2023

Investigating bZip Recognition of DNA Sequences Through a Knob-Socket Perspective

Aaron Tran
University of the Pacific

Investigating bZip Recognition of DNA Sequences Through a Knob-Socket Perspective

By

Aaron H.Q. Tran

A Thesis Submitted

In Partial Fulfillment of the

Requirements for the Degree of

MASTER OF SCIENCE

Thomas J. Long School of Pharmacy and Health Sciences
Pharmaceutical and Chemical Sciences

University of the Pacific
Stockton, California

2023

Investigating bZip Recognition of DNA Sequences Through a Knob-Socket Perspective

By

Aaron H.Q. Tran

APPROVED BY:

Thesis Advisor: Jerry Tsai, Ph. D.

Committee Member: Liang Xue, Ph. D.

Committee Member: Joseph Harrison, Ph. D.

Department Co-Chair: Jianhua Ren, Ph. D.

Department Co-Chair: Jerry Tsai, Ph. D.

Investigating bZip Recognition of DNA Sequences Through a Knob-Socket Perspective

By

Aaron H.Q. Tran

Dedication

This thesis is dedicated to my sisters and parents, who have been my biggest supporters throughout my entire life. Thank you for your constant love and support, which has helped me through the best and worst of times in my graduate school experience.

Acknowledgments

My gratitude goes to the many people, from faculty, family, and friends, who have helped me over the past few years. I would like to thank Professor Jerry Tsai for his constant guidance, support, and patience over the past three years. Your experience and advice have helped me prepare for the next steps in my academic career, and I will hold your lessons close. Much thanks to Dr. Hyun Joo as well. You have been a constant source of advice for both research and technological topics and has greatly influenced my own approach towards computational methods. Thank you for your guidance and patience.

I would also like to thank my thesis committee, Dr. Liang Xue and Dr. Joseph Harrison. Thank you, Dr. Xue, for helping me throughout my undergraduate research and allowing me the opportunity to venture into structural biology. My time in your lab has greatly impacted my research interests and methods. Thank you, Dr. Harrison, for being an amazing teacher and mentor throughout my time as both an undergraduate and graduate student. Our frequent talks and conversations have also greatly impacted my own approach to science, and I would not be here without your support.

To my lab mates Christopher Argueta, Guillermo Ibal, Melina Huey, and Polina Eidelberg. Thank you for the constant advice towards my research project. Huge thanks to Chris and Polina for your constant support and for entertaining my many random ideas and conversation starters.

Finally, a huge thank you to my friends and family for the constant unconditional love and support throughout this process, without which I would not be here today.

Investigating bZip Recognition of DNA Sequences Through a Knob-Socket Perspective

Abstract

By Aaron H.Q. Tran

University of the Pacific
2023

To investigate whether higher order packing interactions confer protein-DNA specificity, a modified Knob-Socket (KS) model was used to analyze the interface of bZIP-DNA crystal structures. The KS analysis identified a nine-residue quadripartite recognition core consisting of four contiguous KS pockets P1, P2, N3, and N4 that each pack one of the four DNA half-site bases in the target sequence. Only one base per base pair packs, and these interactions are split across the DNA strands: the first two positive strand positions 1p and 2p pack into P1 and P2 while the last two negative strand positions 3n and 4n pack into N3 and N4. Amino acid sequence analysis of the four KS pocket regions indicates that the primary mechanism recognition is packing or non-packing of the 5-methyl group of dT as well as 5-methylcytosine. P1 shows little packing of dT; P2 packs dT but including two Asn residues in this pocket seems to block packing in this region; N3 also packs dT, but including a Phe also blocks packing; N4 consistently packs dT. This analysis demonstrates that there is an amino acid code to DNA recognition, allowing for multi-residue recognition and packing of the 5-methyl group.

Table of Contents

List of Tables

Table

List of Figures

Figure

List of Abbreviations

A	alanine
bp	base pair
bZIP domain	basic leucine zipper domain
C	cysteine
CREB protein	CRE-binding protein
CRE	cAMP response element
D	aspartic acid
DNA	deoxyribonucleic acid
dA	deoxyadenosine, adenine
dT	deoxythymidine, thymine
dG	deoxyguanosine, guanine
dC	deoxycytidine, cytosine
5mC	5-methylcytosine
E	glutamic acid
F	phenylalanine
G	glycine
H	histidine
I	isoleucine
K	lysine
KS	Knob-Socket
L	leucine

M	methionine
N	asparagine
P	proline
PDB	protein data bank
Q	glutamine
R	arginine
RPC	relative packing clique
S	serine
T	threonine
V	valine
W	tryptophan
Y	tyrosine

List of Symbols

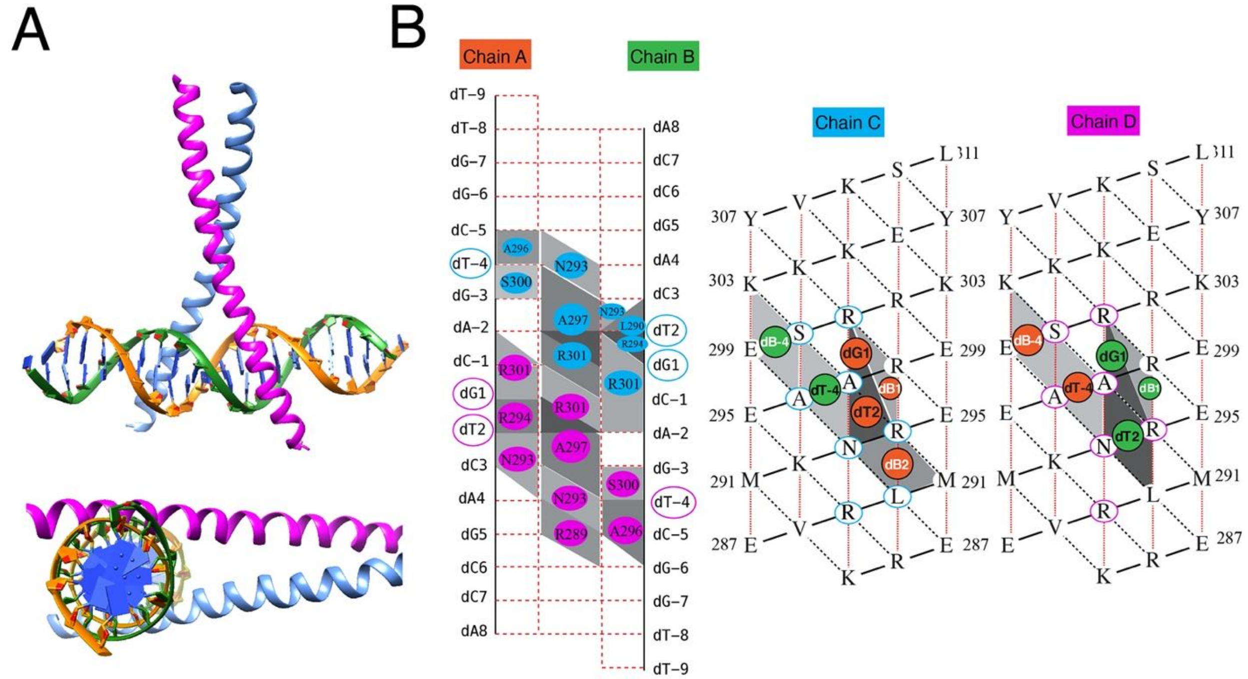
α	alpha
β	beta
5’	five prime
3’	three prime
%	percent

CHAPTER 1: INTRODUCTION

Protein transcription factors regulate gene expression by binding to specific DNA sequences. While DNA-binding domains vary in binding mechanism, sequence length, and recognition sequence to *cis*-regulatory gene sequences (1), a code for protein recognition of DNA has only been found in one system (2). Even with numerous crystal structures of proteins bound to DNA (1), a general code for how the twenty amino acids recognize the four nucleotide bases in duplex DNA has been difficult to uncover for numerous reasons. The goal has been to find a base readout mechanism for nucleotide recognition by protein residues using radial cutoffs to identify pairwise hydrogen bonds and hydrophobic contacts. However, amino acids interacting with DNA through hydrogen bonds demonstrate no preference for any particular nucleotide base, and in many cases, the hydrogen bonding requires a water molecule to mediate the interaction (3). Most of the protein-DNA interactions occur between amino acids and the repetitive sugar-phosphate backbone through hydrogen bonds. Because the DNA backbone is identical at every position, such interactions provide affinity, but no discernable specificity. It is difficult to identify specificity in protein-DNA using pairwise base readout approaches, perhaps due to the challenge of capturing the spatial diversity of shape and packing between amino acids and nucleotides. Instead, protein-DNA specificity may involve the shape readout of bases by amino acids through higher order, multi-body packing interactions. In this work, the Knob-Socket (KS) description of protein residue packing (4-7) has been modified to consider the structure of duplex DNA and applied to investigate the multi-body contributions of packing to specificity between protein and DNA. In particular, the modified KS analysis provides a framework to identify the higher order structural arrangements of groups of amino acids as well as nucleotide bases that determine binding of proteins to a DNA sequence. As an initial study to investigate a packing-based model of protein-DNA binding, the modified KS approach was used in the comprehensive analysis of one of the simplest protein-DNA interfaces: an α-helix binding into the major groove of DNA as found in the basic leucine zipper (bZip) family of transcription factors.

Figure 1

Structure and Knob-Socket Packing Topology Map of a bZIP Transcription Factor Bound to DNA



Note. A) Crystal structure of a bZIP-DNA complex PDB: 1dh3 (35) is depicted from two views for clarity: the top is viewed from the side and the bottom is viewed from down the axis of the DNA duplex. In each, the two protein bZIP monomers are colored in light blue and magenta, respectively. For the single DNA duplex, the strand starting 5' on the left in the top figure and coming out of the page in the bottom figure is colored in orange, and the strand starting 5' on the right in the top figure and into the page on the bottom figure is colored in green. DNA base pairs are colored dark blue. Each bZIP monomer's N-terminal basic region binds only half of the (Continued, Figure 1) binding site, or half-site, in the DNA major groove, while the C-terminal leucine zipper region regulates dimerization of monomers. Images were created using Chimera (59). B) A representative Knob-Socket packing topology map of only the DNA-protein interactions is shown. On the left is a simplified 2D DNA lattice of the cognate recognition sequence, while the right two lattices represent the N-terminal basic regions of the two α-helical bZIP monomers. On each of the lattices, filled circles represent knobs packing into areas of gray that indicate packing sockets, while open circles are residues that act as knobs in the packing interface. Thus, for the DNA helix on the left, the filled sockets originate from one of the two bZIP monomers and are colored respectively: Chain C in light blue and Chain D in magenta. The packing across the entire cognate recognition sequence is divided into two areas of packing or half-sites that are almost identical in pattern. The open circles around nucleotides indicate these bases pack into the corresponding α-helix monomer. On the right, each of the bZIP monomer α-helical socket lattices of chains C and D show the packing of the DNA knobs, where interactions are color coded based on the strand each knob belongs: Chain A in orange and Chain B in green. The packing of each DNA half-site into the individual protein α-helices is very similar between chains C and D. The packing topology map also reveals that half-site packing involves bases originating from both strands of the DNA, although it is commonly assumed that each bZIP monomer recognizes the 4 contiguous half-site bases on only one DNA strand.

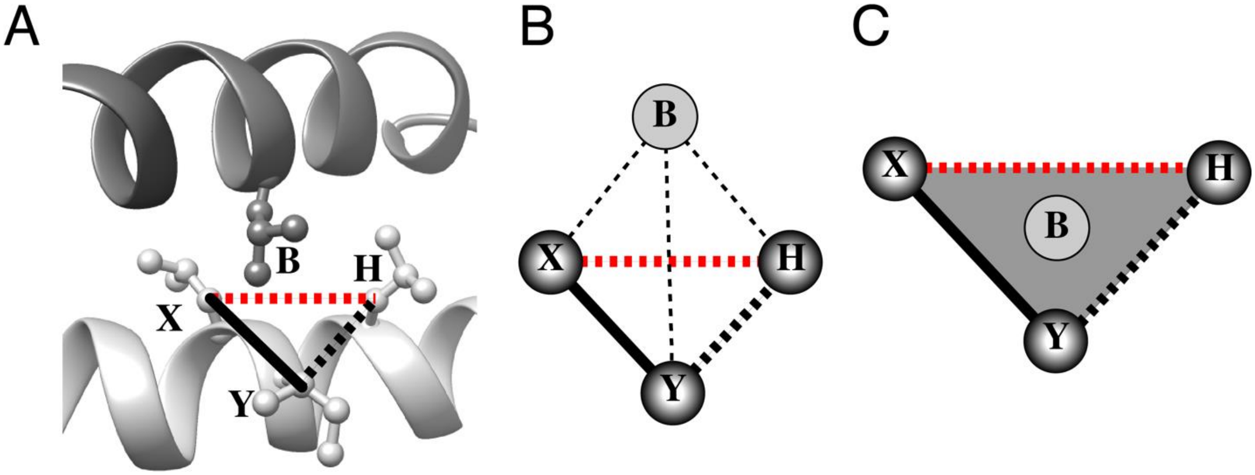
The basic region leucine zipper transcription factors (bZIPs) form homo- and heterodimers to bind a wide range of palindromic and pseudo-palindromic DNA sequences, allowing for higher-order regulation through combinatorial dimerization (8). As shown in Figure 1A, bZip proteins are α-helical dimers, where each 50 to 60 amino acid monomer consists of a basic DNA binding region of 20 to 25 residues and a leucine zipper dimerization region of 30 to

40 residues (9). The basic region of a bZIP α-helix binds into the major groove of a DNA duplex, splitting the DNA cognate site into two half-sites (10), while the leucine zipper region regulates coiled-coil dimerization between bZIP monomers (8). The bZIP dimer binds perpendicularly to the DNA cognate site in a T-like structure, where the protein basic regions clamp into the DNA duplex. In binding DNA, each bZip monomer's basic region recognizes four adjacent nucleotides in the major groove of the double helix structure and makes up one half of the overall recognition site, which is commonly referred to as a half-site (10). Therefore, the full recognition site for a bZIP protein dimer involves eight total DNA bases consisting of two half-sites from each monomer. In homodimers, the two half-sites consist of the same four-base sequence, while in heterodimers, the two half-sites are likely different. The majority of bZip binding sequences are contiguous and therefore eight bases long. However, some recognition sites are seven bases long with a one base pair overlap while other sites are nine bases long with a one base pair separation. In each of these cases, the helices of the bZIP dimer clamp onto opposite sides of the DNA duplex into the major groove, extending from the center of the DNA site outward (Figure 1A).

To investigate the basic regions specificity for DNA, structural studies of bZIP-DNA co-crystals have identified interactions, including hydrogen bond networks and hydrophobic interactions, between protein residues and DNA nucleotides (10-15). These studies identified differing pairwise interactions that may contribute to DNA base/sequence recognition schemes regardless of homologous sequences. Common hydrophobic interactions found within bZIP-DNA complexes describe the recognition of the methyl group of dT and its analogues (10-12,14-21). Specificity of the bZIP-DNA interaction has also been studied through binding assays to determine the effects of altering the bZIP sequence, correlating bZIP residue number and composition to consequential changes in protein-bound DNA sequences. Mutational studies targeted specific bZIP residues to affect DNA specificity, mutating base-contacting and conserved residues in the bZIP domain (11,13,22-25). Previous structural and binding studies have identified residues that contact DNA bases directly, generally ranging from four to five residues per bZIP helix depending on the subfamily (11,26-28). Most identified residues significant to DNA specificity have been proximal to the invariant Asn and Arg, though the number of significant residues differ between bZIP dimers and DNA sites (10-13,17,23-25,29). Analogously, other studies have mutated the DNA cognate site to probe the rigidity of the DNA sequence specificity and observed nucleotide variances at specific positions within the cognate site (14,21,22). Specific positions within a consensus sequence display a higher tolerance for nucleotide mutations than others, implicating a hierarchy between nucleotide positions (8,13,15,30-34). The variable number of DNA-contacting residues appears to limit possible bZIP engineering to each subfamily or sequence, though a general recognition model between bZIP and DNA would expand these sets of residues to apply across the bZIP superfamily. Overall, the recognition of DNA bases depends on the nucleotide position within the cognate site, and the pairwise interactions mentioned are not inherently specific between a single amino acid and nucleotide pair (8,13,31,32,34). While bZIP residues have been shown to interact with both strands of a half-site, no consensus for binding specificity has been found to distinguish which and how DNA bases are recognized in the half-site. These results suggest that bZIP-DNA recognition is more complicated and requires the consideration of higher order, multi-bodied interactions. Therefore, current understanding would benefit from a general recognition model between bZIP dimers and the respective DNA consensus sequences, regardless of bZIP subfamily.

Figure 2

Knob-Socket Model and Motif Involving α-Helical Packing



Note. A) Knob-Socket motif formed in the leucine zipper region of bZIP-DNA complex 1dh3 (35), categorized as a 2:1+1 relative packing clique (RPC). Residues **X**, **Y**, and **H** are in the same local secondary structure; **X** and **Y** are covalently bonded (solid black line), while **X** and **H** are hydrogen bonded (dashed red line) and **Y** and **H** contact via Van der Waals interactions. These three sockets are denoted as a 2:1 socket, which is the most common socket for α-helical packing. **B** packs and interacts with the entire socket to create the tetrahedral 2:1+1 motif, where **B** is represented by "+1." B) A simplified 3D representation of Figure 2A, where residues are represented by circles or nodes, and contacts are represented by the various lines connecting the nodes. C) A 2D representation of the knob-socket motif, where filled sockets are shaded in, with the knob, **B**, found in the center. This socket representation is often used for packing maps.

In this study, the contribution of higher order packing interactions to bZIP-DNA specificity is investigated using the KS model, which simplifies biomolecular packing into easily interpretable maps of three- and four-body motifs. Packing graphs are constructed from residue contacts, which are calculated using Voronoi polyhedral and Delaunay tessellations. Packing surface areas for each residue-to-residue contact are further calculated and associated with each edge between residues. Cliques of residues, or subsets of mutually contacting nodes, are extracted from packing graphs to define protein packing and structure. These relative packing cliques (RPC) can indicate stability or packing preference of three-body sockets. The KS model of packing provides a clear topological depiction of packing between two macromolecules by defining a regular surface for each molecule as repetitive patterns of three-body sockets. Figure 2 supplies an example of the most common packing motif for α-helices, the 2:1+1 clique. The socket is formed from three residues from the same secondary structure, forming a 2:1 socket. The two adjacent residues (**XY**) are represented by "2" as the distal residue (**H**) is by "1"; the distance between the two covalently bonded residues and the distal residue is displayed as ":" to form the 2:1 socket. **B** is another residue or body that is from a separate secondary structure that packs into the 2:1 socket, acting as a socket within the knob-socket motif. The knob packing, indicated by the "+1," fills a socket and forms a common tetrahedral RPC. The majority of α-helical packing can be defined by repetitive and defined 2:1 and 2:1+1 sockets. These three-body sockets pack knob residues specify distinct locations for quaternary packing of protein residues or DNA bases.

As an example, Figure 1B shows an example KS packing topology map of the CREB homodimer bound to its DNA recognition site (35) using the modified analysis. Contrasting the conventional treatment of protein residues in prior KS analyses, the modified KS analysis considers DNA bases and backbone separately as knobs due to the differences in their sizes and roles in protein-DNA recognition. Three lanes are recognized as a result, where sockets form within both positive and negative backbone strands and between both strands. While this modification will be discussed in more detail below, the separation into three lanes can be seen in the lattice representation of DNA on the left of Figure 1B. The two types of packing lattices shown in Figure 1B are the result of the repetitive three-residue socket formation by the macromolecular structures. Packing of protein knobs into the modified DNA lattice is shown on the left of Figure 1B. In a nearly symmetrical pattern, the two CREB α-helices have residues that pack into the DNA duplex lattice and extend outward from the central base pairs of the CREB recognition site. For DNA packing into the protein, the symmetrical and palindromic CREB nucleotide half-site packs into their respective monomer protein helices virtually identically (Figure 1B, right). This symmetry in packing can be clearly seen in both the amino acids packing into the DNA lattice and the nucleotide backbone and bases packing into the two α-helical lattices. The consistent packing patterns between the homodimer and palindromic DNA sequence demonstrates that KS analysis of bZIP-DNA complexes are consistent with sequence composition. Furthermore, the KS analysis identified a regularity of packing for both half-site interfaces in terms of knob positions and the patterns of sockets as well as the composition of amino acids and nucleotide bases. The symmetry and regularity indicate the role of multi-body packing interactions in protein-DNA recognition. The modified KS analysis of both DNA nucleotide knobs packing into protein α-helical sockets and protein α-helical residue knobs packing into DNA sockets provides a framework to relate amino acid and nucleotide packing composition preferences to recognition specificity. Using this approach to characterize the individual half-sites from the 43 crystal structures of bZIP-DNA complexes taken from the PDB database (10-12,15,17,18,29,35-57), a model of bZIP α-helical recognition of DNA was developed that identifies common mutual packing patterns between the two biomolecules with clear regions of affinity and recognition specificity between the protein and DNA duplex.

CHAPTER 2: MATERIALS AND METHODS

Selection, Preparation, and Knob-Socket Analysis of Crystal bZIP-DNA Structure Complexes

The PDB database was scoured through for all crystal structure complexes of bZIP or bZIP-like proteins and DNA duplexes, resulting in a total of 43 crystal structures: 1a02 (37), 1dgc (36), 1dh3 (35), 1fos (18), 1gd2 (10), 1gtw (44), 1gu4 (39), 1gu5 (42), 1h88 (39), 1h89 (39), 1h8a (39), 1hjb (38), 1io4 (38), 1jnm (40), 1nwq (41), 1s9k (45), 1skn (29), 1t2k (43), 1ysa (17), 2c9l (46), 2c9n (46), 2dgc (12), 2e42 (48), 2e43 (48), 2h7h (47), 2wt7 (52), 2wty (52), 3a5t(50), 4auw (49), 4eot (51), 5szx (53), 5t01 (53), 5vpe (54), 5vpf (54), 6mg1 (15), 6mg2 (15), 6mg3 (15), 7aw9 (11), 7aw7 (11), 7nx5 (56), 7l4v (55), 7x5f (57), 7x5e (57), and 7x5g (57). Supplementary Table S1 lists each PDB's protein sequence and DNA recognition sequence. Each PDB file was processed into an index file in which all residues and nucleotides were denoted by single-letter code, number, chain, and secondary structure for all chains. The standardized crystal structure file was input into the Knob-Socket program, outputting an atomic contact file and packing clique file.

Mapping Knob-Socket Packing on Helical and Duplex Lattices

For all bZIP-DNA complexes analyzed utilizing the Knob-Socket model, all Knob-Socket packing between bZIP protein and DNA duplexes were mapped onto the respective lattices. The helical lattices display the four- and five-body cliques that a DNA knob packs into a protein socket or pocket. Similarly, the DNA duplex lattices map all four- and five-body cliques that contain a protein knob packing into a DNA socket or pocket. All maps were collated and compared for common packing trends, especially around the invariant Asn and Arg. Maps are included in the Appendix as Supplemental Figure S1.

Standardization of bZIP-DNA Half-Site Packing

The bZIP proteins were standardized based on the invariant asparagine and arginine, with the two residues renumbered as -4 and 4, respectively; the rest of the residues in the chain were labeled relative to the invariant protein residues. The DNA sequence information was renumbered for each protein chain, allowing for a different numerical translation based on the specific protein chain packing to a half-site. The DNA sequence numbering was modified to consider the central base pair(s) as the first base pair, increasing with each base pair moving away from the center of the DNA sequence. Outside of the main four base pairs, the base pairs were numbered +/- depending on the flanking region: the base pairs before bp 1 were numbered with –, while the pairs after bp 4 were numbered with +. This numbering system follows the numbering convention seen in prior studies of bZIP-DNA complexes. The DNA strand that follows the numbering scheme in a 5' to 3' manner is considered the positive strand ("p") for that interface between a bZIP helix and the half-site, and the other strand is labeled as the negative strand ("n"). Standardizing both the protein and DNA sequence for each packing interface between a bZIP monomer and its respective half-site permits direct comparison of packing interactions and patterns.

Packing Localization of Standardized DNA Nucleotides

The proposed packing regions (P1, P2, N3, and N4) are scanned in each half-site for any DNA knob that packs into a region. All DNA knobs packing into a region are then grouped based on the region it packs, independent of sequence or composition. The resulting heatmaps correlate packing of the standardized nucleotides with specific packing regions to further understand localized packing between bZIP helices and the respective DNA half-site.

Packing of DNA Knobs Involving Protein Sequences and Nucleotide Compositions

All bZIP sequences and packing regions were grouped into a bar plot to compare frequency of sequences found at each region. For each half-site, the packing of the respective nucleotide at each position was recorded based on packing mechanism (e.g., 2p packing into P2, etc.). Base packing DNA knobs were labelled as “Base,” Backbone packing DNA knobs were labelled as “Back,” and if the respective DNA knob did not pack into the corresponding region, the knob was labelled as “Non-Packing.” DNA knobs that pack both base and backbone were considered “Base” as well. The results from all 85 half-sites were collected and grouped based on region, pocket region sequence, and packing mechanism. For each region, three LOGOS plots were created to compare relative packing mechanisms between specific nucleotides and packing region sequence. The sum of nucleotide packing patterns differs based on the region and sequence, since the bar plot displays an uneven occurrence of various packing region sequences.

Analysis of Pairwise Contacts of bZIP-DNA Packing Surfaces

With all bZIP residues and DNA nucleotides standardized, the individual atom contact files were filtered and condensed to produce individual pairwise contacts for a single half-site. Packing between the two macromolecules required selecting pairwise contacts involving both a protein residue and DNA nucleotide, which was then filtered for contacts between residues and DNA bases. The majority of protein contact of DNA bases involves bp 1 to 4. The protein to DNA base pairwise contacts were used to create heatmaps for each base from bp 1 to 4, resulting in eight heatmaps. The heatmaps correlate nucleotide composition to protein pairwise contacts to determine if specific nucleotides have higher packing preferences with varying protein residues.

Frequency Collation of bZIP-DNA Cliques

Each packing clique file was analyzed and parsed for cliques involving both protein and DNA, excluding cliques consisting solely of either bZIP proteins or DNA. The filtered packing cliques were collated and grouped according to clique size and type. Each RPC size was plotted in a stacked bar plot to visualize common RPC sizes, with each column further partitioned based on socket type count; recurrent socket types were compared by stacking the RPC types per RPC size column, with the infrequent socket types further combined under the miscellaneous label "Other." The bar plot was organized and created in R.

Deconstruction of Standardized bZIP-DNA Cliques

As the ratio between protein residues and DNA nucleotides differ per socket size and type, RPC types were modified based on the macromolecule with less contributing residues to consider the clique as interactions between multiple knobs and a residue group, with each knob packing considered independent from one another. Two broad categories were created: protein knobs packing into DNA nucleotide groups and DNA knobs packing into protein residue groups. Therefore, a five-body RPC between two DNA and three protein residues would be considered two four-body packing groups between each DNA knob and the corresponding three protein residues. If the ratio between protein and DNA for a clique equated to one, the clique was included in both categories, maintaining the data to both broader categories. After all half-site packing cliques were modified and subdivided, all duplicates were removed from the dataset to prevent under- and overcounting. Furthermore, smaller packing groups are subsumed into larger cliques if all vertices in a smaller group are found in a larger packing group. The packing groups were then standardized: protein residues based on the invariant asparagine and arginine, and DNA nucleotides based on the bZIP chain packing into the half-site. This process modifies all collated cliques equally, modifying the filtered cliques to analyze the packing patterns of each knob involved in the bZIP-DNA packing interface. All bZIP-DNA half-sites were standardized and analyzed, resulting in 85 half-sites from the 43 crystal structures. The output packing groups were filtered based on knob type, creating two datasets: DNA knob into protein helix and protein knob into DNA duplex.

Packing Analysis of DNA Knobs into Protein Helices

The DNA knob data was grouped by the standardized knob and its packed protein residue group, simplifying the DNA knob by number, chain, and type (e.g., base, backbone, etc.) and the packing group by residue position. The resulting frequency table provides insight to frequently packed groups or sockets on the protein helix occupied by DNA knobs, where the consistently filled sockets were mapped onto two separate composite protein helix lattices depending on whether the DNA knob was a sugar-phosphate backbone or a nitrogenous base. The consistent DNA knobs were found based on packing frequency and average packing area, where the common packing areas were bordered with a color corresponding to the DNA knob. The resulting two lattices compare DNA base packing patterns against DNA backbone packing patterns. The individual DNA knob interaction data were collated into Supplementary Figure S2, displaying all the DNA knobs found to pack with at least two protein residues.

Sorting through the packing data emphasized the importance of base 1p, 2p, 3n, and 4n; by combining the packing areas of these four bases, the recognition core was formed. The quadripartite core region consists of four smaller regions corresponding to a respective DNA base (e.g., 1p packs into P1, whereas 3n packs into N3), with regions P1 and N3 expanded into a 2:2 pocket to resemble the recognition area of P2 and N4. The packing data for each half-site was then filtered based on which DNA knobs interact with the nine-residue recognition core and the peripheral area in every bZIP-DNA half-site. By grouping the packing data of all 85 half-sites, the frequency of a DNA knob packing into the recognition core or peripheral border was compared, limited to a maximum value of 85. The data was then mapped to a modified DNA duplex with the DNA knobs represented by circles, with the packing frequency of the knob represented by a circle's shade and number. Two modified DNA lattices display knob packing frequencies into either the protein helix recognition core or the peripheral region outside the core.

For each half-site, the amino acid sequence for the four smaller regions (P1, P2, etc.) and the nucleotide sequence for the four primary DNA nucleotide positions were recorded to count the times a region is composed of a specific amino acid sequence while the corresponding DNA position is occupied by a specific nucleotide. Collating this data allows for comparison between packing frequency and maximum possible packing frequency between a DNA base knob and its expected packing region depending on sequence data. The half-site data was then analyzed for cliques involving the DNA nucleotide position; each base position is considered packed into its corresponding region if it contacts at least three residues out of the four-residue region. Packing frequency was then compared to the maximum possible packing frequency to determine possible factors determining a region's packing preference for specific nitrogenous bases.

Packing Analysis of Protein Knobs into DNA Duplexes

The protein knob data was grouped by the standardized knob and its packed DNA socket or group, simplifying the protein knob by number and the DNA contact group location and duplex lane. For each half-site, each protein residue number was scanned for which DNA lane with which it packs. The resulting data table was mapped onto the modified protein lattice to correlate protein residues to DNA lanes; each protein number is represented by a circle of varying shade and frequency number, displaying the most common protein knobs per lane. Three modified protein lattices are mapped for the PL, ML, and NL. Based on the results, the protein knob data is split based on *i+4* ridges. The numbering for the ridges is based on the invariant residues, which occupy an *i+4* ridge, with the other three *i+4* ridges named based on the numbering: either a residue before (-1) or after (+1) the invariant asparagine and arginine. As the last ridge points away from the DNA duplex and into the solvent, these residues occupy the solvent ridge. Similar to the DNA knob data, the protein knob data was then grouped based on protein knob number and DNA contact socket, providing insight to commonly packed areas on the duplex. The resulting frequency data was then mapped onto three DNA duplexes corresponding to the -1, 0, and +1 ridges. The duplexes show the frequently packed sockets and areas along with the residue knob that packs it, disregarding sequence data. The common packing patterns for each knob, based on residue number, are bordered with a color corresponding to the bZIP knob. Based on the results and the common residue contacting the four primary DNA base pairs, the protein knob data was further filtered to only include packing interactions with protein residue 0 as the knob.

Scanning the standardized amino acid and DNA sequence for each half-site accounted for all residues found at residue number 0, along with the DNA sequence between adjacent base pairs (e.g., base pair 1 to 2, 2 to 3, and 3 to 4), considering all possible DNA regions in the ML for the protein knob to contact per amino acid occupying position 0. Analyzing each half-site protein knob data for residue 0, the protein knob would be considered to pack into an adjacent base pair region when the residue contacts three of the four nitrogenous bases that make up the ML region. The collated data from all half-sites were compared to the possible combinations between residue 0 and DNA adjacent base pair region; by comparing packing frequency to maximum possible occupancy, possible specific packing patterns may elucidate protein knob preferences within the primary four base pairs of a half-site.

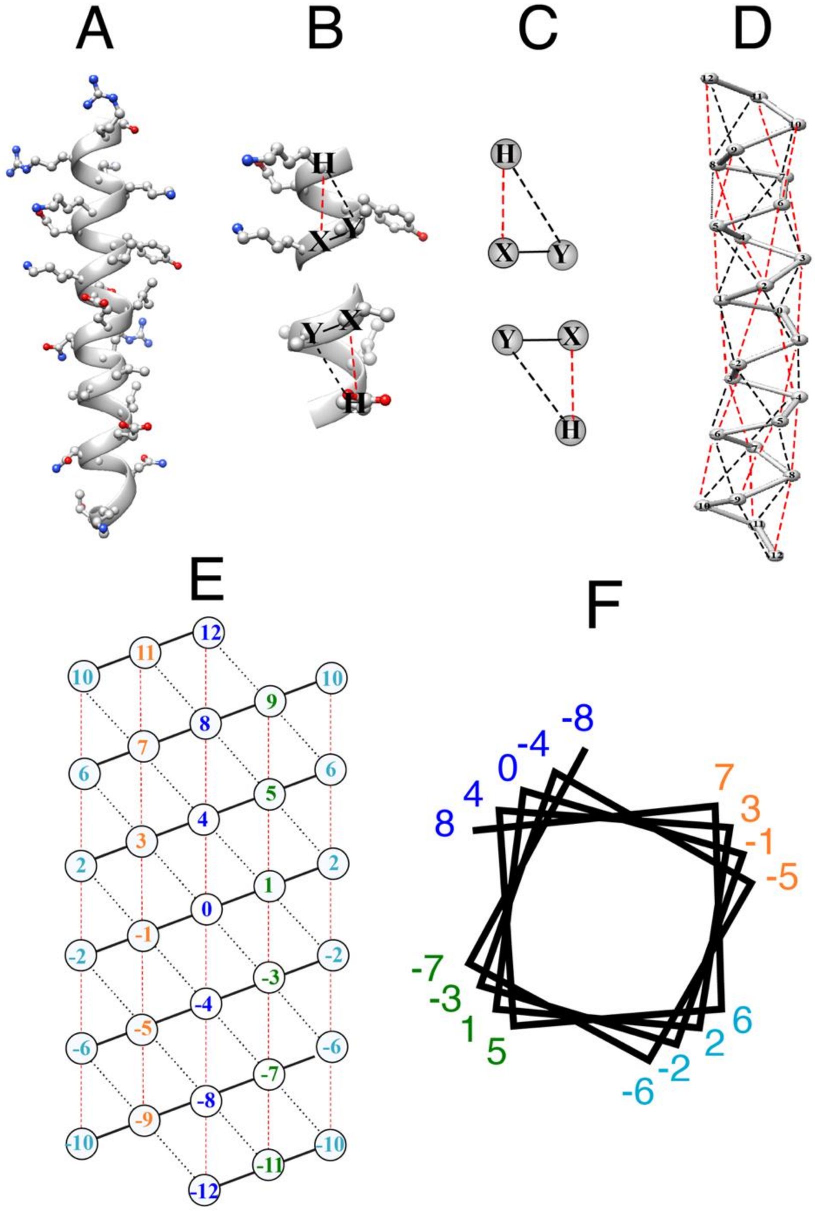
All possible sequences for the four base-recognition regions were collected based on the invariant asparagine and arginine, as well as the middle residue 0. The middle residue in the collated crystal structures either had an alanine, valine, glutamine, and serine. Since all four regions had an invariant residue and the middle residue in its respective pocket region, each region had four separate heatmaps for the four possible middle residues. The x- and y-axes then represented all possible permutations of the remaining two variable residue positions in a region. P1 consists of -3, 0, 1, and 4, where 0 and 4 are set and -3 and 1 are variable. When P1 has the "template" sequence x:Ay:R, residue 4 is R and residue 0 is A. Amino acids "x" and "y" represent -3 and 1, respectively, along with the corresponding x- and y-axes. The heatmap data itself is calculated from the collated packing propensity data gathered from across the PDB. A socket can be free or filled, resulting in a percentage based on the total found socket in the databank. Furthermore, a socket has a "Total" value depending on the occurrence the socket is found, either free or filled. The product of the "Total" and filled value is used to determine possible stability and packing propensities. Each pocket sequence was given a packing propensity based on the Knob-Socket data, which was the sum of the sockets making up the pocket. These propensities were mapped and visualized in the heatmap. The heatmap cells corresponding to all region sequences found in the bZIP-DNA dataset were bordered and colored. Each box was colored according to the most common base found in the region when the region sequence matches the corresponding sequence.

