CHAPTER I

INTRODUCTION: DNA-PROTEIN INTERACTIONS
INTRODUCTION

Nucleic acid-protein interactions form one of the central topics in molecular biology because of the associated important functions. At various stages in the life span of a cell these interactions play an important part in the smooth functioning of the vital processes. These vital processes include regulation of gene expression, protein synthesis, protection and repair mechanisms of nucleic acids. In a broader perspective nucleic acid-protein interactions (Ollis and White 1987, Schleif 1988) can be divided into three categories (1) single strand DNA-protein interactions (Chase and Williams 1986) (2) double strand DNA-protein interactions and (3) RNA-protein interactions. Single strand DNA-protein interactions include recombinase A of E.coli, single strand binding proteins (SSB's) of DNA repair mechanism and gene 5 protein of bacteriophage fd. The proteins such as RNA polymerase, DNA polymerase, repressors, transcription factors, restriction endonucleases and exonucleases belong to the second category. Ribosomal proteins, tRNA aminoacyl synthetase and RNases are some of the examples of proteins interacting with RNA.

For convenience of analysis and understanding, these recognition processes can also be divided into two categories. (1) general recognition and (2) sequence-specific recognition. General or nonspecific recognition involves proteins with proper-
ties which are quite flexible for interaction with different kinds of structures posed by random sequence of the nucleic acids, for example nuclease and DNA-binding protein (HU) of E.coli (Ollis and White 1987). The specific recognition category comprises of proteins which bind to particular nucleic acid sequences in the whole genome to perform specific function. Most of the double stranded DNA binding proteins mentioned above fall into this category. For this kind of interactions, certain combinations of basic structural elements in proteins have evolved in different organisms to do similar functions and these motifs would possibly follow a kind of code (Takeda et.al., 1983, Pabo and Sauer 1984, Matthews 1988). Their chemical and physical properties can also be expected to be stringent and less flexible.

On the basis of the structure and function, reasonably well investigated DNA-protein interaction systems at present are operator-repressors in prokaryotes (Ptashne 1984 and Ptashne 1986 Brennan and Matthews 1989). Information on RNA polymerase-p promoter interactions (von Hippel et.al., 1984), restriction enzymes and transcription factors (Gluzman 1985 and Ptashne 1988, Rangarajan and Padmanabhan 1989 and Struhl 1989) are available currently only to a limited extent. Among these, operator-repressor interactions assume significance because of the
availability of sufficient structural data (Matthews 1988) and the potential possibility of extending the derived general principles to various other systems (Ohlendorf and Mathews 1983 and Pabo and Sauer 1984). Crystal structures of two repressor proteins and an activator protein (Cro - Anderson et al., 1981, CI - Pabo and Lewis 1982, CAP - McKay and Steitz 1981) and co-crystal structures of five sets of operator-repressor complexes (Anderson et al., 1987, Otwinowski et al., 1988 and Aggarwal et al., Jordon et al., Wolberge et al., in press) are known which exhibit some common structural motifs (Takeda et al., 1983). Crystal structure of one restriction enzyme-DNA complex (Fredrick et al., 1984) is known which indicate interesting features in these systems. In the non specific interactions category, structures of about 5 proteins are known and these do not seem to have any commonality (Ollis and White 1987). Even less is known regarding more complex histone-DNA interactions and dynamics (Saenger 1984). Equally important like the X-ray structure studies are the structural studies in solution phase which yield details regarding the dynamics of interacting systems, the pre-requisites for complexation and the resulting conformational effects in the components (Helene and Lancelot 1982). In this chapter, an attempt has been made to summarise salient features of nucleic acid-protein interactions. This is not intended to be an exhaustive review on this topic. The
contents are oriented particularly towards operator-repressor interactions and sometimes referring to their implications for the general understanding of nucleic acid-protein interactions. Various parameters that are involved in studying protein-nucleic interactions, their importance, the methods for their determination and the overall conclusions are described briefly in the following sections.

1.2. STRUCTURAL FEATURES IN PROTEINS FOR DNA RECOGNITION

For a protein to recognise a specific sequence of DNA and to have considerable stability, some structural elements of the protein in three dimension should approach DNA spacially close enough to interact through maximum non covalent-bonds (hydrogen, hydrophobic and electrostatic bonds) (Kim 1983). In other words, there should be a close topological fit between two molecules for making maximum favourable contacts.

