*Accelerated Publications*

# Template-Directed Interference Footprinting of Cytosine Contacts in a Protein–DNA Complex: Potent Interference by 5-Aza-2′-deoxycytidine†

Kathleen C. Hayashibara and Gregory L. Verdine*

*Department of Chemistry, Harvard University, Cambridge, Massachusetts 02138*

*Received July 1, 1992; Revised Manuscript Received September 14, 1992*

ABSTRACT: In the template-directed interference (TDI) footprinting method (Hayashibara & Verdine, 1990), analogs of the naturally occurring DNA bases are incorporated into DNA enzymatically and assayed for interference of sequence-specific binding by a protein. Here we extend this method to include analysis of contacts of amino acid residues to the major groove surface of cytosine residues (TDI-C footprinting). The base analog 5-aza-2′-deoxycytidine, in which the hydrophobic 5-CH of cytosine is replaced by a hydrophilic aza nitrogen, was incorporated into DNA via the corresponding 5′-triphosphate. The analog was found to base pair with guanine during polymerization, resulting in substitution of 2′-deoxycytidine residues. TDI-C footprints of the λ repressor–$O_L1$ operator complex revealed apparent contacts to the cytosines at operator positions 7 and 8. Inspection of the high-resolution X-ray crystal structure of the λ–$O_L1$ complex (Clarke et al., 1992; Beamer & Pabo, 1992) revealed that C8 makes a hydrogen binding contact with the $Lys_3$; C7, on the other hand, makes a previously unnoticed hydrophobic contact with the alkane side chain of $Lys_3$. In only the consensus operator half-site was cytosine interference observed, suggesting that the nonconsensus arm binds DNA very differently if at all. The N-terminal arm represents the archetypal case of a sequence-specific peptide–DNA complex characterized at high resolution; thus, the present studies suggest strategies for design and screening of DNA binding peptides. The finding that 5-aza-2′-deoxycytidine inhibits sequence-specific DNA binding proteins may suggest an alternative rationale for the biological activities of this and related azapyrimidine nucleosides.

Noncovalent interactions between specific amino acid residues and the bases of DNA contribute substantially to sequence discrimination in protein–DNA complexes. These contacts involve predominantly the formation of hydrogen bonds ("HB[1] contacts") or association of hydrophobic elements ("hydrophobic contacts").[2] Underlying the challenge of understanding how such weak interactions confer strong and specific binding is the more pragmatic problem of how to determine which amino acids contact which DNA bases—elucidation of these partners provides a rough blueprint for macromolecular recognition in protein–DNA complexes.

For identification of contacted DNA bases, a promising approach has entailed chemically modifying DNA and then analyzing how the modifications affect protein binding. In the most widely used set of experiments, termed interference footprinting, DNA is modified by treatment with reagents that attack either the base (Siebenlist & Gilbert, 1980; Brunelle & Schleif, 1987) or backbone (Hayes & Tullius, 1989). These methods have generated useful information but have been difficult to generalize for several reasons: (i) many prospective contact positions cannot be selectively modified by known reagents; (ii) many reagents generate drastic lesions in DNA, which may interfere with amino acid contacts to adjacent bases; and (iii) some base modifications involve both major and minor groove contact functionality; consequently, interference cannot be assigned definitively to either groove. In order to circumvent some of the limitations of reagent-based interference footprinting approaches, we have devised a different approach in which principles of molecular design and synthesis are applied to address specific structural questions about a protein–DNA interface. In this procedure—template-directed interference (TDI) footprinting—nucleoside analogs are designed to introduce non-native functionality into the base-contact surface of DNA; the analogs, incorporated into DNA enzymatically, are tested for site-specific interference of protein binding. We have previously demonstrated this approach in studies showing that the analog $N^7$-methyl-2′-deoxyguanosine ($m^7dG$) was incorporated efficiently into DNA via enzymatic polymerization to yield exclusively G-specific interference footprints (Hayashibara & Verdine, 1991). In this report we demonstrate strong and site-specific interference of protein binding by the dC analog 5-aza-2′-deoxycytidine ($a^5dC$) (Piskala & Sorm, 1978).

† This work was supported by grants (to G.L.V.) from the National Science Foundation (Presidential Young Investigator Program), Chicago Community Trust (Searle Scholars Program), and Hoffmann-La Roche (Institute of Chemistry and Medicine). G.L.V. is a Fellow of the Alfred P. Sloan Foundation, a Dreyfus Teacher-Scholar, and a Lilly Grantee. K.C.H. is the recipient of a predoctoral fellowship from the National Science Foundation.

* To whom correspondence should be addressed.

[1] Abbreviations: HB, hydrogen bond(ing); TDI, template-directed interference; TDI-C, template-directed interference footprinting of cytosines; TDI-G, template-directed interference footprinting of guanines; $a^5C$, 5-azacytosine; $a^5dC$, 5-aza-2′-deoxycytidine; $a^5dCTP$, 5-aza-2′-deoxycytidine 5′-triphosphate; H-T-H, helix–turn–helix; $K_d$, thermodynamic dissociation constant.

[2] The terms "hydrophobic contacts" and "hydrophobic interactions" are used here to denote associations between nonpolar atoms that result in reduction of overall solvent-accessible surface. The thermodynamic forces driving formation of these contacts remain a subject of discussion (Ha et al., 1989, and references cited therein).