CHAPTER 3: RESULTS AND DISCUSSION

KS model of protein α-helix

Figure 3

Description of the α-helical Knob-Socket Topology Lattice Map



Note. Protein structure images were created in Chimera (59). A) Protein α-helix structure with the sidechains is displayed. Residues 16-40 are shown from the bZIP Chain D of the 1dh3 crystal

(Continued, Figure 3) structure (35). B) An α-helix binding surface can be simplified based on repetitive pattern of three residue sockets defined by a clique of residues X, Y, and H. There are two types of arrangements of these residues in α-helices: the high-H (upper image) and low-H (lower image) 2:1 sockets. High-H indicates that the H residue is highest in the sequence of three residues **X**, **Y**, and **H**, while low-H indicates that the H residue is lowest. Individually, these sockets are categorized as "free", capable of interacting with another residue. "Filled" sockets pack with another residue from a different secondary structure, also known as a knob, resulting in a tetrahedral four-body packing clique. Both socket types are examples of relative packing cliques, which define mutual contacts between all residues in a subset. The lines represent the varied types of interactions between the residues: the black dashed lines represent van der Waals packing only, solid black lines represent covalent bonds and packing, and dashed red lines represent hydrogen bonds and packing. C) 2D representations of the high-H and low-H 2:1 sockets, with the side chains of residues **X**, **Y**, and **H** protruding out of the page. D) The 2:1 sockets mapped onto the simplified α-helical backbone. In this manner, these free sockets define the discrete binding surface or areas of packing interaction on an **α**-helix. E) A 2D representation of an α-helix mapped as 2:1 sockets, where the cylindrical α-helix is "cut" along an *i* to *i+4* ridge and laid flat. The residues on the edges repeat on each side to account for the cylindrical nature of an α-helix. In the bZIP helix, the cut occurs on the *i+4* solvent ridge (see next section) so that this solvent ridge is on the edge, which places the primary packing *i+4* ridge 0 in the center of the lattice. F) Protein helical wheel representing the *i* to *i+4* ridges, clarifying that the α-helix can be "cut" along a ridge to result in a 2D protein lattice. In addition, the wheel shows the four sides of an α-helix that interact with DNA major groove, and each *i+4* is identified in relation to this packing. Dark blue represents residues that centrally pack into the DNA bases and represents the

(Continued, Figure 3) 0 ridge. Orange and green are residues that interact with the positive and negative strands of the DNA recognition region and so are named the +1 and -1 ridges, respectively. Light blue indicates residues that point primarily out into the solvent, away from the major groove and is named the solvent ridge.

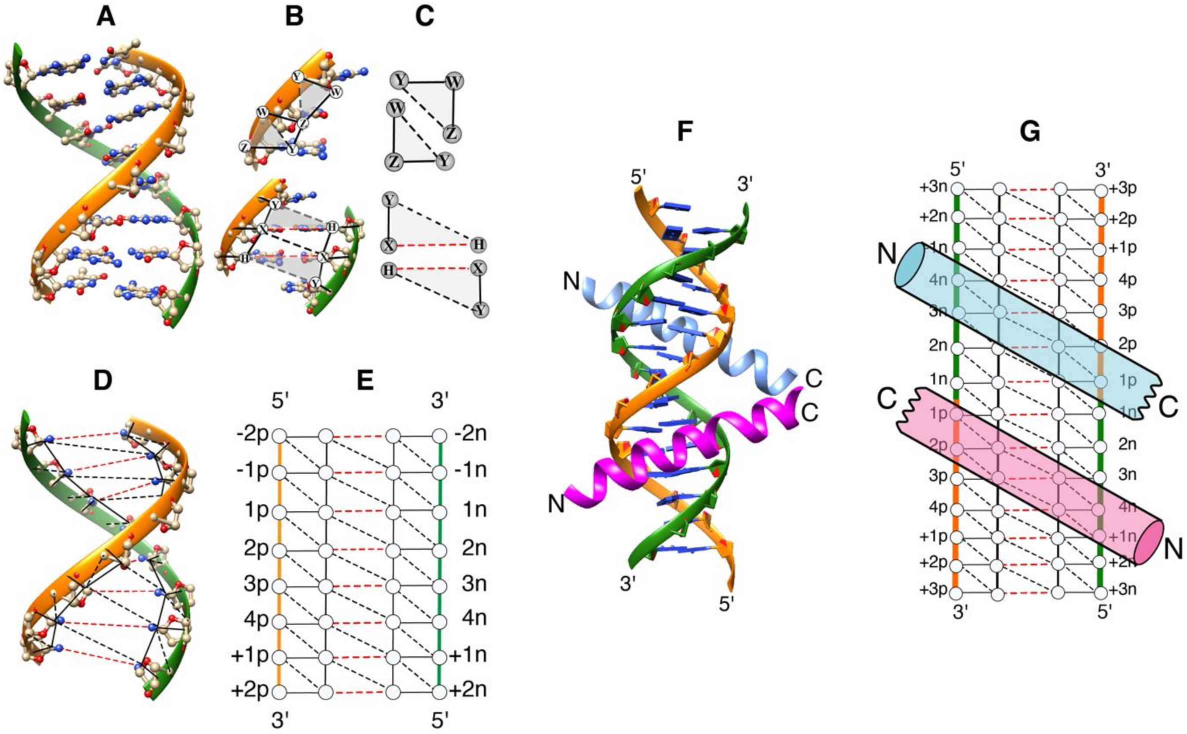
An α-helix is a protein secondary structure formed through a consecutive series of repetitive hydrogen bonds between residues *i* and *i+4* (Figure 3A). Figures 3B-E shows how the KS model provides a simplified topological binding surface map of an α-helix (4). When calculating residue-to-residue contacts within an α-helix, the α-helix binding surface can be represented as a recurring pattern of two types of discrete three-residue sockets. Figure 3B displays the two examples of the standard 2:1 three-residue socket. All three residues are from the same local secondary structure, but each of the three pairs of residues shares a distinct set of interactions as indicated by the line between them. While all three residues contact each other through van der Waals packing, residues **X** and **Y** are adjacent protein residues that also share a peptide bond, which is indicated by a solid black line. Likewise, in addition to van der Waals packing, residues **H** and **X** share the characteristic α-helical hydrogen bond, which is shown by the dashed red line, and are therefore four residues apart. Unlike the previous two pairs, the **Y** and **H** residues only share van der Waals interactions, which is indicated by a dashed black line, and are always three residues apart. For the 2:1 socket nomenclature, adjacent residues **XY** are represented by the "2", while the distant residue **H** is represented by the "1" and separation by a hydrogen bond by the ":" The primary difference between the two types of 2:1 sockets in α-helices is the sequence position of the **X** residue in relationship to the **H** residue. In the lower 2:1 socket of Figure 3B, the **X** residue is highest in sequence, so that **Y** is at position *i*-1 and **H** is lowest at *i*-4. Conversely, in the upper 2:1 socket of Figure 3B, the **X** residue is lowest in sequence, so that **Y** is at position *i*+1 and **H** is highest at *i*+4. With these interaction and position relationships between the socket residues established, the two types of α-helical sockets can simply be abstracted as triangular surfaces, which is shown in Figure 3C. These sockets are discrete binding surfaces which can pack a fourth knob residue or nucleotide in tertiary or quaternary packing. As shown by the backbone trace in Figure 3D, the intrahelical packing of an α-helix is a consistent, repetitive alternating pattern of the two types of sockets. Therefore, the potential binding surface of an α-helix can be clearly depicted in two dimensions by building up the alternating triangular surfaces of Figure 3C. The result is a simple two-dimensional lattice shown in Figure 3E for 25 residues, where each number represents a protein residue in the α-helix. The solid lines represent packing and the covalent *i* to *i*+1 ridge; the broken black lines represent the packing of the *i* to *i*+3 ridge; and the broken red lines represent packing and hydrogen bonding of the *i* to *i*+4 ridge. To account for the cylindrical nature of an α-helix, the residue numbers repeat at the edges. In effect, this α-helical lattice essentially converts the three-dimensional α-helix by cutting down one of the *i* to *i+4* ridges of a helical wheel (Figure 3F) and to lay the surface flat in two-dimensions. In Figure 3E, the protein helix is cut along the *i* to *i+4* ridge consisting of residues -10, -6, -2, 2, 6, and 10. Flattening the three-dimensional α-helix into a plane simplifies visualization and mapping of packing by knobs residues or nucleotides that pack into the discrete areas defined by the three residue sockets, as shown by the nucleotide packing in the right side of Figure 1B.

Because of its simplicity, the α-helical lattice in Figure 3E can be ordered to highlight any particular set of residues as a feature of α-helical structure. Throughout the analysis of crystal structures, every bZIP protein contained a highly conserved Asn and Arg separated by eight residues, and a relatively conserved middle residue directly between the two along the *i+4* ridge. This conservation standardizes each monomer’s residues based on the middle residue as residue 0 and the invariant Asn at -4 and Arg at 4. Therefore, the numbering scheme shown in Figure 3E will be used as a reference for residue positions to compare between the different bZIPs. For orientation based on packing, the middle *i+4* ridge packs deepest in the major groove of the DNA duplex and is called *i+4* ridge 0. The next ridge over to the right of the 0 ridge packs into the positive strand and is called the +1 ridge, while the previous ridge to the right packs the negative strand is called the -1 ridge. The edge *i+4* ridge that repeats on both sides is the solvent ridge and faces away from the major groove.

Modified KS model of the DNA double helix

Figure 4

Explanation of the Knob-Socket Representation of the DNA Duplex Helix Packing Lattice and Orientation of Two Protein Binding Half-site Regions



Note. Example mapping and crystal structures were constructed using 1dh3 (35). Structures imaged using Chimera (59). A) The crystal structure of a DNA duplex with nucleotides shown with the ball-and-stick model for both positive (orange) and negative (green) strands. B) To differentiate between non-specific backbone interactions and specific base interactions, the DNA backbone and DNA nitrogenous base are considered separate vertices in the modified Knob-Socket analysis. Therefore, an outside backbone vertex involves any atom from the phosphate-sugar backbone, whereas the inner nucleotide base vertices include interactions with any atom on a nitrogenous base. Sockets can form between the backbone and base within a DNA strand as well as between base pairs in a duplex, where vertices are designated **X**, **Y** and **H**. Again, these

(Continued, Figure 4) three residue sockets are cliques consisting of vertices that all pack with each other, and lines indicate type(s) of interactions. The dashed black lines indicate only packing interactions, while the solid black lines identify vertices that are packed and covalently bonded with each other. The dashed red lines are residue pairs that pack and hydrogen bond with each other. C) 2D representation of the DNA sockets with the major groove protruding outward. The two primary types of backbone sockets exclusively consist of black solid and dashed lines or only packing and covalent interactions. Sockets for bases consist of all three types. D) Sockets superimposed onto the DNA duplex show three columns or lanes where sockets form: the positive lane (PL) involves the backbone and base vertices of the positive strand, the negative lane (NL) exists similarly within the negative strand, and the middle lane (ML) involves bases from both strands in a DNA duplex. E) 2D DNA duplex KS packing lattice represents the regular pattern of sockets belonging to the major groove, which faces out of the plane. The duplex lattice visualizes the duplex unwound into 2D, creating three lanes where sockets may be filled, and produces a clear definition of the surfaces and areas that can be packed into by quaternary interactions. Vertices of the backbone and base are circles, and their interactions are indicated by the type of line connecting them, following the description in part B. The primary difference is that the solid black lines between residues on the same strand are colored solid orange or solid green to indicate the positive and negative strands, respectively. Numbering is centered around the binding nucleotides of the half-site referenced to the positive strand, where a “p” indicates positive strand and “n” negative strand. In the lattice, the numbering starts two base pairs before the recognized four nucleotide half-site sequence and ends two nucleotides after the half-site sequence. The fourth residue of the recognition sequence is referred to as 4p for the positive strand nucleotide and 4n of the negative strand. Interactions with the backbone at position 4 are

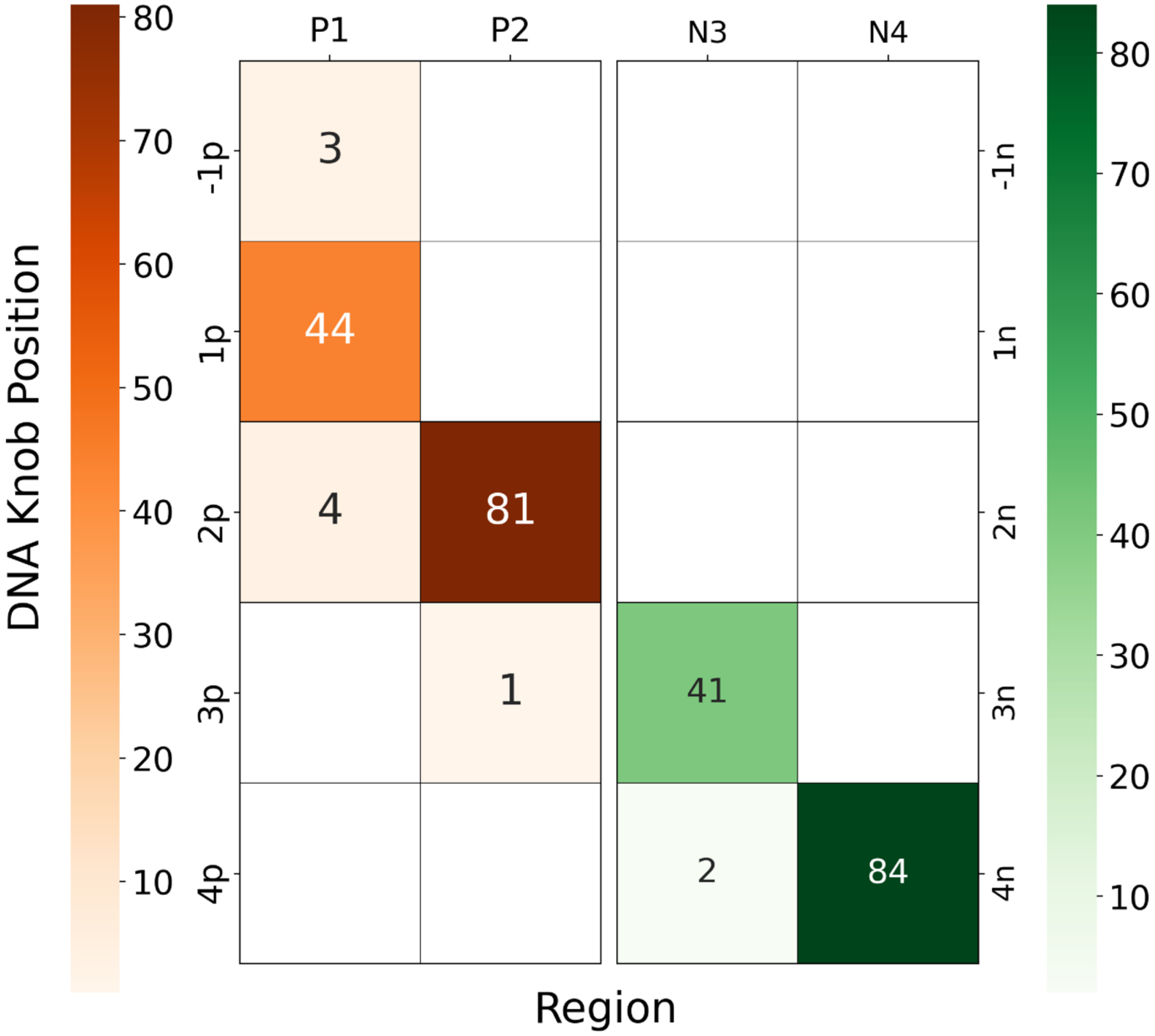
(Continued, Figure 4) referred to as 4pB for the positive strand and 4nB for the negative strand. F) Packing interface between the bZIP dimer and the cognate DNA sequence in duplex is shown. For the two bZIP α-helices in light blue and magenta, correspondingly, only the DNA binding region is shown as a ribbon diagram without their respective leucine dimerization regions. The DNA duplex shows fourteen nucleotide pairs: the eight base pair cognate sequence and an additional three flanking nucleotide pairs on each side. Each α-helix packs into the major groove of a DNA half-site, and both extend outward from the center of the DNA site. G) The KS DNA duplex lattice of a bZIP-DNA complex, where both bZIP proteins pack into the major groove of the duplex. The overlaying protein helices reveal the orientation of the bZIP helices if the DNA duplex was unwound. In contrast to the depiction in part F where the twisting of the DNA shows the dimerization of the two α-helices' C-terminal zipper regions, topologically, the α-helices on the 2D lattice run antiparallel to each other. In addition, each strand flips from between positive and negative strand relative to the orientation of the packing α-helix, as expected with a palindromic recognition sequence.

Duplex DNA (Figure 4A) forms sockets differently than proteins due to the distinct molecular structure of paired nitrogenous bases arrayed between two strands of sugar-phosphate backbones. The double helix also has two sides with the major and minor grooves. Because the bZIP-DNA interactions occur only within the major groove, the KS modeling in this work only considers the major groove. Within proteins, each residue is considered a vertex of a socket as well as a potential packing knob. In duplex DNA, the packing interactions dictate that the sugar-phosphate backbone and nitrogenous base of a single nucleotide should be treated separately as socket vertices and potential packing knobs. As shown in top of Figure 4B, three vertex DNA sockets form between backbone and bases of the same strand classified as a backbone socket, where all the vertices share van der Waals packing interactions and two pairs of vertices are also covalently bonded. A base socket is shown at the bottom of Figure 4B, where the duplex structure forms a three-vertex socket between the pairs of nitrogenous bases. Between the three vertex pairs of the base socket, more standard KS socket interactions occur: only packing, packing with hydrogen bonding, and packing with covalent bonding. Like with protein α-helices, each of these backbone and base sockets are found in two types within duplex DNA. Since these sockets represent interaction surfaces, they can be abstracted from their three-dimensional structure into two-dimensional triangles shown in Figure 4C. Overlaying both types of sockets onto the 3D DNA duplex in Figure 4D illustrates the intramolecular packing surface of the double helix in the major groove. Similarly to how protein α-helices are comprised of repetitive 2:1 sockets (4), the DNA duplex can be represented by a lattice of repetitive sockets that indicates the potential surfaces for protein knobs to pack and bind. Shown for a single half-site, the DNA duplex's binding surface can be easily flattened in two-dimensions (Figure 4E), which clearly shows that the major groove consists of three interaction lanes formed from DNA sockets. The positive orange and negative green strand each form a column of backbone sockets and will be referred to as the positive lane (PL) and negative lane (NL), respectively. Between those two lanes are base sockets involving nitrogenous bases from proximal base pairs that create a middle lane (ML). By figuratively unwinding the DNA duplex, the 3D structure represented as the DNA duplex lattice shown in Figure 4E exhibits the general direction of packing resulting from the duplex twist. Separating the backbone and base sockets allows an explicit classification of packing interactions that contribute only to affinity (backbone sockets) and those that also contribute to specificity (base sockets). Because the phosphodiester backbone is constant for all nucleotides in the DNA strand, the PL and NL formed by backbone sockets involves primarily non-specific affinity interactions and minimal interactions with specific bases. In comparison, packing in the ML formed by base sockets involves contacts that can differentiate between nucleotides and therefore contributes significantly to bZIP-DNA specificity. Therefore, the DNA lattice of sockets shown in Figure 4E represents the duplex with the major groove protruding from the plane, where bZIP residues pack into the major groove of the duplex. This modification of the KS model types of sockets allows a lattice representation of a DNA duplex's major groove in Figure 4E that can now be analyzed with bound protein α-helices from a bZIP dimer. Figures 4F and 4G help orient the depiction of the DNA lattice with respect to two the protein bZIP α-helices, since the flattening out of the DNA duplex twist changes the topological orientation of the protein α-helices. As shown by the crystal structure of the bZIP-DNA complex (35) in Figure 4F, a bZIP dimer packs into the major groove of the DNA cognate site, with each monomer packing into a respective half-site from the center out. Structurally, the twist in the duplex allows for parallel dimerization of the α-helix monomers and therefore, the C-termini to approach each other from the same side of the DNA. As shown in Figure 4G, unraveling of the DNA duplex into a 2D lattice with the major groove pointing out of the plane topologically positions the helices with separated C-termini and anti-parallel to each other. Numbering conventions for the nucleotide bases are also shown in Figure 4F. Half sites conform to 1234 convention, where DNA residues are numbered based on nucleotides pairs in relation to the central pair(s). The cognate site is usually palindromic or pseudo-palindromic and numbered from the central base pair(s) outward. Common palindromes consist of eight nucleotides and are represented in a 4321 | 1234 fashion with each half-site interacting with separate bZIP α-helical monomers. For pseudo-palindromic DNA cognate sites that are 7-bp long, the central base pair is considered position 1 for both half-sites. The 4321 | 1234 numbering scheme flips the positive strand between the two half-sites, which can be seen in the change in orange and green strands on the 2D DNA lattice under the two packing α-helices in Figure 4G. Location is indicated through the label of “p” for positive chain (Figure 4G, orange strands) or “n” for negative chain (Figure 4G, green strands). The positive strand increases from 5’ to 3’ as the base pair number increases, while the opposite goes for the negative strand. In Figure 4G, the protein knobs are packing into the lattice of the major groove. The relative orientation of the two protein α-helices superimposed over the untwisted DNA duplex lattice show that each α-helix monomer packs anti-parallel to each other with the C-termini distal from each other in contrast to the three-dimensional structure. However, the orientation of each α-helix is the same with respect to their half-site nucleotides. Residues proximal to the C-terminus interact with the bases towards the center of the DNA cognate site. As the protein residues continue packing outward from the central base pairs, the protein knobs are more proximal to the N-terminus. This common packing trend supports the standardization of all DNA half-sites to view packing from the central base pairs onward. Using this numbering scheme and labeling the positive and negative strands according to standard conventions allows a consistent comparison between all bZIP-DNA half-sites and extraction of the common features of recognition and specificity between protein α-helices and duplex DNA.

DNA Packing Regions on the bZIP Helices

Figure 5

Heatmaps of DNA Nucleotide Knob Packing into the Four Proposed Protein Packing Regions



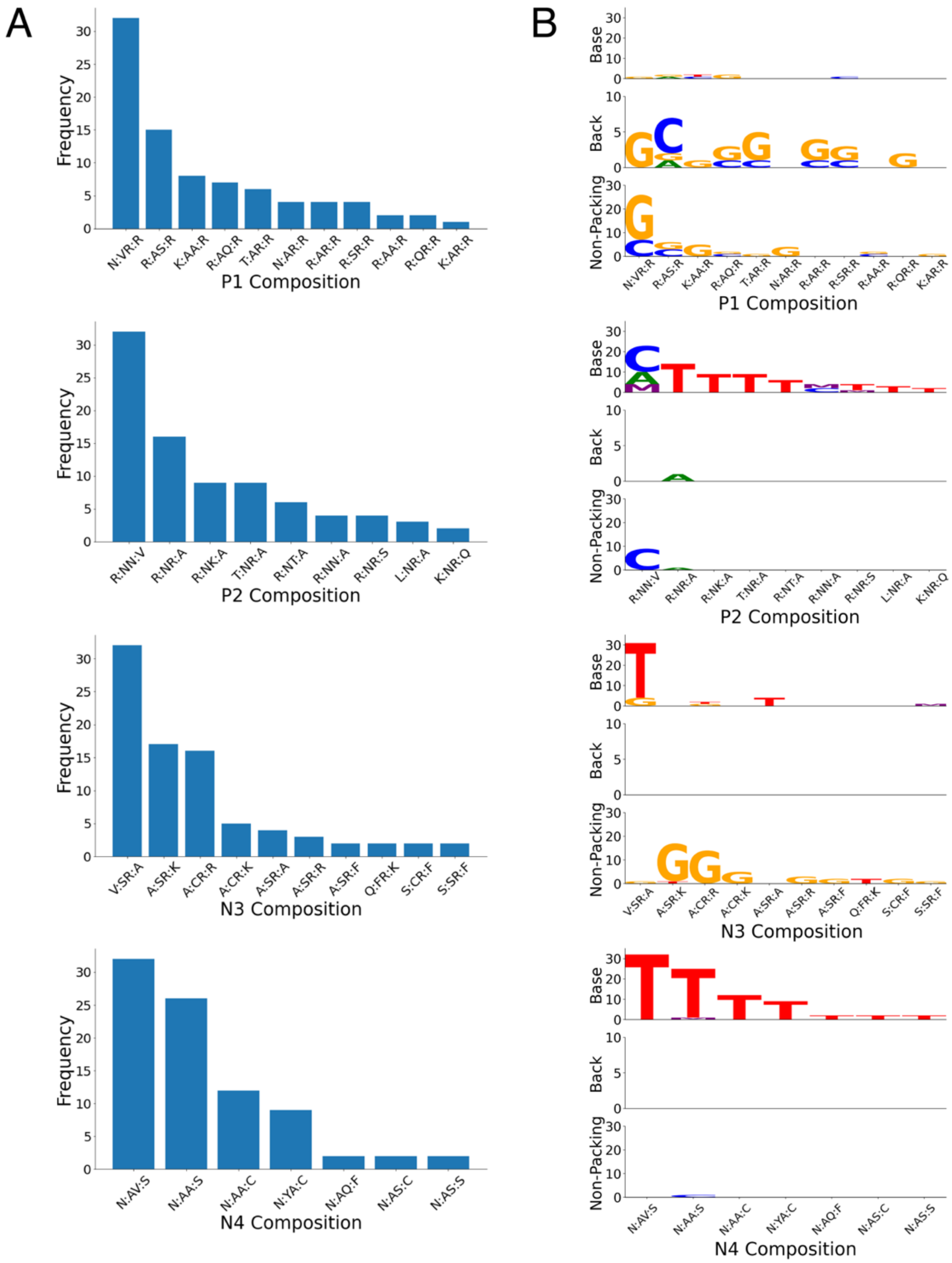
Note. The x-axis displays the four helical pocket regions: P1, P2, N3, and N4; P1 and P2 are colored in shades of orange while N3 and N4 are colored in shades of green, indicating the DNA strand that packs most often in said regions. The y-axis displays all DNA nucleotide positions that pack into these four pocket regions. Both the pocket sequence and nucleotide type are not considered in these heatmaps, emphasizing only the standardized protein residues and DNA

(Continued, Figure 5) nucleotides. The highest packing value will be 85, as a total of 85 half-sites were analyzed from the 43 bZIP-DNA complex structures. P2 and N4 pack consistently with 2p and 4n, respectively. There is little crossover where a different nucleotide position packs into P2, N3, and N4. P1 and P2 shows less consistent DNA knob packing.

After mapping out the knob-socket packing motifs onto the protein helical lattice, the common packing scheme centralizes four DNA knobs around the invariant asparagine and arginine. 1p packs with -3, 0, 1, and 4; 2p packs with -7, -4, -2, and 0; 3n packs with 0, 3, 4, and 7; 4n packs with -4, -1, 0, 3. The packing maps are included in the Supplemental Figures section, displaying the common packing patterns between bZIP helices and the DNA duplex. The packing patterns for all four DNA knobs involves a pocket region, where a knob may fill in a socket (2:1), elongated socket (:3), or a pocket (2(2:1)). These regions were classified as regions P1, P2, N3, and N4, which packs their respective bases 1p, 2p, 3n and 4n. All four regions surround the conserved bZIP helices, sharing edges and residues between the packing regions; the recognition regions, therefore, span nine residues total. With this definition of the four regions, all half-sites were then collated to compare packing locations for the four primary standardized DNA nucleotides across the four distinct helical regions. Figure 5 represents these packing preferences as a heatmap, where the four regions were commonly packed by a single DNA nucleotide. Region P2 has slight packing overlap with nucleotide 2p and 3p, as region N3 has with 3n and 4n. Similarly, P1 has extended overlap with nucleotides -1p, 1p, and 2p. Despite slight overlaps, each region is primarily packed by their respective DNA knobs. Furthermore, P1 and N3 has less packing compared to P2 and N4, indicating possible differences in packing conservation and roles in DNA-bZIP specificity.

Figure 6

Analysis of Packing Mechanisms of DNA Knobs into Region Pockets



(Continued, Figure 6) *Note.* The plots are dependent on nucleotide and pocket composition and sequence. A) Bar plots for each region compare the occurrences of each displayed pocket region composition found throughout the bZIP-DNA half-sites. The number of differing pocket compositions vary for each protein region, with P1 being the most variable and N4 being the least variable. These bar plots assist in understanding Figure 6B, since the small crystal structure dataset results in a skewed composition distribution. B) LOGOS plots for each protein region compares the packing mechanism of a DNA nucleotide knob with specific region compositions. For each region, there are three LOGOS plots, labelled as “Base,” “Back,” and “Non-Packing,” for instances where a DNA knob either packs the base moiety, sugar-phosphate backbone, or neither portions. There were several instances of 5-methylcytosine occurring in the half-site, as is represented by “M” in the LOGOS plots.

The localization of DNA knob packing verifies specific areas on the protein helix that DNA nucleotides pack. However, the packing between 1p and P1, as well as 3n and N3, indicates a lower packing propensity between the DNA knob and the protein pocket. Additionally, Figure 5 only displays the regions where a DNA knob packs into, independent of the specific DNA nucleotide and the protein region sequence. While multiple bZIP proteins may differ in amino acid sequence, consistent and localizes packing regions implicate that conserved pocket sequences for a specific region should maintain DNA nucleotide packing specificity, regardless of variances beyond said protein packing region. Therefore, each half-site was then analyzed for the type or mechanism of packing between each nucleotide position and the respective base. For clarification, a nucleotide can pack the backbone or base portion of the DNA knob into its corresponding protein packing region and is classified as “Base” or “Backbone” packing. If the DNA knob packs both backbone and base, it is classified as “Base” packing as well; if a DNA knob does not pack either backbone or base, it is labelled as a “Non-Packing” base. This labelling scheme was applied across all half-sites and corresponding packing interfaces between a DNA nucleotide and the resulting pocket region. The resulting data will examine the packing mechanism between a nucleotide position and its protein region in an attempt to elucidate the sequence dependencies of packing between a mononucleotide and the pocket region sequence.

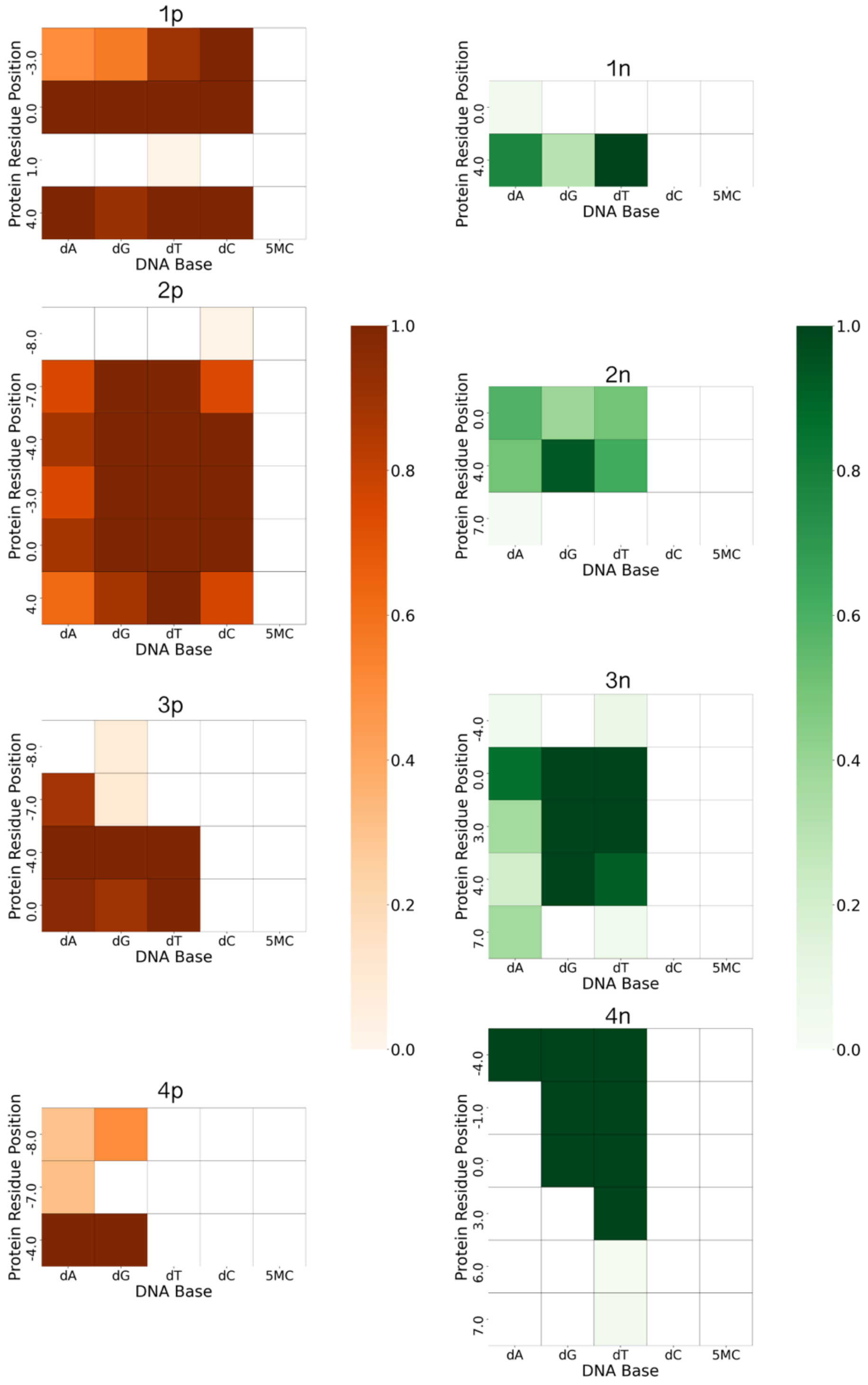
There are several issues to address prior to the analysis. For a complete analysis of packing between a specific protein region and its respective nucleotide base, there should be an even amount of each nucleotide expected to interact with said region; due to the limited bZIP-DNA co-crystal structure dataset collected, the number of sequences at each bZIP region is uneven, with a skewed frequency of nucleotides expected to pack into said region. An example of this difficulty is the large amount of C/EBPβ homodimers packing the DNA sequence 5’-TTGCGCAA-3’ found in the PDB dataset, thus causing a skew in the sequence frequency at each protein region, as well as in the DNA knobs expected to pack into the helical lattice. To emphasize the frequency of packing interactions, the bar plot in Figure 6 shows the frequency of different pocket composition at each recognition region, from P1 to N4. The highest occurring sequence at each region is the result of the skewed dataset, specifically the inclusion of the C/EBPβ homodimers. Since each proposed packing region is a 2(2:1) pocket, the combined high- and low-H socket takes on an **H:XY:H** structure, where **X** and **Y** are adjacent protein residues, with the two **H** residues hydrogen bonding to either **X** or **Y** in an *i* to *i+4* pattern. The LOGOS plots in Figure 6, in contrast, display the packing behavior of each pocket composition at one of the four regions. The added total of each column should equate to the number of occurrences displayed in the bar plot on the left. Each region and its compositions are further split into three LOGOS plots, representing the three possible binding modes previously defined: base-, backbone-, and non-packing. For all four regions, each socket composition was shown to prefer base or backbone packing, as well as the specific packing preference based on a specific nucleotide.

