

## Modular construction of extended DNA recognition surfaces: mutant DNA-binding domains of the 434 repressor as building blocks

Tiebing Liang<sup>1</sup>, Jinqiu Chen<sup>2</sup>, Marie-Louise Tjörnhammar, Sándor Pongor and András Simoncsits<sup>3</sup>

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

<sup>1</sup>Present address: Institute of Botany, Chinese Academy of Sciences, Xiang Shan, Hai Dian Qu, Beijing 100093, China

<sup>2</sup>Present address: Department of Pathology and Laboratory Medicine, School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

<sup>3</sup>To whom correspondence should be addressed.  
E-mail: simoncs@icgeb.trieste.it

**Single-chain derivatives of the 434 repressor containing one wild-type and one mutant DNA-binding domain recognize the general operator ACAA-6 base pairs-NNNN, where the ACAA operator subsite is contacted by the wild-type and the NNNN tetramer by the mutant domain. The DNA-binding specificities of several single-chain mutants were studied in detail and the optimal subsites of the mutant domains were determined. The characterized mutant domains were used as building units to obtain homo- and heterodimeric single-chain derivatives. The DNA-binding properties of these domain-shuffled derivatives were tested with a series of designed operators of NNNN-6 base pairs-NNNN type. It was found that the binding specificities of the mutant domains were generally maintained in the new environments and the binding affinities for the optimal DNA ligands were high (with  $K_d$  values in the range of  $10^{-11}$ – $10^{-10}$  M). Considering that only certain sequence motifs in place of the six base pair spacer can support optimal contacts between the mutant domains and their subsites, the single-chain 434 repressor mutants are highly specific for a limited subset of 14 base pair long DNA targets.**

**Keywords:** HTH motif/protein–DNA interactions/protein engineering/434 repressor/single-chain proteins

### Introduction

Single-chain (sc) proteins that bind double-stranded DNA have recently been constructed from several transcription factors by covalent dimerization of DNA-binding domains (DBDs) or of the whole DNA-binding protein. Examples of covalent dimerization by recombinant peptide linkers include the 434 repressor (Percipalle *et al.*, 1995; Simoncsits *et al.*, 1997), the lac repressor headpiece (Gates *et al.*, 1996), the Arc repressor (Robinson and Sauer, 1996a), the estrogen receptor (Kuntz and Shapiro, 1997), the bHLH domain of MASH-1 (Sieber and Alleman, 1998) and the lambda Cro repressor (Jana *et al.*, 1998). The corresponding parent proteins function naturally as non-covalent dimers or higher oligomers of identical subunits and, in most cases, the covalent dimerization has been shown to cause a general increase in DNA-binding affinity. The sc molecules have been used to study various aspects of

DNA recognition, such as the effect of covalent linkage on subunit association, DNA binding, protein folding and stability (Liang *et al.*, 1993; Robinson and Sauer, 1996b, 1998; Jana *et al.*, 1998; Ruiz-Sanz *et al.*, 1999).

The sc arrangement can also provide new ways to engineer proteins of novel DNA-binding specificities. Contrary to the non-covalently associated natural dimers, the sc dimers can easily accommodate two different DBDs of either identical or different DNA-binding specificities. While the non-covalent dimers generally recognize DNA sequences containing palindromic half-sites, the engineered sc derivatives can recognize either palindromic or non-palindromic sequences. In the case of the 434 repressor, it was shown that covalent dimerization of DBDs did not change the wild-type DNA-binding specificity (Chen *et al.*, 1997; Simoncsits *et al.*, 1997). When rational changes (Wharton and Ptashne, 1985) were introduced into one of the DBDs, the heterodimeric sc repressor mutant recognized non-palindromic sequences (Chen *et al.*, 1997). Combinatorial mutant libraries of the sc 434 repressor containing one wild-type DBD and one partially randomized DBD were also constructed and used in a genetic selection to isolate mutant DBDs that bind to predetermined target sites (Simoncsits *et al.*, 1999). These studies showed that the heterodimeric sc 434 repressors recognize a general, 14 base pair (bp) DNA operator sequence of ACAA-6 bp-NNNN type and that strongly binding mutants can be isolated for defined NNNN targets. The ‘non-contacted’ 6 bp spacer region between the 4 bp contacted operator boxes was also shown to influence the binding affinity strongly and similar, consensus spacer sequences were found to support high affinity binding by the natural, the sc and mutant sc 434 repressors (Chen *et al.*, 1997).

In this study, we show that these findings can be utilized to construct long DNA recognition surfaces of novel specificities by combining previously isolated and characterized mutant DBDs in the sc arrangement. The building blocks used were a designed and previously characterized domain (Chen *et al.*, 1997) as well as three mutant DBDs obtained in a protein selection experiment (Simoncsits *et al.*, 1999). First, the DNA-binding properties of these mutant DBDs were characterized in detail by using binding site selection from randomized DNA pools and by binding affinity studies. In these specificity studies, the mutant DBDs were linked to the wild-type DBD. Several homo- and heterodimeric sc proteins were then constructed from these mutant domains and their DNA-binding properties were tested by using artificial operators. These operators were designed by considering the subsite recognition properties of the constituent mutant DBDs. It is shown that the binding specificities of the DBDs are generally maintained in the engineered, double-mutant sc dimers and in several cases specific, high-affinity interactions could be observed between the newly identified protein–DNA cognate pairs. Thus, the sc framework of the 434 repressor can accommodate selected and characterized, mutant DBDs to engineer novel reagents with defined DNA-binding specificities.

## Materials and methods

### General techniques

Protein expression and HPLC purifications were performed as described (Simoncsits *et al.*, 1999) by using a Resource S column (Pharmacia Biotech). Protein concentrations were determined spectrophotometrically using the extinction coefficient  $12\,660\text{ M}^{-1}\text{ cm}^{-1}$  at 280 nm as described (Gill and von Hippel, 1989). The  $^{32}\text{P}$ -labeled DNA probes were obtained by PCR amplification of the operator regions of the corresponding pRIZ' or pCP8 plasmids as 95–125 bp fragments (Simoncsits *et al.*, 1999). Electrophoretic mobility shift assay (EMSA), data collection and quantitative evaluations were performed as described (Simoncsits *et al.*, 1999). Briefly, binding reactions were performed by using 2-fold serial protein dilutions and  $^{32}\text{P}$ -labeled DNA probe present in a concentration which is significantly lower than the protein concentration in the whole titration range. The binding buffer contained a large excess of nonspecific DNA over the probe DNA. Generally, eight protein concentrations were used and the binding reaction mixtures were analysed by EMSA as described (Simoncsits *et al.*, 1997). Binding affinities ( $K_d$ ) were calculated by plotting the fraction of bound DNA ( $\Theta$ ) as a function of the total protein concentration ( $P_t$ ) and the binding isotherm  $\Theta = 1/(1 + K_d/P_t)$  was evaluated using Kaleidagraph software as described (Robinson and Sauer, 1996a; Simoncsits *et al.*, 1999). Bound DNA was derived from the shifted bands corresponding to 1:1 stoichiometry binding. Bands corresponding to higher stoichiometries can generally be observed with the sc proteins at significantly higher concentrations (starting between 10 and 40 nM) than the titration range and the  $K_d$  of the specific interactions of this study. Methylation protection by dimethyl sulfate (DMS) was performed by following a general protocol (Rhodes and Fairall, 1997).

