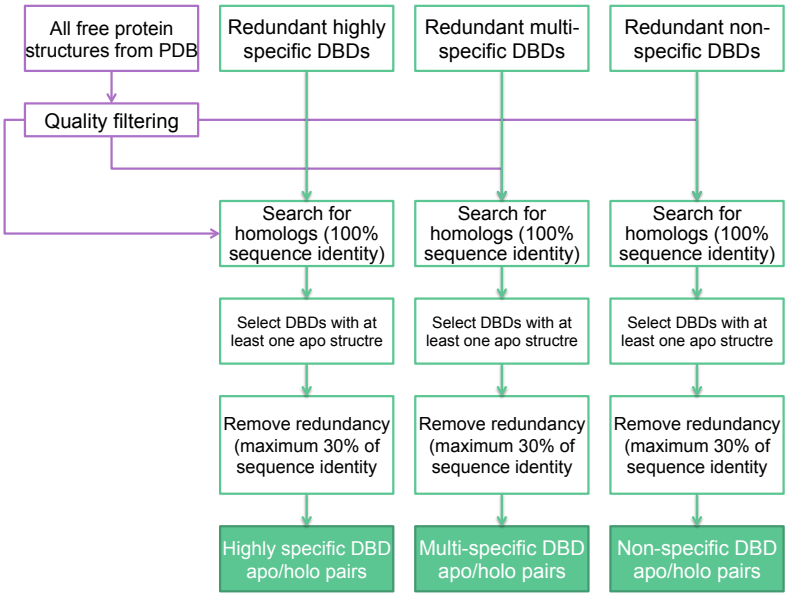


Figure 6: Procedure to compile the non-redundant apo-holo pairs of DNA-binding domains.

The third dataset, svSet has three components: (i) multiHolo, DNA-binding domains with at least 6 PDB structures in complex with cognate DNA; (ii) multiApo, DNA-binding domains with at least 6 structures in the unbound state; and (iii) multiApoHolo, DNA-binding domains with at least 4 structures in both the unbound state

and bound state with cognate DNA. This dataset was used to study the structural variations of DNA-binding domains in free state and in complex with DNA. There are 6 HS, 32 MS, and 24 NS DNA-binding domains in multiHolo dataset (Table S3). Since the number of cases for the HS is small in the multiApo and multiApoHolo sets, we combined the HS and MS cases and compare specific (HS+MS) against non-specific (NS) DNA-binding domains. The multiApo set consists of 9 specific (HS+MS) and 6 non-specific (NS) DNA-binding domains (Table S4) while the multiApoHolo set has 10 specific (HS+MS) and 4 non-specific (NS) DNA-binding domains (Table S5).

2.2.2 Comparison of Structural Features of Protein-DNA Interactions

A comparative analysis of structural features that contribute to DNA-binding specificity was first carried out with the pdNR30 dataset that consists of a non-redundant dataset of DNA-binding domains in complex with DNA (28 HS, 115 MS and 52 NS DNA-binding domains). The structural features for protein-DNA interactions include: 1) protein side-chain/DNA-base binding propensities, 2) protein-DNA contact area (PDCA), 3) number of residue-base contacts (NRBC) [50], 4) the number and geometry of hydrogen bonds, 5) backbone to base contact ratio, 6) minor to major groove contact ratio, and 7) DNA shape.

The DNA binding propensity (p_{ij}) for an amino acid i is calculated as the ratio of the percentage of the amino acid in protein side-chain/DNA base contacts and the percentage of the amino acid in the specific dataset j (Equation 1) [50]. Jackknife resampling was used to estimate the variances and potential bias of the data.

$$p_{ij} = \frac{\frac{F_{ij}}{\sum_{i=1}^{20} F_{ij}}}{\frac{D_{ij}}{\sum_{i=1}^{20} D_{ij}}} \quad (1)$$

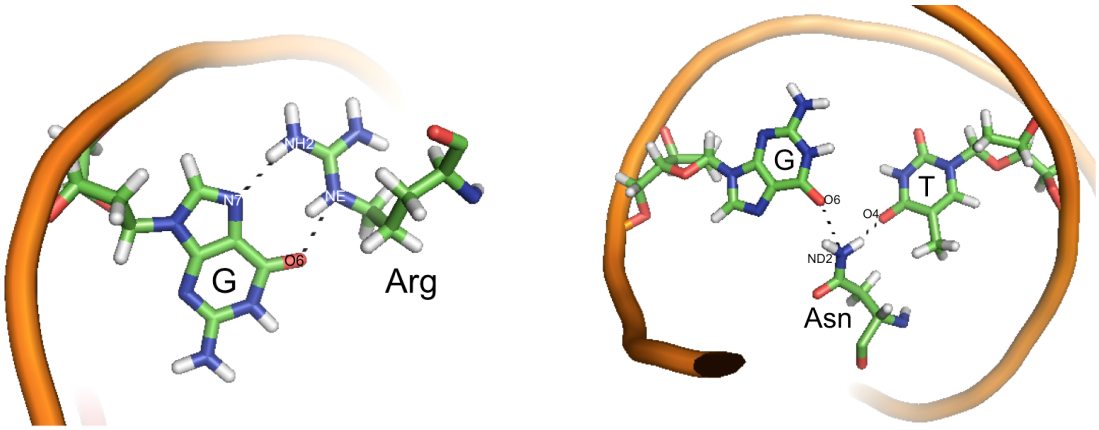
where F_{ij} is the total number of binding residues (whose side-chain atoms are within $3.9\mathring{A}$ of DNA base atoms) of type i in dataset j. D_{ij} is the total number of residues of type i in dataset j, including missing residues. If $p_{ij} > 1$, residue i in dataset j is considered to be enriched in protein side-chain/DNA base contacts.

The PDCA is determined by calculating the difference in solvent accessible surface area (SASA) between the individual protein ($\text{SASA}_{protein}$), DNA structure (SASA_{DNA}) and the corresponding protein-DNA complexe ($\text{SASA}_{complex}$) [50].The solvent accessible surface areas were measured by Naccess with default parameters [39]. Protein-DNA contacts were identified using a distance cutoff of 3.9Å between side-chain heavy atoms and all DNA heavy atoms. These residue-DNA interactions were divided in two non-overlapping sets: (i) residues that are in contact with DNA base (NRBC: number of residue-base contacts) and (ii) residues that are in contact with DNA backbone only. We also calculated the NRBC density, the ratio of NRBC over the PDCA, which represents the number of residue-base contacts per Å^2.

$$PDCA = \frac{SASA_{protein} + SASA_{DNA} - SASA_{complex}}{2} \quad (2)$$

Hydrogen bonds in protein-DNA complexes were identified with HBPLUS [71]. In addition to simple hydrogen bonds, we also analyzed the differences among the three specificity groups in terms of other types of hydrogen bond geometry, *e.g.*, bidentate hydrogen bond that is defined when a residue forms more than one hydrogen bond

with different acceptor and/or donor atoms (Figure 7a).



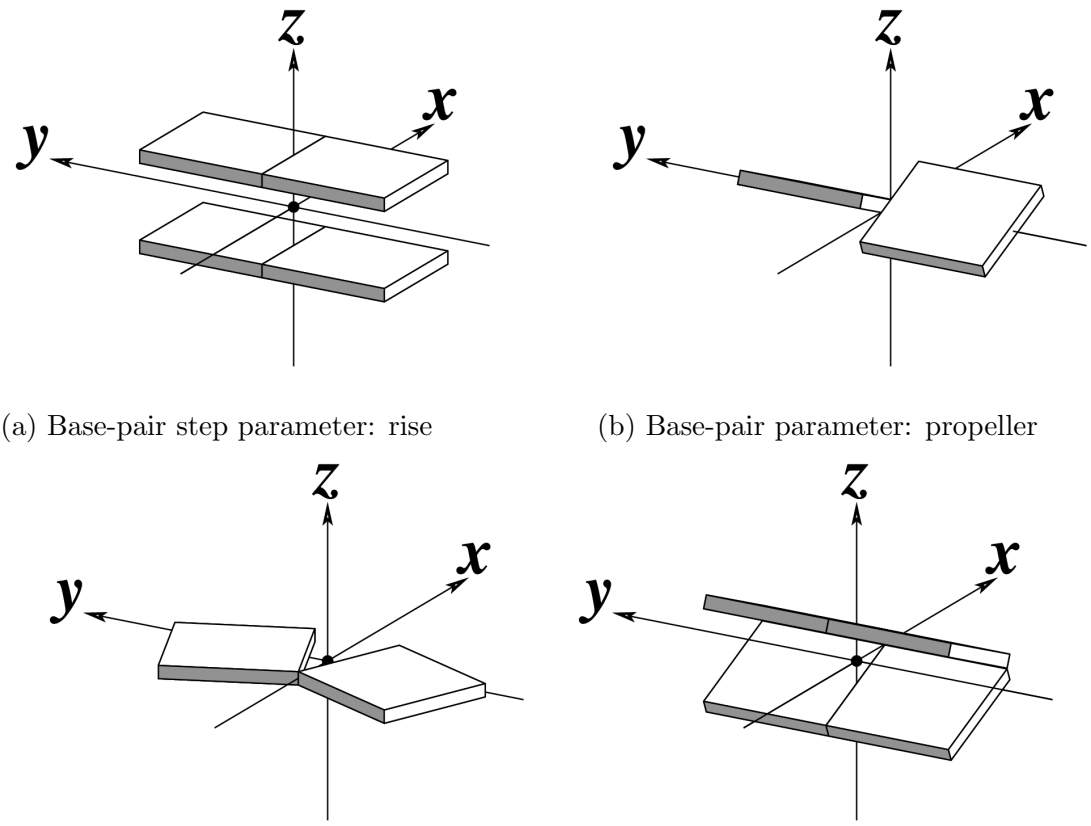
(a) An example of bidentate hydrogen bond. An arginine (Arg) residue forms two different hydrogen bonds with guanine (G).

(b) An example of bifurcarted hydrogen bond. An asparagine (Asn) residue forms two hydrogen bonds by sharing one donor atom (ND2).

Figure 7: Hydrogen bond geometries [61].

The DNA shape features, such as shear, stretch, stagger, shift, slide, rise (Figure 8a), buckle, propeller (Figure 8b), opening (Figure 8c), tilt, roll (Figure 8d), and twist, were measured using 3DNA [60]. We selected nucleotides that are in contact with the protein, plus two more flanking nucleotides on each side, and compared the distributions of the DNA shape features among the three groups of DNA-binding domains. Major and minor groove width were also calculated using 3DNA, which reports the refined P-P distances [60].

The conformational change upon DNA-binding was calculated with two approaches using the pairNR30 dataset. The first approach is to calculate the C_α RMSD (root mean square deviation) (Equation 3) between the unbound ($\mathbf{v}$) and bound ($\mathbf{w}$) conformations for a given DNA-binding protein. The RMSD is calculated by minimizing the C_α RMSD when superimposing two DNA-binding domain structures. In addition



(a) Base-pair step parameter: rise

(b) Base-pair parameter: propeller

(c) Base-pair parameter: opening

(d) Base-pair step parameter: roll

Figure 8: Parameters for describing DNA shape.

to calculating the C_α RMSD for all the residues in the DNA-binding domain, which is a useful measure to assess the overall conformational change, we also calculated the C_α RMSD in DNA-binding pocket, by selecting the binding residues in the bound conformation, using a heavy atom distance cutoff of 3.9 Å. The C_α RMSD of the binding residues can provide more detailed information of conformational adjustment for the pocket residues upon binding to DNA.

$$RMSD(\mathbf{v}, \mathbf{w}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \parallel v_i - w_i \parallel^2} \tag{3}$$

The second approach is to compare $\Delta\chi_1$, the change of side-chain torsion angle χ_1

(the torsion angle for the C_α-C_β axis) between the bound and unbound conformations. We compared the median $\Delta\chi_1$ and the median absolute deviation (MAD) of $\Delta\chi_1$ for each domain among three groups. For residue i, $\Delta\chi_1$ is calculated by Equation 4.

$$\Delta\chi_{1i} = \min(|\chi_{1i}^v - \chi_{1i}^w|, 360 - |\chi_{1i}^v - \chi_{1i}^w|) \tag{4}$$

where χ_{1i}^v and χ_{1i}^w are the χ_1 angles of residue i for the holo (v) and apo (w) structures.

The structural variations of DNA-binding domains were compared in the multiHolo, multiApo and multiApoHolo datasets based on RMSD differences. We calculated the median RMSD and MAD RMSD per DNA-binding domain, and compared the distributions among the three groups of DNA-binding proteins.

2.2.3 Statistical Tests

The Kruskal-Wallis test, a multi-sample non-parametric method, was employed to test whether there are significant differences of each of the features among the three specificity groups, HS, MS and NS. If the p-value of the Kruskal-Wallis test is lower than 0.05, we would carry out a one-sided Mann-Whitney U test, to identify the significant differences between any two of the HS, MS and NS distributions.

2.3 Results

2.3.1 Amino Acid Propensity for DNA-Binding

Arginine and lysine are the two dominant residues in overall protein-DNA contacts (18.4% and 14.9% respectively) as both are positively charged and can bind to negatively charged DNA backbone through electrostatic interactions (Figure 9). Distributions of amino acids that are in contact with DNA, including both backbone

contacts and base contacts, are similar among the three groups except for a relatively higher number of aspartate in the HS group (Figure 9). The catalytic sites in type II restriction endonucleases usually contain aspartate, which may result in the high prevalence of aspartate in the HS group (the percentage changed from 8.4% to 5.9% after removing catalytic residues, which is still higher than those in the MS and NS groups with 1.2% and 3.4%, respectively). Even though amino acid distributions are similar, majority of the residues in the NS group are involved in DNA backbone contacts, while residues in the HS and MS groups participate in more direct residue-base interactions (Figure 10).

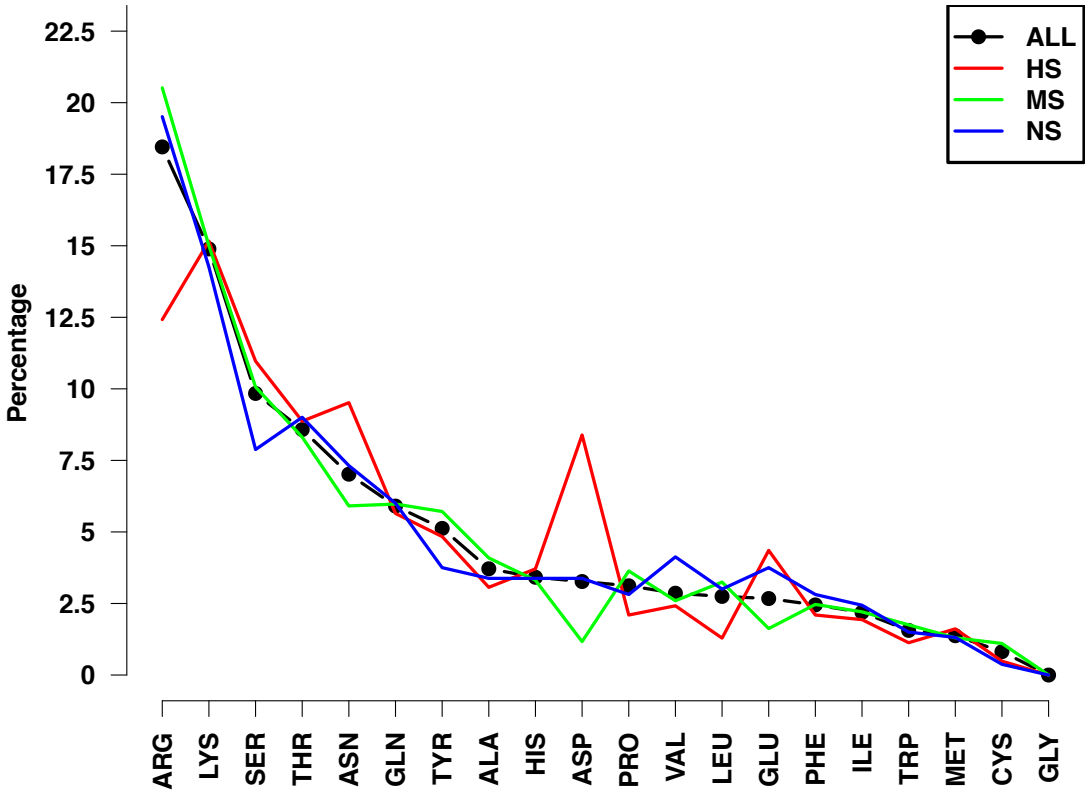


Figure 9: Amino acid distribution of binding residues in the pdNR30 dataset.

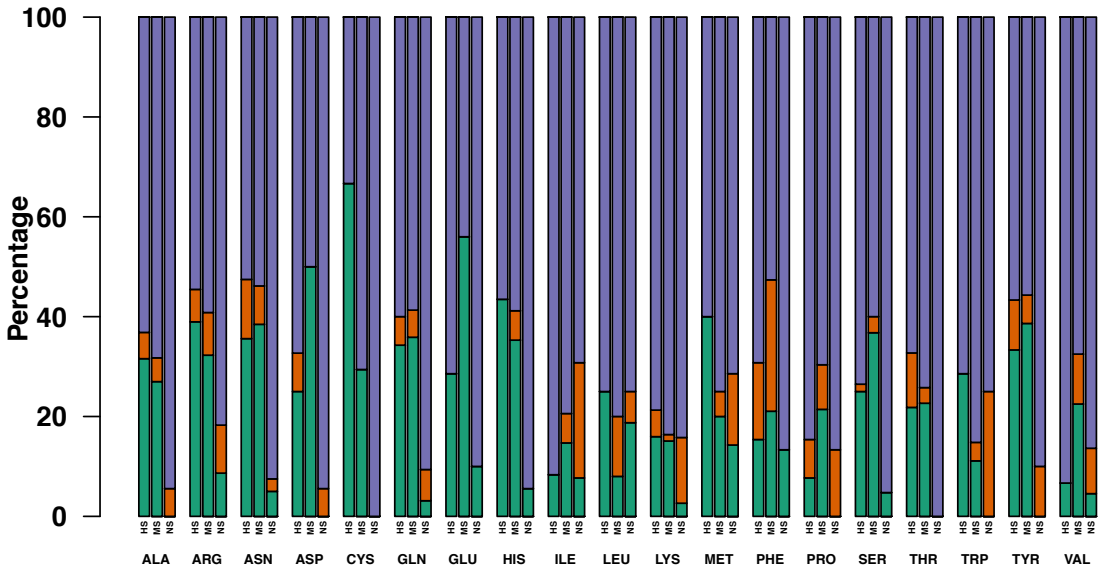


Figure 10: Comparison of DNA backbone/minor groove/major groove contacts. Percentage of DNA-backbone only (blue), minor (orange) and major (green) groove contacts per amino acid for highly specific (HS), multi-specific (MS) and non-specific (NS) DNA-binding proteins.

To study which residues are preferred in specific protein-DNA binding, we compared the residue propensities for interacting with DNA bases among the three groups (see Section 2.2). If the binding propensity of an amino acid is larger than 1, it would suggest that the amino acid is enriched in protein-DNA base interactions. Figure 11A shows that arginine is enriched in all three groups (p_{ARG} is 3.3, 3.2 and 4.8 for the HS, MS and NS groups respectively) while lysine is only highly enriched in the NS group (p_{LYS} is 1.2, 0.9 and 2.5 for the HS, MS and NS groups respectively). Both residues have higher base interacting propensities in the NS group than those in the HS and MS groups. The high propensities of DNA base contact for arginine and lysine in the NS group are rather counter intuitive. A closer look at the data suggests that we need to be careful when interpreting the high propensities of arginine and lysine in the NS group in terms of their contributions to specific protein-DNA interactions. First of

all, there are only 65 total residue-base contacts in the whole NS dataset. Among those contacts, 19 (30%) are arginine-base contacts and 12 (18%) are lysine-base contacts. Secondly, unlike the HS and MS groups, in which arginine and lysine bind predominantly in the major groove, arginine and lysine in the NS group are mainly involved in minor groove contacts (10 out of 19 for arginine and 10 out of 12 for lysine) (Figures 11B and 10). As generally accepted, minor groove contacts do not confer much specificity due to its lack of discriminative pattern for hydrogen bonds, either directly or mediated by water [89, 76], although minor groove interactions with residues may contribute to binding specificity in individual cases (more discussion later) [46].

Asparagine, glutamine, serine, and threonine, which can form hydrogen bonds with DNA bases, are enriched in the HS and MS groups, but not in the NS group, suggesting their important roles in specific protein-DNA interactions. The hydrophobic residues such as alanine, valine, proline, leucine, and isoleucine, are depleted in all cases.

The two negatively charged residues, aspartate and glutamate, have low propensities in protein-DNA base interactions except for aspartate in the HS group (p_{ASP} is 1.37, 0.38 and 0.27 for HS, MS, and NS, respectively) (Figure 11A). In general, negatively charged residues are not favourable in protein-DNA interactions due to the negatively charged DNA backbone and electronegative groups on all the bases except for cytosine [43]. In addition, unlike asparagine and glutamine that can act as both hydrogen bond acceptor and donor, aspartate and glutamate can only serve as hydrogen bond acceptors. Therefore, it is not surprising to see they are depleted

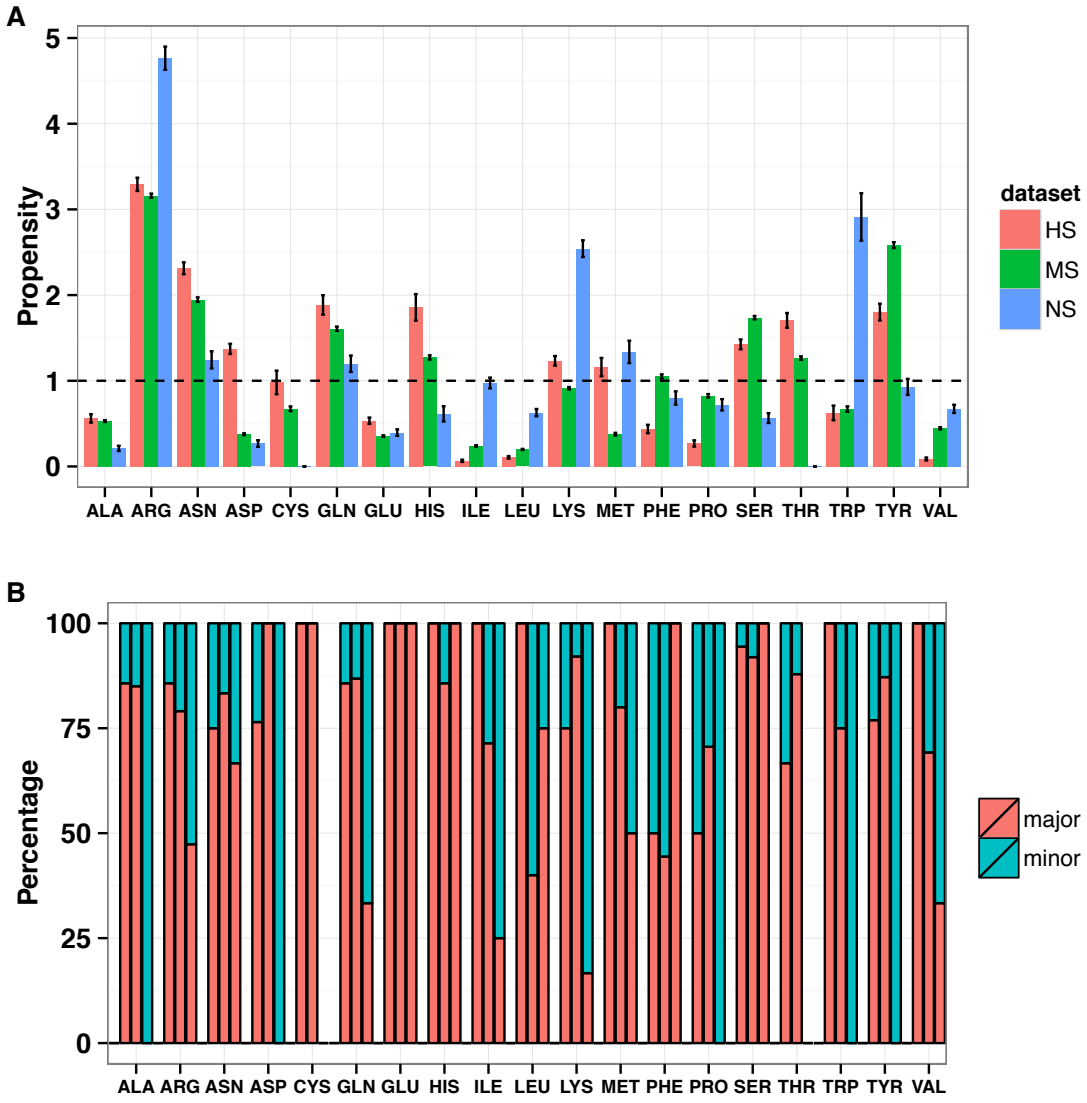


Figure 11: Residue-base contacts in protein-DNA complexes. (A) Amino acid propensities for DNA base interaction in HS (red), MS (green) and NS (blue) groups; (B) Percentage of major (red) and minor groove (cyan) contacts.

in protein-DNA base interactions in general. One interesting exception is the high enrichment of aspartate in the HS group (Figure 11A). Further analysis revealed a striking pattern as shown in Table 3. All the aspartate residues that contact DNA bases are involved in hydrogen bonding with major groove atoms in the highly specific DNA-binding domains. Out of the19 hydrogen bonds, 18 participate in hydrogen

bonding with a cytosine. Though aspartate and glutamate have very low propensities in the MS group and NS group, their major groove contacts are primarily with a cytosine as well. While both cytosine and adenine have one hydrogen bond donor in the major groove, adenine has an electronegative surface, making it unfavourable for interacting with aspartate when compared to cytosine (Figure 12). In the minor groove, except for one case, all other aspartates and glutamates form hydrogen bonds with a guanine, which is not surprising since only guanine can serve as a hydrogen bond donor in the minor groove (Table 3, S6 and S7). More importantly, for aspartate-cytosine specific interactions in the HS group, aspartate form bidentate hydrogen bonds in 5 cases (accounting for 10 of the 19 total atom-level hydrogen bonds) with two consecutive cytosines (Figure 13 and Table S6). The stereochemical properties and hydrogen bond patterns of DNA bases and aspartate make the aspartate-cytosine very specific (Figure 13). There are no bidentate hydrogen bonds for glutamate found in our non-redundant dataset. However, Ecl18kI (PDB ID: 2fqz with a recognition sequence ˆCCNGG), not included in the dataset due to similarity with other enzymes, has a bidentate hydrogen bond between residue Glu187 and two consecutive cytosines [14]. In general, aspartate is preferred over glutamate, probably due to the shorter side-chain of aspartate. The observation of the specific hydrogen bonding between aspartate and glutamate may explain why both amino acids are rarely seen in the MS and NS groups as most transcription factors allow variations at different sites and non-specific binding proteins are not sequence-specific.

