doi:10.1016/j.jmb.2011.08.056 J. Mol. Biol. (2011) 413, 914–928

Contents lists available at www.sciencedirect.com

Journal of Molecular Biology

journal homepage: http://ees.elsevier.com.jmb

# Reengineering Cro Protein Functional Specificity with an Evolutionary Code

**Branwen M. Hall, Erin E. Vaughn, Adrian R. Begaye and Matthew H. J. Cordes***

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

*Received 3 June 2011; received in revised form 13 August 2011; accepted 29 August 2011*
Available online 16 September 2011

***Edited by F. Schmid***

***Keywords:***
transcription factor; recognition code; functional evolution; specificity switch; helix–turn–helix

Cro proteins from different lambdoid bacteriophages are extremely variable in their target consensus DNA sequences and constitute an excellent model for evolution of transcription factor specificity. We experimentally tested a bioinformatically derived evolutionary code relating switches between pairs of amino acids at three recognition helix sites in Cro proteins to switches between pairs of nucleotide bases in the cognate consensus DNA half-sites. We generated all eight possible code variants of bacteriophage λ Cro and used electrophoretic mobility shift assays to compare binding of each variant to its own putative cognate site and to the wild-type cognate site; we also tested the wild-type protein against all eight DNA sites. Each code variant showed stronger binding to its putative cognate site than to the wild-type site, except some variants containing proline at position 27; each also bound its cognate site better than wild-type Cro bound the same site. Most code variants, however, displayed poorer affinity and specificity than wild-type λ Cro. Fluorescence anisotropy assays on λ Cro and the triple code variant (PSQ) against the two cognate sites confirmed the switch in specificity and showed larger apparent effects on binding affinity and specificity. Bacterial one-hybrid assays of λ Cro and PSQ against libraries of sequences with a single randomized half-site showed the expected switches in specificity at two of three coded positions and no clear switches in specificity at noncoded positions. With a few caveats, these results confirm that the proposed Cro evolutionary code can be used to reengineer Cro specificity.



## Introduction

Due to the complex repertoire and context dependence of interactions between amino acids and nucleotides, no deterministic code exists for sequence-specific protein–DNA recognition.[1–3] The most successful efforts at establishing rules of recognition involve codes that apply within a narrow range of binding mode and structural context, such as may exist for a family or a set of close variants of the same DNA-binding protein.[4–7] Some such models are probabilistic rather than deterministic, and most are knowledge based, deriving from experimentally known cognate protein–DNA sequence pairs, sometimes selected using *in vitro* evolution. Structure-based prediction of binding site specificity profiles, including the use of homology models,[8] may provide an alternative route to context-specific rules of recognition.[8–11]

The existence of limited protein–DNA recognition codes prompts two related questions: (1) how broad

*Corresponding author. E-mail address: cordes@email.arizona.edu.

Abbreviations used: RH, recognition helix; EMSA, electrophoretic mobility shift assay; FA, fluorescence anisotropy; hex, hexachlorofluorescein; 3-AT, 3-amino-triazole; EDTA, ethylenediaminetetraacetic acid; SVP, snake venom phosphodiesterase; BSA, bovine serum albumin; dsDNA, double-stranded DNA; SOC, super optimal broth with catabolite repression.

a range of protein and DNA sequence, structure and binding mode can be governed by a common set of recognition rules and (2) whether the natural evolution of transcription factor specificity can be described by "evolutionary codes" involving simple mutational mechanisms. Many important transcription factor families are multispecific in the sense that the family members do not conserve the same binding site profile.[12] In principle, evolution of new specificity in these families can involve both indirect and direct readout effects and can entail complex interdependencies between sequence mutations. Several lines of evidence point to a strong role for mutations in direct contacts, though their effects will not necessarily follow simple deterministic rules. First, with some exceptions, homologous protein–DNA complexes tend to exhibit similar docking geometries, allowing for some conservation of contact patterns.[13] Second, within multispecific families, nonspecific contacts to the backbone are well conserved, while sequence positions making direct contacts to nucleotide bases show high variability.[12] Third, comparisons of protein and cognate DNA sequences within families, including mutual information and evolutionary trace analyses, have revealed protein–DNA sequence covariations and correlations in evolutionary importance that correspond to known or probable direct contacts.[5,7,14,15] Detectable covariations can form the basis for partial evolutionary recognition codes for a given family,[5,7] affording an improved understanding of evolutionary mechanism, as well as applications to the design, engineering and prediction of functional specificity.[16]

Identification of natural protein–DNA sequence covariations depends upon availability of a significant database of homologous cognate pairs. One way of building such databases combines comparative genomics with limited experimental knowledge;[5,7,17] for prokaryotic systems, this type of approach can be aided by local control of gene expression,[18] such that transcription factors and their cognate binding sites are in proximity, and binding site sequences for a given protein can be identified with some confidence in the absence of direct experimental evidence.[5,7] Another route involves the direct experimental determination of binding site profiles for many homologs, for which there are numerous emerging high-throughput methods.[19,20] For some families, such as the TAL (*t*ranscription *a*ctivator-*l*ike) effectors, fewer cognate protein–DNA sequence pairs are required to define recognition rules because of the presence of multiple homologous repeat structures within each transcription factor, each of which recognizes a subsequence of the DNA target.[21,22]

The Cro family of bacteriophage transcription factors exhibits an extreme degree of multispecificity,[7,23] as well as very high diversity at the level of protein sequence and structure.[24,25] As such, it provides an excellent and unique model system for understanding the evolutionary divergence of binding site preference and the evolutionary distance across which code-like relations might apply. Cro proteins are classic helix–turn–helix phage repressors, binding as dimers (Fig. 1a) to a set of three 14- to 20-base-pair pseudosymmetric binding sites within the $O_R$ regulatory region adjacent to the *cro* gene. Although Cro proteins from different phage are in general orthologous and have a conserved gene and binding site position, the protein and binding site properties are widely divergent. For example, Cro proteins from phages N15 and λ have different global folds (all-α *versus* α + β, respectively) and no similarity in protein sequence.[26] Only three of seven base pairs are conserved between the two consensus $O_R$ half-sites.[7,27,28] The full $O_R$ sites also differ in symmetry properties, with the axis of pseudosymmetry for N15 lying between two base pairs and that for λ lying on one base pair.

We previously identified Cro protein binding site pairs from genome sequence information alone by taking advantage of the proximity of the *cro* gene to a set of three cognate $O_R$ binding sites, as well as the symmetry within and similarity between the individual $O_R$ sites.[7] Comparison of 32 Cro proteins with their putative cognate consensus $O_R$ half-sites revealed strong one-to-one sequence correlations between three sites in the third helix of the protein [often called the recognition helix (RH)] and three positions within the half-site (Fig. 1d). At each pair of positions, natural Cro protein and cognate half-site sequence alignments were dominated by two amino acid residue types and two nucleotides, respectively. Each sequence correlation also corresponded to amino acid/nucleotide base contacts in the known crystal structure of λ Cro (Fig. 1a–c) with a consensus binding site.[29] We thus proposed a simple, partial evolutionary code (Fig. 1d) relating sequence variation among Cro proteins to differences in probable binding site specificity.[7]

In the present study, we conduct a key experimental test of this proposed evolutionary code by asking whether it can be used to reengineer the specificity of λ Cro, one member of the family. We report the results of electrophoretic mobility shift assay (EMSA), fluorescence anisotropy (FA) assay and bacterial one-hybrid assay, probing the specificity of λ Cro variants mutated according to the putative code. The results indicate that, with some caveats, the code can indeed be used to reengineer λ Cro's specificity. Because λ Cro has an unusual α + β fold not shared by many other Cro homologs, the results also suggest that some evolutionary recognition rules involving direct contacts might persist across long evolutionary distances and in spite of large changes in protein sequence and structure.

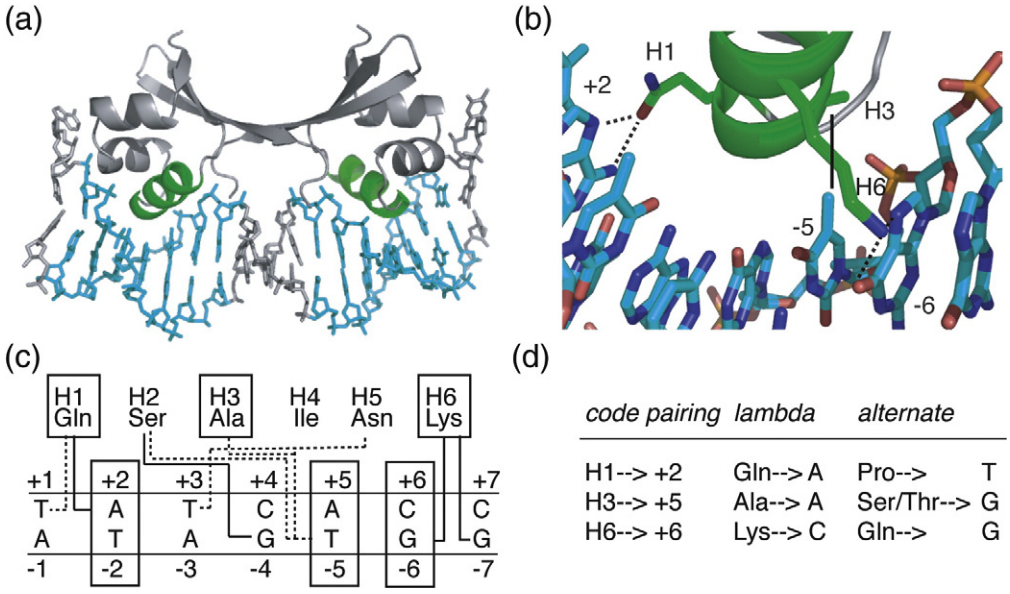


**Fig. 1.** (a) Complex of λ Cro with consensus DNA (Protein Data Bank ID 6CRO) and with Cro RH and DNA half-site highlighted in green and cyan, respectively. (b) Close-up of RH showing interactions of putative coding residues H1, H3 and H6 with bases +2, −5 and −6 in the half-site. (c) Diagram of protein–DNA contacts between the RH and the half-site, with amino acid residues and bases governed by the code shown surrounded by boxes. Continuous lines indicate hydrogen-bonding interactions, while broken lines indicate van der Waals contacts. (d) Table of correlations in the proposed Cro evolutionary code, with linked RH and half-site positions shown on the left-hand column. For each sequence correlation, residue pairings found in λ Cro are shown in one column, and alternate residue pairings are shown in the other.