### Selection of binding sites for RR<sub>TATG</sub> and RR<sub>TRPS</sub>

Selections with the two mutants were performed parallel by using a nitrocellulose filtration technique (Chen *et al.*, 1997). The oligonucleotide TCCGGCTCGTATGTTGCATAC-AATAAAN<sub>9</sub>ATGAGGAAACAGCTATGACCTCC (AT500) contained nine randomized residues (N<sub>9</sub>) and the sequences corresponding to PCR primer sites (AT421 upstream and AT422 downstream) are underlined. Eight selection cycles were performed as described (Chen *et al.*, 1997) with slight modifications and simplifications as follows. The binding reactions were performed in 200  $\mu\text{l}$  of binding buffer (50 mM KCl, 2.5 mM MgCl<sub>2</sub>, 1 mM CaCl<sub>2</sub>, 0.1 mM EDTA, 25 mM Tris-HCl, pH 7.2) containing 10  $\mu\text{g/ml}$  poly(dI-dC), mutant sc protein (25 nM in the first cycle, 10 nM in cycles 2 and 3, then 5 nM in cycles 4–8) and 1 pmol ds DNA for 1 h at room temperature. After filtration and washing with water (200  $\mu\text{l}$ ), the bound DNA was recovered by soaking the nitrocellulose filter in 200  $\mu\text{l}$  of PCR buffer (10 mM Tris-HCl, 1.5 mM MgCl<sub>2</sub>, 50 mM KCl, 0.1% Triton X-100, pH 8.3) for 5 min at room temperature. The eluted DNA (10  $\mu\text{l}$ ) was used directly in the PCR mixture (100  $\mu\text{l}$ ) containing the buffer shown above supplemented with 2.5  $\mu\text{M}$  primers, 0.2 mM dNTP and 3 units of Taq polymerase (Roche Molecular Biochemicals). Twelve amplification cycles (94, 58 and 72°C, 1 min each) followed by a final 10 min incubation at 72°C were performed. The amplified DNA was precipitated with ethanol and ~0.5–1 pmol was used without further purification in the subsequent binding step. After eight selection cycles, an additional enrichment was performed by using the selected

populations as  $^{32}\text{P}$ -labeled probes and EMSA. The shifted bands obtained with 0.5 and 2 nM proteins were used in subsequent analyses.

### Selection of binding sites for RR<sub>TRES</sub>

The starting DNA pool (AT 586, N6 pool) containing six randomized bases had the central sequence –CATACAA-GAAAGNNNNNTTTATG– and the flanking PCR primer regions were identical with those shown above for AT500 (underlined). Four selection cycles based on EMSA were performed by using 2-fold serial protein dilutions in protein titrations. Shifted bands were isolated at protein concentrations when ~5–10% of the  $^{32}\text{P}$ -labeled DNA was shifted. These concentrations were gradually lower as the selection progressed: 1 nM in cycle 1, 0.4 nM in cycle 2, 0.1 nM in cycle 3, 12.5 and 25 pM in cycle 4.

### Cloning of the selected sequences and designed operators

The operator regions of the selected sequences were cloned into the pRIZ'O(–) vector by loop insertion mutagenesis as described previously (Chen *et al.*, 1997). The designed operators were obtained by annealing synthetic oligonucleotide pairs to form duplexes with 5'-TA overhangs which were cloned into the *Nde*I site of either pRIZ'O(–) (Simoncsits *et al.*, 1997) or pCP8 (Simoncsits *et al.*, 1999).

### Construction of single-chain repressors containing one or two mutant DBDs

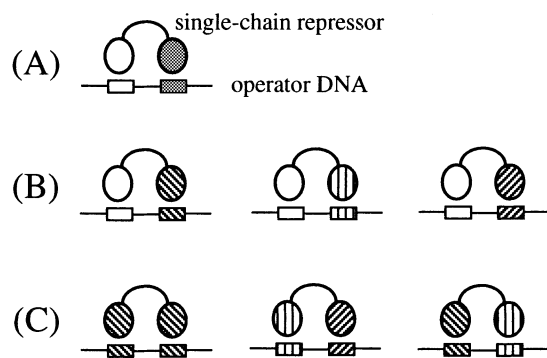
The genes coding for the RR<sub>TATG</sub>, RR<sub>TRPS</sub> and RR<sub>TRES</sub> mutants were cloned into pSET expression vector (Simoncsits *et al.*, 1997) as described (Simoncsits *et al.*, 1999), resulting in pSETRR<sub>TATG</sub>, pSETRR<sub>TRPS</sub> and pSETRR<sub>TRES</sub>. These vectors were used after *Xba*I–*Bam*HI cleavage to replace the coding region of the R wild-type domain with that of the R\* domain of pSETR\*R\*69 (Simoncsits *et al.*, 1997) to obtain pSETR\*R<sub>TATG</sub>, pSETR\*R<sub>TRPS</sub> and pSETR\*R<sub>TRES</sub>. The genes containing two selected mutant domains were also obtained in the pSET vector in two cloning steps. First, the pSETRR<sub>TATG</sub>, pSETRR<sub>TRPS</sub> and pSETRR<sub>TRES</sub> vectors were converted into pSETR<sub>TATG</sub>, pSETR<sub>TRPS</sub> and pSETR<sub>TRES</sub>, respectively, by complete *Eco*RI cleavage followed by vector re-ligation. These vectors were then cleaved with *Xho*I and *Hind*III and were ligated with the *Xho*I (partial)–*Hind*III fragments isolated from pSETR\*R<sub>TATG</sub>, pSETR\*R<sub>TRPS</sub> or pSETR\*R<sub>TRES</sub> to obtain the pSETR<sub>TATG</sub>R<sub>TATG</sub>, pSETR<sub>TATG</sub>R<sub>TRPS</sub>, pSETR<sub>TRPS</sub>R<sub>TRPS</sub> and pSETR<sub>TRES</sub>R<sub>TRES</sub> clones.

The genes coding for substitution mutants of RR<sub>TRPS</sub> and RR<sub>TATG</sub> were constructed by replacing the  $\alpha$ 3 helix coding region of the R\* domain in the pSETRR\*69 with synthetic *Kpn*I–*Xho*I linkers as described for the corresponding pRIZ' vectors (Simoncsits *et al.*, 1997, 1999).

## Results

### General experimental design and origin of the mutant DBDs

The sc derivatives of the 434 repressor contain tandem repeats of two DBDs of the natural repressor to form a single polypeptide chain of 158 amino acids (Percipalle *et al.*, 1995; Simoncsits *et al.*, 1997). The first 89 amino acid residues of the full 434 repressor are fused to a second copy of the first 69 amino acids to form the translational fusion (1–89)–(1–69). In this artificial protein, the two DBDs (residues 1–69) are joined by a peptide linker corresponding to the sequence of 70–89 residues of the full repressor. For simplicity, the prototype of this protein containing two wild-type DBDs



**Fig. 1.** General strategy for building sc proteins that recognize novel DNA target sites. (A) A library of sc repressors containing a mutant DBD (shaded oval) with randomized residues is selected for a target DNA subsite (shaded rectangle). (B) Cognate DBD-subsite pairs are isolated. (C) The mutant DBDs are combined to form novel sc homo- or heterodimers.

in the functional (1–69)–(70–89)–(1–69) arrangement was abbreviated as RR69, where R stands for the DBD and the suffix 69 indicated the length of the second repeat (Simoncsits *et al.*, 1997). Except for the first, designed mutant derivative RR\*69 (Chen *et al.*, 1997; Simoncsits *et al.*, 1997), this suffix is not used in the abbreviations of the selected (Simoncsits *et al.*, 1999) and constructed mutant hetero- and homodimeric sc molecules (this work).

The general scheme of constructing double-mutant sc variants of the 434 repressor with new DNA-binding specificities is shown in Figure 1. In the first step, sc repressor libraries containing one wild-type DBD (empty oval) and one mutant DBD (grey shaded oval) with randomized amino acids at certain, DNA-contacting positions are constructed. The libraries are then selected for interaction with a DNA operator composed of a subsite for the wild-type domain (empty rectangle) and an arbitrarily chosen target subsite (grey shaded rectangle) for the mutant domain. The protein selection experiments provide directly or after further specificity studies a set of mutant DBDs with characterized DNA-binding specificities, i.e. a set of cognate protein–DNA pairs is identified (see Figure 1B, where the components of the cognate pairs are identically striped). These mutant DBDs are finally combined to obtain novel sc molecules which are expected to recognize DNA operators composed of the cognate subsites of the corresponding DBDs (Figure 1C).

