TOWARD UNDERSTANDING PROTEIN-DNA INTERACTIONS

by

Maoxuan Lin

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

Charlotte

2020

Approved by:

____

Dr. Jun-tao Guo

____

Dr. Jennifer W. Weller

____

Dr. Xiuxia Du

____

Dr. Bao-Hua Song

ABSTRACT

MAOXUAN LIN. Toward understanding protein-DNA interactions. (Under the direction of DR. JUN-TAO GUO)

Knowledge of protein-DNA interactions has important implications in understanding biological activities and developing therapeutic drugs. Two types of protein-DNA interactions exist: (1) interactions between double-stranded DNA-binding proteins (DSBs) and double-stranded DNA (dsDNA), and (2) those between single-stranded DNA-binding proteins (SSBs) and single-stranded DNA (ssDNA). DSB-dsDNA interactions have been extensively studied but are still not completely understood. In contrast, less attention has been paid to SSB-ssDNA interactions. To expand our knowledge of DSB-dsDNA interactions, we investigated the roles of individual DNA strands and protein secondary structure types in specific DSB-dsDNA recognition based on side chain-base hydrogen bonds. By comparing the contribution of each DNA strand to the overall binding specificity, we found that highly specific DSBs show balanced hydrogen bonding with each of the two DNA strands, while multi-specific DSBs are generally biased towards one strand. In addition, amino acids involved in side chain-base hydrogen bonds in these two groups of proteins favor different secondary structure types. To advance our understanding of SSB-ssDNA interactions, we performed a comparative structural analysis on known SSB-ssDNA complex structures. Structural features such as DNA binding propensities and secondary structure types of amino acids involved in SSB-ssDNA interactions, protein-DNA contact area, residue-base contacts, protein-ssDNA hydrogen bonding and π-π interactions, were analyzed and compared between specific and non-specific ssDNA-binding proteins. Our results suggest that side chain-base hydrogen bonds play major roles in protein-ssDNA binding specificity, while protein-ssDNA π-π interactions may contribute to binding affinity. In addition, bound and unbound conformations of the same ssDNA-binding domains were compared to investigate the conformational changes upon ssDNA binding, and the results indicate that conformational changes of ssDNA-binding proteins might not be a major contributor in conferring binding specificity. These studies provide new insights into the mechanisms of specific protein-DNA interactions and can help therapeutic drug design.

DEDICATION

To my beloved wife Ning, daughter Yuyan, parents, parents-in-law, and all my family.

ACKNOWLEDGEMENTS

Thank you to my research advisor, Dr. Jun-tao Guo, for great research opportunities and the guidance through each stage of my graduate study. Special thanks to my wonderful committee members: Dr. Jennifer W. Weller, Dr. Xiuxia Du, and Dr. Bao-hua Song, for their time, suggestions, encouragement, and inspiration.

I am grateful for the professional support from Dr. Aaron Brink and Dr. Lynne Harris at the Center for Counseling at Psychological Services (CAPS), the academic writing support from Dr. Lisa Russell-Pinson at the Graduate School and Chris Harrington at the Writing Resources Center, and the academic advising and support from my academic advisors Dr. Anthony Fodor and Dr. Cynthia Gibas and graduate coordinator Ms. Lauren Slane at the Department of Bioinformatics and Genomics. Thanks also to all current members and alumni of GUO lab for the discussions we had throughout the years.

I would like to extend my appreciation to the University of North Carolina at Charlotte, the Graduate School, the National Science Foundation (NSF), the National Institutes of Health (NIH), and my research advisor for the financial support, and to the administrative staff of the Graduate School, the International Students and Scholars Office and the Bioinformatics and Genomics Department for their help.

And my biggest thanks to my family for all the understanding, support and sacrifices through my doctorate study. For my daughter Yuyan, sorry for being away and missing so many milestones in your childhood. And for my parents and parents-in-law, thank you very much for taking great care of the little girl. To my wife Ning, thanks for all your support, encouragement, and discussions about science and life. You are all amazing! I am looking forward to our family reunion!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

DNA, the hereditary material of life, exists primarily in a double-stranded form (known as dsDNA and also referred to as a double helix), which is recognized and bound by proteins to achieve such biological functions as gene regulation and chromosome packaging [2]. As these proteins bind DNA in a double-stranded form, they are called double-stranded DNA-binding proteins (DSBs). In cellular processes, such as DNA replication, recombination, and repair, however, the DNA double helix needs to be unwound, and two complementary DNA strands are exposed in a single-stranded form as metabolic intermediates [3–5]. Single-stranded DNA (ssDNA) is vulnerable to chemical and enzymatic attacks and is prone to form secondary structures, including hairpin via intra-strand pairing. As a consequence, a second group of DNA-binding proteins, single-stranded DNA-binding proteins (SSBs), has evolved to bind to and stabilize these intermediates. In addition, SSBs are essential in the maintenance of genomic stability, especially in telomere end protection [6, 7], and the recruitment of other proteins to modulate DNA metabolic processes [8].

Knowledge of interactions between DNA-binding proteins and DNA forms the basis of our understanding of these biological processes and mechanisms of diseases that are controlled by protein-DNA interactions. Yet, despite efforts over the past several decades, our knowledge of protein-DNA interactions is not complete, especially for those between SSBs and their target ssDNA. Therefore, the focus of this dissertation is to investigate the interactions between DNA-binding proteins and DNA from a structural perspective, with the goals of expanding our knowledge of relatively well-studied DSB-dsDNA interactions and advancing our understanding of less-studied SSB-ssDNA interactions.

1.1 Important features of protein-DNA interactions

1.1.1 Protein-DNA binding specificity

Of particular interest regarding protein-DNA interactions is the specific recognition of proteins and their target DNA sequences, also known as protein-DNA binding specificity. The binding specificity refers to the ability of a protein distinguishing its functional binding sites from the remaining non-functional potential sites in the genome [9]. The calculation of binding specificity depends on another important feature of protein-DNA interactions, the binding affinity. The binding affinity is defined as the dissociation constant K_d, in which products of concentrations of free protein and free DNA are divided by the concentration of protein-DNA complex at equilibrium [9]. To estimate the complete binding specificity, the binding affinity of at least enough, if not the entire set of, potential binding sites should be known first.

To measure the binding affinity of these binding sites, experimental technology advances have been made over the last few decades, including several high-throughput approaches. These technologies can be classified into four groups: (I) Direct affinity measurements, including the mechanically induced trapping of molecular interactions (MITOMI) [10] and surface plasmon resonance [11], (II) Microarray-based methods, such as protein-binding microarrays (PBMs) [12] and the cognate site identifier (CSI) [13], (III) In vitro selection, including systematic evolution of ligands by exponential enrichment (SELEX) [14], SELEX-serial analysis of gene expression (SELEX-SAGE) [15], SELEX combined with massively parallel sequencing (SELEX-seq) [16], and high-throughput SELEX (HT-SELEX) [17, 18], and (IV) Bacterial one-hybrid selections (B1H) [19, 20].

Based on the measurement of binding affinity using these methods, the binding specificity of DNA-binding proteins can be calculated and represented as position weight matrices (PWMs) [21,22], as shown in Table 1.1, which can be better visualized as sequence logos (Figure 1.1) [23]. PWMs and logos show preferences to certain nucleotides at different binding sites. For instance, thymine (T) is the exclusive nucleotide at position 2 of the binding site for MEF2A (Table 1.1). Accordingly, only T is shown at position 2 in the sequence logo (Figure 1.1). These matrices and logos provide an informative way of visualizing binding site conservation and easiness of human interpretation of protein-DNA binding specificity.

Table 1.1: Position weight matrix (PWM) of binding sites of MEF2A (Source: JASPAR [1]).

Position	1	2	3	4	5	6	7	8	9	10
A	1	0	57	2	9	6	37	2	56	6
C	50	0	1	1	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	2	50
T	7	58	0	55	49	52	21	56	0	2

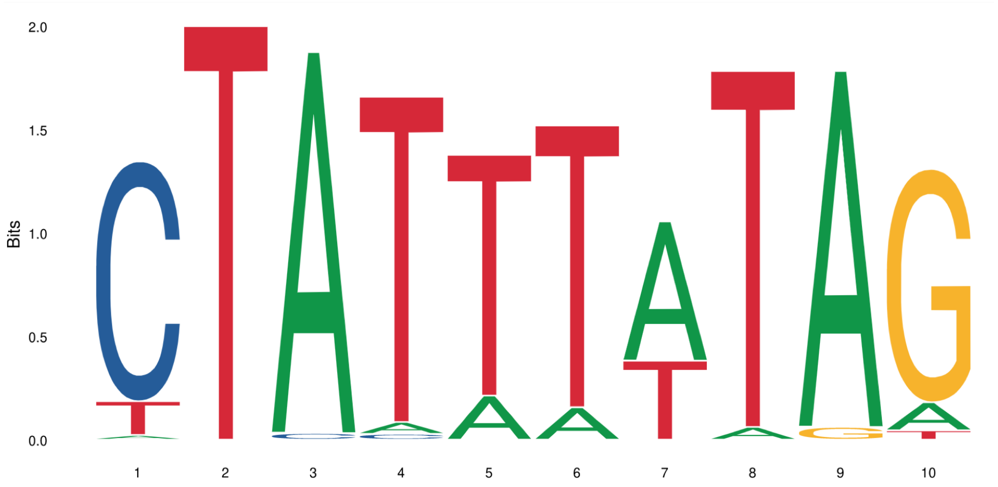


Figure 1.1: Sequence logo of binding sites of MEF2A (Source: JASPAR [1]).

1.1.2 DNA-binding domains

DNA-binding specificity can be investigated at the protein chain or domain level [24,25]. A protein domain is a conserved, independently evolved and folded structural and/or functional unit of a given protein, and domains that bind with DNA are called DNA-binding domains. While the whole chains of many DNA-binding proteins form single DNA-binding domains, chains of other DNA-binding proteins contain additional domains, such as dimerization domains and signal-sensing domains. Therefore, it would be critical to know if chain-based analysis and domain-based analysis of DNA binding specificity are consistent. By analyzing 830 binding profiles of human transcription factors, Jolma *et al.* found that full-length transcription factors and DNA-binding domains bind similar sequences, and they suggested that analysis of DNA-binding domains is sufficient to determine the protein-DNA binding specificity of transcription factors [24]. This study also indicates that future studies of DNA-binding specificity can adopt a domain-based approach, which can capture the overall binding specificity of DNA-binding proteins and avoid the confounding effects from non-DNA-binding domains.

Domain-based analyses rely on accurate annotation of DNA-binding domains. Two most widely used databases for classifying protein structures are SCOP [26] and CATH [27]. Both partition proteins into domains, but these domains are classified into two different hierarchies in these two databases: SCOP sorts domains in a hierarchy from class down to fold, superfamily and family while CATH clusters domains in class, architecture, topology and homologous superfamily. In SCOP, domains in the same class contain the same content of secondary structures, and domains with the same arrangement and topological connections of secondary structures are further assigned into the same fold. Domains in the same superfamily share low sequence identities but sufficient structural similarities that suggest a common evolutionary origin, while domains clustered in the same family are more closely related than superfamilies based on sequence similarity and/or functional evidence. In CATH, domains in the same class have the same secondary structure contents. The second level, architecture, clusters domains in the same class with common orientation of secondary structures but does not take connectivity into account. The topology level, analogous to the fold level in SCOP, considers both the number and topological connections of secondary structures. The last level, homologous superfamily, groups domains with a high structural similarity and similar functions that suggest a common ancestor. Differences between SCOP and CATH also include the curation methods: SCOP is mainly manually curated but starts to apply automated curation methods to structures released since SCOP 1.75, with manual curation performed to correct the errors. CATH, on the other hand, contains more automatic steps and thus has a faster update rate. In addition, CATH in general assigns more but smaller domains than SCOP.

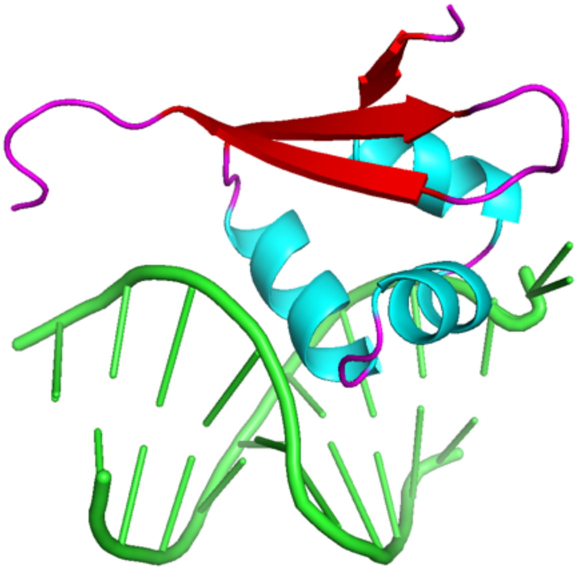


Figure 1.2: Structural example of helix-turn-helix DNA-binding motif in a protein-dsDNA complex (PDBID: 6CRO). Helices are colored in cyan, while sheets and coils are labeled in red and magenta, respectively. The DNA double helix is colored in green.

Despite these differences, both SCOP and CATH classify domains based on secondary structure contents and their topological connections. Secondary structures are defined based on local hydrogen bond patterns between atoms of the backbone. Different secondary structure types have been assigned by different assignment algorithms, and a widely used algorithm is the dictionary of protein secondary structure program (DSSP) [28]. DSSP assigns eight states of secondary structures, which can be classified into three major types—helix, strand, and coil—following the widely used convention: H (α-helix), G (3_{10}-helix) and I (π-helix) states as helix type, E (extended strand) and B (residue in isolated β-bridge) states as strand type, and all the other states from DSSP are considered as coil types [28–31].

A structural example of protein-DNA complex containing a helix-turn-helix DNA-binding motif, a major motif used by transcription regulators and enzymes of prokaryotes and eukaryotes, is shown in Figure 1.2. The Cro protein from bacteriophage lambda contains all three major secondary structure types, and it recognizes the target dsDNA with the helix-turn-helix motif. Similar to the Cro protein, all DNA-binding proteins contain DNA-binding domains featured by different combinations of these secondary structure types, and detailed information of DNA-binding domains of DSBs and SSBs is described in the following sections.

1.2 DSB-dsDNA interactions

1.2.1 Binding specificity of double-stranded DNA-binding proteins

With the advancement of technologies measuring the binding affinity and the increase of high-resolution structures of DSB-dsDNA complexes in Protein Data Bank (PDB) [32, 33], our knowledge of DNA-binding specificity for DSBs has been greatly expanded. Specific recognition of DSBs towards their target binding sites relies on the combination of two readout mechanisms: base readout and shape readout [34, 35]. Base readout refers to the direct interaction between protein and DNA bases, mainly achieved via hydrogen bonds between amino acid side chains and DNA bases and π-interactions between aromatic amino acids and DNA bases [25, 36–42]. Although no simple rules exist for one-to-one correspondence between amino acids and DNA bases, some amino acids show preferences to specific bases, such as arginine with guanine, and both asparagine and glutamine with adenine [42–45]. On the other hand, shape readout refers to both global and local shape of target DNA sequences in protein-DNA recognition [46–52]. DNA shape readout is achieved by both intrinsic and protein-induced DNA deformations in the core binding motifs as well as their flanking regions, especially the A- or T-rich stretch in these regions [24, 47, 53].

Besides these two readout mechanisms, it was found that epigenetic effects, such as CpG methylation [54] and homodimer orientation and spacing [24], also affect protein-DNA binding specificity. Moreover, protein flexibility has been reported to contribute significantly to specific protein-DNA recognition [55–59]. Based on this knowledge, various models have been developed for binding site prediction [24, 39, 48, 60–63]. While performances of these models vary, adding shape features improves prediction accuracy over the sequence-only models.

In addition to these general binding principles, efforts have also been made to explore the binding mechanisms underlying DSBs with different levels of binding specificity [25, 64]. Luscombe and Thornton assigned 21 DSB families to three classes based on their binding specificity and compared the conservation of amino acids in contact with DNA bases [64]. These three classes are: (I) highly specific (HS), where binding is specific and all members of a family bind to the same DNA sequence; (II) multi-specific (MS), where binding is also specific but allows members in the same family to bind different sequences; and (III) non-specific (NS), where the binding of members in the family is independent of sequence. They found different patterns among these three classes: (1) DNA-contacting residues in HS families are highly conserved so that members in these families can bind to the same target sequence; (2) DNA-contacting residues in MS families are frequently mutated to enable members in these families to bind different target sequences; and (3) even though DNA-contacting residues in the NS families are also well conserved, the binding is mainly found in the minor groove where different base types can not be differentiated from each other [64]. While this study provides great insights into DNA-binding specificity from the perspective of the conservation of DNA-contacting residues, little information about the roles of DNA and interactions between proteins and DNA is offered.

To expand the knowledge of binding specificity of DSBs with different degrees of binding specificity, Corona and Guo classified 195 non-redundant DSB binding domains into HS, MS, and NS groups using different definitions and compared structural features contributing to binding specificity among these three groups [25]. Of these three groups, the HS group includes most type II restriction endonucleases, recognizing and cleaving foreign DNA at highly specific target sequences [65]; the MS group is mainly composed of transcription factors, binding to specific DNA sequences while allowing variations at certain positions [34]; and the NS group, consisting proteins like histones and DNA polymerases, does not discriminate DNA sequences for binding. Structural features compared include amino acid propensities, simple and complex hydrogen bonds, major/minor groove and base contacts, and DNA shape. This study shows a clear trend of structural features among these three classes: DSBs with higher binding specificity form more hydrogen bonds, have more major groove and base contacts, and harbor larger DNA shape changes [25]. In addition, specific DSBs, including HS and MS groups, show larger conformational changes upon DNA binding and have larger degree of flexibility [25].

While current studies have greatly expanded our knowledge of DSB-dsDNA binding specificity, these studies consider the dsDNA as a whole binding unit. This strategy, however, may preclude the complete understanding of specific DSB-dsDNA binding. This concern is supported by findings that some dsDNA sequences and their corresponding single strands can be bound by different DNA-binding proteins [66–69]. These findings indicate that two DNA strands may play different roles in specific DSB-dsDNA binding. However, little is known about the roles of single DNA strands of the double helix in specific DSB-dsDNA recognition. Therefore, further investigation of the DSB-dsDNA binding specificity at DNA strand level is needed, and this new perspective would provide new insights into DSB-dsDNA binding specificity.

1.2.2 DNA-binding domains of double-stranded DNA-binding proteins

DSBs show a wide range of DNA-binding domains. Luscombe *et al.* analyzed DNA-binding domains of 240 DSBs and assigned these proteins to eight groups: helix-turn-helix (HTH, including the 'winged' HTH), zinc-coordinating, zipper-type, other α helix, β-sheet, β-hairpin/ribbon, other, and enzymes [2]. These eight groups were further classified into 54 structural families based on pairwise structural alignments [2].

Of these groups, the HTH domains and zinc-coordinating domains are the two most common ones. The HTH domain consists of two almost perpendicular α helices connected by a four-residue long linking turn [70,71]. This motif binds to DNA in the major groove with its second helix, and typically coexists with one to four additional α helices to form a stabilizing hydrophobic core [70]. A structural example of the HTH motif is shown in Figure 1.3A. An extension of the HTH domain is the 'winged' HTH motif, featured by an extra α helix and an additional β sheet, as shown in Figure 1.3B. The 'winged' HTH motif binds to the target DNA similar to the regular HTH motif, with the extra secondary structure elements in contact with the DNA backbone.

In contrast to proteins in the HTH group, structures of zinc-coordinating proteins are more diverse. Zinc-coordinating binding motif distinguishes itself from other DNA-binding domains with the coordination of one or two zinc ions by conserved histidine and cysteine residues [71]. The largest family in this group is the $\beta\beta\alpha$ zinc-finger family. DNA-binding domain of this family consists of an antiparallel β sheet formed by two short β strands and followed by an α helix (Figure 1.3C). The α helix and the second β strand contain two pairs of conserved histidine and cysteine residues, and these conserved residues coordinate a single zinc ion [72]. Similar to the HTH motif, the zinc-coordinating domain also binds DNA with α helix by inserting the α helix into the major groove.

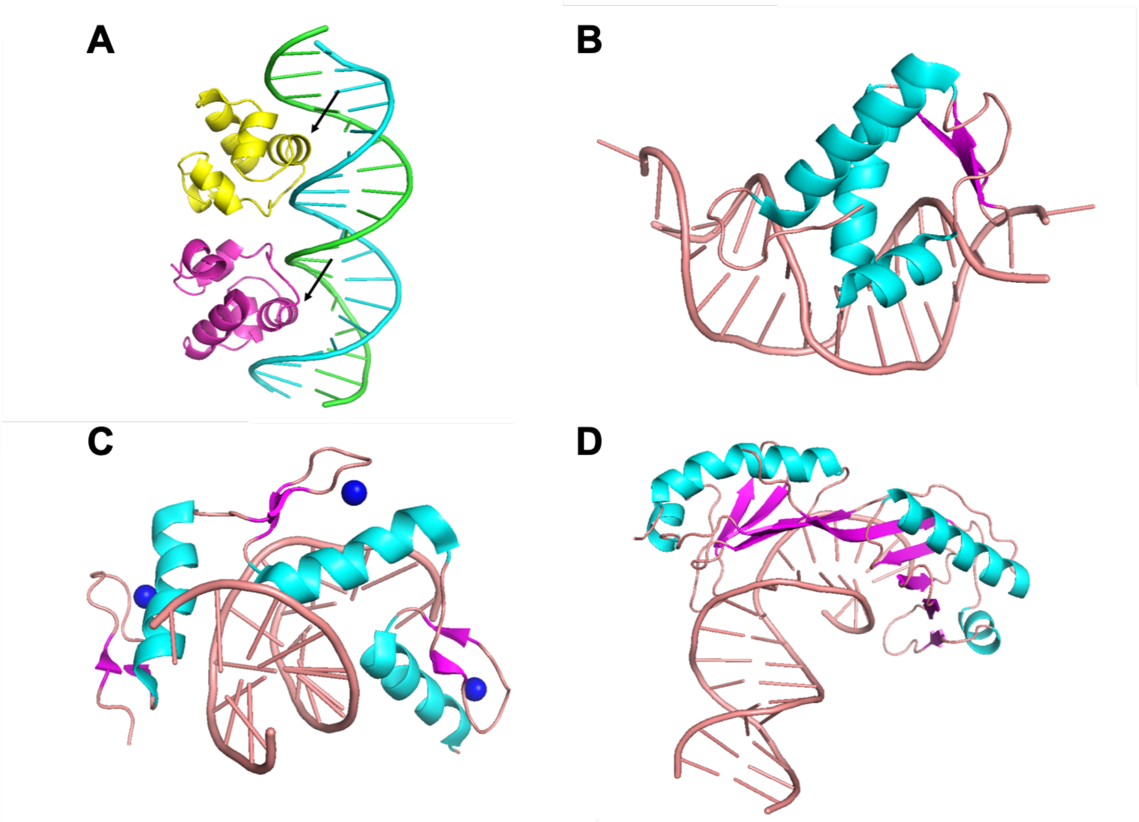


Figure 1.3: Structural examples of common dsDNA-binding domains. (A) The HTH motif. The protein binds as a dimer (colored in yellow and magenta; PDBID: 1PER). Two recognition helices bind in the DNA major groove (black arrows). (B) The ‘winged’ HTH motif (PDBID: 4U0Y). (C) The $\beta\beta\alpha$ zinc-finger motif (PDBID: 4X9J). Blue balls are zinc ions. (D) The β-sheet binding motif (PDBID: 1YTF).

While it is common for DNA-binding proteins to bind DNA with α helices, some proteins bind their target DNA using only β sheets. For instance, a β-sheet protein, containing only the TATA box-binding protein family, binds DNA with a wide β sheet (Figure 1.3D). Compared with the β-sheet proteins, the β hairpin/ribbon proteins bind DNA with smaller two or three-stranded β sheets, either in the major or minor groove [73, 74].

The classification of DSBs into different groups based on their DNA-binding domains provides simple but straightforward standards for comparing DSBs and, more importantly, for predicting and classifying novel DSBs.

1.3 SSB-ssDNA interactions

1.3.1 Binding specificity of single-stranded DNA-binding proteins

Compared with the relatively well-studied binding specificity of DSBs, knowledge of SSBs lags behind. Current knowledge of SSB-ssDNA binding mainly comes from several extensively studied SSBs, such as bacteriophage T4 gene 32 protein (gp32)—the first discovered SSB [75–79], the *E. coli* SSB [80–86], and replication factor A (RPA) [3, 87–92]. Several other studies investigated only a small number of SSBs [6, 93, 94]. These studies show that in general SSBs bind non-specifically to ssDNA, exist in different oligomeric states, and show different binding modes with respect to the quantity of ssDNA substrates [85, 93, 95–98]. For instance, both gp32 and *E. coli* SSB bind ssDNA non-specifically. However, gp32 binds as monomers [76], while *E. coli* SSB functions as a homotetramer [99]. In addition, *E. coli* SSB was found to bind to ssDNA with multiple binding modes: $(\mathrm{SSB})_{35}$, $(\mathrm{SSB})_{56}$, and $(\mathrm{SSB})_{65}$ [100]. These binding modes are affected by many environmental factors, such as salt concentration and type, PH, and temperature [101, 102].

While most SSBs bind ssDNA non-specifically, some SSBs can bind conserved ssDNA with high sequence specificity. Telomere-end protection (TEP) proteins were found to bind specific short, repeated GT-rich ssDNA sequences at the end of enkaryotic telomeres [103,104], likely due to the critical roles of telomeres in the maintenance of chromosomal stability [105]. The achievement of the binding specificity of TEP proteins varies across different organisms. For instance, human POT1 uses both oligonucleotide/oligosaccharide/oligopeptide-binding (OB) folds [106, 107], a characteristic binding unit of SSBs, in the dual-OB fold DNA-binding domain to achieve specific binding, while *Schizosaccharomyces pombe* Pot1 only relies on the first OB fold [108, 109]. OB folds not only contribute to specific binding in SSBs such as TEP proteins, but also involve in non-specific binding in other SSBs including *E. coli* SSB. However, little attention has been paid to the mechanisms underlying the binding discrepancy between SSBs that bind ssDNA independent of sequences and those that bind specific ssDNA sequences. Moreover, previous studies focused on individual SSBs or a small number of SSBs, and to the best of our knowledge, no comprehensive studies of SSB-ssDNA interactions have been conducted.

The lack of comprehensive studies largely limits our understanding of SSB-ssDNA interactions and their roles in genetic activities, and hinders the application of currently derived knowledge into a wider range of SSBs. For instance, an energy-based coarse-grained model was successfully constructed recently for predicting SSB-ssDNA complexes. This model, however, was based on only six SSB-ssDNA complexes, restricting its wider applications, which could be solved by incorporating additional features identified from studies of larger number of SSB-ssDNA interactions [110]. This limitation has been further compounded by the fact that almost all databases and tools built for analyzing protein-DNA complexes explicitly or implicitly exclude SSB-ssDNA complexes. DNAproDB, a well-designed interactive tool for structural analysis of DNA-protein complexes, for example, not only provides users with the automated structural analysis pipeline and visualizations, but also offers powerful searching options. Only until very recently DNAproDB starts to support the analysis of complex structures containing ssDNA [111].

The only database that explicitly curates SSB-ssDNA complex structures is the Nucleic Acid Database (NDB) [112, 113]. The goal of NDB is to archive and distribute three-dimensional structures of nucleic acids and their complexes, manually annotated from primary structural data in PDB [32,33]. In addition to primary data, derived information of nucleic acid structure and function, such as geometric data and classification of structures, and tools and software for analyzing nucleic acids are also available in NDB. With its focus on nucleic acids, NDB provides great services to the community for investigating nucleic acids. However, no information about the interaction between nucleic acids and their binding proteins is provided in NDB, especially about specific SSB-ssDNA interactions. Therefore, a comprehensive study of SSB-ssDNA complex structures in NDB, the largest dataset of SSB-ssDNA complex structures, is needed to gain new insights into SSB-ssDNA interactions.

1.3.2 DNA-binding domains of single-stranded DNA-binding proteins

In contrast to the large number of DNA-binding domains for DSBs, only four common DNA-binding domains have been identified in SSBs: OB folds, K homology (KH) domains, RNA recognition motifs (RRMs), and whirly domains [93, 106, 114–116]. Almost all SSBs contain only one type of these DNA-binding domains and use multiple of these homologous domains to confer full activity [117, 118].

Among these four ssDNA-binding domains, OB folds are the most common ones and are found in many areas of biology and perform multiple functions [106, 107]. These domains range from 70 to 150 amino acids in length and consist of a five-stranded antiparallel β barrel capped by an α helix between the third and fourth β strands. A structural example of the OB-fold is shown in Figure 1.4A. OB fold binds ssDNA with a surface centered on the second and third strands, and the binding shows a conserved polarity, where the 5' end of the ssDNA is closer to the third strand and the 3' end is near the second strand (Figure 1.5). In addition, ssDNA generally binds with the bases toward OB-fold containing proteins while the backbone is exposed to

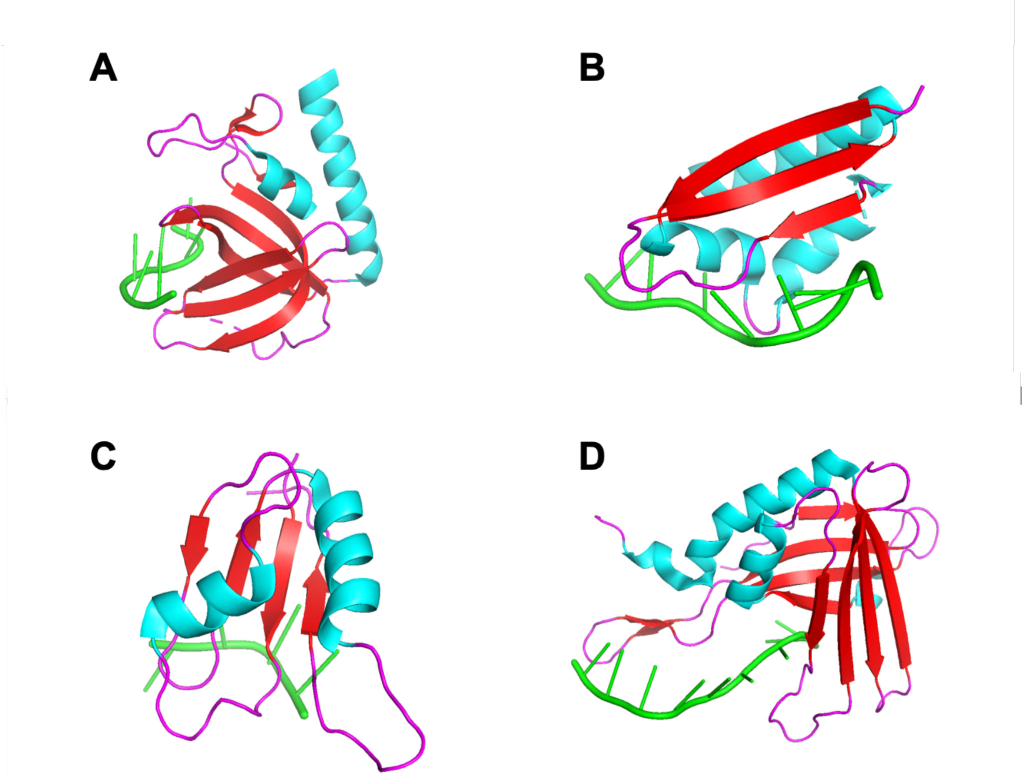


Figure 1.4: Structural representations of common ssDNA-binding domains. Secondary structure types-α helices, β strands, and coils, are colored in cyan, red, and magenta, respectively. The ssDNA is labelled in green color. (A) The OB fold (PDBID: 4GNX). (B) The KH domain (PDBID: 2P2R). (C) The RRM (PDBID: 2L41). (D) The whirly domain (PDBID: 3N1J).

solvent (Figure 1.5).