An initial idea on the kinds of structural elements present in proteins that can fit or recognise specific structures of DNA can be obtained by model building studies (either space filling or computer aided) (Kim 1983 and Rein et.al., 1983). Of the different secondary structural elements found in proteins α-helices, parallel and antiparallel β-sheets, turns and extended polypeptides, model building studies have mostly shown that helices and sheets are important for interactions with nucleic
acids (Fig.1). The physical parameters of α-helices match with those of the major groove of DNA while parallel and antiparallel β-sheets match with the minor groove. When α-helix fits into the major groove (Fig.1a) there can be no symmetry correlations and the peptide backbone being inside the helix do not form any contacts. Contacts are made by the side chains of basic amino acids like Arg, Lys, His, which can either form electrostatic bonds with the phosphate backbone or form hydrogen bond with the bases exposed in the groove. Amino groups of Arg and Lys in position 1 and 4 of an aminoacid chain in a helix can recognise a nucleic acid base pair because these amino acids will be in the plane with the base pair to form hydrogen bonds. The approximate interphosphate distance of 7Å matches with the 7Å distance between interamino groups on adjacent amino acid side chains such as lysines so they can form salt bridges (Kim 1983).

In the case of antiparallel β-ribbon-DNA minor groove recognition (Kim et.al., 1975 and Church et.al., 1977), the symmetry and structural parameters are mutually compatible. Here the peptide backbone or the basic residues of adjacent amino acids can hydrogen bond with the phosphate moieties on the opposite strand (Fig.1c).

These ideas can be utilised to build models when the structure of one of the components is known in detail. Such
Fig. 1. Models for DNA-protein interactions. (a) Helix in major groove of the DNA. No peptide backbone-DNA interaction possible but side chains of basic amino acids 1 and 2 or 2 and 5 can bond to two opposite and two adjacent phosphates respectively. (b) Helix in minor groove of B DNA. Side chains of basic residues have to bend for interaction, so are under strain. (c) Antiparallel $\beta$ ribbon in minor groove of DNA. In this structure alternating -NH groups of peptide can hydrogen bond to the phosphate groups of DNA. Polarity of ribbon is antiparallel to the DNA chains. (d) Antiparallel $\beta$ ribbon in minor groove of DNA - but polarity of the peptide ribbon matches with that of DNA.
predicted models were consistent with the experimentally
determined structures in the case of Cro (Weber and Steitz 1984)
CAP (Ohlendorf et al., 1983) and CI repressor (Lewis et al.,
1982).

1.3. FORCES INVOLVED IN INTERACTION BETWEEN PROTEINS AND DNA:

Three major kinds of molecular interactions that are possible between structural elements of proteins and that of DNA are (1) electrostatic bonds (2) hydrogen bonds and (3) hydrophobic bonds.

1.3.1. Electrostatic bonds:

Most of the nucleic acid binding proteins are basic with positive charges due to the side chain amino groups of lysine, arginine and histidine amino acids. These positively charged side chains can involve in electrostatic interactions with the negatively charged phosphate groups that are in large number along the DNA backbone. Major evidence for this has come from the salt dependence of these interactions. The forces due to electrostatic bonds contribute to the stability and bulk of the free energy of the complex formation. However, they cannot confer specificity as phosphate groups are uniformly distributed over DNA and RNA. The strength of electrostatic bonds are independent of relative orientation of the two constituent polar groups but depend on charge magnitudes.
The number of electrostatic bonds that exist between DNA and protein in the complex molecule can be estimated by determining the association constant as a function of ionic strength (Helene and Lancelot 1982). Typical values that were calculated for lac operator-repressor as $8 \pm 1$ and for RNA polymerase and promoter as 11 (Strauss et al. 1980 and Record et al. 1977). These numbers could further be confirmed by chemical protection experiments using phosphate ethylation interference in the binding (Siebenlist and Gilbert 1980).