Another interesting observation is the high enrichment of two aromatic residues, histidine and tyrosine in the HS ($p_{HIS} = 1.9$, $p_{TYR} = 1.8$) and MS group ($p_{HIS} = 1.3$,

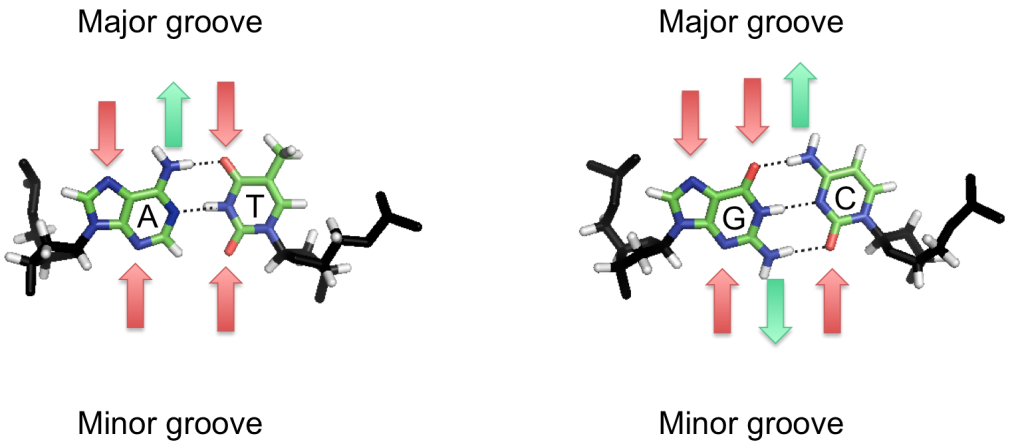


Figure 12: Diagram of hydrogen bond signatures in the DNA major and minor grooves. Red arrows point towards acceptor atoms, and green arrows point away from donor atoms.

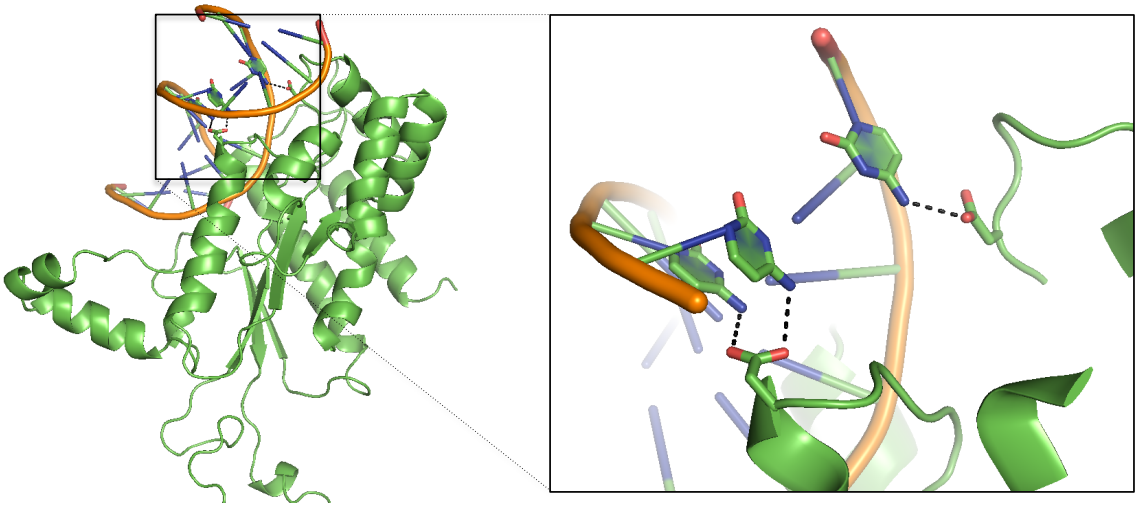


Figure 13: Aspartate forms one bidentate hydrogen bond with two consecutive cytosine bases and one single hydrogen bond with a distant cytosine, via the major groove, in endonuclease NgoMIV (PDB ID: 4abt).

$p_{TYR} = 2.6$), but not in the NS group ($p_{HIS} = 0.6$, $p_{TYR} = 0.9$). But histidine and tyrosine may contribute to specific DNA-binding using different mechanisms. Histidine residues in the HS and MS groups primarily forms hydrogen bonds with guanine (Table S8). The difference between these two groups is that 9 of the 10

histidine-base contacts in the HS group form hydrogen bonds while only half of the histidine-base contacts in the MS group are involved in hydrogen bonding. As for tyrosine, only a small percentage of the base contacts participate in hydrogen bonding (data not shown), suggesting that unlike histidine, hydrogen bonding does not play a major role in specific-protein-DNA binding for tyrosine. Previous studies have shown the importance of aromatic residues and π-π-interactions in protein-DNA complexes [113, 74]. π-interactions occur when the negatively charged electron cloud of an aromatic compound interacts with positively charged atoms or cations [40]. While π-interactions are generally thought to add stability and affinity to macromolecule interactions [113, 74], more recent studies have suggested that aromatic residues may play a major role in determining binding specificity in molecular recognition, such as interaction between carbohydrates and proteins [3]. Wilson *et al.* recently investigated the abundance, structure and strength of π-interactions between aromatic residues and DNA bases and demonstrated that protein-DNA π-interactions are more prevalent than previously thought [113, 112, 7]. Yet, very little is known about the critical role of aromatic-base π-interactions in protein-DNA binding specificity [113, 112, 7]. Our results suggest that tyrosine may play more important roles in conferring specific protein-DNA interactions through π-interactions due to its high propensities in the HS and MS but low propensity in the NS group, and scant of hydrogen bonds. Tryptophan has low occurrences with two residues in each of the three groups. Therefore the high propensity of tryptophan in the NS group is not conclusive due to the small sample size. Moreover, both tryptophan residues in the NS group interact with the minor groove of the DNA (Figure 11B). As for phenylalanine, about 50% of the base

contacts are in the minor groove in both HS and MS, therefore, it is not clear how much contribution it provides for specific protein-DNA interactions.

2.3.2 Interaction Interface

Comparison of interaction surface among the three groups shows a similar trend to their degree of binding specificity. The HS group has the largest protein-DNA contact area (PDCA) while the NS group has the smallest contact area (one-sided two-sample p-values < 0.0002) (Figure 14A). Since interaction surface represents the total contact area between protein and DNA, a combination of both non-specific and specific interactions, we also compared the number of residue-base contact (NRBC) [50], which captures more of specific interactions. Results show a similar decreasing trend for NRBCs to PDCA as the protein-DNA binding specificity decreases (one-sided two-sample p-values < 0.0005) (Figure 14B). In terms of number of residue-base contacts per Å^2 (NRBC density), we found that HS and MS groups have similar NRBC density, while the NS group has much lower NRBC density (one-sided two-sample p-values $< 2 \times 10^{-9}$) (Figure 14C).

The percentage of DNA base contact is much higher in the HS and MS groups than that in the NS group since the contacts between amino acids and DNA-backbone atoms are mainly non-specific (Figure 15A). We also compared the major and minor groove contacts, as major groove contacts represent primary contribution to binding specificity due to the sequence-specific patterns for hydrogen bonds in the major groove. The percentage of major groove contact in the HS and MS groups (81.1% and 82.3% respectively) is more than twice the number in the NS group (35.4%)

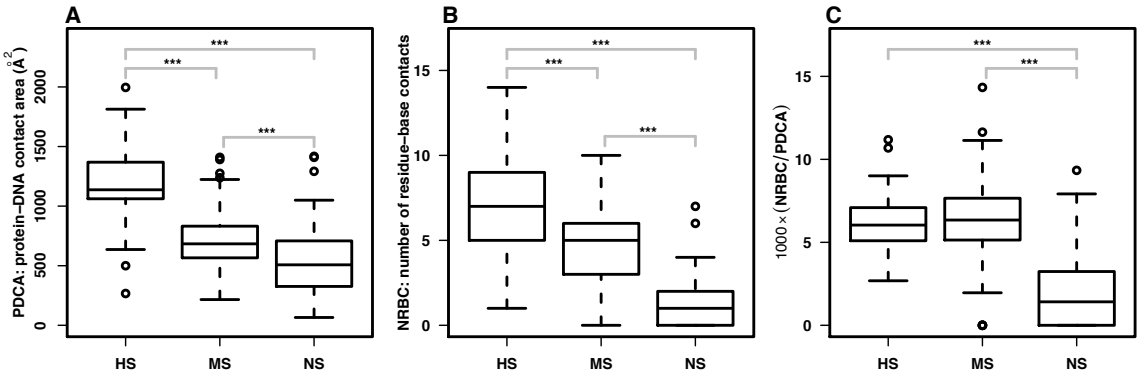


Figure 14: Comparison of protein-DNA interactions. (A) Protein-DNA contact area (PDCA); (B) number of residue-base contacts (NRBC); and (C) NRBC density, NRBC normalized to the total contact area (PDCA). *** for $p < 0.001$; ** for $p < 0.01$; * for $p < 0.05$.

(Figure 15B). In terms of the number of major groove contacts, we observed a clear trend similar to the binding specificity. The HS and MS groups have significantly higher number of major groove contacts than that in the NS group (one-sided two-sample p-values $< 9 \times 10^{-12}$) (Figure 15C). The difference between the HS and MS groups is also significant (one-sided two-sample p-values < 0.005).

The number of minor groove contacts does not have a trend as that in the major groove contacts. Interestingly, there is a statistically significant difference between the HS group and MS/NS groups with HS group having more minor groove contacts (Figure 15D). Even though minor groove contacts are generally considered non-specific, it has been demonstrated that minor groove contacts can contribute to protein-DNA binding specificity. Joshi R *et al.* previously reported that the functional specificity of a Hox protein is mediated by minor groove contacts [46]. More specifically, the minor groove contacts are a result of sequence-dependent DNA shape recognition. It has been reported that the minor groove shape, which deviates from the canonical B-type DNA structure, also plays a role in sequence specific recognition for BsoBI

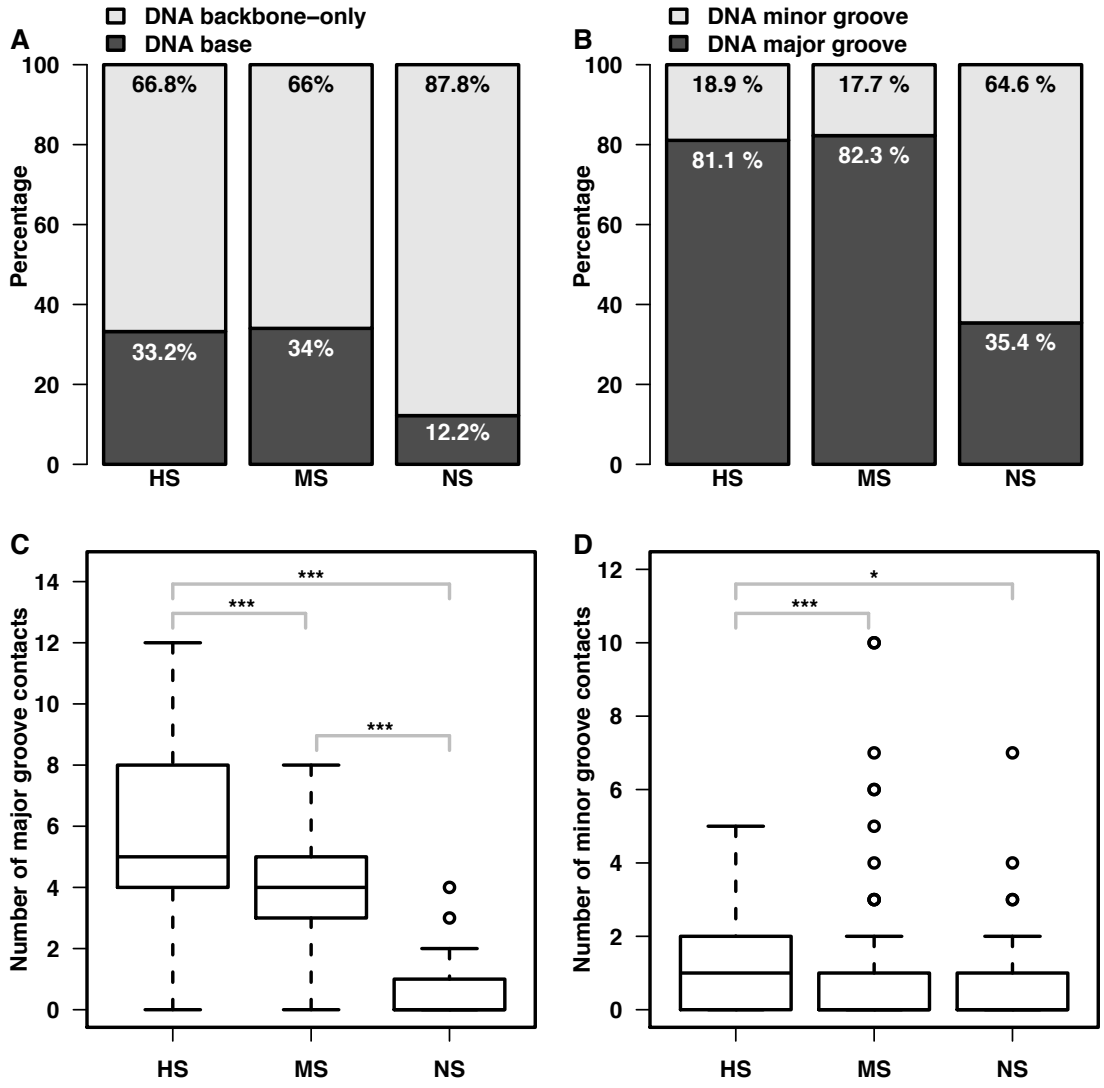


Figure 15: Comparison of DNA base/backbone and major/minor groove contacts. (A) Percentage of DNA backbone-only and DNA base contacts; (B) percentage of major and minor groove contacts; (C) number of major groove contacts; and (D) number of minor groove contacts. *** for $p < 0.001$; ** for $p < 0.01$; * for $p < 0.05$.

endonuclease [114]. Therefore, the relatively large number of minor groove contacts in the HS group may be the result of DNA shape (discussed in next section). Taken together, the HS and MS groups have similar ratios of residue-base contacts and similar percentages of DNA base and major groove contacts, which are significantly larger than those in the NS group. Between the HS and MS groups, HS has larger contact

areas and higher number of DNA base and major groove contacts than those in the MS group, which is consistent with the previous study that shows larger interface in the restriction endonuclease superfamily than the transcription factor superfamilies [18].

Hydrogen bonds have been considered a major factor in protein-DNA binding specificity [61]. Our analysis shows the decreasing pattern from the HS group to the NS group in terms of the total number of hydrogen bonds between protein and DNA (one-sided two-sample p-values < 0.0003) (Figure 16A) as well as between protein and DNA bases (one-sided two-sample p-values $< 8 \times 10^{-9}$) (Figure 16B). While the formation of hydrogen bonds is important for specific protein-DNA binding, the geometry of the hydrogen bonds can also help discern specific and non-specific interactions. The number of bidentate hydrogen bonds between protein and DNA also shows the same trend as the degree of binding specificity (one-sided two-sample p-values $< 2 \times 10^{-3}$) (Figure 16C).

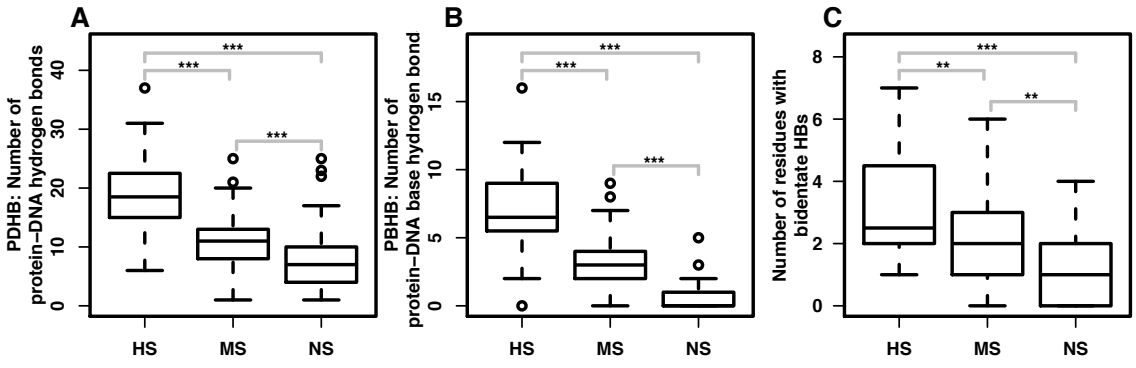


Figure 16: Hydrogen bonds between protein and DNA. (A) Number of hydrogen bonds between protein side-chains and DNA (PDHB); (B) number of hydrogen bonds between protein side-chains and DNA bases (PBHB); and (C) number of residues that form bidentate hydrogen bonds. *** for $p < 0.001$; ** for $p < 0.01$; * for $p < 0.05$.

2.3.3 DNA Shape

To compare the DNA shape in protein-DNA complexes, we used the program 3DNA to derive a number of structural features, including shear, stretch, stagger, buckle, propeller, opening, shift, slide, rise, tilt, roll and twist [60]. We computed the median values for each domain, using only the nucleotides that are in contact with the protein and two flanking bases on each side, and compared the distributions among the three groups. The results show that the median values in each DNA for propeller, opening, rise and roll have significant differences among the HS, MS, and NS groups (Kruskal-Wallis test p-values < 0.02) (Figure 17). Further analysis using Mann-Whitney U test shows that the DNA-binding domains in the HS group have larger propeller (one-sided two-sample p-values < 0.02) and rise (one-sided two-sample p-values < 0.002) median values, and lower opening (one-sided two-sample p-values < 0.05) and roll (one-sided two-sample p-values < 0.004) median values than the MS and NS groups. We also compared the distributions of these four features by pooling all the data within each of the HS, MS, and NS groups and found similar significant differences (data not shown). These results indicate that the HS group has distinct shape features when compared with the other two groups, suggesting a key role of these shape features in the high binding specificity. These shape differences may also explain the number of minor groove contacts in the HS group. The high propeller and rise may make the minor groove more accessible to residues and offer more distinctive patterns for different DNA sequences, thus contributing more to binding specificity.

We also looked at the major and minor groove width of nucleotides in contact with

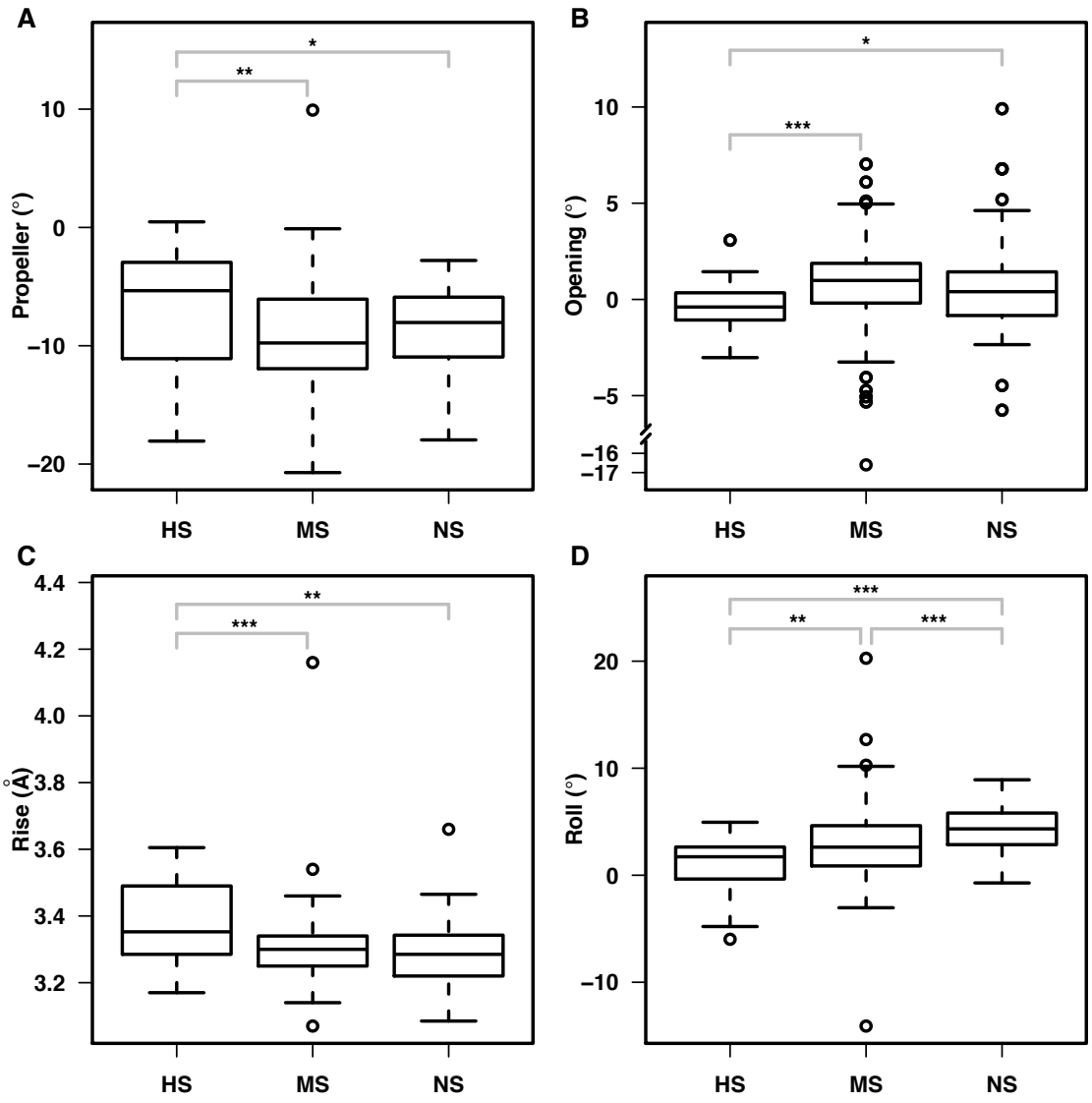


Figure 17: Comparison of DNA shape features. Median (A) propeller, (B) opening, (C) rise, and (D) roll per structure. The shape features are calculated using 3DNA 69. *** for $p < 0.001$; ** for $p < 0.01$; * for $p < 0.05$.

the protein (+2 flanking bases on each side) using 3DNA by comparing the minimum, average, and maximum width for each DNA structure in the pdNR30 dataset. Our analysis shows that there is a similar pattern to the binding specificity in terms of the major groove width, where HS has the highest width, no matter which metric is used (Figure 18A-C). As for the minor groove width, the DNA structures in complex

with highly specific DNA-binding domains have wider minor grooves than those in the multi-specific and non-specific DNA-binding domains (one-sided two sample p-values < 0.05) with the MS group having the smallest minor groove width (Figure 18D-F). This is in part consistent with the observation by Contreras-Moreira *et al.* that restriction endonuclease have a larger proportion of indirectly readout bases [18]. Our data confirms the importance of DNA shape in specific protein-DNA interactions [91, 89].