Of all four regions, P2 and N4 packs the base most often; N3 either packs the base of dT or does not pack with dG, while P1 packs mostly backbone or does not pack neither base nor backbone. Across P2, N3, and N4, the commonly packed base is dT, while other bases are less likely to be packed. Additionally, P2 (R:NN:V and R:NN:A) also packs 5mC, represented by **M**, which is a methyl-containing cytosine analogue. The least consistent packing regions, P1 and N3, also correlate with Figure 5, yet P1 consistently packs the backbone of its respective 1p nucleotide. Phosphate-sugar backbone moieties are constant across the DNA nucleotides and are more involved in stability and electrostatic interactions than specificity. Therefore, the binding mechanism that likely contributes to specific bZIP-DNA packing involves base packing over backbone or non-packing interactions. This emphasis on base packing highlights the importance and consistent packing of dT over dG, dA, and dC. Furthermore, regions P1 and N3 are less significant to DNA-binding specificity, yet the relative conservation of position 1p and 3n seems to contradict the lack of base packing.

Pairwise Packing of bZIP-DNA Interfaces Depending on Nucleotide Composition

Figure 7

Heatmaps of Pairwise Contacts Between DNA Nucleotide Positions and Protein Residues

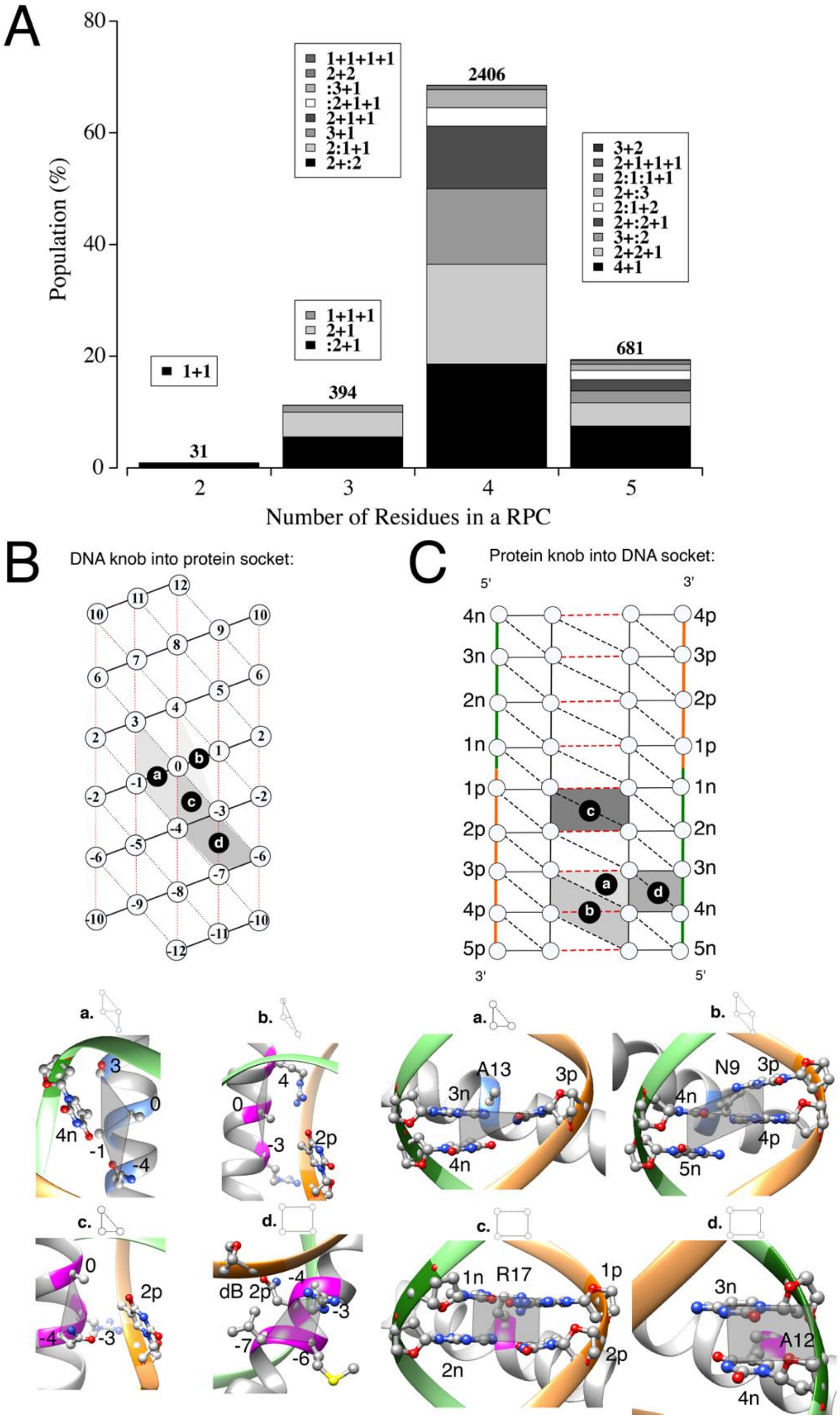


(Continued, Figure 7) *Note.* Each heatmap is grouped from pairwise contacts between a DNA nucleotide knob's base and the bZIP helix, filtering out all backbone contacts. The amino acids occupying each residue position is not considered. The x-axes represent the five nucleotides, consisting of the four conventional deoxyribonucleotides along with 5-methylcytosine. The y-axes involve the protein residue positions that contact the nucleotide position, ranging from base pair 1 to 4 on both the positive and negative strand, resulting in eight heatmaps. The percentage of contact between residues to a specific nucleotide reveals possible packing preferences based on the nucleotide occupying the DNA position. Packing involves the DNA knobs contacting three residues minimum, which may reveal the packing location of knobs as well.

The base packing shown in Figure 6's LOGOS plots seem to lack nucleotide-to-sequence specificity, so nucleotide-to-residue pairwise interactions were collated and observed. Similar to Figure 5, the pairwise interactions were gathered and sorted based on standardized protein residues, regardless of the amino acid occupying a position. The heatmaps of Figure 7, however, are split based on the four primary base pairs, from pairs 1 to 4 spanning both the "p" and "n" strand. Each heatmap has an x-axis for each nucleotide, and the y-axis represents all protein residue numbers that contact the base of the corresponding nucleotide position at least once. While these heatmaps reveal residue contacts for each DNA knob position, the specificity revolving around the protein residue composition is not shown. Each region matches its respective nucleotide position: P1 packs 1p, N4 packs 4n, etc. Most interactions are seen for nucleotides 1p, 2p, 3n, and 4n. These four nucleotide positions contact three or more residues consistently, though specific bases contact said residues more consistently than others. Just as seen in Figures 5 and 6, the most consistent residue contacts involve 2p and 4n, while 3n packs moderately. 1p contacts residues -3, 0, and 4; 3n packs 0, 3, and 4; and 4n packs -4, -1, 0, and 3. All residues contact match the proposed protein packing regions P1, N3, and N4, respectively. 2p, however, contacts residues -7, -4, -3, and 0 frequently, along with residue 4. The first four residue positions match the proposed pocket region P2, yet residue 4 is not considered part of the region. Since 2p packs with -3, 0, and 4, 2p would also be counted as a P1-packing DNA knob in Figure 5 and 6, yet the counted cliques of 2p packing into P1 is minimal. Additionally, the pairwise contacts grouped in Figure 7 involves base contacts only. The heatmap of P1 has consistent base contacts with residue 0 and 4, and moderately frequent contact with residue -3. Packing with three protein residues of region P1 would be considered packing, yet the P1 LOGOS plots of Figure 6 indicate that little to no base packing occurs. These two examples arising from pairwise interaction analysis contradicts the findings of the clique analysis of Figures 5 and 6, indicating possible loss of packing data that may otherwise supply information of packing specificity. Focusing on four-body knob-socket motifs and five-body knob-pocket cliques are insufficient to fully understand bZIP-DNA specificity, therefore requiring a different approach to the packing clique data.

Figure 8

Knob-Socket Quaternary Packing Between bZIP Protein and DNA



(Continued, Figure 8) *Note.* Structure images were created in Chimera (59). A) Stacked bar plot counts frequency of RPC sizes for all cliques involving both bZIP protein and DNA within the dataset. Four-body RPCs are most frequent while three- and five-body cliques significantly contribute to packing as well. Each RPC size is split further into local/non-local packing configurations to display the frequency of the clique configurations. For the nomenclature, residues/bases on the same local structure are grouped together and contiguous groups are summed. Furthermore, a ":" indicates hydrogen bonding between residues/bases and a "+" indicates interactions between two residues/bases from different structures. For example, a 2:1+1 RPC indicates three residues on a local structure, where two are contiguous and one is hydrogen bonded, and these three vertices pack with a residue or base from another non-local structure. 2+:2 RPC's are common 4-body RPC's, forming between two protein residues and two DNA residues. Therefore, a protein residue in 2+:2 cliques can contact only two DNA bases maximum, rather than the conventional three. The 2+1+1 and 3+1 cliques commonly involve a protein knob packing into a DNA socket, with 2+1 sockets forming between base pairs and 3 sockets forming between backbone and base of a single strand. 2:1+1 cliques describe the classic three residue α-helical socket, shown in Figure 3B and 2A, packing with a DNA knob through the common knob-socket motif. Additionally, three- and five-body RPC's contribute to describing bZIP-DNA packing on a higher-level analysis. The :2+1 cliques describe a DNA knob packing into two protein residues, while the 2+1 cliques describe a protein knob packing into two DNA residues. 1+1+1 cliques involve a protein residue contacting two DNA bases on two different strands. While packing and specificity generally involve a knob packing into three or four residues, multiple 3-body RPC's of the same knob may further describe packing mechanisms. Five-body RPC's consist mostly of 4+1 knob-sockets, which is a protein knob packing with two DNA bases and two DNA (Continued, Figure 8) backbones of the same strand; 2+2+1 packing cliques form from a protein knob packing into four DNA bases within the duplex. Infrequent and sparse packing clique types were combined into a general “Other” category and are variable in structure and configuration. B) Examples of DNA knobs packing with protein socket/knobs. The α-helical protein lattice is shown with four representative individual sockets or paired adjacent sockets (pockets) packed with DNA knobs that face into the plane and labeled *a* to *d*. Adjacent sockets packing the same knob form pockets, increasing the packing interface and possible residues involved in specificity. Below this lattice are displayed 3D examples of four types of DNA knobs packing into α-helical sockets/pockets, accordingly labeled from *a* to *d*. Figures show position numbers corresponding to the α-helical and also half-site positions for the DNA. For (a), a classic α-helical diamond pocket formed from adjacent low-H and high-H sockets packs with a DNA base knob at position 4n packs, which is denoted a 2(2:1)+1 knob-socket. For (b), the :3+1 is shown and involves a :3 elongated socket of hydrogen bonded residues -3, 0, and 4 packing with DNA base knob at position 2p. For (c), the knob-socket motif 2:1+1 is shown where a DNA base knob at position 2p packs into the α-helical 2:1 socket. For (d), a five-body clique of 1+2:2 differs from pockets formed from contiguous filled sockets. The DNA knob packs into a four-body protein pocket rather than two separate adjacent sockets. C) Examples of protein knobs packing with DNA sockets/pockets. The DNA duplex lattice representing the major groove is shown with protein knobs packing into DNA sockets and pockets, and the four examples are labeled from *a* to *d*. Some of the examples are five-body RPC’s, where the DNA pocket is not the result of two sockets but rather involved directly in the mutual contact clique. Below the DNA lattice are displayed corresponding 3D examples of the protein knobs packing into the respective socket from PDB 1DH3. Protein residues and positions along with DNA half site positions are noted. In (Continued, Figure 8) (a), the 1+1+2 clique involves a protein knob packing into the 1+2 DNA socket, formed between adjacent base pairs. One such socket consists of the base at position 3p from one strand, its base pair at position 3n, and the base from position 4n on the other strand. In (b), the 1+2+2 clique is built from two adjacent 1+2 sockets packed by the same 1+ knob, resulting in the 2+2 pocket between 3p, 4p, 4n, and 5n. In (c), a protein residue packs into the pocket of adjacent sockets to form a 1+2+2 knob-socket. While similar to the previous pocket, the adjacent 1+2 sockets come from the adjacent base pairs: 1p, 1n and 2p, 2n, and thus form a more rectangular pocket that packs the 1+ protein knob. In (d), a pocket of adjacent DNA sockets forms between a base and backbone of two adjacent DNA nucleotides of the same strand to pack an α-helical knob and forms a 1+4 knob-socket.

From the modified KS analysis of the 43 bZIP-DNA crystal structures, the relative packing clique (RPC) data is grouped and shown in Figure 8A. An RPC is a set of protein residues and DNA backbone and/or base vertices that all contact and pack with each other (4). All RPCs involving both protein and DNA are sorted by size, where the size is the number of vertices within the RPC. Figure 8A shows a histogram of these RPCs between protein and DNA grouped based on clique size and further categorized by contact order of the members in the clique (58). Prior analyses of protein packing identified the importance of three-body and four-body cliques and the lack of two-body pairwise interactions (4-7). Protein interactions with DNA macromolecules also require consideration of three-body and four-body cliques as well as five-body cliques to accurately depict the contact and packing interfaces between the bZIP proteins and the DNA consensus sequences. Besides the two-body cliques, each RPC size consists of multiple clique configurations between protein and DNA based on contact order of interacting clique members. For example, 2+1 cliques represent two adjacent residues or nucleotides packing with one residue or nucleotide from a different structure, while 1+1+1 represents contact between three residues or nucleotides from distinctly separate structures. In Figure 8A, the predominant interactions occur between an α-helical monomer and the DNA. Figures 8B and 8C illustrate the common packing sockets and pockets on respective 2D α-helical and DNA major groove lattices, where pockets are a shorthand term for a set of contiguous sockets packing the same knob. In both protein and DNA socket lattices, sockets or pockets packing a knob residue means that the respective socket or pocket is filled.

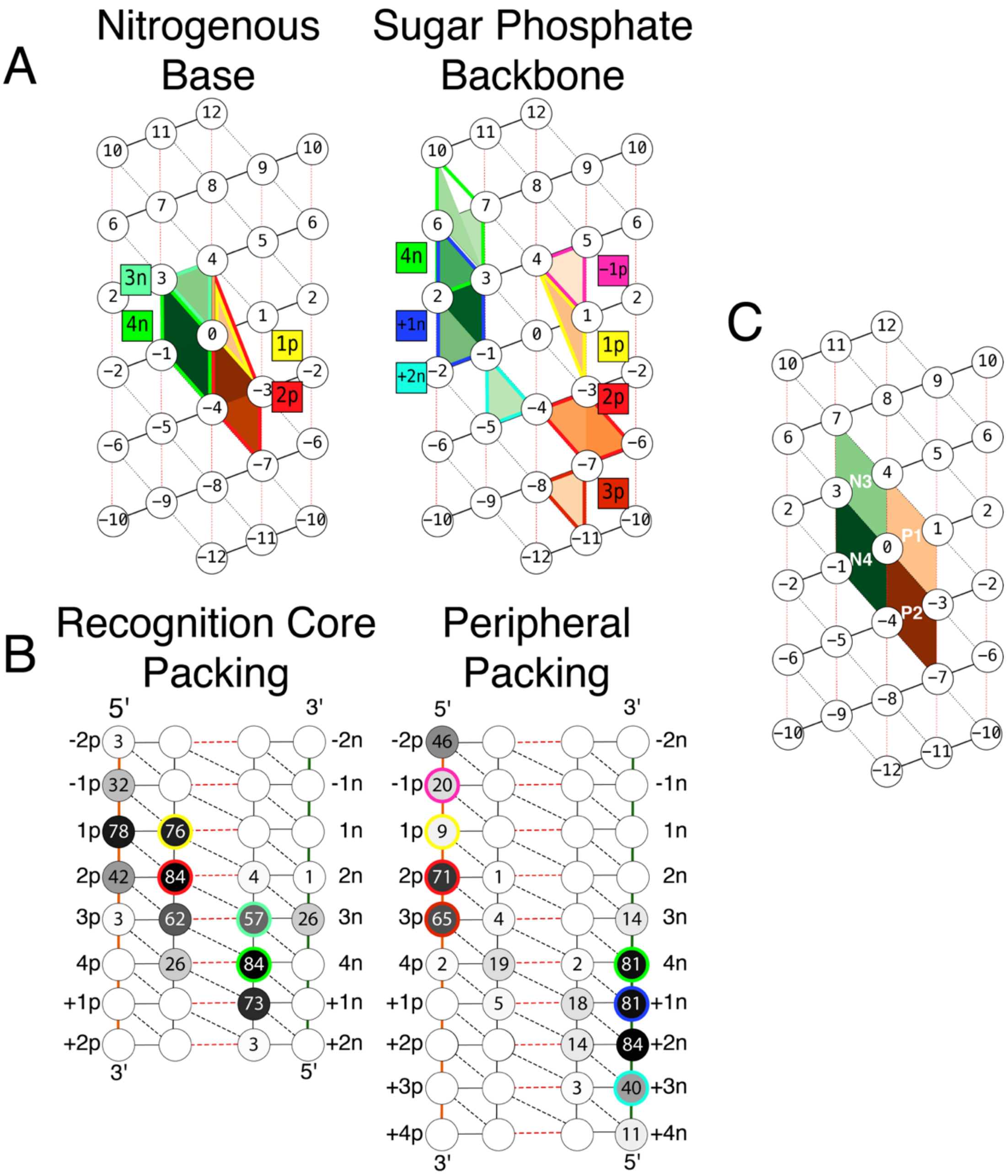
Figure 8B displays the common DNA knobs packing into bZIP sockets and pockets on an α-helix socket lattice. Both Figures 8B.b and 8B.c are examples of sockets, where three residues are in mutual contact, that are filled by a knob, either a nucleotide backbone or base. Figure 8B.b has a filled socket of residues -3, 0, and 4 packing with the DNA knob 2p. This :3 socket type is considered “elongated” since the three residues are not adjacent to one another. Figure 8B.c displays the standard socket, formed from residues -4, -3, and 0, and filled by the DNA knob 2p. The standard 2:1 socket is both common and discrete throughout most protein helices and is categorized as either a high-H or low-H socket, where high and low refer to relative sequence position. The 2:1 socket in Figure 8B.c is a high-H socket, where -4 and -3 are covalently bonded, -4 and 0 are hydrogen bonded, and -3 and 0 contact through van der Waals interactions. Because the distal hydrogen-bonded residue 0, is downstream from the other two residues, this specific example is a high-H, or **XY:H**, socket. Low-H sockets are represented by **H:XY**, where residue **H** comes before residues **X** and **Y**. Besides single filled sockets (Figure 8B.b and c), two adjacent sockets can be filled by the same knob, which is considered a pocket (Figure 8B.a), which mimics the classic coiled-coil packing between two α -helices (4). Note that the two 2:1 sockets between protein residues -4, -1, 0, and 3 are adjacent high- and low-H sockets that form a rhomboidal pocket **H:XY:H** that share residues -1 and 0, while residues -4 and 3 are the two **H** residues that contribute to their respective sockets in this pocket. Another less common pocket packed by a DNA knob is shown in Figure 8B.d, where a low-**H** socket packs the same knob as the next consecutive high-**H** socket. In this case, the rhomboidal pocket **Y:XH:Y** shares **X** and **H** residues -7 and -3 with residue **Y**s at positions -6 and -4. However, not all pockets are formed from two adjacent four-body clique sockets, some are formed by a five-body clique directly where a knob packs into the mutual four-body pocket directly; both types of pockets are considered equally. For example, the DNA nucleotide knob 4n in Figure 8B.a could also pack with all four protein residues instead of two sockets. The same is found in the interactions shown in Figure 8B.d, where a DNA backbone knob 2p could pack into sets of four protein residues. Both examples increase the packing interface between the DNA knob and the bZIP helix. In all four examples in Figure 8B, a DNA knob packs into either sockets or pockets, highlighting the importance of multibody contacts. The knob packs with all residues forming the socket or pocket, while said residues also interact with one another. These higher-order packing interactions indicate how knob-sockets can define specificity when applied to a single DNA knob packing into a protein helix. Because the DNA backbone is standard between all nucleotides, the specificity most likely derives from protein socket and pockets packing DNA base knobs.

Figure 8C shows the common filled sockets and pockets of the DNA duplex lattice with protein knob residues. In Figure 8C.a, an Ala protein knob residue packs with a three-body base socket formed by base residues from positions 3p, 3n, and 4n. Since this is a clique interaction, the protein knob packs with the three DNA bases, which are also in mutual contact with one another. Protein knobs can pack into a single three-body socket or into a four-body pocket. The protein knob may also pack into two adjacent knob-socket motifs that act as a pocket as in Figure 8C.b or directly into the four-body protein pocket displayed in Figure 8C.d. Analogous to Figure 8B, DNA pockets can be formed by combining two sockets, just as a five-body RPC of a protein knob packing into a four-body residue set may be divided into two pseudo-sockets. Figure 8C.b demonstrates how an Asn protein knob packs into two adjacent sockets, where bases 4p and 4n are shared between the sockets formed respectively by 3p and 5n bases positions. These sockets form a pocket where the Asn contacts four DNA bases total. In contrast, Figure 8C.c and 8C.d display five-body cliques where Arg and Ala residues contact four DNA vertices and indicate that the protein knob packs into a set of mutually contacting DNA vertices. These KS clique results allow for specific knob-socket propensities to be analyzed. DNA sockets and pockets form between base pairs of a duplex or between adjacent backbone vertices within a single strand. Packing cliques formed between the base and backbone of adjacent nucleotides within a positive and negative strand involves the positive lane (PL) and negative lane (NL). Cliques formed between base pairs are located in the middle lane (ML). Figure 8C.d is located in the NL, while Figures 8C.a, 8C.b, and 8C.c involve the ML. Since the DNA backbone is standard between all nucleotides, specificity likely depends significantly on base interactions and highlights the importance of the ML. The packing examples are shown structurally below the protein and DNA lattices, emphasizing the variety of packing mechanisms between the two macromolecules.

All RPCs involving both biomolecules were analyzed and classified based on knob and socket combinations as explained in Materials and Methods, and this approach allows a precise investigation of recognition between the protein α-helix and duplex DNA. The DNA nucleotide knobs packing into the protein α-helical socket lattice allows an investigation of protein recognition of DNA bases. Conversely, the knobs from protein α-helix residues packing into the DNA major groove duplex lattice should reveal any DNA recognition of protein residues.

Figure 9

α-helical Recognition Core: Composite Packing Map of the α-helical Packing Surface with Corresponding DNA Knob Position Frequencies



(Continued, Figure 9) *Note.* Protein lattices are shown displaying the common sockets and areas where DNA knobs pack into the helix, categorized by the base number and position. Nucleotide and amino acid sequence composition is not considered for this lattice. The packing groups were limited to common three- or four-residue groups, acting as sockets or pockets. Infrequent or uncommon interactions were excluded from the lattice for clarity. Socket shading corresponds to the highest frequency of a DNA knob packing into the socket, rather than packing overall. Sockets corresponding to bases of the positive strand were shaded orange, while sockets packing with negative strand bases were shaded green. The borders on the lattice represent the base that frequently packs into the area, where the border color corresponds to the colored labeled box. A) The α-helical lattice on the left involves DNA base knob packing only, primarily packs with DNA bases 1p, 2p, 3n and 4n: two bases from the positive strand and two from the negative strand. The DNA base at position 1p packs consistently into the elongated 3 socket region defined by α-helical residues -3, 0, and 4. The DNA at position 2p also packs into this elongated socket inconsistently. Base 2p primarily packs into the pocket region formed by -7, -4, -3, and 0 α-helical residues. The DNA base at position 3n packs definitively into just a socket of residues 0, 3 and 4. The DNA base at position 4n packs very consistently into a pocket made up of -4, -1, 0, and 3 residues on the α-helix. The protein lattice on the right displays the frequent packing interactions where non-specific DNA backbone knobs pack into the α-helix. The representation of the sockets and borders, each corresponding to a backbone knob position, is the same as the lattice previously discussed. DNA backbone knob packing is very inconsistent across areas of the α-helix, resulting in many overlapping borders. B) The left DNA duplex shows the frequencies that each DNA knob packs into the proposed protein recognition core within the 85 packing interfaces between a bZIP helix and the respective DNA half-site, considering a DNA knob as

(Continued, Figure 9) packed if the total packing cliques involve the core region. Significant knobs are bordered by the color corresponding to the protein lattices of part A, showing a localized base preference compared to the backbone knobs. Base 3p and +1n seemingly pack frequently into the recognition core, though most of the packing interactions involve two residues compared to the common three-residue contacts seen in the four primary DNA base positions. The right DNA duplex groups the frequencies of each DNA knob that packs outside of the recognition core within the 85 packing interfaces between bZIP and DNA half-site. The DNA base knobs pack inconsistently, while most of the DNA backbone packs outside of the recognition core. Backbone packing is not localized, as shown by the expanded range of knobs from -2p to +4n. C) The recognition core of the protein α-helix recognizes the DNA half-site by packing four primary bases. The core is formed by the common α-helix packing areas denoted P1, P2, N3, and N4 that correspond to the DNA base knobs from positions 1p, 2p, 3n, and 4n. The regions P2 and N4 pack bases 2p and 4n. The two regions P1 and N3 pack bases 1p and 3n and were expanded by an additional residue to span a pocket region to account for potential influences of residues 1 and 7.

For all the DNA half-sites, the frequencies of packing groups between DNA knobs and protein residue groups were mapped onto protein α-helical lattices without regard to amino acid or DNA sequence, in order to reveal any consistent packing schemes. Because DNA in the modified KS is defined to pack with two types of knobs: backbone and base, each type of knob's packing group was overlaid onto its own protein α-helical lattice, which created the two lattices shown in Figure 9A. A full account of DNA knobs into bZIP α-helix sockets is shown in Supplementary Figure S2. The distinct separation of packing patterns between the base knobs

(Figure 9A, left) and backbone knobs (Figure 9A, right) highlights the importance of modifying the KS analysis for DNA. In particular, specificity for recognizing the base knobs by the bZIP α-helix is clearly identifiable outside of the disperse packing by backbone knobs. The other distinct characteristic of DNA nucleotide packing revealed by the modified KS analysis is that the α-helix does not recognize the half-site all on the positive strand as is commonly assumed. Instead, the modified KS analysis reveals that the α-helix primarily packs the positive strand for the first two nucleotides 1p and 2p and the negative strand for the last two nucleotides 3n and 4n. This splitting across the positive and negative strands is especially evident in the nucleotide base knob packing shown on the left part of Figure 9A and is mostly echoed by the nucleotide backbone knob packing shown on the right part of Figure 9A.

For the nucleotide base packing (Figure 9A, left lattice), the collated DNA base to protein interactions show distinct and localized DNA packing regions on the α-helical lattice for the four half-site bases. The mapping using the aligned residue positions consistent between all bZIPs identifies a clear packing area on the α-helical lattice. For reference, the conserved Asn is at position -4 and the conserved Arg is at position 4. The recognition of the four base DNA half-site consists of a core region in a quadripartite diamond pattern that spans α-helical positions -7 to 7. For each of the four nucleotide bases acting as a knob, the most common packing pattern involved three or four protein residues that formed filled sockets and pockets, respectively. Following DNA base order, 1p packs into a three-residue socket localized around α-helical residues -3, 0, and 4 that also occasionally packs with base 2p. This socket/pocket is in the upper right quadrant of the diamond. In several half-site packing interfaces, 1p also packs two residue pairs involving -3 to 0 and 0 to 4; as the base contacts three local protein residues, the two residue pairs form a “pseudo-socket” with possible levels of specificity. In the bottom lower right quadrant, 2p packs into a pocket formed from adjacent sockets consisting of residues -7, -4, -3, and 0; the low-H socket has lower packing frequencies with 2p then does the high-H socket. As previously mentioned, 2p also packs into the elongated 3 socket with 1p, albeit at a much lower frequency. This overlap may be the result of the orientation of bZIP helix and major groove, and likely does not contribute to specificity as well as the packing between 1p and the elongated socket. From these first two nucleotide bases on the positive strand, packing now switches to the negative strand for the last two nucleotide bases. In the top upper left quadrant, 3n packs into a low-H 2:1 socket including residues 0, 3, and 4 with little to no crossover. In the lower right quadrant, 4n packs into the pocket consisting of residues -4, -1, 0, and 3; similar to the 2n pocket, the 4n pocket is formed from two separate sockets. Packing of 4n has virtually no overlap with other bases and is the most consistent packing base position of all four half-site bases. As indicated by the darker shade of red and green, bases 2p and 4n, respectively, have near constant frequencies, which strongly indicates their roles in DNA sequence recognition. In contrast, bases 1p and 3n have lower packing frequencies compared to 2p and 4n. The lower packing frequencies and smaller packing area of both 1p and 3n implicate lower specificity stringency and lower conservation of knob-to-socket propensities.

In the right lattice of Figure 9A, the packing of nucleotide backbone knobs into the bZIP α-helix clearly shows the importance of modifying the KS analysis to consider the nucleotide base and backbone separately. The common backbone knobs range from DNA positions -1 to +2, which is far more than the four DNA base positions that are commonly considered involved in sequence recognition. The frequency of backbone packing decreases overall while the packing area and socket range increases, which implies that backbone packing is neither localized nor specific to sequence. In other words, the packing of knobs from the backbone is less distinct with more overlaps between residues. Since the sugar-phosphate backbone groups are the same for all nucleotides and therefore would not contribute to specificity, the lack of definition and consistency in packing is to be expected. However, the DNA backbone knobs pack into areas on the bZIP α-helix that are peripheral and complementary to the diamond recognition core of the base knobs with minor overlap between the backbone of 1p and 2p into the core region. If backbone and base knobs were considered together, the bZIP diamond shaped quadripartite recognition core would become lost in the indistinct packing of the backbone knobs.

Figure 9B depicts the frequency distribution of corresponding DNA knobs that pack into bZIP α-helix sockets discussed above. Therefore, the DNA knob frequency was classified based on packing into the recognition core (Figure 9B, left) or into the peripheral socket regions around the core (Figure 9B, right). The DNA lattices in Figure 9B show the frequency at which each possible DNA knob packs into the major groove either in residue pairs, protein sockets, or protein pockets. Both lattices in Figure 9B sort all packing cliques from each half-site to consider if each DNA knob in the sequence packed the core or the peripheral region, where 85 is the maximum frequency value. Consistent with the analysis of α-helix sockets, the majority of DNA knob packing into the recognition core region was primarily due to base knobs 1p, 2p, 3n, and 4n (Figure 9B, left), and again, bZIP recognition of half-sites is split across both positive and negative strands. Also, the packing of base knobs at the first two positions on the negative strand 1n and 2n is negligible. Although the bases of nucleotide position 3p and 4p pack frequently with the protein, they primarily contact two residues rather than three, which does not significantly contribute to specificity (Supplemental Figure S2A). While protein contacts with 3p and 4p are semi-consistent, sockets and pseudo-sockets are not as consistent as the preceding base 2p; a similar pattern occurs with base +1n, where consistent packing with residues -4 and -1 do not confer significant specificity, as residue -4 is invariant. The backbone knobs of -1p and 2p also interact with the recognition core, although usually not into sockets or pseudo-sockets (Supplemental Figure S2B). The backbone knob of 1p, as indicated in Figure 9A, packs in the other elongated socket adjacent to the 1p base; the 1p base knob packs with -3, 0, and 4, while the 1p backbone packs with -3, 1, and 4, essentially completing the rhomboidal region. Additionally, the base knobs of 3p and +1n contact the recognition core frequently as well, although the contacts typically involve protein residue pairs rather than residue trios as previously discussed (Supplemental Figure S2A). These base contacts are likely due to the deep pocket packing of upstream DNA bases 2p and 4n, rather than specific recognition of 3p and 5n. Additionally, although bases 1p and 3n pack into sockets, the general 2(2:1) filled pocket shape of 2p and 4n may also apply to 1p and 3n, where residues 1 and 7 may play a role in specificity that rejects nitrogenous base packing. Packing data additionally reveals the packing frequency outside the recognition core, where the majority of knob packing was limited to DNA backbone, as explained in the corresponding knob frequency DNA duplex lattice (Figure 9B, right). The peripheral packing DNA frequency lattice has little to no base packing, supporting the localized base recognition core model. A large range of DNA backbone knobs pack into the peripheral region around the recognition core, spanning base pair -2 to +3. DNA backbone packing does not have any major overlap with the regions packed by DNA knobs, as such, the packing between DNA base and backbone are complementary to each other. Backbone packing around the recognition core spans beyond the half-site sequence and is inconsistent compared to the packing frequencies of the DNA bases into the recognition core, supporting the proposition that DNA backbone packing contributes more to affinity rather than specificity.

From these analyses of the DNA knobs packing into the bZIP α-helical socket lattice, a simple model of protein α-helix recognition of DNA bases can be made (Figure 9C). The packing of a four base pair DNA half-site is dependent on the bZIP quadripartite recognition core shown in Figure 9C, where the four primary base knobs are packed into a specific region defined by the invariant bZIP α-helical residues Asn and Arg at positions -4 and 4. The proposed nine-residue DNA recognition core defines a quadripartite region that packs knobs from each of the four positions in a DNA half-site. Rather than recognizing base positions along a single strand of a half-site, bZIP α-helices specifically pack the first two bases on the positive strand of a half-site and the last two bases on the negative strand. These base knobs are denoted 1p, 2p, 3n, and 4n, and pack into the corresponding pockets on the bZIP α-helical socket lattice: 1p packs into P1, 2p into P2, 3n into N3, and 4n into N4, as noted in Figure 9C. The four pockets P1, P2, N3, and N4 define the bZIP quadripartite recognition core that surrounds the invariant Asn at -4 and Arg at 4 and mainly consists of residues -7, -4, -3, -1, 0, 3, and 4. While residues 1 and 7 are included in the depiction of the recognition cores, base knobs 1p and 3n have limited interactions with residues 1 and 7, respectively. These nine residues pack the four base positions consistently across all analyzed crystal structures of various protein and DNA sequences. DNA recognition in a bZIP-DNA complex is dependent on these four DNA nucleotides, split across the positive and negative strand, and the nine-residue recognition core on the bZIP α-helix.

