Published online 30 October 2019 *Nucleic Acids Research, 2019, Vol. 47, No. 21* ***11103–11113***
doi: 10.1093/nar/gkz963

New insights into protein–DNA binding specificity from hydrogen bond based comparative study

Maoxuan Lin and Jun-tao Guo*

Department of Bioinformatics and Genomics, The University of North Carolina at Charlotte, Charlotte, NC 28223, USA

Received May 08, 2019; Revised October 06, 2019; Editorial Decision October 10, 2019; Accepted October 10, 2019

ABSTRACT

Knowledge of protein–DNA binding specificity has important implications in understanding DNA metabolism, transcriptional regulation and developing therapeutic drugs. Previous studies demonstrated hydrogen bonds between amino acid side chains and DNA bases play major roles in specific protein–DNA interactions. In this paper, we investigated the roles of individual DNA strands and protein secondary structure types in specific protein–DNA recognition based on side chain-base hydrogen bonds. By comparing the contribution of each DNA strand to the overall binding specificity between DNA-binding proteins with different degrees of binding specificity, we found that highly specific DNA-binding proteins show balanced hydrogen bonding with each of the two DNA strands while multi-specific DNA binding proteins are generally biased towards one strand. Protein-base pair hydrogen bonds, in which both bases of a base pair are involved in forming hydrogen bonds with amino acid side chains, are more prevalent in the highly specific protein–DNA complexes than those in the multi-specific group. Amino acids involved in side chain-base hydrogen bonds favor strand and coil secondary structure types in highly specific DNA-binding proteins while multi-specific DNA-binding proteins prefer helices.

INTRODUCTION

Protein-DNA interactions play crucial roles in many cellular processes, such as transcription, DNA replication, DNA packaging and repair (1). 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 (2). 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 (3). 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 (4,5). 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) (6) and the advancement of technologies for exploring DNA binding motifs, such as ChIP-seq, protein-binding microarrays (PBMs) (7), systematic evolution of ligands by exponential enrichment combined with massively parallel sequencing (SELEX-seq) (8) and high-throughput SELEX (HT-SELEX) (9), 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 (10,11). 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 (12–15). It has been shown that hydrogen bonds between amino acids and bases also provide complex interactions leading to specific recognition (16). Bidentate in-

*To whom correspondence should be addressed. Tel: +1 704 687 7492; Fax: +1 704 687 8667; Email: jguo4@uncc.edu teractions, 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 (12,17,18). Recent studies also suggest that π-interactions between aromatic residues and DNA bases play important roles in specific protein–DNA recognition (17,19–22).

Shape readout refers to both global shape and local shape of target DNA sequences in protein–DNA recognition (10,23–28). 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 (23,29,30). Recently, Rohs group investigated DNA shape changes due to CpG methylation and demonstrated these epigenetic effects on protein–DNA binding (31). 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 (30). Based on these knowledge of protein–DNA binding specificity, various models have been developed for binding site prediction (20,24,30,32–35). 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 (36–38) (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 (37). 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 (38). 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 (36). 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.

We recently carried out a comparative analysis of protein–DNA complex structures with different degrees of binding specificity (17). Our results 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, we 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 (17). Protein flexibility is another key factor in specific protein–DNA recognition (39–43). 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 (17). Based on these observations, we developed a machine learning-based SVM (Support Vector Machine) model for TF (transcription factor)–DNA complex model assessment (44). 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 (44).

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 (45–48). 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 (45,48). 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 (46,47). 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 (10,12,14,18). 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.

MATERIALS AND METHODS

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 (17). Briefly, X-ray crystal structures of protein-dsDNA complexes with resolution ≤3 Å and *R*-factor ≤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 (49). 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% (50). One representative from each cluster was selected to generate the non-redundant dataset (17). 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 (17). 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).

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 (51) and FIRST (Floppy Inclusion and Rigid Substructure Topography) (52) 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 (52). 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 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 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 1C).

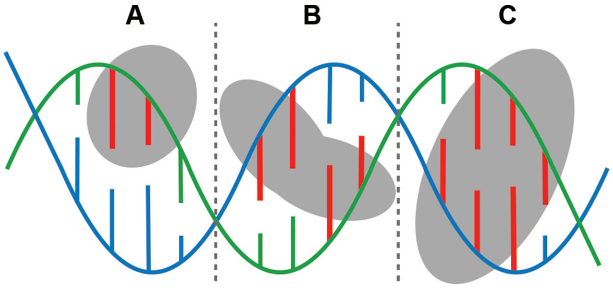

Figure 1. Schematic illustration of different types of side chain-base hydrogen bonds between two DNA strands (green and blue respectively) and a protein. The bases that form hydrogen bonds with protein side chain are colored red. (**A**) Hydrogen bonds between residue side chains and bases from only 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.

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 (53–56).

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 nonparametric Wilcoxon rank-sum test was applied.

RESULTS

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 (10,12,14,18). 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 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 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 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 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 2C) or for the major groove only (Figure 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 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 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 (Supplementary Figure S1).

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 Supplementary Figure S2. Similar patterns to the number of hydrogen bonds were found between the HS and MS groups.