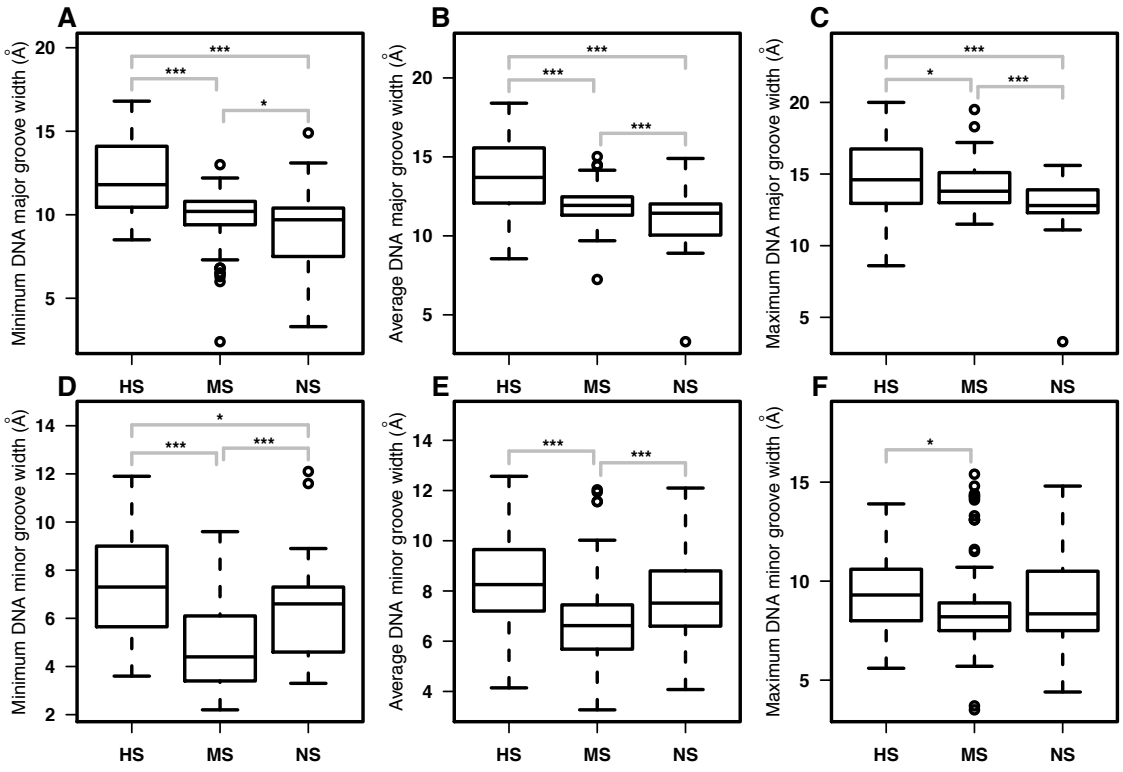


Figure 18: Comparison of DNA major and minor groove width () of protein-contacting DNA bases. Minimum (A), average (B) and maximum (C) major groove width per domain. Minimum (D), average (E) and maximum (F) minor groove width per domain.

2.3.4 Conformational Changes Upon DNA-Binding

We calculated the conformational changes in terms of C_α RMSD of all the residues (Figure 19A) and the residues that are in contact with DNA base (Figure 19B) using pairNR30, a non-redundant dataset of bound/unbound DNA-binding domains. Besides C_α RMSD, which indicates the backbone conformational changes, we also looked at side-chain conformational changes of the binding residues based on χ_1 dihedral angle changes $\Delta\chi_1$, including the distribution of the median $\Delta\chi_1$ per domain (Figure 19C), and the MAD of $\Delta\chi_1$, which shows variances of $\Delta\chi_1$ (Figure 19D). The conformational changes based on RMSDs show that changes are higher in the highly specific group (Figure 19A and 19B). Statistical analysis revealed that the domains in the HS group have significantly higher C_α RMSD for all residues (p-values < 0.02) and DNA-base contacting residues (p-value < 0.004). There is no significant difference between the MS group and the NS group. As for the χ_1 dihedral angle changes, though the median values for the HS group are larger than those in the MS and NS group, the differences are not statistically significant (Figure 19C and 19D). The $\Delta\chi_1$ distributions for all DNA binding residues among the three groups were also compared, but no statistical significant differences were found (data not shown).

Our results suggest that the DNA-binding proteins with higher degree of binding specificity tend to have more conformational changes compared to the non-specific DNA binding proteins and transcription factors. Since the protein-DNA interaction interface for the highly specific proteins is larger, these proteins require backbone flexibility in order to have a precise interface fit for high specificity [48].

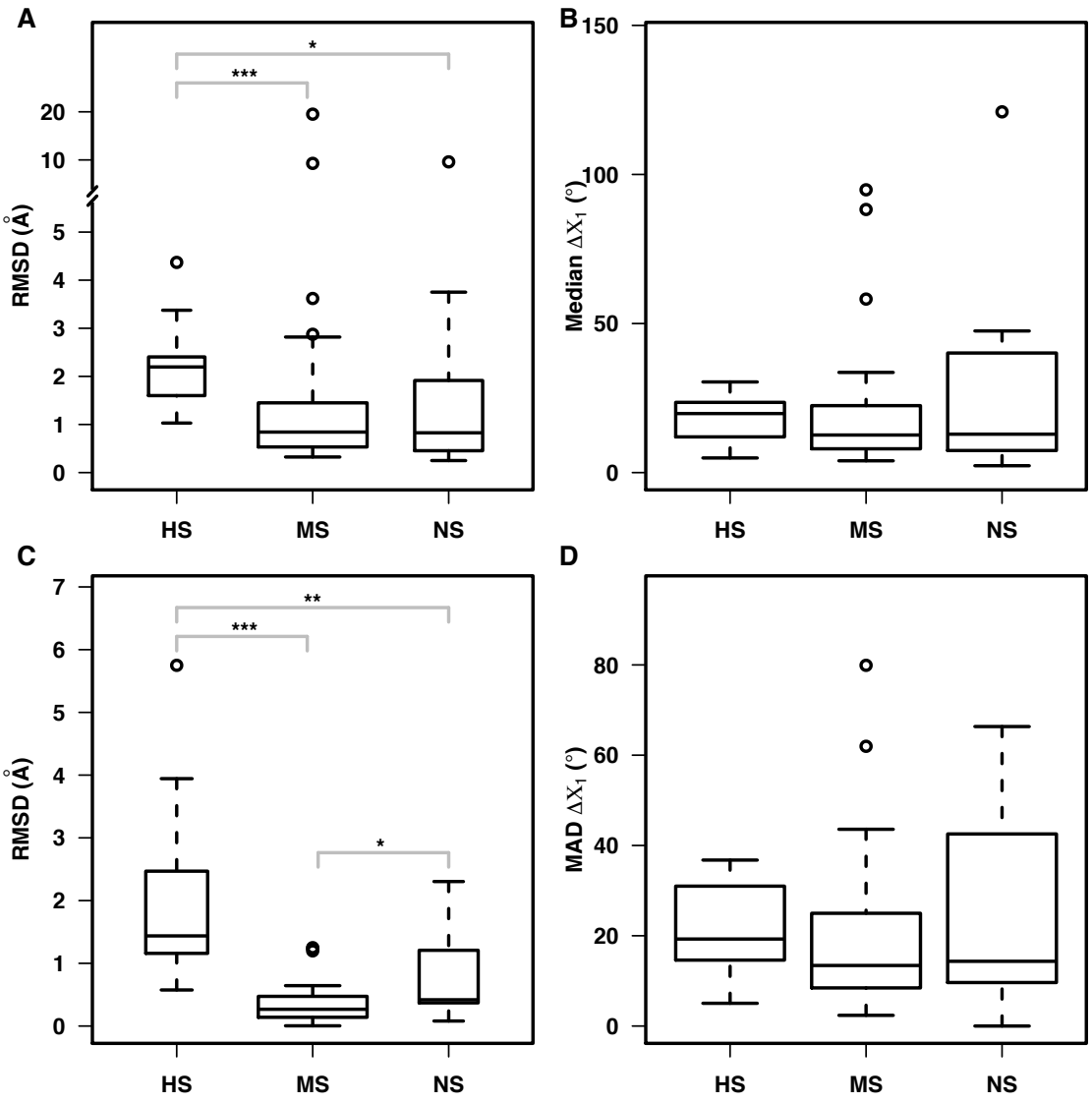


Figure 19: Conformational changes upon DNA-binding. C_α RMSD between the bound and unbound structures in the pairNR30 dataset using (A) all residues and (B) binding residues only. Median $\Delta\chi_1$ (C) and MAD $\Delta\chi_1$ (D) per domain. *** for $p < 0.001$; ** for $p < 0.01$; * for $p < 0.05$.

2.3.5 Structural Variations of DNA-Binding Domains

In addition to studying structural differences between the bound and unbound structures, another way to explore the role of protein flexibility and dynamics to DNA binding specificity is to compare the conformational diversity of DNA-binding proteins

in free state and bound form. Our analysis on three datasets multiHolo, multiApo and multiApoHolo revealed that highly specific and multi-specific DNA-binding domains have a larger range of structural variations in both the bound (Figure 20A and 20B) and free forms (Figure 20C and 20D), when compared to the non-specific DNA-binding domains. The plots based on the multiApoHolo set also show that the NS group has smaller structural variations in terms of median and MAD RMSD than those in the HS and MS groups (Figure 20E and 20F). These results suggest that the flexibility of DNA-binding proteins may contribute to their higher degree of binding specificity, which is consistent with previous findings using different metrics [2].

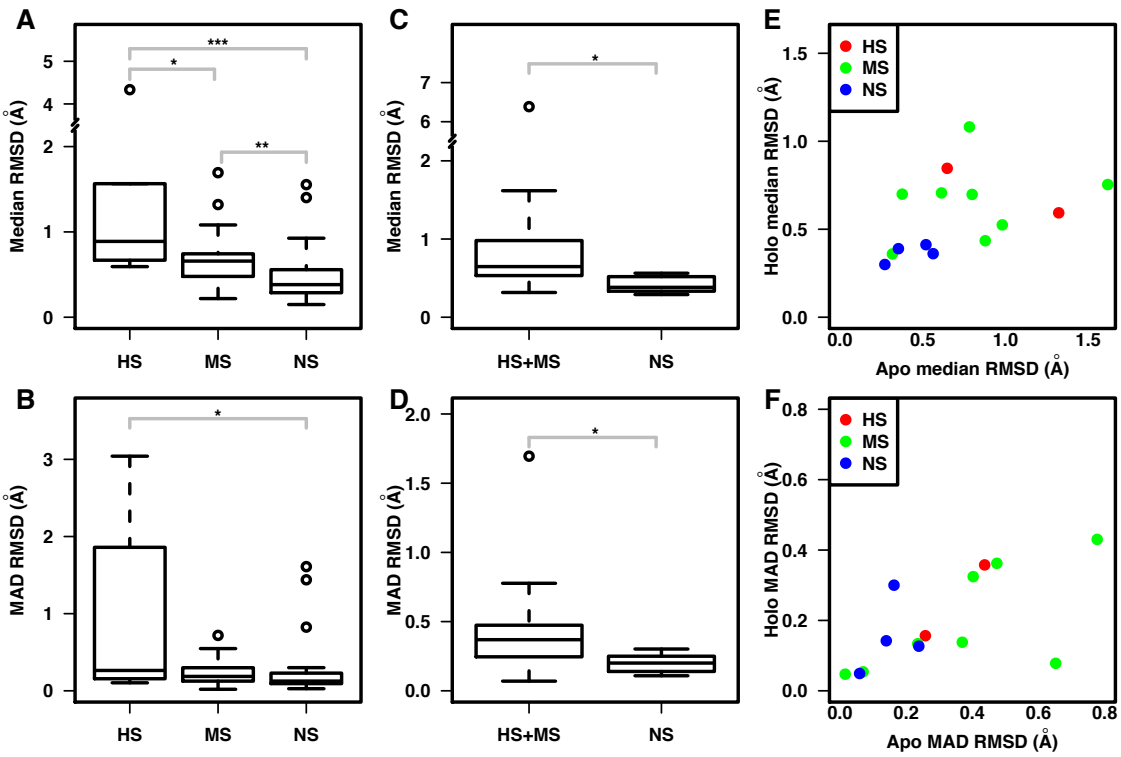


Figure 20: Structural variations in the multiHolo dataset in terms of median RMSD (A) and MAD RMSD (B). Structural variations in the multiApo dataset in terms of median RMSD (C) and MAD RMSD (D). Structural variations in the unbound and bound states of the multiApoHolo dataset in terms of median RMSD (E) and MAD RMSD (F). *** for $p < 0.001$; ** for $p < 0.01$; * for $p < 0.05$.

2.4 Discussion

Knowledge of the structural basis of binding specificity is central to our understanding of protein-DNA interactions, and the evolution and divergence of protein-DNA binding specificity [8]. Such knowledge is also essential to practical applications in rational design of new proteins with novel binding specificity in biotechnology and medicine [5, 107, 87, 109]. Our comparative analyses show a clear trend in terms of both static and dynamic structural features with the degree of protein-DNA binding specificity.

Arginine and lysine have been known to be abundant in protein-DNA interfaces (Figure 9). Though both arginine and lysine can form multiple types of hydrogen bonds with DNA [61], which is a key factor in specific protein-DNA interactions, they also represent two major residues for non-specific interaction between their positively charged side-chains and the negatively charged DNA backbone. In NS group, majority of arginine and lysine residues interact with the DNA backbone. For the residues in the NS group that interact with the DNA bases, the contacts occur primarily in the minor groove. Both the non-specific and specific interactions of arginine and lysine may work together to achieve high specificity in the process of protein-DNA recognition. For specific DNA-binding proteins, the non-specific interactions between arginine/lysine and DNA backbone or minor groove can help search for the target sites very quickly via non-specific electrostatic interactions [48]. Once the target sites are identified, the hydrogen bonds can contribute to sequence-specificity through specific residue-base hydrogen bonding in the major groove.

One important finding from our analysis is the high enrichment of aspartate-base contacts in the group of highly specific (HS) DNA-binding domains. Aspartate is a negatively charged residue and its side-chain atoms can only serve as hydrogen bond acceptor, which makes it unfavourable to interact with DNA due to the negatively charged backbone and electronegative surface, except for cytosine. As such, aspartate interacts with cytosine with high specificity, especially with two consecutive cytosine bases through bidentate hydrogen bonds, as aspartate has two hydrogen bond acceptors (Figure 13). It may also explain why aspartate is rarely seen in DNA base contacts in the MS and NS groups since DNA-binding proteins both groups allow variations to different degrees. In case studies, Jantz and Berg used designed zinc finger proteins and showed that when a residue in one of the fingers is changed from asparagine to aspartate, though the overall affinity decreased, the contacting base changed from adenine to cytosine with higher specificity [43]. Pingoud *et al.* studied SsoII and the evolutionary relationship between different subgroups related to this protein and found that Glu187 in SsoII is highly conserved when aligned to several other restriction enzymes, which can be either an aspartate or a glutamate [86]. To our knowledge, our comparative analysis is the first large-scale study to show the specific recognition of cytosine by aspartate.

Histidine and tyrosine appear to be enriched in highly specific and multi-specific DNA-binding proteins. In addition to their capability to form hydrogen bonds with bases, both aromatic residues can contribute to protein-DNA binding through π-interactions. Our data revealed that histidine contribute to specific DNA binding primarily through hydrogen bonding with guanines while tyrosine uses π-interactions

to achieve the binding specificity. Recent studies demonstrated that π-interactions are more prevalent in protein-DNA recognition than previously thought [113, 112, 7]. However, the role of π-interactions in specific protein-DNA recognition is still not clear. Our data suggest that these two aromatic residues play key roles in specific protein-DNA binding through hydrogen bonds and π-interactions. Based on these results, we have developed an integrative energy function that adds two atomic-level terms, π-interaction energy and hydrogen bond energy, to a knowledge-based multi-body potential for structure-based prediction of transcription factor binding sites. Our results showed that incorporating π-interaction and hydrogen bond energy greatly improved the prediction accuracy of transcription factor binding sites [59, 25].

Not surprisingly, our data show that there are significantly larger base/backbone and major/minor groove contact ratios for DNA-binding proteins in the HS and MS groups when compared to the non-specific DNA-binding proteins. While the contact ratios and density are similar between HS and MS proteins, the total contact number and interaction interface in HS proteins are larger than those in the MS group (Figures 14 and 15). This is consistent with previous results by Contreras-Moreira *et al.* [18]. Similarly, the number of simple and complex hydrogen bonds is another key contributing factor for the degree of DNA-binding specificity (Figure 16).

Since DNA shape has been implicated in protein-DNA binding specificity [91, 89, 90], we also looked for any shape differences among three groups by systematic analysis. However, comparison of the shape features is not as straightforward as examinations of the contact features since there are local and global shape features. Nevertheless, our results showed that the highly specific DNA-binding domains have larger

rise between bases, something that can contribute to more base contacts since the bases can be more exposed [18]. The results on opening, propeller and roll parameters as well as the major and minor groove width are also statistically significant. These differences may be a result of the flexibility of both protein and DNA, which help binding specificity through fitting and fine-tuning to achieve optimal interactions.

In addition to the "static" protein-DNA contact features and the difference in DNA shape, we investigated the dynamic structural features in DNA-binding domains. Currently, there are two widely accepted models for macromolecular recognition, induced-fit and conformational selection [20]. We compared both the conformational changes after DNA-binding (mimicking the induced-fit model) and structural variations of each protein (mimicking the conformational selection model). Based on a limited number of cases in the datasets, we showed that the highly specific and multi-specific DNA-binding domains have larger degree of flexibility in the bound and unbound states, and larger conformational change upon DNA-binding. This is in accordance with the hypothesis that specific DNA-binding proteins need to explore different conformations in order to optimize their binding to the target DNA recognition sites [98, 120], whereas non-specific DNA-binding proteins are not required to explore that many conformations in the process [76]. The flexibility involved in the specific protein-DNA binding process could be a combination of structural variations and induced conformational changes upon binding for both protein and DNA. For example, a very recent work by Chen and Pettitt showed that the flexibility of a specific DNA sequence is about 40% intrinsic and 60% induced while no appreciable non-specific DNA bending is induced [15].

Table 3: Number of hydrogen bonds between DNA base and aspartate (Asp) or glutamate (Glu). In parenthesis, it shows the number of bases that are hydrogen bonded with aspartate or glutamate. For example, there are 19 hydrogen bonds between aspartate and DNA major groove atoms in the highly specific DNA-binding domains with 18 interacting with cytosine (C) and 1 with guanine (G).

Amino Acid	Asp		Glu	
Groove / Dataset	Major	Minor	Major	Minor
HS	19 (18C, 1G)	5 (5G)	2 (2C)	1 (1G)
MS	4 (4C)	0	10 (7C, 3A)	0
NS	0	1 (1G)	0	1 (1T)

In conclusion, protein-DNA recognition is a complex mechanism that can be dissected in terms of static and dynamic structural features that contribute to the degrees of binding specificity. Not only does the knowledge help us better understand the possible mechanisms of specific protein-DNA interactions, these features can also be used to assess the quality of protein-DNA docking predictions.

CHAPTER 3: ASSESSMENT OF PROTEIN-DNA DOCKING PREDICTIONS

3.1 Introduction

DNA-binding proteins play crucial roles in many biological processes. The mechanism of protein-DNA recognition, despite of decades of efforts, is still not fully understood. Protein-DNA complex structures can provide an insight into the molecular mechanisms of DNA recognition and be used as a starting point for structure-based transcription factor (TF) binding site prediction. Although the number of experimentally determined structures in the PDB [9] increases at a higher rate, only a small percentage of them (≈3%) are proteins in complex with DNA.

Computational docking between a protein and DNA, on the other hand, has been considered as a cost-efficient alternative to fill the void in the complex structure landscape. More importantly, it has great potentials in computer-aided drug design. Over the last two decades, several protein-DNA docking algorithms have been developed [52, 21, 50, 106]. They use energy functions, either knowledge- or physics-based, to guide the docking process and ultimately select a protein-DNA complex with the lowest energy. There are two major types of docking methods, rigid and flexible docking algorithms [50]. Rigid docking algorithms do not change the initial conformation of the protein and DNA molecules, they only change the relative position of the protein with respect to the DNA. Flexible docking algorithms consider the conformational

changes of protein and DNA, besides changing the relative positions between protein and DNA. Rigid docking methods are useful to test the validity of energy functions and serve as a staring point to develop flexible docking algorithms. The accuracy of a method is usually reported as the percent of cases to which the algorithm selected a good prediction (in terms of the root mean square deviation (RMSD) of the predicted and the native structure).

Takeda *et al.* [103] developed a novel residue-level, knowledge-based potential and applied it to benchmark dataset of 38 transcription factor-DNA complexes using a rigid-docking algorithm. The algorithm is based on Monte Carlo (MC) simulations. Usually for one protein-DNA docking prediction, 200 independent MC simulations are carried out to increase the coverage of the sampling space. The one with the lowest energy of the 200 simulated complex structures is selected as the predicted model. In general, it is considered a good prediction if the model is within 3 Å of the native structure. The method has a reported accuracy of 55% (21 successful cases out of 38 protein-DNA complexes). However, there are two issues with the current prediction method. First, there are predicted near native structures ($RMSD_{nat,pred} \leq 3$Å) in 13% of the cases (5 out of 38), but these complexes have higher energy, as such, they are not selected as the predicted models. We call these false negative complexes. Secondly, in 32% of the cases (12 out of 38 total cases), the docking algorithm could not produce any good predictions. However, the program will always select the lowest energy complex structure as a predicted model. This is a problem related to false positives.

Model quality estimation is an essential component of protein-DNA docking pre-

dictions, as the accuracy of a model will affect its usefulness for practical applications [13]. To determine the accuracy of predicted complexes when the native complex is unavailable still remains an open problem. The ability to evaluate the quality of protein-DNA docking predictions is urgently needed. Previous docking methods have relied merely on energy scores to rank the predictions [22, 103], however, energy scores have failed to identify the correct from the incorrect solutions [99]. The energy scores used for docking prediction are generally designed to be fast, due to the amount of conformations the algorithms have to explore, and they are not accurate enough for the specific task of identifying good/bad predictions.

In this study, we present a learning model to evaluate the quality of protein-DNA docking solutions. The score indicates the probability of the protein-DNA complex to be a native or near-native structure and is a useful indicator of the quality of the prediction. The goal is to improve the protein-DNA docking prediction and to provide the level of confidence of the prediction, *i.e.*, it will select near native structures if available, and discard bad predictions when the docking algorithm could not produce any near-native structures.

3.2 Materials and Methods

To achieve the above goal, we developed a computational model by training a binary classifier with positive (good predictions) and negative (bad predictions) samples. The good and bad predictions were obtained as follows. Based on a training dataset of 160 native protein-DNA complex structures, we generated 64,000 (400 each) protein-DNA predictions using a rigid-docking algorithm with the orientation

potential [103] and the multi-body potential [59]. Each predicted model is either assigned as positive (or good) if the $RMSD_{nat,pred} \leq 3$Å or negative (or bad) to the remaining predictions. Since only 7.6% of the predictions are near-native structures, we implemented a balanced sampling strategy known as hard negative mining in order to have an unbiased model.

After training a model, it is used to evaluate the quality of the predictions using a testing dataset developed as a benchmark for protein-DNA docking algorithms. We used the Matthews correlation coefficient to assess the quality of the model as binary classifier, and also compare the performance of the docking algorithm with and without the trained model.

3.2.1 The Scoring Function

Features

We applied two groups of features for the model. One consists of protein-DNA interaction energies and the other contains static structural features as described in Chapter 2. The first group includes three knowledge-based energy functions, the multi-body potential (`energyMB`), the orientation potential (`energyOR`), and DDNA3 (`ddna3`). The multi-body potential [59] is a knowledge-based, residue-level potential originally developed to guide the search of a TF-DNA docking algorithm. It was later replaced by the orientation potential [103], another knowledge-based, residue-level potential that takes into account not only the proximity of the nucleotides respect to the protein side-chains, but also the angles between bases and amino acid side-chains. DDNA3 [119] is an atom-level potential that estimates the interaction energy

in a protein-DNA complex.

The static structural features used for this study are: interface size (`pdca`), number of residue-base contacts (NRBC) divided in major (`sc.major`) and minor groove contacts (`sc.minor`), side-chain/DNA backbone only contacts (`sc.bb`), protein backbone-DNA contacts (`bb`), protein-DNA hydrogen bonds (`pdhb`), protein-DNA base hydrogen bonds (`pbhb`), bidentate hydrogen bonds (`bidentateHB`), bifurcated hydrogen bonds (`bifurcatedHB`) and single hydrogen bonds(`singleHB`).

The interface size (PDCA: protein-DNA contact area) was determined by calculating the difference in solvent accessible surface area ($SASA$) between the individual protein ($SASA_{protein}$), the DNA structure ($SASA_{dna}$) and the corresponding protein-DNA complex structure ($SASA_{complex}$) (Equation 2). The solvent accessible surface areas were measured by Naccess [39] v.2.1.1 with default parameters.

Protein-DNA contacts were identified using a distance cutoff of 3.9Å between protein heavy atoms and DNA heavy atoms. The DNA-contacting residues were divided into four non-overlapping sets according to the following hierarchy: (i) `sc.major`, residues that have at least one contact between side-chain atoms and DNA major groove; (ii) `sc.minor`, residues that have at least one contact between side-chain atoms and DNA minor groove; (iii) `sc.bb`, residues that have at least one contact between side-chain atoms and DNA backbone; and (iv) `bb`, all other DNA contacting residues.

Hydrogen bonds (HBs) in protein-DNA complexes were identified with HBPLUS [71] v.3.06. `pdhb` (protein-DNA hydrogen bonds) is the total number of hydrogen bonds between any protein atom and any DNA atom, and `pbhb` (protein-base hy-

drogen bonds) is the subset of hydrogen bonds between any protein atom and DNA base atoms. We also counted residues forming protein-DNA hydrogen bonds by HB geometry: (i) `bidentateHB`, residues that form at least two hydrogen bond with different acceptor and donor atoms (Figure 7a); (ii) `bifurcatedHB`, residues that form two hydrogen bonds by sharing one atom (Figure 7b); and (iii) `simpleHB`, residues that form only one hydrogen bond.

RBF Kernel SVM

Support vector machines (SVM) are supervised learning methods, widely used to train binary classifiers in computational biology [11, 12, 53]. In this case, we trained a non-linear SVM model using the radial basis function (RBF) kernel, with parameter $\gamma = \frac{1}{n_c}$, where n_c (=13) is the number of features selected to develop the model. Platt scaling was used to transform the binary classifier into a scoring function, which applies logistic regression on the SVM scores using the training dataset and cross-validation. The score then, is a probability that estimates the likelihood of a protein-DNA complex to be a near-native structure, or good prediction. We used the package e1071 [73] in R, which has embedded the functionalities to implement the RBF kernel and the Platt scaling while training an SVM model.

