ISAS - INTERNATIONAL SCHOOL FOR ADVANCED STUDIES

CONSTRUCTION AND CHARACTERIZATION OF SINGLE-CHAIN DNA-BINDING PROTEINS DERIVED FROM PHAGE 434 REPRESSOR

Thesis submitted for the degree
"Doctor Philosophiae"

CANDIDATE

Jinqiu CHEN

SUPERVISORS

András SIMONCSITS, Ph.D.
Sándor PONGOR, Ph.D., D.Sc.

Academic Year 1996-1997

SISSA - SCUOLA INTERNAZIONALE SUPERIORE DI STUDI AVANZATI
TRIESTE
Strada Costiera 11

TRIESTE

CONTENTS

Abstract

Single-chain DNA-binding proteins containing covalently dimerized N-terminal domains of the bacteriophage 434 repressor cI have been constructed recently. In the homodimeric molecule RR69, the DNA-binding domains (amino acid residues 1-69) were connected in a head to tail arrangement with a part of the natural linker sequence that connects the N- and C-terminal domains of the intact repressor. This thesis describes the characterization of this molecule and the development of new single chain repressors based on rational design and selection from random protein libraries.

Compared to the isolated N-terminal DNA-binding domain, the single-chain molecule RR69 showed at least 100 fold higher binding affinity *in vitro* and a slightly stronger repression *in vivo*. An engineered heterodimeric molecule, RR*69 contains a wild-type and an engineered DNA binding domain. 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 two single-chain proteins. It is shown that RR69 recognizes DNA sequences containing the consensus subsites of the 434 operator sites in a palindromic arrangement. The heterodimeric single-chain repressor RR*69 recognizes nonpalindromic sequences composed of the respective consensus subsites of the 434 and P22 operators. In both cases, the separation of the subsites (the spacer length between the subsites) is conserved and corresponds to that observed in the 434 operator sites. Analysis of the base composition/sequence of the selected spacers and binding affinity studies suggest that the (binding of) single-chain repressors, similarly to the 434 repressor, is influenced indirectly by the noncontacted, spacer region. RR*69 represents an example for combination of altered direct and unchanged indirect readout mechanisms. The specificity of natural dimeric DNA binding proteins is usually restricted to palindromic operator sites. Random mutagenesis and *in vivo* phenotypic selection were used to develop single-chain repressor molecules that bind to *nonpalindromic sites,* and the binding properties of some of the mutants were characterized.

The thesis is based on the following publications:

Simoncsits, A., **Chen, J.**, Percipalle, P., Wang, S., Törö, I. and Pongor, S. (1997) "Single-chain repressors containing engineered DNA-binding domains of the phage 434 repressor recognize symmetric or asymmetric DNA operators" *J. Mol. Biol.*, 267, 118-131

Chen, J., Pongor, S. and Simoncsits, A. (1997) "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" *Nucleic Acids. Res.*, 25, 2047-2054

Further publications not included in the thesis:

Tuteja, N., Tuteja, R., Ochem, A., Taneja, P. Huang, N.W., Simoncsits, A., Susic, S., Rahman, K., Marusic, L., **Chen, J.**, Zhang, J., Wang, S., Pongor , S., and Falaschi, A (1994):"Human DNA Helicase II: A Novel DNA Unwinding Enzyme identified as the Ku Autoantigen" *EMBO J.*, 13, 4991-5001

1. Introduction

Many important processes of DNA metabolism depend on proteins that can find target DNA sequences within the genome, and then bind to them with a certain strength. Protein/DNA recognition can be characterized by two interrelated properties: target specificity i.e. the ability of the protein to recognize a given site in the presence of a large excess of other DNA sites, and binding parameters i.e. the affinity for the target site and the stability of the protein/DNA complex. These abilities of DNA-binding proteins (DBPs) must be very finely tuned in order to allow a precise regulation in such vital processes as gene-expression, DNA replication and repair.

Engineering DBPs with altered recognition specificity and affinity to DNA targets is an important research area with potential medical or biotechnological applications. In the long run such proteins would allow, for example, to repress undesirable genes or to cut DNA at any target site. Engineering of custom-designed DBPs is not a straightforward task, however. There seems to be no universal protein/DNA recognition code which - similar to base-pairing or translation codes - would allow one to rationally design artificial DBPs in one step. On the other hand, DBPs share a number of general principles, such as the modular design of the protein, the existence of DNA-binding domains (DBDs), and the high specificity conferred by long DNA cognate sites. Based on these principles one can combine rational design and random mutagenesis which will ultimately allow one to design custom-made artificial DBPs.

Sequence-specific DBPs 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 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 (Harrison, 1991; Klug, 1993; Pabo & Sauer, 1992; Rhodes *et al.*, 1996; Wilson & Desplan, 1995). Certain transcription factors contain covalently linked DNA-binding modules, e.g. the classical zinc finger proteins (Schwabe & Klug, 1994), the POU domain containing proteins (Herr & Cleary, 1995) and the c-Myb oncoprotein (Ogata *et al.*, 1994). 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 (Berg & Shi, 1996) for a review] and then these modules can be combined in the covalent framework to recognize longer DNA targets

(Choo *et al.*, 1994; Desjarlais & Berg, 1993). Alternatively, DBDs which are naturally not covalently linked, can be joined with designed or natural linkers to obtain *single-chain DNA-binding proteins* that contain different (Pomerantz *et al.*, 1995) or identical (Percipalle *et al.*, 1995; Robinson & Sauer, 1996) DBDs.

Our group has previously constructed single-chain derivatives of the phage 434 repressor (Figure 1.1), which belongs to the best studied members of the helix-turn-helix (HTH) family of DNA-binding proteins (Ptashne, 1992). 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 (Percipalle et al., 1995). This simplified framework appears to be suitable for designing DBPs with high specificity and affinity.

The aim of this work was

i) to rationally alter the specificity of the homodimeric single-chain protein (RR69), and

ii) to test the specificity of the newly designed proteins by *in vitro* binding-site selection as well as by *in vivo* repression of reporter genes.

As a model system, we choose to apply the example of the "helix-redesign" experiment of Ptashne and coworkers in which the putative DNA-contacting residues of the c2 repressor of phage P22 were grafted into the phage 434 repressor DBD α3 helix (Wharton & Ptashne, 1985). The redesigned repressor showed altered specificity characteristic for the P22 repressor and formed, in equilibrium with the wild-type 434 repressor, a non-covalent heterodimer with mixed specificity (Hollis *et al.*, 1988). In comparison with this classical experiments, the covalently linked framework offers important advantages: First, dimerization is permanent and therefore concentration independent. This makes it possible to characterize the binding properties of the new protein in a clear-cut way. Second, the binding domains are brought to close proximity by the linker - this arrangement is energetically more favourable than the noncovalent active dimer formation through monomer-dimer equilibrium.

The question of specificity is particularly important, both from a theoretical and from a practical point of view. Mutant DNA-binding proteins often show broadened binding specificities. For example, the zinc fingers can bind to a subset of targets, as revealed by rapid assays developed for this class of motifs (Choo & Klug, 1994a; Desjarlais & Berg, 1994). In fact, most specific, natural DNA-binding proteins recognize a set of related sequences (Rhodes et al., 1996) from which a consensus binding site can be

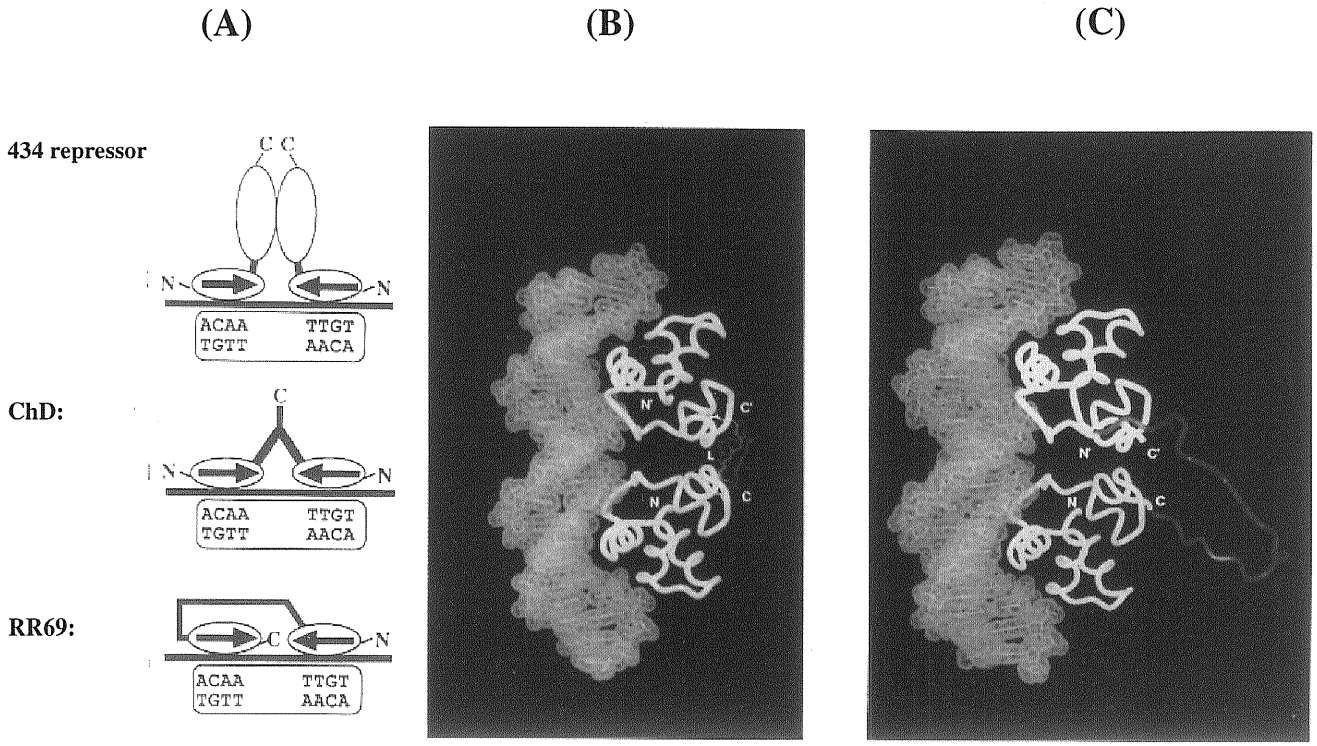


Figure 1.1. Schematic structure of the 434 repressor and the single-chain analogues (from Percipalle *et al* 1995). (A) Scheme of binding of the 434 repressor, ChD (a dimeric single-chain repressor with palindromic symmetry produced by chemical synthesis in which the DBDs are covalently linked through their C-termini) and RR69 (the single-chain repressor characterized in this thesis). The arrows denote the DBD (N to C direction). (B) Predicted 3-D structure of ChD. (C) Predicted 3-D structure of RR69. The linker in RR69 is deemed to be flexible, its position is only symbolically shown.

derived. Both the cI repressor of phage 434 and the c2 repressor of phage P22 recognize six different operator sites of the respective genomes (Wharton *et al.*, 1984; Poteete *et al.*, 1980). The 434 repressor also recognize the operator of the non-related phage 16-3 C repressor, which contains the same consensus sequences as the 434 operators (Dallmann *et al*, 1987). These operators show sequence divergence mainly in the inner, spacer region, which is not in direct contact with the repressor as shown by structural (Aggarwal *et al.*, 1988; Rodgers & Harrison, 1993; Shimon & Harrison, 1993) and/or biochemical studies [see (Koudelka *et al.*, 1996) 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 (Bell & Koudelka, 1993; Bell & Koudelka, 1995; Koudelka et al., 1996; Koudelka & Carlson, 1992; Koudelka *et al.*, 1988; Wu & Koudelka, 1993; Wu *et al.*, 1992).

The complex nature of DNA recognition, which is usually a combination of direct and indirect recognition mechanisms, becomes even more complicated with artificial DNA binding proteins. The consequences of combining and altering DBDs are important to be studied to be able to create new, sequence specific proteins either by rational design or by random methods.

2. Literature review

Protein-DNA interactions play a pivotal role in many important cellular processes, such as transcription, replication, recombination, packaging and restriction. It is therefore necessary to understand the mechanisms of protein-DNA interactions. Many proteins, particularly those involved in chromosome packaging (e.g. the histones) or DNA replication (e.g. DNA polymerase) have low or no sequence specificity, whereas others, such as transcriptional activators, repressors and restriction endonucleases show extensively high specificity for their special target sites. The ability of transcription factors to recognize specific DNA sequences with high fidelity is essential for the regulation of gene expression and hence for the control of cell growth and differentiation.

Owing to the advances in recombinant DNA technology, and also to the improvements in both X-ray crystallographic and nuclear magnetic resonance (NMR) techniques, significant progress has been made in the field of detailed analysis of specific protein-DNA interactions during the past ten years, which has facilitated the understanding of the molecular basis of specificity.

2.1 Principles of specific recognition between protein and DNA

What is the basis for the specific recognition of DNA by proteins? The molecular basis of specificity requires the characterization of the conformational properties of the protein, the DNA target site, and the changes that ensue as a consequence of the interaction. A wealth of biochemical and structural information has been accumulated to illustrate the details of protein-DNA interactions in numerous instances. Though from the structural studies some rules have been derived for a certain protein structure family (Choo & Klug, 1997) or for a group of proteins which interact in similar ways with DNA (Suzuki, 1993; Suzuki *et al.*, 1995a; Suzuki & Yagi, 1994; Suzuki *et al.*, 1995b), the complicated nature of the recognition mechanism precludes a simple recognition code. The general principles of recognition described below are based mainly on the physico-chemical properties of amino acids and nucleotides, including shape recognition and chemical recognition. [see (Pabo & Sauer, 1992; Rhodes et al., 1996; Sinden, 1994) for review]

Chemical rules for recognition

Specificity of protein-DNA interactions relies primarily on the recognition of the linear order of base pairs by protein through hydrogen bonds, hydrophobic interactions, global electrostatic interactions and salt bridge contacts between the amino acid residues and DNA base pairs.

The binding specificity originates mainly from chemical contacts between amino acid side chains and bases in the major grooves. Certain functional groups which serves as hydrogen-bond donor or acceptor, or can participate in van der Waals interaction are located on the edges of base pairs (Figure 2.1). The unique arrangement of these functional groups for each base pair within the major or minor groove provides the specificity utilized by proteins to discriminate regions of a DNA sequence. A number of specific interactions between certain amino acids and DNA base pairs have been demonstrated (Figure 2.2). The most common contacts are the two hydrogen bonds formed between Gln (or Asn) and adenine in the major groove, and similar interactions between Arg and guanine. The 5'-methyl group in thymine is involved in van der Waals contacts with the methyl or methylene groups of amino acid side chain, as well as being a steric hindrance for incorrect binding. (Seeman *et al.*, 1976; Suzuki, 1994)

Besides the intrinsic chemical ability, one must keep in mind that a favourable hydrogen bond can be made only if the interacting groups in both DNA and amino acid side chains are in an appropriate position and orientation. These are affected by the binding geometry between protein and DNA. The DNA phosphate backbone, with relatively uniform shape and negative charge, plays integral role in the site-specific recognition. In principle, any basic or neutral hydrogen-bonding side chain can be used to contact the phosphodiester oxygens, while it seems that short polar side chains and the peptide-amide may provide more stereospecificity than those mediated by the long flexible side chains of Arg and Lys. Backbone contacts may serve to hold the protein against the bases in a fixed arrangement and thereby enhance the specificity of side chain-base interaction, or help to establish the DNA conformation change according to the requirement for protein binding. Therefore, contacts with the DNA backbone are important for the complementary recognition.

The structure complementarity in recognition

Structural studies have shown a remarkable shape complementarity as well as local electrostatic and van der Waals complementarity in the protein-DNA interface. Because, at a rough level, the structure of DNA is essentially uniform, diverse DNA-binding proteins employ similar architectural strategies to achieve interfaces which are

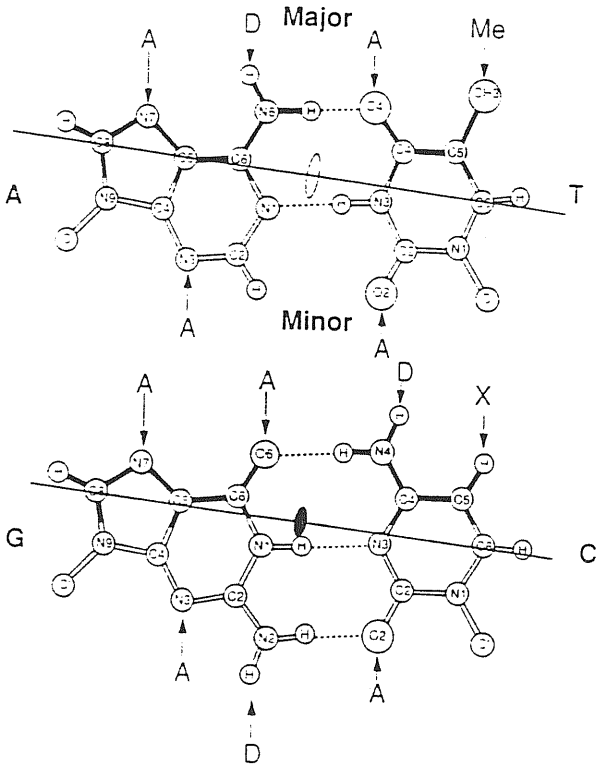


Figure 2.1. Hydrogen-bonding patterns in protein-DNA interactions (from Freemont, 1991). The A.T and G.C base pairs are shown in the same relative orientation. "A" denotes an atom that can act as a hydrogen-bond acceptor, and "D" denotes an atom that can act as a hydrogen-bond donor. Major and minor refer to the two grooves of DNA. The pseudo two-fold axis of the base pairs is indicated. "X" denotes no interaction at this position.

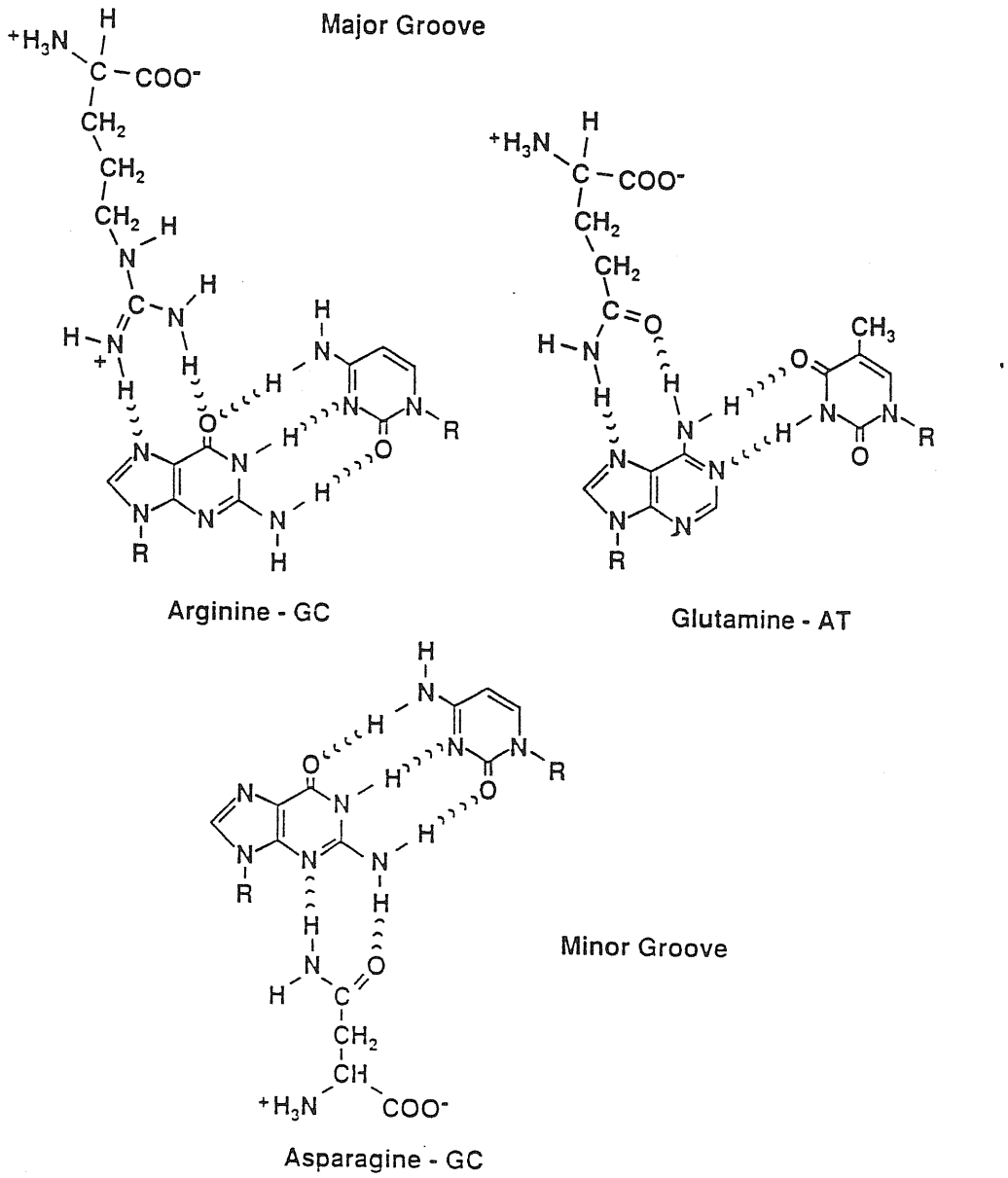


Figure 2.2. Examples of protein-DNA hydrogen bonding (from Sinden, 1994).

complementary in shape. The convex surface of a protein secondary structure, an α helix or a β-sheet is usually docked precisely in the concave surface of the DNA grooves.

The protein α helix in recognition

Most of the well-characterized families of DNA-binding proteins use α-helix to make base contact in the major groove. The overall shape and dimension of an α-helix allows it to fit into the major groove in a number of related but significantly different ways. Some lie in the middle of the major groove and have the axis of the α-helix approximately tangential to the local direction of the major groove. Others are tipped at different angles, and some are arranged so that only the N-terminal portion of the α-helix fits completely into the major groove. The surrounding region of the proteins helps to determine how these α-helices are positioned in the major groove, by maintaining the global architecture of the protein or stabilizing the binding geometry through the non-specific phosphate backbone contacts. (Pabo & Sauer, 1992)

The stereochemical rules

Each family of DNA binding proteins has a specific DNA binding geometry. Some stereochemical charts have been deduced from crystallographic and NMR studies of protein-DNA complexes (Suzuki et al., 1995b). Different families of transcription factors adopt different binding geometries to form the complementary interface. In the specific recognition, the size of amino acid residues should be compatible with the requirement of the contacts. From a fixed position on the interaction surface, a long side chain can reach further and deeper into the DNA groove; whereas at a position very close to the DNA, a small residue can easily fit in but a bulky residue may not.

The DNA structure influence in recognition

The specific interaction between proteins and DNA has most frequently been considered from the viewpoint of protein, due to the large structural diversity of proteins. However, the DNA conformation and configuration also have a profound influence in the process of recognition. In the B-form DNA, the major groove is wider and better suited to accommodate the protein secondary structure than the minor groove. In the major groove the pattern of hydrogen-bond donors and acceptors is unique for each base-pair, whereas in the minor groove it is not possible to distinguish between AT and TA base pairs, nor between GC and CG base pairs (Seeman et al.,

1976). Consequently, a major groove with B-like properties is best suited for allowing direct, sequence-specific interactions. Local or global DNA distortions (i.e. groove width, charge, bending or/and twisting) were observed in many protein-DNA complexes, so that the changed DNA structure can better accommodate a protein secondary structure.

The DNA double helix could adopt discrete conformations depending on the sequence and upon the degree of hydration. When proteins recognize a specific DNA sequences, they "read" the base sequence either through direct interactions, or through recognizing features of the overall DNA structure which is dependent on the base sequence. The latter type of recognition has been termed *indirect readout* or *analogue recognition* in contrast to the *direct readout* or *digital recognition* of individual bases. The primary sequence dictates the local twist angle, the specific tilt and roll of bases, the bends in the DNA, and the width of the major and minor grooves, as well as the ease or flexibility for structural changes. These structure parameters will precisely position in space the hydrogen-bond donor and acceptor sites in the bases and the phosphate backbone. Most DNA-binding proteins are evolved to recognize a particular shape or flexibility of the double helix in addition to a direct readout of individual bases in the recognition site. In many cases, charge distribution in nucleotides within the binding site that are indirectly contacted with a protein strongly influence the binding affinity of the protein. [See (Travers, 1989) for review]

The structural complementarity required for specific recognition is determined by both the geometric placing of the molecules and the energy cost. Protein and DNA molecules will interact if there is a decrease in Gibbs free energy upon the formation of a complex. The change in free energy (ΔG) during complex formation depends on the change in both entropy (ΔS) and enthalpy (ΔH), *i.e.* $\Delta G=\Delta H-(T\Delta S)$, where T represents temperature. Both the enthalpy and entropy terms depend on the shape of the surfaces between the protein and its target DNA. (Rhodes *et al.*, 1996)

Multiple subsite recognition — more specific and precise

Usually one DNA binding domain can access only one side of the DNA and its recognition motif, such as the α-helix, contacts with a short (3-5bp) DNA sequence. This is, however, not sufficient to confer specific and high affinity binding. To overcome this problem, multiple subsite binding by multimeric proteins has been employed. The affinity of a protein for its DNA binding sites is a result of the number and strength of electrostatic and hydrophobic interactions between the protein and DNA. Thus a larger binding site can provide a stronger interaction between the DNA and a protein than a smaller site.

Several strategies have been employed for multiple subsite recognition. The first is to add simply on arms or tails that recognize additional features of the DNA, particularly in the minor groove. The second is to form homo- or hetero-dimers. The third is to employ multiple DNA binding domains, either by using tandem repeats of the same type of DNA-binding motif, or by linking together different types of motifs within a single polypeptide chain. [see (Rhodes et al., 1996) for examples]

Multiple binding sites can also be used to modulate the binding affinity. Instead of a unique DNA sequence, most specific DNA-binding proteins recognize a set of related sequences with varied binding affinity. This can be achieved in two ways: 1) Readjustment of the protein side chain conformation or DNA structures so as to form a different interaction network between the functional groups in DNA and protein. 2) Using the spacer between the DNA subsites to direct the super-structure arrangement of the individual binding motifs.

Influence of other factors

Ordered water molecules have been observed in most protein-DNA interfaces. In some cases, water molecules participate in the network of hydrogen bonds between protein and DNA. Water may act as a kind of lubricant allowing the protein to scan along the DNA for the specific binding site. Water molecules may also play important role in the energetics of protein-DNA interactions. The displacement of the water molecules favours protein-DNA complex formation by providing a favourable entropic contribution to the change in free energy. (Schwabe, 1997)

The formation of a complex is usually associated with local folding events. There are several examples, like the basic region in bZIP proteins that is usually disordered in solution but becomes α–helical upon binding to DNA (Ellenberger. 1994). Percipalle, *et al* (1995) also reported an increase in the α-helix content in artificial single-chain repressor containing the HTH motif. Hence, the formation of a protein-DNA complex involves local folding events and these are coupled to the thermodynamics of binding.

In addition to the thermodynamics upon protein-DNA interaction, the kinetic events of binding and release of protein from their DNA targets are also important for understanding how the transcription factors modulate their regulatory functions.

The understanding of the recognition principles is very important for the design of novel, specific DNA-binding proteins of both biotechnological and pharmaceutical interests.

2.2 DNA recognition by the helix-turn-helix motif

Many DNA-binding proteins achieve specific recognition through small, discrete, independently folded structural units. Most of the structures identified so far fall into a number of different types, each type has a characteristic amino acid sequence and three dimensional structure. The structural families are

1) the helix-turn-helix (HTH) motif, including a large family of prokaryotic transcription factors (Harrison & Aggarwal 1990; Pabo & Sauer, 1992) and the eukaryotic homeodomain (Gehring *et al.*, 1994; Wolberger, 1996);

2) the zinc-binding proteins (Berg & Shi, 1996; Schwabe & Klug, 1994)

3) the leucine zipper (bZIP) and basic helix-loop-helix (bHLH) motifs (Ellenberger, 1994);

4) the β ribbon (Raumann *et al.*, 1994); and

5) the TATA box binding proteins (Burley, 1996).

The helix-turn-helix motif, which is found in the proteins studied in this thesis , is discussed below in detail. For other motifs, see (Freemont *et al.*, 1991; Harrison, 1991; Klug, 1993; Pabo & Sauer, 1992; Sinden, 1994) for reviews.