Compared with OB folds, KH domains are smaller domains with a length of about 70 amino acids. The topology of these domains is featured by three α helices facing a β sheet, which is composed of three strands, as shown in Figure 1.4B [114]. KH domains generally bind to 4-nucleotide (nt) long ssDNA with a core DNA-binding pocket. This core pocket consists of one β strand and two α helices with a loop in between. In addition to the core binding sequence, occasionally one or two additional nucleotides also contact the KH domain [114]. Similar to OB folds, KH domains generally bind ssDNA with nucleotide bases facing the protein with a conserved polarity.

The third ssDNA-binding domain is RRM. RRMs are abundant, presenting in more than 1% of annotated human proteins [119]. RRMs are longer than KH domains ($\sim$

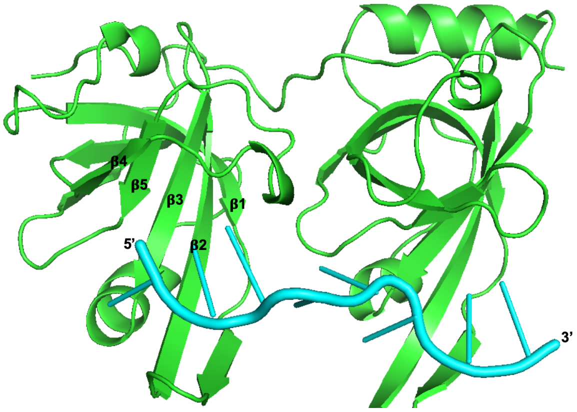


Figure 1.5: OB fold binds ssDNA with polarity. This structure is generated from human replication protein A (RPA70 subunit) in complex with ssDNA (PDBID: 1JMC). The binding shows a conserved polarity with the 5' end of ssDNA closer to strand 3 and the 3' end closer to strand 2.

90 amino acids). Their structures are similar to the OB folds but with a larger β sheet surface, which is composed of four β strands and packed against two α helices (Figure 1.4C). The primary nucleic acid-binding surface is formed by two conserved sequence motifs called ribonucleoprotein domains (RNPs) on the first and third strands, and residues from other parts of the sheet and the loops can also participate in the binding [93]. Similar to the previous two domains, ssDNA also prefers to bind RRMs with bases toward the protein.

Whirly domains, the fourth type of ssDNA-binding domains, are found almost exclusively in a few proteins in plant mitochondria and chloroplasts [116]. Whirly domains are larger domains containing about 180 amino acids, and their structures are characterized by two roughly parallel β sheets (each with four strands) with intervening helices, as shown in Figure 1.4D. Individual whirly domains form tetramers mediated by helices. These tetramers bind 32 nt ssDNA tightly and can also interact further to form hexamers of tetramers [120, 121]. The ssDNA wraps around the tetramers with DNA bases pointing toward protein.

1.4 Summary

As shown in the previous sections, of these two groups of protein-DNA interactions, DSB-dsDNA interactions have been relatively well studied. DSB-dsDNA binding specificity mainly relies on base and shape readout mechanisms, and epigenetic effects such as CpG methylation, as well as homodimer orientation and spacing and protein flexibility also play important roles in the recognition. In addition, significant differences exist between DSBs with different degrees of binding specificity. Knowledge of DSB-dsDNA interactions, however, is limited by the common strategy applied in current studies that considers the dsDNA as a whole binding unit. This limitation precludes our understanding of the roles of single DNA strands of the double helix in DSB-dsDNA recognition. To expand our knowledge of DSB-dsDNA binding specificity, a comparative study of three groups of DSB-dsDNA complexes with different degrees of binding specificity at DNA strand level, was carried out in this dissertation. These three groups of DSBs were updated based on datasets used in [25], and the comparison was focused on the roles of individual DNA strands in specific protein-DNA recognition, based on side chain-base hydrogen bonds. The amount and energy of side chain-base hydrogen bonds, as well as the number of DNA bases and base pairs involved in these hydrogen bonds, were compared to explore the mechanisms underlying binding differences among these groups.

In contrast to DSB-dsDNA interactions, knowledge of SSB-ssDNA interactions lags behind. While most SSBs bind ssDNA non-specifically, some bind their ssDNA substrates with high sequence specificity, such as those binding and protecting the end of telomeres. However, little is known about the mechanisms underlying this binding discrepancy. The lack of this knowledge is mainly due to a shortage of available SSB-ssDNA complex structures. With the increase of SSB-ssDNA complex structures over the past several decades, it is time to perform a comprehensive study of SSB-ssDNA interactions. In this research, we collected all SSB-ssDNA complex structures from

NDB [112,113] and PDB [32,33]. These complex structures were further classified into specific and non-specific groups based on their binding specificity. Next, structural features, such as amino acid propensities, protein-DNA contact area, residue-base contacts, and protein-DNA hydrogen bonding and π-π interactions, were investigated and compared between these two groups. In addition, bound ssDNA-binding domains and their corresponding unbound structures were compared to explore the conformational changes upon ssDNA binding.

Previous studies have demonstrated that DNA-binding proteins bind their target DNA using various DNA-binding domains. However, despite the fact that these domains have been relatively well characterized based on the combinations of secondary structure types, little is known about the distributions and roles of secondary structure types of amino acids involved in protein-DNA recognition. Secondary structure types have been reported to have different levels of resistance to mutations: α helices can tolerate more mutations than β sheets [122]. This finding suggests that secondary structure types may also play different roles in specific protein-DNA recognition and the conservation of the binding specificity. To our knowledge, no studies regarding the roles of secondary structure types in protein-DNA recognition have been reported. To address this shortcoming, we investigated the distributions and roles of secondary structure types of amino acids involved in protein-DNA interactions (including both DSB-dsDNA interactions and SSB-ssDNA interactions) among DNA-binding proteins with different levels of binding specificity. All these studies together provided new insights into protein-DNA binding specificity and may serve as guidance for therapeutic drug design.

CHAPTER 2: NEW INSIGHTS INTO PROTEIN-DNA BINDING SPECIFICITY FROM HYDROGEN BOND BASED COMPARATIVE STUDY

2.1 Copyright Claim

This project has been originally published in the *Nucleic Acids Research* Journal and Oxford University Press is the Publisher (https://academic.oup.com/nar/article/47/21/11103/5609529) [123].

2.2 Introduction

Protein-DNA interactions play crucial roles in many cellular processes, such as transcription, DNA replication, DNA packaging and repair [2]. Of particular interest is the specific recognition between proteins and DNA. Some DNA binding proteins are very specific, which include most type II restriction endonucleases, an important component of the restriction-modification (RM) systems in bacteria. These enzymes recognize and cleave foreign DNA at very specific target sequences while the target sites of the host DNA are protected from cleavage due to methylation [65]. For example, EcoRI and BamHI, two widely used type II restriction endonucleases in molecular cloning, specifically recognize and cut the sequences GAATTC and GGATCC respectively. At the other end of DNA binding specificity spectrum, some DNA binding proteins, such as histone proteins and DNA polymerases, bind DNA non-specifically as they do not discriminate DNA sequences for binding. Transcription factors, a special group of DNA binding proteins, bind to specific and conserved DNA sequences while allowing variations at certain positions [34]. It has been demonstrated that aberrant mutations or genetic variations can alter the binding specificity and thus affect the gene expression, leading to various types of diseases [124, 125]. Therefore, deciphering the protein-DNA recognition codes can not only help us better understand the mechanisms of these specific binding events, but also help explain diseases caused by mutations that affect protein-DNA binding specificity and design therapeutic drugs.

Over the last several decades, with the increasing number of high-resolution structures of protein-DNA complexes in Protein Data Bank (PDB) [32] and the advancement of technologies for exploring DNA binding motifs, such as ChIP-seq, protein-binding microarrays (PBMs) [12], systematic evolution of ligands by exponential enrichment combined with massively parallel sequencing (SELEX-seq) [16] and high-throughput SELEX (HT-SELEX) [18], our knowledge of protein-DNA binding specificity has been greatly expanded. DNA-binding proteins recognize their specific target sites with a combination of two readout mechanisms: base readout and shape readout [35, 46]. Base readout refers to the direct interaction between protein and DNA bases in major groove and minor groove, where the discrimination among bases can be achieved through shape fitting and electrostatic properties, including forming a number of key hydrogen bonds. While there is no simple one-to-one correspondence between amino acids and DNA bases, some particular amino acid-base pairings are enriched, such as arginine with guanine, and asparagine and glutamine with adenine [42–45]. It has been shown that hydrogen bonds between amino acids and bases also provide complex interactions leading to specific recognition [37]. Bidentate interactions, where two or more hydrogen bonds are formed between a residue and a base or a base pair, and complex interactions, where amino acids form hydrogen bonds with more than one base step, have been considered central to specific recognition of single base positions and short DNA sequences and are enriched in highly specific protein-DNA interactions [25, 36, 42]. Recent studies also suggest that π-interactions between aromatic residues and DNA bases play important roles in specific protein-DNA recognition [25, 38–41].

Shape readout refers to both global shape and local shape of target DNA sequences in protein-DNA recognition [46–52]. DNA shape readout relies on both intrinsic and protein-induced DNA deformations in the core binding motifs as well as their flanking regions, especially the A- or T-rich stretch in the flanking regions [24,47,53]. Recently, Rohs group investigated DNA shape changes due to CpG methylation and demonstrated these epigenetic effects on protein-DNA binding [54]. They found that CpG methylation significantly alters local DNA shape, such as roll and propeller twist, and the degree of alterations is affected by the local sequence context. Another study on binding specificity of human transcription factors (TFs) using HT-SELEX and ChIP-seq revealed that homodimer orientation and spacing play a larger role in specific protein-DNA binding than previously thought [24]. Based on these knowledge of protein-DNA binding specificity, various models have been developed for binding site prediction [24,39,48,60–63]. While the performances of these models vary, adding shape features improves prediction accuracy over the sequence-only models.

Several recent studies have also investigated the roles of non-Watson-Crick (WC) base pairs, including Hoogsteen (HG) base pairs and mismatched (MM) base pairs, in protein-DNA recognition [126–128] (and Preprint at https://www.biorxiv.org/content /10.1101/705558v1). The tumor suppressor p53 recognizes diverse DNA response elements (REs) consisting of two continuous or interrupted decameric half-sites. Kitayner *et al.* found that the central A/T doublets of the conserved CATG motifs exhibited non-canonical HG base-pair geometry [127]. This geometry affects the local shape and electrostatic potential of the B-DNA helix and hence the p53-DNA interface, leading to enhanced protein-DNA interactions. The HG geometry of the A/T doublets was also observed by Vainer *et al.* in crystal structure of Lys120-acetylated P53 DNA-binding domain in complex with consensus RE containing CATG motifs [128]. Lys120 acetylation increases the flexibility of loop L1, which is known to increase the DNA-binding specificity of p53, and thus enables the formation of sequence-dependent DNA-binding models. To directly compare the effects of HG and WC base pairs on binding characteristics, Golovenko *et al.* studied p53-DNA crystal structures with designed REs having modified base pairs in either WC or HG form [126]. They found that complexes with REs containing CATG motifs at the center of their half-sites favor the unique HG-induced shape and these complexes are more stable, resulting in enhanced interactions with p53. A very recent study reported the effect of DNA mismatches on DNA binding. The authors found while most MM base pairs within TF binding sites decreased or had no effect on binding affinity, a few MM base pairs increased binding affinity via inducing distortions similar to those induced by TF binding, pre-paying some of the energetic cost associated with DNA distortions contributing to recognition (Preprint at https://www.biorxiv.org/content/10.1101/705558v1). All these studies suggest non-Watson-Crick base pairs play larger roles in protein-DNA recognition than previously thought.

A comparative analysis of protein-DNA complex structures with different degrees of binding specificity was carried out recently [25]. This study revealed a clear trend of structural features among the three DNA-binding protein classes: highly specific (HS), multi-specific (MS), and non-specific (NS). DNA-binding proteins with higher binding specificity form more hydrogen bonds (including both simple and complex hydrogen bonds), have more major groove and base contacts, and the corresponding DNA shape harbors larger propeller and rise. In addition, it was found that aspartate is enriched in highly specific DNA binding proteins and predominately binds to a cytosine through a single hydrogen bond or two consecutive cytosines through complex hydrogen bonds [25]. Protein flexibility is another key factor in specific protein-DNA recognition [55–59]. Highly specific and multi-specific DNA-binding domains tend to have larger conformational changes upon DNA binding and larger degree of flexibility in unbound states [25]. Based on these observations, a machine learning-based SVM

(Support Vector Machine) model for TF (transcription factor)-DNA complex model assessment was developed [129]. The SVM model using structural features of specific protein-DNA interaction significantly improves prediction accuracy of TF-DNA complexes by successfully identifying cases without near-native structural models [129].

Current models for protein-DNA binding specificity primarily focus on interactions between protein and double-stranded DNA (dsDNA). Studies have shown that the double-stranded form of some DNA sequences and their corresponding single strands can serve as binding sites for different DNA-binding proteins [66–69]. For example, the double-stranded form of a 30-bp asymmetric polypurine-polypyrimidine tract serves as a binding site for a transcription enhancer factor-1-related protein, while each single strand binds to two distinct protein factors in regulating the transcriptional activity of the mouse vascular smooth muscle alpha-actin gene in fibroblasts and myoblasts [66,69]. Moreover, it has been reported that several sequence-specific DNA-binding transcription factors bind either the sense or antisense strands of some cis-regulatory elements with enhanced specificity [67,68]. All these findings indicate that two DNA strands may play different roles in specific protein-DNA binding/recognition and the conservation at various binding positions.

We present here an investigation of protein-DNA binding specificity at DNA strand level with a particular focus on side chain-base hydrogen bonds since it has been demonstrated that side chain-base hydrogen bonds are critical to protein-DNA binding specificity [36,42,44,46]. We first performed a comparative analysis at the strand level among DNA-binding proteins with different degrees of binding specificity, HS, MS and NS groups, to explore the contribution of each DNA strand to the overall protein-DNA binding specificity. Our hypothesis is that high binding specificity requires contributions from both DNA strands and thus the bases involved are highly conserved and more sensitive to mutations. In addition, we compared the secondary structure types of residues involved in side chain-base hydrogen bonds in different types of DNA-binding proteins and found distinct patterns. To our knowledge, this is the first large-scale comparative study of protein-DNA binding specificity at the DNA strand level and the role of secondary structure types in specific protein-DNA recognition.

2.3 Materials and Methods

2.3.1 Datasets

The three groups of dsDNA-binding proteins with different degrees of binding specificity, HS, MS and NS, were compiled based on our previous study [25]. Briefly, X-ray crystal structures of protein-dsDNA complexes with resolution $\leq$3 Å and R-factor $\leq$0.3 were selected from PDB. PDA (for protein-DNA complex structure Analyzer) was applied to reconstruct the complete DNA double helix structure via symmetry operations including rotation and translation for complexes with coordinates of only one strand of a double-stranded DNA [130]. These complex structures were then annotated as HS, MS or NS DNA-binding domains based on their binding specificity and function of their DNA-binding domains. Complexes in each group were clustered using CD-HIT with a sequence identity cutoff of 30% [131]. One representative from each cluster was selected to generate the non-redundant dataset [25]. Since the original dataset contains a relatively small number of HS complexes, we expanded the HS dataset by adding four new non-redundant HS protein-DNA complex structures deposited in PDB since our last compilation (Supplementary Table S1). In addition, three DNA-binding domains were updated by either excluding the dimerization domains from the original annotations or by a new PDB ID. More specifically, domain 2e52D01 was changed from 2e52:D to 2e52:D (3-226) and domain 3lsrA01 was changed from 3lsr:A to 3lsr:A (4-53) (Supplementary Table S1). 3qws has been superseded by 6on0 in PDB on 15 May 2019. The final domain-based non-redundant dataset includes 32 HS, 115 MS and 52 NS protein-dsDNA complexes [25]. For comparison purposes, in this study we also generated a corresponding chain-based dataset with

29 HS, 107 MS and 38 NS protein-dsDNA complexes (Supplementary Table S2).

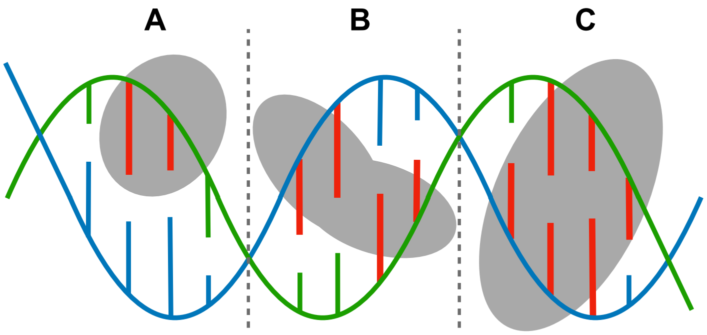


Figure 2.1: Schematic illustration of different types of side chain-base hydrogen bonds between two DNA strands (green and blue respectively) and a protein. The bases that form hydrogen bonds with protein side chain are colored red. (A) Hydrogen bonds between residue side chains and bases from only one DNA strand (green, the dominant strand); (B) equal number of bases that form hydrogen bonds with residue side chains from both DNA strands, also referred as a 50/50 case; (C) another 50/50 case with two base pair side chain-base hydrogen bonds.

2.3.2 Hydrogen bonds and hydrogen bond energy

To assess the contribution of each strand of the DNA double helix to binding specificity, we calculated the number of hydrogen bonds between residue side chains in DNA-binding proteins and DNA bases using HBPLUS [132] and FIRST (Floppy Inclusion and Rigid Substructure Topography) [133] with default parameters. To annotate the hydrogen bonds between protein and DNA with FIRST, we employed an energy cutoff of -0.6 kcal/mol as suggested by the author of FIRST [133]. Percent contribution of each of the two DNA strands in a complex is calculated and the DNA strand with more hydrogen bonds is designated as the dominant strand. For example, the green strand in Figure 2.1A is the dominant strand. If both strands in complexes have equal number of bases forming side chain-base hydrogen bonds, either strand can be the dominant strand and these complexes are referred as 50/50 cases (Figure 2.1B and C). In some cases, both bases of a base pair are involved in forming side chain-base hydrogen bonds with the protein, and these hydrogen bonds are referred as base pair side chain-base hydrogen bonds (Figure 2.1C).

2.3.3 Secondary structure types of DNA interacting residues

An amino acid is defined as a DNA base-contacting residue if it has at least one heavy atom of its side chain within 4.5 Å of any heavy atom of a DNA base. DSSP program was employed to assign three general secondary structure types: helix, strand and coil following the widely used convention: H (α-helix), G (3_{10}-helix) and I (π-helix) states as helix type; E (extended strand) and B (residue in isolated β-bridge) states as strand type and all the other states from DSSP are considered as coil types [28–31].

2.3.4 Statistical analysis

Shapiro-Wilk test was performed to test the normality of the data. If the data is normally distributed, a parametric Student's t-test was carried out. Otherwise, a non-parametric Wilcoxon rank-sum test was applied.

2.4 Results

2.4.1 Comparison of hydrogen bonds between each strand of DNA and DNA-binding domains

It has been demonstrated that hydrogen bonds between amino acid side chains and DNA bases play major roles in specific protein-DNA interactions [36, 42, 44, 46]. It is not surprising that majority of the complexes in the non-specific (NS) DNA-binding group (34 out of 52 complexes) do not have any side chain-base hydrogen bonds and only five complexes have such hydrogen bonds between residues and bases in the major groove. Therefore, we focus on comparing the side chain-base hydrogen bonds between two groups of specific DNA-binding proteins with different degrees of binding specificity: HS and MS.

Percent contributions of single DNA strands in each complex from HBPLUS are shown in Figure 2.2A and B, with the dominant strands shown at the bottom in a descending order. The two DNA strands of the complexes in the HS group tend to have equal or approximately equal contributions to the overall abundance of side chain-base hydrogen bonds. About 34% (11 of 32) of the HS cases have equal number of side chain-base hydrogen bonds from two strands of the DNA double helix and ~91% (29 of 32) of the complexes have no more than 75% of the total contribution from the dominant DNA strand (Figure 2.2A). The MS group, on the other hand, only has ~20% (20 of the total 102 complexes that have at least one side chain-base hydrogen bond) of the cases with equal contributions from the two DNA strands and ~52% (53 of 102) of the complexes have no more than 75% of the total contribution from the dominant DNA strand (Figure 2.2B). Moreover, about 38% (39 of 102) of cases in the MS group only have side chain-base hydrogen bonds from one strand and zero from the other strand while less than 10% (3 of 32) of such cases are found in the HS group (Figure 2.2A, B).

Statistical analysis shows that the distributions of side chain-base hydrogen bonds between the HS and MS groups are significantly different for a combination of both major and minor grooves (Figure 2.2C) or for the major groove only (Figure 2.2D). The side chain-base hydrogen bonds in the minor groove are quite sparse and there are no apparent differences between HS and MS groups as they both skew towards one strand (Figure 2.2E). As a control, we compared distributions in terms of non-side chain-base hydrogen bonds from each strand, which are considered to contribute mainly to protein-DNA binding affinity but not much to specificity. Unlike the more specific side chain-base hydrogen bonds, there are no significant differences between the HS and MS groups, suggesting approximately equal contribution from each strand for hydrogen bonds between protein and DNA backbones in both HS and MS groups

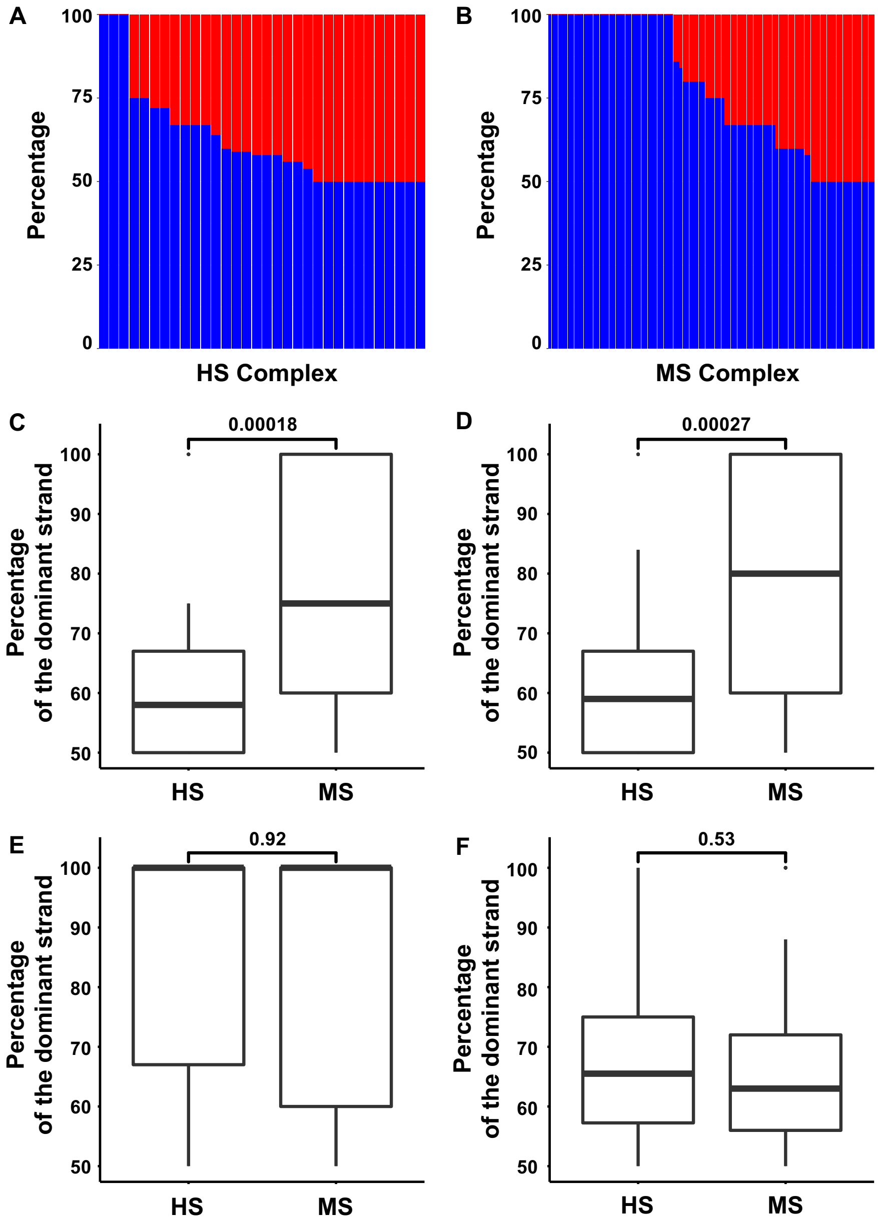


Figure 2.2: Comparison of the number of side chain-base hydrogen bonds of each strand of DNA annotated by HBPLUS between the HS and MS DNA-binding proteins. (A) Percentage contribution of two DNA strands in HS complexes; (B) percentage contribution of two DNA strands in MS complexes. The dominant strands (blue) are shown at the bottom in a descending order. Boxplots and statistical analyses for: (C) both major and minor grooves, (D) major groove only, (E) minor groove only and (F) non-side chain-base hydrogen bonds in both major and minor grooves. P-values are displayed on top of the boxplots.

(Figure 2.2F). To make sure that these observations are robust and not biased results from HBPLUS, we applied a different hydrogen bond identification program, FIRST, using one suggested energy cutoff of -0.6 kcal/mol to determine the number of hydrogen bonds (52). Even though the total number of hydrogen bonds is slightly different from those annotated with HBPLUS due to different hydrogen bond identification algorithms, the results are nevertheless consistent with those from HBPLUS, which is two strands tend to contribute equally to the protein-DNA binding in terms of side chain-base hydrogen bonds in highly specific protein-DNA binding complexes, but the contribution skews towards one strand in the MS group (Figure 2.3).

In addition to comparison of number of hydrogen bonds, we also carried out comparisons of hydrogen bond raw energy between two DNA strands since a hydrogen bond is identified as long as the hydrogen bond energy between two potential hydrogen bond forming atoms is below a cutoff value. The comparison of hydrogen bond energy (below cutoff -0.6 kcal/mol) from FIRST is shown in Figure 2.4. Similar patterns to the number of hydrogen bonds were found between the HS and MS groups.

2.4.2 Chain-based versus domain-based analyses

The above analyses were carried out between DNA-binding domains and DNA double helices. While some protein-DNA complexes only contain DNA-binding domains, other complexes consist of full-chain DNA-binding proteins, which may include signal-sensing or trans-activating domains besides DNA binding domains. These non-DNA-binding domains sometimes provide extra contacts between protein and DNA and contribute to protein-DNA binding affinity and/or binding specificity. It is interesting to see if there are any differences between domain-based and chain-based analyses with respect to the number of side chain-base hydrogen bonds from each DNA strand. While the numbers of hydrogen bonds and hydrogen bond energy are larger in the chain-based comparison, which is expected since some protein chains have two or more

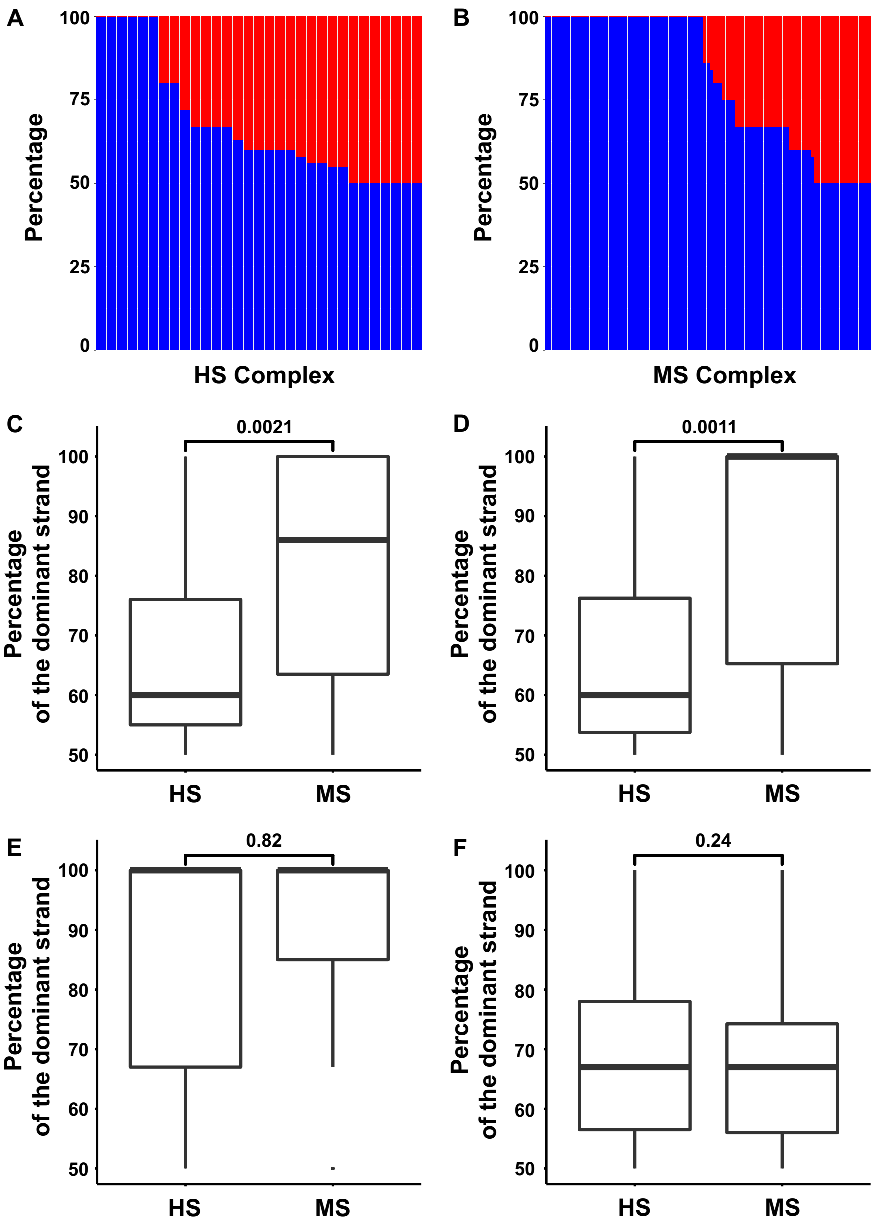


Figure 2.3: Comparison of the number of side chain-base hydrogen bonds annotated by FIRST (-0.6 kcal/mol cutoff) of each strand of DNA between the HS and MS DNA-binding proteins. (A) Percentage contribution of two DNA strands in HS complexes; (B) Percentage contribution of two DNA strands in MS complexes. The dominant strands (blue) are shown at the bottom in a descending order. Boxplots and statistical analyses for: (C) both major and minor grooves, (D) major groove only, (E) minor groove only, and (F) non-side chain-base hydrogen bonds in both major and minor grooves. P-values are displayed on top of the boxplots.