Table 1

Protein Pocket Composition Specificity of DNA Bases

Region	Base Packing Frequency					Region	Base Packing Frequency				
P1			1p			P2			2p		
-3:0 1:4	dA	dT	dG	dC	5mC	-7:-4 -3:0	dA	dT	dG	dC	5mC
K:AA:R		1/1	4/6	1/1		R:NK:A		9/9			
K:AR:R			1/1			R:NN:A				2/2	2/2
N:AR:R			4/4			L:NR:A		3/3			
R:AA:R			1/1	0/1		R:NR:A	0/2	14/14			
R:AQ:R			3/5	0/2		T:NR:A		9/9			
R:AR:R			1/3	1/1		R:NT:A		6/6			
R:AS:R	1/2		2/5	1/8		K:NR:Q		2/2			
T:AR:R			4/5	1/1		R:NR:S		3/3			1/1
R:QR:R			2/2			R:NN:V	6/6			22/22	4/4
R:SR:R			0/2	2/2							
N:VR:R			14/25	2/7							
N3			3n			N4			4n		
0:3 4:7	dA	dT	dG	dC	5mC	-4:-1 0:3	dA	dT	dG	dC	5mC
A:CR:K			0/5			N:AA:C		12/12			
A:CR:R		1/1	2/15			N:YA:C		9/9			
A:SR:A		4/4				N:AA:S		24/24		0/1	1/1
A:SR:F			0/2			N:AQ:F		2/2			
A:SR:K		0/1	2/16			N:AS:C		2/2			
A:SR:R			0/3			N:AS:S		2/2			
Q:FR:K		1/2				N:AV:S		32/32			
S:CR:F			0/2								
S:SR:F			0/1		1/1						
V:SR:A		27/27	4/5								

Note. Packing interactions separated by recognition core region as shown in Figure 9C. For each half-site, packing interactions of DNA positions 1p, 2p, 3n, and 4n were grouped with the respective region. Studies included the 4 standard bases of dA, dT, dG, and dC as well as 5-methylcytosine (5mC). Each nucleotide base knob contacted at least three protein residues in the packing region to be considered significant to specificity, regardless of packing mechanism and orientation. So, when an entry shows 1/7, this indicates that the DNA base packed only one time into a particular packing region's amino acid composition out of the seven instances of that

(Continued, Table 1) specific DNA base and region amino acid composition are found out of 85 half-site instances. The pocket composition corresponds to the given relative residue number position in the bZIP α-helix recognition core. For example, when using the KS pocket nomenclature and region N3 (0:3 4:7) is occupied by V:SR:A, the residues are Val at position 0, Ser at position 3, Arg at position 4, and Ala at position 7, respectively. The packing frequency of the DNA base composition into the corresponding region per sequence composition is compared to the frequency the DNA nucleotide occupies the knob position while the respective region is composed of said amino acid sequence. As a type of amino acid packing code for DNA bases, these tables reveal the propensities a protein packing region is likely to be filled by a specific DNA knob at the relative DNA position, where recognition of a DNA base differs based on region composition. Most frequent and consistent packing involves recognizing dT and thymine analogues, especially in regions P2 and N4. Region N3 also recognizes dT, though packing is less frequent, paralleling the lower packing frequencies seen in Figure 9A.

To investigate protein sequence specificity for DNA bases, Table 1 summarizes all DNA base knob to α-helical region packing frequencies split into four tables corresponding to the four regions making up the quadripartite recognition core: P1, P2, N3, and N4, respectively. Each table compares one of these 4 protein region's amino acid sequence composition to each of the five DNA bases dA, dT, dG, dC, and 5-methylcytosine (5mC). Because these frequencies only come from a non-exhaustive sample of 85 half-sites, the packing frequencies are only suggestive; however, some inferences can be made. In this table, the KS nomenclature is used to simplify the referencing between relative position in the bZIP α-helix recognition core and amino acid. For example, region P1 is composed of residue positions -3, 0, 1, and 4 (see Figure 9C). When

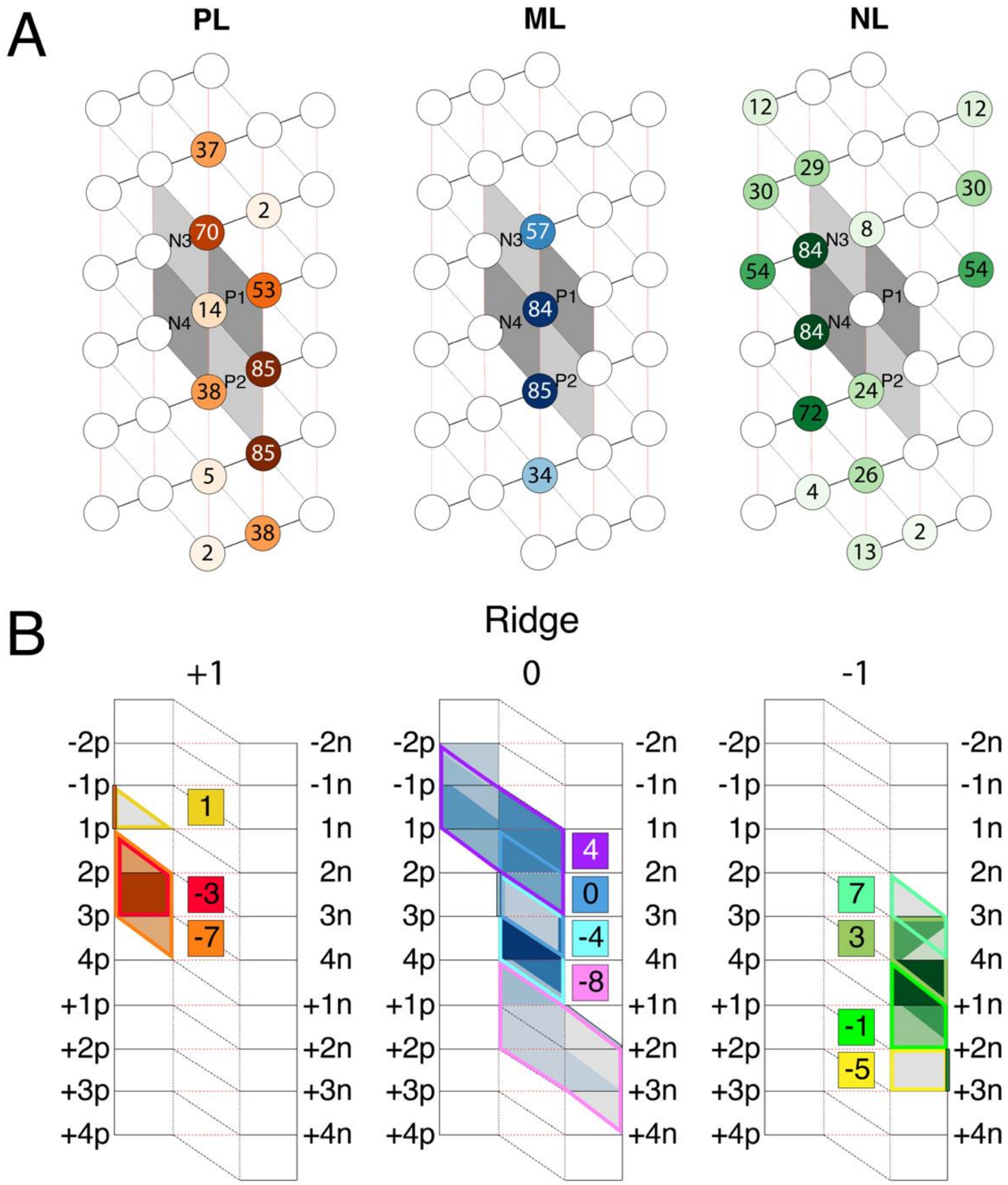
P1's sequence is denoted N:AR:R, the region consists of Asn at position -3, Ala at 0, Arg at 1, and Arg at 4. Also, all the regions include one position with either of the conserved amino acids Asn at -4 and Arg at 4, so specificity for DNA base packing is determined and will be discussed by the other 3 residues of the pocket. In region P1, dG is found predominantly 59 times or 69% in this data set, but only packs into the region P1 pocket 36 times. In general, as shown in the left of Figure 9A, P1 packs the DNA base at 46 of 85 or 54% of the time. The P1 region's variable pocket positions are -3, 0, and 1, but packing doesn't include position (Figure 9C). Position -3 consists primarily of charge or polar residues Lys and Arg or Asn or Thr, respectively, and position 0 consists of hydrophobic or polar residues Ala and Val or Ser and Gln, respectively. However, the sequences do not show any DNA base preference when there is packing. For instance the K:AA:R pocket composition packs dT, dG and dC. In contrast, the P2 region bases pack 83 out of 85 half-sites. The only two instances of no packing are dA for the sequence R:NR:A. Consistently for the P2 region, whenever there is a dT, there is packing. The dT bases packs into seven different P2 -7:-4,-3:0 pocket compositions, where the variable positions consist of the following amino acids: positions -7 and -3 are Arg, Lys, and Thr, and position 0 are Ala, Val, Ser and Gln. Analyzing the two sequences that do not pack dT - R:NN:A and R:NN:V, it seems that having an Asn at position -3 is specific to not packing dT, since having an Arg at position 7 and Ala at position 0 are found in two other region pockets that pack dT. Interestingly, both of these compositions can pack 5mC, which has a methyl group similar to dT. For the N3 region, primarily dG and dT are found at 3p of the half-site sequence in the data set with only one instance of 5mC. Of the 49 instances that dG could pack into the N3 region, packing occurs only 8 times. In contrast, dT packs in 33 of 35 possible instances and 5mC packs in the one instance. These results strongly suggest that dT is recognized, but the dG doesn't demonstrate strong interactions and therefore, less specificity. Comparing N3 pocket 0:3,4:7 region sequences for dT packing, a Phe at position 7 disfavors dT packing, whereas having a hydrophobic at position 0 (Ala or Val) and position 7 (Ala) favors dT binding in the sequence compositions A:SR:A and V:SRA. While the Ser at position 3 is also conserved, it is also found in N3 region sequences that do not pack dT. It is surprising that position 7 in N3 seems exerts influence on packing specificity, since it doesn't directly pack (Figure 9A, left). These results suggest indirect effects such as a bulky Phe at position 7 blocking dT packing and a small Ala allowing packing. In the data set, region N4 almost always packs a dT, and similar to region P2, it packs dT 83 out of 83 instances. For the other two, dC is not packed and 5mC is. Because all the region N4 pocket -4:-1,0:3 sequence compositions pack dT, the only inference about packing is that position -1 is always an Ala. Position 0 varies with Ala, Val, Ser, and Gln as does Position 3 Cys, Ser, and Phe. Overall, the results strongly support that bZIP α-helix specificity for DNA bases primarily occurs primarily through methyl group recognition on dT or 5-methylcytosine by the four regions of the quadripartite recognition core.

By comparing the amino acid composition of sockets used for DNA base methyl recognition to sockets that pack in proteins, insight can be found for the fundamental requirements and possible mechanism of protein recognition of DNA. Supplemental Figure S3 compares the region sequences listed in Table 1 to packing propensities collated from α-helix pocket packing in the PDB (4). The shading indicates the frequency at which a pocket is found filled with another knob. The most common filled sockets and pockets are relatively hydrophobic and are typically packed with another hydrophobic knob; the common amino acid in both knobs and sockets is leucine. When comparing packing preferences in Table 1 to the results and packing propensities in Supplemental Figure S3, the most prevalent and filled sockets are in region N4. Regions P1 and N3 consist of several pocket sequences that are slightly more frequent than the other possible sequences, though the filled propensities are generally lower than those of N4. Overall, no pocket sequence that packs DNA is highly common in protein packing. Because of the conserved Asn at position -4 and the Arg at position 4, these results make sense in the context that proteins binding DNA must bind in the environment of a double helix's charged sugar phosphate backbone. Any pocket including an Asn or Arg is not a strong hydrophobic packing target, yet in the context of the DNA duplex, these are favorable. In addition, hydrophobic pockets on an α-helix promotes protein-protein interactions, so in this way, the composition of the pockets in the basic region of the bZIP α-helix favors DNA interactions by disfavoring protein ones. Lastly, the amino acid composition of the quadripartite region pockets does not strongly favor α-helix formation, which may play a role in the metastable mechanism necessary to sample DNA sequences to find the right target sequence.

Protein Residue Knobs Packing into DNA Backbone and Base Sockets

Figure 10

Protein Knob Packing into Socket of DNA Major Groove



Note. A) The α-helix lattices show the frequency each protein knob packing into the DNA duplex throughout the 85 half-sites grouped according to knob packing into each of the 3 lanes of the DNA major groove lattice: positive backbone strand lane (PL), middle base lane (ML), and

(Continued, Figure 10) negative backbone strand lane (NL), where protein knobs are colored shades of orange, blue and green, respectively, according to their frequency of packing. Knobs in the protein lattice can be grouped according to *i+4* ridges, which each correspond to a face of the α-helix (see Figure 3F). In this model, ridges are categorized relative to the ridge packing into the middle DNA base lane (ML), which is denoted as 0 ridge and is composed of -4, 0, and 4 protein residues. The +1 ridge is composed of the protein residues one position higher than those of 0 ridge, while the -1 ridge is composed of residues one position lower than the 0 ridge. Because of the circular nature of the α-helix, the +2 ridge is the same as the -2 ridge. Since this ridge faces opposite towards the solvent than the centrally packing 0 ridge, it is considered the solvent ridge. As shown in the first protein lattice, packing of protein knobs into the positive lane primarily comes from the +1 ridge with some contributions from the 0 ridge. Similarly, as shown in the third protein lattice, packing of protein knobs into the negative lane primarily comes from the -1 ridge with some contributions from the 0 ridge. The middle lane formed by only DNA base pairs is packed only by the protein knob residues on the 0 ridge: -8, -4, 0, and 4. While these same residues are involved in packing into the positive and negative lanes, the middle lane is packed solely by those four protein knobs. B) Three composite DNA lattice maps are shown based on which ridge protein knobs originate. From left to right, areas packing the knobs from the +1, 0, and -1 ridges are shaded orange, blue, and green, respectively according to frequency. Only the highest packing frequency per socket area as well as how often each area is packed by a specific protein knob is depicted. Packing by overlapping knobs is minimized for clarity. The common areas of the DNA lattice where protein knobs pack are bordered. Clearly, the +1 ridge only packs into the positive lane (left), and the -1 ridge packs only into the negative lane (right). The middle lane (center) is often packed by the 0 ridge residues, but these residues

(Continued, Figure 10) exhibit some overlapping packing area with the positive and negative lanes. The majority of the middle lane packing involves the 0 ridge residues -4, 0 and 4. Note that residues -4 and 4 are the invariant N and R, with residue 0 as the moderately conserved amino acid. Overlap packing occurs with protein residues 4 and -8 that often pack into the positive lane and negative lane.

In the same way that DNA knobs pack into specific areas on the protein lattice, protein residue knobs pack into specific areas on the DNA duplex lattice. Figure 10A illustrates the packing preference for each protein knob between the three DNA lanes, based on the standardized bZIP-DNA half-sites. To help with position references, the recognition core is shown in all the α-helical lattices, where positions can be found by referencing Figure 9C. Protein residues packing into the PL and NL sockets primarily involve non-specific backbone interactions, whereas the ML involves direct DNA base contacts. On the α-helix, the protein knobs can be grouped based on the *i+4* ridge that they occupy. The central ridge packs into the DNA duplex and is denoted the 0 ridge, as it includes the residue at position 0 as well as the invariant Asn at position -4 and Arg at position 4. The *i+4* to the left contains residues -1 in position to the 0 ridge and is denoted the -1 ridge. Likewise, the *i+4* to the right is the +1 ridge. Facing out towards the solvent, the ridges on the edge are the same due to the circular, repetitive nature of an α-helix, and is called the solvent ridge. In general, packing into the DNA PL involves only protein knobs from the 0 and +1 ridges; ML residues from only the 0 ridge; and NL knobs from the 0, -1, and solvent ridges. The PL and NL consists of backbone sockets that do not provide any specificity but are important for binding affinity within the charged phosphate backbone. The PL is commonly packed with residues -7, -3, and 1; these three residues are one position higher than ridge 0 and are considered part of ridge +1. Residues -5, -1, and 3 pack into the NL, therefore the *i+4* ridge on which the three residues are found is considered ridge -1. While the NL is also packed with residue 2, the packing is sparse and is not considered significant. In contrast to the PL and NL, the ML consists of base sockets that allow differentiation between DNA sequence positions. As a somewhat conserved position, residue 0 packs consistently into the ML and sometimes spills into the PL. The conserved residue Asn at position -4 packs across all three lanes with a constant presence in the ML. Conserved residue Arg at position 4 packs frequently into the ML, but also spans the PL just as frequently. Out of residues -4, 0, and 4, residue 0 is the most central and specific to the ML. The conservation of Asn at -4 and Arg at 4 between the different half-site specificities strongly suggests that they do not strongly contribute to protein-DNA sequence specificity. The variation at residue 0 indicates that this position is involved in DNA recognition of protein knobs. Lastly, the residue at position -8 doesn't pack consistently and when it does, it packs upstream of the recognition sequence. The three packing *i+4* ridges were used to split the protein knob data into the three composite maps of Figure 10B. The higher frequency knob-socket cliques were mapped to reveal likely areas where protein knobs recognize and pack into the DNA duplex. The DNA duplex lattices reveal that the packing of the protein knobs is widespread and variable in comparison to the DNA knob cliques. Ridges +1 and -1 pack into PL and NL, respectively, interacting with backbone vertices to likely assist in the protein residue solvation into the DNA duplex major groove. While these protein residues contact the base of the corresponding DNA lane, contacting two DNA bases is not considered significant to base and sequence recognition. Ridge 0 packs primarily into the ML, involving the most protein-base contacts between the macromolecules; residue -8 packs into both ML and NL inconsistently, while residue 4 packs into the ML and PL.

When comparing the overlap between -4, 0, and 4, it seems that the packing of -4 and 4 borders and shifts with the packing mechanism of residue 0. Asn at position -4 and Arg at position 4 are invariant while residue 0 is somewhat variable between 4 amino acids. Changes in specificity and base recognition likely involves residue 0 over residues Asn at -4 and Arg at 4. The amino acid occupying the 0th residue position is more involved in DNA recognition of protein knobs, and ridge 0 is more involved in DNA recognition of protein knobs than ridges +1 and -1. The packing region for residue 0 encapsulates base pairs 1 through 4, the nucleotide positions also involved in DNA packing into bZIP helices.

Table 2

DNA Base Socket Packing of Protein Residue Knobs at Residue 0

Base Pair Sequence		Residue 0 Packing Frequency			
		A	Q	S	V
[1, 2]	AA | TT	0/1			
	AT | TA	1/1			
	CA | GT	0/1			3/3
	CC | GG				4/4
	CT | GA	9/13		2/2	
	GA | CT				1/3
	GC | CG	0/2			8/18
	GM | CG			0/1	
	GM | MG	0/2			0/4
	GT | CA	11/26	2/2	1/1	
	TT | AA	1/1			
[2, 3]	AA | TT				6/6
	AC | TG	1/2			
	CA | GT	2/2			17/17
	CC | GG				5/5
	MA | GT	2/2			4/4
	MG | GM			1/1	
	TA | AT	2/2	2/2		
	TC | AG	33/39		3/3	
[3, 4]	AA | TT	5/5	2/2		27/27
	AG | TC	0/1			
	CA | GT	31/40		3/3	5/5
	CG | GM	0/1			
	GA | MT			1/1	

Note. Packing interactions are grouped by involved base pairs. For each half-site, packing interactions of residue 0 were grouped with the respective adjacent base pair. Each protein knob contacts at least three DNA residues in the base pair packing region to be considered significant to specificity, regardless of packing mechanism and orientation. The base pair pocket composition corresponds to the given base pair sequence (e.g., when [1, 2] has the sequence AA|TT, 1p and 2p are both dA while 1n and 2n are both dT). The packing frequency of the protein residue knob into the respective base pair pocket region per sequence is compared to the

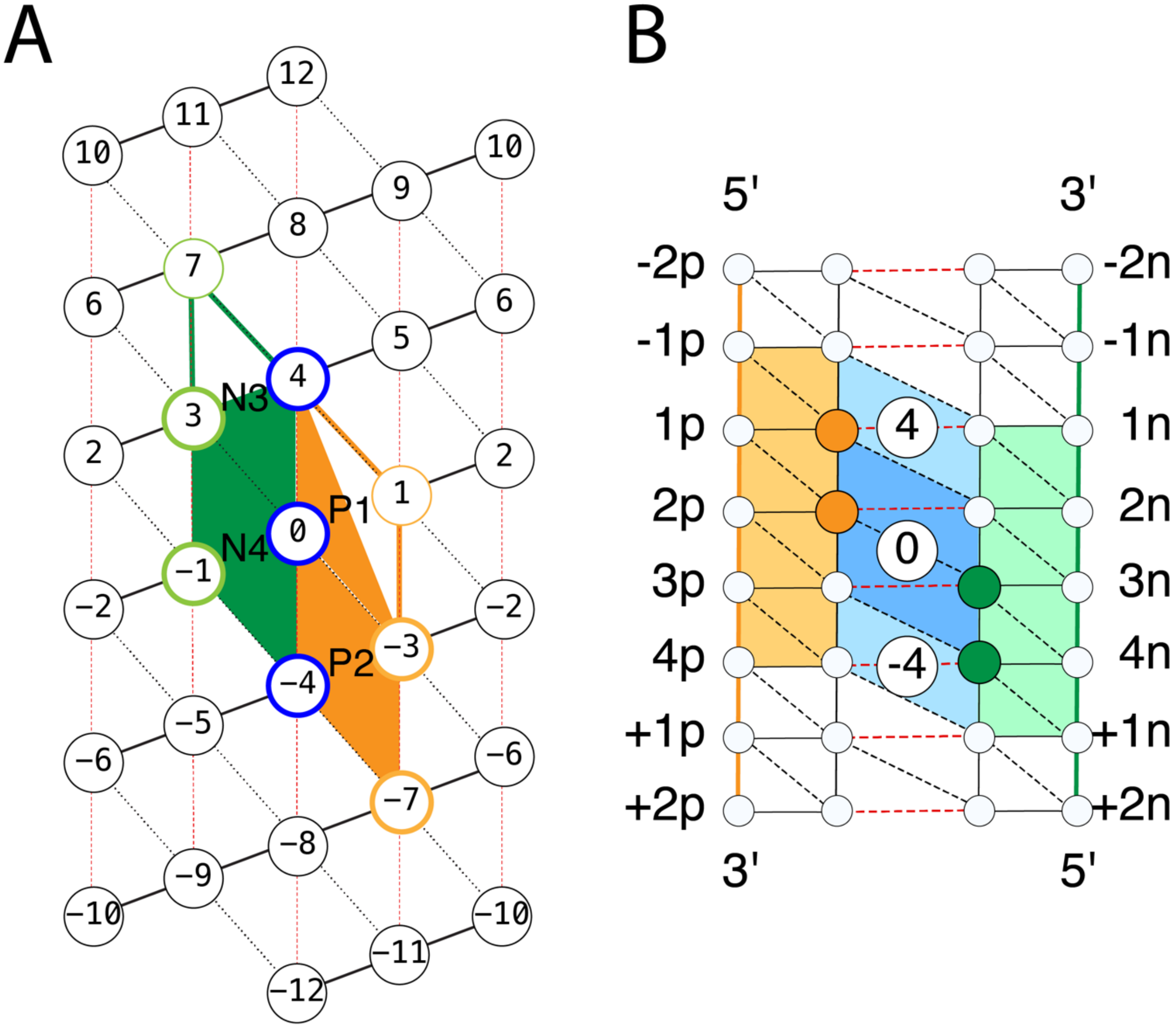
(Continued, Table 2) frequency the amino acid occupies the knob position 0, while the corresponding base pair pocket is composed of said DNA sequence. These tables reveal the propensities a protein packing region is likely to be filled by a specific DNA knob at the relative DNA position, where recognition of a DNA base differs based on region composition.

Table 2 includes all packing frequencies between base pair pocket and protein residue occupying position 0, analyzing how the protein knob packs into the three adjacent DNA pockets: A, V, Q, or S. Each table details contact patterns between a four-body pocket between the bases of adjacent base pairs. For example, [1, 2] describes the socket between 1p, 1n, 2p, and 2n, and is further described by sequence, where [1,2] CA|GT describes the socket formed from dC 1p, dG 1n, dA 2p, and dT 2n. The DNA pocket composition and location are displayed along with all found amino acid knobs for said region, analyzing the packing ratio for all residues at position 0. There is little specificity for particular amino acids in the DNA recognition of protein. This is expected since a 1 to 1 amino acid to nucleotide code has not been found (1,3). Comparing the number of protein residues involved in bZIP-DNA contact interfaces compared to the number of DNA nucleotides, recognition and specificity likely involves a greater number of protein residues over its DNA counterpart.

The bZip Recognition Core: Amino Acid Specific Packing of the Nucleotide Base 5-Methyl Group

Figure 11

Generalized bZIP-DNA Recognition Model



Note. A) The nine-residue bZIP α-helix recognition core of a DNA half-site is shown. The four pocket regions P1, P2, N3, and N4 packing the corresponding primary bases of the half-site 1p, 2p, 3n, and 4n form a quadripartite recognition region defined by nine residues: -7, -4, -3, -1, 0, 1, 3, 4, and 7. Pockets packing positive strand bases are colored in orange, while pockets packing

(Continued, Figure 11) negative strand bases are colored in green. Protein residues of the packing core are colored according to their *i+4* ridge: -1 ridge in green, 0 ridge in blue, and +1 ridge in orange. These colors also correspond to which packing lane on the DNA that these protein residues pack as knobs: residues -1, 3 and 7 pack into the positive lane; residues -4, 0 and 4 in the middle base lane; and residues -7, -3, and 1 in the negative lane. B) The DNA major groove lattice is shown for the half-site. The numbering is based on the four positive strand half-site nucleotides, where labels with "p" are positive strand, "n" are the negative strand, "-" indicate positions before the half-site and "+" indicate positions after the half-site. For further correlation with part A, the backbone is colored orange for the positive strand and negative for the negative strand. The first two bases 1p and 2p pack from the positive strand and are correspondingly colored in orange, and the last two bases 3n and 4n pack from the negative strand and are colored in green. The protein packing propensity into the DNA duplex is dependent on the four half-site base pairs. The packing region into the DNA half-site is where the majority of recognition is dependent on residue 0, the variable residue between -4, 0, and 4. These three protein residue knobs are shown in their general packing pocket regions in the middle base lane on the DNA lattice. Residue 8, despite packing into the middle lane, is far less frequent compared to the conserved three residues and is therefore not shown.

From the KS sequence-dependent packing analysis shown in Figures 9 and 10, a bZIP-DNA recognition model can be developed (Figure 11). The N-terminal basic domain of a bZIP α-helix contains a quadripartite recognition core of four contiguous KS pockets (Figure 11A). The recognition core surrounds the Asn at position -4 and Arg at position 4, two residues conserved throughout the bZIP family. Localizing DNA base packing per half-site, DNA recognition by a bZIP protein is therefore dependent on nine residues total: -7, -4, -3, -1, 0, 3, 4, 7 based on the relative numbering scheme, and includes three residues each from the +1, 0, and -1 *i+4* ridges, none of which are from the solvent ridge. Because they are contiguous to each other, there is a certain symmetry relating the residues in quadripartite sequence to pockets. All the recognition pockets include residue 0. Additionally, four residues are also shared between two adjacent pocket regions: P1 and P2 share residue -3; P2 and N4 residue -4; N4 and N3 residue 3; and N3 and P1 residue 4. Lastly, residues -7, -1, 1, and 7 are unique to pockets P2, N4, N3, and P1. As discussed above, though residues 1 and 7 do not consistently pack, these residues demonstrate influence on dT recognition and are therefore included in the model. Each KS pocket interacts with one of the four bases in the target DNA sequence half-site, but the bZIP α-helix recognizes DNA bases split evenly across both the positive and negative strands: first two bases on the positive strand and the last two bases on the negative strand (Figure 11B). Half-site recognition involves four base positions: 1p, 2p, 3n, and 4n. Bases 1p and 2p are adjacent nucleotides in the positive strand of the half-site, and 3n and 4n are adjacent nucleotides in the opposing negative strand. The primary mechanism of specificity is the packing or non-packing of 5-methyl groups of dT or 5mC. In contrast, bZIP protein residue recognition by the DNA half-site duplex is far more variable and inconsistent, centralizing around the middle residue between the invariant Asn and Arg. Figure 10 and Table 2 reveals inconsistent protein knob packing into DNA duplexes, contrasting the relatively consistent and frequent DNA base knob packing into bZIP helices (Figure 9 and Table 1). The most consistent and localized packing involves residues -4, 0, and 4, but based on their interactions, these residues primarily contribute to binding affinity rather than recognition specificity. Packing of Asn at position -4 and Arg at position 4 bordered and surrounded the packing pattern of residue 0, and likely does not contribute to specificity since both residues are highly conserved. Analysis of the single residue 0 does not yield a clear recognition of protein residues, implicating a far larger contribution to specificity by DNA packing into the recognition core on each bZIP helix than by protein packing into the DNA ML. The analysis does show that a Val at position 0 limits packing of only 2 dT into the recognition core.

DNA base packing interactions for dA, dG, and dC were inconsistent with no clear recognition pattern that discriminately packs any of the three nitrogenous bases, regardless of recognition core region. Due to the small number of crystal structures analyzed and bZIP-DNA combinations, specific packing and recognition of dA, dG, and dC could not be correlated to sequence dependencies between recognition core sequence and DNA half-site sequence. DNA base packing was consistent for dT and 5-methyl containing analogues, such as 5mC. Recognition of half-sites heavily involves recognition of the methyl group of dT and similar analogues, where dT is typically located at the 2p and 4n positions in the DNA sequence and packed into the P2 and N4 region of the recognition core. dT is also occasionally found at the 3n position packing into the N3 region, typically when dT does not occupy the 2p position.

CHAPTER 4: CONCLUSION

Identifying the mechanism of protein-DNA specificity is a well-recognized challenge even with the amount of experimental data and number of protein structures (1,3). As a step forward, the KS model was used to classify quaternary packing in bZIP-DNA co-crystal structures by considering multi-body interactions and lattice maps regularizing packing surfaces. The KS analysis finds that the bZIP-DNA recognition is primarily mediated through packing and non-packing of the 5-methyl group of dT, which is also found on the 5mC nucleotide (Figure 9 and Table 1). Consistent with previous investigations, the DNA does not recognize specific amino acids, but interactions contribute to binding affinity (Figure 10 and Table 2). While many protein residues in the basic region of a bZIP α-helix interact with the DNA backbone to stabilize binding Figure 9A, a quadripartite recognition core defined by nine amino acid residues was identified as the primary packing regions for the target DNA bases (Figures 9A,C and 11). Using the relative positions shown in Figure 9C, the recognition core surrounds invariant residues Asn at position -4 and Arg at position 4 and consists of four KS pockets named P1, P2, N3, and N4 that each bind one of the four DNA bases in a target sequence's half-site (Figure 11). Interestingly, the sockets made do not have a high propensity to be α-helical (Supplemental Figure S3), which may be important to the need for an unstructured state to sample target sequences in the charged DNA duplex. Packing occurs only with one base in the base pair, and the packed bases are split between the positive and negative strands. The first two DNA positions 1p and 2p on the positive strand pack into the P1 and P2 sites on the bZIP α-helix, and the last two DNA positions 3n and 4n on the negative strand pack into the N3 and N4 sites, respectively. Table 1 identifies packing and amino acid sequence preferences. At region P1, none of the bases pack or are significantly recognized. While dG is the primary residue found at P1, it only packs into the KS pocket 36 out of 59 times or 61%, and a similar result is found for dC 8 of 22 or 36%. In region P2, dT and dC are found predominantly, where dT always packs. Interestingly, an Asn at relative α-helix position -3 in this pocket consistently doesn't pack dT. Residue -3 is also shared with P1 pocket packing; however, an Asn does strongly influence P1 packing. KS pocket region N3 only packs dT and dG, where a Phe at position 7 seems not to favor dT packing. While position 7 doesn't directly pack in the N3 pocket, the bulkiness at this position allows recognition of dT. N4 only packs dT. So, dT is often packed into region P2 and N4. Prior studies emphasize the importance of the dT methyl group (15,53,56), which P2 and N4 recognize and pack. Similarly, 5mC also packs into the α-helix region P2 as well as N4. Of the five nucleotide types, dT and 5mC, which both have a methyl group, have a clear recognition mechanism unlike the other nucleotides. Throughout the four tables, dA, dG, and dC are shown to pack infrequently and inconsistently to any of the four regions in the quadripartite packing core. In terms of the packing regions, corroborating the findings shown by the Figure 9A, left, regions P1 and N3 are packed inconsistently. Regions P2 and N4 contribute more to packing and specificity due to their consistent packing patterns compared to the other two regions. Therefore, the primary mechanism found by the KS model is the preference of packing the 5-methyl group from dT.

References

1. Wolberger, C. (2021) How structural biology transformed studies of transcription regulation. *J Biol Chem*, **296**, 100741.
2. Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A. and Bonas, U. (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. *Science*, **326**, 1509-1512.
3. 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.
4. Joo, H., Chavan, A.G., Phan, J., Day, R. and Tsai, J. (2012) An amino acid packing code for α-helical structure and protein design. *J Mol Biol*, **419**, 234-254.
5. Joo, H. and Tsai, J. (2014) An amino acid code for β-sheet packing structure. *Proteins*, **82**, 2128-2140.
6. Joo, H., Chavan, A.G., Fraga, K.J. and Tsai, J. (2015) An amino acid code for irregular and mixed protein packing. *Proteins*, **83**, 2147–2161.
7. Fraga, K.J., Joo, H. and Tsai, J. (2016) An amino acid code to define a protein's tertiary packing surface. *Proteins*, **84**, 201-216.
8. Rodriguez-Martinez, J.A., Reinke, A.W., Bhimsaria, D., Keating, A.E. and Ansari, A.Z. (2017) Combinatorial bZIP dimers display complex DNA-binding specificity landscapes. *Elife*, **6**.
9. Hu, J.C. and Sauer, R.T. (1992) In Eckstein, F. and Lilley, D. M. J. (eds.), *Nucleic Acids and Molecular Biology*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 82-101.
10. Fujii, Y., Shimizu, T., Toda, T., Yanagida, M. and Hakoshima, T. (2000) Structural basis for the diversity of DNA recognition by bZIP transcription factors. *Nat Struct Biol*, **7**, 889-893.
11. Huber, E.M., Hortschansky, P., Scheven, M.T., Misslinger, M., Haas, H., Brakhage, A.A. and Groll, M. (2022) Structural insights into cooperative DNA recognition by the CCAAT-binding complex and its bZIP transcription factor HapX. *Structure*, **30**, 934-946 e934.
12. Keller, W., König, P. and Richmond, T.J. (1995) Crystal structure of a bZIP/DNA complex at 2.2 A: determinants of DNA specific recognition. *J Mol Biol*, **254**, 657-667.
13. Suckow, M., Schwamborn, K., Kisters-Woike, B., von Wilcken-Bergmann, B. and Muller-Hill, B. (1994) Replacement of invariant bZip residues within the basic region of the yeast transcriptional activator GCN4 can change its DNA binding specificity. *Nucleic Acids Res*, **22**, 4395-4404.
14. Syed, K.S., He, X., Tillo, D., Wang, J., Durell, S.R. and Vinson, C. (2016) 5-Methylcytosine (5mC) and 5-Hydroxymethylcytosine (5hmC) Enhance the DNA Binding of CREB1 to the C/EBP Half-Site Tetranucleotide GCAA. *Biochemistry*, **55**, 6940-6948.
15. Yang, J., Horton, J.R., Wang, D., Ren, R., Li, J., Sun, D., Huang, Y., Zhang, X., Blumenthal, R.M. and Cheng, X. (2019) Structural basis for effects of CpA modifications on C/EBPβ binding of DNA. *Nucleic Acids Res*, **47**, 1774-1785.
16. Bogdanove, A.J., Bohm, A., Miller, J.C., Morgan, R.D. and Stoddard, B.L. (2018) Engineering altered protein-DNA recognition specificity. *Nucleic Acids Res*, **46**, 4845-4871.
17. Ellenberger, T.E., Brandl, C.J., Struhl, K. and Harrison, S.C. (1992) The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted alpha helices: crystal structure of the protein-DNA complex. *Cell*, **71**, 1223-1237.
18. Glover, J.N. and Harrison, S.C. (1995) Crystal structure of the heterodimeric bZIP transcription factor c-Fos-c-Jun bound to DNA. *Nature*, **373**, 257-261.
19. Inukai, S., Kock, K.H. and Bulyk, M.L. (2017) Transcription factor-DNA binding: beyond binding site motifs. *Curr Opin Genet Dev*, **43**, 110-119.

