ISAS - INTERNATIONAL SCHOOL
FOR ADVANCED STUDIES

ENGINEERING DNA-BINDING PROTEINS BASED ON THE HELIX-TURN-HELIX MOTIF

Thesis submitted for the degree
"Doctor Philosophiae"

CANDIDATE

Piergiorgio Percipalle

SUPERVISORS

Prof. Sándor Pongor
Dr. András Simoncsits

Academic Year 1994-1995

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

TRIESTE

Mathias, Delphine,
Papà, Mamma, Davide

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

Ahx	6-aminohexanoic acid
Boc	*tert*-butyloxicarbonyl
Bp	base pairs
BSA	bovine serum albumin
Bu, But	butyl
CD	circular dichroism
DCM	dichloromethane
DMF	dimethylformamide
DMS	dimethylsulfate
DMSO	dimethylsulfoxide
dNTP	deoxy-nucleoside 5'-triphosphate
DTT	dithiothreitol
EDT	ethandithiol
EDTA	ethylenediamine tetra acetic acid
ESMS	electrospray mass spectrometry
FABMS	fast atom bombardment mass spectrometry
Fmoc	fluorenylmethoxycarbonyl
HBTU	O-benzotriazolyl-N, N, N', N'-tetramethyluronium hexafluorophosphate
HEPES	N-[2-Hydroxyethyl]piperazine-N-[2-ethanesulfonic acid]
HOBt	hydroxybenzotriazole
IEHPLC	ion-exchange high performance liquid chromatography
OPfp	pentafluorophenyl ester
ODhbt	oxodihydrobenzotriazine
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
PEG	polyethylene glycol
PMC	2, 2, 5, 7, 8-pentamethyl-chroman-6-sulfonyl
PMSF	phenylmethyl sulfonyl fluoride
PS	polystyrene
RPHPLC	reversed phase high performance liquid chromatography
SDS	sodium dodecyl sulfate
SPPS	solid phase peptide synthesis
TA	thioanysol
TBTU	O-benzotriazol-N, N, N', N'-tetramethyluronium tetrafluoroborate
tBu	*tert*-butyl
tBuO	*tert*-butyloxy
TFA	trifluoroacetic acid
TFE	trifluoroethanol
Tris	tris(hydroxymethyl)aminomethane
Trt	trityl

ABSTRACT

The goal of this work was to study the highly specific nature of protein-DNA recognition, using a "rational design approach", i.e. by attempting to construct artificial DNA-binding proteins. The N-terminal domain of the 434 bacteriophage cI repressor, a five-helix domain with a helix-turn-helix motif (HTH) was used as an example. As most known DNA-binding proteins are dimers or contain multiple DNA-binding modules, we decided to design covalently dimerized bipartite DNA-binding proteins using elements of the 434 repressor. Specific DNA-binding was assayed by biochemical methods and protein-DNA interactions studied by biophysical techniques using circular dichroism (CD) difference spectroscopy.

The new polypeptides were designed on the basis of the 3D structure of the 434 repressor, using computer graphics and conformational energy minimization/molecular dynamics techniques. The designed peptides were prepared by solid phase peptide synthesis or by recombinant DNA techniques.

In one set of experiments (based on a "minimalist approach"), the HTH motif and its flanking elements were covalently dimerized with synthetic linkers. The resulting molecules (P4 to P6) were prepared by solid phase peptide synthesis, in the form of dyad-symmetric branched peptides, connected through their C-termini. However, none of these molecules showed specific DNA binding in the presence of competitor DNA.

In another set of experiments, the entire N-terminal domain was dimerized either in a dyad-symmetric fashion, using peptide synthesis (P7), or as direct sequence

repeats produced with recombinant DNA techniques (RR69). In P7, a symmetric linker was used that consisted of two 6-aminohexanoic acid residues connected to both amino groups of a central lysine molecule. RR69 was prepared by recombinant DNA methods, and the two N-terminal domains were connected with the internal 69-89 sequence of the 434 repressor. In gel-mobility shift experiments, both of these new molecules showed specific DNA binding, and they also showed the expected sequence specificity in footprinting, methylation interference and in vitro transcription inhibition experiments. These results show that dimeric architecture is crucial for achieving native-like properties in artificial DNA-binding proteins.

The conformational changes of the peptides occurring on DNA binding were studied using CD difference spectroscopy. Under the experimental conditions used, the monomeric N-terminal domain (P1) does not seem to interact with DNA since α-helical induction was not detected. In contrast, both P7 and RR69 showed strong increase in their α-helical content when exposed to equimolar amounts of cognate DNA. Interestingly, a similar (about 30 % weaker) induction was caused by non-cognate DNA. These results indicate that, in contrast to current views, conformational change is a part of specific DNA recognition by the HTH modules, and may also be crucial in scanning the DNA molecule in search of the specific binding sites.

1. INTRODUCTION

Protein-DNA recognition is at the core of gene regulation. 3D structures are available on many protein-DNA complexes that reveal a variety of binding mechanisms. In Nature, the process of protein-DNA recognition relies on a few structural motifs that are common to many protein families involved in specific DNA binding. Such motifs frequently include helical binding structures and are sometimes flanked with aspecific DNA binding regions, such as stretches of basic amino acids. The helix-turn-helix (HTH) motif is a common structural element found in prokaryotic repressors and also in some eukaryotic transcription factors. Its functions and the basis of its specificity have been studied in great detail in several bacteriophages and a number of 3D structures are currently available.

Generally speaking, interaction of a protein with DNA can be sequence-specific (like interactions with bases) or non-specific (like peptide-backbone/phosphate interactions). In addition, bending propensity of the DNA molecule seems to play a quasi-specific role. Thermodynamic analysis points to the role of the hydrophobic effect that covers the energy expenses of conformational change upon binding. In contrast to most DNA binding domains, HTH proteins are believed to dock on DNA essentially as rigid bodies.

Several proteins undergo conformational changes upon DNA binding, which gives rise to α-helical structures. This phenomenon can be conveniently studied by circular dichroism using difference spectroscopy at low protein-DNA concentrations.

Protein-DNA interactions are involved in many of the fundamental processes that occur in cells, including packaging, replication, recombination, restriction and transcription (Freemont *et al.*, 1991). It is important to understand the nature of these interactions, as they are intimately involved in the control of gene expression, cell division and differentiation. Of the proteins that interact with DNA, restriction enzymes and transcription factors need to discriminate DNA binding sites with specific sequences for correct functioning. Other proteins, such as nucleases and chromosomal proteins, must by definition be able to bind many different and unrelated DNA sequences. Such proteins possess a certain selectivity in the binding process even though they do not bind with equal affinity to different DNA sequences.

DNA binding proteins have been classified according to the small, discrete structural domains that recognize the specific sites on the DNA (for reviews see Ellemberger, 1994; Brennan, 1993; Pabo and Sauer; 1992; Harrison, 1991). In some cases these domains can be interchanged between proteins, showing that they are independently folded units. Several domain groups have been identified on the basis of primary sequence relationships, and three dimensional structures have been determined for members of many of these groups.

Among the various DNA-binding proteins, transcription factors (present both in prokaryotic and eukaryotic organisms) regulate cell development, differentiation and cell growth by binding to specific sequences on the DNA and regulating gene expression. It is therefore of primary importance to investigate the nature of specific protein-DNA recognition in order to gain more insights on how gene expression is regulated.

Now that a number of protein-DNA complexes has been solved, it is possible to take a broader look at the problems of specific protein-DNA recognition. However, this is a very difficult task since specific recognition is the result of a set of contributions of different factors.

Table 1.1. Classes of DNA binding proteins

Class	Example	[a]Subunit	Allosteric Effector	Target DNA
HTH	λcI	α2 (26)	No	TATCACCGCCAGTGGTA
	λcro	α2 (7.5)	No	TATCACCGCCAGTGGTA
	cI434	α2 (7.5)	No	TACAAGAAAGTTTGTT
	cro434	α2 (7.5)	No	TACAAGAAAGTTTGTT
	trp	α2 (12.5)	L-Trp	CGAAGAAAGTTTGTT
	CAP	α2 (22.5)	c-AMP	$TGTGA(N)_6TCACT$
	lac	α4 (35)	allo-Lac	GAATTGTGAGCGCTCACAATT
	ant	α (7)	No	ATTA
	eng	α (7)	No	TAAT
	FIS	α2 (10)	No	Not Applicable
Zn-finger (C_2H_2)	mKR2	α (3)	No	Not Known
	ADR1	α (3)	No	Not Applicable
	SW15	α (3)	No	CCAGCATGCTATAATGC
	HEB	α (3)	No	GGGGAATCCCC
	Xfin	α (3)	No	Not Known
Zn-finger (C_4)	glc	α2 (7.6)	glucorticoid	$AGAACA(N)_3TGTTCT$
	est	α2 (7.6)	oestrogen	$AGGTCA(N)_3TGACCT$
L-zipper	fos	α2 (8)	No	GTGACTCAG
	jun	α2 (8)	No	GTGACTCAG
	GCN4	α2 (8)	No	ATGACTCTT
β-Sheet	arc	α2 (8.5)	No	CATGATAGAAGCACTCTACTA
	metJ	α2 (12)	S-Adenosyl- -Met	AGACGT
	HU	α2 (9.5)	No	Not Applicable
Homeobox	MATa1	α	No	ATTA
Other	TFIID	α	No	TATA
	EcoRI	α2 (31)	No	GAATTC
	EcoRV	α2 (28.5)	No	GATATC
	DNAseI	α (30.5)	No	Not Applicable
	DNA pol I	α (103)	No	Not Applicable

[a] Subunit weights are expressed in KDa

1.1 STRUCTURAL MOTIFS INVOLVED IN SITE-SPECIFIC DNA RECOGNITION

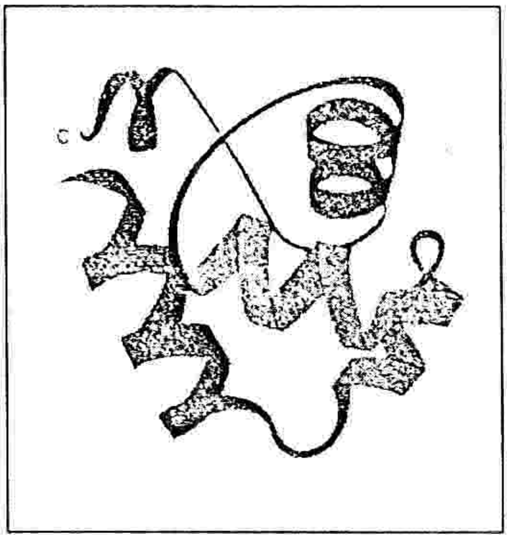
Most sequence-specific DNA binding proteins recognize and bind cognate DNA through well defined structural domains which make sequence-specific contacts with the DNA bases in the major groove. These contacts essentially occur via the H-bonding interactions of amino acid side-chains and the DNA bases in the major groove, and are stabilized by hydrophobic and electrostatic interactions. In contrast, structural studies of the HU class of proteins, DNAseI (Travers and Klug, 1990) and certain DNA binding drugs suggest that the minor groove and phosphate backbone may be primary recognition features for non sequence-specific DNA ligands (Mair *et al.*, 1991; Teng *et al*, 1988; Pjura *et al*, 1987).

At present there are at least five structural classes of major groove DNA binding proteins that have been characterized in some structural detail: (i) the Helix-turn-helix (HTH) proteins, (ii) the Zinc-finger proteins, (iii) the Leucine zipper proteins (bZIP), (iv) the Beta proteins and (v) the basic Helix-loop-helix proteins (bHLH). A list of some well known DNA-binding proteins is shown in Table 1.1 (for recent reviews see: Ellemberger, 1994; Brennan, 1993; Wolberger, 1993; Berg, 1993; Pabo and Sauer, 1992; Freemont *et al*, 1991; Harrison, 1991).

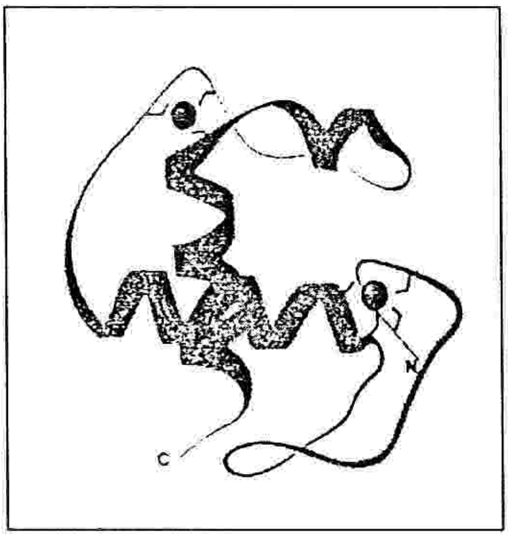
The motifs involved in DNA-binding (Figure 1.1) can be highly conserved in sequence and secondary structure. However it appears that they are by themselves insufficient to impart DNA-binding properties to a protein as it has been shown that in some cases the DNA-binding domain can extend well beyond such structural motifs. In the case of the λcI and trp repressor proteins, for example, flexible terminal extensions are present, while electrostatic complementarity appears to be important for CAP (Warwicker *et al.*, 1987). Furthermore, the DNA molecule often undergoes conformational changes on binding. Thus EcoRI restriction enzyme causes partial unwinding of the DNA double helix, while the CAP protein and the trp and 434 repressor proteins cause substantial bending and kinking of the DNA molecule (Schulz *et al.*, 1990). It is interesting to note that those proteins which cause conformational changes to the DNA molecule do not make use of flexible extensions to create further contacts, whereas proteins that appear to have little effect on DNA conformation in contrast do possess such flexible extensions so as to increase the interaction surface (Freemont *et al.*, 1991).

In this thesis work the helix-turn--helix (HTH) structural motif is dealt with in detail, and section 1.2 describes several regulatory proteins that use this motif to provide sequence-specific binding to DNA. The other DNA-binding motifs (Zinc-finger proteins, bHLH proteins, TATA-box binding proteins, bZIP proteins, Beta-sheet domains) are briefly reviewed in Appendix II.

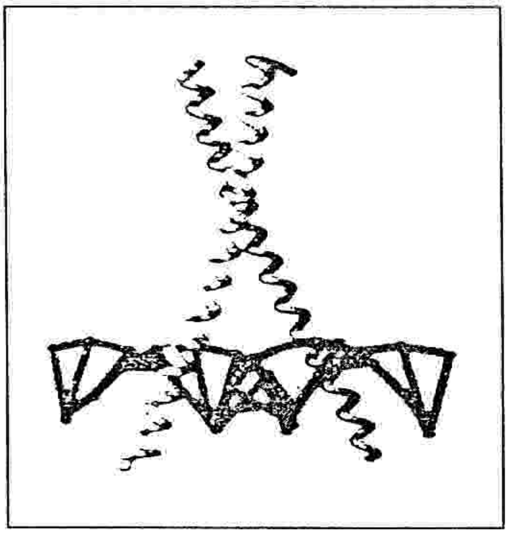
A



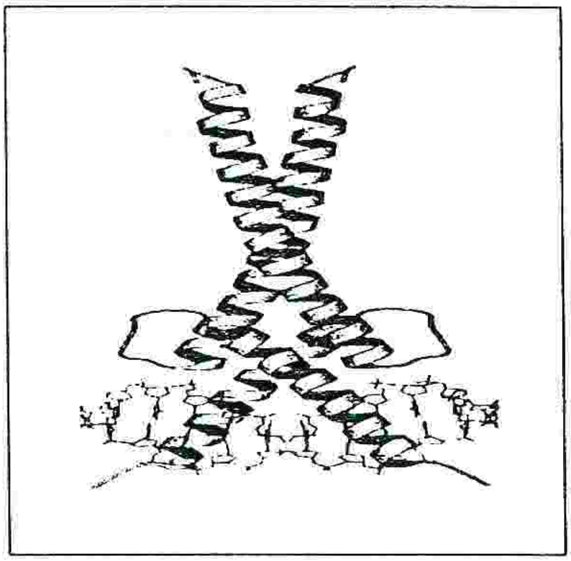
B



C



D



E

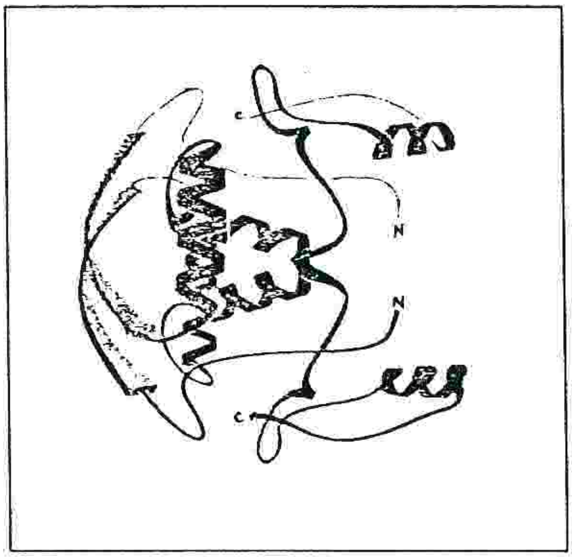


Figure 1.1. Common structural motifs found in DNA binding proteins. (A) HTH; (B) Zinc finger; (C) bZIP; (D) bHLH; (E) Beta-proteins.

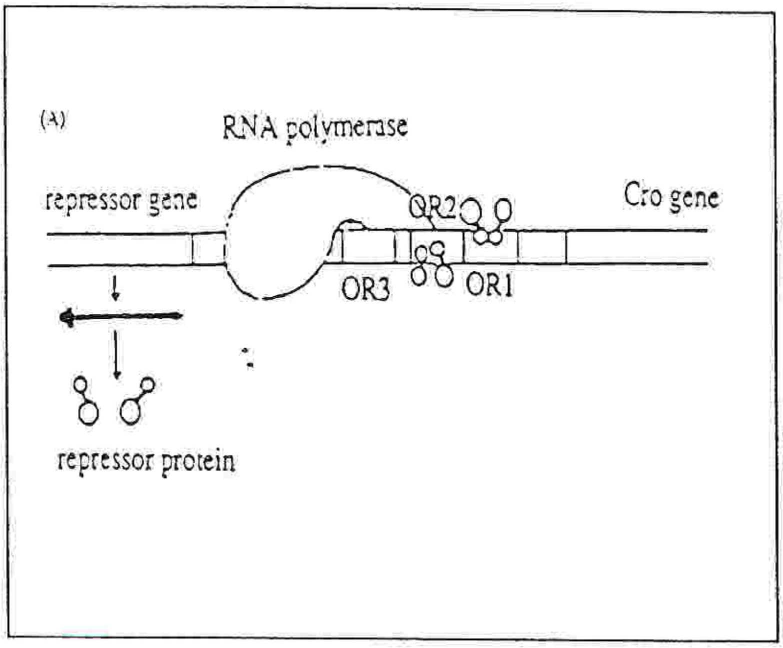
1.2 THE HELIX-TURN-HELIX MOTIF

The helix-turn-helix (HTH) motif has been found in proteins that are directly involved in the regulation of gene expression and also in proteins that serve structural and catalytic roles in other cellular processes (Branden and Tooze, 1991). Many prokaryotic DNA-binding domains contain this motif as well as some eukaryotic transcription factors, like the homeobox gene family, which serve several functions including the genetic control of development (Kornberg, 1993; McGinnis and Krumlauf, 1992; Scott *et al.*, 1989; Gerhing, 1987).

In the HTH motif the two α-helices are roughly perpendicular to each other and are separated by a β-turn. Dimerization results in a rotationally symmetric molecule, in which the motif is contained in each amino-terminal domain, whereas the carboxy-terminal domain, when present, serves for dimerization (Branden and Tooze, 1991).

The most extensively studied prokaryotic regulatory proteins belong to bacteriophage λ and related phages (i.e. phage 434 and p22) which code for two regulatory proteins (Ptashne, 1992), namely the repressor itself (Pabo and Lewis, 1982) and Cro (Ohlendorf *et al.*, 1982; Anderson *et al.*, 1981). Such proteins provide a molecular mechanism for gene control in the form of a genetic on-off switch between the lytic and lysogenic life cycles. The genetic switch is contained in a small portion of the genome comprising the two structural genes that code for the two proteins, and the operator region on which they act. The genes are transcribed in opposite directions and Cro is switched on when RNA polimerase is bound to the right-hand promoter in turn switching on early lytic genes (Figure 1.2). *Vice versa* the λcI repressor is switched on when RNA polymerase interacts with the left-hand promoter, leading to the development of lysogeny during which lytic genes are repressed.

A



B

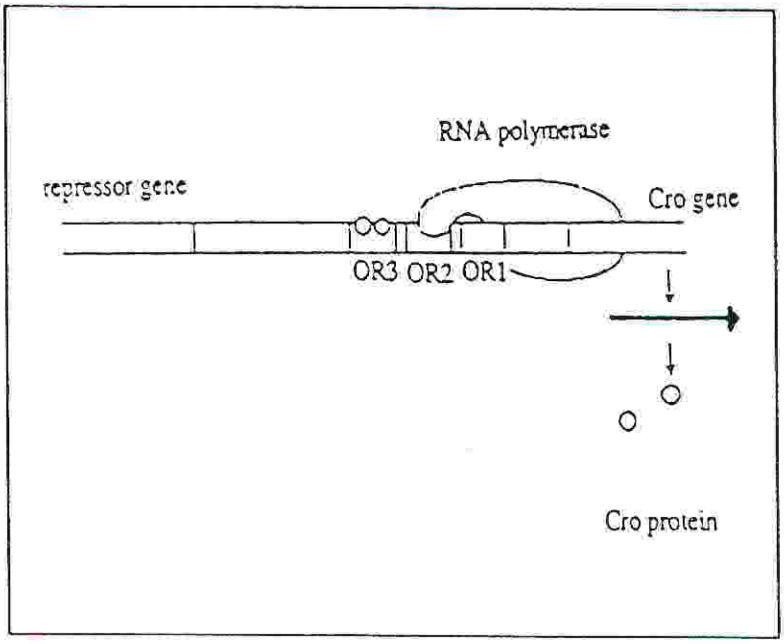


Figure 1.2. Schematic representation of the lytic to lysogeny switch. (A) Repressor proteins are shown as double spheres. They are bound in order to prevent transcription in the direction of the cro gene. (B) Cro proteins (single spheres) are bound and prevent transcription of the repressor gene.

Therefore the lytic-lysogeny decision is taken according to which of the two promoters within the operator is bound by RNA polymerase, this being dependent upon the binding of the Cro or repressor protein to three binding sites O_R1, O_R2 and O_R3 (each 14 base-pairs long) within the operator region. Such binding sites are palindromic sequences with a twofold axis of symmetry providing a dimeric repressor protein with two identical binding sites, one for each subunit. In addition to the operator regions (O_R) the repressor also controls a second set of operators (O_L) that regulate a different set of phage genes (Figure 1.3). It is significant that O_L also has three binding sites (O_L1, O_L2 and O_L3) each also 14 bp long (Ptashne, 1992).

It has been postulated, and partly proven, that since the three operator sites (O_R1, O_R2, O_R3) are separated from each other only by 6-7bp, both dimeric Cro and dimeric repressor proteins upon binding to one O_R site somehow affect the binding of another dimer to the adjacent operator sites (Johnson *et al.*, 1979). Thus when repressor binds O_R1, O_R2 is bound more avidly by a second repressor molecule. This cooperation in the binding however seems to occur only in the case of the entire protein (Johnson *et al.*, 1980) and does not unequivocally appear in the case of the single, DNA-binding, amino-terminal domain. For this reason one is more prone to think that cooperativity is mainly due to non covalent interactions involving residues in the carboxy-terminal domain of the entire protein. It has also been suggested by experimentally inactivating O_R1 through a mutation, that one repressor molecule bound at O_R2 can directly interact with another molecule in O_R3 and that in any case one bound dimer is able to make contacts only with one other dimer regardless of the operator. These observations can be extended both to the case of the right-handed operator region (O_R) and to that of the left-handed one (O_L) [Ptashne, 1992 and references therein].

The crystal structure of λCro complexed with its cognate DNA (Brennan *et al.*, 1990) shows a peculiar arrangement of α-helices (Figure 1.4) similar to that found

Operator sites

OR1	5'	A C A A G A A A G T T T G T	3'
	3'	T G T T C T T T C A A A C A	3'
OR2	5'	A C A A G A T A C A T T G T	3'
	3'	T G T T C T A T G T A A C A	3'
OR3	5'	A C A A G A A A A A C T G T	3'
	3'	T G T T C T T T T T G A C A	5'
OL1	5'	A C A A G G A A G A T T G T	3'
	3'	T G T T C C T T C T A A C T	5'
OL2	5'	A C A A T A A A T A T T G T	3'
	3'	T G T T G T T T A T A A C A	5'
OL3	5'	A C A A T G G A G T T T G T	3'
	3'	T G T T G C C T C A A A C A	5'

Synthetic operator

5'	A C A A T A T A T A T T G T	3'
3'	T G T T G T A T A T A A C A	5'

Figure 1.3. The six operator region which are bound by the 434 cI repressor and Cro protein (Also shown is a synthetic oligonucleotide, based on the operator regions, used in determining the repressor-DNA crystal structure).

in the λ repressor (Branden and Tooze, 1991; Brennan and Matthews, 1989; Jordan and Pabo, 1988). The catabolite gene-activating protein CAP also has a similar arrangement (Branden and Tooze, 1991; McKay and Steitz, 1981). Helix 2 and helix 3 of the λCro amino-terminal domain, form a helix-turn-helix motif. The domain responsible for dimerization is located at the carboxy-terminus of the protein and is separated from the DNA-binding region found in the amino-terminal portion of the protein, from which it can be removed by proteolytic cleavage.

The helix-turn-helix motif is used to bind specific sites on DNA and as already mentioned, utilizes one α-helix to contact the major groove along one face of the DNA double helix. In Cro, each recognition α-helix in the HTH motifs of the two subunits are at opposite ends of the elongated dimer, and are separated by a distance of 34 Å (Figure 1.4). This distance corresponds to one turn of a B-DNA double helix. Thus, the recognition helices bind into the major grooves separated by one turn of the DNA molecule. The second helix in each helix-turn-helix motif, separated from the recognition helix by a β-turn, lies across the major groove, making non specific contacts with DNA (Figure 1.4).

Thus four features of the Cro structure are important for specific binding to the DNA operator, (i) the presence of a HTH motif providing a correctly oriented recognition helix, (ii) the specific aminoacid sequence of this α-helix, (iii) the subunit interactions that provide the correct distance and relative orientation for the two recognition helices of the dimer, (iv) non-specific interactions between the DNA and protein surfaces, which increase the binding affinity. In some cases this requires bending of the DNA to maximize surface interactions. The α-helices forming the DNA binding motif can in fact be rich in basic residues such as lysine and arginine which, being positively charged, can directly interact with the negative phosphates of DNA.

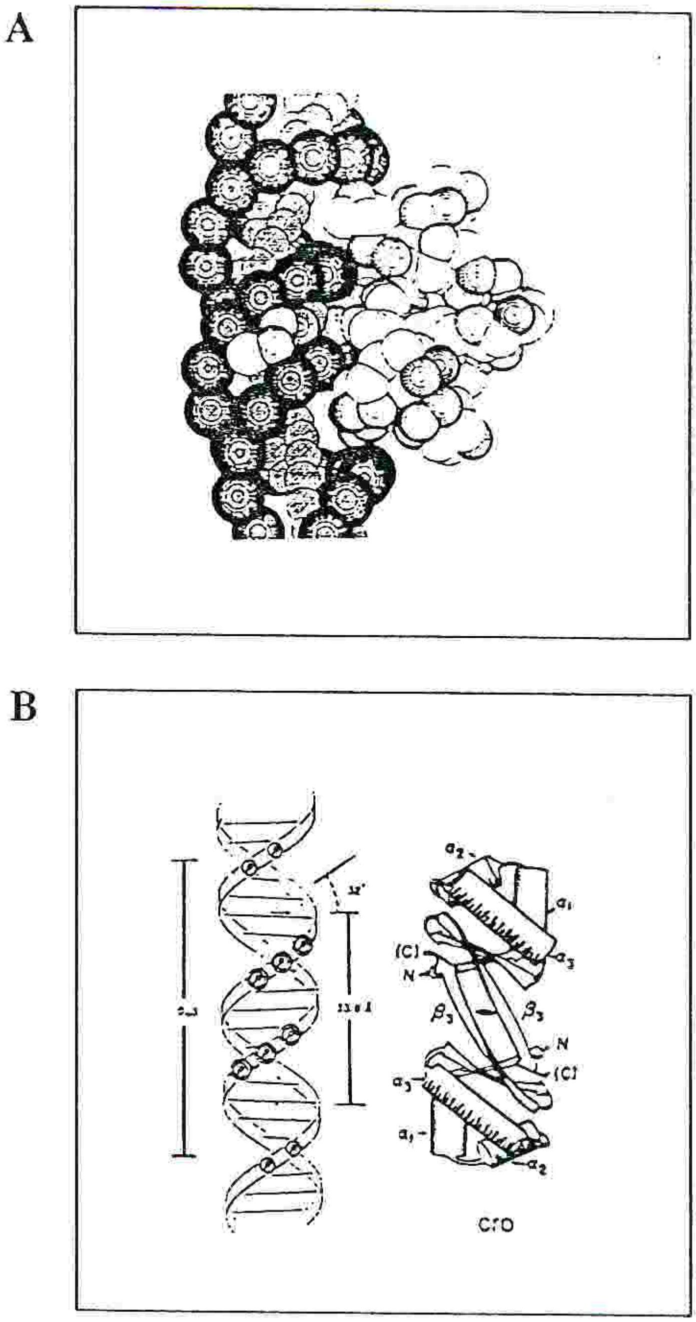
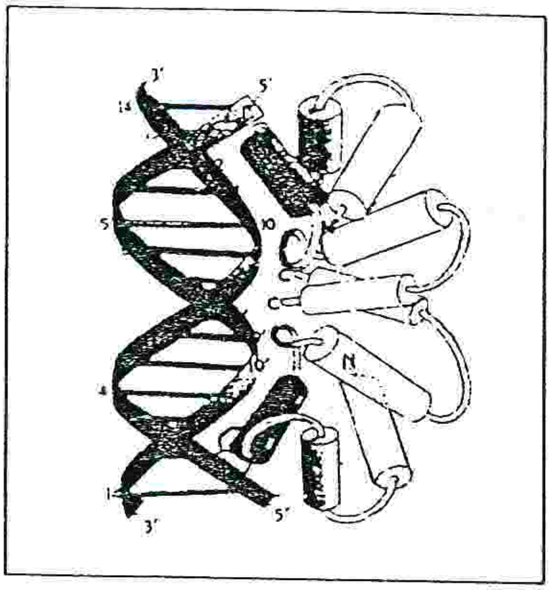


Figure 1.4. Crystal structure of the λ Cro-DNA complex. (A) Space-filling model of Cro. (B) Schematic representation of the Cro protein aligned to DNA.

Genetic studies (Ptashne, 1992) further show that the proposed structural model for DNA binding by λCro and repressor is essentially correct and can be applicable to the Cro proteins and repressors of phages 434 and P22. The amino-terminal domain of the repressor from phage 434, and its complex with the operator, has been crystallized and the structure determined (Figure 1.5 and Figure 3.1) [Aggarwal *et al.*, 1988; Anderson *et al.*, 1987]. It comprises a weak dimer of 69 aminoacid subunits complexed with a 14 bp oligonucleotide containing a palindromic sequence, and thus two identical half sites. Comparison of the amino-terminal domains of 434 Cro (71 aminoacids) [Mondragon *et al.*, 1989a; Wolberger *et al.*, 1988] and repressor (69 aminoacids) shows a homology rating of 48% in terms of aminoacid composition (Branden and Tooze, 1991 and references therein). The homology is higher in the case of the HTH DNA-binding motifs. A high degree of homology also occurs when comparing the aminoacid composition of the DNA-binding motif of 434 Cro and repressor with the corresponding λ phage proteins. Like its λ counterpart the amino-terminal subunit of the 434 repressor consists of five α-helices with H_2 and H_3 in the DNA-binding HTH motif, H_3 being the recognition helix. While the degree of homology between and 434 repressors is only moderate the structures are very similar. The backbones of the two repressors are in fact highly superimposable (Branden and Tooze, 1991). Two HTH motifs are at either ends of the dimer in the amino-terminal domain while the carboxy-terminal domain is responsible for dimerization. Interestingly it was shown that both 434 Cro and 434 repressor are monomers in solution whereas they form dimers only when bound to the cognate DNA (Branden and Tooze, 1991).

It has further been shown that both 434 Cro and repressor proteins induce DNA distortions which are related to the local twist, since DNA is overwound at the palindromic recognition sites and underwound at its ends, causing a narrowing of the minor groove at the center and a widening at the ends (Branden and Tooze, 1991).

A



B

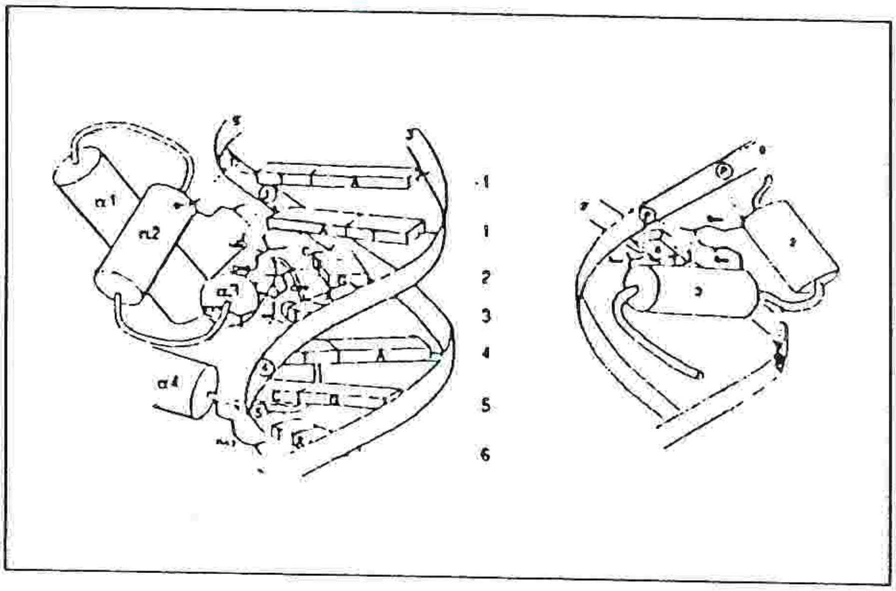


Figure 1.5. (A) Schematic view of the cI434 repressor bound to DNA. (B) Representation of the interactions of side-chains from residues (located both in the HTH motif and in the flanking helices) in cI434 repressor with DNA bases and backbone (Pabo and Sauer, 1992).

The helix-turn-helix motif has also been found in other proteins involved in gene regulation. Similar motifs have been discovered in proteins belonging to completely different and functionally unrelated organisms. It was found that the repressor from the 16-3 phage that infects *Rhizobium meliloti* is able to cross bind the 434 operator site from the repressor of phage 434 which infects *Escherichia coli* and *vice versa* (Dallmann *et al.*, 1987). The 16-3 and 434 operator sites consist of identical half sites separated by six non conserved nucleotides thus showing a related repressor specificity for unrelated phages. The 16-3 repressor shows only a moderate homology in its amino-terminal region to the 434 repressor, except for the HTH motif where it is high.

The tryptophane repressor (trp) and CAP are two DNA-binding proteins whose interaction with DNA is regulated by allosteric control i.e. small molecules whose binding causes conformational changes altering and thus regulating the DNA binding site and its affinity for DNA. The trp repressor controls the operon responsible for the biosynthesis of tryptophane which is regulated through negative feed-back. In fact in absence of L-tryptophane the repressor is inactive and the operon is switched on; in the presence of tryptophane the repressor is switched on and the operon is turned off. As the concentration of tryptophane increases it binds the repressor and induces a conformational change to the active form. The trp repressor has been crystallized (Otwinowski *et al.*, 1988; Shevitz *et al.*, 1985) and is similar to the λ repressor, apart from the arrangement of six amino-terminal α-helices (Branden and Tooze, 1991). The active form is a dimer in which the two monomers pack their helices so as to form a distinct overall strucure in which one can distinguish two heads and one core. The two HTH motifs responsible for the binding to the DNA reside in the two heads of the dimeric molecule. A conformational change occurring upon binding of two L-tryptophane molecules provides the molecular mechanism of the functional switch, the orientations of the recognition helices being altered into positions appropriate for binding (Figure 1.6).

CAP works as a positive control element that binds DNA and assists RNA polymerase in binding to certain promoters (Branden and Tooze, 1991). It becomes a specific DNA-binding protein upon binding of cyclic AMP. CAP controls a set of operons among which the lac operon, responsible for sugar breakdown. When the level of lactose breakdown is too low, the level of c-AMP is increased, activating CAP. Structurally speaking, CAP consists of an eight-stranded jelly roll β-barrel containing a pocket which supposedly acts as the binding site for c-AMP. The barrel-like structure is linked to the DNA-binding region containing the HTH motif. A dimer is formed in such a way that the two recognition α-helices of both HTH motifs are 34 Å apart from each other, in a similar arrangement as that observed for λ Cro.

Crystallographic data has been collected for a number of HTH type DNA-binding proteins. NMR data concerning solution structures is also available. For instance, the 434 repressor protein has been determined by NMR methods (Neri *et al.*, 1992a) and compared to its crystal structure (Mondragon *et al.*, 1989b). Both data are found to be in relatively good agreement especially in the case of the packing of hydrophobic residues in the interior of the protein. Interestingly NMR carried out on urea denatured 434 repressor protein revealed a region from residue 53 to residue 60 rich in hydrophobic aminoacids whose conformation is different to that in the folded protein and may thus serve as a nucleation site in the folding pathway of the protein (Neri *et al.*, 1992b).

The homeobox includes a 60 amino acid residue polypeptide, called the homeodomain, that represents the DNA binding domain of the respective proteins. Some structural features and sequence similarities are found in prokaryotic DNA binding proteins. Since homeodomain proteins have been discovered not only in metazoa but also in plants and fungi, it is assumed that they arose early during the

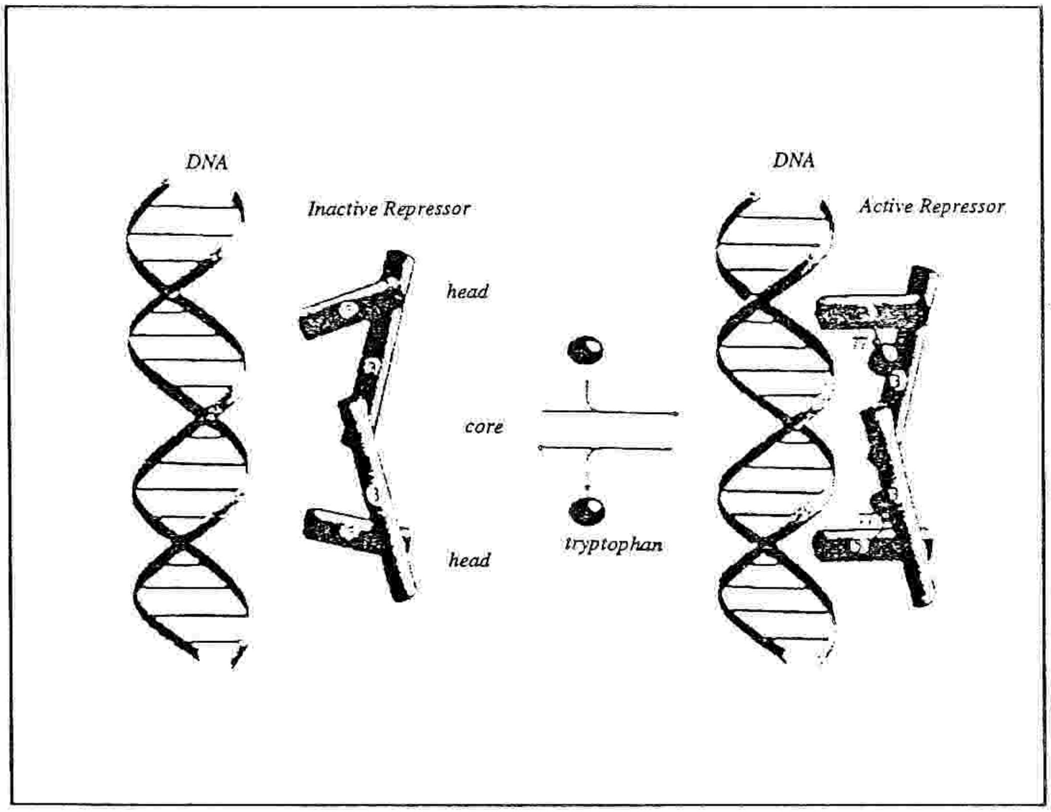
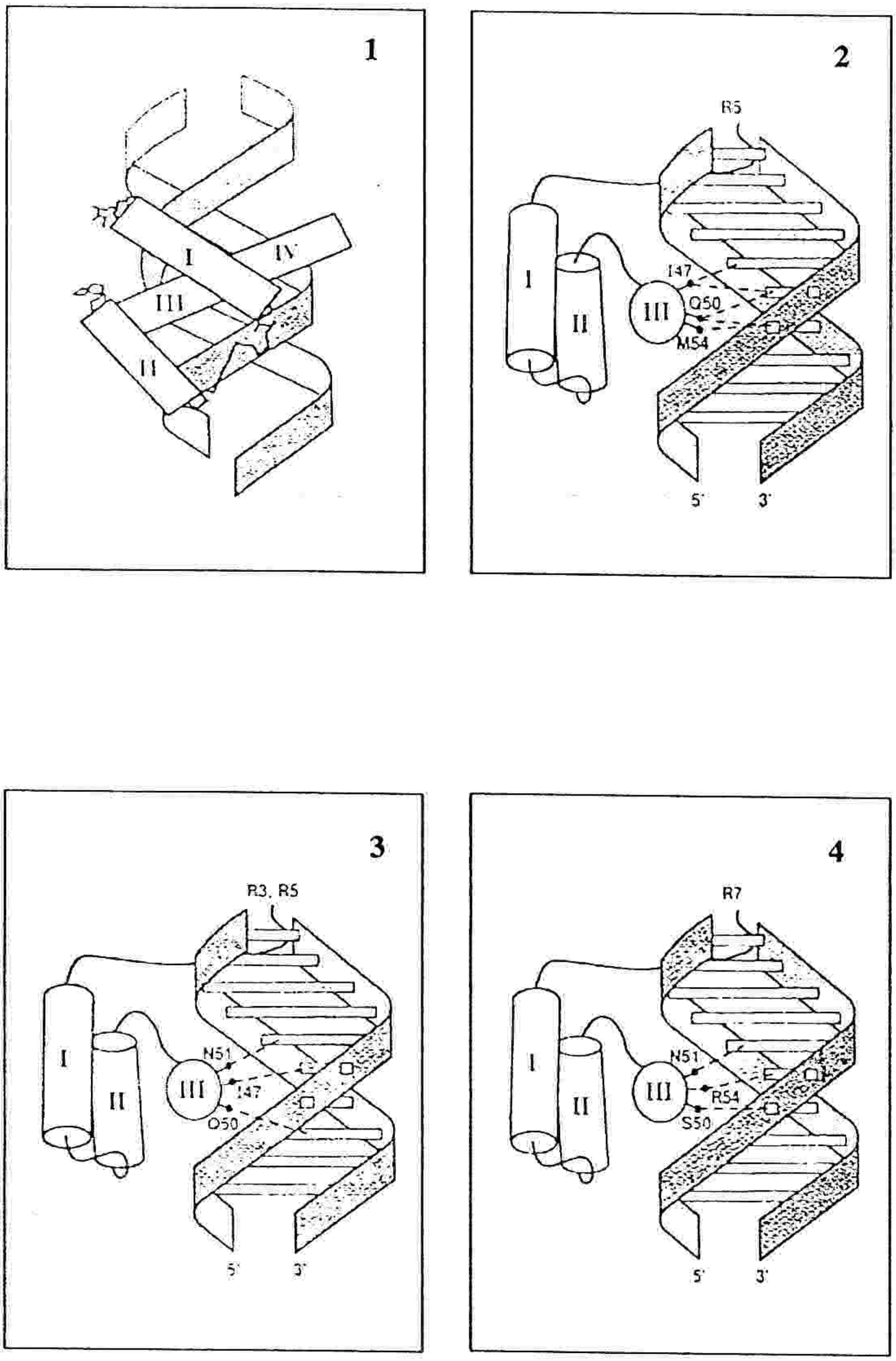


Figure 1.6. Schematic representation of the mechanism of action of the trp repressor upon DNA binding.

evolution of eukaryotes (Gerhing *et al.*, 1994). The amino acid sequence of the homeodomain has been conserved to a high degree (Wuthrich and Gerhing, 1992). For example, the human Hox-A7 homeodomain differs in only 1 out of 60 positions from that of the Antennapedia (Antp) homeodomain, which is putative homolog in Drosophila, even though vertebrates and insects separated more than 500 million years ago. This indicates that there is strong evolutionary pressure to preserve the amino acid sequence of the homeodomain (Gerhing *et al.*, 1994).