The mutant DBDs used in this work were obtained previously by rational design or by selection. The designed DBD (R\*) was obtained (Simoncsits *et al.*, 1997) from the wild-type domain (R) by substituting amino acid residues 27, 28, 29 and 32 as shown in Figure 2B. When this domain was joined to the wild-type domain, the heterodimeric sc repressor RR\*69 recognized the O<sub>R</sub>\*1 type operator sequences and the optimal DNA target of the R\* domain contained the TTAA sequence (Chen *et al.*, 1997) between the 4' and 1' operator positions as shown in Figure 2A. The selected DBDs were obtained from heterodimeric sc repressor libraries containing random mutations in one of the DBDs at the above-indicated positions after a genetic selection for the O<sub>R</sub>\*1 operator or for its 4'-TTAA-1' subsite (Simoncsits *et al.*, 1999). The heterodimeric, selected sc proteins were abbreviated as RR<sub>XXXX</sub>, where R is the wild-type domain and R<sub>XXXX</sub> is the mutant domain containing amino acid substitutions X at the randomized positions. While the *in vivo* selected DBDs were shown to bind the selection target *in vitro*, some of them could also bind to other sequences

(A)

	1	2	3	4	5	6	7	6'	5'	4'	3'	2'	1'	
O <sub>R</sub> 1	<u>A</u>	<u>C</u>	<u>A</u>	<u>A</u>	G	A	A	G	T	<u>T</u>	<u>T</u>	<u>G</u>	<u>T</u>	
O <sub>R</sub> 1-2'A1'C	<u>A</u>	<u>C</u>	<u>A</u>	<u>A</u>	G	A	A	G	T	<u>T</u>	<u>T</u>	<u>A</u>	<u>C</u>	
O <sub>R</sub> *1	<u>A</u>	<u>C</u>	<u>A</u>	<u>A</u>	T	A	A	A	A	C	<u>T</u>	<u>T</u>	<u>A</u>	<u>A</u>
O <sub>R</sub> *1-1'C	<u>A</u>	<u>C</u>	<u>A</u>	<u>A</u>	T	A	A	A	A	C	<u>T</u>	<u>T</u>	<u>A</u>	<u>C</u>

(B)

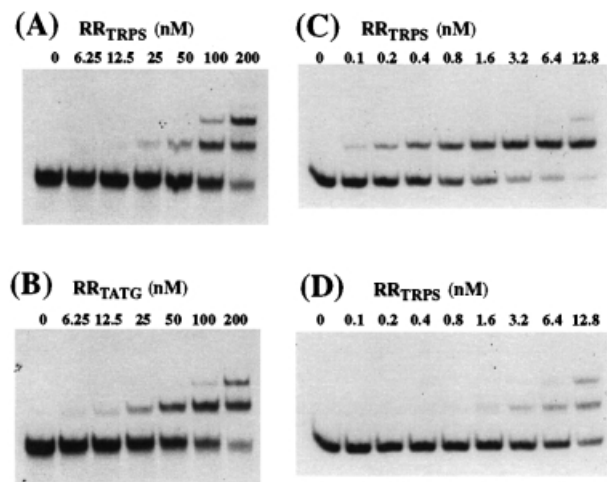
	-1	1	2	3	4	5	6
R	<b>T</b>	<b>Q</b>	<b>Q</b>	S	I	<b>E</b>	Q
R*	<b>S</b>	<b>N</b>	<b>V</b>	S	I	<b>S</b>	Q
R <sub>TATG</sub>	<b>T</b>	<b>A</b>	<b>T</b>	S	I	<b>G</b>	Q
R <sub>TRPS</sub>	<b>T</b>	<b>R</b>	<b>P</b>	S	I	<b>S</b>	Q
R <sub>TRES</sub>	<b>T</b>	<b>R</b>	<b>E</b>	S	I	<b>S</b>	Q
	<b>27</b>	<b>28</b>	<b>29</b>	30	31	<b>32</b>	33

**Fig. 2.** General numbering schemes of the operator bases (A) and of the amino acid residues (B) of the mutant sc 434 repressors. (A) O<sub>R</sub>1 is the natural O<sub>R</sub>1 of the 434 bacteriophage, O<sub>R</sub>\*1 is a designed hybrid operator (Hollis *et al.*, 1988; Simoncsits *et al.*, 1997) and examples for their derivatives are also given. For clarity, only one strand of the double-stranded operators is shown. The regions of O<sub>R</sub>1 (bases 1–4 and the corresponding palindromic 4'–1') which were shown to participate in direct, specific amino acid side chain–base pair contacts in the complex of O<sub>R</sub>1 and the N-terminal DBD (Aggarwal *et al.*, 1988) are underlined. The equivalent, contacted or putatively contacted regions (1–4 and 4'–1') of the listed operator derivatives are also underlined. (B) The amino acid sequences of the wild-type (R) and mutant DBDs used in this work are listed and the residues are numbered according to their positions in the  $\alpha$ 3 helix (above the residues) and in the full 434 repressor (under the residues). R\* is a designed DBD (Wharton and Ptashne, 1985; Simoncsits *et al.*, 1997); the other mutant DBDs were isolated previously (Simoncsits *et al.*, 1999). Residues in the mutated positions are shown in bold.

with even higher affinities. For example, the R<sub>TATG</sub> domain showed preference for the O<sub>R</sub>1 subsite TTGT, the R<sub>TRPS</sub> bound the TTAA target and its close homolog TTAC, while the R<sub>TRES</sub> bound the TTAC with high affinity and specificity (Simoncsits *et al.*, 1999). We chose these DBDs as building blocks to construct double-mutant sc proteins, and therefore their DNA-binding specificities were further studied by binding site selection from random ligand pools and by affinity studies.

#### Selection of binding sites for the R<sub>TRPS</sub>, R<sub>TATG</sub> and R<sub>TRES</sub> domains

Binding site selections for R<sub>TRPS</sub> and R<sub>TATG</sub> were performed by using the sc proteins RR<sub>TRPS</sub> and RR<sub>TATG</sub>, respectively and a random DNA pool containing the ACAATAAAANN-NNNNNNN sequence with nine randomized residues (N9 pool). The binding site of the R domain ACAA is underlined and the binding site of the mutant domain is expected to be 6 bp away from it, within the N9 region. Since this 6 bp spacer or non-contacted region strongly influences the operator binding affinities of the wild-type and mutant sc 434 repressors (Chen *et al.*, 1997; Simoncsits *et al.*, 1999), a major part (5 of 6 bp) of it was kept constant as an 'O<sub>R</sub>\*1-like' spacer (see Figure 2A) in order to make easier comparison between the sequences selected in the putative contacted regions. The selection conditions were not very stringent and this allowed for the isolation of both high and lower affinity binding sites. The average affinity of the random ligands was estimated to be around 100 and 200 nM (Figure 3A and B), while those of the selected ligand populations were at least 100-fold higher for the respective protein with relatively low cross-binding affinity (shown only for RR<sub>TRPS</sub> in Figure 3C and D).



**Fig. 3.** Selection of binding sites. EMSA performed with 2-fold serial protein dilutions (from right to left) shows binding site enrichment. (A) Unselected N9 library-RR<sub>TRPS</sub> interaction; (B) unselected N9 library-RR<sub>TATG</sub> interaction; (C) RR<sub>TRPS</sub> selected population-RR<sub>TRPS</sub> interaction; (D) RR<sub>TATG</sub> selected population-RR<sub>TRPS</sub> interaction.