1.3.2. Hydrogen bonds:

Hydrogen bonds, due to their very specific nature play an important part in protein structure, DNA structure (Watson and Crick 1953) and DNA-protein interactions (von Hippel 1988). A hydrogen bond is formed between an electronegative atom such as oxygen, nitrogen or fluorine acting as an acceptor and a hydrogen attached to another electronegative atom, which is a donor. Polar functional groups in side chains of certain amino acids [OH(Ser, Thr, Tyr), SH(Cys), S-(Met), COOH or COO$^-$(Glu, Asp), CONH$_2$(Gln, Asn), NH$_2$ and NH$_3$ (Arg, Lys) N-H(Trp, His), N:(His)] can form hydrogen bond with polar groups of nucleic acid bases [N:(A, T, G), N-H(T, G), NH$_2$(A, C, G), -C=O(T, C, G)], hydroxyl of ribose sugar and also phosphate groups of the backbone. Thus thirteen of the twenty amino acids and all the four bases of DNA have functional groups which can efficiently form hydrogen bonds.
(Helene and Lancelot 1982). To a lesser extent amide backbone participation is also possible. High specificity of interaction between a protein and a small stretch of DNA that is generally observed can result from the following important considerations. Formation of H-bonds requires the planar configuration of both the donor and acceptor with a maximum variation of 30° out of plane. Thus when such bonds are formed among various acceptor and donors in succession, stringency in recognition occurs. This stringency is favoured for specific protein and DNA combinations but not for non-specific interactions.

Generally, the base pair functional groups with potential hydrogen bonding properties (represented in Fig.2a) project into the major and minor grooves of DNA and form the hydrogen bonds with the protein side chains. Such exposed groups can also be indicated by stick figure representation (Woodbury and von Hippel 1981) as described for four kinds of base pairs (Fig.2b) with the approximate distances between the acceptor and donor groups indicated spacially.

It can be seen from figure 2 that all four kinds of base pairs possible in a DNA can be distinguished in the major groove. In contrast to AT and TA, which can not be differentiated in the minor groove GC and CG base pairs are distinguishable in the minor groove. It can be conceived that various combinations of
Fig.2.a). Functional groups of Watson-Crick basepairs projecting into major groove and minor groove. A=hydrogen bond acceptor, D=Donor, me=methyl group of thymine.(b) Stick figure representation of four kinds of basepairs projecting acceptor and donor groups into two grooves of DNA. MG = major groove; mg = minor groove
base pairs in a DNA molecule can result in a large variety of stereospecific networks of acceptor donor groups for protein recognition and this network can be unique for a specific sequence. It has been suggested (von Hippel 1988) that all the acceptor and donor groups (indicated in Fig. 2) may not be used for the recognition but a few of them may confer specificity. A common pattern of this network is yet to be found from presently existing data but it is most likely that the hydrogen bonding matrix in a specific DNA sequence is unique and complementary to the matrix on the interacting protein. In specific sequences of DNA, the accompanying propeller twist, helical twist and base plane roll might serve to orient the hydrogen bond between acceptors and donors in a characteristic way which would be similar in consensus sequences for related proteins within an organism or for corresponding proteins in different organisms (Dickerson 1989).

1.3.3. Hydrophobic bonds:

Functional groups in amino acids that can form hydrophobic bonds are alkyl side chains of alanine, leucine, isoleucine, valine and aromatic rings of phenylalanine, histidine and tryptophan. The only complementary hydrophobic counterpart on DNA is 5-methyl group of thymine. TA and AT base pairs in a given DNA sequence can project methyl groups (fig 2b) in the major groove
differently in relation to the hydrogen bonds of adjacent base pairs. The arrangement of these methyl groups in relation to the hydrogen bonds of acceptor-donor matrix (described earlier, see page 8) can complement each other in enhancing the specificity for binding a protein.

Entropy changes involving proteins are due to conformational changes, hydrophobic interactions and vibrational effects (Strutvant 1977). If hydrophobic groups contribute to a large number of interactions, binding constant (K) should increase with increase in concentration of salt which displaces water from DNA and enhances the hydrophobic contacts. Since this is not generally observed in the case of protein-nucleic acid interactions (increase in K with salt) hydrophobic contacts may not be present in large numbers. Increase in entropy in protein-nucleic acid interactions is thus largely attributed to the displacement of cations from nucleic acids.

There is evidence for significant contribution from hydrophobic interactions in the stability of at least two operator-repressor systems. In the case of lac operator-represser system, substitution of AT (number 13th) with GC in the operator, reduced the binding affinity of the repressor (Caruthers 1980). This reduced affinity could be restored back using G-C substitution where C has a 5 methyl group. The T-A or
A-U substitution did not restore the affinity. This indicated to a major interaction between 5-methyl group of T and a hydrophobic group of the lac repressor protein in native lac operator and repressor system. A similar major hydrophobic contact between two components was observed in the binding of mutant Mnt protein with mutant operator. In the case of the mutant Mnt protein where His6 was replaced by proline, the affinity of the protein to the native operator was reduced by 1000 times. But the affinity of the mutant protein for the operator was restored when 3rd and 15th GC's were replaced by AT's indicating a possible hydrophobic contact between T-5 CH3 and proline of mutant Mnt (Youderine et.al., 1983).