## Results

The proposed limited Cro code[7] involves two different amino acid residue types at each of three RH sites, and each different residue specifies one of two possible nucleotide bases at one of three half-site positions in the DNA (Fig. 1d). Under the code, it is possible to generate $2^3$ different RH subsequences at positions H1, H3 and H6 (residues 27, 29 and 32, respectively, in the amino acid sequence), and each of these eight sequences has a cognate DNA half-site subsequence at base pairs 2, 5 and 6. To comprehensively test our ability to alter the specificity of λ Cro using the code, we generated all eight possible code variants, along with all eight cognate DNA variants (see Table 1). For convenience, we will refer to the protein and DNA variants by three-letter names corresponding to the protein or DNA subsequence at the three positions related to the code. For example, wild-type λ Cro is called QAK according to the presence of Gln, Ala and Lys at positions H1, H3 and H6 of the RH, respectively, while its cognate site is called AAC according to the presence of the bases Ade, Ade and Cyt at positions +2, +5 and +6 of the half-site (see Table 1). Three of the eight code variants of Cro are single mutants (PAK, QAQ and QSK), three are double mutants (PSK, QSQ and PAQ), one is a triple mutant (PSQ) and one is the wild-type sequence (QAK). The cognate DNA sites studied are variants on a symmetric $O_R$ consensus site for the wild-type protein[29] and differ from the sequence of this site by one, two or three substitutions in each of the two half-sites.

We expressed and purified the eight code variants of λ Cro. Since the mutations were on the solvent-exposed face of an α-helix, they were not expected to affect folding; to confirm folding stability, we performed reversible thermal denaturation experiments on wild-type λ Cro, the three single mutants (PAK, QSK and QAQ) and the triple mutant PSQ (Fig. 2). At 10 μM protein concentration, each variant showed a clear sigmoidal transition with a lower baseline, suggesting that each was largely folded at ambient temperature. Two of the single mutants (QAQ and PAK, corresponding to K32Q and Q27P substitutions, respectively) slightly stabilized λ Cro with $\Delta T_m$ values of +3 and +7 °C, respectively. The third single mutant (QSK, corresponding to an A29S substitution) slightly destabilized λ Cro, with a $\Delta T_m$ of −4 °C. The stability of the triple mutant (PSQ) reflected approximate additivity of the effects of single mutations, giving a $\Delta T_m$ of +6 °C. The double mutants PSK, QSQ and PAQ were not characterized, but based on these results, we assume that their stability is comparable to that of wild-type λ Cro or slightly higher. Both wild-type Cro and PSQ show concentration-dependent

**Table 1.** DNA and protein variant sequences and abbreviated names used in this study

| Cro variant | RH sequence[a] | Site variant | Full site sequence[b] |
|---|---|---|---|
| | HHHHHH | | +++++++ |
| | 123456 | | 1234567 |
| QAK (wild type) | QSAINK | OR1 | A-TATCACCGCCAGAGGTA-B |
| | | OR2 | A-CAACACGCACGGTGTTA-B |
| | | OR3 | A-TATCCCTTGCGGTGATA-B |
| | | AAC(con) | A-TATCACCGGCGGTGATA-B |
| PAK | PSAINK | TAC | A-TTTCACCGGCGGTGAAA-B |
| QSK | QSSINK | AGC | A-TATCGCCGGCGGCGATA-B |
| QAQ | QSAINQ | AAG | A-TATCAGCGGCGCTGATA-B |
| PSK | PSSINK | TGC | A-TTTCGCCGGCGGCGAAA-B |
| QSQ | QSSINQ | AGG | A-TATCGGCGGCGCCGATA-B |
| PAQ | PSAINQ | TAG | A-TTTCAGCGGCGCTGAAA-B |
| PSQ | PSSINQ | TGG | A-TTTCGGCGGCGCCGAAA-B |

[a] RH sequence from 27 to 32 (H1–H6). In code variants, positions of mutations are underlined.
[b] Binding site sequences are listed from 5′ to 3′. Positions of bases +1 to +7 are indicated at the top. For EMSA, sites were flanked by sequences A=TTAGATATT at the 5′-end and B=GATTTAACG at the 3′-end, generating a 35-base-pair overall sequence. For FA experiments, the flanking DNA included only two base pairs at each end rather than eight, for a total of 23 base pairs. In addition, FA constructs had the central base pair inverted. To generate dsDNA, we annealed these sequences to their reverse complements. For the code variants, positions deviating from the wild-type consensus (con) are underlined.

thermal denaturation midpoints (data not shown) consistent with an equilibrium between a folded dimer and an unfolded monomer.[30]

Each Cro variant showed activity toward its putative cognate site in crude initial EMSAs. Addition of each variant protein to cognate DNA yielded a single band shifted to lower mobility (data not shown) in ethidium-bromide-stained EMSA experiments conducted at micromolar concentrations of protein and DNA. Wild-type λ Cro is known to bind cognate consensus DNA cooperatively as a dimer with at least low nanomolar affinity; thus, its binding in this assay should be quantitative and stoichiometric, and the amount of protein required to achieve complete shifting should indicate the percent of the purified protein that is active in binding. We found that all code variants, including wild-type λ Cro, required between 2:1 and 3:1 molar ratios of protein–DNA to achieve complete shifting. If we assume all variants to bind quantitatively to cognate DNA as dimers, the levels of active protein in our preparations varied between 65% and 100%.

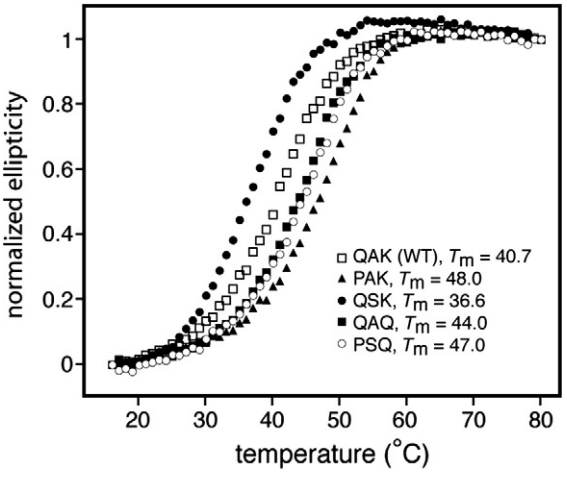


**Fig. 2.** Thermal denaturation curves for single and triple code variants of λ Cro, compared with the wild type, monitored by CD at 222 nm. Ellipticity is normalized based on values at the highest and lowest temperatures.

Next, we conducted EMSA experiments with picomolar concentrations of $^{32}$P-labeled duplex DNA to measure apparent binding affinities of each λ Cro variant for cognate DNA and noncognate DNA. Representative data for wild-type λ Cro (QAK) and the PSQ triple mutant, each bound to its own cognate site, are shown in Fig. 3. We tested each variant protein against its cognate site under the code and also against the wild-type consensus site. Testing of the wild-type protein, on the other hand, was against its own cognate site and all seven noncognate sites, allowing the binding of each variant to its cognate site to be compared with wild-type Cro binding to the same site.

Table 2 summarizes the EMSA data, with each apparent binding affinity represented as the concentration of free protein that achieves half-maximal shifting of the DNA. Variants were studied in three sets consisting of the single variants, the double variants and the triple variant; for each set, the wild-type λ Cro cognate affinity was remeasured, which is why it appears three times in the table. For protein–DNA pairs exhibiting half-maximal protein concentrations of ≤10 nM (including all cognate pairings and all single-base-pair mismatches, as well as the noncognate interactions for double mutants PSK and QSQ), isotherms were best fit by an equilibrium model in which the free protein is monomeric and the bound protein is dimeric. Weaker binding interactions involving higher total protein concentrations required the use of models in