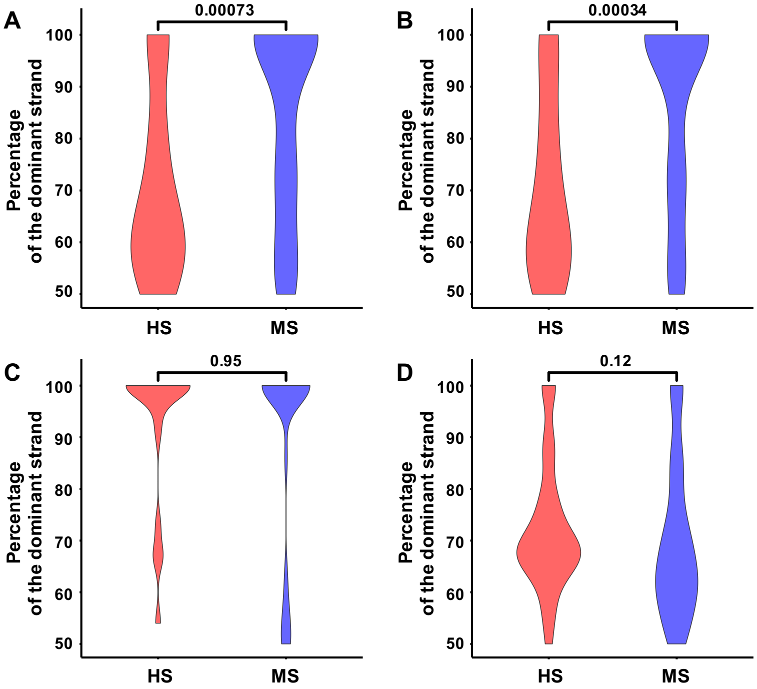


Figure 2.4: Comparison of side chain-base hydrogen bond energy with FIRST (-0.6 kcal/mol cutoff) of each strand of DNA between the HS and MS DNA-binding proteins for: (A) both major and minor grooves, (B) major groove, (C) minor groove, and (D) non-side chain-base hydrogen bonds in both major and minor grooves.

DNA binding domains, similar patterns of differences to the domain-based analyses are found between the HS and MS groups (Figure 2.5). This is also in agreement with the findings reported by Jolma *et al.* that full-length transcription factors and isolated DNA-binding domains bind similar sequences and thus analysis of DNA-binding domains is sufficient to determine the protein-DNA binding specificity [24].

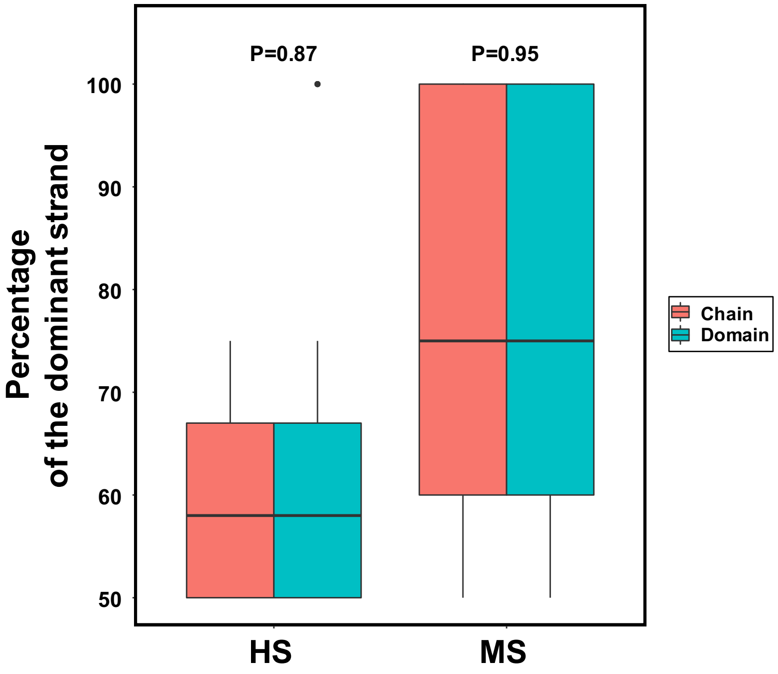


Figure 2.5: Comparison of chain-based and domain-based analyses of the number of side chain-base hydrogen bonds of two strands between the HS and MS groups. Both major and minor grooves were considered.

2.4.3 DNA bases involved in hydrogen bonding with protein side chains from each DNA strand

Since some hydrogen bonds between DNA bases and protein side chains are bidentate and complex interactions, meaning one base can form two hydrogen bonds with one or more residues [42], we next compared the number of DNA bases that are involved in hydrogen bonding with amino acid side chains in DNA-binding domains between two DNA strands. The percentage of bases involved in side chain-base hydrogen bonding from the dominant strands is close to 50% in the HS group while it is larger in the MS group when base contacts in both major and minor grooves are considered (Figure 2.6A) or only base contacts in the major groove are considered (Figure 2.6B). Similar results are observed with FIRST (Figure 2.7A and B). The P-value in Figure 2.6A that compares the number of bases involved in side chain-base hydrogen bonding in both major and minor grooves with HBPLUS is slightly higher (but still <0.05). A closer examination of the data revealed that HBPLUS identifies more complexes and more bases that form hydrogen bonds with side chains in the minor groove than those from FIRST, resulting in a larger percentage of complexes in the MS group with smaller percentage contributions from the dominant strands (data not shown). No apparent differences were found in the minor groove (Figure 2.6C and Figure 2.7C).

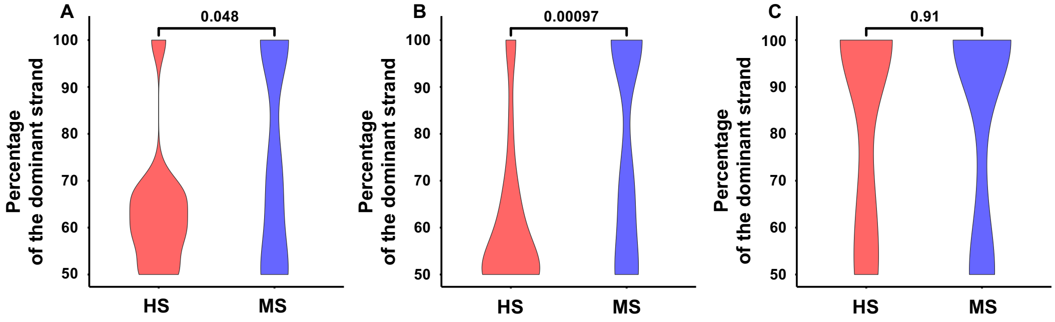


Figure 2.6: Comparison of the number of DNA bases involved in hydrogen bonding with side chains from HBPLUS for: (A) both major and minor grooves, (B) major groove only and (C) minor groove only, between HS and MS DNA-binding proteins.

2.4.4 Side chain-base hydrogen bonding base pairs

Not only does the HS group have much larger percentage of complexes (15/32 ≈ 47%) that have equal number of bases forming side chain-base hydrogen bonds in the major groove from two DNA strands (50/50 cases) than the MS group (30/102 ≈ 29%) (Figure 2.8A), the majority of these 50/50 cases in the HS group have base pair side chain-base hydrogen bonds (12/15 = 80%), while only 3 out of 30 (10%) cases in the MS group have base pairs forming hydrogen bonds with protein side

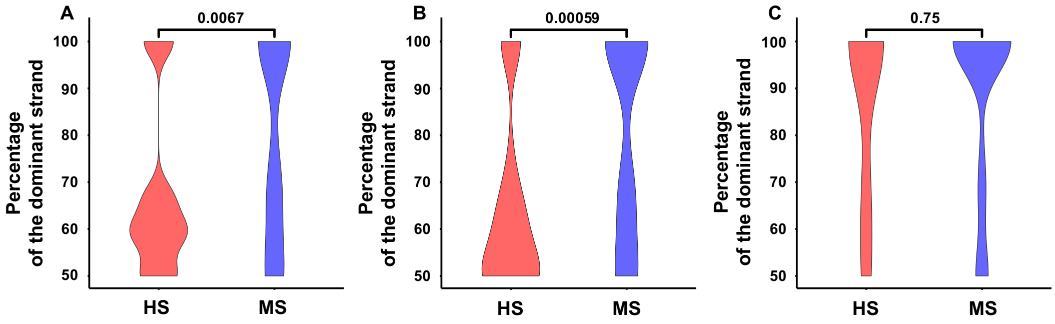


Figure 2.7: Comparison of the number of DNA bases involved in hydrogen bonding with side chains from FIRST (-0.6 kcal/mol cutoff) for: (A) both major and minor grooves, (B) major groove, and (C) minor groove, between HS and MS DNA-binding proteins.

chains (Figure 2.8B and Figure 2.9). For instance, while both restriction endonuclease NgoMIV (PDBID: 4ABT) and transcription factor *Escherichia coli* sigma(E)4 (PDBID: 2H27) form side chain-base hydrogen bonds with equal number of bases from two DNA strands in the major groove, the highly specific DNA binding protein NgoMIV has three continuous base pairs involved in forming hydrogen bonds (Figure 2.10A and Figure 2.9A) but the multi-specific sigma(E)4 forms such hydrogen bonds with unpaired bases (Figure 2.10B and Figure 2.9B). The 50/50 cases from FIRST annotations show similar results with $8/13 \approx 62\%$ in the HS group and $4/24 \approx 17\%$ in the MS group (Figure 2.11). The total amounts of base pairs involved in hydrogen bonding with residues are shown in Figure 2.12 (HBPLUS) and Figure 2.13 (FIRST). The HS group has much larger percentage of complexes that have at least one base pair, two or more base pairs that are involved in side chain-base hydrogen bonding than the MS group. GC base pairs are more prevalent than AT base pairs in both HS and MS groups.

2.4.5 Secondary structure types of DNA interacting residues

DNA-binding proteins recognize their target sites with a number of common binding motifs, such as helix-turn-helix, $\beta\beta\alpha$ zinc finger and zipper-type motifs [2]. The secondary structure types of amino acids involved in specific protein-DNA binding,

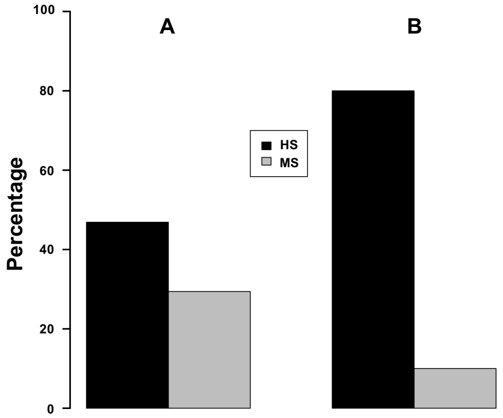
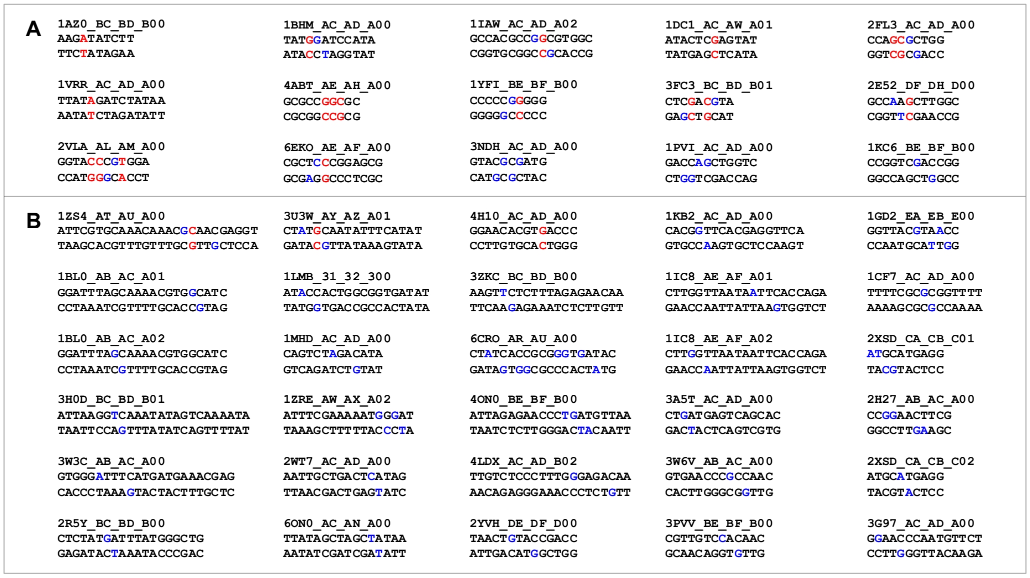


Figure 2.8: Comparison of the number of 50/50 cases (A) and the number of cases with base pairs involved in hydrogen bonding with residue side chains from these 50/50 cases (B) between the HS group and MS group with HBPLUS. (A) The percentage was calculated by dividing the number of 50/50 cases in each group over the total number of complexes forming side chain-base hydrogen bonds in that group; (B) the proportion of these 50/50 cases that have base pairs involved in side chain-base hydrogen bonding.



A	1AZ0_BC_BD_B00 AAGATATCTT TTCTATAGAA	1BHM_AC_AD_A00 TATGGATCCATA ATACCTAGGTAT	1IAW_AC_AD_A02 GCCACGCCGGCGTGGC CGGTGCGGCCGCACCG	1DC1_AC_AW_A01 ATACTCGAGTAT TATGAGCTCATA	2FL3_AC_AD_A00 CCAGCGCTGG GGTCGCGACC
	1VRR_AC_AD_A00 TTATAGATCTATAA AATATCTAGATATT	4ABT_AE_AH_A00 GCGCCGGCGC CGCGGCCGCG	1YFI_BE_BF_B00 CCCCCGGGGG GGGGGCCCCC	3FC3_BC_BD_B01 CTCGACGTA GAGCTGCAT	2E52_DF_DH_D00 GCCAAGCTTGGC CGGTTCGAACCG
	2VLA_AL_AM_A00 GGTACCCGTGGA CCATGGGCACCT	6EKO_AE_AF_A00 CGCTCCCGGAGCG GCGAGGCCCTCGC	3NDH_AC_AD_A00 GTACGCGATG CATGCGCTAC	1PVI_AC_AD_A00 GACCAGCTGGTC CTGGTCGACCAG	1KC6_BE_BF_B00 CCGGTCGACCGG GGCCAGCTGGCC
B	1ZS4_AT_AU_A00 ATTCGTGCAAACAAACGCAACGAGGT TAAGCACGTTTGTTTGCGTTGCTCCA	3U3W_AY_AZ_A01 CTATGCAATATTTCATAT GATACGTTATAAAGTATA	4H10_AC_AD_A00 GGAACACGTGACCC CCTTGTGCACTGGG	1KB2_AC_AD_A00 CACGGTTCACGAGGTTCA GTGCCAAGTGCTCCAAGT	1GD2_EA_EB_E00 GGTTACGTAACC CCAATGCATTGG
	1BL0_AB_AC_A01 GGATTTAGCAAAACGTGGCATC CCTAAATCGTTTTGCACCGTAG	1LMB_31_32_300 ATACCACTGGCGGTGATAT TATGGTGACCGCCACTATA	3ZKC_BC_BD_B00 AAGTTCTCTTTAGAGAACAA TTCAAGAGAAATCTCTTGTT	1IC8_AE_AF_A01 CTTGGTTAATAATTCACCAGA GAACCAATTATTAAGTGGTCT	1CF7_AC_AD_A00 TTTTCGCGCGGTTTT AAAAGCGCGCCAAAA
	1BL0_AB_AC_A02 GGATTTAGCAAAACGTGGCATC CCTAAATCGTTTTGCACCGTAG	1MHD_AC_AD_A00 CAGTCTAGACATA GTCAGATCTGTAT	6CRO_AR_AU_A00 CTATCACCGCGGGTGATAC GATAGTGGCGCCCACTATG	1IC8_AE_AF_A02 CTTGGTTAATAATTCACCAGA GAACCAATTATTAAGTGGTCT	2XSD_CA_CB_C01 ATGCATGAGG TACGTACTCC
	3H0D_BC_BD_B01 ATTAAGGTCAAATATAGTCAAAATA TAATTCCAGTTTATATCAGTTTTAT	1ZRE_AW_AX_A02 ATTTCGAAAAATGGGAT TAAAGCTTTTTACCCTA	4ON0_BE_BF_B00 ATTAGAGAACCCTGATGTTAA TAATCTCTTGGGACTACAATT	3A5T_AC_AD_A00 CTGATGAGTCAGCAC GACTACTCAGTCGTG	2H27_AB_AC_A00 CCGGAACTTCG GGCCTTGAAGC
	3W3C_AB_AC_A00 GTGGGATTTCATGATGAAACGAG CACCCTAAAGTACTACTTTGCTC	2WT7_AC_AD_A00 AATTGCTGACTCATAG TTAACGACTGAGTATC	4LDX_AC_AD_B02 TTGTCTCCCTTTGGGAGACAA AACAGAGGGAAACCCTCTGTT	3W6V_AB_AC_A00 GTGAACCCGCCAAC CACTTGGGCGGTTG	2XSD_CA_CB_C02 ATGCATGAGG TACGTACTCC
	2R5Y_BC_BD_B00 CTCTATGATTTATGGGCTG GAGATACTAAATACCCGAC	6ON0_AC_AN_A00 TTATAGCTAGCTATAA AATATCGATCGATATT	2YVH_DE_DF_D00 TAACTGTACCGACC ATTGACATGGCTGG	3PVV_BE_BF_B00 CGTTGTCCACAAC GCAACAGGTGTTG	3G97_AC_AD_A00 GGAACCCAATGTTCT CCTTGGGTTACAAGA

Figure 2.9: Base pairs involved in hydrogen bonding with residue side chains (red) in 50/50 cases in the HS group (A) and MS group (B) with HBPLUS. Individual bases that are involved in hydrogen bonding with residue side chains are shown in blue font.

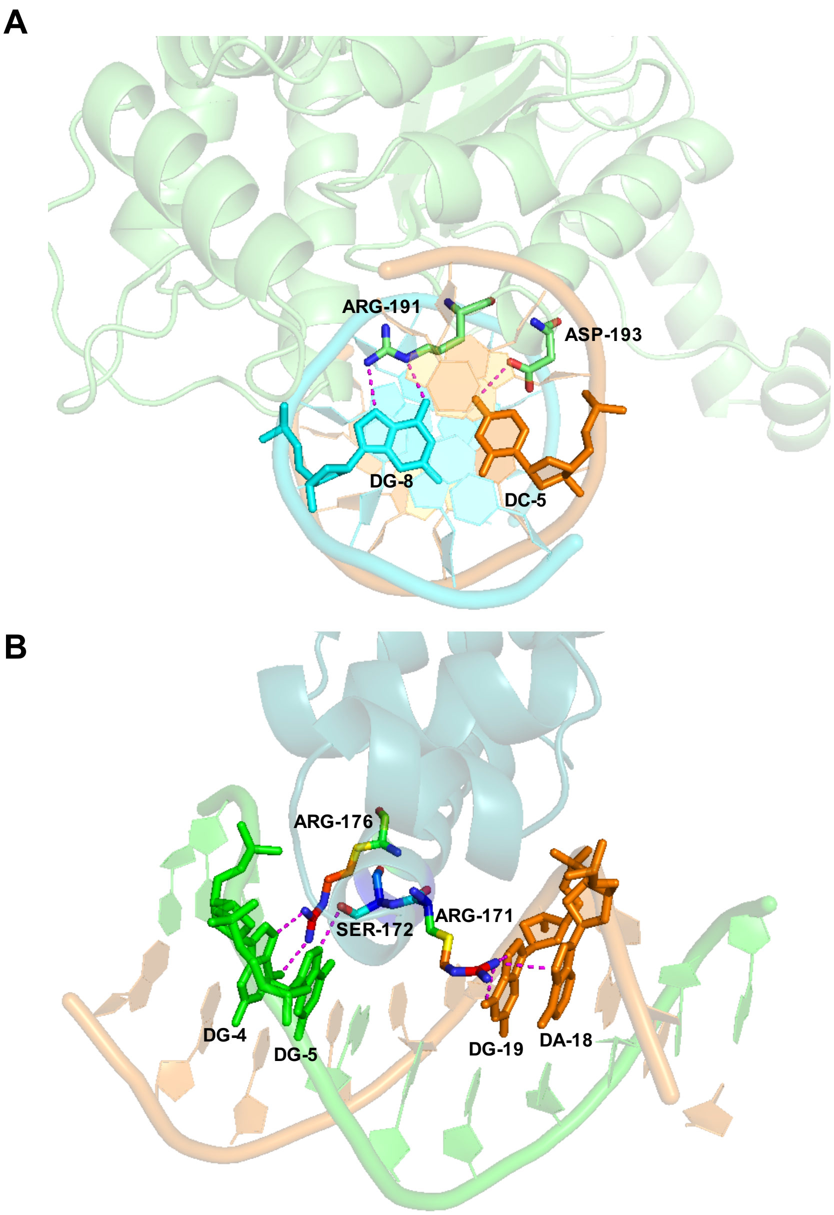


Figure 2.10: Examples of DNA-binding proteins bound to paired bases and unpaired bases. (A) Highly specific DNA-binding protein NgoMIV bound to paired bases (PDBID: 4ABT; protein chain: A; DNA chains: E and H). Only one out of three continuous base pairs involved hydrogen bonding is highlighted. Base pairs DC-9 (chain E) and DG-4 (chain H), DG-7 (chain E) and DC-6 (chain H) are also involved in side chain-base hydrogen bonds. (B) Multi-specific DNA-binding protein sigma(E)4 bound to equal number but unpaired bases with two strands (PDBID: 2H27; protein chain: A; DNA chains: B and C).

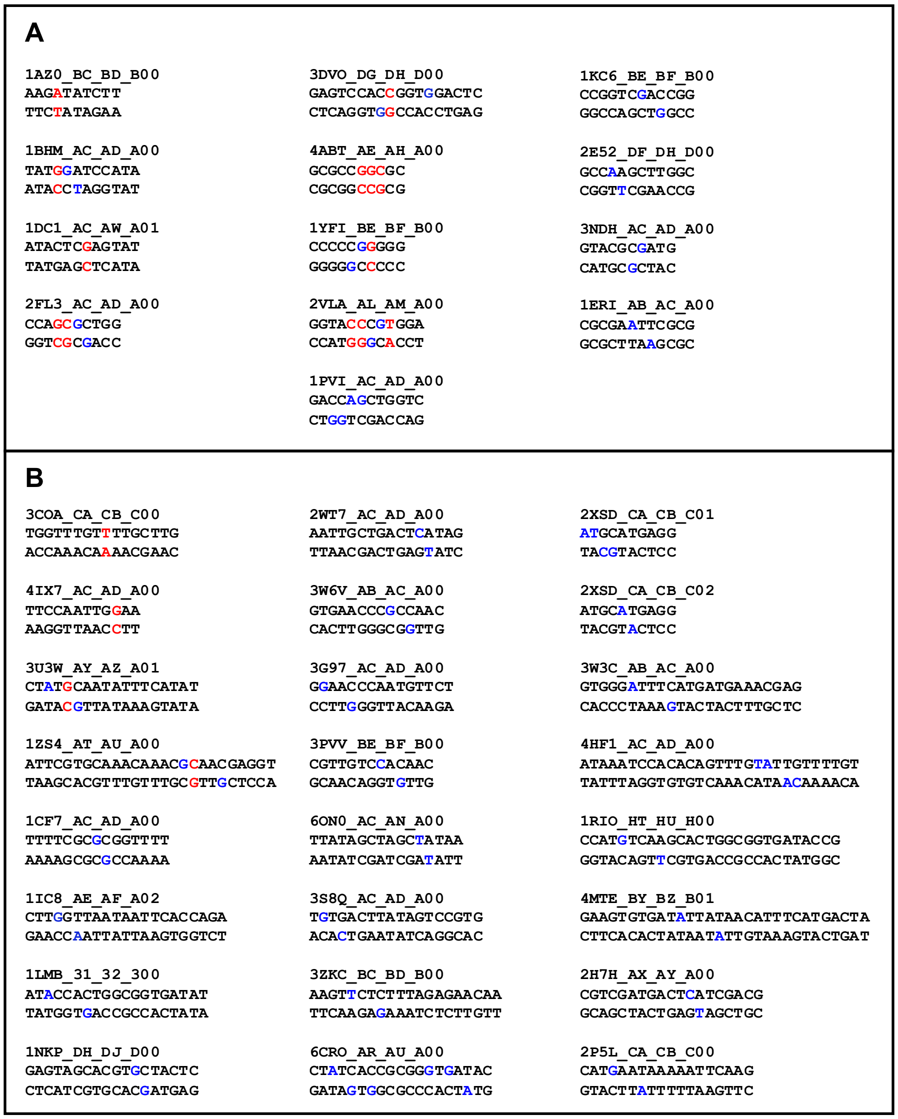


Figure 2.11: Base pairs involved in hydrogen bonding with residue side chains (red) in 50/50 cases in the HS group (A) and MS group (B) identified by FIRST (-0.6 kcal/mol cutoff). Individual bases that are involved in hydrogen bonding with residue side chains are shown in blue font.

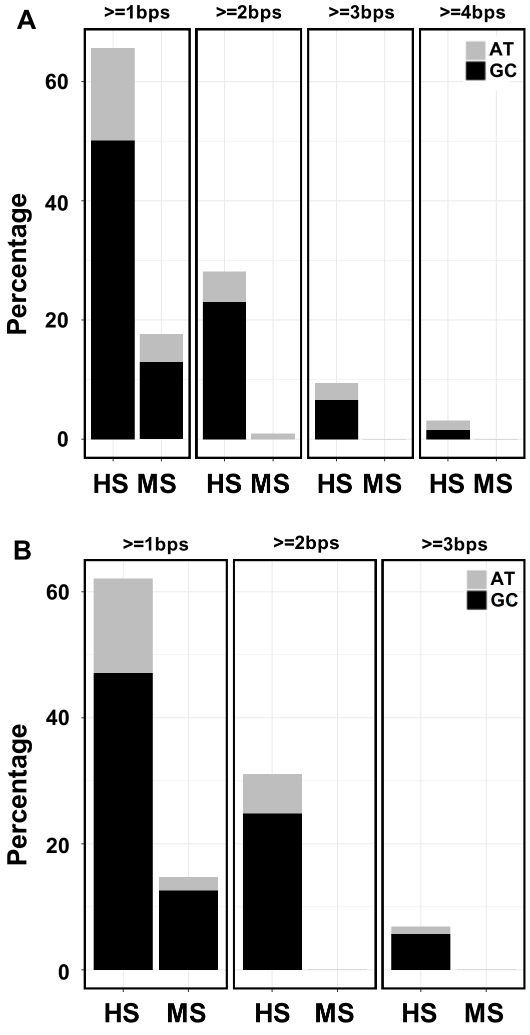


Figure 2.12: Comparison of base pairs that are involved in side chain-base hydrogen-bonding between HS and MS groups with HBPLUS in (A) both major and minor grooves and (B) major groove only.

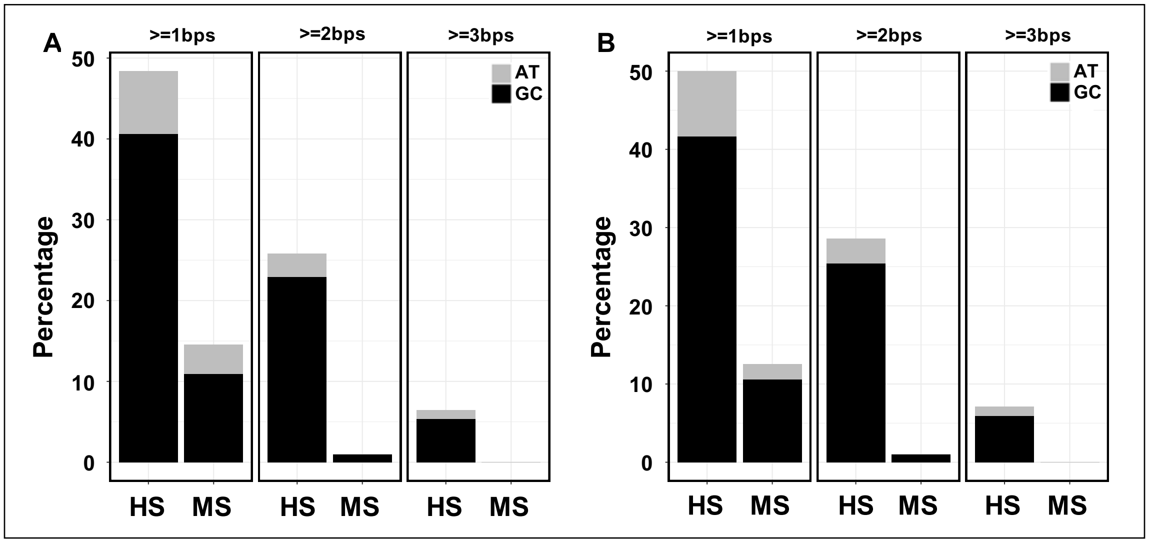


Figure 2.13: Comparison of side chain-base hydrogen-bonding base pairs with FIRST (-0.6 kcal/mol cutoff) between HS and MS groups in (A) both major and minor grooves and (B) major groove.

however, have not been investigated extensively. We first compared the propensities of the secondary structure types of amino acids in DNA-binding domains that are in contact with DNA bases, calculated against the relative frequencies of secondary structure types of residues in respective group of DNA-binding domains. The DNA base-contacting residues in the HS group are enriched in coil conformations while helical secondary structure types are preferred in the MS group (Figure 2.14A). For residues that form hydrogen bonds between their side chains and DNA bases, we used two different background distributions to calculate the propensities: one is the secondary structure type distribution of all base-contacting residues (Figure 2.14B and C) and the other is the secondary structure type distribution of all residues that form hydrogen bonds with DNA including bases and backbone atoms (Figure 2.14D and E).