The helix-turn-helix (HTH) structure was the first discovered DNA-recognition motif, and is to date most thoroughly studied [see (Brennan, 1991; Brennan, 1992; Harrison & Aggarwal, 1990) for review]. The HTH motif of 20 amino acides consists of an 8 amino acid α-helix followed by a 3 amino acid right turn with an angle of about 120° and another α-helix of 9 amino acids. Three amino acids in critical positions are highly conserved and believed to be responsible for the structural stability. These conserved amino acids are at the 5th position Ala in the first α-helix, at the 9th position Gly - the first amino acid of the turn - and either Val or Ile at position 15 in the second α-helix (Sinden, 1994). The amino acids in other positions show a high degree of heterogeneity. Nevertheless, since an α-helix has a repeat length of 3.6 amino acid per turn, certain positions of the HTH face toward the body of the protein (a hydrophobic environment), while other residues face the solvent or DNA (a hydrophilic environment), a similarity in the type of amino acids at these positions has been observed (Suzuki et al., 1995b).

The HTH motif is not a stably folded structure on its own, and usually one or more extra helices from the rest of the protein should be engaged to stabilize this motif. The second helix of HTH lies in the major groove of DNA and carries the main amino acid residues responsible for specific binding, hence this helix is usually called the "recognition helix". However, it is wrong to assume that the recognition involves only this local contacts. Studies have revealed that other regions out of the HTH units can also have significant role in recognition. For example, besides the contacts in the HTH

motif, the λ repressor specifies the operator base pairs by using an extended peptide chain in the loop following the second helix and the N-terminal arm to wrap around DNA. (Pabo & Sauer, 1992)

The HTH motif is evolved in a dimension to fit into the major groove of DNA. The diameter of a typical α–helix is about 12Å, which exactly matches the 12Å wide and 6-8Å deep major groove of B-DNA. If an α-helix is parallel to the direction of the major groove, the straight α-helix can contact 4-6 bp before the bases arrive out of the plane of the amino acids in the α-helix. Thus the short interaction with no more than 4-6 bp of DNA specifies the binding of a particular sequence of amino acid to a unique DNA sequence.

Three types of residues are arranged into the recognition helix, that are

1) the DNA base contacting residues (usually residues 1 to 3, 5 and 6 of the second helix), which are important for the specificity;

2) the phosphate backbone contacting residues, which fix binding geometry;

3) the other residues facing away from the DNA, which limit the rotation of the recognition helix by interacting with the rest of the protein. (Brennan, 1991; Suzuki et al., 1995b).

Crystal structures of prokaryotic HTH protein complexed with their specific DNA, e.g. several phage repressor-operator complexes, Trp repressor, *E.Coli* CAP, *etc.*, have been resolved. These cocrystal structures show some common features of the HTH-protein-DNA interaction (Harrison & Aggarwal, 1990; Pabo & Sauer, 1984; Pabo & Sauer, 1992):

1) The repressors bind as dimers. Each monomer recognizes one half of the binding site, and the approximate symmetry of the DNA binding site is reflected in the approximate symmetry of the protein-DNA complex.

2) The conserved HTH unit contacts the DNA in each half of the operator site. There is no universal mode for docking the HTH motif against the major groove of DNA. This is due to the fact that not only does the precise "angle of attachment" vary from case to case but also the major groove itself has a variable geometry. For example, although 434 repressor (Aggarwal et al., 1988; Mondragon *et al.*, 1989a; Rodgers & Harrison, 1993; Shimon & Harrison, 1993) and Cro (Mondragon & Harrison, 1991; Mondragon *et al.*, 1989b) attach similarly to the DNA backbone, they create different groove structures upon binding. However, despite these variations, there are some important regularities in the mode that HTH elements bind to DNA. The first helix of the HTH unit is somewhat "above" the major groove, but the N-terminus of this helix contacts the DNA backbone. The second helix of the HTH unit fits into the major groove, and the N-terminal part of this helix is closest to the edges of the base pairs.

3) Though there are some local distortions in different protein-DNA complexes, the operator sites are generally B–form DNA.

4) Side chains from the HTH unit make site-specific contacts with base pair groups in the major groove. Direct contacts, polar and non polar, between amino acid side chains and the edges of base pairs in the major groove are the principal sources of specificity. Various side chains (Gln, Asn, Ser, Arg, Lys, *etc.*) donate and accept hydrogen bonds. Many of the contacted base pairs interact with more than one amino acid side chain, and many of these side chains interact with more than one base pair. Changes in directly contacted base pairs generally decrease the affinity by at least one or two orders of magnitude.

5) Each complex has an extensive network of hydrogen bonds between the protein and the DNA backbone. Particularly noteworthy are the hydrogen bonds to non-esterified phosphate oxygens, especially from peptide -NH groups, neutral $-NH_2$ groups of Gln and Asn side chains, and -OH groups of Ser and Thr. These interactions occur in the context of tight van der Waals complementarity that anchors the protein very precisely. Hydrogen bonds between positively charged amino acid side chains (Lys, Arg) and DNA phosphates appear with only modest frequency (Pabo & Sauer, 1992).

Some other factors also contribute to the recognition, which may vary from complex to complex, e.g. in 434 repressor, the free energy of DNA conformation makes an additional contribution to specificity, while the Trp repressor appears to use several water molecules to provide critical contacts. In addition to the DNA binding domains, other domains in the HTH protein also play important roles in regulating activities, such as the N-terminal domain of CAP which allows dimer formation and also binds to cAMP – an allosteric effector of DNA binding. In the case of phage 434, P22 and λ repressors, the C-terminal domains are important for stable dimer formation and high affinity binding, as well as the cooperative binding to the neighbouring operator sites.

2.3 Interactions between the phage 434 repressor and its operators

Role in genetic switch

In temperate bacteriophages, such as lambda and 434, an efficient genetic switch regulates the choice between lysogeny and lytic growth (Ptashne, 1992). The switch requires differential affinity of two proteins, repressor and Cro, for six operator sites, designated as O_R1, O_R2, O_R3 and O_L1, O_L2, O_L3. The O_R operator controls two distinct promoters, known as P_R and P_{RM}. Promoter P_R governs transcription of

Table 2.1. Operator sequences of the 434 (A) and P22 (B) repressors

(A)

	-1	1	2	3	4	5	6	7	7'	6'	5'	4'	3'	2'	1'	-1'
O_R1	T	A	C	A	A	G	A	A	A	G	T	T	T	G	T	T
	A	T	G	T	T	C	T	T	T	C	A	A	A	C	A	A
O_R2	A	A	C	A	A	G	A	T	A	C	A	T	T	G	T	A
	T	T	G	T	T	C	T	A	T	G	T	A	A	C	A	T
O_R3	C	A	C	A	A	G	A	A	A	A	A	C	T	G	T	A
	G	T	G	T	T	C	T	T	T	T	T	G	A	C	A	T
O_L1	T	A	C	A	A	G	G	A	A	G	A	T	T	G	T	A
	A	T	G	T	T	C	C	T	T	C	T	A	A	C	A	T
O_L2	A	A	C	A	A	T	A	A	A	T	A	T	T	G	T	A
	T	T	G	T	T	A	T	T	T	A	T	A	A	C	A	T
O_L3	A	A	C	A	A	T	G	G	A	G	T	T	T	G	T	T
	T	T	G	T	T	A	C	C	T	C	A	A	A	C	A	A

(B)

	1	2	3	4	5	6	7	8	9	9'	8'	7'	6'	5'	4'	3'	2'	1'
O_R1	A	T	T	A	A	A	G	A	A	C	A	C	T	T	A	A	A	T
	T	A	A	T	T	T	C	T	T	G	T	G	A	A	T	T	T	A
O_R2	A	C	T	A	A	A	G	G	A	A	T	C	T	T	T	A	G	T
	T	G	A	T	T	T	C	C	T	T	A	G	A	A	A	T	C	A
O_R3	A	T	T	T	A	A	G	A	T	G	A	C	T	T	A	A	C	T
	T	A	A	A	T	T	C	T	A	C	T	G	A	A	T	T	G	A
O_L1	A	T	T	T	A	A	G	A	C	T	T	C	T	T	A	A	T	T
	T	A	A	A	T	T	C	T	G	A	A	G	A	A	T	T	A	A
O_L2	T	T	T	G	A	A	G	A	A	A	A	C	T	T	A	A	A	T
	A	A	A	C	T	T	C	T	T	T	T	G	A	A	T	T	T	A
O_L3	A	C	T	T	A	A	G	T	T	T	T	T	A	T	T	T	G	A
	T	G	A	A	T	T	C	A	A	A	A	A	T	A	A	A	C	T

genes, including the one for Cro, that are important for initiating lytic growth. P_{RM} is the promoter for repressor transcription. In a lysogen, repressor dimers bind co-operatively to O_R1 and O_R2 blocking the RNA polymerase binding to P_R. Thus the rightward promoter P_R is turned off, while the leftward promoter P_{RM} is activated, which in turn starts the transcription of its own message. High concentration of repressor leads to binding at O_R3 and repression of P_{RM}. In the lytic phase, the synthesis of Cro protein turns off the repressor synthesis through binding to O_R3, that blocks the RNA polymerase binding to P_{RM} .

Structure of the phage 434 repressor and its N-terminal domain (R1-69)/operator complexes

434 repressor binds as a dimer to 14 bp operator sequences by using a helix-turn-helix motif. The carboxy-terminal domain of the repressor mediates the dimerization, while the amino-terminal (1-69 amino acids, denoted as R1-69) is the DNA binding domain. R1-69 is a bundle of five α-helices linked by turns of varying length. The α2 and α3 helices form the HTH motif. The six operator sites to which 434 repressor binds are 14 bp in length, with pseudo two-fold symmetry. The outer four bases of each half site are a conserved ACAA/TTGT sequence, and the inner six base pairs of each operator are variable. Each monomer binds to a half site of the operator. Table 2.1(A) is a list of natural 434 operators.

Crystal structures of the N-terminal domain (R1-69) of the phage 434 repressor complexed with its cognate operators O_R1, O_R2, O_R3 have been solved at 2.5Å resolution.

In the R1-69/O_R1 complex (Aggarwal et al., 1988), the B-type DNA is distorted by bending with variations in twist, and other helical parameters. The bended configuration permits contacts between the sugar-phosphate backbone and the NH_2-terminus of α2 helix (along with Arg10), as well as the interactions of Gln28 and Gln29 with base pairs 1 and 2. A striking feature resulting from the bending is the compression of the minor groove in the center and gradually widening to the ends. The width of the minor groove is compressed to 8.8Å in the center while opened to 14Å at the ends of the 14 bp operator (the normal width of minor groove in B-DNA is 11.5Å). Thus the DNA is overwound in the center and underwound at the ends.

Near the center of the operator between nucleotides 7L and 7R, Arg43 extends from the loop between α3 helix and α4 helix into the minor groove. Arg43 from each monomer forms asymmetric contacts with the bases, sugars and phosphates. The presence of the positively charged side chains in the minor groove is thought to stabilize the minor groove compression, which brings negatively charged phosphates close together.

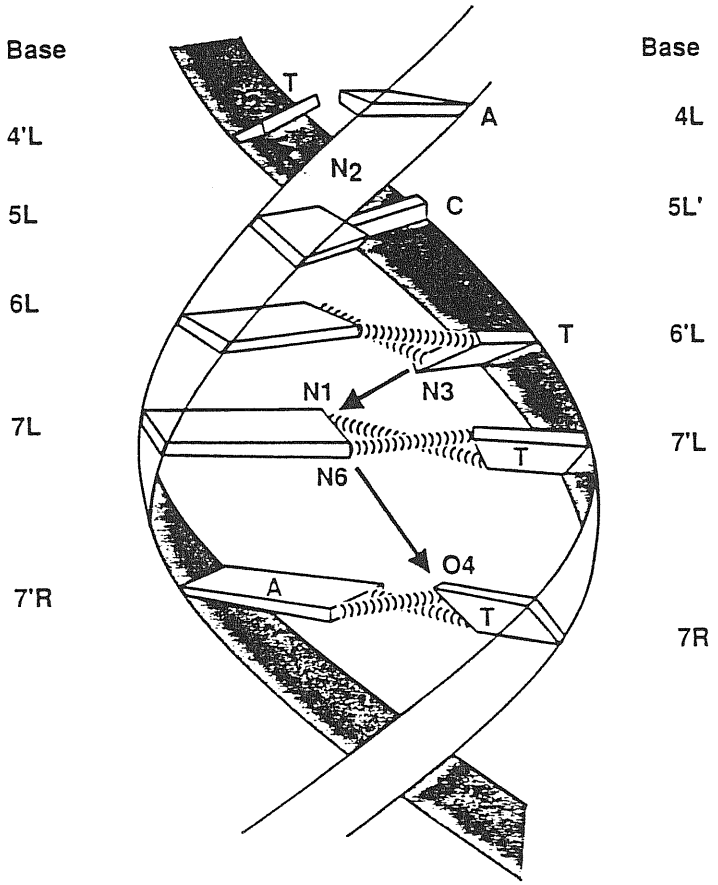


Figure 2.3. Bifurcated hydrogen bonds in the 434 repressor R1-69/O_R1 complex (from Sinden, 1994). Watson-Crick hydrogen bonds between A-T and C-G base pairs in the operator are shown by the standard short curves. The hydrogen bonds between the adjacent dinucleotide pairs are shown as bold arrows.

A high propeller twist was observed in the central base pairs. The propeller twist positions functional groups involved in Watson-Crick hydrogen bonding close to groups on an adjacent base pair. As a result, three bifurcated hydrogen bonds are formed between three adjacent base pairs (O4 of T7R and N6 of A7L, N2 of G5L and O2 of T4L , N3 of T6L and N1 of A7L) at the bend center (Figure 2.3). The capacity to form bifurcated hydrogen bonds (also observed in R1-69/O_R3 complex) is thought to compensate for distortion of the operator by the repressor binding, and it is dependent on the DNA sequence.

The α3 helix lies in a parallel orientation in the major groove. The NH_2-termini of helices 2 and 4 are close to the sugar phosphate backbone, and the loop from helix 3 to helix 4 follows the sugar phosphate backbone at the center of the operator. There is no large scale conformational change of the protein upon binding, but it does reveal significant local adjustment of side chain conformation and a small shift in the turn from α2 helix to α3 helix. The conformation of the side chains of Gln29, Gln32 and Arg43 was rearranged upon binding. All these three amino acides interact with DNA.

The half site interactions are linked by the protein dimer interface. It contains a patch of hydrophobic residues, Leu45, Pro46, Val56 and Leu60. There is also a salt bridge between Arg41 and Gln47.

The protein-DNA interaction involves the hydrogen bonding and hydrophobic interactions, see Table 2.2 for the list of the important contacts, and Figure 2.4 for the illustration.

In the R1-69/O_R2 complex (Shimon & Harrison, 1993), both the protein and the DNA backbone conformations are very similar to the R1-69/O_R1 complex, and the same extensive network of hydrogen bonds anchors the repressor to the DNA backbone. The major groove contacts between the outer four conserved base pairs and critical amino acid side chains are essentially identical in the O_R1 and O_R2 complexes. However the R1-69/O_R2, has relatively coplanar base pairs in the center of the operator, and has no "bifurcated" non-Watson-Crick hydrogen bonds which were observed in the O_R1 and O_R3 complexes. This conformational variation is due to the sequence difference in the two operator center. O_R1 and O_R3 have runs of A-T pair ("A-tract") sequence, while O_R2 has an alternating A-T/T-A pair sequence, which may cause a cross-strand stereo crash if propeller twist is introduced.

The O_R3 operator contains the consensus sequence ACAA in one half-site of the operator, whereas there is one base pair deviation from the consensus in the other half-site (ACAG). The structure of the R1-69/O_R3 (Rodgers & Harrison, 1993) for the consensus half site is essentially identical to that seen in the O_R1 and O_R2 complexes. However, there is an unexpected extensive structure change in the non-consensus half site. The most marked change is at the DNA backbone from position 3' to 4'. The backbone bows out towards the protein, bringing the phosphates closer to residues in

Table 2.2 Contacts between the 434 repressor and operator (from Sinden, 1994)

DNA position	Amino acid	Contact[a]
Major groove hydrogen bond contacts (shown in Figure 8.11)		
#1 A·T	Gln 28	Bidentate hydrogen bonds are formed between the NH_2 of Gln 28 and N7 of A, and between the C=O of Gln 28 and the N6 of A
#2 C·G	Gln 29	The terminal NH_2 of Gln 29 makes bidentate contacts with the O6 carbonyl group and the N7 position of guanine
#4 A·T	Gln 33	The O4 carbonyl of T forms a hydrogen bond with the NH_2 of Gln 33
Major groove van der Waals contacts		
#−1 T·A	Gln 28	The methyl group of thymine at position −1 (outside the 14-bp operator) contacts the methyl groups on the side chain of Gln 28
#3 A·T	Thr 27, Gln 29	Side chain methyl groups of Thr 27 and Gln 29 form a van der Waals pocket to bind the methyl group of thymine
#4 A·T	Gln 29, Ser 30	The methyl group on thymine contacts the side chain methyl groups of Gln 29 and Ser 30
P1, Sugar 1	Asn 16, Gln 17	van der Waals contacts made between the PO_4, and sugar and the CH_2 side chains of the amino acids
Hydrogen bonds to the phosphate backbone		
P−2	Asn 16	The P2 phosphate (on the 5′ side of A 2, between T 3 and A 2) forms a hydrogen bond with the terminal NH_2 of Asn 16
P−1	Gln 17, Arg 10	The P1 phosphate (between T 1 and A 2) hydrogen bonds to the NH_3 of Arg 10 (not shown) and the main chain NH of Gln 17, at the end of helix 2
P1	Gln 17, Asn 36	The terminal NH_2 of Gln 17 forms a hydrogen bond to P1; Asn 36 also forms a hydrogen bond with P1 (not shown)
P5′	Arg 43	The main chain NH group of Arg 43 forms a hydrogen bond with P5′
P6′	Lys 40, Arg 41	The main chain NH groups of Lys 40 and Arg 41 form hydrogen bonds with P6′

[a]Data from Aggarwal *et al.* (1988).

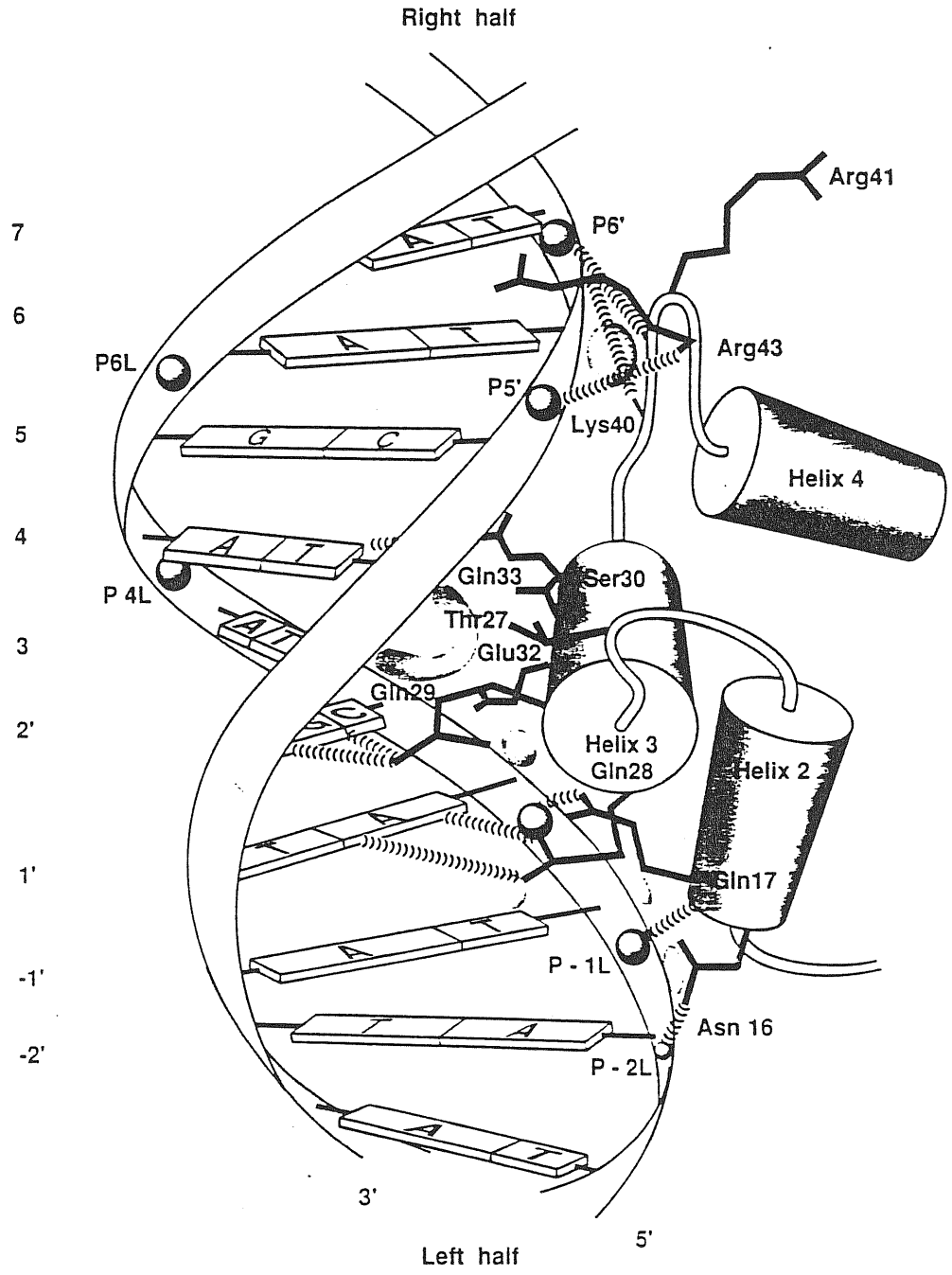


Figure 2.4. Interactions of the 434 repressor with the left half of the 434 operator (from Sinden, 1994). Hydrogen bonds are shown as short curves. Hydrophobic interactions are shown as shaded areas between specific amino acids and the bases or phosphate backbone.

the α2-α3 helix turn, as well as the amino terminus of α3 helix. In addition, the protein monomer rotates relatively to the other monomer that interacts with the consensus half site. This monomer rotation further decreases the gap between the DNA backbone and the α2-α3 helix turn. The substitution of G-C for A-T at position 4 also causes a rearrangement of the Gln33 side chain, the interaction between Gln33 and the base at position 4 is weakened compared with the consensus half site, and there is no direct hydrogen bond is present. Bell *et al* demonstrated that recognition of the base at position 4 by repressor is not an independent event and, moreover, this process is influenced by the sequence of bases not contacted by repressor through altering the global structure of the repressor-operator complex (Bell & Koudelka, 1995).

The recognition properties of 434 repressor

From the structure of the R1-69 /operator complexes, together with biochemical studies (Bell & Koudelka, 1993; Bell & Koudelka, 1995; Koudelka, 1991; Koudelka et al., 1996; Koudelka & Carlson, 1992; Koudelka et al., 1988; Koudelka et al., 1987; Koudelka & Lam, 1993) , Several conclusions could be drawn on the recognition properties of the 434 repressor.

The consensus box recognition: positions 1 to 4

In the natural 434 operators, all twelve half sites contain ACA at position 1 to 3, and eleven have A at position 4. Biochemical studies demonstrated that any substitution for ACAA at base pair 1 to 4 caused a reduction of binding affinity by at least 100 fold (Anderson *et al.*, 1987; Koudelka & Lam, 1993). Both the structural (Aggarwal et al., 1988) and biochemical studies have shown that these conserved positions are specified by the direct contacts between a series of amino acid side chains and the base pairs. No base pair substitution can conserve the complementarity. Some amino acids which do not contact also play important roles in the specificity determination. Koudelka and Lam showed that at least three mutations are needed to eliminate the position 4 base specificity, i.e. mutations on the directly contacting Gln33, and the non-contacting Gln32 and Thr27 (Koudelka & Lam, 1993).

The position 5 recognition

There are no direct amino acid side chain-base pair contacts at position 5. However, there is a solvent mediated network of hydrogen bonds between phosphate 5 and Gln33. It seems that the identity of base pair 5 affects the configuration of Gln 33 and

its interaction with base pair 4. The conformation of the base pair 5 also influence the base pair 6 backbone contacts with the α3 and α4 loop of the protein, resulting in a shift at position 6. The repressor has a base preference at position 5 in the order G>A, T>C (Aggarwal et al., 1988). These affinity differences (approximate five to ten fold) could arise either from the appropriate protein conformational change resulting from repositioning Gln33 or from the way in which other base pairs can adjust to the configuration of base pair 5.

The central base pairs: positions 6 and 7

The 434 repressor binds to the operators with A-T or T-A at the central base pairs more tightly than those with G-C or C-G. This effect is stronger at position 7 than at position 6.

Water mediated Arg43 contacts in the minor groove of the operator center are the only interactions between the repressor and DNA in this region. However, these interactions do not appear to account for the central base pair specificity since similar hydrogen bonds could also be made with other sequences. Koudelka *et al* have reported that an Arg43→Ala mutant has the similar preference for the central sequence as the wild-type repressor (Koudelka et al., 1987).

The conformation of the uncomplexed 434 operators vary with the central sequence, while the conformation of the DNA-phosphate backbone in the protein-DNA complexes is the same and independent of the central base sequence (Koudelka et al., 1996; Koudelka & Carlson, 1992). The central base pair preference is most probably due to the ease for the operators to readjust the local conformation distortion upon repressor binding. The base pairs at the 6th and 7th positions are configured to bring the half-site of the operator into proper alignment with the protein, thus allowing each monomer of the bound dimer to make optimal contacts within each operator half-site. The imposed sugar-phosphate backbone conformations do not appear to vary with nucleotide sequence, but the adjustment of the base pairs does. Therefore, the repressor binding should be favoured by those sequences for which the conformation constrained by the overall structure requirement is energetically least costly. AT-rich sequences seem to accommodate the distortions more readily than GC-rich sequence in the 434 repressor-operator complexes. X-ray studies of a number of DNA structures have shown that at runs of G and C the minor groove of the helix was exceptionally wide or could easily become wide, whereas at certain runs of A and T the minor groove was narrow or could easily become narrow (Drew & McCall, 1990). A sequence dependent likelihood of flexure was established based on the DNase I digestion experiment on nucleosome DNA. There is a strong correlation between the observed affinity of the

repressor for the operator and the predicted likelihood of flexure (Travers & Klug, 1990).

2.4 Design and construction of custom-built DNA-binding proteins

Rational design and construction of novel DNA-binding proteins with user defined specificity and regulatory activities are important for both biological research and biotechnological applications. It is a field that is developing rather fast in the recent years as more and more structural information became available.

One of the construction strategies is to link the existing DNA binding domains together. The best example of this is the artificial protein ZFHD1 which contains zinc fingers 1 and 2 from Zif268, a short polypeptide linker, and the homeodomain from Oct-1. The fusion protein binds optimally to a combined binding site of the zinc fingers and the homeodomain. When fused to an activation domain, this protein was shown to regulate promoter activity *in vivo* in a sequence-specific manner.(Pomerantz et al., 1995)

A more general approach has been used successfully to gain DNA-binding proteins with genetically novel specificity by using the classical zinc finger motif as a frame work. The zinc finger is an independently folded domain in which a zinc ion stabilizes the packing of an antiparallel β–sheet against an α-helix. The crystal structures of zinc finger-DNA complexes show a semiconserved pattern of interactions in which three amino acids from the α-helix contact three adjacent bases (a triplet, in DNA). Fingers with different triplet specificities are combined to give specific recognition of longer DNA sequences (Choo & Klug, 1994b). Modelling, sequence comparison, and phage display have been used to alter the specificity of individual fingers within a multifinger protein (Choo & Klug, 1994b; Jamieson *et al.*, 1994; Rebar & Pabo, 1994; Taylor *et al.*, 1995). For example, fingers that specifically recognize a conserved sequence in the genome of type 1 human immunodeficiency virus have been isolated through phage display selection approach (Wu *et al.*, 1995). Fingers have also been mixed and matched to construct new DNA-binding proteins, Desjarlais and Berg (Desjarlais & Berg, 1993) have designed zinc finger proteins by combining a consensus zinc finger framework sequence with previously characterized recognition regions to generate three novel zinc finger proteins that specifically recognize the predicted sequences. The most comprehensive and efficient technique of the above mentioned approach is the employment of phage display technique to select zinc fingers that recognize desired triplets. The success of this strategy has been strikingly demonstrated through the generation of a novel three-zinc-finger peptide that recognizes the oncogene BCR-ABL, and remarkably inhibits its *in vivo* transcription

(Choo et al., 1994). An improved selection method has been described recently by Greisman and Pabo, which takes into consideration of the context-dependent interaction from neighbouring fingers in the zinc finger-DNA recognition. The approach involves gradually extending a new zinc finger protein across the desired 9- or 10- base pair target site, adding and optimizing one finger at a time so that to ensure that the new finger is always selected in a relevant structural context. Proteins that specifically recognize a TATA box, a p53 binding site, and a nuclear receptor element have been obtained by this approach (Greisman & Pabo, 1997).