0006-2960/92/0431-11265$03.00/0 

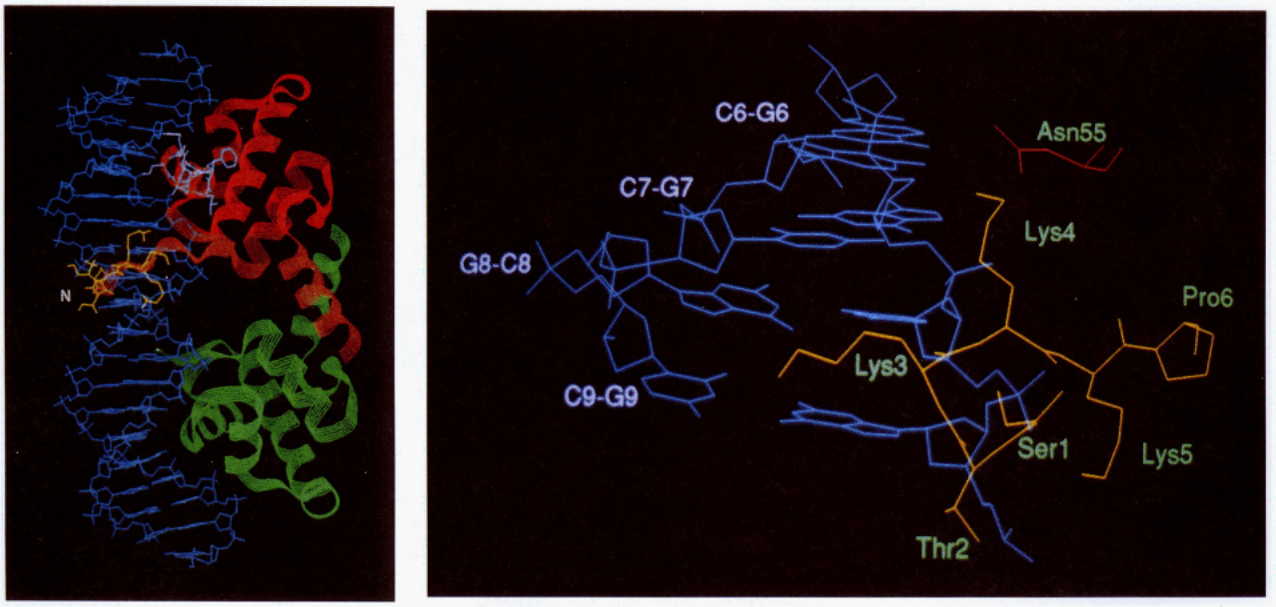



FIGURE 1: Structure of the λ repressor–$O_L1$ complex refined to 1.8-Å resolution (Clarke et al., 1991):[3] (A, top left) entire complex; (B, top right) closeup view of operator positions 6–9 and nearby amino acid residues; (C, bottom) sequence and numbering of $O_L1$. The coloring scheme of (A) and (B) is as follows: the consensus subunit of the λ repressor is shown in red and the nonconsensus subunit in green. Residues of the consensus H-T-H recognition helix [see Aggarwal and Harrison (1988)] are shown explicitly in white and the residues of the N-terminal arm in yellow. N denotes the N-terminus of the consensus subunit. In (B), $Asn_{55}$ is shown, in addition to the residues of the N-terminal arm. Only non-hydrogen atoms are shown.

We chose the bacteriophage λ repressor–operator complex for the present study, owing in part to the wealth of biochemical and structural information available on this system [reviewed in Sauer et al. (1990)]. Of particular relevance are X-ray cocrystal structures of λ repressor bound to the operator $O_L1$ (Jordan & Pabo, 1988), including one recently refined to 1.8-Å resolution (Clarke et al., 1991; Beamer & Pabo, 1992). These studies revealed that the λ repressor belongs to an archetypal class of DNA binding proteins having a "helix–turn–helix" (H-T-H) motif that establishes many of the contacts to DNA bases (Aggarwal & Harrison, 1988). In addition to the canonical H-T-H motif, the λ repressor bears at its N-terminus an unusual hexapeptide segment ($Ser_1$-$Thr_2$-$Lys_3$-$Lys_4$-$Lys_5$-$Pro_6$-)—commonly called the N-terminal "arm"—which wraps around the back of the operator site in the major groove (Figure 1A).[3] Evidence suggests that the N-terminal arm is disordered in solution (Weiss et al., 1984), adopting an ordered conformation only in the presence of its specific DNA recognition site, operator positions 6–9 (Figure 1B). The absence of stable secondary structure in the N-terminal arm results in part from its lack of tertiary interactions with the main body of the λ repressor, even when bound to DNA (Figure 1; Jordan & Pabo, 1988; Clarke et al., 1991). Thus, the N-terminal arm can be viewed as a small, extended DNA binding peptide tethered to the λ repressor via main-chain covalent linkage. Although small peptides generally fail to bind DNA tightly or specifically, the λ repressor apparently represents an exceptional case, since biochemical studies have revealed that DNA contacts involving the N-terminal arm contribute substantially to the strength and specificity of repressor–operator interactions (Pabo et al., 1982; Eliason et al., 1985; Nelson & Sauer, 1986; Brunelle & Schleif, 1987; Hayes & Tullius, 1989; Sarai & Takeda, 1989; Clarke et al., 1991). The N-terminal arm not only appears to bind DNA sequence-specifically but also does so with exceptional economy, using only two amino acid residues ($Lys_3$ and $Lys_4$) to recognize specifically a four base-pair segment of DNA (Figure 1B) (Jordan & Pabo, 1988; Clarke et al., 1991; Beamer & Pabo, 1992).

Small, flexible DNA binding molecules such as peptides, peptidomimetics, and carbohydrates represent prospective candidates for artificial regulators of gene expression. However, principles guiding the design of such molecules are few at present, largely because of the scarcity of well-documented examples. Since the N-terminal arm of the λ repressor represents the archetypal case of a peptide–DNA complex characterized at high resolution, studies of this system may yield general principles that could be used in efforts directed at de novo design of DNA binding peptides. Previous studies, although informative, have involved relatively gross changes in the contact interface (Pabo et al., 1982; Eliason et al., 1985; Nelson & Sauer, 1986; Brunelle & Schleif, 1987; Hayes & Tullius, 1989; Sarai & Takeda, 1989; Clarke et al., 1991), which may thereby have obscured relevant details of the arm–DNA interactions. TDI footprinting, on the other hand, allows atom-level changes to be made in the peptide–DNA interface. Using this technique, we find evidence for a high degree of cooperativity both within the arm–DNA contacts and between the arm and H-T-H elements. Furthermore, we find evidence for a hydrophobic interaction between the side chain of $Lys_3$ and the position 5,6 C=C of C7, which had previously gone unnoticed and which may contribute substantially to the strength of the peptide–DNA interaction.

[3] In this structure (Beamer & Pabo, 1992), all residues of the N-terminal arm except $Ser_1$ and $Thr_2$ are resolved to 1.8 Å; the first two residues, which do not contribute to binding (Clarke et al., 1991), are less ordered in the structure.

The data presented herein show that a$^5$dC residues in DNA can strongly inhibit binding of a sequence-specific regulatory protein. a$^5$dC and the corresponding prodrug 5-azacytidine (Piskala & Sorm, 1964) are well-known agents in cell biology (Riggs, 1983) and cancer chemotherapy (Vogler et al., 1976) and have been shown to exert powerful and pleiotropic effects on cell differentiation [reviewed in Jones (1984) and Holliday (1987)]. Although these gene regulatory effects have primarily been ascribed to inhibition of DNA methyltransferase enzymes by DNA-borne a$^5$dC (Jones, 1984; Santi et al., 1984, 1984; and references cited therein), our results suggest the possibility of an additional mode of action: direct inhibition of transcriptional regulatory proteins by preventing their binding to DNA.

## EXPERIMENTAL PROCEDURES

### *Materials*

Chromatographic resins and naturally occurring dNTPs were from Pharmacia (Piscataway, NJ). *Taq* polymerase was from Promega (Milwaukee, WI). pBS+ was from Stratagene (La Jolla, CA). T4 DNA ligase, T4 polynucleotide kinase, and yeast tRNA were from Bethesda Research Laboratories (Gaithersburg, MD). Exonuclease-free Klenow polymerase, Sequenase 2.0, and the Sequenase 2.0 kit were from United States Biochemicals (Cleveland, OH). 5-Aza-2′-deoxycytidine was obtained from Sigma (St. Louis, MO). *Pst*I was from New England Biolabs (Beverly, MA). Oligonucleotides were synthesized on an Applied Biosystems 380A DNA synthesizer.

### *Methods*

*λ Repressor.* The λ repressor was purified from *Escherichia coli* RB968/pEA305 (Amman et al., 1983) essentially as described (Johnson et al., 1980), with the following modifications: (i) a KCl gradient from 50 to 400 mM was used to elute repressor from QAE-Sephadex (A-25), and (ii) the hydroxylapatite column was omitted.