20. John, M., Leppik, R., Busch, S.J., Granger-Schnarr, M. and Schnarr, M. (1996) DNA binding of Jun and Fos bZip domains: homodimers and heterodimers induce a DNA conformational change in solution. *Nucleic Acids Res*, **24**, 4487-4494.
21. Sayeed, S.K., Zhao, J., Sathyanarayana, B.K., Golla, J.P. and Vinson, C. (2015) C/EBPbeta (CEBPB) protein binding to the C/EBP|CRE DNA 8-mer TTGC|GTCA is inhibited by 5hmC and enhanced by 5mC, 5fC, and 5caC in the CG dinucleotide. *Biochim Biophys Acta*, **1849**, 583-589.
22. Kim, J., Tzamarias, D., Ellenberger, T., Harrison, S.C. and Struhl, K. (1993) Adaptability at the protein-DNA interface is an important aspect of sequence recognition by bZIP proteins. *Proc Natl Acad Sci U S A*, **90**, 4513-4517.
23. Pu, W.T. and Struhl, K. (1991) Highly conserved residues in the bZIP domain of yeast GCN4 are not essential for DNA binding. *Mol Cell Biol*, **11**, 4918-4926.
24. Suckow, M., von Wilcken-Bergmann, B. and Muller-Hill, B. (1993) Identification of three residues in the basic regions of the bZIP proteins GCN4, C/EBP and TAF-1 that are involved in specific DNA binding. *EMBO J*, **12**, 1193-1200.
25. Tzamarias, D., Pu, W.T. and Struhl, K. (1992) Mutations in the bZIP domain of yeast GCN4 that alter DNA-binding specificity. *Proc Natl Acad Sci U S A*, **89**, 2007-2011.
26. Bird, G.H., Lajmi, A.R. and Shin, J.A. (2002) Sequence-specific recognition of DNA by hydrophobic, alanine-rich mutants of the basic region/leucine zipper motif investigated by fluorescence anisotropy. *Biopolymers*, **65**, 10-20.
27. Johnson, P.F. (1993) Identification of C/EBP basic region residues involved in DNA sequence recognition and half-site spacing preference. *Mol Cell Biol*, **13**, 6919-6930.
28. Ray, S., Ufot, A., Assad, N., Singh, J., Durell, S.R., Porollo, A., Tillo, D. and Vinson, C. (2019) The bZIP mutant CEBPB (V285A) has sequence specific DNA binding propensities similar to CREB1. *Biochim Biophys Acta Gene Regul Mech*, **1862**, 486-492.
29. Rupert, P.B., Daughdrill, G.W., Bowerman, B. and Matthews, B.W. (1998) A new DNA-binding motif in the Skn-1 binding domain-DNA complex. *Nat Struct Biol*, **5**, 484-491.
30. Chan, I.S., Fedorova, A.V. and Shin, J.A. (2007) The GCN4 bZIP targets noncognate gene regulatory sequences: quantitative investigation of binding at full and half sites. *Biochemistry*, **46**, 1663-1671.
31. Etheve, L., Martin, J. and Lavery, R. (2017) Decomposing protein-DNA binding and recognition using simplified protein models. *Nucleic Acids Res*, **45**, 10270-10283.
32. Hortschansky, P., Ando, E., Tuppatsch, K., Arikawa, H., Kobayashi, T., Kato, M., Haas, H. and Brakhage, A.A. (2015) Deciphering the combinatorial DNA-binding code of the CCAAT-binding complex and the iron-regulatory basic region leucine zipper (bZIP) transcription factor HapX. *J Biol Chem*, **290**, 6058-6070.
33. Sarkar, A.K. and Lahiri, A. (2019) Dynamical Features of Cognate Site Recognition in bZIP–DNA Interaction. *ACS Omega*, **4**, 292-308.
34. Wang, Z., He, W., Tang, J. and Guo, F. (2020) Identification of Highest-Affinity Binding Sites of Yeast Transcription Factor Families. *J Chem Inf Model*, **60**, 1876-1883.
35. Schumacher, M.A., Goodman, R.H. and Brennan, R.G. (2000) The structure of a CREB bZIP.somatostatin CRE complex reveals the basis for selective dimerization and divalent cation-enhanced DNA binding. *J Biol Chem*, **275**, 35242-35247.
36. König, P. and Richmond, T.J. (1993) The X-ray structure of the GCN4-bZIP bound to ATF/CREB site DNA shows the complex depends on DNA flexibility. *J Mol Biol*, **233**, 139-154.
37. Chen, L., Glover, J.N., Hogan, P.G., Rao, A. and Harrison, S.C. (1998) Structure of the DNA-binding domains from NFAT, Fos and Jun bound specifically to DNA. *Nature*, **392**, 42-48.
38. Tahirov, T.H., Inoue-Bungo, T., Morii, H., Fujikawa, A., Sasaki, M., Kimura, K., Shiina, M., Sato, K., Kumasaka, T., Yamamoto, M. *et al.* (2001) Structural analyses of DNA recognition by the AML1/Runx-1 Runt domain and its allosteric control by CBFbeta. *Cell*, **104**, 755-767.
39. Tahirov, T.H., Inoue-Bungo, T., Sato, K., Sasaki, M., Ogata, K. (2002) Crystal structure of C/EBPBETA BZIP homodimer bound to a high affinity DNA fragment, https://www.rcsb.org/structure/1gu4.

40. Kim, Y., Podust, L.M. (2003) Crystal Structure of the Jun/CRE Complex, https://www.rcsb.org/structure/1jnm.
41. Miller, M., Shuman, J.D., Sebastian, T., Dauter, Z. and Johnson, P.F. (2003) Structural basis for DNA recognition by the basic region leucine zipper transcription factor CCAAT/enhancer-binding protein alpha. *J Biol Chem*, **278**, 15178-15184.
42. Tahirov, T.H., Inoue-Bungo, T., Sato, K., Sasaki, M., Ogata, K. (2003) Crystal structure of C/EBPBETA BZIP homodimer bound to a DNA fragment from the MIM-1 promoter, https://www.rcsb.org/structure/1gu5.
43. Panne, D., Maniatis, T. and Harrison, S.C. (2004) Crystal structure of ATF-2/c-Jun and IRF-3 bound to the interferon-beta enhancer. *Embo j*, **23**, 4384-4393.
44. Tahirov, T.H., Inoue-Bungo, T., Sato, K., Sasaki, M., Ogata, K. (2004) Crystal structure of C/EBPbeta bZip homodimer bound to a DNA fragment from the tom-1A promoter, https://www.rcsb.org/structure/1gtw.
45. Wang, D., Stroud, J.C., Chen, L. (2005) Crystal Structure of Human NFAT1 and Fos-Jun on the IL-2 ARRE1 Site, https://www.rcsb.org/structure/1s9k.
46. Petosa, C., Morand, P., Baudin, F., Moulin, M., Artero, J.B. and Müller, C.W. (2006) Structural basis of lytic cycle activation by the Epstein-Barr virus ZEBRA protein. *Mol Cell*, **21**, 565-572.
47. Kim, Y., Borovilos, M. (2007) Crystal structure of the JUN BZIP homodimer complexed with AP-1 DNA, https://www.rcsb.org/structure/2h7h.
48. Tahirov, T.H., Inoue-Bungo, T., Sato, K., Shiina, M., Hamada, K., Ogata, K. (2007) Crystal structure of C/EBPbeta Bzip homodimer V285A mutant bound to A High Affinity DNA fragment, https://www.rcsb.org/structure/2e42.
49. Textor, L.C., Wilmanns, M. and Holton, S.J. (2007) Expression, purification, crystallization and preliminary crystallographic analysis of the mouse transcription factor MafB in complex with its DNA-recognition motif Cmare. *Acta Crystallogr Sect F Struct Biol Cryst Commun*, **63**, 657-661.
50. Kurokawa, H., Motohashi, H., Sueno, S., Kimura, M., Takagawa, H., Kanno, Y., Yamamoto, M. and Tanaka, T. (2009) Structural basis of alternative DNA recognition by Maf transcription factors. *Mol Cell Biol*, **29**, 6232-6244.
51. Lu, X., Guanga, G.P., Wan, C. and Rose, R.B. (2012) A novel DNA binding mechanism for maf basic region-leucine zipper factors inferred from a MafA-DNA complex structure and binding specificities. *Biochemistry*, **51**, 9706-9717.
52. Pogenberg, V., Consani Textor, L., Vanhille, L., Holton, S.J., Sieweke, M.H. and Wilmanns, M. (2014) Design of a bZip transcription factor with homo/heterodimer-induced DNA-binding preference. *Structure*, **22**, 466-477.
53. Hong, S., Wang, D., Horton, J.R., Zhang, X., Speck, S.H., Blumenthal, R.M. and Cheng, X. (2017) Methyl-dependent and spatial-specific DNA recognition by the orthologous transcription factors human AP-1 and Epstein-Barr virus Zta. *Nucleic Acids Res*, **45**, 2503-2515.
54. Yin, Z., Machius, M., Nestler, E.J. and Rudenko, G. (2017) Activator Protein-1: redox switch controlling structure and DNA-binding. *Nucleic Acids Res*, **45**, 11425-11436.
55. Yang, J., Horton, J.R., Akdemir, K.C., Li, J., Huang, Y., Kumar, J., Blumenthal, R.M., Zhang, X. and Cheng, X. (2021) Preferential CEBP binding to T:G mismatches and increased C-to-T human somatic mutations. *Nucleic Acids Res*, **49**, 5084-5094.
56. Bernaudat, F., Gustems, M., Günther, J., Oliva, M.F., Buschle, A., Göbel, C., Pagniez, P., Lupo, J., Signor, L., Müller, C.W. *et al.* (2022) Structural basis of DNA methylation-dependent site selectivity of the Epstein-Barr virus lytic switch protein ZEBRA/Zta/BZLF1. *Nucleic Acids Res*, **50**, 490-511.
57. Sengoku, T., Shiina, M., Suzuki, K., Hamada, K., Sato, K., Uchiyama, A., Kobayashi, S., Oguni, A., Itaya, H., Kasahara, K. *et al.* (2022) Structural basis of transcription regulation by CNC family transcription factor, Nrf2. *Nucleic Acids Res*, **50**, 12543-12557.
58. Plaxco, K.W., Simons, K.T. and Baker, D. (1998) Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol*, **277**, 985-994.

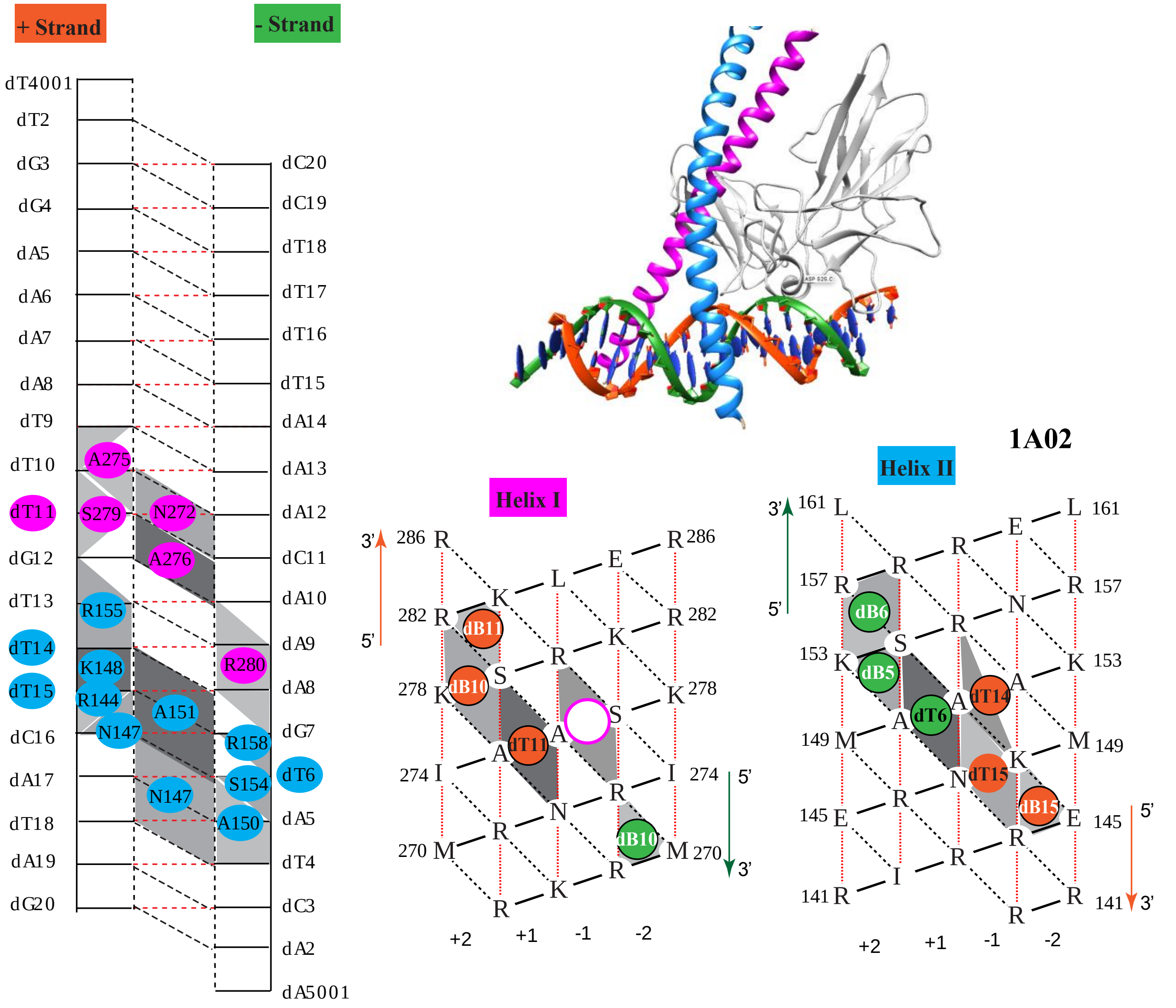
59. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C. and Ferrin, T.E. (2004) UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem*, **25**, 1605-1612.

Appendix

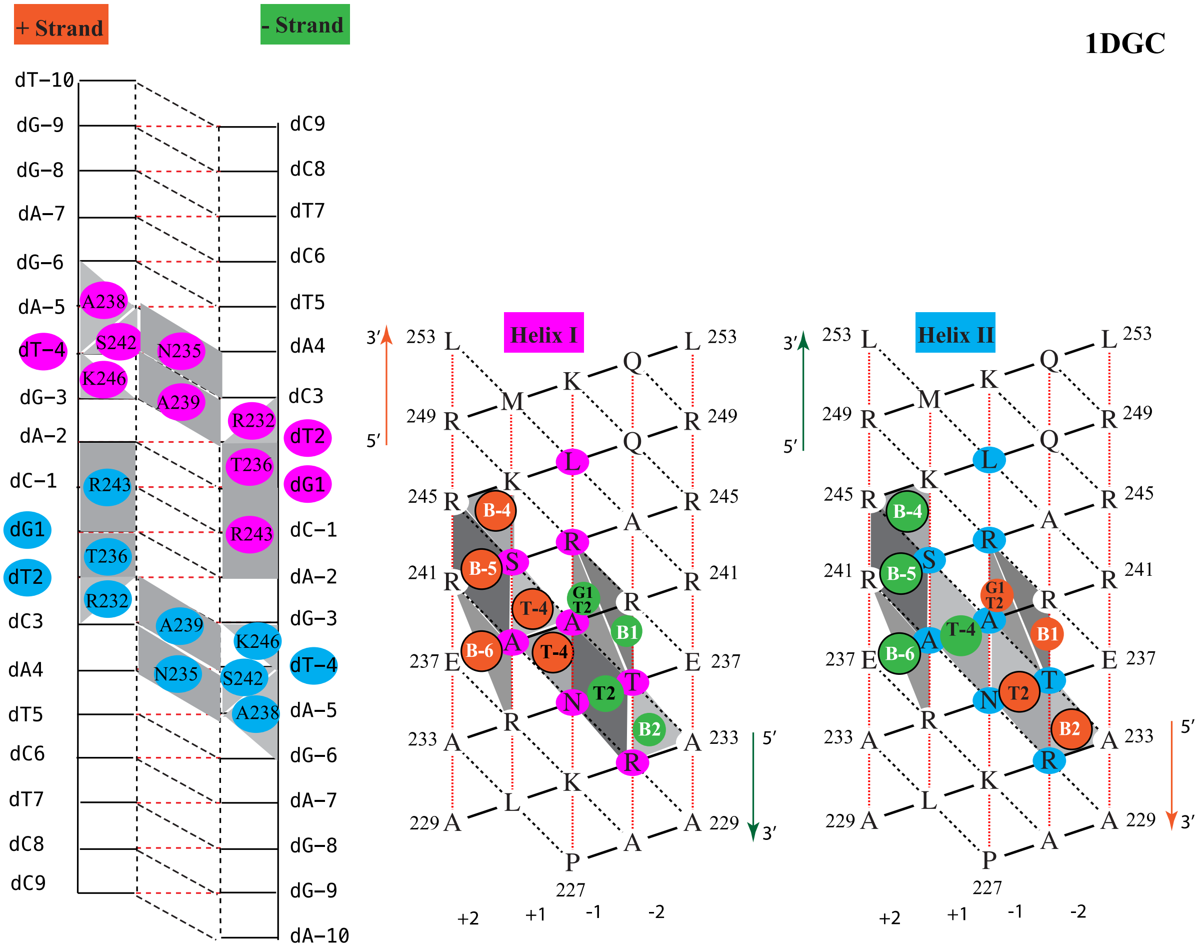
PDB ID	Name	DNA Target Sequence	Reference
1a02	Fos-Jun / AP-1	TGTTTCA	(37)
1dgc	GCN4	CRE	(36)
1dh3	CREB	CRE	(35)
1fos	c-Fos-c-Jun	TRE	(18)
1gd2	PAP1	TTACGTAA	(10)
1gtw	C/EBPβ	TTGCGCCA	(44)
1gu4	C/EBPβ	TTGCGCAA	(39)
1gu5	C/EBPβ	TTGGCCAA	(42)
1h88	C/EBPβ	TGGCGCAA	(39)
1h89	C/EBPβ	TGGCGCAA	(39)
1h8a	C/EBPβ	TGGCGCAA	(39)
1hjb	C/EBPβ	TTTCCAAA	(38)
1io4	C/EBPβ	TTTCCAAA	(38)
1jnm	c-Jun	CRE	(40)
1nwq	C/EBPα	TTGCGCAA	(41)
1s9k	Fos-Jun	TGTGTAA	(45)
1skn	Skn-1	GTCA	(29)
1t2k	ATF-2/c-Jun	TGACATAG	(43)
1ysa	GCN4	TRE	(17)
2c9l	BZLF1	TRE	(46)
2c9n	BZLF1	TRE	(46)
2dgc	GCN4	CRE	(12)
2e42	C/EBPβ (V285A)	TTGCGCAA	(48)
2e43	C/EBPβ (K269A)	TTGCGCAA	(48)
2h7h	Jun	AP-1/TRE	(47)
2wt7	MafB-c-Fos	TGACTCA	(52)
2wty	MafB	T-MARE	(52)
3a5t	MafG	TGAGTCA	(50)
4auw	MafB	C-MARE	(49)
4eot	MafA	MARE	(51)
5szx	Zta	TCGCTCA (methyl)	(53)
5t01	Jun	TGACTCG (methyl)	(53)
5vpe	FosB-Jun-D	TGACTCA	(54)
5vpf	FosB-Jun-D	TGACTCA	(54)

6mg1	C/EBPβ	TTGCGCAA	(15)
6mg2	C/EBPβ	TTGCGCAA	(15)
6mg3	C/EBPβ	TTGCGCAA	(15)

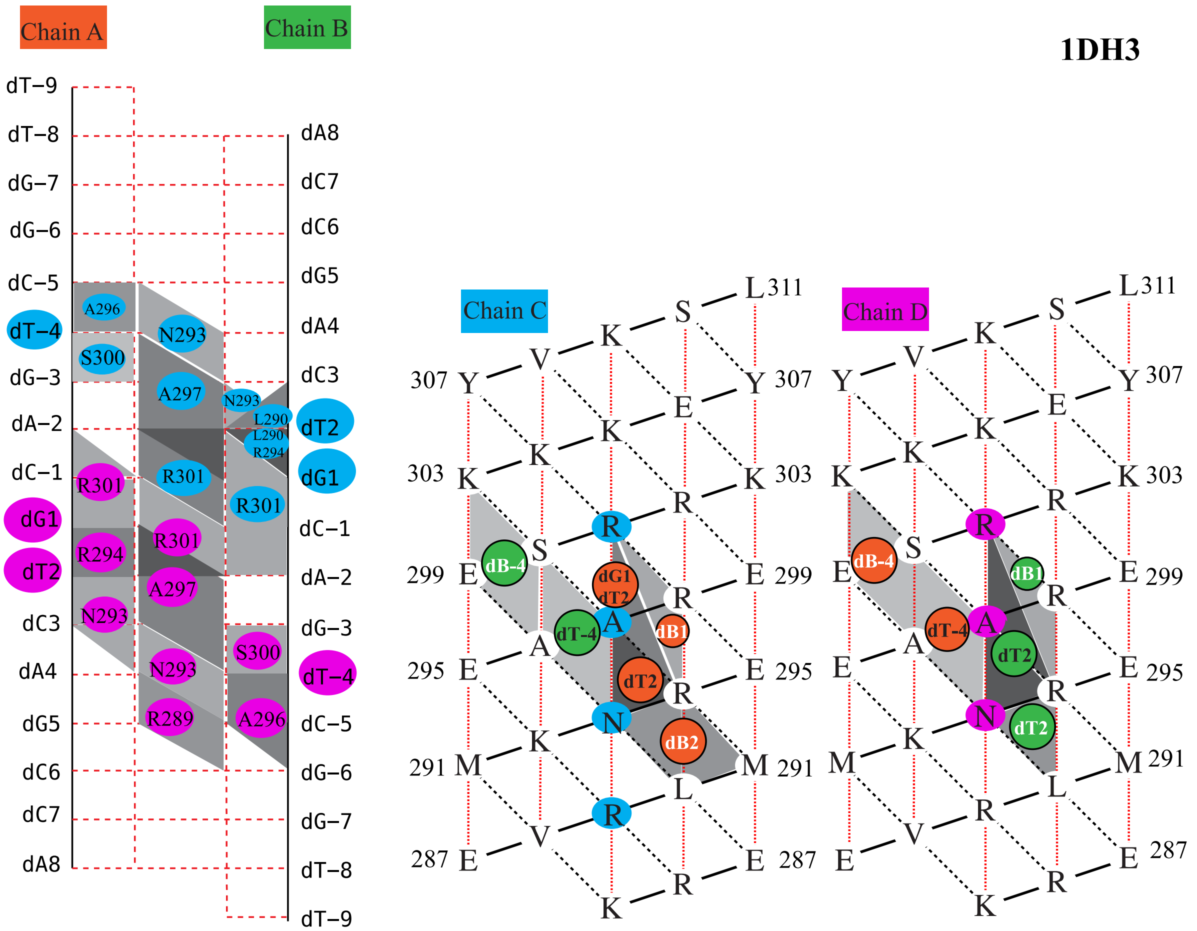
Table S1: List of PDB codes, bZIP names, target sequences, and their references.

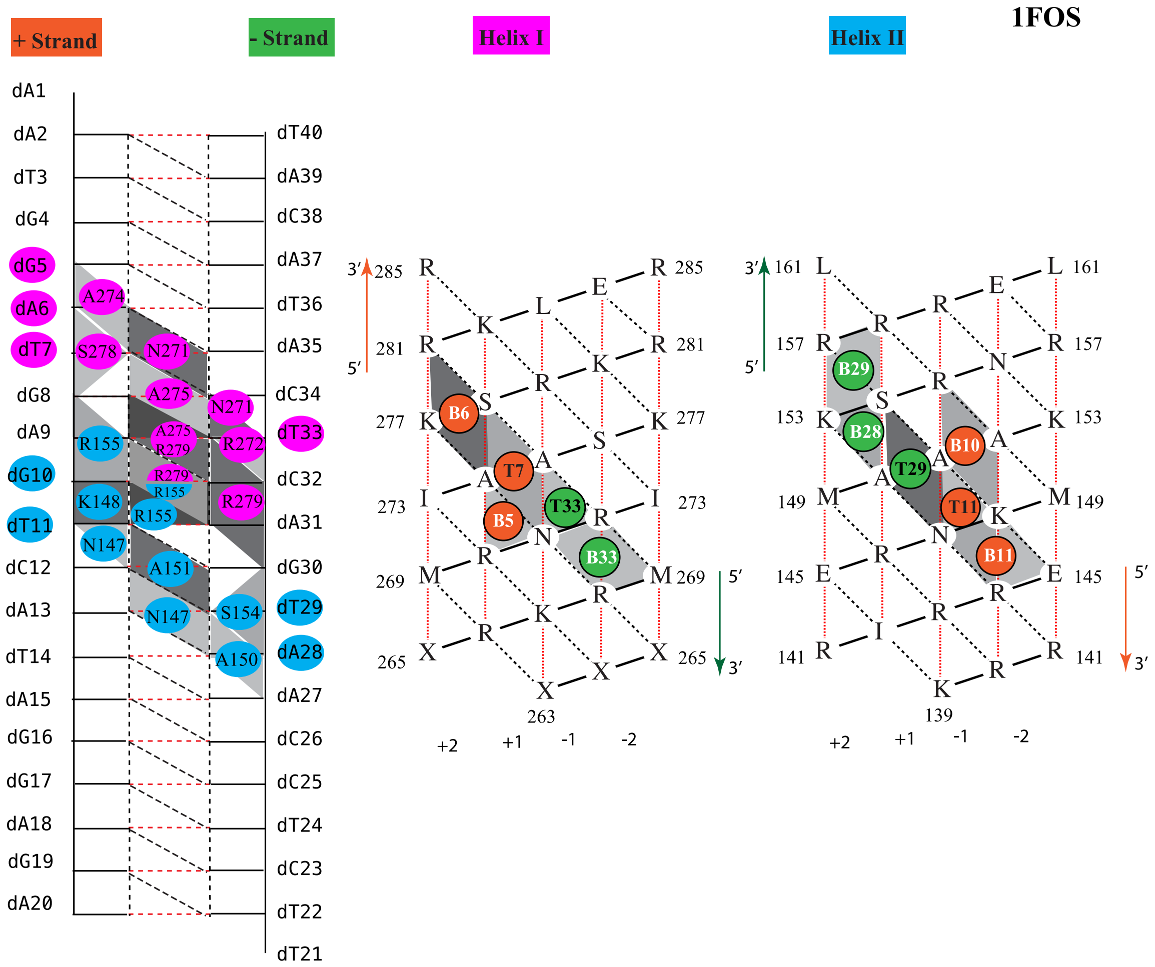

+ Strand
- Strand
dT4001
dT2
dG3
dG4
dA5
dA6
dA7
dA8
dT9
dT10
dT11
dG12
dT13
dT14
dT15
dC16
dA17
dT18
dA19
dG20
dC20
dC19
dT18
dT17
dT16
dT15
dA14
dA13
dA12
dC11
dA10
dA9
dA8
dG7
dT6
dA5
dT4
dC3
dA2
dA5001
A275
S279
N272
A276
R155
K148
R144
N147
A151
R280
R158
S154
N147
A150
ASP 526.C
1A02
Helix I
Helix II
3'
5'
286 R
R 286
K
L
E
282 R
R 282
dB11
S
R
K
278 K
K 278
dB10
A
dT11
A
S
274 I
I 274
R
N
R
270 M
M 270
dB10
R
K
R
+2
+1
-1
-2
161 L
L 161
R
E
R
157 R
R 157
N
dB6
S
R
153 K
K 153
dB5
A
dT14
A
dT6
A
149 M
M 149
K
N
dT15
145 E
E 145
R
dB15
R
141 R
I
R
R
R 141
+2
+1
-1
-2

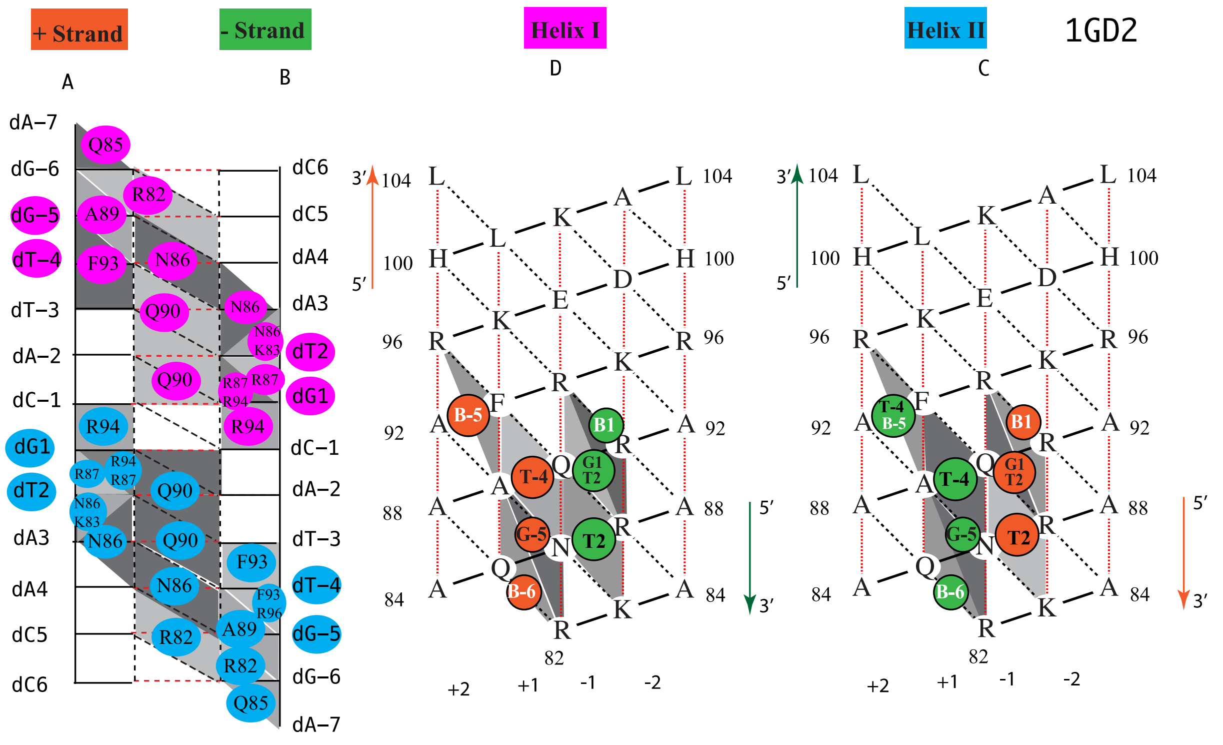
1DGC

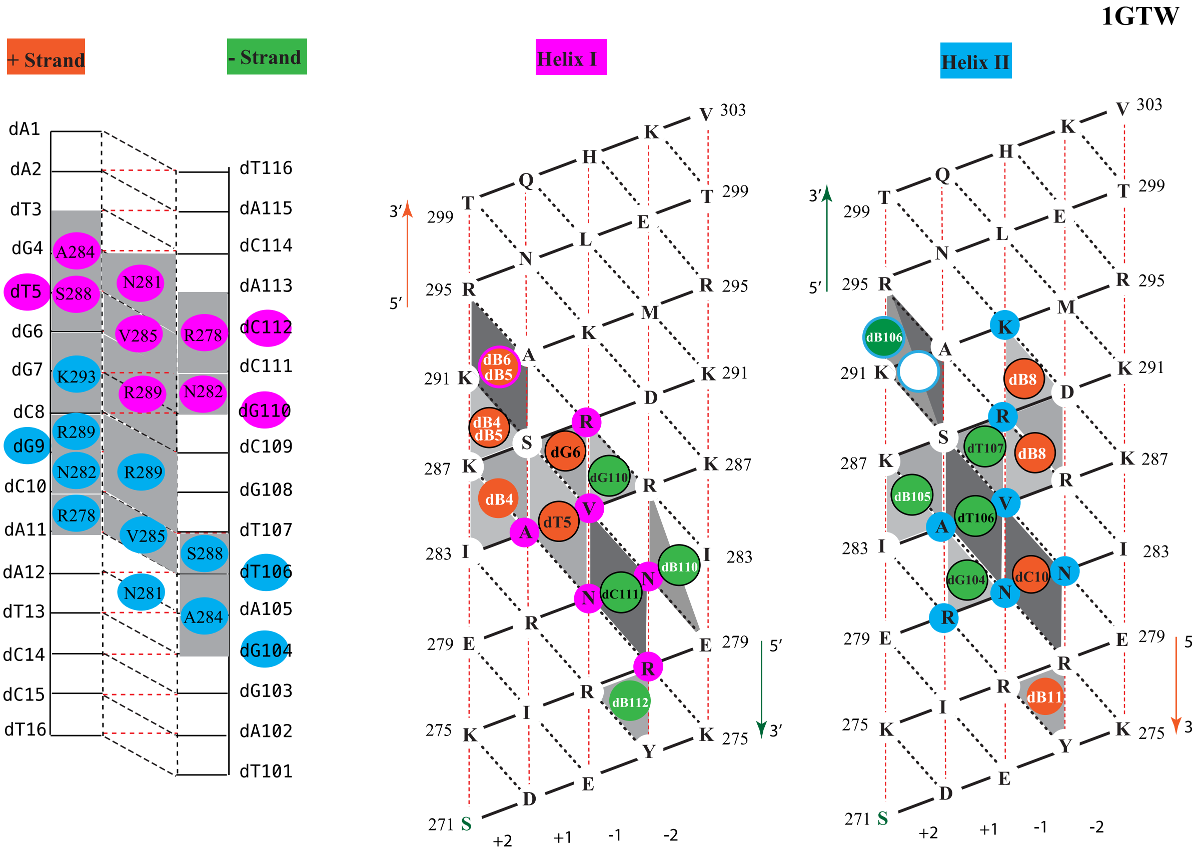

+ Strand
- Strand
dT−10
dG−9
dG−8
dA−7
dG−6
dA−5
dT−4
dG−3
dA−2
dC−1
dG1
dT2
dC3
dA4
dT5
dC6
dT7
dC8
dC9
dA−10
A238
S242
K246
N235
A239
R232
T236
R243
K246
S242
Helix I
Helix II
3′
5′
253
249
245
241
237
233
229
227
L
M
K
Q
R
A
E
S
T
N
P
B-4
B-5
B-6
T-4
G1
T2
B1
B2
+2
+1
-1
-2