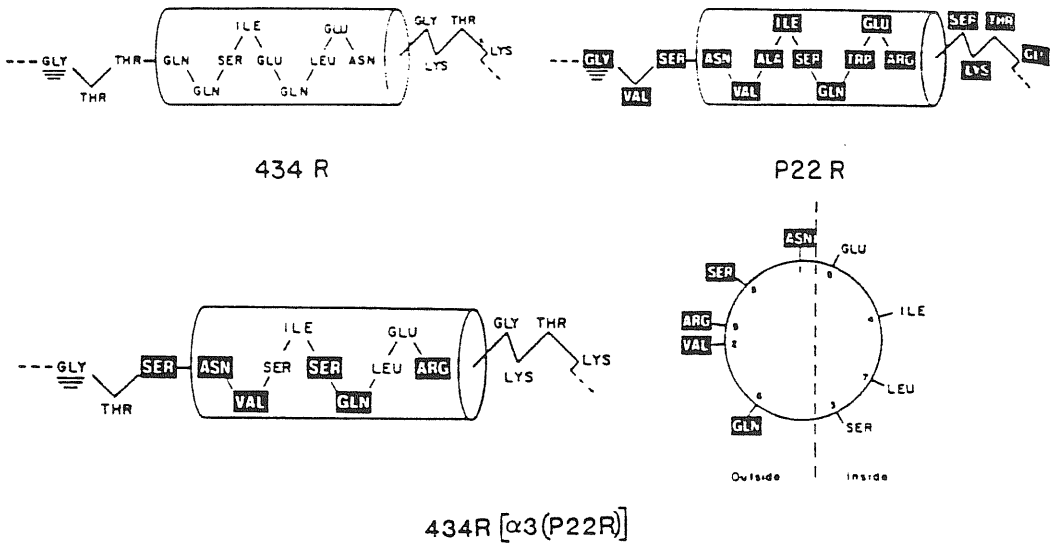
The "helix swap" experiment

Wharton *et al* has substituted the putative recognition helix of 434 repressor with the putative recognition helix of 434 Cro protein to create a hybrid protein named repressor*. The specific DNA contacts made by repressor* are like those of 434 Cro protein. (Wharton et al., 1984)

In a further "helix-swap" experiment of Wharton and Ptashne (Wharton & Ptashne, 1985), the solvent exposed residues of the 434 repressor recognition helix ($\alpha 3$ helix) were replaced with the corresponding residues from the recognition helix of the *Salmonella* phage P22. The resulting protein 434R[$\alpha 3$(P22R)] bound specifically and with high affinity to P22 operators. The subsequent experiments demonstrated that combining the 434 and 434[$\alpha 3$(P22R)] repressor monomers can form a heterodimer, which specifically recognizes a chimeric P22/434 operator that lacks two-fold rotational symmetry (Hollis et al., 1988), that is usually recognized by natural repressor homodimers (Table 2.1). The 434 and 434[$\alpha 3$(P22R)] repressors are also able to form stable heterodimer complex *in vivo,* and efficiently bring about repression through binding to the P22/434 hybrid operator in *E. Coli.* (Webster *et al.*, 1992). The repression level is comparable with that achieved by the 434, 434[$\alpha 3$(P22R)] homodimer binding to their cognate operators. See Figure 2.5 for an illustration of the "helix swap" experiment.

Designed transcription factors will be useful for the target recognition of specific cellular genes. The use of particular DNA binding domains in a hybrid (or the addition of other domains) may allow a protein to interact with other cellular factors or to modulate a particular regulatory pathway. The structure-based design of custom-built DNA-binding proteins should facilitate the development of efficient and specific reagents for both research and biomedical applications.

(A)



(B)

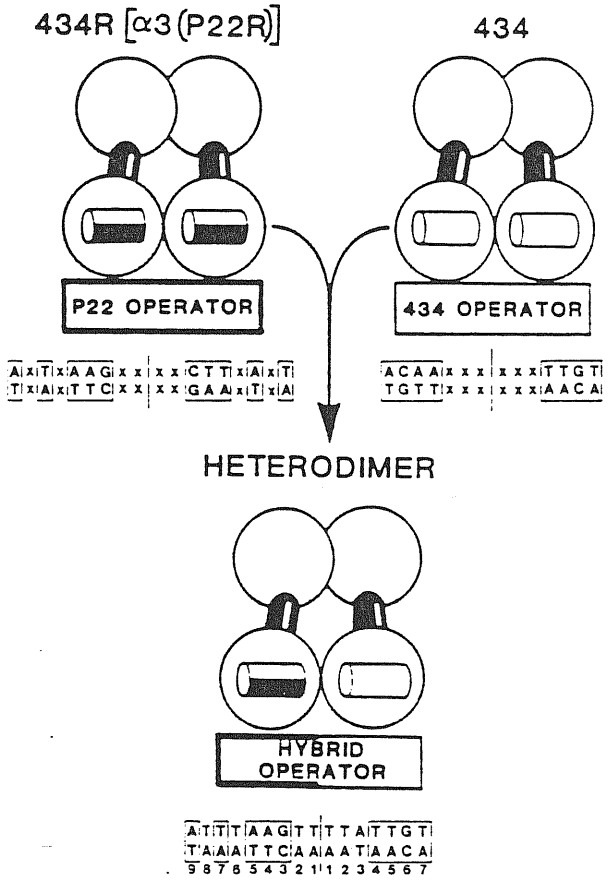


Figure 2.5. Schematic outline of the "helix-swap" experiment. (A) The construction of the 434R[α3(P22R)] (from Wharton *et al*, 1995). (B) The heterodimer recognition of a 434-P22 hybrid operator (Hollis *et al*, 1988)

2.5 Binding site selection from random DNA pools — A useful tool to investigate protein–DNA interaction

Since Kinzler and Vogelstain first combined the *in vitro* selection with *in vitro* amplification to identify the TFIIIA (Kinzler & Vogelstein, 1989) recognition sequence from the human genome DNA pool, a number of different laboratories have developed a variety of methods for expanding this technique [see (Blackwell, 1995; Szostak, 1992; Wright & Walter, 1993) for review and citations therein]. This general approach is usually called as SAAB [Selected and Amplified Binding Sites, (Blackwell & Weintraub, 1990)], CASTing [Cyclic Amplification and Selection of Targets, (Funk & Wright, 1992)], TDA [Target Detection Assay, (Thiesen & Bach, 1990)] or SELEX [Systematic Evolution of Ligands by EXpotential enrichment, which is more generally used for the aptamers selection, (Gold *et al.*, 1995)]. A major simplification and improvement of this method is the use of a synthetic DNA template that contains a degenerate central region flanked by two PCR primer-binding sites. The synthetic oligonucleotide is conversed to double-stranded DNA by priming DNA synthesis with the 3' primer. The resulting pool is incubated with the protein of interest, and DNA-protein complexes are isolated. DNA recovered from the first cycle of selection is then amplified by PCR. This generates a subset of the starting pool, and can be further enriched in binding species by additional cycles of selection and amplification until sufficient specificity is obtained to justify cloning and sequencing. Multiple rounds of selection, with a modest enrichment of 10-100 fold per cycle, makes this protocol possible to isolate specific binding sequences even under circumstances in which these sequences are extremely rare or when small amount of proteins are used, as long as the affinity of the sequence-specific interaction exceeds that of the non-specific binding.

The *in vitro* selection has been proved to be a very powerful approach for new discoveries extending from basic research to the diagnostic and therapeutic purposes. The consensus binding sites or *in vivo* targets for many different proteins have been determined by using this technique. The use of crude cell extract (Funk & Wright, 1992; Pollock & Treisman, 1990), or associated protein complexes (Chittenden *et al.*, 1991) have permitted to identify sequences that interact with one or more factors in a multiprotein complex, that gives further information about the protein-protein interaction in the biological processes. The *in vitro* selection method also makes it possible to identify RNA and single-stranded DNA molecules (termed "aptamers") which form complex structures that are capable of highly specific molecule recognition [see (Gold et al., 1995; Szostak, 1992)] for review. Aptamers have been identified specifically binding to small ligands, non-nucleic acid binding proteins (such as

cytokines and growth factors), as well as small organic molecules (such as ATP and theophylline).

The *in vitro* selection approach is especially useful for studying the DNA recognition property of different protein family members which usually recognize related DNA sequences. The high sensitivity of this method, together with other biochemical techniques (e.g. EMSA, "footprint", functional assay etc.), makes it possible to reveal very subtle differences in the binding sequence preference that is usually the determinant for different regulatory activities. As more structures for representative members of DNA binding protein families are available, coupled with the mutagenesis approach, this strategy is becoming increasingly useful for explaining how different family members recognize their respective cognate sites.

A crucial step in the selection is the isolation of protein-DNA complex. A variety of methods have been developed, like EMSA (Electrophoretic Mobility Shift Assay) to isolate the bound DNA corresponding shifted band (Blackwell, 1995; Blackwell *et al.*, 1990; Blackwell & Weintraub, 1990), filter-binding (Gold et al., 1995; Thiesen & Bach, 1990) to retain the protein together with bound DNA on the nitrocellulose membrane, immunoprecipitation (Funk & Wright, 1992; Pollock & Treisman, 1990), or binding to affinity columns (Chittenden et al., 1991; Pierrou *et al.*, 1995). For an affinity column selection, the protein of interest can be engineered to obtain a selectable "tag", such as glutathione S-transferases (Pierrou et al., 1995).

Different isolation methods have been explored in the single-chain repressor binding site selection. With the pure protein in hand, we mainly concentrated on the EMSA and filter binding methods. As discussed by Blackwell (Blackwell, 1995), the most significant advantage of EMSA is that it allows visualization of each selection step, making apparent the relative ratios of bound and free DNA, specific and non-specific binding. It thus indicates the stringency of selection and reveals whether the selection for specific binding has occurred. The nitrocellulose filtration method is radioactivity eliminated and simpler to perform comparing with EMSA. It is also less time consuming. Whereas this method is not able to give a clear or direct indication of whether selection of specific sites really has occurred. Hence an analytical EMSA was employed as a complement for checking the selection process.

3. Design and construction of single-chain repressor analogs with altered DNA-binding specificity

The first aim of this sub-project was to obtain a chimaeric single-chain repressor analog by introducing the DNA-contacting amino acid residues of the phage P22 repressor into one of the DNA-binding modules of the RR69 framework that contains two 434 repressor DBDs. Casette mutagenesis was used to introduce Thr→Ser, Gln→Asn, Gln→Val and Glu→Ser changes at the -1, 1, 2 and 5 positions of the α3 helix, respectively.The second aim was to obtain a symmetrical single-chain repressor that contains two identical DBDs mutated in the above described manner. The third aim was to characterize the new proteins and to show that they are able to bind to their relevant cognates and also to repress a β-galactosidase reporter gene in vivo.

3.1. Results

Construction of single-chain repressor analogs

The gene encoding two direct repeats of the N-terminal domain of the phage 434 repressor was constructed from two DNA fragments obtained by independent PCR amplifications on λgt10 template (Huynh *et al.*, 1984). This construct can be described as $R_{1\text{-}69}L_{70\text{-}89}R_{1\text{-}69}$, where R refers to the repressor function of the N-terminal DNA-binding domain, L refers to a linker sequence and the subscript numbers identify the amino acids of the natural repressor. For the sake of simplicity, we use the abbreviation RR69 for this homodimeric single-chain repressor. The $L_{70\text{-}89}$ sequence was chosen because it is part of the natural linker that connects the N- and C-terminal domains of the intact repressor (Carlson & Koudelka, 1994). Computer modelling based on the R1-69/operator crystal structure (Aggarwal *et al.*, 1988) showed that, although substantially shorter peptide sequences could span the distance between the C- and N-termini of the two subunits, the chosen long linker should also be suitable as it is likely to adopt a flexible conformation (Percipalle *et al.*, 1995). Silent mutations were then introduced into the RR69 coding gene near the borders of the α3 recognition helix of the second domain to generate unique restriction sites for *Kpn*I and *Xho*I. This enabled us to perform a "helix swap" experiment similar to that described (Wharton & Ptashne, 1985) but which is restricted to the second domain of RR69. Amino acid replacements at positions -1, 1, 2 and 5 of the α3 helix of the 434 repressor with the corresponding residues of the P22 phage c2 repressor resulted in a heterodimeric single-chain repressor, abbreviated as RR*69. Identical amino acid replacements were then performed in the first domain of RR*69 to obtain a homodimeric single-chain repressor

R*R*69, which contains two DNA-binding domains of 434 repressor with the P22 repressor amino acids at the DNA-binding positions. In order to evaluate the *in vivo* function of these single-chain repressor analogs, the control genes coding for the natural 434 repressor cI, its N-terminal 1-69 domain R69, and a frameshifted derivative of the latter, R(-), coding for 16 residues only, were also constructed and cloned into the expression/detection vectors. The repressor analogs used in this study are listed in Figure 3.1(a) and the altered nucleotide and amino acid sequences are shown in Figure 3.1(b).

Construction of a system to detect *in vivo* repressor-operator interaction

The system used to detect *in vivo* repressor-operator interaction in this work is conceptually similar to other systems (Lehming *et al.*, 1987; Wharton & Ptashne, 1987), with the major difference that the repressor gene and the operator-reporter gene fusion are on the same replicon. The repressor analog genes, devoid of any regulatory element of the imm^{434} region, were cloned into derivatives of the pRIZ' *E. coli* expression vector (Simoncsits *et al.*, 1994). This vector contains the *lac* operator controlled *rrnB* P2 promoter (Lukacsovich *et al.*, 1990; Simoncsits *et al.*, 1988) combined with an improved ribosomal binding site (Simoncsits *et al.*, 1994) to establish high level expression of cloned genes. It also contains the *lac*I^q-*lac*pro-*lac*Z'(1-146) region and it is abbreviated here as pRIZ'O_{lac}, where O refers to the operator site. To replace O_{lac}, first it was deleted by a simultaneous creation of a unique *Nde*I site to obtain pRIZ'O(-) (Figure 3.1(d)). Natural or designed operators, as listed in Figure 3.1(d), were then cloned into this site between the *lac* promoter and the *lac*Z' gene. The repressor analogs (Figure 3.1(a)) and operators (Figure 3.1(d)) can be shuffled, by using unique restriction sites (Figure 3.1(c)), to obtain any combination of them in the same vector. Thereby a simplified system is created in which a single expression vector carries all the necessary elements (a regulated repressor gene and its putative operator target placed upstream of a reporter gene) to detect *in vivo* repressor-operator interactions. The *lac*Z'(1-146) reporter gene on the plasmid gives rise β–galactosidase activity only through α–complementation (Ullman *et al.*, 1967) with a defective β-galactosidase encoded by the episome of the host strain XL1-Blue (Bullock *et al.*, 1987). The principle of the detection is based on parallel and consecutive derepression-repression mechanisms taking place at three operator sites. The lactose analog IPTG (isopropyl-β-D-thio-galactoside) derepresses the *rrnB* P2 promoter-*lac* operator controlled repressor and concomitantly the *lac* promoter-operator controlled *lac*ZΔM15 synthesis. Since the expression of the *lac*Z' reporter gene is not under *lac*I control, it is constitutive unless the derepressed, overproduced repressor analog occupies the created operator site. Thus the level of the observed β-galactosidase

activity depends mainly on the interaction of the repressor analog with the operator of the reporter gene.

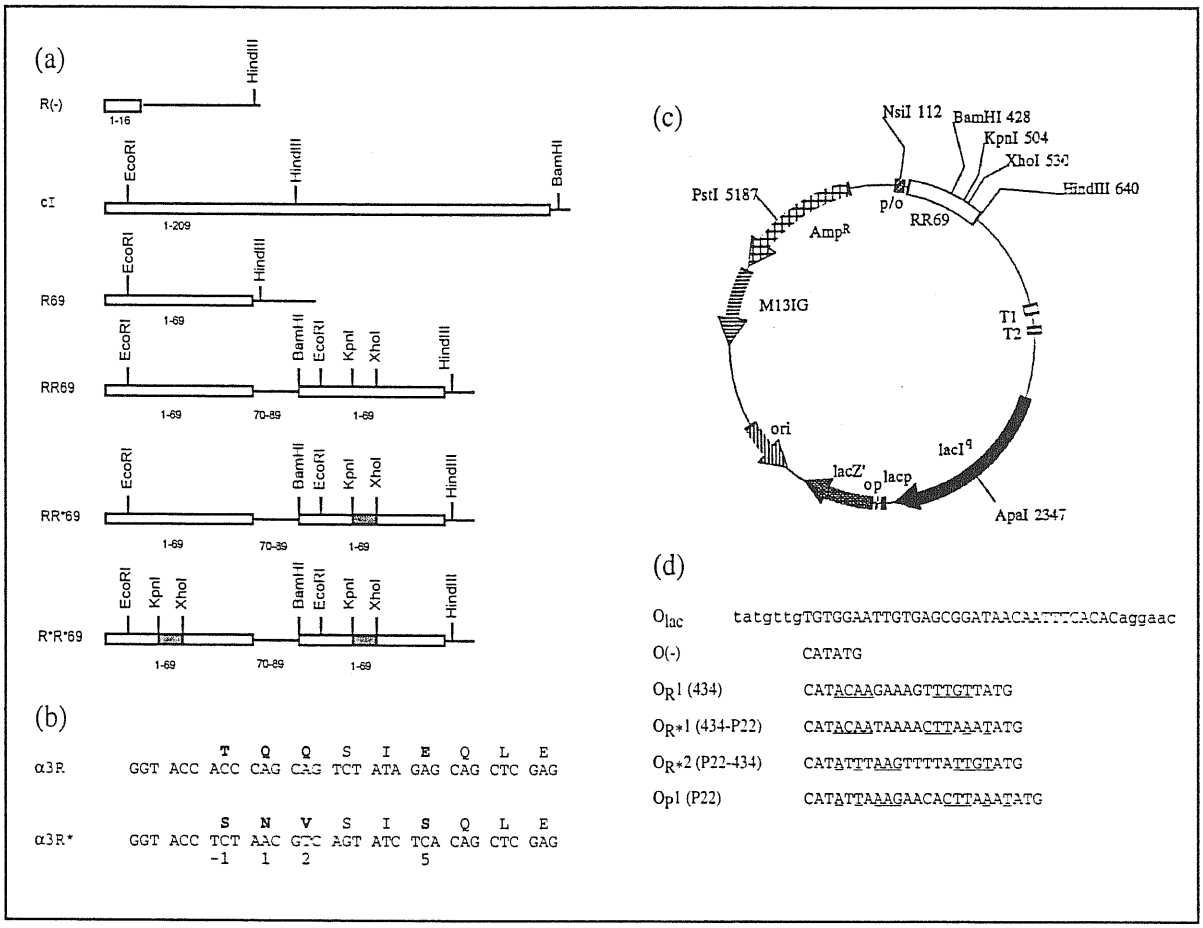


Figure 3.1. Components of pRIZ' vectors used to detect repressor-operator interaction *in vivo*. (a) Linear maps of the genes coding for the repressor analogs. Open boxes represent coding regions for 434 repressor parts, the linker of the single-chain repressors is shown as a solid line between boxes and is labeled 70-89. Shaded areas between *Kpn*I and *Xho*I sites represent altered α3 recognition helices. (b) Nucleotide and amino acid sequences of α3 helix regions: α3R, α3 helix of 434 repressor; α3R*, altered α3 helix in R* domain after replacement of amino acids -1, 1, 2 and 5 of α3 with the corresponding residues of the P22 repressor. (c) Map of the pRIZ' vectors exemplified by pRIZ'O_{lac}RR69. Repressor analog genes (a) are located as shown for RR69 and are placed downstream of the *rrnB* P2 promoter-*lac* operator region (p/o). T1 and T2 represent tandem transcriptional terminators of the *rnnB* operon. (d) Operator sequences at the op site of (c), listed as upper strand sequences in *lac* promoter-op-*lacZ*' arrangement. Lower case letters show operator flanking sequences. The consensus bases of the 434 and P22 operators are underlined.

In vivo interaction of single-chain repressors with cognate, half-cognate and non-cognate operators.

The *in vivo* detection system described above was applied to most of the repressor analog-operator pairs of Figure 3.1 and the results are summarized in Table 3.1. As it was observed that different operator constructs gave rise to different β-galactosidase activities under non-repressed condition, relative β-galactosidase activities were calculated and are shown in Table 3.1. This presentation of the data makes it easier to compare the repression levels, obtained by binding of different repressor analogs to a given operator or by binding of a given repressor analog to different operators.

Table 3.1. Recognition of operators by different repressors *in vivo*

Repressor[a]	% β-galactosidase activity[b] observed with operators[c]			
	O_R1	$O_{R*}1$	$O_{R*}2$	O_P1
$R^{(-)}$	100±14	100±12	100±9	100±16
R69	33±4	115±6	120±4	130±10
RR69	**27±3**[d]	120±11	138±12	152±9
cI	**25±3**	ND[e]	ND	ND
RR*69	55±5	**49±4**	**30±4**	53±3
R*R*69	67±3	64±2	43±2	**43±4**

[a] Repressors or repressor analogs are shown in Figure 3.1(a) and 3.1(b).

[b] Data measured in Miller units (Miller, 1972) are expressed in relative β-galactosidase activity as a percentage of the units obtained with R(-) nonfunctional polypeptide (non-repressed condition). 100% was given for all non-repressed operators although the absolute units were different for different operators (see text for data). Data shown are based on 4-6 independent assays (mean ± standard deviation).

[c] Operator sequences are listed in Figure 3.1(d).

[d] Data for cognate or putative cognate interactions are shown in bold.

[e] Not determined. Vectors containing these repressor-operator combinations were not constructed.

The most important observation based on the data of Table 3.1 is that the homodimeric single-chain repressor RR69 is as active *in vivo* as is the natural counterpart cI when tested on the natural O_R1 operator (27 and 25% relative β-galactosidase activities, respectively, which represent about fourfold repression). The N-terminal domain R69 shows lower, but significant, repression (33% relative value, threefold repression). The mutant single-chain repressors show significantly lower, but clearly detectable repression. These latter observations suggest that the intracellular concentrations of the repressor analogs are high enough to give rise to interaction of R69 with O_R1 as well as interactions of the single-chain repressor analogs with half-cognate and non-cognate operators. *In vitro* studies (see below) showed that these interactions do take place at high protein concentrations. Nevertheless, the repression observed on O_R1 with different repressor analogs follows the expected order (see O_R1 column of Table 3.1). Comparison of the different single-chain repressor analogs on O_R1 shows that the cognate RR69-O_R1 interaction (where two operator subsites may interact with two cognate DNA-binding domains) is the strongest, the half-cognate RR*69-O_R1 interaction (where one binding domain of RR*69 is cognate to either subsites of O_R1) is medium and the non-cognate R*R*69-O_R1 interaction is the weakest.

A similar order of cognate and half-cognate interactions on the hybrid, asymmetric operators $O_{R*}1$ and $O_{R*}2$, and on the symmetric operator O_P1 (O_R1 of P22) can be detected, as shown in the respective columns of Table 3.1. It is interesting to note that on these operators the non-cognate RR69 and the R69 do not show any repression and surprisingly, the β-galactosidase activities are 20-50% higher than those observed with the nonfunctional R(-) control. Our experimental data do not provide a rational explanation for this observation.

We mention here that the pRIZ' vectors contain the *lacL8* mutation which makes the *lac* promoter insensitive to positive regulation by CAP (Ebright *et al.*, 1984) This implicates that the β-galactosidase activities (Miller, 1972) may be too low in our system to detect repression. This is indeed true for the constructs containing the *lac* operator, but substantially higher activities were observed when the *lac* operator was replaced with the operators used in this study. The β-galactosidase activities obtained in a typical experiment were 4.8, 164, 235, 284 and 245 units for O_{lac}, O_R1, $O_{R*}1$, $O_{R*}2$ and O_P1 operators, respectively, when they were tested under similar non-repressed conditions. Compared to O_{lac}, even its deletion derivative, O(-) showed 4-5 times higher β-galactosidase activity. This observation, together with the fact that different synthetic operator sequences cloned in different orientations all gave rise to elevated β-galactosidase activity, makes it unlikely that an insertion of a fortuitous promoter element is responsible for the higher activities.

Expression and purification of single-chain repressors

The single-chain repressors were initially expressed using the pRIZ' expression vectors in XL1-Blue cells in relatively low levels (less than 1% of the total cell protein). For large scale expressions, T7 promoter-based expression systems were chosen (Studier *et al.*, 1990). Vector construction, expression and purification are described in the Materials and Methods section. Figure 3.2 shows the purities of the single-chain repressors used in the *in vitro* studies.

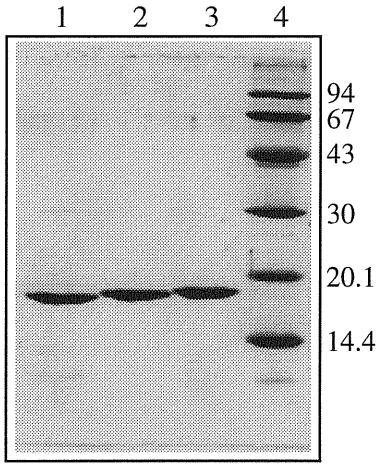


Figure 3.2. SDS-PAGE analysis of the purified single-chain repressors. Lane 1, R*R*69; lane 2, RR*69; lane 3, RR69; lane 4, molecular weight standards (kD). 100 pmol proteins were analyzed, bands were detected by Coomassie Blue staining.

In vitro DNA-binding properties of single-chain repressors

Electrophoretic mobility shift assay (EMSA) was used to determine the protein concentration required for half-maximal binding, which represents the apparent equilibrium dissociation constant (K_d), if the DNA probe concentration is negligible compared to the protein concentration and the binding equilibrium is not perturbed under the assay conditions (Carey, 1991; Fairall *et al.*, 1992). Small increments in protein concentration were used in these titration experiments to be able to define, even by visual inspection of the autoradiograms, a narrow concentration range in which half-maximal binding takes place. Figure 3.3 shows representative experiments performed at the same time with different operator-repressor cognate pairs. DNA probes

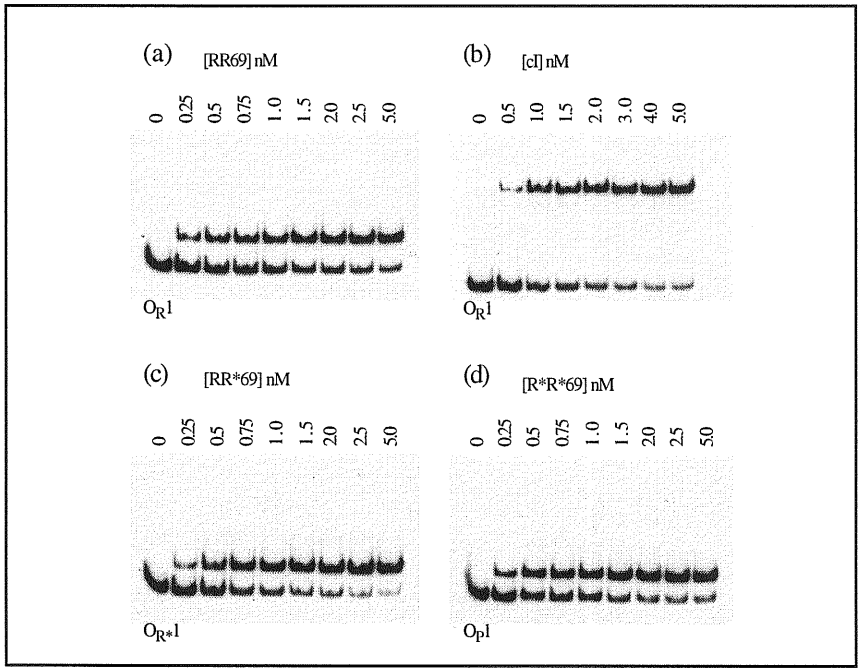


Figure 3.3. Determination of half-maximal binding for cognate protein-DNA interactions by EMSA. The ^{32}P end-labeled DNA probes (157-160 bp) were generated by PCR. The binding reactions and electrophoretic separations were performed as described in Materials and Methods. (a) Interaction of RR69 with O_R1 (O_R1 of 434). (b) Interaction of cI repressor with O_R1. (c) Interaction of the heterodimer RR*69 with $O_{R*}1$ (434-P22 hybrid operator containing the consensus boxes of 434 O_R1 and P22 O_R1). (d) Interaction of the double-mutant homodimer R*R*69 with O_P1 (O_R1 of P22). The DNA probe concentrations were less than 20 pM. Protein concentrations are shown above the respective lanes of the gels. For easy comparison of the affinities of RR69 and cI to O_R1, concentrations of cI in (b) were given as dimer equivalents.

containing the cloned operator sequences were about 160 bp long. Based on these and on many other experiments, RR69 and the natural cI repressor bind to O_R1 with approximately the same affinity (K_d~$1x10^{-9}$ M) under our assay conditions. At the same time, half-maximal binding with the isolated DNA-binding domain R69 was observed between 100-200 nM concentration (expressed in dimer equivalent for better comparison). For the O_R*1-RR*69 and O_P1-R*R*69 interactions the estimated K_d is in the range of $5\text{-}8x10^{-10}$ M. These data are in close agreement with those reported for the naturally dimerized 434, P22 and hybrid repressors and for R69 (Hollis *et al.*, 1988; Wharton & Ptashne, 1985; Koudelka & Lam, 1993; Bell & Koudelka, 1993; Carlson & Koudelka, 1994).