3.2.2 Model Training

Training Dataset

The training dataset was originally used to develop the orientation potential [103]. It consists of 160 TF-DNA complex structures (Table 4) from the Protein Data Bank [9]. The DNA and protein of each native structure is separated first and 200 indepen-

Table 4: Non-redundant dataset of 160 transcription factor-DNA complexes for training the scoring function.

1a02N	1a0aA	1a3qA	1aisA	1am9A	1an4A	1b72A	1b72B	1bdhA	1bdtA
1bf5A	1bg1A	1bl0A	1bvoA	1c9bA	1cf7A	1cf7B	1cqtI	1d3uB	1d5yA
1dh3A	1dp7P	1dszA	1efaA	1fosF	1fzpB	1g2dC	1gd2E	1h0mA	1h88C
1h9dA	1h9tA	1hbxG	1hcqA	1hjbA	1hwtC	1if1A	1ignA	1ihfB	1imhC
1je8A	1jfiA	1jfiB	1k6oB	1k78A	1kb2A	1ku7A	1lb2B	1lq1A	1mdyA
1mjeA	1mjeB	1mnmC	1nkpA	1nlwA	1nvpC	1odhA	1ozjA	1perL	1pp8F
1pueE	1pyiA	1r4iA	1rm1A	1rm1C	1sknP	1svcP	1t2kD	1ttuA	1u8bA
1u8rA	1ubdC	1vtnC	1xbrA	1xsdA	1yo5C	1zlkA	2a07F	2aybA	2bopA
2bsqA	2bsqE	2c9lY	2caxA	2d5vA	2dgcA	2drpA	2er8A	2etwA	2f8xC
2fo1D	2fo1E	2gliA	2h27A	2h8rA	2hanB	2hosA	2hzvA	2i13A	2iieA
2nnyA	2o4aA	2o61A	2p5lC	2prtA	2qfjA	2qhbA	2ql2B	2r1jL	2r5yA
2vz4A	2wbuA	2wt7A	2x6vA	2xroA	2xsdC	3a01A	3a5tA	3bs1A	3c2iA
3clzA	3co7C	3coqA	3croL	3d1nI	3d2wA	3d6yA	3dfvC	3do7A	3dzuA
3dzuD	3ereD	3f27D	3fdqA	3fmtA	3g73A	3gfiA	3h0dA	3htsB	3igmA
3iktA	3iv5A	3jtgA	3ketA	3lspA	3m9eA	3mlpA	3mva0	3mzhA	3o9xA
3odcA	3oqmA	3orcA	3osfA	3q05A	3q0aA	3q5fA	3qmbA	3qsvA	3qymA

dent predictions were generated using the rigid-docking algorithm with the orientation potential and 200 predictions were generated with the multi-body potential, with the objective of obtaining different complex conformations. The core of the docking program is a Monte Carlo (MC) simulation approach, which runs a maximum of 1.5 million steps or until it converges. The predictions are labeled "good" and "bad" based on the RMSD, computed between the DNA backbone heavy atoms of the native ($\mathbf{v}$) and the predicted ($\mathbf{w}$) structures, after superimposing the proteins (Equation 3). If the RMSD is less or equal to 3Å, we consider the prediction as "good" (or positive), and if it is more than 3Å, then it is labeled a "bad" (or negative) prediction.

Balanced Class Selection

After calculation of the RSMD of all the predictions from the training dataset, we observed the small percentage of positive samples (7.6%) and a large percentage of

negative samples (92.4%). Hard negative mining was implemented to address the problem of an unbalanced training dataset. Hard negative mining is an iterative training process that selects an initial training dataset combining all positive cases and a random sample from the negative cases, then trains a model based on the initial training dataset and adds to it the cases that resulted in false positives after the previous training, until the training dataset remains unchanged (Figure 21).

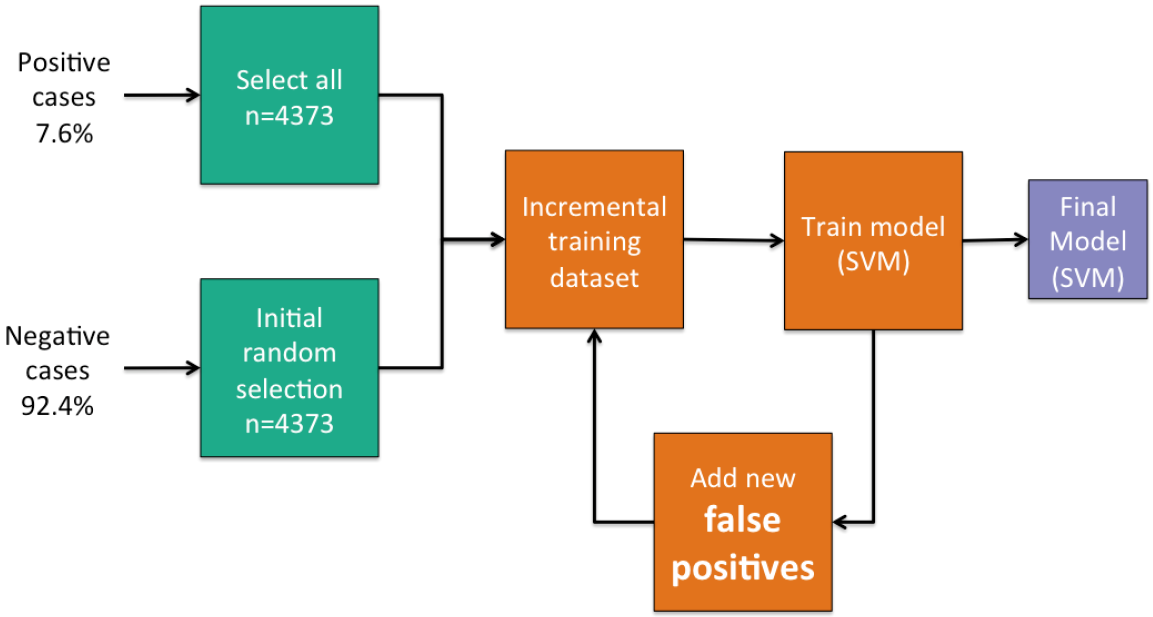


Figure 21: Training of an SVM model using hard negative mining.

3.2.3 Model Performance

Testing Dataset

To test the performance of the scoring function we used a testing dataset consisting of 38 transcription factor-DNA complex structures (Table 5), which was developed as a benchmark for rigid-docking algorithms [50].

Table 5: Non-redundant dataset of 38 transcription factor-DNA complexes for testing.

1aay	1an2	1b01	1by4	1cma	1gxp	1h8a	1hjc	1jj4	1jt0
1lmb	1qn4	1qpi	1r8d	1rio	1sax	1tro	1xpx	1z9c	1zme
1zs4	2ac0	2bnw	2c6y	2cgp	2e1c	2fio	2irf	2it0	2or1
2rbf	2yvh	2zhg	3clc	3dnv	3e6c	3hdd	3gz6		

Matthews Correlation Coefficient (MCC)

The performance of binary classifiers can be measured in multiple ways [85]. The selected measures can have a big impact on the development of the model, due to the biases they present towards minimizing false positive or false negative cases. The Matthews correlation coefficient (MCC) (Equation 5) is a widely used measure for the quality of binary classifiers. It takes into account all cases, true positive (TP), true negative (TN), false positive (FP) and false negative (FN) cases, in contrast with other measures such as precision that are biased towards increasing the number of true positive cases only. It also has the advantage of working in applications where the number of positive and negative cases is unbalanced, which makes it particular useful for this study.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

Performance Evaluation

The main goal of the model developed in this study, is to improve the performance of the protein-DNA docking algorithm by providing quality assessment of the predicted models. Currently, the orientation potential can have a prediction accuracy of 55%. However one of its biggest limitations of the current approach is its capability to

recognize incorrect predictions.

The test cases are classified as true positive (TP_{OR}), false negative (FN_{OR}) and false positive (FP_{OR}) according to the orientation potential. A case is TP_{OR}, if the conformation with the lowest energy (out of 200 predicted conformations) is a "good" prediction, *i.e.*, has an $RMSD_{nat,pred} \leq 3$Å. On the other hand, a case is FN_{OR} if there is a good prediction among the 200, but the conformation with the lowest energy is a "bad" prediction, *i.e.*, has an $RMSD_{nat,pred} > 3$Å. The third class (FP_{OR}) reprsents the cases where all the docked conformations are bad predictions.

When the SVM model is used, the cases are classified as true positive (TP_{SVM}), true negative (TN_{SVM}), false negative (FN_{SVM}), and false positive (FP_{SVM}). For example, the score represents the probability (p) of the protein-DNA complex to be a good prediction. We then set a probability cutoff of 0.5 to predict if the complex is a good ($p \geq 0.5$) or a bad ($p < 0.5$) conformation. If the maximum probability (out of the 200 predictions) is greater or equal than 0.5, and the $RMSD_{nat,pred}$ of the complex with the maximum probability (the "best" predicted conformation) is less or equal than $3\mathring{A}$, then the case is a TP_{SVM}; if the maximum probability is greater or equal than 0.5, but the $RMSD_{nat,pred}$ of the best prediction is greater than 3Å, the case is a FP_{SVM}. On the other hand, if the maximum probability is less than 0.5, and the minimum $RMSD_{nat,pred}$ is greater than 3Å, *i.e.*, the docking algorithm was not able to find a good conformation, the case is classified as TN_{SVM}. However, if the maximum probability is less than 0.5, but there is at least one good prediction, then it is a FN_{SVM} case.

The improvement of the SVM model over the orientation potential on assessing

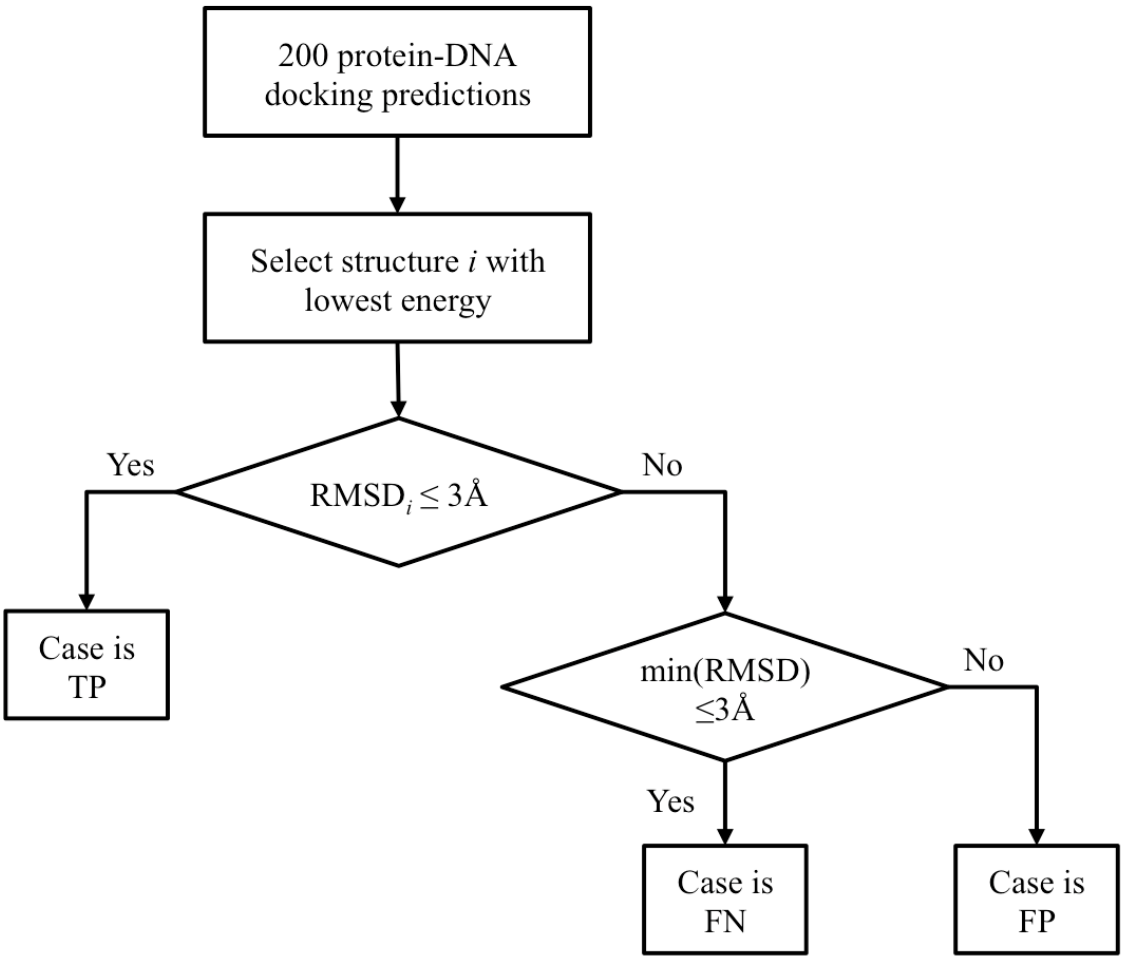


Figure 22: Protein-DNA docking predictions are classified into true positive (TP), false positive (FP), false negative (FN), and true negative (TN=0), using an energy score to select the best conformation.

the quality of the protein-DNA docking predictions is estimated by comparing the accuracy (Equation 6) of each method.

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{6}$$

3.3 Results

The scoring function is trained using hard negative mining, which takes an initial random sample from the training dataset. Due to the randomness of the selection

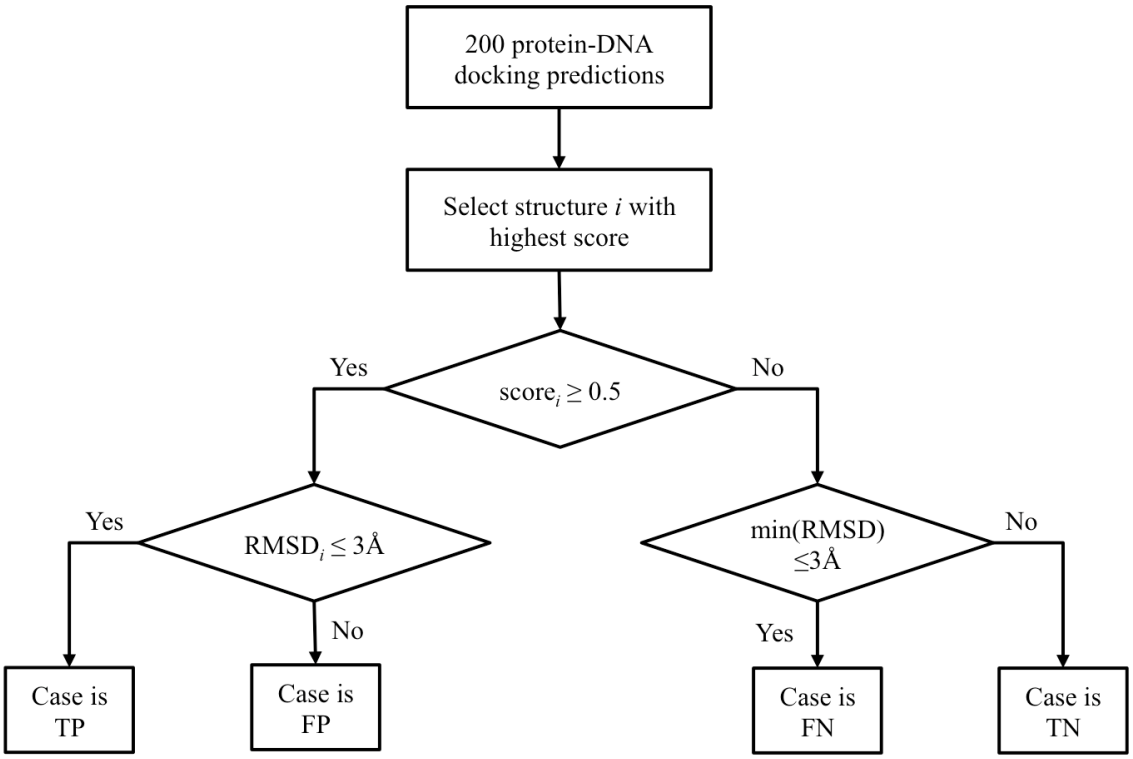


Figure 23: Protein-DNA docking predictions are classified into true positive (TP), true negative (TN), false positive (FP), and false negative (FN) using the SVM scoring function to select the best conformation.

of the dataset to train the SVM model, 30 independent models were generated. The SVM model, as a binary classifier, has an average Matthews correlation coefficient of 0.82 ($s = 0.003$) (Figure 24A).

The orientation potential has a reported accuracy of 0.55. With our implementation, the current accuracy for the orientation potential is 0.61 (=23/38), which is much smaller than any accuracy obtained using either DDNA3, an all-atom potential, or the SVM scoring function (Figure 24B). The median accuracy for the SVM model is 0.79 (=30/38). It was able to consistently recover 21 out of 23 TP_{OR} cases, to correctly predict up to 8 TN_{SVM} cases from the 10 FP_{OR}, and to make the correct selection on up to 3 out of the 5 FN_{OR}.

As shown in Figure 25, the DNA-binding unit of forkhead box protein K2 (FOXK2, PDB ID: `2c6y`) is a case where both, the orientation potential and the SVM model selected a correct structure. The HTH-type transcriptional regulator QacR (PDB IDL: `1jt0`) is an example of a negative case (FP_{OR}), correctly predicted by the SVM model (TN_{SVM}). The omega transcriptional repressor (PDB ID: `2bnw`), for which the orientation potential failed to select a conformation with small RMSD (FN_{OR}), is a true positive case when ranked by the SVM score (TP_{SVM}), which is another example of the improvement made by the SVM model on classifying good and bad structures. The remaining cases from the testing dataset can be found in Figure S1.

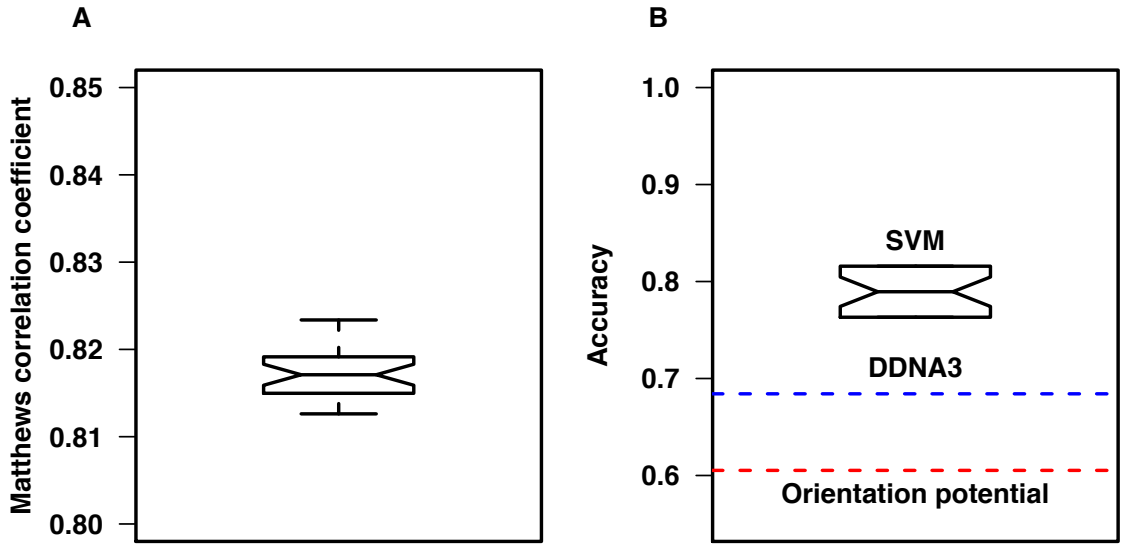


Figure 24: Performance of the SVM model. (A) Distribution of the Matthews correlation coefficient (MCC) of 30 independent SVM models on the testing dataset. (B) Distribution of the accuracy of the SVM model (boxplot), compared to the accuracy of the orientation potential (red dashed line) and the accuracy of DDNA3 (blue dashed line).

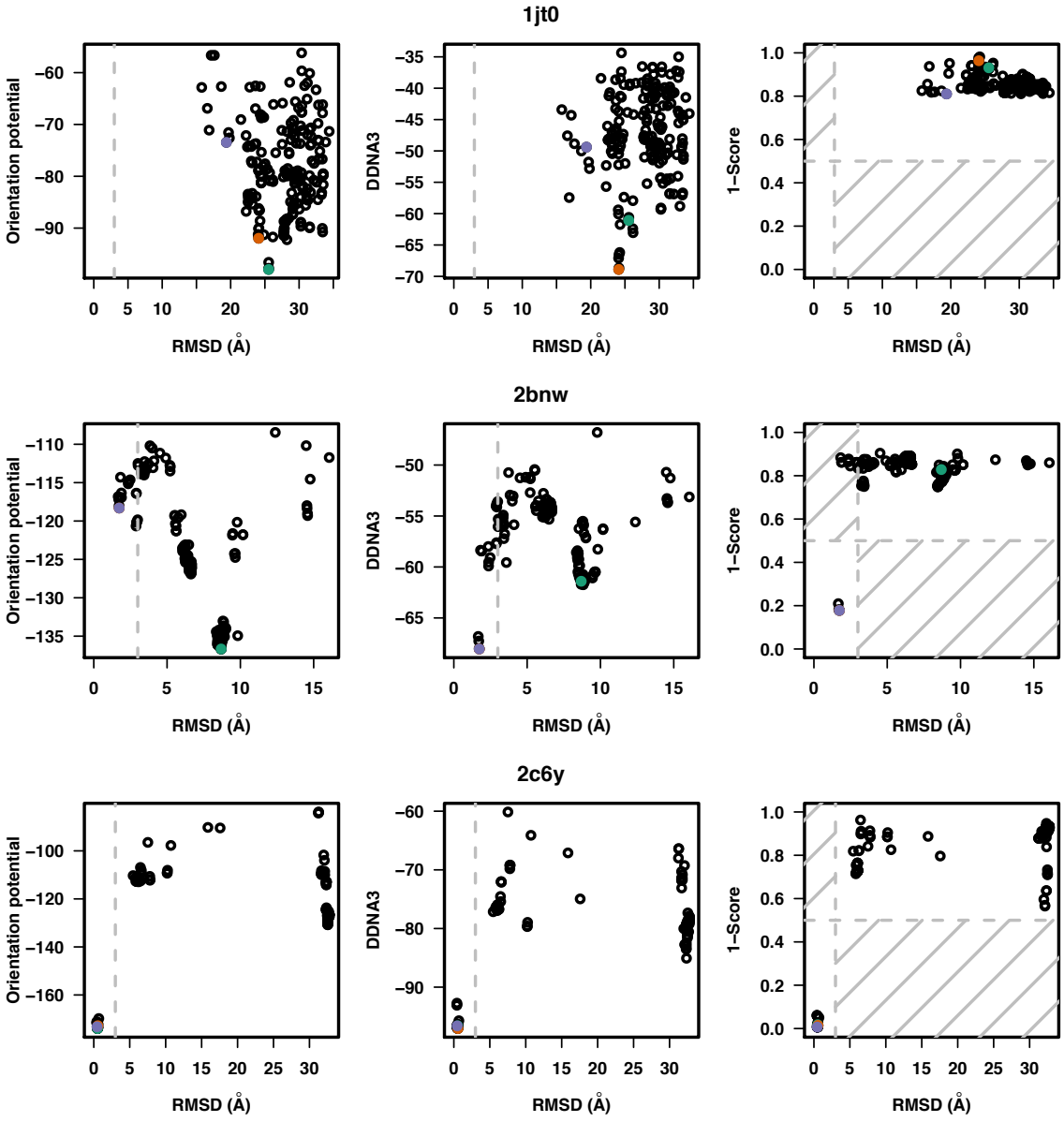


Figure 25: Root mean square deviation (RMSD) vs. orientation potential, DDNA3 potential and predicted SVM quality score for 1jt0, 2bnw and 2c6y from the testing dataset. The conformation with the lowest orientation potential (green), DDNA3 potential (orange) and highest quality score (blue) are highlighted across the three selection methods. The RMSD cutoff is set at 3Å (vertical gray dashed line) and the quality score cutoff value is set at 0.5 (horizontal gray dashed line). False positive and false negative samples, according to the scoring function (rightmost plot), fall in the gray rectangles.

3.4 Discussion

We developed a scoring function that uses three energy functions and static structural features to estimate the quality of a protein-DNA docking prediction. The

scoring function has an average Matthews correlation coefficient of 0.82 and a median accuracy of 0.79, which is a great improvement over the orientation potential (accuracy=0.61) or DDNA3 (accuracy=0.68) to select the best model from a pool of protein-DNA docking predictions.

This new SVM scoring function help us identify the true negatives by lowering the number of false positives, where the docking algorithm failed to produce good predictions, since any energy function by itself, is unable to detect true negative cases. It can be applied as a selection strategy for any docking algorithm, either rigid- or flexible-docking, and potentially to estimate the quality of any protein-DNA complex structure, due to the simplicity of the features selected for the model.

In conclusion, we can envision a fully developed, efficient and accurate pipeline for TF-DNA docking prediction, where the SVM model developed in this study will serve as a confidence measure of the predicted conformations or clusters of conformations. Other steps may include flexible-docking on the protein and DNA structures, and side-chain packing or refinement.

REFERENCES

[1] P. Agback, H. Baumann, S. Knapp, R. Ladenstein, and T. Hrd. Architecture of nonspecific protein-DNA interactions in the Sso7d-DNA complex. *Nature Structural Biology*, 5(7):579–584, July 1998.

[2] M. Andrabi, K. Mizuguchi, and S. Ahmad. Conformational changes in DNA-binding proteins: relationships with precomplex features and contributions to specificity and stability. *Proteins*, 82(5):841–857, May 2014.

