Title

Structure of Aart, a designed six-finger zinc finger peptide, bound to DNA

Permalink
https://escholarship.org/uc/item/7x34c3g6

Journal
Journal of Molecular Biology, 363(2)

ISSN
0022-2836

Authors
Segal, David J
Crotty, Justin W
Bhakta, Mital S
et al.

Publication Date
2006-10-20

Supplemental Material
https://escholarship.org/uc/item/7x34c3g6#supplemental

Peer reviewed

Structure of Aart,
a Designed Six-Finger Zinc Finger Peptide,
bound to DNA

David J. Segal[1*,] Justin W. Crotty[2], Mital Bhakta[1,4], Carlos F. Barbas, III[3] & Nancy C. Horton[4*]

[1]UC Davis Genome Center and Department of Pharmacology, University of California, Davis, CA, 95616

[2]Department of Chemistry, University of Arizona, Tucson, AZ, 85721

[3]Department of Molecular Biology and The Skaggs Institute for Chemical Biology, BCC 550, Scripps Research Institute, La Jolla, CA 92037.

[4]Department of Biochemistry and Molecular Biophysics, University of Arizona, Tucson, AZ, 85721

**Corresponding author*

NCH: Telephone: 520-626-3828 FAX: 520-626-9288 nhorton@u.arizona.edu

DJS: Telephone: 530-754-9134 FAX: 530-754-9658 djsegal@ucdavis.edu

Running Title: Structure of Aart/DNA

Summary (150 words, 150 allowed)

Cys2-His2 zinc fingers are one of the most common types of DNA-binding domains. Modifications to zinc-finger binding specificity have recently enabled custom DNA-binding proteins to be designed to a wide array of target sequences. We present here a 1.96 Å structure of Aart, a designed six-zinc finger protein, bound to a consensus DNA target site. This is the first structure of a designed protein with six fingers, and was intended to provide insights into the unusual affinity and specificity characteristics of this protein. Most protein-DNA contacts were found to be consistent with expectations, while others were unanticipated or insufficient to explain specificity. Several were unexpectedly mediated by glycerol, water molecules or amino acid-base stacking interactions. These results challenge some conventional concepts of recognition, particularly the finding that triplets containing 5'A, C, or T are typically not specified by direct interaction with the amino acid in position 6 of the recognition helix.

Introduction

The Cys2-His2 type of zinc finger domain represents an important class of DNA-binding elements. Their usefulness in biology is indicated by the fact that there are more than 500 zinc finger proteins that collectively contain nearly 5,000 of these domains in the human genome[1]. Their relatively simple structure and binding mode have also proven useful for the engineering of custom DNA-binding proteins[2,3], which have recently led to the production of artificial transcription factors and targeted endonucleases[4-9]. The intense interest in zinc fingers as specific binding elements has led to a wealth of insights regarding their interactions with DNA, making zinc fingers the best understood DNA binding protein. However, there remains no comprehensive recognition code for these or any other DNA-binding protein. Even for designed zinc finger proteins, the interactions that give rise to specificity are often poorly understood.

A single domain or “finger” consists of approximately 30 amino acids with a simple ββα fold stabilized by hydrophobic interactions and the chelation of a zinc ion between two histidines and two cysteines[10,11]. These domains can be found in covalent tandem arrays of up to 37 repeats[12], facilitating recognition of extended DNA sequences. The polydactyl nature of zinc finger proteins distinguishes them from many other DNA binding motifs that typically form homo- or hetero-dimers, such as helix-turn-helix, bZIP, hormone receptors, and homeodomains[13]. Presentation of the zinc finger α-helix into the major groove of DNA allows for sequence specific base contacts, with each domain typically recognizing three nucleotides. The first crystal structure of a zinc finger protein bound to DNA, Zif268, suggested specific roles for each residue in the recognition helix[14,15]. With respect to the start of the α-helix, positions -1, 3, and 6 contacted the 3', middle, and 5' nucleotides, respectively, of the target strand (that is, the strand containing the sequence that is considered to be the binding site for that protein). Positions –2, 1 and 5 were often involved in direct or water mediated contacts to the phosphate backbone. Position 4 was typically a leucine, a highly conserved residue that packs in the hydrophobic core of the domain. Position 2 has been shown to interact with other helix residues and DNA bases depending on the helical and DNA sequences. Cooperative interactions between fingers in Zif268 were limited. Although subsequent cocrystal-structures of natural and designed zinc finger proteins would reveal more complex and cooperative recognition paradigms, many individual zinc fingers were found to follow the same general pattern as Zif268 with regard to the role of positions -1, 2, 3, and 6[16].

The unique properties of zinc finger domains led to the experimental manipulation of their specificity[2,17]. In previous work, rational and combinatorial methods were used to modify residues in the recognition helix. Starting with a wild type 3-finger protein, amino acids in positions –1 through 6 of the central domain were randomized. Proteins that could specifically recognize a new three-nucleotide subsite were selected by phage display, then optimized by site-directed mutagenesis. To date, domains have been reported that bind with high affinity and specificity to members of the 5'-GNN-3', 5'-ANN-3', and 5'-CNN-3' sets of DNA triplets[18-21]. The optimized domains can be modularly assembled into six-finger proteins, which have the potential to recognize an 18-bp target site[22,23]. A site of this length should be unique in the human genome, as well as all other known genomes.

One such six-finger protein is Aart, which was designed to bind an artificial (non-biological) A-rich DNA sequence: 5'-ATG-TAG-AGA-AAA-ACC-AGG-3'[20]. This protein is an interesting case study for zinc finger-DNA recognition for several reasons. Most zinc finger proteins bind G-rich sequences, particularly those composed of 5'-GNN-3' triplets. In fact, of the 65 structures of zinc finger domains complexed with DNA in the Protein Data Bank, there are only eight cases in which a 5'-A appears in the finger's binding site, and only a single example in which the 5'-A is recognized by an amino acid in position 6[16]. There is only one case (appearing identically in 11 structures) in which a 5'-T appears in the finger's binding site, and it is not recognized by the amino acid in position 6[15]. Aart provides five examples of domains designed to recognize 5'A, and one to 5'T. Target site selection experiments were performed to determine if the expected binding specificity would be observed *in vitro*[24]. These experiments revealed that Aart actually preferred binding to a more G-rich site (differences underlined): 5'-ATG-<u>(G/T)</u>AG-<u>(A/G)</u>GA-AAA-<u>G</u>CC-<u>CNN</u>-3'. These results suggested that three of the five domains that had been designed to recognize 5'A in fact preferred 5'G in the context of this particular protein. We were therefore interested to understand the structural basis for these unexpected preferences.

Another unusual property of Aart is that it binds its DNA target with picomolar affinity, despite the fact that all six fingers are joined by canonical linkers (TGEKP). Typically, six-finger proteins with all canonical linkers bind with affinities in the nanomolar range[22,23,25,26]. Others have suggested that affinity could be improved by using longer linkers between fingers 3 and 4

(2x3 format) or between both fingers 2 and 3 as well as 4 and 5 (3x2 format)[27-29]. These modifications are predicated on the hypothesis that canonical linkers do not allow successive fingers to maintain proper register with the DNA bases, thus introducing a strain in the protein-DNA complex that results in decreased binding energy. We were therefore also interested to look for structural evidence of strain or misalignment in the six-finger Aart protein, or to understand how Aart could avoid such strain.

Materials and Methods

Overexpression, purification, crystallization, and native data collection have been described[30]. The sequences of the Aart polypeptide and DNA used in the crystallization are given in Fig. 1a-b.

Electrophoretic mobility shift assays:

Target oligonucleotides were labeled at their 5' termini with [^{32}P] and gel purified. Eleven 3-fold serial dilutions of protein were incubated in 20 ml of binding reactions (10 mM Tris, pH7.5/90 mM KCl/1 mM $MgCl_2$/100 μM $ZnCl_2$/1% BSA/5 mM DTT/0.12 μg/μl sheared herring sperm DNA (Sigma)/10% glycerol/1 pM target oligonucleotide) for 3 hr at room temperature, then resolved on a 5% polyacrlyamide gel in 0.5x TBE buffer (90 mM Tris/65 mM boric acid/2.5 mM EDTA, pH 8.3). Dried gels were exposed to XAR film (Kodak) for 48 hours. Quantification was performed from scanned autoradiograms using ImageJ software (NIH). K_D values are the average of at least two independent experiments. The standard error of the measurements was ±50%.

Preparation of Aart/ DNA cocrystals:

Crystals of Aart and DNA were prepared as previously described[30]. Crystals of Aart bound to brominated DNA, where the C5 methyl position of selected thymine nucleotides was substituted with bromine (5-bromouracil), were prepared similarly. The brominated DNA was obtained synthetically (Yale Keck Center) and purified as described[30] using reverse phase HPLC. A total of six unique subsituted oligonucleotides differing in number and position of the 5-bromouracil bases were prepared and crystallized with Aart. MAD data (although using the zinc edge) was collected using a crystal of Aart bound to brominated DNA with 5-bromouracil substituted for thymine at positions marked with an X in Fig. 1b.

Diffraction data collection

MAD data collection using the zinc absorption edge was collected at Argonne National Labs Structural Biology Center BL-19ID (Table 1). Data collection at three wavelengths was performed, corresponding to the inflection, peak, and remote wavelengths. Exposures were 15 seconds, with 2° oscillations per frame. A total of 200° (100° and its inverse in phi) were collected at the peak wavelength first, then 230° at the remote and inflection wavelengths. Lower completion of the data set collected at the inflection wavelength is due to diminished diffraction from the radiation damaged crystal. Native (non-brominated DNA) was collected at SSRL BL 1-5. All diffraction data was integrated, scaled, and reduced with HKL 2000[31].

Structure solution and refinement

The program SOLVE[32] was used to find and refine the zinc sites, as well as to calculate the initial electron density map, which was subsequently improved using the program RESOLVE[32]. The model was built by fitting the previously created homology model (using RCSB code 1AAY) into the clearly defined electron density, and subsequently refined to 2.6 Å against the peak diffraction data set of the MAD data set using manual model rebuilding, simulated annealing, temperature factor refinement, and positional refinement as implemented in CNS[33]. The model was then refined against a high resolution data set of crystals of Aart to non-brominated DNA, to 1.96 Å.

Structure analysis

The final 1.96 Å structure containing no bromouracil was used in the structural analysis. Helical parameters of the DNA were determined using the program CURVES[34] allowing for the calculation of a bent helical axis. Superpositions were performed with the CCP4 program LSQKAB[35]. The rotation angle of recognition helices relative to that of Finger 3 of molecule A were calculated by first superimposing each finger onto Finger 3 using the triplet base pairs recognized by that finger as well as the adjacent 5' (upper strand direction) base pair, and then finding the rotation to superimpose the recognition helices (using alpha carbon atoms of residues -1 to +11). The translation was also calculated from this superposition. The superpositions were performed with recognition helices from the Aart structure as well as with 30 zinc fingers from Protein Data Bank structures with RCSB codes: 1AAY, 1A1G, 1A1J, 1A1F, 1A1K, 1A1H, 1A1I, 1A1L, 1JK1, 1JK2, 1G2F, 1G2D, 1UBD, 1MEY, 2GLI, 2DRP, 1TF6, 1TF3.

