

Extreme divergence between one-to-one orthologs: the structure of N15 Cro bound to operator DNA and its relationship to the λ Cro complex

Branwen M. Hall, Sue A. Roberts and Matthew H.J. Cordes*

Department of Chemistry and Biochemistry, University of Arizona, Tucson, AZ 85721, USA

Received October 08, 2018; Revised May 25, 2019; Editorial Decision May 28, 2019; Accepted June 01, 2019

ABSTRACT

The gene *cro* promotes lytic growth of phages through binding of Cro protein dimers to regulatory DNA sites. Most Cro proteins are one-to-one orthologs, yet their sequence, structure and binding site sequences are quite divergent across lambdoid phages. We report the cocrystal structure of bacteriophage N15 Cro with a symmetric consensus site. We contrast this complex with an orthologous structure from phage λ , which has a dissimilar binding site sequence and a Cro protein that is highly divergent in sequence, dimerization interface and protein fold. The N15 Cro complex has less DNA bending and smaller DNA-induced changes in protein structure. N15 Cro makes fewer direct contacts and hydrogen bonds to bases, relying mostly on water-mediated and Van der Waals contacts to recognize the sequence. The recognition helices of N15 Cro and λ Cro make mostly nonhomologous and nonanalogous contacts. Interface alignment scores show that half-site binding geometries of N15 Cro and λ Cro are less similar to each other than to distantly related CI repressors. Despite this divergence, the Cro family shows several code-like protein–DNA sequence covariations. In some cases, orthologous genes can achieve a similar biological function using very different specific molecular interactions.

INTRODUCTION

Cro and CI proteins are homodimeric, helix–turn–helix transcription factors with conserved regulatory functions in lambdoid bacteriophages. Cro and CI from phage λ are archetypes of transcriptional regulation, acting as antagonistic players in a genetic switch between phage lifestyles (1). The *cro* and *cI* genes occupy adjacent positions within the phage immunity region, part of which is shown in Figure 1. Cro and CI proteins bind as dimers to three imperfect palin-

dromes (O_{R1} , O_{R2} and O_{R3}) in the *cro-cI* intergenic region, along with sites in the nearby O_L region. At low concentration, Cro binds first to O_{R3} , while CI binds cooperatively to O_{R1} and O_{R2} using interdimer contacts made by a C-terminal domain not present in Cro. When bound to O_{R3} , Cro represses transcription of *cI*; when bound to O_{R1} and O_{R2} , CI represses transcription of *cro* and other genes, and activates transcription of *cI*. At higher concentrations, Cro or CI can occupy the other O_R sites, thereby downregulating transcription of their own and other genes. The predominance of CI or Cro correlates to a binary switch between lysogeny and lytic growth, respectively. This switch is conserved in all lambdoid bacteriophages.

Cro and the N-terminal domain of CI form two diverse orthologous lineages within a large, ancient superfamily of DNA-binding domains. The adjacent gene positions, common binding sites and shared binding motifs of *cro* and *cI* indicate that they are related by gene duplication and are thus paralogs (2–4). *Cro* and *cI* are single-copy genes, suggesting that each one may represent a single lineage consisting of a set of one-to-one orthologs from different phage (see below, however). Despite their homology, the Cro and CI N-terminal domains are extremely diverse in amino acid sequence, suggesting that the duplication is ancient and that both lineages have evolved extensively. In addition, although Cro (and CI) proteins retain characteristic biological roles and binding site locations, they have not retained a characteristic DNA-binding specificity. Instead, they have coevolved with poorly conserved operator sequences (Figure 1) (5–7) to produce lambdoid phages with diverse immunity specificities. This leads to an unusual situation in which the DNA-binding specificity of a given Cro or CI protein may be more similar to that of its intraspecific paralog, with which it competes, than to that of many of its orthologs. Ordinarily, one-to-one orthologs have tightly conserved functions (8).

The Cro proteins have a particularly dynamic evolutionary history that includes extensive divergence in sequence, structure and function (Figure 1). Indeed, they are not even all true orthologs of each other. For example, the apparent *cro* gene from bacteriophage 434 is the product of displace-

*To whom correspondence should be addressed. Tel: +520 626 1175; Fax: +520 626 9204; Email: cordes@email.arizona.edu

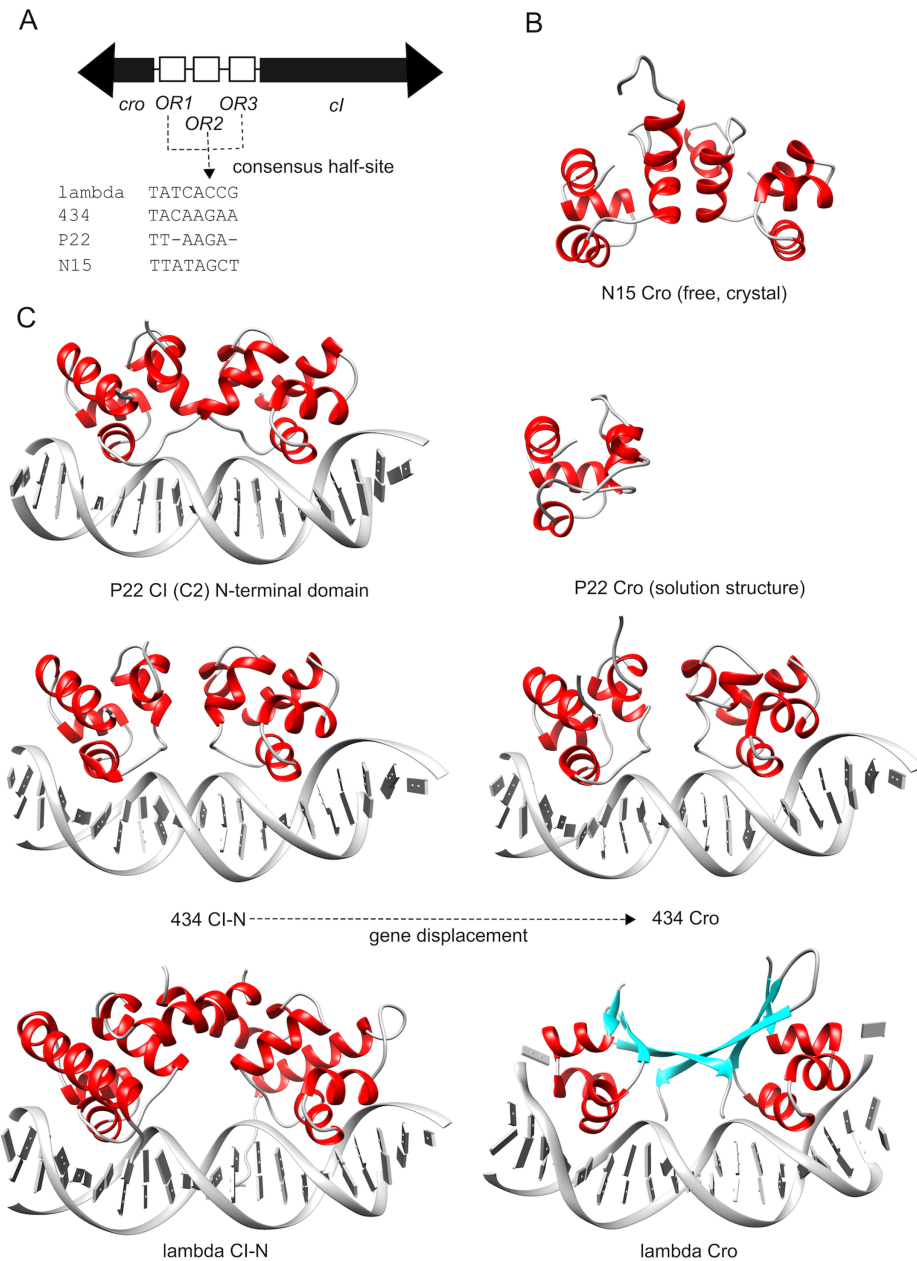


Figure 1. The Cro/CI superfamily. (A) Part of the immunity region of lambdoid phages including divergently transcribed *cro* and *cI* genes, which are related by an ancient gene duplication. Cro and CI proteins bind to three imperfectly palindromic O_R -binding sites in the intergenic region, each containing two similar half sites. Four diverse consensus O_R half-site sequences for different phages are shown. (B) Crystal structure of free N15 Cro, the subject of the current study. (C) Representative protein–DNA complexes for three CI N-terminal domains (P22, 434 and lambda; PDB ID: 2R1J, 2OR1 and 1LMB, respectively) and two Cro proteins (434 and lambda; PDB ID: 3CRO and 6CRO, respectively), one of which (434 Cro) is the product of gene displacement by a portion of the adjacent *cI* gene. Also shown is the solution structure of free P22 Cro (PDB ID: 1RZS).

ment by a duplicated adjacent *cI* DNA-binding domain, and is technically a paralog rather than an ortholog of other *cro* genes (3,4). Many other *cro* genes are true orthologs, but have diverged to the point where many of the proteins are related by transitive rather than direct amino-acid sequence similarity (3). Because O_R sequences (including the O_R3 site) are not well conserved, Cro proteins have diverse DNA-binding specificities. Most strikingly, the structures of orthologous Cro proteins have diverged into two different folds with totally different dimer interfaces (3,9,10).