The yeast Matα2 is a transcriptional activator which binds to DNA through a homeodomain made up of three α-helices (ca. 60 residues, including the recognition helix) and an amino-terminal arm (Wolberger *et al.*, 1991). Such motif, essentially found in eukaryotic organisms, is structurally and functionally related to the helix-turn--helix motif mainly found in bacterial proteins. Studies on the Antennapedia homeodomain (Otting *et al.*, 1990) show that the long α-helix 3 is used for direct binding to the cognate DNA. The helix specifically recognizes the major groove of the DNA, whereas the long, random-like, amino-terminal arm wraps around the minor groove. The determination of the crystal structure of MATα2 at 2.7 Å resolution (Wolberger *et al.*, 1991) has shown that four side chains in the homeodomain (three of them belonging to residues in the recognition helix) provide specific contacts with DNA bases, whereas some eight additional side chains form contacts with the DNA backbone. Sequence comparison of MATα2 with the Drosophila engrailed homeodomain, and structural comparisons of the respective DNA complexes determined by X-ray crystallography (Kissinger *et al.*, 1990), show a high level of homology (rating up to 72%) and a strong structural similarity since the respective recognition helices are spatially positioned in an almost identical manner (Figure 1.7). The recent X-ray structure determination of an Oct-1 POU domain (Figure 1.8) bound to an octamer DNA (Klemm *et al.*, 1994) and its solution structure (Dekker *et al.*, 1993) together with the structure determination of an Oct-2 POU domain (Figure 1.8)

A



1
I
IV
III
II
2
R5
I
II
III
I47
Q50
M54
5'
3'
3
R3, R5
I
II
III
N51
I47
Q50
5'
3'
4
R7
I
II
III
N51
R54
S50
5'
3'

B

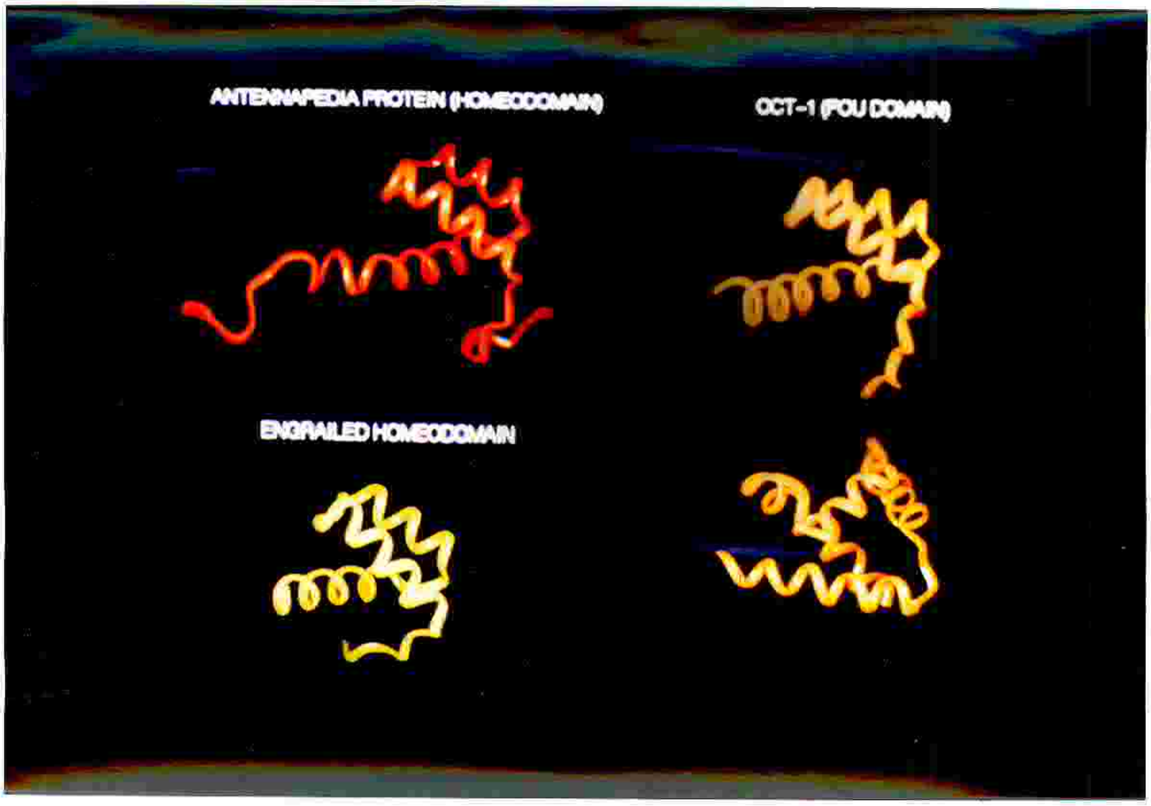


Figure 1.7. (A) Schematic comparison of the homeodomain-DNA complexes of Antp, en and Matα2. (1) Antp. View perpendicular to the axis of the recognition helix (III, IV) located in the major groove. The α-helices I, II and III, IV are indicated by cylinders. (2) Antp; (3) en; (4) Matα2. View along the axis of the recognition α-helix (III). (B) Ribbon representation of homeodomain DNA-binding proteins Antp, en and Oct-1 (Gehring *et al.*, 1994).

in solution (Sivaraja *et al.*, 1994) indicate that the docking of the POU homeodomain is quite similar to that of Antp and Matα2.

In particular the Oct-2 POU homeodomain consists of a bipartite structure with an amino terminal subdomain POUs and a carboxy terminal divergent homeodomain POU$_{HD}$, which are both necessary for high affinity DNA binding, even though the two domains fold independently (Botfield *et al.*, 1992). In fact, the isolated POUs subdomains exhibit only weak specific binding activity (Verrijzer and van der Vliet, 1993). The structure of POUs has been shown to be similar to that of phage repressors (Assa-Munt *et al.*, 1993; Dekker *et al.*, 1993) which bind DNA via a helix-turn-helix motif with a third α-helix flanking the recognition helix (HTH).

In conclusion, the HTH motif is a well-defined supersecondary structural element that has convenient geometric features for DNA recognition. However the remainder of the protein in which the motif is located can vary significantly among different repressors. This suggests that dimerization interfaces can be very different from each other even among functionally related proteins. Dimerization is in fact carried out in a number of ways, including nearly parallel α-helices (CAP), antiparallel α-helices (cI), antiparallel β-sheet (cro repressor) or an intertwined arrangement of helices (trp repressor) [Harrison and Aggarwal, 1990; Dodd and Egan, 1990].

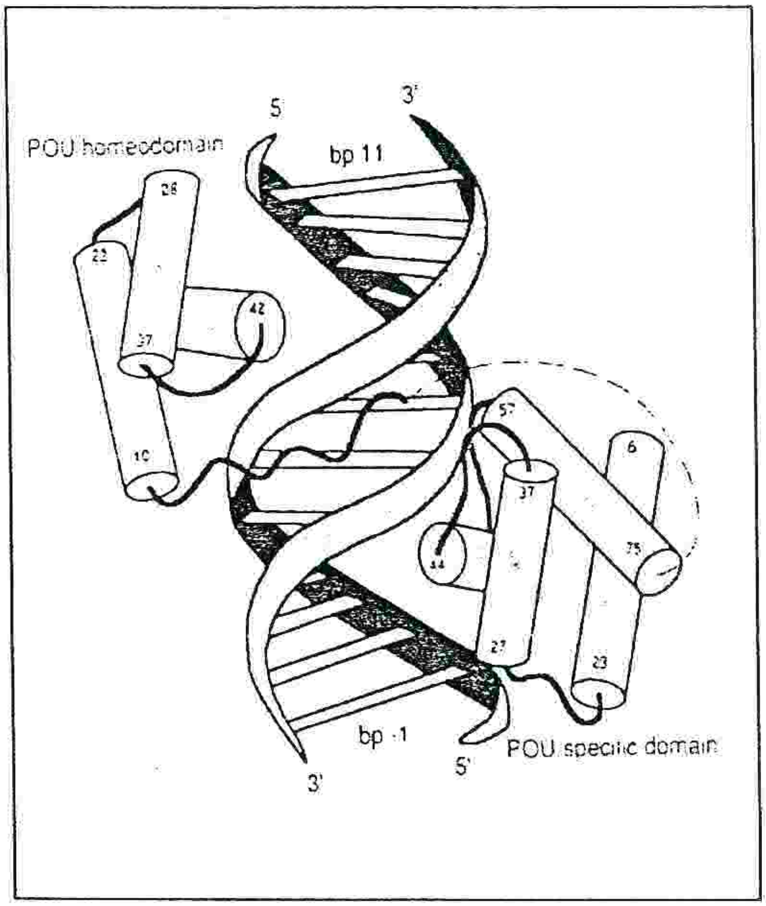


Figure 1.8. Schematic representation of the POU domain-DNA complex for Oct-1 (Klemm *et al.*, 1994).

1.3 PRINCIPLES OF PROTEIN-DNA RECOGNITION

Interaction with Bases

A crucial role in specific protein-DNA interaction is played by specific base contacts. This can occur in different ways including (i) direct hydrogen bond formation between DNA bases and aminoacid residues, (ii) occasional hydrogen bond formation between the polypeptide backbone and the DNA bases, (iii) hydrogen bonds mediated by water molecules and (iv) hydrophobic contacts (Pabo and Sauer, 1992).

Most of these contacts which determine the binding specificity are made with the major groove (although some proteins recognize sequences in the minor groove) essentially because it is larger and more accessible in B-DNA and has more potential sites for hydrogen bond formation and hydrophobic interactions. However, it also appears from known examples of protein-DNA complexes that there is no simple recognition code (i.e. no one-to-one correspondance between side chains and DNA bases) which may help generalizing the nature of specific side chain-base contacts (Pabo and Sauer, 1992 and references therein). These contacts result extremely diverse even if closely related DNA binding proteins are considered. A possible explanation could be that even when the same secondary structure, such as an α-helix, is used for base recognition, the precise position and orientation of the α-helix may determine what set of interactions are possible for a particular residue. Further level of complexity in determining site-specific protein-DNA interaction is due to the presence of (i) side chain-side chain interactions, (ii) residues contacting more than one base pair and (iii) contacts involving water molecules or divalent cations.

However for closely related proteins belonging to the same family or subfamily some organizing principles emerge. Examples are provided by the λ and 434 repressor (Pabo and Sauer, 1992 and references therein) and by the engrailed (en) and a2 homeodomains (Wolberger *et al.*, 1991; Kissinger *et al*, 1990) which respectively belong to the same protein families characterized by common folding structural motifs (Figure 1.7). In both cases conserved residues at the protein-DNA interface make conserved contacts with the DNA. This suggests that a given family of DNA binding proteins may have, in addition to the conserved folding motif, a conserved "docking mechanism" and a conserved set of contacts. Therefore, the position and orientation of the polypeptide backbone with respect to the DNA determine the meaning of a particular side chain (Pabo C.O., 1984) which can be presented to the major groove in different ways by different structural motifs. Since structure is central to recognition both the folding and overall arrangement help to determine which base contacts are possible.

Interactions with the DNA backbone

Also the DNA backbone plays an integral role in site-specific recognition. In the known complexes, approximately half of all the hydrogen bonds involve contacts with the phosphodiester oxigen of the DNA backbone (Jordan and Pabo, 1988; Aggarwal *et al.*, 1988; Otwinowski *et al.*, 1988). However, also in this case there is no simple rule or pattern describing which residues are used to make contacts. In general, it seems likely that most of these contacts are made via any basic or neutral amino acid residues even though both aspartic and glutamic acid residues can contact the DNA backbone through divalent cations such as Mg^{2+} and Ca^{2+} (Pabo and Sauer, 1992 and references therein).

The exact role of DNA backbone contacts is not really understood, however two major roles appear possible: (i) Backbone contacts can serve as "fiducial marks" that help hold the protein against the bases in a fixed arrangement and thereby enhance the specificity of the side chain-base interactions. The protein can therefore be oriented in the major groove and establish a fixed register. (ii) Sequence-dependant variations in the DNA structure may also be important including flexibility and bendability of DNA (Pabo and Sauer, 1992).

The role of DNA conformation in protein-DNA recognition

There are several ways in which the local structure of DNA can influence specificity. (i) Sequence-dependent influences can occur on the structure of a given site on the DNA molecule inducing deviations from the canonical B-DNA form (Otwinowski *et al.*, 1988; Dickerson and Drew, 1981), including intrinsically bent or kinked DNA which somehow enhance the binding specificity. (ii) Sequence-dependent effects on the DNA flexibility can also contribute to the interaction specificity if the protein distorts the DNA upon binding possibly by bending. In fact, several proteins are known to significantly bend DNA, including the Trp repressor (Otwinowski *et al.*, 1988), 434 repressor (Aggarwal *et al.*, 1988), 434 Cro protein (Mondragon *et al.*, 1991), λ repressor (Jordan and Pabo, 1988) and λ Cro protein (Brennan *et al.*, 1990) which all belong to the helix-turn-helix family (see section 1.2) of proteins. In each case the DNA bends as if it were going to wrap around the DNA binding helix-turn-helix motif. An even more dramatic DNA distortion is observed for CAP which apparently causes a 90 degree bend on the DNA upon binding (Schulz *et al.*, 1991). However, even when it is possible to detect DNA structure distortion due to protein binding, it is still difficult to evaluate the contribution of bending to the binding specificity as it is difficult to determine the average structure of the free DNA and the energetic cost of bending the DNA.

A detailed analysis of backbone contacts raises several related problems about recognition including the structural relationship between specific and non-specific binding. The main question when addressing this problem is whether DNA binding proteins make the same backbone contacts while interacting with non-specific DNA.

One of the limitation to the understanding of this problem is due to lack of structural information of aspecific protein-DNA complexes. Luisi *et al.* (Luisi *et al.*, 1991) have crystallized two different structures of the glucocorticoid receptor DNA binding domain complexed with DNA. When the DNA binding domain binds to DNA it dimerizes placing the subunits in adjacent major grooves. In one complex, the DNA has the symmetrical consensus target sequence; in the second, the central spacing between the target's half sites is larger by one base pair. This results in one subunit interacting specifically with the consensus half site and the other non specifically with a non cognate element. The DNA induced dimer fixes the separation of the subunits' recognition surfaces so that the spacing between the half sites becomes a critical feature of the target sequence's identity. This is one of the few known examples of an aspecific protein-DNA complex. However, since biological specificity only results from the free energy differences between specific and non specific binding, the real nature of site-specific binding can not be understood without understanding the nature of non specific interactions. An attractive work was recently carried out by Johnson *et al.* (Johnson *et al.*, 1994) in which the nature of non specific binding was investigated by circular dichroism. In fact, in order to understand the thermodynamic and kinetic bases of specific binding by two different transcription factors (a leucine zipper and a helix-loop-helix protein), they investigated whether their flexible DNA binding motifs can be driven into an α-helical conformation when the protein is bound at non specific DNA sites. They show that small peptides containing important features of the proteins DNA binding domains do indeed have the potential to form α-helices by interacting non specifically with double-stranded DNA, suggesting that helical DNA binding motifs may be generally present in non specific as well as specific complexes between such transcription factors and DNA. The induction of α-helical conformation both at non specific DNA sites may be thermodinamically important for specificity of binding as it can help stabilize the specific protein-DNA complex by decreasing the excess of free energy due to the coil to helix transition. The presence of the same conformation both at specific and aspecific sites may also account for a mechanism of binding in which the regulatory protein slides along the DNA from aspecific to specific sites, docks and recognizes its own regulatory sequence.

It is therefore clear that site-specific protein-DNA recognition is the result of several and significantly different contributions for which no simple rules or pattern of rules can be worked out. This is also due to the broad number of DNA binding protein families which interact with their DNA sites via different structural motifs as reviewed in AppendixII. However it appears that in general aspecific interactions are necessary to stabilize site-specific protein-DNA recognition.

The role of α-helices in protein-DNA recognition

Most of the DNA binding proteins that have been characterized so far use α-helices (although β-sheets or regions of extended polypeptide chain can also make contacts) to make contacts in the major groove of the DNA. However there is no example reported in the literature in which a protein regulates a gene or a set of genes by binding to DNA through an isolated helical peptide (i.e. a single α-helix). It is more likely that the overall binding specificity in protein-DNA recognition is the result of a set of interactions also involving regions of the polypeptide chain which are proximal to the recognition helix (Pabo and Sauer, 1992).

The orientation of the recognition α-helix also appears to be important in determining the specificity of binding even though it varies quite a lot according to the family of DNA binding proteins. The overall shape and dimensions of an α-helix allow it to fit into the major groove in a number of ways which are significantly different from each other. Helices can lie in the middle of the major groove, they can be tipped at different angles or they can be arranged so that only the amino terminal portion contacts the DNA. However regions of the protein proximal to the recognition α-helix apparently play a role in accomodating it into the major groove.

Thermodynamic principles of protein DNA recognition

Macromolecules, like proteins and DNA are quite flexible in solution, i.e. they can be best pictured as being distributed over a wide range of conformational states having similar energies. When two such molecules associate, the structure of the stable complex will be more rigid than any of the previous free states, i.e. the interaction will stabilize a small subset of the originally possible conformational states. The experimenter who can observe the average conformation of the free partners will thus see a conformational change. This mechanism is termed "induced fit" (Koshland, 1958) or "induced conformational complementarity" ("handshake paradigm", Colman, 1988). In contrast, rigid molecules have an *a priori* restricted freedom in solution so they assemble without appreciable conformational change. This is the "key-lock" paradigm, described mainly for small molecular ligands (Amit *et al.*, 1986). Protein-protein associations are frequently accompanied by conformational changes, and are, in many cases, driven by hydrophobic effects. The meaning of this term is that a hydrophobic surface of the free molecule gets into an energetically more favourable buried state upon complex formation. Naturally, in addition to the hydrophobic effect, hydrogen bonds, ionic links, vand der Waals interactions all play a role in stabilizing macromolecular complexes.

The first systematic treatment of the thermodynamics of protein-DNA association was that of Spolar and Record (Spolar and Record Jr., 1994). These authors reviewed the energetics of protein-DNA complexes based mostly on calorimetric data, and applied an approximate formula to determine the number of residues that are folded in the protein molecule upon interaction with DNA. They came to the conclusion, that protein-DNA complexes fall into two large categories. In most of them, a large number of protein residues takes part in the folding, and these numbers are more or less in agreement with the number of residues that are found buried in the protein-DNA complexes. In another group of proteins, namely the HTH-based repressor proteins, the number of residues estimated from calorimetry was either few or zero, so the authors concluded, in agreement with previous suggestions (von Hippel and Berg, 1986), that HTH domains dock as rigid bodies. In addition, it was concluded that conformational changes are predominantly induced by specific binding, and are absent or less pronounced in aspecific binding. Summarizing, it appears that protein-DNA associations are driven by forces very similar to those undrlying protein-protein associations (Johnson *et al.*, 1994). The hydrophobic effect seems to play a role, and the resulting conformational change may be in most cases similar to the induced fit-like association of proteins (Alber, 1993; Frankel and Kim, 1991). A notable exception to this rule seems to be the HTH-containing repressors that are considered to associate with DNA as rigid bodies.

1.4 CD ANALYSIS OF PROTEIN-DNA COMPLEXES

Secondary structure determination from CD data

Proteins and protein-DNA complexes can be studied with different spectroscopic techniques which can give several and significantly different information, but nevertheless complementary to each other.

Circular dichroism (CD) and optical rotatory dispersion (ORD) are two optical phenomena that are based on the asymmetryc nature of biological molecules which are able to rotate the plane of light polarization due to the presence of centres of chirality (carbon atoms). ORD is the property of a molecule to rotate the plane of linearly polarized light as a function of the wavelength. CD gives information about the unequal distribution of left and right-handed circularly polarized light. A combination of the two components gives elliptically polarized light. Both optical phenomena are therefore able to differentiate two enantiomers in solution (Cantor and Shimmel, 1980; Yang *et al.*, 1978; Schmid, 1989).

Since the conformation of a macromolecule affects its optical activity, these properties can be utilized to obtain information about the structural organization of such molecules in solution. Crick and Kendrew in 1957 stated that "there is an encouraging parallelism between the X-ray and optical results". In fact, CD, which has now replaced ORD in the conformational analysis of proteins, provides characteristic bands which differentiate various secondary structure elements within the protein. We are therefore able to detect α-helices, β-sheets and turns against random structures and investigate the transitions that occur in the presence of various inducing agents or in acqueous buffer provided that the protein is in a low salt milieu.

CD bands of proteins occur in two spectral regions. The far UV or amide region (170-250 nm) is dominated by contributions of the peptide bonds, whereas CD bands in the near UV region (250-300 nm) originate from the aromatic amino acids

(Chakrabarty *et al.*, 1993; Yang *et al.*, 1978; Schmid, 1989). The two spectral regions give different types of information about protein structure.

The signals observed in the amide region provide information about the peptide bond and secondary structure of a protein. They are often used to monitor structural transitions due to changes in secondary structures and subdomain folding (Lumb *et al.*, 1994) and also to investigate structural stability of proteins (Thompson *et al.*, 1993). Optical activity in the region of the spectrum between 190 nm and 230 nm is essentially dominated by the protein backbone. A number of experiments has shown that the nature of particularly aliphatic side-chains does not markedly affect the CD spectrum in this region. Therefore, in the first approximation one can consider a protein as a linear combination of backbone regions with α-helical, β-sheets or random coil structures.

The α-helix shows a strong and characteristic CD spectrum in the far UV region. The signal is detected as a typical double minimum at 222 nm and 208 nm and a maximum at 191-193 nm (Figure 1.9). The spectral contributions of the other structural elements are less well defined. In 1966 the CD spectrum of Pauling and Corey's β-pleated sheet (of β-form) was found for $(Lys)_n$ in sodium dodecyl sulfate solution (Sarkar and Doty, 1966), $(Lys)_n$ at alkaline pH (Townend *et al.*, 1966) and silk fibroin in methanol-water (Iizuka and Yang, 1966). It shows a negative band near 216-218 nm and a positive one between 195 and 200 nm, but the intensities of these three bands are different in all three cases. The CD spectrum of the unordered form generally shows a strong negative band near 200 nm and a very weak band around 220 nm. The amplitude of this strong negative band can vary according to the polypeptide and the environmental conditions (Figure 1.9).

For proteins, therefore, the major objective is to deduce average secondary structures by CD studies. There are at least three established methods to determine protein secondary structure from CD data. A semiempirical approach was derived by

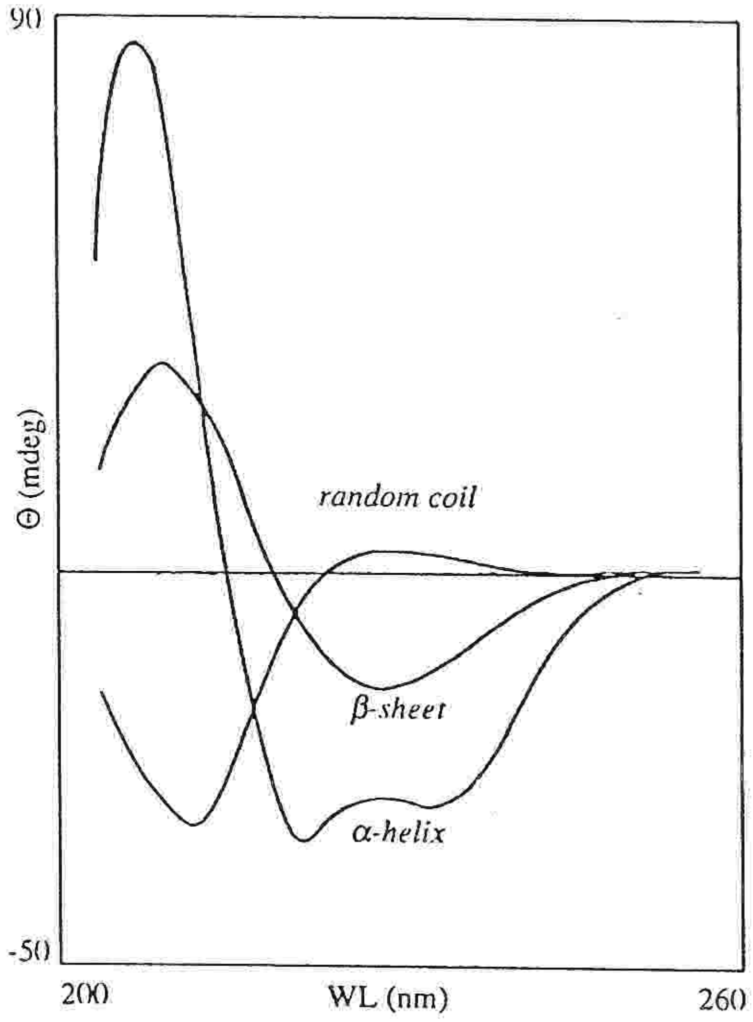


Figure 1.9. Schematic representation of CD spectra for α-helix, β-sheet and random coil conformations

Wu (Wu *et al.*, 1981) and makes use of known structures of synthetic polypeptides as reference spectra. An alternative approach is to compute reference spectra from CD spectra of proteins of known secondary structure (Provencer, 1984). The third method concerns the direct analysis of a protein as a linear combination of the CD spectra of proteins of known secondary structure, thus avoiding the problem of defining reference spectra of individual conformations (Schmid, 1989; see references citated therein).

The semiempirical method (Wu *et al.*, 1981), in particular, is based on the theoretical values of mean residue ellipticity, $[\Theta]_{MRW}$ (based on the concentration of the sum of the amino acids in the protein solution under investigation), calculated for an infinite α-helix and an infinite random coil. This method gives a rough idea about the α-helical content of a given protein, especially if the protein under investigation is made essentially of α-helices.

CD of protein-DNA complexes

Several protein-DNA complexes have been studied by circular dichroism as (i) the technique is very sensitive (a measured ellipticity of 10 mdeg is equivalent to a ΔA of only 0.0003 absorbance units) and (ii) it allows to perform difference spectroscopy. As for protein-DNA complexes, the underlying principle is that frequently the protein (or also the monomeric chromophore if dimerization studies are carried out by CD) has a small but not negligible CD when compared with the complex or DNA alone. In this case it is therefore more convenient to compute the difference spectrum in order to determine the average secondary structure of the protein in the presence of the DNA. A comparison of the difference CD spectrum with that of the protein alone indicates whether structural transitions have occurred in the presence of the inducing agent.

Several examples of structural transitions occurring in DNA binding proteins in the presence of cognate DNA have been already reported in the literature. Most of the studies were carried out on bLZIP/bHLH (reviewed by Ellemberger, 1994) proteins (Auer *et al.*, 1994; Ferrè-D'Amarè *et al.*, 1994; Johnson *et al.*, 1994) or hybrid polypeptides in which different structural motifs are fused together. The spectrum of information that can be retrieved is quite broad and in fact it definitely concerns the intrinsic nature of protein-DNA recognition, including for example the sequence requirements for coiled coils (Hu *et al.*, 1990) and α-helical induction of proteins in the presence of their cognate DNA sequences (O'Neal *et al.*, 1991). Recently, a study of the altered protein conformation upon DNA binding by the Jun and Fos heterodimer has been reported (Patel *et al.*, 1990) together with an investigation of the folding transition in the DNA binding domain of GCN4 on specific DNA binding (Weiss *et al.*, 1990). In most cases, however, short basic peptides (Talanian *et al.*, 1992; Talanian *et al.*, 1990; O'Neal *et al.*, 1990) are used to investigate the induction of α-helices upon addition of the specific DNA in solution. Short basic peptides, in fact, provide optimal probes to study such transitions for several reasons: (i) they are more soluble in low ionic strength buffers, (ii) they bind DNA more easily and, in general, (iii) they are prone to undergo structural transitions in the presence of a number of inducing agents since they may exhibit simpler folding behaviour. In an attempt to explore how the bZIP class of transcriptional activator binds to DNA, O'Neal (O'Neal *et al.*, 1990) applied a "minimalist approach" to protein design to produce two short basic peptides identical to native leucine zipper proteins only at positions hypothesized to be critical for dimerization and DNA recognition. The peptides form dimers that bind specifically to DNA with their basic regions in α-helical conformations. Short intrinsically flexible peptide domains also derived from the bZIP and bHLH families have been designed and used to investigate the nonspecific interactions that underlie the principles of specific DNA recognition (Johnson *et al.*, 1994).

However, most of the CD studies on protein-DNA complexes were performed on eukaryotic transcription factors mainly because they bind the DNA through structural motifs which in general appear more flexible as compared to prokaryotic ones. In fact, very little examples if any at all are reported in the literature concerning CD structure determination in solution of helix-turn-helix containing transcription factors. An attempt at structure determination in solution of a helix-turn-helix containing bacterial repressor was carried out by Shirakawa *et al.* (Shirakawa *et al.*, 1991). They produced an intersubunit disulfide-bonded λCro protein by site-directed mutagenesis. CD studies performed on such mutant protein essentially indicate that it mantains its α-helical conformation in acqueous buffer, whereas NMR data indicates that the protein keeps its correct α-helical folding upon binding to the DNA (O_R3 site). An unsuccesful example has also been reported by Grokhovsky *et al.* (Grokhovsky *et al.*, 1991) of an attempt at designing helix-turn-helix containing peptides which however do not fold into an α-helical structure.

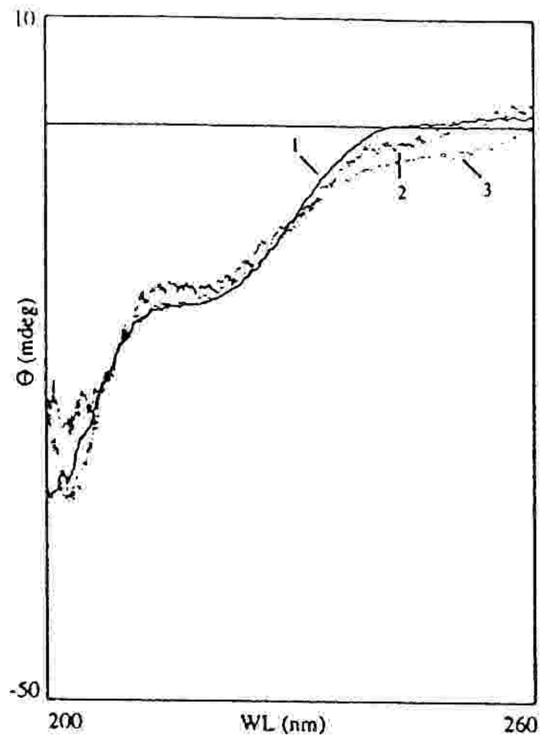
Therefore, the idea of producing a conformational probe to monitor coil-to-helix transitions in solution upon specific DNA binding of a HTH protein is quite attractive because not much is known and moreover it is a difficult system to study in solution due to its intrinsic structural rigidity. But it is even more attractive to consider the hypothesis that such conformational probe can also be used to monitor coil-to-helix transitions that occur in the presence of non-specific DNA. This could help understanding the intrinsic mechanism of site-specific recognition of HTH containing proteins. For this purpose, excellent candidate molecules are given by bacterial repressors most of which are essentially α-helical, provided though that their helix-turn-helix motif involved in the specific binding be not a rigid structure as postulated by Johnson *et al.* (1994).

2 AIM OF THE PROJECT

The long term aim of this project is the further understanding of protein-DNA recognition through *de novo* designed molecules that are capable of specific DNA binding. This is essentially a *design approach to protein-DNA interactions* which is based on the hypothesis that individual structural elements involved in protein-DNA complexes can be built into artificial molecules and their function studied in virtual isolation. The ultimate goal of these studies is the construction of DNA binding proteins in which the specificity is varied at will. Such proteins, apart from providing more insights into the nature of site-specific protein-DNA recognition, may have a variety of practical applications including *in vivo* repression of undesired genes and carriers for specific chemical DNA-cleaving agents.

The immediate aim of this project is the construction of artificial DNA-binding proteins containing HTH motifs, using the bacteriophage 434 cI repressor N-terminal domain as a model system (Mondragon *et al.*, 1989b; Aggarwal *et al.*, 1988). We set out to design proteins that show specific DNA-binding activity *in vitro*, using gel-mobility shift assay and CD difference spectroscopy as tests. Gel mobility shift assays were performed in the presence of a roughly 1000 fold excess of competitor DNA to show specifc DNA binding. CD difference spectroscopy was used to monitor conformational changes induced by DNA binding. When assayed with these standard techniques, the N-terminal domain alone (produced by continuous flow solid phase peptide synthesis) does not show any sign of DNA-binding (Figure 2.1). Neither does its conformation change upon addition of cognate or non-cognate DNA sequences (Figure 2.1 A), nor is there any gel-mobility shift characteristic of protein-DNA association (Figure 2.1 B), which is essentially in keeping with its known weak DNA-binding activity (Ptashne, 1992; Branden and Tooze, 1991 and references therein).

A



B

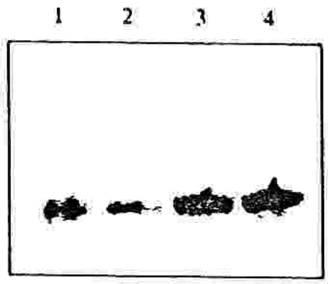


Figure 2.1. (A) CD difference spectroscopy performed with the complete amino-terminal domain (termed P1) in the presence of either cognate [434O_R1 (20 bp)] or non-cognate (CL/WL) DNA. P1 (7.2 µM) alone (line 1), P1 with cognate DNA (line 2), P1 with non-cognate DNA (line 3). (B) Gel mobility shift assay performed with the complete amino-terminal domain (P1) in the presence of either cognate [434O_R1 (20 bp)] or non-cognate (CL/WL) DNA. Lanes 1 and 2 refer to experiments carried out with cognate DNA: lane 1, no protein; lane 2, P1 (final concentration 0.5 µM). Lanes 3 and 4 refer to experiments carried out with non-cognate DNA: lane 1, no protein; lane 2, P1 (final concentration 0.5 µM). In all cases binding reactions were performed with 1000 fold excess of poly[d(I-C)].

We speculated that dimer formation, a salient feature of most natural DNA-binding proteins, might provide a way to achieve high specificity. We thus decided to build covalently dimerized molecules that we termed *single-chain repressors*. These are analogous to bipartite eukaryotic DNA-binding proteins (eg. Oct-1 and Oct-2 POU domains; see Figure 1.3) except that we use two bacterial domains to build them. Therefore, there are two identical and potentially specific binding domains rather than two domains with different specificities as for example is the case with the POU homeodomains (Klemm *et al.*, 1994; Sivaraja *et al.*, 1994; Dekker *et al.*, 1993). In principle, such a single-chain repressor can incorporate either the entire cI434 N-terminal domain, or parts thereof. Such molecules can be built with a complete dyad symmetry, as branched peptides, and produced with peptide synthesis. Alternatively, one can combine the various structural elements present in the native repressor in recombinant molecules as direct repeats. We decided to try a number of these strategies and to test the resulting single chain repressor analogs *in vitro*.

A concomitant goal of our work was to clarify whether HTH domains are really as rigid as they appear from thermodynamic and theoretical studies (Spolar and Record Jr., 1994; von Hippel and Berg, 1986), or the helical conformation is at least partially induced by DNA-binding. At present, bacterial HTH domains are the only type of DNA-binding structures that are supposed to bind through rigid-body docking, rather than through conformational adaptation, or "induced fit" (Alber, 1993; Frankel and Kim, 1991; Koshland, 1958). It has also been suggested, that the "induced fit" mechanism underlying the formation of protein-DNA complexes is due to the specific interactions only, and interactions with non-cognate DNA do not apparently produce such conformational transitions (Spolar and Record Jr., 1994 and references therein). Since the cI434 N-terminal domain is predominantly α-helical, we speculated that it could be a sensitive conformational probe for assaying weak interactions with non-specific DNA.

3. DEVELOPING MODEL PEPTIDES BASED ON THE HTH MOTIF

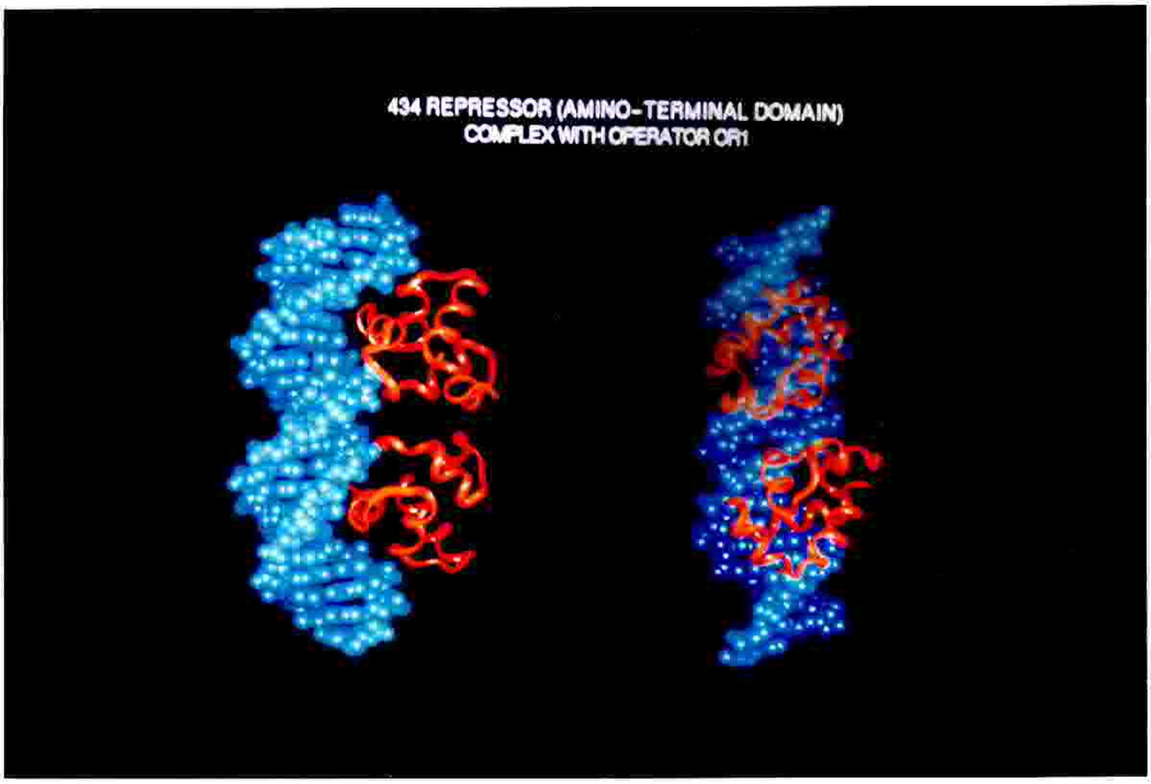
The design is based on the analysis of the 3D structure, with special regard to DNA and intra-domain residue contacts (sections 3.1-2) and to the conservation of the structurally important residues (section 3.3). The importance of individual parts within the N-terminal domain is approached with a "minimalist" strategy, i.e. designing covalently dimerized peptides that contain the HTH motif and its flanking elements (section 3.4). The importance of the dimerization is studied by designing covalent dimers of the whole N-terminal domain that are either dyad-symmetric or contain direct repeats in their sequence (section 3.5). As all of these peptides contain the repressor DNA-binding elements in a single peptide chain, we term them single-chain repressors.

3.1 ANALYSIS OF THE 3D STRUCTURE OF THE cI434 N-TERMINAL DOMAIN

Apart from known residues of the HTH motif, the loop flanking the recognition helix 3 and residues within helix 1 make contacts with DNA.

Complexes between both the cI434 repressor amino-terminal domain and the λ repressor amino-terminal domain and their cognate DNAs were available for modelling. Sequence comparison of 434 repressor and λ repressor highlighted a reasonable level of homology, rating up to 48% in the case of those residues located within the helix-turn-helix binding motif. Structural similarity between the λ repressor and the 434 repressor was tested by superimposition and found to be very high (results not shown). This finding is important, as it indicates that those residues which are conserved between the two repressors in their amino-terminal subunits may play a structurally important role. The X-ray structure of the cI434 repressor amino-terminal domain bound to DNA is shown in Figure 3.1A. Visualization of the molecular model

A



B

|----H_1----|----|--H_2--|-T-|---H_3---|-L->
1 20 40
SISSRVKSKRIQLGLNQAELAQKVGTTQQSIEQLENGKTK

>-L--|--H_4--|---|--H_5--|-----|
60
RPRFLPELASALGVSVDWLLNGTSDSNVR

Figure 3.1. (**A**) Crystal structure (ribbon type representation) of the cI434 repressor protein bound to its cognate DNA. (**B**) Amino acid sequence of the complete amino-terminal domain and position of structural units (amino acid residues 1 to 69).

was carried out with the Insight II programme in the Biosym package, on a Silicon Graphics Power Plus workstation. This programme allows one to reduce the protein to a simplified form, such as the ribbon model, and then introduce the side chains of each residue in the protein, one by one, so as to directly observe their interaction with DNA or with other residues. A schematic representation of the amino-terminal domain and its sequence is shown in Figure 3.1 B.