Results

Affinity of Aart for the designed and consensus DNA target sites

Aart was designed to bind to the target sequence 5'-ATG-TAG-AGA-AAA-ACC-AGG-3' [20]. However, subsequent target site selection experiments showed that Aart actually preferred binding to a more G-rich site (differences underlined), 5'-ATG-(G/T)AG-(A/G)GA-AAA-GCC-CNN-3'[24]. We chose initially to solve the structure of Aart with 5'-ATG-TAG-GGA-AAA-GCC-CGG-3' (hereafter referred to as the consensus target site), to investigate the structural basis for these unexpected preferences. Electrophoretic mobility shift assays performed with the designed and consensus binding sites found only a slightly higher average affinity of Aart for the consensus site (90 pM and 50 pM, respectively). Raw data for this assay can be found in the Supplementary Information (Fig. S1).

Overall structure of Aart bound to DNA

Refinement statistics of the Aart/DNA complex are given in Table 2, with the crystallographic R factor of the final model of 20.3% and R_{free} of 24.9%. A ribbon diagram of one Aart polypeptide bound to one DNA duplex is shown in Fig. 2a, and an example of a simulated annealing Fo-Fc omit map is shown in Fig. 2b. The asymmetric unit contains two complexes of Aart (molecules A and B) bound to DNA (chains C,D and E,F). All 22 nucleotides of the four strands of DNA are visible, as well as all six zinc fingers, although selected residues of the protein chain including in chain A 1-30, 45-47, 185-190, and in chain B: 1-19, 158-172 (the linker between Fingers 5 and 6), and 179-190 are not. The RMSD of the superposition using the alpha carbons of the polypeptide chains of the two molecules of Aart, A and B, are shown in Fig. 2c (blue line) plotted per residue. The refined Debye-Waller Temperature or B factors for alpha carbon atoms of chains A and B, are also shown in Fig. 2c (solid red line, chain A, dashed red line, chain B). Similarly, the RMSD for the superposition of all atoms of the DNA strands, chains C and E (blue solid line), and chains D and F (lower strands, dashed blue line) is shown in Fig. 2d. The B factors averaged all atoms of each nucleotide for all four DNA strands are also plotted in Fig. 2d. The RMSD and B factors are lowest in the middle residues or nucleotides of each chain, indicating the greatest order at these locations.

Recognition helix orientations

The orientations of the recognition helices of the six zinc fingers of Aart, relative to bound DNA, were compared. First, the triplet DNA recognized by each finger, along with a neighboring base pair (5' using the upper strand direction), were superimposed onto those of all other fingers. The neighboring base pair was included since each recognition helix anchors to the DNA via a histidine to phosphate contact at this base pair. Next, the relative positions of the recognition helices were compared. The recognition helix of Finger 3 was found to be the most consistent with the others, and therefore was used as the reference position. The overall RMSD using alpha carbon atoms from each recognition helix and those of Finger 3, following the superposition of DNA, range from 0.2 to 2.0 Å, but show no systematic trend in displacement from one end of the helix to the other (data not shown).

Next, the rotation and translation required to superimpose the recognition helices onto that of Finger 3 were calculated. Fig. 3a plots the rotation angle required for the superposition for each finger, and Fig. 3b plots the translation. Translations of the centers of mass (COM) vary from 0.1 to 2.3 Å, and rotations from 2° to 16°. The rotation and translation of helices from 30 examples of C2H2 zinc fingers were also calculated and shown in Fig. 3c (red circles) to compare with those of Aart (blue diamonds).

DNA recognition in Aart/DNA

The contacts predicted and used in the original design of Aart are shown in Fig. 4a. Triplets are identified as F1-F6, and correspond to the three base pairs of DNA recognized by zinc Fingers 1-6 of Aart. Fig. 4b represents the results from the CAST (cyclic amplification and selection of targets) assay used to determine the actual sequence preference of Aart. An alternative representation of this data is offered in the Supplementary Information. Fig. 4c summarizes the observed contacts between Aart and DNA in the Aart/DNA structure. Distances between the specific atoms of the protein side chains and DNA bases can be found in the Supplementary Information (Table S1), and includes information from both independent copies in the asymmetric unit. Fig. 5 and 6 show stereo diagrams of these contacts for each finger. Most of the contacts occur between Aart and the upper strand in the duplex DNA, therefore the position of a base pair within a triplet is denoted relative to the upper strand, as 3', middle, or 5'. Some contacts do occur to the lower strand, typically from position 2 on one recognition helix and the base-pairing partner of the 5' base pair of a triplet recognized by a neighboring zinc finger.

Finger 1

Finger 1 of Aart contains the sequence RSD-H-LAE in positions -1 to 6 of its recognition helix at residues 33-39 (Fig. 1a), and was designed to recognize the 3'-GGA-5' triplet (upper strand) in DNA. However, it was found by CAST to specify 3'-NNC-5' (where N indicates any base) (Fig. 4b). The triplet sequence used in the crystal structure is 3'-GGC-5' in the upper (U) strand, and 5’-CCG-3’ on the opposite or lower (L) strand, numbered U19-U21/L3-L5. Figure 5a shows the side chain-base contacts between Finger 1 and DNA in the Aart/DNA molecule A. Position -1 of the recognition helix, R33, is within hydrogen bonding distance to the O6 and N7 of Gua U21 (2.8 Å, 2.8 Å in molecule A, 2.8 Å, 2.8 Å in molecule B), the 3' nucleotide of the first DNA triplet 3'-GGC-5' (Fig. 4c, 5a), despite the fact that the CAST data shows no specificity at this position. The electron density for the side chain of R33 is very poor indicating disorder in molecule A; the poor order may be related to the observed lack of specificity (see below). The side chain of D35 hydrogen bonds to R33 in the expected way, which appears to aid in orienting the arginine side chain. The epsilon nitrogen of H36 (helix position 3) is predicted to recognize a G in the middle position of the triplet, but CAST analysis shows virtually no specificity at this position. In one side chain orientation, the epsilon nitrogen of H36 is within hydrogen bonding distance (2.8 Å in molecule A, 2.8 Å in molecule B) of the Gua U20 N7, the middle nucleotide in the 3'-GGC-5' triplet. In this orientation, it is too far (3.7 Å in molecule A, 3.5 Å in molecule B) from the O6 of this base to form a hydrogen bond. If the side chain were flipped from the currently modeled orientation, the epsilon nitrogen would be in a position to hydrogen bond to the O6 of Gua U20, however, the distance between these groups is 3.4 Å, making for a weak hydrogen bond. Therefore side chain orientation has been chosen to allow the more optimal hydrogen bond to the N7. In addition, this orientation allows the delta nitrogen of H36 to hydrogen bond to the side chain of E39 (molecule B only, the E39 side chain is disordered in molecule A). No direct contacts to the 5' nucleotide of the triplet are visible, however, the side chain of K63 (position 2 of the neighboring helix) makes a hydrogen bond to the N7 of Gua L5 in molecule A. Electron density is absent for the terminal amine and methylene group of the side chain in molecule B. The absent electron density indicates that this side chain samples multiple conformations, and could spend part of the time contacting the O6 of Gua L5.

Finger 2

Finger 2 contains the sequence DKK-D-LTR in positions -1 to 6 of its recognition helix, residues 61 to 67 in Aart, and was designed to recognize the triplet 3'-CCA-5', however, CAST analysis indicates the preference for 3'-CCG-5'. The sequence 3'-CCG-5' in the upper (U) strand and 5'-GGC-3' in the lower (L) strand was used in the crystal structure (U16-U18/L6-L8). The -1 position of the recognition helix, D61, was expected to contact the 3'C, Cyt U18, however, no side chain density is evident for D61 in molecule A, and the side chain is 5.6 Å from the N4 of Cyt U18 in molecule B. Instead, a water molecule bridges the D61 side chain and the N4 of Cyt U18. Further, the water molecule is also hydrogen bonded to D64. If both D61 and D64 are ionized, then both will be accepting hydrogen bonds from the water, leaving it with the capacity to only donating hydrogen bonds. This could, in principle, provide specificity for the N4 of Cyt U18. The middle nucleotide, Cyt U17, was expected to be contacted by the position 3 residue, D64. The D64 side chain atoms OD1 or OD2 are 3.2-3.5 Å from its N4 in both molecules A and B. This distance is slightly longer than that of an optimal hydrogen bond (2.6-3.3 Å). The arginine at position 6 of the recognition helix, R67, was predicted to make direct hydrogen bonds to the base pairing partners of the middle and 5' nucleotides, thereby contributing to the specificity for C at the middle and A at the 5' nucleotide[20]. The CAST analysis does show specificity for C at the middle position by Aart, however, shows G rather than A at the 5'[24]. The structure of Aart bound to DNA shows R67 making direct contacts to both the N7 and O6 of the 5' G, Gua U16, explaining the preference for G. R67 also contacts Gua L7, the base pairing partner of the middle nucleotide, however, the contact is water mediated.

Finger 3

Finger 3 contains the sequence QRA-N-LRA in positions -1 to 6 of the recognition helix, at residues 89-95 of Aart, and was expected to recognize the triplet 3'-AAA-5', as it was indeed found by CAST analysis to do. The crystal structure contains 3'-AAA-5' in the upper (U) strand and 5'-TTT-3' in the lower (L) strand at positions U13-U15/L9-L11, respectively. Position -1 of the recognition helix was predicted to contact the 3' A, Ade U15, as it does; Q89 contacts the N6 and N7 (2.9 Å, 3.1 Å in molecule A, and 3.0 Å, 3.0 Å in molecule B), explaining the specificity for A in the 3' position (Fig. 5c). The middle position was predicted to be specified as an adenine, and indeed was borne out by the CAST analysis. Aart residue N92, position 3 of the recognition helix, makes the predicted contacts to the N6 and N7 of the middle A, Ade U14 (3.1

Å, 3.0 Å in molecule A, 3.1 Å, 3.1 Å in molecule B). A91 (position 2 of the recognition helix) was predicted to make a van der Waals contact to the base pairing partner of the 3'A, Thy L9, and we find its methyl group is 4.1 Å (4.0 Å in molecule B) from the 5'methyl. Finally, previous studies found this domain to be able to specify 5'A; however, the mechanism of the specificity was not clear [20]. No direct contacts were predicted to occur to the 5' A, Ade U13, or its base-pairing partner, Thy L11. CAST data confirm a strong preference for A. The structure reveals two contacts to Thy L11: a water mediated contact to Q117 (position -1 of Finger 4) to the O4, and a hydrophobic contact between the methyl of A119 (position 2 of the adjacent zinc finger) to the 5'methyl (3.5 Å in molecule A, 3.7 Å in molecule B). The leucine residue at the 1 position of the adjacent finger, Finger 4, was suspected of making a contact to the thymine at L11, or even L10[16], however, we find that it is pointing away from the DNA and its gamma carbon is 7.4 Å and 7.8 Å (of molecules A and B, respectively) from the 5'methyl groups of these nucleotides. Additional electron density too large for individual water molecules was successfully modeled as glycerol. The glycerol molecule fits snugly between the protein and DNA interface (Fig. 4, 5a, 7a) with its hydrophobic backbone 3.8 Å from the methyl group of A95, and the three hydroxyl groups making hydrogen bonds with the O4 of Thy L10, N6 of Ade U14, N6 and N7 of Ade U13, and the amide oxygen of the side chain of N92. The positioning is identical in the two independent copies of the Aart/DNA complex. A similar interaction between a glycerol molecule and the Aart/DNA complex occurs at F4.