When residues involved in side chain-base hydrogen bonds in the major and minor grooves are combined, DNA-binding proteins in both the HS and MS groups prefer strand types and there are no major differences between the HS and MS groups no matter which background distribution is used (Figure 2.14B and D). However, when

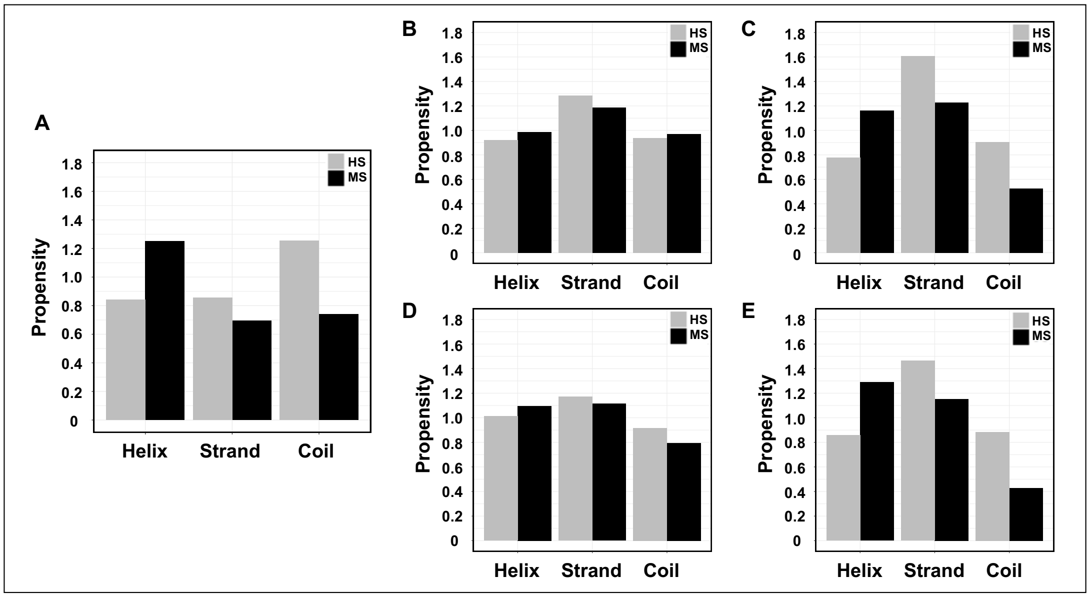


Figure 2.14: Propensities of secondary structure types in the HS and MS groups. (A) Propensities of secondary structure types of DNA base-contacting residues, the background relative frequencies of secondary structure types are calculated using all residues in the DNA-binding domains in each group. Propensities of secondary structure types of residues involved in side chain-base hydrogen bonds with HBPLUS for both major and minor grooves (B, D) and for major groove only (C, E). Propensities are calculated using either the relative frequencies of secondary structure types of base-contacting residues (B, C) or all DNA hydrogen-bonding residues (D, E).

only such contacts in the major groove are considered, there is a distinct pattern. The strand type is highly enriched in the HS group, while proteins in the MS group favor both strand and helical types but are depleted in coil conformations when compared to DNA-binding domains in the HS group (Figure 2.14C and E). For example, residues involved in side chain-base hydrogen bonds in restriction endonuclease BstYI, a highly specific DNA-binding protein, reside in strand and coil secondary structure types (Figure 2.15A) while in hepatocyte nuclear factor 1-alpha (HNF-1alpha) residues in helical conformation are involved in hydrogen bonding with bases (Figure 2.15B). The above results suggest a role of flexibility in conferring different degrees of binding specificity (See detailed discussions in the next section). This observation is consistent between HBPLUS and FIRST results (Figure 2.16). Further investigation revealed that residues in the MS group that are involved in side chain-base hydrogen bonds have ∼70% coils in the minor groove, which may explain the differences of propensities between the major+minor grooves (Figure 2.14B and D) and major groove alone (Figure 2.14C and E).

2.5 Discussion

Understanding the mechanisms of protein-DNA binding specificity is of paramount importance in deciphering gene regulation networks and designing therapeutic drugs. It has been demonstrated that hydrogen bonds between amino acid side chains and DNA bases play major roles in specific protein-DNA recognition [36, 42, 44, 46]. As such, to further understand structural features in protein-DNA binding specificity, we performed a comparative analysis based on side chain-base hydrogen bonds. We first investigated protein-DNA binding specificity at DNA strand level, which has not been explored before. The amounts of side chain-base hydrogen bonds between each DNA strand and DNA-binding domains of two groups of DNA-binding proteins, HS and MS, were compared [25]. Since there are a number of different algorithms for calculating hydrogen bond energy and typically a default energy cutoff is applied for

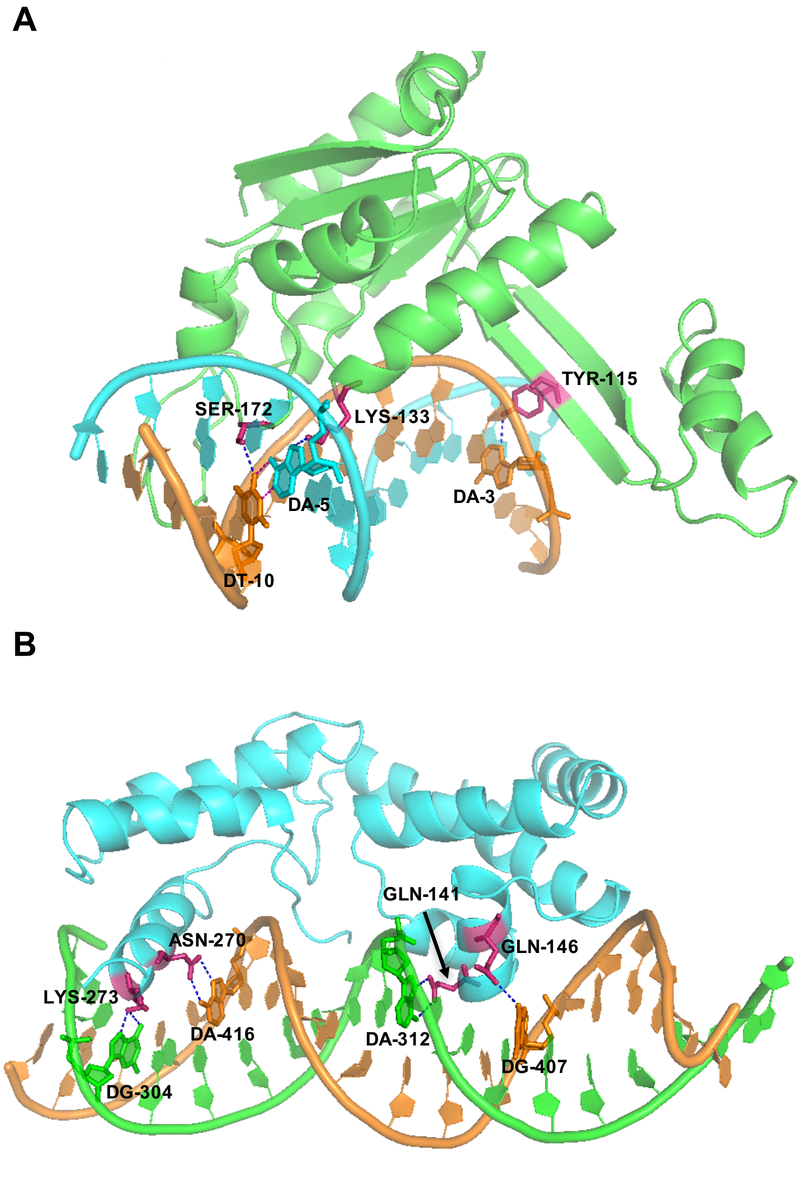


Figure 2.15: Secondary structure preferences of highly specific DNA-binding protein and multi-specific DNA-binding protein. (A) Highly specific DNA-binding protein representative (PDBID: 1VRR; protein chain: A; DNA chains: C and D). Strand and coil secondary structure types (magenta) are involved in side chain-base hydrogen bonds (blue dash line). Two DNA bases involved in hydrogen bonds with protein side chains, A5 and T10, are paired bases; hydrogen bonds between this pair are shown in red dash line. (B) Multi-specific DNA-binding protein representative (PDBID: 1IC8; protein chain: A; DNA chains: E and F). Residues involved in hydrogen bonding are in helical conformation (magenta).

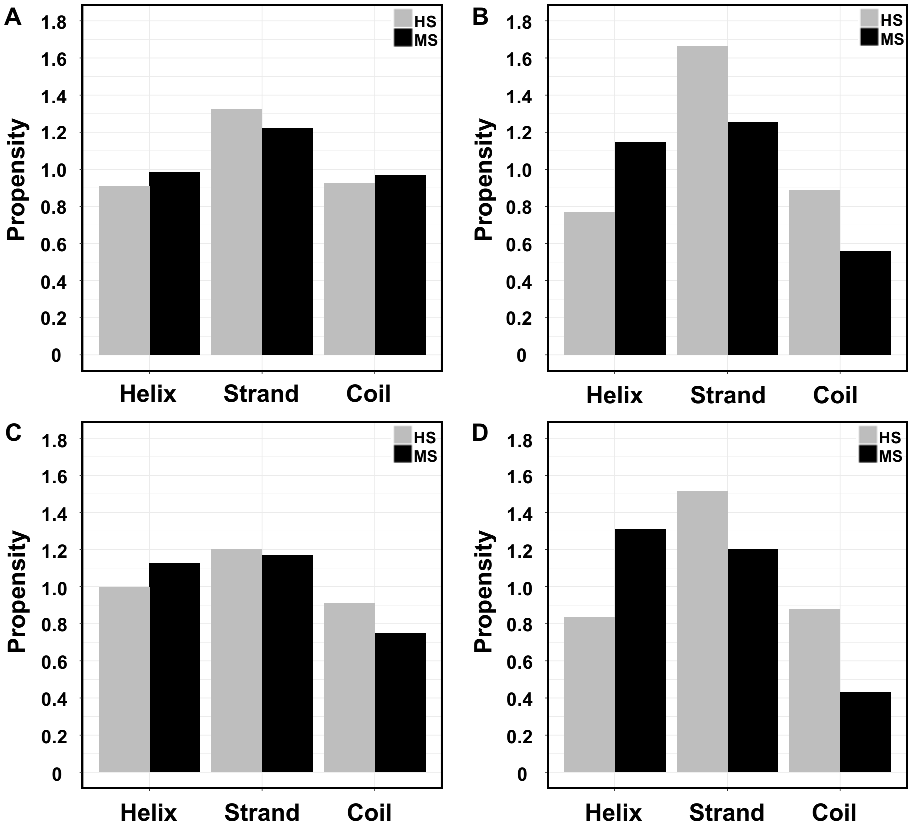


Figure 2.16: Propensities of secondary structure types of residues involved in side chain-base hydrogen bonds with FIRST (-0.6 kcal/mol cutoff). (A, C) both major and minor grooves and (B, D) major groove only. Propensities are calculated over the relative frequencies of secondary structure types of base-contacting residues (A, B) and all DNA hydrogen-bonding residues (C, D).

determining the existence of hydrogen bonds, we applied two widely used hydrogen bond annotation programs HBPLUS and FIRST to ensure our results are robust and the conclusions are independent of hydrogen bond identification programs. Results show that DNA-binding domains with high binding specificity have approximately equal contributions of side chain-base hydrogen bonds from two DNA strands, while a larger percentage of protein-DNA complexes form side chain-base hydrogen bonds with only one DNA strand in the MS group (Figure 2.2, Figure 2.3). Not only are these findings in agreement between HBPLUS and FIRST, they are also consistent between domain-based and chain-based analyses (Figure 2.5).

We also found that highly specific protein-DNA complexes have more base pairs involved in hydrogen bonding with protein side chains than those with lower binding specificity in the MS group (Figure 2.12 and Figure 2.13). These observations, approximately equal distributions from two DNA strands and larger number of base pairs involved in side chain-base hydrogen bonding in the high binding specificity group, help explain why the bases in the high binding specificity group are highly conserved and are very sensitive to mutations. DNA-binding proteins in the HS group are mainly Type II restriction endonucleases. These endonucleases recognize short palindromic sequences of 4-8 bps specifically as homodimers and cleave DNA double helices [134]. This process relies on the concerted recognition of two DNA strands and the communication of this recognition information between two subunits, suggesting this recognition process coordinates efforts from specific interactions between protein and both DNA strands. Transcription factors in the MS group, on the other hand, regulate gene expression by binding to target sequences, called transcription factor binding sites (TFBS) [135]. While the binding between transcription factors and their corresponding binding sites is specific and certain positions are highly conserved, transcription factors generally allow variability at some other base positions. In addition, it has been shown that some transcription factors can bind to two different binding motifs, called primary and secondary binding motifs [136]. If one strand is the primary one for a DNA-binding protein, a base mutation would have less effect than the case that both bases of a base pair get involved in specific interaction. Hydrogen-bonding donor and acceptor patterns in the major groove are unique to specific base pairs, therefore it is impossible to maintain the original hydrogen-bonding patterns if a base of a base pair is mutated when this particular base pair is involved in specific hydrogen bonding, making it more sensitive to mutations and thus more conserved.

Majority of the base pairs involved in hydrogen bonding in both HS and MS groups are GC pairs (Figure 2.12 and Figure 2.13). Nadassy *et al.* analyzed 65 X-ray structures of protein-dsDNA complexes and observed that GC pairs make three times as many hydrogen bonds as AT pairs in the major groove [137]. However, in their study, the occurrence of base pairs was counted in a different way. As long as one base of a pair is involved, it is considered a pair participation. Nikolajewa *et al.* found a significant GC contact in type II restriction enzyme binding sites [138]. These results suggest that GC pairs play critical roles in specific protein-DNA binding. These observations are not surprising since guanine has a strong electronegative character in the major groove and is compatible to the guanidinium group of arginine. In addition, guanine contributes an extra hydrogen bond donor of N2 in the minor groove. Studies have shown that the addition, removal, substitution and relocation of the exocyclic 2-amino group of guanine in the minor groove affect DNA cleavage by DNA-binding proteins, DNA binding with small molecules and antibiotics [139–142]. For instance, by examining base substitutions that affect the presence and location of the 2-amino group of guanine in tyrT(A93) DNA, Bailly *et al.* found these alterations affect both the flexibility of tyrT(A93) DNA and its affinity for its binding protein, the *Escherichia coli* Factor for Inversion Stimulation (FIS) [142].

Statistical analyses show significant differences in the major groove but not in the minor groove between HS and MS groups. This is consistent with the base readout mechanism. In the major groove, every base pair has a unique hydrogen bond acceptor and donor pattern that can be distinguished from other base pairs. In the minor groove, however, the degeneracy of the pattern of hydrogen bond acceptors and donors cannot distinguish A/T from T/A or C/G from G/C. For non-specific DNA-binding proteins, we found more complexes have side chain-base hydrogen bonds in the minor groove than the major groove (data not shown). Although in general hydrogen bonds between proteins and bases in the minor groove play a less role than those in the major groove, in some cases, the minor groove hydrogen bonds are critical especially when the shape readout is considered. Rohs *et al.* demonstrated that arginine prefers to bind narrow minor grooves in AT-rich regions and the role of DNA shape in the protein-DNA recognition, which represents a novel DNA recognition mechanism in many DNA binding protein families [50]. These minor-groove interactions may stabilize the deformed DNA structure and identify incorrectly incorporated non-Watson-Crick base pairs [64]. It has also been reported that amino acid side chain-base hydrogen bonds in the minor groove are important in insertion and extension of base pairs in DNA replication [143–146].

DNA base-contacting residues in highly specific DNA-binding proteins are enriched in coils while multi-specific DNA-binding proteins prefer helices (Figure 2.14A). For residues forming hydrogen bonds with bases in the major groove, the propensity of coil conformations for HS proteins is about two times more than that for the MS proteins (Figure 2.14C and E, Figure 2.16B and D). These results suggest that protein flexibility play important roles in protein-DNA recognition, as reported in previous studies [25,55–59]. For instance, our previous study found that specific DNA-binding domains tend to have larger conformational changes upon DNA-binding and larger degree of flexibility in unbound states [25]. It has been hypothesized that protein flexibility can help speed up DNA recognition [147,148]. The higher flexibility of coils than helices should play important roles in locating DNA-binding proteins to their specific target sites. More importantly, flexibility can enhance the binding specificity via forming larger number of hydrogen bonds with DNA bases due to coil's fine-tuning capability. A recent comparative molecular dynamics simulations on wild-type and F10V mutant P22 Arc repressor in both free and complex conformations demonstrated the role of protein flexibility in protein-DNA binding specificity [58]. The DNA-binding motif of wild-type Arc repressor is more flexible and this flexibility leads to more hydrogen bonds formed with DNA bases upon binding, which results in higher DNA-binding specificity [58]. We also found that while residues involved in hydrogen bonding with DNA major grooves generally prefer strand secondary structure types (HS group shows slightly higher preference), MS group also favors helices (Figure 2.14C and E). Mutation tolerance study of different secondary structure elements of proteins shows that alpha helices are more robust to mutations than beta strands [122]. The preference of strands of highly specific DNA-binding proteins makes them more sensitive to mutations from the perspective of protein conformations. These secondary structure type preferences and the fact that DNA bases are more conserved in highly specific DNA-binding proteins, indicate that the conservation of highly specific DNA-binding proteins requires both conserved protein secondary structures and DNA bases.

While our analyses are based on complexes with targeted DNA bases forming canonical Watson-Crick base pairing geometry, the method can be generalized for studying structures with non-Watson-Crick base pairs, including HG and MM base pairs when large datasets of such cases become available. In addition to DNA shape, the effect of DNA mismatches on protein-DNA binding specificity can be investigated in terms of hydrogen bonds (https://www.biorxiv.org/content/10.1101/705558v1). It would be interesting to see how the mutated bases of those mismatched base pairs from different strands affect the protein-DNA binding affinity and/or specificity by altering the hydrogen bonding patterns or other types of interactions. Anti-syn transitions of

DNA base conformation have been widely observed when base pairing changes from WC geometry to HG and MM base pairing [149–152]. Future studies can reveal if the transitions are biased toward one strand or randomly distributed between two strands. Our results also offer possible clue to the increased mutation rates around transcription factor binding sites (TFBS) [153, 154]. The increased levels of mutations around TFBS have been attributed to the barrier created by DNA-binding proteins to the displacements of DNA synthesized by error-prone polymerase-α [153], and a decrease of nucleotide excision repair (NER) activity caused by interference of DNA-binding proteins with the NER machinery [154].

Our study, for the first time to our knowledge, reports that high protein-DNA binding specificity may require approximately equal contributions from two DNA strands. Investigation of secondary structure types of DNA interacting residues suggests that both secondary structure types and protein flexibility play important roles in specific protein-DNA recognition. Our results not only provide new insights into protein-DNA binding specificity, but also have great potential in further exploration of novel mechanisms of protein-DNA interactions in complexes containing non-Watson-Crick base pairs.

CHAPTER 3: A COMPARATIVE STUDY OF PROTEIN-SSDNA INTERACTIONS

3.1 Introduction

In many essential cellular processes such as DNA replication, recombination, and repair, the DNA double helix is unwound and two complementary DNA strands are exposed in a single-stranded form [3,4]. Single-stranded DNA (ssDNA) is vulnerable to chemical and enzymatic attacks and is prone to form secondary structures that interfere with biological activities such as DNA replication. As a consequence, a specific group of proteins, single-stranded DNA-binding proteins (SSBs), has evolved to bind to and stabilize ssDNA. SSBs are essential in the maintenance of genomic stability, playing critical roles in telomere end protection [6,7], DNA damage repair, control of the cell cycle checkpoint [155], and the recruitment of partner proteins to regulate DNA metabolism [8]. It has been demonstrated that aberrant ssDNA binding leads to genome instability and tumorigenesis [156,157]. Therefore, knowledge of SSB-ssDNA interactions can help understand the mechanisms underlying normal cellular processes and human malignancies. More importantly, it can provide guidance for drug design in targeted cancer therapy.

Current knowledge of SSB-ssDNA interaction, however, lags far behind of other types of protein-nucleic acid interactions. Our understanding of SSB-ssDNA binding mainly comes from several extensively studied individual SSBs, such as bacteriophage T4 gene 32 protein (gp32) [75,78], *E. coli* SSB [80], and replication factor A (RPA) [87,88]. gp32 is the first SSB that has been identified [75], and it exists as monomers in solution without DNA substrates [158]. The central region of the gp32 monomer is the ssDNA-binding domain containing an oligonucleotide-oligosaccharide binding fold (OB fold) [106], while the N-terminal domain participates in the cooperative binding of gp32 monomers and the C-terminal domain mediates protein-protein interactions [76, 159, 160]. Compared with gp32, *E. coli* SSB functions as a homotetramer with each subunit containing a single OB-fold [99]. The ssDNA binding domain of *E. coli* SSB is in the N-terminal, and the C-terminal, featured by a nine conserved amino acid tip, mediates protein-protein interactions, with a non-conserved intrinsically disordered linker in between [99]. gp32 and *E. coli* SSB are the two most widely studied and well-characterized members of the SSB family and serve as the prototypes for many SSB studies to understand their functions in bacteria and higher organisms [8,78]. Generally thought as a eukaryotic homolog of *E. coli* SSB, RPA is a heterotrimeric SSB. RPA is composed of three subunits with different molecular weights of 70, 32, and 14 kDa, named RPA70, RPA32, and RPA14 with four, one, and one OB-folds, respectively [3,161,162]. One OB-fold from each subunit interacts with each other to form a stable trimerization core [90, 163]. In addition to studies of individual SSBs, researchers also investigated small groups of SSBs. Shi *et al.* compared the biological functions of a novel SSB derived from *Thermococcus kodakarensis* KOD1 with three known SSB proteins from *Thermus thermophilus*, *E. coli*, and *Sulfolobus Solfataricus* P2 [94]. They found all these four SSBs bound to ssDNA and viral RNA and affected viral RNA metabolism, but these SSBs showed different levels of resistance to heat treatment. This study provided new guidance for future exploration of novel functions of SSBs. Ashton *et al.* reviewed SSBs in the human genome, including RPA, hSSB1, and hSSB2, and discussed their roles in cellular processes for maintaining genomic stability, such as DNA replication, DNA damage repair and cell cycle-checkpoint activation [164].

Of particular interest in SSB-ssDNA interaction is specific ssDNA recognition, or the binding specificity between SSBs and ssDNA. While many SSBs bind ssDNA with high affinity but independent of sequences, some SSBs bind ssDNA with high sequence specificity, such as Telomere-end protection (TEP) proteins [5, 93]. The mechanisms underlying this binding discrepancy, however, have not been clearly elucidated. Current studies, based on individual SSBs or a small group of SSBs, consider that the binding specificity of SSBs is contributed by the electrostatic, hydrogen-bonding and stacking interactions between SSBs and ssDNA, as well as the flexibility of SSB and/or ssDNA [5, 93]. However, the roles of each of these factors in ssDNA binding specificity seem to be different among these small-scale studies. Shamoo suggested that hydrogen-bonding interactions and small pockets at the protein surface that fit only certain nucleotide bases contribute sequence specificity to TEBP' interaction with ssDNA supplemented by the generalized stacking and electrostatic interactions [5]. However, Dickey *et al.* found that despite the apparent base-specific hydrogen bonds, Pot1pC, one of the two OB-folds in *S. pombe* Pot1 was able to bind various ssDNA sequences with little to no specificity [165]. By comparing structures of Pot1pC in complex with different non-cognate ssDNA ligands, they suggested that the binding promiscuity of Pot1pC is achieved by new binding modes featured by alternate stacking interactions and new hydrogen-bonding networks [165]. In addition to base-mediated hydrogen bonding, the binding specificity also relies on the flexibility of protein and/or ssDNA [93]. The importance of flexibility of protein and ssDNA is also supported by the TEBP : $(T_4G_4)_2$ complex structure, in which the protein and ssDNA bind in a cofolding mode to induce formation of the DNA-binding pockets [5, 103]. An analysis of crystal structures of 10 different non-cognate ssDNAs complexed with the *Oxytricha nova* telomere end-binding protein (*On*TEBP), however, revealed that while the overall protein conformation in all complexes remained nearly identical to that of the cognate complex, the ssDNA exhibited subtle to dramatic conformational changes [166]. Pal and Levy also found that the ssDNA molecules were more flexible than the proteins, but they suggested that the sequence specificity was mostly introduced by the stacking interactions between aromatic residues and DNA bases [110].

While these above studies provide some aspects of protein-ssDNA interactions, to our knowledge, there are no reports of large-scale structural studies of SSB-ssDNA interactions, especially comparative studies for understanding structural features in protein-ssDNA binding specificity as in protein-dsDNA interactions. Studies of binding specificity of double-stranded DNA-binding proteins (DSBs) toward dsDNA demonstrated that hydrogen bonds between amino acid side chains and DNA bases, π-interactions between aromatic residues and DNA bases, and protein flexibility all play important roles in specific DSB-dsDNA binding [25,42,58,123]. Compared with dsDNA, ssDNA is more flexible due to the lack of the steric hindrance of the complementary DNA strand. Therefore, we hypothesize that: (1) SSBs also rely on side chain-base hydrogen bonds and protein-ssDNA π-interactions to achieve specific binding, but the contribution from π-interactions may increase compared with that in DSB-dsDNA interactions because of the increased availability of the bases; and (2) specific SSBs show larger conformational changes upon ssDNA binding than non-specific SSBs. To test our hypotheses, we collected all available protein-ssDNA complex structures from the Nucleic Acid Database (NDB) [112,113] and the Protein Data Bank (PDB) [32,33] and assigned them into specific (SP) and non-specific (NS) groups. We then carried out a comprehensive analysis by comparing the key structural features in protein-ssDNA interaction. These features include the propensities and secondary structure types of ssDNA base interacting residues, side chain-base hydrogen bonds and π-π interactions between protein and ssDNA, interaction interface, and protein conformational changes upon ssDNA binding. To our knowledge, this is the first large-scale comparative study of protein-ssDNA binding specificity, especially the roles of π-π interactions and secondary structure types in specific protein-ssDNA recognition.

3.2 Methods

3.2.1 Datasets

SSB-ssDNA complex structures, defined as any structure containing one or more protein chains and at least one single-stranded DNA, were collected from the November 2019 release (11/20/19) of NDB [112, 113] and PDB [32, 33]. From this set we excluded structures containing false ssDNA (4KMF, 3ER8, 3HZI, confirmed by NDB) and suspicious ssDNA (3G2C, 3G3Y, 3QYX, lack of evidence in primary citations). The major source of false positives in identifying SSB-ssDNA complexes comes from complexes that contain only one strand of the double helix in the asymmetric unit. These cases have been successfully filtered out by NDB, where the coordinates for the complete structure have been reconstructed by applying the transformation matrices provided in the PDB files to the half structure [112, 113]. This resulted in a dataset of 214 protein-ssDNA complexes (Supplementary Table S3).

For comparative analysis, only high-quality X-ray structures with resolution better than 3.0 Å and R-value smaller than 0.3 and NMR structures were selected. As the first step, all ssDNA-contacting chains were identified from these complexes. An ssDNA-contacting chain is a protein chain that has at least one heavy atom within 3.9 Å of any heavy atoms of a nucleotide of the ssDNA. An ssDNA-contacting chain was filtered out if: 1) the ssDNA has nucleotides other than AGCT; 2) the length of ssDNA is shorter than 3 nucleotides; 3) ssDNA is engineered, such as aptamers; and 4) there are mutated residues in the ssDNA-contacting chain.

Of these high-quality ssDNA-contacting protein chains, some only have single ssDNA-binding domains, while others also contain signal-sensing domains or dimerization domains. To avoid any potential biases, we chose ssDNA-binding domains and their target ssDNA as comparison units. CATH [27, 167], one of the most widely used structural classification databases and containing annotation information for all structures in our dataset, was used for protein structural domain annotation. An ssDNA-binding domain was selected for analysis if there are more than one protein-DNA contacts within 3.9 Å, and the domain has 40 or more amino acids. These ssDNA-binding domains were then used to generate two datasets, Dataset I and Dataset II, for comparative analyses (Figure 3.1).

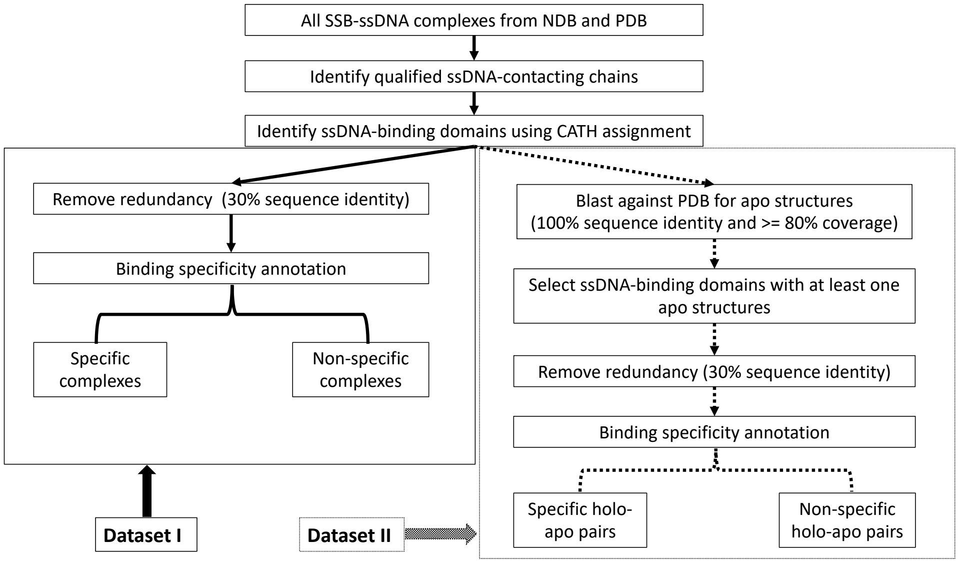

All SSB-ssDNA complexes from NDB and PDB
Identify qualified ssDNA-contacting chains
Identify ssDNA-binding domains using CATH assignment
Remove redundancy (30% sequence identity)
Binding specificity annotation
Specific complexes
Non-specific complexes
Blast against PDB for apo structures (100% sequence identity and >= 80% coverage)
Select ssDNA-binding domains with at least one apo structures
Remove redundancy (30% sequence identity)
Binding specificity annotation
Specific holo-apo pairs
Non-specific holo-apo pairs
Dataset I
Dataset II

Figure 3.1: Flowchart for compiling non-redundant specific (SP) and non-specific (NS) datasets. Dataset I: non-redundant ssDNA-binding domains in complex with their target ssDNA; Dataset II: non-redundant ssDNA-binding domains paired with their unbound structures.

Dataset I contains non-redundant complexes of SP and NS ssDNA-binding domains and their corresponding ssDNA. To generate this dataset, redundant ssDNA-binding domains were first removed using PISCES with a 30% sequence identity cutoff [168]. These non-redundant domain-based SSB-ssDNA complexes were then assigned into two groups, SP and NS, based on their binding specificity. The ssDNA binding specificity was manually annotated by referring the primary references for these structures and/or their homologs in PDB [32,33], as well as relevant information in UniProt [169]. The final non-redundant domain-based dataset contains 22 SP and 42 NS SSB-ssDNA complexes (Supplementary Table S4).

Dataset II includes non-redundant SP and NS ssDNA-binding domains paired with their corresponding apo structures. All redundant ssDNA-binding domains (holo forms) identified in the previous step were searched against PDB using default settings in NCBI BLAST Blastp program [170]. All unbound structures (apo forms) that have 100% sequence identity and at least 80% coverage with the bound form of the complex structures were selected. For X-ray apo structures, only those with resolution better than 3.0 Å and R-value smaller than 0.3 were kept; if multiple apo structures exist, the one with the best resolution and R-value was selected. An NMR apo structure was selected if there are no X-ray apo structures available. Redundant holo-apo structural pairs were then removed using PISCES with a 30% sequence identity cutoff. These non-redundant pairs were assigned into SP and NS holo-apo pairs based on their ssDNA binding specificity annotations. This resulted in 14 SP and 29 NS non-redundant holo-apo ssDNA-binding domain pairs (Supplementary Table S5).