The Cro orthologs from bacteriophages λ , P22 and N15 exemplify the diversity of the family (Figure 1). No two of these three proteins share any direct sequence similarity. Multiple lines of evidence nonetheless support their orthology (9), including conserved gene position, transitive sequence homology and PSI-BLAST connecting P22 Cro to λ Cro (3), and structural similarity and PSI-BLAST connecting N15 Cro to P22 Cro (9). N15 Cro and P22 Cro conserve the canonical repressor fold, shared by the distantly related CI proteins and composed of five or six α -helical elements

(Figure 1C) (3,9), but λ Cro has a mixed $\alpha+\beta$ fold (11). N15 Cro and λ Cro dimerize in crystals and in solution ($K_D \sim 3\text{--}5 \mu\text{M}$) (2,9,12–14), but P22 Cro shows no evidence for dimerization in the absence of DNA (3,15). The homodimer interfaces of N15 Cro and λ Cro have similar contact surface area and affinity but completely different architectures (9). Consensus O_R half-site sequences for the three phage conserve only two of eight base pairs (7) (Figure 1A). The symmetry axis that exchanges the two operator half sites coincides with a base pair in λ (16,17), but rests between two base pairs in N15 and P22 (18,19).

Despite this extreme diversity, Cro proteins obey a partial ‘evolutionary code’, in which changes in the sequence of the Cro recognition helix correlate to changes in the consensus sequence of the O_R half sites. A database of Cro proteins and cognate binding sites showed one-to-one sequence correlations between three positions in the recognition helix (H1, H3 and H6; a universal residue numbering system we will use henceforth) and three base pairs (2, 5 and 6, respectively) in the half site (7). The correlations are binary, wherein a switch between two amino acid residues relates to a switch between two recognized base pairs. The H1/+2 pairing features either Pro/Thy or Gln/Ade; the H3/+5 pairing features either Ala/Ade or (Ser/Thr)/Gua and the H6/+6 pairing features either Lys/Cyt or Gln/Gua. The three pairings in λ Cro correspond to direct protein–DNA contacts observed in the λ Cro–DNA complex (20). We previously used the code to re-engineer the functional specificity of λ Cro (21).

Protein–DNA cocrystal structures have been reported for four CI-like N-terminal domains (including 434 CI, P22 CI (C2), λ CI and 434 Cro), along with one true member of the Cro lineage (λ Cro) (Figure 1C) (20,22–28). No cocrystal structure has been reported for any true Cro ortholog with the helical repressor fold, though structures of free N15 Cro and P22 Cro have been solved using crystallography and NMR, respectively (3,9). P22 Cro is an attractive target for cocrystallization given the existing structure of the P22 CI (c2) repressor bound to DNA, but attempts to cocrystallize P22 Cro with DNA have not been successful. Here, we report the crystal structure of a complex of N15 Cro with a symmetrical consensus O_R site at 1.6 Å resolution. This structure both allows a comparison of DNA binding function by highly divergent Cro orthologs and serves as a source of information on how alternative residue pairings may modulate specificity under the proposed evolutionary code.

MATERIALS AND METHODS

Preparation of protein and DNA samples

N15 Cro was overexpressed and purified as described previously (9) and flash frozen in 500 μl aliquots in 10 mM Tris (pH 8.5). Single aliquots were rapidly defrosted in warm water and then centrifuged at top speed in a microcentrifuge for 20 min. Protein concentration was measured using an estimated extinction coefficient of $7953 \text{ M}^{-1}\text{cm}^{-1}$. For crystallization, HPLC purified DNA with the sequence 5′-TTTATAGCTAGCTATAA-3′ was obtained from Eurogentec (San Diego, CA). DNA was resuspended in a

small volume of buffer (10 mM Tris [pH 8.5]) and annealed in a thermal cycler over ~ 8 h by gradual temperature reduction from 80 to 15°C, to form 16 base pairs of duplex corresponding to the symmetrical consensus O_R site (7,18) plus a one base 5′ thymine overhang on each strand. For electrophoretic mobility shift assays of binding to the O_{R3} site, two oligonucleotides (5′-GCAAAATTATAGCCAGCTATAAAGAGCG-3′ and its reverse complement) were obtained from Integrated DNA Technologies (Coralville, IA) and purified by urea-denatured polyacrylamide gel electrophoresis. End-labeling of one strand with ^{32}P and annealing of the two strands to form a duplex were then performed as described previously for λ Cro (21), resulting in a 28 base-pair singly ^{32}P -labeled duplex consisting of the 16 base-pair N15 O_{R3} site flanked at each end by four base pairs of the natural flanking DNA sequence (18), and then further flanked at each end by two GC base pairs. O_{R2} and O_{R1} sites were constructed similarly using the same flanking DNA as the O_{R3} site.

Electrophoretic mobility shift assay

Electrophoretic mobility shift assays for N15 Cro were performed and imaged as described for λ Cro (21). At each concentration of Cro protein, the apparent fraction of DNA bound by Cro was measured as the intensity of the band representing the bound form divided by the total intensity measured for that lane of the gel. The fraction bound was plotted against added protein concentration and subjected to nonlinear least squares fitting in Kaleidagraph, using an equation that assumes a large excess of free protein over DNA, and a two-state equilibrium between free Cro monomers and bound Cro dimers. Because the fraction bound leveled off at value less than one, the maximal fraction bound was not fixed at one, but was allowed to vary as a parameter in the fit. Due to the dimer stoichiometry, the equilibrium dissociation constants derived from these fits correspond to the square of the protein concentration giving half-maximal DNA binding (i.e. an apparent K_D or $K_{D,app}$). The reported K_D value in the text represents $K_{D,app}$ and was calculated from the mean of four independent measurements. The uncertainty is reported as the standard error of the mean. As an additional note, measured O_{R3} affinities for untagged and C-terminally hexahistidine tagged N15 Cro were indistinguishable ($K_{D,app} = 90 \pm 7 \text{ nM}$ and $91 \pm 11 \text{ nM}$, respectively).

Crystallization of N15 Cro/consensus operator DNA complex

N15 Cro and DNA were combined in a 2:1.2 molar ratio (corresponding to a slight excess of DNA over protein based on a 2:1 protein:DNA binding stoichiometry) and allowed to equilibrate 30 min. The concentration of protein/DNA complex used for crystallization was 14–15 mg/ml. The complex was crystallized using the hanging-drop method, with drops composed of 2 μl of complex and 2 μl of mother liquor (0.4 M monobasic ammonium phosphate). After ~ 1 week, crystals grew as single or clustered rods. Crystals were tetragonal with space group $P4_12_12$, and the asymmetric unit contained one N15 Cro dimer complexed to a single DNA duplex.

Crystal structure determination

Crystals were flash-frozen in liquid nitrogen after being gradually transferred to fresh reservoir solution that contained mother liquor plus 15% glycerol. Data were collected on crystals cooled at 100 K at SSRL beamline 9–2 using a MAR 325 detector. Diffraction images were processed with Mosflm (29) and scaled with Scala (30). Diffraction was highly anisotropic, with a ΔB value of 33.43 Å². To compensate, ellipsoidal truncation and anisotropic scaling were carried out using the Diffraction Anisotropy Server (31). A molecular replacement solution was identified by Molrep (32) using a previously solved crystal structure of the N15 Cro dimer (PDB ID: 2HIN) (9). DNA was manually built into the electron density map using COOT (33). Refinement was carried out with Refmac5 (34) and manual rebuilding using COOT. The data were then later reprocessed using iMosflm and scaled using AIMLESS. The data were submitted to the STARANISO (35) server that produced an ellipsoidally fitted dataset with resolution 1.6–2.17 Å. R_{free} flags were imported from the previously used reflection dataset and generated for reflections outside the resolution limit of the previous data set. The structure was refined and rebuilt using Refmac and COOT. Final refinement included refinement of TLS groups with the optimum TLS groups determined using the TLS motion determination server (36,37). Most programs were accessed through the CCP4 suite (38). Data measurement and refinement statistics are given in Table 1.

Structural analysis

Superposition of proteins and protein–DNA complexes was performed using UCSF Chimera, and superposition of protein–DNA complexes was performed using CLICK (39). Protein–DNA contact surface was evaluated using Swiss-PDB Viewer (40). DNA structure was analyzed with CURVES 5.1 (41,42). Single base overhangs were excluded to avoid any influence on the calculation of the global helical axis. Global rather than local parameters were reported, where the latter measure relationships between base pairs without reference to the overall conformation of the DNA.

