# Recognition of DNA by single-chain derivatives of the phage 434 repressor: high affinity binding depends on both the contacted and non-contacted base pairs

**Jinqiu Chen, Sándor Pongor and András Simoncsits***

International Centre for Genetic Engineering and Biotechnology (ICGEB), Area Science Park, Padriciano 99, I-34012 Trieste, Italy

Received March 7, 1997; Revised and Accepted April 18, 1997

## ABSTRACT

**Single-chain derivatives of the phage 434 repressor, termed single-chain repressors, contain covalently dimerized DNA-binding domains (DBD) which are connected with a peptide linker in a head-to-tail arrangement. The prototype RR69 contains two wild-type DBDs, while RR*69 contains a wild-type and an engineered DBD. In this latter domain, the DNA-contacting amino acids of the α3 helix of the 434 repressor are replaced by the corresponding residues of the related P22 repressor. We have used binding site selection, targeted mutagenesis and binding affinity studies to define the optimum DNA recognition sequence for these single-chain proteins. It is shown that RR69 recognizes DNA sequences containing the consensus boxes of the 434 operators in a palindromic arrangement, and that RR*69 optimally binds to non-palindromic sequences containing a 434 operator box and a TTAA box of which the latter is present in most P22 operators. The spacing of these boxes, as in the 434 operators, is 6 bp. The DNA-binding of both single-chain repressors, similar to that of the 434 repressor, is influenced indirectly by the sequence of the non-contacted, spacer region. Thus, high affinity binding is dependent on both direct and indirect recognition. Nonetheless, the single-chain framework can accommodate certain substitutions to obtain altered DNA-binding specificity and RR*69 represents an example for the combination of altered direct and unchanged indirect readout mechanisms.**

## INTRODUCTION

Sequence-specific DNA-binding proteins usually recognize their target sequences by a combination of direct and indirect mechanisms. The direct readout mechanism is generally mediated by small motifs of the DNA-binding domain (DBD), like α-helical regions, as reading heads. Such small motifs, however, contact only a short (3–5 bp) DNA sequence and cannot, *per se*, confer specific and high affinity binding. This is usually achieved in transcription factors by homo- or heterodimer formation of DBDs and by recognition of closely located subsites of longer DNA targets (1–5). Certain transcription factors contain covalently linked DNA-binding modules, e.g. the classical zinc finger proteins (6), POU domain containing proteins (7) and the c-Myb oncoprotein (8). This natural, covalent linkage strategy can be utilised in different ways to obtain artificial DNA-binding proteins. First, as was shown for the zinc finger proteins, individual modules with altered specificity can be designed or selected for given DNA triplets (see ref. 9 for a review) and then these modules can be combined in the covalent framework to recognize longer DNA targets (10,11). Alternatively, DBDs which are naturally not covalently linked, can be joined with designed or natural linkers to obtain single-chain DNA-binding proteins. Such proteins containing different (12) or identical (13,14) DBDs, or engineered variants of the same DBD (15), were shown to recognize the respective expected DNA sequences.

We have previously constructed single-chain derivatives of the phage 434 repressor, which belongs to the best studied members of the helix–turn–helix (HTH) family of DNA-binding proteins (16). First, a homodimeric single-chain protein (RR69) containing two covalently linked DBDs (residues 1–69) in a head to tail arrangement was obtained by expression of a gene containing direct repeats of the 1–69 coding region (13). This molecular arrangement was then used as a framework to introduce, following the example of the ‘helix-redesign’ experiment (18), amino acid changes into one of the domains to obtain a heterodimeric single-chain protein RR*69 (15). In the engineered domain of RR*69, the DNA-contacting amino acids (–1, 1, 2 and 5 in the α3 helix) were replaced by the corresponding residues of the P22 repressor c2, which shows high degree of structural homology in its DBD to that of the 434 repressor (17). *In vitro* and *in vivo* assays showed that RR69 bound to the natural $O_R1$ site of 434 and RR*69 bound to a chimeric operator (19) containing the consensus operator subsites of the 434 and P22 phages (15). These single-chain proteins were therefore termed single-chain repressors.

Binding to a particular sequence by engineered DNA-binding proteins with a novel framework (RR69) and with amino acid substitutions in the recognition helix (RR*69) does not necessarily mean the exclusive recognition of that target. Mutant DNA-binding

*To whom correspondence should be addressed. Tel: +39 40 37571; Fax: +39 40 226555; Email: simoncs@icgeb.trieste.it proteins often show broadened binding specificities. For example, zinc fingers can bind to a subset of targets, as revealed by rapid assays developed for this class of motifs (20,21). In fact, most specific, natural DNA-binding proteins recognize a set of related sequences (3) from which a consensus binding site can be derived. Examples most relevant to this study are the cI repressor of phage 434 and the c2 repressor of phage P22, which each recognize six different operator sites of the respective genome (22,23). These operators show sequence divergence mainly in the inner, spacer region, which is not in direct contact with the repressor as shown by structural (24–26) and/or biochemical studies (see ref. 27 for a review). The sequence of the non-contacted spacer has an indirect effect on the affinity of the operator for repressor in both the 434 and P22 systems (27–34).

Taken together, it seemed desirable to study further the DNA recognition by RR69 and RR*69. By isolating a set of DNA sequences which bind to these artificial DNA-binding proteins, we sought to answer if: (i) the specific recognition pattern, conferred by direct contacts of the 434 repressor, is maintained in the single-chain framework, (ii) the 434-P22 specificity change, obtained originally in the intact 434 repressor by slightly more extensive redesign (19), could be transferred to a domain of the single-chain repressor without the apparent broadening of the specificity, (iii) the conserved spacer length observed in the natural operators is preserved in the selected DNA ligands, (iv) the DNA-binding of the single-chain repressor is indirectly influenced by the sequence of the non-contacted spacer region. Both the ‘wild-type’ RR69 and the mutant RR*69 were used in this study in order to reveal possible common recognition principles which may be due to the common single-chain 434 repressor framework. Our experimental approach to these problems was based on binding site selection from randomized DNA pools and binding affinity determination of the selected ligands. To complement the repertoire of the selected ligands, certain mutations were also generated by targeted mutagenesis. No binding site selection data that we are aware of are available for the 434 repressor. Therefore, this experiment was also performed and the results are compared with those obtained for the single-chain repressors.

## MATERIALS AND METHODS

### Vectors and proteins

Vector construction, protein expression and purification were as previously described (15).