[3] J. L. Asensio, A. Ard, F. J. Caada, and J. Jimnez-Barbero. Carbohydrate-aromatic interactions. *Accounts of Chemical Research*, 46(4):946–954, Apr. 2013.

[4] J. Ashworth and D. Baker. Assessment of the optimization of affinity and specificity at proteinDNA interfaces. *Nucleic Acids Research*, 37(10):e73, June 2009.

[5] J. Ashworth, J. J. Havranek, C. M. Duarte, D. Sussman, R. J. Monnat, B. L. Stoddard, and D. Baker. Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature*, 441(7093):656–659, June 2006.

[6] J. Ashworth, G. K. Taylor, J. J. Havranek, S. A. Quadri, B. L. Stoddard, and D. Baker. Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. *Nucleic Acids Research*, 38(16):5601–5608, Sept. 2010.

[7] C. M. Baker and G. H. Grant. Role of aromatic amino acids in protein-nucleic acid recognition. *Biopolymers*, 85(5-6):456–470, Apr. 2007.

[8] C. R. Baker, B. B. Tuch, and A. D. Johnson. Extensive DNA-binding specificity divergence of a conserved transcription regulator. *Proceedings of the National Academy of Sciences*, 108(18):7493–7498, May 2011.

[9] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, Jan. 2000.

[10] V. Berthold and K. Geider. Interaction of DNA with DNA-Binding Proteins. *European Journal of Biochemistry*, 71(2):443–449, Dec. 1976.

[11] N. Bhardwaj, R. E. Langlois, G. Zhao, and H. Lu. Kernel-based machine learning protocol for predicting DNA-binding proteins. *Nucleic Acids Research*, 33(20):6486–6493, 2005.

[12] N. Bhardwaj and H. Lu. Residue-level prediction of DNA-binding sites and its application on DNA-binding protein predictions. *FEBS letters*, 581(5):1058–1066, Mar. 2007.

[13] M. Biasini, S. Bienert, A. Waterhouse, K. Arnold, G. Studer, T. Schmidt, F. Kiefer, T. G. Cassarino, M. Bertoni, L. Bordoli, and T. Schwede. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, page gku340, Apr. 2014.

[14] M. Bochtler, R. H. Szczepanowski, G. Tamulaitis, S. Grazulis, H. Czapinska, E. Manakova, and V. Siksnys. Nucleotide flips determine the specificity of the Ecl18ki restriction endonuclease. *The EMBO Journal*, 25(10):2219–2229, May 2006.

[15] C. Chen and B. M. Pettitt. DNA Shape versus Sequence Variations in the Protein Binding Process. *Biophysical Journal*, 110(3):534–544, Feb. 2016.

[16] T. Y. Chiang and G. A. Marzluf. DNA recognition by the NIT2 nitrogen regulatory protein: Importance of the number, spacing, and orientation of GATA core elements and their flanking sequences upon NIT2 binding. *Biochemistry*, 33(2):576–582, Jan. 1994.

[17] P. Chne. Mutations at position 277 modify the DNA-binding specificity of human p53 in vitro. *Biochemical and Biophysical Research Communications*, 263(1):1–5, Sept. 1999.

[18] B. Contreras-Moreira, J. Sancho, and V. E. Angarica. Comparison of DNA binding across protein superfamilies. *Proteins*, 78(1):52–62, Jan. 2010.

[19] P. Cramer, K.-J. Armache, S. Baumli, S. Benkert, F. Brueckner, C. Buchen, G. Damsma, S. Dengl, S. Geiger, A. Jasiak, A. Jawhari, S. Jennebach, T. Kamenski, H. Kettenberger, C.-D. Kuhn, E. Lehmann, K. Leike, J. Sydow, and A. Vannini. Structure of Eukaryotic RNA Polymerases. *Annual Review of Biophysics*, 37(1):337–352, 2008.

[20] P. Csermely, R. Palotai, and R. Nussinov. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends in Biochemical Sciences*, 35(10):539–546, Oct. 2010.

[21] M. v. Dijk, A. D. J. v. Dijk, V. Hsu, R. Boelens, and A. M. J. J. Bonvin. Information-driven proteinDNA docking using HADDOCK: it is a matter of flexibility. *Nucleic Acids Research*, 34(11):3317–3325, Jan. 2006.

[22] C. Dominguez, R. Boelens, and A. M. J. J. Bonvin. HADDOCK: A Protein-Protein Docking Approach Based on Biochemical or Biophysical Information. *Journal of the American Chemical Society*, 125(7):1731–1737, Feb. 2003.

[23] A. K. Dunker and V. N. Uversky. Drugs for protein clouds: targeting intrinsically disordered transcription factors. *Current Opinion in Pharmacology*, 10(6):782–788, 2010.

[24] H. J. Dyson and P. E. Wright. Coupling of folding and binding for unstructured proteins. *Current Opinion in Structural Biology*, 12(1):54–60, Feb. 2002.

[25] A. Farrel, J. Murphy, and J.-t. Guo. Structure-based prediction of transcription factor binding specificity using an integrative energy function. *Bioinformatics*, 2016.

[26] G. N. Filippova, C.-F. Qi, J. E. Ulmer, J. M. Moore, M. D. Ward, Y. J. Hu, D. I. Loukinov, E. M. Pugacheva, E. M. Klenova, P. E. Grundy, A. P. Feinberg, A.-M. Cleton-Jansen, E. W. Moerland, C. J. Cornelisse, H. Suzuki, A. Komiya, A. Lindblom, F. Dorion-Bonnet, P. E. Neiman, H. C. Morse, S. J. Collins, and V. V. Lobanenkov. Tumor-associated zinc finger mutations in the CTCF transcription factor selectively alter tts DNA-binding specificity. *Cancer Research*, 62(1):48–52, Jan. 2002.

[27] J. H. Fong, B. A. Shoemaker, S. O. Garbuzynskiy, M. Y. Lobanov, O. V. Galzitskaya, and A. R. Panchenko. Intrinsic Disorder in Protein Interactions: Insights From a Comprehensive Structural Analysis. *PLoS Comput Biol*, 5(3):e1000316, Mar. 2009.

[28] N. K. Fox, S. E. Brenner, and J.-M. Chandonia. SCOPe: Structural Classification of Proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research*, 42(Database issue):D304–309, Jan. 2014.

[29] S. J. Franklin and J. K. Barton. Differential DNA Recognition by the Enantiomers of 1-Rh(MGP)2phi: A Combination of Shape Selection and Direct Readout. *Biochemistry*, 37(46):16093–16105, Nov. 1998.

[30] M. Fuxreiter, I. Simon, and S. Bondos. Dynamic protein-DNA recognition: beyond what can be seen. *Trends in Biochemical Sciences*, 36(8):415–423, Aug. 2011.

[31] S. Gama-Castro, H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, L. Muiz-Rascado, H. Solano-Lira, V. Jimenez-Jacinto, V. Weiss, J. S. Garca-Sotelo, A. Lpez-Fuentes, L. Porrn-Sotelo, S. Alquicira-Hernndez, A. Medina-Rivera, I. Martnez-Flores, K. Alquicira-Hernndez, R. Martnez-Adame, C. Bonavides-Martnez, J. Miranda-Ros, A. M. Huerta, A. Mendoza-Vargas, L. Collado-Torres, B. Taboada, L. Vega-Alvarado, M. Olvera, L. Olvera, R. Grande, E. Morett, and J. Collado-Vides. RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Research*, 39(Database issue):D98–105, Jan. 2011.

[32] R. Gordn, K. F. Murphy, R. P. McCord, C. Zhu, A. Vedenko, and M. L. Bulyk. Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights. *Genome Biology*, 12:R125, 2011.

[33] X. Guo, M. L. Bulyk, and A. J. Hartemink. Intrinsic disorder within and flanking the DNA-binding domains of human transcription factors. *Pacific*

Symposium on Biocomputing. Pacific Symposium on Biocomputing, pages 104–115, 2012.

[34] T. Ghler, S. Jger, G. Warnecke, H. Yasuda, E. Kim, and W. Deppert. Mutant p53 proteins bind DNA in a DNA structure-selective mode. *Nucleic Acids Research*, 33(3):1087–1100, Jan. 2005.

[35] S. Gnther, K. Rother, and C. Frmmel. Molecular flexibility in proteinDNA interactions. *Biosystems*, 85(2):126–136, Aug. 2006.

[36] H. Haas, K. Angermayr, and G. Stffler. Molecular analysis of a Penicillium chrysogenum GATA factor encoding gene (sreP) exhibiting significant homology to the Ustilago maydis urbs1 gene. *Gene*, 184(1):33–37, Jan. 1997.

[37] S. Hauenstein, C.-M. Zhang, Y.-M. Hou, and J. J. Perona. Shape-selective RNA recognition by cysteinyl-tRNA synthetase. *Nature Structural & Molecular Biology*, 11(11):1134–1141, Nov. 2004.

[38] J. J. Havranek, C. M. Duarte, and D. Baker. A simple physical model for the prediction and design of protein-DNA interactions. *Journal of molecular biology*, 344(1):59–70, Nov. 2004.

[39] S. J. Hubbard and J. M. Thornton. NACCESS, 1993.

[40] C. A. Hunter and J. K. M. Sanders. The nature of .pi.-.pi. interactions. *Journal of the American Chemical Society*, 112(14):5525–5534, July 1990.

[41] A. C. Jamieson, S.-H. Kim, and J. A. Wells. In vitro selection of zinc fingers with altered DNA-binding specificity. *Biochemistry*, 33(19):5689–5695, May 1994.

[42] J. Janin and M. J. Sternberg. Protein flexibility, not disorder, is intrinsic to molecular recognition. *F1000 Biology Reports*, 5, Jan. 2013.

[43] D. Jantz and J. M. Berg. Probing the DNA-Binding Affinity and Specificity of Designed Zinc Finger Proteins. *Biophysical Journal*, 98(5):852–860, Mar. 2010.

[44] T. Jenuwein and C. D. Allis. Translating the Histone Code. *Science*, 293(5532):1074–1080, Aug. 2001.

[45] A. Jolma, J. Yan, T. Whitington, J. Toivonen, K. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, K. Palin, J. Vaquerizas, R. Vincentelli, N. Luscombe, T. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, and J. Taipale. DNA-Binding Specificities of Human Transcription Factors. *Cell*, 152(12):327–339, Jan. 2013.

[46] R. Joshi, J. M. Passner, R. Rohs, R. Jain, A. Sosinsky, M. A. Crickmore, V. Jacob, A. K. Aggarwal, B. Honig, and R. S. Mann. Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell*, 131(3):530–543, Nov. 2007.

[47] C. M. Joyce and T. A. Steitz. Function and Structure Relationships in DNA Polymerases. *Annual Review of Biochemistry*, 63(1):777–822, 1994.

[48] C. G. Kalodimos, N. Biris, A. M. J. J. Bonvin, M. M. Levandoski, M. Guennuegues, R. Boelens, and R. Kaptein. Structure and Flexibility Adaptation in Nonspecific and Specific Protein-DNA Complexes. *Science*, 305(5682):386–389, July 2004.

[49] T. Kaplan, N. Friedman, and H. Margalit. Ab Initio Prediction of Transcription Factor Targets Using Structural Knowledge. *PLOS Comput Biol*, 1(1):e1, June 2005.

[50] R. Kim, R. I. Corona, B. Hong, and J.-t. Guo. Benchmarks for flexible and rigid transcription factor-DNA docking. *BMC structural biology*, 11:45, 2011.

[51] R. Kim and J.-t. Guo. PDA: an automatic and comprehensive analysis program for protein-DNA complex structures. *BMC genomics*, 10 Suppl 1:S13, 2009.

[52] R. M. Knegtel, J. Antoon, C. Rullmann, R. Boelens, and R. Kaptein. MONTY: a Monte Carlo approach to protein-DNA recognition. *Journal of Molecular Biology*, 235(1):318–324, Jan. 1994.

[53] I. B. Kuznetsov, Z. Gou, R. Li, and S. Hwang. Using evolutionary and structural information to predict DNA-binding sites on DNA-binding proteins. *Proteins*, 64(1):19–27, July 2006.

[54] D. S. Latchman. Transcription-Factor Mutations and Disease. *New England Journal of Medicine*, 334(1):28–33, Jan. 1996.

[55] A. Laughon. DNA binding specificity of homeodomains. *Biochemistry*, 30(48):11357–11367, Dec. 1991.

[56] W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, July 2006.

[57] R. Linding, R. B. Russell, V. Neduva, and T. J. Gibson. GlobPlot: Exploring protein sequences for globularity and disorder. *Nucleic acids research*, 31(13):3701–3708, July 2003.

[58] J. Liu, J. R. Faeder, and C. J. Camacho. Toward a quantitative theory of intrinsically disordered proteins and their function. *Proceedings of the National Academy of Sciences*, 106(47):19819 –19823, Nov. 2009.

[59] Z. Liu, F. Mao, J.-t. Guo, B. Yan, P. Wang, Y. Qu, and Y. Xu. Quantitative evaluation of protein-DNA interactions using an optimized knowledge-based potential. *Nucleic Acids Research*, 33(2):546–558, 2005.

[60] X.-J. Lu and W. K. Olson. 3dna: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research*, 31(17):5108–5121, Sept. 2003.

[61] N. M. Luscombe, R. A. Laskowski, and J. M. Thornton. Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic acids research*, 29(13):2860–2874, July 2001.

[62] N. M. Luscombe, R. A. Laskowski, and J. M. Thornton. Amino acidbase interactions: a three-dimensional analysis of proteinDNA interactions at an atomic level. *Nucleic Acids Research*, 29(13):2860–2874, July 2001.

[63] N. M. Luscombe and J. M. Thornton. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *Journal of Molecular Biology*, 320(5):991–1009, July 2002.

[64] N. M. Luscombe and J. M. Thornton. ProteinDNA Interactions: Amino Acid Conservation and the Effects of Mutations on Binding Specificity. *Journal of Molecular Biology*, 320(5):991–1009, July 2002.

[65] Y. Mandel-Gutfreund and H. Margalit. Quantitative parameters for amino acid-base interaction: implications for prediction of protein-DNA binding sites. *Nucleic Acids Research*, 26(10):2306–2312, May 1998.

[66] A. Mathelier, O. Fornes, D. J. Arenillas, C.-y. Chen, G. Denay, J. Lee, W. Shi, C. Shyr, G. Tan, R. Worsley-Hunt, A. W. Zhang, F. Parcy, B. Lenhard, A. Sandelin, and W. W. Wasserman. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 44(D1):D110–D115, Jan. 2016.

[67] A. Mathelier, X. Zhao, A. W. Zhang, F. Parcy, R. Worsley-Hunt, D. J. Arenillas, S. Buchman, C.-y. Chen, A. Chou, H. Ienasescu, J. Lim, C. Shyr, G. Tan, M. Zhou, B. Lenhard, A. Sandelin, and W. W. Wasserman. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, page gkt997, Nov. 2013.

[68] S. W. Matson and K. A. Kaiser-Rogers. DNA Helicases. *Annual Review of Biochemistry*, 59(1):289–329, 1990.

[69] B. W. Matthews. No code for recognition. *Nature*, 335(6188):294–295, Sept. 1988.

[70] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R.

Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science*, 337(6099):1190–1195, Sept. 2012.

[71] I. K. McDonald and J. M. Thornton. Satisfying Hydrogen Bonding Potential in Proteins. *Journal of Molecular Biology*, 238(5):777–793, May 1994.

[72] M. Merika and S. H. Orkin. DNA-binding specificity of GATA family transcription factors. *Molecular and Cellular Biology*, 13(7):3999–4010, July 1993.

[73] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C.-C. C. l. C++-code), and C.-C. L. l. C++-code). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien, Aug. 2015.

[74] M. Michael Gromiha, C. Santhosh, and M. Suwa. Influence of cation interactions in proteinDNA complexes. *Polymer*, 45(2):633–639, Jan. 2004.

[75] M. Michael Gromiha, J. G. Siebers, S. Selvaraj, H. Kono, and A. Sarai. Intermolecular and Intramolecular Readout Mechanisms in ProteinDNA Recognition. *Journal of Molecular Biology*, 337(2):285–294, Mar. 2004.

[76] F. V. Murphy and M. E. Churchill. Nonsequence-specific DNA recognition: a structural perspective. *Structure*, 8(4):R83–R89, Apr. 2000.

[77] P. M. Murphy, J. M. Bolduc, J. L. Gallaher, B. L. Stoddard, and D. Baker. Alteration of enzyme specificity by computational loop remodeling and design. *Proceedings of the National Academy of Sciences*, 106(23):9215–9220, June 2009.

[78] K. Nadassy, S. J. Wodak, and J. Janin. Structural features of protein-nucleic acid recognition sites. *Biochemistry*, 38(7):1999–2017, Feb. 1999.

[79] M. Oda, K. Furukawa, K. Ogata, A. Sarai, and H. Nakamura. Thermodynamics of specific and non-specific DNA binding by the c-myb DNA-binding domain1. *Journal of Molecular Biology*, 276(3):571–590, Feb. 1998.

[80] S. Osada, H. Yamamoto, T. Nishihara, and M. Imagawa. DNA Binding Specificity of the CCAAT/Enhancer-binding Protein Transcription Factor Family. *Journal of Biological Chemistry*, 271(7):3891–3896, Feb. 1996.

[81] C. O. Pabo and L. Nekludova. Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *Journal of Molecular Biology*, 301(3):597–624, Aug. 2000.

[82] C. O. Pabo and R. T. Sauer. Transcription factors: structural families and principles of DNA recognition. *Annual review of biochemistry*, 61:1053–1095, 1992.

[83] G. Paillard, C. Deremble, and R. Lavery. Looking into DNA recognition: zinc finger binding specificity. *Nucleic Acids Research*, 32(22):6673–6682, Jan. 2004.

[84] Y. Pan, C.-J. Tsai, B. Ma, and R. Nussinov. Mechanisms of transcription factor selectivity. *Trends in Genetics*, 26(2):75–83, Feb. 2010.

[85] C. Parker. On measuring the performance of binary classifiers. *Knowledge and Information Systems*, 35(1):131–152, Sept. 2012.

[86] V. Pingoud, E. Kubareva, G. Stengel, P. Friedhoff, J. M. Bujnicki, C. Urbanke, A. Sudina, and A. Pingoud. Evolutionary Relationship between Different Subgroups of Restriction Endonucleases. *Journal of Biological Chemistry*, 277(16):14306–14314, Apr. 2002.

[87] M. H. Porteus and D. Baltimore. Chimeric Nucleases Stimulate Gene Targeting in Human Cells. *Science*, 300(5620):763–763, May 2003.

[88] P. Radivojac. Protein flexibility and intrinsic disorder. *Protein Science*, 13(1):71–80, Jan. 2004.

[89] R. Rohs, X. Jin, S. M. West, R. Joshi, B. Honig, and R. S. Mann. Origins of Specificity in Protein-DNA Recognition. *Annual Review of Biochemistry*, 79(1):233–269, May 2010.

[90] R. Rohs, S. M. West, P. Liu, and B. Honig. Nuance in the double-helix and its role in protein-DNA recognition. *Current Opinion in Structural Biology*, 19(2):171–177, Apr. 2009.

[91] R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig. The role of DNA shape in proteinDNA recognition. *Nature*, 461(7268):1248–1253, Oct. 2009.

[92] P. J. Sapienza, J. M. Rosenberg, and L. Jen-Jacobson. Structural and thermodynamic basis for enhanced DNA binding by a promiscuous mutant EcoRI endonuclease. *Structure (London, England: 1993)*, 15(11):1368–1382, Nov. 2007.

[93] C. Sayou, M. Monniaux, M. H. Nanao, E. Moyroud, S. F. Brockington, E. Thvenon, H. Chahtane, N. Warthmann, M. Melkonian, Y. Zhang, G. K.-S. Wong, D. Weigel, F. Parcy, and R. Dumas. A Promiscuous Intermediate Underlies the Evolution of LEAFY DNA Binding Specificity. *Science*, 343(6171):645–648, Feb. 2014.

[94] J. J. Schott, D. W. Benson, C. T. Basson, W. Pease, G. M. Silberbach, J. P. Moak, B. J. Maron, C. E. Seidman, and J. G. Seidman. Congenital heart disease caused by mutations in the transcription factor NKX2-5. *Science (New York, N.Y.)*, 281(5373):108–111, July 1998.

[95] A. Sebastian and B. Contreras-Moreira. The twilight zone of cis element alignments. *Nucleic acids research*, 41(3):1438–1449, Feb. 2013.

[96] T. W. Siggers and B. Honig. Structure-based prediction of C2h2 zinc-finger binding specificity: sensitivity to docking geometry. *Nucleic Acids Research*, 35(4):1085–1097, 2007.

[97] I. Sillitoe, T. E. Lewis, A. Cuff, S. Das, P. Ashford, N. L. Dawson, N. Furnham, R. A. Laskowski, D. Lee, J. G. Lees, S. Lehtinen, R. A. Studer, J. Thornton, and C. A. Orengo. CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research*, 43(Database issue):D376–381, Jan. 2015.

[98] W. Song and J.-T. Guo. Investigation of arc repressor DNA-binding specificity by comparative molecular dynamics simulations. *Journal of Biomolecular Structure & Dynamics*, 33(10):2083–2093, 2015.

[99] M. J. Sternberg, H. A. Gabb, and R. M. Jackson. Predictive docking of proteinprotein and proteinDNA complexes. *Current Opinion in Structural Biology*, 8(2):250–256, Apr. 1998.

[100] G. D. Stormo and Y. Zhao. Determining the specificity of proteinDNA interactions. *Nature Reviews Genetics*, 11(11):751–760, Nov. 2010.

[101] B. D. Strahl and C. D. Allis. The language of covalent histone modifications. *Nature*, 403(6765):41–45, Jan. 2000.

[102] A. H. Swirnoff and J. Milbrandt. DNA-binding specificity of NGFI-A and related zinc finger transcription factors. *Molecular and Cellular Biology*, 15(4):2275–2287, Apr. 1995.

[103] T. Takeda, R. I. Corona, and J.-T. Guo. A knowledge-based orientation potential for transcription factor-DNA docking. *Bioinformatics (Oxford, England)*, 29(3):322–330, Feb. 2013.

[104] S. K. Thukral, Y. Lu, G. C. Blain, T. S. Harvey, and V. L. Jacobsen. Discrimination of DNA binding sites by mutant p53 proteins. *Molecular and Cellular Biology*, 15(9):5196–5202, Sept. 1995.

[105] D. Turner, R. Kim, and J.-t. Guo. TFinDit: transcription factor-DNA interaction data depository. *BMC bioinformatics*, 13:220, 2012.

[106] I. Tuszynska, M. Magnus, K. Jonak, W. Dawson, and J. M. Bujnicki. NPDock: a web server for proteinnucleic acid docking. *Nucleic Acids Research*, 43(Web Server issue):W425–W430, July 2015.

[107] T. G. Uil, H. J. Haisma, and M. G. Rots. Therapeutic modulation of endogenous gene function by agents with designed DNA-sequence specificities. *Nucleic Acids Research*, 31(21):6064–6078, Nov. 2003.

[108] U. Y. Ulge, D. A. Baker, and R. J. Monnat. Comprehensive computational design of mCreI homing endonuclease cleavage specificity for genome engineering. *Nucleic Acids Research*, 39(10):4330–4339, May 2011.

[109] F. D. Urnov, J. C. Miller, Y.-L. Lee, C. M. Beausejour, J. M. Rock, S. Augustus, A. C. Jamieson, M. H. Porteus, P. D. Gregory, and M. C. Holmes. Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature*, 435(7042):646–651, June 2005.

[110] D. Vuzman and Y. Levy. Intrinsically disordered regions as affinity tuners in proteinDNA interactions. *Molecular BioSystems*, 8(1):47–57, 2012.

[111] G. G. Wilson and N. E. Murray. Restriction and Modification Systems. *Annual Review of Genetics*, 25(1):585–627, 1991.

[112] K. A. Wilson, J. L. Kellie, and S. D. Wetmore. DNAprotein -interactions in nature: abundance, structure, composition and strength of contacts between aromatic amino acids and DNA nucleobases or deoxyribose sugar. *Nucleic Acids Research*, 42(10):6726–6741, June 2014.

[113] R. Wintjens, J. Livin, M. Rooman, and E. Buisine. Contribution of cation-pi interactions to the stability of protein-DNA complexes. *Journal of Molecular Biology*, 302(2):395–410, Sept. 2000.

[114] M. J. v. d. Woerd, J. J. Pelletier, S.-y. Xu, and A. M. Friedman. Restriction Enzyme BsoBIDNA Complex. *Structure*, 9(2):133–144, Feb. 2001.

[115] P. E. Wright and H. J. Dyson. Intrinsically unstructured proteins: reassessing the protein structure-function paradigm. *Journal of Molecular Biology*, 293(2):321–331, Oct. 1999.

[116] H. Xi, E. Davis, N. Ranjan, L. Xue, D. Hyde-Volpe, and D. P. Arya. Thermodynamics of Nucleic Acid Shape Readout by an Aminosugar. *Biochemistry*, 50(42):9088–9113, Oct. 2011.

[117] B. Xu, Y. Yang, H. Liang, and Y. Zhou. An all-atom knowledge-based energy function for protein-DNA threading, docking decoy discrimination, and prediction of transcription-factor binding profiles. *Proteins*, 76(3):718–730, Aug. 2009.

[118] L. Yang, T. Zhou, I. Dror, A. Mathelier, W. W. Wasserman, R. Gordan, and R. Rohs. TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Research*, 42(D1):D148–D155, Nov. 2013.

[119] H. Zhao, Y. Yang, and Y. Zhou. Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function. *Bioinformatics*, 26(15):1857–1863, Aug. 2010.