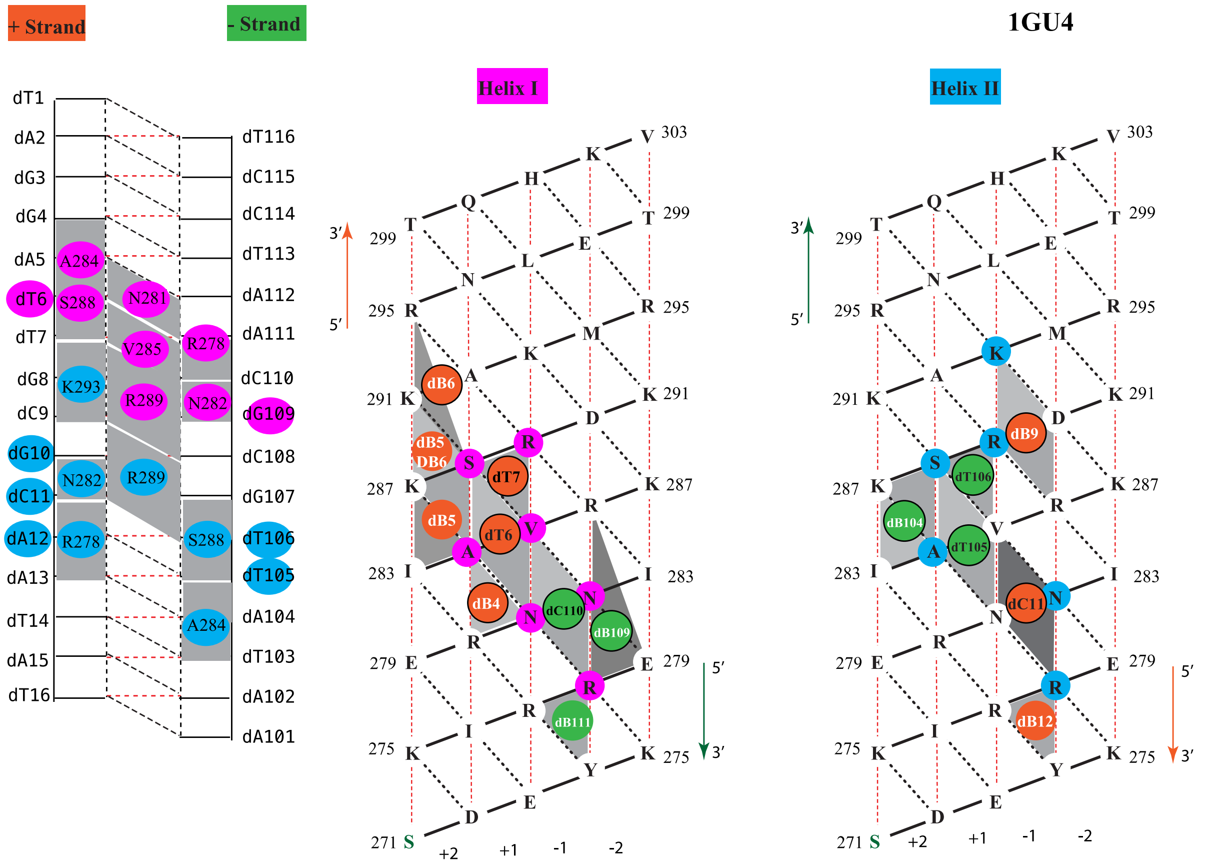
1DH3

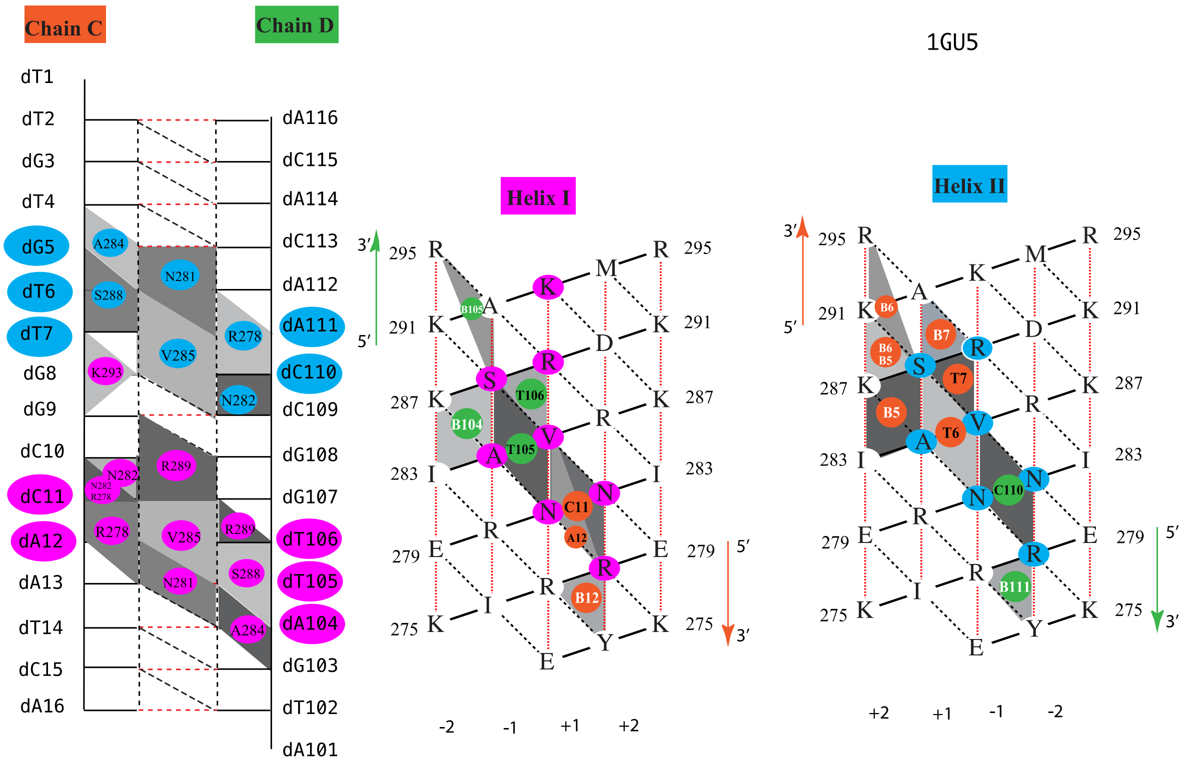

Chain A
Chain B
dT−9
dT−8
dG−7
dG−6
dC−5
dT−4
dG−3
dA−2
dC−1
dG1
dT2
dC3
dA4
dG5
dC6
dC7
dA8
dA8
dC7
dC6
dG5
dA4
dC3
dT2
dG1
dC−1
dA−2
dG−3
dT−4
dC−5
dG−6
dG−7
dT−8
dT−9
A296
S300
N293
A297
N293
L290
L290
R294
R301
R301
R301
R294
R301
A297
N293
N293
R289
S300
A296
Chain C
Chain D
L 311
S
K
V
Y 307
E
K
K 303
R
R
E 299
S
A
dB-4
dG1
dT2
dB1
dT-4
E 295
dT2
N
dB2
M 291
L
R
V
E 287
K
307
303
299
295
291
287
dB1
dT2
dT2

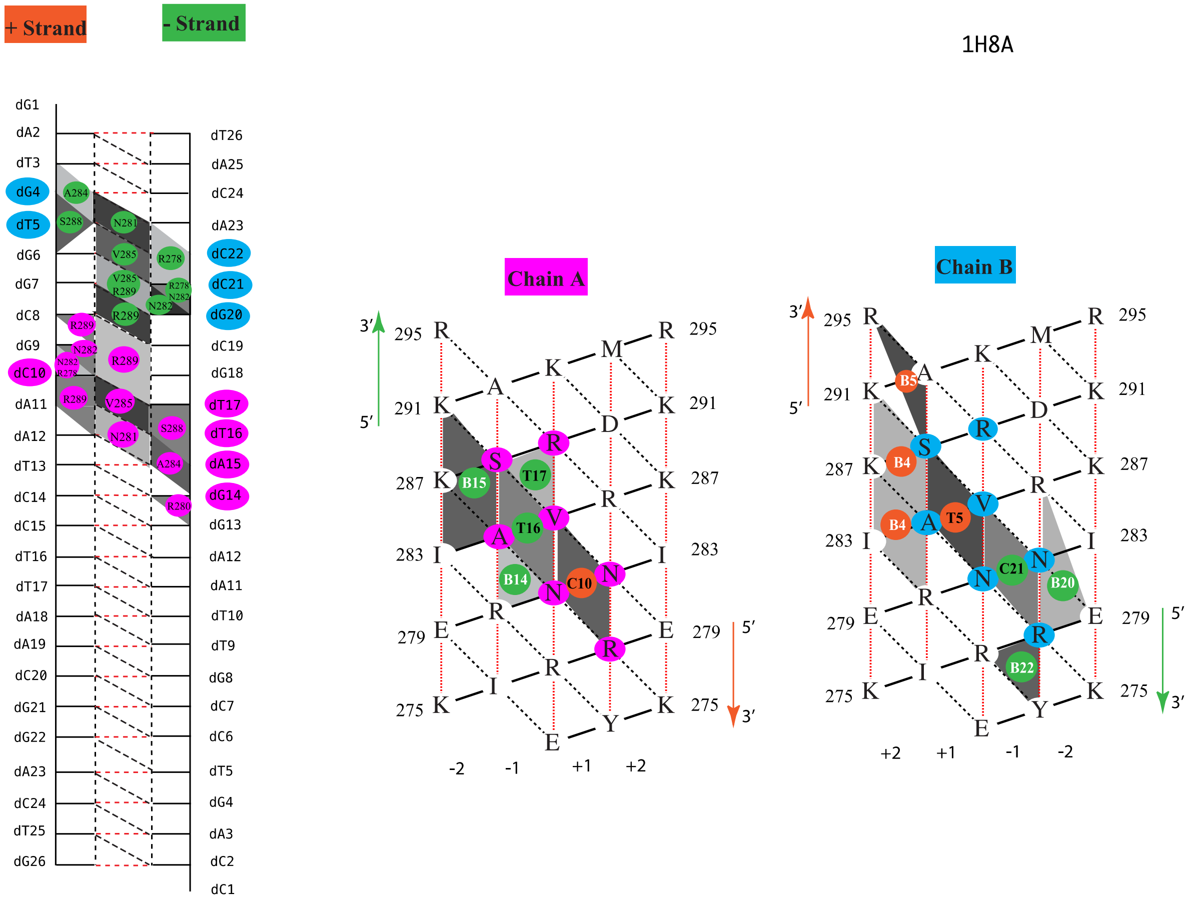

1FOS
+ Strand
- Strand
Helix I
Helix II
dA1
dA2
dT3
dG4
dG5
dA6
dT7
dG8
dA9
dG10
dT11
dC12
dA13
dT14
dA15
dG16
dG17
dA18
dG19
dA20
dT40
dA39
dC38
dA37
dT36
dA35
dC34
dT33
dC32
dA31
dG30
dT29
dA28
dA27
dC26
dC25
dT24
dC23
dT22
dT21
A274
S278
N271
A275
N271
A275
R279
R272
R279
R155
R279
R155
K148
R155
R279
N147
A151
N147
S154
A150
3′
5′
285
281
277
273
269
265
263
R
K
L
E
R
R
K
S
R
K
K
S
A
R
K
A
S
I
I
M
R
N
R
M
X
R
K
R
X
X
X
B6
T7
B5
T33
B33
5′
3′
+2
+1
-1
-2
3′
5′
161
157
153
149
145
141
139
L
R
E
L
R
R
N
R
K
S
R
A
K
M
A
N
K
M
E
R
R
E
R
I
R
R
R
K
R
B29
B28
T29
B10
T11
B11
5′
3′
+2
+1
-1
-2


+ Strand
- Strand
Helix I
Helix II
1GD2
A
B
D
C
dA−7
dG−6
dG−5
dT−4
dT−3
dA−2
dC−1
dG1
dT2
dA3
dA4
dC5
dC6
dC6
dC5
dA4
dA3
dT2
dG1
dC−1
dA−2
dT−3
dT−4
dG−5
dG−6
dA−7
Q85
R82
A89
F93
N86
Q90
N86
N86
K83
R87
R87
R94
R94
R94
R87
N86
K83
N86
Q90
Q90
N86
F93
F93
R96
A89
R82
R82
Q85
3′
5′
104
100
96
92
88
84
82
L
H
R
A
A
A
L
K
E
R
F
A
Q
N
R
K
A
D
K
R
Q
R
R
K
L
H
R
A
A
A
B-5
T-4
G-5
B-6
B1
G1
T2
T2
T-4
B-5
T-4
G-5
B-6
B1
G1
T2
T2
5′
3′
+2
+1
-1
-2

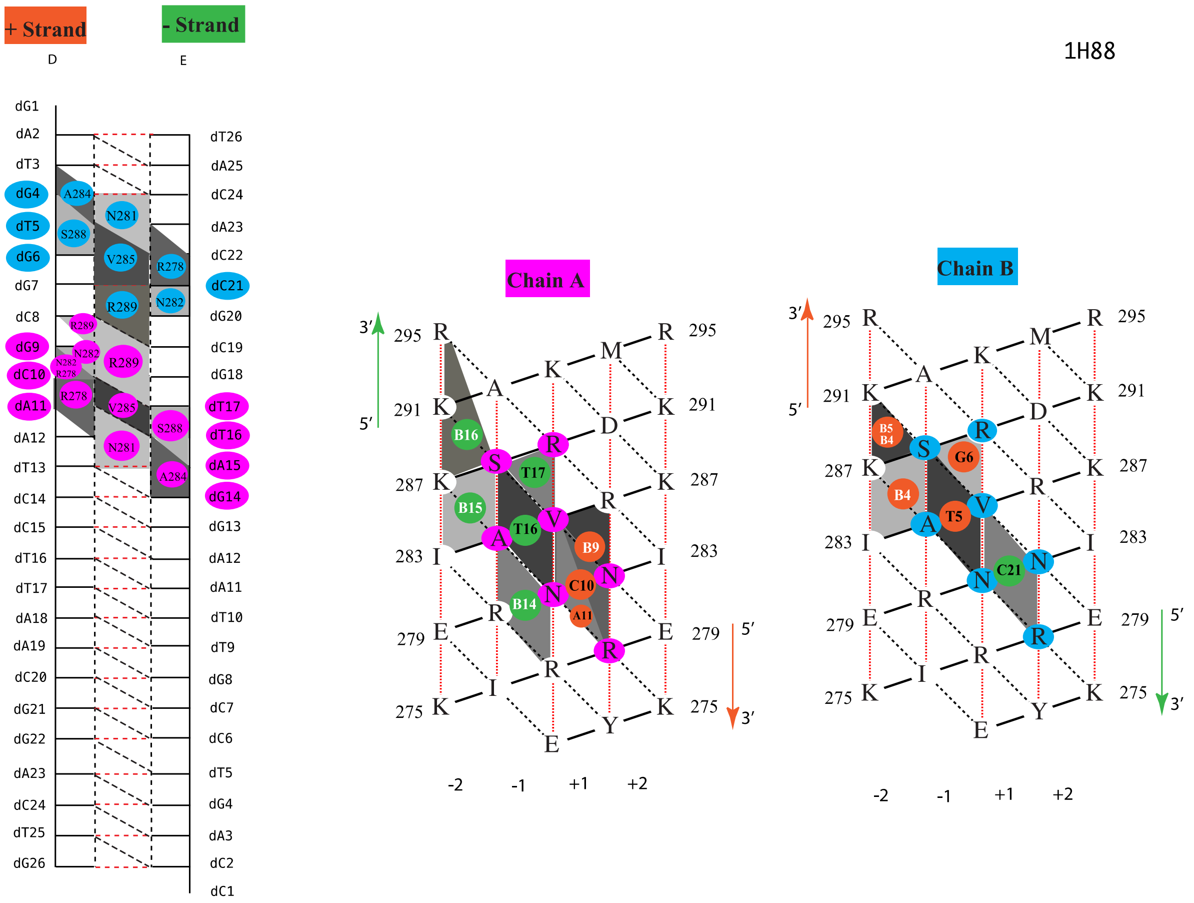

1GTW
+ Strand
- Strand
Helix I
Helix II
dA1
dA2
dT3
dG4
dT5
dG6
dG7
dC8
dG9
dC10
dA11
dA12
dT13
dC14
dC15
dT16
dT116
dA115
dC114
dA113
dC112
dC111
dG110
dC109
dG108
dT107
dT106
dA105
dG104
dG103
dA102
dT101
A284
S288
N281
V285
R278
K293
R289
N282
R289
N282
R289
R278
V285
S288
N281
A284
3′
5′
303
299
295
291
287
283
279
275
271
V
T
R
K
K
I
E
K
S
Q
N
A
S
A
N
R
I
D
H
L
K
R
V
R
E
E
M
D
R
I
N
R
Y
K
T
R
K
K
I
E
K
dB6
dB5
dB4
dB5
dB4
dG6
dT5
dG110
dB110
dC111
dB112
dB106
dB8
dB8
dT107
dB105
dT106
dG104
dC10
dB11
5′
3′
+2
+1
-1
-2

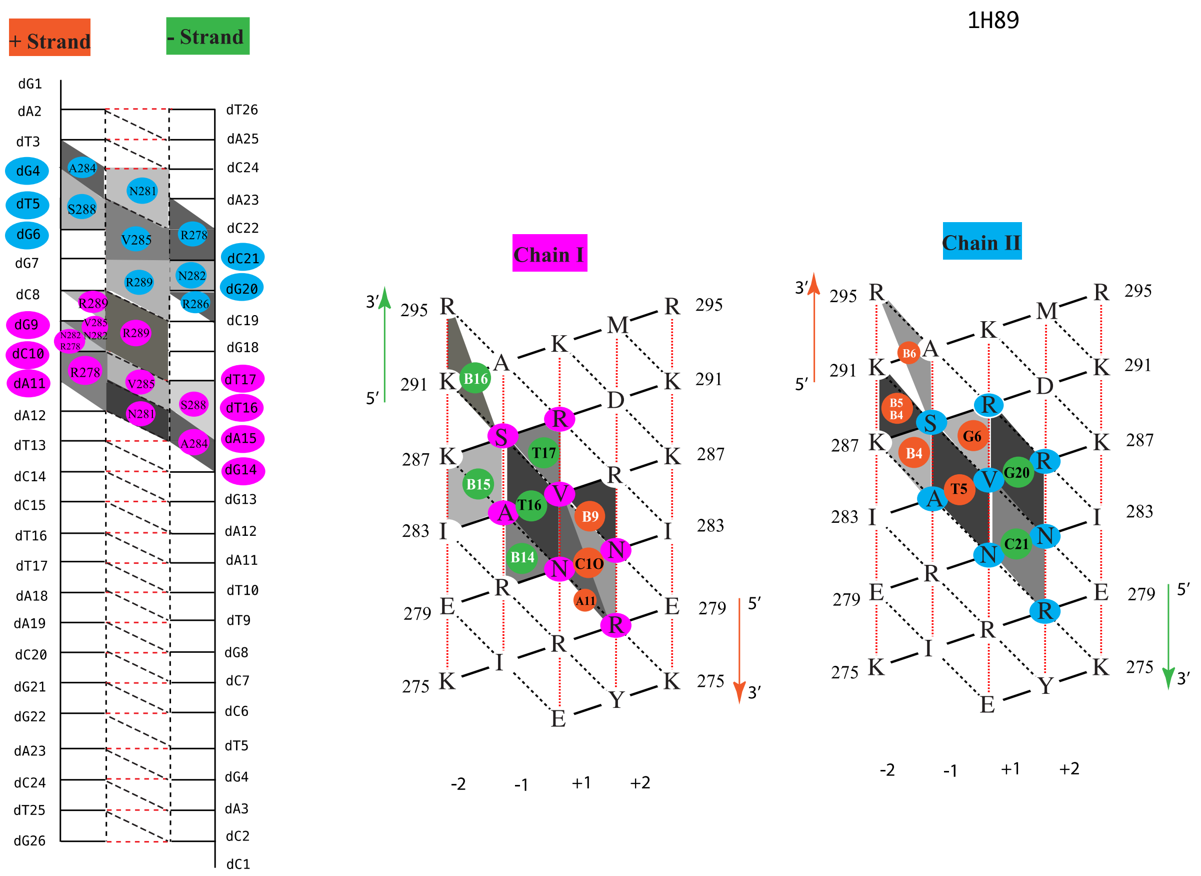

+ Strand
- Strand
1GU4
Helix I
Helix II
dT1
dA2
dG3
dG4
dA5
dT6
dT7
dG8
dC9
dG10
dC11
dA12
dA13
dT14
dA15
dT16
dT116
dC115
dC114
dT113
dA112
dA111
dC110
dG109
dC108
dG107
dT106
dT105
dA104
dT103
dA102
dA101
A284
S288
N281
V285
R278
K293
R289
N282
N282
R289
R278
S288
A284
3′
5′
299
295
291
287
283
279
275
271
303
V
K
H
Q
T
E
L
N
R
M
A
D
S
I
Y
dB6
dB5
DB6
dT7
dB5
dT6
dB4
dC110
dB109
dB111
dB9
dT106
dB104
dT105
dC11
dB12
+2
+1
-1
-2


Chain C
Chain D
1GU5
dT1
dT2
dG3
dT4
dG5
dT6
dT7
dG8
dG9
dC10
dC11
dA12
dA13
dT14
dC15
dA16
dA116
dC115
dA114
dC113
dA112
dA111
dC110
dC109
dG108
dG107
dT106
dT105
dA104
dG103
dT102
dA101
A284
S288
N281
V285
R278
N282
K293
R289
N282
N282
R278
R278
V285
R289
N281
S288
A284
Helix I
3′
5′
295
291
287
283
279
275
R
K
K
I
E
K
A
R
I
S
A
R
N
R
E
K
R
V
N
R
Y
M
D
R
I
E
B105
T106
B104
T105
C11
A12
B12
5′
3′
-2
-1
+1
+2
Helix II
3′
5′
B6
B6
B5
B7
S
R
T7
B5
A
V
T6
N
C110
N
R
B111
D
R
I
E
K
K
+2
+1
-1
-2

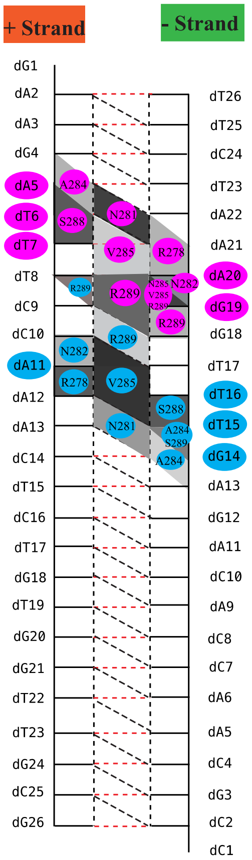

1H8A
+ Strand
- Strand
dG1
dA2
dT3
dG4
dT5
dG6
dG7
dC8
dG9
dC10
dA11
dA12
dT13
dC14
dC15
dT16
dT17
dA18
dA19
dC20
dG21
dG22
dA23
dC24
dT25
dG26
dT26
dA25
dC24
dA23
dC22
dC21
dG20
dC19
dG18
dT17
dT16
dA15
dG14
dG13
dA12
dA11
dT10
dT9
dG8
dC7
dC6
dT5
dG4
dA3
dC2
dC1
A284
S288
N281
V285
R278
V285
R289
R278
N282
R289
R289
N282
N282
R278
R289
R289
V285
S288
N281
A284
R280
Chain A
Chain B
3′
5′
5′
3′
295
291
287
283
279
275
R
K
A
K
M
R
D
S
R
K
B15
T17
V
R
A
T16
I
N
B14
C10
N
E
R
R
E
K
I
R
Y
K
E
-2
-1
+1
+2
B5
B4
S
B4
A
T5
V
C21
N
N
B20
R
B22
Y
+2
+1
-1
-2

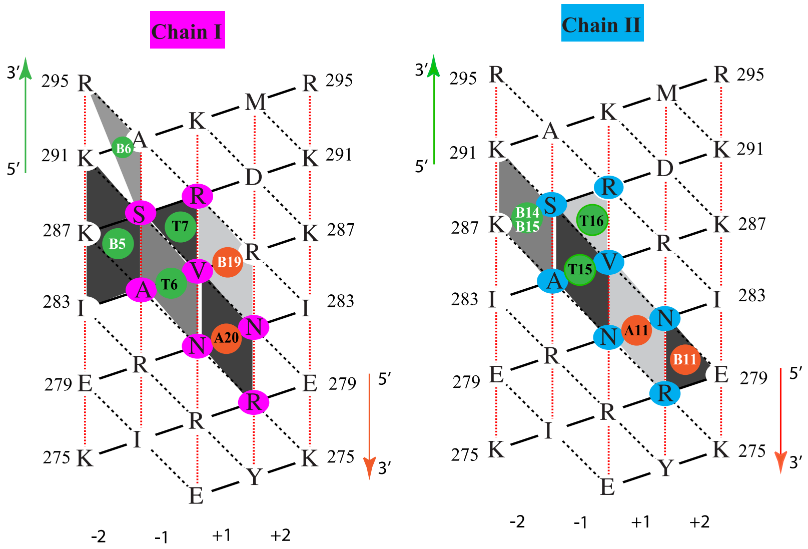
1H88


+ Strand
- Strand
D
E
dG1
dA2
dT3
dG4
dT5
dG6
dG7
dC8
dG9
dC10
dA11
dA12
dT13
dC14
dC15
dT16
dT17
dA18
dA19
dC20
dG21
dG22
dA23
dC24
dT25
dG26
dT26
dA25
dC24
dA23
dC22
dC21
dG20
dC19
dG18
dT17
dT16
dA15
dG14
dG13
dA12
dA11
dT10
dT9
dG8
dC7
dC6
dT5
dG4
dA3
dC2
dC1
A284
N281
S288
V285
R278
R289
N282
R289
N282
R278
R289
R278
V285
S288
N281
A284
Chain A
3′
5′
295
291
287
283
279
275
R
K
K
I
E
K
A
R
I
E
S
A
R
N
R
K
M
D
R
N
R
Y
R
K
I
E
K
B16
B15
B14
T17
T16
B9
C10
A11
5′
3′
-2
-1
+1
+2
Chain B
3′
5′
295
291
287
283
279
275
R
K
K
I
E
K
A
S
A
R
I
E
K
R
V
N
R
Y
M
D
R
N
R
K
K
I
E
K
B5
B4
B4
G6
T5
C21
5′
3′
-2
-1
+1
+2

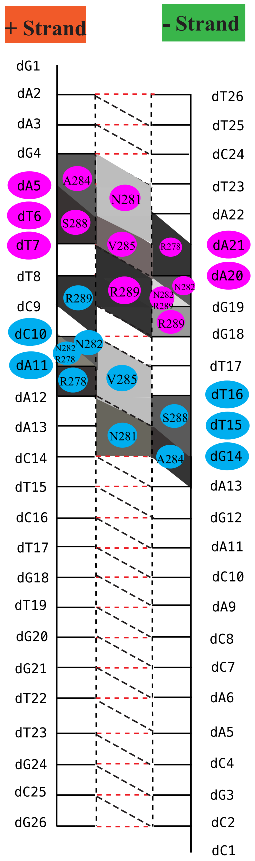

1H89
+ Strand
- Strand
dG1
dA2
dT3
dG4
dT5
dG6
dG7
dC8
dG9
dC10
dA11
dA12
dT13
dC14
dC15
dT16
dT17
dA18
dA19
dC20
dG21
dG22
dA23
dC24
dT25
dG26
dT26
dA25
dC24
dA23
dC22
dC21
dG20
dC19
dG18
dT17
dT16
dA15
dG14
dG13
dA12
dA11
dT10
dT9
dG8
dC7
dC6
dT5
dG4
dA3
dC2
dC1
A284
N281
S288
V285
R278
R289
N282
R286
R289
V285
N282
N282
R278
R289
R278
V285
S288
N281
A284
Chain I
3′
5′
295
291
287
283
279
275
R
A
K
M
R
K
B16
S
R
D
K
K
T17
B15
V
T16
A
R
K
B9
I
N
C10
B14
N
I
E
R
A11
R
E
K
I
R
Y
K
E
295
291
287
283
279
275
5′
3′
-2
-1
+1
+2
Chain II
3′
5′
295
291
287
283
279
275
R
B6
A
K
M
R
K
B5
B4
R
D
K
S
G6
B4
R
K
A
T5
V
G20
I
R
N
C21
N
I
E
R
R
E
K
I
R
Y
K
E
295
291
287
283
279
275
5′
3′
-2
-1
+1
+2

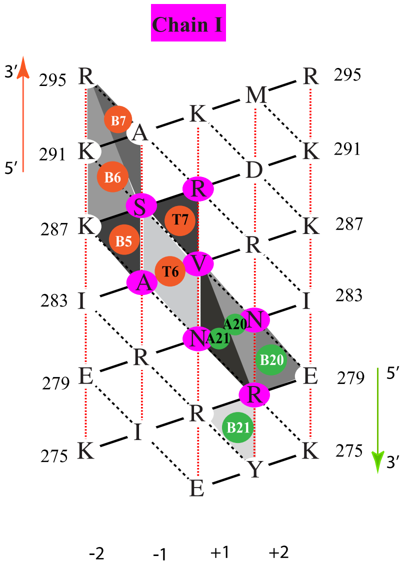
1HJB

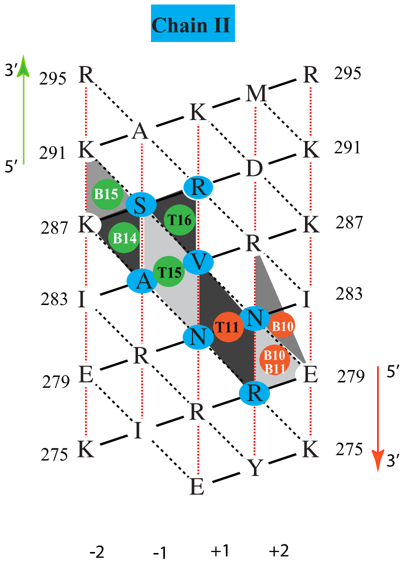

+ Strand
- Strand
dG1
dA2
dA3
dG4
dA5
dT6
dT7
dT8
dC9
dC10
dA11
dA12
dA13
dC14
dT15
dC16
dT17
dG18
dT19
dG20
dG21
dT22
dT23
dG24
dC25
dG26
dT26
dT25
dC24
dT23
dA22
dA21
dA20
dG19
dG18
dT17
dT16
dT15
dG14
dA13
dG12
dA11
dC10
dA9
dC8
dC7
dA6
dA5
dC4
dG3
dC2
dC1
A284
N281
S288
V285
R278
R289
R289
N282
R289
R289
N282
R278
V285
S288
N281
A284
S289
A284

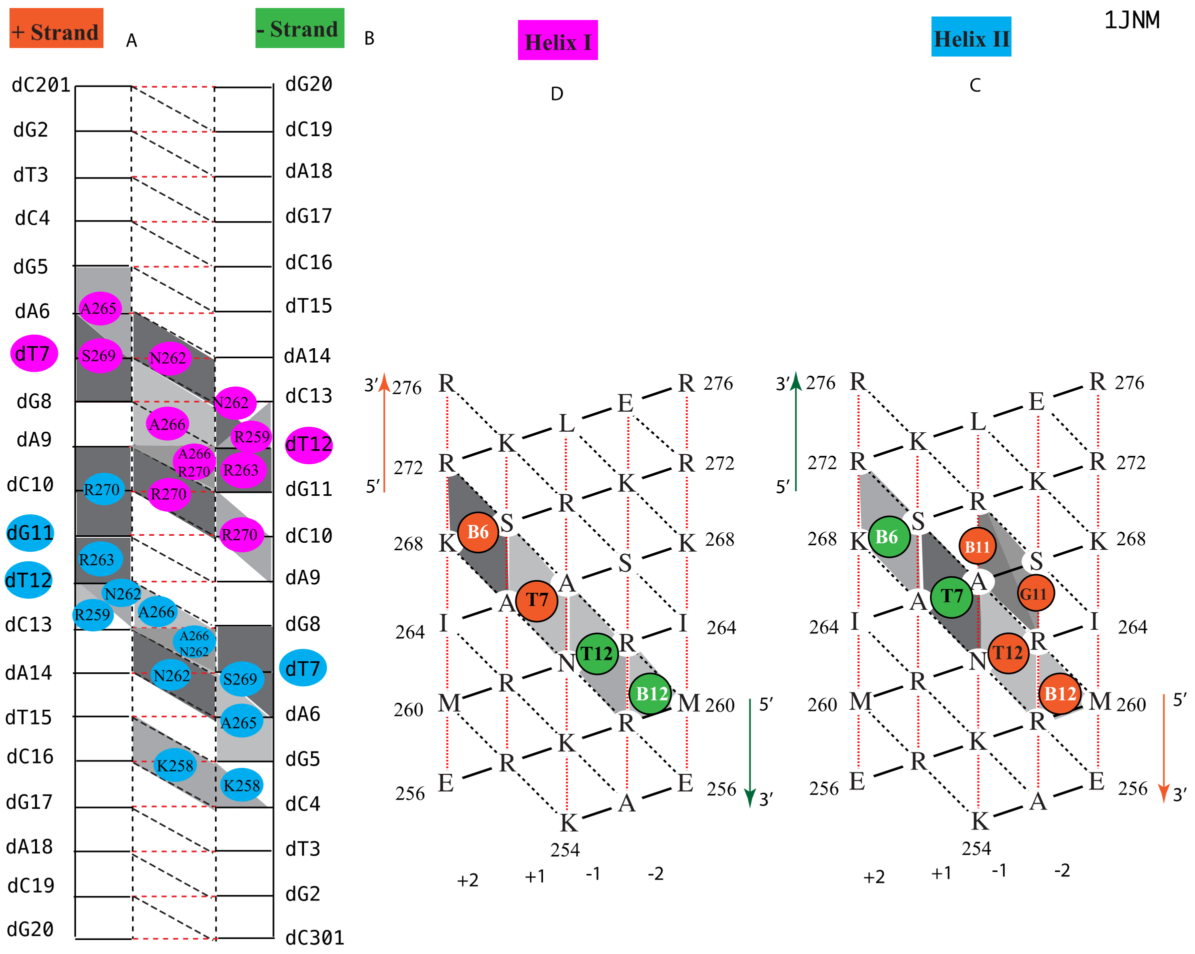

Chain I
Chain II
3′
5′
5′
3′
295
291
287
283
279
275
R
K
K
I
E
K
A
K
M
R
D
K
S
R
T7
B6
B5
A
T6
V
B19
R
N
N
A20
R
I
E
R
I
R
Y
E
-2
-1
+1
+2
T16
B14
B15
T15
A11
B11