The following aspects of the interactions of the cI434 amino-terminal domain with its cognate DNA were analysed: (i) the interaction of the binding helix H_3 with the DNA bases in the binding site, (ii) the effect of the stabilizing helix H_2, (iii) interactions of residues outside the HTH motif with the DNA (including the formation of a complementary surface between the DNA and the protein, which requires conformational changes in both), (iv) interactions between residues in the amino-terminal domain and (v) interactions between the two amino-terminal subunits leading to dimerization. The reviewes by Harrison and Aggarwal (Harrison and Aggarwal, 1990) and Brennan (Brennan, 1993) were of considerable help in this respect.

The recognition helix H_3.

The recognition helix consists of residues 28 to 36. Some of these residues project their side-chains into the major groove of the cognate DNA in order to provide non-covalent bonding with the edges of base-pairs (Figure 1.5). There are six distinct binding sites for the repressor, the three operators O_R1, O_R2 and O_R3 that control the lysogeny/lytic decision (Figure 1.2) and three operator sites involved in the regulation of other phage genes (O_L1 to O_L3). Residues Gln28 and Gln29 are strongly involved in specific hydrogen bond formation. In particular, if the palindrome is given by the sequence 5'-ACAA$(dNu)_6$TTGT-3' (where dNu represents any deoxyribonucleotide) on strand 1, Gln28 interacts with N6 and N7 of the first adenine located at the 5' end whereas Gln29 forms hydrogen bonds with the O6 and N7 of the second guanine residue located at the 3' end of the sequence 3'-TGTT$(dNTP)_6$AACA-5' on strand 2. Interestingly a sort of hydrophobic pocket is formed by the methyl groups of the side chains of Thr27 and Gln29 which is probably important because it accomodates the methyl group of the third T residue on strand 2 of the DNA binding site (Branden and Tooze, 1991).

It is important to observe that the first three residues on the three operators (O_R1, O_R2 and O_R3) recognized by the 434 repressor protein, are identical as one can see in Figure 1.3. This is evidence that interactions between Gln28 and Gln29 with these first three base-pairs cannot be the device that contributes to the discrimination among the various binding sites on the DNA molecule. Rather, one could think of them as providing the general recognition of cI434 for all the operator regions. The pattern of hydrogen bonds and hydrophobic interactions therefore accounts for the recognition of the operator region both by the 434 Cro and the 434 repressor proteins. The role of Gln29 results of particular importance because it interacts with two different base-pairs, namely C-G and A-T, in two different ways, by hydrogen bond formation and through hydrophobic interactions respectively (Branden and Tooze, 1991). Interestingly it has been experimentally proven that replacement of glutamines 28 and 29 produces mutant phages which are no longer viable (Branden and Tooze, 1991 and references therein). Other residues in H_3 which on visual inspection come close to the DNA bases and may be involved in binding are Gln33 and Asn36.

The stabilizing helix H_2.

The helix H_2 lies orthogonally across and behind the binding helix H_3 (with respect to DNA, see Figures 1.5 and Figure 3.1A) and is thought to stabilize it. From the molecular model it is apparent that residue Gln17 comes into contact with residues Glu32 and Glu35 in H_3, and residue Leu20 comes into contact with Ile31 in H_3, as well as hydrophobic residues in the loop L and helix H_4, forming a hydrophobic pocket (Figure 3.2). Also apparent are interactions between Asn16 and G1 and T2 in strand 2 of the DNA operator site (see Figure 1.3 for the sequence of the DNA) and Gln17 with T2 and A3 of the same strand and A1 of strand 1.

Interactions of residues outside the HTH motif with DNA.

Several residues outside the HTH binding motif are involved in nonspecific binding to DNA, via either hydrophobic interactions with DNA bases or electrostatic interactions with the DNA backbone.

The loop L following the binding helix H_3 contains two lysines (positions 38 and 40) and two arginines (positions 41 and 43) which are placed relatively close to the DNA phosphodiester backbone (Figure 3.3), and which may be involved in electrostatic interactions with the phosphate groups. Lys38 from one subunit comes into contact with the DNA in proximity to the bases C4 (strand 2, Figure 1.3) while Lys38 from the other subunit comes into contact with the DNA in the proximity of A10 (strand 1). Regarding the arginines, one subunit points its Arg43 directly at the DNA minor groove, where it comes into contact with the base T10 in strand 1 and T9, T10 and T11 in the other strand of the cognate DNA, while that in the other subunit comes into contact with A8 and G9 in strand 1 and T10 in strand 2. Arg41 points upwards towards the DNA phosphodiester backbone in one subunit, but in the other it points away from the DNA (Figure 3.3 and Figure 3.4).

The formation of preferential complementary surfaces between the DNA molecule itself and the protein are strongly favoured by the ability of the DNA to undergo structural changes, which in turn are favoured by the establishment of non-specific interactions between proteins and the sugar-phosphate backbone. In the case of the cI434 repressor, the structural change essentially involves the bending of DNA and overwinding of the middle regions of the operator (Figure 3.1A). When this does

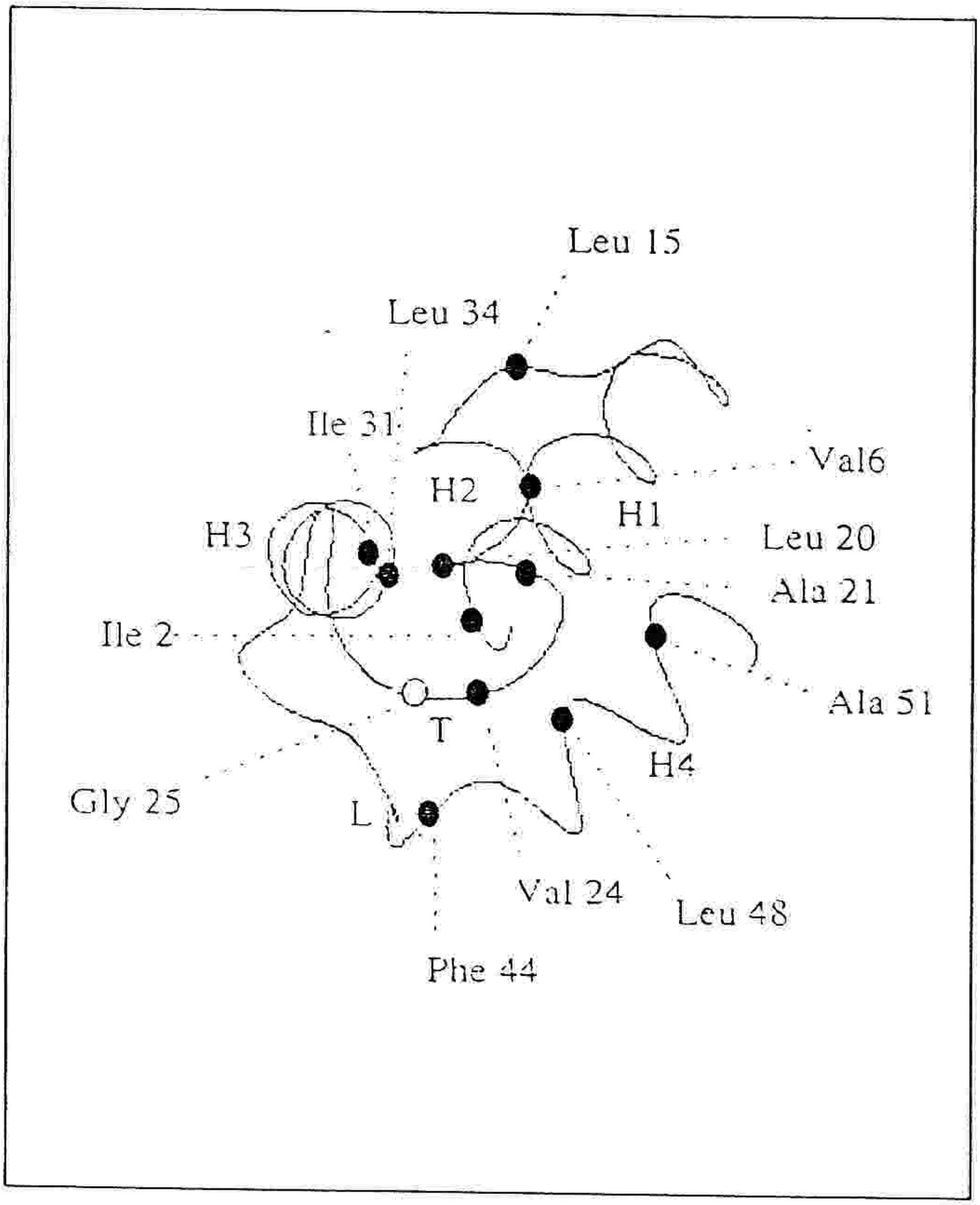


Figure 3.2. Schematic representation of the positions of hydrophobic residues in the hydrophobic core formed by the helices H_1, H_2, H_3 and H_4, the turn T and the loop L in the cI434 repressor amino-terminal domain. Helix H_5 is omitted for clarity. Hydrophobic residues are shown as filled circles, while the highly conserved residues Gly25 (located in the turn) as an open circle.

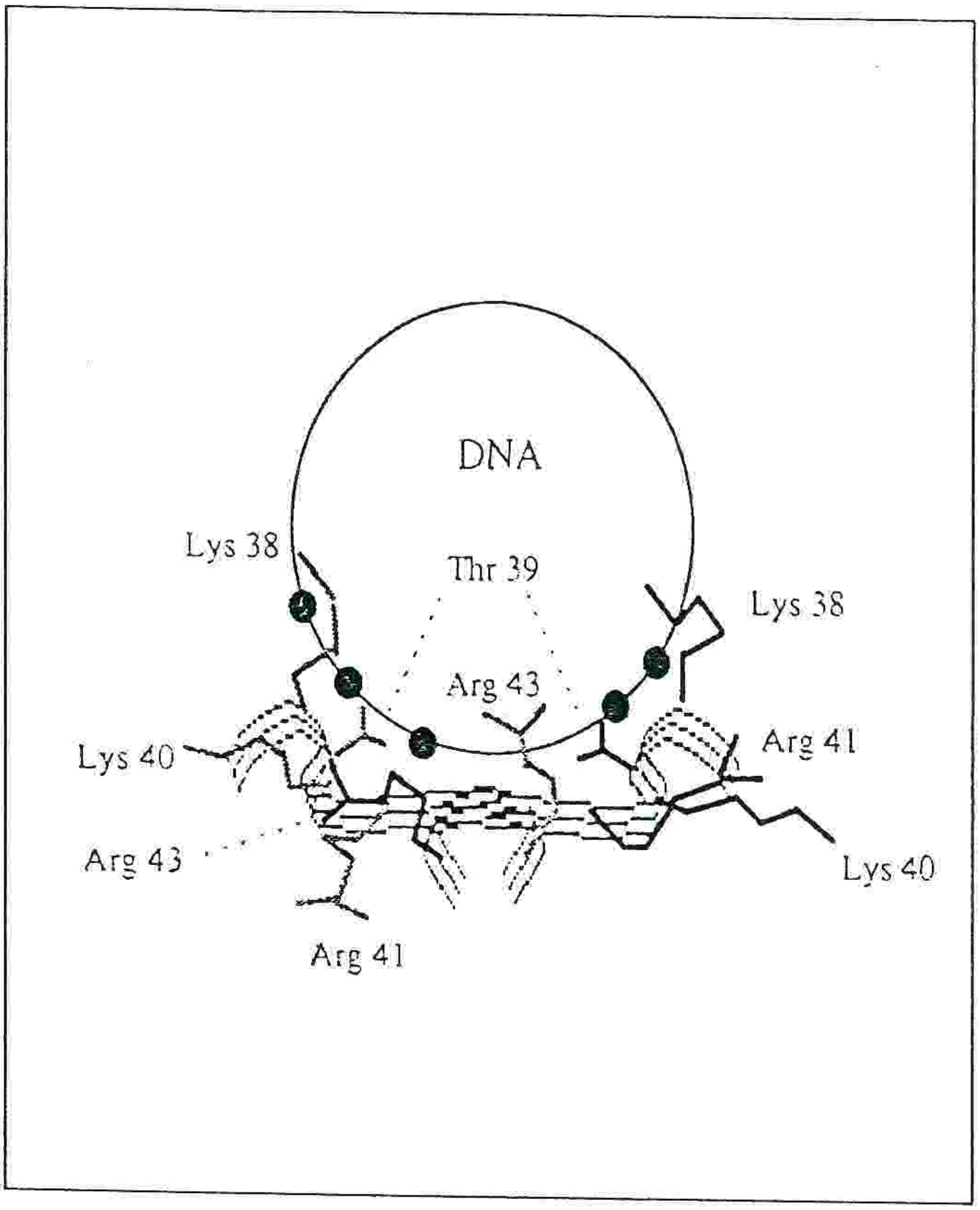


Figure 3.3. Schematic representation of the interactions of side-chains from residues in the loop (L) region of the cI434 amino-terminal domain with DNA. The diagram is based on a computer visualization of the X-ray structure. The DNA is viewed along the major axis and phosphate groups in the general vicinity of the loop are shown as filled circles. Side-chains from one subunit are in grey, those from the other one are in black.

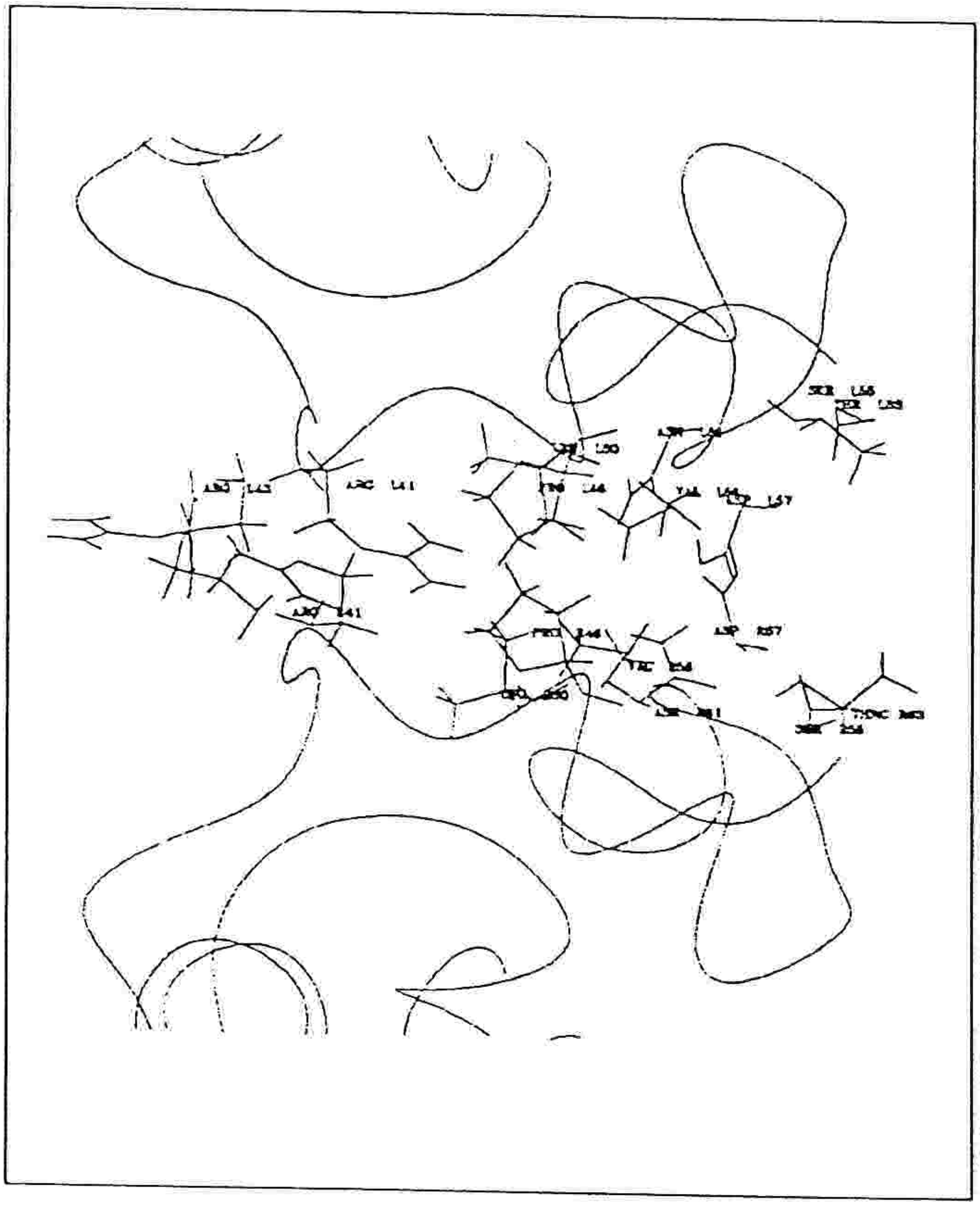


Figure 3.4. Side-chains of residues at the interface between the two subunits in the cI434 amino-terminal domain.

not occur, as in the interaction of the 434 repressor to O_R3, binding is weak. The protein-DNA complementary surface is quite extended and since each monomer of the repressor protein can see only a half operator, the DNA molecule has to undergo the abovementioned structural changes so as to coincide with the dimeric organization of the protein itself and to accomodate the protein in the best manner. The arginine in position 43 projects into the minor groove and provides a positive charge that shields the DNA phosphate groups, thus facilitating the compression of the minor groove. This is a typical non-specific interaction with a multiple role, namely that of (i) directly stabilizing interaction with DNA via electrostatic interaction, (ii) stabilizing the change in DNA conformation and (iii) as a result of the latter effect, increasing the complementary surface between the DNA and the protein dimer.

The arginine at position 10, and thus in the helix H_1 in both subunits also points at the DNA phosphodiester backbone, and may be providing an electrostatic anchoring interaction. Another residue that contacts the DNA is Asn16.

Interactions between residues in the N-terminal domain.

The amino-terminal domain is a compact structure consisting of 5 α-helices connected by turns or loops (Figure 3.1). Hydrophobic interactions at the core of the subunit are important for the stability of the structure and its correct folding. One could detect these interactions by looking at the computer model, but a more efficient method involving a distance matrix is described in section 3.2. Modeling was useful in validating the indications given by the distance matrix, and the results of these studies are described in detail in that section.

Interactions between subunits.

Dimerization in the cI434 repressor is mediated by the carboxy-terminal domain. In its absence, a weak dimer is formed in the presence of DNA on crystallization but does not survive in solution. Examining the molecular model it was found that the residues in the loop L come into close contact with residues in the loop L of the other subunit and with two residues (positions 56 and 60) in the helix H_5 and *vice versa* (Figure 3.4). From the molecular model it was possible to establish that the point of closest approach for the two subunits is at residue Pro46 with $C\alpha$ to $C\alpha$ distance of 9.9Å and $C\gamma$ to $C\gamma$ distance of only 5.85Å.

3.2 ANALYSIS OF INTRA-DOMAIN RESIDUE CONTACTS

An analysis of the inter-residue distance matrix reveals stabilizing interactions between the HTH motif and helix1 and helix5. The hydrophobic residues of the loop between helix3 and helix4 provide a small hydrophobic core in the interior of the domain that might be important for stabilization.

It is logical to suppose that sequentially distant residues within the cI434 repressor amino-terminal subunit that come spatially close to each other may be acting to stabilize the overall structure, by means of non-covalent interactions. It is not possible to efficiently individuate all such interactions by means of a visual inspection of the molecular model. The modelling programme however allows a listing of all interactions as a side product of its energy minimization routine. By definition, the calculation of interaction energies, as used in minimization, requires a measuring of all interatomic distances within the molecule. These are stored in a data file, and it is relatively simple to select only certain interactions (say Cα to Cα) and filter out all those above a cutoff distance (say 7Å). The result of this process is best summarized in a distance matrix (Figure 3.5). Spatially vicinal but sequentially distant residues fall above the diagonal (values on the diagonal can be disregarded as they represent the distance of a Cα atom from itself). Below the diagonal are represented residues from different subunits that come within 7Å.

As far as the HTH motif is concerned, results from the distance matrix confirm that there are a number of interactions between H_2 and H_3, but it also points to several interactions between H_2 and the turn T. What is immediately apparent from the matrix is that a number of interactions occur between the loop L after H_3 with helix H_3 itself and the turn T. Returning to the molecular model, it was in fact found that hydrophobic residues (Val24, Ile31 and Phe44) in T, H_3 and L are part of a hydrophobic pocket, to which Leu48 and Ala51 from the beginning of helix H_4 also contribute.

A number of interactions are also indicated between the binding helix H_3 with helix H_1. In particular, hydrophobic interactions are indicated for Ile2 and Val6 in H_1 with Leu34 in H_3 (Figure 3.2), and electrostatic interactions for Arg10 and Glu35.

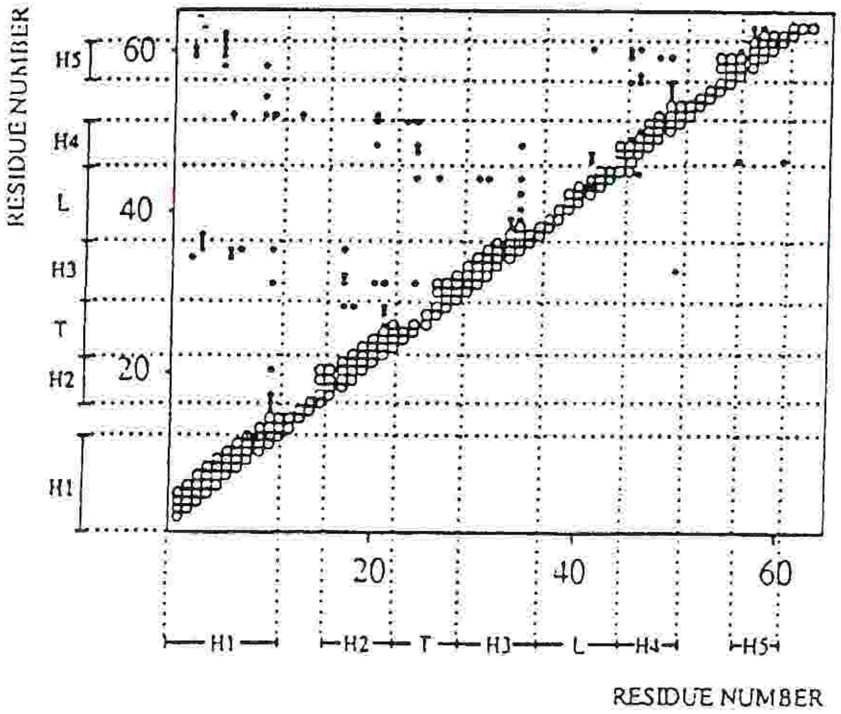


Figure 3.5. Distance matrix of residues in the cI434 amino-terminal domain (7 Å cutoff). The grid divides into its helix (H_1 to H_5), turn (T) and loop (L) secondary structure elements. The DNA binding motif is H_2TH_3.

3.3 CONSERVATION OF STRUCTURALLY IMPORTANT RESIDUES

Comparison of related HTH motifs confirm that residues involved in DNA-contacts and in intradomain contacts are highly conserved.

A data base search was carried out on the Swiss-prot database for all bacterial phage repressors, which structurally resemble the cI434. The aminoacid sequence of the helix-turn-helix motif, flanked by 10-20 residues on each side, was then abstracted from the complete sequences of each repressor, and the Intelligenetics programme Genalign used to perform an alignment. It was thus possible to determine those aminoacid residues that are conserved. As this homology study was carried out in aid of structural determinations, we were concerned with the conservation of physical characteristics, such as hydrophobicity, aromaticity or charge, rather than with exact identities.

The alignment of the cI434 amino-terminal domain sequence with those from other phage repressors is shown in Figure 3.6. Directly conserved residues, or hydrophobic residues that are conserved as such are in bold characters, while conserved charged residues are in italics. It can be seen from the alignment, that hydrophobic residues are conserved at positions 13, 15, 20, 21, 24, 25, 31, 34 and 44/45 (with respect to numbering in the cI434 amino-terminal domain). Inspection of the molecular model and distance matrix for cI434 shows that the leucine in the loop before helix H_2 at position 13 does not appear to interact with other residues, while that at position 15 interacts with the leucine at position 20 in helix H_2. This in turn forms part of a hydrophobic pocket with Val24 in the turn (T), Ile31 in the binding helix H_3 and hydrophobic residues in helix H_4. The alanine at position 21 is highly conserved among all repressors and is probably required in this position to allow packing of the helices H_2 and H_3 (Branden and Tooze, 1991). A glycine at position 25 is also highly conserved and probably is required to allow formation of a tight turn at T. Residue Val24, apart from the abovementioned interactions, also comes into contact with the Phe at position 44, as do Ile31 and Leu34 in the helix H_3. Thus, the conservation of hydrophobic residues in these positions is explained by the need to pack the protein correctly and stabilize it via the formation of a hydrophobic core.

A cluster of conserved basic residues is apparent around position 10. From the model and distance matrix it was found that the arginine in this position contacts the DNA and it appears to be involved in electrostatic interactions and H-bonding to the bases (Harrison and Aggarwal, 1990). This residue is at the end of helix H_1, pointing towards the DNA, and probably has similar role in different repressors. At the other end of the HTH motif, another cluster of conserved basic residues appears in the loop L at around position 40. Conservation is not so high in this case and it is possible that van der Waals complementarity of the loop with the DNA phosphodiester backbone (Harrison and Aggarwal, 1990) is more important than electrostatic interactions. An arginine at position 43 is present only in the 434 repressor and Cro, so it is possible that its role is that of mediating the conformational changes required of the cognate DNA for this particular system.

Also highly conserved are glutamines at positions 17 and 28. In the first case, the residue comes at the begining of the helix H_2, pointing towards DNA and is probably acting in a similar manner as Arg 10. Gln 28 at the beginning of the binding helix is involved in recognition of the binding site for all repressors. Asn 16, which also contacts DNA in cI434, is instead peculiar to this repressor.

Thus, a homology study of different repressors combined with structural studies can give valuable information as to those residues which play either an important role in the recognition and binding of DNA, or in maintaining the stability of the DNA-binding region in the protein. Such studies have also been carried out in an attempt to derive rules for identifying probable HTH motifs in proteins (Brennan, 1993; Branden and Tooze, 1991; Harrison and Aggarwal, 1990; Dodd and Egan, 1990; Brennan and Matthews, 1989).

```
          <---H1---|-------|------H2------|--T-|---------H3-------|-------L-------|-H4->
             10                  20                  30                  40
              |                   |                   |                   |
λ Cro     L K D Y A M R F G Q T K T A K D L G V Y Q S A I N K I H A G R K I F L T I N A
λ Rp      K K K N E L G L S Q E S V A D K M G M G Q S G V G A L L F N G I N A L N A Y N
P22 Cro   K D V I D H F G T Q R A.V A K A L G I S D A A V S Q W K E V I P E K D A Y R L
P22 Rp    A R R K K L K I R Q A A.L G K M V G V S N V A I S Q W Q R S E T E P N G N E L
434 Cro   K R R I A L K M T Q T E.L A T K A G V K Q Q S I Q L I E A G V T K R P R F L F
434 Rp    S K R I Q L G L N Q A E.L A Q K V G T T Q Q S I E Q L E N G K T K R P R F L P
16-3RP    A K L E E A G I T Q A E L A R R V G Q S Q Q A I N N L F A A V R R R Q W F A R
```

Figure 3.6. Sequence alignment of helix-turn-helix motifs found in DNA binding phage repressors (RP) and Cro proteins.

3.4 DESIGN OF DIMERIC PEPTIDES BASED ON THE HTH MOTIF. A MINIMALIST APPROACH.

Peptides are designed that incorporate the HTH motif and its flanking elements into symmetric molecules dimerized via their C-termini with a lysine linker. The peptides include H_2TH_3, H_2TH_3L, H_3L dimerized, H_2TH_3L dimerized, $H_1H_2TH_3L$ dimerized (number of the helices given in subscript).

Having obtained structural information as to which residues are important for binding to DNA, and which are important for stabilizing the structure, one can start to design a peptide which contains the minimum amount of structural features that will allow it to function. Reducing the length of the viable peptide to a minumum is desirable as there is a limit to the length of peptides that may be easily and inexpensively chemically synthesized in good yields. At this point we only consider dimeric peptides, since dimerization is a crucial characteristic of all known repressors.

The following features must be incorporated into the peptide: (i) The HTH motif, (ii) sufficient residues flanking this motif, on either side, to promote folding into the correct structure for binding, (iii) a number of residues apart from the binding helix which will interact with the DNA and improve the affinity of binding, (iv) some means of introducing stable dimerization to give correctly oriented binding helices and (v) from practical considerations, a balanced number of hydrophobic and charged residues should be present, so as to allow the peptide to dissolve and fold in acqueous solution.

Considering the information obtained from the molecular model, the distance matrix and the homology study, it was decided to begin the peptide at Pro46. This is the point of closest aproach for the subunits in the repressor (Figure 3.7), and the distance between the two proline residues in different subunits (about 10Å) is spannable by a lysine linking the carboxylic groups of the two proline residues with its

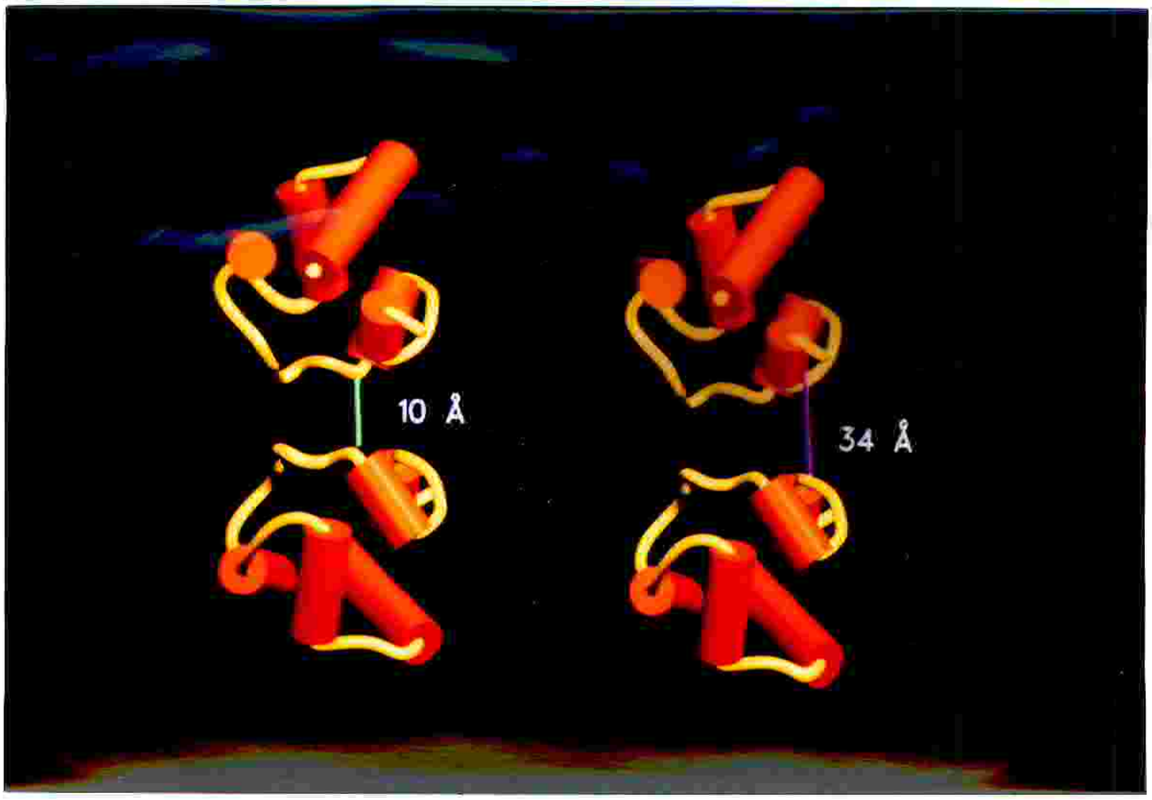


Figure 3.7. Cα-Cα distances measured (A) at the point of closest approach (Pro46 in one monomer to Pro46 in the other monomer) between two complete amino-terminal domains (~ 10 Å) and (B) between their C-termini (Gly62 on one monomer to Gly62 in the other monomer, ~ 35 Å).

α and ε amino groups. Thus a way of introducing a permanent form of dimerization is possible. Continuing the synthesis towards the N-terminus of the repressor, one introduces the loop L and then H_3, the turn T and H_2. It was decided initially to terminate the synthesis at this point. Then, since from modeling studies helix H_1 appears to be involved both in DNA binding (aspecific interactions with the DNA backbone) and protein stabilization, we decided to continue the synthesis to yield a peptide with a higher level of structural complexity.

Thus the syntheses for both branched peptides would require 30 and 42 coupling reactions, including an initial branched lysine, which is well within the capacity of an automated synthesizer, and assuming a reasonable coupling yield of 98% per reaction, would lead to a final yield of 54%. The resulting peptides are termed H_2TH_3Ldim and $H_1H_2TH_3L$dim respectively (Figure 3.8). H_2TH_3Ldim would contain two branches, each with 30 residues, of which 22 neutral or hydrophobic, 5 basic and 3 acidic. Similarly, $H_1H_2TH_3L$dim has two branches both with 42 residues, of which 34 neutral or hydrophobic, 9 basic and 3 acidic.

Before synthesizing the peptides, a modelling study was carried out to test whether they were viable. The molecular model for the cI434 repressor N-terminal domain (Brookhaven entry PDB2OR1, Figure 3.1), consisting of two protein subunits each of 69 residues, bound to a 20 base-pair cognate DNA fragment, was taken and 40 counterions (Na^+) added so as to shield the DNA phosphate groups. The counterions were placed manually roughly in proximity of each phosphate, and allowed to take up a correct position by use of the molecular dynamics programme Discover (10000 cycles), with the dielectric set to water. This was carried out so as to mimic the real environment of the DNA-protein complex more closely. The residues from 47 to 69 in each subunit were then excised and the C-terminal carboxylic groups reconstructed. A lysine was then bound via its α amino group to the C-terminus of the bottom subunit at Pro46 and a push/pull force applied between the ε amino group and the C-terminus of the top subunit (also Pro46) with equilibrium set at a 2 Å distance. The model was then subjected to 10000 cycles of molecular dynamics (with all atoms except those of the last 10 residues in the bottom subunit fixed), and the two groups brought within 2 Å of each other. The programme Insight was then used to create a bond between the ε amino group and the C-terminus of the top subunit, but due to the characteristics of the programme, this dislodges the bottom subunit from the DNA, so as to give a physically correct bond. Editing of the molecular coordinates file, so as to replace back the correct coordinates for a DNA bound subunit then resulted in DNA-bound, covalently linked subunits, but with a physically incorrect bond between the bridging lysines and the top subunit. Energy minimization was then carried out to correct the bond.

At this stage a model for the H_2TH_3Ldim peptide, bound to DNA, had been obtained, but the question remained as to whether this structure was energetically acceptable. All residues except for the bridging lysine and ten residues on each side were then fixed and the structure subjected to 10000 cycles of molecular dynamics. The resulting structure did not show a significant change in this region beyond the first two or three residues flanking the lysine (Figure 3.9), indicating that the dimer bound to DNA is an energetically acceptable structure.

This last step ended the process of designing dimeric peptides based on the HTH motif, namely H_3Ldim, H_2TH_3Ldim, $H_1H_2TH_3$Ldim and with a potential for binding to DNA (Figure 3.9). The next step required the syntheses of the peptides and testing. The possibility of analysing and redesigning a structure in detail by computer greatly increases the confidence with which this next step was approached, considering the large expenditure in time, effort and resources that it would entail. Attempts have in fact already been made at constructing DNA binding peptides based on the HTH motif of 434 Cro. Grokhovsky *et al.* have recently linked this HTH motif to a multiple lysine core. Their four-stranded peptides, with substitutions rendering the HTH more

A

```
1                   20                  40                  60
|                   |                   |                   |
SISSRVKSKRIQLGLNQAELAQKVGTTQQSIEQLENGKTKRPRFLPELASALGVSVDWLLNGT
|-------H1-----|---H2---|-T-|---H3---|---L---|----H4---|----H5--->
```

B

P1 = cI434 (1-63)	**P4** = [cI434 (28-46)]-Lys[cI434 (28-46)]-Gly-OH
P2 = cI434 (16-37)	**P5** = [cI434 (16-46)]-Lys[cI434 (16-46)]-Gly-OH
P3 = cI434 (16-45)	**P6** = [cI434 (1-46)]-Lys[cI434 (1-46)]-Gly-OH
	P7 = [cI434 (1-62)-Ahx$_2$]-Lys[cI434 (1-62)-Ahx$_2$]-Gly-OH

Figure 3.8. (**A**) Sequence of the cI434 repressor amino-terminal domain and position of structural units. (**B**) Peptides obtained by chemical synthesis are termed from P1 to P7. The short model peptides containing only parts of the N-terminal domain are also termed according to the structural elements which are included i.e. **P2** = H_2TH_3, **P3** = H_2TH_3L, **P4** = H_3Ldim, **P5** = H_2TH_3Ldim, **P6** = $H_1H_2TH_3$Ldim.

basic, are quite far removed from the actual structural arrangement in the protein. They find that the peptide does not bind to O_R1, O_R2 or O_R3 of the cognate DNA. Some binding is observed to the pseudo operator O_P1, but via a β-hairpin conformation inserted into the minor groove of DNA (Grokhovsky *et al.*, 1991). It is to avoid a similar outcome, and to attempt to design peptides that might take up structures as faithful as possible to that of the analogous fragments in the repressor, that extensive modeling studies were carried out before synthesis.

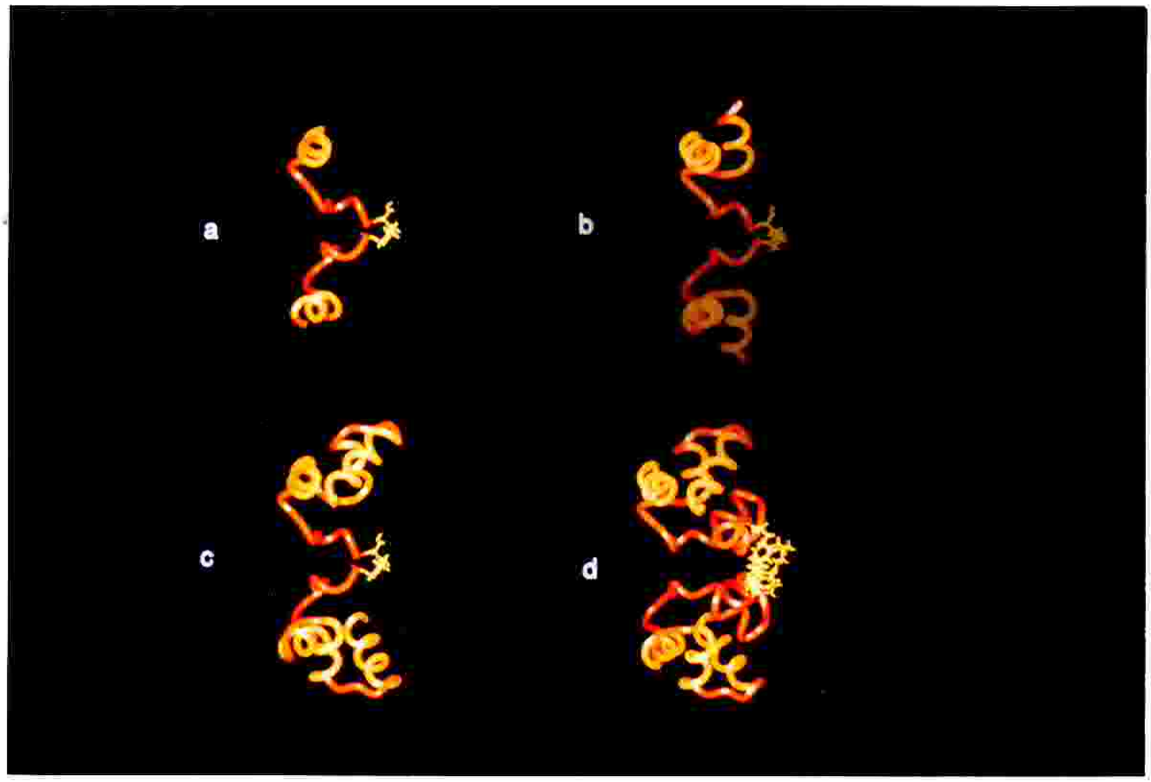


Figure 3.9. Computer models of synthetic dimeric peptides. (A) H_3Ldim; (B) H_2TH_3Ldim; (C) $H_1H_2TH_3$Ldim; (D) "Chemical Dimer", P7.

3.5 DIMERIZATION STUDIES USING THE ENTIRE N-TERMINAL DOMAIN

Dimeric versions of the cI434 N-terminal domain are designed by connecting either the C-terminus with a symmetric linker into dyad symmetric peptides that can be prepared by automated solid phase peptide synthesis, or by connecting the N- and C-termini with a peptide chain, to give direct repeats within a linear protein chain that can be produced using recombinant DNA methods.

Minimalist models, described in the previous section, allow the study of individual elements. If dimerization strategies are to be analyzed, on the other hand, one has to start with an element that will probably bind to DNA, and dimerize that element in different ways. Harrison and Aggarwal in the review on DNA recognition by the HTH motif (Harrison and Aggarwal, 1990) make the statement in their conclusion "Thus the unit that recognizes DNA is really an entire binding domain, appropriately dimerized, and nearly all the protein-DNA contacts in a complex with a correct target site contribute to specificity". So we chose the entire N-terminal domain (1-63) as DNA-binding element, and tried to design dimerization strategies that allow synthesis of a bipartite DNA-binding protein.

Attempts have been already reported in the literature for λCro in which two amino-terminal domains were dimerized via a disulfide bridge occurring between two cysteine residues in lieu of Val55 on both domains (Shirakawa *et al.*, 1991) apparently producing functional recombinant proteins. A similar approach was used to dimerize the basic α-helices of GCN4 responsible for DNA binding and the short peptide resulted fully viable (Talanian *et al.*, 1990). An alternative to the use of disulfide bridges would be to consider large organic cyclic molecules or chelating agents which bridge two peptide chains (Lieberman *et al.*, 1993, Stewart *et al.*, 1993) as for photochemical cleaving agents.

However, the introduction of a linker to build dimerization into a molecule has to (i) keep the monomeric subunits close together (at the correct distance as determined by computer modeling) and (ii) the whole dimeric structure flexible enough for DNA binding to occur.