### Oligonucleotides and random DNA pools

Oligonucleotides were synthesized by the ICGEB oligonucleotide service or by Primm s.r.l. (Milan, Italy). The oligonucleotides containing randomized sequences were TCCGGCTCGTATGTTG**CAT**AN$_3$AAGAAN$_5$RT**ATG**AGGACAGCTATGACC (N8.5) and TCCGGCTCGRATGTTG**CAT**ACAAN$_{14}$**ATG**AGGAAACGACTATFACC (N14); the sequences corresponding to the PCR primers TCCGGCTCGTATGTTG (AT421) and GGTCATAGCTGTTTCCT (AT422) are underlined and interrupted boxes for an *Nde*I site are shown in bold. The underlined and bolded sequences are identical to sequences flanking the unique *Nde*I site of the pRIZ′O (–) vectors (15) and serve to clone the selected sequences by loop insertion mutagenesis. These were purified by electrophoresis using acrylamide–8 M urea gels. The randomized oligonucleotides were made double-stranded by primer extension using AT422, Klenow polymerase and dNTP and the double-stranded fragments were purified by electrophoresis on 10% polyacrylamide gel (19:1 acrylamide/bisacrylamide). Targeted mutations in the operator region of pRIZ′ vectors were obtained either by cloning a mixture of linkers (obtained from the oligonucleotides TACAATAAAANTTAAA and TATTTAANTTTTATTG, where N = A, C, G or T) into the *Nde*I site of pRIZ′O(–) vector or by site directed mutagenesis of cloned operators with primers CATACAATAAAACTTBAATATGAGGAAACA (AT501, B = C, G or T) and CATACAATAAAACTTABATATGAGGAAACAG (AT502).

### Binding site selection and cloning of the selected sequences

Selection of binding sites for RR69 and RR*69 from N8.5 random DNA was performed by a method that uses electrophoresis to isolate the bound DNA (35). Binding reactions (25 µl) were performed in binding buffer A (50 mM NaCl, 5 mM $MgCl_2$, 0.2 mM EDTA, 20 mM HEPES, pH 7.9 and 5% glycerol) containing 0.5–1 pmol $^{32}$P-end-labeled N8.5 DNA, 0.5 µg poly(dI-dC) (Pharmacia) and 200 nM repressor protein for 40 min at room temperature. Eight selection cycles, including binding reaction, bound DNA isolation, PCR (25 cycles of 94°C, 58°C and 72°C, 30 s each, 0.5 µM primer), gel purification of the amplified DNA and 5′-end-labeling, were performed.

Selection from the N14 random DNA with RR69, RR*69 and 434 repressor cI was performed by employing nitrocellulose filtration to separate the bound and unbound DNA (36). Binding was performed in binding buffer B (200 mM KCl, 2.5 mM $MgCl_2$, 1 mM $CaCl_2$, 0.1 mM EDTA, 25 mM Tris–HCl, pH 7.2) containing 0.5–1 pmol N14 double-stranded DNA and repressor proteins. The protein concentration was gradually decreased, as the selection progressed, from the initial 200 nM to the final 10 nM (RR69 and cI) or 6 nM (RR*69). The KCl concentration was 100 mM in the first three cycles, and the binding reactions contained 2 µg/ml poly(dI-dC) from the sixth selection cycle. The binding mixtures were filtered through nitrocellulose membrane (BA85, Schleicher & Schüll) using a slot blot manifold (PR600, Hoefer) and the filter was washed with 1 ml binding buffer B. The bound DNA was recovered by soaking the filter slice in 200 µl 0.1 M NaOH followed by neutralization with 15 µl 3 M NaOAc (pH 5.2) and ethanol precipitation in the presence of 10 µg glycogen (Boehringer, MB grade). PCR was performed as above, but it was limited to 15 cycles and the concentration of the primers was increased to 2.5 µM. The PCR products were purified by using the Qiaquick PCR purification kit (Qiagen). This step caused severe loss (at least 70%) of the 55–57 bp long products, but it allowed for rapid progress (two selection cycles per day). The number of the selection cycles was 12 (RR69 and RR*69) or 10 (cI). The progress of the selection was monitored by testing the selected population after certain selection cycles in electrophoretic mobility shift assay (EMSA). The DNA probes used in these assays were obtained by PCR labeling of the selected pool with $^{32}$P-end-labeled AT422 and unlabeled AT421 primers.

Cloning of the selected sequences was performed by loop insertion mutagenesis. Single-stranded oligonucleotide population was prepared after the final selection cycle by asymmetric PCR (37) using AT421 and AT422 primers in a molar ratio of 50:1. Following gel purification and 5′-phosphorylation, mutagenesis was performed on uracil-containing single-stranded DNA templates (38) of pRIZ′O(–) vectors (15). To minimize the background of the non-mutagenized vector, *Nde*I cleavage was performed before

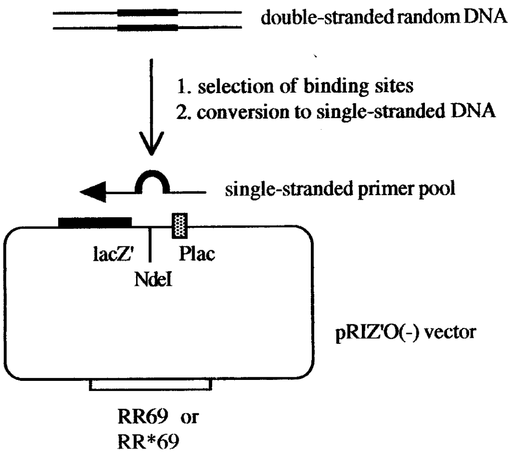

double-stranded random DNA
1. selection of binding sites
2. conversion to single-stranded DNA
single-stranded primer pool
lacZ'
Plac
NdeI
pRIZ'O(-) vector
RR69 or
RR*69

**Figure 1.** Scheme for selection and cloning of the selected sequences into pRIZ′ vectors by loop insertion mutagenesis. Thick lines represent randomized DNA regions.