The binding geometry of N15 Cro with DNA was compared to that of other Cro and CI repressor complexes using Protein-DNA Interface Alignment Software (43) as well as by superposition of half-site complexes using CLICK (39). All Cro and CI structures were divided into half-complexes for comparative purposes, and only the conserved helix-turn-helix portion of the protein structure was included. The Protein-DNA Interface Alignment Software quantifies and compares relative binding geometries using a procedure based on work by Pabo and Nekludova (44). Individual amino acid side chains and bases are first assigned local coordinate systems. The spatial similarity between interacting coordinate systems (or interacting pairs) in each complex is ranked with a similarity score $S(i,j)$, where higher scores indicate more similar geometric relationships. Nucleotides are considered to interact with amino acid side chains if a vector connecting the origins of their coordinate systems has a magnitude <16 Å. The sum of similarity scores for aligned pairs in the two structures gives an overall interface

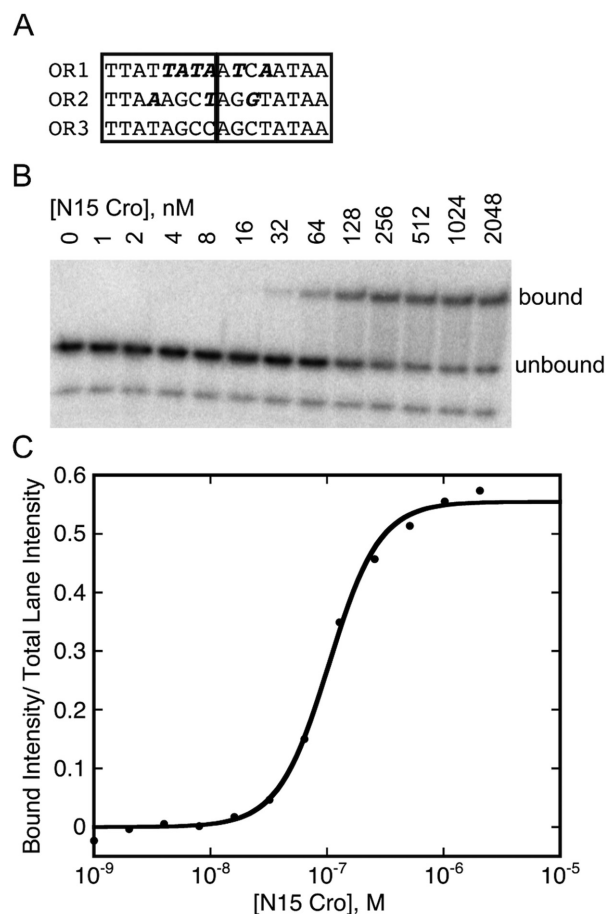


Figure 2. N15 Cro binds operator DNA. (A) Right operator sequences of bacteriophage N15, with the two half sites outlined in boxes. Positions where OR1 and OR2 sequences differ from that of the preferred OR3 site are in bold italic. (B) Representative electrophoretic mobility shift assay using a 28 bp ³²P end-labeled duplex containing the OR3 site from the N15 right operator region. Protein and DNA were incubated for 30 min at ambient temperature in KP200 buffer (20 mM KPO₄ [pH 7], 200 mM KCl, 1 mM EDTA and 5% glycerol) plus 150 μg/ml bovine serum albumin, then loaded onto a 10% native polyacrylamide gel running at 250 V at 4°C (21). Comparable experiments with OR1 and OR2 showed no shifted DNA. (C) Fitting of the electrophoretic mobility shift assay data to a model in which free N15 Cro is exclusively monomeric and DNA-bound N15 Cro is exclusively dimeric. Based on previous assays N15 Cro dimerizes with a K_D of 5 μM (9).

alignment score (IAS). The method examines spatial relationships without taking into account secondary structure or primary sequence, and has been used to describe clusters of similar binding geometries that exist both within and between different DNA-binding protein families.

RESULTS

DNA-binding affinity

Purified recombinant N15 Cro is functional and binds to its predicted cognate DNA sequence with nanomolar affinity. N15 Cro binds to OR3, generally the highest affinity site for Cro, with an apparent $K_D = 90 \pm 7$ nM in an electrophoretic mobility shift assay (EMSA) (Figure 2). We also tested binding to OR2 and OR1 sites and observed no bind-

Table 1. Crystallographic data for N15 Cro/DNA complex

Crystal preparation			
Conditions	0.4 M ammonium phosphate, monobasic		
Cryoprotectant	15% glycerol		
Space group	$P4_12_12$		
Unit cell (Å)	$a = b = 56.56, c = 195.76$		
V_M^a	3.21		
Data collection and reduction ^{b,c}			
X-ray source	SSRL beamline 9–2		
Wavelength (Å)	0.97946		
Resolution ^d	28.7–1.60	(1.63–1.60)	[1.87–1.83]
Observed reflections	540 896		
Unique reflections	42 733		
Average redundancy	12.7	(4.3)	[11.7]
Completeness (%)	98.7	(85.3)	[100.]
R_{merge}^e	0.066	(2.8)	[1.17]
$I/\sigma(I)$	13.2	(0.3)	[1.8]
R_{pim}^f	0.019	(1.7)	[0.35]
CC1/2	1.000	(0.214)	[0.942]
Wilson plot, B factor	42.2		
Anisotropic correction			
Ellipsoid cut-off surfaced	1.60–2.17		
Resolution ^d	28.7–1.60	(1.78–1.60)	
Completeness (%)	99.6	(94.4)	
$I/\sigma(I)$	15.7	(3.08)	
Structure refinement			
R_{cryst}^b	0.2058		
R_{free}^b	0.2172		
RMSD bonds (Å)	0.0109		
RMSD angles (°)	1.871		
Average B , protein	25.6		
DNA	32.1		
Water	32.0		

^a V_M : Matthews coefficient.

^bData reduction statistics before anisotropic correction.

^cOverall, (highest resolution shell), [shell for which $I/\sigma(I) \geq 2$].

^dHighest and lowest resolution of ellipsoidally corrected data.

^e R_{merge} and R_{pim} formulas can be found at <https://strucbio.biologie.uni-kostanz.de/ccp4wiki/index.php/R-factors>.

ing by EMSA up to $\sim 1 \mu\text{M}$ protein. For comparison, ϕKO2 Cro, which has $\sim 90\%$ sequence identity with N15 Cro and an identical cognate O_R3 sequence, binds O_R3 with apparent $K_D = 44 \text{ nM}$ by EMSA under somewhat different conditions, and binds ϕKO2 O_R2 and O_R1 very weakly ($K_D \sim 1\text{--}2 \mu\text{M}$) (45). The cognate affinities of N15/ ϕKO2 Cro for O_R3 DNA are both weaker than that of λ Cro, which binds λ O_R3 with apparent $K_D = 3.6 \pm 0.5 \text{ nM}$ in under EMSA conditions comparable to those used here for N15 Cro (21).

Determination of N15 Cro-operator DNA cocrystal structure

N15 Cro cocrystallized with a 16 base-pair symmetrical consensus O_R duplex that also contained a single thymine overhang at each 5' end. Sixteen base pairs represent the minimal binding site size found for ϕKO2 Cro (45). The symmetrical site differs from O_R3 by a single base pair near the center (see Figure 1), which is not contacted directly by the protein (see below). ϕKO2 Cro binds both O_R3 and this symmetrical site in EMSA experiments (45). The symmetrical site also represents the consensus sequence of the six O_R half sites of N15 (7).

We determined the structure of the N15 Cro-operator DNA complex in space group $P4_12_12$, using molecular replacement with the structure of free N15 Cro (Table 1

and Figure 3A). Representative electron density is shown in Supplementary Figure S1, including a portion of the protein/DNA-binding interface and several ordered water molecules. The asymmetric unit contains a dimer of N15 Cro, with each subunit bound to one half site of the symmetrical operator DNA. We modeled residues 1–62 for one N15 subunit (chain A) and residues 1–63 for the other (chain B), while residues 64–71 appear disordered in both chains. In the structure of free N15 Cro, approximately the same C-terminal region is disordered (9). The refined structure includes 277 modeled water molecules in the asymmetric unit, including many ordered water molecules at the protein–DNA interface. The complex has a noncrystallographic pseudo-twofold axis relating the two N15 subunits in the dimer and the two DNA half sites. The two halves align with RMSD = 0.23 Å for all atoms, excluding water molecules and residue 63 of N15 Cro, which was only present in chain B.