The carboxy-terminus of the cI434 amino-terminal domain was then covalently linked through Gly 62, on both monomers (Figure 3.7 and Figure 3.8), to the α and ε amino groups of a lysine via two 6-aminohexanoic acid residues on each side (Figure 3.9). Computer modeling indicates that such linker (~ 35 Å long) which tightly packs close to the surface of interaction between the monomeric subunits, does not influence the stability of the molecule which apparently mantains the correct conformation necessary for DNA binding (Figure 3.10A and 3.10B).

It was thus decided to begin the dimerized protein termed P7 (Figure 3.8) at Gly 62. The molecule therefore contains all structural determinants important for stability and activity and covalent dimerization, apparently necessary for the protein to be fully capable of DNA binding, built into the protein in lieu of the carboxy-terminal domain of the native molecule.

The choice of the linker is important because it also provides the molecule with a certain degree of symmetry which mimics that of the binding site on the cognate DNA and contributes to render the protein analogue a functional model of the cI434 repressor. The single-chain repressor protein designed by this approach (130 amino acid residues long) contains a palindromic tail-to-tail symmetry which, in the native protein is provided by the carboxy-terminal domain of the cI434 repressor molecule. Such symmetry mimics the palindromic binding site on the cognate DNA and apparently contributes to render the protein analogue a functional model of the repressor (Figure 3.11).

In these conditions the synthesis would require 65 coupling reactions including the linker. This is still within the capacity of the synthesizer and for this reason we decided to attempt the synthesis in a single run with the appropriate precautions (Appendix I). The synthetic protein termed P7 (Figures 3.8, 3.9, 3.10) would contain two branches each with 62 amino acid residues with a similar topology as that described for the shorter dimeric peptides designed according to a minimalist approach.

Before synthesizing the protein, a modeling study was performed in order to test the viability of the design. The molecular model of the cI434 repressor amino terminal domain (Brookhaven entry PDB2OR1; Figure 3.1) was used as mentioned already, consisting of two monomeric subunits of 69 amino acid residues bound to a 20 bp long cognate DNA sequence. The same procedure was essentially followed as that reported in the previous section (see section 3.4) for the minimal peptides, to build dimerization and to energy minimize the whole dimeric structure P7. Also in this case a bond was firstly created between the ε amino group of the linker and the carboxy-terminal residue in the designed protein (Gly62) and then between the α amino group and the carboxy terminus of the other monomer. Energy minimization was carried out after both operations. To test whether the structure was energetically favourable all residues, but the bridging linker were kept fixed and energy minimization performed. The whole structure did not show significant changes as compared to that of the native protein (Figure 3.9 and Figure 3.10).

Dimerization was also built into the complete cI434 amino-terminal domain through a long stretch of amino acid residues connecting the carboxy-terminus of one monomer to the amino-terminus of the other monomer (Figures 3.11 and 3.12). The linker sequence (Figure 3.12) was directly abstracted from the sequence of the polypeptide chain connecting the amino-terminal domain to the carboxy-terminal domain of the native cI434 repressor. In this case the linker provides a different type

A

B

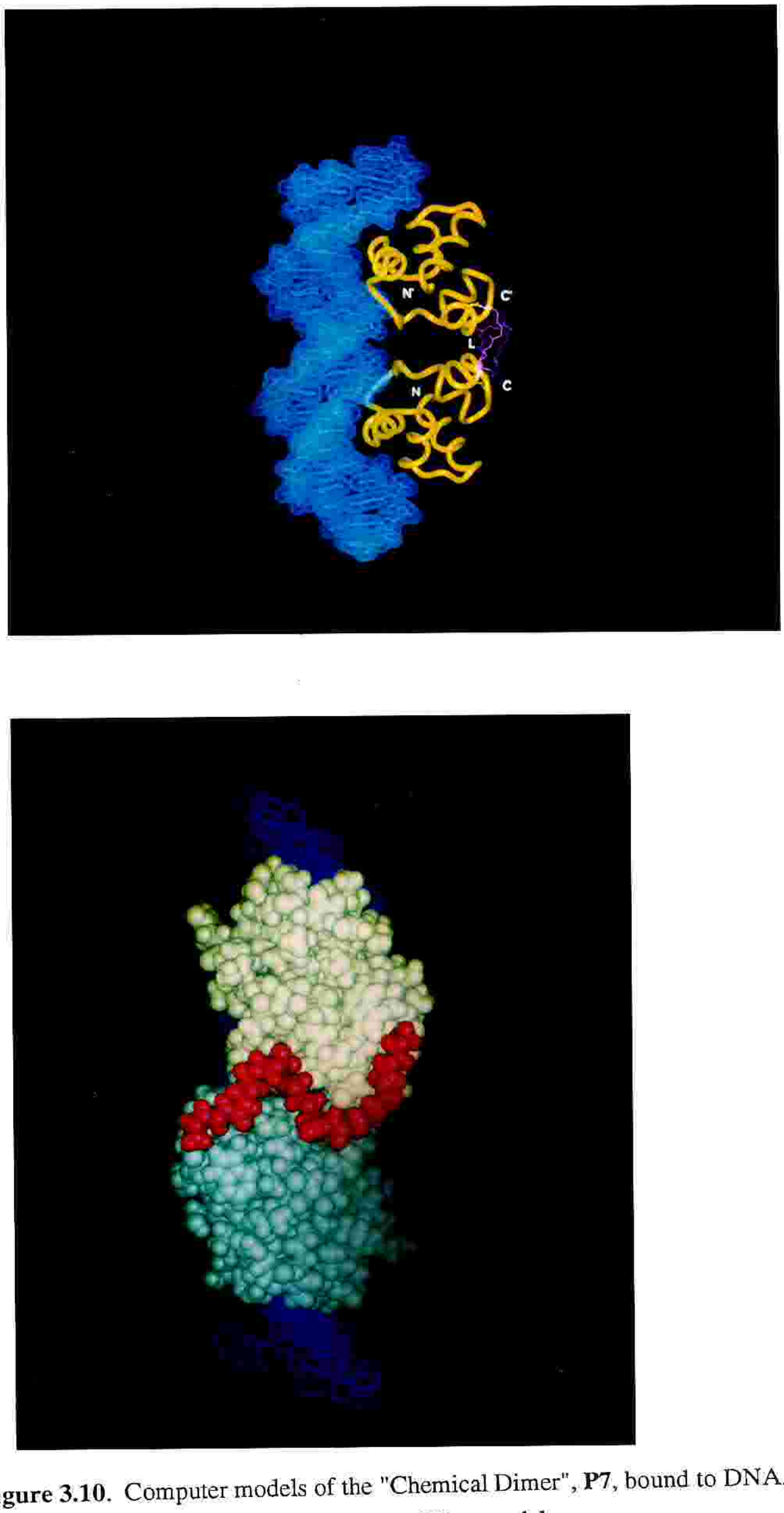


Figure 3.10. Computer models of the "Chemical Dimer", **P7**, bound to DNA. **(A)** Ribbon type representation and **(B)** space-filling model.

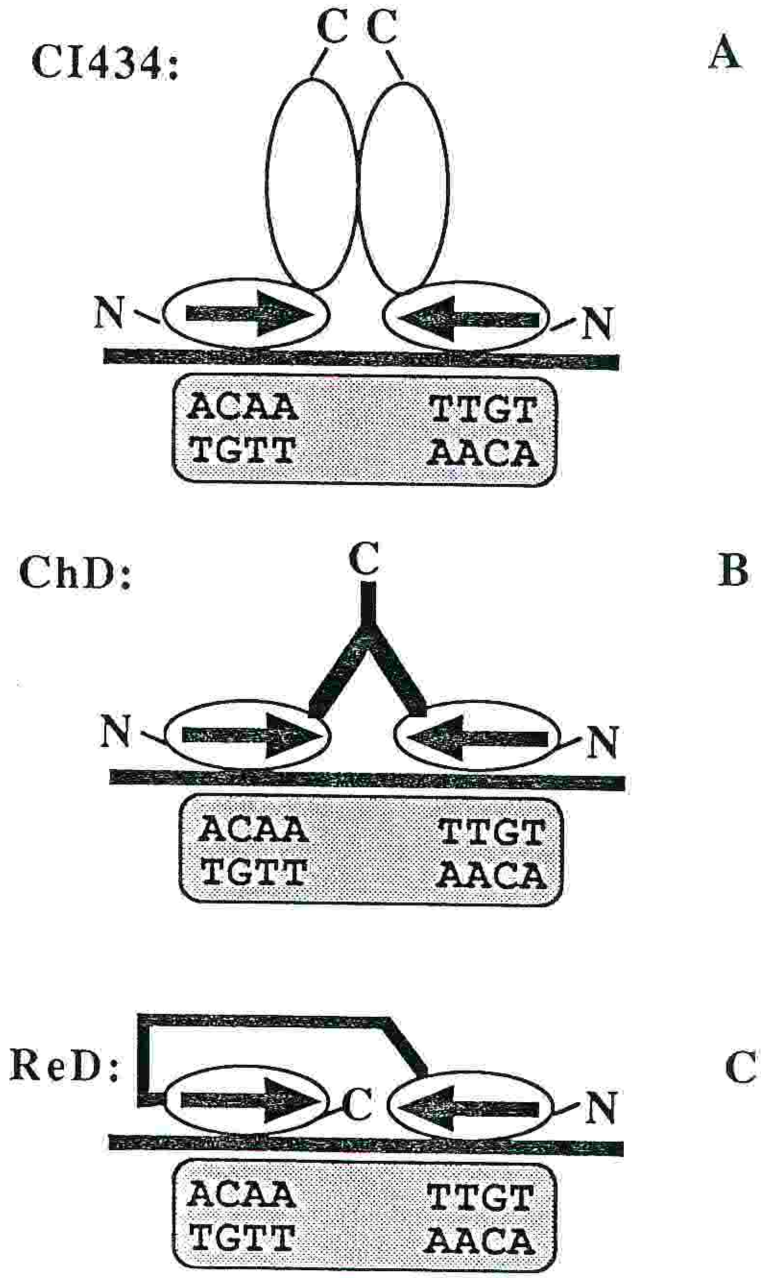


Figure 3.11. Schematic representation of (A) the native cI434 repressor, (B) "Chemical Dimer", P7, and (C) "Recombinant Dimer", RR69.

```
1                   20                  40
|                   |                   |
SISSRVKSKRIQLGLNQAELAQKVGTTQQSIEQLENGKTKRPRFLPELAS
|------H1------|---H2---|-T-|---H3---|---L---|-H4->

         60                  80                  100
         |                   |                   |
ALGVSVDWLLNGTSDSNVRFVGHVEPKGKYPLISMVRAGSISSRVKSKRI
>H4--|------H5-----|-------Linker------|---H1---->

                   120                 140
                   |                   |
QLGLNQAELAQKVGTTQQSIEQLENGKTKRPRFLPELASALGVSVDWLLN
>H1-|---H2---|-T-|---H3---|---L---|----H4----|-H5->

       158
       |
GTSDSNVR
>H5----|
```

Figure 3.12. Complete amino acid sequence of the "Recombinant Dimer", **RR69** and position of structural units.

of symmetry as compared to the one described above: the dimerized amino-terminal domain, termed RR69, is a linear or single chain repressor with a head-to-tail symmetry (Figure 3.11).

Molecular modeling studies performed on RR69 (Figure 3.13) show that the linker wraps around the whole molecule and apparently it does not perturb the structural and functional properties of the repressor. In this case the linker is more stable as compared to the one described above for P7.

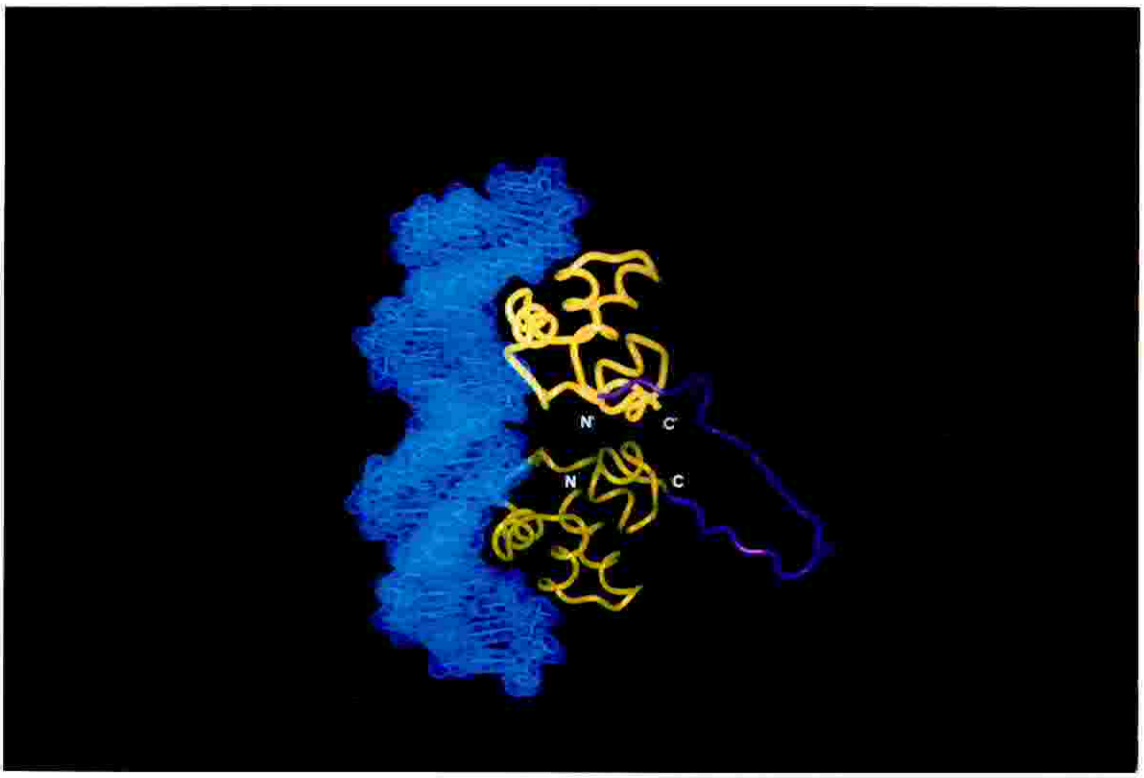


Figure 3.13. Computer model of the "Recombinant Dimer", RR69, bound to DNA.

4 CHARACTERIZATION OF SHORT DIMERIC PEPTIDES CONTAINING PARTS OF THE N-TERMINAL DOMAIN

5 peptides were prepared by automated solid phase synthesis, purified to homogeneity with reverse phase HPLC, and characterized with mass spectrometry and amino acid analysis. CD analysis revealed that α-helical structures are induced by the hydrophobic solvent TFE, but no conformational change was detectable upon addition of cognate DNA. Electrophoretic mobility shift assays (EMSA) showed no association with cognate DNA either. Photochemical crosslinking on the other hand showed some binding between cognate DNA and the H_2TH_3Ldim and $H_1H_2TH_3Ldim$ peptides.

4.1 SYNTHESIS AND CHARACTERIZATION OF SHORT MODEL PEPTIDES

The peptides H_2TH_3, H_2TH_3L, H_2TH_3Ldim and $H_1H_2TH_3Ldim$ all include the recognition helix (H_3) flanked on the N-terminal by a stabilizing helix and turn (H_2T). A loop (L) is added in peptide H_2TH_3L as well as peptide H_3Ldim, which includes only the recognition helix H_3 (Figure 3.9). The peptides were synthesized using the Fmoc chemistry in automated, continous flow, solid phase. In all cases the syntheses presented difficulties because of the strong tendency of such peptides to give conformational collapse on the resin. This problem was partly overcome by using a polyethyleneglycol/polystyrene (PEG-PS) graft resin, in particular in the case of H_2TH_3Ldim and $H_1H_2TH_3Ldim$, allowing a higher degree of solvation and thus increasing the efficiency of each coupling and the total yield of the synthesis (Appendix I). The peptides were purified by RP-HPLC, eluting the column with a linear gradient from A (0.05% TFA in water) to B (90% acetonitrile in A). During the purification process difficulties were faced in solubilizing the peptides. In particular H_2TH_3 was soluble only in a mixture containing 50% acetonitrile in 0.05% TFA. H_2TH_3L, H_3Ldim, H_2TH_3Ldim and $H_1H_2TH_3Ldim$ showed higher solubility in aqueous solvents due to the presence of the basic residues of the loop (L) and helix H_1.

After RP-HPLC purification, H_2TH_3, H_2TH_3L, H_2TH_3Ldim and $H_1H_2TH_3Ldim$ were characterized by amino acid analysis on a custom made system based on the PicoTag method. H_2TH_3L and H_2TH_3Ldim were also characterized by tryptic digestion/FAB-MS to determine the correct sequence, structure and molecular weight. Further analysis on both peptides performed by ES-MS showed that the measured values of the masses were in good agreement and they are as theoretically expected. $H_1H_2TH_3Ldim$ was also characterized by ES-MS and the molecular weight was found to be correct (Appendix I). The analytical data concerning all peptides are shown in Table 4.1.

Table 4.1. Analytical data of peptides based on the bacteriophage 434 cI Repressor.

Peptide (M.w.) Molecular formula	ESMS[#] $(M+H)^+$	Relative retention time (RPHPLC) I[b]	II[c]	"Difficult synthesis" (residues)[a]	Yield RPHPLC-area (%)	Purified (%)	Peptide MS-fingerprinting[e] tryptic\chymotryptic[f] digestion
P1 (6891.95)	6891.20±1.40	1.40	1.80	W^{58}-L^{45};V^{24}-N^{16}	33(39)[k]	15-18	
$C_{301}H_{508}N_{90}O_{94}$							
P2 (2768.28)	2772.34	1.18	1.53	V^{56}-S^{50}	62	48	
$C_{127}H_{207}N_{35}O_{34}$							
P3 (3413.87)	3413.24±0.40	1.12	1.24	V^{24}-E^{19}	-	24	N16-K23; V24-K38; TK; RPR;T39-R43;FL
$C_{146}H_{246}N_{46}O_{48}$	(3500)[d](3413.872)[i]						N16-L20; A21-Q28;Q29-L34; E35-R42;FL
P4 (4706.43)	4706.419[i]	1.28	1.47	-	27	-	
$C_{206}H_{434}N_{65}O_{61}$	(5000)[d]						
P5 (7189.12)	7190.60±1.57[g]	1.36	1.54	V^{24}-Q^{17}	17	12	N16-K23; V24-K38; TK; RPR;T39-R43;FL
$C_{310}H_{519}N_{97}O_{99}$	(14500)[d](7000)[j]						N16-L20; A21-Q28;Q29-L34; E35-R42;FL; FLPK(FLP)G
P6 (10497.19)	10494.38	1.52	1.94	V^{24}-Q^{17}; I^{11}-R^{5}	-	5	
$C_{454}H_{783}N_{145}O_{39}$	(10500)[d]						
P7 (14201.55)	14195.07	1.68[h]	2.23[h]	V^{54}-E^{47}; V^{24}-A^{18}; I^{11}-S^{1}	-	<1.1	
$C_{626}H_{1059}N_{185}O_{189}$	(14500)[d]						

[a]Shrinkage of swollen gel resin and/or broadening of deprotection peak. [b]Lichrospher 5RP18, 0.1M TEA- phosphate buffer pH=2.25, gradient 1% min.$^{-1}$ for 60 min. [c]Lichrospher 5RP18, 20 mM KOAc pH=7.2, gradient 1% min.$^{-1}$ for 60 min. [d]SDS-PAGE [e]enzymatic digestion/RPHPLC separation and FABMS identification of peptides; [f]Main cleavage sites. [g]Second signal 7212.08±1.13 correspond to the cationized **P7**. [h]dissolved in buffer A, containing 6-8M urea. [i]FABMS measured molecular weights. [j]SEHPLC (Shodex WS-802.5F), TEA- phosphate buffer, pH=3.0,dissolved and injected in 6 M guanidine hydrochloride; [k]IEHPLC-area (Shodex SP, borate buffer, gradient NaCl). [#] The values for amino acid residues after acid hydrolysis of all peptides (6N HCl, 24 hrs, 110°C) were ± 10% of the expected.

4.2 DNA-BINDING ASSAYS

Different types of binding assays were attempted, which consisted mainly in determining whether the peptides induced a band-shift in DNA or oligonucleotides containing one or more operator sites. To test for specific binding to DNA, the effect of aspecific competitor DNA (in this case poly[d(I-C)]) was determined. Given that the design of the peptides was approached from a minimalist point of view, and that a considerable part of the protein with its contribution to the overall affinity for DNA had been eliminated, it was assumed that very weak but specific binding might occur that can not be detected under the standard conditions for electrophoretic mobility shift assays. Experiments were thus designed in which the peptides would be permanently covalently linked to the DNA, at the binding site, via photochemical crosslinking. This would trap the peptide-DNA complex and allow observation of a gel retardation.

4.2.1 ELECTROPHORETIC MOBILITY SHIFT ASSAYS

Electrophoretic mobility shift assays (EMSA) were carried out in the presence of H_2TH_3, H_2TH_3L, H_2TH_3Ldim and $H_1H_2TH_3$Ldim with various DNA fragments based on the cI434 operators. These DNA fragments were either oligonucleotides obtained by chemical synthesis or longer DNA fragments obtained by PCR amplification of a template plasmid of phage DNA containing operator sequences (Appendix I).

All gel shift experiments performed with short oligonucleotides or DNA fragments containing one of the three cI434 operators (O_R1) in the presence of both monomeric and dimeric peptides have as yet failed to show any retardation (data not shown). This either indicates that the binding is not occurring in any case, or that in cases where it does occur, binding is weak and does not survive the conditions for band-shift experiments. Care was taken to incubate the peptides with the cognate DNA at 4°C, and to electrophorese them under mild conditions in a cold room. However as the peptides are slightly basic, it is possible that they were swept away from the negatively charged DNA under electrophoretic conditions if binding was weak, or reversible. Control experiments in the presence of all operator models (Figure 4.1) with the cI434 repressor crude extract in the presence of a large excess of competitor DNA (~ 1000 fold as compared to the specific DNA), showed retardation in all cases. Results of band-shift experiments with H_3Ldim were inconclusive as this tended to form an aggregate with the DNA, which did not enter the gel.

Gel shift assays in the presence of a DNA fragment containing the complete cI434 O_R region apparently show a slightly retarded band only for the dimeric peptides (data not shown) which tends to disappear when increasing amounts of competitor DNA, poly[d(I-C)], are added. It is not observed for H_2TH_3 (data not shown) nor is it detected in the case of H_2TH_3L (data not shown). Band-shifts cease to be observed for dimeric peptides only when a greater than 100 fold excess of aspecific competitor is added, indicating that the peptide apparently binds in a specific manner to the cognate DNA.

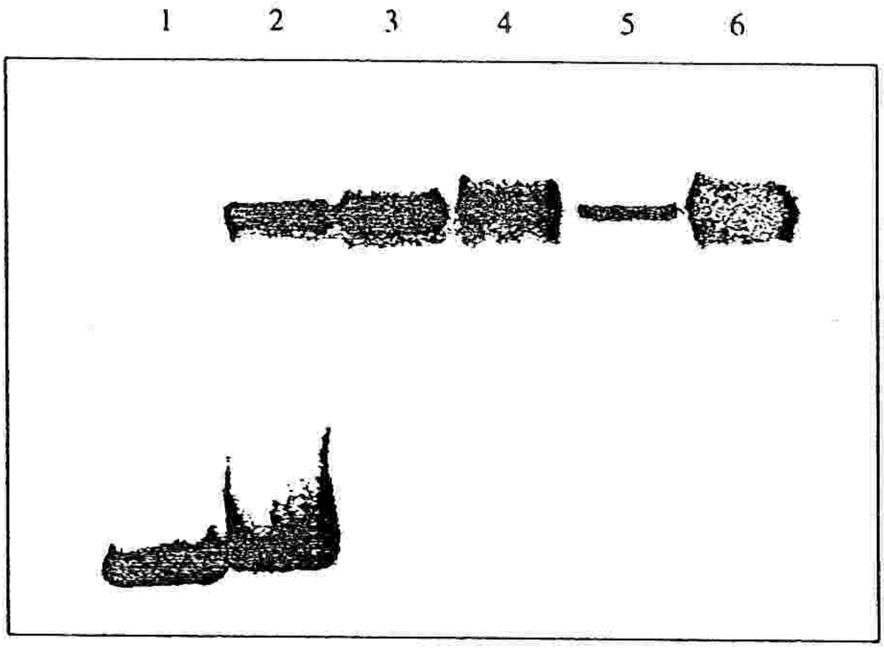


Figure 4.1. Gel mobility shift assay performed with a protein extract containing the cI434 repressor in the presence of $434O_R1$ (150 bp). Lane 1, no protein extract; lanes 2 to 6, incremental amounts (1 to 5 µl) of protein extract (stock solution 1 µg/µl) with ~ 1000 fold excess of poly[d(I-C)].

4.2.2 PHOTOCHEMICAL CROSSLINKING

To test the binding specificity of H_2TH_3Ldim and $H_1H_2TH_3Ldim$ for its cognate DNA while overcoming the problem of weak binding, gel shift assays were carried out with double-stranded oligonucleotides containing either one or two cI434 operators synthesized with bromodeoxyuridine 5'-triphosphate (dUBrTP) in the place of dTTP. These operator analogues were then photochemically crosslinked both to each dimeric peptide and to the monomeric H_2TH_3L peptide. This is a general methodology in order to induce covalent adduct formation (Appendix I).

Strong adduct formation is observed when the binding reaction is performed with the 45 bp long oligonucleotide (with dTTP replaced by dUBrTP residues) with two operator sites in the presence of H_2TH_3Ldim (Figure 4.2) and $H_1H_2TH_3Ldim$ (data not shown). This is direct evidence of the binding of both peptides to the cognate DNA in the case when two operator sites are present.

In accordance with the conventional gel-shift experiments described in the previous section, no adduct formation is observed when the reaction is carried out with the 32 bp long oligonucleotide (with dUBrTP replacing dTTP) with one operator site only in the presence of H_2TH_3Ldim (Figure 4.2) and $H_1H_2TH_3Ldim$ (data not shown). This is direct evidence of the importance of the cooperative effect.

No adduct formation was observed with both DNA models in the presence of H_2TH_3L (data not shown), consistent with the results obtained in conventional gel mobility shift assays. This confirms the importance of subunit dimerization in the binding process.

Furthermore the molecular weight of the shifted bands observed in crosslinking experiments with both dimeric peptides were estimated by comparison with molecular weight standards, after staining with Coomassie. As for H_2TH_3Ldim, the estimated

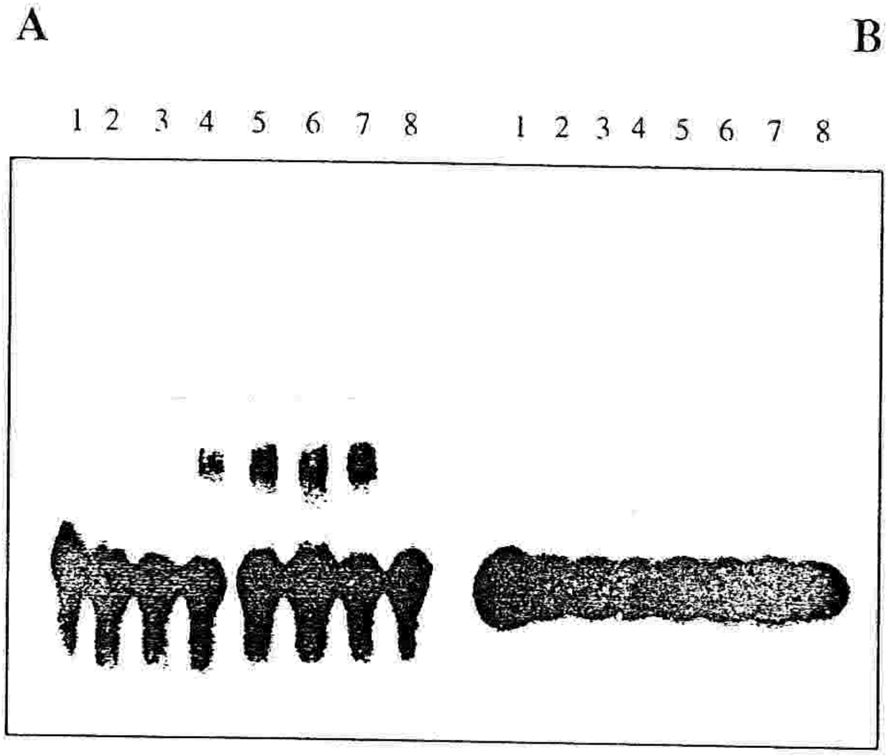


Figure 4.2. Photochemical crosslinking of peptide H_2TH_3Ldim to oligonucleotides containing bromodeoxyuridine 5'-triphosphate in the place of thymidine residues. (A) Experiments were performed with an oligonucleotide (45 bp) containing 434 operators O_R1 and O_R2. Lanes 1 to 7, incremental amounts of peptide (final concentrations ranging between 1.4 μM and 9.5 μM); lane 8, no peptide. (B) Experiments were performed with an oligonucleotide (32 bp) containing only the O_R1 434 operator. Lanes 1 to 7, incremental amounts of peptide (final concentrations 1.4 μM, 2.7 μM, 4 μM, 5.4 μM, 6.7 μM, 8.1 μM and 9.5 μM); lane 8, no peptide.

mass for the complex of two such dimeric peptides with the 45 bp long oligonucleotide containing two operator sites is approximately 40KDa, and the complex band falls at a position with respect to the standards consistent with this molecular weight. This supports a 2:1 stoichiometry for the interaction. Also, the fact that there are two molecules of peptide for two operator sites suggests that the binding of H_2TH_3Ldim to its operator sites is specific. Similarly, this is the case for $H_1H_2TH_3$Ldim which apparently binds the same dUBrTP containing operator DNA with a 2:1 stoichiometry suggesting that specific interaction occurs.

From the results described above, the following considerations can be made:

1) Binding does not occur for the monomeric H_2TH_3 or H_2TH_3L to any of the DNA models including one or more operator sites, as well as the dimeric peptides, in the standard EMSA conditions used in our experiments.

2) The dimeric peptides H_2TH_3Ldim and $H_1H_2TH_3$Ldim, but not the monomeric ones, specifically bind to oligonucleotides (with dUBrTP replacing dTTP) containing at least two operator sites only by photochemical crosslinking.

3) They do not bind oligonucleotides (with dUBrTP replacing dTTP) containing only one operator site.

From these one could conclude that binding only occurs for a dimeric form of the model peptide, and that in any case the peptide-DNA interaction is very weak, although it appears to be specific. A degree of cooperativity is suggested by the fact that binding occurs only if more than one operator site is present in the oligonucleotide sequence. The cooperativity could derive from two distinct effects, i) an interaction of an incoming peptide with one already bound to the DNA, aiding the binding of the former, or ii) a conformational change in the DNA which increases the avidity of the second peptide to the DNA, or from a combination of both these effects. As for the full repressor protein, it is known that binding of one repressor dimer to an operator site (say O_R1) facilitates the binding to the adjacent operator (O_R2), but this does not extend to O_R3. In that case we are however dealing with quite large entities, so that a contact between adjacently bound proteins is plausible and known to occur, especially as the DNA to which they are bound is bent. It is difficult to see how this could be extended to quite small peptides. However, the design of suitable DNA molecules containing multiple operator sites either in-phase (i.e. facing the same side on the DNA duplex) or out-of-phase (i.e. with the binding sites alternating on different sides of the DNA duplex), would probably provide clear-cut evidence about a cooperative effect aiding the actual process of DNA binding.

4.3 CD SPECTROSCOPY

Circular dichroism was carried out on H_2TH_3, H_2TH_3L, H_2TH_3Ldim and $H_1H_2TH_3Ldim$ to study their tendency to fold into the appropriate α-helical conformation required for binding to DNA. We used trifluoroethanol (TFE) to check if the peptides show the conformational transition common to all amphifilic peptides, and oligonucleotides containing the O_R1 binding site, to check if induction of the helical structure is promoted by DNA-binding.

In the case of H_2TH_3 the CD spectrum in aqueous solution (data not shown) indicates that the peptide is in a random coil conformation. This is as expected for a relatively short peptide. At 10% TFE, a solvent that mimics an organic environment such as that found in the DNA major groove, a change in the spectrum may already be observed, whereas one can clearly observe the typical spectrum of an α-helical peptide when the percentage of TFE is raised to above 25%. This indicates that the peptide folds to a structure with a high α-helical content in an organic environment, which is consistent with its structure inside the cI434 amino-terminal domain. Similarly the CD spectra of H_2TH_3L in the presence of TFE (Figure 4.3) show clearly that it also folds into an α-helix at above 25% TFE. Figure 4.3 also shows the CD spectra obtained for H_2TH_3Ldim and $H_1H_2TH_3Ldim$ at increasing concentrations of TFE. The behaviour of dimeric peptides is entirely analogous to that of the other two peptides. However the α-helical propensity at corresponding concentrations of TFE is higher for the dimeric peptides H_2TH_3Ldim and $H_1H_2TH_3Ldim$ as compared to the monomeric ones. Such conformational transitions occurring for both dimeric peptides indicate that the dimerization process, effected by covalent binding of two monomeric units to the α and ε amino groups of a lysine, has not had a deleterious effect on folding (Figure 3.9). That some form of inter-strand interaction might occur precluding a correct folding had been reason for concern. The CD spectra of all peptides indicate that in water or phosphate buffer they are essentially present as random coils. The cI434 repressor *in vivo,* being a fully fledged protein, obviously exists in a folded form, but even in its case, it only achieves a complete three dimensional arrangement of its α-helical units in the presence of cognate DNA (Branden and Tooze, 1991 and references therein).

A change to the CD spectrum of the peptide, in the presence of a cognate DNA supporting an α-helical arrangement, would indicate that an interaction is occurring. CD studies were performed with synthetic peptides in the presence of either DNA fragments (obtained by direct PCR amplification) with the cI434 O_R1 or short oligonucleotides containing the specific binding site or a random sequence. In the first case, technical problems were faced in that the absorption of DNA was too strong at the concentration required to observe a typical α-helix band at 222 nm due to the peptides conformational transitions (from a random coil to an α-helix) upon binding to DNA or in the presence of a folding agent. This is clearly a difficulty as the CD spectra are measured as a difference in the absorption for alternatively right and left polarized light, and the signal to noise ratio becomes very high at increased absorbances. Use of shorter oligonucleotides containing the binding site or random sequences did not provide clear-cut results in that none of the available peptides gave CD spectra with characteristic α-helix bands. The explanation could be that no binding occurs or the peptide-DNA interaction in this case is rather labile and can not be studied by CD. This type of experiment will therefore have to await an optimized peptide which will bind strongly to the cognate DNA, thus enabling binding to shorter DNA fragments and at a lower excess (possibly at equimolar concentrations).

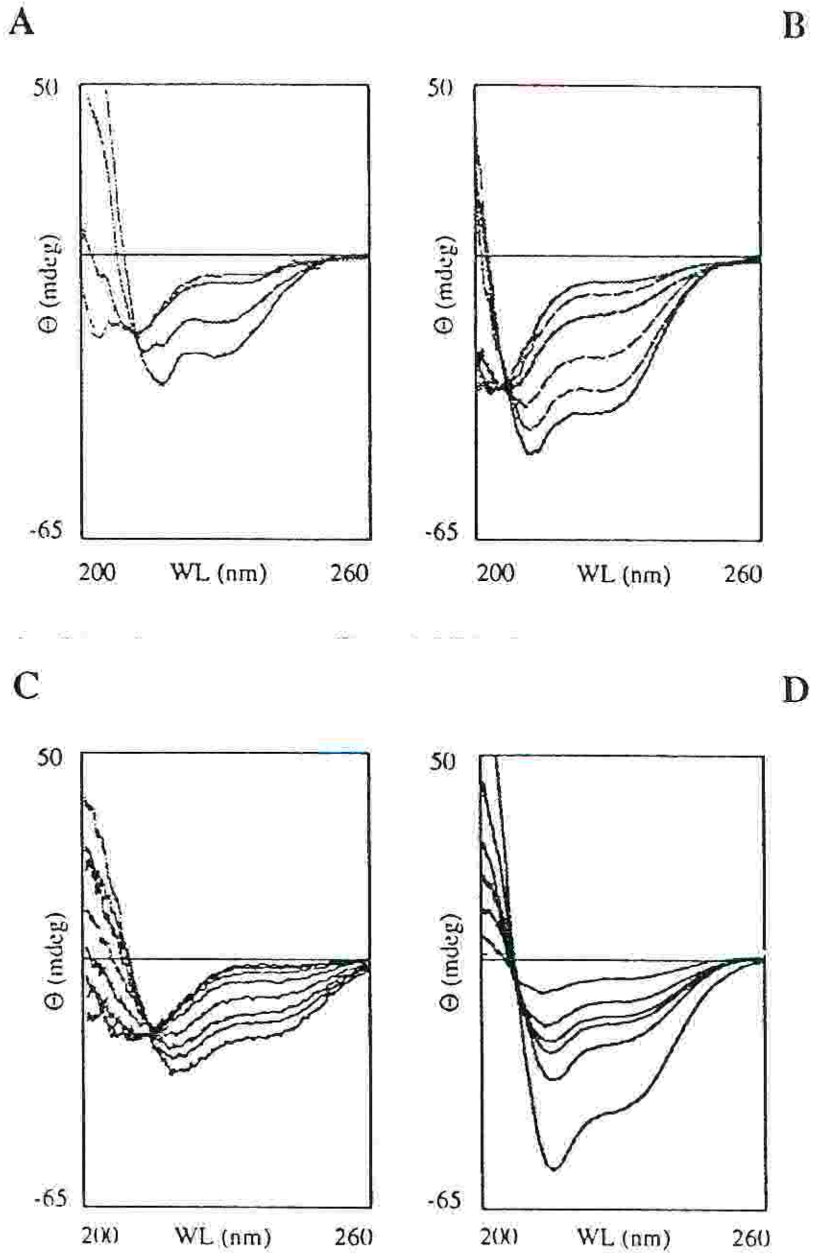


Figure 4.3. CD spectra performed with peptides (A) H_2TH_3L, (B) H_2TH_3Ldim, (C) $H_1H_2TH_3Ldim$ and (D) "Chemical Dimer" P7 in 10 mM phosphate buffer with increasing amounts of TFE (0 to 50 %).

5 CHARACTERIZATION OF SINGLE-CHAIN REPRESSORS OBTAINED BY COVALENT DIMERIZATION OF THE N-TERMINAL DOMAIN

The dyad-symmetric dimer P7 was prepared by chemical synthesis whereas dimer RR69 which contains direct sequence repeats was prepared by recombinant DNA methods. Their properties were compared with those of the monomeric cI434 N-terminal domain (termed P1 and prepared by chemical synthesis). Gel-shift experiments showed that P7 and RR69 can bind cognate DNA (K_d=3.25x10^{-6} and 5x10^{-8}, respectively), while P1 can not. None of the three peptides bound non-cognate DNA gel mobility shift experiments. Dimers P7 and RR69, but not the monomeric P1, showed specific DNA-binding to the entire operator region (O_R1, O_R2, O_R3) in DNAse footprinting and methylation interference experiments and both of them inhibited in vitro transcription to the same extent. CD studies showed that cognate DNA substantaially inceases the α-helix content of the dimers, but not that of the monomer. Interestingly, non-specific DNA gave about 70 % of the α-helical induction produced by cognate DNA. The result was essentially similar with both dimeric molecules.

In Chapter 3, the design of two dimeric molecules was described. One of them, P7, is a branched peptide in which the C-termini of two N-terminal domains are connected via 6-aminohexanoic acid molecules to a central lysine residue, to give a branched peptide with dyad-simmetry. The chemical synthesis and characterization of this "chemical dimer" molecule (P7) is described in section 5.1. The other dimeric molecule, also described in Chapter 3, contains two direct repeats of the N-terminal domain connected with an internal sequence of the cI434 repressor. The production of this "recombinant dimer" molecule (RR69) through expression in *E. coli* is described in section 5.2. Finally, section 5.3. describes the *in vitro* studies conducted on these two molecules (gel mobility shift, footprinting, methylation interference, *in vitro* transcription inhibition).

5.1 CHEMICAL SYNTHESIS AND CHARACTERIZATION OF P7, A BRANCHED PEPTIDE CONTAINING TWO COPIES OF THE cI434 N-TERMINAL DOMAIN.

The synthesis of the branched dimeric molecule P7 (Figure 3.8) was based on the synthesis of the linear monomeric cI434 N-terminal domain, a peptide (P1) which was briefly mentioned in Chapter 2. Peptide P1 was synthesized first and the problematic steps were identified (Percipalle *et al.*, 1994a). P7 (Figure 3.8 and Figure 5.1) was then synthesized in a single run by continuous flow SPPS using the Fmoc chemistry. In both cases, the syntheses presented several difficulties due (i) to the length of the polypeptide chains elongated (63 and 130 amino acid residues respectively for P1 and P7) and (ii) to the strong tendencies of such proteins to fold during the synthesis thus causing conformational collapse of the structure on the resin. Same as for the shorter peptides described in Chapter 4, this problem was partly overcome by using a polyethyleneglycol/polystyrene (PEG-PS) graft resin, to allow a higher degree of solvation and an increse in the efficiency of each coupling step. In both syntheses, however, the coupling reagents Fmoc-amino acid, HOBT, TBTU were dissolved in N-methylpyrrolidone. This latter solvent has excellent solvating properties and therefore it reduces the extent of folding and aggregation, which is particularly important in the case of such long polypeptide chains. Capping was also introduced in both syntheses after residues Lys 40, Leu 34, Gln 28, Val 24, Leu 15, Ile 11, Val 6, Ile 2 as couplings were predicted to be difficult (see Appendix I for materials and methods). Protein P1 was synthesized in a relatively good yield whereas the difficulties encountered during the synthesis of the 130 residue long P7 were reflected in a lower yield (Table 4.1).

After synthesis, P1 was directly purified to homogeneity using RP-HPLC (Table 4.1) with a linear gradient from 0 to 60% acetonitrile. As for P7, some problems of solubility were encountered. For this reason, it was dissolved in 30% acetonitrile containing 6 M guanidine-HCl and then characterized and purified by RP-HPLC (Table 4.1) using 0.1 M triethylammonium phosphate buffer, pH 2.25, with a gradient of acetonitrile (see Appendix I for HPLC profiles).

Purified proteins P1 and P7 were characterized by amino acid analysis as for the shorter peptides described in Chapters 3 and 4, on a custom made system based on the PicoTag method. Both P1 and P7 were characterized by ES-MS and the molecular weights experimentally obtained are in good agreement with the expected ones (Table 4.1).

