https://doi.org/10.1093/nargab/lqac101

Alignment of major-groove hydrogen bond arrays uncovers shared information between different DNA sequences that bind the same protein

Jacklin Sedhom [1], Jason Kinser [2] and Lee A. Solomon [1,*]

[1] Department of Chemistry and Biochemistry, George Mason University, Fairfax, VA 22030, USA and [2] Department of Computational and Data Sciences, George Mason University, Fairfax, VA 22030, USA

Received October 05, 2022; Revised October 28, 2022; Editorial Decision December 01, 2022; Accepted December 12, 2022

ABSTRACT

Protein–DNA binding is of a great interest due to its importance in many biological processes. Previous studies have presented many factors responsible for the recognition and specificity, but understanding the minimal informational requirements for proteins that bind to multiple DNA-sites is still an understudied area of bioinformatics. Here we focus on the hydrogen bonds displayed by the target DNA in the major groove that take part in protein-binding. We show that analyses focused on the base pair identity may overlook key hydrogen bonds. We have developed an algorithm that converts a nucleotide sequence into an array of hydrogen bond donors and acceptors and methyl groups. It then aligns these non-covalent interaction arrays to identify what information is being maintained among multiple DNA sequences. For three different DNA-binding proteins, Lactose repressor, controller protein and λ-CI repressor, we uncovered the minimal pattern of hydrogen bonds that are common amongst all the binding sequences. Notably in the three proteins, key interacting hydrogen bonds are maintained despite nucleobase mutations in the corresponding binding sites. We believe this work will be useful for developing new DNA binding proteins and shed new light on evolutionary relationships.

INTRODUCTION

Protein-DNA binding is critically important for a number of biological processes (e.g. DNA transcription, replication and repair) (1,2). The sequence-specific interaction between proteins and DNA is of particular interest. Understanding the biophysical principles that guide how proteins recognize DNA with high specificity impacts how we study regulatory processes in the living organisms and our ability to develop new gene therapies and therapeutic drugs (1). Many studies have investigated the complementarity of hydrogen bonds presented in the major groove (termed direct readout) (1–7). Luscombe *et al.* reviewed 129 protein-DNA complexes and clarified the roles of hydrogen bonds, van der Waals interactions, and water mediated bonds at the protein-DNA interface (3). Similarly, Garvie and Wolberger described how protein-DNA binding specificity arises from pairing hydrogen bond donors and acceptors between the protein and DNA and the role of van der Waals interaction between the thymine 5-position methyl group and amino acid side chains (3,4). These studies were further corroborated by Emamjomeh *et al.*, who showed that the highest degree of binding specificity is obtained from the complimentary pairing of hydrogen bond donors and acceptors in the major groove with amino acids (2). Recently, Lin and Guo carried out a comparative analysis for different protein-DNA complexes of different degrees of binding specificity (1). These studies all discussed the role of direct readout in recognition specificity, further highlighting the role of major groove hydrogen bonds.

However, these studies did not focus on proteins with multiple DNA-binding sites, what information is shared between them, or the minimal amount of direct readout needed. They were primarily focused on the base pairs themselves and did not seek to address how the information is displayed and if any of it is maintained between sequences. We hypothesize that focusing on specific nucleobases will miss some of the individual hydrogen bonds essential for recognition and binding. For example, the 7-position nitrogen of purine bases will be maintained if Guanine is mutated to an Adenine or vice versa (Figure 1).

In this study, we developed an analysis that can take a new view of direct readout, with a focus on proteins that bind multiple DNA sequences. We investigated the DNA binding specificity determinants of the structurally well-characterized helix-turn-helix DNA binding proteins: Lactose repressor (Lac R), Controller protein (C-protein), which both have three binding sites, and λ-phage repressor protein (CI) which has six. All these protein's functions

[*] To whom correspondence should be addressed. Tel: +1 703 993 6418; Email: lsolomo@gmu.edu

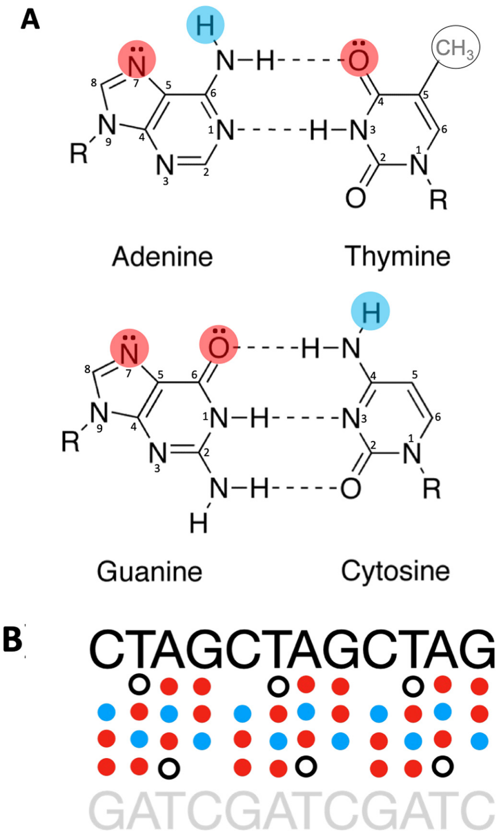


Figure 1. Hydrogen-bond donor/ acceptor pattern exposed in the major groove. (**A**) AT and GC base pairs showing which atoms contribute to the hydrogen bonding pattern. (**B**) DNA sequences displayed as an array of hydrogen bond acceptors (red circle) and donors (blue circle). The white circles represent the thymine methyl groups.

are dependent on their ability to screen different sequences from random DNA with high specificity. Most studies we have found refer to the evolutionarily conserved base pairs as the main reason for recognition and specificity (1,8,9). However, here we wanted to identify the important DNA-protein contacts (hydrogen bonds and methyl groups) that govern sequence recognition by a cognate protein, irrespective of the degenerate DNA sequences to which the three proteins bind. To do this, we wrote an alignment algorithm to analyze DNA-binding sites that focuses on the major groove-exposed hydrogen bond donors/acceptors and thymine methyl groups used for protein interactions. The analysis algorithm depends primarily on the uniqueness of the hydrogen bonding pattern in the major groove which provides the specificity for protein binding. Despite the appreciation of noncovalent interactions made between the protein and the DNA backbone (e.g. Ribose and Phosphate groups) as well as the DNA minor groove, they do not provide the same unique pattern of hydrogen bonding as the major groove. Thus, we did not address them in this work, we chose to focus on the analysis of the variable patterns shown by multiple DNA sequences that bind the same protein.

Figure 1A depicts DNA base pairs showing the atoms that are responsible for making hydrogen bonds to a binding protein. All atoms that can donate a hydrogen in the major groove of DNA are represented as blue circles while the atoms that have lone pairs to accept a hydrogen bond are represented as red circles. Thymine methyl groups, which can make van der Waals interactions, are represented as white circles. The algorithm takes these hydrogen bonds and methyl groups that are exposed in the major groove (i.e. accessible to protein), and converts them into an array (Figure 1B). It then aligns multiple arrays to provide the conserved information between DNA sequences. Any standard Watson-crick hydrogen bonding between DNA strands is not analyzed by the algorithm since they do not contribute to protein binding.

For each of the three proteins Lac R, C-protein and λ-phage CI, a consensus pattern displaying the conserved information between different sequences was developed. All the consensus patterns were then refined from the published crystal or NMR structures, using molecular visualization and MD techniques (10–12), to extract one pattern which we call a ‘distinct pattern’. We hypothesize this distinct pattern represents the minimal amount of direct readout information needed for a protein to contact and recognize its target DNA. For example, Figure 2 shows the workflow of the analysis process applied to Lac R example, which will be discussed deeply in the method and results sections.

The analyses of the three proteins shows that there are hydrogen bonds that are maintained despite the change of the nucleobase itself over the different binding sites of the same binding protein, which has implications to evolution and design of DNA binding proteins.

MATERIALS AND METHODS

Programs used for visualization and alignment

For visualization of the crystal and NMR structures, as well as inspection of bonds and interactions, both UCSF Chimera (13) and UCSF ChimeraX (14,15) were chosen for these studies because they can be learned quickly, and are available free of charge for noncommercial use. The analysis algorithm was developed using Python with packages NumPy (16) and Matplotlib (17). The codes are available for free on our GitHub page.