Finger 4

Finger 4 contains the sequence QLA-H-LRA in the -1 to 6 position of the recognition helix, residues 117-123 of Aart, and was designed to recognize the sequence 3'-AGA-5', however, CAST analysis shows that it prefers 3'-AG(A/G)-5'. The sequence 3'-AGG-5' in the upper (U) strand and 5'-TCC-3' in the lower (L) strand was used in the crystal structure at nucleotides U10-U12/L12-L14, respectively. Q117 at position -1 of the recognition helix contacts the N6 and N7 of the 3'A, Ade U12 (2.9 Å, 2.9 Å, molecule A, 2.9 Å, 2.9 Å in molecule B), as expected (Fig. 6a). The methyl group of A119, the position 2 residue, was predicted to make a van der Waals contact to the 5' methyl of Thy L12, the base-pairing partner of the 3' nucleotide of the 3'-AGG-5' triplet, however we find this distance long (4.7 Å in molecule a, 4.9 Å in molecule B, van der Waals distance is about 3.8-4.0Å). H120, position 3 of the recognition helix, was predicted to contact the middle G, and indeed contacts the O6 of Gua U11, (3.0 Å and 3.1 Å in molecules A and B, respectively) the middle nucleotide of 3'-AGG-5', explaining the CAST specificity for G. The epsilon nitrogen of H120 is also within hydrogen bonding distance from the N7 of Gua U11 (3.1 Å in both copies). The CAST analysis indicates a specificity of G/A for the 5' position; D147, position 2 of the neighboring finger, contacts the N4 of Cyt L14, (2.9 Å and 2.9 Å in molecules A and B respectively) as was predicted. A glycerol molecule was successfully modeled into the electron density at the protein-DNA interface near the middle and 5' base pairs (Fig. 4c, 6a, 7b). Its orientation and interactions at this interface are very similar to those of a glycerol molecule at F3, and is the same in both independent copies of the Aart/DNA complex. The hydrophobic backbone of the glycerol molecule is in van der Waals contact distance from the side chain of A123, and the hydroxyl groups hydrogen bond to the N4 of Cyt L13, O6 and N7 of Gua U10. Each hydroxyl of glycerol can accept two, but donate only one hydrogen bond, therefore, in contrast to the glycerol molecule found in Finger 3, the glycerol in Finger 4 hydrogen bonds to the O6 position of only one of the two neighboring purine bases at the middle and 5' position (Fig. 7b). In Finger 3, the glycerol hydroxyl can accept two hydrogen bonds and therefore contacts both N6 groups of the neighboring adenines (Fig. 7a).

Finger 5

Finger 5 contains the sequence RED-N-LHT in the -1 to 6 position of the recognition helix, residues 145-151 of Aart, and was designed to recognize the sequence 3'-GAT-5'. CAST analysis shows the specificity at these positions as 3'-GA(G/T)-5'. The sequence 3'-GAT-5' was used in the upper (U) strand and 5'CTA-3' in the lower (L) strand at U7-U9/L15-L17. The -1 position of the recognition helix, R145, contacts the O6 and N7 of the 3' nucleotide (2.8 Å, 2.9 Å and 2.7 Å, 2.9 Å in molecules A and B, respectively), Gua U9, as predicted (Fig. 6b). D147, position 2 of the recognition helix, contacts the R145 side chain to orient/stabilize it in the expected manner. Position 3 of the recognition helix, N148, hydrogen bonds to the N6 and N7 of Ade U8 (3.1 Å, 3.0 Å and 3.4 Å and 2.9 Å in molecules A and B, respectively). No direct contacts are visible to the 5' nucleotide, however, D175, at position 2 of the neighboring finger, is 3.0 Å and 3.2 Å from the N6 of Ade L17 in molecules A and B, respectively. We also find that the base of Thy U7 appears to be stacking onto the side chain of N148.

Finger 6

Finger 6 contains the sequence RRD-A-LNV in the -1 to 6 position of the recognition helix, residues 173-179 of Aart, and was designed to recognize the triplet 3'-GTA-5'. CAST analysis shows the expected specificity of 3'-GTA-5', and 3'-GTA-5' in the upper (U) strand and 5'-CAT-3' in the lower (L) strand, were used in the crystal structure at nucleotides U4-U6/L18-L20. The position -1 residue of the recognition helix, R173, contacts the O6 and N7 of Gua U6, the 3' nucleotide of the 3'-GTA-5' triplet in both molecules (2.7 Å, 3.0 Å, molecule A, 2.9 Å, 3.1 Å, molecule B). The side chain of A176, position 3 of the recognition helix, closely approaches the 5-methyl group of the middle base of this triplet, Thy U5 (4.1 Å and 4.8 Å, in molecules A and B, respectively, although this portion of molecule B exhibits poor electron density). In addition, an unusual DNA conformation is seen where the middle nucleotide, Thy U5 stacks onto the R173 side chain. The 5' nucleotide, Ade U4, stacks onto the Thy U5 and continues the helical stacking on the 5' side. No direct contacts are made to the 5' base, Ade U4 or its base pairing partner, Thy L20.

DNA conformation

The overall bend angles in the two duplexes, as calculated by the program CURVES, are 13° and 14°. Bend angles range from 0.4° to 5° throughout the duplex, at each base step, resulting in the relatively even distribution of bending. Bending by roll angle varied from -3° to 8° degrees, with the greatest bends occurring in F1, F2, and F6.

Helical parameters for the two Aart bound duplexes are given in Table S2 of the Supplementary Information. These were compared to those for DNA bound to other zinc finger structures found in the protein data bank. The values for slide between adjacent base pairs found in F2 were notable in that they were significantly more negative (-2.25) than even the most negative found in the other zinc finger bound DNA molecules (-1.5). The inclination of the base planes at these base pairs is also strikingly negative (-6.4) as compared to that of the rest of the Aart DNA and the other zinc finger bound DNA.

The standard values of inclination are 2.4° for B form and 12° for A. Inclination is the angle that the base pair plane makes with the plane normal to the DNA helical axis, rotated about the base pair plane short axis. The base pair planes are relatively perpendicular to the helical axis in B form, but very tilted in A. The average inclination values are 3.1° and 3.4° the two Aart bound duplexes, thus exhibiting overall B form character. However, the values vary between -5.7° and +8° at each base pair. As far as individual base pair detail, the most negative values occur in F2. The overall helical twist values for the duplexes bound to molecules A and B are

32.8° and 32.5°, respectively. The standard values for A and B forms are 32.7° and 36.1°, respectively. Therefore, in terms of helical twist, the DNA bound to Aart looks more like A, than B form DNA. This intermediate, but unique, conformation of zinc finger bound DNA has been previously described[36]. Over 18 base pairs, the DNA is under-wound relative to B form DNA a total of 64° to 70°. Modeling has shown that the unwinding of the duplex is necessary to fit the adjacent recognition helices of multi-finger zinc finger proteins into the major groove[15].

Discussion

Expected vs. observed contacts leading to recognition

Figure 4a-c illustrates expected contacts (Fig. 4a), measured Aart specificity (CAST analysis, Fig. 4b), and contacts observed in the crystal structure of Aart bound to DNA (Fig. 4c). The expected contacts of Fig. 4a are derived from the phage display results and modeling[19,24] , while figure 4b illustrates the experimentally determined specificity of Aart[24]. Figure 4c shows the hydrogen bonds (red or black lines), van der Waals (red or black lines with VDW, also glycerol (gol) to alanine black line), and water mediated contacts (wat) found between Aart and the DNA. The colored boxes (green, yellow, blue and red) represent different types of protein-DNA contacts. A green box indicates that the expected specificity was found in the CAST assay (Fig. 4b) and also the expected contacts are observed in the crystal structure. Yellow indicates weak specificity found in the CAST assay, and the structure provides an explanation for this. Blue indicates either unexpected specificity found by the CAST assay, or unexpected contacts found in the structure, where contacts exist which explain the observed specificity. Finally, red indicates specificity determined by the CAST assay however no contacts are observed between Aart and the DNA which can explain the observed specificity. Some positions are not identified with colored boxes despite the existence of contacts between Aart and the DNA; in these cases the contributions of these contacts to the specificity is unclear.

Contacts that give rise to specificity generally involve the donation or acceptance of hydrogen bonds to the base edges in such a way that could only occur with one of the four base types[37,38]. Hydrophobic contacts to the 5-methyl group of thymine are also commonly observed in sequence specific DNA binding proteins. Contacts between Aart and the DNA that were predicted, and which explain the observed specificity, occur at nine nucleotides (green, Fig. 4c). Three other contacts between Aart and the DNA that may also give rise to expected specificity, and were predicted, were found with distances between the side chain and base atoms longer than expected (Fig. 4c, green boxes, red lines, A119 and Thy L12, D64 and Cyt U17, A176 and Thy U5). The fact that the distances between potentially interacting groups is found to be longer than expected could be due to (in some cases) coordinate error (0.2 Å presently). Another possibility is that the crystallization process has frozen out one of several conformers of the protein-DNA complex; however in solution, the groups might actually be within the optimal contacting distance some fraction of the time. In the case of hydrophobic groups, the van der Waals contribution to interaction energy falls off as $1/R^6$, where the contribution, although diminished, could still be realized even though the distance is longer than optimal. In addition, the burial of these groups from water is expected to be favorable. Therefore, the energetic consequences to the affinity and specificity of the Aart/DNA complex in the cases of 'long contacts' (red lines in Fig. 4c) are expected to be favorable but are less conclusive than those with optimal contact distances. All twelve positions of the expected contacts with expected specificity (green boxes, Fig. 4c) involve the 3' or middle nucleotide of the triplet on the upper strand, or the lower strand 3' or 5' triplet position. None occur in Finger 1, and none occur to the 5' position on the upper strand.