1.4. THERMODYNAMIC PARAMETERS:

The various thermodynamic parameters useful for the binding phenomena are association constants, free energy, enthalpy and entropy.

1.4.1. Association constants:

The measure of stability of a complex is indicated by values of association constants and such values for different protein-nucleic acid interactions vary between $10^6 - 10^{14} \text{ M}^{-1}$ (table 1). These values can be correlated with the nature of the biological functions that result from such interaction. For example, in certain interactions such as tRNA synthetase-tRNA, once the
<table>
<thead>
<tr>
<th>System</th>
<th>Association constant</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Nonspecific tRNA val (yeast) + Ile tRNA synthetase <em>(E.coli)</em></td>
<td>$7 \times 10^5$</td>
</tr>
<tr>
<td>2. Specific tRNA Ile <em>(E.coli)</em> + Leu tRNA synthetase <em>(E.coli)</em></td>
<td>$2.5 \times 10^7$</td>
</tr>
<tr>
<td>3. 5S RNA <em>(E.coli)</em> + Ribosomal proteins L25 L5</td>
<td>$2.25 \times 10^8$ 2.25 $\times 10^6$</td>
</tr>
<tr>
<td>4. RNA polymerase-promoter</td>
<td>$10^{11}-10^{12}$</td>
</tr>
<tr>
<td>5. Operator-repressors</td>
<td>$10^{12}-10^{14}$</td>
</tr>
</tbody>
</table>
function is over the complex should be dissociated. So in these cases the association constants are relatively low \((10^8 \text{M}^{-1})\). Several techniques have been used for measurement of association constants and their choice depends on the strength of binding. For weak binding systems \((K=10^4-10^6 \text{M}^{-1})\), physical techniques such as UV spectroscopy and circular dichroism are appropriate, for intermediate binding \((K=10^6-10^8 \text{M}^{-1})\) fluorescence spectroscopy is suitable (Helene et al., 1971) and nitrocellulose filter binding (Riggs et al., 1970) together with gel retardation analysis (Fried and Crothers 1981 and 1984) can be used to measure high binding values \((K=10^9-10^{15} \text{M}^{-1})\). Difference boundary sedimentation velocity (Lohman et al., 1979) may be used for nonspecific interactions \((K=10^2-10^5 \text{M}^{-1})\) whereas sucrose density gradient band sedimentation method is effective for a broad range of \(K\) (Draper and von Hippel 1979) \((10^3 \text{ to } 10^{11} \text{M}^{-1})\).

1.4.2. Free energy:

Free energy change \((\Delta G)\) is another measure of stability of the complex. It can be calculated for a protein-nucleic acid interaction using association constant \((K)\) by the thermodynamic equation.

\[
\Delta G = -RT \ln K
\]

\(\Delta G\) values obtained for operator-repressor systems are in the range of \(-14\text{Kcal/mole}\) to \(-16\text{Kcal/mole}\). This value is mostly accounted for by hydrogen and electrostatic bonds. Electrostatic
bonds provide free energy by displacement of condensed counter ions from DNA phosphate groups by positively charged amino acid side chains. This may amount to as much as -8Kcal/mole as observed in lac system (8 electrostatic bonds). Even though bond energy associated with a hydrogen bond is 4Kcal/mole, the free energy available for complex from one hydrogen bond on an average is 0.5Kcal/mole (von Hippel 1988). This is because, in absence of the hydrogen bond from DNA, the protein is involved in hydrogen bonds with solvent water which would compensate for most of the energy. For example in a 17 base pair operator, if on an average each base pair forms one hydrogen bond with protein this would account for ΔG of -8kcal/mole.