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 DNA binding domains, similar patterns of differences to the domain-based analyses are found between the HS and MS groups (Supplementary Figure S3). 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 (30).

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 (12), 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 3A) or only base contacts in the major groove are considered (Figure 3B). Similar results are observed with FIRST (Supplementary Figure S4A and B). The *P*-value in Figure 3A 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 3C and Supplementary Figure S4C).

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

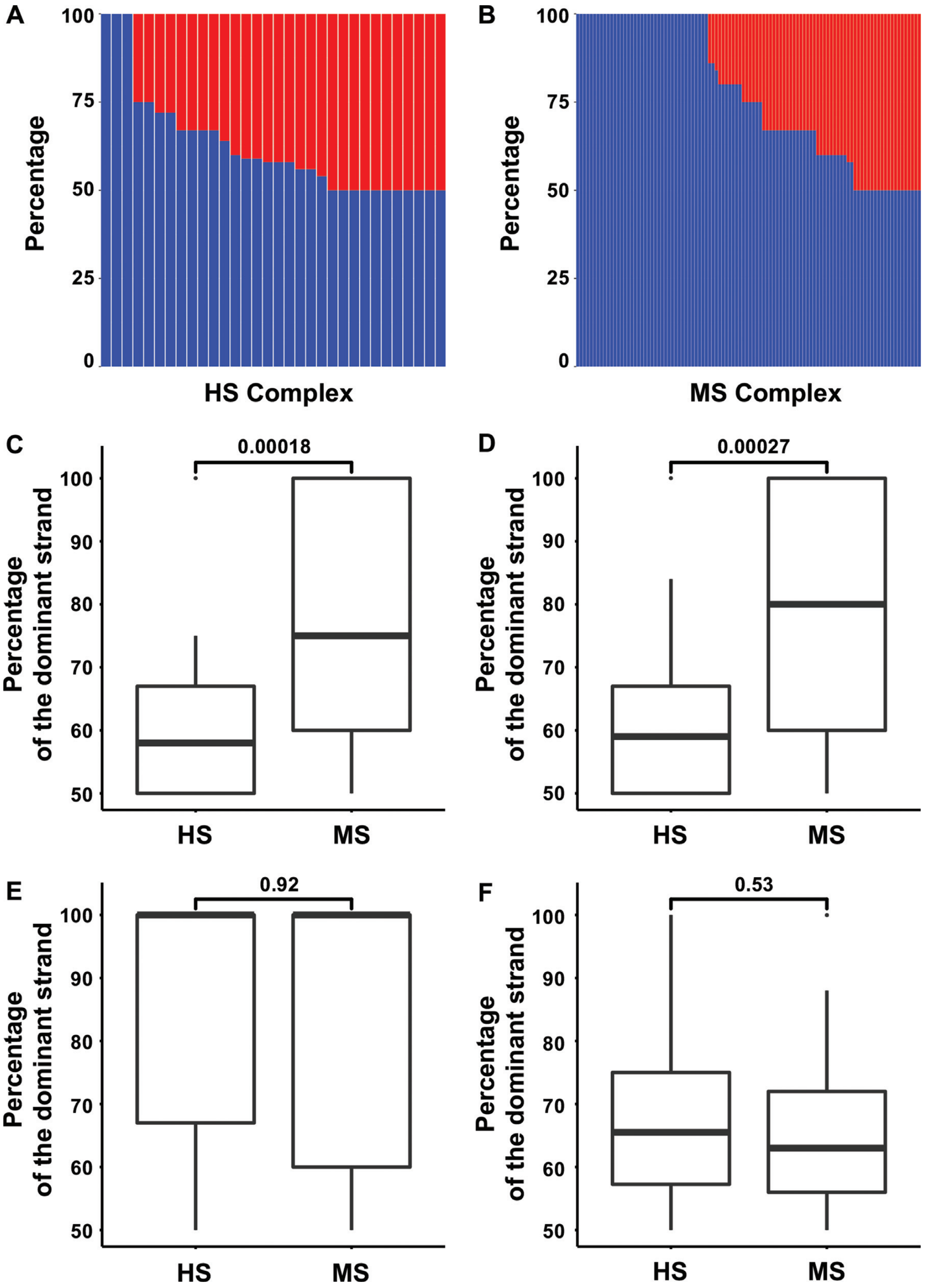

Figure 2. Comparison of the number of side chain-base hydrogen bonds of each strand of DNA annotated by HBPLUS between the HS and MS DNA-binding proteins. (**A**) Percentage contribution of two DNA strands in HS complexes; (**B**) percentage contribution of two DNA strands in MS complexes. The dominant 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.