The general method applied for each DNA-binding protein

The general workflow can be found in Figure 2. In Steps 1 and 2, the sequences of all the corresponding DNA-binding sites were obtained from the literature and used as input into the algorithm to generate their corresponding hydrogen bond patterns. Individual base pairs were converted into a four-slot vertical array of hydrogen bond donors (blue circle), acceptors (red circle), methyl groups (white circle), or left blank if nothing was in that position (i.e. the five-position of cytosine). While methyl groups are only present on thymine nucleotides, they are included for further development of this algorithm which will include methylated nucleotides.

In Step 3, these hydrogen bond patterns were aligned to obtain only one pattern that is shared among all the binding sites (consensus pattern). We held a 100% cutoff, meaning that a specific hydrogen bond had to be present in every sequence or it was not used.

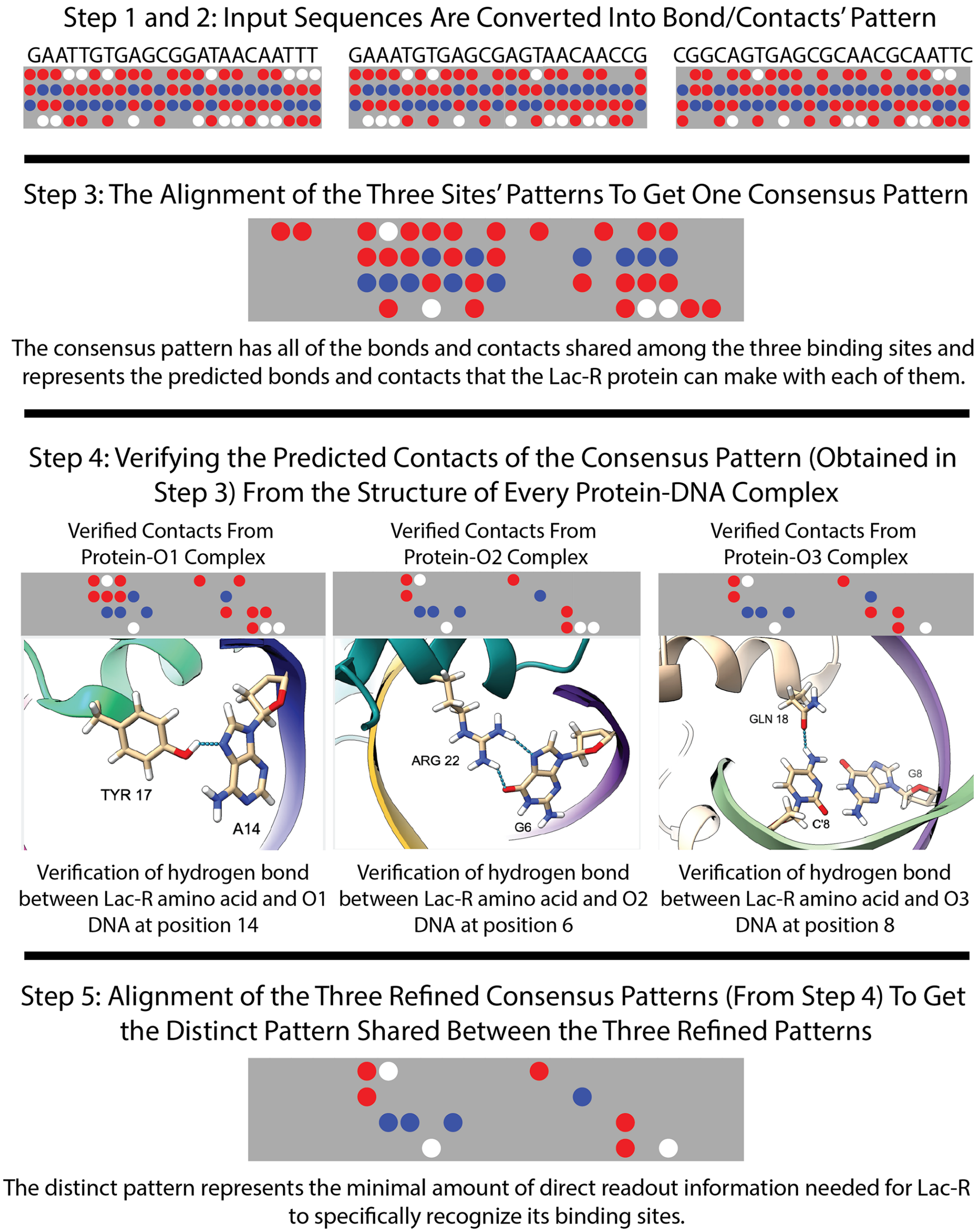

Step 1 and 2: Input Sequences Are Converted Into Bond/Contacts' Pattern
GAATTGTGAGCGGATAACAATTT
GAAATGTGAGCGAGTAACAACCG
CGGCAGTGAGCGCAACGCAATTC
Step 3: The Alignment of the Three Sites' Patterns To Get One Consensus Pattern
The consensus pattern has all of the bonds and contacts shared among the three binding sites and represents the predicted bonds and contacts that the Lac-R protein can make with each of them.
Step 4: Verifying the Predicted Contacts of the Consensus Pattern (Obtained in Step 3) From the Structure of Every Protein-DNA Complex
Verified Contacts From Protein-O1 Complex
Verified Contacts From Protein-O2 Complex
Verified Contacts From Protein-O3 Complex
TYR 17
A14
ARG 22
G6
GLN 18
G8
C'8
Verification of hydrogen bond between Lac-R amino acid and O1 DNA at position 14
Verification of hydrogen bond between Lac-R amino acid and O2 DNA at position 6
Verification of hydrogen bond between Lac-R amino acid and O3 DNA at position 8
Step 5: Alignment of the Three Refined Consensus Patterns (From Step 4) To Get the Distinct Pattern Shared Between the Three Refined Patterns
The distinct pattern represents the minimal amount of direct readout information needed for Lac-R to specifically recognize its binding sites.

Figure 2. Workflow diagram depicting the creation, alignment, verification and final realignment of the Lac repressor protein. Nucleotide numbering is taken from alignment output. Steps 1 and 2: the algorithm takes input sequences and converts them into arrays of hydrogen bonds (blue circle: hydrogen bond donor, red circle: hydrogen bond acceptor, white circle: methyl group). Step 3: the algorithm aligns the various arrays to extract all possible information that is shared among the sequences. Step 4: the PDB structures were used to verify which contacts are present in the actual protein-DNA interactions. All numbering of amino acids is taken from the PDB file's internal sequencing information. Numbering of nucleotides is from the alignment output. Step 5: we realign the verified arrays to produce a final distinct pattern of information from all input sequences that is used by the protein itself.

Table 1. The sequences of the DNA-binding sites

DNA-binding protein	Operator	DNA sequence	PDB ID
Lac repressor	O1	GAATTGTGAGCGGATAACAATTT	2KEI, 1L1M
	O2	GAAATGTGAGCGAGTAACAACCG	2KEJ
	O3	CGGCAGTGAGCGCAACGCAATTC	2KEK
	Symmetrical sequence	GAATTGTGAGCGCTCACAATTC	1CJG
C-protein	O_L	ATGTGACTTATAGTCCGTG	3S8Q, 4IWR
	O_R	CGTGTGATTATAGTCAACA	3CLC
	O_M	ATGTAGACTATAGTCGACA	3UFD
λ-repressor	O_R1	TACCTCTGGCGGTGATA	1LMB (mutated)
	O_R2	TAACACCGTGCGTGTTG	1LMB (mutated)
	O_R3	TATCACCGCAAGGGATA	1LMB (mutated)
	O_L1	TATCACCGCCAGTGGTA	1LMB
	O_L2	CAACACCGCCAGAGATA	1LMB (mutated)
	O_L3	TATCACCGCAGATGGTT	1LMB (mutated)

Step 4: The crystal or NMR structures for the various protein-DNA complexes were obtained from the Protein Data Bank (PDB, www.rcsb.org). These structures were used to verify that the maintained bonds in the alignment are indeed used by the protein for binding and recognition. Any bonds and contacts not detected in the available structures were eliminated from the consensus pattern. The 'H-bonds' structural analysis tool, built into UCSF ChimeraX, was used to identify and analyze the hydrogen bonds that formed between the protein and the DNA. The numbering of amino acids and nucleotides in the manuscript here is taken from the sequence information in the PDB file. The relax distance tolerance was 0.4 Å and the relax angle tolerance was 20.0° (18). We only displayed hydrogen bonds that have at least one atom in the distinct pattern from step 2 of our workflow (Figure 2).