**Table 1.** Sequences selected for RR69 (A and B), for cI (C) and their binding affinities

|  |  |  | $K_d$ (nM) |  |
|---|---|---|---|---|
|  |  |  | RR69 | cI |
| (A) |  |  |  |  |
| $O_R1$ |  | CATACAAGAAAGTTTGTTATG | 0.8 | 0.8 |
| N8.5 |  | -TTGCATANNNAAGAANNNNNRTATGAGG- |  |  |
| a1 |  | AACAAGAAACCTTGT | 0.8 | 0.8 |
| a2 |  | GACAAGAAATCTTGT | 1.0 | 1.0 |
| a3 |  | TACAAGAATACTTGT | 1.0 | 1.0 |
| a4 |  | TACAAGAAATATTGT | 1.0 | 1.0 |
| a5 | 2x | TACAAGAATCATTGT | 1.0 | 1.0 |
| a6 | 2x | GACAAGAATTCTTGT | 1.2 | 1.2 |
| a7 |  | AACAAGGATTCTTGT | 1.8 | 2.0 |
| a8 | 3x | AACAAGAAACTTTGT | 3.2 | 2.4 |
| a9 |  | AACAAGAAGCGTTGT | 200 | >50 |
| a10 |  | TACAAGAATACTAGT | 100 | 100 |
| (B) |  |  |  |  |
| N14 |  | -TTGCATACAANNNNNNNNNNNNNNATGAGG- |  |  |
| b1 | 2x | ACAAGATATCTTGTAATT | 0.3 |  |
| b2 |  | ACAAGATTCCTTGTATCT | 0.4 |  |
| b3 |  | ttaaACAAGTTATCTTGT....ATG | 0.4 |  |
| b4 | 2x | ACAAGAAAGTTTGTATCG | 0.8 |  |
| b5 | 2x | ACAATATTTCTTGTATTA | 0.8 |  |
| b6 |  | ACAAGGAAACTTGTAGGG | 0.8-1.6 |  |
| b7 | 3x | ACAAGATATATTGTTATT | 0.8 |  |
| b8 |  | ACAATATATCTTGTAATG | 0.8 |  |
| b9 |  | ACAAGATATATTGTATAC |  |  |
| b10 |  | ACAAGTAATATTGTATAT |  |  |
| b11 | 2x | ACAAGTAATATTGTATAG | 1.6-3.2 |  |
| b12 |  | ACAATATAATTTGTATTA | 3.2 |  |
| b13 |  | ACAGGATATATTGTTATT | >200 |  |
| (C) |  |  |  |  |
| c1 | 3x | ACAAGAAAACTTGTATTTg |  |  |
| c2 |  | ACAAGATATATTGTATTA |  |  |
| c3 |  | ACAAGATATCTTGTAATTg |  |  |
| c4 |  | ACAAGTTTATTTGTATTT |  |  |
| c5 |  | ACAATCTTTATTGTATTT |  |  |
| c6 | 9x | ACAATCTTTCTTGTATTT |  |  |
| c7 |  | ACAAGAAACATTGTATTT |  |  |
| c8 | 3x | ACAAGAATTCTTGTATTT |  |  |

Lower case letters represent mutation or insertion, dots represent deletions. Underlined regions correspond to the consensus 434 operator boxes. Sequences isolated more than once are marked, e.g., 2x.

transformation. The RR69 and cI selected sequences were inserted into pRIZ′O(–)RR69, while those selected for RR*69 were inserted into the pRIZ′O(–)RR*69 vector, to obtain pRIZ′$O_x$RR69 and pRIZ′$O_x$RR*69, respectively, where $O_x$ stands for the selected operator analog. Randomly picked clones were sequenced by using the T7 sequencing kit (Pharmacia) and the AS181 primer (15).

### EMSA and binding affinity determination

Radioactive DNA probes (~160 bp) were obtained by PCR using the corresponding pRIZ′Ox plasmid template, $^{32}$P-end-labeled AS181 and unlabeled AT404 primers (15). Apparent $K_d$ values were obtained by determining the protein concentrations at half-maximal binding in protein titration experiments as described (15). The ratio of bound and total DNA probe was obtained by the evaluation of fixed and dried gels with InstantImager (Packard) and the data were evaluated by Kaleidagraph software.

## RESULTS AND DISCUSSION

### Selection of binding sites

Selection of binding sites from two different degenerate DNA pools for RR69 and RR*69 was performed by using two selection methods which differ in the technique of separating the bound and unbound DNA fractions. The selection and cloning scheme is outlined in Figure 1. We used the loop insertion mutagenesis method since it allowed us to introduce the selected sequences precisely into the sequence context of reference operators, previously cloned in the same vector (15). Accordingly, the PCR arms of the degenerate oligonucleotides were designed to correspond to vector sequences flanking an operator insertion site located between the *lac* operator and the *lac*Z′ reporter gene. The clones obtained in this way can be used to study, both *in vitro* and *in vivo*, the interaction between repressors and operator analogs in the same way as described for the reference operators (15).

The N8.5 pool, containing two randomized regions was used in the initial experiments. The full sequence is listed in Materials and Methods and the central region is shown in Tables 1 and 2. The degenerate regions together with adjacent residues could provide consensus boxes for both domains (ACAA and CTT.A.T were expected for R and R*, respectively) with a variety of spacing. By using this pool, we wanted to see whether these boxes were present in the selected sequences and that their spacing corresponded to those found in the 434 operators or in the rationally designed (19) 434-P22 hybrid operator $O_{R*}1$ (15). The results of these selections and experiments with operator analogs of altered spacing (15) showed that both the presence of the consensus boxes and their proper spacing are important for high affinity binding. However, two observations prompted further studies. Firstly, sequences containing imperfect P22 consensus boxes that sometimes showed a higher affinity than those with perfect boxes were also found. Secondly, both the sequence of the spacer region and the identity of intervening bases in the discontinuous P22 box seemed to influence the binding affinity. Since in the N8.5 pool parts of these positions were fixed, we designed a new random pool, N14. This pool contains a fixed 434 operator box ACAA followed by a 14 residue long, fully degenerate region for selection of the spacer and the other consensus box. Selections from N14 were performed for RR69, RR*69 and cI by using the