[120] H.-X. Zhou. Intrinsic disorder: signaling via highly specific but short-lived association. *Trends in Biochemical Sciences*, 37(2):43–48, Feb. 2012.

[121] T. Zhou, N. Shen, L. Yang, N. Abe, J. Horton, R. S. Mann, H. J. Bussemaker, R. Gordn, and R. Rohs. Quantitative modeling of transcription factor binding specificities using DNA shape. *Proceedings of the National Academy of Sciences*, 112(15):4654–4659, Apr. 2015.

APPENDIX A: SUPPLEMENTARY TABLES

Table S1: The non-redundant dataset pdNR30 has 28 highly specific, 115 multi-specific and 52 non-specific DNA-binding domains in complex with DNA.

Dataset	Domain ID	Domain definition	Protein name (recognition sequence)
HS	1az0B00	1az0:B	EcoRV (GAT^ATC)
	1bhmA00	1bhm:A	BamHI (G^GATCC)
	1d2iB00	1d2i:B	BglII (A^GATCT)
	1dc1A01	1dc1:A (5-38,127-323)	BsoBI (C^YCGRG)
	1dc1A02	1dc1:A (39-126)	BsoBI (C^YCGRG)
	1eriA00	1eri:A	EcoRI (G^AATTC)
	1iawA01	1iaw:A (10-176)	NaeI (GCC^GGC)
	1iawA02	1iaw:A (177-309)	NaeI (GCC^GGC)
	1kc6B00	1kc6:B	HincII (GTY^RAC)
	1pviA00	1pvi:A	PvuII (CAG^CTG)
	1vrrA00	1vrr:A	BstYI (R^GATCY)
	1wteA01	1wte:A (1-87, 212-272)	EcoO109I (RG^GNCCY)
	1wteA02	1wte:A (88-211)	EcoO109I (RG^GNCCY)
	3dvoD00	3dvo:D	SgrAI (CR^CCGGYG)
	3hqfA00	3hqf:A	EcoRII (^CCWGG)
	4abtA00	4abt:A	NgoMIV (G^CCGGC)
	d1yfib_	1yfi:B	MspI (C^CGG)
	h2e52D0	2e52:D	HindIII (A^AGCTT)
	h3m7kA0	3m7k:A	PacI (TTAAT^TAA)
	h3oqgA0	3oqg:A	Hpy188I (TCN^GA)
	m2fl3A0	2fl3:A	HinP1I (G^CGC)
	m2oaaA0	2oaa:A	MvaI (CC^WGG)
	m3c25A0	3c25:A	NotI (GC^GGCCGC)
	m3fc3B1	3fc3:B (2-107)	Hpy99I (CGWCG^)

Table S1: (Continued)

Dataset	Domain ID	Domain definition	Protein name (recognition sequence)
	m3goxB2	3gox:B (108-189)	Hpy99I (CGWCGˆ)
	m3imbD0	3imb:D	BcnI (CCˆSGG)
	m3ndhA0	3ndh:A	ThaI (CGˆCG)
	m4rdmB0	4rdm:B	NgoAVII (GCCGC)
MS	1b3tA00	1b3t:A	Epstein-Barr nuclear antigen 1
	1bdtD00	1bdt:D	Transcriptional repressor arc
	1bl0A01	1bl0:A (9-64)	Multiple antibiotic resistance protein MarA
	1bl0A02	1bl0:A (65-124)	Multiple antibiotic resistance protein MarA
	1cf7A00	1cf7:A	Transcription factor E2F4
	1cmaA00	1cma:A	Met repressor
	1ea4G00	1ea4:G	Transcriptional repressor CopG
	1exjA02	1exj:A (3-75)	Multidrug-efflux transporter 1 regulator
	1fzpB00	1fzp:B	Transcriptional regulator SarA
	1gd2E00	1gd2:E	AP-1-like transcription factor
	1gxpE00	1gxp:E	Phosphate regulon transcriptional regulatory protein PhoB
	1h6fA00	1h6f:A	T-box transcription factor TBX3
	1hjbB00	1hjb:B	CCAAT/enhancer-binding protein beta
	1hjbC00	1hjb:C	Runt-related transcription factor 1
	1ic8A01	1ic8:A (87-180)	Hepatocyte nuclear factor 1-alpha
	1ic8A02	1ic8:A (203-276)	Hepatocyte nuclear factor 1-alpha
	1jfiA00	1jfi:A	Transcription regulator NC2 alpha chain
	1jfiB00	1jfi:B	Transcription regulator NC2 beta chain
	1k78A01	1k78:A (19-84)	Paired box protein Pax-5
	1k78B00	1k78:B	Protein C-ets-1
	1kb2A00	1kb2:A	Vitamin D3 receptor
	1le5F01	1le5:F (38-241)	Nuclear factor NF-kappa-B p105 subunit
	1lmb300	1lmb:3	Repressor protein cI

Table S1: (Continued)

Dataset	Domain ID	Domain definition	Protein name (recognition sequence)
	1lq1B00	1lq1:B	Stage 0 sporulation protein A
	1mdmA02	1mdm:A (85-139)	Paired box protein Pax-5
	1mhdA00	1mhd:A	Mothers against decapentaplegic homolog 3 (SMAD3)
	1nkpD00	1nkp:D	Myc proto-oncogene protein
	1owrP01	1owr:P (397-569)	Nuclear factor of activated T-cells, cytoplasmic 2
	1pnrA01	1pnr:A (3-59)	HTH-type transcriptional repressor PurR
	1qn3B01	1qn3:B (19-29, 116-197)	TATA-box-binding protein 1
	1qn6A02	1qn6:A (30-115)	TATA-box-binding protein 1
	1qpiA01	1qpi:A (4-66)	Tetracycline repressor protein class D
	1r8dA00	1r8d:A	HTH-type transcriptional activator mta
	1rioH00	1rio:H	RNA polymerase sigma factor SigA
	1saxA01	1sax:A (9-72)	Methicillin resistance regulatory protein MecI
	1sknP00	1skn:P	Protein skinhead-1
	1t2kB01	1t2k:B (7-110)	Interferon regulatory factor 3
	1xpxA00	1xpx:A	Homeobox protein prospero
	1zreA02	1zre:A (138-207)	cAMP-activated global transcriptional regulator CRP
	1zs4A00	1zs4:A	Transcriptional activator II
	2ac0C00	2ac0:C	Cellular tumor antigen p53
	2bopA00	2bop:A	Regulatory protein E2
	2e1cA01	2e1c:A (24-76)	Uncharacterized HTH-type transcriptional regulator PH1519
	2h27A00	2h27:A	ECF RNA polymerase sigma-E factor
	2h7hA00	2h7h:A	Viral jun-transforming protein
	2i9tB02	2i9t:B (546-650)	Nuclear factor NF-kappa-B p105 subunit
	2p5lC00	2p5l:C	Arginine repressor
	2r5yB00	2r5y:B	Homeobox protein extradenticle
	2wt7A00	2wt7:A	Proto-oncogene c-Fos
	2yvhD00	2yvh:D	Transcriptional regulator

Table S1: (Continued)

Dataset	Domain ID	Domain definition	Protein name (recognition sequence)
	2zhgA00	2zhg:A	Redox-sensitive transcriptional activator SoxR
	3a01A00	3a01:A	Homeodomain-containing protein
	3coaC00	3coa:C	Forkhead box protein O1
	3dfxB00	3dfx:B	Trans-acting T-cell-specific transcription factor GATA-3
	3dnvB00	3dnv:B	Antitoxin HipB
	3g97A00	3g97:A	Glucocorticoid receptor
	3hddB00	3hdd:B	Segmentation polarity homeobox protein engrailed
	3iagC01	3iag:C (53-200, 359-380)	Recombining binding protein suppressor of hairless
	3iagC02	3iag:C (201-358)	Recombining binding protein suppressor of hairless
	3iktA01	3ikt:A (0-73)	Redox-sensing transcriptional repressor Rex
	3jtgA01	3jtg:A (273-357)	ETS-related transcription factor Elf-3
	3jxdR00	3jxd:R	Repressor protein C2
	3o9xA02	3o9x:A (59-131)	Antitoxin MqsA
	3p57B01	3p57:B (13-91)	Myocyte-specific enhancer factor 2A
	3pvvB00	3pvv:B	Chromosomal replication initiator protein DnaA
	3qwsA00	3qws:A	Repressor protein Gp39
	3s8qA00	3s8q:A	Regulatory protein
	3u2bC00	3u2b:C	Transcription factor SOX-4
	3zkcB00	3zkc:B	HTH-type transcriptional regulator SinR
	4fthA00	4fth:A	Transcriptional regulator (NtrC family)
	4g92B00	4g92:B	Transcription factor HapC (Eurofung)
	6croA00	6cro:A	Regulatory protein cro
	d1odha_	1odh:A	Chorion-specific transcription factor GCMa
	d2iszd1	2isz:D (1-64)	Iron-dependent repressor IdeR
	d2xsdc1	2xsd:C (247-319)	POU domain, class 3, transcription factor 1
	d2xsdc2	2xsd:C (343-397)	POU domain, class 3, transcription factor 1
	d3coqa1	3coq:A (8-48)	Regulatory protein GAL4

Table S1: (Continued)

Dataset	Domain ID	Domain definition	Protein name (recognition sequence)
	d3e6cc1	3e6c:C (148-233)	Cyclic nucleotide-binding protein
	h2er8C0	2er8:C	Regulatory protein LEU3
	h2vy1A0	2vy1:A	Transcription factor LEAFY
	h3a5tA0	3a5t:A	Transcription factor MafG
	h3gnaA0	3gna:A	V(D)J recombination-activating protein 1
	h3igmA0	3igm:A	Transcription factor with AP2 domain(S), putative
	h3lsrA0	3lsr:A	DesT
	h3mlpE0	3mlp:E	Transcription factor COE1
	h3vebA0	3veb:A	Macrodomain Ter protein
	h3w3cA0	3w3c:A	Virulence regulon transcriptional activator VirB
	h3zplF0	3zpl:F	Putative MarR-family transcriptional repressor
	h4gclD0	4gcl:D	Nucleoid occlusion factor SlmA
	h4h10A0	4h10:A	Aryl hydrocarbon receptor nuclear translocator-like protein 1
	h4hf1A0	4hf1:A	HTH-type transcriptional regulator IscR
	h4ihtC0	4iht:C	HTH-type transcriptional regulator BenM
	h4ix7A0	4ix7:A	RE55538p
	h4jl3A0	4jl3:A	Transcriptional regulator, TetR family
	m3fdqA0	3fdq:A	Motility gene repressor MogR
	m3h0dB1	3h0d:B (3-75)	Transcriptional regulator CtsR
	m3n7qA0	3n7q:A	Transcription termination factor 1, mitochondrial
	m3u3wA1	3u3w:A (3-58)	PlcR associated protein, PapR
	m3w6vA0	3w6v:A	AdpA
	m3zqlA0	3zql:A	Putative repressor SimReg2
	m4g92A0	4g92:A	HAPB protein
	m4g92C0	4g92:C	HapE
	m4jcyB0	4jcy:B	Csp231I C protein
	m4knyA2	4kny:A (124-225)	KDP operon transcriptional regulatory protein KdpE

Table S1: (Continued)

Dataset	Domain ID	Domain definition	Protein name (recognition sequence)
	m4l62P1	4l62:P (7-49)	Probable transcriptional regulator
	m4ldxB2	4ldx:B (121-229)	Auxin response factor 1
	m4llnA0	4lln:A	MarR family regulatory protein
	m4lmgD0	4lmg:D	Iron-regulated transcriptional activator AFT2
	m4mteB1	4mte:B (3-72)	Zinc uptake regulation protein
	m4nnuA1	4nnu:A (44-122)	Transcription factor A, mitochondrial
	m4nnuA3	4nnu:A (153-236)	Transcription factor A, mitochondrial
	m4on0B0	4on0:B	NolR
	m4qtkA0	4qtk:A	White-opaque regulator 1
	m4u0yB0	4u0y:B	HTH-type transcriptional repressor YvoA
	m4ux5A0	4ux5:A	Transcription factor MBP1
NS	1cezA01	1cez:A (8-325)	T7 RNA polymerase
	1f66C00	1f66:C	Histone H2A.Z
	1jeyA02	1jey:A (251-278, 342-439)	X-ray repair cross-complementing protein 6
	1jeyA03	1jey:A (279-341)	X-ray repair cross-complementing protein 6
	1jeyB02	1jey:B (243-443)	X-ray repair cross-complementing protein 5
	1rztA03	1rzt:A (386-508)	DNA polymerase lambda
	1rztI04	1rzt:I (509-575)	DNA polymerase lambda
	1skrA03	1skr:A (415-477, 590-704)	DNA-directed DNA polymerase
	1sxqA02	1sxq:A (167-332)	DNA beta-glucosyltransferase
	1x9wA02	1x9w:A (233-414)	DNA-directed DNA polymerase
	1xslA02	1xsl:A (332-385)	DNA polymerase lambda
	1ya6B01	1ya6:B (998-1176, 1387-1400)	DNA alpha-glucosyltransferase
	2bzfA00	2bzf:A	Barrier-to-autointegration factor
	2dnjA00	2dnj:A	Deoxyribonuclease-1
	2pi4A05	2pi4:A (554-784)	T7 RNA polymerase
	2voaA00	2voa:A	Exodeoxyribonuclease III (XthA)

Table S1: (Continued)

Dataset	Domain ID	Domain definition	Protein name (recognition sequence)
	2wtfA04	2wtf:A (393-509)	DNA polymerase eta
	3aafA00	3aaf:A	Werner syndrome ATP-dependent helicase
	3av2A00	3av2:A	Histone H3.3
	3cwsC02	3cws:C (113-230)	DNA-3-methyladenine glycosylase 2
	3gv5B04	3gv5:B (299-414)	DNA polymerase iota
	3l4jA01	3l4j:A (429-561, 609-691)	DNA topoisomerase 2
	3l4jA03	3l4j:A (692-860, 974-988)	DNA topoisomerase 2
	3l4jA04	3l4j:A (872-973)	DNA topoisomerase 2
	3n4mB00	3n4m:B	DNA-directed RNA polymerase subunit alpha
	3uiqA02	3uiq:A (109-339)	DNA polymerase
	3uiqA06	3uiq:A (775-866)	DNA polymerase
	4eyhB01	4eyh:B (26-36, 99-221)	DNA polymerase iota
	d3jxya_	3jxy:A	Alkylpurine DNA glycosylase AlkD
	d4klua1	4klu:A (11-91)	DNA polymerase beta
	d9icka3	9ick:A (92-148)	DNA polymerase beta
	h1s9fA4	1s9f:A (244-341)	DNA polymerase IV
	h2wwyA0	2wwy:A	ATP-dependent DNA helicase Q1
	h3kxtA0	3kxt:A	Chromatin protein Cren7
	h3raxB3	3rax:B (1167-1233)	DNA polymerase IV
	h4eluA2	4elu:A (423-832)	DNA polymerase I, thermostable
	h4g0vB0	4g0v:B	DNA topoisomerase 2-beta
	h4o0iA2	4o0i:A (491-605)	DNA polymerase
	h4o5eA3	4o5e:A (149-335)	DNA polymerase beta
	h4oinD0	4oin:D	DNA-directed RNA polymerase subunit beta'
	m2o8bA3	2o8b:A (321-855)	DNA mismatch repair protein Msh2
	m2o8bB1	2o8b:B (362-518)	DNA mismatch repair protein Msh6
	m2o8bB3	2o8b:B (728-1335)	DNA mismatch repair protein Msh6

Table S1: (Continued)

Dataset	Domain ID	Domain definition	Protein name (recognition sequence)
	m3f2bA0	3f2b:A	DNA polymerase III PolC-type
	m3l2pA1	3l2p:A (168-336)	DNA ligase 3
	m4c2uA4	4c2u:A (384-561)	DNA helicase
	m4dl4A4	4dl4:A (313-432)	DNA polymerase eta
	m4ir1F1	4ir1:F (0-10, 74-165)	DNA polymerase IV
	m4ir1F4	4ir1:F (236-341)	DNA polymerase IV
	m4o3mA3	4o3m:A (1072-1194)	Bloom syndrome protein
	m4plbB1	4plb:B (417-1033)	DNA gyrase subunit B
	m4plbB2	4plb:B (1034-1376, 1461-1491)	DNA gyrase subunit B

Table S2: The pairNR30 dataset consists of 11 highly specific, 41 multi-specific and 16 non-specific bound-unbound DNA-binding domain pairs.

Dataset	Apo		Holo		Protein name
	Domain ID	Domain definition	Domain ID	Domain definition	
HS	1k0zA00	1k0z:A	1pviA00	1pvi:A	Type-2 restriction enzyme PvuII
	1rveA00	1rve:A	1az0B00	1az0:B	Type-2 restriction enzyme EcoRV
	1qc9A00	1qc9:A	1cl8A00	1cl8:A	Type-2 restriction enzyme EcoRI
	h4rctA0	4rct:A	m4rdmB0	4rdm:B	Restriction endonuclease R.NgoVII
	1ev7A01	1ev7:A (10-176)	1iawA01	1iaw:A (10-176)	Type-2 restriction enzyme NaeI
	1wtdA01	1wtd:A (1-86, 212-272)	1wteA01	1wte:A (1-87, 212-272)	EcoO109IR
	h3bvqA0	3bvq:A	m3c25A0	3c25:A	NotI restriction endonuclease
	h2odhA0	2odh:A	m3imbD0	3imb:D	R.BcnI
	1sdoA00	1sdo:A	1vrrA00	1vrr:A	BstYI
	1bamA00	1bam:A	1bhmA00	1bhm:A	Type-2 restriction enzyme BamHI
	h1ynmA0	1ynm:A	m2fl3A0	2fl3:A	R.HinP1I restriction endonuclease
MS	2gzwC02	2gzw:C (138-206)	1zreA02	1zre:A (138-207)	cAMP-activated global transcriptional regulator CRP
	3g5gD00	3g5g:D	3s8qA00	3s8q:A	Regulatory protein
	2znzB01	2znz:B (25-76)	2e1cA01	2e1c:A (24-76)	Uncharacterized HTH-type transcriptional regulator PH1519
	1md0A00	1md0:A	1k78B00	1k78:B	Protein C-ets-1
	1e50G00	1e50:G	1hjbC00	1hjb:C	Runt-related transcription factor 1
	h4omzG0	4omz:G	m4on0B0	4on0:B	NolR
	2wv0I01	2wv0:I (9-79)	m4u0yB0	4u0y:B	HTH-type transcriptional repressor YvoA
	1xwrD00	1xwr:D	1zs4A00	1zs4:A	Transcriptional activator II
	2yveA00	2yve:A	2yvhD00	2yvh:D	Transcriptional regulator

Table S2: (Continued)

Dataset	Apo		Holo		Protein name
	Domain ID	Domain definition	Domain ID	Domain definition	
	1cmcA00	1cmc:A	1cmaA00	1cma:A	Met repressor
	1bftA00	1bft:A	1leiA02	1lei:A (191-285)	Transcription factor p65
	d5croo_	5cro:O	6croA00	6cro:A	Regulatory protein cro
	1okrB01	1okr:B (9-72)	1saxA01	1sax:A (9-72)	Methicillin resistance regulatory protein MecI
	3fmyA00	3fmy:A	3o9xA02	3o9x:A (59-131)	Antitoxin MqsA
	h3lisA0	3lis:A	m4jcyB0	4jcy:B	Csp231I C protein
	1jjhB00	1jjh:B	2bopA00	2bop:A	Regulatory protein E2
	1xcbA01	1xcb:A (2-73)	3iktB01	3ikt:B (1-73)	Redox-sensing transcriptional repressor Rex
	h4g4kA0	4g4k:A	h3bs1A0	3bs1:A	Accessory gene regulator protein A
	d4abza1	4abz:A (2-67)	1qpiA01	1qpi:A (4-66)	Tetracycline repressor protein class D
	h2y2zA0	2y2z:A	m3zqlA0	3zql:A	Putative repressor SimReg2
	1vokA02	1vok:A (30-115)	1qn6A02	1qn6:A (30-115)	TATA-box-binding protein 1
	1vokA01	1vok:A (19-29, 116-197)	1qn3B01	1qn3:B (19-29, 116-197)	TATA-box-binding protein 1
	3e7lB00	3e7l:B	4fthA00	4fth:A	Transcriptional regulator (NtrC family)
	3fisA00	3fis:A	d4ihxa_	4ihx:A	DNA-binding protein Fis
	3ecoA00	3eco:A	m4llnA0	4lln:A	MarR family regulatory protein
	2frhA00	2frh:A	1fzpB00	1fzp:B	Transcriptional regulator SarA
	2p5kA00	2p5k:A	2p5lC00	2p5l:C	Arginine repressor
	d3a02a_	3a02:A	d3a01b_	3a01:B	Homeobox protein aristaless
	h3m1eA0	3m1e:A	h4ihtC0	4iht:C	HTH-type transcriptional regulator BenM

Table S2: (Continued)

Dataset	Apo		Holo		Protein name
	Domain ID	Domain definition	Domain ID	Domain definition	
	h4jkzA0	4jkz:A	h4jl3A0	4jl3:A	Transcriptional regulator, TetR family
	1b0nA00	1b0n:A	3zkcB00	3zkc:B	HTH-type transcriptional regulator SinR
	d4g91b_	4g91:B	4g92B00	4g92:B	Transcription factor HapC (Eurofung)
	h4g91C0	4g91:C	m4g92C0	4g92:C	HapE
	1gxqA00	1gxq:A	1gxpE00	1gxp:E	Phosphate regulon transcriptional regulatory protein PhoB
	1r69A00	1r69:A	1perL00	1per:L	Repressor protein CI
	1mijA00	1mij:A	1xpxA00	1xpx:A	Homeobox protein prospero
	1iknA01	1ikn:A (19-186)	1ramB01	1ram:B (19-187)	Transcription factor p65
	d4a3na_	4a3n:A	3f27D00	3f27:D	Transcription factor SOX-17
	d3zq7a_	3zq7:A	m4knyA2	4kny:A (124-225)	KDP operon transcriptional regulatory protein KdpE
	1ci6B00	1ci6:B	1hjbB00	1hjb:B	CCAAT/enhancer-binding protein beta
	1jbgA00	1jbg:A	1r8dA00	1r8d:A	HTH-type transcriptional activator mta
NS	d3u8uc_	3u8u:C	m4iemD0	4iem:D	DNA-(apurinic or apyrimidinic site) lyase
	1jg6A02	1jg6:A (167-332)	1m5rB02	1m5r:B (167-332)	DNA beta-glucosyltransferase
	d1zqua1	1zqu:A (91-148)	d1huza3	1huz:A (92-148)	DNA polymerase beta
	d1zqua2	1zqu:A (149-335)	d2bpfa4	2bpf:A (149-335)	DNA polymerase beta
	1jg6A01	1jg6:A (1-166, 333-349)	1ixyA01	1ixy:A (1-166, 333-349)	DNA beta-glucosyltransferase
	2a40E00	2a40:E	2dnjA00	2dnj:A	Deoxyribonuclease-1

Table S2: (Continued)

Dataset	Apo		Holo		Protein name
	Domain ID	Domain definition	Domain ID	Domain definition	
	1mpgA02	1mpg:A (113-230)	3cwsC02	3cws:C (113-230)	DNA-3-methyladenine glycosylase 2
	h2v1xA0	2v1x:A	h2wwyA0	2wwy:A	ATP-dependent DNA helicase Q1
	d4rnpa_	4rnp:A	2pi4A05	2pi4:A (554-784)	T7 RNA polymerase
	1ih7A02	1ih7:A (107-339)	3uiqA02	3uiq:A (109-339)	DNA polymerase
	1jihA04	1jih:A (393-509)	2wtfA04	2wtf:A (393-509)	DNA polymerase eta
	d3uxna1	3uxn:A (10-91)	d1huza1	1huz:A (10-91)	DNA polymerase beta
	d3eqld_	3eql:D	h4oinD0	4oin:D	DNA-directed RNA polymerase subunit beta'
	1xv5A02	1xv5:A (1177-1386)	1y6fB02	1y6f:B (1177-1386)	DNA alpha-glucosyltransferase
	1wajA06	1waj:A (782-865)	3uiqA06	3uiq:A (775-866)	DNA polymerase
	d4hgab_	4hga:B	3av2A00	3av2:A	Histone H3.3

Table S3: The multiHolo dataset has 6 highly specific (HS), 32 multi-specific (MS), and 24 non-specific (NS) DNA-binding domains.

multiHolo				
Dataset	Cluster ID	Number of domains	Representative domain	Protein name
HS	1	39	1az0A00	Type-2 restriction enzyme EcoRV
	2	24	1kc6A00	Type-2 restriction enzyme HincII
	3	22	3dpgA00	SgraIR restriction enzyme
	4	8	m2odiA0	R.BcnI
	5	8	1esgB00	Type-2 restriction enzyme BamHI
	6	6	1eyuA00	Type-2 restriction enzyme PvuII
MS	1	29	3mfkA00	Protein C-ets-1
	2	28	3g6pB00	Glucocorticoid receptor
	3	27	1qn3A01	TATA-box-binding protein 1
	4	27	1qn3A02	TATA-box-binding protein 1
	5	23	3iglA00	Cellular tumor antigen p53
	6	20	1zrfA02	cAMP-activated global transcriptional regulator CRP
	7	19	1h88B00	CCAAT/enhancer-binding protein beta
	8	16	m4l62A1	Probable transcriptional regulator
	9	15	3clcB00	Regulatory protein
	10	14	3iv5B00	DNA-binding protein Fis
	11	12	1ea4F00	Protein CopG
	12	10	3kovA01	Myocyte-specific enhancer factor 2A
	13	10	1le5A02	Transcription factor p65
	14	9	1le5B01	Nuclear factor NF-kappa-B p105 subunit
	15	9	h4l18A0	Runt-related transcription factor 1
	16	9	m3tmmA1	Transcription factor A, mitochondrial
	17	9	m3tmmA3	Transcription factor A, mitochondrial
	18	8	3jxbC00	Repressor protein C2
	19	8	m4jqdE0	Csp231I C protein