All the hydrogen bonds that were detected using the previous criteria were kept, even if two hydrogen bonds were detected from the same atom, we left it displayed to avoid any user-bias of the results. The 'Contacts' structural analysis tool was used to detect van der Waals interactions between the methyl group of thymine and hydrophobic groups on amino acids (19). The main focus was on amino acid residues that are directly contacting the DNA. Interchain interactions that were not included in DNA binding were ignored, however, these bonds were often identified by the software. Again, to avoid bias in our results we left those hydrogen bonds in but did not consider them in further analyses.

Step 5: We obtained the final refined patterns by aligning the verified contacts that were detected in the published structures for each binding site complex. This provided the final 'distinct pattern' of the common hydrogen bonds and van der Waals contacts that formed between a cognate protein and its different binding sites.

Criteria for PDB structure selection

All the selected structures from PDB should satisfy the following conditions: high resolution crystal structures (up to 3.0 Å) which provides detailed information about protein-DNA interaction or NMR structures, the DNA strands have the sequence of the known binding sites, and non-mutated structures except for Lac R NMR structures which were mutated to link the dimeric Lac R headpiece covalently to facilitate the NMR studies (20). Structures with consensus sequences, palindromic DNA sequences, or any mutated DNA sequences were excluded from the analysis since the algorithm is built on analysis of the real and exact binding sites' sequences. Also, structures that have inducers or factors that affect the natural binding were excluded since the study is mainly concerned with analysis the absolute conditions of binding that happens in nature without the presence of any external influences.

Nucleic acid mutations and energy minimization

The 'Swapna' command in UCSF Chimera mutates one nucleic acids base to another. After making the required mutations for the DNA strands in the protein-complex, the energy minimization function in UCSF Chimera was used to relax the entire complex structure. UCSF Chimera uses the AMBER forcefield to minimize protein structures. First, it performs *Steepest descent* minimization to relieve highly unfavorable clashes. Then, it performs *conjugate gradient* minimization to reach an energy minimum. The parameters for energy minimization were steepest descent steps: 100, steepest descent step size: 0.02 Å, conjugate gradient steps: 10, conjugate gradient step size: 0.02 Å, update interval: 10 and no atoms were fixed.

RESULTS

Lac R-DNA specific binding results and data analysis

The Lac R protein controls the transcription of lactose metabolizing genes (20–22). Transcription is repressed by Lac R binding, as a dimer, to its operator site O1(20,23). Repression is further enhanced by binding to the two auxiliary operator sites O2 or O3 (20). The binding affinity of Lac R is highest for O1 followed by O2, and finally O3 (20). The three sequences of the operator sites were obtained from the literature (Table 1) (20). The contacts arrays derived from these sequences were aligned to produce an initial pattern (Figure 3B), which shows that the bases in positions 6–12 and 18–20 are entirely conserved throughout the three operator sequences. Some positions had no common information and appeared as empty columns (locations 1, 4, 5, 13, 15 and 23, Figure 3A, B). However, we noticed that in locations 2, 3, 14, 16, 17, 21 and 22, the hydrogen bonds are maintained, but the base pair identity is different (Figure 3A, B).

The available structures for Lac R operator complexes were obtained from the PDB. Four NMR structures were

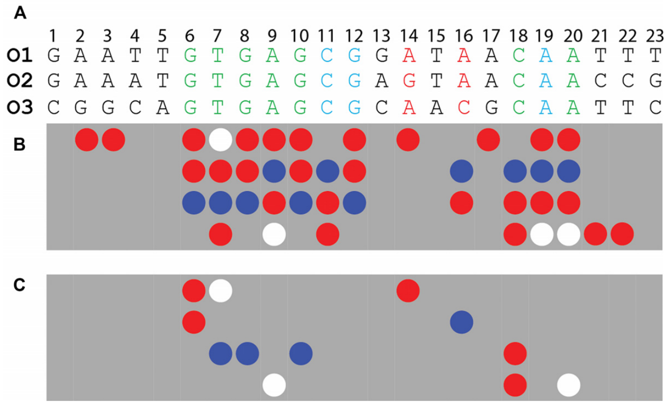

A
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
O1 G A A T T G T G A G C G G A T A A C A A T T T
O2 G A A A T G T G A G C G A G T A A C A A C C G
O3 C G G C A G T G A G C G C A A C G C A A T T C
B
C

Figure 3. Lac operators' sequences, their consensus pattern and their final distinct pattern. (**A**) The three operators' sequences are color coded. Black letters indicate base pairs that are different among the three operators and do not have any maintained hydrogen bonds in the final distinct pattern shown in (C). Green letters indicate the conserved base pairs which have direct bonds and van der Waals contacts with the Lac R shared among the three operators. Blue letters are conserved base pairs which do not have any bonds or contacts shared among the three operators in the final distinct pattern shown in (C). Red letters indicate different base pairs in the three operators which contribute the same hydrogen bond to Lac R binding, and appear in the final distinct pattern in (C). (**B**) The consensus pattern that resulted from aligning the three sequences' hydrogen bond donors (blue circles) and acceptors (red circles) pattern, white circles represents methyl groups. (**C**) The distinct pattern of bonds and contacts shared among the three operators' complexes.

used: two structures for Lac R-O1 complex (PDB ID: 2KEI and 1L1M)(20,23), one structure for Lac R-O2 complex (PDB ID: 2KEJ) (20) and one structure for Lac R-O3 complex (PDB ID: 2KEK)(20). The predicted bonds and contacts were verified from the consensus sequence using the NMR structures.

The structure of lac R-O1 complex showed 20 bonds and interactions in the major groove in the consensus pattern (Supplementary Figure S1A), while the NMR structures of O2 and O3 protein–DNA complexes show 13 hydrogen bonds and methyl group interactions (Supplementary Figure S1B, C). By comparing all the refined hydrogen bonds and interactions that lac repressor can make with each of the three operator sequences, a distinct pattern was extracted (Figure 3C). We believe the distinct pattern represents the minimal number of specific bonds and contacts Lac R needed to recognize these three binding sites.

This distinct pattern was then analyzed to see what base pairs could make these bonds and interactions to Lac R protein. We found that most of the interactions came from conserved base pairs among the three operators (Figure 3A, green colored). However, in locations 14 and 16 (red colored), Lac R made hydrogen bonds despite different base pairs being present at these positions. In addition, three base pairs were maintained in the three operator sites in positions 11, 12 and 19, but we do not observe common hydrogen bonding to Lac R protein maintained in the three operators (blue colored). These results indicate that Lac R recognizes specific hydrogen bonds in the same location of all three operators regardless of the base pair identity.

A deeper analysis aligning two binding sites together was run to see how the information changes between individual operators and if that can shed any light on the order of binding. We compared O1 and O2, O1 and O3, and finally O2 and O3. We first chose the indispensable operator O1 to the auxiliary operator O2. These two sequences are the same except for four nucleobases at locations 4, 13, 14 and 23 (Supplementary Figure S2A). The verified pattern of hydrogen bonds for operators O1 and O2 was extracted (Supplementary Figure S2B). The distinct pattern indicates that all the bonds and contacts are made from the conserved nucleobases in the two binding sites except for one bond at location 14 that was maintained despite the change in the base pair identity from A in binding site O1 to G in binding site O2 (Supplementary Figure S2C).

Similarly, the operator O1 was aligned to operator O3 to see which contacts both sequences have in common. We observed many differences between the O3 sequence relative to O1. However, most of the conserved nucleobases in both operators make the same bonds and contacts with Lac R protein. Additionally, we see that there are two hydrogen bonds maintained despite the difference of the nucleobase identity from A-T in O1 to C-G in O3 at location 16 (Supplementary Figure S3).

Next, operators O2 and O3 were aligned together. We found that most of bonds and contacts originate from the conserved base pairs in the two operators. Interestingly, there are three hydrogen bonds maintained in the two operators regardless the identity of the nucleobases in three different locations: 13, 14 and 16 (Supplementary Figure S4).