Features of the N15 Cro–DNA complex and comparison to λ Cro

N15 Cro shows no evidence of major induced-fit changes in protein structure upon DNA binding. The structure of the DNA-bound N15 Cro dimer is very similar to that of the

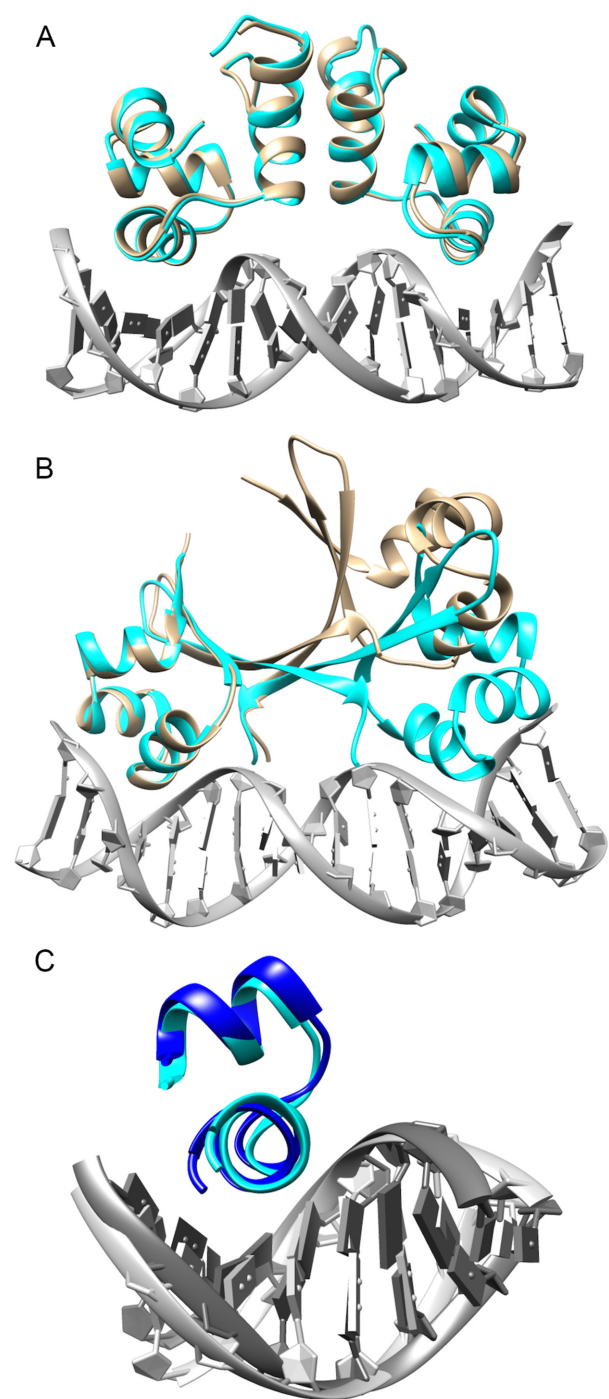


Figure 3. Comparison of N15 Cro and λ Cro and DNA complexes. (A) N15 Cro free (tan) or bound (cyan) to symmetric consensus operator DNA (gray). (B) λ Cro with same color scheme as in panel (A). (C) Superposition of N15 (dark blue and gray) and λ (light blue and gray) half-site complexes with only the helix-turn-helix motif of the protein included.

free dimer (backbone RMSD = 0.6 Å, residues 1–62; Figure 3A). The DNA-bound and free crystal structures of the wild-type λ Cro dimer, meanwhile, are quite different (backbone RMSD = 4.0–4.1 Å, residues 4–60; Figure 3B). Note, however, that free λ Cro variant structures differ widely and the free wild-type crystal structure may not represent the

lowest energy solution structure (46–49). Overall, the N15 Cro dimer appears relatively rigid, while the λ Cro dimer is clearly flexible, owing in part to a ‘ball-and-socket’ joint in the dimer interface (11,20).

Both N15 and λ operator duplexes show some deviations from canonical B-form geometry in the Cro complexes, suggesting that the protein induces changes in DNA structure (20). Both structures show DNA bending, but the duplex is significantly more bent in the λ Cro than in the N15 Cro complex (39 versus 25 degrees, respectively; Figure 3A versus Figure 3B). Both complexes show the largest (and most negative) values of ‘roll’ between adjacent base pairs at the center of the operator, indicating that the DNA bends most strongly near the dyad axis of symmetry (Supplementary Tables S1 and S2). Base pairs in the N15 operator on average show greater deviations from planarity, as measured by propeller twist, than those in the λ operator (Supplementary Tables S1 and S2). Overall, however, changes in both DNA and protein structure upon binding appear larger in the λ Cro complex.

The recognition helices of N15 Cro and λ Cro rest at rather different angles in the major groove with respect to the overall DNA double helix structure (compare Figure 3A and B). However, the geometry of major groove recognition is less different if one removes the more global DNA and protein deformations by superimposing only the half-complexes (protein/DNA backbone RMSD of 1.5 Å; Figure 3C). We return to the half-complex structures below in the context of a more extensive comparison between all Cro and CI complexes.

Protein–DNA contacts

N15 Cro makes relatively few direct contacts to nucleotide bases compared to λ Cro, contacting only five base pairs (Thy +1, Thy +2, Thy +4, Thy -5 and Cyt -6) compared to seven in λ Cro (Figure 4). N15 Cro makes no direct contacts to base pairs 3 and 7, while λ Cro recognizes base pair 3 using a Van der Waals contact from Asn H4 (recognition helix numbering), and base pair 7 with a hydrogen bond from Lys H6. Gln H6 makes the only two direct hydrogen bonds to DNA bases seen in the N15 Cro complex (both to Thy bases), while three side chains in λ Cro make a total of seven hydrogen bonds to four different bases. N15 Cro relies more heavily on Van der Waals contacts as well as water-mediated contacts in base recognition (see Figure 4; see also Figures 5 and 6 below). The overall contact surface area between protein and DNA is 432 Å² for the N15 complex compared to 708 Å² for the λ Cro complex.

Most base contacts in the N15 complex occur between positions on the protein and DNA that do not correspond to aligned positions in the λ complex (Figure 4). Both N15 and λ have a Van der Waals contact between Ala H3 of the recognition helix and the methyl group of Thy at -5 in the half site (Figures 4 and 6). In both proteins, the side chain at H6 (Lys or Gln) contacts the base at -6 (Gua or Cyt). However, in N15 Cro Gln H6 barely contacts Cyt at -6 (3.4–3.5 Å distance between the side chain oxygen of Gln H6 and C5 of Cyt -6) while also making hydrogen bond contacts to base pairs 4 and 5 (Figures 4 and 5). In λ Cro Lys H6 makes a hydrogen bond not only to Gua at -6, but also to

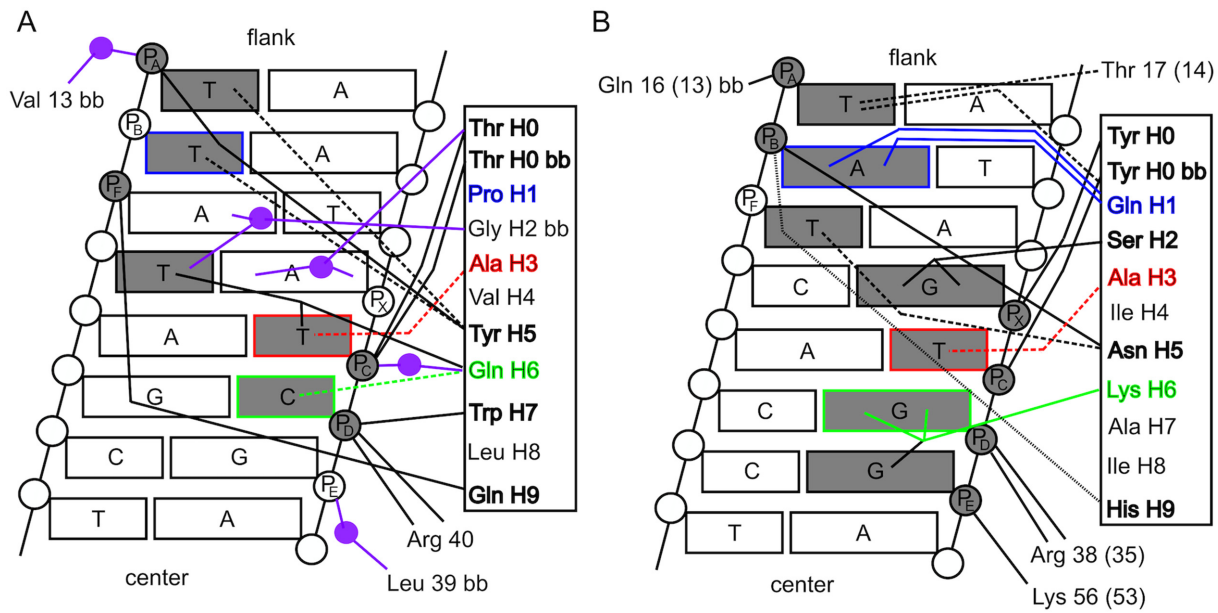


Figure 4. Protein–DNA contact maps. (A) N15 Cro and (B) λ Cro complexes with DNA. One DNA half site is shown in each case, with directly contacted bases and phosphate groups shaded gray, and phosphate nomenclature based on that of reference (20). The recognition helix sequence of each protein is contained in a box, with residues numbered H0 to H9 beginning with the N-cap. Hydrogen bond contacts are shown with solid lines, Van der Waals contacts with broken lines and presumed electrostatic interactions with finely dashed lines. Residues and bases with sequence covariations according to the previously described Cro evolutionary code are colored blue, red and green. Additional contacts made by residues outside the recognition helix are also shown, with absolute residue numbers shown for both proteins, but aligned N15 residue numbers given in parentheses in panel (B). Selected water-mediated contacts discussed in the text are shown in purple. The water-mediated contact labeled as originating from Gly H2 is part a network of two to three water molecules (depending on which half site is considered) that also hydrogen bonds to the side chain of Gln H9 (see also Figure 6).

base pair 7, while direct contacts to base pairs 4 and 5 are made by Ser at H2. Base pairs 1 and 2 in the λ consensus site are contacted by Gln H1 and Thr 17 from helix 2, while the same positions in the N15 consensus site are recognized by Van der Waals contacts from Tyr H5 to methyl groups on thymine bases (Figures 4 and 6). In summary, N15 Cro and λ Cro exhibit very different patterns of direct contacts to DNA bases, despite a qualitatively similar positioning of the recognition helix in the major groove (Figure 3C).