Thus hydrogen bonds individually may not confer much stability in terms of free energy. An incorrect or broken hydrogen bond (either by mutation of a base pair or amino acid) may lead to a situation in which two acceptor or donor groups face opposite to each other without solvation (von Hippel 1988). One hydrogen bond uncompensated this way will produce a loss of +5 Kcal/mole. Thus effect of a proper hydrogen bond on free energy change is simply additive whereas loss in free energy due to the absence of hydrogen bond is magnified many fold. This loss of free energy from unfavourable hydrogen bonds may confer specificity indirectly.
1.4.3. Enthalpy (ΔH) and Entropy (ΔS):

ΔH, ΔS values can be obtained by determining temperature dependence of association constant(K) and using the thermodynamic equation,

\[ K = \exp\left(-\frac{\Delta G}{RT}\right) \text{ where } \Delta G = \Delta H - T \Delta S \]

Most of the protein-DNA interactions have large entropy changes and often have positive or zero enthalpies. A study of interaction of 2'- and 3' CMP with RNase A (Flogel et al., 1975) has provided information on relative contributions of ΔH and ΔS for different interactions. Entropy changes predominantly arise from phosphate binding by basic amino acid side chain residues, and the accompanying release of cations from the DNA backbone. Aromatic bases of nucleic acids interacting with proteins provide ΔH value. Protonation of the side groups of amino acids during binding further contributes to the ΔH value.

1.5. METHODS IN STUDY OF PROTEIN-NUCLEIC ACID INTERACTIONS

1.5.1. Genetic and Recombinant DNA methods:

Genetic methods originally formed the basis for identification of control elements and proteins involved in gene regulation. Further, genetic methods helped in identification of important amino acids involved in DNA binding in the case of lac repressor (Miller 1979). Important base pairs in λ and lac operators for protein binding were determined by isolation of operator constitutive mutants (Gilbert et al., 1976 and Jobe
et al., 1974). Genetic methods combined with recombinant DNA technology form one of the most important and powerful tools in DNA-protein interactions. Helix swap experiments between 434 repressor and 434 Cro (Wharton et al., 1984 and Wharton and Ptashne 1986) provided unequivocal importance of helix-turn-helix motif in DNA binding proteins. Domain swap experiment in which one of the DNA binding domains of Gal4 of yeast was replaced with that of Lex A of E.Coli and its interaction with an engineered promoter element studied gave an insight into how a protein factor can activate gene expression in eucaryotes (Brent and Ptashne 1985). Further, use of similar approaches have yielded important information such as requirements for tetramerisation of repressor proteins (Knight and Sauer 1988), looping of DNA (Majumdar and Adhya 1984) and conformation of DNA for gene regulation.

1.5.2. Chemical methods:

Development of synthetic methods of DNA (Khorana 1979) had tremendous applications in various fields of molecular biology including protein-DNA interactions. Easier methods of synthesis, made the synthetic DNA available for non-chemists. Oligonucleotides (DNA fragments) can be assembled from monomer units using either solution phase or solid phase. Smaller amounts of long oligonucleotides (30-60 bases) find use in molecular
biology (Narang 1987) whereas large amounts of short oligonucleotides are required in structural studies of DNA and DNA-protein interactions. Oligonucleotides as linkers and adapters are useful in recombinant DNA technology (Lathe et al., 1983). As hybridization probes, they are invaluable in sequencing, site directed mutagenesis and DNA probe technology (for diagnostic purposes) (Itakura et al., 1984). Structures of A, B, Z DNA at atomic resolution (Saenger 1984) with detailed structural parameters were possible only after development of chemical synthesis of DNA. Chemical synthesis of biologically active gene and gene control regions (Caruthers 1980, Kawashima et al., 1977) and use of them in protein-DNA interactions demonstrated the remarkable applications of chemical methods. In gene regulatory elements where particular bases were replaced with others or modified bases, the code for protein-DNA interactions was investigated (Goeddel et al., 1978, Caruther 1980). Caruther et al., 1980 introduced 5-Bromo 2' deoxyuracil, 5-methyl 2' deoxy cytosine, hypoxanthine bases in lac operator for studying repressor interactions. They found that in this system certain base pairs influence binding to a large extent and the binding phenomena of lac repressor is asymmetric to a partially symmetric operator. DNA-protein co-crystal structure studies (Anderson et al., 1987, Otwinowski 1988) were done using short stretches of synthetic DNA.
Site-specific chemical reagents such as dimethyl sulfate (DMS), hydroxyl radical generated by [Fe(EDTA)²⁻] and ethyl nitroso urea (ENU) in addition to some other reagents are being routinely used to probe protein-DNA complexes (Tullius 1988).