3.2.2 Structural features of SSB-ssDNA interactions

Comparative analyses of structural features in SSB-ssDNA interactions include: 1) propensities of amino acids in contact with ssDNA, amino acids involved in side chain-base hydrogen bonds and protein-ssDNA π-π interactions, 2) protein-ssDNA contact area (PDCA), 3) number of residue-base contacts (NRBC), 4) percentages of base contacts, and 5) secondary structure types of residues involved in protein-ssDNA interactions. All structural features were extracted with widely used tools in structural bioinformatics, including PyMOL (The PyMOL Molecular Graphics System, Version 2.3.2 Schrodinger, LLC), DSSP [28, 31], HBPLUS [132], pdb-tools [171], and FreeSASA [172], and analyzed and visualized with in-house Python and R scripts.

Propensity of an amino acid that interacts with ssDNA was calculated as the ratio of the percentage of this amino acid in contact with ssDNA over the percentage of this amino acid in each data set (Equation 3.1):

$$P_{ij} = \frac{\frac{N_{ij}}{\sum_{i=1}^{20} N_{ij}}}{\frac{M_{ij}}{\sum_{i=1}^{20} M_{ij}}} \tag{3.1}$$

where P_{ij} is the propensity of amino acid i in dataset j; N_{ij} represents the total number of amino acid i in contact with DNA in dataset j; M_{ij} is the total number of amino acid i in dataset j. If $P_{ij} > 1$, amino acid i is said to be enriched in protein-DNA contacts in dataset j. When computing the propensity of an amino acid interacting with a specific nucleotide, the equation is updated as shown in Equation 3.2:

$$P_{abj} = \frac{\frac{N_{abj}}{\sum_{a=1}^{20} \sum_{b=1}^{4} N_{abj}}}{\frac{M_{aj}}{\sum_{a=1}^{20} M_{aj}} \frac{K_{bj}}{\sum_{b=1}^{4} K_{bj}}} \tag{3.2}$$

where P_{abj} is the propensity of the interacting pair of amino acid a and nucleotide b in dataset j; N_{abj} represents the total number of amino acid a in contact with nucelotide b in dataset j; M_{aj} is the total number of amino acid a in dataset j and K_{bj} is the total number of nucleotide b in dataset j . If $P_{abj} > 1$, contact between amino acid a and nucleotide b is said to be enriched in protein-DNA contacts in dataset j.

PDCA was calculated by subtracting the solvent accessible surface area (SASA) of a protein-ssDNA complex from the sum of solvent SASAs of its protein and DNA components and divided by two (Equation 3.3). The solvent accessible surface area was measured by FreeSASA [172].

$$PDCA = \frac{SASA_{protein} + SASA_{DNA} - SASA_{complex}}{2} \tag{3.3}$$

Protein-ssDNA contacts were defined using a distance cutoff of 3.9 Å between side chain heavy atoms of an amino acid and all heavy atoms of a nucleotide. These protein-ssDNA contacts were further divided into two subsets depending on the parts of DNA involved: 1) NRBC for the number of residue and DNA base contacts, and

2) NRBbC for the number of residue and DNA backbone contacts.

To investigate the roles of secondary structure types in protein-ssDNA binding specificity, DSSP program was applied to assign residues involved in protein-DNA interactions into three general secondary structure types: helix, strand, and coil, where H (α-helix), G (3_{10}-helix) and I (π-helix) states from DSSP are assigned as helix type, E (extended strand) and B (residue in isolated β-bridge) states from DSSP are classified as strand type, and all the other states are considered as coil type [28–31, 123]. An important background for evaluating the enrichment of secondary structure types of residues involved in side chain-base hydrogen bonds and protein-DNA π-π interactions is the secondary structure type distribution of DNA base-contacting residues. An amino acid is defined as a DNA base-contacting residue if it has at least one heavy atom of its side chain within 3.9 Å of any heavy atom of a DNA base.

In addition, conformational changes of SSBs upon ssDNA binding were measured by comparing both mainchain root mean square deviation (RMSD) and interface RMSD (IRMSD) between bound (holo) ssDNA-binding domains and their unbound (apo) structures. Interface residues are residues that have at least one heavy atom within 10 Å of any heavy atoms of DNA. RMSD and IRMSD were calculated using the PyMOL *align* command for all heavy atoms in ssDNA-binding domains and in-house python scripts for heavy atoms of interface residues aligned with TM-align respectively [173].

3.2.3 Statistical tests

For statistical analyses between two groups, Shapiro-Wilk test was performed first to test the normality of the data. If the data is normally distributed, a parametric Student's t-test was carried out. Otherwise, a non-parametric Wilcoxon rank-sum test was applied. To test the association of interplanar angle distributions of protein-ssDNA π-π interactions between two groups, chi-square test or Fisher's exact test was performed based on the sample size and expected values.

3.3 Results

3.3.1 Amino acid propensity for protein-ssDNA interaction

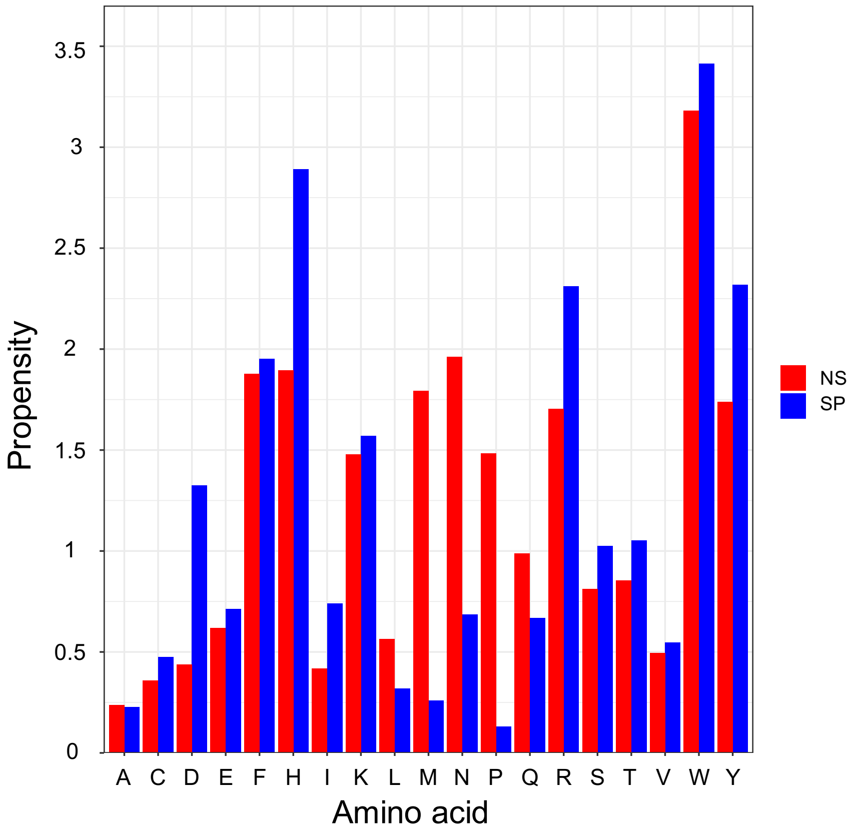


Figure 3.2: Propensities of amino acids contacting DNA bases in SP and NS groups.

Overall propensity of residues involved in side chain-base contacts. As shown in Figure 3.2, aromatic and positively charged amino acids phenylalanine (F), histidine (H), tryptophan (W), tyrosine (Y), lysine (K), and arginine (R) are enriched in both SP and NS groups. Of these six residues, histidine, tyrosine, and arginine are more enriched in the SP group than those in the NS group. Four aromatic residues (F,H,W,Y) are likely involved in protein-ssDNA π-π interactions, and two positively charged residues (K,R) can both form hydrogen bonds with DNA bases and interact with negatively charged DNA backbone. In addition, the SP group is distinctly enriched in aspartate, while the NS group shows a high enrichment of methionine, proline, and asparagine. Most nonpolar residues, except for methionine, and poplar residue cysteine do not show enrichment in either group. To further explore the underlying mechanisms of these enrichment patterns, propensities of residues interacting with different nucleotides were calculated.

Propensity of aromatic residues involved in protein-ssDNA π-π interactions with different nucleotides. Of all interactions between aromatic residues and nucleotides, we identified protein-ssDNA π-π interactions by visually inspecting each nucleobase-aromatic amino acid dimer using PyMOL. While aromatic residues in contact with ssDNA are all enriched in both groups (Figure 3.2), those involved in protein-ssDNA π-π interactions show different preferences for nucleotides between two groups. Figure 3.3A shows that while tryptophan-thymine is enriched in both NS and SP groups, it is even more enriched in the SP group. Tryptophan also shows preference to cytosine in the SP group. Another enriched residue, histidine, prefers guanine and adenine in the NS group but favors thymine in the SP group. To investigate if there are any differences in geometry types of π-π interactions between these two groups, we measured the interplanar angle (ω) between the two aromatic planes using the *angle_between_helices* command in PyMOL psico module. These π-π interactions were classified as stacked ($0 \leq \omega \leq 20°$), inclined ($20° < \omega < 70°$), and T-shaped ($70° \leq \omega \leq 90°$) types as described in [41]. Since the expected values of some cells are less than five, Fisher's exact tests were performed to compare the interplanar angle distributions between two groups for individual aromatic residues as well as group-wise comparisons. No significant differences were found between these two groups in terms of the geometry of protein-ssDNA π-π interactions (p-value=0.08654, Table 3.1).

Propensity of residues involved in side chain-base hydrogen bonds with different nucleotides. Figure 3.3B shows that the SP group has a larger number of residues

Table 3.1: Frequency of geometry types of protein-ssDNA π-π interactions and Fisher's exact test results between specific and non-specific groups. (Stacked: [0,20°]; inclined: (20°-70°); T-shaped: [70°,90°])

Amino acid	Specific			Non-specific			P-value
	stacked	inclined	T-shaped	stacked	inclined	T-shaped	
Phenylalanine	8	8	1	6	6	5	0.8367
Histidine	6	3	2	1	7	3	0.1063
Tryptophan	4	4	2	5	5	5	0.7733
Tyrosine	7	3	3	4	4	3	0.5045
All	25	18	8	16	22	16	0.08654

enriched in side chain-base hydrogen bonds, but the most enriched pairs are in the NS group, led by histidine-guanine and followed by asparagine, arginine with guanine and histidine with adenine. Raw counts of these most enriched pairs, however, are relatively small: only one histidine-guanine, one asparagine-guanine, one histidine-adenine, and two arginine-guanine pairs were observed in the NS group, and one tryptophan-thymine pair was found in the SP group. Therefore, these propensity values are not robust enough to be considered as significant.

Two enriched pairs, aspartate-guanine (propensity=5.620) and lysine-guanine (propensity=5.439), however, do have relatively large raw counts (seven and eight pairs respectively) and are exclusively detected in the SP group. Of all 11 side chain-base hydrogen bonds formed between seven aspartate-guanine pairs, eight are bidentate (six, where two hydrogen bonds formed with a DNA base via two pairs of hydrogen bond donors and acceptors) or bifurcated (two, where one hydrogen bond acceptor/donor is shared by two hydrogen bonds) hydrogen bonds formed with the same guanine bases [42]. Interestingly, DNA base atoms involved in these hydrogen bonds are those typically form Watson-Crick base pairs in dsDNA (designated as WC atoms in this study). For instance, OD2 atom (acceptor) of ASP223 on chain A of the *Oxytricha nova* telomere end binding protein (*On*TEBP, PDBID: 1OTC) forms bifurcated hydrogen bonds with H(N1) and H(N2) atoms (donors) of guanine 4 on a single strand telomeric DNA, where these two donor atoms are WC atoms (Fig-

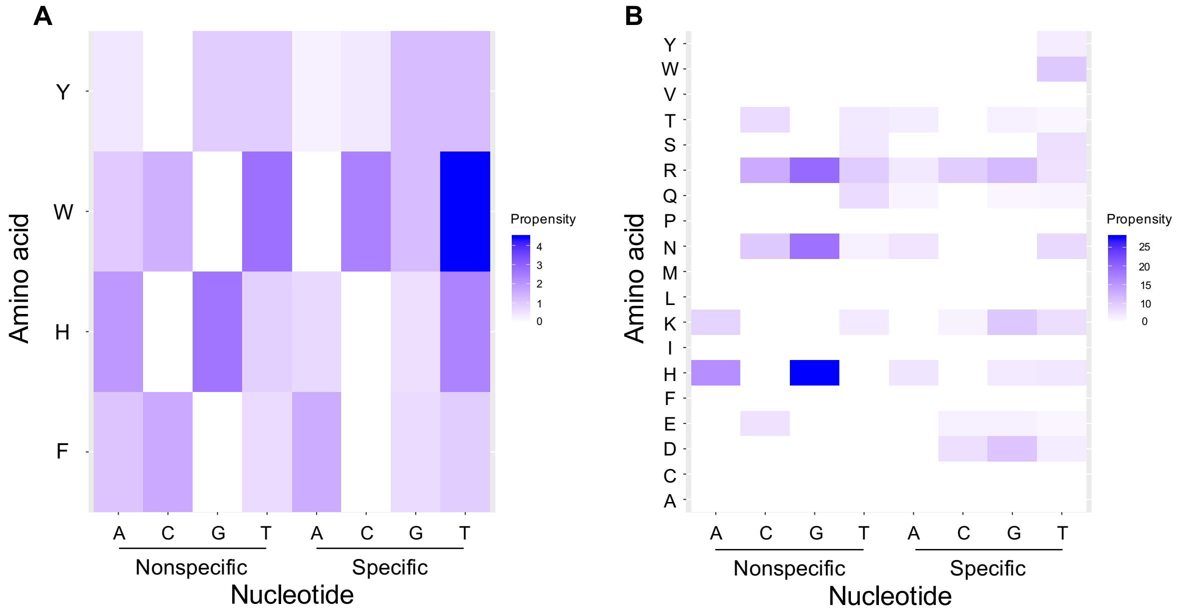


Figure 3.3: Propensities of aromatic residues involved in protein-ssDNA π-π interactions (A) and residues involved in side chain-base hydrogen bond with different DNA bases (B) in SP and NS groups.

ure 3.4A). On the other hand, ASP42 on chain A of the unwinding protein (UP1) forms bidentate hydrogen bonds using OD1 and OD2 atoms as acceptors with H(N2) and H(N1) atoms (donors) of guanine 205 respectively on a human telomeric repeat (PDBID: 1PGZ, Figure 3.4B). The second enriched pair with a large number of raw counts, lysine-guanine, also forms four bidentate hydrogen bonds while the remaining six are simple hydrogen bonds. Unlike the aspartate-guanine pair, most of DNA base atoms involved in side chain-base hydrogen bonding (N7, 7 out of 10) in these lysine-guanine pairs are non-WC atoms.

3.3.2 Protein-ssDNA side chain-base hydrogen bonds

Given these above findings based on side chain-base hydrogen bonds, we further investigated the percentages of side chain-base hydrogen bonds in all protein-DNA hydrogen bonds. Percentages of side chain-base hydrogen bonds in each complex from HBPLUS are shown in Figure 3.5A and B, with percentages of side chain-base hydrogen bonds shown at the bottom in a descending order. Complexes in the

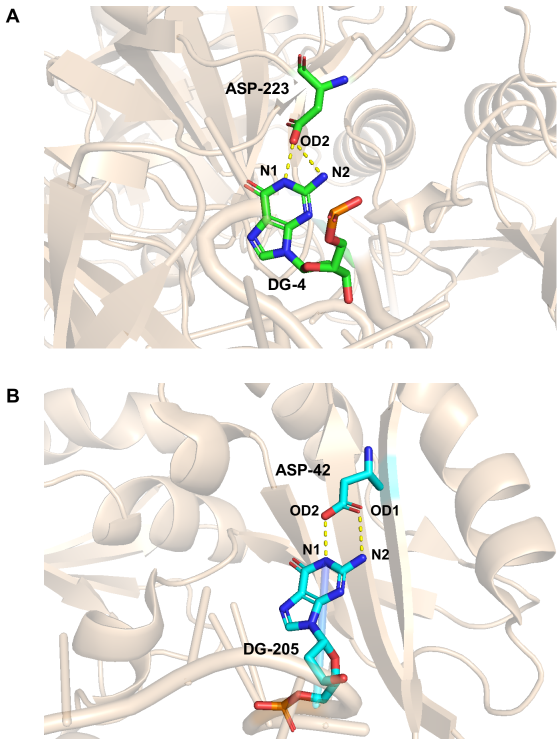


Figure 3.4: Examples of aspartate forming bifurcate (A) and bidentate (B) hydrogen bonds with the same guanine in the SP group. Hydrogen bonds are represented by yellow dashed lines. (A) Bifurcate hydrogen bonds. OD2 atom (acceptor) of ASP223 on the *Oxytricha nova* telomere end binding protein (*On*TEBP, PDBID: 1OTC; protein chain: A; DNA chain: D) forms bifurcated hydrogen bonds with H(N1) and H(N2) atoms (donors) of guanine 4. These two donor atoms are Watson-Crick atoms. (B) Bidentate hydrogen bonds. OD1 and OD2 atoms (acceptors) of ASP42 on the unwinding protein (UP1) form bidentate hydrogen bonds with H(N2) and H(N1) atoms (donors) of guanine 205 on a human telomeric repeat (PDBID: 1PGZ; protein chain: A; DNA chain: B).

SP group generally show large contributions of side chain-base hydrogen bonds to the total number of protein-DNA hydrogen bonds. Of note are two NS complexes, domains 3kqlA03 and 4j1jA02, have exclusively side chain-base hydrogen bonds with their bound ssDNA. Their raw counts, however, are very small with only one for each complex. About 82% (18 of 22) of the SP complexes form side chain-base hydrogen bonds and ∼55% (12 of 22) of complexes have percentages of side chain-base hydrogen bonds equal to or above 50% (Figure 3.5B). The NS group, on the other hand, only has ∼47% (18 of the total 38 complexes that have at least one protein-DNA hydrogen bond) of the cases form side chain-base hydrogen bonds and ∼13% (5 of 38) of the complexes have percentages of side chain-base hydrogen bonds equal to or above 50% (Figure 3.5A). Wilcoxon rank-sum test shows that the difference between these two groups is significant with a p-value of 0.00023 (Figure 3.5C).

As the enriched aspartate-guanine and lysine-guanine pairs in the SP group show different degrees of involvement of WC atoms, we also compared the percentages of WC atom-based hydrogen bonds in all side chain-base hydrogen bonds between these two groups. Percentages of WC atom-based hydrogen bonds in each complex from HBPLUS are shown in Figure 3.5D and E, with percentages of WC atom-based hydrogen bonds shown at the bottom in a descending order. Overall, complexes in the SP group show large percentages of WC atom-based hydrogen bonds. About 94% (17 of the total 18 complexes that have side chain-based hydrogen bonds) of the SP complexes form WC atom-based hydrogen bonds and all these 17 complexes (100%) have percentages of side chain-base hydrogen bonds larger than or equal to 50% (Figure 3.5E). The NS group, on the other hand, only has ∼39% (7 of the total 18 complexes that have at least one side chain-base hydrogen bond) of the cases form WC atom-based hydrogen bonds and ∼28% (5 of 18) of the complexes have percentages of side chain-base hydrogen bonds no less than 50% (Figure 3.5D). Wilcoxon rank-sum test shows that the percentages of WC atom-based hydrogen bonds between the SP and NS groups are significantly different (p-value=0.013) (Figure 3.5F).

3.3.3 Protein-ssDNA interaction interface

Figure 3.6 shows that there is no significant PDCA difference between the SP and NS groups (Figure 3.6A, p-value=0.22), but the SP complexes have larger number of NRBC than the NS cases and the difference is statistically significant (Figure 3.6B, p-value=0.012). The difference becomes bigger after normalizing the raw NRBC counts with PDCA, and the SP group has more residue-base contacts per 1000 Å^2 contact area (Figure 3.6C, p-value=0.0032). In addition, for most cases, more than half of

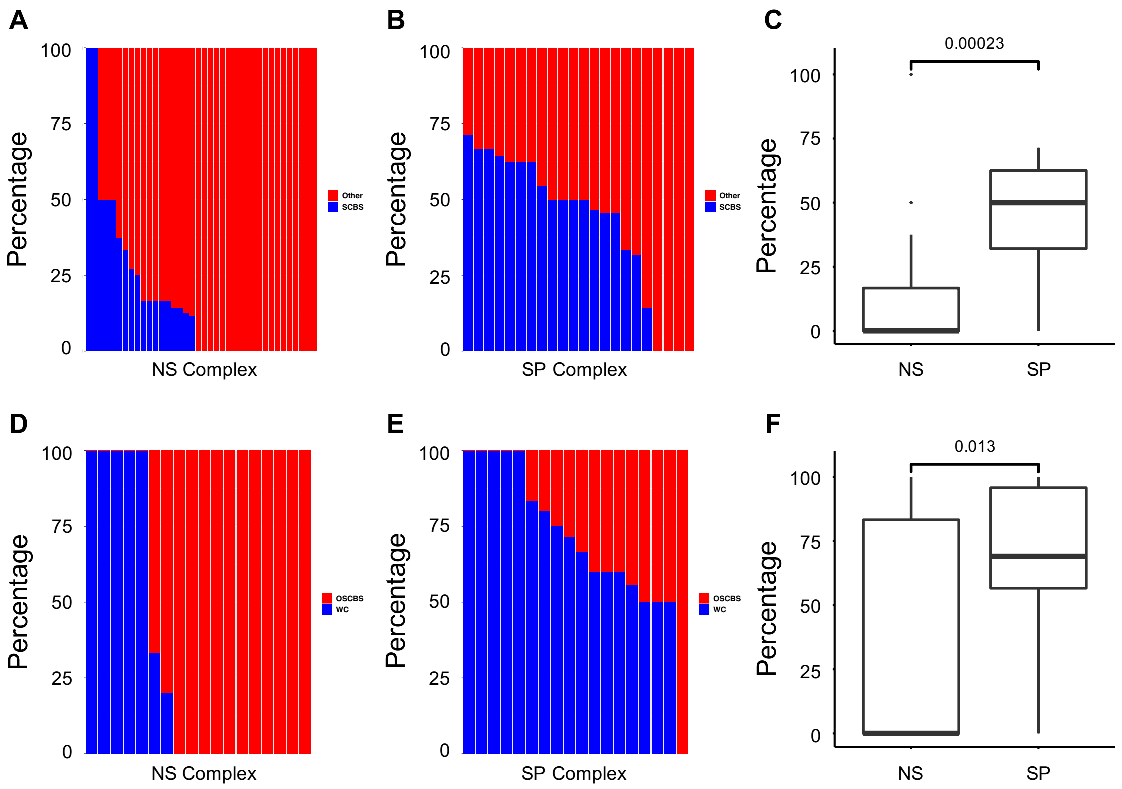


Figure 3.5: Comparisons of percentages of side chain-base hydrogen bonds (A-C) and Watson-Crick atom-based side chain-base hydrogen bonds (D-F) annotated by HBPLUS between the SP and NS ssDNA-binding proteins. (A) Percentages of side chain-base hydrogen bonds in all protein-DNA hydrogen bonds in NS complexes. (B) Percentages of side chain-base hydrogen bonds in all protein-DNA hydrogen bonds in SP complexes. Side chain-base hydrogen bonds (SCBS, colored in blue) are shown at the bottom in a descending order, while all other protein-DNA hydrogen bonds (Other, colored in red) are on the top. (C) Boxplot and statistical analysis for comparison between two groups. (D) Percentages of Watson-Crick atom-based side chain-base hydrogen bonds in all side chain-base hydrogen bonds in NS complexes. (E) Percentages of Watson-Crick atom-based side chain-base hydrogen bonds in all side chain-base hydrogen bonds in SP complexes. Watson-Crick atom-based side chain-base hydrogen bonds (WC, colored in blue) are shown at the bottom in a descending order, while all other side chain-base hydrogen bonds (OSCBS, colored in red) are on the top. (F) Boxplot and statistical analysis for comparison between two groups. P-values are displayed on top of the boxplots.

protein-ssDNA contacts are residue-base contacts, with larger percentages of NRBC in the SP group (Figure 3.6D, p-value=0.052).

Compared with our previous study of protein-dsDNA binding specificity [25], DNA-binding domains interact with dsDNA and ssDNA in a similar PDCA range, but

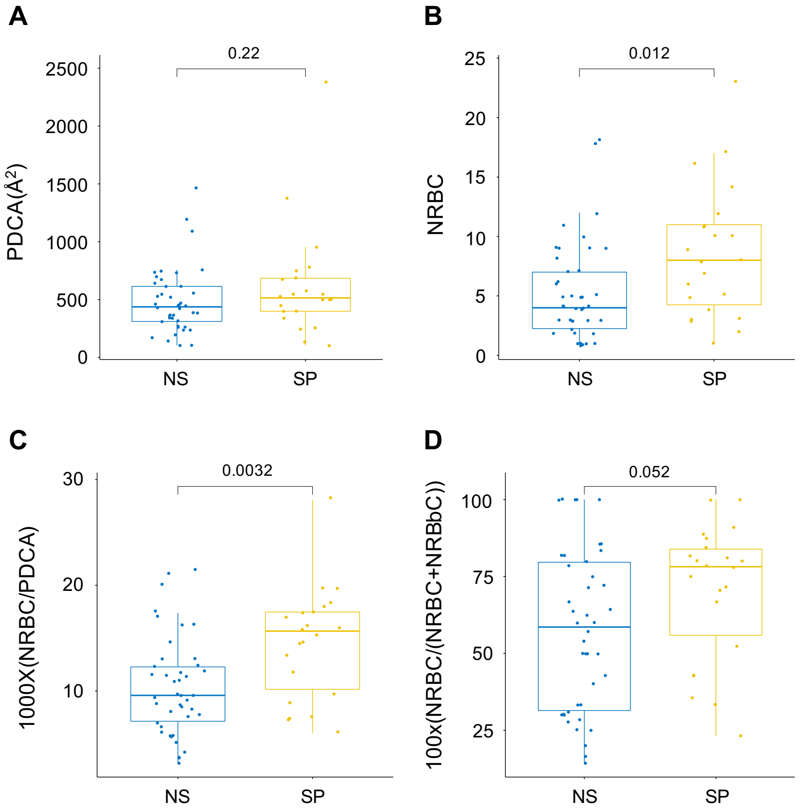


Figure 3.6: Comparison of protein-ssDNA interactions. (A) Protein-DNA contact area (PDCA); (B) number of residue-base contacts (NRBC); (C) NRBC density, NRBC normalized to PDCA; and (D) percentage of NRBC in all protein-DNA contacts, the sum of NRBC and NRBbC (number of residue-DNA backbone only contacts). P-values are displayed on top of the boxplots.

protein-ssDNA interactions have more contacts between residues and DNA bases, in terms of both raw and normalized NRBC counts (detailed information of protein-dsDNA binding specificity can be found in Figure 4 in [25]). One major difference between protein-dsDNA and protein-ssDNA interactions is that in protein-ssDNA complexes, residue-base contacts dominate the interaction between residue and ssDNA, while protein-dsDNA interactions show the opposite trends that most of residue-DNA contacts are formed between residues and DNA backbone (see Figure 5A in [25] for information about protein-dsDNA interaction). These differences are largely due to the increased accessibility of ssDNA base atoms compared to dsDNA.

3.3.4 Protein conformational changes upon ssDNA binding

To explore protein conformational changes upon ssDNA binding, we first calculated all heavy-atom RMSD between ssDNA-binding domains and their corresponding apo conformations. As can be seen in Figure 3.7A, most ssDNA-binding domains do not change dramatically upon ssDNA binding, and majority of the RMSD values are less than 2 Å. However, there are a few individual cases that show relatively large conformational changes. The largest change comes from a pair in the SP group, domain 3c2pA07 in Coliphage N4 virion-encapsidated RNA polymerase (vRNAP) and its apo structure 2po4 (RMSD: 6.532 Å). Figure 3.7B shows that the motif B of N4 vRNAP rearranges its structure from a loop (apo form, green) to a short anti-parallel α-hairpin (holo form, magenta). This change and other conformational changes transit the polymerase from the inactive state to an active form to accommodate the binding of incoming DNA [174]. Statistical analysis shows no significant RMSD difference between these two groups (Figure 3.7A, p-value=0.37). Similarly, no significant IRMSD difference was found between these two groups (p-value=0.89) (Figures 3.7C).

3.3.5 Secondary structure types of ssDNA interacting residues

ssDNA-binding proteins recognize their target sites with four major structural topologies: OB folds, K homology (KH) domains, RNA recognition motifs (RRMs), and whirly domains [93]. The secondary structure types of amino acids involved in protein-ssDNA interactions, however, have not been investigated extensively. We first compared the propensities of secondary structure types of amino acids in ssDNA-binding domains that are in contact with DNA bases, calculated against the relative frequencies of secondary structure types of all residues in the respective group of ssDNA-binding domains. DNA base-contacting residues in both groups are enriched in strand conformations with a higher enrichment in the SP group, while coil secondary structure types are also preferred in the NS group (Figure 3.8A).

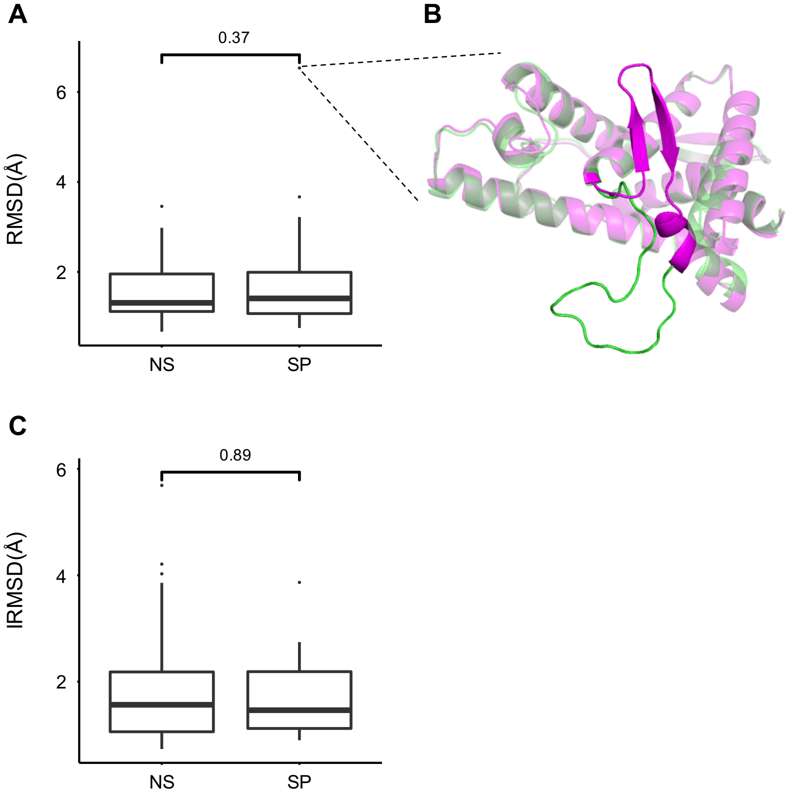


Figure 3.7: Conformational changes upon ssDNA binding. (A) RMSD of heavy atoms of all residues between bound (holo) and unbound (apo) structures. (B) Structural alignment of domain 3c2pA07 (magenta) in Coliphage N4 virion-encapsidated RNA polymerase (vRNAP) and its apo structure 2po4 (green; overall RMSD: 6.532 Å). (C) RMSD of heavy atoms of all interface residues (IRMSD) between bound (holo) and unbound (apo) structures.