There are two positions where the expected specificity is not observed, or is weak, and where the structure of Aart bound to DNA provides some explanation (yellow, Fig. 4c). The CAST data shows that the expected specificity at the 3' and middle base pairs of F1 are not observed; rather, very little specificity occurs at these positions (Fig. 4b). The structure shows the expected contacts between R33 and the 3' nucleotide, as well as a contact from H36 to the middle G. This latter contact is to the N7 of the G, and therefore would specify either G or A. These contacts would predict the specificity of Aart for 3'-G(G/A)-5' at the 3' and middle of F1. One reason for the observed lack of specificity may be related to the observed poor order of the Finger 1 residues. As described above, the temperature factors are largest at these residues, and the RMSD between structures also highest. In addition, the side chain electron density for R33 is almost non-existent in molecule A. Poor ordering of atoms within a crystal structure is due to positional heterogeneity. Therefore, the observed lack of specificity at the 3' and middle positions of the F1 triplet could be due to poor overall binding of this region to the DNA, and consequently multiple conformations of the protein at the protein-DNA interface, and poor ordering in the crystal structure. Indeed, CAST data obtained from a number of proteins suggest that the F1 position of zinc finger proteins have more liabilities with respect to specificity than other fingers[24].

A third class of contacts are colored blue in Fig. 4c, and these include positions where either the expected specificity is indicated by the CAST assay and unexpected contacts are found in the crystal structure explaining the observed specificity, or the specificity was unexpected and contacts are found explaining the observed specificity. Such contacts occur at the 3' of F2, and the 5' positions of F2, F3, and F4. Specificity for C at the 3' of F2 was expected from the design, and although not as specific as other sites, the CAST results do show specificity for C over other nucleotides. Modeling suggested that the side chain of D61 (position -1 of the recognition helix) might interact with the cytosine base edge, however, the structure shows that this side chain is too distant for direct hydrogen bonding. Instead, a water molecule appears to mediate the interaction between D61 and Cyt U18. The same water is also hydrogen bonding to the side chain of D64 (position 3 of the recognition helix). If both carboxylates of D61 and D64 are deprotonated, they will only be able to accept hydrogen bonds, and the water molecule would be oriented such that it could only accept hydrogen bonds from the base edge at the 3' position. Only cytosine and adenine have hydrogen bond donors in the major groove. If the molecule were positioned to interact optimally with cytosine, then the observed specificity for C would be explained by this interaction. A C at the 3' position of a triplet is found in four other cases, two of which appear to be recognized by an aspartic acid at the -1 position of the recognition helix (DSNR-F1, PDB code 1A1F[39], and GLI-F5, PDB code 2GLI). In two other structures (GLI-F4, PDB code 2GLI, RADR-F1, PDB code 1A1K[39]), no contacts to the zinc finger protein are visible.

A second contact in this class occurs at the 5' of F2. Aart was predicted to specify an A, although CAST indicated a preference for G. The structure shows R67, position 6 of Finger 2, hydrogen bonding to the N7 and O6 of Gua U16, at the 5' position of F2. Although this observed interaction could be considered ‘expected’ for an arginine in position 6 based on previous structural observations, previous studies showed this domain to have very good specificity for 5’A[20]. It should be noted, however, that the binding affinity to the designed site is similar within error to binding to the consensus site.

Another unexpected contact which explains Aart specificity occurs at the 5' position of F3. Previous studies determined this domain had moderate specificity for 5’A, although the mechanism was not clear[20]. The methyl group of the alanine in position 6 was anticipated to be too far from the 5’ base for interaction, and no other interactions could be rationalized. The crystal structure shows an unexpected contact of the 5-methyl group of Thy L11 to the methyl group of A119, a residue from the recognition helix of the neighboring finger. A water mediated interaction occurs between Q117 and Thy L11, although the contribution to specificity is unclear. In addition, a glycerol molecule was found hydrogen bonding to the 5'A of F3, and may influence specificity (see below).

L118 (position 1 of Finger 4) did not appear to have any interaction with Thy L11 or any other base in Aart, in contrast to the structure of protein TATA_{ZF}[16]. Aart and TATA_{ZF} are the only designed zinc finger proteins for which CAST analysis showed robust recognition of 5’A and structural information is available. Despite the similarities of their recognition sequence in this region, L75 in $\mathrm{TATA}_{\mathrm{ZF}}$ is oriented towards the bases while L118 in Aart adopts a more traditional orientation away from the bases. Such examples illustrate our lack of predictive ability even in well-defined cases that are structurally similar. The structural features underlying these differences deserve greater study.

CAST indicates the specificity for G or A at the 5' upper strand nucleotide of F4 (Gua U10). D147, position 2 of the neighboring finger, is within hydrogen bonding distance of the N4 of the base pairing partner of the 5' nucleotide of F4. We find a similar contact to the N6 of Ade L17 (base pairing partner of 5' nucleotide of F5, Thy U7) from D175, position 2 of Finger 6. The influence of aspartate in position 2 on the neighboring base, referred to as target site overlap, has been well studied[40]. In the structure of protein Zif268[15], aspartates in position 2 of Fingers 2 and 3 contact the extracyclic amines of A or C on the lower strand of the neighboring site, analogous to the arrangement of D175 and Ade L17 in Aart. Consequently, T or G is specified as the 5’ base on the upper stand of the neighboring site[41]. CAST data for Aart show T and G are the most frequently specified 5’ bases in the F5 subsite (Fig. 4b), consistent with this explanation. However, in addition to the target site overlap at the 5' of F4, the upper strand base is also contacted by glycerol, which may affect specificity (see below).

The last group of nucleotide positions, colored red in Fig. 4c, are those in which the observed specificity is difficult to explain from the structural data. Two positions are found in this category: the 5' of F1 and the 5' of F6, Ade U4. Aart was predicted to specify an A at the 5' position of F1, however, CAST indicated a preference for C. In principle, the preference for C can be rationalized by a direct interaction with E39 at position 6. However, the E39 carboxylate oxygens are 5.1 Å from the N4 of Cyt U19 and thus beyond hydrogen bonding distance. In fact, no contacts are visible to the 5' nucleotide of the upper strand, however in molecule A the side chain of K63 (position 2 of Finger 2) is within hydrogen bonding distance to the N7 of Gua L5, its base pairing partner. In molecule B, the last two atoms of the K63 side chain are disordered. A contact to the N7 could provide specificity for a purine at this position, although the flexibility of a lysine side chain suggests that it could accommodate also the O4 of a thymine, and if uncharged, the N4 of cytosine. Thus the expected specificity of the F1 5' position would be any base C, T, A or G. The origin of the apparent preference for C is not clear. Within other structures, two examples of C in the 5' position exist. In one (GLI-F4, PDB code 2GLI[42]) it is contacted by an aspartic acid at position 3 of a recognition helix, although the docking orientation of this helix is non-canonical. In the other case (YY1-F4, PDB code 1UBD[43]), no contacts to the protein are visible.

The other position of irreconcilable specificity occurs in the 5' position of F6. CAST data indicate a preference for A at this position. However, no contacts are seen between Aart and Ade U4 or its base pairing partner, Thy L20. The base step between the middle and 5' base pairs possesses a very low twist (29°). Within the structural database only two occurrences of adenine at the 5' position exist ($TATA_{ZF}$ F2 in PDB codes 1G2F and 1G2D[16]) which are the same triplets having T at the middle position. The step between the middle and 5' positions in these structures also possess low twist (29.6° and 28.5°).

At the 5' of F5, we also found that Thy U7 stacks on top of N148. In Zif268, a similar interaction occurs with the 5’ thymine stacking onto the histidine at position 3. This was recognized by the authors[14,39] as possibly contributing to the preference for thymine[44]. The Thy U7-N148 stacking interaction in Aart may therefore additionally contribute to the observed specificity (Fig. 6b). Similarly, the thymine at the middle position of F6 appears to stack onto the side chain of R173 (Fig. 6c). Such a contact was predicted[45] and may contribute to the specificity at this position. Only two other structures provide examples of a thymine in the middle triplet position ($TATA_{ZF}$ F2 in PDB codes 1G2F, and 1G2D[16]). In both cases a glutamine is found at the -1 position of the recognition helix. Inspection of the structures also shows considerable stacking of the thymine onto the -1 position side chain. Again, further investigation of the Aart interactions is required. An intriguing possibility is that indirect readout is occurring at these positions, but this will require further investigation.

The role of glycerol

A surprising finding was the presence of molecules of glycerol at the protein-DNA interface at the 5' positions of F3 and F4. The molecule is oriented in exactly the same way at both positions (and in the other molecule of the asymmetric unit). It makes hydrogen bonds to the 6 (O in Gua and N in Ade) and N7 positions of the purine bases Ade U13 and Gua U10, as well as to the 4 position of the base pairing partner (Fig. 7). In the case of the glycerol molecules at Ade U13, hydrogen bonds are also made to the middle base of the triplet, Ade U14, and the amino acid side chain contacting that base, N92. In all cases, the carbon backbone of the glycerol molecules make van der Waals contacts to the methyl group of an alanine at position 6 of the respective recognition helices, namely A95 and A123. Glycerol (30% final) is used in preparing the crystals for flash freezing, just before diffraction data collection. Although glycerol was not added to the phage display experiments, polyethylene glycol, or possibly the dried milk used in the assay may conceivably contain glycerol or a similar small molecule. Of course, it is readily possible that glycerol may displace water molecules normally present at the parts of the protein-DNA interface, and its presence in the structure is merely an artifact of the crystal preparation. In addition, we have no data, at this time, indicating whether or not the observed glycerol at the protein-DNA interface in the crystal structure is present in the binding studies or if it contributes to recognition. However, the presence of a short hydrophobic side chain at the 6 position of the recognition helix leaves a cavity at the protein-DNA interface and appears perfectly suited to stabilize a molecule such as glycerol. Given the two different sequences contacted by glycerol in F3 and F4, one where the preference is A,A at the middle and 5' positions and another where it is G,A/G, it may be possible that the glycerol is involved in specifying purine bases at these positions. In addition, the target site overlap of an aspartic acid from position 2 of one finger to the base pairing partner of the 5' position of a preceding triplet is believed to specify either G or T in the upper strand position. Such appears to be the case in F5, however, the 5' of F4, where glycerol binds, shows specificity for either G or A, and has the same target site overlap. The difference may derive from the interactions with glycerol in this position. The interactions of a small molecule like glycerol at the protein-DNA interface can serve as a model for drug mediated DNA recognition or zinc fingers to be used as sensors of small molecules.

Rethinking recognition of the 5’ base.

It comes as little surprise that the existing dataset is most sparse concerning interactions the 5' position, and that interactions at this position are atypical. The 5' base is difficult to contact by side chains from the recognition helix, since the recognition helix is more distant from the DNA at this end of the triplet. This was shown most convincingly by the use of the artificial amino acid, citrulline[46]. Citrulline contains the same hydrogen bond donors and acceptors as glutamine, but has the same side chain length as arginine. The long side chain of arginine at position 6 is long enough to contact the DNA, however recognizes exclusively G at the 5' position. Citrulline in position 6 was able to specify 5’A, while glutamine could not. (An exception to this is found in F2 of TATA$_{ZF}$* (PDB code 1G2D), in which Q(6) is able to hydrogen bond to 5’A[12]. The contacts made by this finger also are distinctive in that they allow a significant number of base contacts with the "lower" strand of the DNA. The structural features that enable the contact to 5'A in this case deserve further study.) A large part of the motivation for creating and solving the structure of the Aart protein originated in its ability to recognize bases other than G at the 5' triplet position by Fingers 1, 3, 5, and 6.