DNaseI is used in identifying the region of the nucleic acid that is occupied by a protein in a complex. This procedure is called foot-printing analysis of protein-DNA complex (Galar and Schmitze 1978 and Johnson et.al., 1979). A labelled DNA molecule complexed with a particular protein is subjected to partial DNaseI digestion. The denaturing urea-polyacrylamide gel analysis of resulting DNA yields a band pattern where bands corresponding to protein bound region of nucleic acid will be absent. One disadvantage of DNaseI method is its size because of which the resolution is limited.

Dimethyl sulfate (DMS) methylates guanosine at N7 in major groove and adenosine at N3 position in minor groove. On methylation of protein-DNA complex with DMS (protection experiment) only G's and A's which are outside the protein bound region and G's and A's within the bound region which are not protected by protein are methylated. Cleavage and analysis of such modified DNA indicates topology of the protein binding to a specific DNA (Ogata and Gilbert 1978 and Simpson 1980). Partial ethylation of phosphate groups (by ethyl nitroso urea) on DNA or
partial methylation of G's and A's and subsequent binding of the protein, separating bound and unbound DNA by electrophoresis followed by analysis of the resultant cleaved DNA indicate which phosphate groups (ethylation interference - Siebenlist and Gilbert 1980) or which G's and A's (methylation interference) might interact with the protein in the complex.

Hydroxyl radical generated by $[\text{Fe(EDTA)}^{2-}]$, in similar experiments can be used to probe different regions of DNA in the complex (Tullius 1987). Hydroxy radical is produced by the reaction of hydrogen peroxide with the EDTA complex of Fe(II) in $[\text{Fe(EDTA)}^{2-}]$. The Fe(II)-EDTA is negatively charged and the actual probe is neutral hydroxy radical which cleaves DNA at each phosphodiester bond. Hydroxy radical footprints can be monitored to find out the solvent accessibility regions of the DNA back bone in the presence of the bound protein (Shanblatt and Rezvin 1987). Photo cross linking (Shetlar 1980) is used when the complexes are less stable in the above modification conditions. This method is used to identify which groups in the protein are associated with the different groups of nucleic acid, by cross linking and further treatment with proteases and nucleases. The method was used for complexes of tRNA with aminoacyl tRNA synthetase, pancreatic RNase A with inhibitor pUp, pCp, rRNAs and
RNA polymerase with the T7 promoters (Helene and Lancelot 1982 and references therein).

1.5.3. Biophysical methods

Biophysical methods that yield valuable information regarding macromolecule structure and interactions at atomic resolution are X-ray crystallography in solid state and NMR in solution state. X-ray crystallography has been used to obtain the structures of repressor proteins and later on the structures of cocrystals of protein-DNA complexes. The other important technique has been NMR spectroscopy which has an advantage of being used in solutions. Till now it has been used to obtain structures of small DNA and protein molecules. CD and fluorescence are other spectroscopic techniques that yield information regarding dynamic structures of the molecules in solution. The data from these can be complemented with that of X-ray and NMR to obtain a complete picture about structure-function relationship in biological molecules.

1.6. STRUCTURE OF NUCLEIC ACIDS INVOLVED IN PROTEIN RECOGNITION

Most of the DNA sequences that are recognised by the proteins are palindromic and two fold symmetric (Sobell 1976). They can be classified as follows:

1. Palindromic, completely symmetric eg; restriction enzyme recognition sites.
(2) Palindromic, partially two-fold symmetric eg; operators, operator DNA sequences of Gal, Arc etc.

(3) Palindromic, completely symmetric passing through central base pair eg; Mnt operator

(4) Asymmetric, eg; promoter sequences.

The size of these symmetric sequences vary to a large extent. They are very small in restriction enzyme sites (4-6 bps), medium size in operator DNA (14-24 bps) and about 30 bps in site specific ter endonucleases (Weigel et al. 1973 and Murray and Murray 1973).

Symmetry (partial and complete) seem to have been used for the recognition of the specific sequences in a large non-specific DNA sequence (17 bps in \(4 \times 10^6\) bps) (Sobell 1976). But surprisingly the interactions of the protein to DNA in most of these cases seem to be asymmetric. Symmetric groups on either half of the operator are not identified by the protein with equal ease (Caruthers 1980). Further support for asymmetric recognition is provided by the operator constitutive mutants falling on one side of the two symmetric halves (Jobe et al., 1974). This was also confirmed by the protection and interference experiments in case of lac operator (Ogata and Gilbert 1978 and 1979) and \(\lambda\) operators (Humayun et al., 1977 and Johnson et al., 1978).
While symmetry in nucleic acid sequences may be responsible for recognition and stability, asymmetry observed in protein contacts cannot be explained. Asymmetry in promoter-RNA polymerase interaction can be explained in terms of polarity of transcription that is to occur.