For residues that form hydrogen bonds between their side chains and DNA bases, we used two different background distributions to calculate the propensities: one is the secondary structure type distribution of all base-contacting residues (Figure 3.8B) and the other is the secondary structure type distribution of all residues that form hydrogen bonds with DNA including bases and backbone atoms (Figure 3.8C). A similar trend is seen no matter what background distribution is used: residues involved in side chain-base hydrogen bonds in both groups are enriched in strand conformations. However, between these two groups, these residues in the NS group have higher propensity for strands than those in the SP group. The propensity for

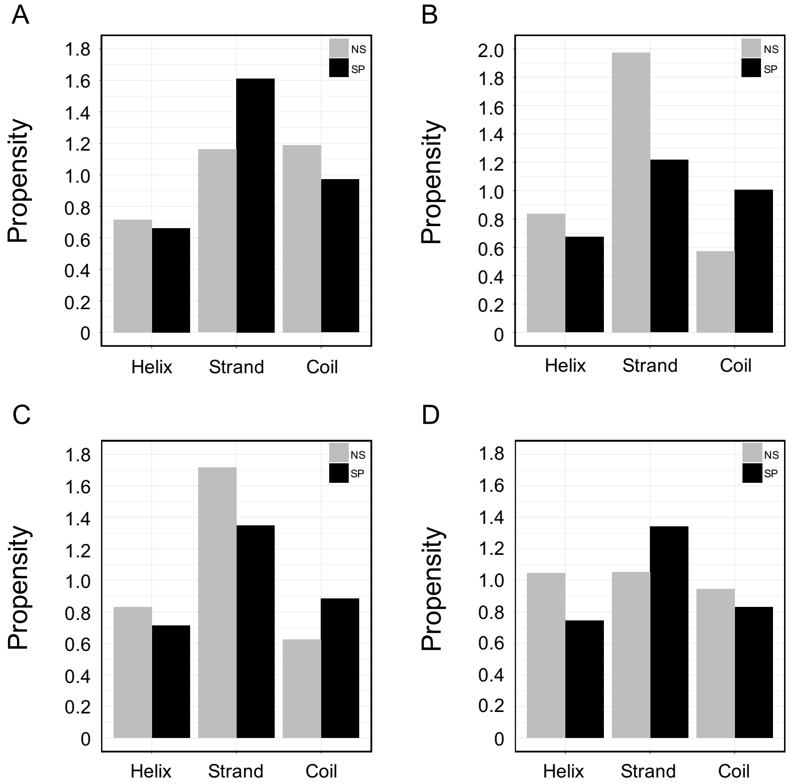


Figure 3.8: Propensities of secondary structure types in the SP and NS groups. (A) Propensities of secondary structure types of DNA base-contacting residues. The background distributions of secondary structure types were calculated using all residues in the ssDNA-binding domains in each group. Propensities of secondary structure types of residues involved in side chain-base hydrogen bonds (B, C) and residues involved in protein-ssDNA π-π interactions (D) with HBPLUS. Propensities were calculated using either the distribution of secondary structure types of base-contacting residues (B, D) or all DNA hydrogen-bonding residues (C).

coil types in the SP group is larger than that in the NS group (Figure 3.8B and C). These results suggest that for residues involved in side chain-base hydrogen bonds, relatively more such residues in the SP group are adopting coil conformation, which represents a more flexible conformation. However, we would like to point out that results in Figure 3.8B and 3.8C need to be interpreted with caution as the raw counts of secondary structure types in the NS group are relatively small (Helix: 6, strand: 13, and coil: 7).

In addition, we calculated the propensities of residues involved in protein-ssDNA

π-π interactions using the secondary structure type distribution of all base-contacting residues as the background. Figure 3.8D shows that the SP group prefers strand types, while both strand and helix conformations are slightly enriched in the NS group.

3.4 Discussion

We presented here the first large-scale comparative study of SSB-ssDNA interactions with a focus on the binding specificity. Our results suggest that side chain-base hydrogen bonds play the major role in protein-ssDNA binding specificity, while protein-ssDNA π-π interactions may contribute to binding affinity. Although conformational changes of both ssDNA and protein are important in SSB-ssDNA binding, our results indicate that conformational changes of ssDNA might play a larger role than those of protein in specific ssDNA recognition.

Our comparative analyses show that the SP group forms more contacts with DNA bases than the NS group (Figure 3.6), and propensities of amino acids that involved in protein-ssDNA contacts show that both groups prefer aromatic residues and positively charged residues, but H, Y, R are more enriched in the SP group (Figure 3.2). These findings are consistent with previous studies [110, 175, 176]. The enrichment of all aromatic residues can be attributed to the increased accessibility of ssDNA. Without the steric hindrance from the complementary strand, ssDNA can change conformation relatively easily so that DNA bases become more accessible to aromatic residues for π-π interactions, which are suggested to be determinants of specific ssDNA recognition [110]. The distributions of π-π geometry types between SP and NS groups, however, do not show significant differences in this comparative analysis (Table 3.1), suggesting in general protein-ssDNA π-π interactions may mainly contribute to binding affinity though in individual cases π-π interactions can contribute to specific protein-ssDNA recognition. While there are no π-π geometry studies in protein-DNA binding specificity, using 428 protein-DNA complexes, Wilson *et al.* found that the stacked orientation (58%) is more common than the inclined configuration (29%), with the T-shaped interaction as the least frequent one (13%) [41]. By combining the frequencies of each geometry type of π-π interactions in both NS and SP groups in Table 3.1, we found that the stacked orientation represents 39% in protein-ssDNA π-π interactions, similar to the inclined configuration (38.1%) with the T-shaped interaction at 22.9%.

The most interesting structural feature is the enrichment of aspartate in the SP group, but not in the NS group (Figure 3.2). This difference is more distinct at the side chain-base hydrogen bond level comparison, where aspartate forms side chain-base hydrogen bonds with all nucleotides except for adenine in the SP group but none such hydrogen bonds were found in the NS group (Figure 3.3B). Out of three types of contacting nucleotides, the preference of aspartate to guanine is the most dominant one. Structural inspections show that this preference is achieved via bidentate and/or bifurcate hydrogen bonds between aspartate and WC atoms of DNA bases of the same guanines (Figure 3.4). Aspartate was also reported to be enriched in specific protein-dsDNA binding but with different binding characteristics: in the dsDNA study, aspartate favoured cytosine, and most (10 out of total 19) of aspartate-cytosine side chain-base hydrogen bonds were bidentate hydrogen bonds formed with two consecutive cytosines in the major groove [25].

In addition to guanine, aspartate also shows preferences to cytosine (propensity=3.042) and thymine (propensity=1.865), and this is consistent with the critical role of aspartate in mutagenesis driven by the cytidine deaminase APOBEC in cancer [177]. APOBEC mutations, especially those driven by APOBEC3A and APOBEC3B, are sensitive to cytosines in TpC sites in hairpin (stem-loop) DNA structures formed while transiently single-stranded in a sequence specific manner [156, 177, 178]. For instance, Shi *et al.* solved crystal structures of human APOBEC3A and a chimera of human APOBEC3B and APOBEC3A in complex with ssDNA to study the structural basis for targeted DNA cytosine deamination and mutagenesis by APOBEC3A and APOBEC3B [177]. Their results suggested the loop regions of DNA stem-loop structures may be hotspots for mutagenesis. Particularly, they found aspartate 131 (D131) strongly influenced the preference of the upstream nucleotide of the target cytosine in the TpC sites: substitution by a small non-polar alanine (D131A) decreased selectivity, and glutamate substitution (D131E) converted the preference to cytosine from thymine, while threonine substitution (D131T) retained selectivity [177] (Figure 3.9). This indicates that aspartate could be a potential drug target for treating APOBEC-mediated mutagenesis. Additionally, they found that two neighboring tyrosine residues (Y130, Y132) were also important in conferring the selectivity (78) (Figure 3.9). These findings might suggest the combination of side chain-base hydrogen bonds and protein-ssDNA π-π interactions and thus short oligomers of residues forming side chain-base hydrogen bonds and/or aromatic residues are likely to be a signature for conferring ssDNA-binding specificity. Despite all three APOBEC3(A/B)-ssDNA complex structures in PDB (PDBID: 5KEG, 5SWW, 5TD5) were excluded in this study due to different numbers of mutations in the protein, preferences of aspartate to cytosine and thymine only in the SP group suggest these patterns are likely a general feature of the SP group rather than a unique feature of the APOBEC family.

The enrichment of aspartate suggests the important role of side chain-base hydrogen bonds in protein-ssDNA binding specificity, which is further supported by significant larger percentages of overall and WC atom-based side chain-base hydrogen bonds in the SP group than the NS group (Figure 3.5). Without the complementary strand in ssDNA, atoms normally forming Watson-Crick base pairs in the dsDNA offer more hydrogen bond donors/acceptors to facilitate binding specificity and/or affinity. Comparisons based on interacting aromatic amino acids and protein-ssDNA π-π interactions, on the other hand, do not show significant differences between two groups. Taken together, our results suggest that side chain-base hydrogen bonds,

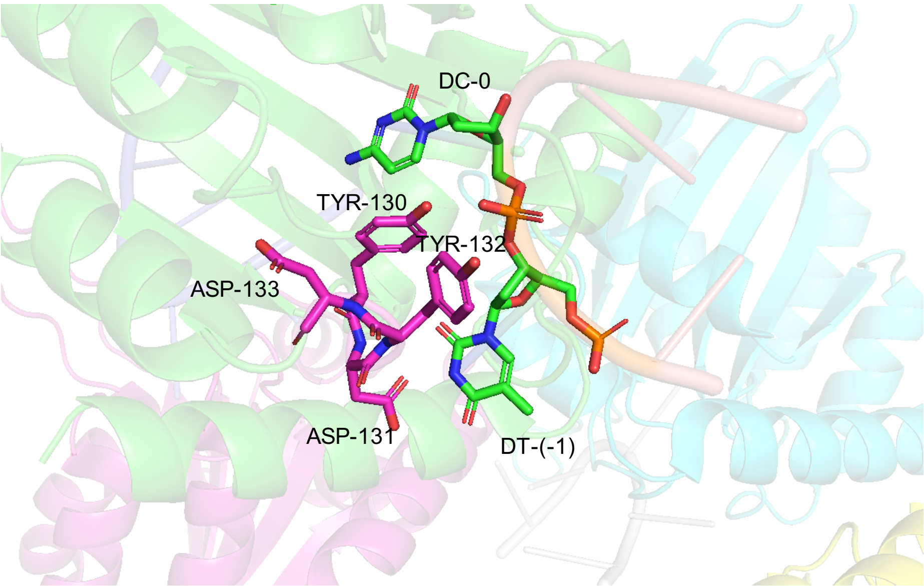


Figure 3.9: Structural example of the involvement of aspartate in the mutagenesis on hairpin TpC sites driven by APOBEC3A (PDBID: 5SWW). Asp131 and two neighboring tyrosine residues (TYR130, TYR132) play important role in determining the preference on the nucleotide (Thymine -1) upstream of the target cytosine (DC-0). In addition, A second aspartate (ASP133) locates right next to TYR132.

especially bidentate and/or bifurcated hydrogen bonds, are major determinants in specific protein-ssDNA binding, while protein-ssDNA π-π interactions might mainly contribute to binding affinity.

Upon ssDNA binding, conformational changes of ssDNA-binding domains between two groups in terms of heavy atom RMSD of both all residues and interface residues only, do not show significant differences (Figure 3.7). This indicates that while specific protein-ssDNA recognition relies on the flexibility of protein and/or ssDNA in some cases [93], the conformational change of ssDNA may play a larger role than that of protein. This is consistent with what Theobald and Schultz found from structural and thermodynamics comparisons of the cognate *Oxytricha nova* telomere end-binding protein (*On*TEBP)-ssDNA complex with 10 different non-cognate ssDNAs complexed with the *On*TEBP [166]. They found that while protein conformations in all non-cognate complexes remained nearly identical to that in the cognate complex, the ssDNA exhibited dramatic differences in three non-cognate complexes and subtle conformational changes in other seven non-cognate complexes [166]. This study revealed the great plasticity of the *On*TEBP in accommodating non-cognate ssDNA sequences, especially via a phenomenon they named nucleotide shuffling—conformational rearrangements via shifts in the ssDNA register of various number of nucleotides [166].

DNA base-contacting residues in specific ssDNA-binding proteins are enriched in strands while non-specific ssDNA-binding proteins show preferences to strands and coils (Figure 3.8A). The secondary structure type preferences of specific SSBs are different from those of specific DSBs, including both highly specific (HS) and multi-specific (MS) DSBs, where HS DSBs prefer coils and MS DSBs favor helices [123]. These results are largely in agreement with the conformational studies in both specific protein-dsDNA and protein-ssDNA interactions. In protein-ssDNA complexes, there is less conformational change after binding DNA that is consistent with the larger strand propensity while in specific protein-dsDNA interactions, proteins are more flexible and relatively more coil conformations are enriched. The SP group has a larger propensity of coils than the NS group (Figure 3.8B), indicating protein flexibility plays a role in the binding specificity. Protein flexibility impacts DNA recognition likely via speeding up locating DNA-binding proteins to their target sites [25, 56, 58, 59, 147, 148]. In addition, it is suggested that the loops/linkers connecting ssDNA-binding domains, especially their lengths, are responsible for binding specificity [93, 175, 179]. For residues involved in protein-ssDNA π-π interactions, the SP group shows higher level of enrichment in strands than the NS group while the latter also slightly favours helices. This might indicate that specific ssDNA-binding proteins are more sensitive to mutations, as helices are more robust to mutations than strands [122].

To our knowledge, this is the first large-scale comparative structural study of protein-ssDNA interactions. This study expands our knowledge of SSB-ssDNA interactions, especially the mechanisms underlying the binding discrepancy between SSBs with different degrees of binding specificity. Findings from this study can also improve SSB-ssDNA binding prediction models. By incorporating hydrogen bond especially side chain-base hydrogen bond information, the energy-based coarse-grained model developed by Pal and Levy may improve prediction accuracy [110]. Moreover, this study can enrich current databases that analyze protein-DNA interactions. Recently, DNAproDB has taken SSB-ssDNA interactions into consideration by mainly focusing on individual SSB-ssDNA complexes [111]. While our study is limited by the relatively small number of currently available protein-ssDNA complexes, we believe that the trends and general conclusions will not change when more protein-ssDNA complex structures become available in the future, just as the initial protein-dsDNA studies and a number of other structural bioinformatics investigations in their early days.

REFERENCES

[1] O. Fornes, J. A. Castro-Mondragon, A. Khan, R. van der Lee, X. Zhang, P. A. Richmond, B. P. Modi, S. Correard, M. Gheorghe, D. Baranašić, *et al.*, "Jaspar 2020: update of the open-access database of transcription factor binding profiles," *Nucleic acids research*, 2019.

[2] N. M. Luscombe, S. E. Austin, H. M. Berman, and J. M. Thornton, "An overview of the structures of protein-dna complexes," *Genome biology*, vol. 1, no. 1, pp. reviews001–1, 2000.

[3] M. S. Wold, "Replication protein a: a heterotrimeric, single-stranded dna-binding protein required for eukaryotic dna metabolism," *Annual review of biochemistry*, vol. 66, no. 1, pp. 61–92, 1997.

[4] D. J. Richard, E. Bolderson, L. Cubeddu, R. I. Wadsworth, K. Savage, G. G. Sharma, M. L. Nicolette, S. Tsvetanov, M. J. McIlwraith, R. K. Pandita, *et al.*, "Single-stranded dna-binding protein hssb1 is critical for genomic stability," *Nature*, vol. 453, no. 7195, p. 677, 2008.

[5] Y. Shamoo, "Single-stranded dna-binding proteins," *Encyclopedia of life sciences*, pp. 1–7, 2002.

[6] D. J. Richard, E. Bolderson, and K. K. Khanna, "Multiple human single-stranded dna binding proteins function in genome maintenance: structural, biochemical and functional analysis," *Critical reviews in biochemistry and molecular biology*, vol. 44, no. 2-3, pp. 98–116, 2009.

[7] P. Gu, W. Deng, M. Lei, and S. Chang, "Single strand dna binding proteins 1 and 2 protect newly replicated telomeres," *Cell research*, vol. 23, no. 5, pp. 705–719, 2013.

[8] R. D. Shereda, A. G. Kozlov, T. M. Lohman, M. M. Cox, and J. L. Keck, "Ssb as an organizer/mobilizer of genome maintenance complexes," *Critical reviews in biochemistry and molecular biology*, vol. 43, no. 5, pp. 289–318, 2008.

[9] G. D. Stormo and Y. Zhao, "Determining the specificity of protein–dna interactions," *Nature Reviews Genetics*, vol. 11, no. 11, p. 751, 2010.

[10] S. J. Maerkl and S. R. Quake, "A systems approach to measuring the binding energy landscapes of transcription factors," *Science*, vol. 315, no. 5809, pp. 233–237, 2007.

[11] C. T. Campbell and G. Kim, "Spr microscopy and its applications to high-throughput analyses of biomolecular binding events and their kinetics," *Biomaterials*, vol. 28, no. 15, pp. 2380–2392, 2007.

[12] M. F. Berger, A. A. Philippakis, A. M. Qureshi, F. S. He, P. W. Estep III, and M. L. Bulyk, "Compact, universal dna microarrays to comprehensively determine transcription-factor binding site specificities," *Nature biotechnology*, vol. 24, no. 11, p. 1429, 2006.

[13] C. L. Warren, N. C. Kratochvil, K. E. Hauschild, S. Foister, M. L. Brezinski, P. B. Dervan, G. N. Phillips, and A. Z. Ansari, "Defining the sequence-recognition profile of dna-binding molecules," *Proceedings of the National Academy of Sciences*, vol. 103, no. 4, pp. 867–872, 2006.

[14] A. R. Oliphant, C. J. Brandl, and K. Struhl, "Defining the sequence specificity of dna-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast gcn4 protein.," *Molecular and cellular biology*, vol. 9, no. 7, pp. 2944–2949, 1989.

[15] E. Roulet, S. Busso, A. A. Camargo, A. J. Simpson, N. Mermod, and P. Bucher, "High-throughput selex–sage method for quantitative modeling of transcription-factor binding sites," *Nature biotechnology*, vol. 20, no. 8, p. 831, 2002.

[16] M. Slattery, T. Riley, P. Liu, N. Abe, P. Gomez-Alcala, I. Dror, T. Zhou, R. Rohs, B. Honig, H. J. Bussemaker, *et al.*, "Cofactor binding evokes latent differences in dna binding specificity between hox proteins," *Cell*, vol. 147, no. 6, pp. 1270–1282, 2011.

[17] Y. Zhao, D. Granas, and G. D. Stormo, "Inferring binding energies from selected binding sites," *PLoS computational biology*, vol. 5, no. 12, p. e1000590, 2009.

[18] A. Jolma, T. Kivioja, J. Toivonen, L. Cheng, G. Wei, M. Enge, M. Taipale, J. M. Vaquerizas, J. Yan, M. J. Sillanpää, *et al.*, "Multiplexed massively parallel selex for characterization of human transcription factor binding specificities," *Genome research*, vol. 20, no. 6, pp. 861–873, 2010.

[19] X. Meng, M. H. Brodsky, and S. A. Wolfe, "A bacterial one-hybrid system for determining the dna-binding specificity of transcription factors," *Nature biotechnology*, vol. 23, no. 8, p. 988, 2005.

[20] X. Meng and S. A. Wolfe, "Identifying dna sequences recognized by a transcription factor using a bacterial one-hybrid system," *Nature protocols*, vol. 1, no. 1, p. 30, 2006.

[21] G. D. Stormo and G. W. Hartzell, "Identifying protein-binding sites from unaligned dna fragments," *Proceedings of the National Academy of Sciences*, vol. 86, no. 4, pp. 1183–1187, 1989.

[22] G. D. Stormo, "Dna binding sites: representation and discovery," *Bioinformatics*, vol. 16, no. 1, pp. 16–23, 2000.

[23] T. D. Schneider and R. M. Stephens, “Sequence logos: a new way to display consensus sequences,” *Nucleic acids research*, vol. 18, no. 20, pp. 6097–6100, 1990.

[24] A. Jolma, J. Yan, T. Whitington, J. Toivonen, K. R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, *et al.*, “Dna-binding specificities of human transcription factors,” *Cell*, vol. 152, no. 1-2, pp. 327–339, 2013.

[25] R. I. Corona and J.-t. Guo, “Statistical analysis of structural determinants for protein–dna-binding specificity,” *Proteins: Structure, Function, and Bioinformatics*, vol. 84, no. 8, pp. 1147–1161, 2016.

[26] N. K. Fox, S. E. Brenner, and J.-M. Chandonia, “Scope: Structural classification of proteinsâextended, integrating scop and astral data and classification of new structures,” *Nucleic acids research*, vol. 42, no. D1, pp. D304–D309, 2014.

[27] N. L. Dawson, T. E. Lewis, S. Das, J. G. Lees, D. Lee, P. Ashford, C. A. Orengo, and I. Sillitoe, “Cath: an expanded resource to predict protein function through structure and sequence,” *Nucleic acids research*, vol. 45, no. D1, pp. D289–D295, 2017.

[28] W. Kabsch and C. Sander, “Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features,” *Biopolymers: Original Research on Biomolecules*, vol. 22, no. 12, pp. 2577–2637, 1983.

[29] R. Kim and J.-t. Guo, “Systematic analysis of short internal indels and their impact on protein folding,” *BMC structural biology*, vol. 10, no. 1, p. 24, 2010.

[30] M. Lin, S. Whitmire, J. Chen, A. Farrel, X. Shi, and J.-t. Guo, “Effects of short indels on protein structure and function in human genomes,” *Scientific reports*, vol. 7, no. 1, p. 9313, 2017.

[31] W. G. Touw, C. Baakman, J. Black, T. A. te Beek, E. Krieger, R. P. Joosten, and G. Vriend, “A series of pdb-related databanks for everyday needs,” *Nucleic acids research*, vol. 43, no. D1, pp. D364–D368, 2014.

[32] 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*, vol. 28, pp. 235–242, 2000.

[33] S. K. Burley, H. M. Berman, C. Bhikadiya, C. Bi, L. Chen, L. Di Costanzo, C. Christie, K. Dalenberg, J. M. Duarte, S. Dutta, *et al.*, “Rcsb protein data bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy,” *Nucleic acids research*, vol. 47, no. D1, pp. D464–D474, 2018.

[34] Y. Pan, C.-J. Tsai, B. Ma, and R. Nussinov, “Mechanisms of transcription factor selectivity,” *Trends in Genetics*, vol. 26, no. 2, pp. 75–83, 2010.

[35] M. Slattery, T. Zhou, L. Yang, A. C. D. Machado, R. Gordân, and R. Rohs, "Absence of a simple code: how transcription factors read the genome," *Trends in biochemical sciences*, vol. 39, no. 9, pp. 381–399, 2014.

[36] N. C. Seeman, J. M. Rosenberg, and A. Rich, "Sequence-specific recognition of double helical nucleic acids by proteins," *Proceedings of the National Academy of Sciences*, vol. 73, no. 3, pp. 804–808, 1976.

[37] V. E. Angarica, A. G. Pérez, A. T. Vasconcelos, J. Collado-Vides, and B. Contreras-Moreira, "Prediction of tf target sites based on atomistic models of protein-dna complexes," *BMC bioinformatics*, vol. 9, no. 1, p. 436, 2008.

[38] C. M. Baker and G. H. Grant, "Role of aromatic amino acids in protein–nucleic acid recognition," *Biopolymers: Original Research on Biomolecules*, vol. 85, no. 5-6, pp. 456–470, 2007.

[39] A. Farrel, J. Murphy, and J.-t. Guo, "Structure-based prediction of transcription factor binding specificity using an integrative energy function," *Bioinformatics*, vol. 32, no. 12, pp. i306–i313, 2016.

[40] R. Wintjens, J. Liévin, M. Rooman, and E. Buisine, "Contribution of cation-π interactions to the stability of protein-dna complexes," *Journal of molecular biology*, vol. 302, no. 2, pp. 393–408, 2000.

[41] K. A. Wilson, J. L. Kellie, and S. D. Wetmore, "Dna–protein π-interactions in nature: abundance, structure, composition and strength of contacts between aromatic amino acids and dna nucleobases or deoxyribose sugar," *Nucleic acids research*, vol. 42, no. 10, pp. 6726–6741, 2014.

[42] 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*, vol. 29, no. 13, pp. 2860–2874, 2001.

[43] Y. Mandel-Gutfreund, O. Schueler, and H. Margalit, "Comprehensive analysis of hydrogen bonds in regulatory protein dna-complexes: in search of common principles," *Journal of molecular biology*, vol. 253, no. 2, pp. 370–382, 1995.

[44] C. O. Pabo and R. T. Sauer, "Transcription factors: structural families and principles of dna recognition," *Annual review of biochemistry*, vol. 61, no. 1, pp. 1053–1095, 1992.

[45] M. Suzuki, "A framework for the dna–protein recognition code of the probe helix in transcription factors: the chemical and stereochemical rules," *Structure*, vol. 2, no. 4, pp. 317–326, 1994.

[46] 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*, vol. 79, pp. 233–269, 2010.

[47] R. N. Azad, D. Zafiropoulos, D. Ober, Y. Jiang, T.-P. Chiu, J. M. Sagendorf, R. Rohs, and T. D. Tullius, “Experimental maps of dna structure at nucleotide resolution distinguish intrinsic from protein-induced dna deformations,” *Nucleic acids research*, vol. 46, no. 5, pp. 2636–2647, 2018.

[48] A. Mathelier, B. Xin, T.-P. Chiu, L. Yang, R. Rohs, and W. W. Wasserman, “Dna shape features improve transcription factor binding site predictions in vivo,” *Cell systems*, vol. 3, no. 3, pp. 278–286, 2016.

[49] Z. Otwinowski, R. Schevitz, R.-G. Zhang, C. Lawson, A. Joachimiak, R. Marmorstein, B. Luisi, and P. Sigler, “Crystal structure of trp represser/operator complex at atomic resolution,” *Nature*, vol. 335, no. 6188, p. 321, 1988.

[50] R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig, “The role of dna shape in protein–dna recognition,” *Nature*, vol. 461, no. 7268, p. 1248, 2009.

[51] Z. Shakked, G. Guzikevich-Guerstein, F. Frolow, D. Rabinovich, A. Joachimiak, and P. Sigler, “Determinants of repressor/operator recognition from the structure of the trp operator binding site,” *Nature*, vol. 368, no. 6470, p. 469, 1994.

[52] A. A. Travers, “Dna conformation and protein binding,” *Annual review of biochemistry*, vol. 58, no. 1, pp. 427–452, 1989.

[53] R. Gordân, N. Shen, I. Dror, T. Zhou, J. Horton, R. Rohs, and M. L. Bulyk, “Genomic regions flanking e-box binding sites influence dna binding specificity of bhlh transcription factors through dna shape,” *Cell reports*, vol. 3, no. 4, pp. 1093–1104, 2013.

[54] S. Rao, T.-P. Chiu, J. F. Kribelbauer, R. S. Mann, H. J. Bussemaker, and R. Rohs, “Systematic prediction of dna shape changes due to cpg methylation explains epigenetic effects on protein–dna binding,” *Epigenetics & chromatin*, vol. 11, no. 1, p. 6, 2018.

[55] D. Badia, A. Camacho, L. Pérez-Lago, C. Escandón, M. Salas, and M. Coll, “The structure of phage ϕ29 transcription regulator p4-dna complex reveals an n-hook motif for dna binding,” *Molecular cell*, vol. 22, no. 1, pp. 73–81, 2006.

[56] M. Fuxreiter, I. Simon, and S. Bondos, “Dynamic protein–dna recognition: beyond what can be seen,” *Trends in biochemical sciences*, vol. 36, no. 8, pp. 415–423, 2011.

[57] 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*, vol. 131, no. 3, pp. 530–543, 2007.

[58] W. Song and J.-T. Guo, "Investigation of arc repressor dna-binding specificity by comparative molecular dynamics simulations," *Journal of Biomolecular Structure and Dynamics*, vol. 33, no. 10, pp. 2083–2093, 2015.

[59] H.-X. Zhou, "Intrinsic disorder: signaling via highly specific but short-lived association," *Trends in biochemical sciences*, vol. 37, no. 2, pp. 43–48, 2012.

[60] M. L. Bulyk, "Computational prediction of transcription-factor binding site locations," *Genome biology*, vol. 5, no. 1, p. 201, 2003.

[61] J. Li, J. M. Sagendorf, T.-P. Chiu, M. Pasi, A. Perez, and R. Rohs, "Expanding the repertoire of dna shape features for genome-scale studies of transcription factor binding," *Nucleic acids research*, vol. 45, no. 22, pp. 12877–12887, 2017.

[62] A. V. Morozov, J. J. Havranek, D. Baker, and E. D. Siggia, "Protein–dna binding specificity predictions with structural models," *Nucleic acids research*, vol. 33, no. 18, pp. 5781–5798, 2005.

[63] T. Zhou, N. Shen, L. Yang, N. Abe, J. Horton, R. S. Mann, H. J. Bussemaker, R. Gordân, and R. Rohs, "Quantitative modeling of transcription factor binding specificities using dna shape," *Proceedings of the National Academy of Sciences*, vol. 112, no. 15, pp. 4654–4659, 2015.

[64] 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*, vol. 320, no. 5, pp. 991–1009, 2002.

[65] A. Pingoud, M. Fuxreiter, V. Pingoud, and W. Wende, "Type ii restriction endonucleases: structure and mechanism," *Cellular and molecular life sciences*, vol. 62, no. 6, p. 685, 2005.

[66] J. G. Cogan, S. Sun, E. S. Stoflet, L. J. Schmidt, M. J. Getz, and A. R. Strauch, "Plasticity of vascular smooth muscle alpha-actin gene transcription. characterization of multiple, single-, and double-strand specific dna-binding proteins in myoblasts and fibroblasts," *Journal of Biological Chemistry*, vol. 270, no. 19, pp. 11310–11321, 1995.

[67] T. Davis-Smyth, R. C. Duncan, T. Zheng, G. Michelotti, and D. Levens, "The far upstream element-binding proteins comprise an ancient family of single-strand dna-binding transactivators," *Journal of Biological Chemistry*, vol. 271, no. 49, pp. 31679–31687, 1996.