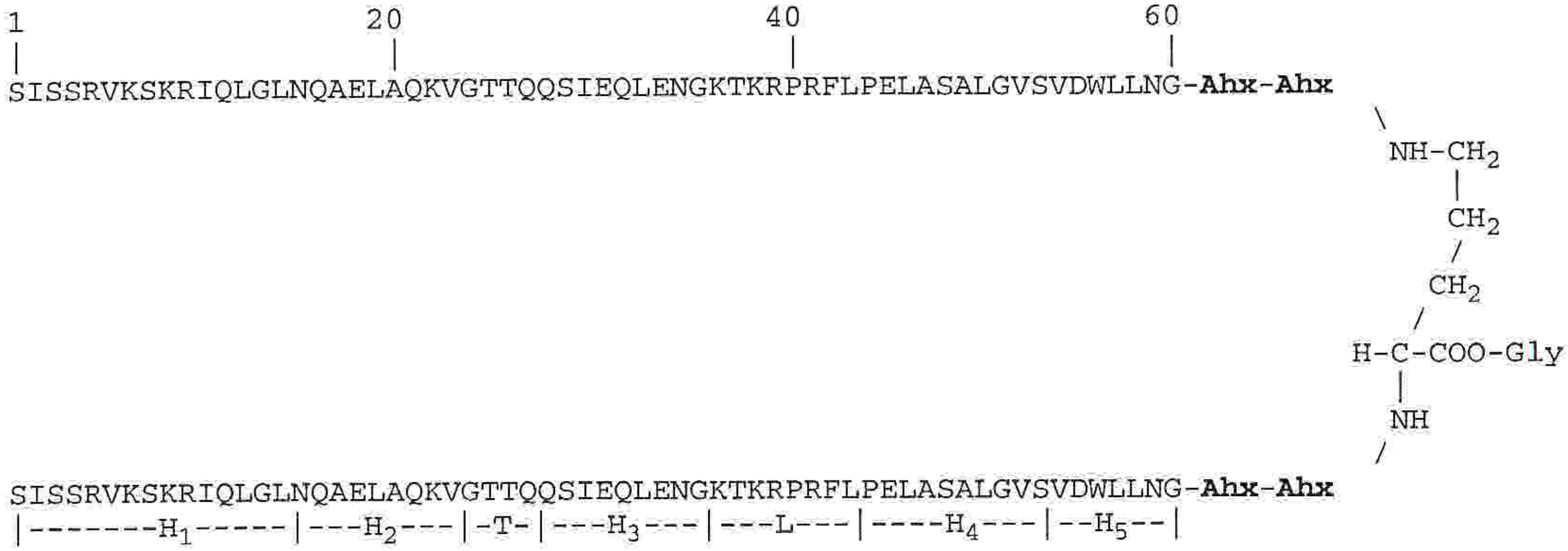


Figure 5.1. Schematic representation of the "Chemical Dimer", **P7** and position of structural units. Ahx indicates the 6-aminohexanoic acid residues used in the linker.

5.2 PREPARATION OF RR69, A SINGLE CHAIN REPRESSOR WITH TWO DIRECT REPEATS OF THE cI434 N-TERMINAL DOMAIN USING RECOMBINANT DNA TECHNIQUES.

Head-to-tail dimerization of the cI434 amino-terminal domain (1-69 amino acid residues in this case) was built in via a long stretch of amino acid residues between the two domains. The peptide linker is a part of the natural linker found between the N- and C-terminal domains of the full repressor, and corresponds to amino acids 70-89 in the native protein. The arrangement of the single chain repressor is therefore R(1-89)R(1-69) or, in shorter form, RR69 (Figure 3.12 and Figure 3.13), and includes 158 amino acid residues.

RR69 was obtained through *E. coli* expression of the corresponding artificially duplicated gene. The artificial gene arrangement was constructed from DNA fragments coding for amino acids 1-89 and 1-69, respectively, which were in turn obtained by PCR amplification of these fragments using λgT10 DNA template. The artificial tandem gene was cloned into pET *E. coli* expression vector (Simoncsits *et al.*, manuscript in preparation) and the RR69 protein was purified by cation exchange column chromatography using S-Sepharose to over 95% homogeneity as analyzed on SDS polyacrylamide gel electrophoresis. The protein concentration was estimated by using the calculated molar extinction coefficient at 280nm.

5.3 DNA BINDING STUDIES

Binding assays were designed and performed according to standard protocols (see Appendix I for "Materials and methods") with P1 and P7, as well as RR69, with all available operator DNA models. Proteins were assayed for specific DNA binding activity by electrophoretic mobility shift experiments (section 5.3.1), DNAse footprinting (5.3.2), DMS methylation interference (5.3.3) and inhibition of *in vitro* transcription (5.3.4). Dissociation constants (K_d) were calculated for P7 and RR69 which gave specific binding to DNA in the electrophoretic assays.

5.3.1 GEL MOBILITY SHIFT ASSAYS

Gel mobility shift assays are used to demonstrate the association of protein ligands with labeled DNA (Fried and Crothers, 1981; Garner and Revzin, 1981). We used "cognate" oligonucleotides containing specific 434 repressor interaction sites, as well as aspecific ones ("non-cognate"). The double-stranded DNA molecules were prepared either by chemical synthesis (phosphoramidite chemistry) or by PCR amplification of the binding site(s) (Appendix I) and are summarized below.

1) 434 O_R1 (150bp)	It contains only the O_R1 rightward operator of phage 434 and it was used in all DNA binding experiments.
2) 434 O_R (180bp)	It contains the full rightward operator, comprised of three adjacent sites (O_R1, O_R2, O_R3). It was used in some DNA binding experiments.
3) 434 O_R1' (20bp)	It is a short double-stranded oligonucleotide containing the O_R1 site. It was used in circular dichroism studies to investigate cognate protein-DNA interactions.

4) CL/WL (20bp)	It is a short double-stranded oligonucleotide with a sequence not related to any 434 operators. It was used in circular dichroism studies to investigate non-cognate protein-DNA interaction ("non-cognate" DNA).

The sequences of the above oligonucleotides as well as of 434 O_R1 (150bp) and 434 O_R (180bp) are shown in Figure 5.2.

The gel-mobility experiments were carried out in the presence of a 1000 fold excess of competitor DNA (Appendix I); the results are summarized in Figures 5.3 through 5.6. The monomer P1 showed no detectable DNA-binding activity towards any of the DNA-fragments, cognate or non-cognate alike. The chemical dimer P7 showed specific binding towards all cognate DNA samples; the results for O_R1 are shown in Figure 5.3. The recombinant dimer RR69 also showed specific binding to all cognate DNA molecules, the results for O_R1 are shown in Figure 5.4. Figure 5.5 shows that increasing concentrations of the recombinant dimer RR69 produce three association complexes with 434 O_R (180bp). This is in agreement with the incremental occupation of the 3 operator sites included in the DNA fragment. The chemical dimer P7 produced identical results (data not shown).

Binding to cognate and non-cognate DNA is compared in Figure 5.6. The results illustrate that none of the peptides bind non-cognate DNA, and the two dimers P7 and RR69 both bind to cognate DNA. As the assay is conducted in the presence of a large excess of competitor DNA, the binding is considered specific. Control experiments were made to test if the peptides bind to non-cognate DNA in the absence of competitor, but no binding was detected (data not shown).

Gel-mobility experiments were used to derive dissociation constants (K_d values) for 434 O_R1 complexes formed with P7 and RR69, respectively. The K_d values were determined as the protein concentration necessary to bind 50 % of the labelled DNA (Williams *et al.*, 1993; Carey, 1988) using concentration series similar to

A

5' TTACCCTGGAAGAAATACTCATAAGCCACCTCTGTTATTTACCCCCAATCTT

C**ACAA**GAAAAA**CTGT**ATTTGACAA**ACAA**GATACA**TTGT**ATGAAAAT**ACAA**GA
|----O_R3-----| |----O_R2-----| |-O_R1

AAGT**TTGT**TGATGGAGGCGATATGCAAACTCTTTCTGAACGCCTCAAGAAGA
>-O_R1--|

GGCGAATTGCGTTAAAAATGACGCAAACCG

B

5' TAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTCGTA

TGTTGCAT**ACAA**GAAAGT**TTGT**TGAGGAAACAGCTATGACCATGATTACGGA
|----O_R1-----|

TTCACTGGCCGTCGTTTTACAACGTCGTGACTGGGAAAACCCTGGCGTTAC

C

5' **TACAA**GAAAGT**TTGT**
|----O_R1-----|

D

5' AGAGGAGCGCCTGTCATGCA

Figure 5.2. Sequences of the various 434 operator models used in the binding assays and CD studies. In (**A**) and (**B**) the DNA fragments, obtained by PCR amplification, respectively contain all 434 operators (O_R1 to O_R3) and O_R1 only. In (**C**) a short oligonucleotide was designed for CD studies which contains only operator O_R1. In (**D**) the DNA was designed to contain an unrelated sequence and used in CD studies.

those shown in Figures 5.3 and 5.4. In this assay, the chemical dimer P7 gave a K_d value of 3.25×10^{-6} M, whereas a K_d value of 5×10^{-8} M was found for the recombinant dimer RR69 (K_d values for non-cognate DNA and for P1-cognate DNA complexes could not be determined this way as there was no binding detected in the entire concentration range allowed by the solubility of the samples).

A dissociation constant of 3×10^{-9} M is reported for the native 434 repressor protein-O_R1 operator complex based on quantitative footprinting and gel mobility shift experiments (Ptashne, 1992 and references therein). The K_d values of RR69 and P7 are one and three orders of magnitude less. At present we do not know the significance of this difference; we plan to express the native repressor in order to compare the binding properties within the same experiment. It is apparent, however, that both artificial dimers bind less strongly than the native repressor, and that, of the two molecules, the chemical dimer has the lower binding affinity.

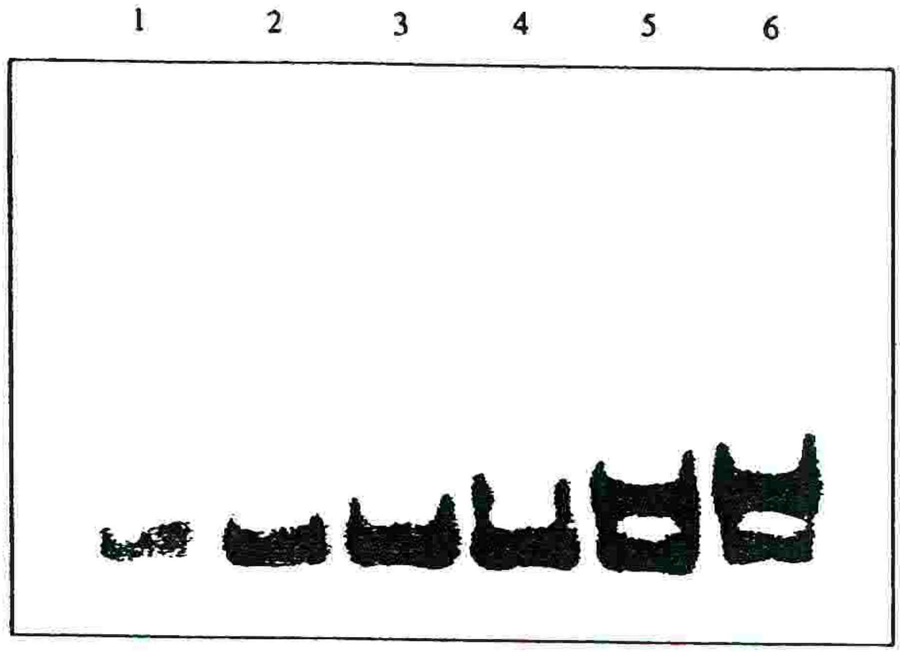


Figure 5.3. Gel mobility shift assay performed with P1, P7 in the presence of $434O_R1$ (150 bp). Lane 1, P1 (final concentration 3 μM); lane 2, no proteins; lanes 3 to 6, incremental amounts of P7 (final concentrations of 0.325 μM, 0.65 μM, 1.5 μM, 3.5 μM, respectively) with1000 fold excess of poly[d(I-C)].

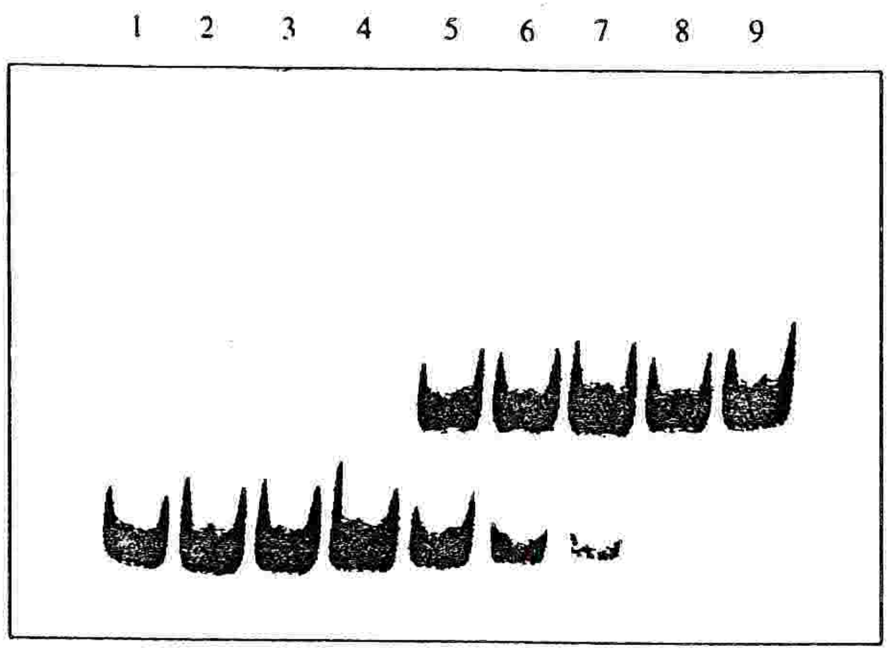


Figure 5.4. Gel mobility shift assay performed with RR69 in the presence of $434O_R1$ (150 bp). Lane 1, no protein; lanes 2 to 9, incremental amounts of RR69 (final concentrations 1.2 nM, 12 nM, 24 nM, 48 nM, 120 nM, 240 nM, 0.6 μM, 1.2 μM, respectively) with 1000 fold excess of poly[d(I-C)].

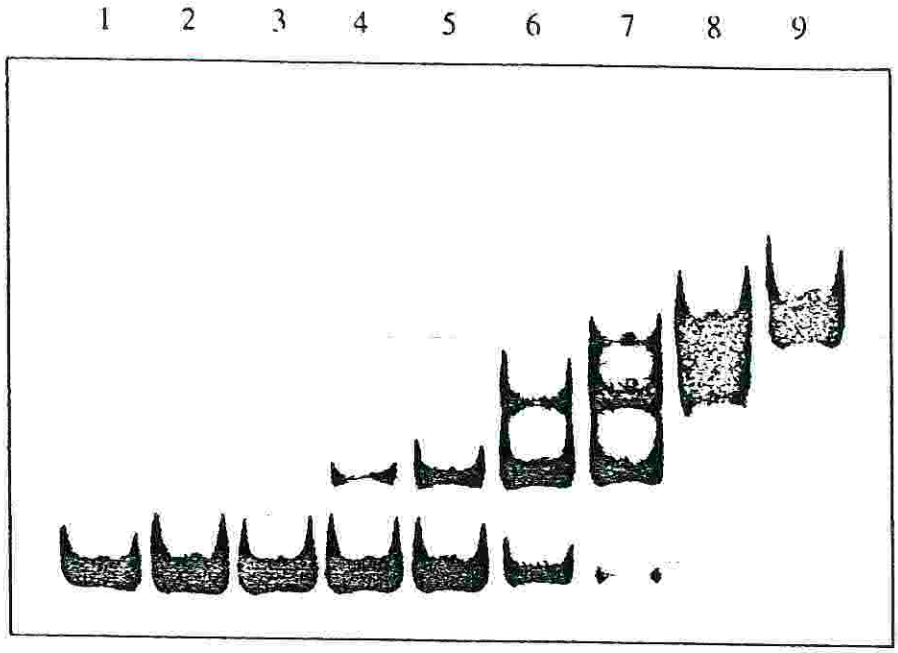


Figure 5.5. Gel mobility shift assay performed with RR69 in the presence of $434O_R$ (180 bp). Lane 1, no protein; lanes 2 to 9, incremental amounts of RR69 (final concentrations 1.2 nM, 12 nM, 24 nM, 48 nM 120 nM, 240 nM, 0.6 μM, 1.2 μM, respectively) with 1000 fold excess of poly[d(I-C)].

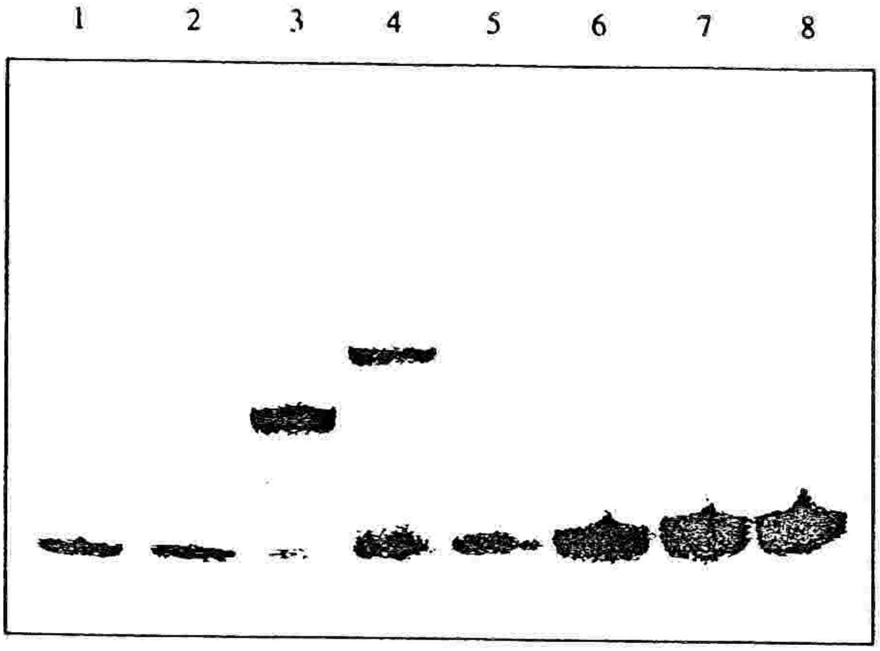


Figure 5.6. Gel mobility shift assay performed with P1, P7 and RR69 in the presence of either cognate [434O_R1 (20 bp)] or non-cognate (CL/WL) DNA. Lanes 1 to 4 refer to experiments carried out with cognate DNA: lane 1, no proteins; lane 2, P1 (final concentration 0.5 μM); lane 3, P7 (final concentration 0.48 μM); lane 4, RR69 (final concentration 0.25 μM). Lanes 5 to 8 refer to experiments carried out with non-cognate DNA: lane 5, no proteins; lane 2, P1 (final concentration 0.5 μM); lane 3, P7 (final concentration 0.48 μM); lane 4, RR69 (final concentration 0.25 μM). In all cases binding reactions were performed with 1000 fold excess of poly[d(I-C)].

5.3.2 DNAseI FOOTPRINTING

The most common method for determining protein binding sites on DNA is DNAseI footprinting (Galas and Schmitz, 1978), which is based on determining the DNAseI cleavage pattern of a DNA fragment in the presence of DNA-binding proteins. In this analysis the bonds protected by bound protein can be identified with chemical DNA sequencing (Maxam and Gilbert, 1980). DNAse footprinting is based on the relatively low sequence specificity of the enzyme, and on the ability of DNA-bound proteins to prevent cleavage of the DNA backbone in the region of the DNA-protein interaction.

DNAse footprinting was carried out to probe the binding specificity of P7 and RR69 against various cognate DNA sequences. For this purpose, 5' end-labelled DNA fragments were prepared by PCR amplification using a 5' ^{32}P-labelled primer (Appendix I).

434 O_R1 (150bp) and 434 O_R (180bp) [see section 5.3.1, for description of DNA models, and Figure 5.2 for their sequences] were used in the experiments (Appendix I). The results are summarized in Figures 5.7 and 5.8.

Figure 5.7 shows the footprinting pattern of P7 and RR69 on 434 O_R1 (150bp). The footprinting patterns (lanes 2 and 4) are similar in every respect and correspond to the protection of the 14 bp O_R1 operator site. In both cases, the complete disappearence of the band corresponding to the central G residue of the TTGT half cI434 O_R1 operator is very clear (see the corresponding Maxam-Gilbert reactions in the same Figure 5.7 and Figure 5.2 for sequence of the DNA fragment). This indicates specific protein interaction between P7 and RR69 with their cognate DNA.

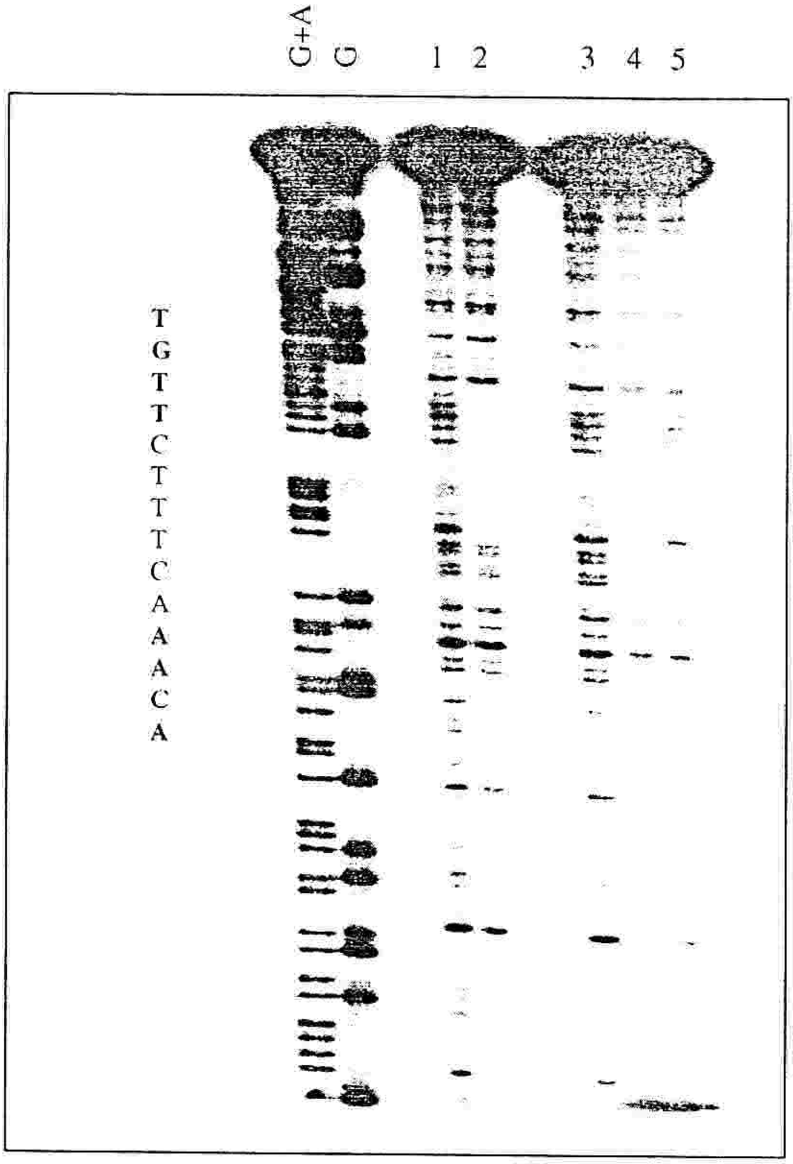


Figure 5.7. The DNAseI footprint of $434O_R1$ (150 bp) with P7 and RR69. Lanes 1,3 and 5, DNAseI control (no proteins present); lane2, P7 (final concentration 5.8 μM); lane 4, RR69 (final concentration 3.6 μM). The corresponding G and G+A Maxam and Gilbert sequencing reactions are also shown.

One drawback of the footprinting method is that DNAseI does not cleave all phosphodiester bonds to the same extent. The O_R1 region is apparently not a good substrate for DNAseI cleavage, therefore no strong bands are detected even without repressors (P7 or RR69). Nevertheless, the disappearence of the weak bands after binding clearly indicates sequence-specific interactions.

Figure 5.8 shows a footprinting experiment, carried out with the complete 434 O_R (180bp) segment with RR69. The figure shows that incremental addition of RR69 produces a gradual protection of the 3 operator sites. This is in parallel with the band shift experiments described in the previous section (Figure 5.5). Similar results were obtained with the chemical dimer P7 (data not shown). Incremental addition of proteins (final concentrations of P7 were from 0.325 μM to 6.2 μM; final concentrations of RR69 were from 0.12 μM to 3.6 μM) respectively showed occupation of the site of highest affinity (O_R1), followed by those for which the protein has a lower affinity (respectively $O_R2 > O_R3$). The result appears in full agreement with the gel mobility shift assays performed with the full operator DNA (434 O_R) and P7 (data not shown) or RR69 (Figure 5.5) in which several levels of occupancy are observed.

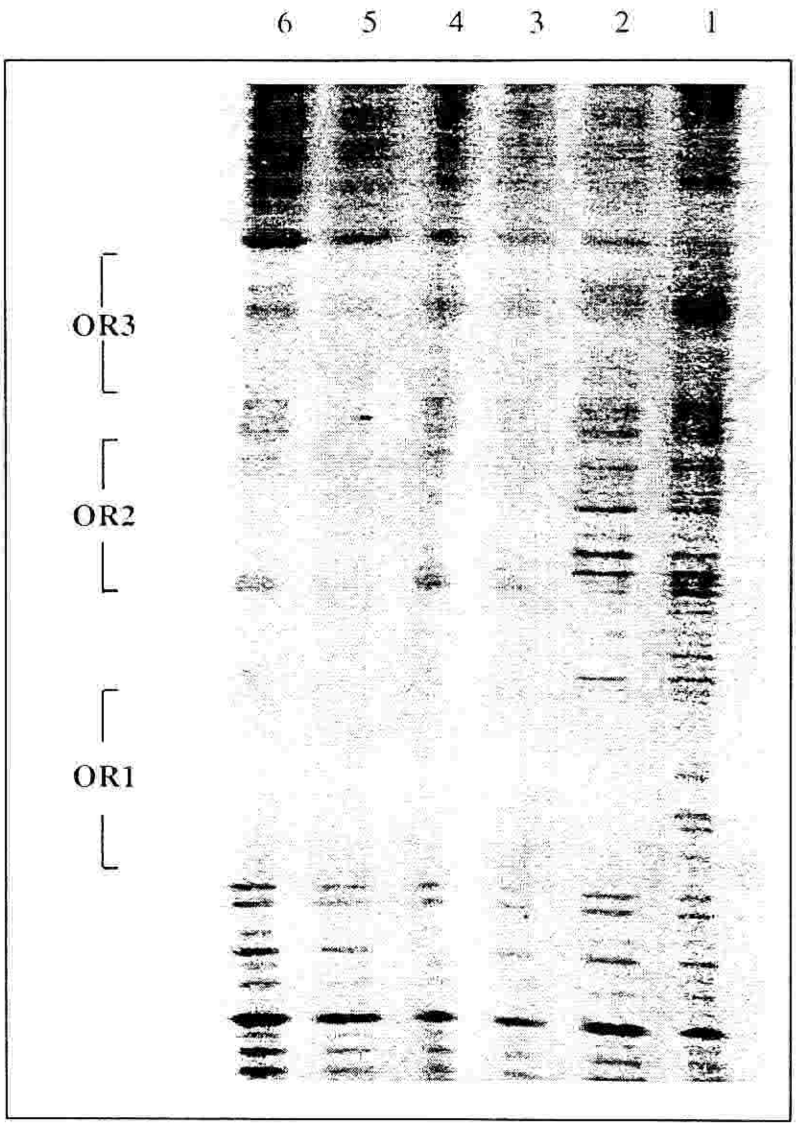


Figure 5.8. The DNAseI footprint of 434O_R (180 bp) with RR69. Lane 1, DNAseI control (no protein present); lane 2 to 6, incremental amounts of RR69 (final concentrations 0.12 μM, 0.24 μM, 0.6 μM, 1.2 μM, 3.6 μM, respectively)

5.3.3 DMS METHYLATION INTERFERENCE

Methylation with dimethylsulfate (DMS) renders phosphodiester bonds labile to cleavage by piperidine or alkali. Specifically DMS methylates (among other positions) the 7-nitrogen of guanine, which opens between carbon 8 and carbon 9 in a base-catalyzed reaction, and piperidine then displaces the ring-opened 7-methylguanine from its sugar. Piperidine goes on to catalyze β-elimination of phosphates from the empty sugar to finally break the DNA strand, leaving the sugar free (Maxam and Gilbert, 1980). When carried out in the presence of DNA-binding proteins, cleavage within the protected area will be less intensive so a specific footprint is obtained. We carried out methylation interference reactions in order to identify the points at which proteins P7 and RR69 contact their specific target sequence 434 O_R1 (150 bp).

The experiment was carried in such a manner that the protein-DNA complex was formed first and then methylated *in situ* (Papavassiliou, 1993 and references therein). For this purpose, gel mobility shift assays were performed with 434 O_R1 (150 bp) both in the presence and in the absence of peptide (either P7 or RR69, gels not shown), and subsequently, after pre-equilibration of the slab gel with phosphate buffer pH 8.0, both complexed and non-complexed DNA samples were *in situ* methylated. The *in situ* methylation reaction was allowed to proceed for a period of time sufficient to obtain, statistically, one methylation event per DNA strand. This time period is inversely proportional to the size of the DNA probe and was found to be 5 min. in our case. After quenching DMS with β-mercaptoethanol and DNA extraction from the gel, the modified samples were cleaved at high temperature (90°C) either with piperidine treatment or with alkali treatment. The methods give complementary information since piperidine cleaves only at G residues, whereas alkali cleaves both at G and A residues, with a preference to G (G > A).

As shown in Figure 5.9, both P7 and RR69 produce a clear protection pattern in the expected 14 bp operator region, indicating that both proteins impede methylation of the specific sequence to which they bind. In particular, the strong protection of the central G residue found in the half 434 O_R1 operator sequence TTGT, is especially conspicuous both with piperidine treatment and with alkali. This residue is apparently protected both by P7 and RR69. This is in good agreement with the DNAse digestion pattern discussed for both proteins in the previous section (Figure 5.7 and Figure 5.8).

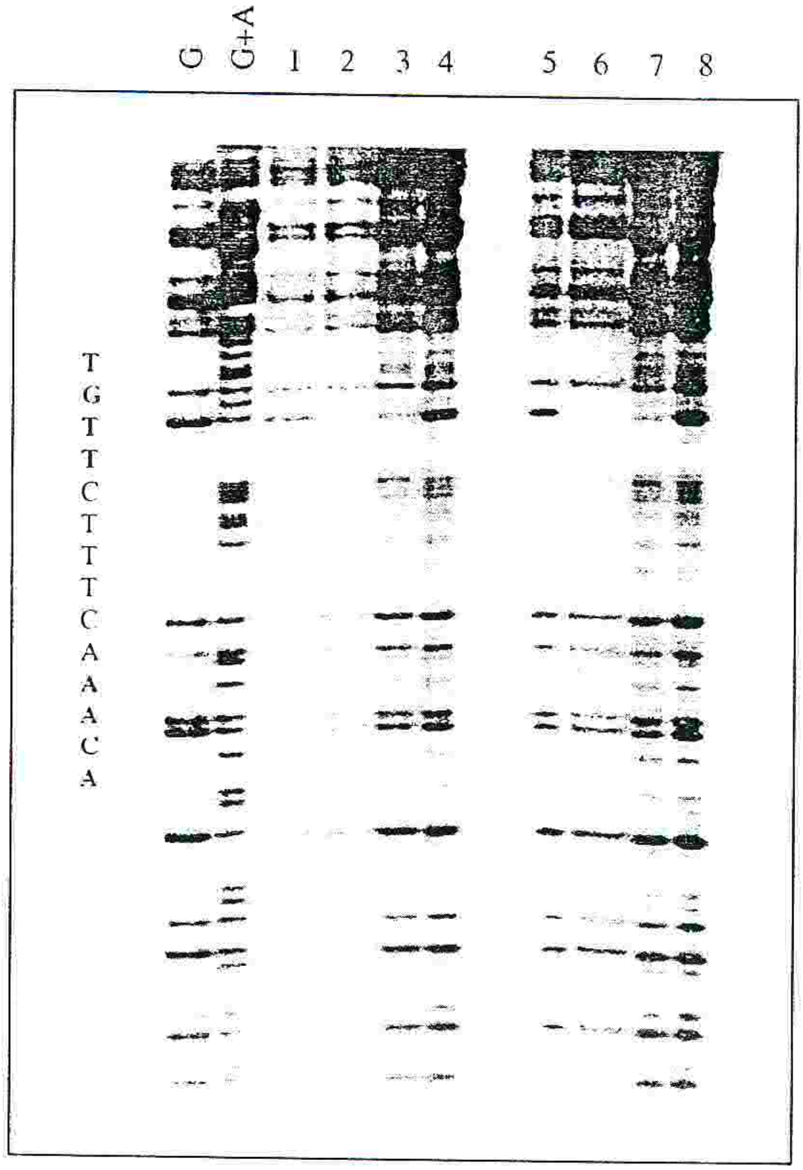


Figure 5.9. The methylation interference cleavage pattern of $434O_R1$ (150 bp) with P7 and RR69. Lanes 1 and 5, no proteins present (DNA cleavage carried out with alkali); lanes 4 and 8, no proteins present (DNA cleavage carried out with piperide); lanes 2 and 3 (DNA cleavage with alkali and piperidine respectively), P7 (final concentration 5.8 μM); lanes 6 and 7 (DNA cleavage with alkali and piperidine respectively), RR69 (final concentration 3.6 μM).

5.3.4 IN VITRO TRANSCRIPTION EXPERIMENTS

In vitro transcription experiments can be used to detect proteins that can inhibit the action of RNA-polymerase by binding to promoter regions (Fong *et al.*, 1993 and references therein). Since the specific binding sites O_R1, O_R2, O_R3 overlap with the promoters of cI and cro genes (Figure 1.2), this experiment can be used to test the action of the artificial single chain repressors. The DNA fragment 434 O_R (180bp) was obtained by PCR amplification (Appendix I). This fragment contains all the essential regulatory elements which are found in the bacteriophage 434 genome. It includes promoters P_R (rightward promoter) and P_{RM} (repressor maintenance promoter) which overlap the operator region and truncated forms of the cI and cro genes respectively. *In vitro* transcription assays were performed in two steps, (i) allowing first the formation of open complexes by preincubating RNA polymerase with the template DNA and then (ii) starting the reaction by addition of substrates (Fong *et al.*, 1993 and references therein).

Experiments performed without any repressor protein, yielded two RNA transcripts with a molecular weight corresponding to the expected truncated forms of the cI and cro genes (Figure 5.10). In the presence of both chemical dimer P7 and recombinant dimer RR69 no transcription was detectable (Figure 5.10, lanes 3 and 5, respectively). The monomeric P1 protein, on the other hand, was apparently not capable of inhibiting transcription in our system (Figure 5.10, lane 4). This is a clear-cut result which indicates that both dimeric proteins but not the monomeric N-terminal domain P1 interferes with transcription.

In conclusion, both P7 and RR69, but not P1, can be considered as functional models of the cI434 repressor protein as they have been shown to specifically interact with their cognate DNA and repress transcription *in vitro*.

As a control, *in vitro* transcription experiments were also performed with all short synthetic dimeric peptides. None of them was found to repress and therefore compete with RNA polymerase for the binding to the template DNA (Figure 5.10, lanes 6 to 7). These results are consistent with the binding studies carried out with all synthetic peptides (Chapter 3.0).

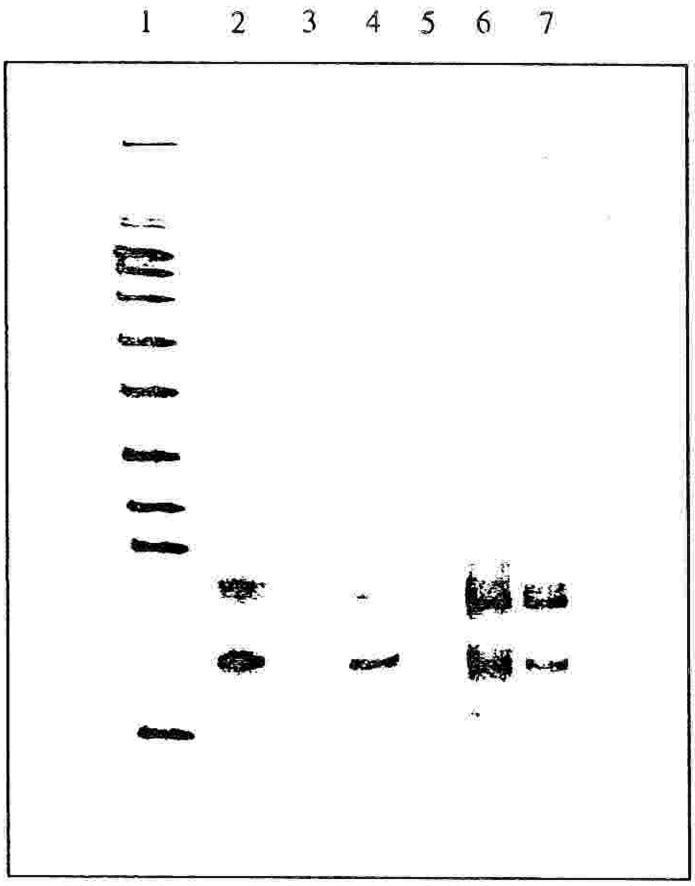


Figure 5.10. *In vitro* transcription assays at P_{RM} and P_R with $434O_R$ (180 bp). Lane 1, molecular weight marker; lanes 2, no protein or peptide present; lane 3, RR69 (final concentration 3.6 μM); lane 4, P1 (final concentration 6.6 μM); lane 5, P7 (final concentration 5.8 μM); lane 6, $H_1H_2TH_3$Ldim (final concentration 6.2 μM); lane 7, H_2TH_3Ldim (final concentration 6.5 μM).

5.4 CD SPECTROSCOPY

CD difference spectroscopy can be used to monitor the induction of α-helical structures occurring upon DNA-binding. CD studies were carried out on synthetic proteins P1 and P7, as well as on the recombinant RR69 protein. The oligonucleotide cI434 O_R1'(20bp) containing the O_R1 operator site was chosen as cognate DNA, while an unrelated oligonucleotide was chosen as non-cognate DNA. Sequences are shown in Figure 5.2.

CD spectra of P1, P7 and RR69 were obtained in phosphate buffer, both alone, and in the presence of cognate and non-cognate DNA. The CD difference spectra are shown in Figure 5.11. A drastic difference is apparent between the behaviour of the monomer P1 on the one hand, and the dimers P7 and RR69 on the other. Addition of cognate or non-cognate DNA hardly changes the spectra of the monomer. In contrast, the α-helical signal (222 nm) sharply increases upon DNA addition in both dimers, the cognate causing a larger increase than the non-cognate oligonucleotide. This shows that dimerization alone provides a basis for strong binding.

For a quantitative analysis of the spectra, the mean residue weight ellipticity, $[\Theta]_{MRW}$, was calculated from the observed values (Θ in mdeg) at 222 nm, according to the formula (Schmid, 1989 and references therein):

$$[\Theta]_{MRW} = \Theta \times 100 \times MRW \times c^{-1} \times d^{-1} \qquad (1)$$

where Θ is the measured ellipticity in degrees, c is the protein concentration in mg/ml, d is the path length in cm, and MRW is the mean residue weight. $[\Theta]_{MRW}$ has the units degrees x cm^2 x $dmol^{-1}$ and the factor 100 in the equation originates from the

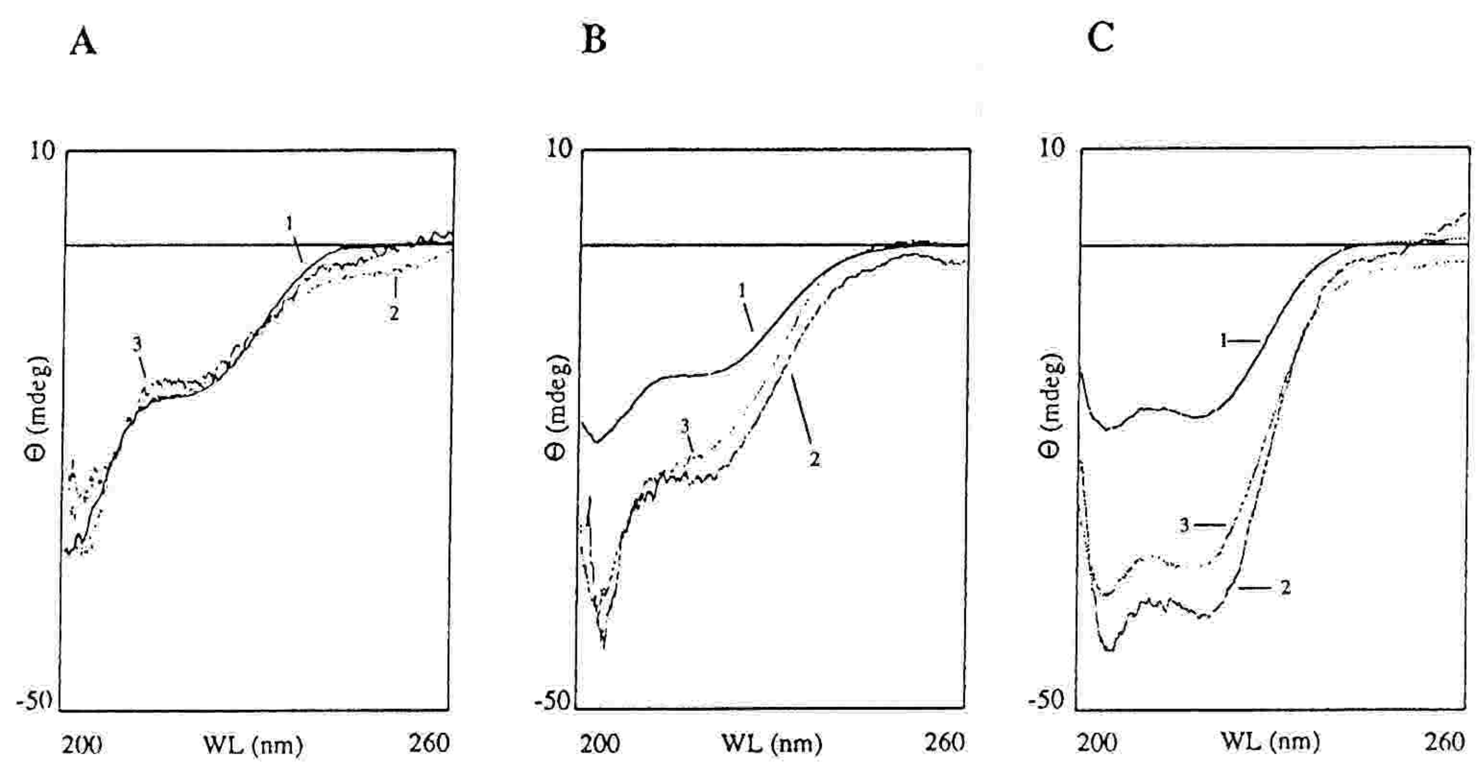


Figure 5.11. CD difference spectroscopy performed with P1, P7 and RR69 in the presence of either cognate [434O_R1' (20bp)] or non cognate (CL/WL) DNA. (**A**) P1 (7.2 μM) alone (line 1), P1 with cognate DNA (line 2), P1 with non-cognate DNA (line 3). (**B**) P7 (35 μM) alone (line 1), P7 with cognate DNA (line 2), P7 with non-cognate DNA (line 3). (**C**) RR69 (4.8 μM) alone, RR69 with cognate DNA, RR69 with non-cognate DNA.

conversion of the molar concentration to the dmol/cm^3 concentration unit. The percentage of induction can then directly be calculated according to the formula:

$$\% \text{ induction} = \frac{[\Theta]_{MRW,\ complex} - [\Theta]_{MRW,\ peptide}}{[\Theta]_{MRW,\ peptide}} \times 100 \qquad (2)$$

The data are shown in Table 5.1. The percentage of α-helical content was calculated according to the approximate formula (Wu *et al.*, 1981)

$$\%\alpha_H = \frac{\Theta_{Obs} + \Theta_C}{\Theta_H + \Theta_C} \qquad (3)$$

where Θ_{Obs} is the calculated mean residue ellipticity, Θ_C is the mean residue ellipticity calculated for a coil at 222 nm (- 2000 degrees x cm^2 x dmol^{-1}) and Θ_H represents the mean residue ellipticity of an infinite α-helix (- 37000 degrees x cm^2 x dmol^{-1}). This formula is approximate, because it is based on synthetic peptides, so it is used here only on a comparative basis.