Our results suggest that 5’A, C, and T are typically not specified by direct interaction with the position 6 residue. Many attempts to derive recognition codes frequently assume a model in which the 5’ base is specified by the position-6 residue[47-50]. Our results challenge such a canonical model of zinc finger-DNA recognition, and suggest the specification of the 5’ base by position 6 might be more often the exception (primarily in the case of 5’G) rather than the rule. The results strongly support the earlier efforts of several groups to incorporate some aspect of inter-domain or context-dependant interactions into their design strategy[16,40,51,52]. These findings also highlight the need for more structural studies containing examples of 5’A, C, and T recognition. In the Aart structure, 5' specificity is apparently influenced most frequently by a target site overlap interaction from the amino acid in position 2 of the neighboring finger (occurring in F1, F3, F4, and F5). Unexpected stacking interactions within a domain also were postulated to influence specificity, such as at the 5'T of F5 stacking on the N(3) (resulting in G/T specificity), the middle T of F6 stacking onto R(-1), and perhaps the 5'A of F6. Surprisingly, glycerol appears to mediate much of the interaction between the 6 position of the recognition helix and the middle and 5' base pairs of F3 and F4. The 6 position is an alanine, a small, hydrophobic side chain ideal for accommodating a small molecule such as glycerol. The role of glycerol in specificity is not known at this time but may be to interact more favorably with purine bases at the middle and 5' position of the upper strand. Future work will be needed to clarify its importance to sequence specificity by Aart.

Evidence for strain induced by the use of canonical linkers.

Past studies have suggested that proteins containing multiple fingers may suffer from strain imposed by suboptimal linker lengths between fingers, leading to decreased affinity and/or specificity[27-29]. Such strain could be evidenced in the Aart/DNA complex by distorted DNA, mispositioned fingers, or poor contacts between fingers and DNA. Inspection of the Aart/DNA structure shows no large bends in the DNA, but rather a continuum of smaller bends at each base step. Since the six fingers of Aart wrap around the DNA completely, perhaps DNA bending would not be expected to relieve strain. Some unusual helical parameters, relative to B form DNA and to DNA bound to other zinc fingers, occurs in the form of slide in F2. This may be evidence of strain, but the origin and consequences of the unusual slide are unknown. The DNA is underwound by 64°-70°, which may lead to strain within the DNA molecule. If the DNA were closed circular, or had fixed ends, the underwinding would incur a cost or possibly be favorable, depending on the state of supercoiling. In linear, unrestrained DNA, the unwinding could result in unfavorable base stacking or phosphate-phosphate repulsion. The unwinding is a consequence of widening the major groove for the close approach of multiple alpha helices to the DNA[36]. Modeling experiments suggest that a longer linker might diminish the unwinding[15]. However, it is not clear if the energy gained by relaxing the DNA closer to B form would compensate for the loss of domain packing[53], interdomain hydrogen bonds[15], and target site overlap interactions important for specificity and affinity. This conclusion is consistent with the report that tandem bound Zif268 proteins could accommodate a canonical TGEKP linker in the middle of the complex without obvious structural distortion[54], and a more recent report that longer linkers in multi-finger proteins produced no obvious benefit[55]. With respect to finger positioning relative to the DNA, we found no evidence of gross misplacement of the zinc fingers (Fig. 3d). Most of the contacts identified have reasonable interaction distances, also inconsistent with mispositioning of the finger helices.

The evidence of strain that has been found includes the systematic trend of slight rotation and translation of recognition helices relative to the orientation Finger 3 takes with its bound triplet of DNA, where these values increase in fingers more distant from Finger 3. In addition, higher B factors are found at the first and last fingers, as well as their bound DNA. High B factors indicate either atomic motion within, or positional displacement among, the different molecules that make up the crystal. Since the diffraction data were collected at 100K, atomic motion is not expected to be significant, therefore, fingers at either the end of the molecule occupy somewhat different positions within the many copies in the crystal. The ends of the Aart bound DNA contact the DNA of neighboring complexes, forming a pseudo-continuous helix throughout the crystal. Such contacts are likely to be stabilizing to the crystal, however, we find that the disorder within the Aart/DNA complex is greatest at these locations. Finger 1 of molecule A exhibits so much disorder that the electron density on these atoms is patchy and absent for many residues. Also, the linker between Fingers 5 and 6, as well as much of Finger 6 in one of the two copies of

Aart bound to DNA, is found to be so disordered that it cannot be modeled. This disorder may be related to the lack of specificity observed by Aart at particular bases, in it is possible that direct contacts are made in only a fraction of the complexes at any one time leading to multiple orientations and consequently conformational disorder and poor electron density in the crystal structure.

Acknowledgements

CFB was supported by The Skaggs Institute for Chemical Biology and GM065059. JWC was supported in part by NSF IGERT 9870659. Portions of this research were carried out at the Advanced Photon Source, a national user facility operated by Argonne National Laboratory, as well as at Stanford Synchrotron Radiation Laboratory, a national user facility operated by Stanford University on behalf of the U.S. Department of Energy, Office of Basic Energy Sciences. The SSRL Structural Molecular Biology Program is supported by the Department of Energy, Office of Biological and Environmental Research, and by the National Institutes of Health, National Center for Research Resources, Biomedical Technology Program, and the National Institute of General Medical Sciences.

Figure legends

Figure 1. Protein and DNA sequences used in this study. **a.** Aart sequence. Recognition helix residues of each zinc finger shown in bold. **b.** Sequence of DNA duplex cocrystallized with Aart. X indicates 5-bromouracil in the brominated DNA, and thymine in the native DNA. The structure of Aart with nonbrominated DNA was used in all structural analyses.

Figure 2. The Aart/DNA complex. **a.** Ribbon diagrams of one complex. Blue ribbon: trace through backbone of Aart protein. Pink ball: zinc ions. Green sticks: DNA. **b.** Stereo view of simulated annealing 2Fo-Fc (grey, 1σ), and Fo-Fc (pink, 2.5σ) omit electron density map at Ade U13/Thy L11 with selected side chains, glycerol and a water molecule. **c.** RMSD at each alpha carbon following superposition of chains A and B of Aart (blue) and average Debye-Waller Temperature Factor (red solid, chain A, red dashed, chain B) of each residue. **d.** RMSD following superposition of the duplex DNA using all atoms (blue solid, chains C and E, blue dashed, chains D and F), and average Debye-Waller Temperature Factor (red solid, chain C, red dashed, chain D, red loose dotted, chain E, red tight dotted, chain F) of each nucleotide.

Figure 3. Orientation of the recognition helix with respect to the DNA. **a.** Angle required to superimpose recognition helices onto helix of Finger 3 of molecule A after superposition of DNA. **b.** Distance between centers of mass (COM) of recognition helices following DNA superposition. **c.** Plot of translation vs. rotation of C2H2 zinc fingers from the PDB (red circles) and Aart fingers (blue diamonds) superimposed onto Aart F3, as in Fig. 3a-b.

Figure 4. Protein-DNA contacts. **a.** Expected contacts between recognition helices of Aart and DNA. **b.** Aart CAST analysis results[24]. Boxed nucleotides differ from designed target sequence. **c.** Schematic of observed interactions between Aart and DNA bases. Green boxes, contacts predicted and which explain observed specificity. Blue boxes, contacts or specificities which were unexpected. Yellow boxes, weak specificity. Red boxes, specificity cannot be explained. Red lines, weak (longer than optimal) interactions. Black lines, hydrogen-bonds or Van der Waal interactions to glycerol. VDW, Van der Waal hydrophobic interaction. wat, water-mediated interaction. stk, stacking interaction, gol, glycerol molecule.

Figure 5. Stereo figures of Aart/DNA interactions. Colored by atom type (green=carbon, blue=nitrogen, red=oxygen). Spheres show the position of methyl groups. Water molecules are shown as small red spheres. Hydrogen bonds between the protein and DNA are shown as dotted lines. **a.** Finger 1 of molecule A. **b.** Finger 2 of molecule A. **c.** Finger 3 of molecule A. Gol1, glycerol.

Figure 6. Stereo figures of Aart/DNA interactions. Colored by atom type (green=carbon, blue=nitrogen, red=oxygen). Spheres show the position of methyl groups. Water molecules are shown as small red spheres. Hydrogen bonds between the protein and DNA are shown as dotted lines. **a.** Finger 4 of molecule A. Gol2, glycerol. **b.** Finger 5 of molecule B. **c.** Finger 6 of molecule A.

Figure 7. Contacts between glycerol and the protein-DNA interface at **a.** F3, and **b.** F4.

References

1. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., et al. (2001). The sequence of the human genome. *Science* 291, 1304-51.
2. Blancafort, P., Segal, D. J. & Barbas, C. F., 3rd. (2004). Designing transcription factor architectures for drug discovery. *Mol Pharmacol* 66, 1361-71.
3. Wolfe, S. A., Nekludova, L. & Pabo, C. O. (2000). DNA recognition by Cys2His2 zinc finger proteins. *Annu Rev Biophys Biomol Struct* 29, 183-212.
4. Carroll, D. (2004). Using nucleases to stimulate homologous recombination. *Methods Mol Biol* 262, 195-207.
5. Beerli, R. R. & Barbas, C. F., 3rd. (2002). Engineering polydactyl zinc-finger transcription factors. *Nat Biotechnol* 20, 135-41.
6. Beumer, K., Bhattacharyya, G., Bibikova, M., Trautman, J. K. & Carroll, D. (2006). Efficient Gene Targeting in Drosophila with Zinc Finger Nucleases. *Genetics* 172, 2391-2403.
7. Alwin, S., Gere, M. B., Guhl, E., Effertz, K., Barbas, C. F., 3rd, Segal, D. J., Weitzman, M. D. & Cathomen, T. (2005). Custom zinc-finger nucleases for use in human cells. *Mol Ther* 12, 610-7.
8. Jamieson, A. C., Miller, J. C. & Pabo, C. O. (2003). Drug discovery with engineered zinc-finger proteins. *Nat Rev Drug Discov* 2, 361-8.
9. Urnov, F. D., Miller, J. C., Lee, Y. L., Beausejour, C. M., Rock, J. M., Augustus, S., Jamieson, A. C., Porteus, M. H., Gregory, P. D. & Holmes, M. C. (2005). Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* 435, 646-51.
10. Lee, M. S., Gippert, G. P., Soman, K. V., Case, D. A. & Wright, P. E. (1989). Three-dimensional solution structure of a single zinc finger DNA- binding domain. *Science* 245, 635-7.
11. Miller, J., McLachlan, A. D. & Klug, A. (1985). Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. *Embo J* 4, 1609-14.
12. Rhodes, D. & Klug, A. (1993). Zinc Fingers: They play a key part in regulating the activity of genes in many species, from yeast to humans. Fewer than 10 years ago no one knew they existed. *Scientific American*, 56-65.
13. Pabo, C. O. & Nekludova, L. (2000). Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *J Mol Biol* 301, 597-624.