N15 Cro makes several water-mediated contacts to bases, at least one of which may contribute to sequence specificity. First, there is a gap between the protein and DNA near base pair 3 due in part to the lack of a side chain for Gly H2 (Figure 6). Networks of ordered water molecules fill this gap. One chain of two to three water molecules (depending on half site) extends from the backbone of Gly H2 to the side chain of Gln H9, while also hydrogen bonding to the Ade base at +3 and the Thy base at +4. Second, Thr H0, which acts as the N-cap on the recognition helix, makes hydrogen bonds to both a phosphate oxygen and a water molecule that is positioned to hydrogen bond to Ade -4 (Figures 5 and 6). The water molecule occupies part of the space left by the lack of a side chain for Gly H2. Thr H0 donates its hydrogen to the phosphate oxygen, and therefore must act as a hydrogen bond acceptor toward the water, engaging one of its hydrogen atoms. The other hydrogen in the water molecule would then be donated to N7 of Ade -4, while the oxygen atom of the water molecule acts as an acceptor toward the Ade NH₂ group. Importantly, the water molecule is oriented by Thr H0 to interact as both a donor and acceptor toward Ade, but would be unable to interact equiv-

alently with Gua, which has two hydrogen bond acceptors adjacent to the water molecule's position. We propose that Thr H0 (23) and Gly H2 (25), along with Gln H6, which makes a hydrogen bond to Thy at +4, collude to specify base pair 4. The involvement of water-mediated contacts in DNA recognition by λ Cro is difficult to compare to that of N15 Cro due to the smaller number of modeled waters (14 versus 277) and lower resolution of the λ structure (3.0 Å versus 1.6 Å).

N15 Cro makes the same total number of hydrogen bonds (7) to backbone phosphate groups as λ Cro (see Figure 4), but in a different pattern. N15 Cro concentrates most of its contacts at two phosphates and contacts a smaller total number of phosphates (four versus six for λ Cro). At least one contact is completely homologous in the two complexes, between the backbone of residue H0, at the N terminus of the recognition helix, and the phosphate labeled P_C. But most phosphate contacts in the two complexes are not between homologous positions. In λ Cro, there is a direct contact between the protein backbone at the N terminus of helix 2 (Gln 16) and the first phosphate at the flank of the binding site (P_A). This contact, which anchors the N-terminal end of the helix-turn-helix motif, is formally absent in N15 Cro, and supplanted by a contact from the side chain of Tyr H5. Nonetheless, the backbone of helix 2 in N15 Cro does make a water-mediated contact to this phosphate group (see Figure 4). Similarly, Lys 56 in λ Cro contacts phosphate P_E near the center of the operator, while N15 Cro does not contact this phosphate directly but instead utilizes a water-mediated contact from the backbone amide group of Leu 39 (see Figure 4).

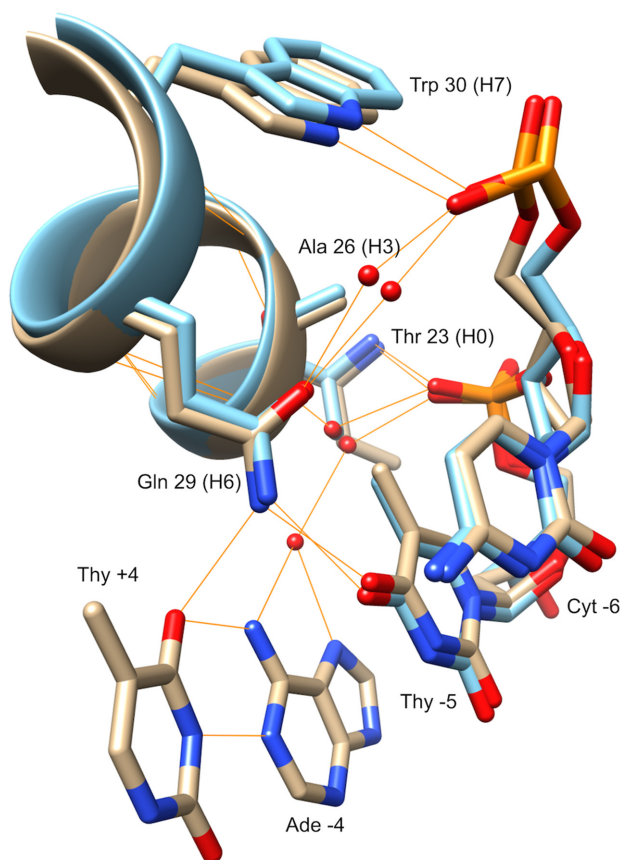


Figure 5. Recognition of base pairs 4 through 6. The N15 Cro complex (tan) includes both direct and water-mediated contacts to base pairs 4 through 6. Direct base contacts include two hydrogen bond contacts from Gln H6 to Thy +4 and Thy -5, Van der Waals contact between Ala H3 and Thy -5, and Van der Waals contact between Gln H6 and Cyt -6. A water molecule is oriented by Thr H0 to interact with Ade -4. The P22 CI (c2) complex with operator DNA (light blue) has precisely the same pattern of contacts to base pairs 5 and 6, including direct and water-mediated contacts to the backbone from Thr H0, Gln H6 and Trp H7.

Comparison of Cro and CI half-complexes

A broader comparison of Cro/CI binding geometry, including complexes for three CI N-terminal domains in addition to the two Cro cocrystal structures, reinforces the impression of dissimilarity between the two Cro complexes (Table 2). To make a universally applicable comparison, we used only partial half complexes consisting of the helix-turn-helix motif of one protein chain and its associated eight base-pair DNA half site. For two measures of similarity between the binding geometry of different complexes, we used interface alignment scores (IAS) and combined protein/DNA backbone RMSD from a best-fit superposition. The highest IAS and lowest RMSD occurred in comparisons of different halves of the same complex (IAS: mean 150 out of 190 possible, s.d. 50; RMSD: mean 0.5 Å, s.d. 0.3 Å), while the lowest IAS and highest RMSD occurred between N15 Cro and λ Cro half complexes (IAS: mean 30, s.d. 4; RMSD 1.5 Å). Similarities between half complexes of different CI proteins were on average higher than those between the Cro half complexes (IAS: mean 88, s.d.

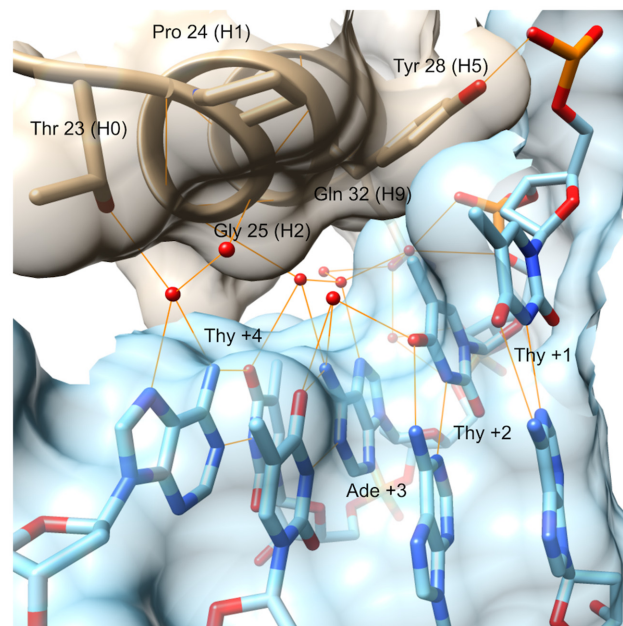


Figure 6. Recognition of base pairs 1 through 3. N15 Cro (tan; chain A) makes only Van der Waals contacts from Tyr H5 to base pairs 1 and 2 in the DNA (light blue). Pro H1, which is correlated to Thy +2 in the Cro evolutionary code, does not contact DNA at all but could help position Tyr H5. N15 Cro makes no direct specific contact to base pair 3. A gap near Gly H2 is filled by ordered water molecules that interact with base pairs 2, 3 and 4, including a network that makes hydrogen bonds with Ade +3/Thy +4 and a water molecule oriented by Thr H0 to interact with Ade -4. This gap may account for the smaller contact surface area between protein and DNA in the N15 Cro complex than the λ Cro complex. Note that the water network in this gap is slightly different (though similar) in the other half of the complex (N15 chain B; not shown).

20; RMSD: mean 1.0 Å, s.d. 0.2 Å), with the caveat that these comparisons involve different phage sets. In comparisons between Cro and CI proteins, N15 Cro has mostly higher IAS scores with the CI proteins (IAS: mean 72, s.d. 18) than λ Cro does (IAS: mean 42, s.d. 9), but the same average RMSD (mean 1.3 Å; s.d. < 0.1 Å). The N15 Cro complex has particularly high IAS with P22 CI (93–95), and the two proteins make the same contacts to base pairs 5 and 6 of the half site (see Figure 5). We conclude that N15 Cro and λ Cro half complexes are quite geometrically divergent from each other, more so than either one is from its distant CI cousins; and that λ Cro may have diverged more strongly than N15 Cro from the CI lineage, possibly related to loss of the ancestral repressor fold in λ Cro.