Substantially higher single-chain repressor concentrations than those shown in Figure 3.3 were also tested with the three operators (O_R1, O_R*1 and O_P1) in all possible operator-repressor combinations in order to detect non-specific binding or cross-reactivity. The appearance of shifted bands in non-cognate interactions may indicate either non-specific binding or lower affinity specific binding due to sequence homology between different operator subsites. The appearance of multiple shifted bands in both cognate and non-cognate interactions may be due either to non-specific binding outside the operator region or to protein aggregation at the operator site. At 200 nM concentration, RR69 did not give detectable bandshift with O_R*1 and O_P1, and gave only a single, specific shift with its cognate O_R1 (not shown). RR*69 gave a single, specific bandshift with its cognate O_R*1 operator in a broad concentration range and only weak bands of slower mobilities were seen at 200 nM concentration (Figure 3.4(c)). With the half-cognate O_R1 and O_P1 operators, however, RR*69 was shown to interact at relatively low concentrations. In these cases, both single- and double-shifted bands could be detected at 10-25 nM and increasing protein concentration gave rise to a predominantly double-shifted band at around 200 nM concentration (Figure 3.4(a) and 3.4(b)). It is possible that at higher concentrations, RR*69 forms protein dimer (or oligomer) at the half-cognate binding site of the long DNA probe. Similar binding and protein aggregation under identical conditions was not detected on a DNA probe which did not contain an operator site (see O(-) of Figure 3.1(d); data not shown), indicating that the initial DNA binding by RR*69 is mainly due to the presence of half-cognate operator sites and not to non-specific interactions. These observations are in agreement with the *in vivo* behaviour of RR*69 which showed nearly twofold repression on the half-cognate operators (see Table 3.1). R*R*69 was shown to bind with single stoichiometry to both its cognate O_P1, and to the half-cognate O_R*1, up to about 100 nM concentration, but the protein concentrations at half-maximal binding were substantially different: 20 nM for O_R*1 (Figure 3.4(d)) and 0.5-0.8 nM for O_P1 (Figure 3.3(d)), indicating about 30 fold higher affinity to the cognate O_P1 site over the

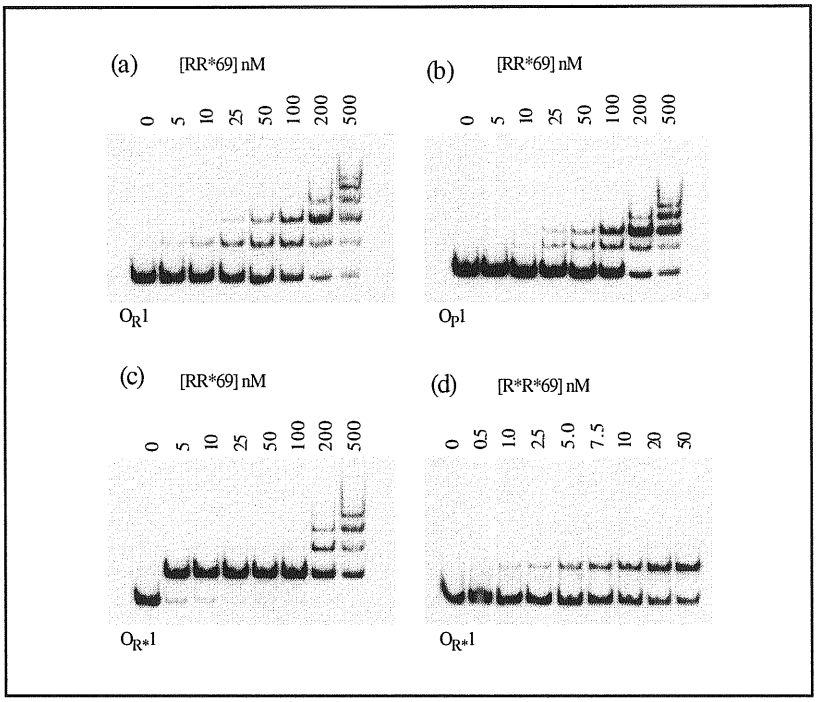


Figure 3.4. Detection of half-cognate interactions by EMSA. (a) RR*69-O_R1 interaction, the R domain of RR*69 is cognate to O_R1 half-sites. (b) RR*69-O_P1 interaction, the R* domain is half-cognate to O_P1 half-sites. (c) Interaction of RR*69 with its cognate operator $O_{R*}1$ under the same conditions as in (a) and (b), showing a significantly higher affinity as seen from the disappearance of the free DNA. See also Figure 3.3(c) for lower RR*69 concentration binding. (d) Interaction of R*R*69 homodimer with the half-cognate $O_{R*}1$ operator, showing a half-maximal binding with single stoichiometry at or slightly under 20 nM protein concentration. Only trace amounts of higher complexes could be detected at even higher (100 nM) concentration (not shown). For comparison with the cognate operator interaction see also Figure 3.3(d).

half-cognate site. On the non-cognate O_R1 site, R*R*69 did not show significant binding under 200 nM concentration.

Competition assays were also performed in order to support the results of the direct binding experiments, which suggested that the operator sites of the long (160 bp) DNA probes were the targets of the specific binding. In these assays, the radioactively labelled long DNA probes were allowed to form complexes with the respective cognate repressors, then large excess of short, double-stranded DNA competitors containing only the operator sequences were added. Figure 3.5 shows that the cognate operator competitors caused complete dissociation of the preformed complexes, in contrast to the half-cognate or non-cognate competitors which caused only partial or no dissociation. These results are in qualitative agreement with those obtained in the direct binding experiments.

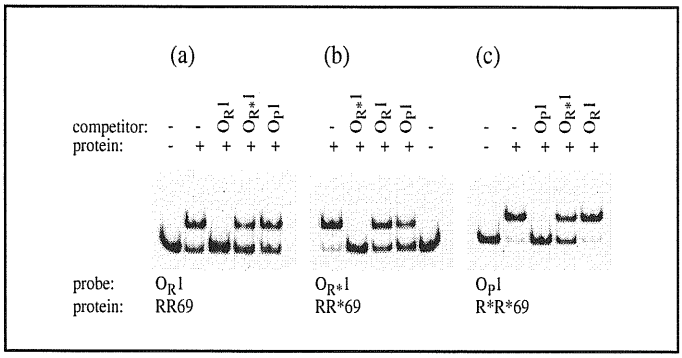


Figure 3.5. Specificity of the cognate interactions as studied by qualitative competition assays. Complexes formed between operator probes and repressors such as O_R1-RR69 (a), $O_{R*}1$-RR*69 (b) and O_P1-R*R*69 (c) pairs were treated with large excesses of unlabeled double-stranded oligonucleotide competitors and the reaction mixtures were analyzed by EMSA. Operator probes were 157-160 bp long, ^{32}P end-labeled PCR fragments. The competitor operators were 17 bp (O_R1), 18 bp ($O_{R*}1$) and 20 bp (O_P1) synthetic double-stranded oligonucleotides. The final concentrations were: <20 pM probe DNA, 25 nM non-specific carrier DNA, 5 nM single-chain repressor and 100 nM competitor oligonucleotides.

The kinetic stabilities of the single-chain repressor interactions with their cognate operators were studied by detecting the dissociation rate constant (k_{off}). The half life time of the wild-type 434 repressor cI, RR69, RR*69 and R*R*69 dissociation from the approximate 60 bp DNA probe was 72 sec, 528 sec, 6hr and >10 hr, respectively. The much quicker dissociation rate of the DNA complex of cI than that of RR69 is probably due to the dissociation of the cI dimer to its subunits at low protein

concentrations. It seems that the amino acid substitutions make the R* domain more stable on binding to its cognate sequence than the R domain.

DNase I protection experiments, performed in all possible operator-repressor combinations are shown in Figure 3.6. Relatively high repressor concentrations (100 and 200 nM) were used in order to be able to detect both the specific and the possible non-specific or non-cognate interactions. As expected, strong protection could be detected only at the operator sites with the corresponding cognate repressors. The natural cI and the single-chain RR69 repressors showed identical footprints at the O_R1 site. RR69 protected neither $O_{R*}1$ nor O_P1. RR*69 and R*R*69 showed strong protection at their respective cognate operator sites $O_{R*}1$ and O_P1. RR*69 showed very weak protection at 200 nM concentration on its half-cognate operators, O_R1 and O_P1 (see Figure 3.6(a) and 3.6(c), respectively). Under the same conditions, EMSA showed the disappearance of over 50% of the free O_R1 and O_P1 probes and mainly double-shifted bands were detected (Figure 3.4(a) and 3.4(b)). The assumption based on the EMSA results that, at high protein concentrations, RR*69 could dimerize at the half-cognate operator sites could be supported by the weak DNase I protection patterns which were restricted to the operator regions of the long DNA probes. The difference between the extent of the DNase I protection and the extent of binding observed by EMSA could be due to different kinetic stabilities of the complexes under the conditions of these two different methods. R*R*69 showed partial protection of the O_R1 non-cognate operator, and strong protection of the $O_{R*}1$ half-cognate operator. In this latter case, however, the strongly protected area of the footprint was approximately 6 nucleotides shorter (see lane R*R*69 of Figure 3.6(b)) than that obtained with the cognate RR*69, indicating an asymmetric occupancy of the half-sites of the hybrid $O_{R*}1$ operator. For the same interaction, EMSA showed a single stoichiometry binding with a K_d of ~$2x10^{-8}$ M. It is possible that one of the binding domains of R*R*69 may initially approach and interact with the consensus CTT.A.T P22 operator half-site (Poteete *et al.*, 1980), causing a local increase in the effective concentration of the covalently joined second domain, which can then recognize with lower affinity the homologous aTTgtaT sequence of the other, 434 specific half of the $O_{R*}1$ operator. This mode of binding is somewhat similar to that observed, for example, in the glucocorticoid receptor-DNA complex which shows that one monomer of the DNA-induced homodimer binds to a specific half-site and the other one binds to a non-specific sequence (Luisi *et al.*, 1991). The Arc repressor provides a further example for such binding, detected by EMSA, to a DNA fragment containing only one specific half-site (Brown & Sauer, 1993). It is likely, however, that the second domain of R*R*69 interacts with a correctly spaced and homologous half-site sequence as shown above.

The single-chain architecture of the covalently dimerized N-terminal domains may, in principle, allow the recognition of operator half-sites with different

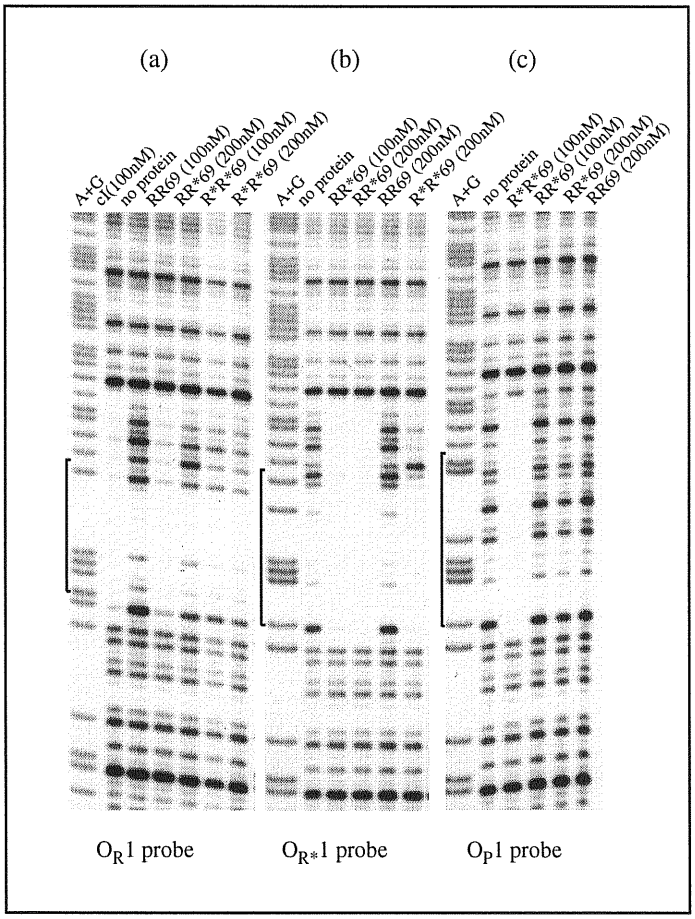


Figure 3.6. Analysis of single-chain repressor-operator interactions by DNase I protection assay. Operator probes were 157-160 bp long PCR products and the 5' -end of the lower strands (complementary to the sequences listed in Figure 3.1(d) were ^{32}P end-labeled. Protection of the O_R1 region (a), the $O_{R*}1$ region (b) and the O_P1 region (c) by different repressors are shown. The operator regions are marked by vertical lines and aligned with the A+G control lanes.

juxtapositions assuming that the linker is really flexible and that it allows the independent free movement of the individual binding domains within the constraint of the linker length. We have tested this possibility with mutant O_R1 operators in which the consensus recognition boxes were spaced by 5, 7, 8 and 10 nucleotides instead of the natural 6 nucleotides spacing. As detected by EMSA, no binding took place between RR69 and the mutant operators at a protein concentration which is 100 fold higher than that required for half-maximal binding to O_R1 (not shown). A similar result was obtained with RR*69, which did not show significant binding at 50 nM concentration to an $O_{R*}1$ hybrid operator analog which contained 6 nucleotides between the consensus operator boxes of 434 (ACAA) and P22 (CTT.A.T) instead of the 5 nucleotides spacing of the designed $O_{R*}1$ operator (not shown). These findings and binding site selection experiments performed with RR69 and RR*69 (Chen *et al.*, 1997) show that a strict spacing of the contacted half-operator boxes is required for high affinity binding to take place. The spacing preferred by the single-chain repressors is the same as that observed for the naturally dimerized repressor counterparts: see RR69 vs. 434 repressor (Harrison & Aggarwal, 1990 and references therein) and RR*69 vs. the 434+434R[α3(P22R)] heterodimer complex (Hollis *et al.*, 1988). Therefore it is likely that the covalently joined binding domains of the DNA-bound single-chain repressors are in the same spatial arrangements as the R1-69 domains in the complexes with DNA. X-ray crystallography studies for the latter complexes revealed significant interdomain contacts at the dimer interface (Anderson *et al.*, 1987; Aggarwal *et al.*, 1988) which are believed to be the same in the intact repressor (Aggarwal *et al.*, 1988). We assume that the same dimer contacts primarily determine the orientation of one monomer with respect to the other in the DNA-bound single-chain repressors and therefore the linker plays no role in the juxtapositioning of the binding domains. In other words, the natural linker seems to be flexible enough to undergo conformational changes, probably enforced by the contacts between the two domains.

3.2 Discussion: Properties of the single-chain repressor analogs

We have constructed single-chain repressor analogs containing either a wild type and an engineered DNA-binding domain (RR*69) or two identically altered domains (R*R*69) , respectively. To the best of our knowledge, these are the first examples of covalent dimerization of engineered DNA-binding domains of a transcription factor, effected by a recombinant linker. The mutant zinc finger domains of the Cys_2His_2 type, obtained by rational or random mutagenesis (for reviews, see Berg & Shi, 1996; Rhodes *et al.*, 1996), also contain engineered, covalently linked DNA-binding

modules. These are, however, naturally linked in an existing framework. The difference in this work is that, for similar mutagenesis studies, a new single-chain framework has been created by covalent attachment of DNA-binding domains that naturally function in noncovalent dimers. For covalent dimerization an arbitrarily chosen, 20 amino acids long natural linker, present in the same intact repressor was used and proved to be suitable. It is likely that designed, flexible peptide linkers of suitable length such as those used, for example, in single-chain antibodies (Huston *et al.*, 1988) or in the P22 Arc repressor dimer (Robinson & Sauer, 1996) could also have been used in our constructs. Nevertheless, we chose the natural linker because practical considerations in the gene construction favoured the use of this domain-contiguous linker, and previous modelling studies indicated that it could undergo flexible conformational changes (Percipalle *et al.*, 1995).

It was previously shown that covalent joining of the DNA-binding domains of the 434 repressor by different strategies gave rise to an increased DNA-binding affinity compared to that of the isolated domain and that the single-chain molecules could be used to study conformational changes upon interaction with DNA (Percipalle *et al.*, 1995). It is shown here that the single-chain RR69 and the natural cI bind to O_R1 with the same affinity, therefore the covalent linker has the same binding enhancement effect as the noncovalent dimerization domain. RR69 recognizes operators only with the natural 6 bp spacing between the consensus half-sites. This suggests that the relative orientation of the two domains in the DNA-bound RR69 is the same as in the natural repressor and it is primarily determined by the interdomain contacts observed between the R69 monomers in the complexes with DNA (Aggarwal *et al.*, 1988). DNase I footprinting, *in vivo* repression assay and binding site selection experiments (Chen et al., 1997) further confirmed that RR69 and cI recognize DNA in a similar manner. The functional replacement of the noncovalent dimerization domain with the covalent linker resulted in a single-chain molecule with significantly reduced (by over 60%) molecular mass but with the DNA-binding properties of the parent molecule.

The single-chain repressors of this study selectively recognized their respective cognate operators with high affinity. Half-maximal binding was observed around or slightly under 1 nM repressor concentrations. Significant binding to non-cognate or to half-cognate sites took place at least at a hundred fold higher concentrations, with the exception that R*R*69 showed only 30 fold discrimination between the cognate O_P1 and the half-cognate $O_{R*}1$ operators. The interesting observation that, at higher repressor concentration (200 nM), the single-chain heterodimer RR*69 binds to the half-cognate operators O_R1 and O_P1 with predominantly double stoichiometry may serve as a starting point to elucidate a mechanism of DNA recognition by the single-chain repressors. Based on our experimental data, we assume that at high protein concentrations and at symmetric operator sites of long DNA probes the heterodimeric

single-chain repressors may dimerize. Our hypothesis is that in these cases an asymmetric interaction with single stoichiometry can initially take place: one domain makes specific, strong contacts with one half-site and the other one interacts weakly with the other half-site of the operator. This is probably the case since single-shifted band could be detected by EMSA at relatively low protein concentrations. At higher protein concentrations, this complex may rearrange by the replacement of the weakly interacting domain with the specific domain of a second molecule, which approaches the operator site by a sliding mechanism. The new complex can be stabilized by the stronger protein-DNA contacts and, possibly, by the formation of intermolecular protein-protein contacts. The involvement of the sliding mechanism in this hypothetical pathway seems to be essential, since the domain replacement, as above, by direct entry is very unlikely. By using short DNA probes containing only the operator sequences, dimerization could not be observed at even higher protein concentrations. However, the dimerization observed *in vitro* on longer DNA is likely to take place *in vivo* and this may explain the discrepancies occasionally observed *in vivo* .

The *in vivo* interaction of the single-chain repressors with operator DNA was studied in a simplified detection system. The one-plasmid system used here may be advantageous over the two-plasmid systems as no selective loss of either the repressor gene or the operator-reporter gene fusion can take place under the assay conditions. In this system the basic single-chain repressor RR69 was just as active as the natural 434 repressor. The fact that the isolated DNA-binding domain R69 caused only a slightly lower repression than either the covalent or the natural dimer does not contradict our *in vitro* binding data. Overexpression of the isolated N-terminal domain of lambdoid repressors (Sauer *et al.*, 1979) and the phage 16-3 C repressor (Dallman *et al.*, 1991) was shown to give rise to detectable activity *in vivo*. The mutant single-chain repressors RR*69 and R*R*69 were most effective when they interacted with their respective cognate operators. This indicates that our system is capable of detecting specific interactions, although the accurate data evaluation is somewhat obscured by the high level repressor expression in this system, and consequently, by the possible repressor dimerization at closely related or half-cognate operators, as discussed above. The level of repression in this system was generally about 2-4 fold. This low value may be due to "titration out" of the repressors by the multicopy operators. However, the level of repression in artificial assay systems seems to depend mainly on the repressor itself: the *lac* repressor showed a minimum of 200 fold repression in a two-plasmid system (Lehming *et al.*, 1987), while the 434 repressor (Wharton & Ptashne, 1987), the Arc repressor and its covalent dimer (Robinson & Sauer, 1996) showed only a modest few fold repression of single-copy reporter genes. Nevertheless, such a low repression level proved to be sufficient to isolate a new specificity 434 repressor mutant showing at least 50-150 fold operator discrimination *in vitro* (Wharton and Ptashne,

1987). The single-chain repressors of this study showed a similar degree of *in vitro* discrimination.

The major finding of this work is that both the wild-type and engineered DNA-binding domains of the 434 repressor can be covalently dimerized to form functional single-chain repressors which show selective, high affinity recognition of DNA operators composed of the respective subsites of the joined domains. To obtain engineered DNA-binding domains in the single-chain framework, we used the principle of the α helix redesign experiment (Wharton and Ptashne, 1985) and showed that the obtained homo- and heterodimeric single-chain repressors behaved similarly to their naturally dimerized counterparts (Wharton and Ptashne, 1985; Hollis *et al.*, 1988; Webster *et al.*, 1992). A potential advantage of the single-chain heterodimer is that, unlike the noncovalent heterodimer which exists in a mixture with two homodimers, it is homogeneous. This permits a better study of DNA-binding properties, for example by binding site selection from randomized DNA pools (Chen *et al.*, 1997). Moreover, the single-chain framework allows new approaches to be applied in the search for altered recognition specificities. A previous study showed that a combinatorial mutant library of the natural 434 repressor selected against natural or symmetrically altered mutant operators did not provide mutant repressors with new, high affinity recognition properties (Hu *et al.*, 1994). It is possible that weaker, but specific interactions are overlooked in this system, as simultaneous binding of two weakly interacting subunits may form an unstable complex with DNA. This problem may be overcome by using the single-chain framework, as it provides the possibility of altering only one of the domains whilst keeping the other one unchanged. The unchanged domain can provide a supporting interaction with operators containing a half-site cognate to it. Selection of a single-chain repressor library containing one unchanged domain and one partially randomized domain against operators containing a cognate half-site to the unchanged domain can identify single-chain repressor mutants with high overall affinity even if the intrinsic affinity of the isolated mutant domain toward the target half-site is not very high. For the selection of such libraries, it is possible to use the *in vivo* detection system described here to detect phenotypic differences on indicator plates (A. Simoncsits, S. Wang, I. Törö, J. Chen & S. Pongor, manuscript in preparation). Alternatively, more straightforward, direct genetic selection techniques (Elledge *et al.*, 1989; Mossing *et al.*, 1991) could be applied. The possibility to create combinatorial libraries with amino acid changes in one of the domains of the single-chain dimer, and the availability of *in vivo* selection methods together may provide a system which, similarly to the zinc-finger phage display systems (Rebar & Pabo, 1994; Jamieson *et al.*, 1994; Choo & Klug, 1994a, b; Wu *et al.*, 1995), could be used to derive further recognition rules and to generate new DNA-binding specificities in a given domain framework.

4. Recognition properties of the single-chain repressor analogs

*The aim of this sub-project was to define the optimum DNA recognition sequence for the single-chain repressor RR69 and RR*69 through binding site selection, targted mutagenesis and binding affinity studies. 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 nonpalindromic 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 base pairs. 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.*

4.1. Results

Selection of binding sites

Selèctions of binding sites from two different degenerate DNA pools for RR69 and RR*69 were 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 4.1. 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 (Simoncsits *et al.*, 1997). 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 *lacZ'* 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 (Simoncsits et al., 1997).

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 4.1(A) and 4.3(A). 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

1. Synthesize the random DNA pool oligo and convert it to double -stranded DNA by Klenow polymerase.

2. Incubate the single-chain repressors with the random DNA library to form the protein -DNA complex.

5. Repeat the selection cycle.

3. Isolate the protein-DNA complex and elute the bound DNA.

4. Amplify the bound DNA by PCR.

6. Convert the selected population to single-stranded DNA pool by asymmetric PCR with excess amount of primer AT421a.

7. Clone the selected population into the pRIZ' vector by loop insertion mutagenesis for sequencing and further characterization.

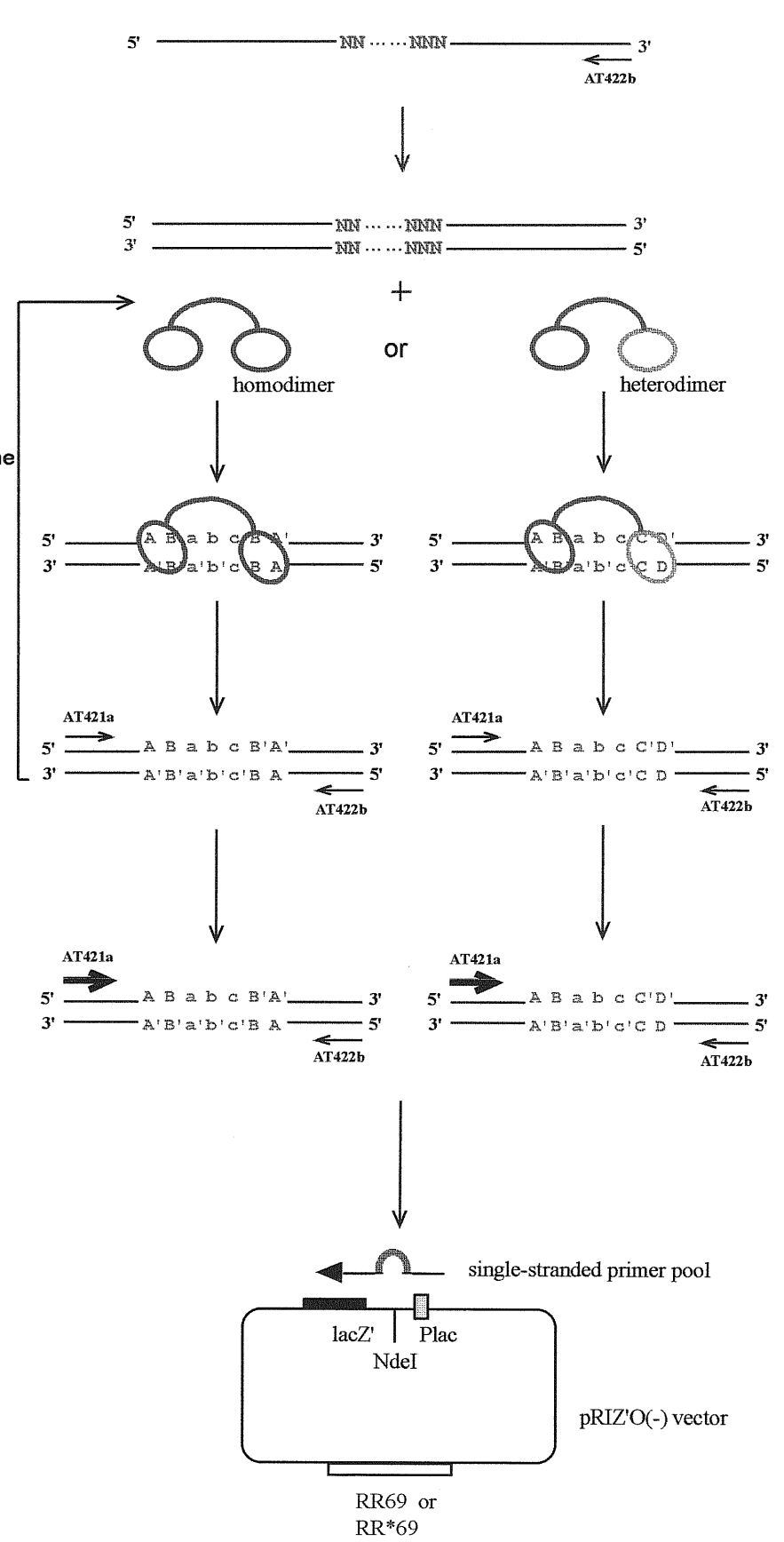


Figure 4.1. Scheme for the binding site selection and cloning of the selected sequences into pRIZ' vector by loop insertion mutagenesis.