This work can shed new light on previous studies that investigated the binding interface of Lac R protein, and its DNA binding sites. There were four amino acids noted to be responsible for the recognition of target DNA: Arg^{22}, Gln^{18}, Tyr^{7} and Tyr^{17} which agreed with previous studies (Supplementary Figures S2–S4). The Tyr^{17} hydroxyl group is responsible for the hydrogen bonding to location 14 in all operators (Supplementary Figures S2–S4). It was previously observed that Tyr^{17} makes hydrogen bond to the 7-position-N in either A or G.

Kalodimos *et al.* emphasized the importance of Tyr^{17} hydroxyl group in the specific binding of Lac R. They showed that mutating Tyr^{17} to Phe (Y17F) dropped the affinity ∼100-fold (24). They also showed that the mutant repressor has 10-fold reduction in binding affinity to nonspecific sequences relative to the wild-type repressor. Through the lens of our data, we interpret this 100-fold affinity reduction to have been, in part, due to the protein losing one of the key contacts used to identify its sequence: the Tyr-OH group that contacts G/A at position 14. Even though the base pair changed, the hydrogen bonding pattern was maintained allowing the protein to recognize the site without having to mutate itself. Our findings affirm that Lac R could recognize specific distinct pattern of contacts and highlight some interactions that may have been lost to evolutionary analyses made based on the base pair identity.

To further validate our hypothesis, we next investigated the Lac R binding a symmetrical sequence. The hydrogen bonds and contacts pattern were verified using the NMR structure of the Lac R protein and this sequence, taken from Spronk *et al.* (25). The Lac R headpiece consists of three helices in a canonical helix-turn-helix DNA binding motif plus nine more residues at the C-terminal that form the so-called hinge region α-helix upon binding to its specific DNA sequence (20). In case of non-specific binding of Lac-R or the absence of the DNA, these nine residues remain unstructured, which helps in distinguishing the specific binding mode of Lac R from the non-specific binding mode (20,24). Although this symmetrical sequence is not one of the known Lac R binding sites, Lac R binds to it and forms the hinge region α-helix which used to be seen in the specific binding mode (24,25).

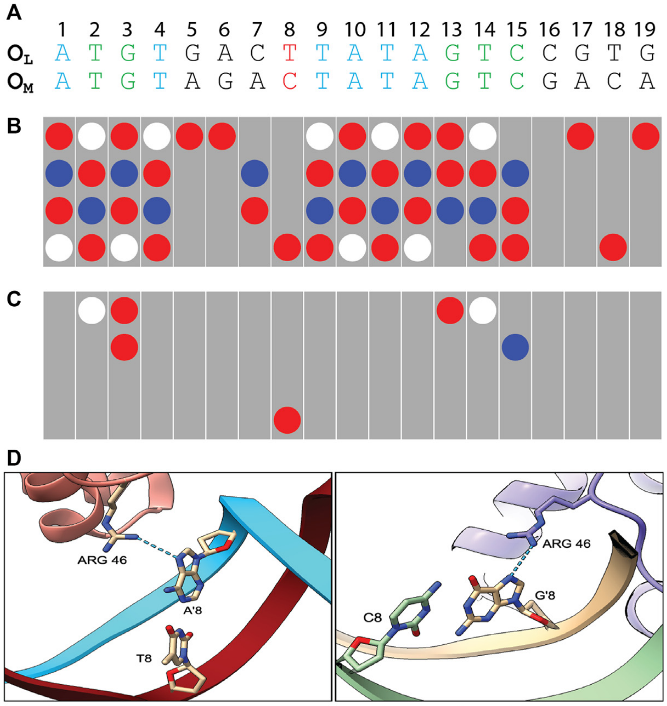
The symmetrical sequence includes 22 bp (Table 1). The first 11 bp are identical to the first 11 bp of the Lac R binding site O1, but the second half has different sequence. We inspected the binding pattern for the symmetrical sequence to understand how Lac R could identify and bind it, forming the hinge region, even though it is not one of its known binding sites. For the binding pattern inspection, we used the published NMR structure by Spronk *et al.* (PDB ID: 1CJG) to verify the hydrogen bonds and contacts for the symmetrical sequence. Then, the binding pattern of the symmetrical sequence was compared to the binding pattern of Lac R indispensable operator O1 since they share the same sequence in the first 11 base pairs.

During our alignment, we noticed that O1 operator is longer than the symmetrical sequence by one base pair. Adding a blank space to account for this, we aligned the 2 sequences and found 18 common bonds and contacts (Supplementary Figure S5). The blank space was entered in position 12 to avoid impacting any area where Lac R should bind. These 18 bonds and contacts represent 60% of the hydrogen bonds and contact, that potentially can be made, shown in O1 which may be enough for the protein to define the symmetrical operator as a binding site and allow formation of the hinge region despite the missing 40% of contacts. We believe this 60% sequence is the minimal information required and note that the certain position 16 shows hydrogen bonds formed despite changes in base pair identity.

Controller protein–DNA specific binding results and data analysis

The restriction-modification (RM) system is considered a primitive immune system in bacteria that protects them from bacteriophage infection (8,26). The proteins that regulate this system are called Controller proteins (8). The operator sequence includes two binding sites: O_L binds with a higher affinity, compared to O_R (8). Martin *et al.*, showed the crystal structure of C-protein binds O_R only as a dimer and O_L + O_R as a tetramer (26). Surprisingly, C-protein doesn't bind O_R with a helix-turn-helix (HTH) motif, it binds 'end-on' to O_R making very few interactions (26). The protein structure in this complex closely matches the free protein structure.(26) It was also shown that O_L binding increases the affinity of C-protein binding at O_R by two orders of magnitude by opening the major groove of O_R to bind another C-protein dimer (26).

C-protein recognizes three DNA sequences, which were used to make a consensus pattern (Figure 5B) (8). However, we took into consideration that C-protein doesn't bind O_R independently, it requires O_L binding first. Thus, a second consensus pattern of only O_L and O_M (OLM consensus, Figure 4B) was made. As predicted, this consensus shows more interactions because the C-protein can identify both operators independently and only two DNA sequences are compared. In the OLM consensus pattern, we see nucleotide positions where the whole base pair is preserved; ones where hydrogen bonds are preserved but the base pair themselves are different, and ones where nothing is preserved.


A
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
O_L A T G T G A C T T A T A G T C C G T G
O_M A T G T A G A C T A T A G T C G A C A
B
C
D
ARG 46
A'8
T8
ARG 46
G'8
C8

Hydrogen bond between A'8 of O_L and Arg^{46} of C-protein

Hydrogen bond between G'8 of O_M and Arg^{46} of C-protein

Figure 4. (**A**) The sequences of O_L and O_M binding sites are color coded. Color coding is the same as in Figure 3A. (**B**) OLM consensus pattern of the aligned two operators that C-protein could recognize, O_L and O_M. The representation of colored circles is the same as in Figure 3B. (**C**) The distinct pattern of bonds and contacts shared between the two operators. The representation of colored circles is the same as in Figure 3B. (**D**) the hydrogen bonds at locations eight verified from the crystal structures of the two operators.

The crystal structures of O_L and O_M were used to refine the hydrogen bonds in the OLM consensus. Four crystal structures are available: two crystal structure for C-protein-O_L complex (PDB IDs: 3S8Q and 4IWR)(26–28), one crystal structure for protein–O_L + O_R complex (PDB ID: 3CLC) (28), and one structure for protein–O_M complex (PDB ID: 3UFD) (8).