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 4A), 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 chains (Figure 4B and Supplementary Figure S5). 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 5A and Sup-

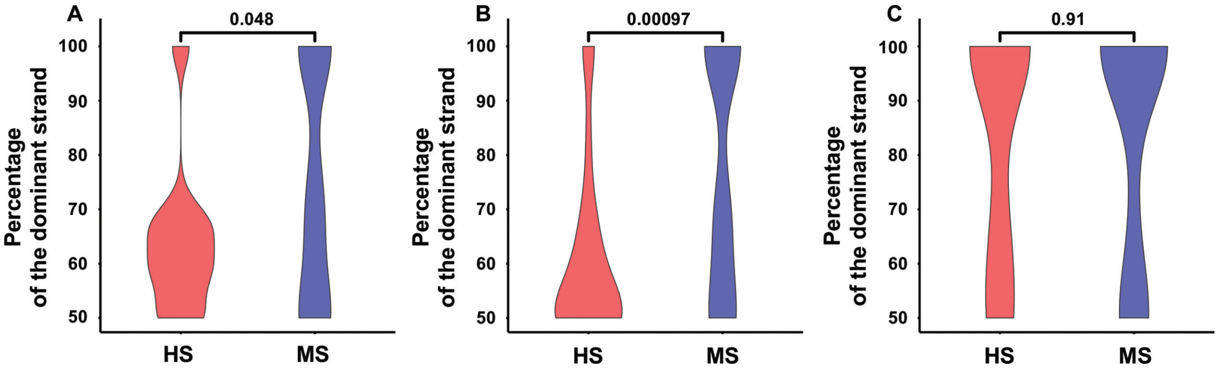

Figure 3. Comparison of the number of DNA bases involved in hydrogen bonding with side chains from HBPLUS for: (**A**) both major and minor grooves, (**B**) major groove only and (**C**) minor groove only, between HS and MS DNA-binding proteins.

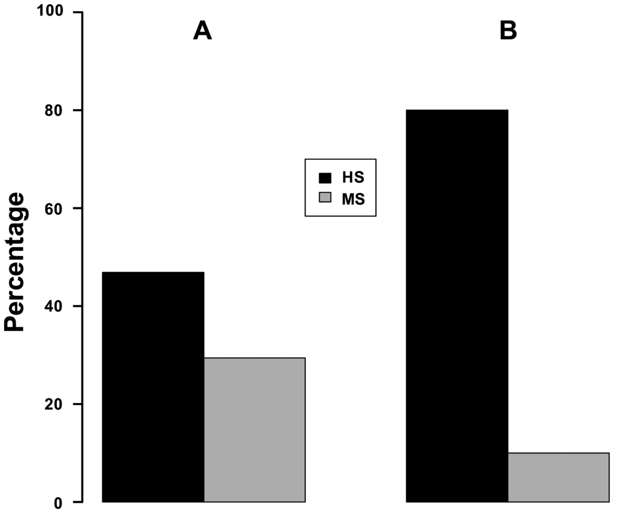

Figure 4. Comparison of the number of 50/50 cases (**A**) and the number of cases with base pairs involved in hydrogen bonding with residue side chains from these 50/50 cases (**B**) between the HS group and MS group with HBPLUS. (**A**) The percentage was calculated by dividing the number of 50/50 cases 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.

plementary Figure S5A) but the multi-specific sigma(E)4 forms such hydrogen bonds with unpaired bases (Figure 5B and Supplementary Figure S5B). The 50/50 cases from FIRST annotations show similar results with 8/13 ≈ 62% in the HS group and 4/24 ≈ 17% in the MS group (Supplementary Figure S6). The total amounts of base pairs involved in hydrogen bonding with residues are shown in Figure 6 (HBPLUS) and Supplementary Figure S7 (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.

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, ββα zinc finger and zipper-type motifs (1). The secondary structure types of amino acids involved in specific protein–DNA binding, 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 7A). 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 7B 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 7D 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 7B and D). However, when 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 7C 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 (Supplementary Figure S8A) while in hepatocyte nuclear factor 1-alpha (HNF-1alpha) residues in helical conformation are involved in hydrogen bonding with bases (Supplementary Figure S8B). 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 (Supplementary Figure S9). 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 ma-