14. Pavletich, N. P. & Pabo, C. O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å. *Science* 252, 809-817.

15. Elrod-Erickson, M., Rould, M. A., Nekludova, L. & Pabo, C. O. (1996). Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. *Structure* 4, 1171-80.

16. Wolfe, S. A., Grant, R. A., Elrod-Erickson, M. & Pabo, C. O. (2001). Beyond the "recognition code": structures of two Cys2His2 zinc finger/TATA box complexes. *Structure* 9, 717-723.

17. Pabo, C. O., Peisach, E. & Grant, R. A. (2001). Design and Selection of Novel Cys2his2 Zinc Finger Proteins. *Annu Rev Biochem* 70, 313-340.

18. Dreier, B., Fuller, R. P., Segal, D. J., Lund, C. V., Blancafort, P., Huber, A., Koksch, B. & Barbas, C. F., 3rd. (2005). Development of zinc finger domains for recognition of the 5'-CNN-3' family DNA sequences and their use in the construction of artificial transcription factors. *J Biol Chem* 280, 35588-97.

19. Dreier, B., Segal, D. J. & Barbas III, C. F. (2000). Insights into the molecular recognition of the 5'-GNN-3' family of DNA sequences by zinc finger domains. *J Mol Biol* 303, 489-502.

20. Dreier, B., Beerli, R. R., Segal, D. J., Flippin, J. D. & Barbas III, C. F. (2001). Development of zinc finger domains for recognition of the 5'-ANN-3' family of DNA sequences and their use in the construction of artificial transcription factors. *J Biol Chem* 276, 29466-78.

21. Segal, D. J., Dreier, B., Beerli, R. R. & Barbas III, C. F. (1999). Toward controlling gene expression at will: selection and design of zinc finger domains recognizing each of the 5'-GNN-3' DNA target sequences. *Proc Natl Acad Sci U S A* 96, 2758-2763.

22. Liu, Q., Segal, D. J., Ghiara, J. B. & Barbas III, C. F. (1997). Design of polydactyl zinc-finger proteins for unique addressing within complex genomes. *Proc Natl Acad Sci U S A* 94, 5525-5530.

23. Beerli, R. R., Segal, D. J., Dreier, B. & Barbas III, C. F. (1998). Toward controlling gene expression at will: specific regulation of the erbB-2/HER-2 promoter by using polydactyl zinc finger proteins constructed from modular building blocks. *Proc Natl Acad Sci U S A* 95, 14628-14633.

24. Segal, D. J., Beerli, R. R., Blancafort, P., Dreier, B., Effertz, K., Huber, A., Koksch, B., Lund, C. V., Magnenat, L., Valente, D. & Barbas, C. F., 3rd. (2003). Evaluation of a modular strategy for the construction of novel polydactyl zinc finger DNA-binding proteins. *Biochemistry* 42, 2137-48.

25. Beerli, R. R., Dreier, B. & Barbas III, C. F. (2000). Positive and negative regulation of endogenous genes by designed transcription factors. *Proc Natl Acad Sci U S A* 97, 1495-1500.

26. Segal, D. J., Goncalves, J., Eberhardy, S., Swan, C. H., Torbett, B. E., Li, X. & Barbas, C. F., 3rd. (2004). Attenuation of HIV-1 replication in primary human cells with a designed zinc finger transcription factor. *J Biol Chem* 279, 14509-19.

27. Moore, M., Choo, Y. & Klug, A. (2001). Design of polyzinc finger peptides with structured linkers. *Proc Natl Acad Sci U S A* 98, 1432-6.

28. Moore, M., Klug, A. & Choo, Y. (2001). Improved DNA binding specificity from polyzinc finger peptides by using strings of two-finger units. *Proc Natl Acad Sci U S A* 98, 1437-41.

29. Kim, J. S. & Pabo, C. O. (1998). Getting a handhold on DNA: design of poly-zinc finger proteins with femtomolar dissociation constants. *Proc Natl Acad Sci U S A* 95, 2812-7.

30. Crotty, J. W., Etzkorn, C., Barbas, C.F., Segal, D.J., Horton, N.C. (2005). Crystallization and preliminary X-ray crystallographic analysis of Aart, a designed six-finger zinc-finger peptide, bound to DNA. *Acta Cryst*. F61, 573-576.
31. Otwinowski, Z. a. M., W. (1997). Processing of X-ray Diffraction Data Collected in Oscillation Mode. In *Methods in Enzymology* (Carter, C. W., Jr. & Sweet, R.M., ed.), Vol. 276, pp. 307-326. Academic Press, New York.
32. Terwilliger, T. C. & Berendzen, J. (1999). Automated MAD and MIR structure solution. *Acta Crystallogr D Biol Crystallogr* 55 (Pt 4), 849-61.
33. Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallographica. Section D: Biological Crystallography* 54, 905-21.
34. Ravishanker, G., Swaminathan, S., Beveridge, D. L., Lavery, R. & Sklenar, H. (1989). Conformational and helicoidal analysis of 30 PS of molecular dynamics on the d(CGCGAATTCGCG) double helix: "curves", dials and windows. *J Biomol Struct Dyn* 6, 669-99.
35. Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Cryst.*, 992-923.
36. Nekludova, L. & Pabo, C. O. (1994). Distinctive DNA conformation with enlarged major groove is found in Zn-finger-DNA and other protein-DNA complexes. *Proc Natl Acad Sci U S A* 91, 6948-52.
37. Seeman, N. C., Rosenberg, J. M. & Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci U S A* 73, 804-808.
38. Luscombe, N. M., Laskowski, R. A. & Thornton, J. M. (2001). Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res* 29, 2860-74.
39. Elrod-Erickson, M., Benson, T. E. & Pabo, C. O. (1998). High-resolution structures of variant Zif268-DNA complexes: implications for understanding zinc finger-DNA recognition. *Structure* 6, 451-464.
40. Isalan, M., Choo, Y. & Klug, A. (1997). Synergy between adjacent zinc fingers in sequence-specific DNA recognition. *Proc Natl Acad Sci U S A* 94, 5617-21.
41. Isalan, M., Klug, A. & Choo, Y. (1998). Comprehensive DNA recognition through concerted interactions from adjacent zinc fingers. *Biochemistry* 37, 12026-33.
42. Pavletich, N. P. & Pabo, C. O. (1993). Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. *Science (Washington, D. C., 1883-)* 261, 1701-7.
43. Houbaviy, H. B., Usheva, A., Shenk, T. & Burley, S. K. (1996). Cocrystal structure of YY1 bound to the adeno-associated virus P5 initiator. *Proc Natl Acad Sci U S A* 93, 13577-82.
44. Swirnoff, A. H. & Milbrandt, J. (1995). DNA-binding specificity of NGFI-A and related zinc finger transcription factors. *Mol. Cell. Biol.* 15, 2275-87.
45. Lamoureux, J. S., Maynes, J. T. & Glover, J. N. (2004). Recognition of 5'-YpG-3' sequences by coupled stacking/hydrogen bonding interactions with amino acid residues. *J Mol Biol* 335, 399-408.
46. Jantz, D. & Berg, J. M. (2003). Expanding the DNA-recognition repertoire for zinc finger proteins beyond 20 amino acids. *J Am Chem Soc* 125, 4960-1.

47. Kaplan, T., Friedman, N. & Margalit, H. (2005). Ab initio prediction of transcription factor targets using structural knowledge. *PLoS Comput Biol* 1, e1.

48. Sera, T. & Uranga, C. (2002). Rational design of artificial zinc-finger proteins using a nondegenerate recognition code table. *Biochemistry* 41, 7074-81.

49. Wolfe, S. A., Greisman, H. A., Ramm, E. I. & Pabo, C. O. (1999). Analysis of zinc fingers optimized via phage display: evaluating the utility of a recognition code. *J Mol Biol* 285, 1917-1934.

50. Choo, Y. & Klug, A. (1994). Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc Natl Acad Sci U S A* 91, 11168-72.

51. Jamieson, A. C., Wang, H. & Kim, S. H. (1996). A zinc finger directory for high-affinity DNA recognition. *Proc Natl Acad Sci U S A* 93, 12834-9.

52. Greisman, H. A. & Pabo, C. O. (1997). A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science* 275, 657-61.

53. Foster, M. P., Wuttke, D. S., Radhakrishnan, I., Case, D. A., Gottesfeld, J. M. & Wright, P. E. (1997). Domain packing and dynamics in the DNA complex of the N-terminal zinc fingers of TFIIIA. *Nat Struct Biol* 4, 605-8.

54. Peisach, E. & Pabo, C. O. (2003). Constraints for zinc finger linker design as inferred from X-ray crystal structure of tandem Zif268-DNA complexes. *J Mol Biol* 330, 1-7.

55. Neuteboom, L. W., Lindhout, B. I., Saman, I. L., Hooykaas, P. J. & van der Zaal, B. J. (2006). Effects of different zinc finger transcription factors on genomic targets. *Biochem Biophys Res Commun* 339, 263-70.

a.

```
1   |   10    |    20     |  30      |  40
ISEFGSSSSVAQAALE PGEKP YACPECGKSFSR SDHLAEHQRTH
                 |    50    |  60      |  70
                 TGEKP YKCPECGKSFSD KKDLTRHQRTH
                   |    80    |   90     | 100
                 TGEKP YKCPECGKSFSQ RANLRAHQRTH
                     |   110    |  120     |
                 TGEKP YACPECGKSFSQ LAHLRAHQRTH
                130     |  140    |  150     |
                 TGEKP YKCPECGKSFSR EDNLHTHQRTH
                  160     |  170     | 180     |  190
                 TGEKP YKCPECGKSFSR RDALNVHQRTH TGKKTS
                                 -1 123456
                ______ ____________ ___________
                linker   β-sheet     α-helix
```

b.

```
          22 20      15   10      5    1
          |  |      |     |      |    |
Upper 3'  G GGC CCG AAA AGG GAT GTA GAC 5'
Lower 5' GC CCG GGC XTT TCC CTA CAX CT 3'
         |    |      |      |     |  |
         1    5     10     15    20 22
```

Figure 1. Segal et al.