**Table 2**. Sequences selected for RR*69 and their binding affinities

| | | | Kd(nM) |
|---|---|---|---|
| (A) | | | |
| $O_R$*1 | | CATACAATAAAACTTAAATATG | 0.8 |
| N8.5 | | lacZ'-CCTCATAYNNNNNTTCTTNNNTATGCAA-$P_{lac}$ | |
| a*1 | | CATACAATATTTCTTAATTATG | 0.6 |
| a*2 | | ACAAGGTTTCTTTATT | |
| a*3 | | ACAAGTATTCTTAACT | 1.6 |
| a*4 | | ACAAATATTCTTTACT | 2.0 |
| a*5 | | ACAAATATTCTTTATTg | 2.0 |
| a*6 | | ACAAATATTCTTCATT | 15 |
| a*7 | | ACAAAGATTCTTTAAT | |
| a*8 | | ACAATTATTCTTAACT | |
| a*9 | | cACAAGCATTCTTAAGTg | 5.0 |
| a*10 | | ACAACCATTCTTAAAT | 15 |
| a*11 | | ACAAGAATTCTTCAAT | |
| a*12 | | cACAAGAATTCTTCATT | |
| a*13 | 2x | ACAATAATTCTTTATT | |
| a*14 | | ACAAGGATTCTTAAGT | |
| a*15 | | ACAAAGCTTCTTAAGTg | 20 |
| a*16 | | ACAAGATTTCTTCGCT | 15 |
| a*17 | | ACAAGTATTCTTCGCT | 50 |
| a*18 | | ACAAACATTCTTAGTT | 40 |
| a*19 | | ATACAAGAAATGTTATATG | 5.0 |
| a*20 | | ATACAAGAATAATTATATG | 10 |
| a*21 | | ACACAAGAATGGTTATATG | 35 |
| a*22 | | AACAAAGAAAGTTAATATG | 1.6 |
| a*23 | | AACAACGAATATTAATATG | 20 |
| N8.5 | | $P_{lac}$-TTGCATANNNAAGAANNNNNRTATGAGG-lacZ' | |
| (B) | | | |
| N14 | | -TTGCATACAANNNNNNNNNNNNNNATGAGG- | |
| b*1 | | CATACAAGATATATTAACTAAATG | 0.40 |
| b*2 | 5x | ACAAGATATATTAATTTT | 0.29 |
| b*3 | | ACAAGATATGTTAAATAT | 0.38 |
| b*4 | | ACAAGATAAGTTAATATT | |
| b*5 | 2x | ACAAGATAAGTTAAATTT | |
| b*6 | | ACAAGATAAGTTAAATTA | |
| b*7 | | ACAAGATAAATTAAATTA | 0.30 |
| b*8 | | ACAAGATAAATTAATTCT | 0.32 |
| b*9 | | ACAAGATAATTTAAATTT | 1.0 |
| b*10 | | ACAAGAAAGTTTAATATT | |
| b*11 | | ACAAGAAAGATTAAAAAT | 0.29 |
| b*12 | | ACAAGAAAGATTAAACAA | 0.28 |
| b*13 | | ACAAGAAACATTAAATAT | |
| b*14 | | ACAAGAAATATTAAGTGA | 0.45 |
| b*15 | | ACAAGAAATATTAATTTG | 0.26 |
| b*16 | | ACAAGAAATATTAAATG. | 0.25 |
| b*17 | 2x | ACAAGAAATATTAAATCC | |
| b*18 | | ACAAGAAATATTAAATT. | 0.17 |
| b*19 | | ACAAGAAATATTAAACT. | |
| b*20 | | ACAAGAAATATTAAACTT | 0.27 |
| b*21 | | ACAAGAAATATTAAAATT | 0.25 |
| b*22 | 2x | ACAAGAAATGTTAATATT | 0.50 |
| b*23 | | ACAAGAAATGTTAAGTT. | |
| b*24 | | ACAATAAAGATTTGTTAA | >12.8 |
| b*25 | | ACAATAAAAGTTAAATCCg | |
| b*26 | | ACAAGAAAAGTTAATAC. | |
| b*27 | | ACAAGAAAAGTTAACAGG | 0.80 |
| b*28 | | ACAAGAAAAATTAATTAC | 0.34 |
| b*29 | | ACAAGAAAAATTAAATTC | 0.20 |
| b*30 | | ACAAGAAAAATTAATTAT | |
| b*31 | | ACAAGTTAAATTAATTCT | |
| b*32 | | ACAAGTAATGTTAATATT | |
| b*33 | | ACAAGATTTCTTAAATTG | 0.60 |
| b*34 | | ACAAATTTACTTTAGTTT | 1.2 |
| b*35 | | ACAACTTATCTTAATATT | 1.6 |
| b*36 | | ACAATATTAATTAAATAA | |
| b*37 | | ACAAACAAGATTAATTAA | |

Underlined bases correspond to the consensus 434 (ACAA) or P22 (CTT.A.T) operator boxes.

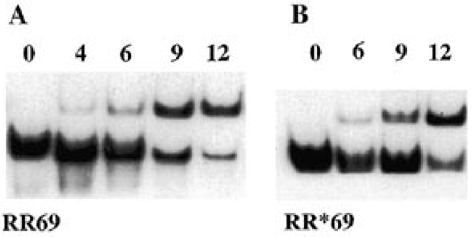


**Figure 2.** Progressive enrichment of the binding sites during selection from the N14 pool for RR69 (**A**) and for RR*69 (**B**). EMSA was performed by using the selected pools as $^{32}$P-labeled probes and the corresponding protein at 10 nM concentration. The number of the selection cycles are shown over the lanes. Lane 0, shift with the starting N14 library; only trace amount of shifted band was seen even by using 500 nM repressors (not shown).

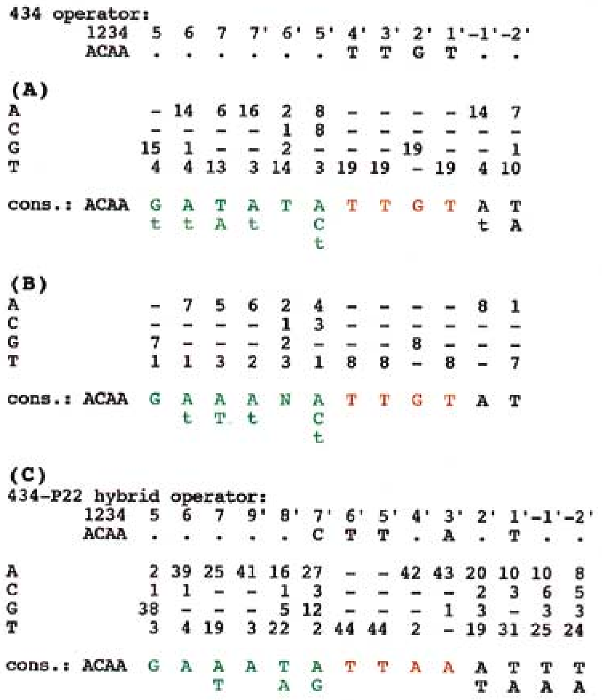
nitrocellulose filtration technique originally applied to the selection of Sp1 binding sites (36). By gradually increasing the stringency of the binding conditions in the subsequent cycles, high affinity ligand populations could be isolated for all three proteins. The progress of the selection for RR69 and RR*69 binding sites is shown in Figure 2. Similar results were obtained with cI. In this case the selected population showed equally high binding affinity for cI and RR69 (not shown). The selected operator analogs and their binding affinities for the corresponding protein(s) are listed in Tables 1 and 2.