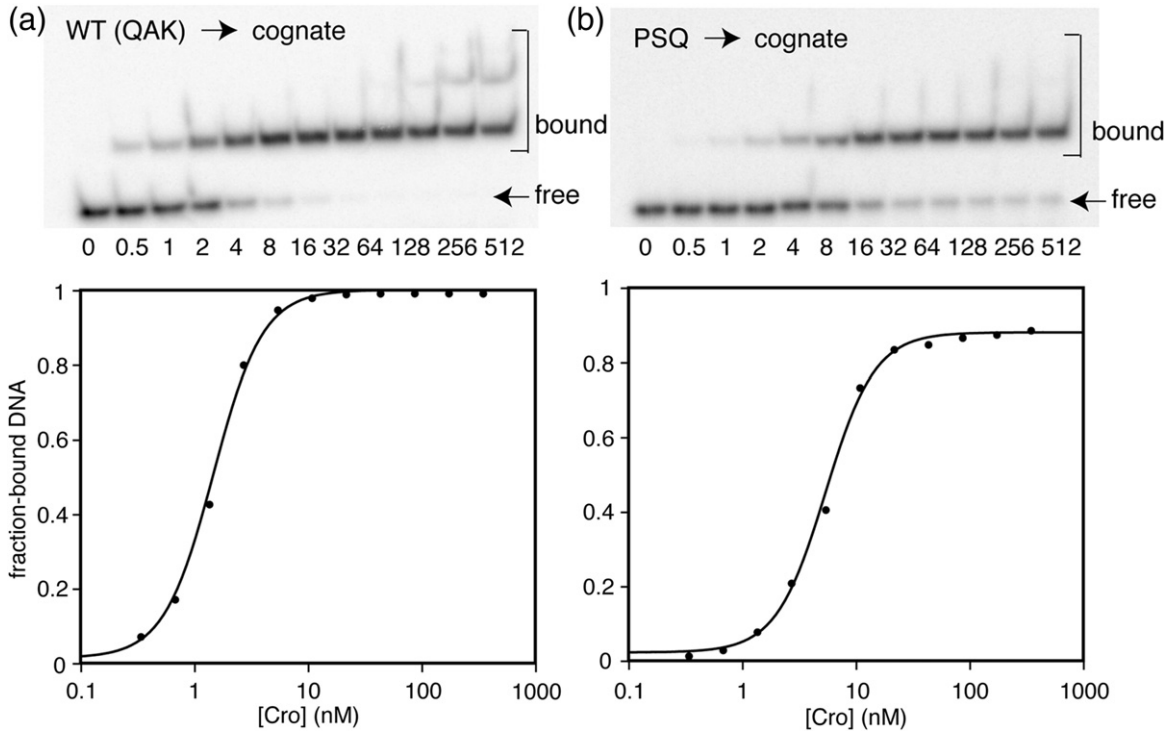


**Fig. 3.** Representative EMSA images (top) and the corresponding fitted binding isotherms (bottom) for (a) wild-type λ Cro binding to cognate consensus DNA and (b) PSQ binding to its putative cognate consensus DNA. Concentrations of protein range from 0 to 512 nM in KP200 buffer containing both glycerol and BSA. Note that, in the case of PSQ, the curve reaches a plateau at less than complete shifting, an effect observed frequently in this study for weaker binding interactions.

which the free protein was assumed to be partly or even completely dimeric (see Materials and Methods). For some protein–DNA pairs, especially those in which weaker binding was observed, the fraction of shifted DNA reached a plateau level significantly below 1 (the experiment for PSQ in Fig. 3b shows a minor version of this effect).

**Table 2.** Nanomolar affinities derived from EMSAs

| Singles | AAC | TAC | AGC | AAG |
|---|---|---|---|---|
| QAK | 1.0 ± 0.3 | 8.6 ± 0.5 | 10.2 ± 1.8 | 4.9 ± 2.4 |
| PAK | 0.7 ± 0.3 | 1.0 ± 0.2 | | |
| QSK | 3.7 ± 1.0 | | 1.4 ± 0.1 | |
| QAQ | 4.5 ± 1.6 | | | 1.6 ± 0.9 |
| **Doubles** | **AAC** | **TGC** | **AGG** | **TAG** |
| QAK | 1.0 ± 0.4 | 69.2 ± 2.9 | 38.0 ± 16.1 | 138 ± 5 |
| PSK | 2.7 ± 0.6 | 1.9 ± 0.8 | | |
| QSQ | 10.1 ± 3.5 | | 4.7 ± 0.2 | |
| PAQ | 21.1 ± 1.2 | | | 1.8 ± 0.5 |
| **Triple** | **AAC** | **TGG** | | |
| QAK | 1.9 ± 0.2 | 205 ± 27 | | |
| PSQ | 54.1 ± 11.3 | 4.5 ± 0.8 | | |

QAK and AAC represent wild-type λ Cro protein and its cognate consensus site, respectively. Values shown are in units of nM and correspond to the total protein concentration yielding half-maximal shifting of the DNA (see Materials and Methods for details of data analysis and fitting).

The EMSA results agree qualitatively with the proposed evolutionary code,[7] with some caveats to be considered in Discussion. The variant λ Cro proteins showed 1–5 nM apparent affinity for their cognate sites and, in most cases, exhibited a preference for this site relative to the noncognate (wild-type) site. Two variants, PAK and PSK, failed to show any significant preference between cognate DNA and noncognate DNA and bound to both quite strongly. Wild-type Cro bound its cognate site significantly tighter than any noncognate site, with stronger specificities seen against more highly mutated sites. In all cases, the variants bound their cognate sites significant more strongly than wild-type λ Cro bound the same site. Effects of mutations in either the protein or the binding site did not show a strict additivity.

The code variants are somewhat diminished in the quality of their DNA-binding function, as measured by their cognate affinity and specificity relative to wild-type λ Cro. Cognate affinities are in general slightly lower for the code variants, an effect more readily apparent for the double and triple mutants than for the single mutants (Table 2). The variants also showed a smaller preference for cognate DNA

*versus* noncognate DNA when compared to the inverse preference of wild-type λ Cro for the same two sites.

One aspect of the EMSA data caused concern. We calculated the apparent relative affinity of wild-type λ Cro for cognate DNA compared to the three sites with each half-site changed at a single position (top section of Table 2) from the fitted $K_d$ values, which are equal to the square of the free protein concentration yielding half-maximal binding. The wild-type protein showed apparent reductions in binding free energies of +2.5, +2.6 and +1.9 kcal/mol for sites mutated at base pairs 2, 5 and 6, respectively, compared to cognate binding. In a filter-binding experiment with a similar set of symmetric sites containing mutations in both half-sites, Takeda *et al.* measured much larger effects of +5.0, +3.0 and +4.5 kcal/mol, respectively.[31] This suggests that either the EMSA or the filter-binding experiments might not reflect thermodynamic reality. Interestingly, the filter-binding results also showed much higher absolute affinities for cognate DNA (in the low picomolar range) in comparison to our experiments and other EMSA and footprinting studies (nanomolar range).

To explore this discrepancy further, we measured binding of wild-type λ Cro to the natural $O_R1$, $O_R2$ and $O_R3$ sites, which have been thoroughly studied by footprinting methods. Johnson *et al.* measured half-maximal binding values of 24, 24 and 3 nM for these sites, respectively,[32] while Darling *et al.* measured similar values of 18, 35 and 3 nM, respectively.[33] By EMSA, we measured values of 7.7 ± 1.8, 6.5 ± 0.5 and 3.6 ± 0.5 nM, respectively. Our results qualitatively reproduce the order of binding affinities ($O_R3 > O_R1 \sim O_R2$) in the footprinting experiments, and the absolute affinity of wild-type λ Cro for $O_R3$ is almost identical. However, the apparent binding of $O_R1$ and $O_R2$ is somewhat stronger in our EMSA experiment, such that the range of affinities toward the different sites observed is smaller in comparison to the footprinting experiments. In terms of $\Delta\Delta G$, our affinity differences for the three sites are only about 1 kcal/mol, while those in the footprinting experiments are roughly 2.5 kcal/mol. Combined with the comparison to the filter-binding assay, the comparison to footprinting suggests an approximately twofold discrepancy (in kilocalories per mole) between energetic effects of λ Cro binding site mutations measured by our EMSA experiments relative to those measured by a second corroborating method.

While the above observations suggest that EMSA qualitatively reproduces binding order of a λ Cro variant with respect to different sites, they cast doubt on the quantitative thermodynamic significance of the apparent affinities. To reinforce our certainty about the utility of the code in altering specificity, we turned to the alternate technique of FA using sites 5′-labeled with hexachlorofluorescein (hex) on one strand. FA is a less sensitive technique than EMSA, and binding affinities of ~1 nM or below cannot easily be measured in direct titrations, so that this method might be of limited utility for a full comparative study of our eight variants. However, FA has the advantage of measuring a true binding constant in aqueous solution and would be useful at the very least in confirming the large specificity change observed for the triple mutant PSQ.

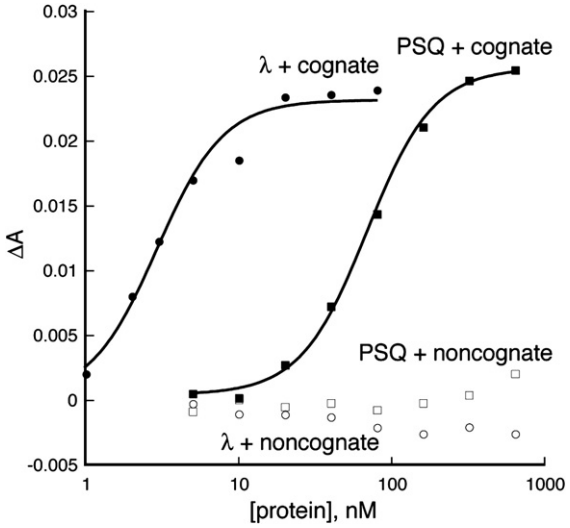


**Fig. 4.** FA of wild-type λ Cro and PSQ variant binding to consensus sites for each. Both proteins were titrated against both putative cognate consensus sites at a concentration of 5 nM singly hex-labeled dsDNA at ambient temperature in KP200 buffer lacking glycerol or BSA. The *y*-axis shows the change in anisotropy. The fit shown for λ Cro is unlikely to be meaningful in terms of dissociation constant measurement but, rather, likely reflects quantitative binding of added protein to DNA in a 2:1 stoichiometric ratio.