The extent of α-helical induction, i.e. the change in mean residue ellipticity values (Table 5.1) confirms the same tendencies as discussed previously. No significant induction is shown in the monomer P1 and substantial increases in P7 and RR69. It is apparent, that helical induction in RR69 is stronger than in P7. The effect of cognate and non-cognate DNA can also be quantitatively compared by calculating a ratio of induction percentages. In the case of P7 we find that the induction caused by non-cognate DNA is 69 % of the cognate induction, while for RR69 the same ratio is

74 %. So, in both cases, non-cognate DNA induces around 70 % of the change caused by cognate DNA (Percipalle *et al.*, 1994; manuscript in preparation).

The percentages of α-helical content were calculated only as a comparison. It appears that the monomer P1 is somewhat more helical than the dimer P7, and the dimer RR69 has the highest apparent helical content. However, we are not certain if these differences can be evaluated on a structural basis. First, the formula used to calculate elipticity is of approximate nature. Second, P1 and P7 were prepared by chemical synthesis, while RR69 was prepared by recombinant DNA methods, so the relatively low α-helical percentage of P1 and P7 may be due to the harsh reaction and purification conditions used in chemical synthesis.

Table 5.1. CD data obtained for proteins P1, P7 and RR69 in aqueous buffer with and without cognate or non-cognate DNAs.

[a]Sample	[b]Θ_{Obs}	[c]$[\Theta]_{MRW}$	[d]% helical induction	[e]$\%\alpha_H$
*P1	- 0.01940	$- 2.48 \times 10^3$	0	11.5%
†P1+cognate	- 0.02090	$- 2.68 \times 10^3$	4	12.0%
‡P1+aspecific	- 0.01970	$- 2.52 \times 10^3$	1	11.6%
*P7	- 0.01472	$- 1.62 \times 10^3$	0	9.3%
†P7+cognate	- 0.03181	$- 3.51 \times 10^3$	52	14.1%
‡P7+aspecific	- 0.02720	$- 2.90 \times 10^3$	38	12.8%
*RR69	- 0.008212	$- 5.37 \times 10^3$	0	18.9%
†RR69+cognate	- 0.0188	$- 1.10 \times 10^4$	76	33.3%
‡RR69+aspecific	- 0.01608	$- 8.84 \times 10^3$	47	27.8%

[a] Peptides are denoted according to Figure 3.8. [P1]=7.2μM; [P7]=35μM; [RR69]=4.8μM

[b] Θ_{Obs} is the observed value of ellipticity calculated in degrees.

[c] $[\Theta]_{MRW}$ is the mean residue ellipticity calculated according to equation (1) in degrees x cm^2 x $dmol^{-1}$.

[d] Percentage α-helical induction is calculated according to equation (2).

[e] $\%\alpha_H$ is the percentage α-helical content calculated according to equation (3).

* Proteins alone in 10mM phosphate buffer pH 7.0, 100mM NaCl.

† Proteins in the presence of cognate DNA in the same buffer as above.

‡ Proteins in the presence of non-cognate DNA in the same buffer as above.

6 DISCUSSION AND CONCLUSIONS

Synthetic peptides capable of sequence specific DNA binding have been designed by several groups, based on the bZIP motif (Cregut *et al.*, 1993; Saudek *et al.*, 1991; Patel *et al.*, 1990; O'Neal *et al.*, 1990; Talanian *et al.*, 1990; Vinson *et al.*, 1989), the bHLH motif (Anthony-Cahill *et al.*, 1992) as well as on the zinc finger motif (Kriwacki *et al.*, 1992). Attempts to prepare similar peptides on the basis of HTH containing DNA binding proteins have not yet been successful (Benevides *et al.*, 1991, Grokhovsky *et al.*, 1991)

In this work, our goal was to design novel dimeric peptides based on the HTH motif and to study their behaviour towards specific and non-specific DNA sequences. As these peptides contain two repressor-derived DNA-binding domains within one polypeptide chain, we termed them single-chain repressors. As a criterion for specific DNA binding we have used a positive gel mobility shift assay in the presence of a large excess of aspecific competitor DNA together with DNAse footprinting, methylation interference and inhibition of *in vitro* transcription. CD difference spectroscopy was used to study helical transitions occurring upon DNA binding.

As the monomeric cI434 amino-terminal domain does not bind to its half operator site, we decided to design covalently dimerized molecules. For this purpose we have considered two design strategies. In the first, "minimalist" approach, we combined various structural elements of the cI434 amino-terminal domain into branched symmetric peptides. Essentially, none of these showed specific binding in the presence of competitor DNA, nor was there any detectable conformational change in the presence of cognate or non-cognate DNA. In some cases, weak binding could be detected using photochemical crosslinking experiments, but this affinity is definitely negligeable as compared to the high binding affinity of natural DNA binding proteins.

These results are negative, but they do not exclude the possibility that other linker constructs (eg. more flexible or longer linker elements) will eventually allow the synthesis of short HTH containing peptides capable of specific binding to DNA with high affinity.

The second strategy involved the dimerization of the entire cI434 amino-terminal domain through two different linkers. First, a short synthetic linker (~ 35 Å) was used to connect the carboxy-termini of two cI434 amino-terminal domains (amino acid residues 1 to 62). The resulting branched molecule, obtained by continuous flow solid phase peptide synthesis and termed **P7**, possesses a dyad symmetry in its sequence which mimics the palindromic DNA site. Second, two cI434 amino-terminal domains were connected as direct repeats with a longer linker (approximately 96 Å), obtained from the sequence of the same native cI434 repressor (amino acid residues 63 to 96). This molecule, termed **RR69**, was prepared by recombinant DNA methods. The bipartite structure of **P7** and **RR69** mimics the modular organization of the POU specific homeodomain of Oct-1 (Klemm *et al.*, 1994; Dekker *et al.*, 1993) and Oct-2 (Sivaraja *et al.*, 1994), except that they contain two identical, specific DNA-binding modules, while the two modules in POU proteins show different DNA binding specificities.

Both **P7** and **RR69** showed specific DNA binding activity as detected by standard biochemical techniques. Their behaviour in DNAse footprinting, methylation interference and *in vitro* transcription experiments was indistinguishable, and corresponded to the behaviour of the cI 434 repressor. Gel-mobility shift assays with cognate DNA showed a K_d of 3.25×10^{-6} M for **P7** and 5×10^{-8} M for **RR69**. In these experiments, no association could be detected with non-cognate DNA.

CD data showed an increase in the α-helical content of both **P7** and **RR69** upon exposure to cognate DNA, and a similar, albeit 30% smaller conformational change was found with non-specific DNA (Percipalle *et al.*, 1994; manuscript in preparation).

The specific DNA binding activity found for both bipartite, single chain repressors we designed is noteworthy itself. The fact that dimerized cI434 amino-terminal domains allowed the monitoring of conformational changes upon DNA binding is also relevant since the monomeric amino-terminal domain, which can be co-crystallized with the DNA, does not permit such measurements. It is also believed that HTH containing repressors are more rigid than other DNA binding proteins (eg. bZIP and bHLH proteins) and do not undergo substantial conformational changes upon DNA binding (Johnson *et al.*, 1994; Harrison and Aggarwal, 1990; Berg and von Hippel, 1988; Berg and von Hippel, 1987; von Hippel and Berg, 1986; von Hippel *et al.*, 1974). Particularly, it has been suggested that non-specific DNA binding by HTH proteins is essentially a rigid-body docking (Spolar and Record Jr., 1994 and references therein). In contrast, our data show that α-helical induction does take place upon both specific and non-specific DNA binding. We speculate that this flexibility is in fact necessary when a protein uses the scanning mechanism for approaching a target DNA site using one-dimensional linear diffusion (von Hippel and Berg, 1989; Berg and von Hippel, 1985; Richter and Eigen, 1974; Adam and Delbruck, 1968). This mechanism is different from the ATP-driven one adopted by type I restriction endonucleases (Studier and Bandyopadyay, 1988) and it has already been described for some HTH proteins such as the lac repressor (Barkley, 1981; Winter and von Hippel, 1981; Winter *et al.*, 1981) and the λ Cro protein (Kim *et al.*, 1987). Whatever the underlying mechanism is, our results suggest that induced conformational complementarity is characteristic also to the HTH based domains (and not only to the bZIP and bHLH ones) and it therefore appears to be a general phenomenon in protein-DNA recognition. The fact that other workers did not pay attention to this phenomenon may be simply due to the fact that their conformational studies were conducted on proteins which undergo a major change upon DNA-binding, so the concomitant stabilization of the binding helix could easily go undetected.

The fact that non-specific DNA produces α-helical induction which is 70 % of that produced by cognate DNA can be interpreted, in principle, in two ways: i) similar binding affinity but weaker conformational change, or ii) weaker binding and similar conformational change. Since the binding to aspecific DNA must be much weaker, as shown by electrophoretic mobility shift assays, in our view the second possibility seems more plausible.

We can therefore try to outline a coherent scenario for the binding of HTH motifs to DNA. First, the protein approaches DNA guided by electrostatic interactions, and displaces the water and counter-ion molecules around the DNA molecule. This first, approximate docking, takes place at an aspecific site, with the binding helix already exploring the major groove (cf. the observed conformational change). The interactions (hydrophobic effect, aspecific backbone-to-phosphate links) are not strong enough to fix the molecule in place, so the protein will dissociate but still remaining in the electrostatic shell of DNA. The next docking takes place in a similar fashion, up to a point, when specific hydrogen bonds (well known from X-ray crystallographic analyses) stabilize a specific binding at a cognate site. This mechanism requires a great "conformational plasticity" from the HTH binding helix, i.e. it has to be able to explore all kinds of DNA sequences in a search for stabilizing hydrogen bonds. This flexibility can be seen by the CD experiments hereby reported. Even though this model is theoretical, it has some common features with other ligand-binding phenomena. Binding of ligands is described, generally speaking, either as key-lock interactions or rigid body docking (Amit *et al.*, 1986), or as an "induced fit" (Koshland, 1958), or "handshake" mechanism (Colman, 1986). The former is an attractive model for the recognition of rigid, small molecular substrates by enzymes. The latter seems more rational to protein-protein and antibody-hapten interactions.

Spolar and Record Jr. (1994) convincingly argue that most protein-protein interactions and many DNA-protein recognition processes fall into the latter, "induced fit" category. Our present data here suggest, that the conformational flexibility of a recognition domain, like the HTH domain, may be a prerequisite for a molecule to be able to scan a large number of sites or conformations. In antibodies, this ability is given by the conformational flexibility of the binding loops. In DNA-binding proteins, it comes from the flexibility of the binding modules, which is substantial even when the module's overall structure is essentially rigid.

APPENDIX I

Materials and methods

I.1 Preparation and characterization of DNA fragments corresponding to cI434 operator sites

Oligonucleotide synthesis

Oligonucleotides were synthesized on an Applied Biosystem model 380A DNA synthesizer and purified by gel electrophoresis on polyacrylamide gel containing 8 M urea. Concentrations were determined using the calculated molar extinction coefficients (Sambrook *et al.*, 1989) at 260 nm. Double-stranded oligonucleotides were obtained after mixing equimolar amounts of the corresponding single-stranded oligonucleotides and annealing was performed at 65°C followed by slow cooling (approximately 1°C/min) to room temperature.

The sequences of 434 O_R1 (150bp), 434 O_R (180bp) as well as those of the short oligonucleotides 434 O_R1' (20bp) and CL/WL used for CD studies are reported in Figure 5.2.

Preparation of a Dna fragment containing phage 434 O_R1 site

This DNA fragment (150 bp) was produced by PCR amplification of a region of plasmid pRIZ'O_R1, which contains the cloned synthetic 14 bp O_R1 operator of phage 434. PCR primers 5'-TAGCTCACTCATTAGGCACC (AT404) and 5'-CCAGGGTTTTCCCAGT (ol. 181) define a 150 bp target on the plasmid template and the resulting DNA fragment contains the cloned O_R1 operator site approximately in the middle. The length and base composition of the PCR primers were chosen to achieve identical T_m annealing temperatures for a primer pair. T_m values were calculated as described (Saiki *et al.*, 1988; Mullis and Faloona, 1987; Saiki *et al.*, 1985).

Specific labeling of one of the strands of the DNA was achieved by performing the PCR amplification with one labeled and one unlabeled primer. Therefore primer 181 was labeled with ^{32}P-phosphate using T4 polynucleotide kinase before the PCR amplification was performed. 10 pmol of primer 181 was phosphorylated in 5 μl reaction volume containing 2.5 μl γ-^{32}P-ATP (> 5000 Ci/mmol, 10 $\mu Ci/\mu l$) as described (Sambrook *et al.*, 1989). After heat inactivation of the kinase at 90°C for 5 minutes, the reaction volume was adjusted to 20 μl with water to obtain 0.5 μM 5'-^{32}P-primer 181.

PCR amplification was performed in 100 μl reaction mixture containing 0.4 ng plasmid DNA, 2.5 mM dNTPs, 4% DMSO, 10 mM Tris-HCl pH=8.3, 50 mM KCl, 8 mM $MgCl_2$, 0.001% gelatin, 0.02 μM 5'-^{32}P-primer 181, 0.4 μM primer 181, 0.4 μM primer AT404 and 5 units of TAQ DNA polymerase. The mixture was overlayed with mineral oil. Initial denaturation was performed at 95°C for 5 minute, then 25 cycles of 95°C for 1 minute, 58°C annealing for 1 minute and 72°C chain elongation for 1 minute were performed followed by chain elongation after the last cycle at 72°C for 10 minutes.

Analysis of a sample of the PCR product was performed by electrophoresis on either 1% agarose followed by ethidium bromide staining or on polyacrylamide followed by autoradiography. The reaction mixture was then extracted with phenol/chloroform and the DNA was precipitated with 2.5 volumes of ethanol followed by centrifugation. The DNA pellet was washed with 70% ethanol, dried and dissolved in water to obtain a concentration of approximately 0.3 mM.

Preparatioin of a DNA fragment containing the full O_R region of phage 434

The full rightward operator region containing all three operator sites O_R1, O_R2 and O_R3 was obtained as a part of a 186 bp DNA fragment, which was prepared by PCR amplification of a corresponding region of the phage vector λgT10 (Boehringer Mannheim) which contains the immunity region of the 434 phage (Huynh *et al.*, 1985) using PCR primers 5'-CGGTTTGCGTCATTTTTAAC (ol. 434) and 5'-TTACCCTGGAAGAAAT (AT429).

5'-end labeling of one of the strands of the PCR product, PCR amplification, analysis and purification were performed as described above. In this case primer ol.434 was labeled prior to PCR amplification. Annealing temperature for this primer pair was 54°C.

Maxam-Gilbert Sequencing of the 434 O_R1 and 434 O_R fragments

This was performed as described by Maxam and Gilbert (1980)

I.2 Synthesis of short, dimeric peptides

I.2.1 Synthesis and characterization of H_2TH_3

Synthesis protocol

The synthesis of H_2TH_3 (Figure 3.8) was performed by the automated continuous flow solid phase method (SPPS) on a Milligen 9050 Pepsynthesizer. *In situ* activated chemistry (HOBt/TBTU) was used (Atherton and Sheppard, 1989) in the presence of a 4 molar excess of Fmoc-AA-OH and double coupling at every step. The peptide chain was elongated on 1.2 g of the acid labile Fmoc-Gly-Pepsyn-KA resin (Kieselguhr, Milligen).

Cleavage and deprotection

After completion of the synthesis, the resin was washed with isoamyl alcohol, acetic acid, isoamyl alcohol and diethyl ether and then freeze-dried. After liophylization the peptide was cleaved from the resin by treatment with a solution containing TFA/Phenol/EDT/TA/Tris-isopropylsilane/water (90:2:2:2:2:2 %) and the reaction was allowed to proceed for about two hours at room temperature. The resin was then filtered on a sintered funnel and washed with the cleavage mixture and then few times with TFA. The filtrate was collected in a round-bottomed flask and evaporated at 40°C under diminished pressure. The dry crude material was dissolved in water/acetic acid (1:1), extracted several times with diethyl ether and the aqueous phase freeze-dried.

Quantification of the yield of the synthesis was carried out spectrophotometrically. For this purpose the absorbance of a diluted solution of the peptide was measured at different wavelengths (i.e. λ=280 nm and λ=205 nm) in order to evaluate the extinction coefficient (1 mg/ml) at 205 nm according to the equation ε=27.0/1-3.85(A280/A205), where A280 and A205 are the absorbances measured, respectively, at 280 nm and 205 nm (Scopes, 1974). The calculated extinction coefficient was then used to calculate the total actual yield which was estimated at 83 mg.

Purification and characterization

The crude peptide was analyzed and purified by RP-HPLC on a Waters HPLC system equipped with a Waters 484 UV detector. For analytical runs a Delta Pak C-18 300Å (3.9mm x15cm) column and a linear gradient was used from buffer A (0.05% TFA in water) to buffer B (90% acetonitrile in buffer A) in 30 minutes with a flow rate of 1 ml/min. The HPLC spectrophotometer was set at 214 nm. The same gradient was used for the purification of the crude material on a semipreparative column, Delta Pak C-18 300Å (19mm x30cm), with a flow rate of 10 ml/min (Table 4.1).

Amino acid analysis on H_2TH_3 was performed using a custom made reactor, using the Waters PicoTag system according to the protocol reported in the Waters instruction manual.

I.2.2 Synthesis and characterization of H_2TH_3L

Synthesis protocol

H_2TH_3L (Figure 3.8) was synthesized on a solid phase support using the Fmoc strategy (Atherton and Sheppard, 1989) on a Milligen 9050 Pepsynthesizer by automated continuos flow. The synthesis was carried out on the acid labile Fmoc-Leu-Pepsyn-KA resin (strong acid labile Kieselguhr resin from Milligen) using 0.94 g of resin (0.08 mmole/g) and a 6 fold excess of each Fmoc-AA-OH protected on the side chains so as to avoid secondary reactions. A mixture of active ester (OPfp or ODhbt) and in situ activated (HOBt/TBTU) chemistry was used.

Cleavage and deprotection

After the synthesis the resin was washed with isoamyl alcohol, acetic acid, isoamyl alcohol and diethyl ether and then liophylized. The peptide-resin was treated with a mixture containing TFA/EDT/Phenol (92.5:5:2.5%) and the cleavage and deprotection reaction allowed to proceed for 3 hours at room temperature. The cleavage mixture was separated from the resin by filtration on a sintered funnel and the resin further washed with fresh cleavage solution and consecutively with small aliquots of TFA. The TFA was then removed by evaporation, the precipitate was dissolved in a mixture containing water/acetic acid (1:1) and finally extracted several times with equal volumes of diethyl ether. After extraction the aqueous phase was lyophilized, the dry material dissolved in a minimum volume of water and spectrophotometrically quantitated (Scopes, 1974). The total actual yield was thus evaluated to be 68mg.

Purification and characterization

Analysis of the crude H_2TH_3L peptide was carried out on a RP-HPLC analytical column, Delta Pak C-18 300Å (3.9mm x15cm), and eluted by running a linear gradient from buffer A (0.05% TFA) to buffer B (90% acetonitrile in buffer A) in 45 minutes with a flow rate of 1 ml/min and detected on the HPLC spectrophotometer at 214 nm. The same gradient as above was utilized for the purification on a semipreparative Delta Pak C-18 300Å (19mm x30cm) RP-HPLC column with a flow rate of 10 ml/min (Table 4.1). The fraction corresponding to the main peak was collected and used to carry out amino acid analysis on a Waters PicoTag workstation. H_2TH_3L was also characterized by mass spectrometry. Its exact molecular weight was determined according to the standard methodology described in details in the case of H_2TH_3Ldim (Percipalle *et al.*, 1994b; Moore, 1993)

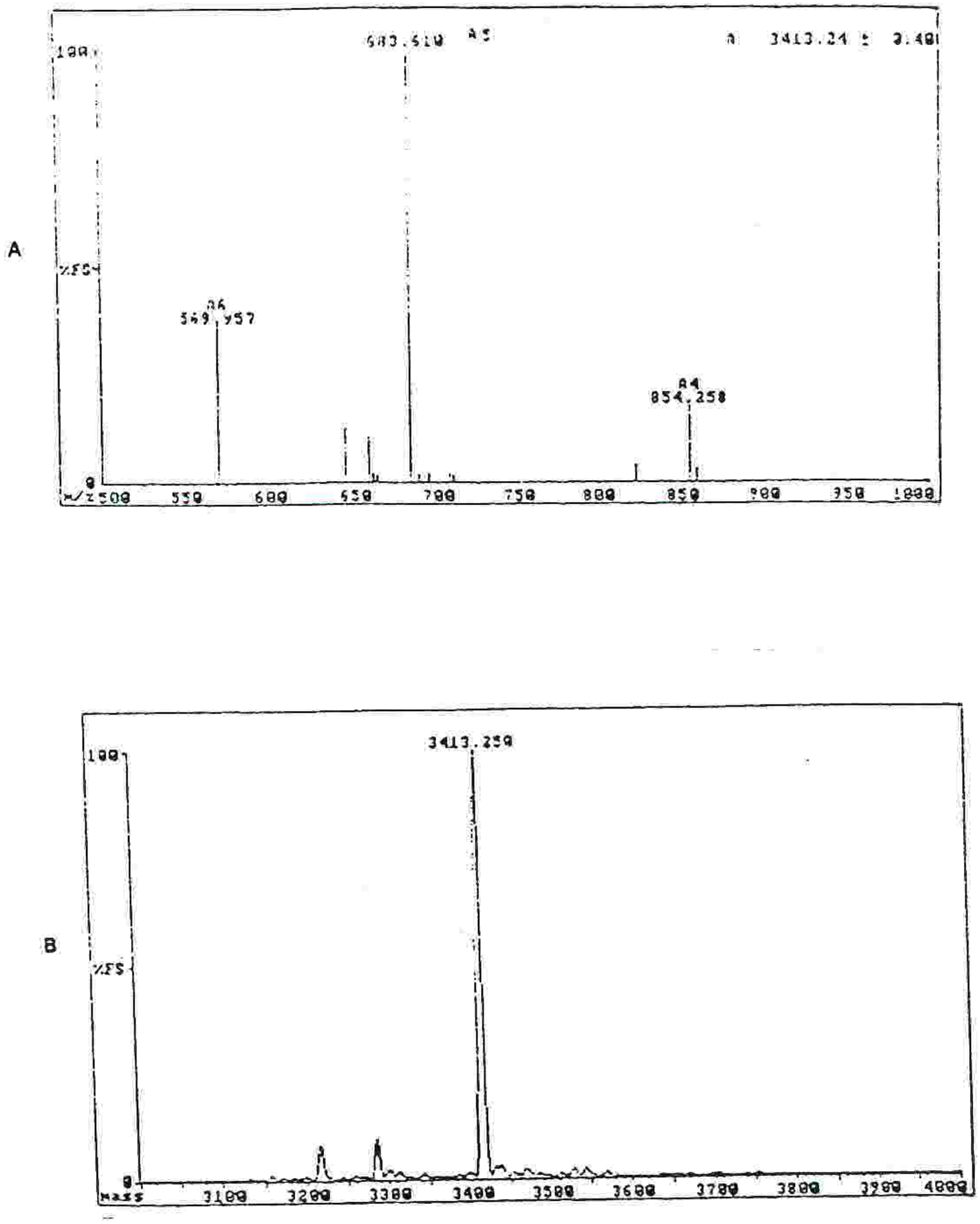


Figure I.1. ES ionization mass spectrum of H_2TH_3L. (A) Spectrum produced by the mass analyzer; (B) deconvoluted spectrum (see section I.2.2).

consisting of the tryptic digestion of H_2TH_3L and FAB-MS on the various fragments. The peptide was also characterized by ES-MS (Figure I.1 and Table 4.1).

I.2.3 Synthesis and characterization of H_2TH_3Ldim

Synthesis protocol

H_2TH_3Ldim (Figure 3.8) was synthesized on a Milligen 9050 Pepsynthesizer by using the Fmoc chemistry (Atherton and Sheppard, 1989). As H_2TH_3Ldim is relatively long (30 residues per strand), the synthesis was carried out on a PAC-PEG-PS resin which permits greater accessibility to incoming aminoacids and reagents, and has better solvation properties than Kieselguhr-based resins. 0.400 g of Fmoc-Gly-PEG-PS resin (0.190 meq/g) were used for the synthesis, which was carried out on a 0.152 mM scale. A six-fold aminoacid excess with equimolar quantities of HOBT and TBTU was utilized according to three differentiated sets of automated cycles. During the first, Fmoc-Lys(Fmoc)-OH was anchored to the solid support to provide the core for the branching of H_2TH_3Ldim. The second set consisted in the coupling of the first two Pro residues. For this purpose the deprotection step, concerning the removal of the two Fmoc groups protecting the α and ε amino groups of the Lys core, was very fast to avoid the formation of dichetopiperazine as non-desired product (given the presence of Gly as the first residue). The third set consisted in the elongation of the branched peptide. The synthesis was interrupted at the level of the helix H_3 and a small amount of resin removed, so as to obtain the truncated peptide H_3Ldim.

Cleavage and deprotection

After the synthesis the PEG-PS resin was washed several times with methanol and then diethyl ether and vacuum dried. The resin was then treated with the cleavage mixture TFA/Phenol/EDT/TA/Tris-isopropylsilane/H_2O (90:2:2:2:2:2 %), and the

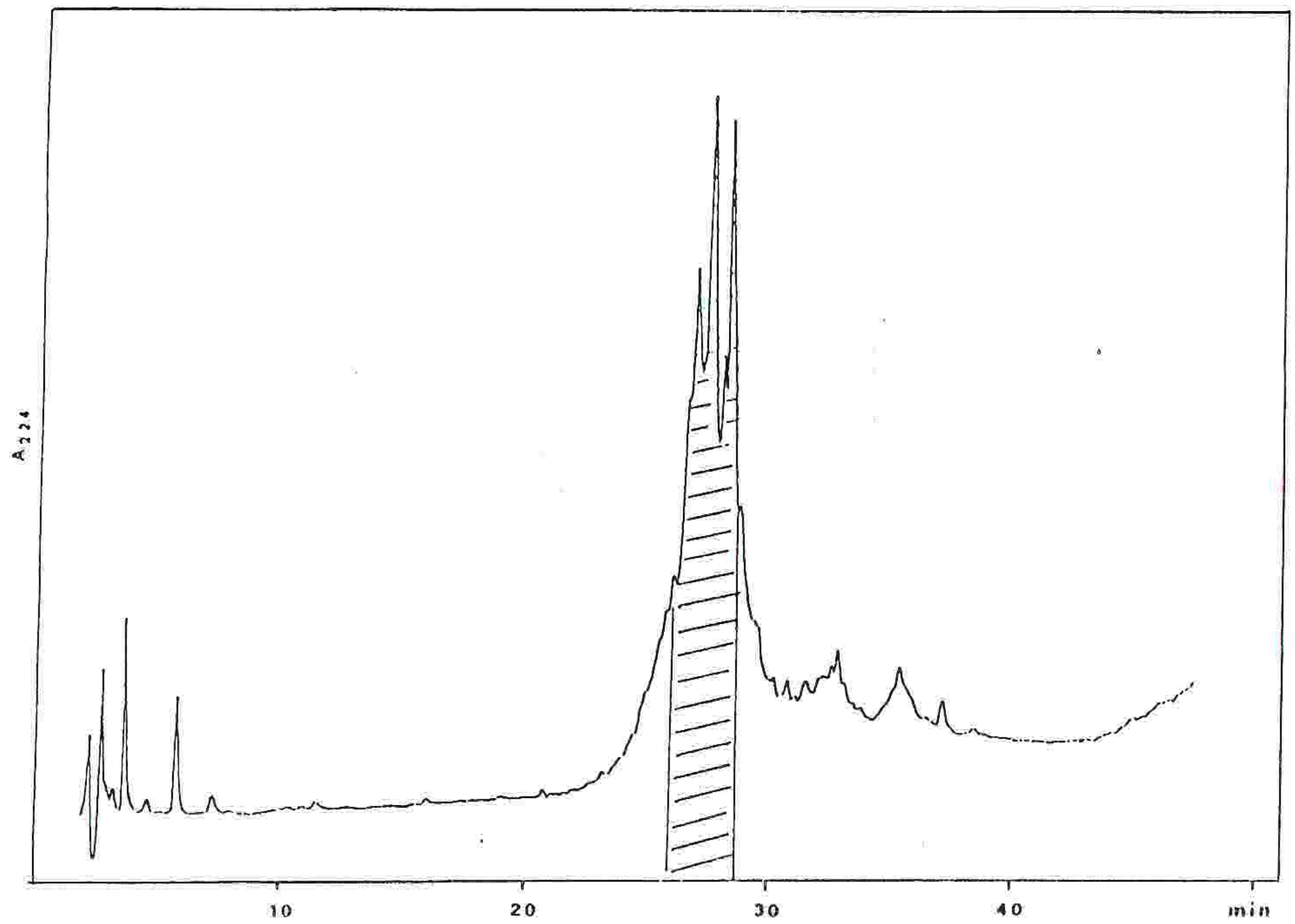


Figure I.2. HPLC detector traces of the crude H_2TH_3Ldim peptide (see section I.2.3).

reaction allowed to proceed for about 4 hours at room temperature. The resin was then filtered on a sintered funnel and further washed with cleavage mixture and few times with TFA. The filtrate was collected in a round-bottomed-flask and evaporated at 40°C. The dry crude material was then dissolved in a mixture of water/acetic acid (1:1) and extracted several times with diethyl ether. The aqueous phase was collected and lyophilized. The yield of the synthesis was then spectrophotometrically assayed (Scopes, 1974).

Purification and characterization

The HPLC purification of H_2TH_3Ldim was carried out on a Varian 9010 chromatographer equipped with a Varian 9010 detector. The sample (8 mg) was dissolved in 0.1% TFA, filtered on Millex-HV filters (Millipore) and injected in a Vydac 5RP18 (1.0x25cm), 300Å pore size RP-HPLC column (Table 4.1). The column was initially eluted at room temperature with a mixture of 10% B (acetonitrile/isopropanol0.05% TFA) in 90% A (0.05% TFA) for 5 minutes. The sample was then eluted with a linear gradient until 45% B in 40 minutes with a flow rate of 3 ml/min. The detector was set at 224 nm. The three main peaks were collected together, as shown in Figure I.2, lyophilized and subjected to tryptic digestion. Similar conditions were used to purify H_3Ldim (Table 4.1).

Dry sample (4.7 mg) was then dissolved in 0.02 M ammonium acetate containing 0.001 M calcium chloride, pH=8.3. Trypsin was added with an enzyme/substrate ratio of 1:50 (mol/mol) and the sample incubated for 4 hours at 37° C. The enzymatic reaction was stopped by freezing the sample which was then lyophilized and redissolved in 0.1% TFA, filtered and injected. The HPLC separation of the tryptic fragments was carried out on the same Vydac 5RP18 (1.0x25cm) column used for the sample purification. The column was first eluted at room temperature with a mixture of 5% B (acetonitrile/5% water/0.08% TFA) in 95% A

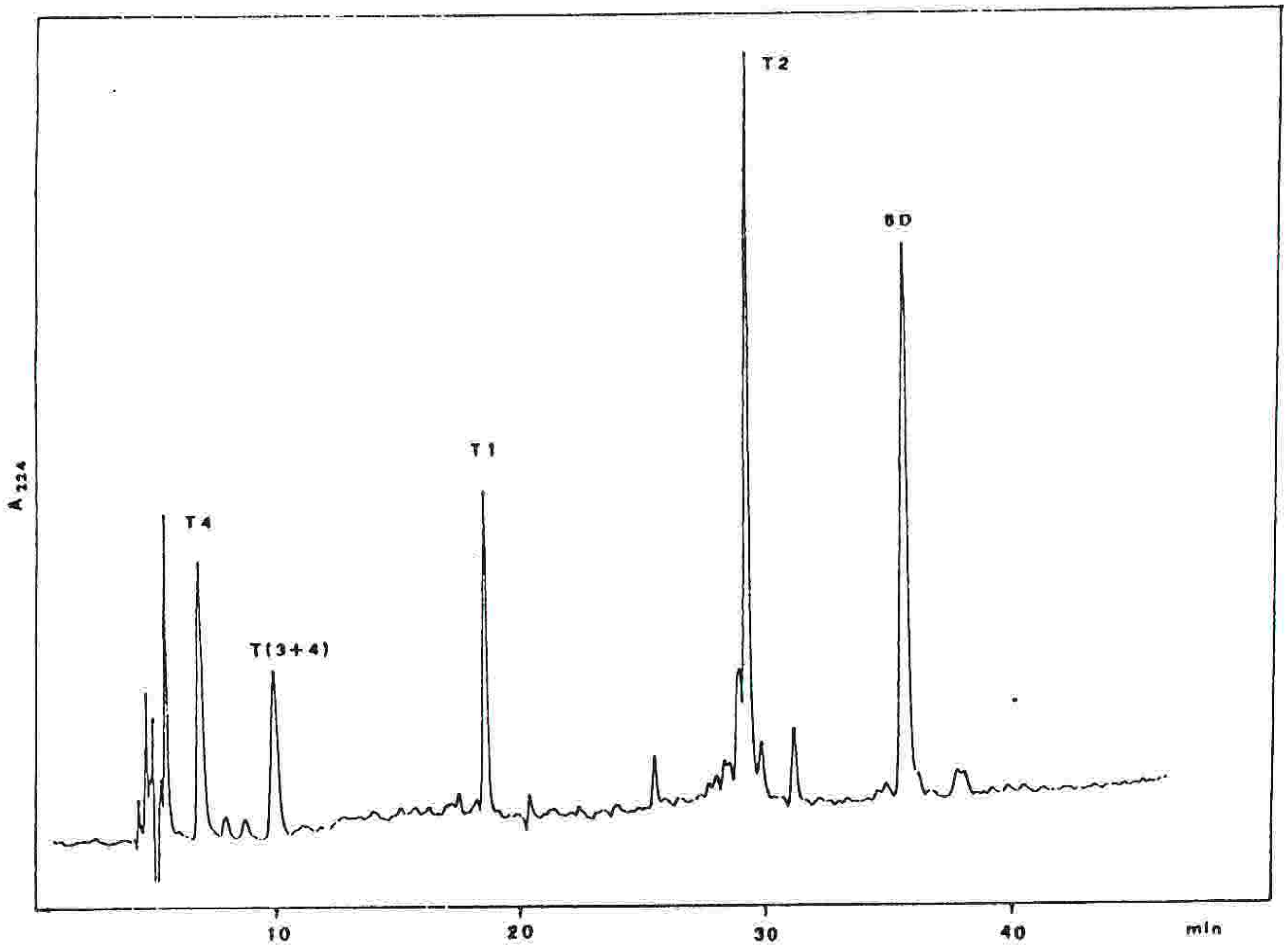


Figure I.3. HPLC separation of the tryptic digest of H_2TH_3Ldim peptide (see section I.2.3)

(0.1% TFA) for 5 minutes and then with a linear gradient until 47% B in 55 minutes. The flow rate was 3 ml/min and the wavelength was set at 224 nm. The chromatographic profiles are shown in Figure I.3. All five peaks collected were freeze-dried and analyzed by FAB-MS.

The mass spectra were obtained by a VG ZAB-2SE mass spectrometer equipped with a FAB source emitting cesium ions. Lyophilized samples were placed on top of the probe for evaporation and then about 2 μl of glycerol were added to solubilize them. Samples were then bombed with a cesium ion beam of 35 KeV. The spectra show that all of the five peaks collected during the HPLC runs correspond to the theoretically predicted tryptic fragments. H_2TH_3Ldim was also characterized by ES-MS and Figure I.4 shows the ES ionization mass spectrum (Percipalle *et al.*, 1994b; Moore, 1993).

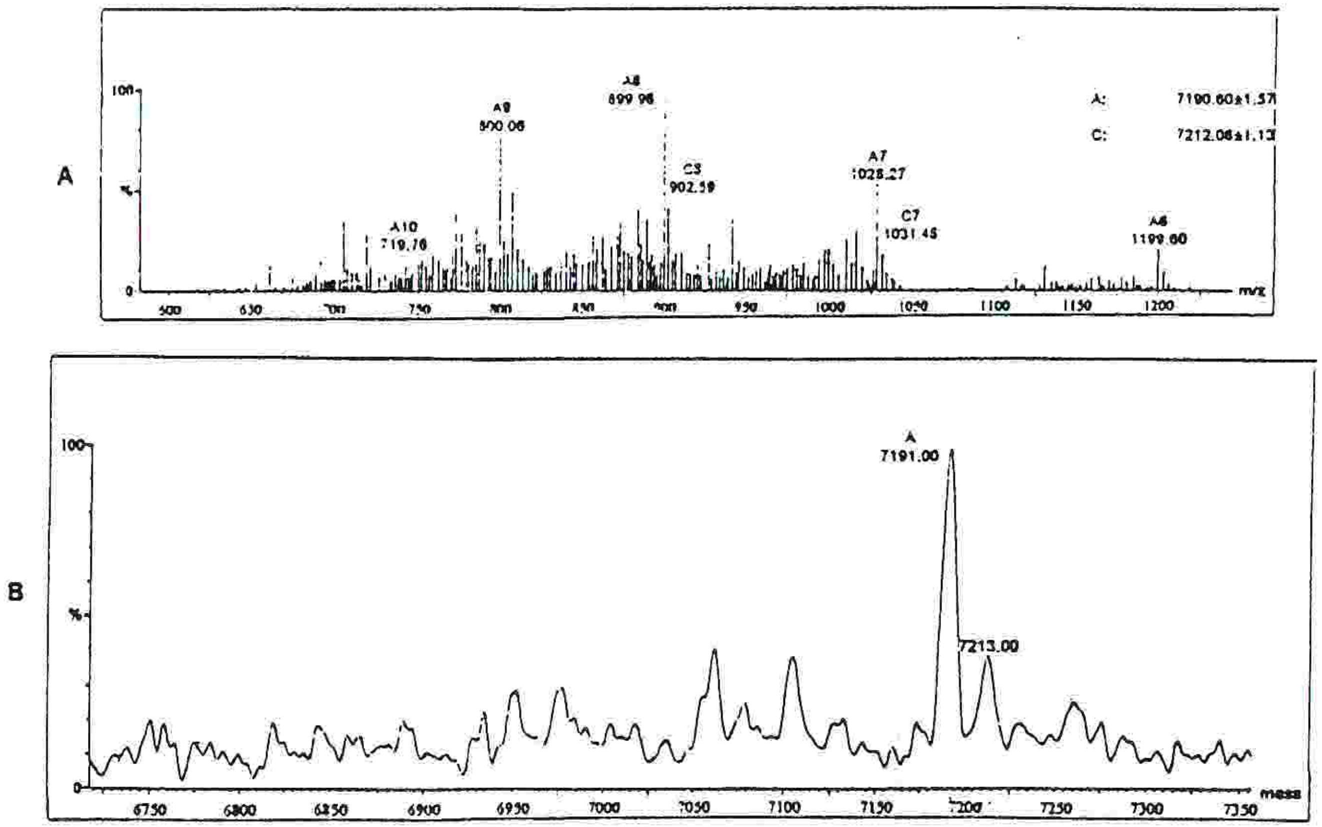


Figure I.4. ES ionization mass spectrum of H_2TH_3Ldim. (A) Spectrum produced by the mass analyzer; (B) deconvoluted spectrum (see section I.2.3).

I.2.4 Synthesis and characterization of $H_1H_2TH_3$Ldim

Synthesis protocol

The synthesis was performed on a Milligen 9050 Pepsynthesizer by using the Fmoc chemistry (Atherton and Sheppard, 1989). $H_1H_2TH_3$Ldim (Figure 3.8) was synthesized on a PAC-PEG-PS resin with a 10 fold excess (5 fold per each branch) of incoming amino acid on a 0.1 mM scale. The synthesis was carried out by using 1:1:1 Fmoc-AA/HOBT/TBTU as coupling mixture dissolved in N-methyl pyrrolidone containing 1.7-2 equivalents of N,N-diisopropylethylamine. The long branching peptide (94 amino acid residues) was synthesized on a Gly-substituted resin to which Fmoc-Lys(Fmoc)-OH had been coupled in a single run. Side-chain protecting groups were: PMC for Arg, Trt for Gln and Asn, Boc for Lys and Trp, tBuO for Glu and Asp, and tBu for Ser and Thr. Capping with acetic anhydride was carried out (6-fold excess for 15 minutes) after Lys40, Leu34, Gln28, Val24, Leu15, Ile11, Val6, Ile2, as these couplings were predicted to be difficult.

Cleavage and deprotection

After completion of the synthesis cycles, the peptide-resin was washed several times with methanol and diethylether and dried overnight under vacuum. The peptide was then cleaved from the resin with a mixture containing 82.5% TFA, 2.5% H_2O, 2.5% tri-isopropylsilane, 5% TA, 5% phenol and 2.5% EDT at room temperature for 5 hours. The post-cleavage work up was essentially the same as reported for the previous peptides. In fact, after resin washings with TFA on a sintered funnel, the lyophilized crude material from the synthesis was dissolved in water/acetic acid (1:1) and then extracted several times with diethyl ether. The aqueous phase was collected and lyophilized. The yield of the synthesis was then spectrophotometrically assayed as reported in the previous sections (Scopes, 1974).