1.7. PROTEIN STRUCTURAL MOTIFS FOR SPECIFIC DNA BINDING: OBSERVATIONS

In the systems studied to date, most important and widely present motif found in protein-DNA interactions is helix-turn-helix (Takeda et al. 1983). It has got powerful evidence from X-ray crystal structure (Anderson et al., 1981, McKay and Steitz 1981, Pabo and Lewis 1982) supported by equally powerful recombinant DNA helix swap experiments (Wharton et al., 1984). It is found in three crystal structures determined for repressors (Anderson et al. 1981, McKay and Steitz 1981, Pabo and Lewis 1982) and cocrystal structures solved (Jordon et al., Aggarwal et al., and Wolberger et al., all in press) for three other operator-repressor complexes. Sequence alignment studies (Pabo and Sauer 1984 and Sauer et al., 1982) indicated that a similar motif exists in large number of proteins that bind to DNA. Other protein structures for DNA binding that have come to light and have been implicated in number of systems are "zinc finger motif" (Klug and Rhodes 1987) and "leucine zipper" (Landschulz et al., 1988).
"Helix-turn-helix" motif is best explained by considering Cro and CI repressor proteins (Ptashne 1986) of phage λ because of two reasons; (1) their crystal structure is known (2) they bind to two kinds of three operators each OL1, OL2, OL3 and OR1, OR2, OR3. The latter reason interestingly elucidates the novelty present in the two proteins as they bind to the three operators in reverse order of affinity (Johnson et al., 1978).

Cro protein as found in X-ray structure (Anderson et al., 1981) has three α-helices (1, 2 and 3) and three β-sheets. CI repressor's amino terminal domain has five α helical structures namely 1, 2, 3, 4, and 5. Both proteins exist as dimers indicating the possibility of symmetry in recognition. Even though the overall structure is different, the two proteins show remarkable similarity in projecting one helix each (helix 3 in both cases) from two monomer units in similar geometry. The two helices that are projected from dimeric protein bodies are 34Å separated and tilted 30°, the dimensions of which exactly match with the two adjacent major grooves of a B DNA (Ohlendorf et al., 1982) (Fig. 3). These helices are called "recognition helices" which bind into the two symmetric halves of the palindromic operator, the side chains of the amino acids in the helix making contacts with base extensions in the form of H bonds. The inside part of the recognition helix of protein is faced with another helix, helix 2, which in turn helps to orient the recognition .
Fig. 3. Binding of "helix turn helix" motif into major groove of operator DNA.

Two units from two monomers bind the bases of two consecutive major grooves of symmetric operator. Only carbons of the aminoacids in the motif represented.
helix. This helix lies across the major groove. These two helices together with a turn in between, give the name "helix-turn-helix".

How the two proteins cro and cI repressor bind to three operators in reverse order depends on the amino acids distributed in their recognition helices (Ptashne 1984). In both of them gln28, Ser29 are conserved and bind to base pairs 2nd AT and 4th GC of operator by hydrogen bonds. These 2nd AT and 4th GC base pairs are invariant in all operators and are probably used in recognition. Other surface projected amino acids (33 ala in cI repressor and asp 32, lys 33 in cro) differ in both proteins and probably interact with variant base pairs in different operators giving different strengths.

The crystal structures were used to predict models (Ohlendorf et.al., 1982, Lewis et.al., 1982 and Weber and Steitz 1984) using energy minimization procedures along with other spectroscopic and genetic evidences. The general features as predicted from models are confirmed by the co-crystal structural studies of 434 repressor-operator (Aggarwal et.al., - in press) 434 cro repressor-operator(Wolberger et.al., in -press), and cI repressor-operator (Jordon and Pabo - in press). An activator DNA binding protein, CAP when crystallised shows similar bihelical motif with a variation in the helix orientation to DNA (Weber and
Steitz 1984). The trp repressor-operator co-crystal structure shows deviation from general operator-repressor interactions and in this case indirect contacts between aminoacids and phosphate backbone have been observed which are mediated by water molecule in contrast to direct H-bonds. Salient features found in "helix-turn-helix" motif can be summarised as follows (Mathews 1988).
(1) Operator-repressor complexes are 2-fold symmetric, the 2-fold axis of protein coincides with the 2-fold axis of DNA.
(2) Alignment of recognition helix in major groove is similar (with exception of CAP).
(3) Readout and specificity is achieved by interactions of aminoacid side chains to base extensions in grooves of DNA (with exception of trp repressor).
(4) Affinity is due to the protein and DNA phosphate back bone contacts.