*Preparation of Template DNA.* A phagemid vector containing $O_L1$ in its polylinker, pBS-$O_L1$, was constructed by insertion of a duplex oligodeoxynucleotide into the *Pst*I restriction site of pBS+ (Figure 3A). This vector was then used as a source of $O_L1$ in asymmetric PCR (Gyellenstein & Erlich, 1988), along with commercially available sequencing primers, to generate single-stranded templates containing the top or bottom strand of $O_L1$ (Figure 3). For each template strand, ten 50-μL PCRs were run in parallel. The reactions contained 50 mM KCl, 10 mM Tris-HCl (pH 8.4), 1.5 mM $MgCl_2$, 100 μg/mL gelatin (Fisher Scientific, purified grade, 275 bloom), 0.1% Triton X-100 (Sigma), 200 μM each of dNTP, 1 unit of *Taq* polymerase, 10 fmol of pBS-$O_L1$, 50 pmol of primer for template strand synthesis, 0.5 pmol of primer for complementary strand synthesis, and a mineral oil cap. A 100:1 ratio of primers A (NEB no. 1233) and D (NEB no. 1224) was used to generate template strand A (top strand) and a 1:100 ratio of primers A and D was used to obtain template strand D (bottom strand). After 35 cycles, the aqueous phases were combined, extracted with phenol, phenol/chloroform, chloroform, and ether, and then washed with 3 × 300 μL of TE (10 mM Tris-HCl, pH 8.0, 1 mM EDTA) in an Ultrafree MC filter (Millipore, 10-kDa nominal exclusion limit). The retentate compartment was washed twice with 50 μL of TE to recover the DNA. The combined washes were typically found to contain ~10 pmol of template, as determined by sequencing with a known amount of primer. This amount of template is sufficient for 50 TDI footprinting lanes.

*Synthesis of 5-Aza-dCTP.* The phosphorylation of a$^5$dC was carried out by a modification of a published procedure (Ruth & Cheng, 1981). 5-Aza-2′-deoxycytidine (25 mg) was dissolved in 220 μL of trimethyl phosphate (Aldrich) with magnetic stirring under a nitrogen atmosphere at room temperature. After being stirred for 15 min, most of the nucleoside had dissolved, and 20 mL of phosphorus oxychloride was added. The solution became clear and then from it precipitated a white solid. The reaction was allowed to proceed 30 min, and the resultant mixture was removed by syringe and then added immediately to another stirred flask containing 660 μL of 1 M tris(tributylammonium) pyrophosphate in DMF under nitrogen. The reaction was allowed to proceed 5 min, during which time the solution became clear. The reaction mixture was removed by syringe and quickly added to a 4 °C mixture of 0.15 mL of triethylamine plus 2.2 mL of doubly distilled $H_2O$ and placed on ice. The precipitate that formed initially dissolved upon swirling. The reaction mixture was stoppered and stored at 4 °C overnight.

Purification of a$^5$dCTP was carried out on DEAE-Sephadex A-25 using a 1000-mL gradient of trimethylammonium formate from 0 to 2 M. Peak fractions were desalted on a Sephadex G-10 column in 0.1 M triethylammonium bicarbonate (pH 7), lyophilized, redissolved in TE, and quantified by UV using an extinction coefficient of 6026 at 246 nm (Piskala & Sorm, 1964).

*Template-Directed Incorporation of a$^5$dC into DNA.* The following quantities are required for four lanes of a TDI-C footprint.

End labeling of primers: primers B and C were used in combination with template strands A and D, respectively (Figure 3). In separate reactions, 1 μL of 0.2 pmol/μL primers B and C were end labeled in a 5-μL volume by adding 0.5 μL of 400 mM Tris-HCl (pH 7.0), 200 mM $MgCl_2$, 1 unit of T4 polynucleotide kinase, and 0.4 μL of 6000 Ci/mmol [γ-$^{32}$P]-ATP (New England Nuclear). This mixture was incubated at 37 °C for 30 min, heated at 90 °C for 5 min, allowed to cool to room temperature, microcentrifuged briefly, and then used directly in the annealing step.

Annealing: DNA obtained by asymmetric PCR was diluted to 0.04 pmol/μL in TE. A 5-μL volume of this template solution was added to the end-labeled primer (5 μL), and the mixture was heated to 65 °C for 5 min and then allowed to cool to room temperature. Condensate was recombined by brief microcentrifugation.

Primer extension: 10 μL of annealed primer/template obtained above (0.2 pmol with respect to primer) was combined with 3.5 μL of polymerization buffer (60 mM Tris free base, 33 mM $MgCl_2$, 170 mM NaCl) and 2.0 μL of Klenow enzyme [diluted to 1.25 units/μL in a buffer containing 10 mM Tris-HCl (pH 7.5), 5 mM DTT, and 0.5 mg/mL BSA]. The resulting solution was added to 11 μL of a mixture containing 78.95 μM a$^5$dCTP, 1.05 μM dCTP, and 80 μM each of dATP, dGTP, and TTP and then to 50 mM NaCl. The extension was allowed to proceed at 37 °C for 30 min, then stopped by addition of 7.5 M sodium acetate (pH 7.0), and precipitated at −70 °C with 90 μL of 100% ethanol and 1 μL of 5 mg/mL yeast tRNA. The pellet obtained after microcentrifugation at 13 000 rpm for 10 min was washed with 80% aqueous ethanol, dried, and resuspended in 16 μL of TE prior to incubation with protein.

*Interference Binding Assays.* The following quantities are those required for one lane of a TDI-C footprint. A 4.0-μL aliquot (0.05 pmol) of extended template DNA was incubated with varying amounts of repressor in 20 μL of buffer consisting

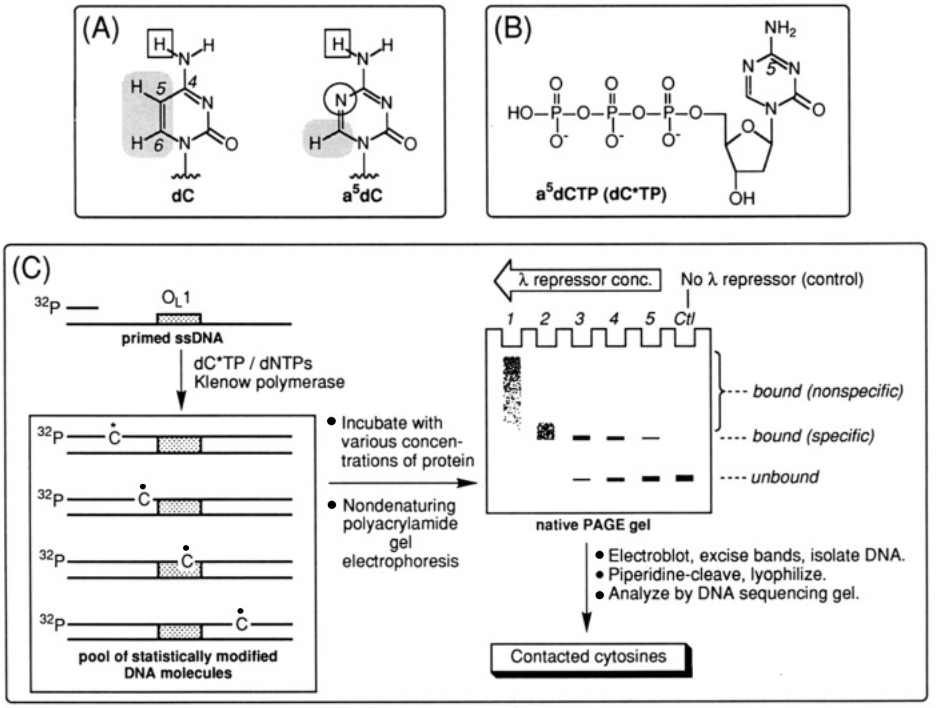



FIGURE 2: Overview of TDI-C footprinting: (A) structural comparison of the base moieties of dC and the analog $a^5dC$, with hydrogen-bonding donors boxed, acceptors circled, and hydrophobic elements shaded; (B) structure of $a^5dCTP$; (C) scheme illustrating individual steps of the TDI-C footprinting protocol. C* denotes $a^5dC$. Concentrations of protein (based on dimer) used in the binding reactions are as follows: lane 1, $1 \times 10^{-7}$ M; lane 2, $5 \times 10^{-8}$ M; lane 3, $1 \times 10^{-8}$ M; lane 4, $5 \times 10^{-9}$ M; lane 5, $1 \times 10^{-9}$ M; lane Ctl, 0 M (control). The DNA concentration in the binding experiment is $\sim 1 \times 10^{-8}$ M.