## Analysis of the selected sequences

Consensus sequences were derived from the sequences selected from the N14 pool (Fig. 3). The individual sequences can be analyzed by making a correlation between their sequence similarities to the consensus and their observed binding affinities. The sequences obtained from the N8.5 pool are not included in the consensus calculations because certain positions were fixed in this pool. However, these sequences are equally important in our analysis since certain members are identical or very similar to some N14 derived sequences and certain other members show such sequence deviations from the consensus which are not found in the N14 selections. We focus our discussion on the following points: (i) the symmetrically arrayed outer four bases, or contacted regions obtained in RR69 and cI selections; (ii) the length and common sequence features of the spacer, or non-contacted region obtained in all selections; (iii) the outer or contacted bases selected by the R* domain of RR*69.

## The homodimeric RR69 and the natural cI repressor contact identical bases at the outer regions of the operators

In the natural 434 operators the contacted region is, with one exception, ACAA or its 2-fold rotationally symmetric TTGT (22). Substitution of any of these bases, in the context of the 14mer reference operator (39), was previously shown to reduce the binding affinity by at least 100-fold (24,40). Two RR69 selected sequences (a10 and b13 in Table 1) with such substitutions were found, which showed the corresponding affinity decrease relative to $O_R1$ or to other selected sequences containing the same spacer sequence as these mutants but perfect outer boxes (see a10 versus a3 and b13 versus b7). The 4A→4G mutation in b13, which is probably PCR related, results in an outer box present at one side of $O_R3$. The affinity of b13 is much lower than that of $O_R3$ or of



```
434 operator:
        1234   5   6   7  7'  6'  5'  4'  3'  2'  1' -1' -2'
        ACAA   .   .   .   .   .   .   T   T   G   T   .   .

(A)
A              -  14   6  16   2   8   -   -   -   -  14   7
C              -   -   -   -   1   8   -   -   -   -   -   -
G             15   1   -   -   2   -   -   -  19   -   -   1
T              4   4  13   3  14   3  19  19   -  19   4  10

cons.: ACAA    G   A   T   A   T   A   T   T   G   T   A   T
               t   t   A   t       C                   t   A
                                   t

(B)
A              -   7   5   6   2   4   -   -   -   -   8   1
C              -   -   -   -   1   3   -   -   -   -   -   -
G              7   -   -   -   2   -   -   -   8   -   -   -
T              1   1   3   2   3   1   8   8   -   8   -   7

cons.: ACAA    G   A   A   A   N   A   T   T   G   T   A   T
                   t   T   t       C
                                   t

(C)
434-P22 hybrid operator:
        1234   5   6   7  9'  8'  7'  6'  5'  4'  3'  2'  1' -1' -2'
        ACAA   .   .   .   .   .   C   T   T   .   A   .   T   .   .

A              2  39  25  41  16  27   -   -  42  43  20  10  10   8
C              1   1   -   -   1   3   -   -   -   -   2   3   6   5
G             38   -   -   -   5  12   -   -   -   1   3   -   3   3
T              3   4  19   3  22   2  44  44   2   -  19  31  25  24

cons.: ACAA    G   A   A   A   T   A   T   T   A   A   A   T   T   T
                       T       A   G                   T   A   A   A
```

**Figure 3.** Analysis of the sequences selected from the N14 pool for RR69 (**A**), for cI (**B**) and for RR*69 (**C**). Sequences for a general 434 operator and for a 434-P22 hybrid operator, together with the numbering scheme used in Results and Discussion, are shown. In (B), all sequences counted only once, and the c5 and c6 sequences were complemented and reversed for better comparison. The consensus bases of the spacer region and of the contacted regions are shown in green and red, respectively. Under the consensus sequences, bases selected with lower probability are also shown. Lower case letters show bases which were selected relatively infrequently (10–25%) but apparently did not impair the binding in the context of the tested sequences.

the $O_R1$ derivative, $O_R1$-4G, which contains the same mutation in position 4 (32). This lower than expected affinity is probably due to the conformational rigidity of the GG base step, present in positions 4 and 5 of b13. A similar effect is likely to be responsible for the low affinity of the a9 sequence: it contains perfect outer boxes but also a run of GCG in positions 7′ to 5′. All other RR69 selected sequences contain the consensus outer boxes and show a somewhat variable, but high affinity for RR69. The affinity changes of the selected ligands for RR69 and for cI were roughly parallel (Table 1A). These results show that the ACAA outer boxes are required for operator recognition by RR69 and no substitution can be made without substantial loss of binding affinity.

### The sequences selected by the single-chain repressors and the cI repressor share common features in the spacer or non-contacted region

The outer, contacted boxes in the 434 operator sites are separated by 6 bp (22). Previously we showed that altered length spacers (5–10 bp) caused a drastic binding affinity decrease for RR69 and concluded that the interdomain interaction in the DNA-bound RR69 apparently overrules any orientational flexibility allowed by the relatively long linker (15). The present study confirms these results and also shows that both RR69 and RR*69 require 6 bp separation between the respective contacted operator boxes. The hybrid 434-P22 reference 16mer operator (19), here and in the previous study (15) termed $O_{R*}1$, contains 5 bp between the 434 and P22 consensus boxes. The selection and directed mutational studies (see below) showed that the innermost C base of the P22 consensus box (position 7′ in Fig. 3C) is not contacted by RR*69, therefore the spacer in the RR*69 selected DNA ligands is also 6 bp long.

As summarized in Figure 3, the spacer in the selected sequences shows conserved features. At position 5, a preference for G by all three proteins is observed. Structural studies showed that the identity of this base affects the repressor interaction with base pair 4 (24,25), and systematic changes at this position also showed repressor preference for 5G (24). On the right half-site of the RR69 or the cI selected sequences, a less clear preference for C or A is seen at the corresponding 5′ position. Binding affinity data also support these preferences: sequences with both 5G and 5′C are usually the strongest binders (see, e.g., b1, b2 and b3) and stronger than those with only 5G or only 5′C in the same sequence context (see b1 versus b7 or b8).