Using the available crystal structures of O_L, 10 bonds and interactions in the OLM consensus were verified while the available crystal structure of O_M only verified eight (Supplementary Figure S6). By aligning the two refined patterns together, a distinct pattern of seven bonds and interactions were found (Figure 4C). Analyzing the base pairs that contribute to this distinct pattern, we found six bonds and interactions that come mainly from conserved base pairs in the two operators and there is one bond in location eight coming from different base pairs in the two operators (T-A and G-C) (Figure 4A). Interestingly, the center TATA se-

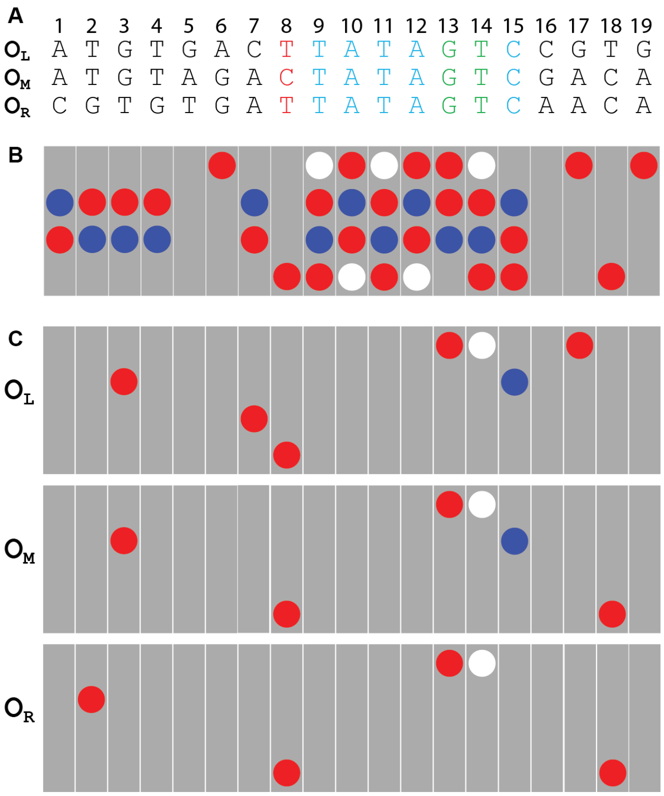

A
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
OL A T G T G A C T T A T A G T C C G T G
OM A T G T A G A C T A T A G T C G A C A
OR C G T G T G A T T A T A G T C A A C A
B
C
OL
OM
OR

Figure 5. The consensus and distinct patterns of bonds and interactions of the three operator complexes for the controller protein. (**A**) The three operators' sequences are color coded. Color coding is the same as in Figure 3A. (**B**) The consensus pattern of the aligned three operators that C-protein can bind. The representation of colored circles is the same as in Figure 3B. (**C**) The pattern of the bonds and the interactions of each binding site verified from each crystal structure of the corresponding complex in the consensus pattern. The representation of colored circles is the same as in Figure 3B.

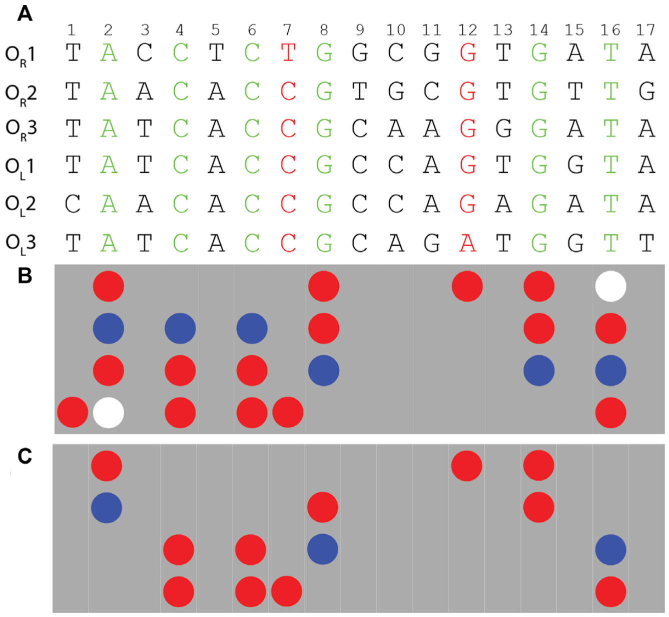

A
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
OR1 T A C C T C T G G C G G T G A T A
OR2 T A A C A C C G T G C G T G T T G
OR3 T A T C A C C G C A A G G G A T A
OL1 T A T C A C C G C C A G T G G T A
OL2 C A A C A C C G C C A G A G A T A
OL3 T A T C A C C G C A G A T G G T T
B
C

Figure 6. The consensus and distinct patterns of bonds and interactions of the six λ-phage operators. (**A**) The six operators' sequences are color coded. Color coding is the same as in Figure 3A. (**B**) The consensus pattern that results from aligning the six sequences' hydrogen bond donors and acceptors pattern. The representation of colored circles is the same as in Figure 3B. (**C**) The final distinct pattern of the bonds and the interactions shared among the six binding sites, verified from the corresponding crystal structures. The representation of colored circles is the same as in Figure 3B.

quence did not contribute directly with specific bonds from the major groove to the specific recognition and binding. We believe this is due to structural aspects, indirect readout, and other considerations as none of the hydrogen bonds are used in any of the structures.

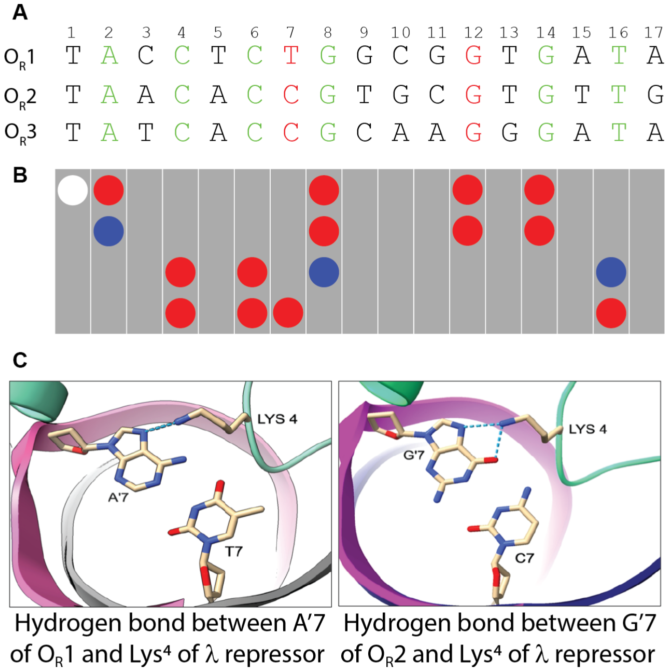
The verified hydrogen bond patterns from the three operators were compared to see how much the binding pattern of O_R matches O_L and O_M. The results showed that O_R has a distorted distinct pattern compared to the other two operators. However, the hydrogen bond in location 8 that comes from different base pairs is maintained in the low affinity O_R (Figure 5). We believe this distorted pattern is what contributes to the lower affinity, O_L binding is thus required to help position the C-protein and assist in deforming the DNA to properly form the required contacts for DNA binding.

λ-phage repressor-DNA specific binding results and data analysis

Bacteriophage λ is a virus that infects *Escherichia. coli*. Upon infection the phage can enter into either a silent life cycle or a virulent life cycle (29). This decision is, in part, controlled by a transcriptional repressor protein named CI (30,31). CI binds in two different promoter regions of the phage genome P_R and P_L (31). Each of these promoters comprises three different operator sites where CI binds as a dimer (30,31). The six operators are termed as O_R1, O_R2, O_R3, O_L1, O_L2, and O_L3. The genomic sequences of λ-phage's six operator sites are available online in the NCBI taxonomy database (32,33). The consensus pattern was made using the six operator sequences (Figure 6B). There are three positions where hydrogen bonding is preserved despites variation in the base pair identity, six positions where the nucleotides are preserved, and eight positions where nothing is preserved (Figure 6A).

The published crystal structure for CI-O_L1 complex from Beamer and Pabo (1LMB) was used for our analysis due to its high resolution (34). Since there are no crystal structures available for the other protein-operator complexes, we mutated the DNA sequence of 1LMB in UCSF Chimera to the other five operator sequences. We minimized the complex structures to relax the mutated DNA complexes before detecting the hydrogen bonds. The λ-CI protein has a flexible arm that interacts with specific DNA nucleobases (9). We noticed this arm is cut off in one of the protein monomers. Therefore, the sequence of each of the six operators were mutated into the crystal structure twice, once running forward and one running in the reverse direction such that each monomer is contacting the DNA close to the 5′ and 3′ end of the same strand. This allowed us to approximate the interaction between the flexible arm and all of the nucleotides.