[68] S. Haas, A. Steplewski, L. D. Siracusa, S. Amini, and K. Khalili, "Identification of a sequence-specific single-stranded dna binding protein that suppresses transcription of the mouse myelin basic protein gene," *Journal of Biological Chemistry*, vol. 270, no. 21, pp. 12503–12510, 1995.

[69] S. Sun, E. S. Stoflet, J. G. Cogan, A. R. Strauch, and M. J. Getz, "Negative regulation of the vascular smooth muscle alpha-actin gene in fibroblasts and myoblasts: disruption of enhancer function by sequence-specific single-stranded-dna-binding proteins.," *Molecular and cellular biology*, vol. 15, no. 5, pp. 2429–2436, 1995.

[70] S. C. Harrison and A. K. Aggarwal, "Dna recognition by proteins with the helix-turn-helix motif," *Annual review of biochemistry*, vol. 59, no. 1, pp. 933–969, 1990.

[71] S. C. Harrison, "A structural taxonomy of dna-binding domains," *Nature*, vol. 353, no. 6346, p. 715, 1991.

[72] G. H. Jacobs, "Determination of the base recognition positions of zinc fingers from sequence analysis.," *The EMBO journal*, vol. 11, no. 12, pp. 4507–4517, 1992.

[73] K. Kamada, T. Horiuchi, K. Ohsumi, N. Shimamoto, and K. Morikawa, "Structure of a replication-terminator protein complexed with dna," *Nature*, vol. 383, no. 6601, p. 598, 1996.

[74] P. A. Rice, S.-w. Yang, K. Mizuuchi, and H. A. Nash, "Crystal structure of an ihf-dna complex: a protein-induced dna u-turn," *Cell*, vol. 87, no. 7, pp. 1295–1306, 1996.

[75] B. M. Alberts and L. Frey, "T4 bacteriophage gene 32: a structural protein in the replication and recombination of dna," *Nature*, vol. 227, no. 5265, p. 1313, 1970.

[76] Y. Shamoo, A. M. Friedman, M. R. Parsons, W. H. Konigsberg, and T. A. Steitz, "Crystal structure of a replication fork single-stranded dna binding protein (t4 gp32) complexed to dna," *Nature*, vol. 376, no. 6538, p. 362, 1995.

[77] K. Pant, R. L. Karpel, I. Rouzina, and M. C. Williams, "Salt dependent binding of t4 gene 32 protein to single and double-stranded dna: single molecule force spectroscopy measurements," *Journal of molecular biology*, vol. 349, no. 2, pp. 317–330, 2005.

[78] D. Jose, S. E. Weitzel, W. A. Baase, and P. H. von Hippel, "Mapping the interactions of the single-stranded dna binding protein of bacteriophage t4 (gp32) with dna lattices at single nucleotide resolution: gp32 monomer binding," *Nucleic acids research*, vol. 43, no. 19, pp. 9276–9290, 2015.

[79] B. Camel, A. Dang, K. Meze, D. Jose, and P. H. von Hippel, "Mapping interactions of single-stranded (ss) dna with the ss-dna binding protein (gp32) of the t4 dna replication complex at specific nucleotide residue positions," *Biophysical Journal*, vol. 114, no. 3, p. 442a, 2018.

[80] N. Sigal, H. Delius, T. Kornberg, M. L. Gefter, and B. Alberts, "A dna-unwinding protein isolated from escherichia coli: its interaction with dna and with dna polymerases," *Proceedings of the National Academy of Sciences*, vol. 69, no. 12, pp. 3537–3541, 1972.

[81] L. B. Overman, W. Bujalowski, and T. M. Lohman, "Equilibrium binding of escherichia coli single-strand binding protein to single-stranded nucleic acids in the (ssb) 65 binding mode. cation and anion effects and polynucleotide specificity," *Biochemistry*, vol. 27, no. 1, pp. 456–471, 1988.

[82] S. Raghunathan, C. S. Ricard, T. M. Lohman, and G. Waksman, "Crystal structure of the homo-tetrameric dna binding domain of escherichia coli single-stranded dna-binding protein determined by multiwavelength x-ray diffraction on the selenomethionyl protein at 2.9-å resolution," *Proceedings of the National Academy of Sciences*, vol. 94, no. 13, pp. 6652–6657, 1997.

[83] L. Hamon, D. Pastre, P. Dupaigne, C. L. Breton, E. L. Cam, and O. Pietrement, "High-resolution afm imaging of single-stranded dna-binding (ssb) proteinâdna complexes," *Nucleic acids research*, vol. 35, no. 8, p. e58, 2007.

[84] S. Suksombat, R. Khafizov, A. G. Kozlov, T. M. Lohman, and Y. R. Chemla, "Structural dynamics of e. coli single-stranded dna binding protein reveal dna wrapping and unwrapping pathways," *Elife*, vol. 4, p. e08193, 2015.

[85] E. Antony and T. M. Lohman, "Dynamics of e. coli single stranded dna binding (ssb) protein-dna complexes," in *Seminars in cell & developmental biology*, vol. 86, pp. 102–111, Elsevier, 2019.

[86] L. M. Spenkelink, J. S. Lewis, S. Jergic, Z.-Q. Xu, A. Robinson, N. E. Dixon, and A. M. van Oijen, "Recycling of single-stranded dna-binding protein by the bacterial replisome," *Nucleic acids research*, vol. 47, no. 8, pp. 4111–4123, 2019.

[87] M. S. Wold and T. Kelly, "Purification and characterization of replication protein a, a cellular protein required for in vitro replication of simian virus 40 dna," *Proceedings of the National Academy of Sciences*, vol. 85, no. 8, pp. 2523–2527, 1988.

[88] M. P. Fairman and B. Stillman, "Cellular factors required for multiple stages of sv40 dna replication in vitro.," *The EMBO Journal*, vol. 7, no. 4, pp. 1211–1218, 1988.

[89] A. Bochkarev, R. A. Pfuetzner, A. M. Edwards, and L. Frappier, "Structure of the single-stranded-dna-binding domain of replication protein a bound to dna," *Nature*, vol. 385, no. 6612, p. 176, 1997.

[90] A. Bochkarev and E. Bochkareva, "From rpa to brca2: lessons from single-stranded dna binding by the ob-fold," *Current opinion in structural biology*, vol. 14, no. 1, pp. 36–42, 2004.

[91] R. Chen and M. S. Wold, “Replication protein a: single-stranded dna’s first responder: dynamic dna-interactions allow replication protein a to direct single-strand dna intermediates into different pathways for synthesis or repair,” *Bioessays*, vol. 36, no. 12, pp. 1156–1161, 2014.

[92] L. A. Yates, R. J. Aramayo, N. Pokhrel, C. C. Caldwell, J. A. Kaplan, R. L. Perera, M. Spies, E. Antony, and X. Zhang, “A structural and dynamic model for the assembly of replication protein a on single-stranded dna,” *Nature communications*, vol. 9, no. 1, p. 5447, 2018.

[93] T. H. Dickey, S. E. Altschuler, and D. S. Wuttke, “Single-stranded dna-binding proteins: multiple domains for multiple functions,” *Structure*, vol. 21, no. 7, pp. 1074–1084, 2013.

[94] H. Shi, Y. Zhang, G. Zhang, J. Guo, X. Zhang, H. Song, J. Lv, J. Gao, Y. Wang, L. Chen, *et al.*, “Systematic functional comparative analysis of four single-stranded dna-binding proteins and their affection on viral rna metabolism,” *PLoS One*, vol. 8, no. 1, p. e55076, 2013.

[95] T. M. Lohman and M. E. Ferrari, “Escherichia coli single-stranded dna-binding protein: multiple dna-binding modes and cooperativities,” *Annual review of biochemistry*, vol. 63, no. 1, pp. 527–570, 1994.

[96] J. W. Chase and K. R. Williams, “Single-stranded dna binding proteins required for dna replication,” *Annual review of biochemistry*, vol. 55, no. 1, pp. 103–136, 1986.

[97] J. Kur, M. Olszewski, A. Dlugolecka, and P. Filipkowski, “Single-stranded dna-binding proteins (ssbs)-sources and applications in molecular biology,” *ACTA BIOCHIMICA POLONICA-ENGLISH EDITION-*, vol. 52, no. 3, p. 569, 2005.

[98] C. Perales, F. Cava, W. J. Meijer, and J. Berenguer, “Enhancement of dna, cdna synthesis and fidelity at high temperatures by a dimeric single-stranded dna-binding protein,” *Nucleic acids research*, vol. 31, no. 22, pp. 6473–6480, 2003.

[99] S. Raghunathan, A. G. Kozlov, T. M. Lohman, and G. Waksman, “Structure of the dna binding domain of e. coli ssb bound to ssdna,” *Nature Structural & Molecular Biology*, vol. 7, no. 8, p. 648, 2000.

[100] W. Bujalowski, L. B. Overman, and T. Lohman, “Binding mode transitions of escherichia coli single strand binding protein-single-stranded dna complexes. cation, anion, ph, and binding density effects.,” *Journal of Biological Chemistry*, vol. 263, no. 10, pp. 4629–4640, 1988.

[101] T. M. Lohman, L. B. Overman, and S. Datta, “Salt-dependent changes in the dna binding co-operativity of escherichia coli single strand binding protein,” *Journal of molecular biology*, vol. 187, no. 4, pp. 603–615, 1986.

[102] W. Bujalowski and T. M. Lohman, “Escherichia coli single-strand binding protein forms multiple, distinct complexes with single-stranded dna,” *Biochemistry*, vol. 25, no. 24, pp. 7799–7802, 1986.

[103] M. P. Horvath, V. L. Schweiker, J. M. Bevilacqua, J. A. Ruggles, and S. C. Schultz, “Crystal structure of the oxytricha nova telomere end binding protein complexed with single strand dna,” *Cell*, vol. 95, no. 7, pp. 963–974, 1998.

[104] M. P. Horvath, “Structural anatomy of telomere ob proteins,” *Critical reviews in biochemistry and molecular biology*, vol. 46, no. 5, pp. 409–435, 2011.

[105] D. E. Gottschling and B. Stoddard, “Telomeres: structure of a chromosome’s aglet,” *Current biology*, vol. 9, no. 5, pp. R164–R167, 1999.

[106] A. G. Murzin, “Ob (oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences.,” *The EMBO journal*, vol. 12, no. 3, pp. 861–867, 1993.

[107] D. L. Theobald, R. M. Mitton-Fry, and D. S. Wuttke, “Nucleic acid recognition by ob-fold proteins,” *Annual review of biophysics and biomolecular structure*, vol. 32, no. 1, pp. 115–133, 2003.

[108] S. E. Altschuler, T. H. Dickey, and D. S. Wuttke, “Schizosaccharomyces pombe protection of telomeres 1 utilizes alternate binding modes to accommodate different telomeric sequences,” *Biochemistry*, vol. 50, no. 35, pp. 7503–7513, 2011.

[109] M. Lei, E. R. Podell, and T. R. Cech, “Structure of human pot1 bound to telomeric single-stranded dna provides a model for chromosome end-protection,” *Nature structural & molecular biology*, vol. 11, no. 12, p. 1223, 2004.

[110] A. Pal and Y. Levy, “Structure, stability and specificity of the binding of ssdna and ssrna with proteins,” *PLoS computational biology*, vol. 15, no. 4, p. e1006768, 2019.

[111] J. M. Sagendorf, N. Markarian, H. M. Berman, and R. Rohs, “Dnaprodb: an expanded database and web-based tool for structural analysis of dna–protein complexes,” *Nucleic acids research*, vol. 48, no. D1, pp. D277–D287, 2020.

[112] H. M. Berman, W. K. Olson, D. L. Beveridge, J. Westbrook, A. Gelbin, T. Demeny, S.-H. Hsieh, A. Srinivasan, and B. Schneider, “The nucleic acid database. a comprehensive relational database of three-dimensional structures of nucleic acids.,” *Biophysical journal*, vol. 63, no. 3, p. 751, 1992.

[113] B. Coimbatore Narayanan, J. Westbrook, S. Ghosh, A. I. Petrov, B. Sweeney, C. L. Zirbel, N. B. Leontis, and H. M. Berman, “The nucleic acid database: new features and capabilities,” *Nucleic acids research*, vol. 42, no. D1, pp. D114–D122, 2014.

[114] R. Valverde, L. Edwards, and L. Regan, "Structure and function of kh domains," *The FEBS journal*, vol. 275, no. 11, pp. 2712–2726, 2008.

[115] A. Cléry, M. Blatter, and F. H. Allain, "Rna recognition motifs: boring? not quite," *Current opinion in structural biology*, vol. 18, no. 3, pp. 290–298, 2008.

[116] D. Desveaux, A. Maréchal, and N. Brisson, "Whirly transcription factors: defense gene regulation and beyond," *Trends in plant science*, vol. 10, no. 2, pp. 95–102, 2005.

[117] M. Zeeb, K. E. Max, U. Weininger, C. Löw, H. Sticht, and J. Balbach, "Recognition of t-rich single-stranded dna by the cold shock protein bs-cspb in solution," *Nucleic acids research*, vol. 34, no. 16, pp. 4561–4571, 2006.

[118] I. Letunic, T. Doerks, and P. Bork, "Smart 7: recent updates to the protein domain annotation resource," *Nucleic acids research*, vol. 40, no. D1, pp. D302–D305, 2011.

[119] U. Consortium, "Reorganizing the protein space at the universal protein resource (uniprot)," *Nucleic acids research*, vol. 40, no. D1, pp. D71–D75, 2011.

[120] L. Cappadocia, A. Maréchal, J.-S. Parent, É. Lepage, J. Sygusch, and N. Brisson, "Crystal structures of dna-whirly complexes and their role in arabidopsis organelle genome repair," *The Plant Cell*, vol. 22, no. 6, pp. 1849–1867, 2010.

[121] L. Cappadocia, J.-S. Parent, E. Zampini, E. Lepage, J. Sygusch, and N. Brisson, "A conserved lysine residue of plant whirly proteins is necessary for higher order protein assembly and protection against dna damage," *Nucleic acids research*, vol. 40, no. 1, pp. 258–269, 2011.

[122] G. Abrusán and J. A. Marsh, "Alpha helices are more robust to mutations than beta strands," *PLoS computational biology*, vol. 12, no. 12, p. e1005242, 2016.

[123] M. Lin and J.-t. Guo, "New insights into protein–dna binding specificity from hydrogen bond based comparative study," *Nucleic acids research*, 2019.

[124] D. S. Latchman, "Transcription-factor mutations and disease," *New England Journal of Medicine*, vol. 334, no. 1, pp. 28–33, 1996.

[125] 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*, vol. 281, no. 5373, pp. 108–111, 1998.

[126] D. Golovenko, B. Bräuning, P. Vyas, T. E. Haran, H. Rozenberg, and Z. Shakked, "New insights into the role of dna shape on its recognition by p53 proteins," *Structure*, vol. 26, no. 9, pp. 1237–1250, 2018.

[127] M. Kitayner, H. Rozenberg, R. Rohs, O. Suad, D. Rabinovich, B. Honig, and Z. Shakked, “Diversity in dna recognition by p53 revealed by crystal structures with hoogsteen base pairs,” *Nature structural & molecular biology*, vol. 17, no. 4, p. 423, 2010.

[128] R. Vainer, S. Cohen, A. Shahar, R. Zarivach, and E. Arbely, “Structural basis for p53 lys120-acetylation-dependent dna-binding mode,” *Journal of molecular biology*, vol. 428, no. 15, pp. 3013–3025, 2016.

[129] R. I. Corona, S. Sudarshan, S. Aluru, and J.-t. Guo, “An svm-based method for assessment of transcription factor-dna complex models,” *BMC bioinformatics*, vol. 19, no. 20, p. 506, 2018.

[130] R. Kim and J.-t. Guo, “Pda: an automatic and comprehensive analysis program for protein-dna complex structures,” *BMC genomics*, vol. 10, no. 1, p. S13, 2009.

[131] W. Li and A. Godzik, “Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences,” *Bioinformatics*, vol. 22, no. 13, pp. 1658–1659, 2006.

[132] I. K. McDonald and J. M. Thornton, “Satisfying hydrogen bonding potential in proteins,” *Journal of molecular biology*, vol. 238, no. 5, pp. 777–793, 1994.

[133] D. J. Jacobs, A. J. Rader, L. A. Kuhn, and M. F. Thorpe, “Protein flexibility predictions using graph theory,” *Proteins: Structure, Function, and Bioinformatics*, vol. 44, no. 2, pp. 150–165, 2001.

[134] A. Pingoud and A. Jeltsch, “Structure and function of type ii restriction endonucleases,” *Nucleic acids research*, vol. 29, no. 18, pp. 3705–3727, 2001.

[135] A. R. Sonawane, J. Platig, M. Fagny, C.-Y. Chen, J. N. Paulson, C. M. Lopes-Ramos, D. L. DeMeo, J. Quackenbush, K. Glass, and M. L. Kuijjer, “Understanding tissue-specific gene regulation,” *Cell reports*, vol. 21, no. 4, pp. 1077–1088, 2017.

[136] G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, *et al.*, “Diversity and complexity in dna recognition by transcription factors,” *Science*, vol. 324, no. 5935, pp. 1720–1723, 2009.

[137] K. Nadassy, S. J. Wodak, and J. Janin, “Structural features of protein- nucleic acid recognition sites,” *Biochemistry*, vol. 38, no. 7, pp. 1999–2017, 1999.

[138] S. Nikolajewa, A. Beyer, M. Friedel, J. Hollunder, and T. Wilhelm, “Common patterns in type ii restriction enzyme binding sites,” *Nucleic acids research*, vol. 33, no. 8, pp. 2726–2733, 2005.

[139] C. Bailly, N. E. Møllegaard, P. E. Nielsen, and M. Waring, “The influence of the 2-amino group of guanine on dna conformation. uranyl and dnase i probing of inosine/diaminopurine substituted dna.,” *The EMBO journal*, vol. 14, no. 9, pp. 2121–2131, 1995.

[140] C. Baily and M. J. Waring, “Transferring the purine 2-amino group from guanines to adenines in dna changes the sequence-specific binding of antibiotics,” *Nucleic acids research*, vol. 23, no. 6, pp. 885–892, 1995.

[141] C. Bailly and M. J. Waring, “The purine 2-amino group as a critical recognition element for specific dna cleavage by bleomycin and calicheamicin,” *Journal of the American Chemical Society*, vol. 117, no. 28, pp. 7311–7316, 1995.

[142] C. Bailly, M. J. Waring, and A. A. Travers, “Effects of base substitutions on the binding of a dna-bending protein,” 1995.

[143] S. Doublie, S. Tabor, A. M. Long, C. C. Richardson, and T. Ellenberger, “Crystal structure of a bacteriophage t7 dna replication complex at 2.2 å resolution,” *Nature*, vol. 391, no. 6664, p. 251, 1998.

[144] J. R. Kiefer, C. Mao, J. C. Braman, and L. S. Beese, “Visualizing dna replication in a catalytically active bacillus dna polymerase crystal,” *Nature*, vol. 391, no. 6664, p. 304, 1998.

[145] J. C. Morales and E. T. Kool, “Minor groove interactions between polymerase and dna: More essential to replication than watson- crick hydrogen bonds?,” *Journal of the American Chemical Society*, vol. 121, no. 10, pp. 2323–2324, 1999.

[146] H. Pelletier, M. R. Sawaya, A. Kumar, S. H. Wilson, and J. Kraut, “Structures of ternary complexes of rat dna polymerase beta, a dna template-primer, and ddctp,” *Science*, vol. 264, no. 5167, pp. 1891–1903, 1994.

[147] Y. Levy, J. N. Onuchic, and P. G. Wolynes, “Fly-casting in protein- dna binding: Frustration between protein folding and electrostatics facilitates target recognition,” *Journal of the American Chemical Society*, vol. 129, no. 4, pp. 738–739, 2007.

[148] B. A. Shoemaker, J. J. Portman, and P. G. Wolynes, “Speeding molecular recognition by using the folding funnel: the fly-casting mechanism,” *Proceedings of the National Academy of Sciences*, vol. 97, no. 16, pp. 8868–8873, 2000.

[149] A. Granzhan, N. Kotera, and M.-P. Teulade-Fichou, “Finding needles in a basestack: recognition of mismatched base pairs in dna by small molecules,” *Chemical Society Reviews*, vol. 43, no. 10, pp. 3630–3665, 2014.

[150] E. N. Nikolova, H. Zhou, F. L. Gottardo, H. S. Alvey, I. J. Kimsey, and H. M. Al-Hashimi, “A historical account of hoogsteen base-pairs in duplex dna,” *Biopolymers*, vol. 99, no. 12, pp. 955–968, 2013.

[151] G. Rossetti, P. D. Dans, I. Gomez-Pinto, I. Ivani, C. Gonzalez, and M. Orozco, "The structural impact of dna mismatches," *Nucleic acids research*, vol. 43, no. 8, pp. 4309–4321, 2015.

[152] C. Yang, E. Kim, and Y. Pak, "Free energy landscape and transition pathways from watson–crick to hoogsteen base pairing in free duplex dna," *Nucleic acids research*, vol. 43, no. 16, pp. 7769–7778, 2015.

[153] M. A. Reijns, H. Kemp, J. Ding, S. M. de Procé, A. P. Jackson, and M. S. Taylor, "Lagging-strand replication shapes the mutational landscape of the genome," *Nature*, vol. 518, no. 7540, p. 502, 2015.

[154] R. Sabarinathan, L. Mularoni, J. Deu-Pons, A. Gonzalez-Perez, and N. Lopez-Bigas, "Nucleotide excision repair is impaired by binding of transcription factors to dna," *Nature*, vol. 532, no. 7598, p. 264, 2016.

[155] Y. Wu, J. Lu, and T. Kang, "Human single-stranded dna binding proteins: guardians of genome stability," *Acta biochimica et biophysica Sinica*, vol. 48, no. 7, pp. 671–677, 2016.

[156] R. Buisson, A. Langenbucher, D. Bowen, E. E. Kwan, C. H. Benes, L. Zou, and M. S. Lawrence, "Passenger hotspot mutations in cancer driven by apobec3a and mesoscale genomic features," *Science*, vol. 364, no. 6447, p. eaaw2872, 2019.

[157] N. J. Haradhvala, P. Polak, P. Stojanov, K. R. Covington, E. Shinbrot, J. M. Hess, E. Rheinbay, J. Kim, Y. E. Maruvka, L. Z. Braunstein, *et al.*, "Mutational strand asymmetries in cancer genomes reveal mechanisms of dna damage and repair," *Cell*, vol. 164, no. 3, pp. 538–549, 2016.

[158] R. L. Karpel, "T4 bacteriophage gene 32 protein," *The Biology of Non-Specific DNA-Protein Interactions. Boca Raton: CRC Press. p*, pp. 103–130, 1990.

[159] J. R. Casas-Finet, K. R. Fischer, and R. L. Karpel, "Structural basis for the nucleic acid binding cooperativity of bacteriophage t4 gene 32 protein: the (lys/arg) 3 (ser/thr) 2 (last) motif.," *Proceedings of the National Academy of Sciences*, vol. 89, no. 3, pp. 1050–1054, 1992.

[160] N. Lonberg, S. C. Kowalczykowski, L. S. Paul, and P. H. von Hippel, "Interactions of bacteriophage t4-coded gene 32 protein with nucleic acids: Iii. binding properties of two specific proteolytic digestion products of the protein (g32pâ i and g32pâ iii)," *Journal of molecular biology*, vol. 145, no. 1, pp. 123–138, 1981.

[161] E. Fanning, V. Klimovich, and A. R. Nager, "A dynamic model for replication protein a (rpa) function in dna processing pathways," *Nucleic acids research*, vol. 34, no. 15, pp. 4126–4137, 2006.

[162] C. Iftode, Y. Daniely, and J. A. Borowiec, "Replication protein a (rpa): the eukaryotic ssb," *Critical reviews in biochemistry and molecular biology*, vol. 34, no. 3, pp. 141–180, 1999.

[163] E. Bochkareva, S. Korolev, S. P. Lees-Miller, and A. Bochkarev, “Structure of the rpa trimerization core and its role in the multistep dna-binding mechanism of rpa,” *The EMBO journal*, vol. 21, no. 7, pp. 1855–1863, 2002.

[164] N. W. Ashton, E. Bolderson, L. Cubeddu, K. J. OâByrne, and D. J. Richard, “Human single-stranded dna binding proteins are essential for maintaining genomic stability,” *BMC molecular biology*, vol. 14, no. 1, p. 9, 2013.

[165] T. H. Dickey, M. A. McKercher, and D. S. Wuttke, “Nonspecific recognition is achieved in pot1pc through the use of multiple binding modes,” *Structure*, vol. 21, no. 1, pp. 121–132, 2013.

[166] D. L. Theobald and S. C. Schultz, “Nucleotide shuffling and ssdna recognition in oxytricha nova telomere end-binding protein complexes,” *The EMBO journal*, vol. 22, no. 16, pp. 4314–4324, 2003.

[167] T. E. Lewis, I. Sillitoe, N. Dawson, S. D. Lam, T. Clarke, D. Lee, C. Orengo, and J. Lees, “Gene3d: extensive prediction of globular domains in proteins,” *Nucleic acids research*, vol. 46, no. D1, pp. D435–D439, 2018.

[168] G. Wang and R. L. Dunbrack Jr, “Pisces: a protein sequence culling server,” *Bioinformatics*, vol. 19, no. 12, pp. 1589–1591, 2003.

[169] U. Consortium, “Uniprot: a worldwide hub of protein knowledge,” *Nucleic acids research*, vol. 47, no. D1, pp. D506–D515, 2019.

[170] M. Johnson, I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T. L. Madden, “Ncbi blast: a better web interface,” *Nucleic acids research*, vol. 36, no. suppl_2, pp. W5–W9, 2008.

[171] J. P. Rodrigues, J. M. Teixeira, M. Trellet, and A. M. Bonvin, “pdb-tools: a swiss army knife for molecular structures,” *F1000Research*, vol. 7, 2018.

[172] S. Mitternacht, “Freesasa: An open source c library for solvent accessible surface area calculations,” *F1000Research*, vol. 5, 2016.

[173] Y. Zhang and J. Skolnick, “Tm-align: a protein structure alignment algorithm based on the tm-score,” *Nucleic acids research*, vol. 33, no. 7, pp. 2302–2309, 2005.

[174] M. L. Gleghorn, E. K. Davydova, L. B. Rothman-Denes, and K. S. Murakami, “Structural basis for dna-hairpin promoter recognition by the bacteriophage n4 virion rna polymerase,” *Molecular cell*, vol. 32, no. 5, pp. 707–717, 2008.

[175] G. Mishra and Y. Levy, “Molecular determinants of the interactions between proteins and ssdna,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 16, pp. 5033–5038, 2015.

[176] C. Maffeo and A. Aksimentiev, "Molecular mechanism of dna association with single-stranded dna binding protein," *Nucleic acids research*, vol. 45, no. 21, pp. 12125–12139, 2017.

[177] K. Shi, M. A. Carpenter, S. Banerjee, N. M. Shaban, K. Kurahashi, D. J. Salamango, J. L. McCann, G. J. Starrett, J. V. Duffy, Ö. Demir, *et al.*, "Structural basis for targeted dna cytosine deamination and mutagenesis by apobec3a and apobec3b," *Nature structural & molecular biology*, vol. 24, no. 2, p. 131, 2017.

[178] D. L. Faden, S. Thomas, P. G. Cantalupo, N. Agrawal, J. Myers, and J. DeRisi, "Multi-modality analysis supports apobec as a major source of mutations in head and neck squamous cell carcinoma," *Oral oncology*, vol. 74, pp. 8–14, 2017.

[179] J. M. Flynn, I. Levchenko, R. T. Sauer, and T. A. Baker, "Modulating substrate choice: the sspb adaptor delivers a regulator of the extracytoplasmic-stress response to the aaa+ protease clpxp for degradation," *Genes & development*, vol. 18, no. 18, pp. 2292–2301, 2004.

APPENDIX A: SUPPLEMENTARY TABLES

Table S1: Domain-based non-redundant DNA-binding domains in HS, MS and NS groups. (*: DNA-binding domain was updated by excluding dimerization domain; #: 6on0 superseded 3qws on 2019-05-15; **Bold**: New HS DNA-binding domains.)

Dataset	Domain ID	Domain definition	Domain ID	Domain definition
HS	1az0B00	1az0:B	d1yfib_	1yfi:B
	1bhmA00	1bhm:A	2e52D01*	2e52:D (3-226)
	1d2iB00	1d2i:B	h3m7kA0	3m7k:A
	1dc1A01	1dc1:A (5-38,127-323)	h3oqgA0	3oqg:A
	1dc1A02	1dc1:A (39-126)	m2fl3A0	2fl3:A
	1eriA00	1eri:A	m2oaaA0	2oaa:A
	1iawA01	1iaw:A (10-176)	m3c25A0	3c25:A
	1iawA02	1iaw:A (177-309)	m3fc3B1	3fc3:B (2-107)
	1kc6B00	1kc6:B	m3goxB2	3gox:B (108-189)
	1pviA00	1pvi:A	m3imbD0	3imb:D
	1vrrA00	1vrr:A	m3ndhA0	3ndh:A
	1wteA01	1wte:A (1-87, 212-272)	m4rdmB0	4rdm:B
	1wteA02	1wte:A (88-211)	**m2vlaA0**	**2vla:A**
	3dvoD00	3dvo:D	**m4zsfA1**	**4zsf:A(70-272)**
	3hqfA00	3hqf:A	**m5dwaA0**	**5dwa:A**
	4abtA00	4abt:A	**m6ekoA0**	**6eko:A**
MS	1b3tA00	1b3t:A	1nkpD00	1nkp:D
	1bdtD00	1bdt:D	1owrP01	1owr:P (397-569)
	1bl0A01	1bl0:A (9-64)	1pnrA01	1pnr:A (3-59)
	1bl0A02	1bl0:A (65-124)	1qn3B01	1qn3:B (19-29, 116-197)
	1cf7A00	1cf7:A	1qn6A02	1qn6:A (30-115)
	1cmaA00	1cma:A	1qpiA01	1qpi:A (4-66)
	1ea4G00	1ea4:G	1r8dA00	1r8d:A
	1exjA02	1exj:A (3-75)	1rioH00	1rio:H

Table S1: Domain-based non-redundant DNA-binding domains in HS, MS and NS groups (Continued). (*: DNA-binding domain was updated by excluding dimerization domain; #: 6on0 superseded 3qws on 2019-05-15; **Bold**: New HS DNA-binding domains.)