of 50 mM KCl, 10 mM Tris-HCl (pH 7.8), 1 mM $MgCl_2$, 1 mM $CaCl_2$, 100 μg/mL bovine serum albumin, 10% (v/v) glycerol, and 2.5 μg/mL salmon sperm DNA for 30 min at room temperature. Protein-bound DNA was separated from unbound as described (Chodish, 1989). Briefly, the binding reactions were electrophoresed on a 4% nondenaturing polyacrylamide gel, the gel was electroblotted onto a DEAE membrane (Schleicher & Schuell NA 45), and bands were visualized by autoradiography and cut out with scissors. The DNA was eluted, extracted with phenol/chloroform, and precipitated with ethanol.

The DNA pellet was heated to 90 °C for 30 min in 50 μL of 10% (v/v) aqueous piperidine, cooled to room temperature, and microcentrifuged briefly. Samples were lyophilized and then twice resuspended in 15 μL of distilled $H_2O$ and lyophilized. Samples were dissolved in loading dye containing 45% formamide, 10 mM EDTA, 0.025% bromphenol blue, and 0.025% xylene cyanol and electrophoresed on a 10% acrylamide sequencing gel.

*Sanger Sequencing.* In two separate reactions, single-stranded templates A and D were annealed to $^{32}P$ end-labeled primers B and C, respectively (Figure 3) on a 10-μL scale as described above (Primer extension). This was combined with 3.5 μL of a solution containing 60 mM Tris-HCl (pH 7.8), 33 mM $MgCl_2$, and 170 mM NaCl, followed by 2.0 mL of Sequenase 2.0 diluted to 1.4 units/μL in buffer containing 10 mM Tris-HCl (pH 7.5), 5 mM DTT, and 0.5 mg/mL bovine serum albumin. The termination reactions were carried out by adding 3.5-μL aliquots (0.05 pmol with respect to primer) of the above mixture to a 2.5-μL aliquot of each dideoxy termination mix from a Sequenase 2.0 sequencing kit according to the manufacturer's protocol.

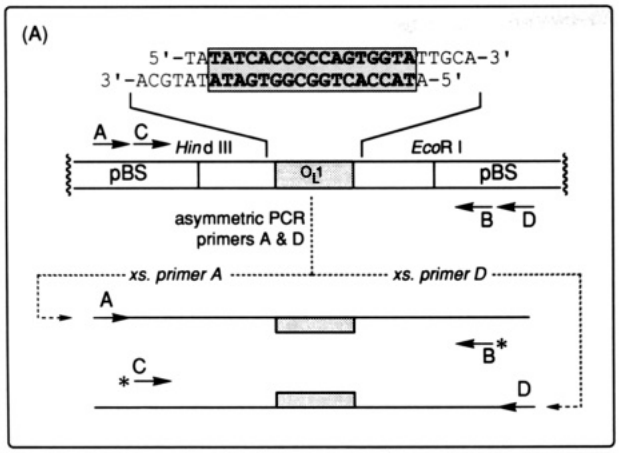
## RESULTS

For TDI footprinting of cytosine contacts (TDI-C footprinting), we chose the analog $a^5dC$ (Piskala & Sorm, 1978), in which a hydrophilic aza nitrogen replaces the hydrophobic 5-CH of cytosine (Figure 2A). This analog is expected to interfere by either of two modes: (i) disruption of hydrogen-bonding interactions to the $N^4$-H of dC, by electron-pair repulsion between the hydrogen bond acceptor and the aza lone pair, and (ii) disruption of hydrophobic interactions[2] involving the 5,6 double bond of dC, by introduction of the hydrophilic aza N. Since the aza N of $a^5dC$ is analogous electronically, sterically, and positionally to the aza N (i.e., $N^7$) of purines, $a^5dC$ is expected to perturb DNA secondary structure no more than that caused by changing the sequence.[4]

Enzymatic incorporation of $a^5dC$ into DNA was accomplished via the corresponding 5′-triphosphate, $a^5dCTP$ (Figure 2B). In preliminary primer extension experiments (not shown), various ratios of $a^5dC$ to dC were tested for their ability to support polymerization. The products of extension were treated with piperidine to cause specific strand scission at $a^5dC$ sites, thereby revealing the extent of incorporation of the analog. These experiments showed that (i) $a^5dCTP$ is efficiently incorporated into DNA by Klenow polymerase (more so than modified T7 polymerase; data not shown), (ii) $a^5dCTP$ pairs predominantly with dG residues in the template, and (iii) a dCTP:$a^5dCTP$ ratio of 1:75 directs incorporation of approximately one $a^5dC$ residue during polymerization of a ~200-nt template.

The protocol used for TDI-C footprinting of the λ repressor–$O_L1$ complex is illustrated in Figure 2C. A single-stranded template containing $O_L1$, annealed to a $^{32}P$ end-labeled primer, was extended using a 1:75 ratio of dCTP to afford duplex DNA molecules substituted statistically with $a^5dC$ on one

[4] Modest differences in bond lengths are also expected between C and $a^5C$: the C=N bond has an average length of 2.8 Å; C=C bond, 1.34 Å; $sp^2$-N—$sp^2$-C bond, 1.36 Å; and $sp^2$-C—$sp^2$-C bond, 1.48 Å (March, 1985). In addition, differences may exist between the electropositivity (hydrogen bond donating strength) of the $N^4$-H in C and $a^5C$. These factors may also contribute to interference by $a^5dC$ residues in DNA.




(B)

| Primer | Sequence |
|---|---|
| A | 5'-d(AGCGGATAACAATTTCACACAGGA) |
| B | 5'-d(GTAAAACGACGGCCAGT) |
| C | 5'-d(AACAGCTATGACCATG) |
| D | 5'-d(CGCCAGGGTTTTCCCAGTCACGAC) |

(C)

| Experiment | Nucleotide components used in primer extension |
|---|---|
| Sanger sequencing | dNTPs, ddNTPs |
| TDI-C footprinting | $a^5$dCTP, dNTPs |

FIGURE 3: Details of templates and primers used in this study. (A) Generation of single-stranded templates containing either the $O_L1$ (shaded) top or the $O_L1$ bottom strand. Oligonucleotide primers are denoted by arrows, with the arrowhead corresponding to the 3′ end; • denotes 5′-$^{32}$P end label. At the top is the sequence of a duplex oligonucleotide containing $O_L1$ (shaded) that was inserted into pBS to furnish an $O_L1$-containing phagemid used in asymmetric PCR, along with primers A and D. (B) Sequences of primers. (C) Nucleoside triphosphate components used in extension of the primed templates shown in (A).

strand. This pool was then incubated with various concentrations of the λ repressor, and the protein-bound DNA was separated from the unbound DNA. After electroblotting of the gel to a DEAE membrane and exposure of the blot to film, the individual bands were recovered from the membrane, treated with piperidine, and analyzed on a DNA sequencing gel to reveal the locations of the analog.