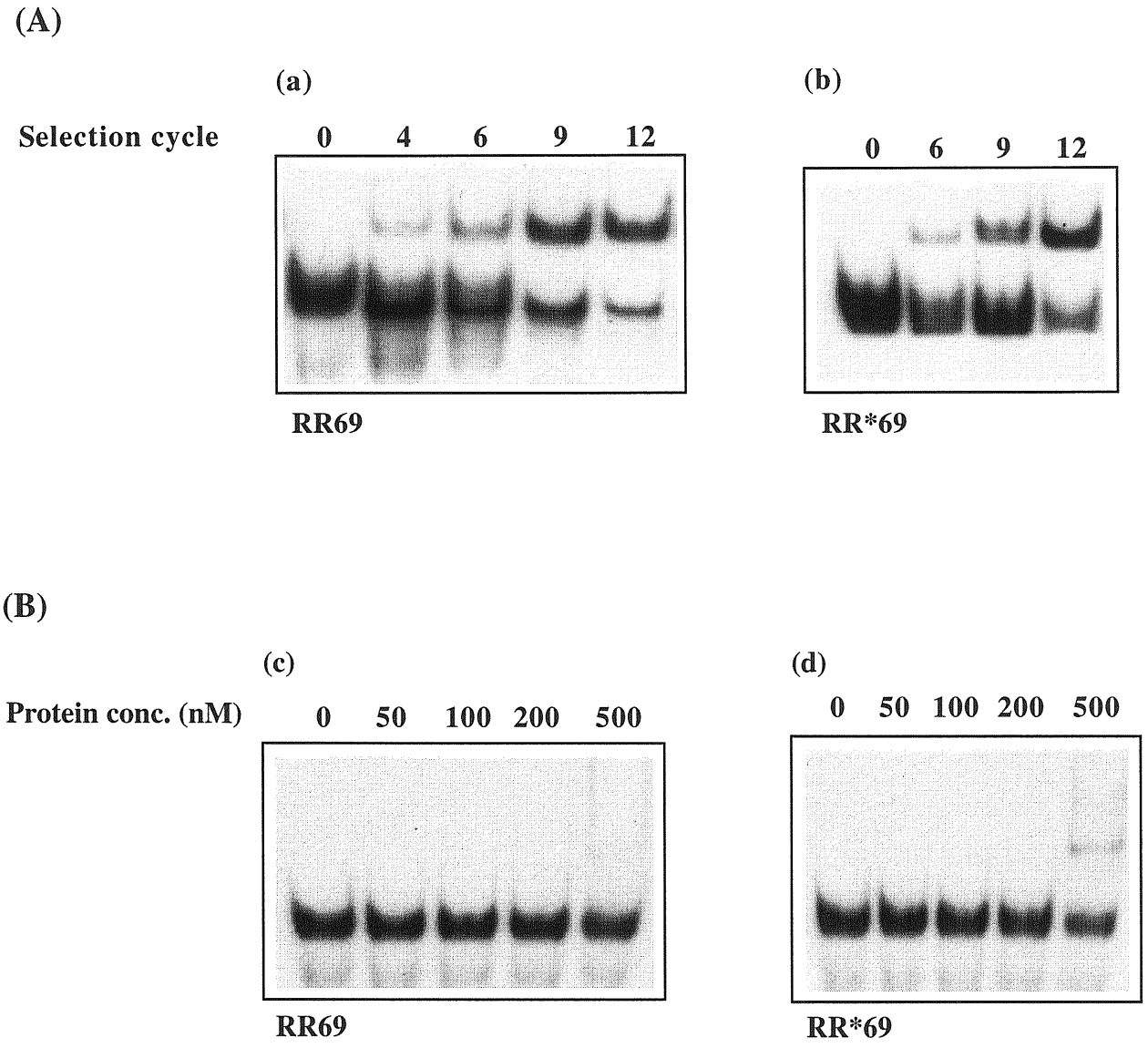
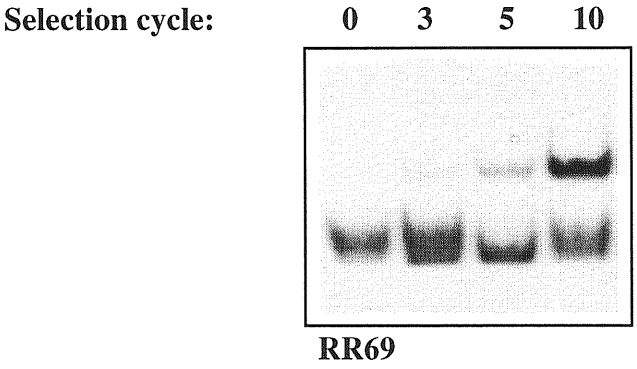


Figure 4.2. (A) Progressive enrichment of the binding sites during selection from the N14 pool for RR69 (a) and for RR*69 (b). EMSA was performed by using the selected pools as ^{32}P-labeled probes and the corresponding protein at 10 nM concentration. (B) In the binding of RR69 (c) and RR*69 (d) with the starting N14 library, only trace amount of shifted band was seen even by using 500 nM repressors.

(A)



(B)

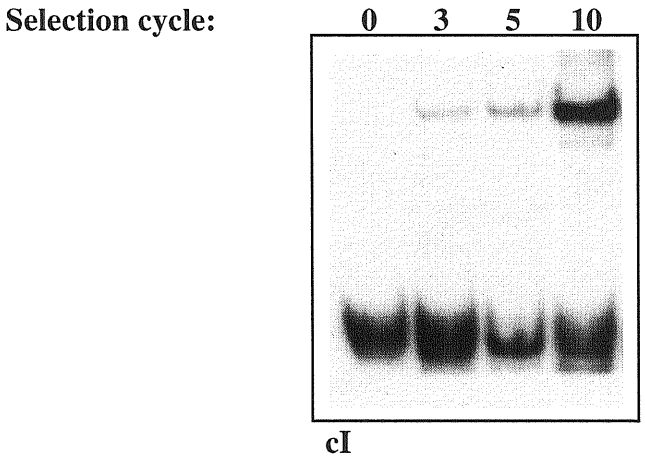


Figure 4.3. Progressive enrichment of the binding sites during selection from the N14 pool for cI. EMSA was performed by using the selected pools as ^{32}P-labeled probes and 10 nM RR69 (A) or 10nM cI (B).

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 434-P22 hybrid operator $O_{R*}1$ (Simoncsits et al., 1997). The results of these selections and experiments with operator analogs of altered spacing (Simoncsits et al., 1997) 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 residues 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 nitrocellulose filtration technique originally applied to the selection of Sp1 binding sites (Thiesen & Bach, 1990). 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 4.2. Similar results were obtained with cI. In this case the selected population showed equally high binding affinity for cI and RR69 (Figure 4.3). The selected operator analogs and their binding affinities for the corresponding protein(s) are listed in Tables 4.1 and 4.3.

Analysis of the selected sequences

Consensus sequences were derived from the sequences selected from the N14 pool (Figure 4.4). 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 4 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.

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	-	-	-	-	15	7
C		-	-	-	-	1	8	-	-	-	-	-	-
G		15	1	-	-	2	-	-	-	19	-	-	1
T		4	4	13	3	14	3	19	19	-	19	3	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	23
cons.:	ACAA	G	A	A	A	T	A	T	T	A	A	A	T	T	T
				T		A	G					T	A	A	A

Figure 4.4. Analysis of the sequences selected from the N14 pool for RR69 (A) (data from Table 4.1(B)), for cI (B) (data from Table 4.1(C)) and for RR*69 (C) (data from Table 4.3(B)). Sequences for a general 434 operator and for a 434 - P22 hybrid operator, together with the numbering scheme used in result analysis, 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 region 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 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 two-fold rotationally symmetric TTGT (Wharton et al., 1984). Substitution of any of these bases, in the context of the 14-mer reference operator (Anderson et al., 1987), was previously shown to reduce the binding affinity by at least 100 fold (Aggarwal et al., 1988; Koudelka & Lam, 1993). Two RR69 selected sequences (a10 and b13 in Table 4.1) 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 *vs* a3 and b13 *vs* b7). The 4A to 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. 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 4.1(A)]. 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 (Wharton et al., 1984, and Table 2.1(A)). Previously we showed that altered length spacers (5 to 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 (Simoncsits et al., 1997). 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 16-mer operator (Hollis et al., 1988), here and in the previous study (Simoncsits et al., 1997) 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 Figure 4.4(C)) is not contacted by RR*69, therefore the spacer in the RR*69 selected DNA ligands is also 6 bp long.

Table 4.1. Sequences selected for RR69 (A and B), cI (C) and their binding affinities

			Kd(nM) RR69	cI
O_R1		CATACAAGAAAGTTTGTTATG	0.8	0.8
(A)				
N8.5		-TTGCATANNNAAGAANNNNNNRTATGAGG-		
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		-TTGCATACAANNNNNNNNNNNNNNNATGAGG-		
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.

As summarized in Figure 4.4, 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 (Aggarwal et al., 1988; Shimon & Harrison, 1993), and systematic changes at this position also showed repressor preference for 5G (Aggarwal et al., 1988). 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 (Table 4.2).

Table 4.2. Higher binding affinities of RR69 for the operators containing 5G or/and 5'C

		Kd(nM)
b2	ACAA**G**ATTC**C**TTGT	0.4
b1	ACAA**G**ATAT**C**TTGT	0.3
b7	ACAA**G**ATATATTGT	0.8
b8	ACAATATAT**C**TTGT	0.8
a1	ACAA**G**AAAC**C**TTGT	0.8
a8	ACAA**G**AAACTTTGT	3.2
b3	ACAA**G**TTAT**C**TTGT	0.4
b12	ACAATATAATTTGT	3.2

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' (for the R recognized half-site) or 9' (for the R* recognized half-site)] 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 (Koudelka et al., 1987). This latter study showed that any change in the inner ATAT sequence of the 14-mer 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 (Travers & Klug, 1990), was established (Drew & McCall, 1990). The sequence dependent effects, on the other hand, are explained on the basis of the different "twistabilities" (Koudelka et al., 1988) or torsional flexibilities (Koudelka et al., 1996; Koudelka & Carlson, 1992) 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 (Aggarwal et al., 1988; Anderson et al., 1987; Rodgers & Harrison, 1993; Shimon & Harrison, 1993) and show different conformational adaptations of the base pairs of these motifs to a common DNA backbone conformation, imposed upon by the repressor binding.

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 (Simoncsits et al., 1997) 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 nonspecific 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 (Aggarwal et al., 1988; Shimon & Harrison, 1993). 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 (Aggarwal et al., 1988; Shimon & Harrison, 1993), 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 (Poteete et al.,

Table 4.3(A) RR*69 N8.5 pool selected sequences and their binding affinities

			Kd(nM)
O_R*1		CATACAATAAAACTTAAATATG	0.8
N8.5		*lac*Z'-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		ACAAACATTGTTAGTT	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-*lac*Z'	

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

Table 4.3(B) RR*69 N14 pool 14 cycle selected sequences and 9 cycle selected strong binder (b*9, b*11, b*27 and b*34) sequences and their binding affinities

			Kd(nM)
O_R*1		CATACAATAAAACTTAAATATG	0.8
N14		-TTGCATACAANNNNNNNNNNNNNNNATGAGG-	
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

Table 4.3(C) RR*69 N14 pool 9 cycle selected sequences and their binding affinities

		Kd(nM)
O_R*1	CATACAATAAAACTTAAATATG	0.8
N14	-TTGCATACAANNNNNNNNNNNNNNNATGAGG-	
b*11	CATACAAGAAAGATTAAAAATATG	0.29
b*27	CATACAAGAAAAGTTAACAGGATG	0.8
b*9	CATACAAGATAATTTAAATTTATG	1.0
b*34	CATACAAATTTACTTTAGTTTATG	1.2
	CATACAAACAAGTTAAGTTAGAT G	<5
	CATACAAATAAAATTAAGAAAATG	10
	CATACAATATTAATTAGATT ATG	7.5-10
	CATACAAATTAAATTAATGACATG	10-15
	CATACAACGTATCTTAAACACATG	15-20
	CATACAAGAATTATTAGTTTAATG	50
	CATACAAGAATAATTACTCCAATG	75
	CATACAAGATTAGTTCTGTAGATG	100
	CATACAATTATAATTTAGAATATG	100-150
	CATACAAATAAAACTAATTAAATG	20-30
	CATACAAAAAAACTATATATTATG	30-40
	CATACAAGATATGTAGATTAAATG	50
	CATACAAATAATCATTAATAAATG	75
	CATACAATACAAATATTATATATG	75
	CATAACATCTAATTAAGTATAATG	100
	CATACAAAAAGTTAAGTATCAATG	150

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

1982, and Table 2.1(B)). It was shown previously that 434 repressor analogs containing the redesigned α3 helix (Wharton & Ptashne, 1985) in the whole repressor (Hollis et al., 1988; Wharton & Ptashne, 1985) and in the single-chain framework (Simoncsits et al., 1997) 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 4.3(A), a*1 - a*18] and P_{lac} - (434-P22) - l*ac*Z' [Table 4.3a, 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 4.3(B) and Figure 4.4(C)). 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 lacking one of these T residues, that were obtained at earlier stage of the selection (Table 4.3(C)), showed strongly reduced (25 to 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 (Table 4.4), or by comparing the binding affinities of those selected sequences which differed only at the 2' or at the 1' position (Table 4.5). The results of the mutational analysis are summarized in Table 4.4. 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

Table 4.4. Affinities of O_R*1 mutants for RR*69

operator	sequence[a]	affinity[b]
	7'6'5'4'3'2'1'	
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 consensus sequence of the P22 operator half-site (CTT. A.T) is underlined.

[b] Relative affinities are given. Value 1 corresponds to an apparent K_d of 8×10^{-10} M.

Table 4.5. Roles of the 2' and 1' residues on the RR*69 binding

		Kd(nM)
	7'6'5'4'3'2'1'	
O_R*1:	CATACAATAAAACTTAAATATG	0.8
1) 2' groups:		
b*1	CATACAAGATATATTAACTAA	0.40
b*2	CATACAAGATATATTAATTTT	0.29
b*7	CATACAAGATAAATTAAATTA	0.30
b*8	CATACAAGATAAATTAATTCT	0.32
b*14	CATACAAGAAATATTAAGTGA	0.45
b*15	CATACAAGAAATATTAATTTG	0.26
b*16	CATACAAGAAATATTAAATG.	0.25
b*28	CATACAAGAAAAATTAATTAC	0.34
b*29	CATACAAGAAAAATTAAATTC	0.20
2) 1' groups:		
b*18	CATACAAGAAATATTAAATT	0.17
b*20	CATACAAGAAATATTAAACTT	0.27
b*21	CATACAAGAAATATTAAAATT	0.25
b*12	CATACAAGAAAGATTAAACAA	0.28
b*11	CATACAAGAAAGATTAAAAAT	0.29

residues at the 7' position. Of the 4' mutants, the 4'T derivative showed only one and a half 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 (Table 2.1(B)). The results of the selection (see Figure 4.4(C)) 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 (Table 4.5) by using various groups of the N14 selected sequences as described above. These and a few other examples taken from Table 4.3b showed that the sequences containing A or T residues at these positions had only slightly (less than twofold) 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.

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, suggest 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 nonspecific 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 (Figure 4.5). 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

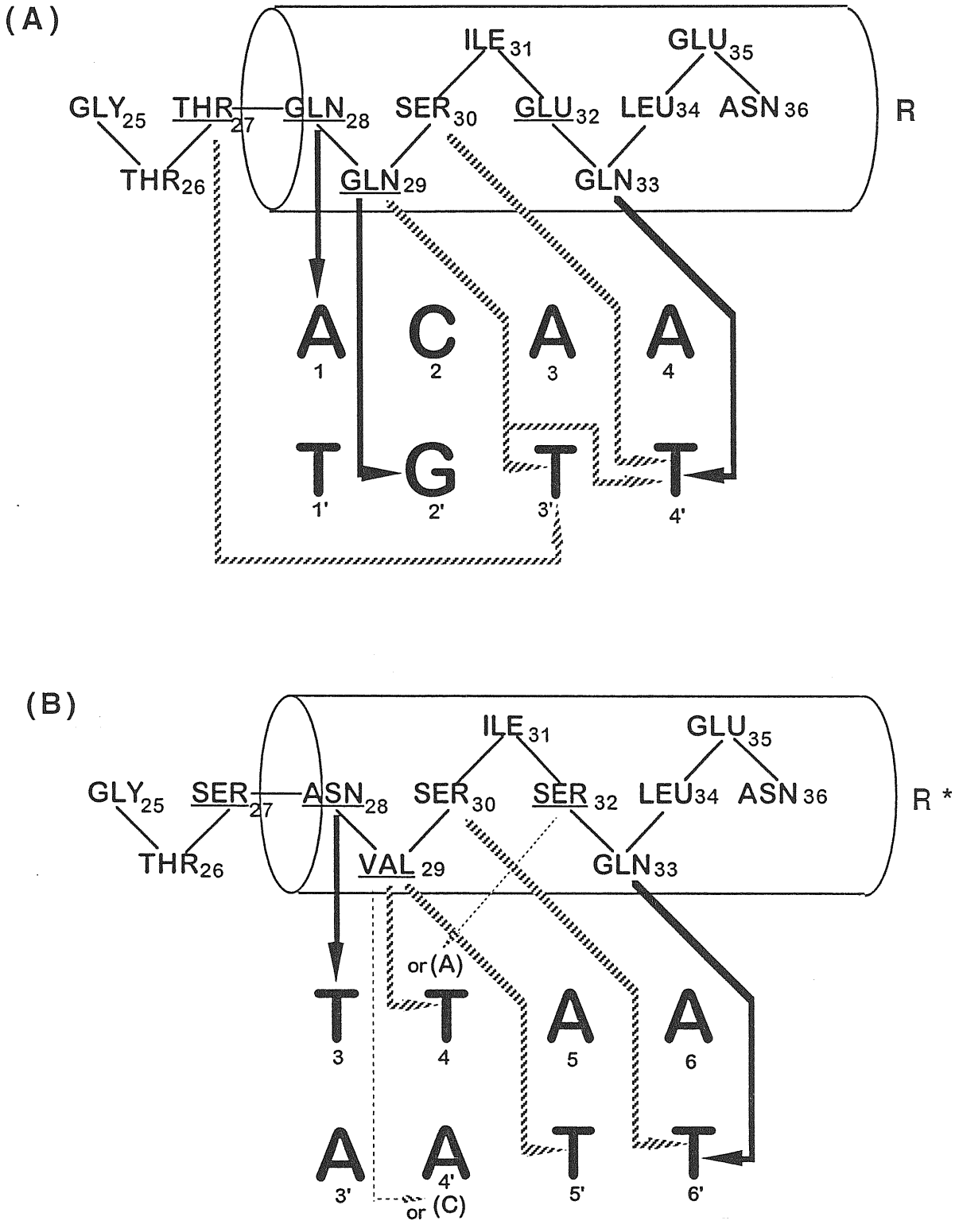


Figure 4.5. Proposed protein-DNA contacts between the α3 helix, of the 434 repressor (A) and the R* (B), and their corresponding operators. Solid lines indicate proposed hydrogen bonds, hatched lines indicate proposed hydrophobic or van der Waals interactions, and discontinuous lines indicate the alternative interactions.

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 (Hilchey *et al.*, 1995; Lehming *et al.*, 1988; Suzuki et al., 1995b) the predictions of this study are in best agreement with the model based on mutational analysis of both the operator and the repressor (Hilchey et al., 1995). 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 (Rodgers & Harrison, 1993), 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 (Figure 4.1) and were tested as described (Simoncsits et al., 1997). *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 (Table 4.6(A)), we observed that generally the repression levels varied between 2 to 4 fold, meanwhile the apparent K_d values varied between 0.8 to 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 (Table 4.6(B)). 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 (Simoncsits et al., 1997).

Table 4.6(A). *In vivo* recognition of the RR69 selected operators

	β-Galactosidase activity[a] observed with operators[b]						
repressor	O_R1	b5	b7	b12	a9	a8	a10
$R_{(-)}$	100	100	100	100	100	100	100
RR69	30	25	30	40	70	25	50

[a] Relative activities are given as described in Table 3.1

[b] Operator sequences are listed in Table 4.1

Table 4.6(B) *In vivo* recognition of the RR*69 selected operators

	β-Galactosidase activity observed with operators[a]				
repressor	O_R*1	O_R*2	a*15	a*21	a*23
$R_{(-)}$	100	100	100	100	100
RR*69	50	30	55	30	60

[a] Operator sequences are listed in Table 4.3a

4.2. Discussion: Recognition properties of the single-chain repressor analogs

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 six 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 noncovalent heterodimer (Hollis et al., 1988) 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 (Berg & Shi, 1996; Choo & Klug, 1994a; Choo et al., 1994; Desjarlais & Berg, 1993; Desjarlais & Berg, 1994; Gogos *et al.*, 1996; Jamieson et al., 1994; Rebar & Pabo, 1994; Rhodes et al.,

1996; Wu et al., 1995) although a prototype with A + T rich binding sequences has also been reported (Gogos et al., 1996). 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 (Simoncsits et al., 1997), it may be possible to isolate single-chain repressor mutants which recognize long DNA targets of more balanced base composition.

5. Single-chain repressor analogs from random protein libraries: Selection and preliminary characterization

Rational design *versus* selection from random protein libraries

The architecture of the single-chain repressor framework and the bidentate DNA-binding mode of this new class of artificial DBPs allows the introduction of independent changes into the individual DBDs. This possibility opens the way to design or select from combinatorial mutant libraries new DBPs with altered specificity. As discussed earlier, many natural DBPs form homodimers, and the non-covalently dimerized units bind to identical or very similar DNA subsites. Therefore the recognition is restricted to palindromic DNA sequences. Successful specificity change (which itself is not trivial and has been achieved only in a very few cases) in the natural, non-covalent arrangement could lead to new homodimers which would recognize new palindromic sequences. A covalently linked heterodimer, if the individual DBDs recognize different DNA subsites, may overcome this limitation and could be used to target asymmetric, non-palindromic DNA sequences. The rationally designed RR*69, as shown previously, is a prototype of such heterodimeric single-chain repressors.

As an alternative to the rational design, construction of combinatorial protein expression libraries and their selection for DNA ligands could provide a more general approach to obtain DBPs with desired specificities. We have tried to exploit this possibility and constructed single-chain repressor libraries in *E. coli* expression vectors. In these libraries, one DBD of the RR69 was kept unchanged meanwhile the other DBD was partially or fully randomized at amino acid positions -1, 1, 2 and 5 of the $\alpha 3$ recognition helix (these correspond to amino acids 27, 28, 29 and 32 of the intact 434 repressor). These are the same residues which were rationally changed in the second domain of RR69 to obtain RR*69 (see Figure 3.1(C)). Such a random library, with a general formula of RR'69 (where the R' stands for the randomized domain as described above), can be targeted for a general operator ACAA - 6bp - NNNN. In this operator, the ACAA sequence is recognized by the unchanged 434 domain R of RR'69, thereby it provides a supporting interaction for the other R' domain to approach and interact with a particular, user-defined sequence in the place of NNNN. Interaction of certain members of the RR'69 library with a given operator can be revealed by using the *in vivo* detection system developed for repressor-operator interaction in this study (Figure 3.1(c)) . Instead of measuring the expression level of the β-galactosidase reporter gene, the expression library is screened by plating onto indicator plates and the expression level is detected by visual inspection of color development in the individual clones (Figure 5.1).

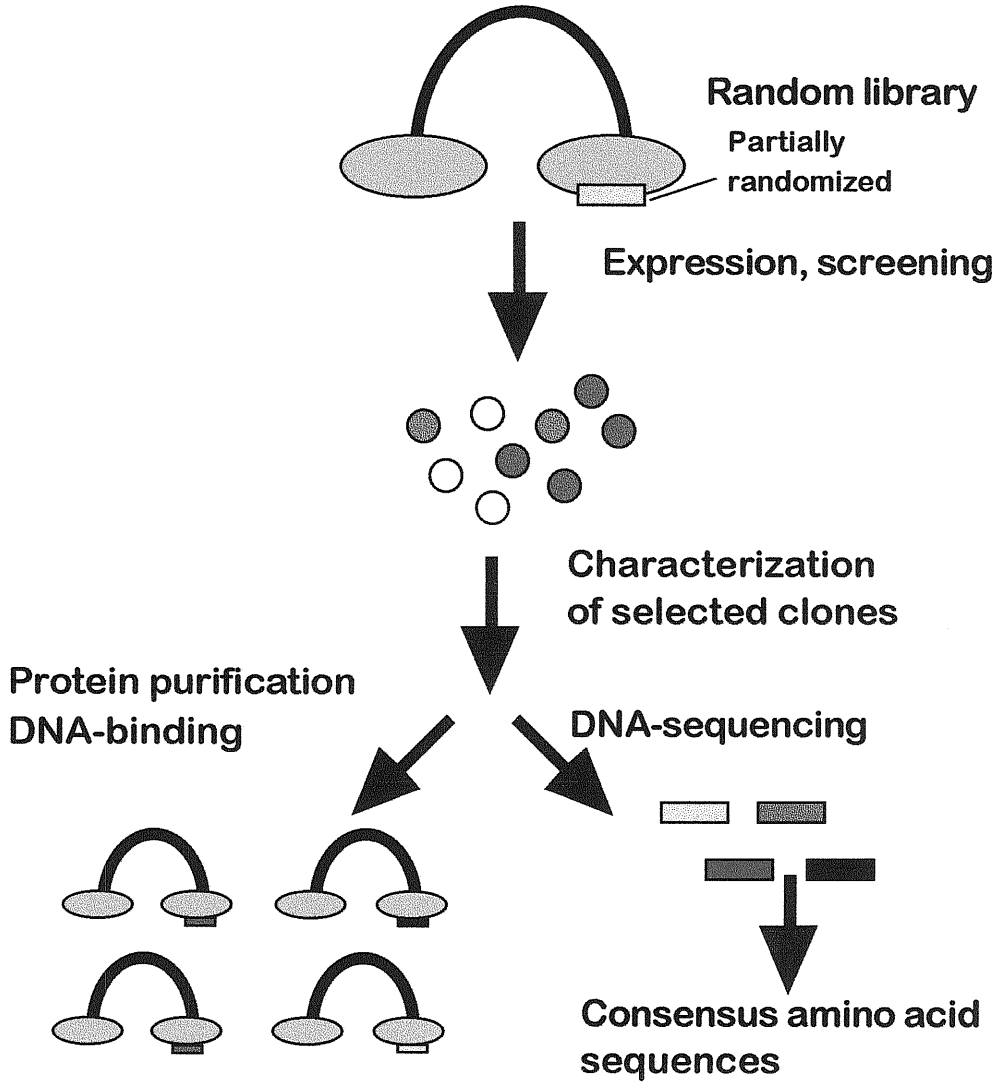


Figure 5.1. Cognate-driven *in vivo* selection of artificial single-chain repressors in *E. coli*

Selection of a single-chain repressor library for the 434-P22 hybrid operator O_R*1

To test the applicability of the combinatorial randomization approach and of the *in vivo* detection/selection system, we have constructed several libraries in the pRIZ' vectors. One of these libraries, pRIZ'O_R*1RR'69 was used to select single-chain repressor mutants for the 434-P22 hybrid operator O_R*1. About 100 clones were selected and nucleotide sequencing revealed that different amino acid combinations could give rise to *in vivo* repression. Somewhat surprisingly, the rationally designed RR*69, which has been shown to bind with high affinity to O_R*1 *in vitro*, was not found among the selected members, but some of its close homologs were isolated. Based on homology studies, a few consensus groups could be defined and some 30 mutant proteins were expressed, purified and partially characterized. Detailed description of the available results of the selection and binding studies with these proteins exceeds the scope and limits of this dissertation. For this reason, only some general properties of these mutants is briefly described here, which prompted more detailed specificity studies to be performed with two members of them. Most mutant proteins were shown to bind to O_R*1 with variable, but generally lower affinity than that observed for RR*69. DNase I footprinting revealed that the isolated proteins bound to the O_R*1 site of long DNA probes. However, many of the mutants showed binding also to the symmetric O_R1 of 434, and this binding was often coupled with dimerization at the O_R1 site of long DNA probes at high protein concentrations. Such aggregation on short oligonucleotide probes, which contained only the operators (either O_R1 or O_R*1) was not observed with any of the isolated mutants. This indicates that the mutant single-chain repressors bind to DNA in a bidentate manner, i.e. the two domains each bind to one half-site of the operators, and this binding mechanism is also prevalent on the long DNA probes at low protein concentrations. The fact that the selection target site TTAA in O_R*1 is closely homologous to the TTGT sequence at the corresponding position of O_R1, which was also recognized by most mutants, could mean that the isolated proteins have somewhat relaxed specificities. The results of the *in vivo* selection, *in vivo* repression assays and *in vitro* binding studies together indicated that a combination of direct and indirect mechanisms determine the DNA-binding properties of the single-chain repressor mutants. To better understand the contribution of these mechanisms to the target recognition, the DNA recognition property of two selected mutants was studied in more detail by DNA binding site selection.