The inner four bases (6, 7, 7′ and 6′ in the RR69- or 6, 7, 9′ and 8′ in the RR*69-selected sequences) are predominantly, while the central two bases (7 and 7′ or 9′) are exclusively A or T in the N14 derived sequences. At the same time, the N8.5 derived sequence a9, which contains a GCG sequence from 7′ to 5′ positions, shows very low affinity for both RR69 and cI. These results are in agreement with the observed effect of the non-contacted bases on the operator affinity for the 434 repressor (28). This latter study showed that any change in the inner ATAT sequence of the 14mer reference operator for C or G reduced the operator affinity for both the intact repressor and R1-69 and that this negative effect was especially large when the central bases 7 and 7′ were changed. A correlation between the observed binding affinities and the predicted likelihood of DNA-flexure, based on sequence dependent bending preferences in the nucleosome (41), was established (42). The sequence dependent effects, on the other hand, are explained on the basis of the different ‘twistabilities’ (29) or torsional flexibilities (27,30) of the central sequences. The structure of R1-69/operator complexes show that the operator DNA is distorted and different operators take up a particular DNA backbone conformation upon repressor binding. This conformation can generally be characterized by a slight, two-centered bending toward the DBDs and an overwound central region with a compressed minor groove. The affinity of the operator for the repressor seems to depend on the ease with which this conformation can be achieved, independent to the differential changes of various helical parameters upon complex formation with different operators. Our results indicate that, similar to the 434 repressor, RR69 and RR*69 can easily bring about these changes in operators containing either alternating A-T/T-A pairs or runs of three or more A-T pairs in the central region. Structural data for complexes containing such operator motifs exist (24–26,39) and show different conformational adaptations of the base pairs of these motifs to a common DNA backbone conformation, imposed upon by the repressor.

In summary, the present results that the intact repressor and the single-chain repressors select operators with a conserved spacer length and similar spacer sequence motifs further support our previous results (15) which indicated that the spatial arrangement and the interdomain contacts of the covalently joined DBDs in the DNA-bound single-chain repressors should be very similar to those observed for the isolated DBDs in the R1-69/operator complexes. The results also suggest that the set of non-specific contacts between the DBDs of the single-chain repressors and the sugar–phosphate backbone are very similar to those observed in the R1-69/operator complexes (24,25). Thus both the protein–protein and protein–DNA backbone contacts, which were observed for the isolated DBD and shown to cause conformational changes in the operator DNA (24,25), seem to be maintained when the single-chain repressors bind to operator DNA.

### The R* domain of RR*69 selects a consensus TTAA sequence but mutational analysis shows it has a relaxed specificity

Analysis of the P22 operator sites for the c2 repressor and genetic data established a discontinuous CTT.A.T consensus sequence at the operator half-sites (23). It was shown previously that 434 repressor analogs containing the redesigned α3 helix (18) in the whole repressor (18,19) and in the single-chain framework (15) recognized this sequence, but these studies did not reveal whether the whole sequence was required for recognition. The results of this study, as detailed below, suggest that the optimum recognition sequence is 6′-TTAA-3′ and that the corresponding base pairs, with the possible exclusion of the 4′ pair, are contacted by the amino acid residues of the α3 helix of the R* domain.

The results of the N8.5 selection showed that the individual bases in the 7′-CTT.A.T-1′ sequence may not equally contribute to the high affinity binding. In this experiment, the two domains of RR*69 could select hybrid operators in two different orientations with respect to the *lac* promoter: $P_{lac}$–(P22-434)–*lac*Z′ (see Table 2A, a*1–a*18) and $P_{lac}$–(434-P22)–*lac*Z′ (Table 2A, a*19–a*23). The 7′C and 1′T bases may not be major recognition determinants, since their presence in the first group did not require selection, and they were both absent in sequence a*22, which showed high affinity binding. On the other hand, a slight preference for A at the 4′ position could be observed. It was also seen that the non-contacted region, which was partly derived from fixed bases, influenced the binding affinity.

A more stringent selection from the N14 pool provided a population of high binding affinity, which contained a better consensus sequence for both the non-contacted and the contacted regions (Table 2B and Fig. 3C). The highly consensus 6′-TTAA-3′ sequence was found in the putative contacted region. The dinucleotide 6′-TT-5′ was present in all selected sequences and seemed to be absolutely required for high affinity binding, since all those sequences that were obtained at earlier stages of the selection (after six and nine cycles, results not shown) and that lacked one of these T residues showed strongly reduced (25–100 nM) binding affinities. It could be concluded again that the 7′ and 1′ residues are probably not specifically contacted. At the 7′ position, predominantly A and G residues were found, but the few sequences with 7′C or 7′T also showed reasonably high binding affinities. T residue was mainly selected at the 1′ position, but sequences with 1′A or 1′C were also found.

The roles of the individual residues in the 7′ to 1′ region, with the exception of 6′-TT-5′, were further defined by directed mutational analysis of certain residues (7′, 4′ and 3′) in the sequence context of the $O_{R*}$1 operator, or by comparing the binding affinities of those selected sequences which differed only at the 2′ or at the 1′ position. The results of the mutational analysis are summarized in Table 3. This shows that there is no strongly preferred residue at the 7′ position. We have noticed that the affinity order for the 7′ mutants of $O_{R*}$1 does not correlate exactly with the statistical occurrence of the selected 7′ residues: this could be due to differential context effects on the affinity and/or kinetic stability of complexes with different residues at the 7′ position. Of the 4′ mutants, the 4′T derivative showed only 1.5-fold, and the 4′C mutant 6-fold lower affinity than the reference $O_{R*}$1 with 4′A. These residues are also present at the corresponding position in a few P22 operator sites (23). The results of the selection (Fig. 3C) showed a stronger bias for 4′A than could be expected on the basis of the affinity data. At the 3′ position, replacement of A with any other residue led to substantially lower binding affinities. The roles of the 2′ and the 1′ residues were analyzed in constant sequence contexts by using various groups of the N14 selected sequences as described above. Such groups are, for example, the b*14, b*15 and b*16 sequences (difference at the 2′ position) and the b*18, b*20 and b*21 sequences (difference at the 1′ position) of Table 2B. These and a few other examples taken from Table 2B showed that the sequences containing A or T residues at these positions had only slightly (< 2-fold) higher binding affinities than those containing C or G residues. Thus it can be concluded that the 2′ and 1′ residues are not specifically contacted by the R* domain of RR*69.