In Figure 4, each panel contains a Sanger sequence and TDI-C footprint for each strand of $O_L1$: left panel, top strand; right panel, bottom strand. Next to the Sanger sequence in each panel is a control lane (Ctl) in which the entire pool of $a^5$dC-containing DNA was directly cleaved with piperidine. Each band represents a position in the sequence at which $a^5$dC had been incorporated during polymerization (cf. Figure 2C). Taking into account the slower mobility of dideoxy-terminated DNA versus piperidine-cleaved DNA (which is one nucleoside unit shorter), the fragments in the Sanger C lane correspond in a 1:1 fashion with the fragments in the TDI-C control lane Ctl. Thus, as stated earlier, $a^5$dC pairs with dG during enzymatic polymerization, resulting in replacement of dC residues with the analog.

In each TDI-C footprint (Figure 4), five lanes represent cleavage products of protein-bound DNA isolated from the native PAGE gel (Figure 2C) and five lanes represent cleavage products of unbound DNA isolated from the native PAGE gel. Lanes with the same number represent DNA samples obtained at the same protein concentration, with 1 being the highest protein concentration and 5 the lowest. At each dC in the sequence, the band intensity in each of the five lanes represents the tendency of DNA containing $a^5$dC at that position to bind the λ repressor at that concentration. What results, in comparing horizontally from lane 1 to lane 5 at each dC position, is a titration profile of that $a^5$dC-substituted position with the λ repressor. Most dC residues in the sequence lie outside the specific binding site ($O_L1$), where the analog should have no effect on binding by the λ repressor; these positions yield horizontal bands with similar titration profiles (even though the absolute intensities vary among them, because of sequence-dependent differences in rate of incorporation of $a^5$dC). However, when the analog is present at a contact site, it interferes with binding by the protein, and this interference can only be overcome by increasing the protein concentration. Interfering positions can thus be identified in a TDI footprint as dC positions that have a shorter titration profile (bands in fewer lanes) in the bound lanes and a longer titration profile in the unbound lanes, compared to the majority of positions in the bound and unbound lanes, respectively.

In the bound lanes of Figure 4, most C positions have the same titration profile, with a band weakly visible in lane 5, strongly visible in lane 4, and fully present in lane 3. As expected, most of the unbound lanes show the opposite behavior: the bands are strongest in lane 5, weaker in lane 4, and weaker yet in lane 3. A markedly different titration profile (pattern of band intensities) is observed at only two positions in the footprint, and both of these are located in $O_L1$: C7 in the top strand and C8 in the bottom strand (Figure 4, arrows). At these two positions, significant radioactivity is found only in bound lanes 1 and 2 and is largely absent in lanes 3–5, indicating that the presence of the analog interferes with binding at lower protein concentrations. Since the protein concentrations used to generate lanes 1 and 2 were above the thermodynamic dissociation constant ($K_d$) for nonspecific binding to DNA (as judged by upward smearing on native PAGE—cf. Figure 2C—which is characteristic of nonspecific binding), the appearance of bound radioactivity in these lanes indicates only that $a^5$dC does not perturb nonspecific binding to DNA. On the other hand, the protein concentration used to generate lane 3 was below the nonspecific $K_d$, and under these conditions the bound radioactivity at C7 and C8 was virtually absent. Taken together, these data suggest that $a^5$-dC at position C7 or C8 virtually abolishes sequence-specific DNA binding by the λ repressor.

In the unbound lanes of Figure 4, the titration profile of C7 and C8 is different from the other positions; however, the effect is not as striking as that observed in the bound lanes. This phenomenon not only is observed in TDI footprinting but is commonly observed in all interference footprinting methods. Nonetheless, interference is still clearly evident in the unbound fractions at C7 and C8.

The X-ray structure of the λ repressor–$O_L1$ complex reveals an apparent HB contact between C8 and the main-chain carbonyl of $Lys_3$ (Figure 5A). Presumably, $a^5$dC interferes with this contact by electrostatic repulsion between its aza lone pair and the electrons on the amide carbonyl oxygen. Interference at C7 is less apparent at first glance, since the only amino acid functionality nearby is the alkane side chain of $Lys_3$. However, addition of van der Waals surfaces to the 5,6 double bond of C7 and the side chain of $Lys_3$ reveals that the two associate to make a hydrophobic contact (Figure 5C). In this case, we believe that $a^5$dC causes interference by introducing a hydrophilic aza N into what would otherwise be a hydrophobic interface. It is unlikely that $a^5$dC causes interference mainly by alteration of DNA secondary structure,