In the ligand selection for RR<sub>TRES</sub>, a random DNA pool containing the ACAAGAAAGNNNNNNTTT sequence (N6 pool) was used. Here, higher affinity ligands can generally be expected owing to context effects. First, the O<sub>R</sub>1-like spacer (GAAAGN) confers high binding affinity for both wild-type and mutant sc 434 repressors, partly owing to the first G residue (Simoncsits *et al.*, 1999). Second, operators with NTTT flanking sequence were often isolated in previous selections (Chen *et al.*, 1997) and a 3-fold affinity increase compared with O<sub>R</sub>1 was observed with such an operator when tested with RR69 ( $K_d \approx 5$  pM, not shown). Correspondingly, the average affinity of the random N6 pool for RR<sub>TRES</sub> was ~4–8 nM (not shown). Four selection cycles were performed, but selected pools of earlier stages were also used to obtain lower affinity ligands.

#### DNA-binding specificities of the R<sub>TRPS</sub>, R<sub>TATG</sub> and R<sub>TRES</sub> domains: binding site selection reveals consensus operator regions

The selected DNA pools were cloned by loop insertion mutagenesis (Chen *et al.*, 1997). A number of clones were sequenced and their binding affinities for the corresponding protein were determined by EMSA. The results of the selection and affinity studies are summarized in Tables I (R<sub>TRPS</sub>), II (R<sub>TATG</sub>) and III (R<sub>TRES</sub>). When a certain sequence was obtained more than once, the numbers of occurrences are indicated (x). At the bottom part of each table, data obtained with several reference or designed operators are also included.

Generally, consensus sequences can be found between positions 4' and 1', which region is separated by a 6 bp spacer sequence from the ACAA binding site of the wild-type DBD. This region (underlined in the tables) is likely to be in direct contact with the mutant DBDs. At the right side of this region, A + T-rich sequences were selected (for R<sub>TRPS</sub> and R<sub>TATG</sub>), which is in accord with previous selections for the R\* domain binding sites (Chen *et al.*, 1997). The putative contacted regions generally contain the TT sequence in the 4'3' positions. For the R<sub>TRPS</sub> domain, TTAC and TTAA sequences (between the 4' and 1' operator positions) were selected most frequently, but several other sequences with high binding affinities were also obtained. Three main groups with TTTA, TTGA and

**Table I.** Sequences selected for RR<sub>TRPS</sub> and their binding affinities (nM)

Designation	Sequence	Affinity
N9 random	ACAATAAAAN <u>NNNN</u> NNNNATG	
a1	ACAATAAAAG <u>TTAC</u> TATCATG	0.310
a2	ACAATAAAAG <u>TTAC</u> ATTAAATG	
a3	ACAATAAAAG <u>TTAC</u> ATATATG	0.140
a4	ACAATAAAAA <u>TTAC</u> CATAATG	0.340
a5 2×	ACAATAAAAA <u>TTAC</u> ATTTATG	0.134
a6	ACAATAAAAA <u>TTAC</u> TTTGATG	0.331
a7	ACAATAAAAT <u>TTAC</u> ACAATG	0.116
a8	ACAATAAAAT <u>TTAC</u> TTCTATG	0.113
a9	ACAATAAAAT <u>TTAC</u> TTGTATG	
a10	ACAATAAAAT <u>TTAC</u> TGTTATG	
a11	ACAATAAAAT <u>TTAC</u> TATAATG	
a12	ACAATAAAAT <u>TTAA</u> TCTAATG	0.368
a13	ACAATAAAAT <u>TTAA</u> TATGATG	
a14	ACAATAAAAT <u>TTAA</u> TACGATG	
a15	ACAATAAAAA <u>TTAA</u> CTACATG	0.578
a16	ACAATAAAACT <u>TTAA</u> CATTATG	0.190
a17	ACAATAAAACT <u>TTAA</u> TTTAAATG	0.108
a18	ACAATAAAACT <u>TTAA</u> TCTAATG	0.230
a19	ACAATAAAAA <u>TTAT</u> CATAATG	2.88
a20	ACAATAAAAG <u>TTA</u> CATAATG	0.670
a21	ACAATAAAAA <u>TATA</u> TGTAATG	3.75
a22	ACAATAAAAG <u>TATA</u> TTTTATG	0.600
a23 2×	ACAATAAAACT <u>TATC</u> TTGTATG	0.140
a24	ACAATAAAACT <u>TAAC</u> ATTAAATG	0.065
a25	ACAATAAAAA <u>TTAC</u> ATATATG	0.487
a26	ACAATAAAAG <u>TAAA</u> TATGGTG	
a27	ACAATAAAAG <u>TAAA</u> TTATATG	1.10
a28	ACAATAAAAG <u>GTTAA</u> ACTTATG	
a29	ACAATAAAAT <u>TTAC</u> TATGATG	7.91
a30	ACAATAAAAG <u>TAT</u> TTAACATG	
O <sub>R</sub> *1	ACAATAAAACT <u>TTAA</u> TATG	0.770
O <sub>R</sub> 1	ACAAGAAAG <u>TTTGT</u> TATG	4.14
O <sub>R</sub> *1-1'C	ACAATAAAACT <u>TTAC</u> ATATG	0.025
O <sub>571</sub>	ACAAGAAACT <u>TTAC</u> ATTATG	0.007

TTGT sequences were obtained for R<sub>TATG</sub>, while the R<sub>TRES</sub> selections resulted in two high-affinity groups with TTAC and TTCC sequences. For a better understanding of the selection results, they were complemented by using the following techniques: (i) binding with designed operator sequences to clarify certain context effects, mainly the effect of the base at the operator 5' position on the binding affinity, (ii) by studying the effect of amino acid substitutions of putatively contacting residues on the binding affinities, (iii) methylation protection of DNA to identify amino acid-bp contacts.

#### Identification of the putatively contacted operator subsites at the 4' to 1' positions and effect of the 5' base on the binding affinity

The major recognition sites of the R<sub>TRPS</sub> domain are TTAA and TTAC (Table I). This domain was originally selected *in vivo* for the TTAA target of the O<sub>R</sub>\*1 operator (Simoncsits *et al.*, 1999). The TTAA containing sequences obtained in this study are higher affinity binding ligands than O<sub>R</sub>\*1 (a12–a18, Table I). The other major group (TTAC sequences) also contained high or even higher affinity ligands (a1–a11) and it was also noted that the rarely selected TATC and TAAC containing ligands (a23–a25) were also high binders. While the highest binders in this small group (and in the TTAA group) contained C at the 5' operator position, this residue was not found in the TTAC group. To test the role of this residue, a set of operators containing all four possible bases in the 5' position was compiled for the TTAC, TAAC and

**Table II.** Sequences selected for RR<sub>TATG</sub> and their binding affinities (nM)

Designation	Sequence	Affinity
	5'4'3'2'1'	
N9 random	ACAATAAAAANNNNNNNNNATG	
b1	ACAATAAAACTTTAATTTATG	0.49
b2	ACAATAAAAGTTTAATTTATG	0.42
b3	ACAATAAAAGTTTAATACATG	
b4	ACAATAAAAGTTTAGTATATG	
b5	ACAATAAAACTTTAGATAATG	
b6	ACAATAAAAGTTTAATAATATG	
b7	ACAATAAAAGTTTAATTATATG	0.27
b8	ACAATAAAAGTTGATAATATG	
b9	ACAATAAAAGTTGATATTATG	0.16
b10	ACAATAAAAGTTGAGTCTATG	0.29
b11	ACAATAAAAATTGAATTTATG	1.44
b12	ACAATAAAAATTGTTATAATG	0.74
b13	ACAATAAAAGTTGTTTATATG	0.32
b14	ACAATAAAAGTTGTTATTAATG	0.27
b15 2×	ACAATAAAACTTGTAAATAATG	
b16 4×	ACAATAAAACTTGTAAATATG	0.11
b17 3×	ACAATAAAACTTGTAAATATG	
b18	ACAATAAAACTTATATCTATG	1.53
b19	ACAATAAAACTAGTAAAAATG	1.50
b20	ACAATAAAAATTATCATTATG	5.00
b21	ACAATAAAAGTATATTTAATG	0.52
b22	ACAATAAAACTCGTATTTATG	1.60
O <sub>R</sub> 1	ACAAGAAAGTTTGTATATG	0.140
O <sub>R*</sub> 1	ACAATAAAACTTTAAATATG	6.00
O <sub>R*</sub> 1-2'T	ACAATAAAAGTTTAATATG	0.60
O <sub>R*</sub> 1-2'G	ACAATAAAACTTGAATATG	0.82
O <sub>R*</sub> 1-1'T	ACAATAAAACTTTATATATG	1.24
Ocl(1)	ACAAGAAAACTTGTATTTTGTG	0.031