Studies of the DNA-binding properties of two selected single-chain repressor mutants, B31 and B94 by binding site selection from random DNA pool

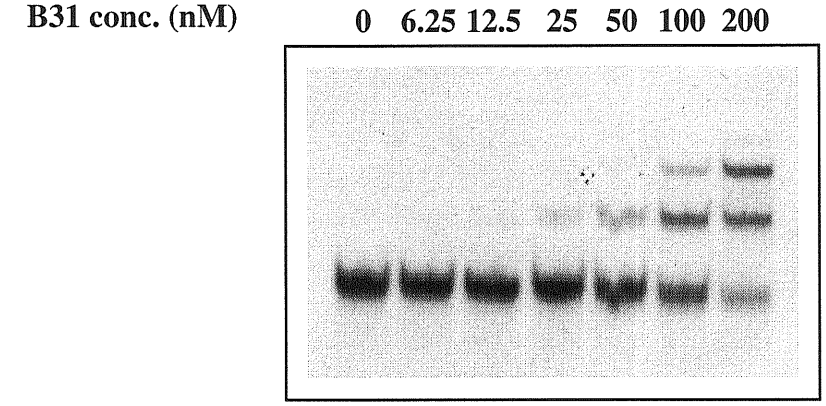
Two *in vivo* selected single-chain repressor mutants were chosen for more detailed specificity studies. They represent extreme examples of the *in vivo* selected analogs. One of the mutants, $RR'_{B31}69$ showed a relatively strong binding to O_R*1 and a high degree of *in vitro* discrimination of O_R*1 over O_R1, but also showed a high tendency to dimerize at the O_R1 site of long DNA probes. The other mutant, $RR'_{B94}69$ bound strongly to the O_R1 site of long DNA probes without apparent aggregation tendency in a broad concentration range, and showed strong discrimination of O_R1 over O_R*1. These mutants are abbreviated as B31 and B94, respectively, in the further discussions. The corresponding nucleotide and amino acid sequences for the α3 helix region are listed as follow:

amino acid positions in α3	-1	1	2			5	
	S/T	X	X	S	I	X	Q
α3 in R'	GGT ACC	WCT	NNS	NNS	AGT ATC	NNS	CAG CTC
		T	R	P		S	
α3 in B31	GGT ACC	ACT	CGC	CCG	AGT ATC	TCG	CAG CTC
		T	A	T		G	
α3 in B94	GGT ACC	ACT	GCC	ACG	AGT ATC	GGC	CAG CTC

Selection of binding sites was performed by using the N9 random library and the nitrocellulose filtration technique. The full sequence of the N9 random pool is shown in the Materials and Methods and the relevant central region for operator selection is shown in Tables 5.1 and 5.2. The central region contains the ACAA box followed by the O_R*1 spacer sequence TAAAA and the 9 bases long, fully randomized region N9. The ACAA box is recognized by the R domain in the single-chain repressors and according to previous selection results the O_R*1 spacer sequence is appropriate for the high affinity binding by the single-chain repressors.

The nitrocellulose filtration technique was used in these selections, but the conditions of the binding, washing and elution steps were modified compared to the previous selection conditions. These technical improvements are detailed in the Materials and Methods section. The binding affinities (the protein concentrations at half-maximal binding) of B31 and B94 were in the range of 100 to 200 nM for the starting

(A)



(B)

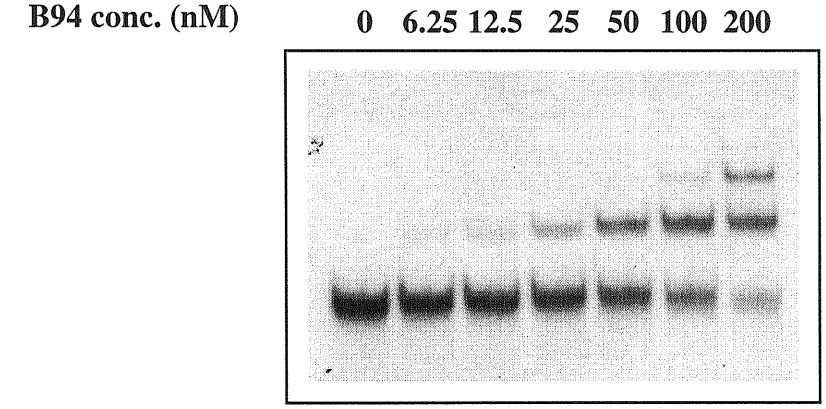
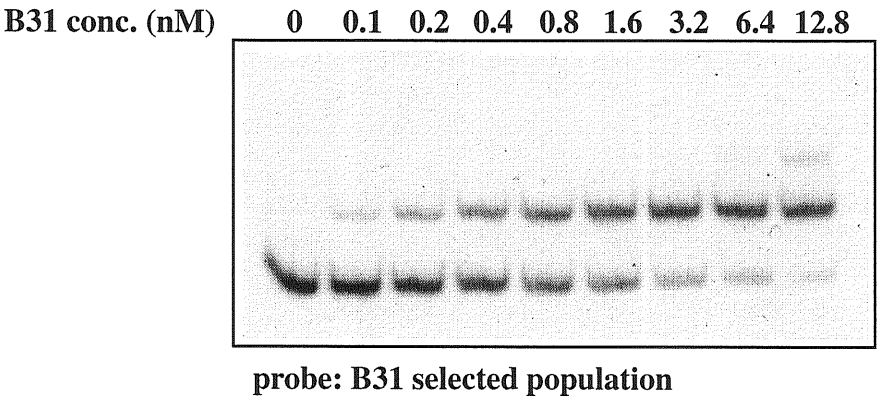


Figure 5.2 Binding affinities of B31 (A) and B94 (B) for the starting N9 library.

(A)



(B)

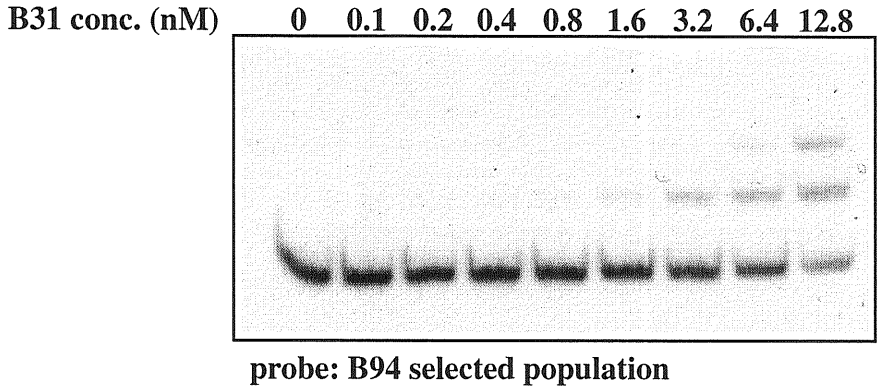


Figure 5.3 The B31 binding affinities to the 0.5 nM B31 (A) and 0.5 nM B94 (B) selected populations.

N9 library (Figure 5.2). After 8 selection cycles, the selected populations showed significantly higher affinities (Kd of 3-5 nM) for the respective proteins in protein titration experiments followed by EMSA. The shifted bands observed at the 0.5 and 2nM protein concentration steps of this analysis were isolated and used for further analysis by cloning, sequencing and binding affinity determination. See Figure 5.3 for the B31 binding to the 0.5 nM protein selected populations. B31 has higher affinity to its own selected population (Kd of 0.8 nM) than to the B94 selected population. Similar result was observed for B94.

The statistical analysis of the sequences selected for B31 and B94 is summarized in Figure 5.4. The lists of the selected sequences for B31 and B94 are shown in Tables 5.1 and 5.2, respectively. These tables also show the binding affinities of certain sequences to the respective protein mutant for which they were selected, and to other proteins described in this thesis. It is to be noted here that the conditions of the binding affinity assay were modified (by reducing the KCl concentration from 200 to 50 mM and by including the non-ionic detergent Triton X100 at 0.02% concentration in the binding buffer) and therefore the binding affinities reported in this section are significantly higher than those shown earlier.

The B31 mutant selected predominantly the TTAA and TTAC sequences in the putative contacted region (from the 4' to the 1' positions, see Table 5.1. for the base numbering scheme). The B31 mutant itself was selected for the TTAA target sequence; thus, the results of the *in vivo* protein selection and the *in vitro* DNA binding site selection are in agreement. The TTAC motif was found more frequently and the tested sequences containing this motif showed high affinity for B31. However, B31 recognizes some other sequences (TAAC, TATC and TATA) equally well or even with higher affinity. Other sequences containing the TT dinucleotide at the 4' and 3' positions were also obtained (a-19: TTAT and a-20: TTTA), but they showed significantly lower affinities to B31 (not shown, but see the similar b-18 and b-2 sequences in Table 5.2). The group of a-21 to a-27 sequences contain the common TA dinucleotide at the 4' and 3' positions and again C or A at the 1' position (TANC and TANA sequences). While in TANC (a-23 to a-25) either A or T can stand for N, and in in TANA (see a-22 versus a-27) only T can stand in place of N to confer high affinity binding to these sequences. Sequences without the 4'T residue were rarely obtained (a-28 to a-30): they do not or are not likely to bind to B31 with high affinities. The 4'T is contacted by the Gln33 of the 434 repressor, and this amino acid was kept unchanged in all repressors of this study (RR69, RR*69, B31 and B94). It seems that the putatively non-contacted 5' residue does not have a significant effect on the interaction of B31 domain with the putatively contacted bases at the 4' to 1' positions. In summary, the B31 domain seems to have a relaxed specificity, but it can bind with high

(A)

	1234	5	6	7	7'6'	5'	4'	3'	2'	1'	-1'	-2'	-3'	-4'
	ACAA	T	A	A	A A	.	.	.	.	.	.	.	.	.
A						9	1	8	27	13	7	11	6	11
C						6	-	-	-	17	5	4	2	3
G						8	2	-	-	-	-	2	3	5
T						9	29	24	5	2	20	15	21	13
cons.:	ACAA	T	A	A	A A	N	T	T	A	C	T	T	T	T
								a	t	A	a	A	a	A
											c			

(B)

	1234	5	6	7	7'6'	5'	4'	3'	2'	1'	-1'	-2'	-3'	-4'
	ACAA	T	A	A	A A	.	.	.	.	.	.	.	.	.
A						3	-	3	1	12	17	16	10	7
C						14	-	1	-	-	1	-	2	1
G						11	-	-	18	-	3	-	-	-
T						-	28	24	9	16	7	12	16	20
cons.:	ACAA	T	A	A	A A	C	T	T	G	T	A	A	T	T
						G		a	T	A	t	T	A	a

Figure 5.4. Analysis of the sequences selected from the N9 pool for B31(A) (data from Table 5.1), and B94 (B) (data from Table 5.2). The numbering scheme used in result analysis and discussion is shown. The consensus bases of the spacer region and of the contacted regions are shown in green and red respectively. Lower case letters in the consensus sequences represent bases which were selected with lower frequency but did not impair the binding in the context of the tested sequences.

Table 5.1. B31 N9 pool selected sequences and their binding affinities

		Kd for different proteins(nM)			
		B31	B94	RR*69	RR69
O_R*1:	ACAATAAAACTTAAATATG	0.5		0.05	
	5'4'3'2'1'				
N9	-ACAATAAAANNNNNNNNNATG-				
a-1#	ACAATAAAAGTTACTATCATG				
a-2#	ACAATAAAAGTTACATTAATG				
a-3	ACAATAAAAGTTACATATATG	0.20	1		
a-4#	ACAATAAAAATTACCATAATG	0.26	10	6.4	>10
a-5 2x	ACAATAAAAATTACATTTATG				
a-6	ACAATAAAAATTACTTTGATG				
a-7#	ACAATAAAATTTACACAAATG	0.20			
a-8	ACAATAAAATTTACTTCTATG				
a-9#	ACAATAAAATTTACTTGTATG				
a-10#	ACAATAAAATTTACTGTTATG				
a-11#	ACAATAAAATTTACTATAATG				
a-12	ACAATAAAATTTAATCTAATG				
a-13	ACAATAAAATTTAATATGATG				
a-14	ACAATAAAATTTAATACGATG				
a-15#	ACAATAAAAATTAACTACATG	0.60	5	<0.5	>10
a-16#	ACAATAAAACTTAACATTATG				
a-17#	ACAATAAAACTTAATTTAATG	0.25			
a-18	ACAATAAAACTTAATCTAATG				
a-19#	ACAATAAAAATTATCATAATG				
a-20	ACAATAAAAGTTTACATAATG				
a-21	ACAATAAAAATATATGTAATG				
a-22	ACAATAAAAGTATATTTTATG	0.52	0.26		
a-23# 2x	ACAATAAAACTATCTTGTATG	0.19			
a-24#	ACAATAAAACTAACATTAATG				
a-25#	ACAATAAAAATAACATATATG	0.17		6.4	
a-26	ACAATAAAAGTAAATATGGTG				
a-27	ACAATAAAAGTAAATTATATG	10			
a-28	ACAATAAAAAGTAAACTTATG				
a-29	ACAATAAAATGTACTATGATG	20			
a-30	ACAATAAAAGATATTAACATG				

0.5 nM B31 selected sequences, the others are the 2 nM B31 selected sequences.

The 4' to 1' position residues are underlined.

Table 5.2. B94 N9 pool selected sequences and their binding affinities

		Kd of different proteins(nM)			
		B94	B31	RR69	RR*69
O_R1:	ACAAGAAAGTTTGTATG	0.15		0.03	
	5'4'3'2'1'				
N9	-ACAATAAAANNNNNNNNNATG-				
b-1#	ACAATAAAACTTTAATTTATG				
b-2#	ACAATAAAAGTTTAATTTATG	0.24	5-10	>10	<0.5
b-3#	ACAATAAAAGTTTAATACATG				
b-4#	ACAATAAAAGTTTAGTATATG				
b-5#	ACAATAAAACTTTAGATAATG				
b-6	ACAATAAAAGTTTATAATATG				
b-7	ACAATAAAAGTTTATTATATG				
b-8#	ACAATAAAAGTTGATAATATG				
b-9#	ACAATAAAAGTTGATATTATG	0.16	10	>10	1
b-10	ACAATAAAAGTTGAGTCTATG				
b-11	ACAATAAAAATTGAATTTATG	0.84	10	10	0.5
b-12	ACAATAAAAATTGTTATAATG				
b-13#	ACAATAAAAGTTGTTTATATG				
b-14	ACAATAAAAGTTGTATTAATG	0.18	>1	<0.5	>1
b-15# 2x	ACAATAAAACTTGTAATAATG				
b-16# 4x	ACAATAAAACTTGTAATTATG	0.2			
b-17# 3x	ACAATAAAACTTGTAAATATG				
b-18	ACAATAAAACTTATATCTATG	0.38	4	>10	1-10
b-19	ACAATAAAACTAGTAAAAATG	0.67			
b-20	ACAATAAAAATATTCATTATG				
b-21	ACAATAAAAGTATATTTAATG				
b-22	ACAATAAAACTCGTATTTATG	1.6			

0.5 nM B94 selected sequences, the others are the 2 nM B94 selected sequences.

The 4' to 1' position residues are underlined.

affinity to only certain tetranucleotides. These are TTAC, TTAA, TAAC, TATC and TATA.

The B94 mutant was previously characterized as a low affinity binder to the TTAA target sequence of the *in vivo* protein selection experiment. At the same time, it showed high affinity binding to the O_R1 operator, which contains the TTGT sequence in the same position. The results of the *in vitro* DNA binding site selection with B94 show that most of the isolated sequences contain the tetranucleotides TTGT, TTGA and TTTA at the 4' to 1' region (Table 5.2.) and that these sequences bind to B94 with high affinities. Other sequences which were not isolated frequently but bound to B94 with reasonable high affinities are TTAT (b-18), TAGT (b-19) and TATA (b-21, data not shown but see the related a-22 sequence in Table 5.1). This latter sequence seems to be recognized by both B31 and B94. Similarly to the selected ligands of B31, the sequences recognized by B94 contain the TT or TA dinucleotide at the 4' and 3' positions, but the dinucleotide preferences of the two mutants at positions 2' and 1' are different. Besides these differences, B94 seems to prefer G or C at the presumably non-contacted 5' position. In addition, both B94 and B31 prefer A and T rich sequences outside of the putative contacted region (positions -1 to -4 of the N9 sequence).

The conclusions from the results of the *in vivo* protein selection experiments and from the analysis of the sequence recognition of the selected protein mutants B31 and B94 can be summarized at this stage of the work as follows. The *in vivo* selection technique is capable of detecting functional protein-DNA interactions at the predetermined DNA target site. Different amino acid combinations in the DNA-contacting positions of the α3 recognition helix of the 434 repressor can recognize the same target sequence of a general operator ACAA - 6bp - NNNN (where NNNN stands for the target region of the selection, in the present case for TTAA). The different selected proteins, however, show different affinities and different degree of specificities for the target DNA. Similar results emerged from the *in vitro* selection of phage display zinc finger libraries (Choo & Klug, 1994b; Choo et al., 1994; Desjarlais & Berg, 1993; Jamieson et al., 1994; Rebar & Pabo, 1994; Wu et al., 1995). The relatively simple recognition mode of the zinc fingers allows easy screening of the selected mutants to reveal their binding specificities or preferences (Choo & Klug, 1994a; Desjarlais & Berg, 1994). In many cases, the selected proteins recognize also other trinucleotide sequences than they were selected for, and sometimes they bind better or more specifically to these trinucleotides than to the original target one. Due to the more complicated DNA recognition by the HTH proteins and to their longer target subsites (4-5 base pairs for HTH and generally 3 bases for zinc finger motif), testing the specificities of the mutant HTH proteins is also more complicated. The DNA binding site selection techniques should be used for HTH proteins, and the starting random DNA pools should also be substantially more diverse than the "minilibraries" applicable

to the zinc fingers. With advantageous modifications of the PCR based selection techniques, a rapid procedure was developed, in a parallel manner, to select DNA binding sites for a large number of mutant single-chain repressors. The results obtained with the B31 and B94 mutants show that, similar to the selected zinc finger proteins, they have relaxed specificities i.e. they recognize a limited set of nucleotide sequences. However, these results also show that the *in vivo* selection approach developed for this new protein class in our laboratory can be used to obtain high affinity DNA-binding proteins. Analysis of the *in vitro* DNA binding site selection results can further help to better understand the interaction of mutant single-chain repressors with DNA and to develop improved selection conditions for such protein libraries. Further studies may prove that the HTH motif containing proteins can also be targeted to used-defined specific, long DNA sequences. The extensive research work with the zinc fingers has already resulted in potentially useful mutant proteins. These were obtained by using designed or selected fingers and the "mix and match" approach (Choo et al., 1994; Desjarlais & Berg, 1993) or by a newly reported "add and optimize" assembly technique which gradually extends a new protein across the DNA target site (Greisman & Pabo, 1997). The single-chain framework of HTH proteins, in principle, allows similar "domain-swap" and gradual "walking to the target site" approaches to be applied in search of new specific proteins for a given target DNA.

6. Conclusions and Perspectives

This work is part of a long term project aimed at developing artificial DNA-binding proteins that can bind to arbitrary DNA targets, for potential medical or biotechnological uses. We used a simplified single-chain repressor architecture that contains two copies of the phage 434 repressor DNA-binding domain which was developed in the previous phases of this project. Our results can be summarized as follows:

1. Rational design was used to change the specificity of an artificial single chain repressor framework based on the phage 434 repressor. Recognition-helix redesign yielded a construct binding to a long non-palindromic cognate (the 434/P22 hybrid).

2. The rationally designed artificial molecules exhibited *in vivo* and *in vitro* DNA-binding activity which was similar or slightly greater than that of the natural phage 434 repressor.

3. *In vitro*-selected cognate sequences correspond to the expected specificities both in the protein-contacted subsites and in the spacer elements. The protein-contacted subsites show a very high sequence conservation. Larger variation occurs in the non-contacted spacer regions, even though only a small set of DNA motifs were selected, suggesting that there are well defined structural constraints.

4. *In vivo* selection of partially randomized repressor molecules against a particular operator gave well-defined consensus amino acid sequences, with strong *in vivo* repression and *in vitro* DNA-binding. Preliminary studies on the selected mutants showed a high affinity but relaxed sequence specificity in the protein contacted regions.

5. The kinetic properties of the single-chain repressors (on rate, off rate) showed a large variation as compared to the natural 434 repressor.

6. The data confirm that a combination of rational, modular design and random selection methods may allow one to develop helix-turn-helix type DNA binding proteins specific for given DNA target sequences.

7. Materials and Methods

7.1 Materials

Enzymes, chemicals and purification kits

Restriction endonucleases and DNA modification enzymes were obtained from New England Biolabs, Inc., Boehringer Mannheim, Pharmacia Biotech, GIBCO BRL (Life Technologies, Inc.), Promega and Perkin Elmer. Radiochemicals were from Amersham. Protein and DNA molecular weight markers were from Pharmacia Biotech. All other chemicals were from Merck, Sigma, Aldrich or Boehringer Mannheim. The plasmid purification kit, QIAquick PCR purification kit, QIAquick nucleotide removal kit, QIAquick gel extraction kit were from QIAGEN.

Oligonucleotides

Oligonucleotides were synthesised by the ICGEB oligonucleotide service or by Primm s.r.l. (Milan, Italy). The following is a list of some oligonucleotides used in this project, the others will be described in the corresponding methods.

PCR primers

AT421a: 5'TCCGGCTCGTATGTTG3'
AT422b: 5'GGTCATAGCTGTTTCCT3'
AT477: 5'AGGCTTTACACTTTATGCTTCC3'
LMB2: 5'GTAAAACGACGGCCAGT3'
AT404: 5'TAGCTCACTCATTAGGCACC3'
AS181: 5'GTAACGCCAGGGTTTTCCCAGT3'

Mutagenesis oligonucleotides

a) For the O_R*1 7' position mutation:
AT495: 5'TACAATAAAANTTAAA3'
AT496: 5'TATTTAANTTTTATTG3'
The two oligos were annealed (in 25 mM NaCl, 65°C for 5 min, then slowly cooled down to room temperature) to form double-stranded DNA, then the double-stranded DNA linker was cloned into the *Nde*I site of the pRIZ'O(-) vector.
b) For the O_R*1 4', 3' position mutation:

AT501: 5'CATACAATAAAACTTBAATATGAGGAAACAG3', (B = C, G, or T)
AT502: 5'CATACAATAAAACTTABATATGAGGAAACA3'
Site directed mutagenesis was performed on pRIZ'$O_{R*}1$.

Bacterial strains and vectors

XL1-Blue (from Stratagene) was used as host during the vector construction steps and in the *in vivo* repressor-operator interaction studies, the genotype is *sup*E44, *hsd*R17, *rec*A1, *end*A1, *gyr*A96, *thi*, *rel*A1, *lac*, F'[*pro*AB, *lac*I^{q}ZΔM15, Tn*10*(*tet*^{r})] (Bullock *et al.*, 1987). BL21(DE3)pLysS expression host (Studier *et al.*, 1990) was obtained from Novagen. CJ236, used to isolate uracil-containing DNA template for site-directed mutagenesis was from Bio-Rad, helper phage VCSM13 was from Stratagene. The pRIZ' (named pRIZ'O_{lac} in this work) key vector for the construction of repressor and operator clones was previously constructed (Simoncsits *et al.*, 1994). The polycloning region of this vector is *Nsi*I (at the -10 box of the *rrn*B P2 promoter)- *Nco*I (containing translational initiation codon)-*Bam*HI-*Sal*I-*Eco*RI-*Cla*I-*Pst*I-*Xba*I-*Bgl*II-*Hin*dIII, the sites used for repressor gene construction and cloning are underlined. The *Nsi*I and *Hin*dIII sites were used for repressor shuffling between different operator clones. M13mp18 (Yanisch-Perron *et al.*, 1985) was from Pharmacia Biotech. Phagemid pKZ152 (Tjörnhammar & Simoncsits, 1991) was used for cloning and mutagenesis of the *lac*I^q-*lac*pro-*lac*Z'(1-146) cluster at the *lac* operator site. The source of this cluster was pMC9 (Miller *et al.*, 1984) isolated from strain Y1089 (Huynh *et al.*, 1984), which was obtained from Stratagene. pRSET expression vectors were donated by R. Schoepfer (Schoepfer, 1993) and pET expression vectors (Studier *et al.*, 1990) were from Novagen. The source of the 434 repressor gene was the λgt10 vector (Boehringer Mannheim), which contains the *imm*^{434} region (Huynh *et al.*, 1984).

Solutions

General Buffers

1) TE buffer: 10 mM Tris-HCl, 1 mM EDTA, pH 8.0
2) 10x TM buffer: 100 mM Tris-HCl, 100 mM $MgCl_2$, pH 7.5
3) 10x PCR buffer: 100 mM Tris-HCl, 15-20 mM $MgCl_2$, 500 mM KCl, 1% Triton, pH 8.3-9.0 (20-25 °C), store at -20°C
4) 10x TBE buffer: 900 mM Tris base, 900 mM boric acid, 20 mM Na_2EDTA, pH 8.3
5) 10x SDS running buffer: 250 mM Tris base, 1.9 M glycine, 1% SDS, pH 8.3
6) SDS gel staining buffer: 150 ml methanol + 150 ml water + 21 ml acetic acid + 9 ml 1% Coomassie Brilliant Blue R250 (in water : methanol (1:1 v/v))

7) SDS gel destaining buffer: 1 litre buffer contains 200 ml methanol and 70 ml acetic acid

Gel-loading buffers

1) Sequencing gel-loading buffer: 0.06% each bromophenol blue and xylene cyanol FF, 10mM EDTA (pH7.5-8.0), and 90% deionized formamide.
2) Agarose gel loading buffer: 20% Ficoll, 0.2% bromophenol blue in water
3) SDS gel-loading buffer: 125 mM Tris-HCl, pH 6.8, 30% (v/v) glycerol, 2% SDS, 6M urea, 1 M β-mercaptoethanol
4) EMSA gel-loading buffer: 30% (v/v) glycerol, 20-50 mM Tris-HCl, pH 7.2, low concentration bromophenol blue

Antibiotic stocks:

1) Ampicillin: 50 mg/ml in sterile water, store at -20 °C.
2) Tetracyclin: 5 mg/ml in EtOH-water (1:1), store at -20 °C.
3) Chloramphenicol: 30 mg/ml in EtOH, store at -20 °C.
4) Kanamycin: 50 mg/ml in sterile water, store at -20 °C.

7.2 Methods

General techniques

The guidelines of published methods: recombinant DNA techniques (Sambrook *et al.*, 1989), site-directed mutagenesis (Kunkel et al., 1987; Kunkel *et al.*, 1991), DNA sequencing (Sanger *et al.*, 1977), β-galactosidase assay (Miller, 1972), EMSA or gel retardation (Carey, 1991), DNase I protection assay (Johnson *et al.*, 1979) and base-specific chemical cleavage of DNA (Maxam & Gilbert, 1980) were followed.