The binding of both wild-type λ Cro (QAK) and PSQ to the AAC site (wild-type cognate site) and to the triple mutant site TGG (cognate site of PSQ), monitored by FA, is shown in Fig. 4. Binding of wild-type λ Cro to AAC is essentially quantitative and stoichiometric at 5 nM labeled DNA, while no binding is observed to TGG even at ~500 nM concentrations. By contrast, binding of PSQ to TGG is weaker than stoichiometric and yields a protein concentration for half-maximal binding of 50 nM, while only a hint of binding is observed, even at high concentration, to the ACC site.

These results reinforce the qualitative specificity change observed in the EMSA experiments. Again, however, significant discrepancies are observed in the apparent affinities measured by the two methods. For the noncognate interactions, essentially no binding is observed in the FA experiments, while

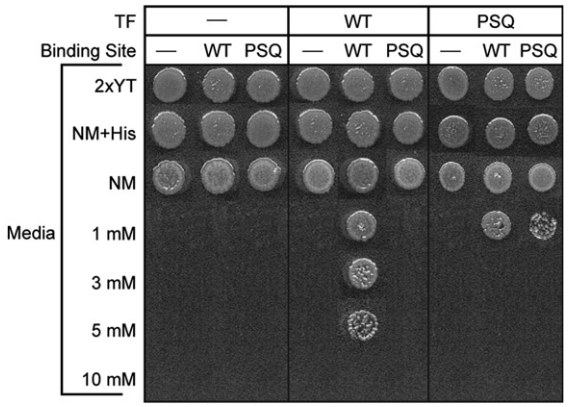


**Fig. 5.** Spot testing of α-TF fusion vectors containing either no transcription factor or wild-type λ Cro or PSQ, against binding site selection vectors containing either no binding site insert or consensus sites for wild-type λ Cro or PSQ. All combinations show survival on rich media (2× YT) or on minimal media (NM) with or without added histidine, while only certain combinations show survival on NM containing increasing concentrations of the competitive HIS3 inhibitor 3-AT. The figure is a composite of spot images from multiple plates; for the sake of the appearance of the figure, however, the same region of imaged plate was used in all cases where zero growth was clearly observed.

50–200 nM affinities were observed by EMSA. For the cognate PSQ interaction, clear binding is observed by FA, but it is somewhat weaker than that observed by EMSA (50 nM *versus* 5 nM). For the cognate wild-type interaction, essentially quantitative stoichiometric binding is observed; thus, under these conditions (5 nM DNA), we can only estimate that the binding affinity is probably less than 5 nM. However, for the two noncognate interactions and the PSQ cognate interaction, FA clearly suggests weaker binding affinity compared with the EMSA experiment. Overall, the FA experiments on wild-type λ Cro and PSQ clearly show that the code mutations switch the specificity and suggest that the thermodynamic effects might be more dramatic than as suggested by the EMSA results.

Finally, we used a bacterial one-hybrid assay[34,35] to conduct a global assessment of the half-site specificity of wild-type λ Cro and the PSQ variant. In this selection, a transcription factor fusion leads to transcriptional activation of a histidine biosynthesis gene (*HIS3*) upon binding of an upstream site, allowing survival in the presence of the competitive HIS3 inhibitor 3-amino-triazole (3-AT). First, we ran spot tests in which each of λ Cro and PSQ was paired with each of the two consensus sites (Fig. 5). The λ Cro/λ consensus DNA pairing showed survival up to 5 mM 3-AT, while no survival even at 1 mM 3-AT was seen when λ Cro was paired with the noncognate PSQ consensus DNA. PSQ protein paired with cognate PSQ consensus DNA showed survival up to 1 mM 3-AT; surprisingly, PSQ paired with the noncognate λ consensus DNA also showed survival at 1 mM 3-AT. No survival in the presence of 3-AT resulted from pairings that included blank transcription factor and/or binding site vectors. These initial findings essentially accord with our *in vitro* binding assays, in the sense that both proteins recognize the putative cognate sites, but PSQ has both lower apparent affinity and lower binding site specificity than wild-type λ Cro.

We next generated two binding site libraries, one for each protein, in which the seven principal base pairs of one half-site were fully randomized, while the bases of the other half of the cognate site, along with the central three base pairs between the two half-sites, were kept constant. This experimental design leads to measurement of half-site specificity of one subunit when binding of the other subunit is templated by the cognate sequence. Sequences of 23 isolates for the wild-type λ Cro library and 20 from the PSQ library, from plates selected at 6 mM 3-AT, contained between one and six substitutions relative to the expected consensus but showed a clear statistical bias toward the putative consensus half-site sequence in each case (Table S2 and Fig. S1).

Sequence logos[36] generated from the wild-type library isolate sequences (Fig. 6) showed a fairly weak consensus but no clear conflicts with the wild-type consensus sequence. The strongest specificities occurred for positions +2, +3 and +6, with fairly weak preferences at other positions including, somewhat surprisingly, the code position +5. Sequence logos for PSQ showed a stronger consensus and also agreed with the expected PSQ consensus at almost every position, with particularly strong and anticipated preferences seen at half-site positions +1, +2, +3, +4 and +6. However, as with the λ Cro library, the apparent specificity of PSQ was very weak at the code position +5 and, in this case, slightly favored the base present in the wild-type consensus site. We conclude that the bacterial one-hybrid library results (a) offer clear support for the proposed evolutionary code at positions +2 and +6, (b) do not offer clear support for the code at position +5 where both variants show weak base preferences and (c) suggest that the code mutations do not alter specificity at noncoded sites such as +3 and +4. The stronger consensus for the PSQ library clones may seem surprising given the lower apparent specificity of PSQ in the spot test assays (Fig. 5). One possible explanation is that the weaker binding affinity of PSQ may place a more stringent selective pressure on this variant relative to the wild type; however, the wild-type and PSQ libraries showed very similar survival rates at equivalent levels of 3-AT (see Materials and Methods).

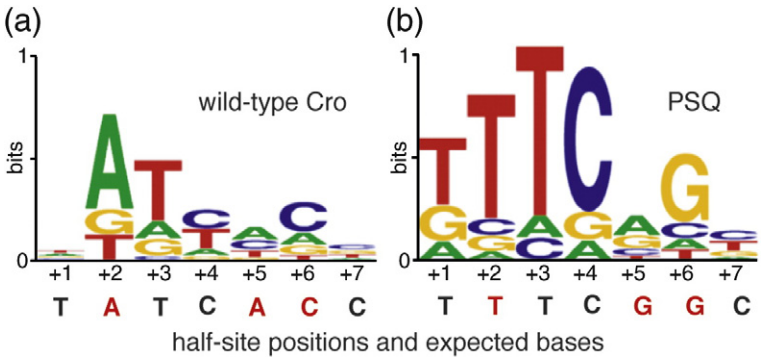


**Fig. 6.** Sequence logos of half-site sequences generated from (a) 23 clones from a wild-type λ Cro library, selected on 6 mM 3-AT plates, and from (b) 20 clones from a PSQ library, selected on 6 mM 3-AT plates. Sequences at the bottom show expected bases at each positions based on the known wild-type λ Cro consensus sequence and the putative PSQ consensus sequence based on the evolutionary code. The three positions specified by the code are highlighted in red.

## Discussion

We have presented results from EMSA, FA assay and bacterial one-hybrid assay relevant to an evolutionary code for Cro DNA-binding specificity derived from genome sequence information.[7] On the whole, mutations at three RH positions appear to achieve reengineering of binding site specificity at three coded DNA half-site positions. Experimental support for the code is accompanied by some caveats, however. The code variants, in particular PSQ, exhibit reduced affinity and specificity. In addition, two of three individual recognition rules are supported by some but not all of the data: EMSA shows very low specificity with respect to the H1→+2 pairing, while the bacterial one-hybrid assay shows very low specificity with respect to the H3→+5 pairing. Below, we discuss these caveats and limitations, as well as the broader context and significance of the work.

We can imagine two models to explain the reduced cognate affinity and specificity observed for PSQ. One possibility is that the larger global contexts of the protein and DNA sequence and structure have coevolved to support a particular combination of direct contacts, namely, those found in the wild-type complex. Movement of the specific coding and coded positions away from the wild-type sequences, without making any other changes, then moves the system away from evolutionary optimality. A second possibility is that the code itself has an inherent asymmetry, and certain contact combinations, such as those found in wild-type λ Cro, are superior in providing high affinity and specificity. In principle, these two different models (contextual coevolution and code asymmetry) could be distinguished by inverse mutagenesis studies on a Cro protein with the opposite code pattern. Studies of this kind are in progress in our laboratory.

EMSA studies provided little support for the H1→+2 sequence correlations under the proposed code. By contrast, this is the most strongly supported protein–DNA sequence linkage in the bacterial one-hybrid library studies. Because of the possible compression of mutational effects on the EMSA work (see above), larger effects of the H1 mutations on the thermodynamics of binding might exist than are evident from our data, which show no significant change for the single substitution at H1 (Table 2). We also note that mutational models based on the λ Cro complex suggest that a van der Waals contact may explain the correlation between Pro at H1 and Thy at +2, but in fact, some structural adjustment would be required to form such a contact.[7] Further clouding the picture, the recently determined structure of N15 Cro with consensus DNA (B. M. Hall, M. S. Dubrava, S. A. Roberts and M. H. Cordes, unpublished results) contains a Pro/Thy sequence pairing at H1→+2 but shows no direct contact between these residues in the complex. The basis and nature of the Pro/Thy code pairing at H1→+2 remain somewhat mysterious.