**Table III.** Sequences selected for RR<sub>TRES</sub> and their binding affinities (pM)

Designation	Sequence	Affinity
	5'4'3'2'1'	
N9 random	ACAAGAAAGNNNNNNNTTTATG	
c1 3×	ACAAGAAAGTTTACGTTTATG	1.02
c2 3×	ACAAGAAAGTTTACATTTATG	1.35
c3 5×	ACAAGAAAGTTTACCTTTATG	2.04
c4 3×	ACAAGAAAGTTTACTTTTATG	3.37
c5 2×	ACAAGAAAGCTTACGTTTATG	1.41
c6 6×	ACAAGAAAGCTTACATTTATG	2.50
c7 4×	ACAAGAAAGCTTACCTTTATG	2.94
c8 3×	ACAAGAAAGCTTACTTTTATG	3.08
c9 3×	ACAAGAAAGCTTACGTTTATG	2.94
c10 2×	ACAAGAAAGCTTACATTTATG	8.63
c11	ACAAGAAAGCTTACATTTATG	10.7
c12 4×	ACAAGAAAGCTTCCATTTATG	10.0
c13	ACAAGAAAGCTTCCGTTTATG	6.70
c14	ACAAGAAAGCTTCCATTTATG	5.80
c15 2×	ACAAGAAAGCTACCCTTTATG	46.0
c16 3×	ACAAGAAAGCTAACGTTTATG	80.0
c17	ACAAGAAAGCTAACATTTATG	
c18	ACAAGAAAGATCACCTTTATG	850
c19 2×	ACAAGAAAGCTCACGTTTATG	
c20	ACAAGAAAGCTCACCTTTATG	
c21 3×	ACAAGAAAGCTTATCGTTTATG	850
c22	ACAAGAAAGCTTTCCTTTATG	
O <sub>R</sub> 1	ACAAGAAAGTTTGTATATG	1200
O <sub>R*</sub> 1	ACAATAAAACTTTAAATATG	2120
O <sub>R*</sub> 1-1'C	ACAATAAAACTTTACATATG	27.0
O <sub>571</sub>	ACAAGAAAACTTACATTTATG	2.80

**Table IV.** Effect of the 5' operator base (bold) on the binding affinity of RR<sub>TRPS</sub> for the TTAC, TAAC and TATC subsites (underlined)

Sequence	Affinity (pM)
- <b>CTTACA</b> -	25
- <b>TTTACA</b> -	116
- <b>GTTACA</b> -	140
- <b>ATTACA</b> -	134
- <b>CTAACA</b> -	65
- <b>TTAACA</b> -	160
- <b>GTAACA</b> -	176
- <b>ATAACA</b> -	487
- <b>CTATCA</b> -	57
- <b>TTATCA</b> -	676
- <b>GATCA</b> -	284
- <b>ATATCA</b> -	706

TATC sequences by complementing the selected sequences with synthetic ones. In these collections, the other flanking base (downstream of the shown tetramer sequences) or a short flanking region was generally constant. Comparison of the binding affinities showed a general, strong preference for C at the 5' operator position (Table IV) in all three groups and the optimal pentamer sequence between the 5' and 1' positions was found to be CTTAC. The binding affinity could further be increased by performing a symmetrical change at the 5 position, i.e. by introducing a G residue next to the operator subsite (ACAA) contacted by the wild-type R domain of RR<sub>TRPS</sub> (compare O<sub>R\*</sub>1-1'C and O<sub>571</sub> in Table I).

The data in Table II show that the optimal operator subsite of the R<sub>TATG</sub> domain is among the TTTA, TTGA and TTGT tetramers. The affinities are not very high and a clear preference for any of these sites, even after considering the context effect by using several designed analogs (bottom part of the Table II), could not be observed. The highest affinity was observed with the Ocl(1) operator, which was selected for the wild-type domain (Chen *et al.*, 1997). It should be noted that the TTGT subsite represents the specificity of the R domain, therefore the TTGT containing operators can form two differently oriented complexes with RR<sub>TATG</sub>. Thus the affinity data shown in Table II probably overestimate the intrinsic affinity of the R<sub>TATG</sub> domain for the TTGT subsite. These results together with binding data obtained with several substitution analogs of RR<sub>TATG</sub> (see below) show that R<sub>TATG</sub> is a relaxed specificity mutant.

The high-affinity ligands obtained in the RR<sub>TRES</sub> selection (Table III) contained the TTAC or the TTCC subsite. The affinities for the TTAC sequences were several-fold higher than those for the TTCC sequences when they were compared in identical sequence contexts (see the c1/c13, c2/c12 and c6/c14 pairs). Several other sequences containing the consensus 4'T and 1'C residues (c15-c22) exhibited substantially lower affinities. The best ligand in this group contained the TACC sequence (c15 in Table III), but its affinity was still about 30-fold lower than that of the corresponding (with the same flanking bases) TTAC ligand (c5). The affinities varied from 1 to 10 pM in the TTAC group (c1-c11) which again could be due to context effects. Examples in Table III show that the preferred 5' flanking residue is T or C, whereas the preferred residue at the other side of TTAC is G or A. Similar preference for 5' C was also observed (data not shown) when the operators contained the O<sub>R\*</sub>1-like spacer sequence (see O<sub>R\*</sub>1-1'C in

**Table V.** Effect of substitutions of Arg28 of RR<sub>TRPS</sub> on the binding affinities (nM) to O<sub>R</sub>\*1 operator derivatives containing TTAN subsites

Subsite	Protein		
	RR <sub>TRPS</sub>	RR <sub>TAPS</sub>	RR <sub>TGPS</sub>
-TTAC-	0.025	0.22	0.25
-TTAA-	0.77	0.11	0.10
-TTAT-	5.2	0.13	0.12
-TTAG-	>10.0	0.92	0.25

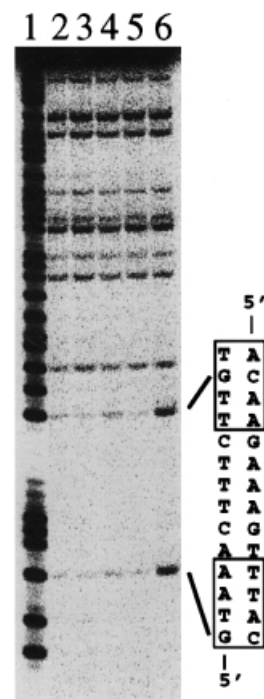
Table III). Introducing G at the symmetrical 5 position again caused a significant affinity increase (see O<sub>571</sub> in Table III).

#### Identification of possible amino acid–base pair contacts

To understand the results of the selection and binding affinity studies, we attempted to delineate possible interactions between certain amino acid residues of the mutant DBDs and the selected consensus operator sites. In one approach, the putative DNA-contacting residues were substituted and the effects on the binding affinity for a set of operators were tested. In the other, a footprinting technique was used to identify the contacted base which was protected in the complex against chemical modification. In addition, the data from a previous specificity study (Simoncsits *et al.*, 1999) obtained with other protein and operator mutants were also used in the evaluations.