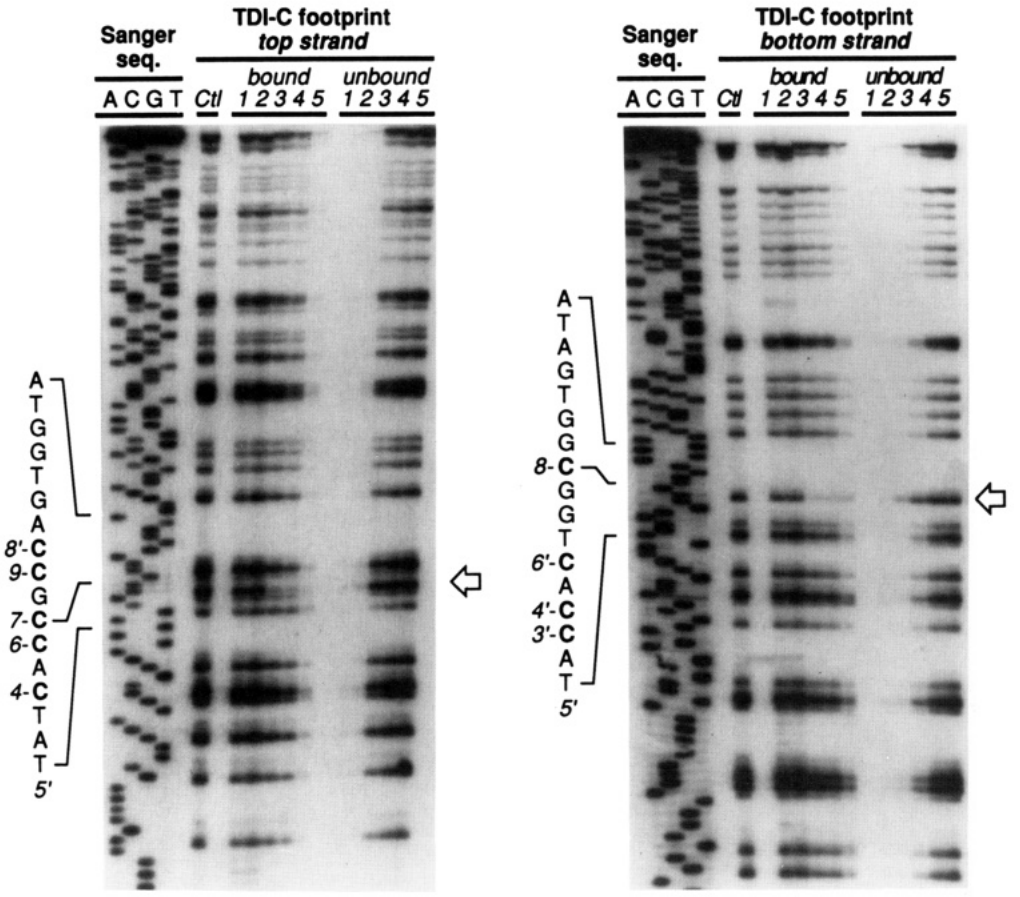



FIGURE 4: TDI-C footprints of the λ repressor–$O_L1$ complex. Each panel contains data for one strand of $O_L1$ (as defined by Figure 1C): left, top strand; right, bottom strand. The region of the gel corresponding to $O_L1$ is denoted at the left of each panel, adjacent to the Sanger sequencing lanes. In the TDI-C footprinting lanes, the Ctl lanes are controls in which the $a^5$dC-containing DNA was cleaved directly without having been incubated with the λ repressor; bound lanes represent protein-bound DNA fractions from nondenaturing PAGE; unbound lanes represent unbound (free) DNA from nondenaturing PAGE; lane numbers (and, therefore, protein concentrations) correspond to those in the PAGE gel illustrated in Figure 2C. Arrows denote positions at which interference is evident (see text).

not only because it so closely resembles dC but also because it does not cause interference at any other cytosine in $O_L1$, including those that are base paired to contact dG residues (C4, 4′, 6, 6′, 9). $a^5$dC also causes no interference when paired to a contact dG in the 434 repressor–operator complex.[5] Finally, the lack of interference at C8′ suggests that in the nonconsensus half-site the N-terminal arm is either not bound to DNA at all or is not bound in the same way as in the consensus half-site.

In summary, at either of two sites in the contact interface between $O_L1$ and the N-terminal arm of the λ repressor, the conservative change of a vinyl methine function to an aza nitrogen (Figure 2A) renders the DNA incapable of being recognized specifically by the protein.

[5] All DNA binding proteins can be considered a collection of peptides, each of which is in itself incapable of binding to DNA. However, in the overwhelming majority of these cases, the DNA binding element is part of an α-helix, the formation of which is associated with a large decrease in conformational entropy. In order for two DNA binding residues on successive turns of an α-helix to bind DNA, they must lie on the same helical face—that is, separated by three to four residues. For small peptides, the conformational entropy cost of ordering such a large number of residues in order to make each contact is presumably overwhelming. This situation is to be distinguished from the λ repressor N-terminal arm, in which two successive residues—part of an extended structure—make base contacts to four base pairs of DNA. Since most random peptide screening methods necessarily involve small peptides, the molecules being screened are more likely to be present in an extended structure such as that observed in the N-terminal arm.

## DISCUSSION

X-ray analysis of the λ repressor–$O_L1$ complex (Clarke et al., 1991; Beamer & Pabo, 1992) has revealed intimate contacts between the N-terminal arm and a four base-pair segment of the consensus half-site (Figure 1). In the nonconsensus half-site, electron density for the N-terminal arm is undefined, suggesting the absence of a stable interaction between the nonconsensus arm and its DNA site. The question of whether arm–DNA interactions are important for the overall stability and specificity of the λ repressor–operator complexes has been studied extensively by analyzing the effects of modifications in the protein–DNA interface. While these studies have strongly supported the notion that the consensus N-terminal arm binds DNA tightly, most have involved relatively gross structural changes such as mutation or removal of contact residues ordinarily present in the DNA (Brunelle & Schleif, 1987; Hayes & Tullius, 1989; Sarai & Takeda, 1989) or protein (Pabo et al., 1982; Eliason et al., 1985; Nelson & Sauer, 1986; Clarke et al., 1991). Hence, these prior studies may have left obscure subtle yet important aspects of the arm–DNA interaction. As a means of potentially overcoming some of these limitations, we have applied an alternative strategy—template-directed interference (TDI) footprinting—in which designed analogs of the naturally occurring DNA bases are used to introduce specific contact probes into the protein–DNA interface. Here, in a duplex DNA molecule containing $O_L1$, we have individually replaced the vinyl 5-CH of each cytosine with an aza N and analyzed at each modified site the

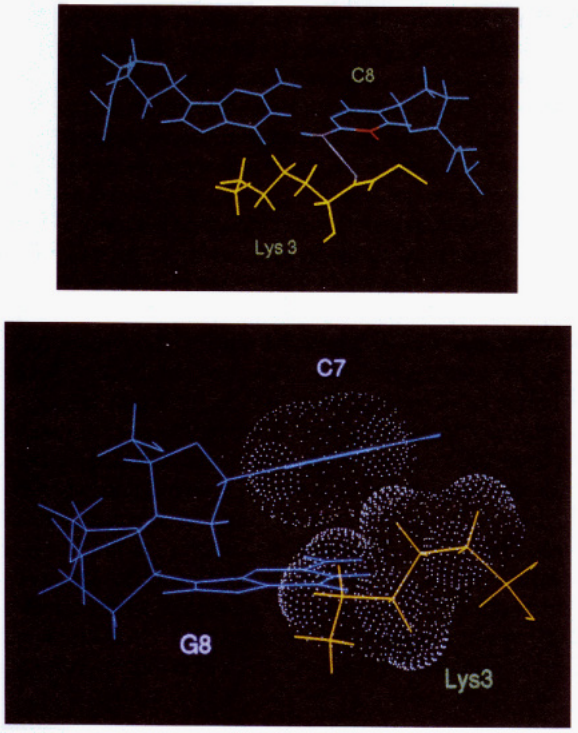



FIGURE 5: Cytosine contacts disrupted by substitution of dC for $a^5$dC in the λ repressor–$O_L1$ complex: (A, top) HB contact (white line) between the exocyclic amino group of C8 and the main-chain carbonyl of $Lys_3$; (B, bottom) hydrophobic contact between the 5,6-double bond of C7 and the side-chain methylenes of $Lys_3$, with van der Waals radii indicated by stippling. In (A), the position that becomes substituted by the aza-N of $a^5$dC is shown in red and, in (B), is in blue, directly under the caption heading C7. All atoms, including hydrogens, are shown explicitly in both framework and van der Waals representations.

effect on binding by the λ repressor. At two positions in the consensus half-site of $O_L1$, namely, C7 and C8, the 5-aza functionality prevents the operator from being recognized specifically by the λ repressor. On the other hand, interference is not observed in the nonconsensus half-site. These studies therefore confirm and extend earlier findings that (i) the N-terminal arm of the consensus-half site contributes substantially to the overall thermodynamics of the repressor–operator interaction and (ii) the N-terminal arm of the nonconsensus half-site binds DNA weakly, if at all, or binds in a much different way from the consensus arm (Pabo et al., 1982; Eliason et al., 1985; Nelson & Sauer, 1986; Brunelle & Schleif, 1987; Hayes & Tullius, 1989; Sarai & Takeda, 1989; Clarke et al., 1991).