**Table 3.** Affinities of $O_R$*1 mutants for RR*69

| Operator | Sequence[a] | Affinity[b] |
|---|---|---|
| $O_R$*1(7′C) | ACAATAAAACTTAAAT | 1 |
| 7′A | A | 1 |
| 7′T | T | 2 |
| 7′G | G | 0.75 |
| 4′T | T | 1.5 |
| 4′G | G | 64 |
| 4′C | C | 6 |
| 3′T | T | 32 |
| 3′G | G | 16 |
| 3′C | C | >64 |

[a]The 434 (ACAA) and P22 (CTTAAAT) operator half-sites are underlined. The base numbering scheme is shown in Figure 3C.

[b]Relative affinites are given. Value 1 corresponds to an apparent $K_d$ of $8 \times 10^{-10}$ M.

The results of the selection and mutagenesis studies show that the base pairs contacted by the R* domain of RR*69 are located in the 6′-TTAA-3′ region. Structural predictions for the amino acid side chain–base pair interactions should be largely speculative. The relaxed specificity, observed especially at the 4′ base pair, suggests that alternative networks of direct or solvent-mediated contacts could be formed at the protein–DNA interface. Structural data for the P22 repressor/operator complexes, which may help to elucidate a network of such contacts are not available. On the other hand, the structure of the 434 repressor/operator complexes should be mainly considered, since the observed, surprisingly similar effects of the non-contacted operator sequence on the binding affinities of RR69 (or cI) and RR*69 suggest that the non-specific contacts provided by the 434 domain residues are also similar. These contacts should influence the positioning of the changed α3 helix into the major groove of the DNA. It is also to be considered that all changed amino acid residues in the R* domain have shorter side chains than the corresponding 434 amino acid residues, which implicate a more intimate insertion of the changed α3 helix into the major groove. Based on these considerations, we presume that Gln33 (unchanged 434 residue) interacts with 6′T in a way similar to that seen in the R1-69/operator complexes (for an overview of the contacts between the DBD of the 434 repressor and $O_R1$, see figure 6 of ref. 24 and figure 1A of ref. 40). The Gln28 equivalent in the R* domain, Asn28 may interact with the 3′ A-T base pair, probably by donating a hydrogen bond to the O4 of the T residue. The Val29 residue is likely to make a hydrophobic contact with 5′T and could also contribute to the formation of a hydrophobic environment for the T residue of the 4′A-T base pair. Alternative interactions including Ser32 may also be possible; for example, its interaction with the A residue of the 4′T-A base pair of the 4′T mutant operator could explain the relaxed specificity observed at this position. The 7′ residue is conserved in the P22 operators and is probably specifically contacted by the P22 repressor, but it does not seem to be contacted by the R* domain. The relaxed specificity of the R* domain in RR*69 could be partly due to the lack of this contact. These predictions are based entirely on studies of operator variants and are not complemented by mutational analysis of the putative DNA-contacting amino acid residues. Compared to the proposed models for the P22 repressor–operator interaction (43–45) the predictions of this study are in best agreement with the model based on mutational analysis of both the operator and the repressor (45). Predicted and actual structures may show substantial differences. The structure of the R1-69/$O_R3$ complex shows how a single base pair replacement (at the non-consensus half-site of the operator) can lead to such extensive changes in the protein–DNA interface (26), that could hardly have been predicted. In the present study, there are multiple changes in both interactive partners compared to the R1-69/operator complexes, which were used as ‘templates’ for prediction.

### *In vivo* function of the selected operator analogs

The selected sequences were cloned into a vector which contains the respective gene coding for the single chain repressor (Fig. 1) and were tested as described (15). *In vivo* interaction between repressors and operator analogs could be detected for all tested operator analogs, but no quantitative correlation between the observed *in vitro* binding affinities and the *in vivo* repression levels could be established. By testing some of the operator analogs selected by RR69, we observed that the repression levels varied between 2.5- and 4-fold, meanwhile the apparent $K_d$ values varied between 0.4 and 3.2 nM. Even the low affinity (200 nM) a9 sequence showed detectable, 1.3-fold repression. Similar results were obtained with the members of the RR*69 selected sequences (not shown). The main reason for the lack of more accurate *in vivo* discrimination is probably due to the high intracellular concentration of the single-chain repressors, as previously discussed (15).

## CONCLUSIONS

Single-chain derivatives of the 434 repressor (RR69 and RR*69) recognize highly consensus DNA sequences containing a 14 bp long core sequence. The outer four bases of this sequence are contacted by the amino acid residues of the α3 ‘recognition’ helices and are separated by a 6 bp long spacer or non-contacted region.

The homodimeric RR69 recognizes the general sequence ACAA–6 bp–TTGT, which is identical with the consensus of the natural 434 operator sites and with that of the operators selected for the natural 434 repressor in this work.

The heterodimeric RR*69 recognizes the general sequence ACAA–6 bp–TTAA. The mutant R* domain shows relaxed specificity compared to the wild-type R domain, as base substitutions in the consensus TTAA box cause less dramatic affinity decrease for RR*69 than substitutions in the ACAA box cause for the 434 repressor. The R* domain in RR*69 is also likely to be less specific than the DBD of the wild-type P22 repressor. Detailed specificity studies for the R* domain in the corresponding non-covalent heterodimer (19) are not available, but we assume that the R* domain has the same specificity in RR*69 as in the whole 434 repressor.

The non-contacted regions selected for both single-chain repressors and for the 434 repressor show remarkably similar common features. The sequence-dependent indirect effect of the non-contacted region on the affinity of repressor binding, observed for the 434 repressor, seems to be maintained in the interaction with the single-chain repressors. In addition, all three repressors prefer a short, A + T rich stretch at the outer side of the contacted regions.

The combination of the maintained indirect effects of the non-contacted region and the altered specificity direct contacts (as shown for RR*69) can lead to highly specific recognition of long DNA targets. Such a recognition was previously demonstrated for the zinc finger proteins. The best studied members of the class I zinc finger proteins and their mutants recognize G + C rich sequences (3,9–11,20,21,46–49) although a prototype with A + T rich binding sequences has also been reported (50). The single-chain repressors of this study also prefer A + T rich sequences. However, by using combinatorial libraries and *in vivo* selection techniques as proposed previously (15), it may be possible to isolate single-chain repressor mutants which recognize long DNA targets of more balanced base composition.

## ACKNOWLEDGEMENTS

Thanks are due to Ms S. Kerbavcic and K. Vlahovicek for help with the manuscript preparation. J.C. was supported by a predoctoral fellowship of SISSA (International School of Advanced Studies, Trieste). A.S. is on leave from the Department of Neurochemistry, The Arrhenius Laboratories for Natural Sciences, Stockholm University, Sweden.