It was suggested previously (Simoncsits *et al.*, 1999) that the Arg28 residue in the R<sub>TRPS</sub>, R<sub>TRES</sub> and other similar domains could contribute significantly to the binding affinity by forming a contact with the G residue of the 1'C–G base pair of the TTAC subsite. Such a contact is frequently observed in protein–DNA complexes (Seeman *et al.*, 1976; Pabo and Sauer, 1992; Suzuki, 1994; Mandel-Gutfreund *et al.*, 1995) and may explain the preferential recognition of TTAC over TTAA by such domains. We used the principle of the 'loss of contact' approach (Ebright, 1991) and constructed mutants of the R<sub>TRPS</sub> domain by substituting the Arg28 residue by Ala and by Gly to obtain RR<sub>TAPS</sub> and RR<sub>TGPS</sub>. The effects of these substitutions were tested by using a set of operators containing TTAN subsites, where N is A, C, G or T. The results (Table V) show that both mutants bound to the TTAC operator with about 10-fold reduced affinity and they also lacked the ability to discriminate the 1' base pairs. At the same time, these interactions were relatively strong, indicating that important contacts may be maintained or become even more pronounced between other, unchanged residues. For example, the Pro29 residue may make hydrophobic contacts with the 3' T–A and 2' A–T base pairs in a manner similar to that proposed for the Tet repressor (Baumeister *et al.*, 1992). Support for this assumption can be provided by comparing the data from previous affinity studies (Simoncsits *et al.*, 1999) which showed that RR<sub>TRPS</sub> bound the TTAA and TTAC sequences with higher affinities than its RR<sub>TRVS</sub> and RR<sub>TRSS</sub> homologs and that RR<sub>TRPS</sub>, compared with these homologs, also exhibited a significantly stronger preference for TTAC within the TTNC ligand series. It was also observed in other studies (Y.Lin, J.Gál and A.Simoncsits, unpublished data) that further substitutions of Pro29 of RR<sub>TAPS</sub> (by Val or Ala) caused a significant affinity decrease for the TTAA ligands (not shown).

Similar substitution mutants of the R<sub>TATG</sub> domain were also constructed. Changing Ala28 to Gly and Thr29 to Gly resulted in RR<sub>TGTG</sub> and RR<sub>TAGG</sub> proteins, respectively, which were

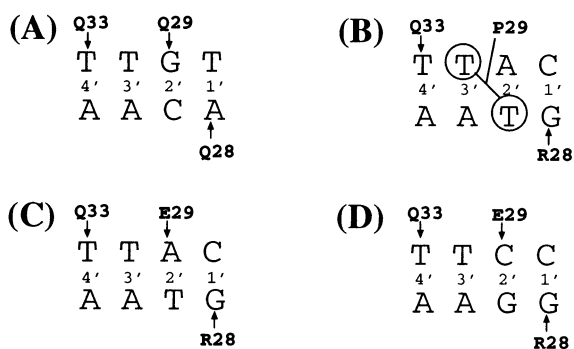


**Fig. 4.** Footprint analysis showing protection of G residues from methylation by DMS. Lane 1, G + A reaction; lanes 2 and 3, RR<sub>TRPS</sub> complex (0.5 and 5 nM protein); lanes 4 and 5, RR<sub>TRES</sub> complex (0.5 and 5 nM); lane 6, no protein. Boxed regions indicate subsites for the R (upper) and the mutant DBDs (lower box).

tested for binding with the b16 and the designed O<sub>R</sub>\*1 derivatives of Table II. It was observed that the A28G substitution caused either a slight (up to 2-fold) increase or no change in the affinities, whereas the T29G mutation caused a general (5–10-fold) decrease. Thus Thr29 plays a certain role in the recognition of the 2'1' residues, but its combination with Ala28 results in relatively low affinity, relaxed specificity binding.

The present study confirms the results of previous specificity studies with the R<sub>TRES</sub> domain (Simoncsits *et al.*, 1999), which suggested that the optimal binding site is TTAC followed by TTCC. We suppose that a contact between Arg28 and the 1'C–G base pair, as suggested for the R<sub>TRPS</sub> domain above, also exists in the R<sub>TRES</sub> interactions with both the TTAC and TTCC subsites. The Glu29 residue probably accepts an H-bond from either the 2'A or 2'C residue. This assumption could be supported by using Glu29 substitution analogs. Several mutants of the R<sub>TRES</sub> domain, R<sub>TRPS</sub>, R<sub>TRVS</sub> and R<sub>TRSS</sub>, were available and binding data obtained with them showed that the Val29 and Ser29 substitutions resulted in 4- and 10-fold reduced affinity for the TTAC subsite and all three substitutions lead to at least a 20-fold affinity decrease for the TTCC subsite (Simoncsits *et al.*, 1999). These data indicate the loss of a favourable contact between Glu29 and the 2'A or 2'C residue.

Methylation protection by dimethyl sulfate (DMS) was used to confirm the postulated contact between the G residue of the operator 1'C–G pair and the Arg28 of the R<sub>TRPS</sub> and R<sub>TRES</sub> domains. Figure 4 shows that the extent of this protection in both interactions is comparable to that observed for the 2G residue in the TTGT box, and this G is shown to be in contact with Gln29 of the wild-type domain in the O<sub>R</sub>1 complex (Aggarwal *et al.*, 1988). The major protein–DNA contacts



**Fig. 5.** Major amino acid–bp contacts observed in the wild-type DBD–operator complex (A) and proposed for the interactions of  $R_{TRPS}$  (B) and  $R_{TRES}$  (C, D) domain with their respective operator subsites.

observed in the wild-type complex and proposed for the mutant  $R_{TRPS}$  and  $R_{TRES}$  domain interactions are shown in Figure 5.

#### Construction of sc proteins containing two mutant DBDs and characterization of their DNA-binding properties

The above-characterized  $R_{TRPS}$ ,  $R_{TATG}$  and  $R_{TRES}$  domains were used together with the previously designed  $R^*$  domain as building blocks to construct several homo- and heterodimeric sc proteins. These proteins were purified and their DNA-binding properties were tested by using a collection of reference, selected or newly designed operators (the latter are labeled with subscript numbers) as shown in Table VI. The operator subsites which are either cognate or not to a given domain combination are underlined. These sites are separated by a spacer sequence of 6 bp, the sequence of which was conserved within this group. The affinity data are underlined when they are considered to represent cognate interactions regarding both of the protein domain–operator subsite pairs. The list of these cognate subsites is TTAA and TTTA for the  $R^*$ , TTAC and TTAA for the  $R_{TRPS}$ , TTGT, TTGA and TTTA for the  $R_{TATG}$  and TTAC for the  $R_{TRES}$  domain (operators containing the TTCC subsite of  $R_{TRES}$  are shown in Table VII). Owing to operator symmetry, the corresponding palindromic sequences are shown for the left subsite of the operators.

The data in Table VI show that the subsite recognition preferences of the mutant domains, observed in combination with the wild-type R domain, are generally maintained in novel combinations with other mutant domains. The interpretation of the binding affinity data may be complicated in certain cases when the domains of the double mutant protein have overlapping specificities and a common subsite is present in the test operator. For easier data comparison, several data obtained with  $RR_{TATG}$  and  $RR_{TRPS}$  are also included in Table VI. Comparison of data in the  $RR_{TATG}$  and  $R_{TATG}R_{TATG}$  columns shows qualitatively similar tendencies. It is also shown that  $R_{TATG}R_{TATG}$  prefers the operators containing two of the  $R_{TATG}$  binding subsites over those which contain only one, but this homodimer does not seem to be capable of high-affinity (subnanomolar) interactions. The  $R_{TATG}$  domain can, however, participate in higher affinity interactions in the  $R^*R_{TATG}$  and  $R_{TATG}R_{TRPS}$  heterodimers. The highest affinity ligands for  $R^*R_{TATG}$  are  $O_{528c}$  and  $O_{538}$ . The data obtained with  $R^*R_{TATG}$  also reflect the  $R^*$  domain preference for TTAA over TTTA (compare  $O_{528c}$  and  $O_{528b}$ ) subsite. The  $R_{TATG}R_{TRPS}$  heterodimer bound most strongly to the  $O_{550}$ ,  $O_{568}$  and  $O_{571}$  operators, which agrees well with the established subsite recognition preferences of both domains and also the previously observed

effects of the 5' or 5' operator base on the binding affinity. Comparison of the data for common ligands in the  $RR_{TRPS}$  and  $R_{TATG}R_{TRPS}$  columns also shows that the  $R_{TRPS}$  domain exhibits similar subsite preferences in these two sc molecules. Both the  $R^*R_{TRPS}$  heterodimer and the  $R_{TRPS}R_{TRPS}$  homodimer were capable of high-affinity binding to operators containing the optimal subsites of the corresponding DBDs ( $O_{546A}$  for  $R^*R_{TRPS}$  and  $O_{565}$  for  $R_{TRPS}R_{TRPS}$ ).