These findings, taken in conjunction with earlier X-ray and biochemical studies, suggest that the consensus N-terminal arm of the λ repressor can be considered to be a sequence-specific DNA binding peptide, with the nominal recognition site 5′-d(CCGC) (operator positions 6–9; Figure 1). The economy of this peptide–DNA interaction is exceptional, considering that only two amino acid residues—$Lys_3$ and $Lys_4$—establish all of the arm contacts to the four base-pair site. X-ray analysis[3] (Clarke et al., 1991) had pointed to the importance of numerous HB contacts in the arm–DNA interface: the ϵ-amino group of $Lys^3$ contacts the position 6 C=O of G8 and G9; the main-chain carbonyl of $Lys_3$ contacts the $N^4$-H of C8; and the ϵ-amino group of $Lys_4$ contacts the position 6 C=O and $N^7$ of G7 and the position 6 C=O of G6. Although the $Lys_4$–G6 contact is made in conjunction with $Asn_{55}$ (Figure 1B), it would be G-specific even in the absence of $Asn_{55}$. In addition to these hydrogen-bonding contacts, our studies have suggested the presence of a previously unrecognized hydrophobic interaction between the hydrocarbon side chain of $Lys_3$ and the olefinic edge of C7, which may contribute substantially to the stability of the arm–DNA interaction. The aza N of $a^5$dC is in a similar region of major groove space as that occupied by the $N^7$ atom of purines; thus, one would expect purines to interfere when present at position 7 in the top strand of $O_L1$. This hypothesis cannot be tested directly, since the complementary change of G7 to a pyrimidine would disrupt the interaction with $Lys_4$. Nonetheless, the analogy between $a^5$dC and purines, combined with the observation of strong interference by $a^5$dC at C7, supports the notion that hydrophobic interactions cannot only affect the strength of protein–DNA interactions but also direct the specificity (Ha et al., 1989). At this stage it is unclear to what degree the $Lys_3$ side-chain contact to C7 would discriminate between C and T. It is remarkable that $Lys_3$ alone makes specific base contacts to three base pairs of DNA, operator positions 7–9. We have noted the presence of a similar contacts in the GAL4–DNA complex (Marmorstein et al., 1992), in which a single Lys (residue 18) recognizes three base pairs of DNA (operator residues 6–8). However, the sequences recognized by the two residues are not identical; $Lys_3$ of the λ repressor reorganizes 5′-CGC, whereas $Lys_{18}$ of GAL4 recognizes 5′-CGG.

The N-terminal arm represents a notable exception to the general failure of small, flexible peptides to bind DNA tightly and specifically. The origins of these effects may hold clues to the design of small-molecule DNA binding agents. The failure of small peptides to bind most likely results from an insufficient sum of thermodynamically favorable interactions to overcome the highly unfavorable loss of rotational and translational entropy that accompanies specific binding. Inspection of the X-ray structure of the λ repressor–$O_L1$ complex (Figure 1A) suggests show the N-terminal arm has overcome this entropic problem. The N-terminal core domain of the λ repressor, upon binding to the helix–turn–helix recognition site, presents the tethered N-terminal arm into the major groove at a high local concentration and in an orientation that is suitable for binding. Stated otherwise, the association of the core (H-T-H) domain with DNA substantially reduces the (translational) entropy cost for association of the N-terminal arm. The powerful advantage conferred by such cooperative binding in the case of the λ repressor may suggest strategies for design or screening of small, flexible DNA binding molecules. Specifically, stable binding of flexible molecules should be favored by their attachment to a (sequence-specific or nonspecific) DNA-anchoring segment such as a polycation or intercalator. A related though distinct strategy for engineering cooperative binding systems includes enforced dimerization of peptide–DNA binding elements (O'Neill et al., 1990; Talanian et al., 1990).[5]

Our finding that interference is undoubtedly stronger than the contact being interrupted suggests a great deal of cooperativity in the protein–DNA interaction. Such a high degree of cooperativity might be expected among the arm–DNA contacts, because they are so highly interconnected. For example, perturbation of the hydrophobic contact between C7 and $Lys_3$ would also be expected to destabilize the other contacts made by the same amino acid residue: C8, G8, and G9 (Figure 1B). What is less apparent is why perturbation of the arm–DNA contacts also seems to affect the H-T-H contacts, as evidenced by the lack of residual sequence-specific binding in the DNA molecules with $a^5$dC at C8 or C9. However, the arm and H-T-H binding elements are not structurally independent in the protein–DNA complex, since

$Lys_4$ and $Asn_{55}$ form a coordinate contact with base pair 6. This bridging interaction thus provides a link between the arm and H-T-H contact networks, allowing for direct communication between the two. The kind of negative cooperative effects caused by perturbation of arm contacts has previously been noted in methylation-interference footprints of the λ repressor and a variant lacking the N-terminal arm (Eliason et al., 1985).

Amino acid contacts to dC residues were statistically underrepresented in early crystallographic structures of protein–DNA complexes, which seemed to raise the possibility that C was somehow less competent than A, G, or T for making such contacts. That possibility has now been conclusively excluded with more recent findings suggesting that cytosine contacts are widespread, being present inter alia in *Eco*RI endonuclease (Kim et al., 1990; Heitman & Model, 1990), *Eco*RV endonuclease (Winkler, 1992, and personal communication), catabolite activator protein (Schultz et al., 1991), GAL4 (Marmorstein et al., 1992), and the subject of this study, the λ repressor (Clarke et al., 1991; Beamer & Pabo, 1992). It is interesting to note that of all crystallographically characterized hydrogen-bonding contacts to cytosines, a significant majority are of a single type, in which a main-chain amide carbonyl hydrogen bonds to the pyrimidine $N^4$-H function; this is even more surprising when one considers that analogous main-chain contacts to the $N^6$-H of A function have not been observed thus far. The results reported herein suggest that $a^5$dC is an effective probe of both hydrogen bonding and hydrophobic contacts to C and should thus be of general value in studies of protein–DNA interactions.

The axis of pseudodyad symmetry in $O_L1$ runs through a base pair (Figure 1C, dot) rather than between a base pair, which renders the operator (and all λ operators) incapable of perfect $C_2$ symmetry. Since the N-terminal arm contacts the guanine residue of base pair 9 (Figure 1B), which is present in only the consensus half-site, it follows that the consensus and nonconsensus arms cannot bind identically. In our TDI-C footprints, the lack of detectable interference at C8′, which is related by symmetry to the contact residue C8, suggests that the N-terminal arm binds much differently if at all in the nonconsensus half-site. This asymmetry in arm contacts results in thermodynamic asymmetry of the entire repressor–operator complex: Sarai and Takeda (1989) first noted that base substitutions in the consensus half-site generally have more deleterious effects on binding than corresponding mutations in the nonconsensus half-site, suggesting that the former is more tightly bound to the λ repressor than the latter. We have noted similar effects in TDI-G footprints (Hayashibara & Verdine, 1991) of the λ–$O_L1$ complex;[6] namely, interference by $N^7$-methylguanine is as much as 10 times stronger in the consensus half-site than in the nonconsensus. This asymmetry appears to be an intrinsic feature of the repressor–operator interaction, rather than a consequence of the nonconsensus base pairs in $O_L1$, since an operator symmetrized to possess two consensus half-sites (except base pair 9) binds the λ repressor more weakly than natural operators having nonconsensus residues (Sarai & Takeda, 1989). Taken together, these data suggest that the interaction of the N-terminal arm with consensus base pairs 6–9 preferentially stabilizes protein–DNA interactions within the entire consensus half-site.