The bacterial one-hybrid library selections probed the overall half-site specificity of wild-type λ Cro and PSQ rather than just the effect of specific substitutions in the code positions. Analysis of selected sequences showed no effects of the coding mutations on putatively noncontacted, noncoded base pairs, suggesting that the code may represent a set of one-to-one sequence relationships to a first approximation. The results also showed general agreement with specificity predictions for the proposed evolutionary code pairings themselves. A striking exception was the absence of specificity seen at base pair 5 for both wild-type λ Cro and PSQ. One possible explanation is that specificity at base pair 5 hinges critically on correct formation of other specific interactions and only applies when the overall half-site is very close to the consensus sequence. Significantly, our current proposed model[7] for the structural basis of specificity at base pair 5 is unusual in that it involves subtle water-mediated interactions and interactions between the protein backbone and DNA, rather than specific side chain/base contacts. If the interactions affording specificity at base pair 5 depend critically on

subtleties of geometry, they might not lead to selection in the present assay, where most of the selected site sequences differ from the consensus by at least two substitutions (Fig. S1b). Some support for this explanation comes from the fact that selected PSQ half-site sequences containing the expected base at +5 (G) differ by an average of 0.2 bases from the consensus at other positions between +1 and +6, while those containing the expected base at other positions differ by an average of 1.2–1.3 bases from the consensus at other positions; however, a comparable effect is not seen among wild-type isolates (Fig. S1b).

The fact that Cro functions as a dimer bears some discussion with respect to our *in vitro* affinity measurements. DNA binding for the wild type occurs at low nanomolar concentrations where the free protein in solution is predominantly monomeric.[37] Under these conditions, the apparent DNA binding constant includes both the dimerization equilibrium of the protein, which has a dissociation constant of 0.3–7 μM as measured in several studies,[37–39] and the affinity of protein dimer for DNA, which, by at least one estimate, is very strong (~4 pM).[40] Consequently, differences in apparent binding of different Cro variants to the same site (e.g., the fact that the code variants bind their cognate sites better than wild-type Cro does) could partly reflect differences in propensity to dimerize, in addition to differences in protein–DNA interactions. Although our mutations are not in the Cro dimer interface, they do affect folding stability (Fig. 2). Subunit folding of λ Cro may occur independently of dimerization, based on a comparative hydrodynamic and guanidine denaturation analysis by Jana *et al.* that gives a monomer folding stability of 2.1 kcal/mol with an uncertainly of about 1 kcal/mol.[37] However, this stability is sufficiently marginal that folding and dimerization may be partly coupled processes, and mutational effects on folding stability may indirectly contribute to dimerization strength.

Thus, we cannot assume *a priori* that dimerization does not vary among the Cro variants studied and does not affect binding comparisons. Cognate protein–DNA binding curves for all variants were well described by a free monomer/bound dimer model, but wild-type Cro, PAQ and PSQ all deviated from this model for weak noncognate DNA interactions, consistent with significant population of a Cro dimer at ~100 nM total protein concentrations; this deviation was more severe for PAQ and PSQ, consistent with the possibility of stronger dimerization for these variants, which have higher folding stability (Fig. 2). On the other hand, in sedimentation equilibrium experiments, no significant difference was evident between measured dimer dissociation constants of wild-type Cro and PSQ [data not shown; $K_d$ values were near 0.5 μM for both variants, for 5 μM protein loading concentrations in 50 mM Tris (pH 7.5), 250 mM KCl and 0.2 mM ethylenediaminetetraacetic acid (EDTA), at 23,000 and 30,000 rpm].

Cro dimerization behavior has at least one other potential effect on affinity measurements. The monomer–dimer equilibrium is established slowly in solution, and Jana *et al.* and Jia *et al.* have noted that some inconsistencies between reported Cro–DNA affinities in the literature may represent artifacts related to slow dimer dissociation rates (0.02–0.04 $s^{-1}$).[40,41] Specifically, the Cro dimer has a very high intrinsic affinity for DNA (~4 pM by one estimate), and dimer–DNA complex dissociation can be rather slow (a half-life of ~15 $min^{-1}$ has been measured for several Cro variants under the same conditions as our experiments).[40] If protein samples are diluted from a fairly high concentration solution (>1 μM) and mixed rapidly with DNA, excess dimer–DNA complexes could be formed. If analysis of binding occurs before dissociation of the excess complexes restores equilibrium, the result will be an artifactually strong measured binding constant.

In principle, such behavior might explain the apparent compression of relative apparent binding constants of wild-type Cro for different sites in our EMSA experiments. Compression of mutational effects on binding would be observed if the effect of excess complex formation were stronger at the higher protein concentrations necessary to measure the isotherms for the weaker binders. In our experiments, we incubated all protein dilutions for 30 min at ambient temperature prior to mixing with DNA and subsequently incubated protein–DNA mixtures for 30 min before gel loading; these steps should minimize any excess complex formation. Moreover, in test EMSA experiments (data not shown) where protein–DNA mixtures were allowed to stand 3 h before gel loading, apparent binding constants were not weakened and actually became a bit tighter, contrary to expectation if undissociated dimer were rapidly and tightly binding to DNA upon mixing. If artificially strong binding were expected anywhere in our studies, it should appear mostly in the FA assay, where protein was titrated directly into a DNA solution from a relatively high concentration protein solution. However, we see significantly *weaker* cognate site binding for PSQ in the FA assay relative to the EMSA experiment. We believe that the unexpectedly weak mutational effects seen in EMSA do not relate to incomplete dimer dissociation prior to protein–DNA mixing but instead have some other source.

This work is one of a few recent studies that apply comparative prokaryotic genomics to deduce apparent recognition rules within a functionally diverse family and then use these codes for reengineering of protein function or verify them

against existing mutational data.[5,7,16,17] Each of these efforts combines limited experimental data and comparative genomic analysis to assemble databases of homologous protein–DNA cognate pairs. Our studies on Cro,[7] and those of Camas *et al*. on LacI,[5] are aided by the fact that the *lac* and *cro* repressors exert transcriptional control locally in the genome; systems of this kind, which are relatively common in prokaryotes, may be the most convenient for this type of analysis. High-throughput methods for experimental binding site identification offer an alternative route to obtaining protein–DNA cognate pairs for a large number of members of a family.[20] For example, bacterial one-hybrid selections have been used recently to identify optimal binding sites for 84 natural homeodomains from *Drosophila melanogaster*, as one example.[19] Comparisons with the corresponding homeodomain RH sequences produced several clear rules of binding for this family. The increasing availability of both genome sequences and experimental functional information promises more discoveries of family-specific evolutionary codes in the future, along with applications to design and engineering.

Among these examples, the Cro family is arguably the most diverse at the level of binding site sequence, protein sequence and protein structure. Most particularly, several Cro proteins are known to have a classic five-helix repressor fold, while others, including λ Cro, have an unusual mixed α+β fold that is not known outside of this family and descended from the all-helical structure by a relatively continuous accumulation of small mutational events.[24,25] The change in fold completely remodels the dimer interface[24,26,42] and could easily modulate specificity through indirect readout. In this light, the fact that protein–DNA sequence correlations persist across the Cro family[7] and can be used to reengineer the specificity of λ Cro is quite remarkable and suggests that certain simple evolutionary recognition rules may survive major contextual changes during evolution. A more complete picture awaits a comparison of the effects of code mutations between λ Cro and other Cro proteins differing in fold, sequence and binding site. In addition, we have recently determined the structure of N15 Cro bound to consensus cognate DNA (B. M. Hall, M. S. Dubrava, S. A. Roberts and M. H. Cordes, unpublished results), which will allow comparison of docking geometries and contacts for Cro proteins of different fold. Finally, incorporation of *in vitro* or *in vivo* evolution (see, e.g., Gaj *et al.*[43]), including randomization of the coding residues and/or other sites in Cro proteins, might yield improved models of direct and/or contextual effects on specificity and could yield insights not accessible to pure comparative methods or to structure-based design.

## Materials and Methods

### Mutagenesis and protein purification

The previously constructed plasmid pMC140[39] contains the λ Cro open reading frame along with sequence encoding a C-terminal -LEHHHHHH sequence tag for nickel-affinity purification. Mutations were introduced sequentially into pMC140 with a QuikChange site-directed mutagenesis kit (Stratagene), resulting in a family of singly, doubly and triply mutated sequences at positions 27, 29 and 32, corresponding to H1, H3 and H6 of the RH. A listing of λ Cro variants used, along with shorthand names, is given in Table 1. Tagged λ Cro variants were overexpressed from the *Escherichia coli* strain BL21(λDE3) and purified to >95% homogeneity by chromatography on nickel-affinity and SP-Sephadex columns essentially as previously described[44] except that 10 mM imidazole was included in all load and wash steps for the nickel-affinity column. Protein concentrations were quantified by absorbance at 280 nm using a molar extinction coefficient of 4040 $M^{-1}$ $cm^{-1}$ determined for wild-type λ Cro.[45,46] For wild-type λ Cro and the triply mutated variant (PSQ), untagged versions were also generated, and the proteins were expressed and purified according to a previously described procedure.[47]

### Circular dichroism spectroscopy

Thermal denaturation of wild-type λ Cro and variants PAK, QSK, QAQ and PSQ was monitored by changes in circular dichroism (CD) at 222 nM using a Jasco J-810 CD spectrometer. Purified protein was dialyzed into SB250 buffer [50 mM Tris (pH 7.5), 250 mM KCl and 0.2 mM EDTA], diluted to 10 μM and placed in a 1-cm-pathlength cell. Data points were collected from 15–80 °C in 1 °C intervals with 1-min equilibration times at each point. All melts were >95% reversible. Thermal denaturation midpoints ($T_m$) were obtained by nonlinear least-squares fitting essentially as described previously[44] with the program KaleidaGraph (Synergy Software, Reading, PA).