Generally, the sc molecules containing two mutant DBDs exhibited lower affinities than the corresponding heterodimers containing one wild-type DBD. This was also the case when the  $R_{TRES}$  domain was combined with  $R^*$ , but the  $R_{TRES}R_{TRES}$  homodimer exhibited very strong binding to some of the test operators in Table VI. Several operators containing TTAC and/or TTCC subsites were constructed for  $R_{TRES}R_{TRES}$  and the affinities are summarized in Table VII. High-affinity binding was observed with operators containing any combination of these subsites (two TTAC or two TTCC or one TTAC and one TTCC) connected by GAAAGT (as in  $O_{R1}$ ) or GAAAAN type spacers. The data also show previously observed preferences for 5'C over 5'G in both the TTAC and the TTCC operator groups.

The role of the central, non-contacted operator bases in the high affinity binding has also been demonstrated for sc molecules containing two mutant DBDs (Table VI). When the 7' base was changed from A to C, a significant affinity decrease was observed for both hetero- and homodimeric mutants (compare operator pairs  $O_{538}$ – $O_{445}$  and  $O_{540}$ – $O_{498}$  for  $R^*R_{TATG}$ ,  $R^*R_{TRPS}$  and  $R^*R^*$  interactions).

## Discussion

Artificial DNA-binding proteins that recognize long sequences and exhibit novel binding specificities have been constructed mainly from the Cys2His2 type of zinc finger motifs by using design and selection principles (Desjarlais and Berg, 1993; Choo and Klug, 1994; Jamieson *et al.*, 1994; Rebar and Pabo, 1994; Wu *et al.*, 1995; Greisman and Pabo, 1997). In a simplified view, the Zn fingers recognize contiguous 3 bp subsites or overlapping 4 bp subsites, therefore a three-finger protein can target 9–10 bp long sequences. Recognition of even longer (continuous or/and discontinuous) sequences could be achieved by linking the Zn fingers to other DBDs (Pomerantz *et al.*, 1995; Kim *et al.*, 1997) or by linking two three-finger proteins (Liu *et al.*, 1997; Kim and Pabo, 1998). Other DNA-binding motifs, such as the helix–turn–helix (HTH) motif, have received much less attention.

In this work, we used mutant DBDs of the phage 434 repressor, which belongs to the HTH protein family, as building blocks to construct sc proteins that recognize relatively long (up to 14 bp) DNA sequences with high affinity and specificity. Several sc mutants containing one wild-type and one mutant domain were used to determine the subsite recognition specificities of the mutant DBDs. These sc proteins were shown to recognize the ACAA–6 bp–NNNN general sequence where the NNNN subsite is contacted by and characteristic for the mutant domain. Four characterized mutant DBDs ( $R^*$ ,  $R_{TATG}$ ,  $R_{TRPS}$  and  $R_{TRES}$ ) were linked in several combinations to obtain homo- and heterodimeric sc proteins, which were shown to recognize the NNNN–6 bp–NNNN general operator sequence. The mutant DBDs exhibited their subsite specificities in all tested novel combinations and the strong preference for the 5' or 5' operator base pair (the outer bases of the 6 bp spacer) was also maintained in the  $R_{TRPS}$  and  $R_{TRES}$  interactions.

**Table VI.** Summary of the DNA-binding affinities (nM) of mutant homo- or heterodimeric single-chain 434 repressors

Operator designation	Sequence	Protein										
		RR <sub>TATG</sub>	RR <sub>TRPS</sub>	R <sub>TATG</sub> R <sub>TATG</sub>	R* <sub>R</sub> T <sub>ATG</sub>	R <sub>TATG</sub> R <sub>TRPS</sub>	R* <sub>R</sub> T <sub>TRPS</sub>	R <sub>TRPS</sub> R <sub>TRPS</sub>	R* <sub>R</sub> T <sub>TRPS</sub>	R <sub>TATG</sub> R <sub>TATG</sub>	R* <sub>R</sub> T <sub>TRPS</sub>	R* <sub>R</sub> T <sub>TRES</sub>
Or1	CATAGAAAGAAAGTTTGTTATG	0.14	4.14	4.35	5.8	>21.0	>40.0	>40.0	>40.0			
Or*1	CATACAAATAAAACTTTAAATATG	6.00	0.770	19.0	0.62	6.4	8.0					
b2	CATAGAAATAAAAGTTTAAATTTATG	0.42	4.00	3.85	1.20							
b9	CATACAAATAAAAGTTTGATATTTATG	0.16		1.71	>12.8							
Oc1(1)	CATAGAAAGAAACTTGTATTTTGTG	0.031		1.25								
O528a	CATATTCAGAGAAAGTTTGAATTTATG			1.50	1.61							
O528b	CATATTCAGAGAAAGTTTAAATTTATG			1.60	0.358							
O528c	CATATTCAGAGAAAGTTTAAATTTATG			6.00	0.108							
Or*1-C	CATAGAAATAAAACTTTACATATG		0.025			4.5						
O571	CATACAGAAACCTTACATTTATG		0.007			0.400						
a1	CATAGAAATAAAAGTTTACTATCATG		0.310			0.280						
a16	CATAGAAATAAAACCTTAAACATTTATG		0.190			2.00						
a17	CATAGAAATAAAACCTTAAATTTAATG		0.108			1.76						
O546C	CATATTCAGAGAAAGTTTACATTTATG					0.880						
O546A	CATATTCAGAGAAAGTTTAAATTTATG								0.220			0.100
O550	CATATTCAGAGAAAGTTTAAATTTATG							0.142	0.550			
O568	CATATTCAGAGAAAGTTTAAATTTATG					0.183						
O538	CATATTCAGAGAAAGTTTAAATTTATG					0.208						
O445	CATATTCAGAGAAACCTTAAATTTATG					2.50						
O540	CATATTCAGAGAAACCTTAAATTTATG							0.360	4.20		0.133	
O498	CATATTCAGAGAAACCTTAAATTTATG					12.4		3.75	2.30		0.490	
O565	CATATTCAGAGAAACCTTAAATTTATG							2.60	0.170		0.460	
								2.74				0.008



**Table VII.** Binding of the R<sub>TRES</sub>R<sub>TRES</sub> homodimer to designed operators containing combinations of TTAC and TTCC subsites

Operator designation	Sequence	Affinity (pM)
<i>Two TTAC subsites:</i>		
O <sub>682</sub>	TAAATGTAAGAAAAGTTTACGTTT	4.8
O <sub>684T</sub>	TAAATGTAAGAAAACCTTACGTTT	5.6
O <sub>565T</sub>	ATATTGTAAGAAAACCTTACATTG	8.0
O <sub>546C</sub>	ATATTGTAAGAAAAGTTTACATTA	19.0
<i>Two TTCC subsites:</i>		
O <sub>664C</sub>	TAAATGGAAGAAAACCTTCCGTTT	7.4
O <sub>654</sub>	TAAATGGAAGAAAACCTTCCATTT	10.0
O <sub>664G</sub>	TAAATGGAAGAAAAGTTTCCGTTT	36.0
O <sub>684</sub>	TAAATGGAAGAAAAGTTTCCGTTT	13.0
<i>One TTAC and one TTCC subsite:</i>		
O <sub>648G</sub>	TAAATGGAAGAAAACCTTACGTTT	7.7