Table S3: (Continued)

Dataset	Cluster ID	Number of domains	Representative domain	Protein name
	20	8	1bdtB00	Transcriptional repressor arc
	21	8	m4mtdD1	Zinc uptake regulation protein
	22	8	h4gclB0	Nucleoid occlusion factor SlmA
	23	7	1bdiA01	HTH-type transcriptional repressor PurR
	24	7	1owrM01	Nuclear factor of activated T-cells, cytoplasmic 2
	25	6	m4wwcB1	HTH-type transcriptional repressor YvoA
	26	6	3co6C00	Forkhead box protein O1
	27	6	d2isza1	Iron-dependent repressor IdeR
	28	6	3rn2A01	Interferon-inducible protein AIM2
	29	6	3rn2A02	Interferon-inducible protein AIM2
	30	6	1kb6B00	Vitamin D3 receptor
	31	6	m4llnD0	MarR family regulatory protein
	32	6	h2er8A0	Regulatory protein LEU3
NS	1	149	d1bpxa1	DNA polymerase beta
	2	149	d1bpxa3	DNA polymerase beta
	3	143	h1jx4A3	DNA polymerase IV
	4	143	h1jx4A4	DNA polymerase IV
	5	48	h1jxlA2	DNA polymerase IV
	6	34	1dizA02	DNA-3-methyladenine glycosylase 2
	7	32	1t3nA01	DNA polymerase iota
	8	31	1rztA02	DNA polymerase lambda
	9	29	1rztA03	DNA polymerase lambda
	10	28	d1bpya4	DNA polymerase beta
	11	27	h3lwmA2	DNA polymerase I, thermostable
	12	24	h4k4iA4	DNA polymerase lambda
	13	11	2xo7A02	DNA polymerase
	14	10	h4b9lA3	DNA polymerase

Table S3: (Continued)

Dataset	Cluster ID	Number of domains	Representative domain	Protein name
	15	10	2alzA04	DNA polymerase iota
	16	10	h3qx3B0	DNA topoisomerase 2-beta
	17	10	1skrA03	DNA-directed DNA polymerase
	18	8	1ixyA02	DNA beta-glucosyltransferase
	19	7	2xy5A04	DNA polymerase
	20	7	d1pjia2	Formamidopyrimidine-DNA glycosylase
	21	7	d1pm5a1	Formamidopyrimidine-DNA glycosylase
	22	6	m3mr5A4	DNA polymerase eta
	23	6	d1pjia3	Formamidopyrimidine-DNA glycosylase
	24	6	1y6fA02	DNA alpha-glucosyltransferase

Table S4: The multiApo dataset consists of 9 specific (HS+MS) and 6 non-specific (NS) DNA-binding domains.

multiApo				
Dataset	Cluster ID	Number of domains	Representative domain	Protein name
HS	1	6	1k0zA00	Type-2 restriction enzyme PvuII
MS	1	16	2gzwC02	cAMP-activated global transcriptional regulator CRP
	2	15	3g5gD00	Regulatory protein
	3	9	2znzB01	Uncharacterized HTH-type transcriptional regulator PH1519
	4	8	1md0A00	Protein C-ets-1
	5	8	1e50G00	Runt-related transcription factor 1
	6	8	h4omzG0	NolR
	7	6	2wv0I01	HTH-type transcriptional repressor YvoA
	8	6	1xwrD00	Transcriptional activator II
NS	1	16	d3u8uc_	DNA-(apurinic or apyrimidinic site) lyase
	2	12	1jg6A02	DNA beta-glucosyltransferase
	3	12	d1zqua1	DNA polymerase beta
	4	9	d1zqua2	DNA polymerase beta
	5	8	1jg6A01	DNA beta-glucosyltransferase
	6	7	2a40E00	Deoxyribonuclease-1

Table S5: The multiApoHolo dataset has 10 specific (HS+MS) and 4 non-specific (NS) DNA-binding domains.

multiApoHolo					
Dataset	Cluster ID	Representative domain	# of apo domains	# of holo domains	Protein name
HS	1	1h56A00	6	6	Type-2 restriction enzyme PvuII
	2	1rveA00	4	39	Type-2 restriction enzyme EcoRV
MS	1	d3rypb2	16	20	cAMP-activated global transcriptional regulator CRP
	2	3g5gD00	15	15	Regulatory protein
	3	1gvjB00	8	29	Protein C-ets-1
	4	1e50C00	8	9	Runt-related transcription factor 1
	5	h4omzC0	8	4	NolR
	6	2wv0A01	6	6	HTH-type transcriptional repressor YvoA
	7	2yveA00	4	4	Transcriptional regulator
	8	1my5A00	4	10	Transcription factor p65
NS	1	1e9nA00	16	4	DNA-(apurinic or apyrimidinic site) lyase
	2	1c3jA02	12	8	DNA beta-glucosyltransferase
	3	1jejA01	8	4	DNA beta-glucosyltransferase
	4	1mpgA02	4	34	DNA-3-methyladenine glycosylase 2

Table S6: List of protein-DNA hydrogen bonds between aspartate (Asp) and DNA bases (major and minor groove) in highly specific (HS), multi-specific (MS) and non-specific (NS) DNA-binding domains.

Dataset	DNA groove	HB geometry	Domain ID	Donor	Acceptor
HS	Major	Bidentate	1bhmA00	D0008- DC N4	A0154-ASP O
				D0009- DC N4	A0154-ASP OD2
			1iawA01	F0008- DC N4	A0146-ASP OD1
				F0009- DC N4	A0146-ASP OD2
			3dvoD00	H0008- DC N4	D0248-ASP OD1
				H0009- DC N4	D0248-ASP OD2
			4abtA00	H0005- DC N4	A0193-ASP OD1
				H0006- DC N4	A0193-ASP OD2
			m4rdmB0	E0011- DC N4	B0279-ASP OD2
				B0279-ASP N	E0010- DG N7
		Single	1iawA02	D0008- DC N4	A0226-ASP OD2
			4abtA00	E0009- DC N4	A0034-ASP OD2
			h2e52D0	H0007- DC N4	D0123-ASP OD1
			m2fl3A0	D0017- DC N4	A0226-ASP OD1
			m2oaaA0	D-001- DC N4	A0207-ASP OD2
			m2oaaA0	C-002- DC N4	A0224-ASP O
			m3c25A0	D0012- DC N4	A0187-ASP OD2
			m3imbD0	L-001- DC N4	D0200-ASP OD2
			m3imbD0	K-002- DC N4	D0215-ASP O
	Minor	Single	1bhmA00	D0005- DG N2	A0196-ASP OD2
			1pviA00	D0008- DG N2	A0034-ASP OD1
			m3goxB2	C0002- DG N2	B0162-ASP OD1
			m3imbD0	K0000- DG N2	D0032-ASP OD1
			m3imbD0	K0001- DG N2	D0033-ASP OD2
MS	Major	Single	1hjbC00	H0008- DC N4	C0171-ASP OD2
			h3vebA0	N0007- DC N4	A0108-ASP OD2
			h4ix7A0	D0004- DC N4	A0351-ASP OD1
			m4lmgD0	H0022- DC N4	D0078-ASP OD2
NS	Minor	Single	1cezA01	T0005- DG N2	A0240-ASP OD2

Table S7: List of protein-DNA hydrogen bonds between glutamate (Glu) and DNA bases (major and minor groove) in highly specific (HS), multi-specific (MS) and non-specific (NS) DNA-binding domains.

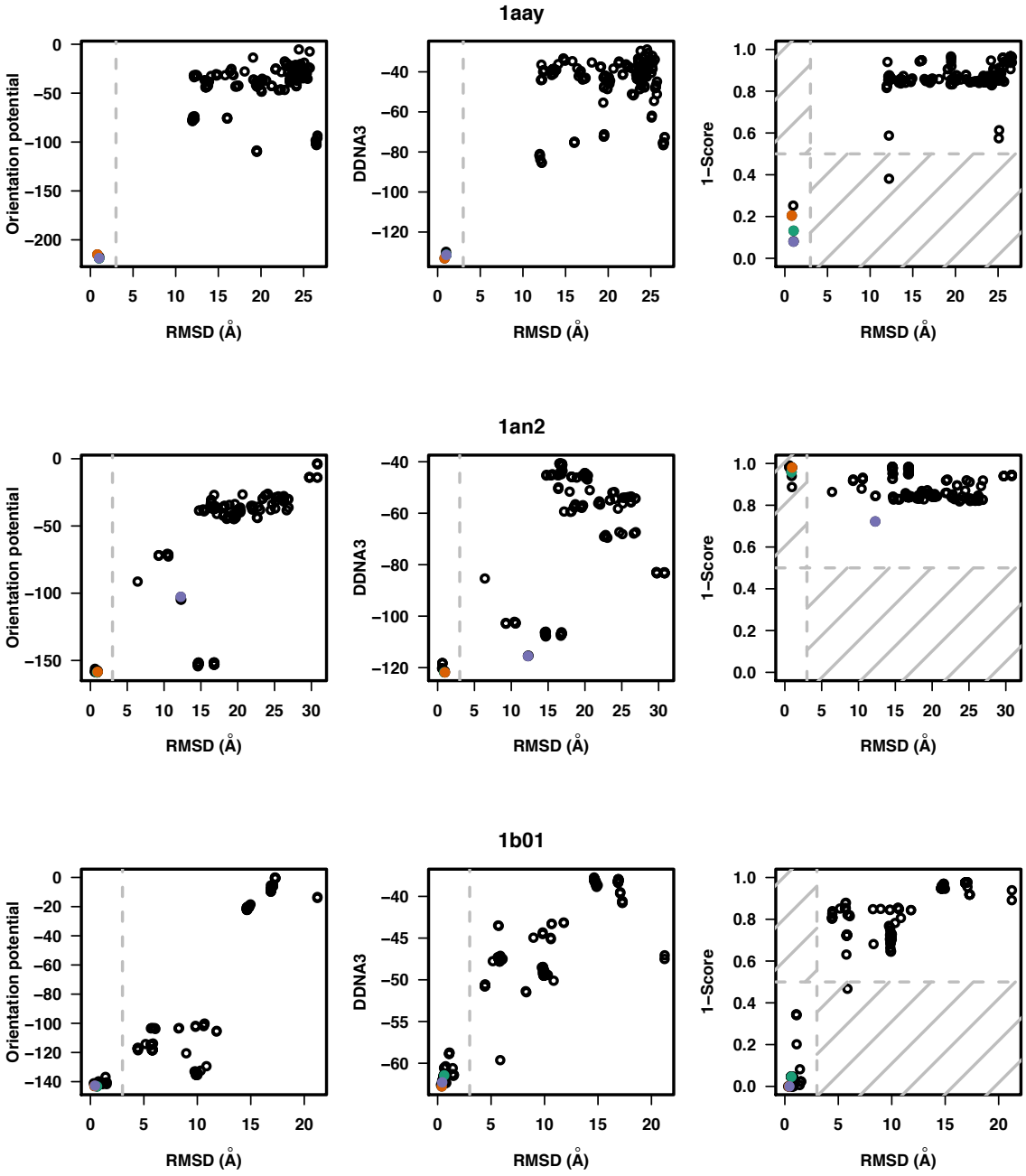
Dataset	DNA groove	HB geometry	Domain ID	Donor	Acceptor
HS	Major	Single	1dc1A01	W0006- DC N4	A0252-GLU OE2
			3hqfA00	C-002- DC N4	A0096-GLU OE1
	Minor		m3ndhA0	D0006- DG N2	A0048-GLU OE1
MS	Major	Single	1le5F01	H0022- DC N4	F0060-GLU OE1
			1nkpD00	J0808- DC N4	D0510-GLU OE1
			1owrP01	F5011- DC N4	P0427-GLU OE1
			1zreA02	X0006- DC N4	A0181-GLU OE1
			h4gclD0	Z0034- DC N4	D0045-GLU OE1
			h4h10A0	D0308- DC N4	A0081-GLU OE2
			h4hf1A0	D0007- DC N4	A0043-GLU OE2
			1kb2A00	D0431- DA N6	A0042-GLU OE1
			1rioH00	T0008- DA N6	H0410-GLU OE2
			m4lmgD0	H0020- DA N6	D0075-GLU O
NS	Minor	Single	m2o8bB1	F0023- DT N3	B0434-GLU OE2

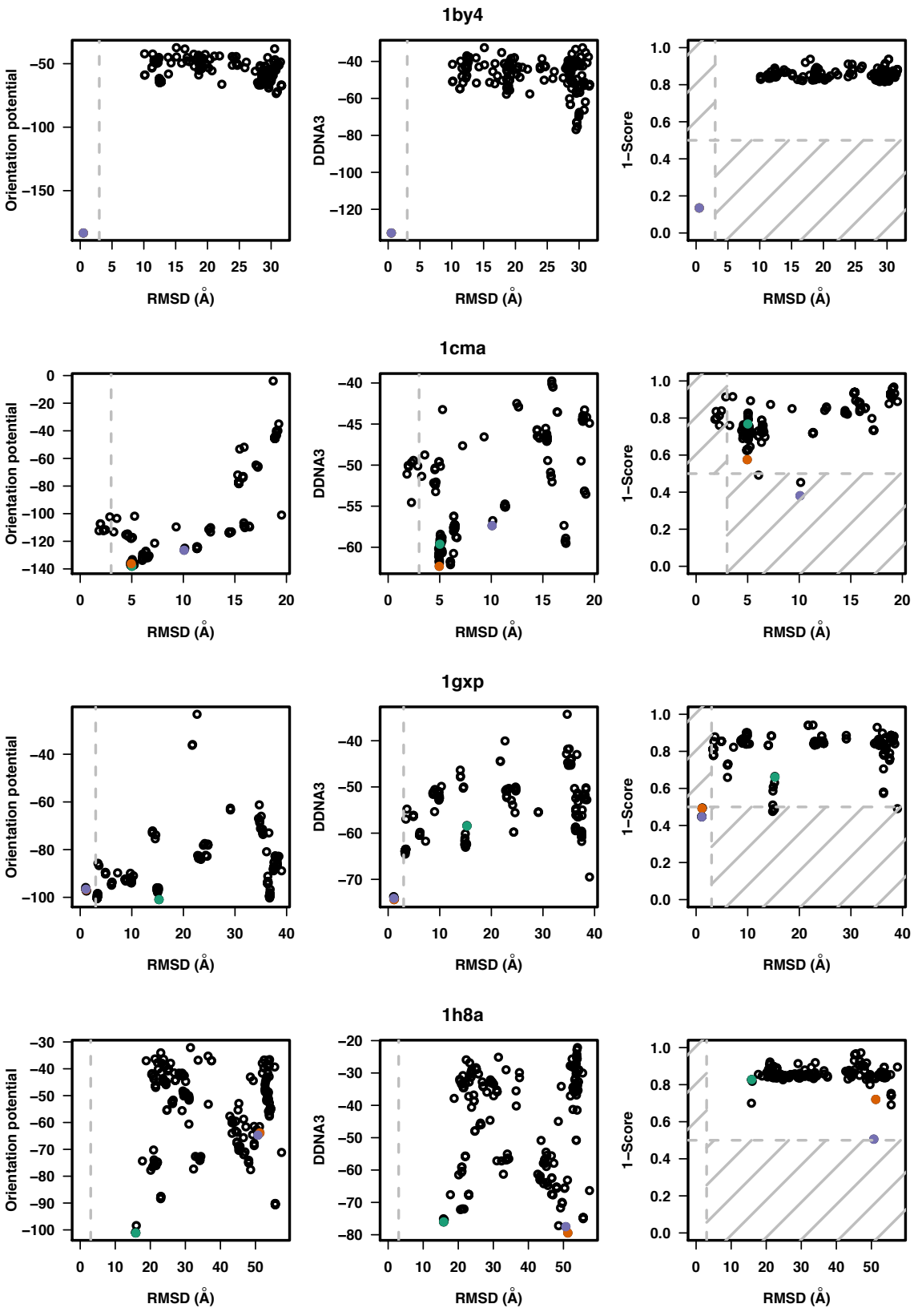
Table S8: List of protein-DNA hydrogen bonds between histidine (His) and DNA bases (major and minor groove) in highly specific (HS), multi-specific (MS) and non-specific (NS) DNA-binding domains.

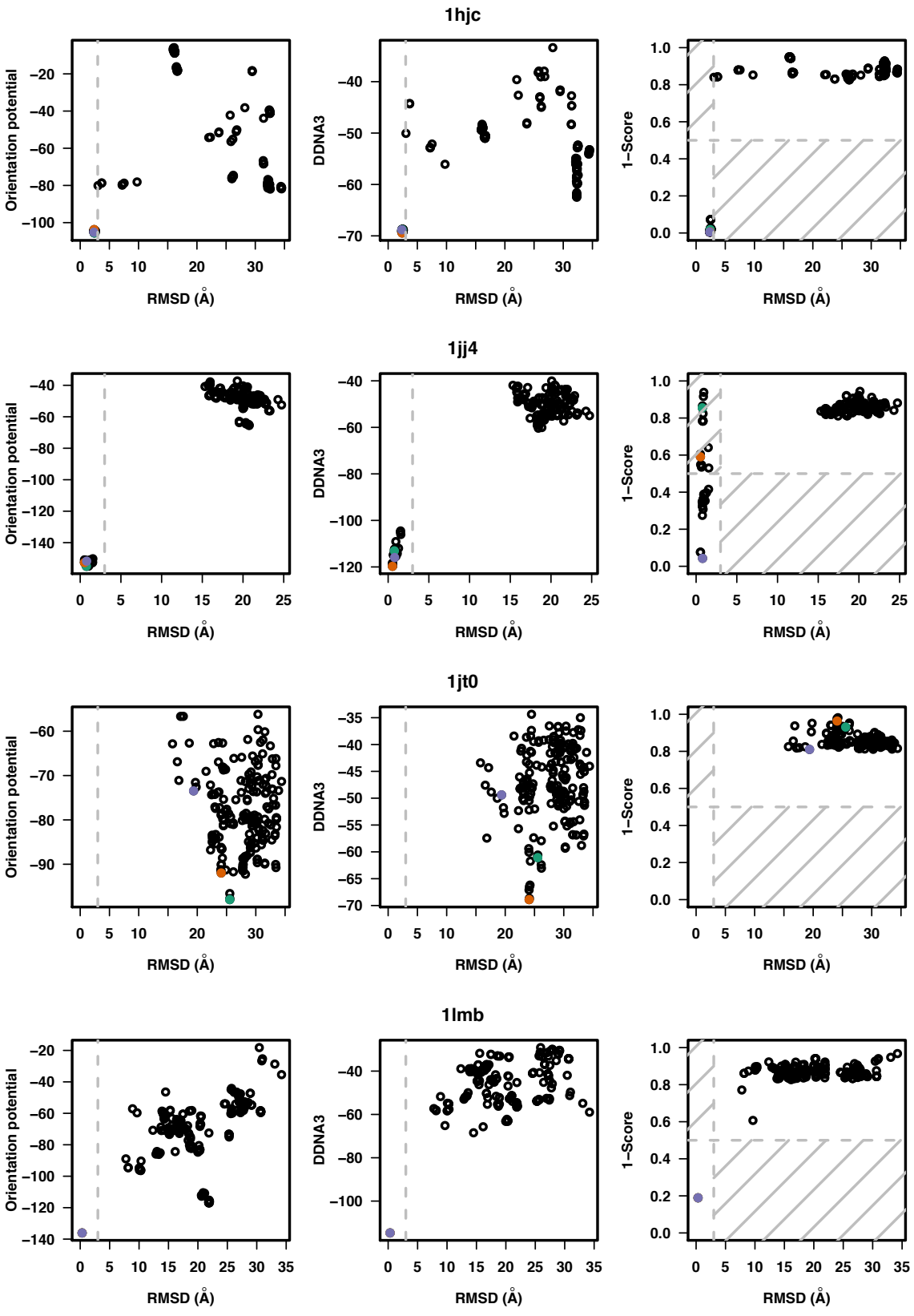
Dataset	DNA groove	HB geometry	Domain ID	Donor	Acceptor
HS	Major	Bidentate	3hqfA00	A0036-HIS N	C0001- DG N7
				B-002- DC N4	A0036-HIS O
				A0036-HIS ND1	C0002- DG O6
			m2oaaA0	C-001- DC N4	A0225-HIS O
				A0225-HIS ND1	D0001- DG O6
		Single	1dc1A01	A0253-HIS NE2	C0008- DG N7
			1pviA00	A0084-HIS ND1	C0008- DG O6
			m2oaaA0	A0223-HIS NE2	D0002- DG O6
			m3c25A0	A0189-HIS ND1	C0008- DG O6
			m3imbD0	D0077-HIS NE2	K0000- DG N7
				D0214-HIS NE2	L0002- DG O6
				D0219-HIS NE2	K0000- DG O6
MS	Major	Bifurcated	1le5F01	F0064-HIS ND1	G0002- DG N7
					G0002- DG O6
		Single	1k78A01	A0062-HIS NE2	C0011- DG O6
			1nkpD00	D0506-HIS NE2	H0613- DG O6
			3coaC00	C0215-HIS ND1	A0005- DT O4
			3pvvB00	B0470-HIS NE2	F0204- DG N7
			d1odha_	A0067-HIS NE2	C1009- DG O6
			h3zplF0	F0077-HIS NE2	H0008- DG O6
			h4h10A0	A0077-HIS NE2	C0113- DG O6
			m4jcyB0	C0014- DC N4	B0043-HIS NE2
			m4ldxB2	B0136-HIS ND1	D0014- DG O6
	Minor		3a01A00	A0175-HIS NE2	C0014- DG N3
			3u2bC00	C0029-HIS NE2	B0012- DG N3
			m4g92A0	A0276-HIS NE2	E0011- DA N3
NS	Minor	Single	d4klua1	T0005- DC N1	A0034-HIS ND1