Insights into the proposed Cro evolutionary code

The previously proposed Cro evolutionary code (7,21) and the relevant amino-acid residues and bases in the N15 and λ Cro complexes are shown in Figure 7. At position H3 and base pair 5, Ala H3 correlates to Ade +5 while Ser/Thr H3 correlates to Gua +5. λ Cro and N15 Cro conserve a nearly identical relationship between the methyl group of Ala H3 in the recognition helix and the methyl group of Thy -5, though the distance is closer in N15 Cro. The alternate Ser/Thr H3 to Gua +5 correlation has yet to be observed in a cocrystal structure between a true Cro lineage mem-

Table 2. Interface Alignment Scores (IAS) and RMSD between Cro/CI complexes

	N15 Cro A	N15 Cro B	λ Cro	λ CI 3 _{nc}	λ CI 4 _c	434 CI R _{nc}	434 CI L _c	P22 CI R _{nc}	P22 CI L _c
N15 Cro	–	188	33	65	78	55	52	93	95
A		100	63	79	84	58	68	79	79
N15 Cro	0.4	–	27	62	72	60	45	94	95
B	100		58	74	89	68	74	79	79
λ Cro	1.5	1.5	–	56	45	44	33	37	34
	84	86		84	79	63	58	74	74
λ CI	1.2	1.2	1.3	–	140	70	95	110	112
3 _{nc}	99	100	97		100	79	89	100	100
λ CI	1.5	1.3	1.3	0.9	–	46	60	85	93
4 _c	98	95	94	97		68	84	100	100
434 CI	1.2	1.3	1.3	0.9	1.2	–	79	87	87
R _{nc}	91	97	99	100	94		79	89	89
434 CI	1.3	1.4	1.3	0.9	1.1	0.6	–	103	103
L _c	97	98	96	100	95	100		100	100
P22 CI	1.2	1.2	1.2	0.8	1.0	1.0	0.9	–	190
R	96	96	97	100	96	100	100		100
P22 CI	1.3	1.3	1.3	0.8	1.3	0.9	0.9	0.2	–
L	99	97	97	100	100	100	100	100	

Interface Alignment Scores (upper right; out of 190 possible, with scores listed above in bold and percentage of matched helix-turn-helix residues below) and backbone RMSD (lower left; with RMSD listed above in bold and percentage of aligned backbone atoms below) are shown for the 19-residue helix-turn-helix motif in a given Cro/cI protein chain in association with a cognate eight base-pair DNA half site: N15 Cro (PDB ID: 6ON0) chain A or B bound to consensus DNA; λ Cro (PDB ID: 6CRO) bound to consensus DNA; λ cI repressor (PDB ID: 1LMB) chain 3 or 4 bound to O_L1; 434 CI repressor (PDB ID: 2OR1) chain R or L bound to O_R1, and P22 CI (c2) repressor (PDB ID: 2R1J) chain R or L bound to a synthetic operator sequence. In the case of λ and 434 one half site in each complex corresponds to a consensus (c) sequence and the other to a nonconsensus (nc) sequence. Backbone RMSD was computed using N, CA and C atoms in the protein and C3', C5' and P atoms in the DNA.

ber and operator DNA. At position H6, Lys H6 correlates to Cyt +6 while Gln H6 correlates to Gua +6. λ Cro has a Lys residue at H6 that makes hydrogen bonding contacts to Gua -6, as well as Gua -7. Gln H6 of N15 Cro makes hydrogen bonds to Thy +4 and Thy -5, but also makes a limited Van der Waals interaction with Cyt -6 and a water-mediated contact to the phosphate backbone at base pair 6. Similar contact patterns to base pairs 5 and 6 are present in the consensus half sites of complexes of 434 CI with O_R1 and O_R3, as well as both halves of the complex of P22 CI with a synthetic operator (comparison to one of the half sites in the P22 CI complex is shown in Figure 5). The other halves of the 434 CI complexes (called the nonconsensus halves) lack Cyt at -6 and also lack the equivalent water-mediated contact from Gln H6 to the backbone. Koudelka *et al.* have proposed that Gln H6 in the P22 CI complex influences the specificity at base pair 6 in part through solvent organization and shape complementarity (27). If correct, this argument presumably applies to N15 Cro and its similar contact pattern. At position H1 and base pair 2, Pro H1 correlates to Thy +2 while Gln H1 correlates to Ade +2. Gln H1 of λ Cro shows a pair of hydrogen bonding contacts to Ade +2. Pro H1 of N15 Cro does not contact Thy +2 and in fact does not interact with DNA at all. Instead, Tyr H5 contacts the methyl group of Thy +2 (Figure 4). Pro H1 does make contact with Tyr H5, and it is possible that Pro H1 helps position it and may contribute to a general nonpolar surface that favors Thy at +2. On the whole, although it is possible to rationalize some of the observed sequence correlations based on the N15 and λ Cro complexes, it seems unlikely that one would propose all three binary one-to-one relationships *a priori*, if one had only these two divergent structures to go on.

Structural basis of the preference of N15 Cro for the O_R3 site

Apart from providing a comparison to DNA recognition by λ Cro, the N15 Cro-DNA structure suggests a basis for preferential recognition of the bacteriophage N15 O_R3 site over O_R2 or O_R1. O_R2 and O_R1 both differ from O_R3 by at least one directly contacted base pair in each half site: base pair 4 or 6 in the case of O_R2, and base pairs 5 (both half sites) and 6 (one half site) in the case of O_R1 (see Figure 2A). Substitutions in three of four half sites of O_R2 and O_R1, at either base pair 4 or 5, convert an A-T base pair to a T-A base pair or vice versa. These changes effectively place a hydrogen bond donor from an adenine base in the position of one of the pair thymine hydrogen bond acceptor groups recognized by the NH₂ group of Gln H6 (see Figure 5). Conversion of A-T to T-A at position 5 also disrupts a contact between Ala H3 and the methyl group of Thy -5 (see Figure 5). At position -6, on the basis of the 434 CI-DNA complexes discussed above, bases other than Cyt may lead to loss of a water-mediated contact between Gln H6 and the backbone. We suggest that the contact patterns made by Ala H3 and Gln H6 to base pairs 4–6 (see Figure 5) are the primary basis for specific recognition of O_R3. A water-mediated contact involving Thr H0, as discussed above, probably also contributes to specification of base pair 4. Finally, it should not be discounted that indirect readout mechanisms, possibly involving differences in noncontacted base pairs 7 and 8 near the center of the site (see Figure 2A) may contribute to specific recognition.

DISCUSSION

N15 Cro and λ Cro are distant orthologs that have evolved to achieve a similar biological result using very different interactions at the detailed molecular level. The two proteins

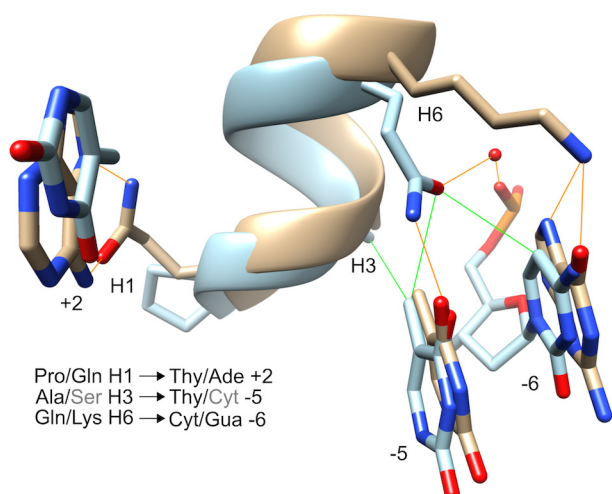


Figure 7. The Cro evolutionary code as represented in N15 Cro and λ Cro complexes with DNA. The proposed Cro evolutionary code is shown at lower left. The λ Cro complex (tan) has direct protein–DNA contacts corresponding to each of the three code pairings: Gln H1 makes hydrogen bond contacts (orange) to Ade +2; Ala H3 is in Van der Waals contact with the methyl of Thy -5 (though only the backbone, not shown); and Lys H6 makes hydrogen bond contacts to Gua at -6. In the N15 Cro complex (light blue), the Ala H3/Thy -5 contact is conserved, and includes Van der Waals contact (green) from the Ala side chain. The alternate code pairing of Ser H3/Cyt -5 (gray) is not present in any Cro–DNA structure. For the other two position pairs, the N15 Cro complex features the alternate residue pairings in the code. Pro H1 does not contact Thy +2, leaving this sequence correlation mysterious. Gln H6 makes closer contact with base pair 5 but does make Van der Waals contact with Cyt -6 (green) as well as a water-mediated contact to the phosphate group (see text for discussion).

play similar biological roles in promoting lytic growth (50–54). Both bind O_R3 and N15 Cro shares λ Cro’s preference for O_R3 over other O_R sites (45). But N15 Cro and λ Cro lack significant sequence similarity, adopt different folds and form completely different homodimer interfaces (9,11). The O_R sites to which they bind have different half-site sequences and spacing. The N15 Cro and λ Cro protein–DNA complexes compared here show different levels of apparent induced fit, very different contact patterns and have the most divergent binding geometries among a set of pairwise comparisons that includes both Cro complexes and those of the distantly related CI paralogs.