The two outer pentamers, a total 10 of the 14 bp sequence are therefore the major determinants of the sequence specificity. Previous studies also showed that the central four or non-contacted, operator base pairs influenced the binding affinities of both the natural (Koudelka *et al.*, 1987) and the sc repressors (Chen *et al.*, 1997) and the spacers of the affinity selected ligands generally contained either alternating A–T/T–A pairs or runs of at least three A–T pairs in these positions (Chen *et al.*, 1997). The sc 434 repressors containing two mutant DBDs seem to share these properties: several such homo- and heterodimeric proteins of this study showed significantly weaker binding when one of the central (7 or 7') operator bases was changed from A to C. Thus, all 14 bp of the operator are important in the interaction with mutant sc repressors. In addition, previous (Chen *et al.*, 1997) and the present binding site selection results revealed that the mutant derivatives also prefer the A + T-rich operator flanking regions. DNA recognition by the 434 repressor itself is a result of direct and indirect mechanisms. The double-mutant sc derivatives in this work, similar to the mutant prototype RR\*69 (Chen *et al.*, 1997), seem to combine the characteristic indirect effects with altered specificity direct readout in their DNA recognition mechanism.

The mutant DBDs, apart from the relaxed specificity R<sub>TATG</sub>, were able to form high-affinity sc proteins in combination with the wild-type R domain: the RR\*69, RR<sub>TRPS</sub> and RR<sub>TRES</sub> heterodimers showed half-maximal binding to their corresponding optimal operators at or under 10 pM concentration, which is comparable to the data observed for the 'wild-type' RR69. When the mutant DBDs were combined, the affinities of the double-mutant sc derivatives were generally in the range 100–200 pM. These affinities are lower than expected and of the tested combinations, only the homodimeric R<sub>TRES</sub>R<sub>TRES</sub> exhibited high-affinity binding (5–20 pM). The wild-type 434 DBD–operator complexes show a network of interactions besides the direct contacts between the  $\alpha 3$  'recognition' helix and the operator subsites and suggest important protein–protein interactions between the DBDs (Aggarwal *et al.*, 1988; Rodgers and Harrison, 1993; Shimon and Harrison, 1993). The mutations introduced into the  $\alpha 3$  helix may differentially influence these interactions, including those at the interface of the DBDs and thereby the cooperativity of DBDs in DNA binding. It is reasonable to suppose that different DBD pairs cooperate to different extents and that the flexibilities of the different test operators also influence the cooperative binding

process. Owing to such effects, quantitative interpretations of the binding data are complicated. Nevertheless, this work shows that altered specificity mutant DBDs of the 434 repressor can be combined in the sc arrangement to engineer extended recognition surfaces of expected, novel specificities.

## Acknowledgements

We thank I.Törö and C.Guarnaccia for help with protein purifications. T.L. and J.C. were supported by fellowships from ICGEB and SISSA, respectively. M.-L.T. and A.S. are on leave from the The Arrhenius Laboratories for Natural Sciences, Stockholm University, Stockholm, Sweden.

## References

- Aggarwal,A.K., Rodgers,D.W., Drottar,M., Ptashne,M. and Harrison,S.C. (1988) *Science*, **242**, 899–907.
- Baumeister,R., Helbl,V. and Hillen,W. (1992) *J. Mol. Biol.*, **226**, 1257–1270.
- Chen,J., Pongor,S. and Simoncsits,A. (1997) *Nucleic Acids Res.*, **25**, 2047–2054.
- Choo,Y. and Klug,A. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 11168–72.
- Desjarlais,J.R. and Berg,J.M. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 2256–2260.
- Ebright,R.H. (1991) *Methods Enzymol.*, **208**, 620–640.
- Gates,C.M., Stemmer,W.P.C., Kaptein,R. and Schatz,P.J. (1996) *J. Mol. Biol.*, **255**, 373–386.
- Gill,S.C. and von Hippel,P.H. (1989) *Anal. Biochem.*, **182**, 319–326.
- Greisman,H.A. and Pabo,C.O. (1997) *Science*, **275**, 657–661.
- Hollis,M., Valenzuela,D., Pioli,D., Wharton,R. and Ptashne,M. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 5834–5838.
- Jamieson,A.C., Kim,S.H. and Wells,J.A. (1994) *Biochemistry*, **33**, 5689–5695.
- Jana,R., Hazbun,T.R., Fields,J.D. and Mossing,M.C. (1998) *Biochemistry*, **37**, 6446–6455.
- Kim,J.-S. and Pabo,C.O. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 2812–2817.
- Kim,J.-S., Kim,J., Cepek,K.L., Sharp,P.A. and Pabo,C.O. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 3616–3620.
- Koudelka,G.B., Harrison,S.C. and Ptashne,M. (1987) *Nature*, **326**, 886–888.
- Kuntz,M.A. and Shapiro,D.J. (1997) *J. Biol. Chem.*, **272**, 27949–27956.
- Liang,H., Sandberg,W.S. and Terwilliger,T.C. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 7010–7014.
- Liu,Q., Segal,D.J., Ghiara,J.B. and Barbas,C.F.,III. (1997) *Proc. Natl Acad. Sci. USA*, **94**, 5525–5530.
- Mandel-Gutfreund,Y., Schueler,O. and Margalit,H. (1995) *J. Mol. Biol.*, **253**, 370–382.
- Pabo,C.O. and Sauer,R.T. (1992) *Annu. Rev. Biochem.*, **61**, 1053–1095.
- Percipalle,P., Simoncsits,A., Zakhariiev,S., Guarnaccia,C., Sanchez,R. and Pongor,S. (1995) *EMBO J.*, **14**, 3200–3205.
- Pomerantz,J.L., Sharp,P.A. and Pabo,C.O. (1995) *Science*, **267**, 93–96.
- Rebar,E.J. and Pabo,C.O. (1994) *Science*, **263**, 671–673.
- Rhodes,D. and Fairall,L. (1997) In Creighton,T.E. (ed.), *Protein Function: a Practical Approach*. 2nd edn. IRL Press, Oxford, pp. 215–244.
- Robinson,C.R. and Sauer,R.T. (1996a) *Biochemistry*, **35**, 109–116.
- Robinson,C.R. and Sauer,R.T. (1996b) *Biochemistry*, **35**, 13878–13884.
- Robinson,C.R. and Sauer,R.T. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 5929–5934.
- Rodgers,D.W. and Harrison,S.C. (1993) *Structure*, **1**, 227–240.
- Ruiz-Sanz,J., Simoncsits,A., Törö,I., Pongor,S., Mateo,P.L. and Filimonov,V.V. (1999) *Eur. J. Biochem.*, **263**, 246–253.
- Seeman,N.C., Rosenberg,J.M. and Rich,A. (1976) *Proc. Natl Acad. Sci. USA*, **73**, 804–808.
- Shimon,L.J. and Harrison,S.C. (1993) *J. Mol. Biol.*, **232**, 826–838.
- Sieber,M. and Alleman,R.K. (1998) *Nucleic Acids Res.*, **26**, 1408–1413.
- Simoncsits,A., Chen,J., Percipalle,P., Wang,S., Törö,I. and Pongor,S. (1997) *J. Mol. Biol.*, **267**, 118–131.
- Simoncsits,A., Tjörnhammar,M.-L., Wang,S. and Pongor,S. (1999) *Nucleic Acids Res.*, **27**, 3474–3480.
- Suzuki,M. (1994) *Structure*, **2**, 317–326.
- Wharton,R.P. and Ptashne,M. (1985) *Nature*, **316**, 601–605.
- Wu,H., Yang,W.P. and Barbas,C.F.,III. (1995) *Proc. Natl Acad. Sci. USA*, **92**, 344–348.

Received June 8, 2000; revised April 17, 2001; accepted May 14, 2001