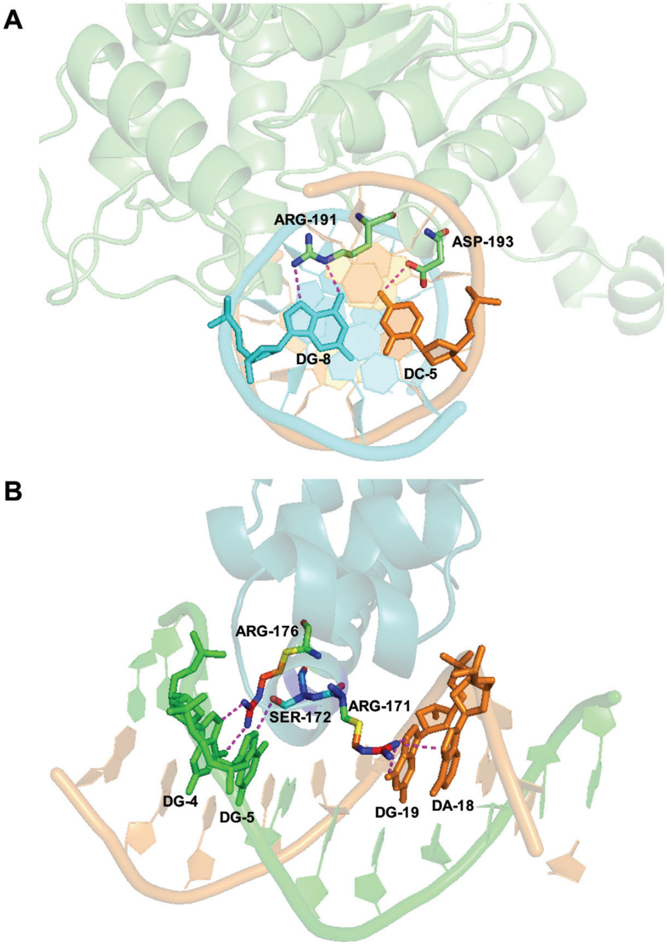

Figure 5. Examples of DNA-binding proteins bound to paired bases and unpaired bases. (**A**) Highly specific DNA-binding protein NgoMIV bound to paired bases (PDBID: 4ABT; protein chain: A; DNA chains: E and H). Only one out of three continuous base pairs involved hydrogen bonding is highlighted. Base 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).

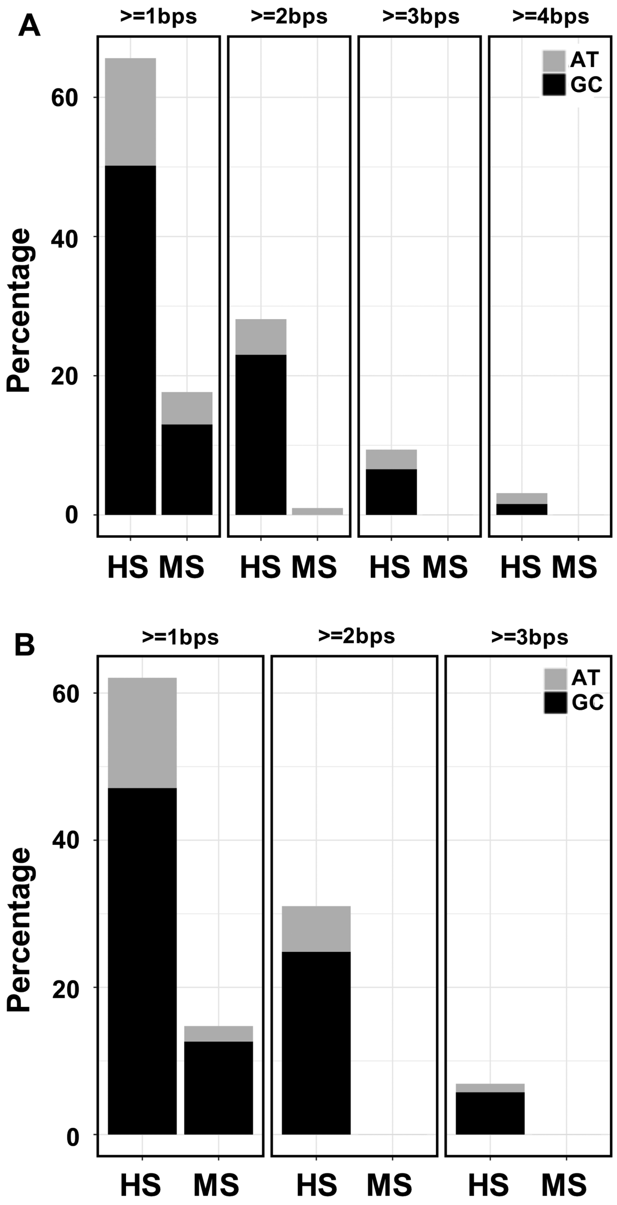

Figure 6. Comparison of base pairs that are involved in side chain-base hydrogen-bonding between HS and MS groups with HBPLUS in (**A**) both major and minor grooves and (**B**) major groove only.

jor+minor grooves (Figure 7B and D) and major groove alone (Figure 7C and E).

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 (10,12,14,18). 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 (17). Since there are a number of different algorithms for calculating hydrogen bond energy and typically a default energy cutoff is applied for 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, Supplementary Figure S1). Not only are these findings in agreement between HBPLUS and FIRST, they are also consis-