Purification and characterization

The crude $H_1H_2TH_3$Ldim was characterized by analytical RP-HPLC both on a C4 and C18 column. The peptide was then dissolved in 30% acetonitrile containing 6 M guanidine-HCl and injected on a semipreparative Delta Pak C-18 300Å (19mm x30cm) RP-HPLC column. The sample gave a broad, but symmetric band (data not shown) on a C18 column using 0.1 M triethyl ammonium phosphate buffer, pH 2.25 (33-58.5% acetonitrile in 50 minutes with a flow rate of 1 ml/min, detected on the HPLC spectrophotometer set at 214 nm). Fractions from the centre of this band were collected and characterized by ES-MS. The fraction corresponding to the correct peptide was purified to homogeneity on a LiChrospher 100 RP-8 column and by ion exchange HPLC (Table 4.1).

I.3 Synthesis of the 434 repressor N-terminal domain and its branched dimer

First, the complete amino-terminal domain was synthesized , including amino acid residues 1 to 63 from the native protein (69 amino acid residues long). Based on the experiences of this synthesis, P7, the branched dimer of the domain was prepared with one synthesis (1.3.2). In this molecule, the carboxy-terminus (Gly62) of the complete amino-terminal domain is linked to both the α and ε amino groups of a lysine via two 6-aminohexanoic acid residues on each side (35 Å bridge), to give a branched peptide of two covalently dimerized 434 amino-terminal domains (Percipalle *et al.*, 1994a).

I.3.1 Synthesis and characterization of the complete cI434 amino-terminal domain

Synthesis protocol

The long monomeric cI434 amino-terminal domain (P1; Figure 3.8) was synthesized by automated continuos flow SPPS using the Fmoc chemistry (Atherton and Sheppard, 1989) on a Milligen 9050 Pepsynthesizer. The synthesis was performed on Fmoc-Thr(tBu)-PAC-PEG-PS resin with a 6 fold excess of incoming amino acid on a 0.1 mM scale. The coupling solution was an equimolar mixture of Fmoc-AA/HOBT/TBTU dissolved in DMF. As for the previous peptides, side-chain protecting groups included: PMC for Arg, Trt for Gln and Asn, Boc for Lys and Trp, tBuO for Glu and Asp, and tBu for Ser and Thr. Capping with acetic anhydride was carried out (6-fold excess for 15 minutes) after Lys40, Leu34, Gln28, Val24, Leu15, Ile11, Val6, Ile2, as couplings were predicted to be difficult.

Cleavage and deprotection

After completion of the synthesis cycles the peptide-resin was washed several times with methanol and then diethyl ether and dried overnight under vacuum. P1 was cleaved from the resin by using a cleavage mixture containing 82.5% TFA, 2.5% H2O, 2.5% triisopropylsilane, 5% TA, 5% phenol, 2.5% EDT. The reaction was left for about 5 hours at room temperature. The cleaved peptide was then separated from the resin by filtration under vacuum on a sintered funnel. The resin was further washed with the same cleavage mixture and few times with TFA. The filtrate was evaporated at 40°C and the dry material was then extracted several times with diethyl ether. The aqueous phase was collected and lyophilized.

Purification and characterization

P1 (crude material from the synthesis; Figure I.5 for HPLC traces) was characterized by RP-HPLC both on a C4 and C18 column. The peptide was then purified on a semipreparative Delta Pak C-18 300Å (19mm x30cm) RP-HPLC column, using 0.05% TFA as buffer A and 90% acetonitrile (in buffer A) as buffer B, with a linear gradient from 0-60% acetonitrile in 50 minutes and a flow rate of 1 ml/min (the sample was detected on the HPLC spectrophotometer set at 214 nm). The fraction collected from the semipreparative HPLC run was then confirmed to be correct by ES-MS (Table 4.1) and homogeneous by SDS gel electrophoresis (data not shown). The fraction corresponding to the correct peptide was then reinjected on a LiChrospher 100 RP-8 column and further purified.

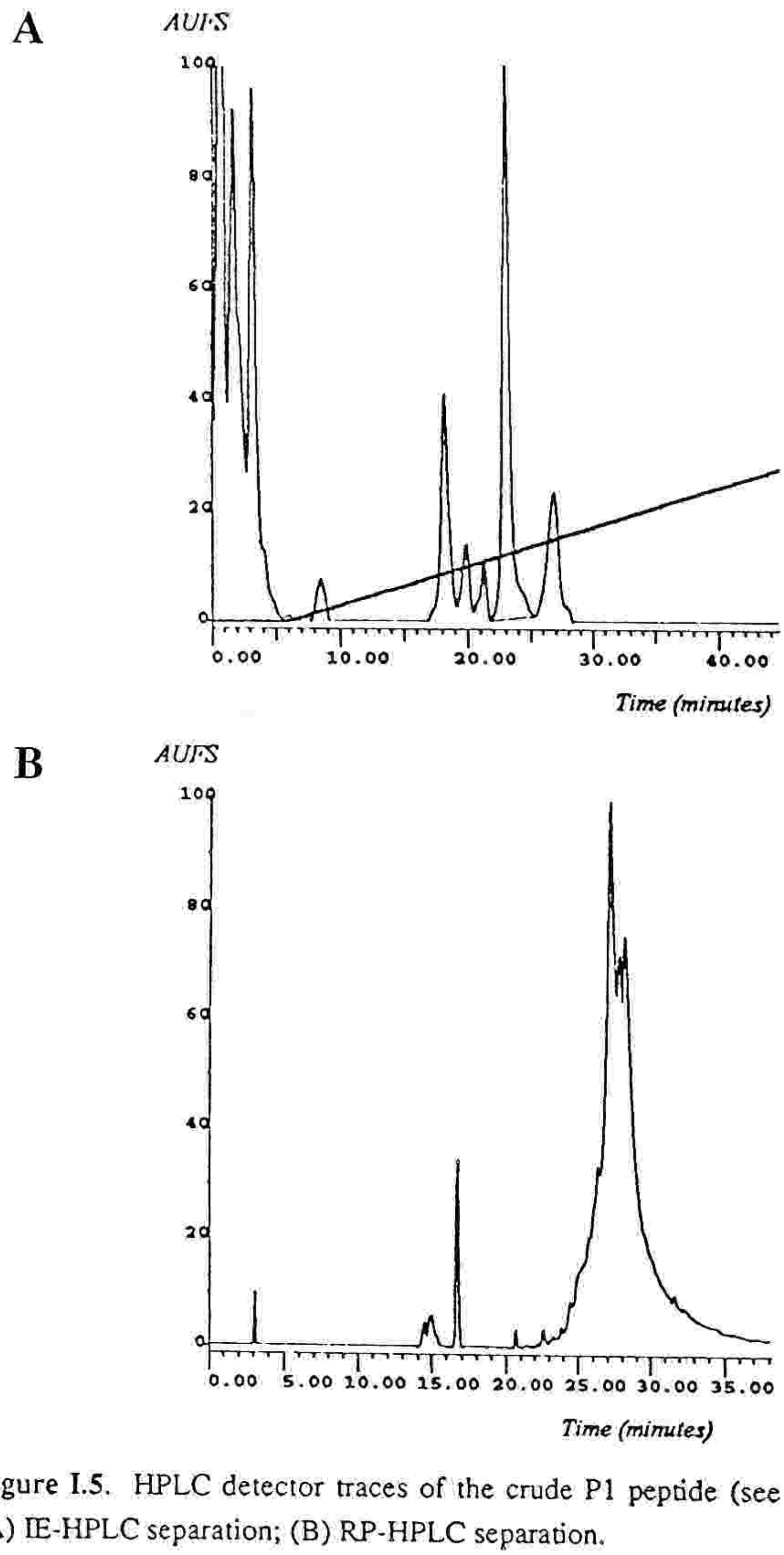


Figure I.5. HPLC detector traces of the crude P1 peptide (see section I.3.1). (A) IE-HPLC separation; (B) RP-HPLC separation.

I.3.2 Synthesis and characterization of P7, a covalent branched dimer of two 434 repressor amino-terminal domains.

Synthesis protocol

The synthesis of two cI434 amino-terminal domains covalently dimerized (P7; Figure 3.8) was carried out by continuous flow SPPS on a Milligen 9050 Pepsynthesizer in a single run (Atherton and Sheppard, 1989). The synthetic protein was synthesized on an Fmoc-Gly-PAC-PEG-PS resin on a 0.1 mM scale, using a 10 fold excess of incoming amino acid (5 fold per each branch). A coupling mixture containing Fmoc-amino acid, HOBT and TBTU in equimolar concentrations was utilized, with the components dissolved in N-methylpyrrolidone and 1.7-2 equivalents of N,N-diisopropylethylamine. Side-chain protecting groups included, as for the other peptides: PMC for Arg, Trt for Gln and Asn, Boc for Lys and Trp, tBuO for Glu and Asp, and tBu for Ser and Thr. Capping was also carried out with acetic anhydride, using a 6-fold excess for 15 minutes, after Lys40, Leu34, Gln28, Val24, Leu15, Ile11, Val6, Ile2, as couplings were predicted to be difficult (Percipalle *et al.*, 1994a).

Cleavage and deprotection

After several washings with methanol, diethyl ether and vacuum drying, the peptide was cleaved from the resin by using a cleavage mixture containing 82.5% TFA, 2.5% H_2O, 2.5% triisopropylsilane, 5% TA, 5% phenol, 2.5% EDT. The reaction was left for about 5 hours at room temperature. The cleaved peptide was then separated from the resin by filtration under vacuum on a sintered funnel. The resin was further washed with the same cleavage mixture and few times with TFA. The filtrate was evaporated at 40°C, the dry material was dissolved in water/acetic acid (1:1) and then extracted several times with diethylether. The aqueous phase was collected and lyophilized and the sample spectrophotometrically quantitated (Scopes, 1974).

Purification and characterization

After cleavage from the resin, the crude material coming directly from the synthesis was dissolved in a solution containing 30% acetonitrile and 6 M guanidine-HCl and analyzed both by C4 and C18 RP-HPLC. It was then purified by using a semipreparative Delta Pak C-18 300Å (19mm x30cm) RP-HPLC column using 0.1 M triethyl ammonium phosphate buffer, pH 2.25, (33 to 60% acetonitrile in 50 minutes), flow rate 3 ml/min. The chromatogram (Figure I.6) shows a broad, but symmetric band on the C18 column. For this reason, fractions corresponding to the very top of the same band were collected (total of four fractions) and the ones with the correct molecular weight as determined by ES-MS (Table 4.1) were further purified to homogeneity. The purification was carried out both by anion-exchange and RP HPLC using a Shodex QA825 column and a Lichrospher 100 RP-18 (5 μm) column respectively (Figure I.6).

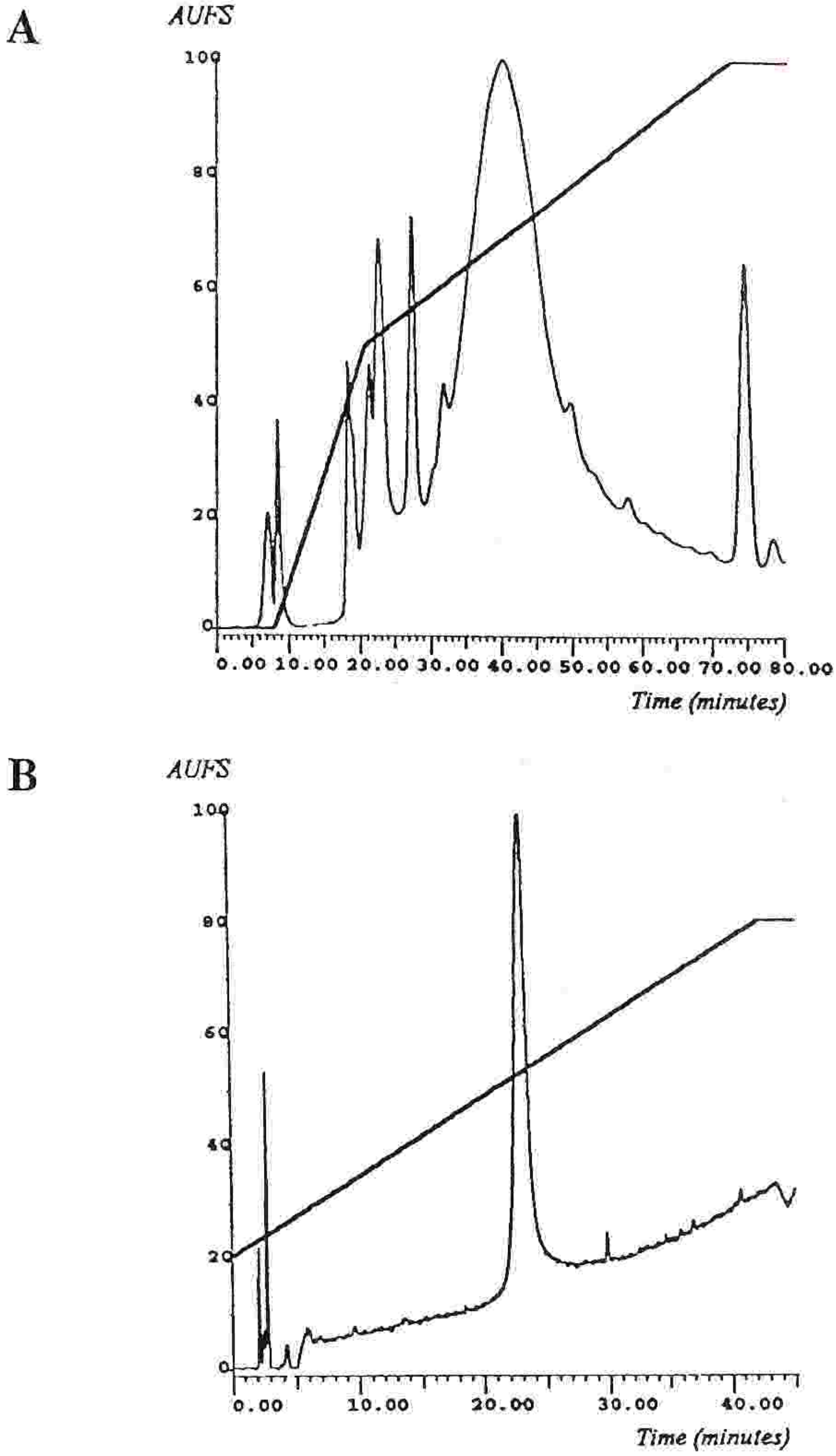


Figure I.6. HPLC detector traces of P7. (A) Crude P7 from the synthesis; (B) RP-HPLC of pure P7

I.4 Preparation of a crude extract containing the cI434 repressor protein

E.coli DH5 bacteria infected with lysogen imm434 (kindly donated by Dr R. Weinberg) were grown overnight in LB plates at 37°C. A small colony was then picked up and further grown overnight in 250 ml of LB medium at 37°C and then centrifuged at 5000 RPM for 30 minutes at 4°C. The pellet was resuspended in a buffer containing 30 mM Tris-HCl pH=7.4, 0.15 mM sodium chloride, 5 mM magnesium chloride, 1 mM PMSF (protease inhibitor) and then sonicated for 5-10 minutes. The suspension was then centrifuged for 30 minutes at 15000 RPM at 4°C and the pellet discarded. The supernatant was then treated with 25% ammonium sulphate (w/v) overnight at 4°C and then centrifuged at 10000 RPM for 30 minutes again at 4°C. The pellet was then dissolved in TM buffer (i.e. 50 mM Tris-HCl pH=7.4, 10 mM $ZnCl_2$, 12.5 mM $MgCl_2$, 1 mM EDTA, 10% glycerol) and tested for the presence of the 434 repressor protein by band shift and footprinting experiments.

I.5 Binding assays

I.5.1 Band shift reactions

Gel shift experiments were carried out with all the available DNA fragments prepared as already described, double-stranded oligonucleotides with the specific cI434 binding sites and non-specific oligonucleotides. Gel shift assays were also performed with DNA probes suitably prepared for photocrosslinking experiments. The reactions were carried out using all synthetic peptides (H_2TH_3, H_2TH_3L, H_2TH_3Ldim and $H_1H_2TH_3$Ldim) and synthetic (P1 and P7) and recombinant protein (RR69) as well as the cI434 repressor containing crude extract (Chapters 4 and 5).

Binding reactions were performed with ~ 5 fmoles of a 5' ^{32}P-end-labelled double-stranded DNA probe incubated with peptides, proteins or protein extract in a buffer containing 20 mM HEPES (N-[2-Hydroxyethyl]piperazine-N-[2-ethanesulfonic acid]) pH=7.9, 50 mM sodium chloride, 5 mM magnesium chloride, 2 mM DTT, 0.2 mM EDTA and 5% glycerol. Synthetic peptides were present in the reaction mixtures in concentrations ranging between 1.4 μM and 9.5 μM; synthetic proteins were present at final concentrations ranging between 0.3 μM and 3.5 μM, whereas recombinant proteins were assayed at concentrations between 1.2 nM and 1.2 μM. Incubation was carried out on ice for 30 minutes in the case of the cI434 repressor protein extract, 40 minutes for the synthetic and recombinant proteins whereas it was prolonged to 1 hour for all the synthetic shorter peptides. Since highly specific binding and affinity was tested, poly[d(I-C)] (1 g/l) was used in large excess (1000 fold excess as compared to the labelled DNA) to remove any non-specific binding to the DNA probes. At the end of the incubation period samples were loaded on a 6% non-denaturing polyacrylamide gel and electrophoresed at low voltage (100 V/cm), in TBE running buffer, at 4°C. For gel shift experiments with short DNA probes, such as the 20bp long oligonucleotides, samples were always loaded on a 8% non-denaturing polyacrylamide gel and electrophoresed at 170-180 V, in TBE running buffer, for approximately 2 hours at 4°C. In this case higher voltage is preferred to avoid any problem of sample diffusion throughout the gel.

I.5.2 Photochemical peptide-DNA crosslinking

Preparation of the DNA binding sites for UV crosslinking

1) 32bp oligonucleotide containing one O_R site

The 32-mer oligonucleotide 5'-CTTAGTGACTCGTACAACTACAATTGTTGAAG (PC1) was chemically synthesized on an Applied Biosystem 380A DNA Synthesizer and designed so as to comprise the palindromic half operator sites of the cI434 operator O_R1. It was purified by n-butanol precipitation, then electrophoresed on a denaturing polyacrylamide gel and the band corresponding to its size excised. After extraction from polyacrylamide, it was phenol/chloroform extracted and precipitated with 100% ethanol. The oligonucleotide was then used as template for *E. coli* polymerase I Klenow fragment in order to enzymatically prepare the double-stranded oligonucleotide to be used in the binding reactions. In a total reaction volume of 25 µl containing 2 µl of the oligogonucleotide (PC1) (0.3 mM), 2.5 µl of 10 times concentrated reaction buffer (130 mM potassium phosphate, pH=7.4, 6.5 mM magnesium chloride, 1 mM DTE, 32 g/ml BSA), 1.25 µl of dATP (100 mM), 1.25 µl of dGTP (100 mM), 2 µl of γ-^{32}P-dCTP (25 µCi/µl), 4 µl of dUBrTP (i.e. 25 mM bromodeoxyuridine 5'-triphosphate), 1 µl of Klenow Fragment (2 U/µl) of DNA Polymerase I was added. The polymerization reaction was carried out for 1 hour at 37 °C. The double-stranded oligonucleotide was then purified on a G25 Sephadex column by eluting the sample with TE buffer (i.e. 10 mM Tris.HCl pH=7.5, 1 mM EDTA) [Sambrook *et al.*, 1989].

2) 45bp oligonucleotide containing two adjacent O_R sites

Partially complementary oligonucleotides PC2, 5'-CAAACAATGTACTTGACGACTCATCAC, and PC3, 5'-TCTACAACTACAATTGTTGTGATGAGTCG, were synthesized on an Applied Biosystem 380A DNA Synthesizer. They were precipitated with n-butanol, purified by electrophoresis on 8 M urea-polyacrylamide gel and their concentrations calculated as described before. Equimolar amounts of PC1 and PC2 were then annealed in H_2O by heating them at 95°C and then slowly cooling to room temperature. The partially double-stranded oligonucleotide was converted into fully double-stranded form by filling in the single-stranded region with Klenow polymerase. For this purpose, in a total volume of 50 μl containing 20 μl of sample DNA, 5 μl of 10 times concentrated Klenow buffer (i.e. 130 mM potassium phosphate, pH=7.4, 6.5 mM magnesium chloride, 1 mM DTE, 32 g/ml BSA), 1.25 μl of dATP (100 mM), 1.25 μl of dGTP (100 mM), 2 μl of γ-^{32}P-dCTP (25 μCi/μl), 4 μl of dUBrTP (i.e. 25 mM bromodeoxyuridine 5'-triphosphate) and 1 μl of *E. coli* DNA Polymerase I Klenow fragment (2 U/μl) was added. The reaction was carried out for 1 hour at 37°C. The double-stranded oligonucleotide was purified through a G25 Sephadex column by eluting it with TE buffer (i.e. 10 mM Tris.HCl, pH=7.5, 1 mM EDTA) [Sambrook *et al.*, 1989].

Binding reactions

Binding reactions were carried out essentially as already described. The only protocol difference is due to the light sensitivity of bromodeoxyuridine-containing oligonucleotides, and therefore the operations had to be carried out in a dark room.

Photochemical crosslinking

The crosslinking reaction was carried out with all synthetic peptides i.e. H_2TH_3, H_2TH_3L, H_2TH_3Ldim and $H_1H_2TH_3Ldim$. Samples were loaded on a strip of parafilm placed on ice and then irradiated with a UV lamp (λ = 254nm) placed at 5 cm from the samples. UV irradiation was carried out for about 1 hour. Samples were then loaded on a 12% reducing SDS polyacrylamide gel and then electrophoresed at 20 mA and 160 V. The gel was then stained with Coomassie and exposed.

I.5.3 DNAse Footprinting

DNAse footprinting reactions (Galas and Schmitz, 1979) were carried out with DNA fragments containing the wild-type 434 operator O_R region (180 bp) and the O_R1 (150 bp) operator prepared as already described. Preliminary footprinting reactions were performed also in the presence of the shorter peptides, but experiments were not succesful. The digestion reactions were mainly carried out on the DNA fragments in the presence of the synthetic proteins P1 (data not shown) and P7 and recombinant protein RR69. In order to have control experiments, digestion reactions with DNAseI were also carried out on the naked template DNA. Both synthetic proteins (P1 and P7 respectively) were incubated with a 5' ^{32}P-end-labelled DNA fragment (10000 CPM) containing operator site O_R1 for 60 minutes in ice water.

The binding reactions were assembled as schematically reported in the table below.

	[a] Crude Extract	[b] P1	P7	RR69
Labelled DNA	1 µl	1 µl	1 µl	1 µl
3x Solution D	8 µl	8 µl	8 µl	8 µl
Water	6 µl	8.5 µl	6 µl	10 µl
Protein(s)	5 µl	2.5 µl	5 µl	1 µl

a Data not shown.

b Data not shown as P1 does not give a retardation in gel shift experiments and therefore it can not protect the DNA.

3x Solution D contains 60 mM HEPES, pH=7.9, 150 mM potassium chloride, 0.3 mM EDTA, 1.5 mM DTT and 30% glycerol. 2 μl of Mg/Ca mix (125 mM magnesium chloride and 25 mM calcium chloride) were added and the samples incubated for 1 minute at 20°C in a water bath. Then 2 μl of DNAse I (Boehringer Mannheim) freshly diluted from a stock solution (3 mg/ml) 1000 fold with 1x Solution D (containing 100 μg/ml BSA) were added to the reaction mixture (final volume 24 μ l) and then incubated again for 1 minute at 20°C. The digestion reaction was stopped by adding 200 μl of a stop solution (20 mM Tris, pH=7.5, 0.1 M sodium chloride, 1% SDS, 5mM EDTA, 50 g/ml proteinase K, 25 μg/ml yeast RNA) and incubating for 30 minutes at 45°C. The samples were then precipitated with 750 μl of 100% ethanol, washed with 70% ethanol and dried. The pellets were then dissolved in formamide containing dye solution, denatured for 3 minutes at 95°C and resolved at high voltage (1800-1900 V) and constant power (41 W) by electrophoresis on a 6% denaturing polyacrylamide gel (Chapter 5).

I.5.4 DMS Methylation interference

In this experiment the DMS treatment was carried out *in situ* as this substantially alleviates problems related to (i) the intrinsic solubility of DMS itself in the reaction buffer and (ii) to the clarity of the results, highly dependant on the extent of occupancy of the binding sites on the DNA. The method allows the analysis of kinetically labile complexes since (i) the cleavage background is greatly reduced and (ii) the "caging effect" is enhanced (Papavassiliou, 1993 and references therein).

Resolving DNA-protein complexes by gel shift assays

Binding reactions and electrophoresis assembly were performed as already described. Synthetic protein P7 and recombinant RR69 were included in the reaction mixtures in the presence of their 5' ^{32}P-end-labelled cognate DNA template cI434 O_R1 (150bp) prepared as described above. In this case the reactions were 5 fold scaled up as compared to the corresponding gel shift assays.

In situ DMS methylation

At the end of the electrophoretic run, the polyacrylamide gel was soaked in 300 ml of a buffer containing 10 mM Tris-HCl, pH=8.0 for 10 minutes. After equilibration the gel slab was transferred into a clean plastic tray containing 100 ml of freshly prepared 0.2% (v/v) DMS (reagent grade) solution in 10 mM Tris-HCl, pH=8.0. The *in situ* DNA methylation was allowed to proceed without shaking of the tray for 5 minutes. At the end of the incubation period, the slab gel was transferred into a new plastic tray containing 300 ml of freshly made DMS-stop buffer (0.5 M β-mercaptoethanol, 150 mM Tris-HCl, pH=7.3, 5 mM EDTA) and shaked vigorously on a rocking platform for 10 minutes at 37°C (at this temperature DMS is rapidly hydrolysed). The slab gel was then removed from the stop solution, placed on a suitable backing material (i.e. overhead photocopy film), covered with Saran wrap and subjected to autoradiography at 37°C for about 2 hours to detect the position of the retarded and unretarded DNA species. Radioactive bands were then excised at minimal sizes using disposable razor blades. Each gel slice was resuspended in 0.6 ml of elution buffer (0.5 M ammonium acetate, pH 7.5, 0.1% SDS, 1 mM EDTA) including 0.1 M β-mercaptoethanol to ensure inhibition of residual DMS, and the DNA samples are allowed to diffuse out by shaking the tubes at 42°C overnight. The supernatants were then recovered and the sample DNA precipitated with 100% ethanol in the presence of 10 µg of carrier glycogen. The pellet was then rinsed with 70% ethanol and dried.

Piperidine cleavage and NaOH cleavage of the DMS methylated DNA samples

The DNA samples were then subjected to standard chemical cleavage both with piperidine and NaOH (Maxam and Gilbert, 1980; Perbal, 1988), which respectively promote strand scission at the level of guanine residues and strand scission at the level of guanine and adenine residues of the methylated DNA.

(i) *Cleavage with piperidine.* 100 µl of 1 M piperidine were added to each sample and the cleavage reactions carried out at 90°C for 30 minutes. The reactions and the quenching were essentially performed as already reported. At the end the lyophilized pellets were resuspended in formamide sample buffer, heat-denatured and resolved on a 6% sequencing denaturing polyacrylamide gel for comparative electrophoretic analysis.

(ii) *Cleavage with NaOH.* The dried pellets were resuspended in 20 µl of 10 mM phosphate buffer, pH=7.2, containing 1 mM EDTA and incubated at 90°C for 15 minutes. Samples were immediately transferred to ice and spun down. 2 µl of 1 M NaOH (freshly prepared) were then added and the DNA samples were left at 90°C for

30 minutes. During the incubation period, samples were spun down every 10 minutes for few seconds. At the end of the NaOH treatment, 130 μl of 0.3 M sodium acetate (pH=7.3) was added and the DNA precipitated with 450 μl of 100% ethanol in the presence of 10 μg of carrier glycogen. After microfuging at 14000 RPM for 15 minutes, supernatants were removed and pellets precipitated again by resuspending them in 100 μl of 0.3 M sodium acetate (pH=7.3) and 300 μl 100% ethanol. After centrifugation, pellets were washed with 70% ethanol and then vacuum-dried. After lyophilization, pellets were resuspended in formamide sample buffer, heat-denatured and resolved on a 6% sequencing denaturing polyacrylamide gel for comparative electrophoretic analysis.

I.5.4 In vitro transcription experiments

In vitro transcription experiments (Fong *et al.*, 1993 and references therein) were carried out in the presence of a template DNA fragment containing the full cI434 O_R operator region containing the O_R1, O_R2 and O_R3 sites. Transcription was assayed both in the absence and in the presence of all available synthetic peptides H_2TH_3, H_2TH_3L, H_3Ldim, H_2TH_3Ldim and $H_1H_2TH_3Ldim$, synthetic proteins P1 and P7, and recombinant protein RR69.

In vitro transcription

Transcription in the absence of peptides or proteins was assayed using as template 180 bp cI434 O_R fragment (containing all the three operators O_R1, O_R2, O_R3) prepared as already described. Products of the reaction are terminated transcripts initiated at P_{RM} and P_R, respectively. In a total volume of 15 µl, 1 µl template DNA was preincubated with 1 µl *E. coli* RNA polymerase and 1.5 µl 10 times transcription buffer (to a final concentration of 40 mM Tris-HCL, pH=8.0, 0.1 M KCl, 10 mM $MgCl_2$, 1 mM dithiothreitol, 100 µg BSA per reaction volume) at 37° C for 15 minutes to allow formation of open complexes (McClure, 1985). At the end of the preincubation period, 5 µl of transcription buffer containing heparin and substrates for RNA polymerase were added to yield the final concentrations: heparin, 50 µg/ml; ATP, CTP and GTP, 200 µM and 40 µM [α-^{32}P]UTP (2.5 Ci/mmol). After 15 minutes reactions were terminated by addition of 4 µl of transcription stop solution (7 M urea, 0.1 M EDTA, 0.4 % SDS, 40 mM Tris-HCl, pH=8.0, 0.5% bromophenol blue, 0.5% xylene cyanol) and samples loaded on a 6% denaturing polyacrylamide gel and then autoradiographed.

Inhibition of in vitro transcription

Abortive initiation assays were performed with the same DNA template as before in the presence of synthetic peptides and proteins and recombinant protein. Synthetic peptides H_2TH_3, H_2TH_3L, H_3Ldim, H_2TH_3Ldim and $H_1H_2TH_3Ldim$, synthetic proteins P1 and P7, and recombinant protein RR69 were preincubated with template DNA as above described and left for 15 minutes at 37°C to allow formation of open complexes. After preincubation, 5 µl of transcription buffer containing heparin and substrates for RNA polymerase were added to yield the final concentrations as described in the previous section. Reactions were terminated after 15 minutes by addition of transcription stop solution as already mentioned and samples electrophoresed on a 6% denaturing polyacrylamide gel at 1800 V (W=41). Final concentratiions were: RNA polymerase, 20 nM; DNA template, 2 nM; synthetic peptides H_2TH_3, H_2TH_3L, H_3Ldim (data not shown), H_2TH_3Ldim, $H_1H_2TH_3Ldim$ were 7.3 µM, 5.8 µM, 7.2 µM, 6.5 µM, 6.2 µM respectively; the concentration of synthetic proteins P1 and P7 were 6.6 µM, 5.8 µM, respectively; the concentration of the recombinant protein RR69 was 3.6 µM.

I.6 Circular Dichroism spectroscopy

CD spectra were measured on a JASCO 600 spectropolarimeter in a thermostated environment at room temperature, using 0.2 cm or 1 cm path-length cells. Samples were dissolved in 10 mM phosphate buffer pH=7.0 and 100 mM NaCl (when present) and spectra were recorded in the presence of various agents such as trifluoroethanol (TFE) or cognate and non-cognate DNA templates.

CD spectra measured in the presence of TFE

The experiments were carried out with all synthetic peptides H_2TH_3 (data not shown), H_2TH_3L, H_3Ldim (data not shown), H_2TH_3Ldim, $H_1H_2TH_3Ldim$, synthetic proteins P1 (data not shown), P7, and recombinant protein RR69 (data not shown).

Samples were dissolved in 10 mM phosphate buffer, pH 7.0, without any NaCl as this may precipitate in the cell upon addition of TFE. Increasing quantities of the α-helical inducing agent TFE were then directly added in the cell and spectra recorded for each sample at different concentrations. CD spectra were scaled up according to the various dilutions made for each measurement so as to have comparable dichrograms to be overlayed in the same plot for each single sample peptide or protein. Concentrations of peptides were $[H_2TH_3]=140\mu M$, $[H_2TH_3L]=38\mu M$, $[H_3Ldim]=10\mu M$, $[H_2TH_3Ldim]=45\mu M$, $[H_1H_2TH_3Ldim]=50\mu M$; concentrations of synthetic proteins P1 and P7 were respectively 7.2 μM and 35 μM; concentration of recombinant protein RR69 was 4.8 μM. All spectra were averaged three times and smoothed with the software provided by JASCO Inc. (Easton, Maryland).

CD spectra measured in the presence of oligonucleotides

CD studies were performed with synthetic proteins and recombinant proteins with short oligonucleotides containing both the cI434 OR1 operator (cognate DNA) and completely unrelated sequences (non-cognate DNA). Also in this case, all spectra were averaged three times and smoothed with the software provided by JASCO Inc. (Easton, Maryland). The effect of oligonucleotides on the conformation of proteins was detected by calculating the difference spectra, obtained by subtracting the protein spectrum from the corresponding peptide-DNA spectrum and data were collected and smoothed as described above.

(i) *CD studies on synthetic and recombinant proteins with cognate DNA (cI434 O_R1 site).* Two complementary 20 bp long oligonucleotides, 5'-TACAAGAAAGTTTGT (AT402) and 5'-TAACAAACTTTCTTG (AT403), were annealed in equimolar concentrations and used in CD studies. Each sample contained 10 mM phosphate buffer, pH 7.0, 100 mM NaCl and either proteins P1 and P7 (7.2 μM and 35 μM) or recombinant protein RR69 (4.8 μM). The double-stranded oligonucleotide containing the cI434 O_R1 was added in equimolar amounts when present in the reaction mixture. Samples (proteins and DNA) were incubated for 30 minutes at room temperature before each CD measurement (Chapter 5).

(ii) *CD studies on synthetic and recombinant proteins with non-cognate DNA (unrelated sequence).* Oligonucleotides 5'-AGAGGAGCGCCTGTCATGCA (CL20) and 5'-GCGCTGCATGACAGGCGCTC (WL20), were annealed in equimolar amounts and used in CD studies. The same conditions as previously described for cognate DNA were utilized in this case. The unrelated, non-cognate double-stranded oligonucleotide was also added in equimolar amounts when present in the reaction mixture.

I.7 Computer methods

The X-ray structure coordinates for the cI434 repressor amino-terminal domain bound to its cognate DNA (Aggarwal *et al.*, 1988) was abstracted from the Brookhaven Database. The structure was loaded into the Biosym Technologies Insight II program and standard procedures from the Insight module used for the visualization process.

All energy minimization and molecular dynamics experiments were carried out using the Discover module of the same program, using a standard protocol taken from the Biosym Insight II/Discover instruction manual. The structure of peptides were modelled by manually inserting the linker sequences to the DNA-N-terminal domain complex built using the published 3D coordinates (Aggarwal *et al.*, 1988), and then subjecting the linker structures first to a 100 ps Molecular Dynamics at 300 K with a distance dependent dielectric constant and then to 200 cycles steepest descent energy minimization. The known part of the structure was kept fixed during the entire simulation. The linker in RR69 is deemed to be flexible, its position is only symbolically shown.

Database searching was carried out using the Findseq module in the Intelligenetics program, on the Swissprot 28 Database. Alignments of repressor sequences were performed using the Genealign module of the same program, on sequences containing only the HTH motif with short flanking regions.

The distance matrix was calculated by first running a full cycle of energy minimization (100 ps Molecular Dynamics at 300 K with a distance dependent dielectric constant and then to 200 cycles steepest descent energy minimization.) on the cI434 amino-terminal domain structure and abstracting all the Cα-Cα distances from the resulting molecular definition files.

APPENDIX II

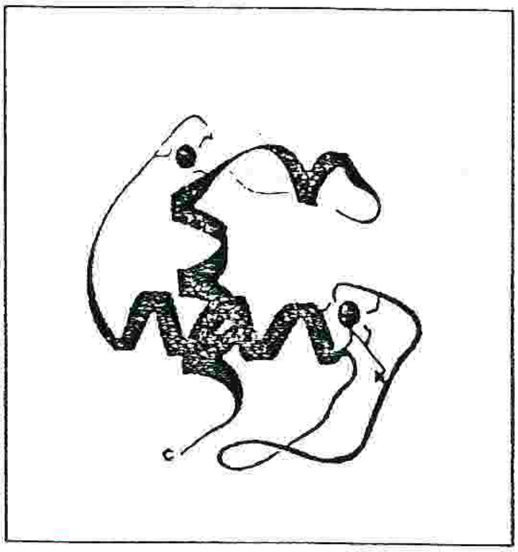
DNA binding proteins

Zinc finger proteins

Transcription factors are proteins that bind DNA and regulate the transcription of mRNA and therefore gene expression. They are known to contain a DNA-binding domain through which they bind in a sequence-specific manner (Berg, 1993). Such highly conserved DNA-binding and dimerization domains have been characterized both by structural studies and by sequence comparisons and each structural motif may be considered to represent the evolution of different solutions to the design of proteins that need to be fully compatible to their target DNA, both from a structural and chemical point of view (Pabo and Sauer, 1992; Harrison, 1991).

GAL4 is a yeast transcriptional activator whose DNA-binding specificity and ability to dimerize is retained in the 65 residue long amino-terminal domain of the protein (881 aminoacids). Structural studies, both by X-ray crystallography and NMR (Baleia *et al.*, 1992; Kraulis *et al.*, 1992; Marmorstein *et al.*, 1992) show that dimerization is affected by the carboxy end of the 65 residue terminus, in particular by a two-stranded coiled-coil structure formed by weak hydrophobic interactions between two aphipathic α-helices from each monomer. The core of the DNA-binding domain (residues 8-40 in the N-terminal domain) forms a compact globular structure folded around two Zn ions which are tetrahedrically coordinated to six cystein residues in a binuclear-type cluster. The metal-binding domain probably mediates the sequence-specific contacts with the bases via two α-helices, one per each monomeric subunit. The dimer binds a 19 bp DNA sequence by inserting side chains from residues in these appropriately positioned α-helices into the major groove on opposite faces of DNA.

A



B

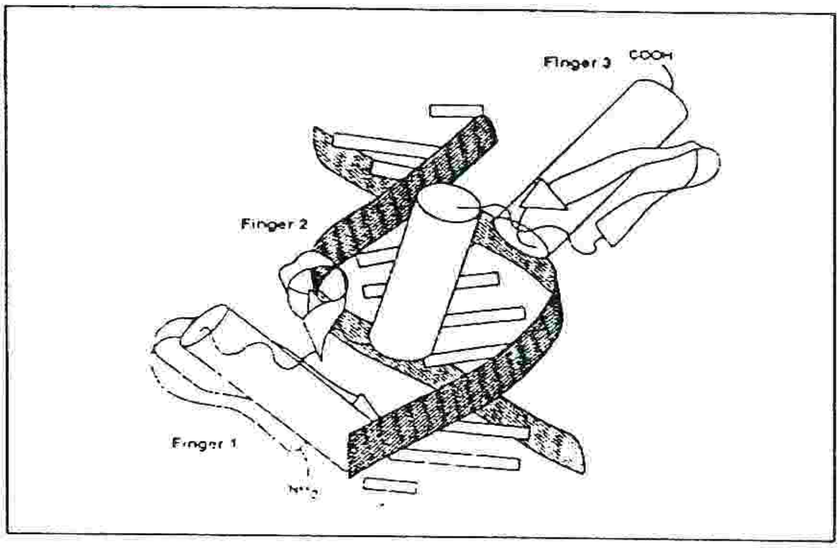


Figure II.1. (A) Ribbon diagram representing the glucocorticoid-receptor DNA-binding domain (Luisi *et al.*, 1991). Two loop-helix modules are shown, each with four cysteines coordinating a Zn^{2+} ion. (B) Overall arrangement of the Zif268 zinc finger-DNA complex (Pavletic and Pabo, 1991). The α-helices are shown as cylinders; β-sheets are shown as ribbons; and the zinc ions have been omitted from this sketch. Each finger makes contacts in a three-base-pair subsite.

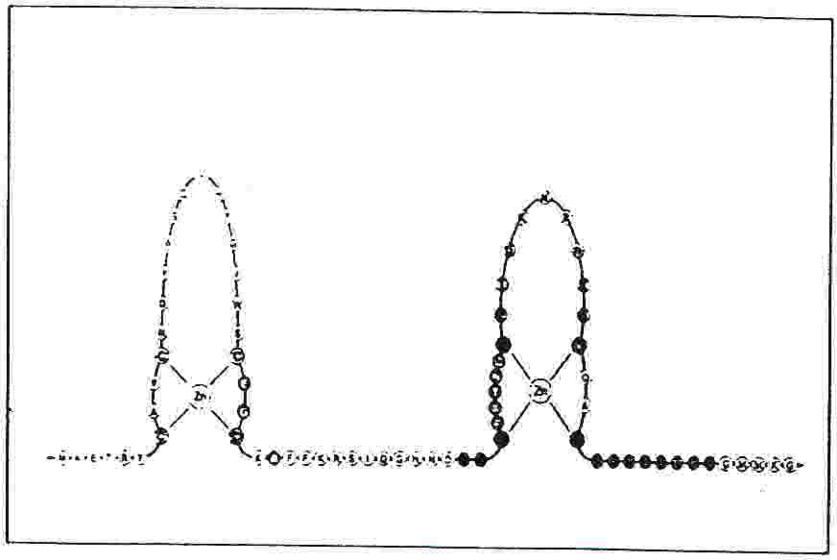
GAL4 is thus characterized by a binuclear cluster and falls into a class of zinc-finger proteins with a generalized Cys-X-X-Cys sequence (X may be any aminoacid) at the beginning of the DNA-binding α-helices. This is also observed in the case of the steroid receptor class of zinc-fingers. More generally, zinc-finger containing DNA-binding proteins have been divided into two main classes according to the aminoacid residues that are involved in metal binding; i.e., (i) Cys2His2 zinc fingers and (ii) Cys3His domains.