## REFERENCES

1 Pabo,C.O. and Sauer,R.T. (1992) *Annu. Rev. Biochem.*, **61**, 1053–1095.
2 Klug,A. (1993) *Gene*, **135**, 83–92.
3 Rhodes,D., Schwabe,J.W.R., Chapman,L. and Fairall,L. (1996) *Phil. Trans. R. Soc, Lond.* B **351**, 501–509.
4 Harrison,S.C. (1991) *Nature*, **353**, 715–719.
5 Wilson,D.S. and Desplan,C. (1995) *Curr. Biol.*, **5**, 32–34.
6 Schwabe,J.W.R. and Klug,A. (1994) *Nature Struct. Biol.*, **1**, 345–349.
7 Herr,W. and Cleary,M.A. (1995) *Genes Dev.*, **9**, 1679–1693.
8 Ogata,K., Morikawa,S., Nakamura,H., Sekikawa,A., Inoue,T., Kanai,H., Sarai,A., Ishii,S. and Nishimura,Y. (1994) *Cell*, **79**, 639–648.
9 Berg,J.M. and Shi,Y. (1996) *Science*, **271**, 1081–1085.
10 Desjarlais,J.R. and Berg,J.M. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 2256–2260.
11 Choo,Y., Sánchez-Garcia,I. and Klug,A. (1994) *Nature*, **372**, 642–645.
12 Pomerantz,J.L., Sharp,P.A. and Pabo,C.O. (1995) *Science*, **267**, 93–96.
13 Percipalle,P., Simoncsits,A., Zakhariev,S., Guarnaccia,C., Sanchez,R. and Pongor,S. (1995) *EMBO J.*, **14**, 3200–3205.
14 Robinson,C.R. and Sauer,R.T. (1996) *Biochemistry*, **35**, 109–116.

15 Simoncsits,A., Chen,J., Percipalle,P., Wang,S., Törö,I. and Pongor,S. (1997) *J. Mol. Biol.*, **267**, 118–131.

16 Ptashne,M. (1992) *A Genetic Switch.* Cell Press and Blackwell Scientific Publications, Cambridge, MA, USA.

17 Sevilla-Sierra,P., Otting,G. and Wütrich,K. (1994) *J. Mol. Biol.*, **235**, 1003–1020.

18 Wharton,R.P. and Ptashne,M. (1985) *Nature*, **316**, 601–605.

19 Hollis,M., Valenzula,D., Pioli,D., Wharton,R. and Ptashne,M. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 5834–5838.

20 Choo,Y. and Klug,A. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 11168–11172.

21 Desjarlais,J.R. and Berg,J.M. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 11099–11103.

22 Wharton,R.P., Brown,E.L. and Ptashne,M. (1984) *Cell*, **38**, 361–369.

23 Poteete,A.R., Ptashne,M., Ballivet,M. and Eisen,H. (1982) *J. Mol. Biol.*, **157**, 21–48.

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

25 Shimon,L.J.W. and Harrison,S.C. (1993) *J. Mol. Biol.*, **232**, 826–838.

26 Rodgers,D.W. and Harrison,S.C. (1993) *Structure*, **1**, 227–240.

27 Koudelka,G.B., Bell,A.C. and Hilchey,S.P. (1996) In Sharma,R.H. and Sharma,M.H. (eds), *Biological Structure and Dynamics*, Adenine Press, Inc., NY., Vol. I, pp. 135–153.

28 Koudelka,G.B., Harrison,S.C. and Ptashne,M. (1987) *Nature*, **326**, 886–888.

29 Koudelka,G.B., Harbury,P., Harrison,S.C. and Ptashne,M. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 4633–4637.

30 Koudelka,G.B. and Carlson,P. (1992) *Nature*, **355**, 89–91.

31 Bell,A.C. and Koudelka,G.B. (1993) *J. Mol. Biol.*, **234,** 542–553.

32 Bell,A.C. and Koudelka,G.B. (1995) *J.Biol. Chem.*, **270**, 1205–1212.

33 Wu,L., Vertina,A. and Koudelka,G.B. (1992) *J.Biol. Chem.*, **267**, 9134–9139.

34 Wu,L. and Koudelka,G.B. (1993) *J.Biol. Chem.*, **268**, 18975–18981.

35 Blackwell,T.K.and Weintraub,H. (1990) *Science*, **250**, 1104–1110.

36 Thiesen,H.-J. and Bach,C. (1990) *Nucleic Acids Res.*, **18**, 3203–3209.

37 McCabe,P.C. (1990) In Innis,M.A., Gelfand,D.H., Sninsky,J.J. and White,T.J. (eds), *PCR Protocols.* Academic Press, Inc., San Diego, CA, pp.76–83.

38 Kunkel,T.A., Bebenek,K. and McClary,J. (1991) *Methods Enzymol.*, **204**, 125–139.

39 Anderson,J.E., Ptashne,M. and Harrison,S.C. (1987) *Nature*, **326**, 846–852.

40 Koudelka,G.B. and Lam,C.-Y. (1993) *J. Biol. Chem.*, **268**, 23812–23817.

41 Travers,A.A. and Klug,A. (1990) In Cozzarelli,N.R. and Wang,J.C. (eds.), *DNA Topology and its Biological Effects*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. pp. 57–106.

42 Drew,H.R., McCall,M.J. and Calladine,C.R.(1990) In Cozzarelli,N.R. and Wang,J.C. (eds), *DNA Topology and its Biological Effects*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. pp. 1–56.

43 Lehming,N., Sartorius,J., Oehler,S., Wilcken-Bergman,B.v. and Müller-Hill,B. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 7947–7951.

44 Suzuki,M., Yagi,N. and Gerstein,M. (1995) *Protein Eng.*, **8**, 329–338.

45 Hilchey,S.P., Wu,L. and Koudelka,G.B. (1995) *J. Biomol. Struct. Dyn.*, **12**, a092.

46 Choo,Y. and Klug,A. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 11163–11167.

47 Jamieson,A.C., Kim,S.H. and Wells,J.A. (1994) *Biochemistry*, **33**, 5689–5695.

48 Rebar,E.J. and Pabo,C.O. (1994) *Science*, **263**, 671–673.

49 Wu,H., Yang,W.-P. and Barbas,C.F., III. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 344–348.

50 Gogos,J.A., Jin,J., Wan,H., Kokkinidis,M. and Kafatos,F.C. (1996) *Proc. Natl. Acad. Sci. USA*, **93**, 2159–2164.