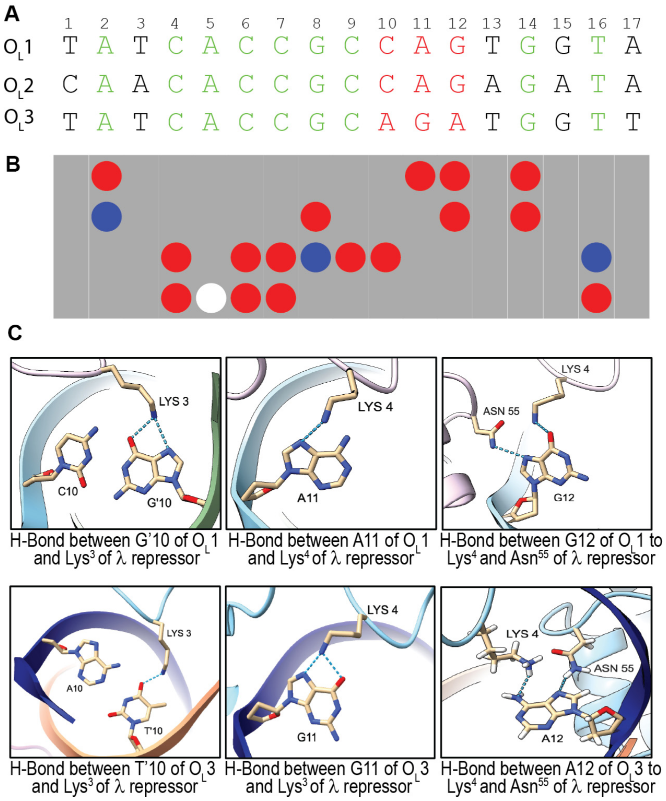
Then, all the DNA–protein complexes were prepared, and the hydrogen bonds were verified to generate the refined


A
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
O_R1 T A C C T C T G G C G G T G A T A
O_R2 T A A C A C C G T G C G T G T T G
O_R3 T A T C A C C G C A A G G G A T A
B
C
LYS 4
A'7
T7
LYS 4
G'7
C7

Hydrogen bond between A′7 of O_R1 and Lys⁴ of λ repressor

Hydrogen bond between G′7 of O_R2 and Lys⁴ of λ repressor

Figure 7. The distinct pattern of bonds and interactions common among the three binding sites of the λ-phage O_R region. (**A**) The three operators' sequences are color coded. Color coding is the same as in Figure 3A. (**B**) The distinct pattern of bonds and contacts shared among the three operators. The representation of colored circles is the same as in Figure 3B. (**C**) the hydrogen bonds at location 7 verified from the crystal structures of the two operators O_R1 and O_R2.


A
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
O_L1 T A T C A C C G C C A G T G G T A
O_L2 C A A C A C C G C C A G A G A T A
O_L3 T A T C A C C G C A G A T G G T T
B
C
LYS 3
C10
G'10
LYS 4
A11
LYS 4
ASN 55
G12
LYS 3
A10
T'10
LYS 4
G11
LYS 4
ASN 55
A12

H-Bond between G′10 of O_L1 and Lys³ of λ repressor

H-Bond between A11 of O_L1 and Lys⁴ of λ repressor

H-Bond between G12 of O_L1 to Lys⁴ and Asn⁵⁵ of λ repressor

H-Bond between T′10 of O_L3 and Lys³ of λ repressor

H-Bond between G11 of O_L3 and Lys³ of λ repressor

H-Bond between A12 of O_L3 to Lys⁴ and Asn⁵⁵ of λ repressor

Figure 8. The distinct pattern of bonds and interactions common among the three binding sites of the λ-phage O_L region. (**A**) the three operators' sequences are color coded. Color coding is the same as in Figure 3A. (**B**) The distinct pattern of bonds and contacts shared among the three operators. The representation of colored circles is the same as in Figure 3B. (**C**) The hydrogen bonds at locations 10, 11 and 12 verified from the crystal structures of the two operators O_L1 and O_L3.

distinct pattern, showing what information is maintained among the six sequences (Figure 6C). We found that most of the hydrogen bonds are from base pairs that were conserved in the six operators except for two hydrogen bonds at locations 7 and 12 (Figure 6A).

We then aligned the individual left and right operators to investigate how CI can tell them apart. The first alignment was for the three binding sites of the right operator (O_R), and it showed that most of the hydrogen bonds shown in the distinct pattern are from conserved base pairs except for hydrogen bonds at location 7 which are maintained despite the change of the base pair identity (Figure 7).

In the next step we wanted to see if the left and right operators have unique information that assists the CI protein in recognizing one set of sequences over the other. The binding sites from the left operators were aligned together. Most of the base pairs are conserved and showing the same pattern of hydrogen bonding. However, three non-conserved base pairs show the same pattern of hydrogen bonding at locations 10, 11 and 12 (Figure 8). In addition, we observe that the amino group of Lys^4 is unexpectedly donating a hydrogen bond to the N atom of 6-position of adenine 12 in O_L3 (Figure 8C) (35).

The alignment of λ-phage's binding sites with other strains' binding sites

λ-phage is one strain of lambdoid phages family that is known to produce Shiga toxins (36). To better understand how information transfers through evolution, a comparative analysis of the λ phage's binding sites and the binding sites of other evolutionary related phages was run. We chose to include Enterobacteria phage VT2-Sakai (VT2-SA) and Stx2 converting phage I (Stx2 I) due to their sequence availability, close evolutionary relation, and the fact that they produce Shiga toxins. Each strain has six binding sites, the same as λ phage. Six alignments were run, one for each operator site for each of the three strains. For the verification step, the contacts from λ-phage were used as the other two strains do not have published structures of CI bound to DNA. The bonds verified from the λ-phage crystal structure 1LMB, were kept while the other bonds were removed.

From this analysis we found that almost all the hydrogen bonds conserved between the three strains arise from different nucleotides (Figure 9). These results reveal some information is hidden if the nucleobase identity is only considered in a comparative analysis. Interestingly, the O_R1 sequence had the least amount of overlap among phage strains, but is noted to have the highest affinity (37). We hypothesize that a more selective O_R1 binding allows the phage to screen for its own DNA from co-infecting phages in the same bacterium. However, further analysis is needed to confirm that, which will be addressed in future studies.

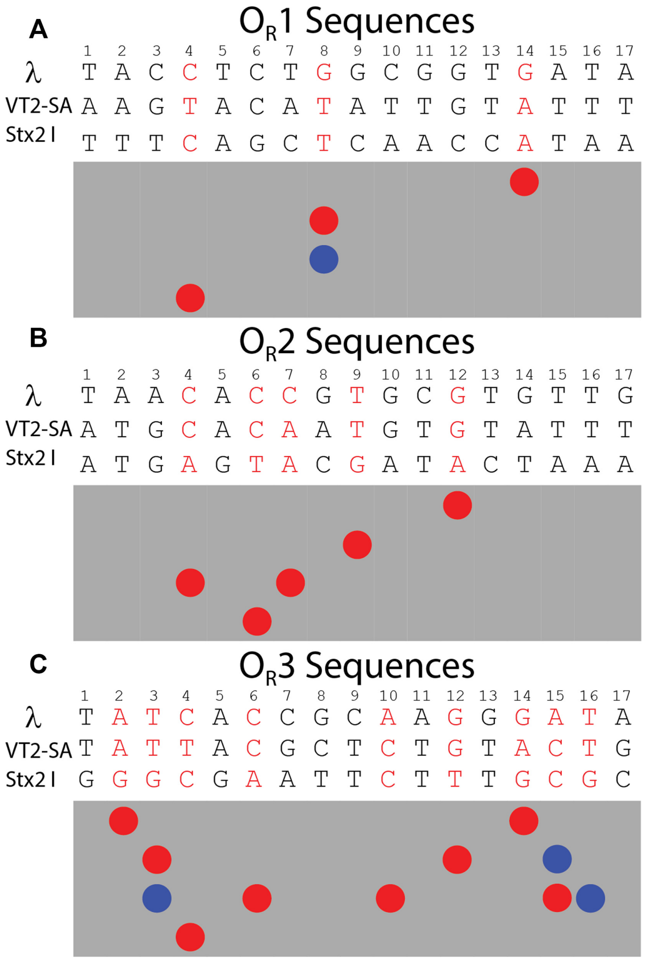

A
O_R1 Sequences
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
λ T A C C T C T G G C G G T G A T A
VT2-SA A A G T A C A T A T T G T A T T T
Stx2 I T T T C A G C T C A A C C A T A A
B
O_R2 Sequences
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
λ T A A C A C C G T G C G T G T T G
VT2-SA A T G C A C A A T G T G T A T T T
Stx2 I A T G A G T A C G A T A C T A A A
C
O_R3 Sequences
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
λ T A T C A C C G C A A G G G A T A
VT2-SA T A T T A C G C T C T G T A C T G
Stx2 I G G G C G A A T T C T T T G C G C

Figure 9. A comparative analysis among the O_R sites of the lambdoid phages: λ-phage, VT2-SA and Stx2 I. (**A**) The alignment of O_R1 sequences from the three strains. (**B**) the alignment of O_R2 sequences from the three strains. (**C**) The alignment of O_R3 from the three strains. See Supplementary Figure S7 for O_L sites' alignments. The representation of colored circles is the same as in Figure 3B.