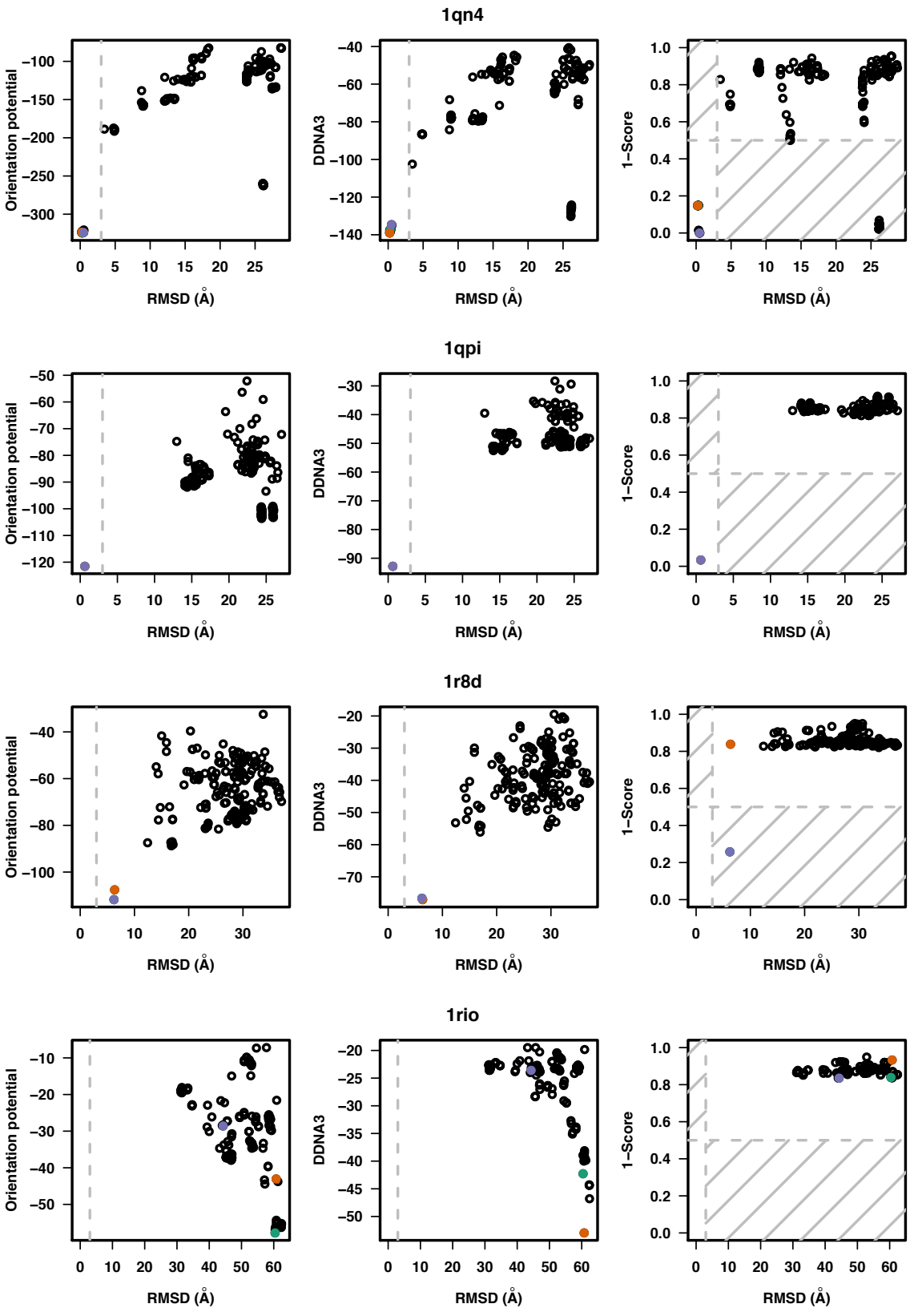
APPENDIX B: SUPPLEMENTARY FIGURES

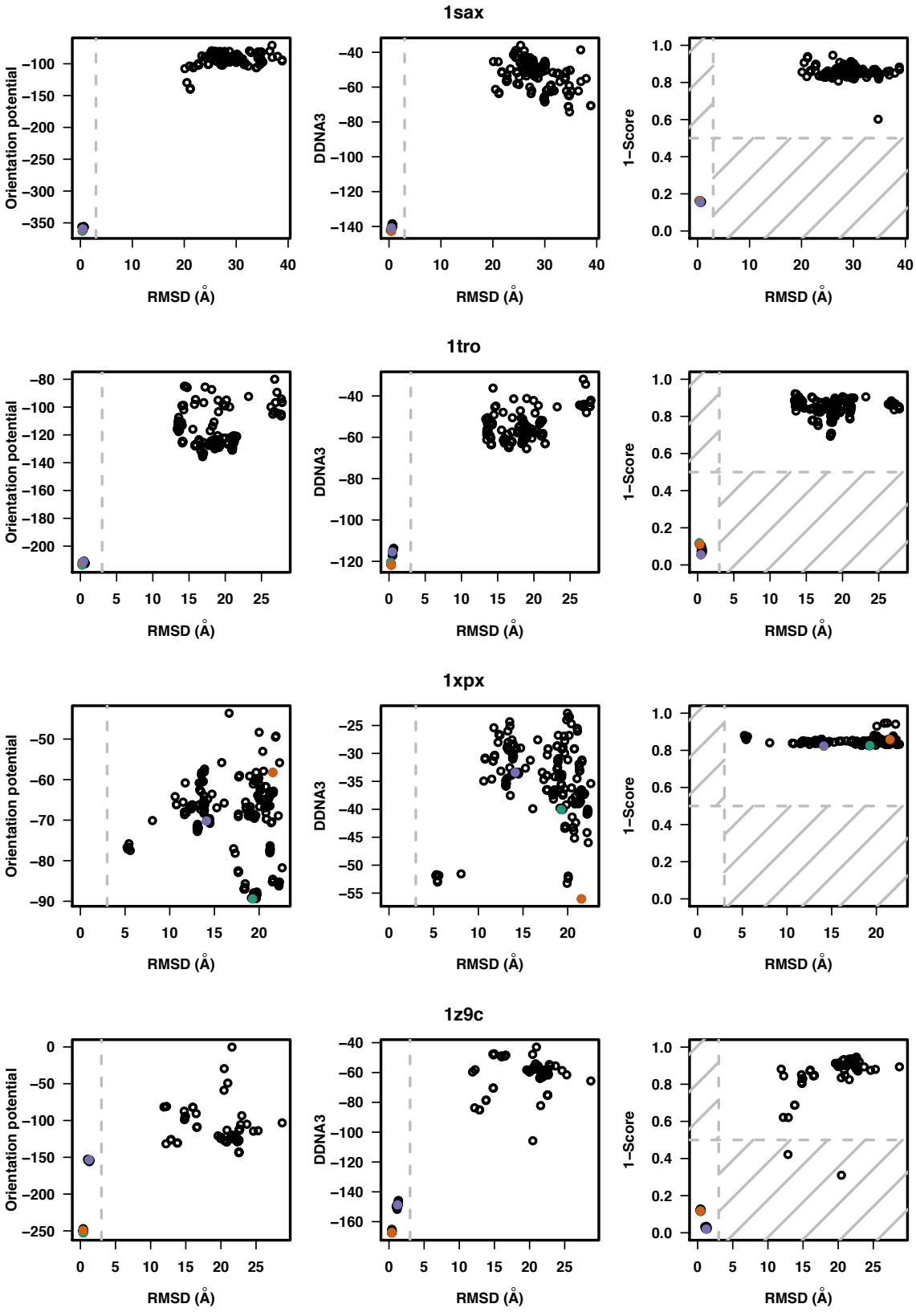
Figure S1: Root mean square deviation (RMSD) vs. orientation potential, DDNA3 potential and predicted SVM quality score for the testing dataset. The conformation with the lowest orientation potential (green), DDNA3 potential (orange) and highest quality score (blue) are highlighted across the three selection methods. The RMSD cutoff is set at 3Å (vertical gray dashed line) and the quality score cutoff value is set at 0.5 (horizontal gray dashed line). False positive and false negative samples, according to the scoring function (rightmost plot), fall in the gray rectangles.

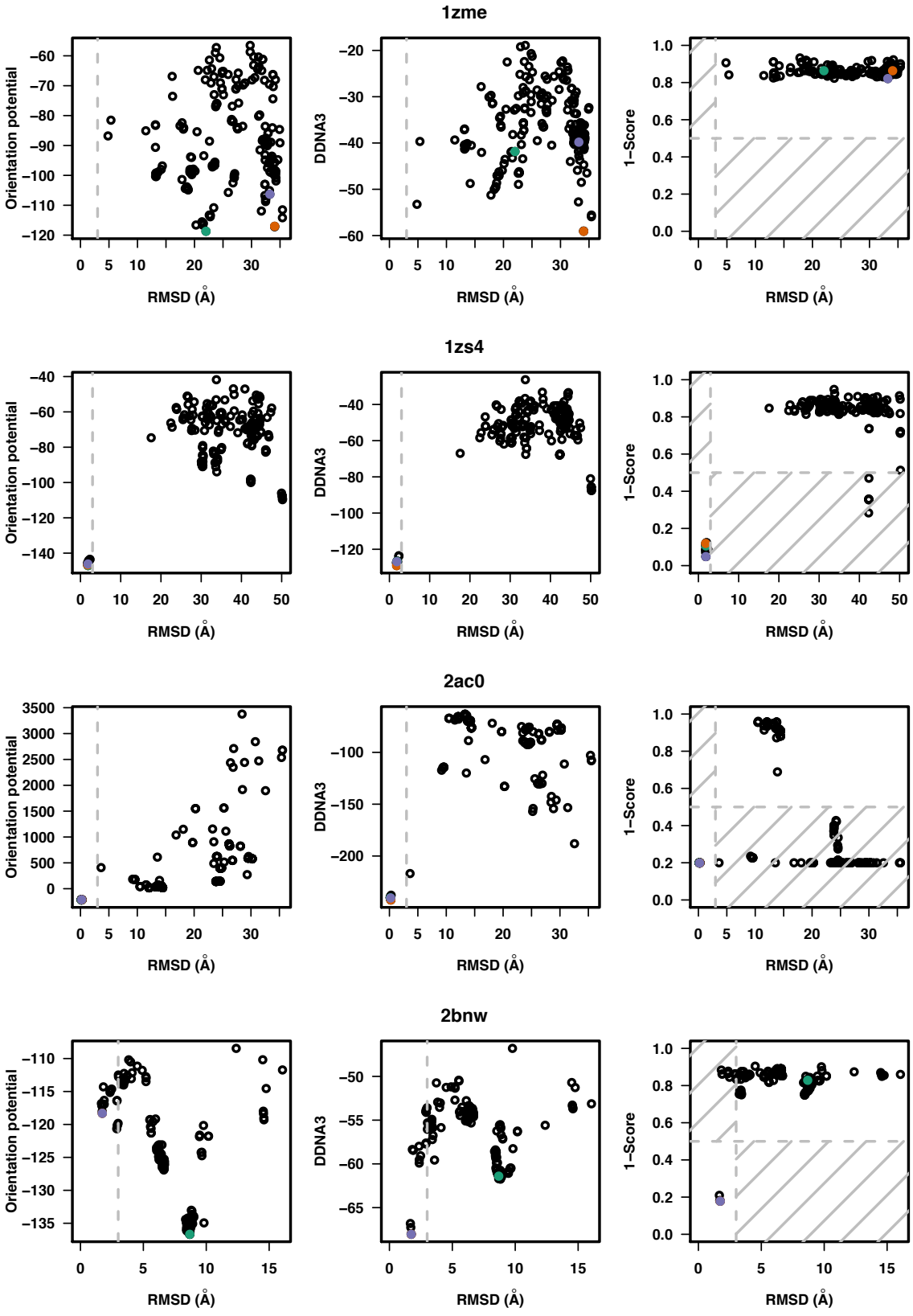


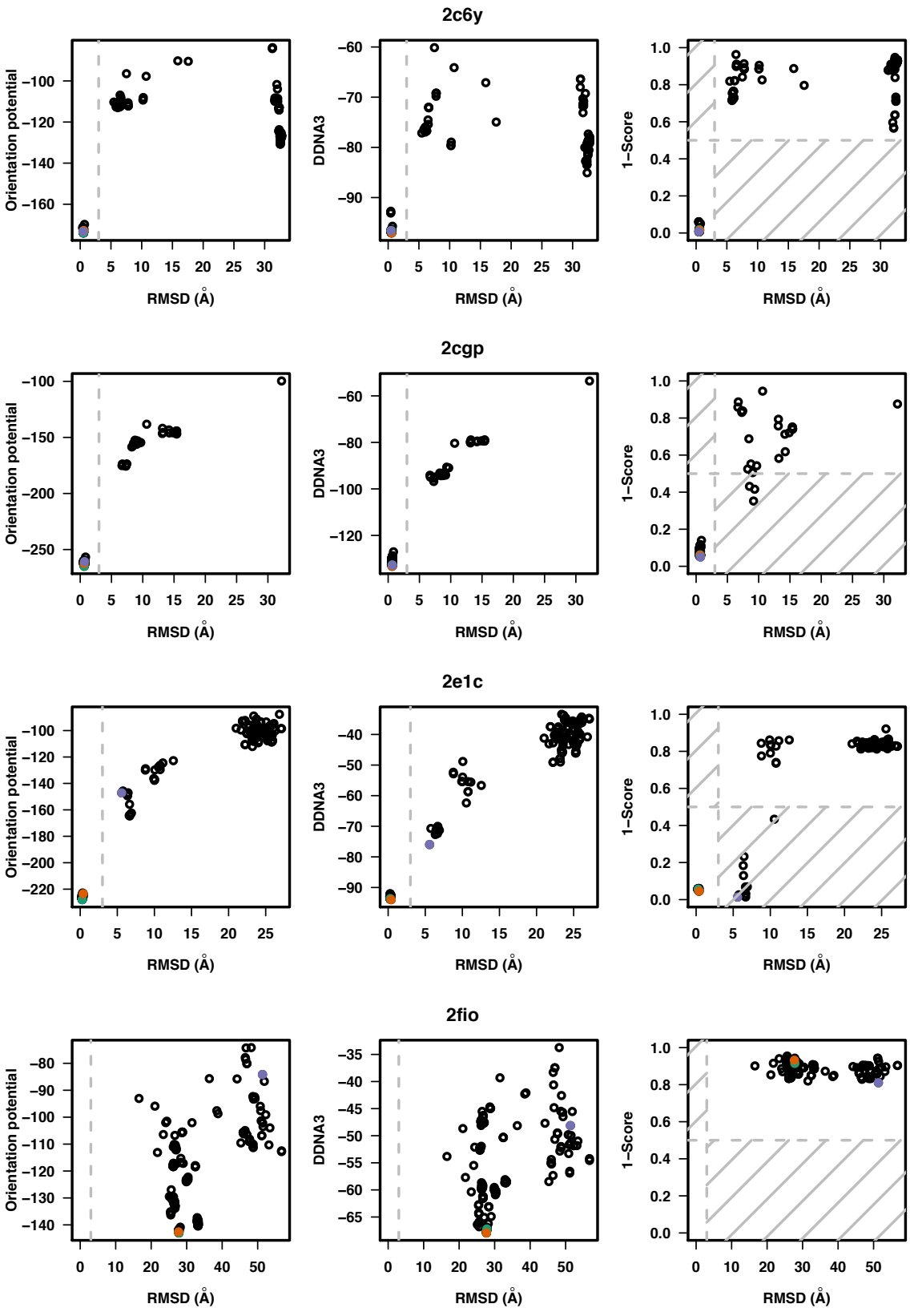

1by4
1cma
1gxp
1h8a
Orientation potential
DDNA3
1−Score
RMSD (Å)

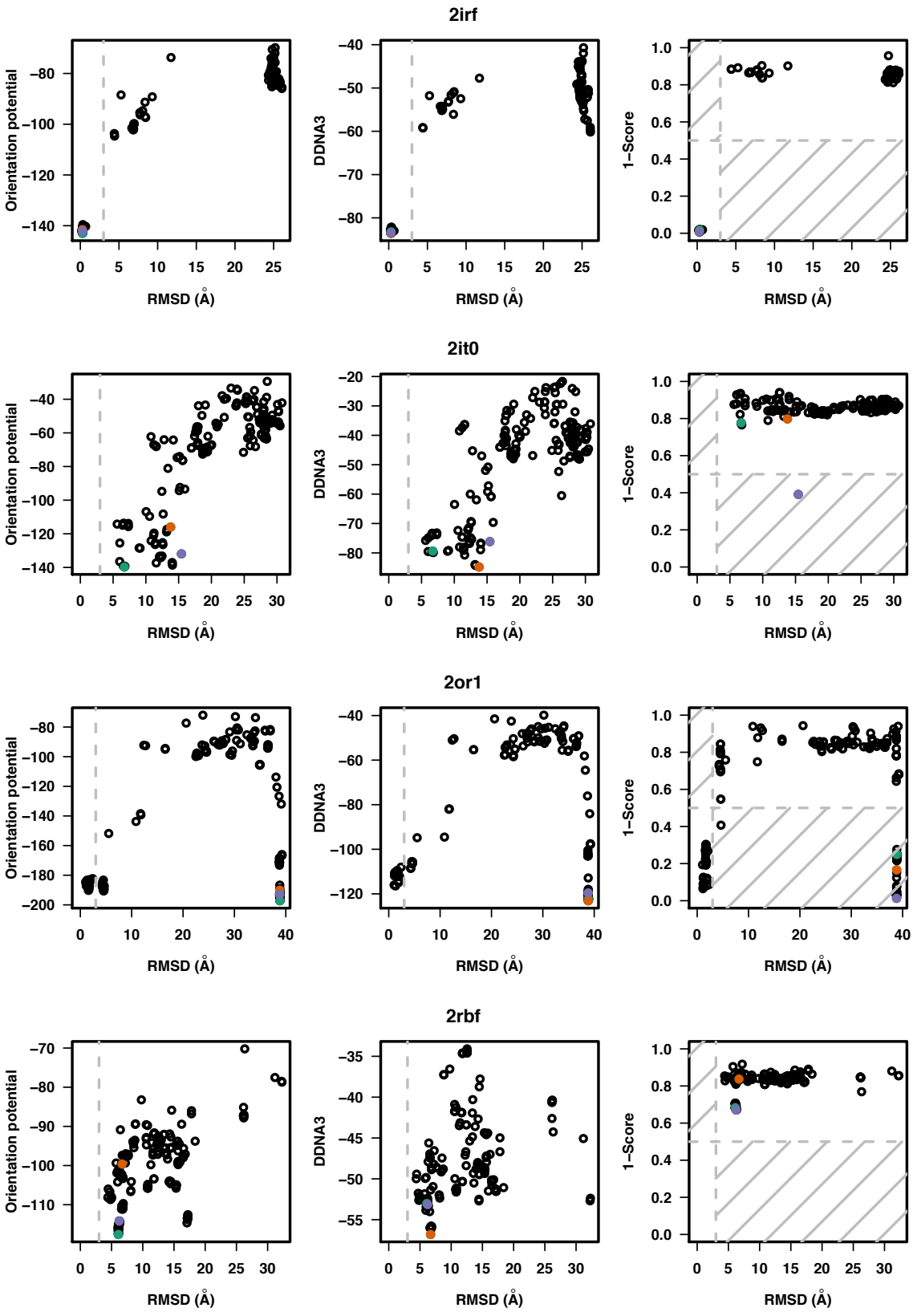

1hjc
Orientation potential
DDNA3
1−Score
RMSD (Å)
1jj4
Orientation potential
DDNA3
1−Score
RMSD (Å)
1jt0
Orientation potential
DDNA3
1−Score
RMSD (Å)
1lmb
Orientation potential
DDNA3
1−Score
RMSD (Å)

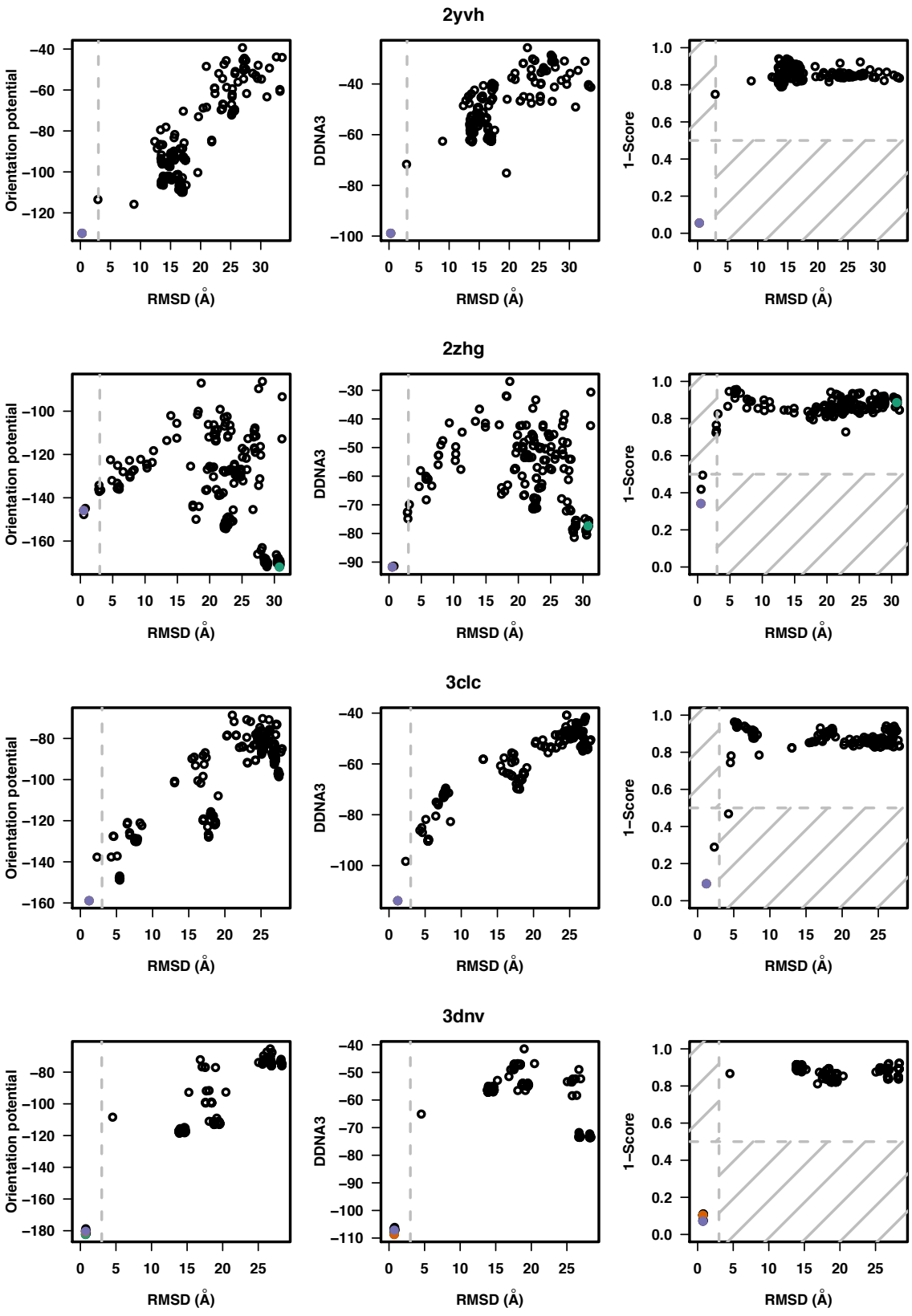

1qn4
Orientation potential
DDNA3
1−Score
RMSD (Å)
1qpi
Orientation potential
DDNA3
1−Score
RMSD (Å)
1r8d
Orientation potential
DDNA3
1−Score
RMSD (Å)
1rio
Orientation potential
DDNA3
1−Score
RMSD (Å)

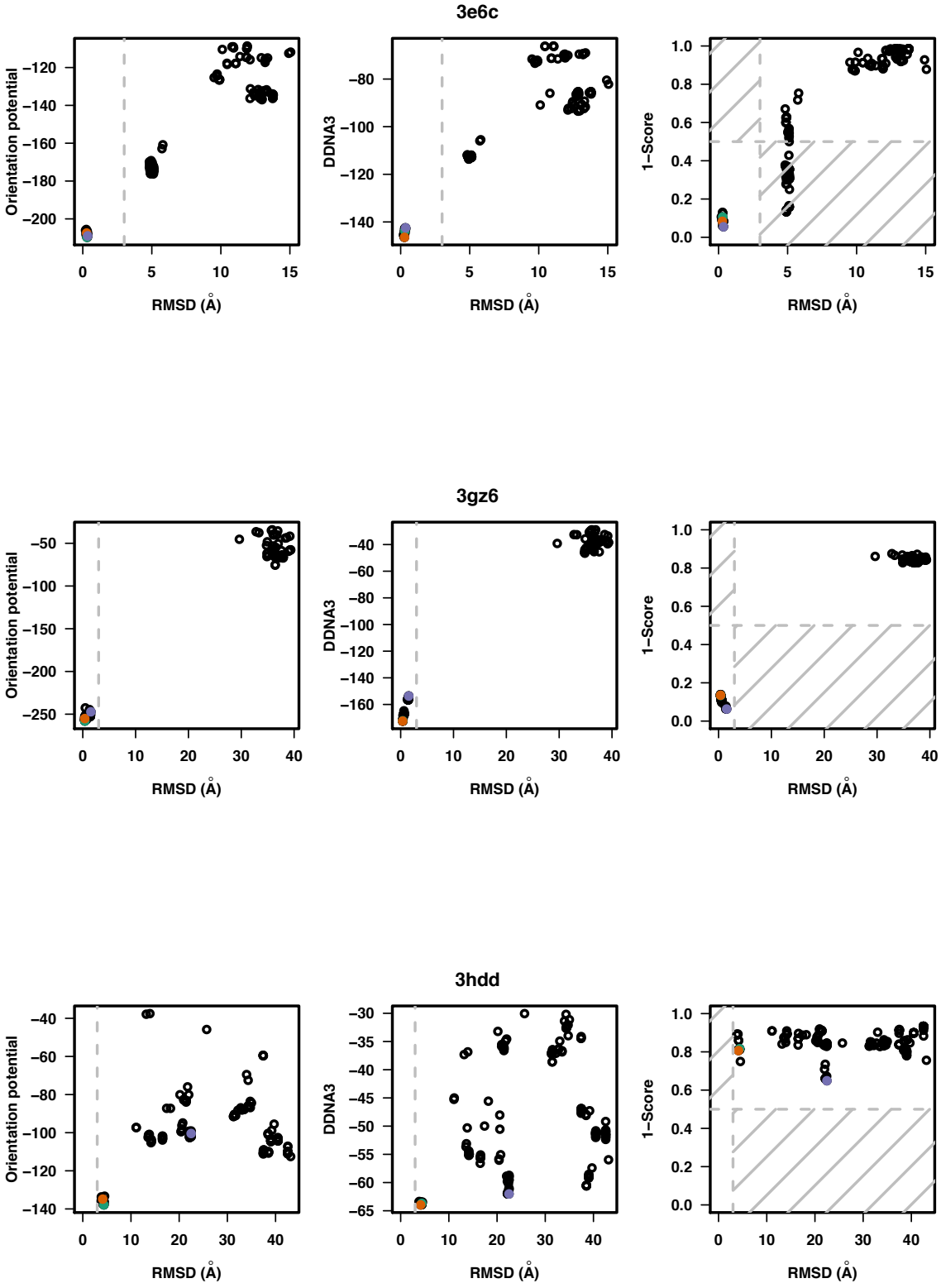

1sax
Orientation potential
DDNA3
1−Score
RMSD (Å)
1tro
Orientation potential
DDNA3
1−Score
RMSD (Å)
1xpx
Orientation potential
DDNA3
1−Score
RMSD (Å)
1z9c
Orientation potential
DDNA3
1−Score
RMSD (Å)


1zme
Orientation potential
DDNA3
1−Score
RMSD (Å)
1zs4
Orientation potential
DDNA3
1−Score
RMSD (Å)
2ac0
Orientation potential
DDNA3
1−Score
RMSD (Å)
2bnw
Orientation potential
DDNA3
1−Score
RMSD (Å)


2c6y
Orientation potential
DDNA3
1−Score
RMSD (Å)
2cgp
Orientation potential
DDNA3
1−Score
RMSD (Å)
2e1c
Orientation potential
DDNA3
1−Score
RMSD (Å)
2fio
Orientation potential
DDNA3
1−Score
RMSD (Å)


2irf
2it0
2or1
2rbf
Orientation potential
DDNA3
1−Score
RMSD (Å)


2yvh
Orientation potential
DDNA3
1−Score
RMSD (Å)
2zhg
Orientation potential
DDNA3
1−Score
RMSD (Å)
3clc
Orientation potential
DDNA3
1−Score
RMSD (Å)
3dnv
Orientation potential
DDNA3
1−Score
RMSD (Å)


3e6c
Orientation potential
DDNA3
1−Score
RMSD (Å)
3gz6
Orientation potential
DDNA3
1−Score
RMSD (Å)
3hdd
Orientation potential
DDNA3
1−Score
RMSD (Å)