1I04

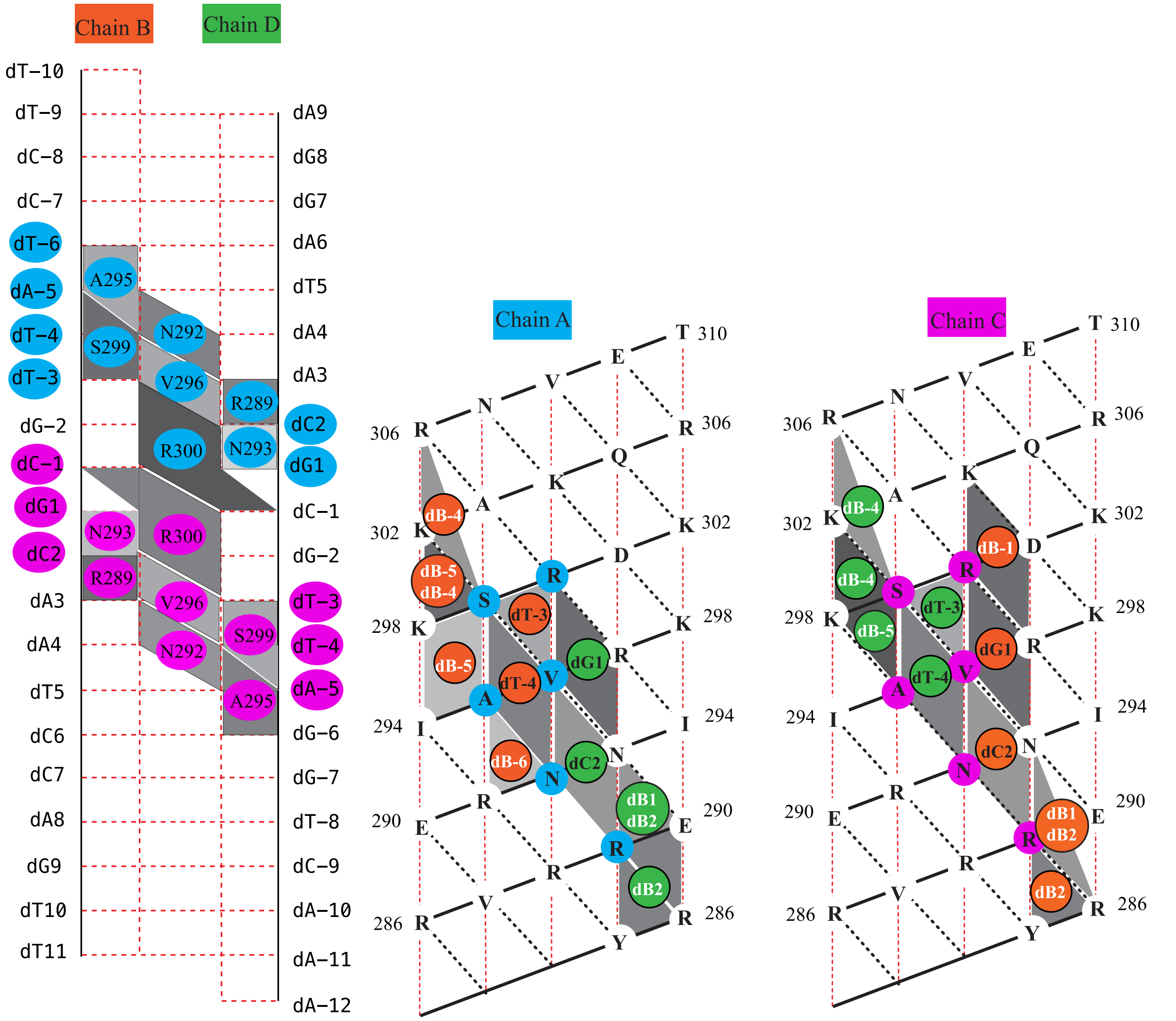

+ Strand
- Strand
dG1
dA2
dA3
dG4
dA5
dT6
dT7
dT8
dC9
dC10
dA11
dA12
dA13
dC14
dT15
dC16
dT17
dG18
dT19
dG20
dG21
dT22
dT23
dG24
dC25
dG26
dT26
dT25
dC24
dT23
dA22
dA21
dA20
dG19
dG18
dT17
dT16
dT15
dG14
dA13
dG12
dA11
dC10
dA9
dC8
dC7
dA6
dA5
dC4
dG3
dC2
dC1
A284
N281
S288
V285
R278
R289
N282
R289
N282
R278
R289
R289
N282
N282
R278
R278
V285
S288
N281
A284

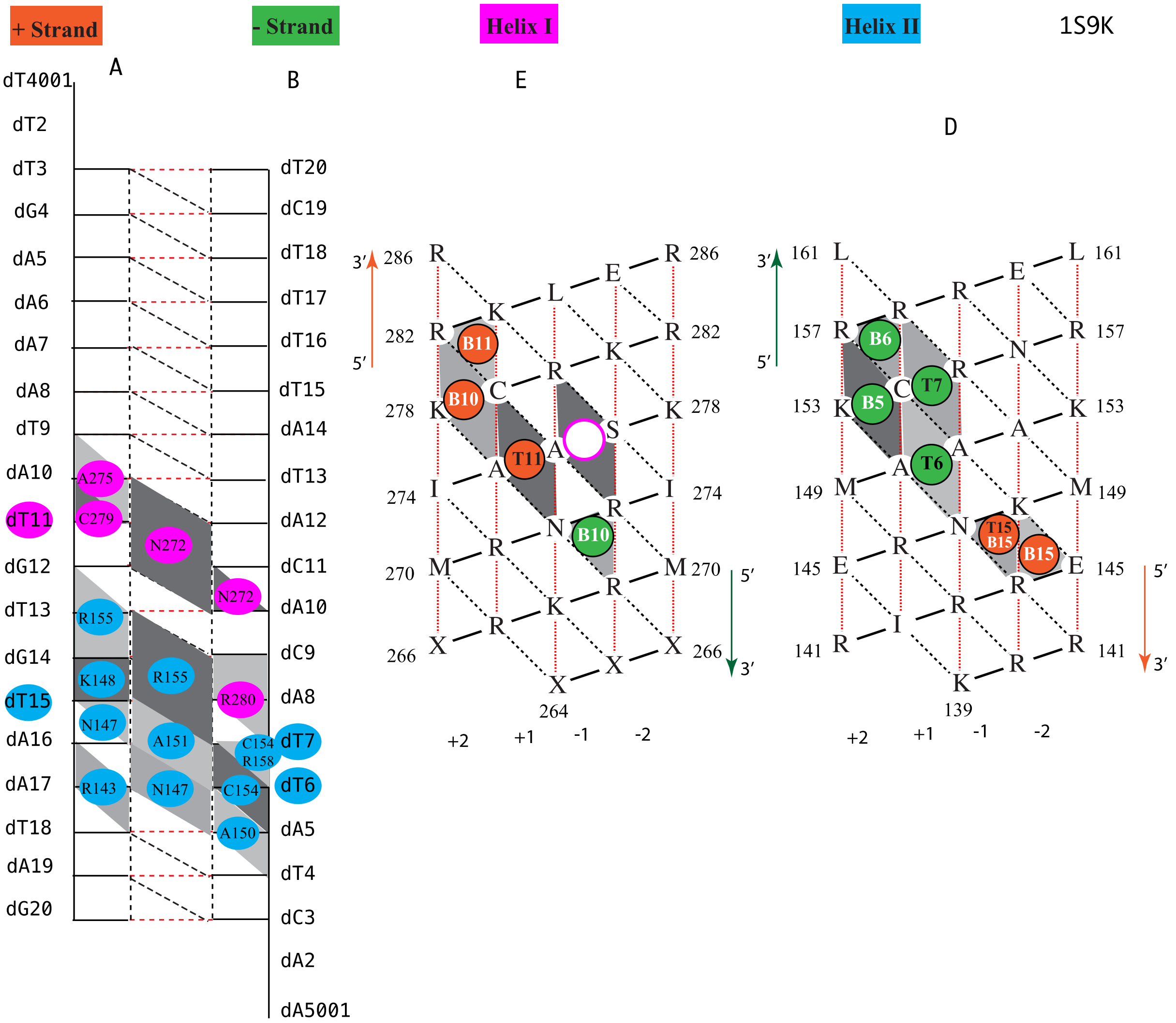

Chain I
3′
5′
295
291
287
283
279
275
R
K
K
I
E
K
A
K
M
R
K
D
K
S
R
R
K
A
V
I
N
N
R
E
R
E
I
R
Y
K
B7
B6
B5
T7
T6
A20
A21
B20
B21
295
291
287
283
279
275
5′
3′
-2
-1
+1
+2

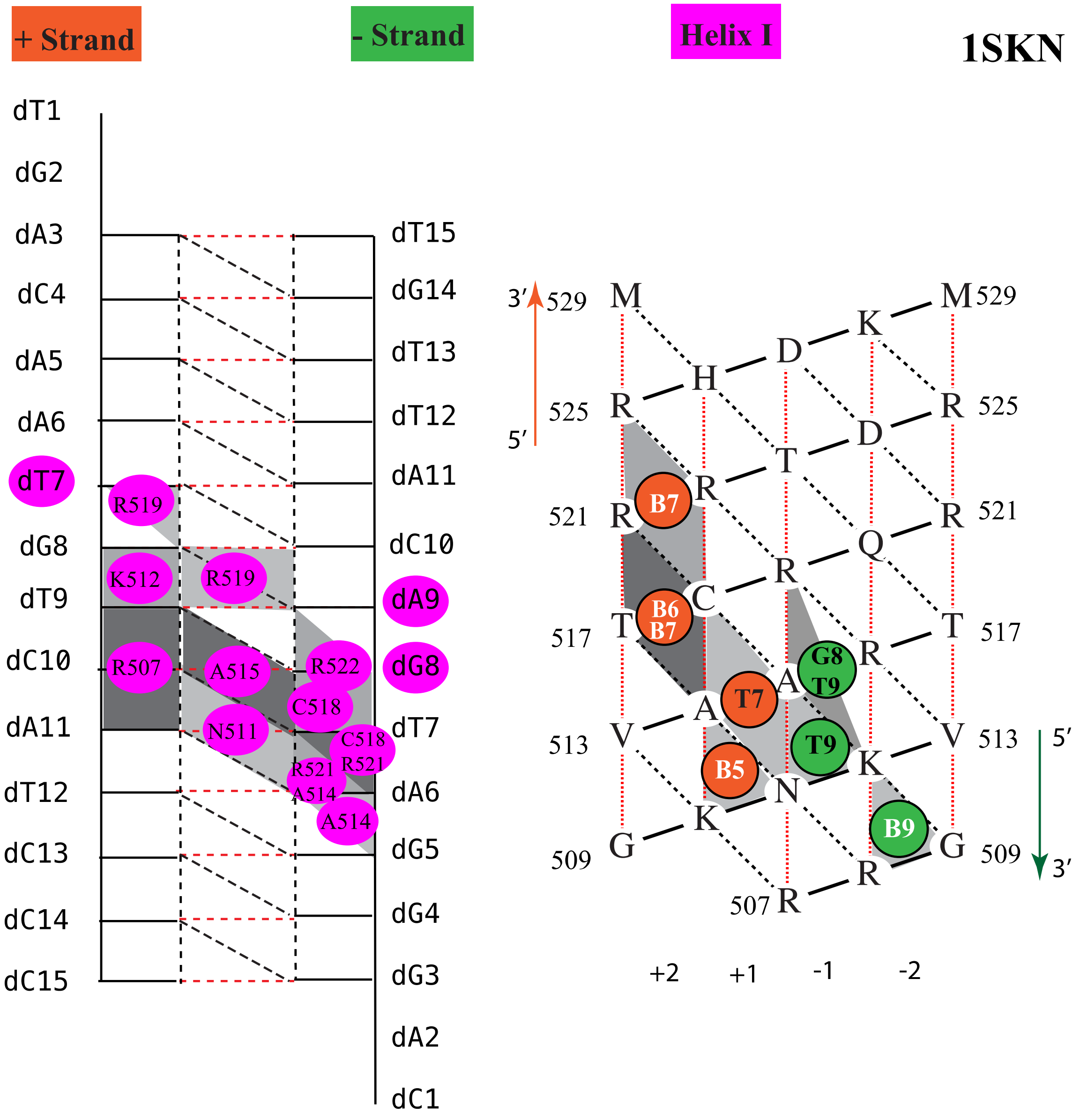

Chain II
3′
5′
295
291
287
283
279
275
R
K
K
I
E
K
A
K
M
R
K
D
K
S
R
R
K
A
V
I
N
N
R
E
R
E
I
R
Y
K
B15
B14
T16
T15
T11
B10
B10
B11
295
291
287
283
279
275
5′
3′
-2
-1
+1
+2


+ Strand
A
- Strand
B
Helix I
D
Helix II
C
1JNM
dC201
dG2
dT3
dC4
dG5
dA6
dT7
dG8
dA9
dC10
dG11
dT12
dC13
dA14
dT15
dC16
dG17
dA18
dC19
dG20
dG20
dC19
dA18
dG17
dC16
dT15
dA14
dC13
dT12
dG11
dC10
dA9
dG8
dT7
dA6
dG5
dC4
dT3
dG2
dC301
A265
S269
N262
N262
A266
R259
A266
R270
R263
R270
R270
R270
R263
N262
R259
A266
A266
N262
N262
S269
A265
K258
K258
3′
276
272
5′
268
264
260
256
254
5′
3′
+2
+1
-1
-2
R
K
L
E
R
R
K
R
K
S
A
S
K
B6
T7
T12
B12
I
A
N
R
I
M
R
R
M
E
R
K
A
E
K
B11
G11

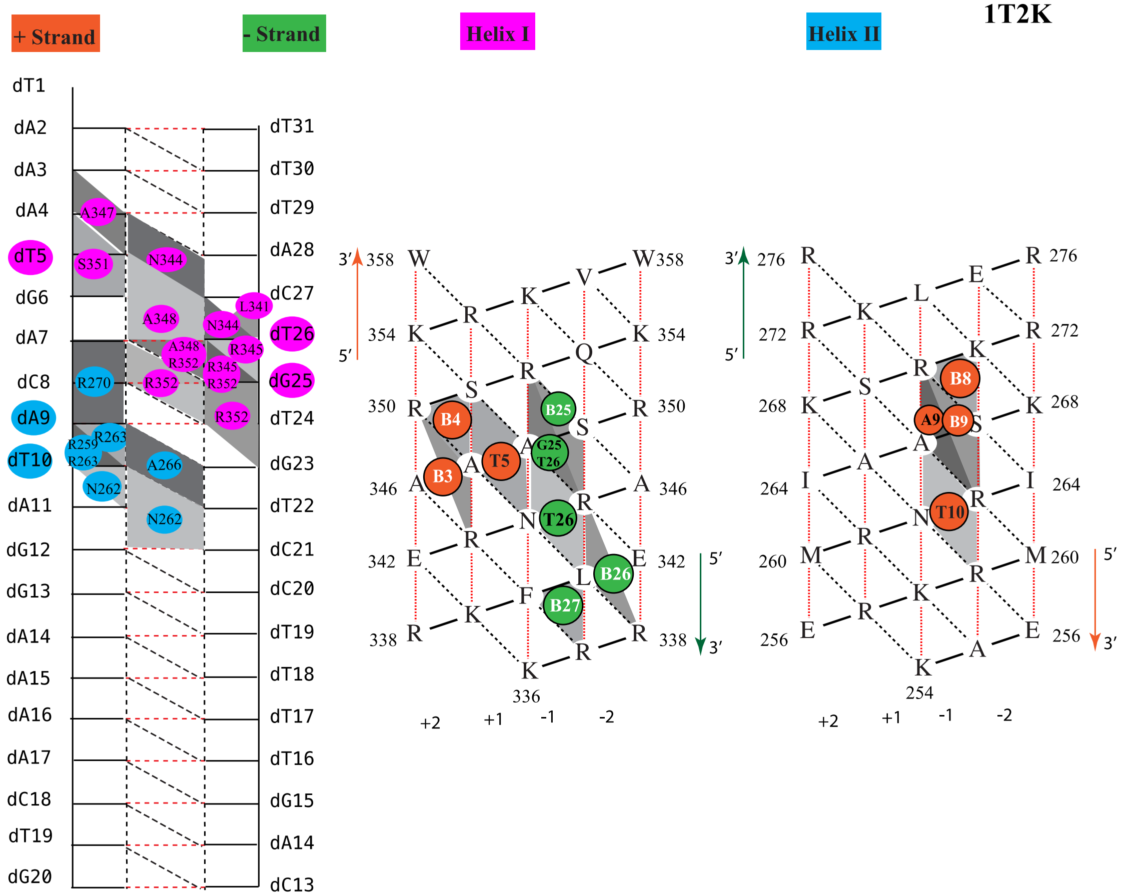
1NWQ


Chain B
Chain D
dT−10
dT−9
dC−8
dC−7
dT−6
dA−5
dT−4
dT−3
dG−2
dC−1
dG1
dC2
dA3
dA4
dT5
dC6
dC7
dA8
dG9
dT10
dT11
dA9
dG8
dG7
dA6
dT5
dA4
dA3
dC2
dG1
dC−1
dG−2
dT−3
dT−4
dA−5
dG−6
dG−7
dT−8
dC−9
dA−10
dA−11
dA−12
A295
S299
N292
V296
R300
R289
N293
N293
R300
R289
V296
N292
S299
A295
Chain A
Chain C
T 310
E
V
N
R 306
306 R
Q
K
A
302 K
K 302
D
R
S
298 K
K 298
V
A
294 I
I 294
N
R
290 E
E 290
R
V
286 R
R 286
Y
dB-4
dB-5
dB-4
dT-3
dB-5
dT-4
dG1
dB-6
dC2
dB1
dB2
dB2
dB-4
dB-1
dB-4
dT-3
dB-5
dG1
dT-4
dC2
dB1
dB2
dB2

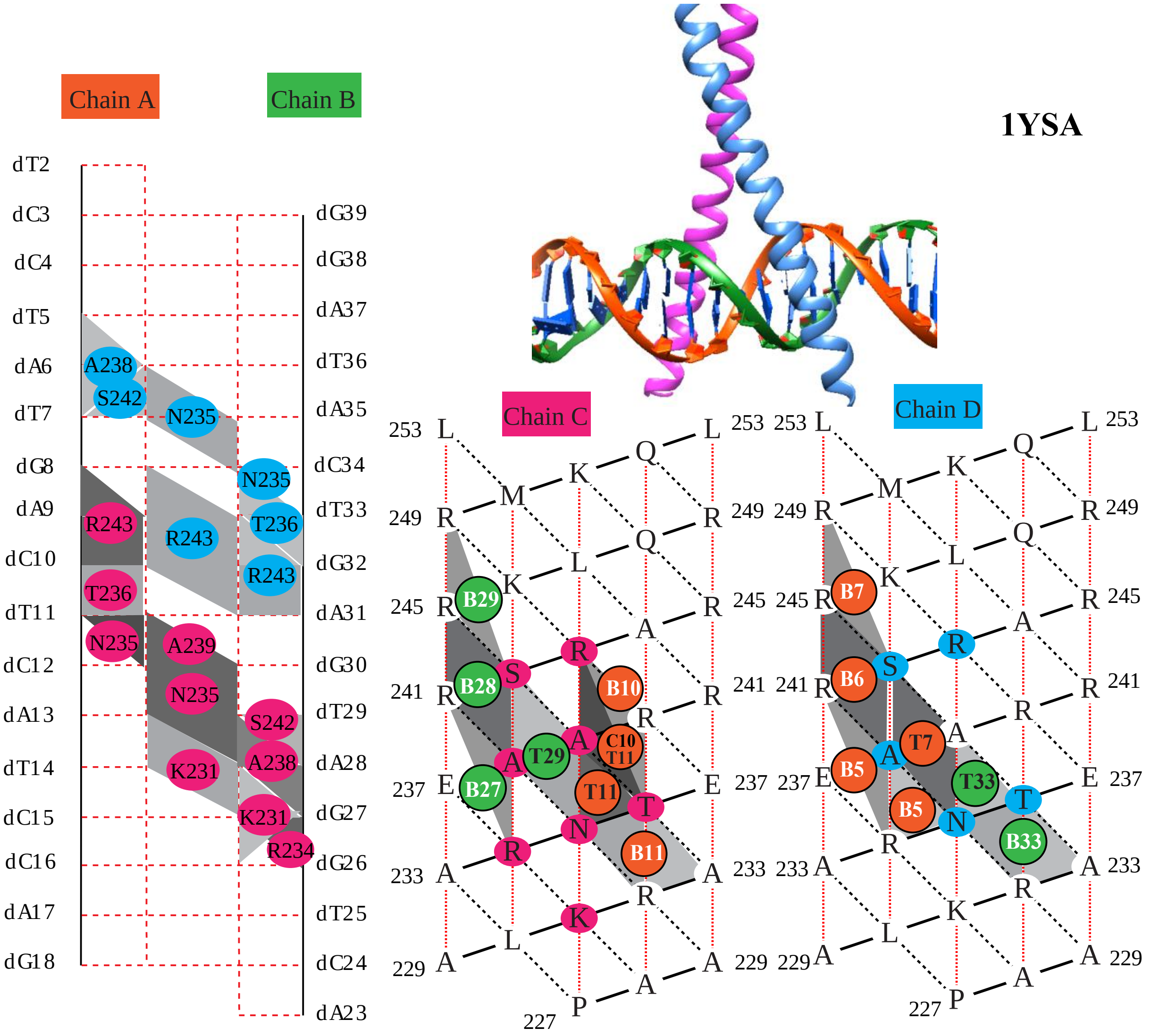

+ Strand
- Strand
Helix I
Helix II
1S9K
A
B
E
D
dT4001
dT2
dT3
dG4
dA5
dA6
dA7
dA8
dT9
dA10
dT11
dG12
dT13
dG14
dT15
dA16
dA17
dT18
dA19
dG20
dT20
dC19
dT18
dT17
dT16
dT15
dA14
dT13
dA12
dC11
dA10
dC9
dA8
dT7
dT6
dA5
dT4
dC3
dA2
dA5001
A275
C279
N272
N272
R155
K148
R155
N147
R280
A151
C154
R158
R143
N147
C154
A150
3′
5′
286
282
278
274
270
266
264
R
K
L
E
R
R
K
R
K
R
C
K
K
A
A
S
I
R
I
M
R
N
M
R
K
R
X
X
X
X
B11
B10
T11
B10
5′
3′
+2
+1
-1
-2
3′
5′
161
157
153
149
145
141
139
L
R
E
L
R
R
N
R
K
C
R
A
K
A
A
M
K
M
E
R
N
E
R
R
I
R
R
K
R
B6
T7
B5
T6
T15
B15
B15
5′
3′
+2
+1
-1
-2

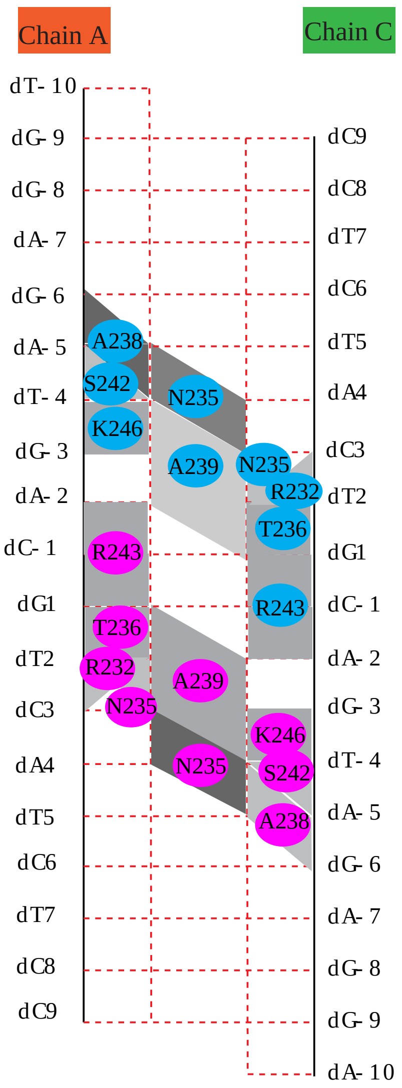

+ Strand
- Strand
Helix I
1SKN
dT1
dG2
dA3
dC4
dA5
dA6
dT7
dG8
dT9
dC10
dA11
dT12
dC13
dC14
dC15
dT15
dG14
dT13
dT12
dA11
dC10
dA9
dG8
dT7
dA6
dG5
dG4
dG3
dA2
dC1
R519
K512
R519
R507
A515
R522
C518
N511
C518
R521
R521
A514
A514
3′
5′
529
525
521
517
513
509
507
M
H
D
K
M
R
D
R
T
R
R
R
Q
R
C
R
T
B7
B6
B7
A
G8
T9
R
T
A
T7
V
K
V
B5
T9
N
K
B9
G
R
G
R
529
525
521
517
513
509
5′
3′
+2
+1
-1
-2

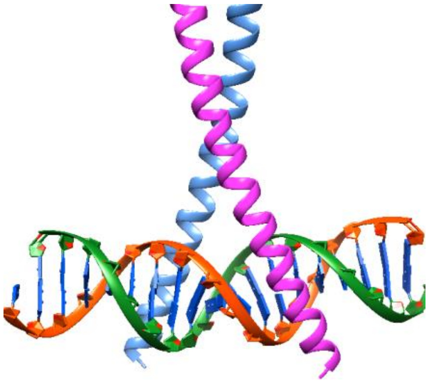
1T2K



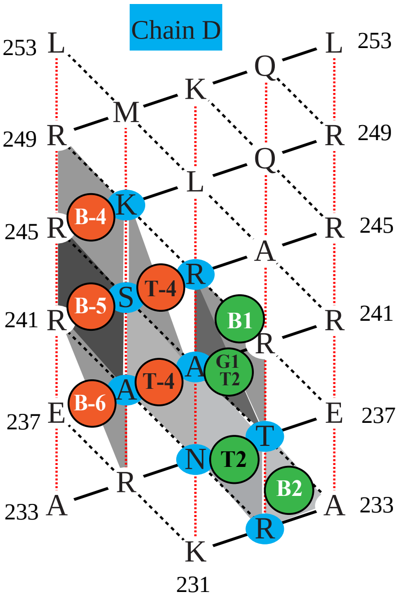
+ Strand
- Strand
Helix I
Helix II
dT1
dA2
dA3
dA4
dT5
dG6
dA7
dC8
dA9
dT10
dA11
dG12
dG13
dA14
dA15
dA16
dA17
dC18
dT19
dG20
dT31
dT30
dT29
dA28
dC27
dT26
dG25
dT24
dG23
dT22
dC21
dC20
dT19
dT18
dT17
dT16
dG15
dA14
dC13
A347
S351
N344
L341
A348
N344
A348
R352
R345
R345
R352
R270
R352
R352
R352
R259
R263
R263
A266
N262
N262
3′
5′
358
354
350
346
342
338
336
W
K
R
A
E
R
K
B4
B3
T5
B25
G25
T26
T26
B26
B27
5′
3′
+2
+1
-1
-2
276
272
268
264
260
256
254
B8
A9
B9
T10

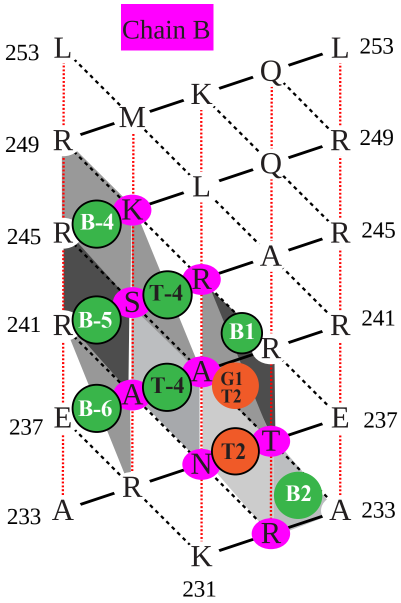

1YSA
Chain A
Chain B
dT2
dC3
dC4
dT5
dA6
dT7
dG8
dA9
dC10
dT11
dC12
dA13
dT14
dC15
dC16
dA17
dG18
dG39
dG38
dA37
dT36
dA35
dC34
dT33
dG32
dA31
dG30
dT29
dA28
dG27
dG26
dT25
dC24
dA23
A238
S242
N235
N235
T236
R243
R243
R243
T236
N235
A239
N235
S242
A238
K231
K231
R234
Chain C
Chain D
253
249
245
241
237
233
229
227
L
M
K
Q
R
B29
B28
B27
T29
B10
C10
T11
T11
B11
S
A
N
T
E
B7
B6
B5
B5
T7
T33
B33
P

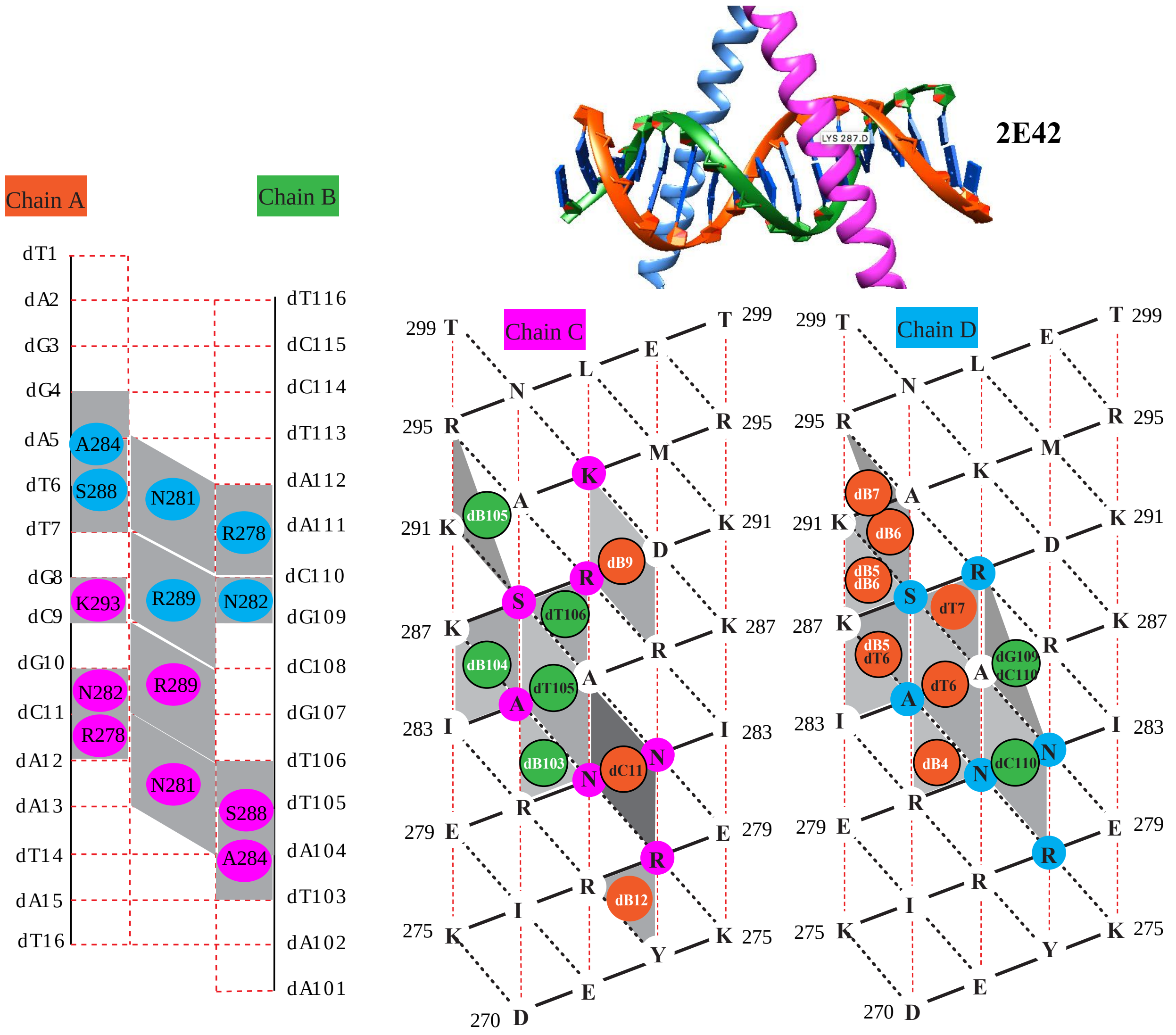

Chain A
Chain C
dT- 10
dG- 9
dG- 8
dA- 7
dG- 6
dA- 5
dT- 4
dG- 3
dA- 2
dC- 1
dG1
dT2
dC3
dA4
dT5
dC6
dT7
dC8
dC9
dC9
dC8
dT7
dC6
dT5
dA4
dC3
dT2
dG1
dC- 1
dA- 2
dG- 3
dT- 4
dA- 5
dG- 6
dA- 7
dG- 8
dG- 9
dA- 10
A238
S242
N235
K246
A239
N235
R232
T236
R243
R243
T236
R232
A239
N235
K246
N235
S242
A238



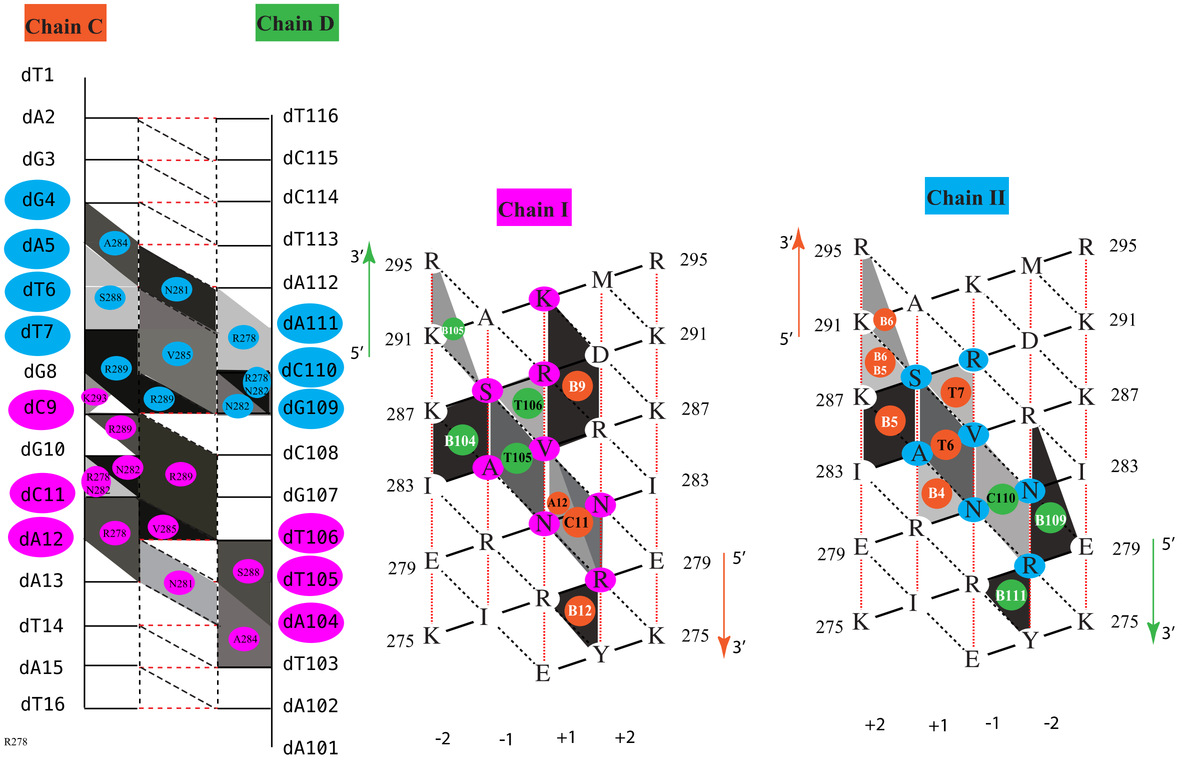
2DGC


Chain D
253 L
L 253
Q
K
M
249 R
R 249
Q
L
245 R
R 245
K
A
R
241 R
R 241
S
R
237 E
E 237
A
A
T
233 A
A 233
R
N
R
K
231
B-4
B-5
B-6
T-4
T-4
B1
G1
T2
T2
B2