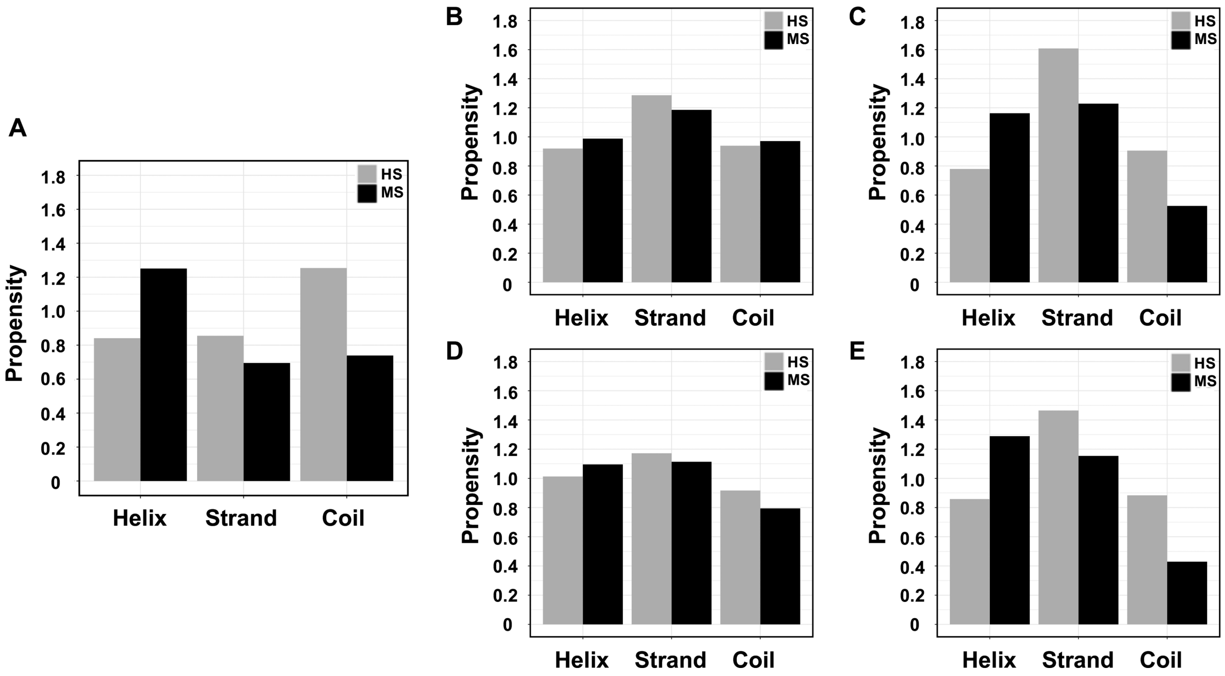

Figure 7. Propensities of secondary structure types in the HS and MS groups. (**A**) Propensities of secondary structure types of DNA base-contacting residues, the background relative frequencies of secondary structure types are calculated using all residues in the DNA-binding domains in each group. 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).

tent between domain-based and chain-based analyses (Supplementary Figure S3).

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 6 and Supplementary Figure S7). 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 (57). 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) (58). 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 (59). 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 6 and Supplementary Figure S7). 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 (60). 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 (61). 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 (62–65). For instance, by examining base substitutions that affect the presence and location of the 2-amino group of guanine in *tyr*T(A93) DNA, Bailly *et al.* found these alterations affect both the flexibility of *tyr*T(A93) DNA and its affinity for its binding protein, the *Escherichia coli* Factor for Inversion Stimulation (FIS) (65).

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 (26). These minor-groove interactions may stabilize the deformed DNA structure and identify incorrectly incorporated non-Watson-Crick base pairs (66). 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 (67–70).

DNA base-contacting residues in highly specific DNA-binding proteins are enriched in coils while multi-specific DNA-binding proteins prefer helices (Figure 7A). 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 7C and E, Supplementary Figure S9B and D). These results suggest that protein flexibility play important roles in protein–DNA recognition, as reported in previous studies (17,39–43). 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 (17). It has been hypothesized that protein flexibility can help speed up DNA recognition (71,72). 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 (42). 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 (42). 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 7C and E). Mutation tolerance study of different secondary structure elements of proteins shows that alpha helices are more robust to mutations than beta strands (73). 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 (74–77). 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) (78,79). 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 (78), and a decrease of nucleotide excision repair (NER) activity caused by interference of DNA-binding proteins with the NER machinery (79).

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.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank the editor and anonymous reviewers for their critical reading of the manuscript. Their constructive suggestions/comments helped improve the manuscript.

FUNDING

National Institutes of Health [R15GM110618 to J.G]; National Science Foundation [DBI1356459 to J.G]. Funding for open access charge: National Institutes of Health [R15GM110618].

Conflict of interest statement. None declared.