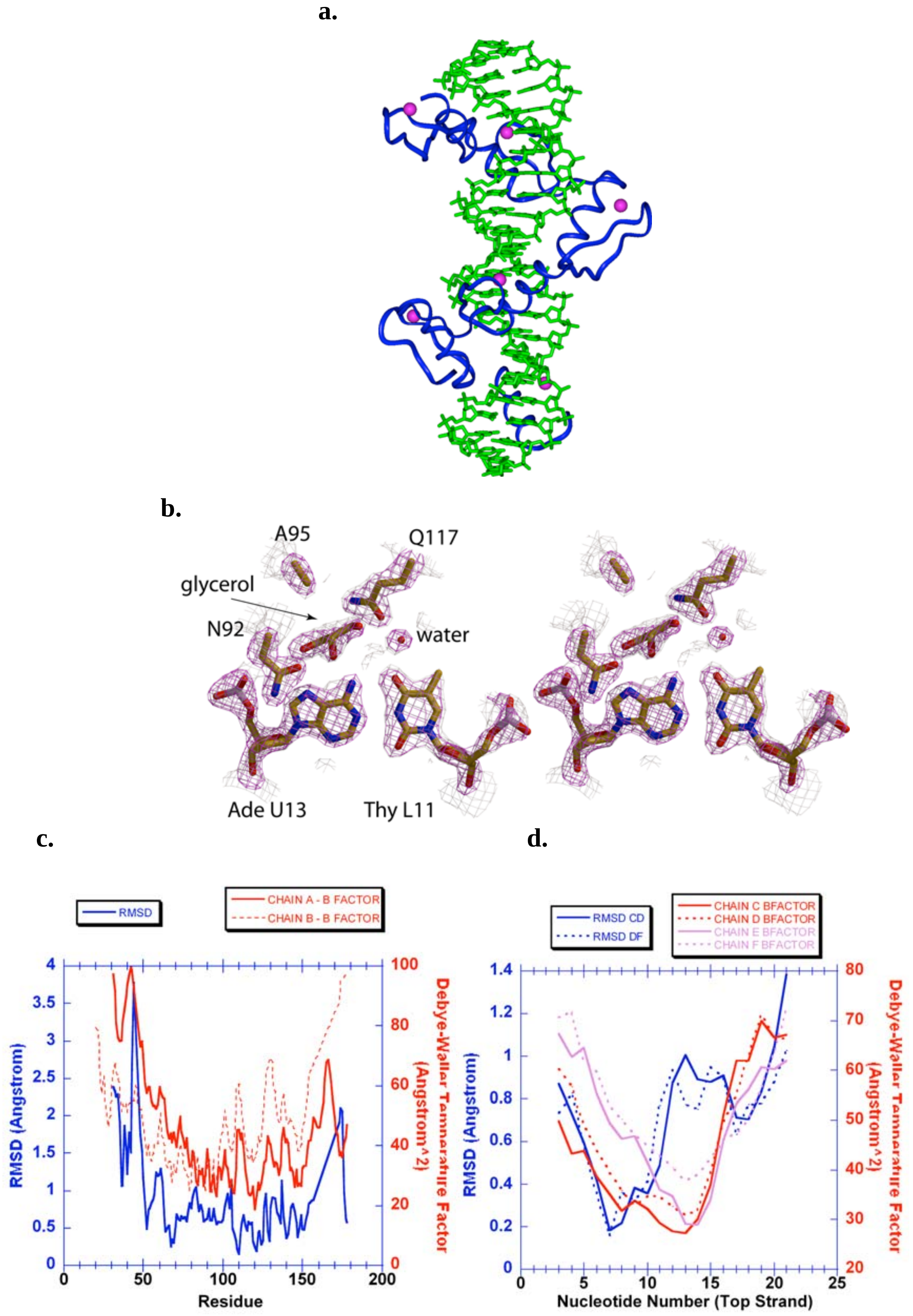

Figure 2. Segal et al.

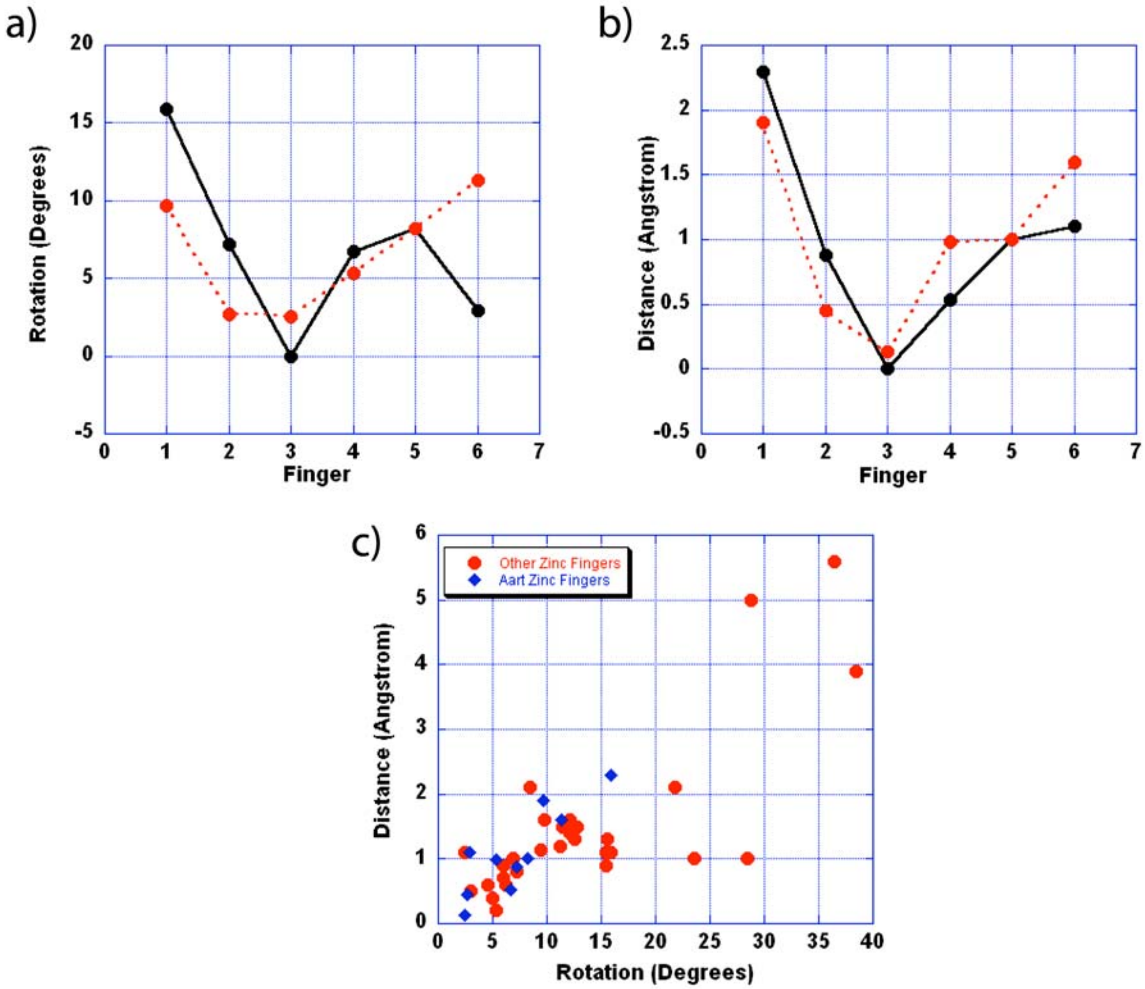

Figure 3. Segal et al.

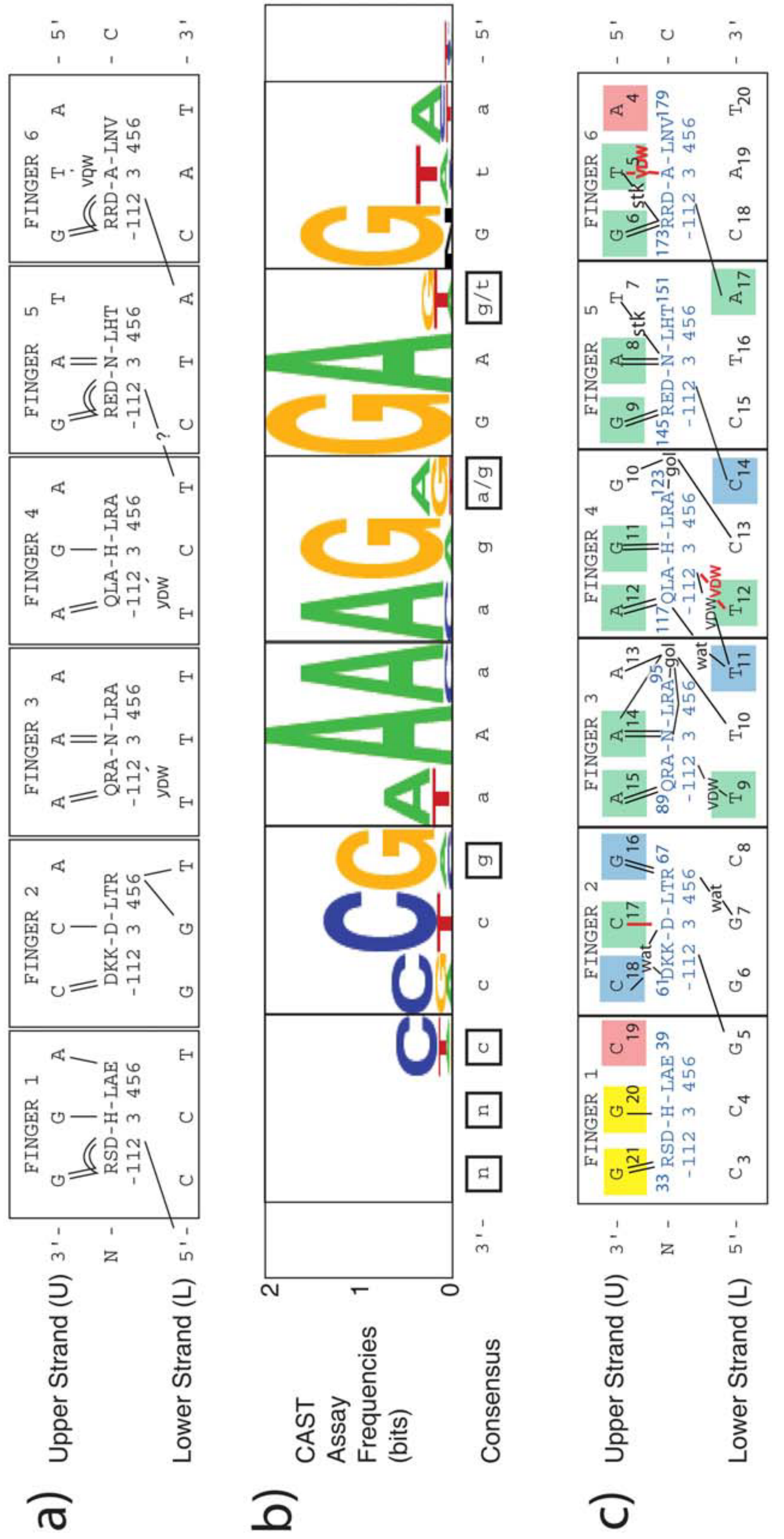
a)
Upper Strand (U) 3'- G G A C C A A A A A G A G A T G T A -5'
N - RSD-H-LAE DKK-D-LTR QRA-N-LRA QLA-H-LRA RED-N-LHT RRD-A-LNV - C
-112 3 456
Lower Strand (L) 5'- C C T G G T T T T T C T C T A C A T -3'
FINGER 1
FINGER 2
FINGER 3
FINGER 4
FINGER 5
FINGER 6
vdw
?
b)
CAST Assay Frequencies (bits)
2
1
0
Consensus 3'- n n c c c g a A a a g a/g G A g/t G t a -5'
c)
Upper Strand (U) 3'- G21 G20 C19 C18 C17 G16 A15 A14 A13 A12 G11 G10 G9 A8 T7 G6 T5 A4 -5'
N - 33 RSD-H-LAE 39 61 DKK-D-LTR 67 89 QRA-N-LRA 95 117 QLA-H-LRA 123 145 RED-N-LHT 151 173 RRD-A-LNV 179 - C
Lower Strand (L) 5'- C3 C4 G5 G6 G7 C8 T9 T10 T11 T12 C13 C14 C15 T16 A17 C18 A19 T20 -3'
wat
gol
stK
vdw

Figure 4. Segal, et al.