In light of this divergence, it may seem surprising that a broad database of Cro protein and cognate DNA sequences yielded a partial ‘evolutionary code’, consisting of three pairs of one-to-one protein/DNA sequence covariations (7,21). The N15 Cro–DNA complex provides some insights into how this proposed code might operate, but leaves other aspects mysterious. First, a clearer rationale now exists for the binary pairings linking residue H6 to base pair 6. However, both Lys and Gln at H6 also contact other base pairs. In particular, Gln H6 may specify Cyt at -6 through water-mediated contacts that depend on shape complementarity, but it is anchored to make that contact by hydrogen-bonding interactions with Thy -5. In our original analysis, about two-thirds of the Cro–DNA pairs with Gln H6/Cyt -6 also had Thy at -5. On the other hand, several also had Cyt at -5, and it is unclear how Gln H6 would specify Cyt -6 in

such cases. Second, the Pro H1/Thy +2 correlations in the code are mysterious given that Pro H1 does not contact Thy +2 at all in the N15 Cro–DNA complex, while a ‘noncoding’ residue, Tyr H5, contacts it strongly. In a binding site selection experiment with Gln H1- and Pro H1-containing variants of λ Cro, position +2 showed a strong switch from a preference for Ade to a preference for Thy; on the other hand, a Gln-to-Pro substitution at H1 showed little thermodynamic preference for Thy versus Ade at position +2 in EMSA experiments (21). In sum, although it is tempting to envision strong binary sequence covariations as resulting from simple evolutionary swaps between two alternative pairwise interactions, the underlying reality could easily be more complicated than that. No universal code exists for protein–DNA recognition (44,55), and the evolutionary rules within multispecific families (56) of transcription factors may also be nonstraightforward, particularly if binding geometry is not well conserved. More sophisticated recognition models, such as those applied to homeodomains and zinc fingers (57,58), might help illuminate the mechanisms by which Cro has evolved diverse specificity.

The apparent affinity of N15 Cro for its cognate O_R3 site is 25- to 30-fold weaker than that of λ Cro (16,21,59). Both proteins are largely monomeric at the protein concentrations used in DNA-binding experiments, and the effective protein–DNA dissociation constant combines the equilibrium describing protein dimerization with that describing binding of dimers to DNA. The dimerization constants of λ Cro and N15 Cro are comparable (2,12–14), suggesting that dimer–DNA interactions are weaker for N15 Cro. Weaker binding is consistent with the present observation that the N15 Cro–DNA complex has lower protein–DNA contact surface area and fewer direct base contacts than the λ Cro–DNA complex. Unusually, weak interactions between N15 Cro and DNA may also explain the similarity between the cognate DNA affinity of N15 Cro and the strongest reported affinity for P22 Cro (15), despite the fact that the latter is monomeric even at millimolar concentrations (3). Dimerization strength and protein–DNA interactions are both important factors determining apparent affinity for cognate DNA.

One caveat in interpreting the N15 Cro–DNA complex relates to crystal packing. In the N15 Cro–DNA complex, the crystal lattice is held together in part by interactions between adjacent duplexes of DNA. In each duplex, 15 pairs of DNA nucleotides are involved in canonical Watson–Crick base pairs. However, at each end Thy and Ade bases come together to form short stretches of DNA triplex. These interactions do not appear to directly interfere with protein–DNA interactions, as no atom in the third DNA strand comes within 7.5 Å of any atom in N15 Cro. Even so, we cannot rule out the possibility that triplex formation may distort the complex from its native conformation, particularly at the flanks.

Two transcription factors that are one-to-one orthologs are generally functionally equivalent (8), yet even so may diverge in DNA-binding specificity due to coevolution with their cognate binding sites (60–63), among other mechanisms. Thus, although gene duplication with functional divergence probably accounts for most of the diverse DNA-binding specificity found within some transcription factor

families (64–68), it is not the only source. Over long evolutionary distances, protein–DNA coevolution in orthologs can lead to large changes in binding site preference, macromolecular structure and binding geometry. In cases such as Cro, the divergence of one-to-one orthologs can rival or even exceed that observed between paralogs.

DATA AVAILABILITY

Atomic coordinates and structure factors for the reported crystal structure have been deposited with the Protein Data bank under accession number 6ON0.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

National Science Foundation [MCB0643790 to M.H.J.C.]. Funding for open access charge: The University of Arizona. *Conflict of interest statement.* None declared.

REFERENCES

- Ptashne, M. (2004) *A genetic switch: phage lambda revisited*. 3rd ed. Cold Spring Harbor Laboratory Press, NY.
- LeFevre, K.R. and Cordes, M.H. (2003) Retroevolution of lambda Cro toward a stable monomer. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 2345–2350.
- Newlove, T., Konieczka, J.H. and Cordes, M.H. (2004) Secondary structure switching in Cro protein evolution. *Structure*, **12**, 569–581.
- Sauer, R.T., Yocum, R.R., Doolittle, R.F., Lewis, M. and Pabo, C.O. (1982) Homology among DNA-binding proteins suggests use of a conserved super-secondary structure. *Nature*, **298**, 447–451.
- Casjens, S., Hatfull, G. and Hendrix, R. (1992) Evolution of dsDNA tailed-bacteriophage genomes. *Semin. Virol.*, **3**, 383–397.
- Fattah, K.R., Mizutani, S., Fattah, F.J., Matsushiro, A. and Sugino, Y. (2000) A comparative study of the immunity region of lambdaoid phages including Shiga-toxin-converting phages: molecular basis for cross immunity. *Genes Genet. Syst.*, **75**, 223–232.
- Hall, B.M., Lefevre, K.R. and Cordes, M.H. (2005) Sequence correlations between Cro recognition helices and cognate O(R) consensus half-sites suggest conserved rules of protein–DNA recognition. *J. Mol. Biol.*, **350**, 667–681.
- Koonin, E.V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, **39**, 309–338.
- Dubrava, M.S., Ingram, W.M., Roberts, S.A., Weichsel, A., Montfort, W.R. and Cordes, M.H. (2008) N15 Cro and lambda Cro: orthologous DNA-binding domains with completely different but equally effective homodimer interfaces. *Protein Sci.*, **17**, 803–812.
- Roessler, C.G., Hall, B.M., Anderson, W.J., Ingram, W.M., Roberts, S.A., Montfort, W.R. and Cordes, M.H. (2008) Transitive homology-guided structural studies lead to discovery of Cro proteins with 40% sequence identity but different folds. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 2343–2348.
- Ohlendorf, D.H., Tronrud, D.E. and Matthews, B.W. (1998) Refined structure of Cro repressor protein from bacteriophage lambda suggests both flexibility and plasticity. *J. Mol. Biol.*, **280**, 129–136.
- Darling, P.J., Holt, J.M. and Ackers, G.K. (2000) Coupled energetics of lambda cro repressor self-assembly and site-specific DNA operator binding I: analysis of cro dimerization from nanomolar to micromolar concentrations. *Biochemistry*, **39**, 11500–11507.
- Jana, R., Hazbun, T.R., Mollah, A.K. and Mossing, M.C. (1997) A folded monomeric intermediate in the formation of lambda Cro dimer–DNA complexes. *J. Mol. Biol.*, **273**, 402–416.
- Jia, H., Satumba, W.J., Bidwell, G.L. 3rd and Mossing, M.C. (2005) Slow assembly and disassembly of lambda Cro repressor dimers. *J. Mol. Biol.*, **350**, 919–929.
- Poteete, A.R., Hehir, K. and Sauer, R.T. (1986) Bacteriophage P22 Cro protein: sequence, purification, and properties. *Biochemistry*, **25**, 251–256.
- Johnson, A.D., Meyer, B.J. and Ptashne, M. (1979) Interactions between DNA-bound repressors govern regulation by the lambda phage repressor. *Proc. Natl. Acad. Sci. U.S.A.*, **76**, 5061–5065.
- Maniatis, T., Ptashne, M., Backman, K., Kield, D., Flashman, S., Jeffrey, A. and Maurer, R. (1975) Recognition sequences of repressor and polymerase in the operators of bacteriophage lambda. *Cell*, **5**, 109–113.
- Lobočka, M.B., Svarechvsky, A.N., Rybchin, V.N. and Yarmolinsky, M.B. (1996) Characterization of the primary immunity region of the Escherichia coli linear plasmid prophage N15. *J. Bacteriol.*, **178**, 2902–2910.
- Poteete, A.R., Ptashne, M., Ballivet, M. and Eisen, H. (1980) Operator sequences of bacteriophages P22 and 21. *J. Mol. Biol.*, **137**, 81–91.
- Albright, R.A. and Matthews, B.W. (1998) Crystal structure of lambda-Cro bound to a consensus operator at 3.0 Å resolution. *J. Mol. Biol.*, **280**, 137–151.
- Hall, B.M., Vaughn, E.E., Begaye, A.R. and Cordes, M.H. (2011) Reengineering Cro protein functional specificity with an evolutionary code. *J. Mol. Biol.*, **413**, 914–928.
- Aggarwal, A.K., Rodgers, D.W., Drottler, M., Ptashne, M. and Harrison, S.C. (1988) Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science*, **242**, 899–907.
- Anderson, J.E., Ptashne, M. and Harrison, S.C. (1987) Structure of the repressor–operator complex of bacteriophage 434. *Nature*, **326**, 846–852.
- Mondragon, A. and Harrison, S.C. (1991) The phage 434 Cro/OR1 complex at 2.5 Å resolution. *J. Mol. Biol.*, **219**, 321–334.
- Rodgers, D.W. and Harrison, S.C. (1993) The complex between phage 434 repressor DNA-binding domain and operator site OR3: structural differences between consensus and non-consensus half-sites. *Structure*, **1**, 227–240.
- Shimon, L.J. and Harrison, S.C. (1993) The phage 434 OR2/R1-69 complex at 2.5 Å resolution. *J. Mol. Biol.*, **232**, 826–838.
- Watkins, D., Hsiao, C., Woods, K.K., Koudelka, G.B. and Williams, L.D. (2008) P22 c2 repressor–operator complex: mechanisms of direct and indirect readout. *Biochemistry*, **47**, 2325–2338.
- Wolberger, C., Dong, Y.C., Ptashne, M. and Harrison, S.C. (1988) Structure of a phage 434 Cro/DNA complex. *Nature*, **335**, 789–795.
- Battye, T.G., Kontogiannis, L., Johnson, O., Powell, H.R. and Leslie, A.G. (2011) iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta Crystallogr. D Biol. Crystallogr.*, **67**, 271–281.
- Evans, P. (2006) Scaling and assessment of data quality. *Acta Crystallogr. D Biol. Crystallogr.*, **62**, 72–82.
- Strong, M., Sawaya, M.R., Wang, S., Phillips, M., Cascio, D. and Eisenberg, D. (2006) Toward the structural genomics of complexes: crystal structure of a PE/PPE protein complex from Mycobacterium tuberculosis. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 8060–8065.
- Vagin, A. and Teplyakov, A. (2010) Molecular replacement with MOLREP. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 22–25.
- Emsley, P. and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 2126–2132.
- Murshudov, G.N., Vagin, A.A. and Dodson, E.J. (1997) Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D Biol. Crystallogr.*, **53**, 240–255.
- Tickle, I.J., Flensburg, C., Keller, P., Paciorek, W., Sharff, A., Vonrhein, C. and Bricogne, G. (2018) *STARANISO*. Global Phasing, Ltd., Cambridge.
- Painter, J. and Merritt, E.A. (2005) A molecular viewer for the analysis of TLS rigid-body motion in macromolecules. *Acta Crystallogr. D Biol. Crystallogr.*, **61**, 465–471.
- Painter, J. and Merritt, E.A. (2006) Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallogr. D Biol. Crystallogr.*, **62**, 439–450.
- Collaborative Computational Project, N. (1994) The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D Biol. Crystallogr.*, **50**, 760–763.
- Nguyen, M.N., Tan, K.P. and Madhusudhan, M.S. (2011) CLICK–topology-independent comparison of biomolecular 3D structures. *Nucleic Acids Res.*, **39**, W24–W28.