DISCUSSION

Protein–DNA binding is vital, underpinning many biological processes such as replication, transcription, and more in all known organisms. Thus, understanding how DNA-binding proteins recognize and bind specifically to their target DNA can contribute to the development of new gene therapies and drugs.

Many factors that contribute to specific recognition and binding are represented in direct and indirect readout of DNA by the protein. Most studies agreed on direct readout as the main factor for recognition and specificity with consideration to the other indirect readout factors. In this work, we only sought to address elements of direct readout, namely the hydrogen bond pattern exposed to proteins in the major groove. We are aware that this is only one part of the entire DNA-binding process, but other considerations (e.g. shape readout, hydrogen bonds to the ribose and phosphate in the backbone, etc.) are beyond the scope of this present work.

In this work, we looked at a class of hydrogen bonds and van der Waals interactions that may be overlooked with standard alignment methods and developed an algorithm that can extract them from sequence information. This study had a specific focus on those DNA-binding proteins that can recognize and bind more than one sequence. Our hypothesis is that each specific protein binds its corresponding DNA sequences through a network of hydrogen bonds and contacts in the major groove, and analyses focused on the base pair identity may overlook key interactions. The study comprised three proteins that are known of multiple binding sites, Lac R, C-protein and λ-CI. The different DNA sequences of each protein were analyzed through the designed algorithm to extract the hydrogen bonds and non-covalent contacts maintained in these different sequences to reveal any overlooked key interactions.

From our studies, many of these key interaction bonds were highlighted. All the examples used in this study have positions where DNA base pairs are variable, but the hydrogen bonds that connect the protein with DNA, are maintained. Interestingly, in Lac R and C-protein, some conserved nucleotides did not contribute to the network of hydrogen bonds as was expected. We suspect these may take part in indirect readout or other structural aspects of DNA recognition which, as noted, is beyond the scope of this work.

To test our hypothesis, we were fortunate that published data exists that we could use to evaluate whether or not a protein can recognize a different sequence that maintains the same hydrogen bond pattern. We investigated a symmetrical sequence binding to Lac-R for this analysis (25). Interestingly, Lac-R could recognize and specifically bind this symmetrical sequence, forming the hinge region, although it retains 60% of the contacts that potentially made by Lac R-O1. On the other side, Lac R could not show specific binding to a sequence that does not maintain the same hydrogen bond pattern and instead it bound non-specifically without forming the hinge region (24) which further supports our hypothesis that DNA binding proteins recognize their DNA target through a network of hydrogen bonds and contacts in the major groove and analyses of base pair identity may overlook some important key interactions for recognition and specificity.

Similarly, we believe our work adds a new perspective to the work of Lin and Guo. Their paper showed that certain proteins only read information from one strand of DNA. In those situations, the effect of maintaining a hydrogen bond can further reduce specificity. A to G mutations maintain the 7-position nitrogen, therefore proteins making that contact could not screen these two nucleotides from one-another based solely on the 7-position lone pair. That leaves only one hydrogen bond available to discern the sequence (the 6-position amino or carbonyl group). We show that the information can be even more variable in that case thereby lessening their specificity more.

We also show some possibly new evolutionary relationships between different phage strains and ways that viruses can screen genomes to bind the correct operator site. Our algorithm indicated the presence of hydrogen bonds that are shared among the binding sites of the three strains. The consideration of the hydrogen bond pattern presented by the nucleobase in the analysis revealed some hidden information which might be ignored when considering only the base pair identity. It is possible that this information may have a hand in the evolutionary trajectory of phages. Based on our results, we suspect that If an operator site mutates, the CI protein will have to mutate accordingly to regain proper binding affinity. However, if the mutation does not change the information (as described here) then no CI mutations would be required. Thus, it is possible that some mutations are benign and allow for other mutations elsewhere to accumulate. In addition, we suspect that the CI repressor of λ-phage might bind the operator sites of either VT2-SA or Stx2 I with fair affinity. However, the *in vitro* data are currently being run and are anticipated for presentation in a future study.

From this study, we find that the most common nucleotide change that maintains hydrogen bonds come from purine to purine. In this case, the 7-position nitrogen provides a lone pair of electrons for hydrogen bonding. This is responsible for the majority of the flexibility we see and is a common target for DNA binding proteins. We also see adenine to cytosine mutations retain a hydrogen bond donor from the amino group of adenine or cytosine and one hydrogen bond acceptor from the Carbonyl group on either thymine or guanine on the complement DNA strand. These combinations provide a lot of information-retention when DNA is mutating. Each base pair has a mutant that can retain the hydrogen bonding character. These interactions are often overlooked if one is only considering the identity of the base pairs themselves. We noticed that the change in the nucleobases is not limited to the typical change between the purine bases (A and G) or the pyrimidine bases (C and T), but also it happens to be a change from purine base to pyrimidine base and vice versa.

Although each of the three proteins, Lac R, C-protein and λ-CI, could recognize and specifically bind to multiple binding sites, we believe that the changes in the base pairs among these different binding sites are responsible for the variation of its affinity of binding that we discussed in each protein's respective results section. Also, the variation of the base pairs from G-C to A-T could affect the structure of the DNA which, in part, contributed to the different binding affinities among the operators of the three proteins.

Future studies will address how other factors affect binding interactions and will be incorporated into the algorithm, as well as look into *in vitro* testing of our hypothesis, search for new ways to apply this work, and expand the algorithm to incorporate chemically modified bases (e.g. methylation), and other structural factors that affect DNA-protein binding.

CONCLUSIONS

From the results, we conclude that DNA binding proteins recognize their DNA target through a network of hydrogen bonds and contacts in the major groove. The focus solely on the identity of the nucleobases can lead analyses to overlook some important key interactions for recognition and specificity. We believe that this work will have a multitude of applications. For example, protein design groups seeking to develop artificial transcription factors (ATFs) could use our approach to better screen out the minimal required information and target those hydrogen bond partners when looking at the interface. This could lead to ATFs with specificity toward multiple sequences as well as a deeper understanding of how existing ones recognize their target DNA. Similarly, structural biologists will benefit from this work by better identifying hydrogen bonds that could be made between proteins and their corresponding DNA binding sites.

We also believe those studying evolution will benefit from this new type of analysis. Our work seeks to better identify the information itself within the DNA. Focusing in on this can help researchers trace how certain mutations can arise first and why some mutations cause more noticeable effects than others. As discussed above, our work can help those groups identify which pieces of the information displayed are more or less important, and from there how interactions with different proteins can be more or less affected by evolutionary changes.

Importantly, we see our work complimenting existing studies that generate consensus sequences to examine DNA binding to multiple sites. Our work can help identify which specific nucleotide positions are important, and hopefully uncover new ones that were missing in previous analyses.

DATA AVAILABILITY

HBondAlign is an open source code available in the GitHub repository at https://github.com/SolomonLabGMU/HBondAlign.

We have also deposited our data into Code Ocean, which can be found here: https://doi.org/10.24433/CO.3089765.v1.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

Molecular graphics and analyses performed with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases.
Molecular graphics and analyses performed with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311.
We would like to thank Drs Gerald Koudelka and Chris Bialas for their help preparing the manuscript.

FUNDING

George Mason University and the Department of Chemistry and Biochemistry.
Conflict of interest statement. None declared.