Chain B
253 L
L 253
Q
K
M
249 R
R 249
Q
L
245 R
R 245
K
A
R
241 R
R 241
S
R
A
237 E
E 237
A
T
N
233 A
A 233
R
R
K
231
B-4
B-5
B-6
T-4
T-4
B1
G1
T2
T2
B2

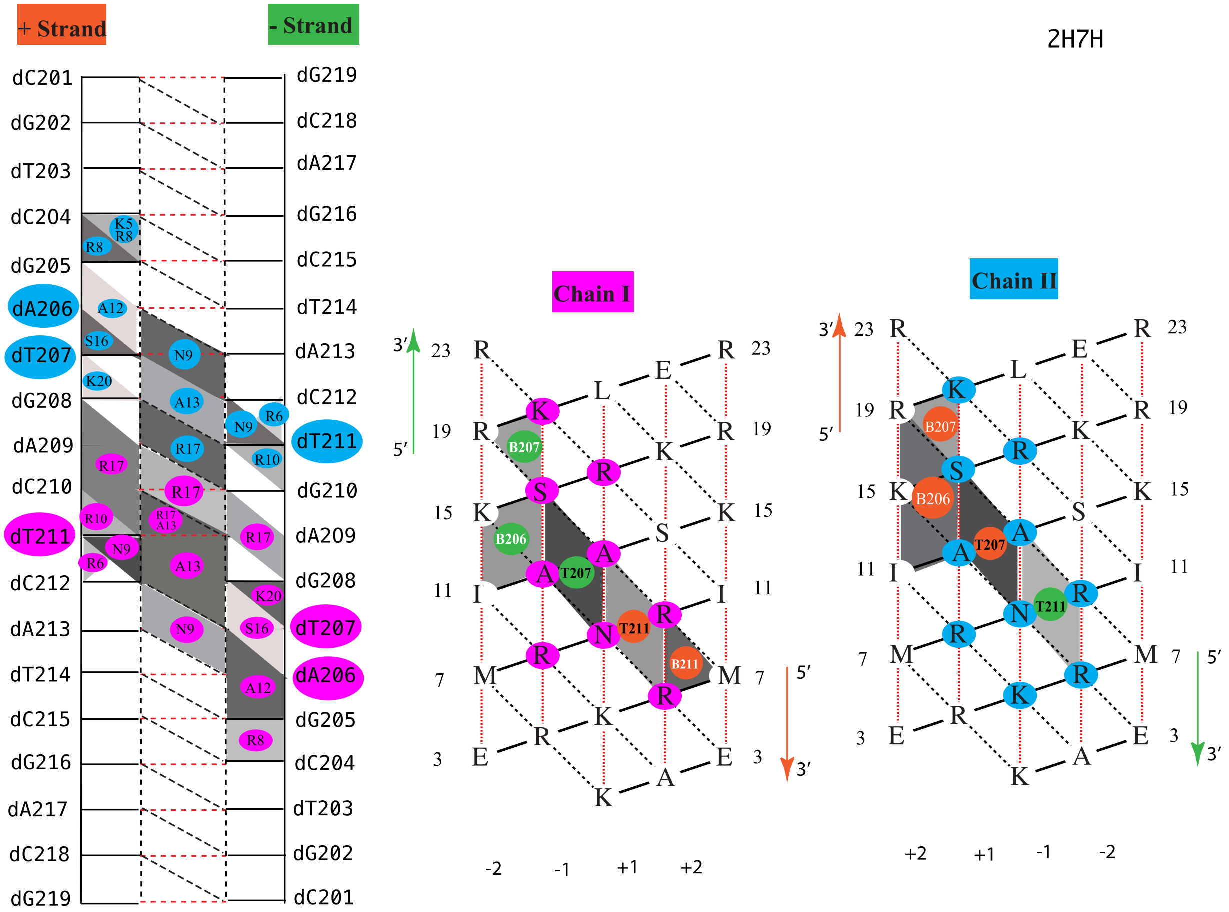

2E42
LYS 287.D
Chain A
Chain B
dT1
dA2
dG3
dG4
dA5
dT6
dT7
dG8
dC9
dG10
dC11
dA12
dA13
dT14
dA15
dT16
dT116
dC115
dC114
dT113
dA112
dA111
dC110
dG109
dC108
dG107
dT106
dT105
dA104
dT103
dA102
dA101
A284
S288
N281
R278
K293
R289
N282
N282
R289
R278
N281
S288
A284
Chain C
Chain D
299 T
295 R
291 K
287 K
283 I
279 E
275 K
270 D
T 299
R 295
K 291
K 287
I 283
E 279
K 275
N
L
E
M
A
K
D
R
S
A
R
N
I
R
Y
dB105
dB9
dT106
dB104
dT105
dB103
dC11
dB12
dB7
dB6
dB5
dB6
dB5
dT6
dT7
dG109
dC110
dT6
dB4
dC110

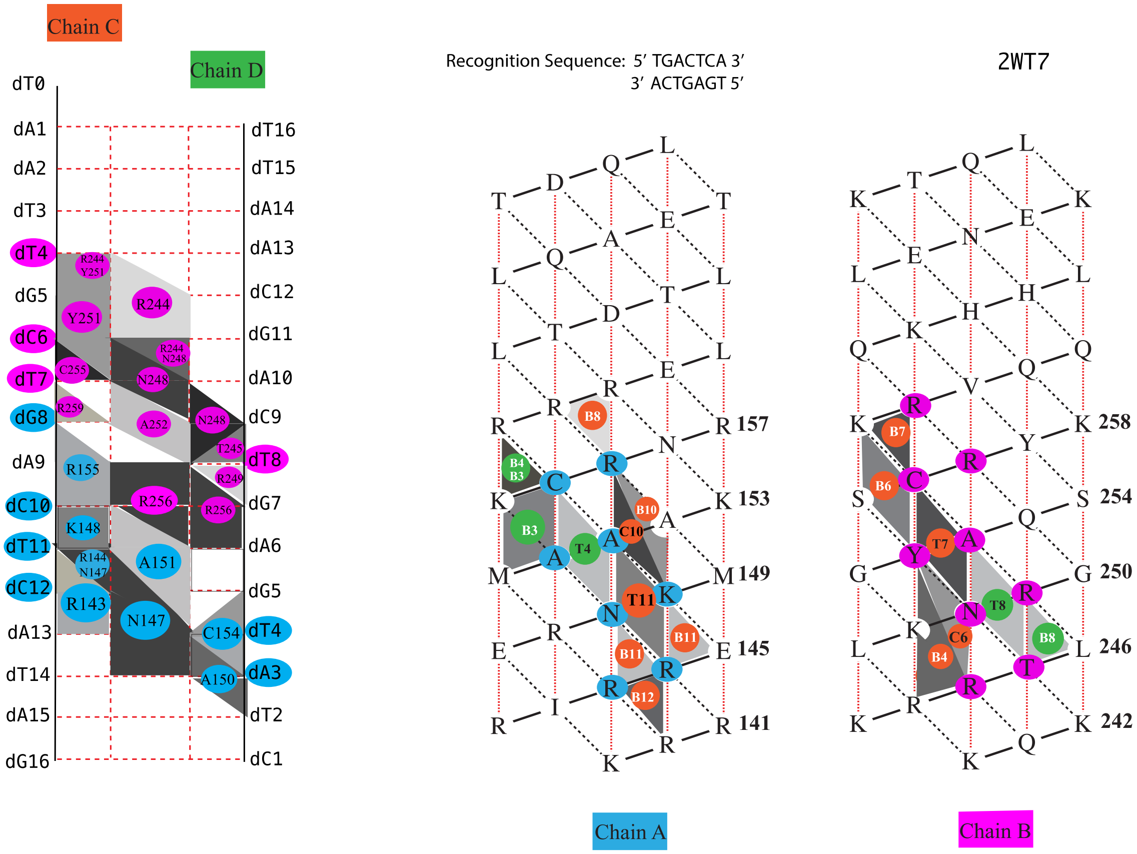
2E43


Chain C
Chain D
dT1
dA2
dG3
dG4
dA5
dT6
dT7
dG8
dC9
dG10
dC11
dA12
dA13
dT14
dA15
dT16
dT116
dC115
dC114
dT113
dA112
dA111
dC110
dG109
dC108
dG107
dT106
dT105
dA104
dT103
dA102
dA101
A284
S288
N281
V285
R278
R289
K293
N282
R278
R289
N282
R289
R278
V285
N281
S288
A284
R278
Chain I
3′
5′
295
291
287
283
279
275
R
K
K
I
E
K
A
S
A
R
I
K
R
V
N
R
E
M
D
R
N
R
Y
R
K
K
I
E
K
295
291
287
283
279
275
5′
3′
B105
T106
B9
B104
T105
A12
C11
B12
-2
-1
+1
+2
Chain II
3′
5′
295
291
287
283
279
275
R
K
K
I
E
K
A
S
A
R
I
E
K
R
V
N
R
M
D
R
N
R
Y
R
K
K
I
E
K
295
291
287
283
279
275
5′
3′
B6
B6
B5
T7
B5
T6
B4
C110
B109
B111
+2
+1
-1
-2

R278

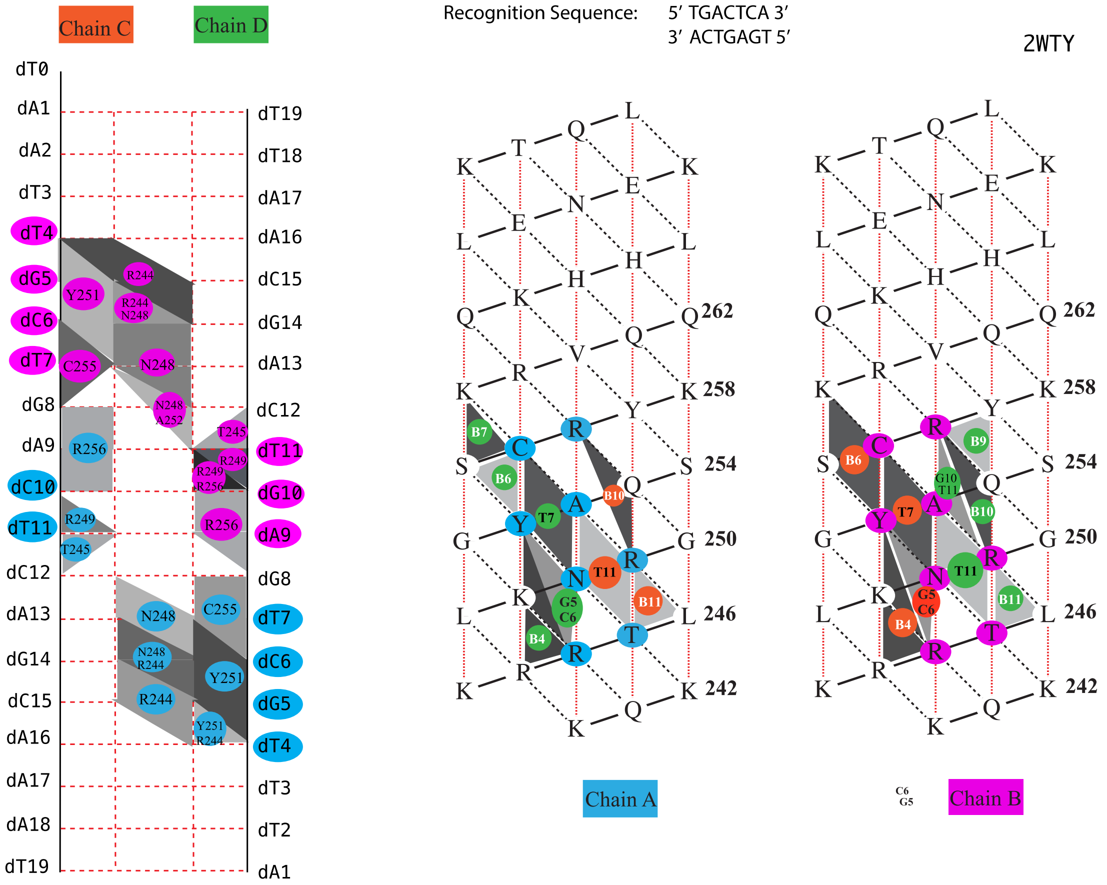
2H7H

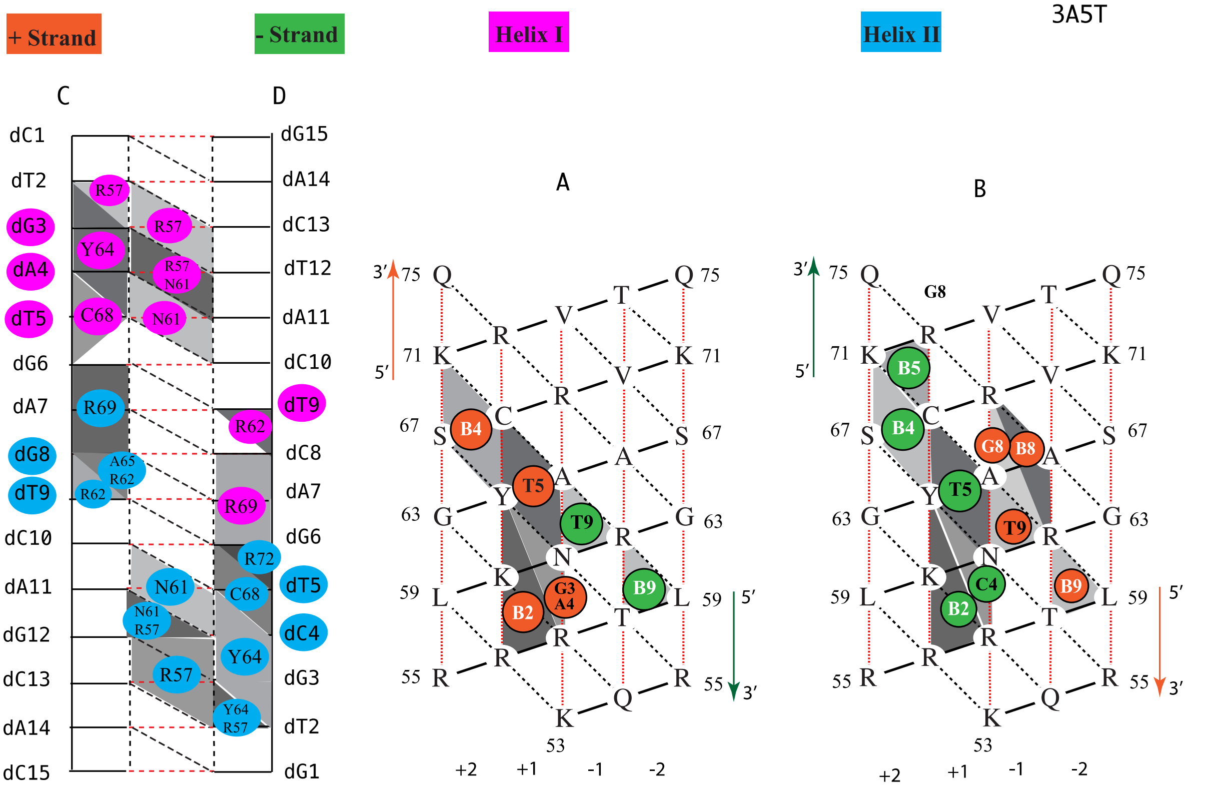

+ Strand
- Strand
dC201
dG202
dT203
dC204
dG205
dA206
dT207
dG208
dA209
dC210
dT211
dC212
dA213
dT214
dC215
dG216
dA217
dC218
dG219
dG219
dC218
dA217
dG216
dC215
dT214
dA213
dC212
dT211
dG210
dA209
dG208
dT207
dA206
dG205
dC204
dT203
dG202
dC201
K5
R8
R8
A12
S16
K20
N9
A13
R17
N9
R6
R10
R17
R17
R10
R17
A13
N9
R6
A13
R17
K20
S16
N9
A12
R8
Chain I
Chain II
3′
5′
5′
3′
23
19
15
11
7
3
-2
-1
+1
+2
B207
B206
T207
T211
B211
B206
B207
T207
T211
+2
+1
-1
-2

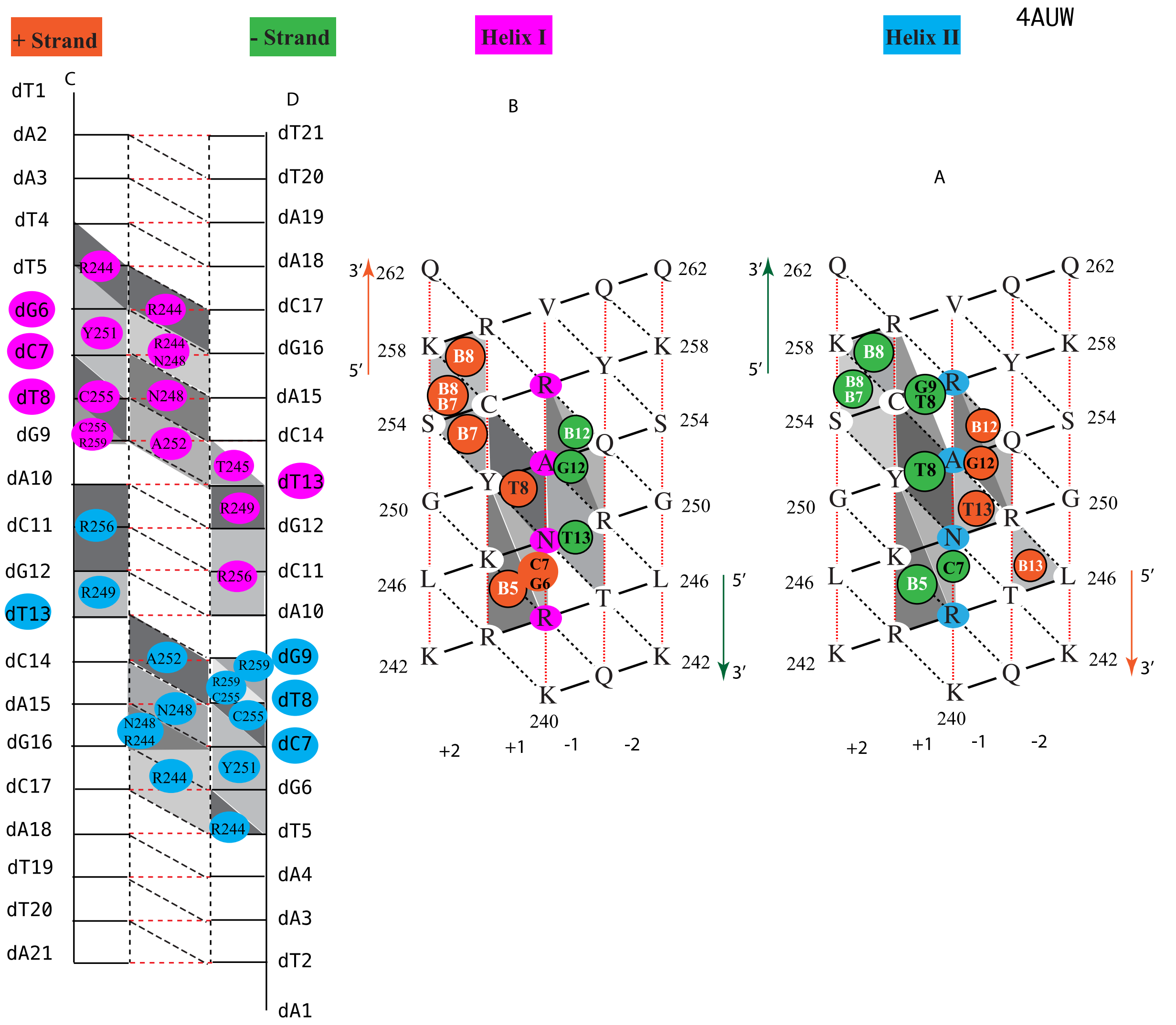

Recognition Sequence: 5′ TGACTCA 3′
3′ ACTGAGT 5′
2WT7
Chain C
Chain D
dT0
dA1
dA2
dT3
dT4
dG5
dC6
dT7
dG8
dA9
dC10
dT11
dC12
dA13
dT14
dA15
dG16
dT16
dT15
dA14
dA13
dC12
dG11
dA10
dC9
dT8
dG7
dA6
dG5
dT4
dA3
dT2
dC1
R244
Y251
R244
Y251
R244
N248
C255
N248
R259
A252
N248
T245
R249
R155
R256
R256
K148
R144
N147
A151
R143
N147
C154
A150
B8
B4
B3
B3
T4
B10
C10
T11
B11
B11
B12
157
153
149
145
141
Chain A
B7
B6
T7
T8
B8
C6
B4
258
254
250
246
242
Chain B

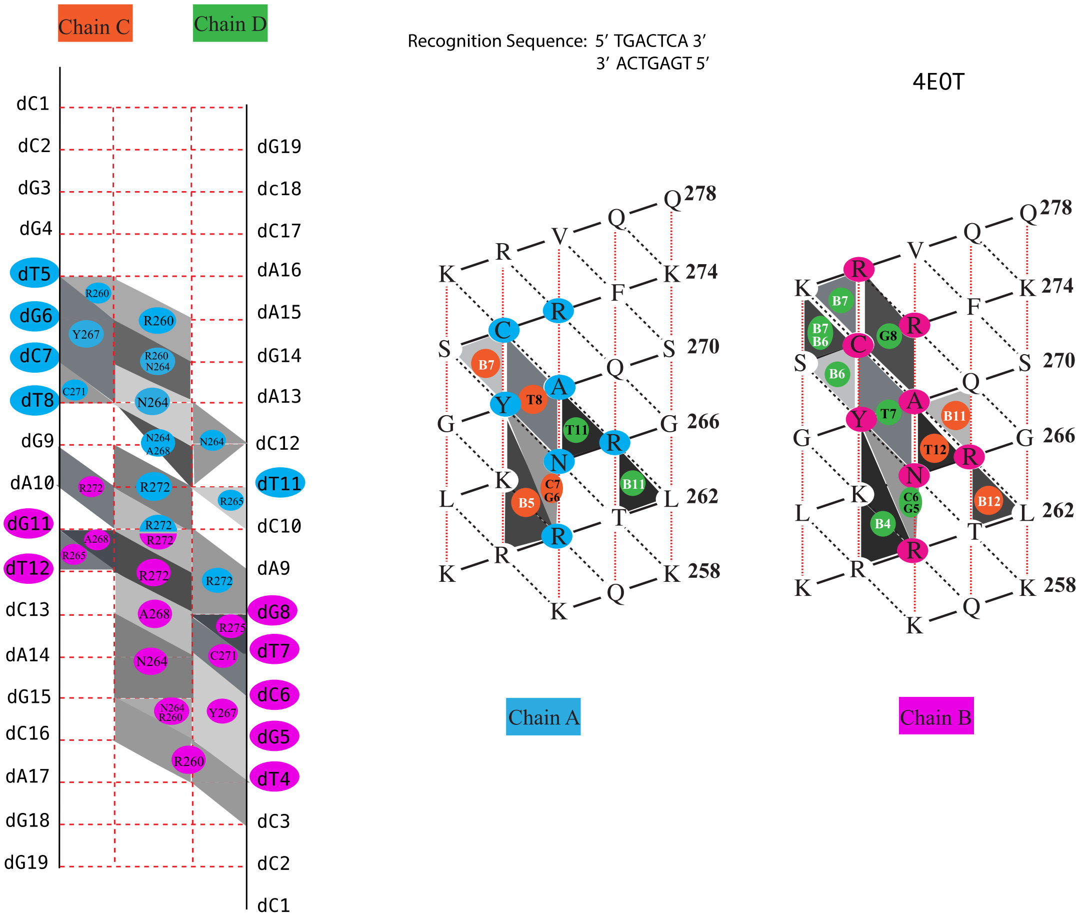
Recognition Sequence: 5′ TGACTCA 3′
3′ ACTGAGT 5′

2WTY

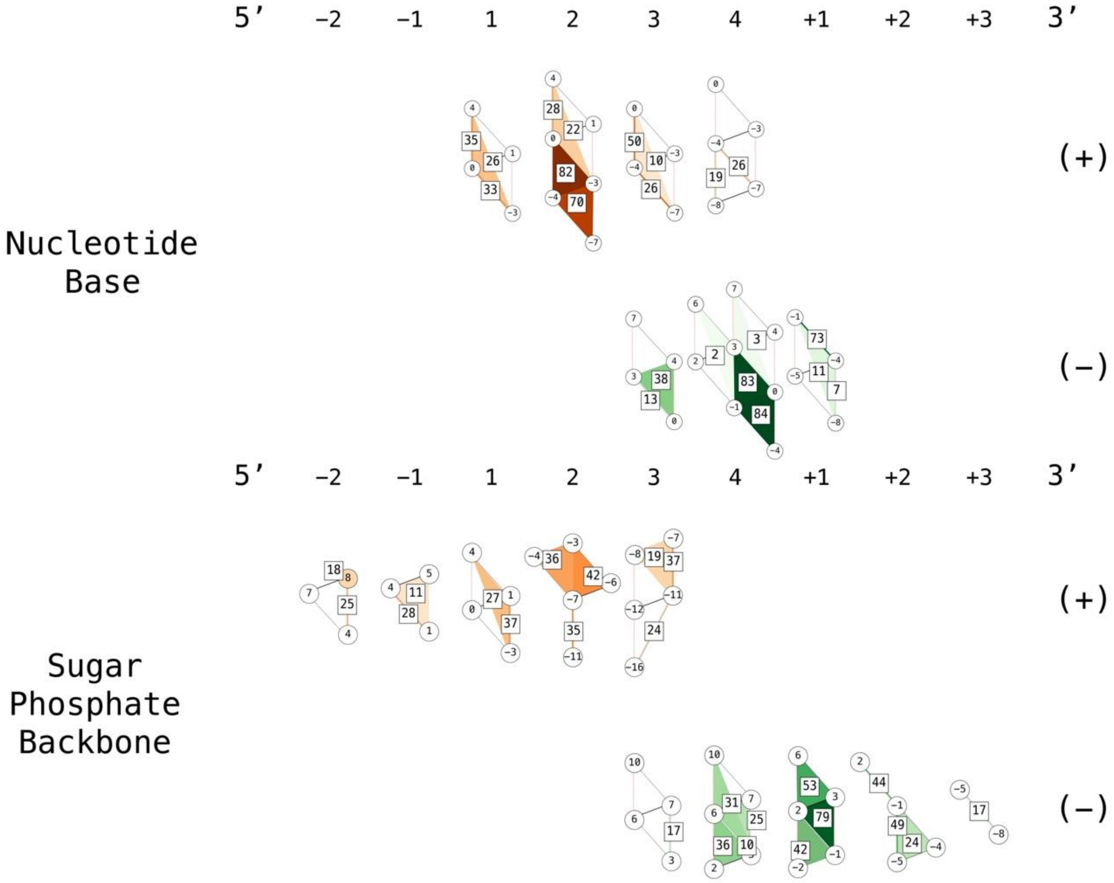



+ Strand
- Strand
Helix I
Helix II
3A5T
C
D
dC1
dT2
dG3
dA4
dT5
dG6
dA7
dG8
dT9
dC10
dA11
dG12
dC13
dA14
dC15
dG15
dA14
dC13
dT12
dA11
dC10
dT9
dC8
dA7
dG6
dT5
dC4
dG3
dT2
dG1
R57
R57
Y64
R57 N61
C68
N61
R69
R62
A65 R62
R62
R69
R72
N61
C68
N61 R57
Y64
R57
Y64 R57
A
B
3′
5′
75
71
67
63
59
55
53
Q
R
V
T
K
C
S
A
G
Y
N
L
+2
+1
-1
-2
B4
T5
T9
B2
G3 A4
B9
5′
3′
G8
B5
B4
G8
B8
T5
T9
C4
B2
B9

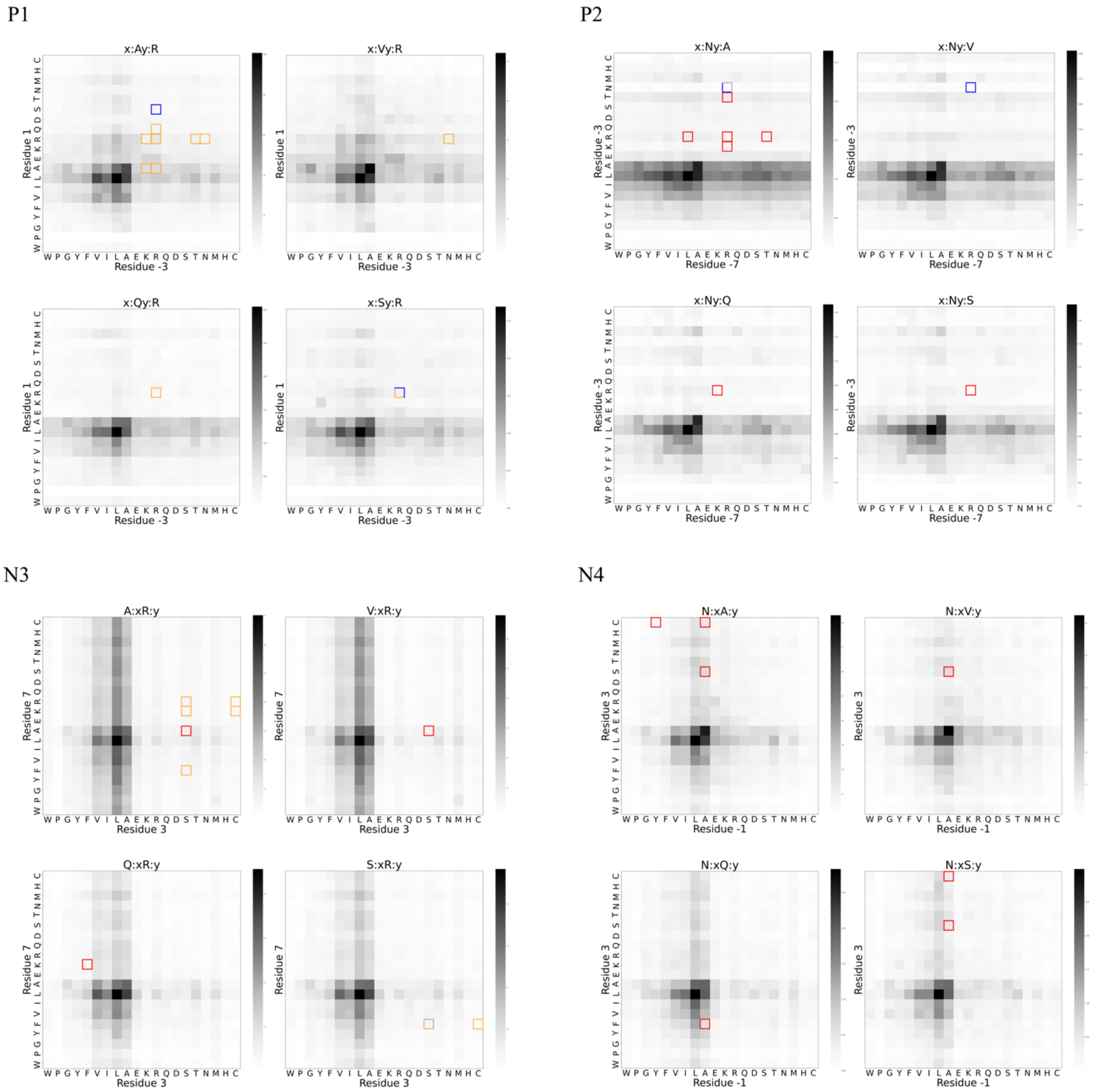

4AUW
+ Strand
- Strand
Helix I
Helix II
C
D
B
A
dT1
dA2
dA3
dT4
dT5
dG6
dC7
dT8
dG9
dA10
dC11
dG12
dT13
dC14
dA15
dG16
dC17
dA18
dT19
dT20
dA21
dT21
dT20
dA19
dA18
dC17
dG16
dA15
dC14
dT13
dG12
dC11
dA10
dG9
dT8
dC7
dG6
dT5
dA4
dA3
dT2
dA1
R244
Y251
C255
C255
R259
R244
R244
N248
N248
A252
T245
R249
R256
R256
R249
A252
R259
R259
C255
N248
N248
R244
C255
R244
Y251
R244
3′
5′
262
258
254
250
246
242
240
Q
R
V
Q
K
Y
S
G
L
K
R
T
N
C
A
B8
B8
B7
B7
T8
C7
G6
B5
B12
G12
T13
G9
T8
B12
G12
T13
B13
C7
B5
+2
+1
-1
-2



Supplemental Figure S1: Packing maps of bZIP-DNA complexes. 28 of the PDB files were used in these maps, displaying packing trends throughout the bZIP-DNA complexes and half-sites from four- and five-body packing cliques.



Supplemental Figure S2: Individual Packing Patterns of the α-helical Packing Surface Corresponding to DNA Knob Position Frequencies. Excised portions of protein lattices are shown displaying the common sockets and areas where DNA knobs pack into the helix, categorized by the base number and position. Nucleotide and amino acid sequence composition is not considered for this figure. The packing groups are displayed before superimposition in Figure 9, where infrequent or uncommon interactions were excluded from the composite protein lattice. Socket shading corresponds to the highest frequency of a DNA knob packing into the socket, rather than packing overall. Additionally, shaded bolded lines indicate contact groups with two protein residues. Sockets and residue pairs corresponding to knobs of the positive strand were shaded orange, while sockets packing with negative strand bases were shaded green. A) The individual packing patterns of DNA base knobs are separated by both position and strand, allowing for direct comparison for similar packing mechanisms prior to superimposition. B) Similar to part A, the individual packing patterns display the frequency and area of DNA backbone packing.



Supplemental Figure S3: Filled propensities of the four recognition regions. All possible pocket sequences were found in previous filled-free propensity data, where the individual sockets' filled propensities were summed to display the stability and frequency of said pocket sequence throughout the PDB database. The darker the shade of the heatmap cell, the more likely the pocket is found and packed by a knob. The set of heatmaps are separated based on region: P1, P2, N3, and N4. Each region has four separate heatmaps, representing the four amino acids (i.e., A, V, Q, and S) found at residue 0 which is shared by all four recognition regions. The four P1 heatmaps, for example, are titled x:Ay:R, x:Vy:R, x:Qy:R, x:Sy:R, corresponding to the four residues of the P1 pocket: -3, 0, 1, 4. The last residue of the socket sequences is the invariant arginine, while the second residue of the sequences is the highly conserved residue 0. The "x" and "y" residues of the P1 socket sequence are highly variable and are represented by the x- and y-axes of the heatmaps. A similar numbering scheme is applied across all four regions. Within the heatmaps of a single region, colored squares represent the region sequences found in the crystal structure dataset. The border squares are colored according to the nucleotide most frequently found in the respective DNA position: dA is green, dT is red, dC is blue, dG is yellow, and 5mC is purple.