Electrophoretic mobility shift assay (EMSA) and binding affinity determination

Radioactively labeled DNA probes of approximately 160 bp long containing operator sites were generated by PCR on the corresponding pRIZ' operator vectors with primers TAGCTCACTCATTAGGCACC (AT404, located upstream of the *lac* promoter -35 box) and GTAACGCCAGGGTTTTCCCAGT (AS181, backward primer, located in *lac*Z) of which primers one (usually AS181) was ^{32}P end-labeled. Other DNA probes

of approximately 60 or 100 bp long were generated with primers AT421a and AT422b, or AT477 and LMB2, respectively. Usually primer AT422b and LMB2 were radioactively labeled. 15 cycles (94°C, 58°C and 72°C, 30 sec each) were performed and the radioactive PCR products were purified using QIAquick PCR purification kit (Qiagen). The maximal DNA probe concentrations used in EMSA and DNase I protection experiments were calculated on the basis of the molar amounts of PCR primers, assuming 100% incorporation and purification yields.

Binding reactions were performed in 1x binding buffer (200 mM KCl, 2.5 mM $MgCl_2$, 1mM $CaCl_2$, 0.1 mM EDTA, 25 mM Tris-HCl (pH 7.2), 6% (v/v) glycerol) containing 2.5 μg/ml sonicated salmon sperm DNA, 100 μg/ml BSA, <20 pM labeled DNA probe and bromophenol blue at the minimal visible concentration, and repressors in different concentrations. A varied binding buffer with KCl at 50 mM and additional 0.02% Triton, was used for the B31, B94 selected operator binding affinity determination (Table 5.1 and 5.2). For titration experiments, large volumes of binding stocks were prepared (without repressor), and to 45 μl aliquots of this stock, 5 μl single-chain repressor (diluted into 1x binding buffer to obtain 10x higher protein concentrations than required for each step) was added. In case of the natural 434 repressor cI, however, the highest dilution was 50 nM and different volumes were added to the binding stocks to obtain the indicated final concentrations in Figure 3.3(B). The mixtures were incubated at room temperature for 1 hr, then identical volumes were loaded onto running 8% polyacrylamide gel (29:1 acrylamide:bisacrylamide), prerun at 4°C in 0.5x TBE buffer, and the electrophoresis was performed with 25V/cm at 4°C for 2 hr. The gels were fixed in 10% acetic acid and dried before autoradiography.

Apparent K_d values were obtained by determining the protein concentrations at half-maximal binding in protein titration experiments as described (Simoncsits et al., 1997). 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.

Competition experiments

Binding reactions containing 5 nM repressors were performed as above. After 1 hr, aliquots were taken out and double-stranded oligonucleotide competitors containing O_R1, $O_{R*}1$ or O_P1 operator sequences were added to the final concentration of 100 nM. The reaction mixtures were incubated at room temperature for 30 min before analysis on gels as above. Competitor oligonucleotide duplexes were prepared from the listed cohesive end duplexes (see operator cloning section) by Klenow polymerase fill-in reaction followed by purification on nondenaturing polyacrylamide gel.

The k_{off} determination

Kinetic analysis of the dissociation of repressors from operators was carried out as described (Brown & Sauer, 1993; Robinson & Sauer, 1996). Binding reaction was performed as described above in 50 mM KCl. When the reaction reached the equilibrium state, dissociation was initiated by a 20 fold dilution with 1x binding buffer containing 2.5 μg/ml sonicated salmon sperm DNA, 100 μg/ml BSA. At appropriate time points samples were loaded on 8% acrylamide gel (19:1) for electrophoresis at 25 V/cm.

In the initial binding reaction, more than 90% DNA was bound, and the final concentration of the repressors after dilution results in less than 5% bound DNA in an equilibrium experiment.

The dissociation rate constant (k_{off}) can be calculated as:

$-k_{off}t = \ln [RO]/[RO]_0 = \ln \theta/\theta_0$ (Kim *et al.*, 1987; Brown & Sauer, 1993)

$k_{off} = 0.693 / t_{1/2}$

Where θ represents the fraction of the bound operator to the total operator, and $t_{1/2}$ is the half life time of the dissociation.

The length of DNA fragment is important for the k_{off} determination. As Kim *et al* reported (Kim *et al.*, 1987), Kd is essentially constant and independent of DNA length, while k_{on} and k_{off} increase as the DNA length increases.

DNase I protection assay

Binding reactions containing 100 or 200 nM repressors were performed as above in 100 μl volumes. 10 ng DNase I was added and the mixtures were incubated at room temperature for 5 min. 100 μl cold stop solution (4 M ammonium acetate, 40 mM EDTA containing 200 μg/ml glycogen) was added and further steps were as described (Johnson *et al.*, 1979) using 6% acrylamide-8M urea gel. A (A+G) base-specific chemical cleavage of DNA reaction described originally by Maxam & Gilbert (Maxam & Gilbert, 1980) was employed as marker.

Competent cell preparation

Buffers

RF1: 100 mM RbCl, 50 mM $MnCl_2$-$4H_2O$, 30 mM potassium acetate, 10 mM $CaCl_2$-$2H_2O$, 15 (w/v) glycerol, pH 5.8 (adjusted with 0.2 M acetic acid)

RF2: 10 mM MOPS, 10 mM RbCl, 75 mM $CaCl_2$-$2H_2O$, 15% (w/v) glycerol, pH6.8 (adjusted with NaOH).

Sterilize the buffers by filtering through 0.22 μM membrane, store at 4 °C.

The competent cells were prepared as described by (Hanahan, 1985). *E. coli* cells were grown in 2 ml SOB medium with 37°C overnight shaking. The overnight culture was 200 fold diluted in SOB medium, and incubated with 37°C shaking to a cell density of about 4-7 x 10^7 cfu/ml. The culture was centrifuged at 750-1000g for 12-15 min at 4 °C. The supernant was drained, and the cell pellet was resuspended with 1/3 of the original culture volume of RF1 buffer and incubated on ice for 30 min. The cells were pelleted and resuspended with 1/12.5 of the original culture volume of RF2 buffer, incubated on ice for 15 min and distributed as aliquots into chilled 1.5 ml eppendorf tubes. The competent cells were kept frozen at -80 °C.

Site-directed mutagenesis

Site-directed mutagenesis was performed as described by Kunkel *et al* (Kunkel et al., 1987; Kunkel et al., 1991) with several modifications.

Uracil-containing DNA template preparation

CJ236 *E. coli* cells, transformed with pRIZ' vectors, were grown in LB medium containing 50 mg/l ampicillin, 30 mg/l chloramphenicol at 37°C with shaking. The overnight culture was 100 fold diluted into 10 ml LB medium containing antibiotics as described above and incubated with shaking at 37°C. At a cell density of approximately 10^8 cfu/ml (A_{600} of 0.5-0.7), the VCSM13 helper phage was added to obtain a M.O.I of 10-20. The cell culture was left to stand at room temperature for 10 min, further incubated with 37 °C shaking for 1 hr, then kanamycin (for selecting of cells infected with helper phage) was added to a final concentration of 70 mg/l, and the incubation is continued for 6-24 hr. Cell culture was centrifuged at 5,000g at 4°C for 30 min. Add 3 ml 20% PEG8000-2.5 M NaCl to the 10 ml clear supernant which contains the phagemid particles, mixed well, and put on ice for 1 hr. The precipitate was collected by centrifuging at 5,000g at 4 °C for 30 min, and the pellet was resuspended in 400 μl TE buffer, and extracted with 1 volume phenol. The uracil-containing single-stranded DNA was obtained by ethanol precipitation and could be further purified with Qiagen plasmid purification tips.

Mutagenic primer preparation

The mutagenic primers were phosphorylated. A 10 µl reaction mixture, containing 1x T4 kinase buffer, 0.03 mM ATP, 30 pmol oligonucleotides, 5 units of T4 polynucleotide kinase (Biolabs), was incubated at 37°C for 45-60 min, followed by heat inactivation of the kinase at 65°C for 10 min.

Mutagenesis reaction

1) Buffers:

10x Annealing buffer: 200 mM Tris-HCl (pH7.4), 20 mM $MgCl_2$, 500 mM NaCl, store at -20 °C.

10x synthesis buffer: 100 mM Tris-HCl (pH7.4), 50 mM $MgCl_2$, 20 mM DTT, 5 mM dNTPs, and 10 mM ATP, store at -20 °C.

2) Annealing of the primer to the template:

0.1 pmol uracil containing DNA was annealed with 3-5 pmol mutagenic primer in 10 µl 1x annealing buffer. The reaction mixture was placed at 70°C for 3-5 min, followed by cooling down at a rate of approximately 1°C/min to the calculated Tm of the oligonucleotide (Tm = [(2 x AT nucleotide numbers) + (4 x CG nucleotide numbers) - 3°C]), then put on ice.

3) Synthesis of the complementary DNA strand:

Add 1 µl 10x synthesis buffer, 1-2 unit T4 DNA ligase (GIBCO BRL, or Boehringer), 1 unit diluted T7 DNA polymerase (Phamacia) to the above annealed sample in the order as listed. The reaction mixture was incubated on ice for 5 min, at room temperature for 5 min, and finally at 37°C for 90 min. 2 µl 100 mM TE buffer (100 mM Tris-HCl, 100 mM EDTA, pH8.0) was added to stop the reaction.

Transformation

The mutagenesis reaction samples were transformed to XL1-Blue *E. coli* cells.

Plasmid Preparation

The alkaline lysis method (Sambrook et al., 1989) was employed for plasmid preparation.

Buffers

RNase A Stock: 11 mg/ml in 10 mM Tris·HCl (pH 7.5), 15 mM NaCl, incubated in boiling water for 15 minutes, store at -20 °C

Solution 1: 100 μg/ml RNase A, 50 mM Tris-HCl, 10 mM EDTA, pH8.0, store at 4 °C

Solution 2: 200 mM NaOH, 1% SDS, store at room temperature.

Solution 3: 3.0 M potassium acetate, pH 5.5, store at 4 °C

3 ml culture scale

Bacteria grown up in 3 ml overnight culture were pelleted, and resuspended in 0.3 ml solution 1 by vortexing; 0.3 ml solution 2 was added, mixed gently and incubated at room temperature for 5 min; then 0.3 ml chilled solution 3 was added , mixed gently, and incubated on ice for 10 min. The mixture was centrifuged at 14,000rpm for 15-20 min. The plasmid DNA could be isolated by two methods:

1) Load the supernant to Qiagen plasmid purification tip5 (Qiagen), followed by washing, elution and precipitation as described by Qiagen plasmid purification kit manual.

2) Precipitate the DNA with 0.7 volume isopropanol. The pellet was resuspended in 90 μl TE buffer and 10 μl solution 1, incubated at room temperature for 30 min for RNA degradation. Then 0.5 μl sample was loaded on 0.6% agarose gel for electrophoresis to check the purity and amount. The plasmid isolated by this step can be directly used for sequencing (see bellow), or further purified by phenol extraction and ethanol precipitation in the presence of 0.3M sodium acetate (pH5.2).

10 ml culture scale

10 ml overnight bacteria culture were pelleted, resuspended and incubated in 0.6 ml each solution 1, 2, 3 as described above. The plasmid DNA was purified by loading the supernant on Qiagen plasmid purification tip20 (Qiagen), followed by washing, elution and precipitation as described by Qiagen plasmid purification kit manual.

Sequencing

The Sanger dideoxynucleotide sequencing method (Sanger et al., 1977) was used, and sequencing was performed as described by the T7sequencing™ kit manual (Pharmacia).

Buffers

Annealing buffer: 1 M Tris-HCl (pH7.6), 100 mM $MgCl_2$, 160 mM DTT

Mix-dATP: 1.375 µM each dCTP, dGTP and dTTP, 333.5 mM NaCl

"A" mix: 840 µM each dCTP, dGTP, and dTTP, 93.5 µM dATP, 14 µM ddATP, 40 mM Tris-HCl (pH7.6), and 50 mM NaCl

"C" mix: 840 µM each dATP, dGTP, and dTTP, 93.5 µM dCTP, 17 µM ddCTP, 40 mM Tris-HCl (pH7.6), and 50 mM NaCl

"G" mix: 840 µM each dCTP, dCTP, and dTTP, 93.5 µM dGTP, 14 µM ddGTP, 40 mM Tris-HCl (pH7.6), and 50 mM NaCl

"T" mix: 840 µM each dCTP, dGTP, and dATP, 93.5 µM dTTP, 14 µM ddTTP, 40 mM Tris-HCl (pH7.6), and 50 mM NaCl

24 µl plasmid temperlate (prepared from 3 ml culture, after the RNA degradation by TE and solution 1 incubation step, see above) was mixed with 6 µl 2M NaOH and incubated at room temperature for 10 min. The denatured DNA was neutralized with 9 µl 3 M sodium acetate (pH5.2), and precipitated with 2.5-3 volume ethanol. 2 µl annealing buffer and 2 µl 2-3 µM sequencing primer were added to 10 µl water resuspended denatured DNA template. The annealing mixture was incubated at 65°C for 5 min, 37°C for 15 min and room temperature for more than 10 min. An "Enzyme Premix" was prepared by mixing 1xn µl water, 3xn µl labelling Mix-dATP, 2xn µl 1.5 unit/µl diluted T7 DNA polymerase (Phamacia) and 1xn µl α-^{35}S-ATP (10 µCi/µl), n = number of templates. 6 µl "Enzyme Premix" was added to the annealed mixture, incubated at room temperature for 5 min, then 4.5 µl reaction mixture was mixed with 2.5 µl each of the sequencing "A" mix, "C" mix, "G" mix and "T" mix pre-warmed in microtiter plate, and incubated at 37°C for 5 min. 5µl sequencing gel-loading buffer was added to stop the reaction, and samples were heated at 80°C for 2 min before loading on 8 M urea-6% polyacrylamide gel for electrophoresis at 38 W (approximately 1500-1800 V).

Vector constructions for *in vivo* repressor-operator interaction studies: operator cloning and construction of repressor genes in pRIZ' vectors.

Operator cloning

First, the *lac* operator region of pRIZ'O_{lac} was replaced by *Nde*I site to obtain pRIZ'O(-), which was then used to clone synthetic oligonucleotides containing operator sequences. The pRIZ'O_{lac} phagemid contains two *lac* operator sites, therefore it could not be converted to pRIZ'O(-) by site-directed mutagenesis. pRIZ'O(-) was therefore obtained as follows. The 1.72 kbp *lac*I^q-*lac*pro-*lac*Z'(1-146) region of pMC9 (Miller *et*

al., 1984) was cloned as *Eco*RI fragment into pKZ152 phagemid, which does not contain *lac* regions (Tjörnhammar & Simoncsits, 1991), then the *lac* operator region was replaced by *Nde*I site using the mutagenic oligonucleotide TTCCGGCTCGTATGTTGCATATGAGGAAACAGCTATGACCAT (*Nde*I site is underlined). The altered cluster was then cloned as blunt-end fragment (obtained after *Eco*RI cleavage and Klenow polymerase + dNTP treatment) into the precursor of the pRIZ' vector as described (Simoncsits *et al.*, 1994) to obtain pRIZ'O(-), in which the gene cluster was in the same orientation as in pRIZ'O_{lac} (Figure 1c). The following synthetic operators with *Nde*I compatible cohesive ends were cloned into the unique NdeI site of pRIZ'O(-):

TACAAGAAAGT**TTGT** **GTT**CTTTCA**AACA**AT	O_R1 (of 434)
TACAATAAAA**CTTAAA** **GTT**ATTTT**GAATTTAT**	O_R*1 (434-P22) or O_R*2 (P22-434), in opposite orientation
TATTAAAGAACA**CTTAAA** **AATTTCTTGTGAATTTAT**	O_P1 (O_R1 of P22)

where bold letters show the respective consensus operator boxes. The vectors obtained are abbreviated as pRIZ'O_R1, pRIZ'O_R*1, pRIZ'O_R*2 and pRIZ'O_P1. Altered spacer operator mutants were cloned similarly, the spacer sequences of the O_R1 mutants (shown for upper strand) were GAAAG (O_R1, 5 bp), GAAAGAT (O_R1, 7 bp), GAAAGTAT (O_R1, 8 bp) and GAAAGTATAT (O_R1, 10 bp), the spacer sequence of the O_R*1 mutant was TAAAAT (O_R*1, 6 bp). The operator clones were verified by sequencing; in most cases both possible orientations with respect to the *lac* promoter were obtained. The orientations shown in Figure 1d or those corresponding to the above listed mutant spacer sequences were used in this study.

Construction of repressor genes in pRIZ' vectors

In the initial stages of this work $R_{1-69}L_{70-89}R_{1-90}$ (or RR90) and R_{1-90} (or R90) were constructed. These elongated proteins showed similar DNA-binding properties to that of the shorter counterparts but they showed enhanced proteolytic sensitivity during expression and purification and were not used in this study. Their genes, however, were used as intermediers during the construction of the shortened versions as described below.

For gene construction and cloning, regions of the cI gene of λgt10 vector (Huynh *et al.*, 1984) were amplified by PCR, using primers designed according to the published nucleotide sequence (Nikolnikov *et al.*, 1984) PCR products, when blunt-end cloning was used, were treated with Klenow polymerase and dNTP. Primer

sequences are written in 5'-3' direction. The R90 region was amplified by using two different pairs of primers. To obtain the first copy of the tandem repeats, primers TCCTTTCATGAGTATTTCTTCCAGGGT (AT405, *Rca*I site underlined) and TCAGGATCCAGCTCTAACCATGCTAAT (AT406, *Bam*HI site underlined, TCA stop codon complementer in bold) were used and the PCR product was cloned as *Rca*I - blunt-end fragment into *Nco*I - blunt-end(*Bam*HI) pRIZ'O_R1 to obtain pRIZ'O_R1R90. The second copy was obtained by PCR using primers TACTTGGATCCATTTCTTCCAGGGTAAAAAGC (AT407, *Bam*HI site underlined) and CTGCTCAAGCTTCACGAACCAGCTCTAACCAT (AT408, *Hin*dIII site underlined, stop codon complementer in bold), which was cloned as *Bam*HI - *Hin*dIII fragment into M13mp18. Site-directed mutagenesis of the mp18 clone was performed with primers CTTAGTTTTACCGTTCTCGAGCTGCTCT (AT409, *Xho*I site underlined, mismatched base in bold) and TAGACTGCTGGGTGGTACCCACCTT TTGAG (AT410, *Kpn*I site is underlined, mismatched bases in bold) in one step as these two primers could anneal contiguously to the template. This second, mutant copy was then cloned as *Bam*HI - *Hin*dIII fragment into pRIZ'O_R1R90 to obtain pRIZ'O_R1RR90, which contains *Kpn*I and *Xho*I sites in the second copy, near the borders of the α3 helix coding region (at the *Bam*HI fusion site between the repeats, Ser90 of the first R90 coincides with Ser1 of the second R90, therefore the RR90 abbreviation stands for $R_{1\text{-}89}R_{1\text{-}90}$). The R69 coding region of the mutated mp18 clone was PCR amplified with primers AT407 and TCATCTAACATTCGAATCAGAGGT (AT414, stop codon complementer in bold), the PCR product was cleaved with *Bam*HI and cloned into *Bam*HI - blunt-end (*Hin*dIII) pRIZ'O_R1R90 vector to obtain pRIZ'O_R1RR69, in which the *Hin*dIII site downstream of the stop codon is regenerated. The α3 helix coding region in the second repeats of pRIZ'O_R1RR90 and pRIZ'O_R1RR69 was replaced with a *Kpn*I - *Xho*I linker, (coding for amino acid substitutions as in Figure 3.1(b))

```
      CTCTAACGTCAGTATCTCACAGC
CATGGAGATTGCAGTCATAGAGTGTCGAGCT
```

to obtain pRIZ'O_R1RR*90 and pRIZ'O_R1RR*69, respectively. To obtain the R*R*69 gene, the second repeat of RR*90 was used as follows. Complete *Eco*RI cleavage (*Eco*RI site is present in both repeats at amino acids 10-12, see Figure 3.1(a) of pRIZ'O_R1RR*90 followed by religation eliminated the first repeat and resulted in pRIZ'O_R1R*90. This vector was used as PCR template with primers TGTAGCGGGAAGGCGTATTAT (AS107, vector specific primer at the *rrnB* P2 promoter overlapping the unique *Nsi*I site of pRIZ') and AT406, then the *Nsi*I - *Bam*HI cleaved PCR product was cloned into pRIZ'O_R1RR*69, replacing the first domain coding region of RR*69 with R* and providing pRIZ'O_R1R*R*69. The

pRIZ'O_R1R69 vector was obtained from pRIZ'O_R1RR69 after complete *Eco*RI cleavage and religation. pRIZ'O_R1R(-) was also obtained from *Eco*RI cleaved pRIZ'O_R1RR69, but Klenow polymerase treatment in the presence of dNTP preceded the religation. The full 434 repressor coding region was obtained from λgt10 by PCR performed with primers AT405 and TCTCTGGATCCTCATACGAATTTTACCCT CGCT (AT453, *Bam*HI site underlined, stop codon complementer in bold), cleaved with *Rca*I and *Bam*HI and cloned into *Nco*I - *Bam*HI cleaved pRIZ' to obtain pRIZ'O_R1cI. All repressor coding regions obtained by PCR were verified by nucleotide sequencing, using AS107 and GGCAGTTTCCCAGACATTACTC (AT419, backward vector specific primer located downstream of *Hin*dIII site) flanking primers and sometimes internal primers. The desired repressor-operator combinations in the same pRIZ' vector were obtained by subcloning the repressor coding regions into other operator vector (shuffling) by using the unique restriction site of the pRIZ' vectors (see Figure 1(c), usually *Nsi*I - *Hin*dIII or sometimes *Pst*I - *Hin*dIII sites).

In vivo detection of repressor-operator interactions: β-galactosidase assay

XL1-Blue *E. coli* cells containing pRIZ' vectors were grown in LB medium containing 100 mg/l ampicillin and 10 mg/l tetracycline at 37°C with shaking. Overnight cultures were diluted 50 fold into fresh medium and grown for 2 hr (A_{600} is approx. 0.3-0.4). IPTG was then added to a final concentration of 1 mM and the cultures were grown for 2.5 hr. β-Galactosidase assay was performed as described (Miller, 1972).

Construction of T7 promoter-based expression vectors

The pRSETRR69 vector was described previously (Percipalle et al., 1995). Large scale expression proved to be difficult with this vector, but even more serious problem was encountered in case of the pRSETRR*69 vector, which did not provide productive colonies upon transformation into BL21(DE3)pLysS or pLysE expression strains. Therefore we constructed an improved vector, called pSET5a from pRSET5a (Schoepfer, 1993) and pET16b (Novagen) vectors by cloning the *Sca*I - *Xba*I fragment of pRSET5a into *Sca*I - *Xba*I cleaved pET16b. The new pSET5a vector combines the advantageous properties of the parents pET16b (lower copy number and tightly controlled T7*lac* promoter) and pRSET5a (versatile multiple cloning site and phagemid properties). The RR69 region of pRSETRR69 was cloned as *Xba*I - *Hin*dIII fragment into pSET5a to obtain pSETRR69. pSETRR*69 was obtained by replacing R69 of pSETRR69 with R*69 in a *Bam*HI - *Hin*dIII cloning. To obtain pSETR*R*69, similar intermediate clones were used as described for the corresponding pRIZ' construct.

Briefly, the R*90 coding region from pRIZ'O_R1RR*90 replaced the R69 part of pSETRR69 by *Bam*HI - *Hin*dIII cloning to provide pSETRR*90, which was converted to pSETR*90 by complete *Eco*RI cleavage followed by religation. PCR was then performed on pSETR*90 template using the T7 promoter primer (Novagen) and AT406, then the *Xba*I - *Bam*HI cleaved PCR product was cloned into pSETRR*69 to obtain pSETR*R*69. The pSETR69 vector was obtained from pSETRR69 after complete EcoRI cleavage and religation. The full-length 434 repressor gene in pET vector was obtained by PCR performed on λgt10 vector with primers AT405 and AT453, followed by cloning the *Rca*I - *Bam*HI cleaved PCR product into *Nco*I - *Bam*HI cleaved pET16b.

Expression and purification of single-chain repressors

pSETRR69 was freshly transformed into BL21(DE3)pLysS strain to obtain about one to two thousands small colonies on LB plates containing 75 mg/l ampicillin and 25 mg/l chloramphenicol after 12-14 hr incubation at 37 °C. The colonies were suspended and grown in 3.6 l of LB medium containing antibiotics as above. Induction with IPTG was performed as described (Studier *et al.*, 1990). After 2 hr induction, cells were harvested by centrifugation, resuspended in 120 ml of TE buffer (10 mM Tris-HCl, 2 mM EDTA, pH 8.0), frozen at -80 °C and thawed. The suspension was sonicated briefly to reduce the viscosity, centrifuged and batch adsorption was performed on the supernatant by adding 25 ml 50 % suspension of SP-Sepharose in TE followed by gentle shaking of the suspension for 20 min. Short column was then prepared which was washed with 150 mM KCl in TE followed by 350 mM KCl in TE to elute the highly enriched RR69. The eluate was diluted threefold with TE and purified (in two portions) on Mono S HR 10/10 column (Pharmacia Biotech) using linear gradient of KCl in TE. The isolated yield was 40 mg per liter culture. RR*69 and R*R*69 were similarly purified using Resource S column in 20 and 6 mg per liter culture yields, respectively. In these latter cases the expression levels were lower and a fraction of the repressors was found to be insoluble. Only the soluble fractions were used in the further purification steps. The single-chain analogs B31 and B94 were similarly expressed on 200 ml scale and purified by using Resource S column.

The 434 repressor and its DNA-binding domain R69 were also purified as above with minor modifications.

***In vitro* binding site selection and cloning of the selected sequences**

The random DNA pools used in this project include:

a) AT420, the N8.5 pool:

5'TCCGGCTCGTATGTTGCATAN$_3$AAGAAN$_5$RTATGAGGAAACAGCTATGACC3'

N = A, C, G or T; R = A or G

b) AT460, the N16 pool:

5'TCCGGCTCGTATGTTGCATN$_{16}$ATGAGGAAACAGCTATGACC3'

c) AT490, the N14 pool:

5'TCCGGCTCGTATGTTGCATACAAN$_{14}$ATGAGGAAACAGCTATGACC3'

d) AT500, the N9 pool:

5'TCCGGCTCGTATGTTGCATACAATAAAAN$_9$ATGAGGAAACAGCTATGACC3'

Bold sequences are interrupted boxes for the *Nde*I site, and PCR primers (AT421a, AT422b) corresponding sequences are underlined. The underlined and bolded sequences are identical to sequences flanking the unique *Nde*I site of the pRIZ'O(-) vectors and serve to clone the selected sequences by loop insertion mutagenesis (Figure 4.1). These were purified by electrophoresis using acrylamide - 8M urea gels. The randomized oligonucleotides were made double-stranded by primer extension using AT422b, Klenow polymerase and dNTP and the double-stranded fragments were purified by electrophoresis on 10% polyacrylamide gel (19:1 acrylamide/bisacrylamide). The double-stranded DNA library can be illustrated as:

→ primer AT421a

5 ' TCCGGCTCGTATGTTGCATNN • • • NNNATGAGGAAACAGCTATGACC3 '

3 ' AGGCCGAGCATACAACGTANN • • • NNNTACTCCTTTGTCGATACTGG5 '

primer AT422b ←

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 (Blackwell,1995; Blackwell *et al.,* 1990). 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: 94 °C, 58 °C and 72 °C, 30 sec 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 (Thiesen & Bach, 1990). 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 repressor proteins and 0.5 - 1 pmol N14 double-stranded DNA. 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 2 µg/ml poly(dI-dC) was added

to the binding reactions 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 to reduce multiple shifted bands in the PCR products. 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) (Figures 4.2 and 4.3). The DNA probes used in these assays were obtained by PCR labeling of the selected pool with ^{32}P end-labeled AT422b and unlabeled AT421a primers.

Selection from the N9 random DNA with B31 and B94 was performed by using the nitrocellulose filtration technique. Several modifications were introduced into the method: 1) the binding reaction was carried out in binding buffer B with 50 mM KCl, and the binding volume was kept constant (200 µl) during the whole selection; 2) 2 µg non-specific competitor DNA poly(dI-dC) was added to the binding reaction mixture from the first selection cycle; 3) instead of 1x binding buffer, membrane washing was simplified by water filtration, as other experiment indicated that water has the same efficiency as the binding buffer on washing; 4) other experiment also showed that 0.1 N NaOH incubation is not so efficient to elute the bound DNA, while 1x PCR buffer can efficiently elute the bound DNA (probably together with protein) from the membrane. This is probably due to the 0.1% Triton in the PCR buffer. DNA eluted in this way can be used directly for PCR amplification, thus eliminating the precipitation step used previously in the NaOH elution; 5) the amplified DNA was precipitated after PCR and used directly for the next cycle selection without any other purification steps, because the coprecipitated primers didn't influence with the binding reaction and can be removed after filtration and washing.