REFERENCES

1. Luscombe,N.M., Austin,S.E., Berman,H.M. and Thornton,J.M. (2000) An overview of the structures of protein–DNA complexes. *Genome Biol.*, **1**, doi:10.1186/gb-2000-1-1-reviews001.
2. Pingoud,A., Fuxreiter,M., Pingoud,V. and Wende,W. (2005) Type II restriction endonucleases: structure and mechanism. *Cell. Mol. Life Sci.*, **62**, 685–707.
3. Pan,Y., Tsai,C.J., Ma,B. and Nussinov,R. (2010) Mechanisms of transcription factor selectivity. *Trends Genet.*, **26**, 75–83.
4. Latchman,D.S. (1996) Transcription-factor mutations and disease. *N. Engl. J. Med.*, **334**, 28–33.
5. Schott,J.J., Benson,D.W., Basson,C.T., Pease,W., Silberbach,G.M., Moak,J.P., Maron,B.J., Seidman,C.E. and Seidman,J.G. (1998) Congenital heart disease caused by mutations in the transcription factor NKX2-5. *Science*, **281**, 108–111.
6. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
7. Berger,M.F., Philippakis,A.A., Qureshi,A.M., He,F.S., Estep,P.W. 3rd and Bulyk,M.L. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.
8. Slattery,M., Riley,T., Liu,P., Abe,N., Gomez-Alcala,P., Dror,I., Zhou,T., Rohs,R., Honig,B., Bussemaker,H.J. *et al.* (2011) Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins. *Cell*, **147**, 1270–1282.
9. Jolma,A., Kivioja,T., Toivonen,J., Cheng,L., Wei,G., Enge,M., Taipale,M., Vaquerizas,J.M., Yan,J., Sillanpaa,M.J. *et al.* (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. *Genome Res.*, **20**, 861–873.
10. Rohs,R., Jin,X., West,S.M., Joshi,R., Honig,B. and Mann,R.S. (2010) Origins of specificity in protein–DNA recognition. *Annu. Rev. Biochem.*, **79**, 233–269.
11. Slattery,M., Zhou,T., Yang,L., Dantas Machado,A.C., Gordan,R. and Rohs,R. (2014) Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.*, **39**, 381–399.
12. Luscombe,N.M., Laskowski,R.A. and Thornton,J.M. (2001) Amino acid-base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucleic. Acids. Res.*, **29**, 2860–2874.
13. Mandel-Gutfreund,Y., Schueler,O. and Margalit,H. (1995) Comprehensive analysis of hydrogen bonds in regulatory protein DNA-complexes: in search of common principles. *J. Mol. Biol.*, **253**, 370–382.
14. Pabo,C.O. and Sauer,R.T. (1992) Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.*, **61**, 1053–1095.
15. Suzuki,M. (1994) A framework for the DNA-protein recognition code of the probe helix in transcription factors: the chemical and stereochemical rules. *Structure*, **2**, 317–326.
16. Angarica,V.E., Perez,A.G., Vasconcelos,A.T., Collado-Vides,J. and Contreras-Moreira,B. (2008) Prediction of TF target sites based on atomistic models of protein–DNA complexes. *BMC Bioinformatics*, **9**, 436.
17. Corona,R.I. and Guo,J.T. (2016) Statistical analysis of structural determinants for protein–DNA-binding specificity. *Proteins*, **84**, 1147–1161.
18. Seeman,N.C., Rosenberg,J.M. and Rich,A. (1976) Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci. U.S.A.*, **73**, 804–808.
19. Baker,C.M. and Grant,G.H. (2007) Role of aromatic amino acids in protein-nucleic acid recognition. *Biopolymers*, **85**, 456–470.
20. Farrel,A., Murphy,J. and Guo,J.T. (2016) Structure-based prediction of transcription factor binding specificity using an integrative energy function. *Bioinformatics*, **32**, i306–i313.
21. Wilson,K.A., Kellie,J.L. and Wetmore,S.D. (2014) DNA-protein pi-interactions in nature: abundance, structure, composition and strength of contacts between aromatic amino acids and DNA nucleobases or deoxyribose sugar. *Nucleic Acids Res.*, **42**, 6726–6741.
22. Wintjens,R., Lievin,J., Rooman,M. and Buisine,E. (2000) Contribution of cation-pi interactions to the stability of protein–DNA complexes. *J. Mol. Biol.*, **302**, 395–410.