TFIIIA is the first member discovered for the Cys2His2 class of zinc finger proteins. It contains nine zinc fingers but it is still not known precisely how it interacts with DNA, considering that it is probably able to bind both DNA and RNA with high specificity and affinity. The role of each zinc finger (Desjarlais and Berg, 1992a; Desjarlais and Berg, 1992b) has been better elucidated for Zif268 (Figure II.1), for which the Zif268-DNA complex crystal structure has been determined (Pavletich and Pabo, 1991), and that of SP1 (Desjarlais and Berg, 1992; Thukral *et al.*, 1992) in which case the middle zinc finger domain out of three is involved in interacting with the preferred DNA-binding site.

A similar structural organization as that of TFIIIA is found in the DNA binding domain of the nuclear receptor superfamily. The glucocorticoid receptor (Figure II.1), whose solution structure has already been determined (Hard *et al.*, 1990) belongs to this group of proteins. It is characterized by a pattern of eight cysteine residues which coordinate two Zn ions each with a tetrahedral geometry. The Zn fingers in the glucocorticoid receptor fold together as part of a larger, unified globular domain as well as the estrogen receptor (Swabe *et al.*, 1990) [Figure II.2], whereas in TFIIIA the fingers act as conformationally independent and stable units, each contributing to DNA binding (Luisi *et al.*, 1991 and references therein).

Retroviral nucleocapsid proteins, on the other hand, belong to the Cys3His class of zinc finger domains. In fact it has been shown that large quantities of Zn copurify with retroviruses such as HIV-I and HTLV-I (Bess *et al.*, 1992).

A



B

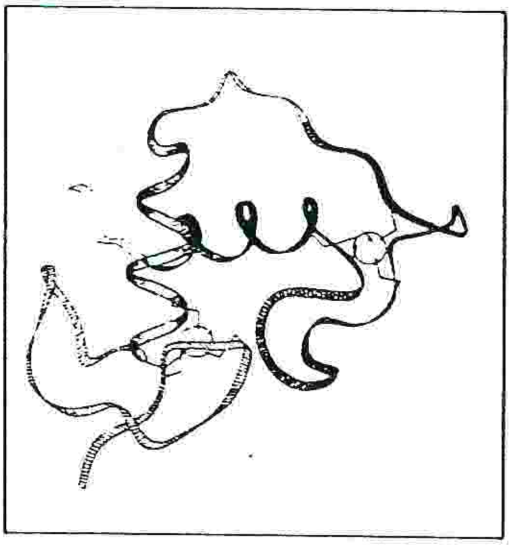


Figure II.2. DNA binding domain of the estrogen receptor. (A) The domain contains two zinc-finger-like units; the first one makes the main contacts with DNA; the second serves as dimerizing unit enabling two receptor molecules to combine and attach securely to DNA. (B) Three-dimensional view of the DNA-binding domain of the estrogen receptor. Binding of zinc by the cysteines ties the terminal segments of the irregular region to the base of the recognition helix (Swabe *et al.*, 1990).

Extensive use of synthetic peptides coupled with X-ray absorption fine structure studies have shown that the sequence Cys-X-X-Cys-X-X-X-X-His-X-X-X-X-Cys is indeed responsible for coordinating Zn (II) ions in the case of the HIV-I nucleocapsid protein.

TATA-box binding protein

The TFIID TATA-box binding protein is one of the factors that plays a role in initiating transcription of mRNA in eukaryotes. In particular it is a site-specific DNA-binding protein that is required for assembly of the pre-initiation complex and moreover interacts with a set of proteins called TATA-associated factors (Thukral *et al.*, 1992; Pavletich and Pabo, 1991). The structure of a 200 residue long polypeptide corresponding to its carboxy-terminal domain has recently been determined by X-ray crystallography (Niklov *et al.*, 1992) and shown to be built up from two 88 residue long domains related to each other by an intramolecular pseudo-dyad. Each domain is made up of five -strands (S1, S2, S3, S4, S5) and two α-helices (H1, H2) which are arranged in the order S1-H1-S2-S3-S4-S5-H2. A linker connects the C-terminal group of H2 with the aminon-terminus of S1' on the other monomer (as indicated by the "prime" index). Interestingly, in such a saddle-like structure H2 and H2' are facing DNA in a concave-type fashion whereas H1 and H1' are respectively perpendicular to H2 and H2'. TFIID is a rare example of a transcription factor that binds DNA in a sequence-specific manner but interacts with the minor groove (Summers *et al.*, 1992; Starr and Hawley, 1991). More insights on the residues involved in direct binding to the TATA-box will be provided by the crystal structure of TFIID with its cognate DNA.

Leucine zipper proteins

The yeast transcriptional activator GCN4 contains a basic region called leucine-zipper (bZIP), a conserved element involved in dimerization and DNA-binding which has been found in a number of eukaryotic transcription factors. The recently determined 2.9 Å resolution crystal structure (Ellenberger *et al.*, 1992) of the bZIP with its cognate DNA AP1 site, as well as NMR structures in solution (Saudek *et al.*, 1991) show that the protein binds DNA as a Y shaped dimer that recognizes binding sites on DNA of directly abutted, dyad-symmetric half sites (Figure II.3). The stem of such a Y-shaped dimer is a coiled pair of amphipathic α-helices adhering together through interchain hydrophobic Leu-Leu interactions occurring every seven repeats i.e. the leucine zipper. Theoretical observations coupled to experimental evidence are consistent with the α-helical arrangement of the leucine zipper. In fact (i) no α-helix destabilizing residues (Pro and Gly) are present in the leucine zipper and (ii) there is a high occurrence of oppositely charged residues configured three or four aminoacids apart, thus stabilizing the helix. From the experimental point of view, disulfide crosslinking experiments (O'Shea *et al.*, 1989), spectroscopic studies (Landschulz *et al.*, 1989) and mutational studies substantiate the hydrophobic spine model punctuated every seven residues by a Leu side chain, where α-helices adhere to one another in parallel forming a coiled coil (Figure II.3). The bifurcating arms are a linked set of DNA contact surfaces in the form of α-helices, each arm being positioned along each half of the recognition site in opposite directions (Figure II.3) i.e. the basic motifs. The bifurcation point is located on the rotational axis of the dimer and it contacts the major groove at the very centre of the dyad symmetric binding site. *In vitro* binding studies with chimeric proteins (Agre et *al.*, 1989) demonstrate that the region directly responsible for DNA binding is immediately amino-terminal to the leucine-zipper and is rich in basic residues. The GCN4 basic α-helices lie nearly parallel to the major groove, one on each side of the DNA duplex and this mode of binding thus supports the "induced helical fork" model (O'Neal *et al.*, 1990). An alternative model, the "scissors grip", predicts a kinking of the α-helices and pronounced bending of cognate DNA (Vinson *et al.*, 1989). Sequence

A B

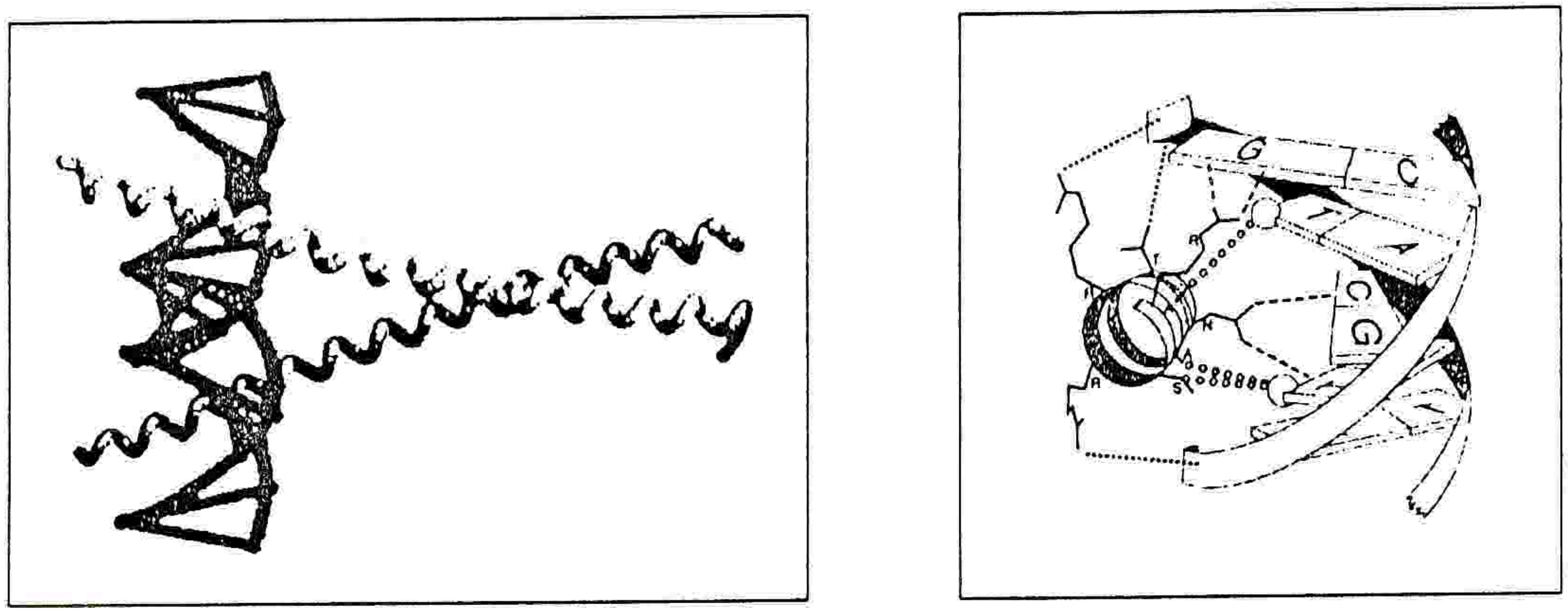


Figure II.3. Structure of the GCN4/DNA complex. (A) Cα backbone structure of GCN4 bound to DNA. (B) Contacts formed in a half-site between residues in a GCN4 monomer basic helix and DNA bases and backbone. Dashed lines indicate H-bonding to bases, dotted lines show H-bonds to phosphate groups, open circles indicate van der Waals interactions (Ellembergere, 1994 and references therein).

alignment of proteins containing the leucine zipper basic motif domain show a high level of conservation of this motif (Busch and Sassone-Corsi, 1990).

In vivo, GCN4 is functional only when it forms a homodimer. On the other hand other regulatory proteins that carry out their roles through the typical bZIP domain are functional only upon heterodimer formation, one example given by the interaction between Fos and Jun which produce a functional dimer only when they are bound to each other (Patel *et al.*, 1990). Direct interaction of the Fos and Jun nuclear oncoproteins has been shown to be via the leucine zipper domain. Sassone-Corsi *et al.* (1988) report that the Fos protein directly modulates Jun function in this manner. The Fos leucine zipper domain is necessary for the DNA binding of the heterodimer whereas a distinct domain, localized in the carboxy-terminal region of the Fos protein, is responsible for transcriptional regulation. This heterodimeric model is also supported by the evidence that the Fos-Jun complex binds the AP-1 site with an approximately 30 times greater affinity than Jun alone, primarily because of stabilization of the protein-DNA interaction. A mutagenesis analysis (Gentz *et al.*, 1989) shows that Fos and Jun do indeed dimerize via their respective leucine zipper domains and the association appears to occur in a parallel rather than an antiparallel configuration.

Beta Proteins.

There are two classes of prokaryotic proteins that interact with DNA through β-ribbon recognition elements (Phillips, 1991). The first class includes the Met J repressor (Figure II.5) of *E. coli* and the *arc* and *mnt* repressors of Salmonella phage P22. The second class includes the protein HU (Table 1.1), which forms condensed nucleoprotein structures in many prokaryotes, the *E. coli* integration host factor IHF, and the *B. subtilis* phage transcription factor TF1. The latter two bind to specific sites.

The structure of Met J has been determined by X-ray crystallography (Rafferty *et al.*, 1989), and the structure of Arc by NMR (Berg *et al.*, 1990).

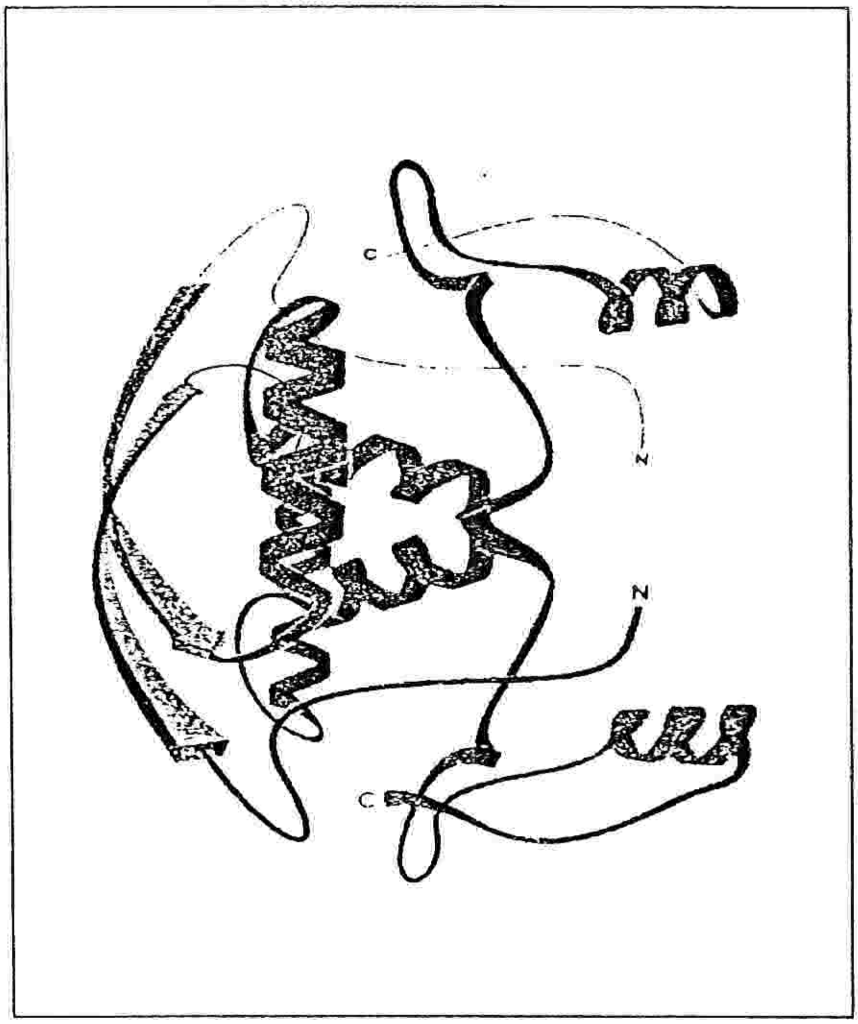


Figure II.4. The Met J dimer (Rafferty *et al.*, 1989). The β-ribbon at the left lies in the major groove of DNA when the protein binds to a met box (Phillips, 1991). The bound DNA would be oriented roughly vertically at the left of the diagram. The helix in the foreground interacts with a corresponding helix from another dimer, when the protein is bound cooperatively at adjacent sites.

These proteins are dimers with a core composed of four α-helices, two from each subunit; Met J contains an additional C-terminal helix (Figure II.5). An antiparallel β-ribbon, formed from an N-terminal segment contributed by each monomer, protrudes from the core. In the specific DNA complex of Met J that has been studied crystallographically, the β-ribbon lies in the major groove (Rafferty *et al.*, 1989). This recognition element requires dimerization for its folded structure. Naturally occurring operators contain at least two tandem repeats of an 8 bp "met box". The consensus met box sequence is twofold symmetric, and one repressor dimer binds to each 8 bp element (Phillips *et al.*, 1989). The crystal structure of the complex shows that the first of the core helices is so positioned that it contacts in antiparallel fashion a corresponding helix from the adjacent dimer (Phillips, 1991). Binding is therefore cooperative.

HU is believed to contact the minor groove. The X-ray structure has been determined (Phillips, 1991 and references therein). It has a core composed of two helices from each monomer, and two projecting β-ribbons, each formed by the C-terminal third of the polypeptide chain. The tips of these ribbons are disordered, presumably so that they can wrap around the DNA helix. Model building can not distinguish between wrapping in the major or minor groove interactions for the homologous IHF (Harrison, 1991 and references therein).

The Helix-loop-Helix structure.

This domain comprises a carboxy-terminal dimerization segment and an amino-terminal portion (basic region) which consists of two α-helices rich in basic residues (Ellemberger, 1994). The basic region is unstructured in solution but it becomes α-helical when bound to cognate DNA. The elements responsible for dimerization on the other contribute to the specificity of binding essentially by proper positioning the DNA recognition basic region. The precise orientation of the binding unit depends on the complementarity of the amino acid side chains with the DNA bases as well as some constraints of α-helical stability. Due to its extended stucture which makes the basic region well exposed to the solvent, the bHLH as well as the bZIP proteins have less constraints as compared to the quite rigid structure of the helix-turn-helix motif (HTH). This adaptability of the bHLH motif at the protein-DNA interface may represent an important aspect of sequence recognition as it may contribute to the different binding specificity of the members of the family (as well as for the bZIP protein family) (Kim *et al.*, 1993). Several transcription factors belonging to this family have now been characterized and their structure determined. Interestingly some of them like Max (Kato *et al.*, 1992; Blackwood and Eisenmann, 1991), Myc (Amati *et al.*, 1993; Amati *et al.*, 1992; Davis and Halazonetis, 1992; Prendergast and Ziff, 1992), TFE3 (Beckmann and Kadesch, 1991) and TFEB (Fisher *et al.*, 1991) require a zipper segment in addition to the HLH for efficient dimerization and DNA binding. Most bHLH and bHLH/ZIP proteins are active only upon heterodimerization with another member of the same family which often occurs through a leucine zipper element. Myc homodimerizes poorly, but will readily form heterodimers with Max (Figure II.6) that bind to DNA with high affinity (Ellenberger, 1994; references citated therein).

Many DNA-binding proteins therefore recognize specific sites through small discrete domains. In some cases they can be interchanged between proteins showing that they are independently folded units. Moreover, eukaryotic transcription factors also contain domains that mediate the formation of homo- and heterodimers. The widespread occurrence of such domains suggests that they are important to the genetic regulatory apparatus.

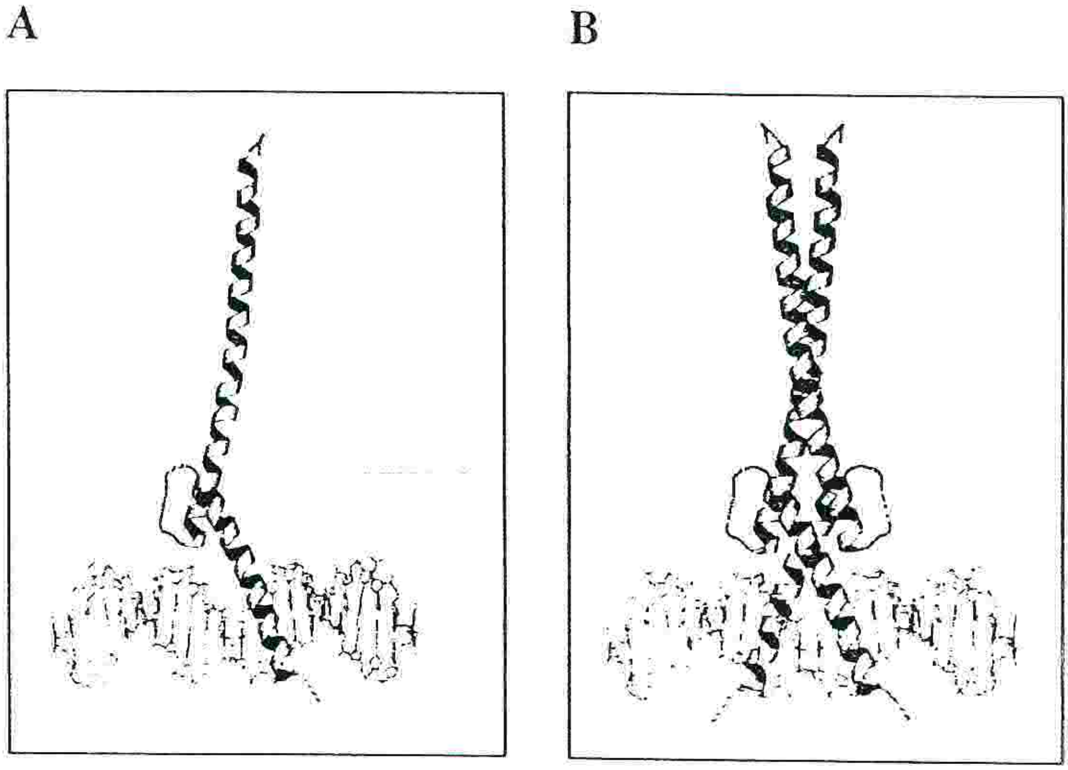


Figure II.5. Schematic representations of the (Max 22-113)-DNA complex. (A) View of a monomer interacting with 22 bp of DNA (Ferre´-D'Amare´ *et al.*, 1993). (B) Same view as in (A), showing the homodimer-DNA interaction.

BIBLIOGRAPHY

Adam G., Delbruck M. (1968) In *Structural Chemistry and Molecular Biology*, Rich A. and Davidson N. (eds), pp. 198-215, Freeman and Co., San Francisco

Aggarwal A. K., Rodger D. W., Drottar M., Ptashne M., Harrison S. C. (1988) *Science* **242**, 899-907

Agre P., Johnson P.F., McKnight S.L.(1989) *Science* **246**, 922

Alber T. (1993) *Curr. Biol.* **3**, 182

Amati B., Brooks M. W., Levy N., Littlewood T. D., Evan G. I., Land H. (1993) *Cell* **72**, 233-245

Amati B., Dalton S., Brooks M. W., Littlewood T. D., Evan G. I., Land H. (1992) *Nature* **359**, 423-426

Amit A. G., Mariuzza R. A., Phillips S. E. V., Poljak R. J. (1986) *Science* **233**, 747-753

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

Anderson W. F., Ohlendorf D. H., Takeda Y., Matthews B. W. (1981) *Nature* **290**, 754-758

Anthony-Cahill S. J., Benfield P. A., Fairman R., Wasserman Z. R., Brenner S. L., Stafford W. S. III, Altenbach C., Hubbel W. L., DeGrado W. F. (1992) *Science* **255**, 979-983

Assa-Munt N., Mortishire-Smith R., Aurora R., Herr W., Wright P. E. (1993) *Cell* **73**, 193-205

Atherton E., Sheppard R. C. (1989) *Solid Phase Peptide Synthesis*, IRL PRESS, Oxford Washington D. C.

Auer M., Gremlich H-U., Seifert J-M., Daly T. J., Parslow T. G., Casari G., Gstach H. (1994) *Biochemistry* **33**, 2988-2996

Baleia J. D., Marmorstein R., Harrison S. C., Wagner G. (1992) *Nature* **356**, 450-453

Barkley M. D. (1981) *Biochemistry* **20**, 3833-3842

Beckmann H., Kadesh T (1991) *Genes Dev.* **5**, 1057-1066

Benevides J. M., Weiss M. A., Thomas G. J. Jr. (1991) *Biochemistry* **30**, 4381-4388

Berg J. M. (1993) *Curr. Op. Struct. Biol.* **3**, 11-16

Berg J. N., van Opheusden J. H. J., Burgering M. J. M., Boelens R., Kaptein R. (1990) *Nature* **346**, 586-589

Berg O. G., von Hippel P. H. (1988) *J. Mol. Biol.* **200**, 700-723

Berg O. G., von Hippel P. H. (1987) *J. Mol. Biol.* **193**, 723-750

Berg O. G., von Hippel P. H. (1985) *Annu. Rev. Biophys. Biophys. Chem.* **14**, 131-160

Bess J. W. Jr., Powell P., Issq H. J., Shumak L. J., Grimes M. K., Henderson L. E., Arthur L. O. (1992) *J. Virol.* **66**, 840-847

Blackwood E. M., Eisenmann R. N. (1991) *Science* **251**, 1211-1217

Botfield M. C., Jancso A, Weiss M. A. (1992) *Biochemistry* **31**, 5841-5848

Branden C., Tooze J. (1991) *Introduction to Protein Structure*, Garland Publishing, Inc., New York and London

Brennan R. G. (1993) *Cell* **74**, 773-776

Brennan R. G., Roderick S. L., Takeda Y., Matthews B. W. (1990) *Proc. Natl. Acad. Sci. U.S.A.* **87**, 8165-8169
Brennan R. G., Matthews B. W. (1989) *Trends Biochem.* **14**, 286-290
Busch S. J., Sassone-Corsi P. (1990) *Trends Gen.* **6**, 36-40
Cantor C. R., Shimmel P. R. (1980) *Biophysical Chemistry Part II: Techniques for the Study of Biological Structure and Function,* W. H. Freeman and Company, San Francisco
Carey J. (1988) *Proc. Natl. Acad. Sci. U. S. A.* **85**, 975-979
Chakrabarty A., Kortemme T., Padmanabhan S., Baldwin R. L. (1993) *Biochemistry* **32**, 5560-5565
Colman P. M. (1988) *Adv. Immunol.* **43**, 99-132
Cregut D., Liautard J. P., Heitz F., Chiche L. (1993) *Protein Eng.* **6**, 51-58
Crick F. H. C., Kendrew J. C. (1957) *Adv. Protein Chem.* **12**, 133
Dallmann G., Papp P., Orosz L. (1987) *Nature* **330**, 398-401
Davis L. J., Halozenetis T. D. (1992) *Oncogene* **7**, 125-132
Dekker N., Cox M., Boelens R., Verrijzer C. P., van der Vliet P. C., Kaptein R. (1993) *Nature* **362**, 852-855
Desjarlais J.R., Berg J.M. (1992a) *Proc. Natl. Acad. Sci. U.S.A.* **89**, 7345-7349
Desjarlais J.R., Berg J.M. (1992b) *Proteins* **12**, 101-104
Dickerson R. E., Drew H. R.(1981) *J. Mol. Biol.* **149**, 761-786
Dodd I. B., Egan B. J. (1990) *Nucl. Acids Res.* **18**, 5019-5026
Ellenberger T.E. (1994) *Curr. Opin. Struc. Biol.* **4**, 12-21
Ellenberger T.E., Brandl C.J., Struhl K., Harrison S.C. (1992) *Cell* **71**, 1223-1237
Ferre`-D'Amare` A. R., Prendergast G. C., Ziff E. B., Burley S. K. (1994) *Nature* **363**, 38-45
Fisher D. E., Carr C. S., Parent L. A., Sharp P. A. (1991) *Genes Dev.* **5**, 2342-2352
Fong R. S-C., Woody S., Gussin G. N. (1993) *J. Mol. Biol.* **232**, 792-804
Frankel A. D., Kim P. S. (1991) *Cell* **65**, 717-719
Freemont, P.S., Lane, A.N. and Sanderson,M.R. (1991) *Biochem. J.* **278**, 1-23
Fried M., Crothers D. M. (1981) *Nucleic Acids Res.* **9**, 6505-6525
Galas D. J., Schmitz A. (1978) *Nucleic Acids Res.* **5**, 3157
Garner M. M., Revzin A. (1981) *Nucleic Acids Res.* **9**, 3047-3060
Gentz R., Rauscher F. J. III, Abate C., Curran T. (1989) *Science* **243**, 1695-1699
Gerhing W. J., Quian Y. Q., Billeter M., Furukubo-Tokunaga K., Schier A. F., Resendez-Perez D., Affolter M., Otting G., Wuthrich K. (1994) *Cell* **78**, 211-223
Gerhing W. J. (1987) *Science* **236**, 1245-1252
Grokhovsky S. L., Surovaya A. N., Brussov R. V., Chernov B. K., Sidorova N. Yu., Gursky G. V. (1991) *J. Biomol. Struct. Dynamics* 8, 989-1025
Hard T., Kellenbach E., Boelens R., Maler B.A., Dahlman K., Freedman L.P., Carlstedt Duke J., Yamamoto K.R., Gustafson L. A., Kaptein R. (1990) *Science* **249**, 157-160
Harrison S.C. (1991) *Nature* **353**, 715-719
Harrison S.C., Aggarwal A. (1990) *Annu. Rev. Biochem.* **59**, 933-969
Hedge R.S., Grossman S.R., Laimins L.A., Sigler P.B. (1992) *Nature* **359**, 505-512
Hu J. C., O'Shea E. K., Kim P. S., Sauer R. T (1990) *Science* **250**, 1400-1403

Huynh T. V., Young R. A., Davis R. W. (1985) In *DNA Cloning Vol. I*, Glover D. M. (ed.), IRL PRESS, Oxford Washington D. C., pp. 49-78
Johnson N. P., Lindstrom J., Baase W. A., von Hippel P. H. (1994) *Proc. Natl. Acad. Sci. U. S. A.* **91**, 4840-4844
Johnson A.D., Pabo C.O., Sauer R.T. (1980) *Methods Enzymol.* **65**, 839
Johnson A.D., Meyer B.J., Ptashne M. (1979) *Proc. Natl. Acad. Sci. U.S.A.* **76**, 5061
Jordan S. R., Pabo C. O. (1988) *Science* **242**, 893-899
Kato G. J., Lee W. M. F., Chen L., Dang C. V. (1992) *Genes Dev.* **6**, 81-92
Kim J., Tzamarias D., Ellemberger T., Harrison S. C. (1993) *Proc. Natl Acad. Sci. U. S. A.* **90**, 4513-4517
Kim J. G., Takeda Y., Matthews B. W., Anderson W. F. (1987) *J. Mol. Biol.* **196**, 149-158
Kissinger C.R., Lju B., Martin-Blance E., Kornberg T.B., Pabo C.O. (1990) *Cell* **63**, 579-590
Klemm J. D., Rould M. A., Aurora R., Herr W., Pabo C. O. (1994) *Cell* **77**, 21-32
Klevit R.E., Herriott J.R., Horvath S.J. (1990) *Proteins* **7**, 215-226
Kornberg T. B. (1993) *J. Biol. Chem.* **268**, 26813-26816
Koshland D. E. Jr (1958) *Proc. Natl. Acad. Sci. U. S. A.* **44**, 98
Kraulis P.J., Raine A.R.C., Gadhavi P.L., Laue E.D. (1992) *Nature* **356**, 448-450
Kriwacki R. W., Schulz S. C., Steitz T. A., Caradonna J. P. (1992) *Proc. Natl. Acad. Sci. U.S.A.* **89**, 9759-9763
Iizuka E., Yang J. T. (1966) *Proc. Natl. Acad. Sci. U. S. A.* **55**, 1175
Landshulz W.H., Johnson P.F., McKnight S.L. (1989) *Science* **243**, 1681-1688
Lee M.S., Gippert G.P., Soman K.V., Case D.A., Wright P.E. (1989) *Science* **245**, 635-637
Lieberman M., Tabet M., Sasaki T. (1991) *Peptides 1991 (Proceedings of the 12th American Peptide Symposium)*, Smith J. A. and Rivier J. A. (eds.), ESCOM Leiden 1992, pp. 332-334
Luisi B. F., Xu W. X., Otwinowski Z., Freedman L. P., Yamamoto K. R., Sigler P. B. (1991) *Nature* **352**, 497-505
Lumb K. J., Carr C. M., Kim P.S. (1994) *Biochemistry* **33**, 7361-7367
Mair E.A., Churchill, Travers, A.A. (1991) *TIBS* **16**, 92-97
Marmorstein R., Carey M., Ptashne M., Harrison S.C. (1992) *Nature* **356**, 408-414
Markovitz A. (1972) *Biochim. Biophys. Acta* **281**, 522-534
Maxam A.M., Gilbert W. (1980) *Methods Enzymol.* **65**, 499
McClure W. R. (1985) *Annu. Rev. Biochem.* **54**, 171-204
McGinnis W., Krumlauf R. (1992) *Cell* **68**, 283-302
McKay D. B., Steitz T. A. (1981) *Nature* **290**, 744-749
Mondragon A., Harrison S. C. (1991) *J. Mol. Biol.* **219**, 321-334
Mondragon A., Wolberger C., Harrison S. C. (1989a) *J. Mol. Biol.* **205**, 179-188
Mondragon A., Subbiah T., Almo S.C., Drottar M., Harrison S.C. (1989b) *J. Mol. Biol.* **205**, 189-201
Moore W. T. (1993) *Biol. Mass Spectrometry* **22**, 149-162
Mullis K.B., Faloona F. (1987) *Methods Enzimol.* **155**, 335-350
Neri D., Billeter M., Wuthrich K. (1992a) *J. Mol. Biol.* **233**, 743-767

Neri D., Billeter M., Wider G., Wuthrich K. (1992b) *Science* **257**, 1559-1563

Niklov D.B., Hu S.H., Lin J.P., Gasch A., Hoffmann A., Horikoshi M., Chua N.H., Roeder R.G., Burley S.K. (1992) *Nature* **360**, 40-46

Ohlendorf D. H., Anderson W. F., Fisher R. G., Takeda Y., Matthews B. W., (1982) *Nature* **298**, 718-723

O′ Neal K.T., Shuman J. D., Ampe C., DeGrado W. F. (1991) *Biochemistry* **30**, 9030-9034

O′ Neal K. T., Hoess R. H., DeGrado W. E. (1990) *Science* **249**, 774-778

O′Shea E. K., Rutkowski R., Kim P. S. (1989) *Science* **243**, 538

Otting G., Qian Y. Q., Billeter M., Mueller M., Affolter M., Gerhing W. J., Wutrich K. (1990) *EMBO J.* **9**, 3085-3092

Otwinowski Z., Shevitz R. W., Zang R. G., Lawson C. L., Joachimiak A., Marmorstein R. Q., Luisi B. F., Sigler B. P. (1988) *Nature* **335**, 321-329

Pabo C. O., Sauer R. T (1992) *Annu. Rev. Biochem.* **61**, 1053-1095

Pabo C. O., Aggarwal A. K., Jordan S. R., Beamer L. J., Obeysekare U. R. et al. (1990) *Science* **247**, 1210-1213

Pabo C. O. (1984) *Specificity in Protein-DNA interactions.* Proc. Robert A. Welch Found. Conf. Chem. Res. XXVII, Stereospecificity in Chemistry and Biochemistry, pp. 222-255

Pabo C. O., Lewis M. (1982) *Nature* **298**, 443-447

Papavassiliou A. G. (1993) *Nucleic Acids Res.* **21**, 757-758

Patel L., Abate C., Curran T. (1990) *Nature* **347**, 572-575

Pavletich N.P., Pabo C.O. (1991) *Science* **252**, 809-817

Perbal B. (1988) *A Practical Guide to Molcular Cloning*, 2nd edn., John Wiley & Sons, Inc., New York

Percipalle P., Zakhariev S., Guarnaccia C., Tossi A., Cserzo M., Simocsits A., Pongor S. (1994a) *Peptides 1994 (Proceedings of the 23rd European Peptide Symposium)*, in press

Percipalle P., Saletti R., Pongor S., Foti S., Tossi A., Fisichella S. (1994b) *Biol. Mass Spectrometry*, in press

Percipalle P. *et al.* (1994), manuscript in preparation

Phillips S. E. V. (1991) *Curr. Opin. Struct. Biol.* **1**, 89-98

Phillips S. E. V., Manfield I., Parsons I., Davidson B. E., Rafferty J. B., Somers W. S., Margarita D., Cohen G. N., Saint-Girons I., Stockley P. G. (1989) *Nature* **341**, 711-714

Pjura, P.E., Grzeskowiak, K. and Dickerson, R.E. (1987) *J. Mol. Biol.* **197**, 257-271

Prendergast G. C., Ziff E. B. (1992) *Trends Genetics* **8**, 91-96

Provencer S. W. (1984) EMBL technical report DA07

Ptashne M. (1992) *A Genetic Switch*, Cell Press & Blackwell Scientific Publications

Qian Y.Q., Billeter M., Otting G., Muller, Gehring W.J., Wuthrich (1989) *Cell* **59**, 573-580

Rafferty J. B., Somers W. S., Saint-Girons I., Phillips S. E. V. (1989) *Nature* **341**, 705-710

Richter P. H., Eigen M. (1974) *Biophys. Chem.* **9**, 255-263

Saiki R. K., Gelfand D. H., Sttoffel S., Sharf S., Higuchi R., Horn G. T., Mullis K. B., Erlich H.A. (1988) *Science* **239**, 487-491

Saiki R. K., Scharf S., Faloona F., Mullis K. B., Horn G. T., Erlich H. A., Arnheim N. (1985) *Science* **230**, 1350-1354

Sambrook J., Fritsch E. F., Maniatis T. (1989) *Molecular Cloning*, Cold Spring Harbor Laboratory Press

Sarkar P. K., Doty P. (1966) *Proc. Natl. Acad. Sci. U. S. A.* **55**, 981

Sassone-Corsi P., Ransone L. J., Lamph W.W., Verma I.M (1988) *Nature* **336**, 692-695

Saudek V., Pastore A., Morelli M. A., Frank R., Gausepohl H., Gibson T. (1991) *Protein Eng.* **4**, 519-529

Schmid F. X. (1989) In *Protein Structure*, Creighton T. E. (ed.), IRL PRESS, Oxford New York Tokyo, pp. 251-284

Schevitz R. W., Otwinowski Z., Joachimiak A., Lawson C. L., Sigler P. B. (1985) *Nature* **317**, 782-786

Schulz S. C., Shields G. C., Steitz T. A. (1991) *Science* **253**, 1001-1007

Schulz S. C., Shields G. C., Steitz T. A. (1990) *J. Mol. Biol.* **213**, 159-166

Scopes T. (1974) *Anal. Biochem.* **59**, 277

Scott M. P., Tamkun J., Hartzell G. W. III (1989) *Biochim. Biophys. Acta* **989**, 25-48

Shirakawa M., Matsuo H., Kyogoku Y. (1991) *Protein Eng.* **4**, 545-552

Simoncsits A. *et al.* (1994) manuscript in preparation

Sivaraja M., Botfield M. C., Mueller M., Jancso A., Weiss M. A. (1994) *Biochemistry* **33**, 9845-9855

Spolar R. S., Record T. M. Jr. (1994) *Science* **263**, 777-784

Starr D. B., Hawley D. K. (1991) *Cell* **67**, 1231-1240

Stewart J. M., Cann J. R., Hahn K. W., Klis W. A. (1991) *Peptides 1991 (Proceedings of the 12th American Peptide Symposium)*, Smith J. A. and Rivier J. A. (eds.), ESCOM Leiden 1992, pp. 335-337

Studier F. W., Bandyopadyay P. K. (1988) *Proc. Natl. Acad. Sci. U. S. A.* **85**, 4677-4681

Suck, D., Lahn, A. and Oefner, C. (1988) *Nature* **322**, 464-468

Summers M. F., Henderson L. E., Chance M. R., Bess J. W. Jr., South T. L., Blake P. R., Sagi I., Peret Alvarado G., Sowder R. C. III, Hare D. R., Arthur L. O. (1992) *Protein Sci.* **1**, 563-574

Swabe J. W. R., Neuhaus D., Rhodes D. (1990) *Nature* **328**, 458-461

Talanian R. V., McKnight J. C., Rutkowski R., Kim P. S. (1992) Biochemistry **31**, 6871-6875

Talanian R. V., McKnight J. C., Kim P. S. (1990) *Science* **249**, 769-778

Teng, M., Usman, N., Frederick, C. A. and Wang, A. H-J. (1988) *Nucleic Acids Research* **16**, 2671-2690

Thompson K. S., Vinson C. R., Freire E. (1993) *Biochemistry* **32**, 5491-5496

Thukral S. K., Morrison M. L., Young E. T. (1992) *Mol. Cell. Biol.* **12**, 2784-2792

Townend R., Kumosinski R. F., Timasheff S. N., Fasman G. D., Davidson B. (1966) *Biochem. Biophys. Res. Commun.* **23**, 163

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

Travers, A.A. (1988) In *Nucleic Acids and Molecular Biology*, Eckstein F., Lilley D. M. J. (eds), pp. 136-148, Springer-Verlag Berlin Heidelberg

Verrijzer C. P., van der Vliet P. C. (1993) *Biochim. Biophys. Acta* **1173**, 1-21

Vinson C. R., Sigler P. B., McKnight S. L. (1989) *Science* **246**, 911-916

von Hippel P. H., Berg O. G. (1989) *J. Biol. Chem.* **264**, 675-678

von Hippel P. H., Berg O. G. (1986) *Proc. Natl. Acad. Sci. U. S. A.* **83**, 1608-1612

von Hippel P. H., Revzin A., Gross C. A., Wang A. C. (1974) *Proc. Natl. Acad. Sci. U. S. A.* **71**, 4808-4812

Yang J. T., Wu C-S. C., Martinez H. (1986) *Methods Enzymol.* **130**, 208

Warwicker J., Engelman B. P., Steitz T. A. (1987) *Proteins* **2**, 283-289

Weiss M. A., Ellemberger T., Wobbe C. R., Lee J. P., Harrison S. C., Struhl K. (1990) *Nature* **347**, 575-578

Williams J. S., Dixon J. E., Andrisani O. M. (1993) *DNA Cell Biol.* **12**, 183-190

Winter R. B., von Hippel P. H. (1981) *Biochemistry* **20**, 6948-6960

Winter R. B., Berg O. G., von Hippel P. H. (1981) *Biochemistry* **20**, 6961-6977

Wolberger C. (1993) *Curr. Op. Struct. Biol.* **3**, 3-10

Wolberger C., Vershon A.K., Lju B., Johnson A.D., Pabo C.O. (1991) *Cell* **67**, 517-528

Wolberger C., Dong Y., Ptashne M., Harrison S. C. (1988) *Nature* **335**, 789-795

Wuthrich K., Gerhing W. (1992) In *Transcriptional Regulation*, McKnight S. L. and Yamamoto K. R. (eds.), Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York

Wu C.-S. C., Ikeda K., Jang J. T. (1981) *Biochemistry* **20**, 566-570

ACKNOWLEDGEMENTS

This thesis work was carried out at the International Centre for Genetic Engineering and Biotechnology, Protein Structure and Function Group, under the supervision of Prof. Sándor Pongor and Dr. András Simoncsits. Without their help and advice this work could not have been accomplished.

I wish to thank Dr. Sotir Zakhariev and Dr. Corrado Guarnaccia for advices and suggestions concerning the syntheses of the various peptides. I would also like to thank Prof. Salvatore Foti, University of Catania, for providing mass spectra of the synthetic peptides and for his deep interest in collaborating with me as well as Prof. Sebastiano Sciuto, my former thesis supervisor at the University of Catania. Moreover, I wish to thank my former collegues at ICGEB, Dr. Alessandro Tossi (now at the University of Trieste, Dept. of Biochemistry) and Dr. Miklos Cserzo (now at the Universite` de Nancy, France) for their work on the initial phases of the project and for their advices. I am also grateful to Roberto Sanchez (ICGEB) for his invaluable help with computer graphics.

I would also like to thank all the people of the Protein Structure and Function Group as well as all my SISSA collegues and friends with whom I really had a nice time during my staying in Trieste.

Finally I wish to express my thanks to Prof. Arturo Falaschi, Director General of ICGEB, for his encouragement throughout my studies and to Prof. Alberto Valvassori, Director General of AssoBiotech, Milan, for his enthusiasm in supporting me during these years.