### DNA synthesis, purification and annealing

Wild-type λ Cro and variant sequences used for EMSAs are listed in Table 1. All oligonucleotides for these assays were obtained from Integrated DNA Technologies (Coralville, IA) and purified by urea-denatured polyacrylamide gel electrophoresis (PAGE). Concentrations of purified single-stranded DNA were estimated from $A_{260}$ measurements using an extinction coefficient ($\varepsilon_{calc}$) based on a sum of individual nucleotide extinction coefficients. Hyperchromicity factors were obtained by monitoring spectrophotometric changes in sample absorption following complete digestion with snake venom phosphodiesterase (SVP; Worthington Science).[48] Rather than determining hyperchromicity for every oligonucleotide, we made measurements on four representative oligonucleotides corresponding to both strands of the wild-type λ Cro and PSQ variant cognate sites (Table 1). The hyperchromatic shift of absorbance at 260 nm was monitored for oligonucleotide in digestion buffer [0.1 M Tris–HCl

(pH 9.0) and 15 mM $MgCl_2$] following addition of SVP. When the reaction reached completion at approximately 2 h, the final $A_{260}$ value was corrected for SVP absorption, and the ratio of initial to final $A_{260}$ values was then used to obtain the hyperchromicity factor. For the four representative oligonucleotides, hyperchromicity factors between 0.743 and 0.755 were found, with a mean of 0.748. Because of the small spread in these values, all oligonucleotide concentrations were determined using $\varepsilon_{corr} = 0.748 * \varepsilon_{calc}$, reflecting the average hyperchromicity. In interpreting the stoichiometric EMSA experiments, these hyperchromicity corrections were applied retroactively. For EMSAs, double-stranded DNA (dsDNA) annealed with 35 base pairs was prepared by heating mixtures of two complementary strands (10 μM each) in KP200 buffer [20 mM $KH_2PO_4$ (pH 7), 200 mM KCl, 1 mM EDTA and 5% glycerol] to 75 °C for 5 min and slowly cooling to room temperature.

DNA for FA assays was synthesized, purified and annealed similarly, with several exceptions. First, a 23-base-pair rather than 35-base-pair site construct, with six fewer base pairs of flanking DNA on each side of the consensus site, was used to maximize the observable change in anisotropy upon protein binding. Second, one strand was ordered with a 5′-hex label and purified by the vendor (IDT) using reverse phase HPLC rather than denaturing PAGE to avoid damage to the fluorophore. Third, estimated extinction coefficients for hex-labeled oligonucleotides included a contribution for the fluorophore of 31.6 $mM^{-1}$ $cm^{-1}$, and hyperchromicity tests were performed on both oligonucleotides for each site, yielding slightly higher values for hex-labeled oligonucleotides (0.79–0.80) compared to the unlabeled strands (0.69–0.74). Fourth, annealing of sites for FA was conducted in a thermal cycler by gradual cooling from 80 °C to 22 °C in 0.3 °C intervals per minute in STE buffer [10 mM Tris (pH 8.5), 50 mM NaCl and 1 mM EDTA].

### Electrophoretic mobility shift assays

The percent activity of each variant protein preparation was first assessed using EMSA runs conducted at DNA concentrations in the micromolar range, where binding was assumed to be quantitative according to a 2:1 protein–DNA stoichiometry. Each protein was tested for activity against its putative cognate site by mixing 4 μM dsDNA in EMS buffer [KP200 plus 100–200 μg/mL bovine serum albumin (BSA)] with an equal volume of protein solution (0, 4, 8, 12 and 16 μM concentrations) in the same buffer and allowing the mixture to equilibrate for 30 min. Samples were then loaded onto cooled 12.5% acrylamide/0.5× Tris–borate–EDTA gels running at a constant voltage of 300 V, and the voltage was then lowered to 150 V after 5 min. Shifted and unshifted DNA bands were visualized by ethidium bromide staining and ultraviolet transillumination. Protein activity levels estimated in this manner ranged from 65% to 100%.

Putatively thermodynamic measurements of affinity for each protein–DNA pairing were determined by EMSA carried out at picomolar concentrations of DNA. Single-stranded oligonucleotides were end-labeled by 5′-phosphorylation with [γ-$^{32}$P]ATP using T4 polynucleotide kinase. Labeled strands were then annealed with complementary unlabeled strands as described above and diluted to 200 pM in EMS buffer. Protein was initially diluted in EMS buffer to either 1024 nM or 4096 nM, with twofold serial dilutions then made down to 1 nM or 4 nM, respectively. After 30 min, equal volumes of diluted protein and labeled DNA were mixed and allowed to equilibrate at ambient temperature for 30 min before loading onto a 10% 0.5× Tris–borate–EDTA polyacrylamide gel running at 250 V at 4 °C. Gels were electrophoresed until a tracking lane containing bromophenol blue ran most of the way down the gel or approximately 2 h. Gels were exposed to a phosphorimager plate (GE Healthcare) at 0 °C overnight and imaged with a Typhoon scanner (courtesy of the Arizona Proteomics Consortium), and bands were quantified with ImageQuant (GE Healthcare).

EMSA isotherms for stronger binding interactions (see the text) were fit using nonlinear least-squares regression to Eq. (1), where $\nu$ is the fraction of shifted (bound) DNA, $\nu_{max}$ is the maximal fraction of bound DNA, $K_d$ is the apparent dissociation constant, and $D_{total}$ and $P_{total}$ are total DNA and protein concentrations, respectively. This relation assumes that the free protein is a monomer and that the bound protein is a dimer but does not assume that protein is in large excess over DNA. Protein concentrations giving half-maximal shifting are reported in Table 2 as the square root of the fitted $K_d$.

$$\nu = \frac{\left[\left(\frac{\nu}{\nu_{max}}\right)^3\left(4D_{total}^2\right) - \left(\frac{\nu}{\nu_{max}}\right)^2\left(4D_{total}P_{total} + 4D_{total}^2\right) - P_{total}^2\right]\nu_{max}}{\left(P_{total}^2 + 4D_{total}P_{total} + K_d\right)} \quad (1)$$

For weaker binding interactions (wild-type sites to doubly and triply mutated sites, PAQ and PSQ to wild-type site), Eq. (2) was used, in which $K_a$ is the association constant for dimer binding to DNA, $P_{dim}$ is the protein dimer concentration, and $K_{d,dim}$ is the dissociation constant for protein dimerization. Equation (2) assumes that protein concentration is very high relative to DNA and that the bound protein is a dimer but does not assume that the free protein is purely monomeric.

$$\nu = \frac{P_{dim}K_a\nu_{max}}{1 + P_{dim}K_a}$$

where

$$P_{dim} = \frac{4P_{total} + K_{d,dim} - \sqrt{K_{d,dim}^2 + 8P_{total}K_{d,dim}}}{8} \quad (2)$$

Approximately 80% of individual isotherms for wild-type Cro bound to doubly and triply mutated sites gave intermediate values of $K_{d,dim}$ in the range of 0.01–0.4 μM when fitted to Eq. (2) with $K_{d,dim}$ allowed to vary; the others fit best to a pure dimer model. Because the error in $K_{d,dim}$ is intrinsically high in these fits, a single, fixed value of $K_{d,dim} \sim 0.1$ μM was chosen for all the wild-type Cro fits to avoid introducing relative systematic error into fitted $K_a$ values. This dimerization dissociation constant is slightly stronger than the strongest measured Cro dimerization strength (0.3 μM) in the literature[38] and also stronger than estimates from our own sedimentation equilibrium experiments (~0.5 μM; see the text). For the PAQ and PSQ noncognate interactions, a pure free dimer model

($P_{dim} = P_{total}/2$) gave the best fits. For all fits to Eq. (2), the $P_{total}$ giving half-maximal shifting of DNA (reported in Table 2) was back-calculated from the half-maximal $P_{dim}$, which is equal to the reciprocal of the fitted $K_a$ value.

All data fits with Eqs. (1) and (2) were carried out with the program R† and gave $P$ values $< 0.005$. $\nu_{max}$ values ranged from 0.67 to 0.99, with stronger binding also generally giving higher $\nu_{max}$ and with all cognate protein–DNA pairings having $\nu_{max}$ values of 0.87 or higher. Values reported in Table 2 represent analysis of three independent experiments. Tests of untagged λ Cro and PSQ variants gave $K_d$ values within experimental uncertainty of the tagged variants, affirming that hexahistidine tags did not impact activity.

### Fluorescence anisotropy

FA assays were performed at ambient temperature using an ISS PC1 photon counting spectrofluorimeter in L-format with excitation monochromator set at 539 nm and emission detected through a 550-nm-cutoff filter (Oriel FSR-OG550). Samples contained 5 nM double-stranded, singly hex-labeled annealed target DNA (see above) in 3 mL KP200 buffer lacking any glycerol or BSA. A 1-cm-pathlength cuvette was used. Protein was titrated into the DNA solution from a 3-μM stock solution in the same buffer. After each addition, the protein–DNA mixture was allowed to equilibrate for 5–10 min with stirring prior to making anisotropy readings.

### Bacterial one-hybrid assays

A previously described bacterial one-hybrid system[34,35] was adapted to select preferred half-site DNA sequences for wild-type λ Cro and the PSQ variant. For the putative consensus sites used in preliminary spot tests, oligonucleotides were designed in which the 17-base-pair sequence of each site was surrounded with four to six base pairs of natural flanking sequence from the $O_R3$ site (see Table S1). In turn, this sequence was flanked by an upstream AscI site and a downstream NotI site to allow ligation into the pH3U3 reporter plasmid following primed second strand synthesis (see Table S1) and restriction digestion. In the resulting binding site constructs, the near end of the consensus site was 18 base pairs upstream of the −35 box of the promoter, a spacing optimized through initial test selections on wild-type λ Cro. Genes encoding wild-type λ Cro and the PSQ variant were cloned into the pB1H1 expression vector, resulting in α-TF fusion constructs.