23. Azad,R.N., Zafiropoulos,D., Ober,D., Jiang,Y., Chiu,T.P., Sagendorf,J.M., Rohs,R. and Tullius,T.D. (2018) Experimental maps of DNA structure at nucleotide resolution distinguish intrinsic from protein-induced DNA deformations. *Nucleic Acids Res.*, **46**, 2636–2647.
24. Mathelier,A., Xin,B., Chiu,T.P., Yang,L., Rohs,R. and Wasserman,W.W. (2016) DNA Shape features improve transcription factor binding site predictions In Vivo. *Cell Syst.*, **3**, 278–286.
25. Otwinowski,Z., Schevitz,R.W., Zhang,R.G., Lawson,C.L., Joachimiak,A., Marmorstein,R.Q., Luisi,B.F. and Sigler,P.B. (1988) Crystal structure of trp repressor/operator complex at atomic resolution. *Nature*, **335**, 321–329.
26. Rohs,R., West,S.M., Sosinsky,A., Liu,P., Mann,R.S. and Honig,B. (2009) The role of DNA shape in protein–DNA recognition. *Nature*, **461**, 1248–1253.
27. Shakked,Z., Guzikevich-Guerstein,G., Frolow,F., Rabinovich,D., Joachimiak,A. and Sigler,P.B. (1994) Determinants of repressor/operator recognition from the structure of the trp operator binding site. *Nature*, **368**, 469–473.
28. Travers,A.A. (1989) DNA conformation and protein binding. *Annu. Rev. Biochem.*, **58**, 427–452.
29. Gordan,R., Shen,N., Dror,I., Zhou,T., Horton,J., Rohs,R. and Bulyk,M.L. (2013) Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.*, **3**, 1093–1104.
30. Jolma,A., Yan,J., Whitington,T., Toivonen,J., Nitta,K.R., Rastas,P., Morgunova,E., Enge,M., Taipale,M., Wei,G. *et al.* (2013) DNA-binding specificities of human transcription factors. *Cell*, **152**, 327–339.
31. Rao,S., Chiu,T.P., Kribelbauer,J.F., Mann,R.S., Bussemaker,H.J. and Rohs,R. (2018) Systematic prediction of DNA shape changes due to CpG methylation explains epigenetic effects on protein–DNA binding. *Epigenet. Chromatin*, **11**, 6.
32. Bulyk,M.L. (2003) Computational prediction of transcription-factor binding site locations. *Genome Biol.*, **5**, 201.
33. Li,J., Sagendorf,J.M., Chiu,T.P., Pasi,M., Perez,A. and Rohs,R. (2017) Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. *Nucleic Acids Res.*, **45**, 12877–12887.
34. Morozov,A.V., Havranek,J.J., Baker,D. and Siggia,E.D. (2005) Protein-DNA binding specificity predictions with structural models. *Nucleic. Acids. Res.*, **33**, 5781–5798.
35. Zhou,T., Shen,N., Yang,L., Abe,N., Horton,J., Mann,R.S., Bussemaker,H.J., Gordan,R. and Rohs,R. (2015) Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 4654–4659.
36. Golovenko,D., Brauning,B., Vyas,P., Haran,T.E., Rozenberg,H. and Shakked,Z. (2018) New Insights into the Role of DNA Shape on Its Recognition by p53 Proteins. *Structure*, **26**, 1237–1250.
37. Kitayner,M., Rozenberg,H., Rohs,R., Suad,O., Rabinovich,D., Honig,B. and Shakked,Z. (2010) Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. *Nat. Struct. Mol. Biol.*, **17**, 423–429.
38. Vainer,R., Cohen,S., Shahar,A., Zarivach,R. and Arbely,E. (2016) Structural Basis for p53 Lys120-Acetylation-Dependent DNA-Binding Mode. *J. Mol. Biol.*, **428**, 3013–3025.
39. Badia,D., Camacho,A., Perez-Lago,L., Escandon,C., Salas,M. and Coll,M. (2006) The structure of phage phi29 transcription regulator p4-DNA complex reveals an N-hook motif for DNA. *Mol. Cell.*, **22**, 73–81.
40. Fuxreiter,M., Simon,I. and Bondos,S. (2011) Dynamic protein–DNA recognition: beyond what can be seen. *Trends Biochem. Sci.*, **36**, 415–423.
41. Joshi,R., Passner,J.M., Rohs,R., Jain,R., Sosinsky,A., Crickmore,M.A., Jacob,V., Aggarwal,A.K., Honig,B. and Mann,R.S. (2007) Functional specificity of a Hox protein mediated by the recognition of minor groove structure. *Cell*, **131**, 530–543.
42. Song,W. and Guo,J.T. (2015) Investigation of arc repressor DNA-binding specificity by comparative molecular dynamics simulations. *J. Biomol. Struct. Dyn.*, **33**, 2083–2093.