In this study, we have shown that the nucleoside analog $a^5$dC, upon incorporation into DNA, can cause potent inhibition of sequence-specific binding by a transcriptional regulatory protein. Apart from its use here, $a^5$dC has been widely employed as a tool in cell biology, owing to its profound effects on cell differentiation [reviewed in Jones (1984) and Holliday (1987)]. These effects have generally been ascribed to the ability of $a^5$dC to inhibit DNA (cytosine-5) methyltransferases, the enzymes that mediate methylation of DNA (Jones, 1984; Santi et al., 1983, 1984; and references cited therein). Our results suggest that, in addition to inhibition of DNA methylation, $a^5$dC may exert some of its effects through another mechanism—inhibition of DNA binding by transcription factors. This latter activity could also help to account for some of the effects of $a^5$dC that seem unrelated to DNA methylation (Hori, 1983; Zimmerman & Scheel, 1984).

The results reported herein, combined with earlier reports (Hayashibara & Verdine, 1992; Pu & Struhl, 1992), now allow interactions of proteins with the major groove surface of three of the four DNA bases—G, C, and T—to be analyzed by TDI footprinting.

[6] K. C. Hayashibara and G. L. Verdine, unpublished results.

## ADDED IN PROOF

In recent work (K. C. Hayashibara, J. L. Mascareñas, and G. L. Verdine, unpublished results) the analogs 5-hydroxy-2′-deoxyuridine and 7-deaza-2′-deoxyadenosine have been used in TDI footprinting of T and A contacts, respectively. Thus, TDI footprinting of contacts to all four bases in DNA can now be performed.

## ACKNOWLEDGMENT

We thank Anna Astromoff and Mark Ptashne for providing pEA305, Lesa Beamer and Carl Pabo for providing X-ray coordinates of the 1.8-Å λ repressor–$O_L1$ complex prior to publication and for informative discussions, Martin Karplus and Polygen/Molecular Simulations for a grant to license the program QUANTA, and George Trainor for expert advice on nucleoside triphosphate synthesis.

## REFERENCES

Aggarwal, A. K., Rodgers, D. W., Drottar, M., Ptashne, M., & Harrison, S. C. (1988) *Science 242*, 899–907.

Amman, E., Brosius, J., & Ptashne, M. (1983) *Gene 25*, 167–178.

Beamer, L. J., & Pabo, C. O. (1992) *J. Mol. Biol. 227*, 177–196.

Brunelle, A., & Schleif, R. F. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 6673–6676.

Chodish, L. A. (1989) in *Current Protocols in Molecular Biology* (Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Smith, J. A., Seidman, J. G., & Struhl, K. Eds.) Vol. 1, pp 12.2.1–12.2.10, Wiley, New York.

Clarke, N. D., Beamer, L. J., Goldberg, H. R., Berkower, C., & Pabo, C. O. (1991) *Science 254*, 267–270.

Eliason, J. L., Weiss, M. A., & Ptashne, M. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 2339–2343.

Gyellenstein, U. B., & Erlich, H. A. (1988) *Proc. Natl. Acad. Sci. U.S.A. 85*, 7652–7656.

Ha, J.-H., Spolar, R., & Record, M. T., Jr. (1989) *J. Mol. Biol. 209*, 801–816.

Harrison, S. C., & Aggarwal, A. (1990) *Annu. Rev. Biochem. 59*, 933–969.

Hayashibara, K. C., & Verdine, G. L. (1991) *J. Am. Chem. Soc. 113*, 5104–5106.

Hayashibara, K. C., & Verdine, G. L. (1992) *Nucleic Acids Res.* (submitted for publication).

Hayes, J. J., & Tullius, T. D. (1989) *Biochemistry 28*, 9521–9527.

Hecht, M. H., Nelson, H. C. M., & Sauer, R. T. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 2676–2680.
Heitman, J., & Model, P. (1990) *EMBO J. 9*, 3369–3378.
Holliday, R. (1987) *Science 238*, 163–170.
Hori, T. A. (1983) *Mutat. Res. 121*, 47–52.
Johnson, A. D., Pabo, C. O., & Sauer, R. T. (1980) *Methods Enzymol. 65*, 839–856.
Jones, P. A. (1984) in *DNA Methylation: Biochemistry and Biological Significance* (Razin, A., Cedar, H., & Riggs, A. D., Eds.) pp 165–187, Springer-Verlag, New York.
Jones, P. A., & Taylor, S. M. (1980) *Cell 20*, 85–93.
Jordan, S., & Pabo, C. (1988) *Science 242*, 893–899.
Kim, Y., Grable, J. C., Love, R., Greene, P. J., & Rosenberg, J. M. (1990) *Science 249*, 1307–1309.
March, J. (1985) *Advanced Organic Chemistry*, 3rd ed., p 19, Wiley-Interscience, New York.
Marmorstein, R., Carey, M., Ptashne, M., & Harrison, S. C. (1992) *Nature 356*, 408–414.
Nelson, H. C. M., & Sauer, R. T. (1986) *J. Mol. Biol. 192*, 27–38.
O'Neill, K. T., Hoess, R. H., & DeGrado, W. F. (1990) *Science 249*, 774–778.
Pabo, C. O., Krovatin, W., Jeffrey, A., & Sauer, R. T. (1982) *Nature 298*, 441–443.
Piskala, A., & Sorm, F. (1964) *Collect. Czech. Chem. Commun. 29*, 2060–2076.
Piskala, A., & Sorm, F. (1978) in *Nucleic Acid Chemistry* (Townsend, L. B., & Tipson, R. S., Eds.) Part I, pp 443–449, Wiley, New York.
Pu, W. T., & Struhl, K. (1992) *Nucleic Acids Res. 20*, 771–775.
Riggs, A. D. (1983) *Adv. Cancer Res. 40*, 1–30.
Ruth, J. L., & Cheng, Y.-C. (1981) *Mol. Pharmacol. 20*, 415–422.
Santi, D. V., Garrett, C. E., & Barr, P. J. (1983) *Cell 33*, 9–10.
Santi, D. V., Norment, A., & Garrett, C. E. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 6993–6997.
Sarai, A., & Takeda, Y. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 6513–6517.
Sauer, R. T., Jordan, S. R., & Pabo, C. O. (1990) *Adv. Protein Chem. 40*, 1–61.
Schultz, S. C., Shields, G. C., & Steitz, T. A. (1991) *Science 253*, 1001–1007.
Siebenlist, U., & Gilbert, W. (1980) *Proc. Natl. Acad. Sci. U.S.A. 77*, 122–126.
Talanian, R. V., McKnight, C. J., & Kim, P. S. (1990) *Science 249*, 769–771.
Vogler, W. R., Miller, D. S., & Keller, J. W. (1976) *Blood 48*, 331–337.
Weiss, M. A., Eliason, J. L., & States, D. J. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 6019–6023.
Winkler, F. K. (1992) *Curr. Biol. 2*, 93–99.
Zimmerman, F. K., & Scheel, I. (1984) *Mutat. Res. 139*, 21–24.