Preliminary spot tests were carried out in the following manner: 10 ng of wild-type or PSQ λ Cro or empty (pB1H1) α-TF fusion plasmids and 10 ng of wild-type or PSQ or empty (pH3U3) binding site reporter plasmids was simultaneously electroporated into 50 μL electrocompetent US0Δ*hisB*Δ*pyrF E. coli* cells (the selection strain). The transformed cells were recovered in 1 mL super optimal broth with catabolite repression (SOC) medium for 1 h at 37 °C, pelleted by centrifugation at 3500*g* at room temperature for 10 min, resuspended in 1 mL NM medium with kanamycin (30 μg/mL) with chloramphenicol (30 μg/mL) and incubated at 37 °C for 1 h to adapt to growth in minimal media.[35] The cells were then pelleted at 18,000*g* at room temperature for 30 s, washed twice with 1 mL $H_2O$, washed once in 1 mL NM medium (centrifuging for 2 min each time) and resuspended in 100 μL NM medium. Serial dilutions were then spotted in triplicate 4-μL spots onto 2× yeast extract/tryptone (YT), NM medium with or without 0.1% histidine and 3-AT selection plates containing 1, 3, 5 and 10 mM 3-AT. All plates contained kanamycin (30 μg/mL) and chloramphenicol (30 μg/mL). Plates were incubated for 24–36 h at 37 °C until colonies formed.

The randomized half-site libraries were designed identically to those used in the preliminary spot tests, except that the first seven base pairs of the half-site nearer to the promoter were replaced with random sequence (see Table S1). Half-site libraries were constructed as follows: consensus site test plasmids were constructed by a simplified version of the same protocol. Binding site oligonucleotides were purified by urea-denaturing PAGE, then subjected to second strand synthesis in a reaction containing 1× ThermoPol buffer (New England Biolabs), 0.4 mM dNTPs, 2.5 mM $MgCl_2$, 50 μM template library oligonucleotide, 50 μM primer (see Table S1) and 5 U Taq DNA polymerase (New England Biolabs). A single thermal cycle was run with a 5-min denaturation step at 95 °C, a 5-min annealing step at 50 °C and a 2-h extension step at 72 °C. The dsDNA was purified by electrophoresis through a 3% TAE UltraPure Agarose 1000 (Invitrogen) gel at 100 V for 1 h, recovered from the gel by electroelution and ethanol precipitation as previously described[34] and resuspended in 50 μL EB buffer (10 mM Tris, pH 8.5). Concentrations of dsDNA achieved were typically ~200–300 ng/μL. The dsDNA was then digested overnight at 37 °C in a 100-μL reaction containing 1× New England Biolabs buffer 4, 100 μg/mL BSA, 5 μg of library dsDNA and 30 U each of AscI and NotI-HF (New England Biolabs). The digested dsDNA was then repurified as above and resuspended in 60 μL EB. Concentrations of digested dsDNA achieved were typically ~25 ng/μL. The pH3U3 reporter vector was digested overnight at 37 °C in a 200-μL reaction containing 1× NEB buffer 4, 100 μg/mL BSA, 20 μg plasmid and 30 U each of AscI and NotI-HF. The digested pH3U3 vector was purified using a QIAquick Gel Extraction Kit (Qiagen) with elution in 60 μL EB. Digested plasmid and insert were ligated in a 30-μL reaction containing 1 μg digested pH3U3 vector, 100 ng digested dsDNA insert and 400 U T4 DNA ligase (New England Biolabs). This corresponded to an optimized 1:10 molar ratio of plasmid to insert. The ligation reaction was incubated overnight at 16 °C, and the DNA was purified using a MinElute Reaction Cleanup Kit (Qiagen), with elution in 20 μL $H_2O$. Electrocompetent XL1-Blue *E. coli* (Stratagene) were then transformed by electroporation of 1 μL of ligation reaction mixed with 60 μL of cells. Cells from four electroporations were pooled in 20 mL SOC medium and recovered with shaking at 37 °C for 1 h. The library was then expanded by growth for 3 h in 30 mL of 2× YT medium with kanamycin (30 μg/mL). The library plasmid DNA was recovered using a GeneJET Plasmid Miniprep Kit (Fermentas), with elution in 80 μL EB. Concentrations of binding site library DNA were typically ~20 ng/μL. Transformation tests prior to expansion gave library sizes of $\sim 5 \times 10^5$ colony-forming units, more than sufficient for

† http://www.r-project.org

coverage of the $1.6\times10^4$ possible sequences for a seven-base-pair randomization.

Each library was subjected to counter-selection to remove self-activating sequences. Electrocompetent US0Δ*hisB*Δ*pyrF E. coli* cells were electroporated with 20 ng of each library (~$10^7$ transformants) with recovery in 1 mL SOC medium at 37 °C for 1 h. The cells were then pelleted by centrifugation at 3500*g* at room temperature for 10 min, resuspended in 1 mL YM medium with kanamycin (30 μg/mL) and incubated at 37 °C for 1 h to adapt to growth in minimal media.[35] The cells were pelleted by centrifugation at 18,000*g* at room temperature for 30 s, washed twice with 1 mL $H_2O$ (increasing centrifugation time to 2 min), washed once with 1 mL YM medium (centrifuging for 30 s) and resuspended in 1.1 mL YM medium. An additional 5 mL YM medium was added to each cell suspension, and the full volume was plated on a single 245 mm × 245 mm 5-fluoroorotic acid counter-selection plate with kanamycin (30 μg/mL) and incubated overnight at 37 °C. The counter-selected binding site libraries were recovered from the selection plate in 10 mL of 2× YT medium, pelleted by centrifugation at 3500*g* for 5 min at 4 °C and purified using a GeneJET Plasmid Miniprep Kit (Fermentas), eluting in 40 μL EB. Concentrations of DNA were typically ~110 ng/μL. Isolates from the counter-selected libraries were sequenced to assess library composition prior to selection (Fig. S1).

The wild-type and PSQ λ Cro α-fusion plasmids were transformed into the US0Δ*hisB*Δ*pyrF* selection strain, and electrocompetent cells were prepared, typically yielding competencies of $10^5$–$10^6$ colony-forming units per nanogram of DNA. Between 100 and 500 ng counter-selected binding site library DNA was electroporated into 60 μL US0Δ*hisB*Δ*pyrF* electrocompetent cells containing either the empty pB1H1 vector or pB1H1 containing wild-type λ Cro or PSQ α-TF fusions. The amount of library DNA used was varied depending on the competency of the cell preparations, so as to approximately equalize the number of transformants selected for the different libraries. The transformed cells were recovered in SOC medium for 1 h at 37 °C, pelleted by centrifugation at 3500*g* at room temperature for 10 min, resuspended in 1 mL NM medium with kanamycin (30 μg/mL) and chloramphenicol (30 μg/mL) and incubated at 37 °C for 1 h to adapt to growth in minimal media. The cells were then pelleted at 18,000*g* at room temperature for 30 s, washed twice with 1 mL $H_2O$, washed once in 1 mL NM medium (centrifuging for 2 min each time) and resuspended in 750 μL NM medium. The cell suspensions were then plated in 100 μL portions onto round, 100 mm × 15 mm 3-AT selection plates containing 0, 1, 2, 3, 4, 5 and 6 mM 3-AT. Each plate had sufficient area for a selection of ~$10^7$ transformants. Plates were then incubated for 36 h at 37 °C. Colonies were isolated from the 6-mM 3-AT plate from each library/TF pair as this concentration provided the most ideal selection stringency. Survival rates for the two cognate library/TF pairings at 6 mM 3-AT were comparable and were 20–200× higher than those for the mismatched library/TF pairings or libraries paired with blank TF vector. Based on these survival differences, we estimated our false positive rate at 0.5–5%.

Isolates were propagated overnight at 37 °C in cultures containing 10 mL of 2× YT medium with kanamycin (30 μg/mL). Plasmids were purified from these cultures using QIAprep Spin Miniprep Kits (Qiagen), eluted in 50 μL $H_2O$ and sequenced using the HU100 primer (Table S1). Subjection of selected isolates to counter-selection tests on fluoroorotic acid, including sequences with four or more substitutions relative to the expected consensus, did not reveal any false positives. Sequences were aligned in ClustalW2, and the aligned half-sites were used to generate DNA sequence logos‡.[36]

## Acknowledgements

This research was supported by the National Science Foundation CAREER Award MCB-0643790 (to M.H.J.C.). Gel scan images were acquired by the Arizona Proteomics Consortium supported by the National Institute of Environmental Health Sciences grant ES06694 to the Southwest Environmental Health Sciences Center, the National Institutes of Health/National Cancer Institute grant CA023074 to the Arizona Cancer Center and the BIO5 Institute of the University of Arizona. We thank Dr. Scot Wolfe for providing vectors, protocols and valuable advice for the bacterial one-hybrid assays.