40. Guex, N. and Peitsch, M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.
41. Lavery, R. (2005) Recognizing DNA. *Q Rev. Biophys.*, **38**, 339–344.
42. Lavery, R. and Sklenar, H. (1988) The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dyn.*, **6**, 63–91.
43. Siggers, T.W., Silkov, A. and Honig, B. (2005) Structural alignment of protein–DNA interfaces: insights into the determinants of binding specificity. *J. Mol. Biol.*, **345**, 1027–1045.
44. Pabo, C.O. and Neklyudova, L. (2000) Geometric analysis and comparison of protein–DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol.*, **301**, 597–624.
45. Hammerl, J.A., Jackel, C., Lanka, E., Roschanski, N. and Hertwig, S. (2016) Binding Specificities of the Telomere Phage varphiKO2 Prophage Repressor CB and Lytic Repressor Cro. *Viruses*, **8**, E213.
46. Ahlstrom, L.S. and Miyashita, O. (2011) Molecular simulation uncovers the conformational space of the lambda Cro dimer in solution. *Biophys. J.*, **101**, 2516–2524.
47. Ahlstrom, L.S. and Miyashita, O. (2014) Packing interface energetics in different crystal forms of the lambda Cro dimer. *Proteins*, **82**, 1128–1141.
48. Hall, B.M., Roberts, S.A., Heroux, A. and Cordes, M.H. (2008) Two structures of a lambda Cro variant highlight dimer flexibility but disfavor major dimer distortions upon specific binding of cognate DNA. *J. Mol. Biol.*, **375**, 802–811.
49. Matsuo, H., Shirakawa, M. and Kyogoku, Y. (1995) Three-dimensional dimer structure of the lambda-Cro repressor in solution as determined by heteronuclear multidimensional NMR. *J. Mol. Biol.*, **254**, 668–680.
50. Echols, H., Green, L., Oppenheim, A.B., Oppenheim, A. and Honigman, A. (1973) Role of the cro gene in bacteriophage lambda development. *J. Mol. Biol.*, **80**, 203–216.
51. Eisen, H., Brachet, P., Pereira da Silva, L. and Jacob, F. (1970) Regulation of repressor expression in lambda. *Proc. Natl. Acad. Sci. U.S.A.*, **66**, 855–862.
52. Eisen, H., Georgiou, M., Georgopoulos, C.P., Selzer, G., Gussin, G. and Herskowitz, I. (1975) The role of gene cro in phage development. *Virology*, **68**, 266–269.
53. Folkmanis, A., Maltzman, W., Mellon, P., Skalka, A. and Echols, H. (1977) The essential role of the cro gene in lytic development by bacteriophage lambda. *Virology*, **81**, 352–362.
54. Hammerl, J.A., Jackel, C., Funk, E., Pinnau, S., Mache, C. and Hertwig, S. (2016) The diverse genetic switch of enterobacterial and marine telomere phages. *Bacteriophage*, **6**, e1148805.
55. Matthews, B.W. (1988) Protein–DNA interaction. No code for recognition. *Nature*, **335**, 294–295.
56. Luscombe, N.M. and Thornton, J.M. (2002) Protein–DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J. Mol. Biol.*, **320**, 991–1009.
57. Christensen, R.G., Enuameh, M.S., Noyes, M.B., Brodsky, M.H., Wolfe, S.A. and Stormo, G.D. (2012) Recognition models to predict DNA-binding specificities of homeodomain proteins. *Bioinformatics*, **28**, i84–i89.
58. Gupta, A., Christensen, R.G., Bell, H.A., Goodwin, M., Patel, R.Y., Pandey, M., Enuameh, M.S., Rayla, A.L., Zhu, C., Thibodeau-Beganny, S. *et al.* (2014) An improved predictive recognition model for Cys(2)–His(2) zinc finger proteins. *Nucleic Acids Res.*, **42**, 4800–4812.
59. Darling, P.J., Holt, J.M. and Ackers, G.K. (2000) Coupled energetics of lambda cro repressor self-assembly and site-specific DNA operator binding II: cooperative interactions of cro dimers. *J. Mol. Biol.*, **302**, 625–638.
60. Baker, C.R., Tuch, B.B. and Johnson, A.D. (2011) Extensive DNA-binding specificity divergence of a conserved transcription regulator. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 7493–7498.
61. Gasch, A.P., Moses, A.M., Chiang, D.Y., Fraser, H.B., Berardini, M. and Eisen, M.B. (2004) Conservation and evolution of cis-regulatory systems in ascomycete fungi. *PLoS Biol.*, **2**, e398.
62. Kuo, D., Licon, K., Bandyopadhyay, S., Chuang, R., Luo, C., Catalana, J., Ravasi, T., Tan, K. and Ideker, T. (2010) Coevolution within a transcriptional network by compensatory trans and cis mutations. *Genome Res.*, **20**, 1672–1678.
63. Sayou, C., Monniaux, M., Nanao, M.H., Moyroud, E., Brockington, S.F., Thevenon, E., Chahtane, H., Warthmann, N., Melkonian, M., Zhang, Y. *et al.* (2014) A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. *Science*, **343**, 645–648.
64. Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X. *et al.* (2009) Diversity and complexity in DNA recognition by transcription factors. *Science*, **324**, 1720–1723.
65. Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Pena-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T. *et al.* (2008) Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*, **133**, 1266–1276.
66. Enuameh, M.S., Asriyan, Y., Richards, A., Christensen, R.G., Hall, V.L., Kazemian, M., Zhu, C., Pham, H., Cheng, Q., Blatti, C. *et al.* (2013) Global analysis of Drosophila Cys(2)–His(2) zinc finger proteins reveals a multitude of novel recognition motifs and binding determinants. *Genome Res.*, **23**, 928–940.
67. McKeown, A.N., Bridgham, J.T., Anderson, D.W., Murphy, M.N., Ortlund, E.A. and Thornton, J.W. (2014) Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. *Cell*, **159**, 58–68.
68. Rogers, J.M. and Bulyk, M.L. (2018) Diversification of transcription factor–DNA interactions and the evolution of gene regulatory networks. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, **10**, e1423.