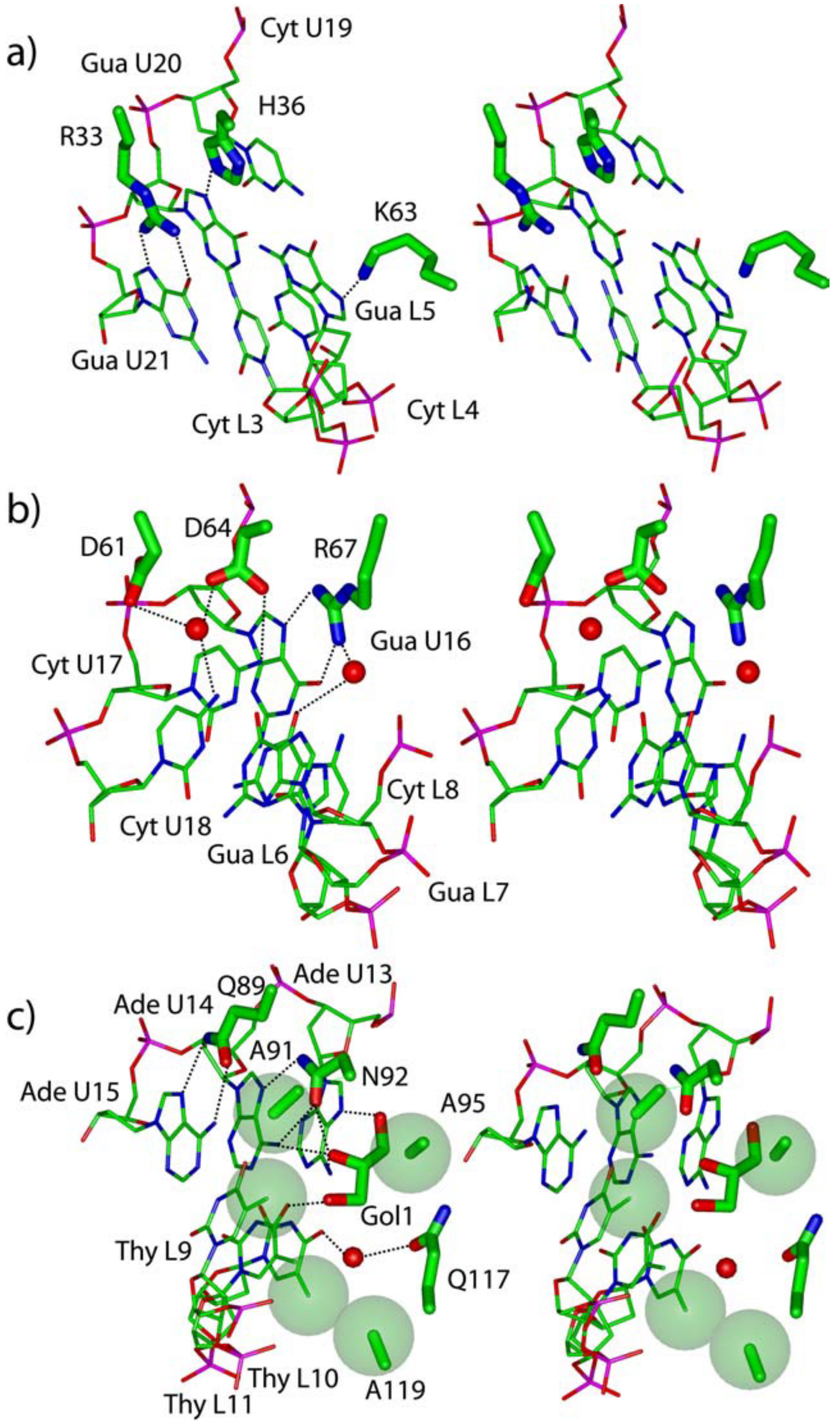

Figure 5. Segal, et al. (2006)

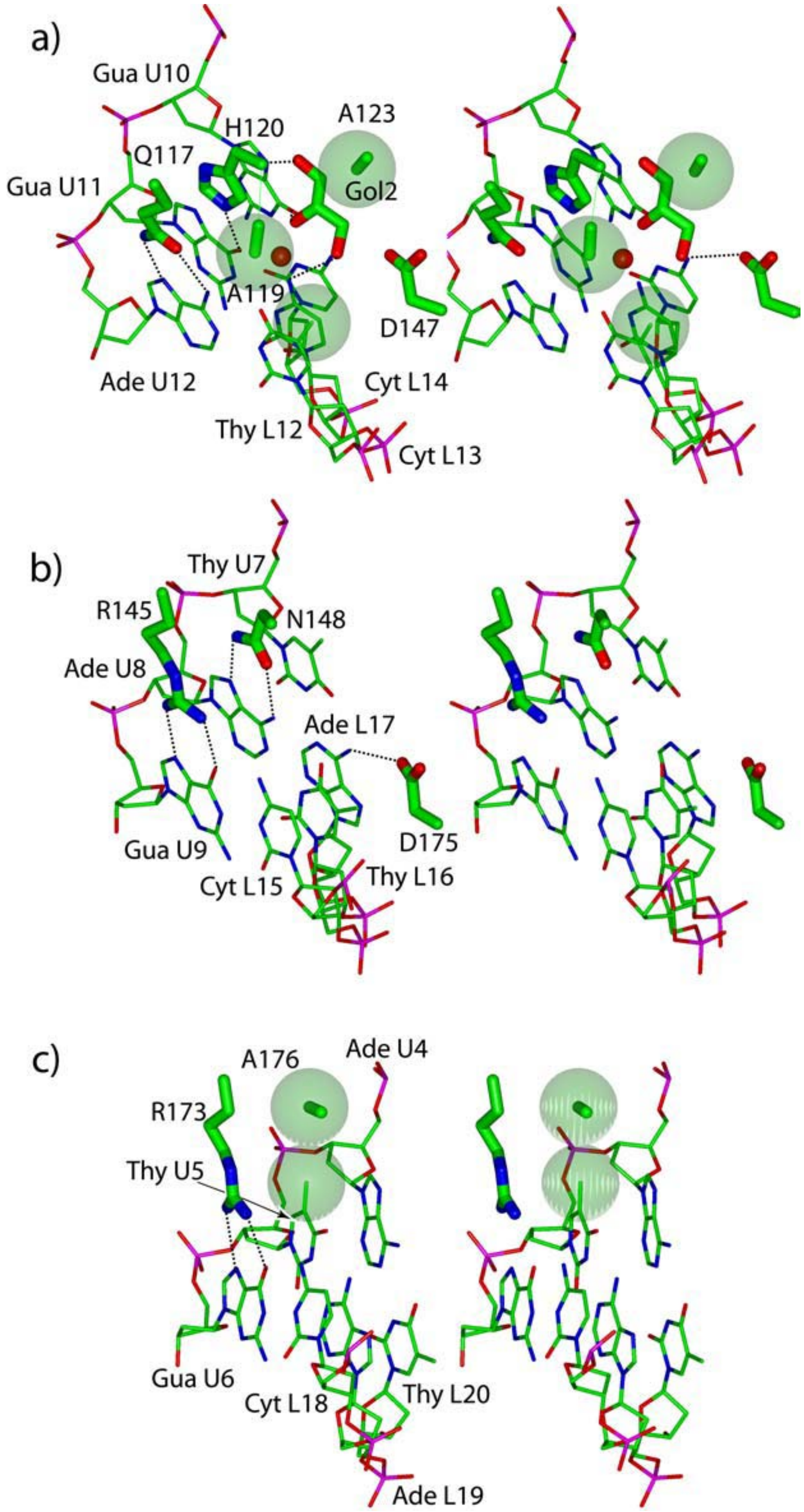

Figure 6. Segal, et al. (2006)

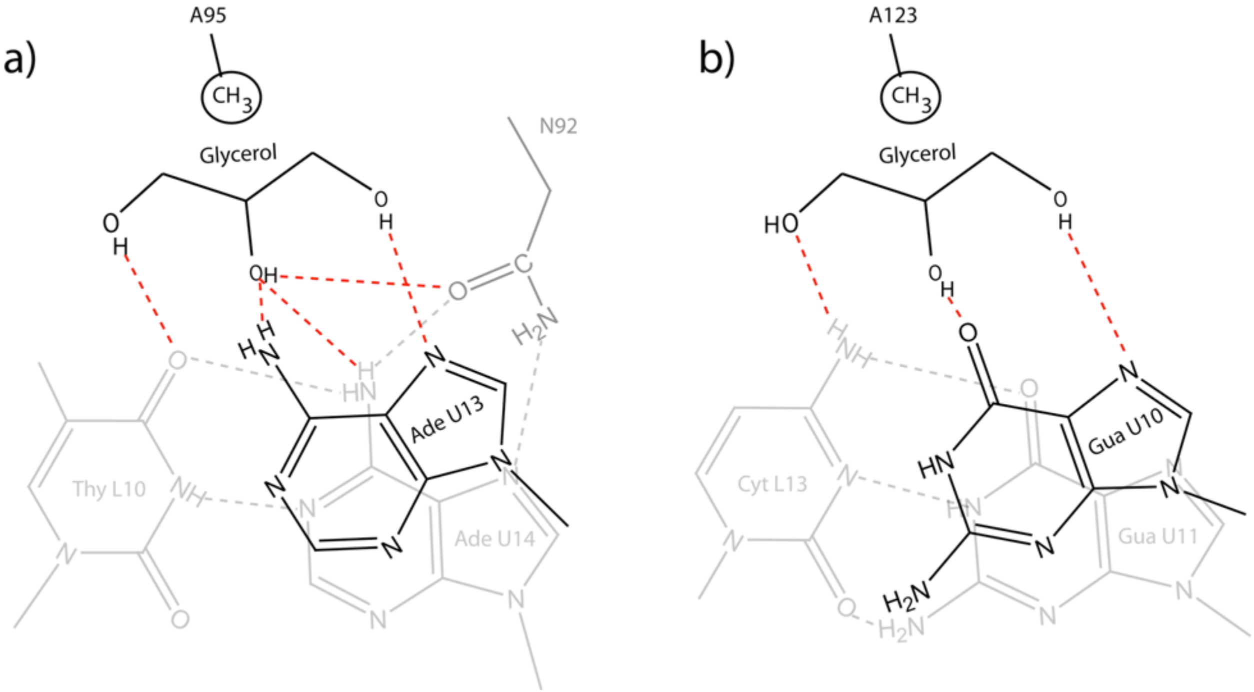

Figure 7 Segal et al.