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 (McCabe, 1990) using AT421a and AT422b primers in a molar ratio of 50:1. Following gel purification and 5'-phosphorylation, mutagenesis was performed on uracil-containing single-stranded DNA templates (Kunkel et al., 1991) of pRIZ'O(-) vectors (Simoncsits et al., 1997). To minimize the background of the non-

mutagenized vector, *Nde*I cleavage was performed before transformation. The RR69, cI and B94 selected sequences were inserted into pRIZ'O(-)RR69, while those selected for RR*69 and B31 were inserted into the pRIZ'O(-)RR*69 vector, to obtain pRIZ'O_XRR69 and pRIZ'O_XRR*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 (Simoncsits et al., 1997).

Construction of the expression library of single-chain repressor mutants and *in vivo* selection

The single-stranded, partially randomized oligonucleotide TATTCTCTGGTACCWC TNNSNNSAGTATCNNS**CAGCTCGAGCTG** (AT443, W = A or T; N = A, C, G or T; S = G or C) containing sequences with restriction sites of *Kpn*I and *Xho*I (underlined) and a 12 nucleotides self-complementary 3'-terminal region (bolded) was converted to a mixture of homoduplexes by self-annealing followed by Klenow polymerase fill-in reaction in the presence of dNTP. The resulting duplex was cleaved with *Kpn*I and *Xho*I, the product containing the randomized was purified by electrophoresis on non-denaturing 16% acrylamide gel. This randomized *Kpn*I - *Xho*I linker:

```
      C WCT NNS NNS AGT ATC NNS CAG C
CA TGG WGA NNS NNS TCA TAG NNS GTC GAG CT
```

was cloned into the corresponding sites of the pRIZ'O_R*1RR(KOX)69 vector. This vector contains a long (about 1.1 kbp) "stuffer" fragment between the *Kpn*I and *Xho*I sites of the region coding for the second domain. After electroporation into XL1-Blue cells, a library of 1.1 x 10^5 independent transformants was obtained

The *in vivo* screening for clones containing single-chain repressor mutants that reduce the expression of the α-fragment of the β-galactosidase reporter gene was performed briefly as follow. The library was transformed into XL1-Blue and the transformed cells in aliquots were plated onto LB agar plates containing 75 mg/l ampicillin and 10 mg/l tetracyclin to obtain about one to two thousand colonies per plate. The agar was covered with nitrocellulose filter (BA 85 type, Schleicher & Schüll) and the plates are incubated at 37°C for 10 to 12 hr. The nitrocellulose filter was then placed (colonies facing up) onto LB agar plates containing antibiotics as above, 1 mM IPTG and 25 mg/l X-gal. The plates were further incubated for 8 to 12 hr at 37°C. Most of the colonies turned slightly blue, but for better color discrimination, the plates were usually kept for two to four days at 4°C. Colonies which were strikingly paler blue than the average were picked for β-galactosidase assay, plasmid preparation and sequencing.

8. References

Aggarwal, A. K., Rodgers, D. W., Drottar, M., Ptashne, M. & Harrison, S. C. (1988). Recognition of a DNA operator by the repressor of phage 434: a view at high resolution. *Science* **242**, 899-907.

Anderson, J. E., Ptashne, M. & Harrison, S. C. (1987). Structure of the repressor-operator complex of bacteriophage 434. *Nature* **326**, 846-52.

Bell, A. C. & Koudelka, G. B. (1993). Operator sequence context influences amino acid-base-pair interactions in 434 repressor-operator complexes. *J Mol Biol* **234**, 542-53.

Bell, A. C. & Koudelka, G. B. (1995). How 434 repressor discriminates between OR1 and OR3. The influence of contacted and noncontacted base pairs. *J Biol Chem* **270**, 1205-12.

Berg, J. M. & Shi, Y. (1996). The galvanization of biology: a growing appreciation for the roles of Zinc. *Science* **271**, 1081-85.

Blackwell, T. K. (1995). Selection of protein binding sites from random nucleic acid sequences. *Methods Enzymol* **254**, 604-18.

Blackwell, T. K., Kretzner, L., Blackwood, E. M., Eisenman, R. N. & Weintraub, H. (1990). Sequence-specific DNA binding by the c-Myc protein. *Science* **250**, 1149-51.

Blackwell, T. K. & Weintraub, H. (1990). Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. *Science* **250**, 1104-10.

Brennan, R. G. (1991). Interactions of the helix-turn-helix binding domain. *Curr Opin Struct Biol* **1**, 80-88.

Brennan, R. G. (1992). DNA recognition by the helix-turn-helix motif. *Curr Opin Struct Biol* **2**, 100-8.

Brown, B. M. & Sauer, R. T. (1993). Assembly of the Arc repressor-operator complex: cooperative interactions between DNA-bound dimers. *Biochemistry* **32**, 1354-63.

Bullock, W. O., Fernandez, J. M. & Short, J. M. (1987). A high efficiency plasmid transforming *recA Escherichia coli* strain with beta-galactosidase selelction. *Biotechniques* **5**, 376-80.

Burley, S. K. (1996). The TATA box binding protein. *Curr. Opin. Struct. Biol.* **6**, 69-75.

Carey, J. (1991). Gel retardation. *Methods Enzymol* **208**, 103-17.

Carlson, P. A. & Koudelka, G. B. (1994). Expression, purification, and functional characterization of the carboxyl-terminal domain fragment of bacteriophage 434 repressor. *J Bacteriol* **176**, 6907-14.

Chen, J., Pongor, S. & Simoncsits, A. (1997). 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. *Nucleic Acids Res*, **25**, 2047-2054.

Chittenden, T., Livingston, D. M. & Kaelin, W. G., Jr. (1991). The T/E1A-binding domain of the retinoblastoma product can interact selectively with a sequence-specific DNA-binding protein. *Cell* **65**, 1073-82.

Choo, Y. & Klug, A. (1994a). Selection of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc Natl Acad Sci U S A* **91**, 11168-72.

Choo, Y. & Klug, A. (1994b). Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage [published erratum appears in Proc Natl Acad Sci U S A 1995 Jan 17;92(2):646]. *Proc Natl Acad Sci U S A* **91**, 11163-7.

Choo, Y. & Klug, A. (1997). Physical basis of a protein-DNA recognition code. *Curr Opin Struct Biol* **7**, 117-25.

Choo, Y., Sanchez-Garcia, I. & Klug, A. (1994). In vivo repression by a site-specific DNA-binding protein designed against an oncogenic sequence. *Nature* **372**, 642-5.

Dallman, G., Marincs, F., Papp, P., Gaszner, M. & Orosz, L. (1991). The isolated N-terminal domain of the c repressor of bacteriophage 16-3 is functional in DNA binding *in vivo* and *in vitro*. *Mol. Gen. Genet.* **227**, 106-12.

Dallmann, G., Papp, P., Orosz, L. (1987) Related repressor specificity of unrelated phages. *Nature* **330**, 398-401.

Desjarlais, J. R. & Berg, J. M. (1993). Use of a zinc-finger consensus sequence framework and specificity rules to design specific DNA binding proteins. *Proc Natl Acad Sci U S A* **90**, 2256-60.

Desjarlais, J. R. & Berg, J. M. (1994). Length-encoded multiplex binding site determination: application to zinc finger proteins. *Proc Natl Acad Sci U S A* **91**, 11099-103.

Drew, H. R. & McCall, M. J. (1990). New approaches to DNA in the crystal and in solution. *DNA Topology and its Biological Effects* (Cozzarelli, N. R., Ed.), Vol. , pp. 1-56, Cold Spring Harbor Laboratory Press, Cold Spring Habor, NY.

Ebright, R. H., Cossart, P., Gicquel-Sanzey, B. & Beckwith, J. (1984). Mutations that alter the DNA sequence specificity of the catabolite gene activator protein of *E. Coli*. *Nature* **311**, 232-35.

Elledge, S. J., Sugiono, P., Guarente, L. & Davis, R. W. (1989). Genetic selection for genes encoding sequence-specific DNA-binding proteins. *Proc. Natl. Acad. Sci.* **86**, 3689-93.

Ellenberger, T. (1994). Getting a grip on DNA recognition: structures of the basic region leucine zipper, and the basic region helix-loop-helix DNA-binding domains. *Curr. Opin. Struct. Biol.* **4**, 12-21.

Fairall, L., Harrison, S. D., Travers, A. A. & Rhodes, D. (1992). Sequence-specific DNA binding by a two zinc-finger peptide from the *Drosophila melanogaster* Tramtrack protein. *J Mol Biol* **226**, 349-66.

Freemont, P. S., Lane, A. N. & Sanderson, M. R. (1991). Structural aspects of protein-DNA recognition. *Biochem J* **278**, 1-23.

Funk, W. D. & Wright, W. E. (1992). Cyclic amplification and selection of targets for multicomponent complexes: myogenin interacts with factors recognizing binding sites for basic helix-loop-helix, nuclear factor 1, myocyte-specific enhancer-binding factor 2, and COMP1 factor. *Proc Natl Acad Sci U S A* **89**, 9484-8.

Gehring, W. J., Qian, Y. Q., Billeter, M., Furukubo-Tokunaga, K., Schier, A. F., Resnedez-Perez, D., Affolter, M., Otting, G. & Wüthrich, K. (1994). Homeodomain-DNA recognition. *Cell* **78**, 211-23.

Gogos, J. A., Jin, J., Wan, H., Kokkinidis, M. & Kafatos, F. (1996). recognition of diverse sequences by class I zinc fingers: Asymmetries and indirect effects on specificity in the interaction between CF2II and A+T rich sequence elements. *Proc Natl Acad Sci USA* **93**, 2159-64.

Gold, L., Polisky, B., Uhlenbeck, O. & Yarus, M. (1995). Diversity of oligonucleotide functions. *Annu. Rev. Biochem.* **64**, 763-97.

Greisman, H. A. & Pabo, C. O. (1997). A general strategy for selecting high-affinity zinc finger proteins for diverse DNA target sites. *Science* **275**, 657-61.

Hanahan, D. (1985). Techniques for transformation of *E. coli. DNA cloning: a practical approach* (Golver, D. M., Ed.), Vol. I, pp. 109-36, IRL press, Oxford.

Harrison, S. C. (1991). A structural taxonomy of DNA-binding. *Nature* **353**, 715-9.

Harrison, S. C. & Aggarwal, A. K. (1990). DNA recognition by proteins with the helix-turn-helix motif. *Annu. Rev. Biochem.* **59**, 933-69.

Herr, W. & Cleary, M. A. (1995). The POU domain: versatility in transcriptional regulation by a flexible two-in-one DNA-binding domain. *Genes Dev* **9**, 1679-93.

Hilchey, S. P., Wu, L. & koudelka, G. B. (1995). DNA binding specificity of the recognition helix of P22 repressor protein. *J Biomol Struct Dyn* **12**, a092.

Hollis, M., Valenzuela, D., Pioli, D., Wharton, R. & Ptashne, M. (1988). A repressor heterodimer binds to a chimeric operator. *Proc Natl Acad Sci U S A* **85**, 5834-8.

Hu, L., Sera, T. & Schultz, P. G. (1994). A permutational approach toward protein-DNA recognition. *Proc. Natl. Acad. Sci.* **91**, 3969-73.

Huston, J. S., Levinson, D., Mudgett-Hunter, M., Tai, M.-S., Novotny, J., Margolies, M. N., Ridge, R. J., Bruccoleri, R. E., Haber, E., Cream, R. & Opperman, H. (1988). Protein engineering of antibody binding sites: recovery of specific activity in an anti-digoxin single-chain Fv analogue produced in *Escherichia coli*. *Proc. Natl. Acad. Sci.* **85**, 5879-83.

Huynh, T. V., Young, R. A. & Davis, R. W. (1984). Constructing and screening cDNA libraries in λgt10 and λgt11. *In DNA Cloning: A Practical Approach* (Glower, D. M., Ed.), Vol. 1, pp. 49-78, IRL Press, Oxford.

Jamieson, A. C., Kim, S. H. & Wells, J. A. (1994). *In vitro* selection of zinc fingers with altered DNA-binding specificity. *Biochemistry* **33**, 5689-95.

Johnson, A. D., Meyer, B. J. & Ptashne, M. (1979). Interaction between DNA-bound repressors govern regulation by the λ phage repressor. *Proc Natl Acad Sci USA* **76**, 5061-65.

Kim, J. G., Takeda, Y., Matthews, B. W. & Anderson, W. F. (1987). Kinetic studies on Cro repressor-operator DNA interaction. *J Mol Biol* **196**, 149-58.

Kinzler, K. W. & Vogelstein, B. (1989). Whole genome PCR: application to the identification of sequences bound by gene regulatory proteins. *Nucleic Acids Res* **17**, 3645-53.

Klug, A. (1993). Co-chairman's remarks: protein designs for the specific recognition of DNA. *Gene* **135**, 83-92.

Koudelka, G. B. (1991). Bending of synthetic bacteriophage 434 operators by bacteriophage 434 proteins. *Nucleic Acids Res* **19**, 4115-9.

Koudelka, G. B., Bell, A. C. & Hilchey, S. P. (1996). Indirect effects of DNA sequence on protein-DNA interactions. *Biological Structure and Dynamics* (Sharma, R. H. & Sharma, M. H., Eds.), Vol. I, pp. 135-53, Adenine Press, Inc., New York.

Koudelka, G. B. & Carlson, P. (1992). DNA twisting and the effects of non-contacted bases on affinity of 434 operator for 434 repressor. *Nature* **355**, 89-91.

Koudelka, G. B., Harbury, P., Harrison, S. C. & Ptashne, M. (1988). DNA twisting and the affinity of bacteriophage 434 operator for bacteriophage 434 repressor. *Proc Natl Acad Sci U S A* **85**, 4633-7.

Koudelka, G. B., Harrison, S. C. & Ptashne, M. (1987). Effect of non-contacted bases on the affinity of 434 operator for 434 repressor and Cro. *Nature* **326**, 886-8.

Koudelka, G. B. & Lam, C. Y. (1993). Differential recognition of OR1 and OR3 by bacteriophage 434 repressor and Cro. *J Biol Chem* **268**, 23812-7.

Kunkel, T., Roberts, J. & Zakour, R. (1987). Rapid and efficient site specific mutagenesis without phenotypic slection. *Methods Enzymol* **154**, 367-82.

Kunkel, T. A., Bebenek, K. & McClary, J. (1991). Efficient site-directed mutagenesis using uracil-containing DNA. *Methods Enzymol* **204**, 125-39.

Lehming, N., Sartorius, J., Niemöller, M., Genenger, G., v. Wilcken-Bergmann, B. & Müller- Hill, B. (1987). The interaction of the recognition helix of *lac* repressor with *lac* operator. *EMBO J* **6**, 3145-53.

Lehming, N., Sartorius, J., Oehler, S., von Wilcken-Bergmann, B. & Müller-Hill, B. (1988). Recognition helices of *lac* and lambda repressor are oriented in opposite directions and recognize similar DNA sequences. *Proc Natl Acad Sci U S A* **85**, 7947-51.

Luisi, B. F., Xu, W. X., Otwinowski, Z., Freedman, L. P., Yamamoto, K. R. & Siegler, P. B. (1991). Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature* **352**, 497-505.

Lukacsovich, T., Orosz, A., Baliko, G. & Venetianer, P. (1990). A family of expression vectors based on the *rrnB* P2 promoter of Escherichia coli. *J. Biotechnol.* **16**, 49-56.

Maxam, A. M. & Gilbert, W. (1980). Sequencing end-labeled DNA with base-specific chemical cleavages. *Methods Enzymol* **65**, 499-560.

McCabe, P. C. (1990). . *PCR Protocols* (Innis, M. A., Gelfand, D. H., Sninsky, J. J. & White, T. J., Eds.), Vol. , pp. 76-83, Academic Press, Inc., San Diego, CA.

Miller, J. H. (1972). *Experiments in Molecular Genetics*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Miller, J. H., Lebkowski, J. S., Greisen, K. S. & Calos, M. P. (1984). Specificity of mutations induced in transfected DNA by mammalian cells. *EMBO J.* **3**, 3117-21.

Mondragon, A. & Harrison, S. C. (1991). The phage 434 Cro/OR1 complex at 2.5 Å resolution. *J Mol Biol* **219**, 321-34.

Mondragon, A., Subbiah, S., Almo, S. C., Drottar, M. & Harrison, S. C. (1989a). Structure of the amino-terminal domain of phage 434 repressor at 2.0 Å resolution. *J Mol Biol* **205**, 189-200.

Mondragon, A., Wolberger, C. & Harrison, S. C. (1989b). Structure of phage 434 Cro protein at 2.35 A resolution. *J Mol Biol* **205**, 179-88.

Mossing, M. C., Bowie, J. U. & Sauer, R. T. (1991). A streptomycin selection for DNA-binding activity. *Methods Enzymol.* **208**, 604-19.

Nikolnikov, S., Posfai, G. & Sain, B. (1984). The construction of a versatile plasmid vector that allows direct slection of fragments cloned into six unique sites of the cI gene of coliphage 434. *gene* **30**, 261-65.

Ogata, K., Morikawa, S., Nakamura, H., Sekikawa, A., Inoue, T., Kanai, H., Sarai, A., Ishii, S. & Nishimura, Y. (1994). Solution structure of a specific DNA complex of the Myb DNA-binding domain with cooperative recognition helices. *Cell* **79**, 639-48.

Pabo, C. O. & Sauer, R. T. (1984). Protein-DNA recognition. *Annu. Rev. Biochem.* **53**, 293-321.

Pabo, C. O. & Sauer, R. T. (1992). Transcription factors: Structural Families and Principles of DNA Recognition. *Annu. Rev. Biochem.* **61**, 1053-95.

Percipalle, P., Simoncsits, A., Zakhariev, S., Guarnaccia, C., Sanchez, R. & Pongor, S. (1995). Rational designed helix-turn-helix proteins and their conformational changes upon DNA binding. *EMBO J* **14**, 3200-5.

Pierrou, S., Enerbäck, S. & Carlsson, P. (1995). Selection of high-affinity binding sites for sequence-specific, DNA binding proteins from random sequence oligonucleotides. *Anal. Biochem.* **229**, 99-105.

Pollock, R. & Treisman, R. (1990). A sensitive method for the determination of protein-DNA binding specificities. *Nucleic Acids Res* **18**, 6197-204.

Pomerantz, J. L., Sharp, P. A. & Pabo, C. O. (1995). Structure-based design of transcription factors. *Science* **267**, 93-96.

Poteete, A. R., Ptashne, M., Ballivet, M. & Eisen, H. (1980). Operator sequences of bacteriophages P22 and 21. *J Mol Biol* **137**, 81-91.

Poteete, A. R., Ptashne, M., Ballivet, M. & Eisen, H. (1982). Control of transcription by the bacteriophage P22 repressor. *J Mol Biol* **157**, 21-48.

Ptashne, M. (1992). 2nd edit. *A Genetic Switch*, Cell Press and Blackwell Scientific Publications, Cambridge, MA.

Raumann, B. E., Brown, B. M. & T., S. R. (1994). Major groove DNA recognition by β-sheets: the ribbon-helix-helix family of gene regulatory proteins. *Current Opinion in Structural Biology* **4**, 36-43.

Rebar, E. J. & Pabo, C. O. (1994). Zinc finger phage: affinity selection of fingers with new DNA-binding specificities. *Science* **263**, 671-3.

Rhodes, D., Schwabe, J. W., Chapman, L. & Fairall, L. (1996). Towards an understanding of protein-DNA recognition. *Philos Trans R Soc Lond B Biol Sci* **351**, 501-9.

Robinson, C. R. & Sauer, R. T. (1996). Covalent attachment of Arc repressor subunits by a peptide linker enhances affinity for operator DNA. *Biochemistry* **35**, 109-16.

Rodgers, D. W. & Harrison, S. C. (1993). The complex between phage 434 repressor DNA-binding domain and operator site OR3: structural differences between consensus and non-consensus half-sites. *Structure* **1**, 227-40.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). 2nd edit. *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring harbor, NY.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* **74**, 5463.

Sauer, R. T., Pabo, C. O., Meyer, B. J., Ptashne, M. & Backman, K. D. (1979). Regulatory functions of λ repressor residue in the amino-terminal domain. *Nature* **279**, 396-400.

Schoepfer, R. (1993). The pRSET family of T7 promoter expression vectors for Escherichia coli. *Gene* **124**, 83-5.

Schwabe, J. W. R. & Klug, A. (1994). Zinc mining for protein domains. *Nature Struct Biol* **1**, 345-9.

Schwabe, W. R. (1997). The role of water in protein-DNA interactions. *Curr Opin Struct Biol* **7**, 126-34.

Seeman, N. C., Rosenberg, J. M. & Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc Natl Acad Sci USA* **73**, 804-8.

Shimon, L. J. & Harrison, S. C. (1993). The phage 434 OR2/R1-69 complex at 2.5 A resolution. *J Mol Biol* **232**, 826-38.

Simoncsits, A., Bristulf, J., Tjornhammar, M. L., Cserzo, M., Pongor, S., Rybakina, E., Gatti, S. & Bartfai, T. (1994). Deletion mutants of human interleukin 1 beta with significantly reduced agonist properties: search for the agonist/antagonist switch in ligands to the interleukin 1 receptors. *Cytokine* **6**, 206-14.

Simoncsits, A., Chen, J., Percipalle, P., Wang, S., Törö, I. & Pongor, S. (1997). Single-chain repressors containing engineered DNA-binding domains of the phage 434 repressor recognize symmetric or asymmetric DNA operators. *J Mol Biol* **267**, 118-31.

Simoncsits, A., Tjörnhammar, M. L., kálmán, M., Cserpán, I., Gafvelin, G. & Bartfai, T. (1988). Synthesis, cloning and expression in *Escherichia coli* of artificial genes coding for biologically active elongated precusors of the vasoactive intestinal polypeptide. *Eur. J. Biochem.* **178**, 343-50.

Sinden, R. R. (1994). DNA-protein interactions. *DNA Structure and Function*, pp. 287-325, Academic Press Inc., San Diego, CA, USA.

Studier, F. W., Rosenberg, A. H., Dunn, J. J. & Dubendorff, J. W. (1990). Use of T7 RNA polymerase to direct expression of cloned genes. *Methods Enzymol* **185**, 60-89.

Suzuki, M. (1993). Common features in DNA recognition helices of eukaryotic transcription factors [published erratum appears in EMBO J 1993 Oct;12(10):4042]. *EMBO J* **12**, 3221-6.

Suzuki, M. (1994). A framework for the DNA-protein recognition code of the probe helix in transcription factors: the chemical and stereochemical rules. *Structure* **2**, 317-26.

Suzuki, M., Brenner, S. E., Gerstein, M. & Yagi, N. (1995a). DNA recognition code of transcription factors. *Protein Eng* **8**, 319-28.

Suzuki, M. & Yagi, N. (1994). DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families. *Proc Natl Acad Sci U S A* **91**, 12357-61.

Suzuki, M., Yagi, N. & Gerstein, M. (1995b). DNA recognition and superstructure formation by helix-turn-helix proteins. *Protein Eng* **8**, 329-38.

Szostak, J. W. (1992). In vitro genetics. *TIBS* **17**, 89-93.

Taylor, W. E., Suruki, H. K., Lin, A. H., Naraghi-Arani, P., Igarashi, R. Y., Younessian, M., Katkus, P. & Vo, N. V. (1995). Designing zinc-finger ADR1 mutants with altered specificity of DNA binding to T in UAS1 sequences. *Biochemistry* **34**, 3222-30.

Thiesen, H. J. & Bach, C. (1990). Target Detection Assay (TDA): a versatile procedure to determine DNA binding sites as demonstrated on SP1 protein. *Nucleic Acids Res* **18**, 3203-9.

Tjörnhammar, M. L. & Simoncsits, A. (1991). Direct selection vectors, promoter-active adaptors and their use for gene assembly by a single-stranded cloning method. *Nucl. Acids Res. Symp. Series* **24**, 252.

Travers, A. A. (1989). DNA conformation and protein binding. *Annu Rev Biochem* **58**, 427-52.

Travers, A. A. & Klug, A. (1990). Bending of DNA in nucleoprotein complexes. *DNA Topology and its Biological Effects* (Cozzarelli, N. R. & Wang, J. C., Eds.), pp. 57-106, Cold Spring Harbor Laboratory Press.

Ullman, A., Jacob, F. & Monod, J. (1967). Characterization by in vitro complementation of a pepeide corresponding to an operator-proximal segment of the β-galactosidase structural gene of *Escherichia coli*. *J. Mol. Biol.* **24**, 339-43.

Webster, C., Merryweather, A. & Brammar, W. (1992). Efficient repression by a heterodimeric repressor in *Escherichia coli*. *Mol Microbiol* **6**, 371-77.

Wharton, R. P., Brown, E. L. & Ptashne, M. (1984). Substituting an alpha-helix switches the sequence-specific DNA interactions of a repressor. *Cell* **38**, 361-69.

Wharton, R. P. & Ptashne, M. (1985). Changing the binding specificity of a repressor by redesigning an alpha- helix. *Nature* **316**, 601-5.

Wharton, R. P. & Ptashne, M. (1987). A new-specificity mutant of 434 repressor that defines an amino acid- base pair contact. *Nature* **326**, 888-91.

Wilson, D. S. & Desplan, C. (1995). Cooperating to be different. *Curr Biol* **5**, 32-34.

Wolberger, C. (1996). Homeodomain interaction. *Curr. Opin. Struct. Biol.* **6**, 62-68.

Wright, W. E. & Walter, D. F. (1993). CASTing for multicomponent DNA-binding complexes. *TIBS* **18**, 77-80.

Wu, H., Yang, W. P. & Barbas, C. F. r. (1995). Building zinc fingers by selection: toward a therapeutic application. *Proc Natl Acad Sci U S A* **92**, 344-48.

Wu, L. & Koudelka, G. B. (1993). Sequence-dependent differences in DNA structure influence the affinity of P22 operator for P22 repressor. *J Biol Chem* **268**, 18975-81.

Wu, L., Vertino, A. & Koudelka, G. B. (1992). Non-contacted bases affect the affinity of synthetic P22 operators for P22 repressor. *J Biol Chem* **267**, 9134-39.

Yanisch-Perron, C., Viera, J. & Messing, J. (1985). Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene* **33**, 103-19.

Appendix I: List of abbreviations

bHLH	basic helix-loop-helix
bp	base pair
bZip	leucine zipper
CASTing	cyclic amplification and selection of targets
CD	circular dichroism spectroscopy
DBD	DNA-binding domain
DBP	DNA-binding protein
EMSA	electrophoretic mobility shift assay
HEPES	N-(2-Hydroxyethyl)piperazine-N-(2-ethanesulfonicacid)
HTH	helix-turn-helix
IPTG	isopropyl-β-D-thiogalactoside
Kd	apparent equilibrium dissociation constant
k_{off}	dissociation rate constant
k_{on}	association rate constant
MOPS	3-(N-mopholino)propanesulfonic acid
NMR	nuclear magnetic resonance
PCR	polymerase chain reaction
PEG	polyethylene glycol
SAAB	selected and amplified binding sites
SDS	sodium dodecyl sulfate
SELEX	systematic evolution of ligands by expotential enrichment
TDA	target detection assay
X-gal	5-bromo-4-chloro-3-indolyl-β-D-galactoside

Appendix II:

List of Figures

List of Tables

ACKNOWLEDGEMENT

This thesis work was carried out in the Protein Structure and Function Group at the International Centre for Genetic Engineering and Biotechnology.

First, I wish to express my thanks to my supervisors Dr. András Simoncsits, who has provided me with extensive instructions and patient guidance and Prof. Sándor Pongor, who has given me many helpful advices and consistent encouragement. Without their help this work could not have been accomplished.

Many thanks go to Alexander Athanasiadis and Kristian Vlahovicek for their constructive suggestions and help. I would like to thank Corrado Guarnaccia and all the other people in the Protein Structure and Function Group, as well as my ICGEB and SISSA colleagues, for their kindly support in the past three years. I am also thankful to Marie Luise Tjörnhammar and Rajalakshmi Pariyarathuparambil for the critical reading of some parts of the manuscript.

I am particularly grateful to Prof. Arturo Falaschi, Director General of ICGEB, Prof. Xiaocheng Gu, Scientific Advisor of ICGEB, Prof. Zengquan Zhou, Dean of the college of life science of Beijing University, and Prof. Zhangliang Chen, Vice President of Beijing University, for providing me the opportunity for this Ph.D. study.

I have warm memories of the wonderful time shared with my friends during my stay in Trieste, Italy.

I want to dedicate this thesis to my husband, father, mother and sister, their deep love made this work possible.