Another characteristic motif was observed in transcription factor TFIIIA (Klug and Rhodes 1987) which is involved in regulation of the synthesis of 5S rRNA of Xenopus oocyte. In this case Cys-Cys and His-His residues placed at regular intervals coordinate with a zinc atom and form stable structures called "zinc fingers". The tertiary folds of these zinc fingers are presumed to be involved in DNA binding. This motif awaits X-ray crystal structure analysis but several zinc finger forming sequences
(Klug and Rhodes 1987) are reported in drosophila, mammalian systems (spl transcription factor), rat (glucocorticoid receptor) and yeast (Gal4 protein).

1.8. PRESENT SYSTEM AND OBJECTIVES

1.8.1. P22 Mnt operator-repressor system:

Of all the lamboid temperate phages studied, λ and P22 are the most interesting. The λ phage was studied as a model system for gene expression (Ptashne 1986). P22 phage shows an additional complex mode of control of gene expression (Susskind and Botstein 1978 and Susskind and Youderian 1983). The prophage (inactive, lysogenic state in host bacterium) of λ is controlled by a single repressor (ci) which exerts negative control at two neighbouring operator sites (OL1, OL2, OL3 and OR1, OR2 and OR3). ci repressor along with cro repressor in the immunity region exerts immunity against super infecting phages and they together decide the fate of the phage for lytic or lysogenic state. In the case of phage P22 of Salmonella, the control of prophage and immunity are conferred by two repressors. c2 repressors (analogous to ci of λ) and Mnt repressor. They lie in two distinctly separated regions ImmC and ImmI respectively (Fig.4). While C2 repressor acting on neighbouring two sets of operators (OR1,OR2,OR3;OL1,OL2,OL3) is directly responsible for lysogeny, Mnt repressor by controlling gene expression of another repressor (antirepressor) is indirectly responsible for lysogeny (Fig.4). This is because
Fig. 4: P22 Repression model; Importance of Mnt repressor in Lysogeny
(From Susskind and Botstein 1978)
antirepressor, if expressed inactivates c2 repressor, whereby genes required for lytic phase are expressed. Mnt repressor prevents antirepressor expression and hence c2 repressor function in lysogeny is uninhibited. There is also another repressor called Arc repressor which turns off the antirepressor in late lytic phage. Mnt repressor, antirepressor and Arc repressor together in immI region exert a secondary control of lysogenic state maintenance of phage P22. Genetic map of P22 and the secondary control model are depicted in figure 5.

Mnt repressor is an 82 aminoacid long protein which binds to 21 base pair operator Omnt and prevents the transcription of antirepressor gene. This 21 base pair operator comprises a 17 bp perfectly 2-fold symmetric sequence with which actual base contacts of protein are observed (Vershon et.al., 1987). Mnt shows little aminoacid homology with known helix-turn-helix proteins (Sauer et.al., 1983) and was predicted to possess different structure (Knight et.al., 1989) for DNA recognition. Perfectly 2-fold symmetric sequences in operator DNA are very rare and from this point of view also the system is interesting.

Synthetic oligonucleotides find wide applications in molecular biology and in the investigation of structure-function relationship of biological macromolecules, DNA and protein.
Fig. 5: P22 Phage Genetic Map
Chemical synthesis methods made available large amounts of DNA molecules required for the biophysical studies.

In the present study, synthetic oligonucleotides are used in the sequence dependent structural studies of DNA and DNA-protein interactions. Chapter II describes the chemical synthesis methods that were followed to obtain oligonucleotides. Chapter III describes nuclear magnetic resonance and circular dichroism spectroscopy as applied to oligonucleotides $d(\text{CACGTG})_2$, $d(\text{CACCGTG}) \cdot d(\text{CACGGTG})$ for their solution structure determination. Interaction studies of synthetic native and modified operator DNA fragments (37-mers) to Mnt repressor as monitored by gel retardation, dimethylsulfate methylation protection and interference experiments and Circular dichroism spectroscopy are depicted in IV Chapter.