Dataset	Domain ID	Domain definition	Domain ID	Domain definition
MS	1fzpB00	1fzp:B	1saxA01	1sax:A (9-72)
	1gd2E00	1gd2:E	1sknP00	1skn:P
	1gxpE00	1gxp:E	1t2kB01	1t2k:B (7-110)
	1h6fA00	1h6f:A	1xpxA00	1xpx:A
	1hjbB00	1hjb:B	1zreA02	1zre:A (138-207)
	1hjbC00	1hjb:C	1zs4A00	1zs4:A
	1ic8A01	1ic8:A (87-180)	2ac0C00	2ac0:C
	1ic8A02	1ic8:A (203-276)	2bopA00	2bop:A
	1jfiA00	1jfi:A	2e1cA01	2e1c:A (24-76)
	1jfiB00	1jfi:B	2h27A00	2h27:A
	1k78A01	1k78:A (19-84)	2h7hA00	2h7h:A
	1k78B00	1k78:B	2i9tB02	2i9t:B (546-650)
	1kb2A00	1kb2:A	2p5lC00	2p5l:C
	1le5F01	1le5:F (38-241)	2r5yB00	2r5y:B
	1lmb300	1lmb:3	2wt7A00	2wt7:A
	1lq1B00	1lq1:B	2yvhD00	2yvh:D
	1mdmA02	1mdm:A (85-139)	2zhgA00	2zhg:A
	1mhdA00	1mhd:A	3a01A00	3a01:A
	3coaC00	3coa:C	h3mlpE0	3mlp:E
	3dfxB00	3dfx:B	h3vebA0	3veb:A
	3dnvB00	3dnv:B	h3w3cA0	3w3c:A
	3g97A00	3g97:A	h3zplF0	3zpl:F
	3hddB00	3hdd:B	h4gclD0	4gcl:D
	3iagC01	3iag:C (53-200, 359-380)	h4h10A0	4h10:A
	3iagC02	3iag:C (201-358)	h4hf1A0	4hf1:A
	3iktA01	3ikt:A (0-73)	h4ihtC0	4iht:C
	3jtgA01	3jtg:A (273-357)	h4ix7A0	4ix7:A

Table S1: Domain-based non-redundant DNA-binding domains in HS, MS and NS groups (Continued). (*: DNA-binding domain was updated by excluding dimerization domain; #: 6on0 superseded 3qws on 2019-05-15; **Bold**: New HS DNA-binding domains.)

Dataset	Domain ID	Domain definition	Domain ID	Domain definition
MS	3jxdR00	3jxd:R	h4jl3A0	4jl3:A
	3o9xA02	3o9x:A (59-131)	m3fdqA0	3fdq:A
	3p57B01	3p57:B (13-91)	m3h0dB1	3h0d:B (3-75)
	3pvvB00	3pvv:B	m3n7qA0	3n7q:A
	6on0A00#	6on0:A	m3u3wA1	3u3w:A (3-58)
	3s8qA00	3s8q:A	m3w6vA0	3w6v:A
	3u2bC00	3u2b:C	m3zqlA0	3zql:A
	3zkcB00	3zkc:B	m4g92A0	4g92:A
	4fthA00	4fth:A	m4g92C0	4g92:C
	4g92B00	4g92:B	m4jcyB0	4jcy:B
	6croA00	6cro:A	m4knyA2	4kny:A (124-225)
	d1odha_	1odh:A	m4l62P1	4l62:P (7-49)
	d2iszd1	2isz:D (1-64)	m4ldxB2	4ldx:B (121-229)
	d2xsdc1	2xsd:C (247-319)	m4llnA0	4lln:A
	d2xsdc2	2xsd:C (343-397)	m4lmgD0	4lmg:D
	d3coqa1	3coq:A (8-48)	m4mteB1	4mte:B (3-72)
	d3e6cc1	3e6c:C (148-233)	m4nnuA1	4nnu:A (44-122)
	h2er8C0	2er8:C	m4nnuA3	4nnu:A (153-236)
	h2vy1A0	2vy1:A	m4on0B0	4on0:B
	h3a5tA0	3a5t:A	m4qtkA0	4qtk:A
	h3gnaA0	3gna:A	m4u0yB0	4u0y:B
	h3igmA0	3igm:A	m4ux5A0	4ux5:A
	h3lsrA01*	3lsr:A(4-53)		
NS	1cezA01	1cez:A (8-325)	1ya6B01	1ya6:B (998-1176, 1387-1400)
	1f66C00	1f66:C	2bzfA00	2bzf:A
	1jeyA02	1jey:A (251-278, 342-439)	2dnjA00	2dnj:A
	1jeyA03	1jey:A (279-341)	2pi4A05	2pi4:A (554-784)

Table S1: Domain-based non-redundant DNA-binding domains in HS, MS and NS groups (Continued). (*: DNA-binding domain was updated by excluding dimerization domain; #: 6on0 superseded 3qws on 2019-05-15; **Bold**: New HS DNA-binding domains.)

Dataset	Domain ID	Domain definition	Domain ID	Domain definition
NS	1jeyB02	1jey:B (243-443)	2voaA00	2voa:A
	1rztA03	1rzt:A (386-508)	2wtfA04	2wtf:A (393-509)
	1rztI04	1rzt:I (509-575)	3aafA00	3aaf:A
	1skrA03	1skr:A (415-477, 590-704)	3av2A00	3av2:A
	1sxqA02	1sxq:A (167-332)	3cwsC02	3cws:C (113-230)
	1x9wA02	1x9w:A (233-414)	3gv5B04	3gv5:B (299-414)
	1xslA02	1xsl:A (332-385)	3l4jA01	3l4j:A (429-561, 609-691)
	3l4jA03	3l4j:A (692-860, 974-988)	h4o0iA	4o0i:A (491-605)
	3l4jA04	3l4j:A (872-973)	h4o5eA3	4o5e:A (149-335)
	3n4mB00	3n4m:B	h4oinD0	4oin:D
	3uiqA02	3uiq:A (109-339)	m2o8bA3	2o8b:A (321-855)
	3uiqA06	3uiq:A (775-866)	m2o8bB1	2o8b:B (362-518)
	4eyhB01	4eyh:B (26-36, 99-221)	m2o8bB3	2o8b:B (728-1335)
	d3jxya_	3jxy:A	m3f2bA0	3f2b:A
	d4klua1	4klu:A (11-91)	m3l2pA1	3l2p:A (168-336)
	d9icka3	9ick:A (92-148)	m4c2uA4	4c2u:A (384-561)
	h1s9fA4	1s9f:A (244-341)	m4dl4A4	4dl4:A (313-432)
	h2wwyA0	2wwy:A	m4ir1F1	4ir1:F (0-10, 74-165)
	h3kxtA0	3kxt:A	m4ir1F4	4ir1:F (236-341)
	h3raxB3	3rax:B (1167-1233)	m4o3mA3	4o3m:A (1072-1194)
	h4eluA2	4elu:A (423-832)	m4plbB1	4plb:B (417-1033)
	h4g0vB0	4g0v:B	m4plbB2	4plb:B (1034-1376, 1461-1491)

Table S2: Chain-based non-redundant protein-dsDNA complexes in HS, MS and NS groups.

Dataset	PDBID (Protein-chain_DNA-chains)
HS	1AZ0(B_CD), 1BHM(A_CD), 1D2I(B_CD), 1DC1(A_CW), 1ERI(A_BC), 1IAW(A_EF), 1IAW(A_CD), 1KC6(B_EF), 1PVI(A_CD), 1VRR(A_CD), 1WTE(A_XY), 3DVO(D_GH), 3HQF(A_BC), 4ABT(A_EH), 1YFI(B_EF), 2E52(D_FH), 3M7K(A_BC), 3OQG(A_CD), 2FL3(A_CD), 2OAA(A_CD), 3C25(A_CD), 3FC3(B_CD), 3IMB(D_KL), 3NDH(A_CD), 4RDM(B_EF), 2VLA(A_LM), 4ZSF(A_BD), 5DWA(A_CD), 6EKO(A_EF)
MS	1B3T(A_CD), 1BDT(D_EF), 1BL0(A_BC), 1CF7(A_CD), 1CMA(A_CD), 1EA4(G_WX), 1EXJ(A_BD), 1FZP(B_KW), 1GD2(E_AB), 1GXP(E_GH), 1H6F(A_CD), 1HJB(B_GH), 1HJB(C_GH), 1IC8(A_EF), 1JFI(A_DE), 1JFI(B_DE), 1K78(A_CD), 1K78(B_CD), 1KB2(A_CD), 1LE5(F_GH), 1LMB(3_12), 1LQ1(B_GH), 1MHD(A_CD), 1NKP(D_HJ), 1OWR(P_EF), 1PNR(A_BD), 1QN3(B_EF), 1R8D(A_CD), 1RIO(H_TU), 1SAX(A_CD), 1SKN(P_AB), 1T2K(B_EF), 1XPX(A_CD), 1ZRE(A_WX), 1ZS4(A_TU), 2AC0(C_GH), 2BOP(A_BC), 2E1C(A_BD), 2H27(A_BC), 2H7H(A_XY), 2P5L(C_AB), 2R5Y(B_CD), 2WT7(A_CD), 2YVH(D_EF), 2ZHG(A_BC), 3A01(A_CD), 3COA(C_AB), 3DFX(B_XY), 3G97(A_CD), 3HDD(B_CD), 3IAG(C_AB), 3IKT(A_CD), 3JTG(A_BC), 3JXD(R_AB), 3O9X(A_EF), 3P57(B_EF), 3PVV(B_EF), 6ON0(A_CN), 3S8Q(A_CD), 3U2B(C_AB), 3ZKC(B_CD), 4FTH(A_CD), 4G92(B_DE), 6CRO(A_RU), 1ODH(A_CD), 2ISZ(D_EF), 2XSD(C_AB), 3COQ(A_DE), 3E6C(C_AB), 2ER8(C_GH), 2VY1(A_CW), 3A5T(A_CD), 3IGM(A_CD), 3LSR(A_BD), 3MLP(E_GH), 3VEB(A_MN),

Table S2: (Continued)

Dataset	PDBID (Protein-chain_DNA-chains)
	3W3C(A_BC), 3ZPL(F_GH), 4GCL(D_WZ), 4H10(A_CD), 4HF1(A_CD), 4IHT(C_GH), 4IX7(A_CD), 4JL3(A_EF), 3FDQ(A_CD), 3H0D(B_CD), 3N7Q(A_BC), 3U3W(A_YZ), 3W6V(A_BC), 3ZQL(A_EF), 4G92(A_DE), 4G92(C_DE), 4JCY(B_CD), 4KNY(A_YZ), 4L62(P_WX), 4LDX(A_CD), 4LLN(A_GH), 4LMG(D_GH), 4MTE(B_YZ), 4NNU(A_CD), 4ON0(B_EF), 4QTK(A_CD), 4U0Y(B_EF), 4UX5(A_CD), 1QPI(A_BM), 3DNV(B_ET), 3GNA(A_DE)
NS	1CEZ(A_NT), 1F66(C_IJ), 1JEY(A_CD), 1JEY(B_CD), 1RZT(A_BC), 1RZT(A_NO), 1SXQ(A_CE), 1X9W(A_CD), 1YA6(B_CD), 2BZF(A_BC), 2DNJ(A_BC), 2VOA(A_CD), 2WTF(A_OP), 3AAF(A_CD), 3AV2(A_IJ), 3CWS(C_GH), 3GV5(B_PT), 3L4J(A_BC), 3N4M(B_DE), 3UIQ(A_PT), 3JXY(A_BC), 4KLU(A_DT), 2WWY(A_PQ), 3KXT(A_BC), 3RAX(B_HJ), 4ELU(A_BC), 4G0V(B_DE), 4O0I(A_BC), 4OIN(D_GH), 2O8B(A_EF), 2O8B(B_EF), 3F2B(A_PT), 3L2P(A_BD), 4C2U(A_XY), 4DL4(A_PT), 4IR1(F_GH), 4O3M(A_PT), 4PLB(B_EF)

Table S3: Dataset of 214 protein-ssDNA complexes.

NDBID	PDBID	Classification
NA2201	4IQV	TRANSFERASE/TRANSFERASE INHIBITOR/DNA
NA2202	4IQW	TRANSFERASE/TRANSFERASE INHIBITOR/DNA
NA2415	4KI2	Transcription/DNA
NA2128	4I27	TRANSFERASE/DNA
NA2129	4I28	TRANSFERASE/DNA
NA2130	4I29	TRANSFERASE/DNA
NA2131	4I2A	TRANSFERASE/DNA
NA2132	4I2B	TRANSFERASE/DNA
NA2133	4I2C	TRANSFERASE/DNA
NA2134	4I2E	TRANSFERASE/DNA
NA2135	4I2F	TRANSFERASE/DNA
NA2136	4I2G	TRANSFERASE/DNA
NA2137	4I2H	TRANSFERASE/DNA
NA2322	4JRP	HYDROLASE/DNA
NA2323	4JRQ	HYDROLASE/DNA
NA2324	4JS4	HYDROLASE/DNA
NA2325	4JS5	HYDROLASE/DNA
NA2247	4J1J	VIRAL PROTEIN/DNA
NA1506	3VAF	RNA BINDING PROTEIN/DNA
NA1512	3VAG	RNA BINDING PROTEIN/DNA
NA1532	3VAH	RNA BINDING PROTEIN/DNA
NA1533	3VAI	RNA BINDING PROTEIN/DNA
NA1534	3VAJ	RNA BINDING PROTEIN/DNA
NA1535	3VAK	RNA BINDING PROTEIN/DNA
NA1536	3VAL	RNA BINDING PROTEIN/DNA
NA1537	3VAM	RNA BINDING PROTEIN/DNA

Table S3: (Continued)

NDBID	PDBID	Classification
NA1812	4FF1	TRANSFERASE/DNA
NA1813	4FF2	TRANSFERASE/DNA
NA1814	4FF3	TRANSFERASE/DNA
NA1816	4FF4	TRANSFERASE/DNA
NA2084	4HID	DNA BINDING PROTEIN
NA2083	4HIK	DNA BINDING PROTEIN
NA2082	4HIM	DNA BINDING PROTEIN
NA2080	4HIO	DNA BINDING PROTEIN
NA2079	4HJ5	DNA BINDING PROTEIN
NA2078	4HJ7	DNA BINDING PROTEIN
NA2077	4HJ8	DNA BINDING PROTEIN
NA2076	4HJ9	DNA BINDING PROTEIN
NA2075	4HJA	DNA BINDING PROTEIN
NA1993	4GOP	DNA BINDING PROTEIN/DNA
NA2114	4HQU	HORMONE/DNA
NA2115	4HQX	HORMONE/DNA
NA2028	4H5Q	VIRAL PROTEIN/DNA
NA1877	4FPV	Hydrolase/DNA
NA1757	4ESV	HYDROLASE/DNA
NA2001	4A75	CHAPERONE/DNA
NA2002	4A76	CHAPERONE/DNA
NA1520	3V9S	HYDROLASE/DNA
NA1521	3V9U	HYDROLASE/DNA
NA1522	3V9W	HYDROLASE/DNA
NA1523	3V9X	HYDROLASE/DNA
NA1524	3V9Z	HYDROLASE/DNA

Table S3: (Continued)

NDBID	PDBID	Classification
NA1525	3VA0	HYDROLASE/DNA
NA1526	3VA3	HYDROLASE/DNA
NA1883	4A8F	TRANSFERASE
NA1884	4A8K	TRANSFERASE/DNA
NA1885	4A8M	TRANSFERASE/DNA
NA1886	4A8Q	TRANSFERASE/DNA
NA1887	4A8S	TRANSFERASE/DNA
NA1888	4A8W	TRANSFERASE/DNA
NA1889	4A8Y	TRANSFERASE/DNA
2LTT	2LTT	TRANSCRIPTION, DNA BINDING PROTEIN/DNA
NA1402	3UPU	HYDROLASE/DNA
NA1375	3UDG	DNA BINDING PROTEIN/DNA
NA1472	3ULP	DNA BINDING PROTEIN/DNA
NA1593	3VDY	DNA BINDING PROTEIN/DNA
NA1384	3VKE	DNA BINDING PROTEIN/DNA
NA1248	3SZ5	HYDROLASE/DNA
NA1564	4A15	HYDROLASE
NA0978	3AUO	TRANSFERASE/DNA
NA1323	3U3Y	HYDROLASE/DNA
NA1355	3U6F	HYDROLASE/DNA
2L45	2L45	VIRAL PROTEIN/DNA
2L46	2L46	VIRAL PROTEIN/DNA
NA1373	3UGO	TRANSCRIPTION/DNA
NA1374	3UGP	TRANSCRIPTION/DNA
NA1376	3ZVM	HYDROLASE/TRANSFERASE/DNA
NA1039	3R8F	REPLICATION ACTIVATOR/DNA

Table S3: (Continued)

NDBID	PDBID	Classification
NA1040	3RA0	DNA BINDING PROTEIN/DNA
NA0556	3MXI	HYDROLASE/DNA
NA0558	3MXM	HYDROLASE/DNA
NA0883	3PX7	ISOMERASE/DNA
NA1014	2Y35	HYDROLASE/DNA
NA0891	3Q0A	TRANSCRIPTION/DNA
NA0594	3NGZ	HYDROLASE/DNA
NA0595	3NH0	HYDROLASE/DNA
NA0596	3NH1	HYDROLASE/DNA
NA0895	3Q22	TRANSFERASE/DNA
NA0898	3Q23	TRANSFERASE/DNA
NA0899	3Q24	TRANSFERASE/DNA/RNA
NA0122	3A5U	DNA BINDING PROTEIN
NA0565	3N1I	DNA BINDING PROTEIN/DNA
NA0566	3N1J	DNA BINDING PROTEIN/DNA
NA0567	3N1K	DNA BINDING PROTEIN/DNA
NA0568	3N1L	DNA BINDING PROTEIN/DNA
2KN7	2KN7	HYDROLASE/DNA
NA0603	3NGO	HYDROLASE/DNA
NA0369	3KQH	HYDROLASE/DNA
NA0370	3KQK	HYDROLASE/DNA
NA0371	3KQL	HYDROLASE/DNA
NA0372	3KQN	HYDROLASE/DNA
NA0373	3KQU	HYDROLASE/DNA
NA0289	3KJO	DNA BINDING PROTEIN/DNA
NA0290	3KJP	DNA BINDING PROTEIN/DNA

Table S3: (Continued)

NDBID	PDBID	Classification
NA0058	3I2O	OXIDOREDUCTASE/DNA
NA0059	3I3M	OXIDOREDUCTASE/DNA
NA0060	3I49	OXIDOREDUCTASE/DNA
PD1283	3H15	REPLICATION/DNA
PD1273	3GP8	HYDROLASE/DNA
PD1274	3GPL	HYDROLASE/DNA
NA0014	2VTB	LYASE/DNA
PD1138	3D2W	DNA/RNA BINDING PROTEIN
NA0339	2VYE	HYDROLASE/DNA
PD1083	3C2P	transferase/DNA
PD1084	3C3L	transferase/DNA
PD1086	3C46	transferase/DNA
PD1153	3DLB	Nucleic Acid Binding Protein/dna
PD1155	3DLH	Nucleic Acid Binding Protein/dna
PD1028	2Z70	HYDROLASE
PD1108	3CMU	recombination/DNA
PD1109	3CMW	recombination/DNA
PD1021	2QFJ	Transcription repressor/DNA
PD1045	3B39	transferase/DNA
PD1010	2PY5	REPLICATION, TRANSFERASE/DNA
PD0966	2OK0	IMMUNE SYSTEM/DNA
PD0915	2O19	ISOMERASE/DNA
PD0940	2O54	ISOMERASE/DNA
PD0941	2O59	ISOMERASE/DNA
PD0942	2O5C	ISOMERASE/DNA
PD0943	2O5E	ISOMERASE/DNA

Table S3: (Continued)

NDBID	PDBID	Classification
PD0948	2O4I	HYDROLASE/DNA
PD0959	2OA8	HYDROLASE/DNA
PD0890	2NMV	HYDROLASE/DNA
PD0785	2FR4	IMMUNE SYSTEM/DNA
PD0855	2DWL	HYDROLASE/DNA
PD0856	2DWM	HYDROLASE/DNA
PD0726	2ES2	GENE REGULATION
PD0858	2I5S	HYDROLASE/DNA
PD0850	2I0Q	structural protein/DNA
PD0675	1ZTG	DNA, RNA BINDING PROTEIN/DNA
PD0746	2D7D	HYDROLASE/DNA
PD0669	1ZM5	DNA BINDING PROTEIN/DNA
PD0766	2FDC	DNA BINDING PROTEIN/DNA
PD0767	2FD8	OXIDOREDUCTASE/DNA
PD0768	2FDF	OXIDOREDUCTASE/DNA
PD0769	2FDG	OXIDOREDUCTASE/DNA
PD0770	2FDH	OXIDOREDUCTASE/DNA
PD0771	2FDI	OXIDOREDUCTASE/DNA
PD0772	2FDK	OXIDOREDUCTASE/DNA
PD0745	2F55	HYDROLASE/DNA
PD0688	2A0I	hydrolase/DNA
PD0695	2A6O	TRANSCRIPTION/DNA
PD0715	2AXY	DNA BINDING PROTEIN/DNA
PD0680	1ZZI	STRUCTURAL PROTEIN/DNA
PD0681	1ZZJ	STRUCTURAL PROTEIN/DNA
PD0517	1S6M	DNA BINDING PROTEIN/DNA

Table S3: (Continued)

NDBID	PDBID	Classification
PD0596	1XF2	IMMUNE SYSTEM/DNA
PD0597	1XJV	TRANSCRIPTION/DNA
PD0599	1XHZ	TRANSFERASE/DNA
PD0603	1XI1	TRANSFERASE/DNA
PD0485	1RFF	HYDROLASE/DNA
PD0486	1RFI	HYDROLASE/DNA
PD0487	1RG1	HYDROLASE/DNA
PD0488	1RG2	HYDROLASE/DNA
PD0489	1RGT	HYDROLASE/DNA
PD0490	1RGU	HYDROLASE/DNA
PD0491	1RH0	HYDROLASE/DNA
PD0493	1RC8	TRANSFERASE/DNA
PD0499	1RPZ	TRANSFERASE/DNA
PD0500	1RRC	TRANSFERASE/DNA
PD0394	1OMH	TRANSFERASE/DNA
PD0402	1OSB	TRANSFERASE/DNA
PD0467	1QX0	TRANSFERASE/DNA
PD0469	1QZH	DNA Binding Protein/DNA
PD0468	1QZG	DNA BINDING Protein/DNA
PD0445	1PGZ	DNA BINDING PROTEIN/DNA
PD0451	1PO6	RNA Binding Protein/DNA
PD0352	1MW8	ISOMERASE
1OVF	1OVF	DNA/ANTIBIOTIC
PD0391	1NOP	HYDROLASE/DNA
PD0326	1M07	HYDROLASE/DNA
PD0276	1KEG	IMMUNE SYSTEM/DNA

Table S3: (Continued)

NDBID	PDBID	Classification
PD0275	1KDH	TRANSFERASE/DNA
1L1V	1L1V	DNA/ANTIBIOTIC
PD0278	1KIX	DNA BINDING PROTEIN/DNA
PD0266	1K8G	DNA BINDING PROTEIN/DNA
PD0206	1I8M	IMMUNE SYSTEM/DNA
PD0203	1I7D	ISOMERASE/DNA
PD0150	1F0V	hydrolase/DNA
PD0144	1EYG	replication/DNA
PD0106	2KZZ	TRANSFERASE/DNA
PD0072	1D8Y	TRANSFERASE/DNA
PR0016	1D9D	TRANSFERASE/DNA, RNA
PD0080	1D9F	TRANSFERASE/DNA, RNA
PD0087	2UP1	GENE REGULATION/DNA
PD0064	1QSL	TRANSFERASE/DNA
PD0074	1OTC	PROTEIN/DNA
PD0043	2PJR	HYDROLASE/DNA
PD0014	2KFN	TRANSFERASE/DNA
PD0015	2KFZ	TRANSFERASE/DNA
PDE0129	1UAA	HYDROLASE/DNA
PDE0137	1KFS	TRANSFERASE/DNA
PDE0138	1KRP	TRANSFERASE/DNA
PDE0136	1KSP	TRANSFERASE/DNA
PDO001	1JMC	REPLICATION/DNA
PDE090	1NOY	TRANSFERASE/DNA
PDE026	1LAU	HYDROLASE/DNA
PDE023	1RBJ	HYDROLASE/DNA

Table S3: (Continued)

NDBID	PDBID	Classification
PDE0117	1RCN	HYDROLASE/DNA
PDE013	1HUT	HYDROLASE/HYDROLASE INHIBITOR/DNA
PDA001	1CBV	IMMUNE SYSTEM/DNA
1HVO	1HVO	Viral protein/DNA
PDE0116	1RTA	HYDROLASE/DNA

Table S4: Domain-based non-redundant ssDNA-binding domains in SP and NS groups. Domain boundaries are assigned by CATH, and residue numbers are consistent with those in PDB.

Dataset	Domain ID	Domain definition	Domain ID	Domain definition
SP	1cbvH01	1cbv:H (1-121)	2o19A04	2o19:A (289-416)
	1cbvL01	1cbv:L (1-113)	2o5cA01	2o5c:A (1-156)
	1otcA01	1otc:A (37-210)	2o5cA02	2o5c:A (157-216,489-659)
	1otcA02	1otc:A (211-326)	4a75A00	4a75:A
	1pgzA01	1pgz:A (8-94)	3c2pA07	3c2p:A (609-803)
	1qzgA00	1qzg:A	3c2pA08	3c2p:A (804-882,978-1052)
	1rbjA00	1rbj:A	3d2wA00	3d2w:A
	1xjvA02	1xjv:A (151-299)	3q0aB01	3q0a:B (5-311,883-916)
	1zziA00	1zzi:A	3ugoA00	3ugo:A
	1osbA00	1osb:A	3v9xA00	3v9x:A
	2i0qB00	2i0q:B	3c2pA02	3c2p:A (312-338,411- 454,546-562,917-977)
NS	1jmcA01	1jmc:A (183-296)	3kqhA01	3kqh:A (189-326)
	1jmcA02	1jmc:A (297-420)	3kqhA02	3kqh:A (327-430,452-483)
	1lauE00	1lau:E	3kqlA03	3kql:A (431-451,484-625)
	1noyB02	1noy:B (107-338)	3n1iA00	3n1i:A
	1uaaA01	1uaa:A (2-107,185-273)	3sz5A00	3sz5:A
	1uaaA02	1uaa:A (108-181)	3ulpC00	3ulp:C
	1uaaB03	1uaa:B (274-376,533-640)	3zvmA01	3zvm:A (143-337)
	2kn7A00	2kn7:A	4a15A04	4a15:A (410-615)
	2lttA00	2ltt:A	4a8mQ02	4a8m:Q (37-91,518-606)
	2pjrA04	2pjr:A (387-543)	4fpvA00	4fpv:A
	3a5uA00	3a5u:A	4h5qB00	4h5q:B
	3cmwA01	3cmw:A(37-268,1001-1036)	4hqbA00	4hqb:A
	3dlbA02	3dlb:A (21-96)	4i2fA03	4i2f:A (302-440)
	3dlbA03	3dlb:A (178-266)	4j1jA02	4j1j:A (133-235)
	3gp8A01	3gp8:A (153-297)	4j1jC01	4j1j:C (2-132)

Table S4: (Continued)

Dataset	Domain ID	Domain definition	Domain ID	Domain definition
	3gp8A02	3gp8:A (298-492,703-716)	4jrpA01	4jrp:A (7-201)
	3gp8A03	3gp8:A (493-571,641-702)	4jrpB02	4jrp:B (214-313,314-357)
	3gplA04	3gpl:A (572-640)	4jrpB03	4jrp:B (358-420)
	3h15A00	3h15:A	3dlhB05	3dlh:B (320-466)
	2d7dA01	2d7d:A (3-89,117-148,320- 338,388-410)	3dlhA01	3dlh:A (3-20,97-177,267- 315,580-584)
	2d7dA02	2d7d:A (90-116,248- 319,339-387)	3dlhA04	3dlh:A (316-319,467- 579,585-685)

Table S5: The non-redundant pair dataset consists of 14 specific and 29 non-specific ssDNA-binding domains paired with their unbound structures.

Dataset	Holo domain ID	Holo domain definition	Apo homolog
SP	1cbvL01	1cbv:L (1-113)	1nbv:L (1-113)
	1pgzA01	1pgz:A (8-94)	1l3k:A (8-94)
	1rbjA00	1rbj:A	1kf5:A
	1zziA00	1zzi:A	1zzk:A
	2fr4B01	2fr4:B (1-120)	1xf3:B (1-120)
	2o19A01	2o19:A (1-156)	1d6m:A (1-156)
	2o19A02	2o19:A (157-216,489-658)	1d6m:A (157-216,489-653)
	2o19A04	2o19:A (289-416)	1d6m:A (289-416)
	3c2pA02	3c2p:A (312-338,411-454,546- 562,917-977)	2po4:A (312-338,411-454,546- 562,917-977)
	3c2pA07	3c2p:A (609-803)	2po4:A (609-803)
	3c2pA08	3c2p:A (804-882,978-1052)	2po4:A (804-882,978-1052)
	3q0aB01	3q0a:B (5-311,883-916)	2po4:A (5-311,883-916)
	3ugoA00	3ugo:A	1ku2:A
	4a75A00	4a75:A	3ulj:A
NS	1jmcA01	1jmc:A (183-296)	1fgu:A (183-296)
	1jmcA02	1jmc:A (297-420)	1fgu:A (297-420)
	1lauE00	1lau:E	1udg:A
	1noyB02	1noy:B (107-338)	1noz:B (107-338)
	2kn7A00	2kn7:A	2aq0:A
	2lttA00	2ltt:A	2ltd:A
	2pjrA01	2pjr:A (11-115,194-282)	1pjr:A (11-115,194-282)
	2pjrA03	2pjr:A (283-385,546-548)	1pjr:A (283-385,546-548)
	2pjrA04	2pjr:A (387-543)	1pjr:A (387-543)
	2pjrB01	2pjr:B (571-646)	1pjr:A (571-646)
	3a5uA00	3a5u:A	1x3e:A
	3cmwA09	3cmw:A (4037-4268)	2reb:A (37-268)

Table S5: (Continued)

Dataset	Holo domain ID	Holo domain definition	Apo homolog
	3gp8A01	3gp8:A (153-297)	3e1s:A (153-297)
	3gp8A02	3gp8:A (298-492,703-716)	3e1s:A (298-492,703-716)
	3gp8A03	3gp8:A (493-571,641-702)	3e1s:A (493-571,641-702)
	3gplA04	3gpl:A (572-640)	3e1s:A (572-640)
	3h15A00	3h15:A	3ebe:A
	3kqhA01	3kqh:A (189-326)	5e4f:A (189-326)
	3kqhA02	3kqh:A (327-430,452-483)	5e4f:A (327-430,452-483)
	3kqlA03	3kql:A (431-451,484-625)	5e4f:A (431-451,484-625)
	3n1iA00	3n1i:A	3n1h:A
	3sz5A00	3sz5:A	3syy:A
	3zvmA01	3zvm:A (143-337)	3zvl:A (143-337)
	4a8mQ02	4a8m:Q (37-91,518-606)	1hhs:A (37-91,518-606)
	4h5qB00	4h5q:B	4h5m:B
	4i2fA03	4i2f:A (302-440)	1jms:A (302-440)
	4jrpA01	4jrp:A (7-201)	3c95:A (7-201)
	4jrpB03	4jrp:B (358-420)	3c95:A (358-420)
	4jrqB02	4jrq:B (214-313,314-359)	3c95:A (214-313,314-359)