## Supplementary Data

Supplementary data to this article can be found online at doi:10.1016/j.jmb.2011.08.056

## References

1. Pabo, C. O. & Nekludova, L. (2000). Geometric analysis and comparison of protein–DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol.* **301**, 597–624.
2. Matthews, B. W. (1988). Protein–DNA interaction. No code for recognition. *Nature*, **335**, 294–295.
3. Wolfe, S. A., Grant, R. A., Elrod-Erickson, M. & Pabo, C. O. (2001). Beyond the "recognition code": structures of two $Cys_2His_2$ zinc finger/TATA box complexes. *Structure*, **9**, 717–723.
4. Benos, P. V., Lapedes, A. S. & Stormo, G. D. (2002). Probabilistic code for DNA recognition by proteins of the EGR family. *J. Mol. Biol.* **323**, 701–727.
5. Camas, F. M., Alm, E. J. & Poyatos, J. F. (2010). Local gene regulation details a recognition code within the LacI transcriptional factor family. *PLoS Comput. Biol.* **6**, e1000989.
6. Desjarlais, J. R. & Berg, J. M. (1992). Toward rules relating zinc finger protein sequences and DNA binding site preferences. *Proc. Natl Acad. Sci. USA*, **89**, 7345–7349.
7. Hall, B. M., Lefevre, K. R. & Cordes, M. H. (2005). Sequence correlations between Cro recognition helices

‡ http://weblogo.berkeley.edu

and cognate $O_R$ consensus half-sites suggest conserved rules of protein–DNA recognition. *J. Mol. Biol.* **350**, 667–681.
8. Morozov, A. V., Havranek, J. J., Baker, D. & Siggia, E. D. (2005). Protein–DNA binding specificity predictions with structural models. *Nucleic Acids Res.* **33**, 5781–5798.
9. Angarica, V. E., Perez, A. G., Vasconcelos, A. T., Collado-Vides, J. & Contreras-Moreira, B. (2008). Prediction of TF target sites based on atomistic models of protein–DNA complexes. *BMC Bioinformatics*, **9**, 436.
10. Paillard, G. & Lavery, R. (2004). Analyzing protein–DNA recognition mechanisms. *Structure*, **12**, 113–122.
11. Robertson, T. A. & Varani, G. (2007). An all-atom, distance-dependent scoring function for the prediction of protein–DNA interactions from structure. *Proteins*, **66**, 359–374.
12. Luscombe, N. M. & Thornton, J. M. (2002). Protein–DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J. Mol. Biol.* **320**, 991–1009.
13. Siggers, T. W., Silkov, A. & Honig, B. (2005). Structural alignment of protein–DNA interfaces: insights into the determinants of binding specificity. *J. Mol. Biol.* **345**, 1027–1045.
14. Mahony, S., Auron, P. E. & Benos, P. V. (2007). Inferring protein–DNA dependencies using motif alignments and mutual information. *Bioinformatics*, **23**, i297–304.
15. Raviscioni, M., Gu, P., Sattar, M., Cooney, A. J. & Lichtarge, O. (2005). Correlated evolutionary pressure at interacting transcription factors and DNA response elements can guide the rational engineering of DNA binding specificity. *J. Mol. Biol.* **350**, 402–415.
16. Desai, T. A., Rodionov, D. A., Gelfand, M. S., Alm, E. J. & Rao, C. V. (2009). Engineering transcription factors with novel DNA-binding specificity using comparative genomics. *Nucleic Acids Res.* **37**, 2493–2503.
17. Rodionov, D. A., Dubchak, I. L., Arkin, A. P., Alm, E. J. & Gelfand, M. S. (2005). Dissimilatory metabolism of nitrogen oxides in bacteria: comparative reconstruction of transcriptional networks. *PLoS Comput. Biol.* **1**, e55.
18. Camas, F. M. & Poyatos, J. F. (2008). What determines the assembly of transcriptional network motifs in *Escherichia coli*? *PLoS ONE*, **3**, e3657.
19. Noyes, M. B., Christensen, R. G., Wakabayashi, A., Stormo, G. D., Brodsky, M. H. & Wolfe, S. A. (2008). Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell*, **133**, 1277–1289.
20. Stormo, G. D. & Zhao, Y. (2010). Determining the specificity of protein–DNA interactions. *Nat. Rev., Genet.* **11**, 751–760.
21. Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S. *et al.* (2009). Breaking the code of DNA binding specificity of TAL-type III effectors. *Science*, **326**, 1509–1512.
22. Moscou, M. J. & Bogdanove, A. J. (2009). A simple cipher governs DNA recognition by TAL effectors. *Science*, **326**, 1501.
23. Fattah, K. R., Mizutani, S., Fattah, F. J., Matsushiro, A. & Sugino, Y. (2000). A comparative study of the immunity region of lambdoid phages including Shiga-toxin-converting phages: molecular basis for cross immunity. *Genes Genet. Syst.* **75**, 223–232.
24. Roessler, C. G., Hall, B. M., Anderson, W. J., Ingram, W. M., Roberts, S. A., Montfort, W. R. & 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. USA*, **105**, 2343–2348.
25. Newlove, T., Konieczka, J. H. & Cordes, M. H. (2004). Secondary structure switching in Cro protein evolution. *Structure*, **12**, 569–581.
26. Dubrava, M. S., Ingram, W. M., Roberts, S. A., Weichsel, A., Montfort, W. R. & Cordes, M. H. (2008). N15 Cro and λ Cro: orthologous DNA-binding domains with completely different but equally effective homodimer interfaces. *Protein Sci.* **17**, 803–812.
27. Lobocka, M. B., Svarchevsky, A. N., Rybchin, V. N. & Yarmolinsky, M. B. (1996). Characterization of the primary immunity region of the *Escherichia coli* linear plasmid prophage N15. *J. Bacteriol.* **178**, 2902–2910.
28. Maniatis, T., Ptashne, M., Backman, K., Kield, D., Flashman, S., Jeffrey, A. & Maurer, R. (1975). Recognition sequences of repressor and polymerase in the operators of bacteriophage lambda. *Cell*, **5**, 109–113.
29. Albright, R. A. & Matthews, B. W. (1998). Crystal structure of λ-Cro bound to a consensus operator at 3.0 Å resolution. *J. Mol. Biol.* **280**, 137–151.
30. Mossing, M. C. & Sauer, R. T. (1990). Stable, monomeric variants of lambda Cro obtained by insertion of a designed beta-hairpin sequence. *Science*, **250**, 1712–1715.
31. Takeda, Y., Sarai, A. & Rivera, V. M. (1989). Analysis of the sequence-specific interactions between Cro repressor and operator DNA by systematic base substitution experiments. *Proc. Natl Acad. Sci. USA*, **86**, 439–443.
32. Johnson, A. D., Meyer, B. J. & Ptashne, M. (1979). Interactions between DNA-bound repressors govern regulation by the λ phage repressor. *Proc. Natl Acad. Sci. USA*, **76**, 5061–5065.
33. Darling, P. J., Holt, J. M. & Ackers, G. K. (2000). Coupled energetics of λ *cro* repressor self-assembly and site-specific DNA operator binding II: cooperative interactions of *cro* dimers. *J. Mol. Biol.* **302**, 625–638.
34. Meng, X. & Wolfe, S. A. (2006). Identifying DNA sequences recognized by a transcription factor using a bacterial one-hybrid system. *Nat. Protoc.* **1**, 30–45.
35. Meng, X., Brodsky, M. H. & Wolfe, S. A. (2005). A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. *Nat. Biotechnol.* **23**, 988–994.
36. Schneider, T. D. & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* **18**, 6097–6100.
37. Jana, R., Hazbun, T. R., Mollah, A. K. & Mossing, M. C. (1997). A folded monomeric intermediate in the formation of lambda Cro dimer–DNA complexes. *J. Mol. Biol.* **273**, 402–416.
38. Darling, P. J., Holt, J. M. & Ackers, G. K. (2000). Coupled energetics of λ *cro* repressor self-assembly and site-specific DNA operator binding I: analysis of *cro* dimerization from nanomolar to micromolar concentrations. *Biochemistry*, **39**, 11500–11507.

39. LeFevre, K. R. & Cordes, M. H. (2003). Retroevolution of λ Cro toward a stable monomer. *Proc. Natl Acad. Sci. USA*, **100**, 2345–2350.
40. Jana, R., Hazbun, T. R., Fields, J. D. & Mossing, M. C. (1998). Single-chain lambda Cro repressors confirm high intrinsic dimer–DNA affinity. *Biochemistry*, **37**, 6446–6455.
41. Jia, H., Satumba, W. J., Bidwell, G. L., 3rd & Mossing, M. C. (2005). Slow assembly and disassembly of λ Cro repressor dimers. *J. Mol. Biol.* **350**, 919–929.
42. Ohlendorf, D. H., Tronrud, D. E. & Matthews, B. W. (1998). Refined structure of Cro repressor protein from bacteriophage λ suggests both flexibility and plasticity. *J. Mol. Biol.* **280**, 129–136.
43. Gaj, T., Mercer, A. C., Gersbach, C. A., Gordley, R. M. & Barbas, C. F., 3rd (2011). Structure-guided reprogramming of serine recombinase DNA sequence specificity. *Proc. Natl Acad. Sci. USA*, **108**, 498–503.
44. Milla, M. E., Brown, B. M. & Sauer, R. T. (1993). P22 Arc repressor: enhanced expression of unstable mutants by addition of polar C-terminal sequences. *Protein Sci.* **2**, 2198–2205.
45. Pace, C. N., Vajdos, F., Fee, L., Grimsley, G. & Gray, T. (1995). How to measure and predict the molar absorption coefficient of a protein. *Protein Sci.* **4**, 2411–2423.
46. Edelhoch, H. (1967). Spectroscopic determination of tryptophan and tyrosine in proteins. *Biochemistry*, **6**, 1948–1954.
47. Hall, B. M., Roberts, S. A., Heroux, A. & Cordes, M. H. (2008). Two structures of a λ Cro variant highlight dimer flexibility but disfavor major dimer distortions upon specific binding of cognate DNA. *J. Mol. Biol.* **375**, 802–811.
48. Kallansrud, G. & Ward, B. (1996). A comparison of measured and calculated single- and double-stranded oligodeoxynucleotide extinction coefficients. *Anal. Biochem.* **236**, 134–138.