REFERENCES

1. Lin,M. and Guo,J. (2019) New insights into protein–DNA binding specificity from hydrogen bond based comparative study. *Nucleic Acids Res.*, **47**, 11103–11113.
2. Emamjomeh,A., Choobineh,D., Hajieghrari,B., MahdiNezhad,N. and Khodavirdipour,A. (2019) DNA–protein interaction: identification, prediction and data analysis. *Mol. Biol. Rep.*, **46**, 3571–3596.
3. Luscombe,N.M. (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res.*, **29**, 2860–2874.
4. Garvie,C.W. and Wolberger,C. (2001) Recognition of specific DNA sequences. *Mol. Cell*, **8**, 937–946.
5. 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.
6. Jayaram,B. and Jain,T. (2004) The role of water in protein-DNA recognition. *Annu. Rev. Biophys. Biomol. Struct.*, **33**, 343–361.
7. Lejeune,D., Delsaux,N., Charloteaux,B., Thomas,A. and Brasseur,R. (2005) Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. *Proteins*, **61**, 258–271.
8. Ball,N.J., McGeehan,J.E., Streeter,S.D., Thresh,S.-J. and Kneale,G.G. (2012) The structural basis of differential DNA sequence recognition by restriction–modification controller proteins. *Nucleic Acids Res.*, **40**, 10532–10542.
9. Hochschild,A., Douhan,J. and Ptashne,M. (1986) How λ repressor and λ cro distinguish between OR1 and OR3. *Cell*, **47**, 807–816.
10. Kumar,S., Bhardwaj,V.K., Singh,R., Das,P. and Purohit,R. (2022) Identification of acridinedione scaffolds as potential inhibitor of DENV-2 C protein: an in silico strategy to combat dengue. *J. Cell. Biochem.*, **123**, 935–946.
11. Rajendran,V., Purohit,R. and Sethumadhavan,R. (2012) In silico investigation of molecular mechanism of laminopathy caused by a point mutation (R482W) in lamin A/C protein. *Amino Acids*, **43**, 603–615.
12. Bhardwaj,V.K., Oakley,A. and Purohit,R. (2022) Mechanistic behavior and subtle key events during DNA clamp opening and closing in T4 bacteriophage. *Int. J. Biol. Macromol.*, **208**, 11–19.
13. 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.
14. Pettersen,E.F., Goddard,T.D., Huang,C.C., Meng,E.C., Couch,G.S., Croll,T.I., Morris,J.H. and Ferrin,T.E. (2021) UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci.*, **30**, 70–82.
15. Goddard,T.D., Huang,C.C., Meng,E.C., Pettersen,E.F., Couch,G.S., Morris,J.H. and Ferrin,T.E. (2018) UCSF ChimeraX: meeting modern challenges in visualization and analysis: UCSF chimeraX visualization system. *Protein Sci.*, **27**, 14–25.
16. Harris,C.R., Millman,K.J., van der Walt,S.J., Gommers,R., Virtanen,P., Cournapeau,D., Wieser,E., Taylor,J., Berg,S., Smith,N.J. *et al.* (2020) Array programming with NumPy. *Nature*, **585**, 357–362.
17. Hunter,J.D. (2007) Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95.
18. Mills,J.E.J. and Dean,P.M. (1996) Three-dimensional hydrogen-bond geometry and probability information from a crystal survey. *J. Comput.-Aided Mol. Des.*, **10**, 607–622.
19. Li,A.J. and Nussinov,R. (1998) A set of van der Waals and coulombic radii of protein atoms for molecular and solvent-accessible surface calculation, packing evaluation, and docking. *Proteins*, **32**, 111–127.
20. Romanuka,J., Folkers,G.E., Biris,N., Tishchenko,E., Wienk,H., Bonvin,A.M.J.J., Kaptein,R. and Boelens,R. (2009) Specificity and affinity of lac repressor for the auxiliary operators O2 and O3 are explained by the structures of their protein–DNA complexes. *J. Mol. Biol.*, **390**, 478–489.
21. Kalodimos,C.G., Boelens,R. and Kaptein,R. (2004) Toward an integrated model of protein−DNA recognition as inferred from NMR studies on the *L ac* repressor system. *Chem. Rev.*, **104**, 3567–3586.
22. Kopke Salinas,R., Folkers,G.E., Bonvin,A.M.J.J., Das,D., Boelens,R. and Kaptein,R. (2005) Altered specificity in DNA binding by the lac repressor: a mutant lac headpiece that mimics the gal repressor. *ChemBioChem*, **6**, 1628–1637.
23. Kalodimos,C.G. (2002) Plasticity in protein-DNA recognition: lac repressor interacts with its natural operator O1 through alternative conformations of its DNA-binding domain. *EMBO J.*, **21**, 2866–2876.
24. Kalodimos,C.G., Biris,N., Bonvin,A.M.J.J., Levandoski,M.M., Guennuegues,M., Boelens,R. and Kaptein,R. (2004) Structure and flexibility adaptation in nonspecific and specific protein-DNA complexes. *Science*, **305**, 386–389.
25. Spronk,C.A., Bonvin,A.M., Radha,P.K., Melacini,G., Boelens,R. and Kaptein,R. (1999) The solution structure of Lac repressor headpiece 62 complexed to a symmetrical lac operator. *Structure*, **7**, 1483–1492.
26. Martin,R.N.A., McGeehan,J.E., Ball,N.J., Streeter,S.D., Thresh,S.-J. and Kneale,G.G. (2013) Structural analysis of DNA–protein complexes regulating the restriction–modification system *Esp* 1396I. *Acta Crystallogr.*, **69**, 962–966.
27. McGeehan,J.E., Ball,N.J., Streeter,S.D., Thresh,S.-J. and Kneale,G.G. (2012) Recognition of dual symmetry by the controller protein C.Esp1396I based on the structure of the transcriptional activation complex. *Nucleic Acids Res.*, **40**, 4158–4167.
28. McGeehan,J.E., Streeter,S.D., Thresh,S.-J., Ball,N., Ravelli,R.B.G. and Kneale,G.G. (2008) Structural analysis of the genetic switch that regulates the expression of restriction-modification genes. *Nucleic Acids Res.*, **36**, 4778–4787.
29. Salmond,G.P.C. and Fineran,P.C. (2015) A century of the phage: past, present and future. *Nat. Rev. Microbiol.*, **13**, 777–786.
30. Stayrook,S., Jaru-Ampornpan,P., Ni,J., Hochschild,A. and Lewis,M. (2008) Crystal structure of the λ repressor and a model for pairwise cooperative operator binding. *Nature*, **452**, 1022–1025.
31. Gao,N., Shearwin,K., Mack,J., Finzi,L. and Dunlap,D. (2013) Purification of bacteriophage lambda repressor. *Protein Expression Purif.*, **91**, 30–36.
32. Schoch,C.L., Ciufo,S., Domrachev,M., Hotton,C.L., Kannan,S., Khovanskaya,R., Leipe,D., Mcveigh,R., O'Neill,K., Robbertse,B. *et al.* (2020) NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database*, **2020**, baaa062.
33. Sayers,E.W., Cavanaugh,M., Clark,K., Ostell,J., Pruitt,K.D. and Karsch-Mizrachi,I. (2019) GenBank. *Nucleic Acids Res.*, **47**, D94–D99.
34. Beamer,L.J. and Pabo,C.O. (1992) Refined 1.8 Å crystal structure of the λ repressor-operator complex. *J. Mol. Biol.*, **227**, 177–196.
35. Kagra,D., Prabhakar,P.S., Sharma,K.D. and Sharma,P. (2020) Structural patterns and stabilities of hydrogen-Bonded pairs involving ribonucleotide bases and arginine, glutamic acid, or glutamine residues of proteins from quantum mechanical calculations. *ACS Omega*, **5**, 3612–3623.
36. Fattah,K.R., Mizutani,S., Fattah,F.J., Matsushiro,A. and Sugino,Y. (2000) A comparative study of the immunity region of lambdoid phages including Shiga-toxin-converting phages. Molecular basis for cross immunity. *Genes Genet. Syst.*, **75**, 223–232.
37. Bell,C.E., Frescura,P., Hochschild,A. and Lewis,M. (2000) Crystal structure of the λ repressor C-Terminal domain provides a model for cooperative operator binding. *Cell*, **101**, 801–811.