43. Zhou,H.X. (2012) Intrinsic disorder: signaling via highly specific but short-lived association. *Trends Biochem. Sci.*, **37**, 43–48.
44. Corona,R.I., Sudarshan,S., Aluru,S. and Guo,J.T. (2018) An SVM-based method for assessment of transcription factor-DNA complex models. *BMC Bioinformatics*, **19**, 506.
45. Cogan,J.G., Sun,S., Stoflet,E.S., Schmidt,L.J., Getz,M.J. and Strauch,A.R. (1995) Plasticity of vascular smooth muscle alpha-actin gene transcription. Characterization of multiple, single-, and double-strand specific DNA-binding proteins in myoblasts and fibroblasts. *J. Biol. Chem.*, **270**, 11310–11321.
46. Davis-Smyth,T., Duncan,R.C., Zheng,T., Michelotti,G. and Levens,D. (1996) The far upstream element-binding proteins comprise an ancient family of single-strand DNA-binding transactivators. *J. Biol. Chem.*, **271**, 31679–31687.
47. Haas,S., Steplewski,A., Siracusa,L.D., Amini,S. and Khalili,K. (1995) Identification of a sequence-specific single-stranded DNA binding protein that suppresses transcription of the mouse myelin basic protein gene. *J. Biol. Chem.*, **270**, 12503–12510.
48. Sun,S., Stoflet,E.S., Cogan,J.G., Strauch,A.R. and Getz,M.J. (1995) 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. *Mol. Cell. Biol.*, **15**, 2429–2436.
49. Kim,R. and Guo,J.T. (2009) PDA: an automatic and comprehensive analysis program for protein–DNA complex structures. *BMC Genomics*, **10**, S13.
50. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
51. McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.
52. Jacobs,D.J., Rader,A.J., Kuhn,L.A. and Thorpe,M.F. (2001) Protein flexibility predictions using graph theory. *Proteins*, **44**, 150–165.
53. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
54. Kim,R. and Guo,J.T. (2010) Systematic analysis of short internal indels and their impact on protein folding. *BMC Struct. Biol.*, **10**, 24.
55. Lin,M., Whitmire,S., Chen,J., Farrel,A., Shi,X. and Guo,J.T. (2017) Effects of short indels on protein structure and function in human genomes. *Sci. Rep.*, **7**, 9313.
56. Touw,W.G., Baakman,C., Black,J., te Beek,T.A., Krieger,E., Joosten,R.P. and Vriend,G. (2015) A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.*, **43**, D364–D368.
57. Pingoud,A. and Jeltsch,A. (2001) Structure and function of type II restriction endonucleases. *Nucleic Acids Res.*, **29**, 3705–3727.
58. Sonawane,A.R., Platig,J., Fagny,M., Chen,C.Y., Paulson,J.N., Lopes-Ramos,C.M., DeMeo,D.L., Quackenbush,J., Glass,K. and Kuijjer,M.L. (2017) Understanding Tissue-Specific Gene Regulation. *Cell Rep.*, **21**, 1077–1088.
59. Badis,G., Berger,M.F., Philippakis,A.A., Talukder,S., Gehrke,A.R., Jaeger,S.A., Chan,E.T., Metzler,G., Vedenko,A., Chen,X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
60. Nadassy,K., Wodak,S.J. and Janin,J. (1999) Structural features of protein-nucleic acid recognition sites. *Biochemistry*, **38**, 1999–2017.
61. Nikolajewa,S., Beyer,A., Friedel,M., Hollunder,J. and Wilhelm,T. (2005) Common patterns in type II restriction enzyme binding sites. *Nucleic Acids Res.*, **33**, 2726–2733.
62. Bailly,C., Mollegaard,N.E., Nielsen,P.E. and Waring,M.J. (1995) The influence of the 2-amino group of guanine on DNA conformation. Uranyl and DNase I probing of inosine/diaminopurine substituted DNA. *EMBO J.*, **14**, 2121–2131.
63. Bailly,C. and Waring,M.J. (1995) Transferring the purine 2-amino group from guanines to adenines in DNA changes the sequence-specific binding of antibiotics. *Nucleic Acids Res.*, **23**, 885–892.
64. Bailly,C. and Waring,M.J. (1995) The purine 2-amino group as a critical recognition element for specific DNA cleavage by bleomycin and calicheamicin. *J. Am. Chem. Soc.*, **117**, 7311–7316.
65. Bailly,C., Waring,M.J. and Travers,A.A. (1995) Effects of base substitutions on the binding of a DNA-bending protein. *J. Mol. Biol.*, **253**, 1–7.
66. Luscombe,N.M. and Thornton,J.M. (2002) Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J. Mol. Biol.*, **320**, 991–1009.
67. Doublie,S., Tabor,S., Long,A.M., Richardson,C.C. and Ellenberger,T. (1998) Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 A resolution. *Nature*, **391**, 251–258.
68. Kiefer,J.R., Mao,C., Braman,J.C. and Beese,L.S. (1998) Visualizing DNA replication in a catalytically active Bacillus DNA polymerase crystal. *Nature*, **391**, 304–307.
69. Morales,J.C. and Kool,E.T. (1999) Minor groove interactions between polymerase and DNA: More essential to replication than Watson-Crick Hydrogen Bonds? *J. Am. Chem. Soc.*, **121**, 2323–2324.
70. Pelletier,H., Sawaya,M.R., Kumar,A., Wilson,S.H. and Kraut,J. (1994) Structures of ternary complexes of rat DNA polymerase beta, a DNA template-primer, and ddCTP. *Science*, **264**, 1891–1903.
71. Levy,Y., Onuchic,J.N. and Wolynes,P.G. (2007) Fly-casting in protein–DNA binding: frustration between protein folding and electrostatics facilitates target recognition. *J. Am. Chem. Soc.*, **129**, 738–739.
72. Shoemaker,B.A., Portman,J.J. and Wolynes,P.G. (2000) Speeding molecular recognition by using the folding funnel: the fly-casting mechanism. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 8868–8873.
73. Abrusan,G. and Marsh,J.A. (2016) Alpha Helices Are More Robust to Mutations than Beta Strands. *PLoS Comput. Biol.*, **12**, e1005242.
74. Granzhan,A., Kotera,N. and Teulade-Fichou,M.P. (2014) Finding needles in a basestack: recognition of mismatched base pairs in DNA by small molecules. *Chem. Soc. Rev.*, **43**, 3630–3665.
75. Nikolova,E.N., Zhou,H., Gottardo,F.L., Alvey,H.S., Kimsey,I.J. and Al-Hashimi,H.M. (2013) A historical account of Hoogsteen base-pairs in duplex DNA. *Biopolymers*, **99**, 955–968.
76. Rossetti,G., Dans,P.D., Gomez-Pinto,I., Ivani,I., Gonzalez,C. and Orozco,M. (2015) The structural impact of DNA mismatches. *Nucleic Acids Res.*, **43**, 4309–4321.
77. Yang,C., Kim,E. and Pak,Y. (2015) Free energy landscape and transition pathways from Watson-Crick to Hoogsteen base pairing in free duplex DNA. *Nucleic Acids Res.*, **43**, 7769–7778.
78. Reijns,M.A.M., Kemp,H., Ding,J., de Proce,S.M., Jackson,A.P. and Taylor,M.S. (2015) Lagging-strand replication shapes the mutational landscape of the genome. *Nature*, **518**, 502–506.
79. Sabarinathan,R., Mularoni,L., Deu-Pons,J., Gonzalez-Perez,A. and Lopez-Bigas,N. (2016) Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature*, **532**, 264–267.