Introduction

Biochemical instructions for life on Earth are stored in each organism's deoxyribonucleic acid (DNA). The massive international effort to sequence the human genome underscores the importance of the information contained in the roughly 30,000 genes, each of which encodes different protein or ribonucleic acid (RNA) products (Venter et al., 2001). Proper cellular function depends on specific protein-DNA interactions necessary for regulating gene expression, and it is the misregulation of gene expression that is responsible for many disease states, including certain cancers. Synthetic molecules that bind to predetermined DNA sequences and regulate gene expression would therefore offer great benefit to human medicine (Gottesfeld et al., 1997; Dickinson et al., 1998; Mapp et al., 2000).

DNA Structure. Double-helical DNA is composed of two polydeoxy-ribonucleotide strands aligned in an antiparallel fashion and associated through specific hydrogen bonds between the heterocyclic bases, adenine (A), thymine (T), guanine (G), and cytosine (C), such that A pairs with T and G with C (Figure 1) (Saenger, 1984; Neidle, 1999). Common B-form DNA is characterized by a wide and shallow major groove and a narrow and deep minor

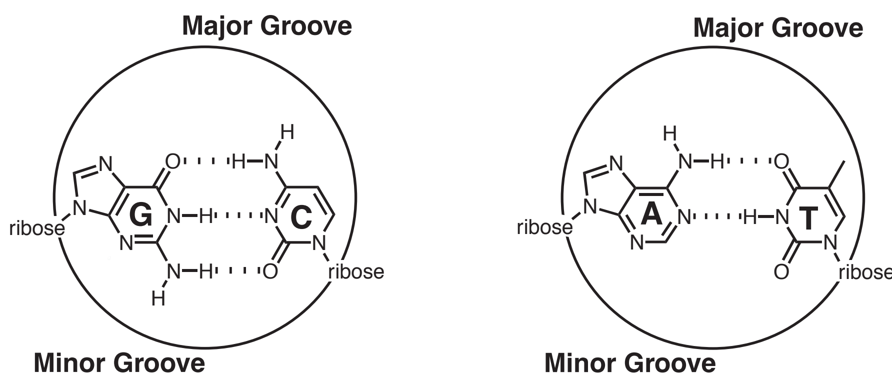

Figure 1 Chemical structure of G•C (left) and A•T (right) base pairs. The major and minor grooves are indicated as the regions spanning between the ribose backbones.

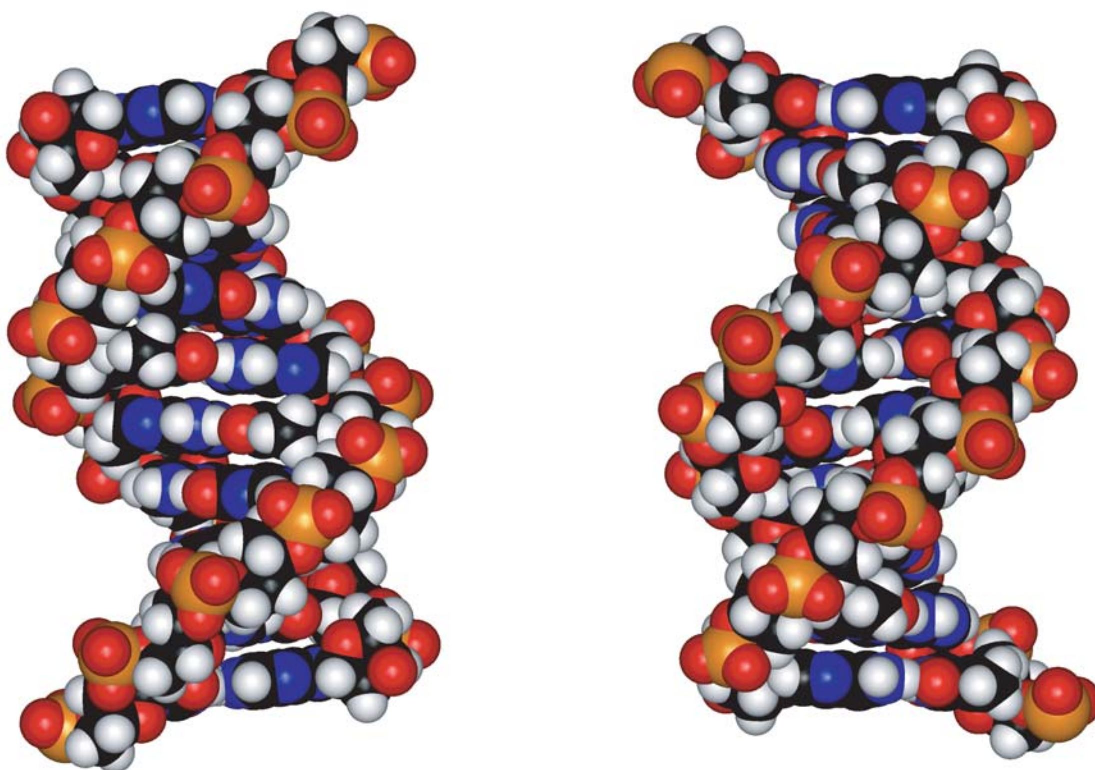

Figure 2 Space-filling model of a ten base-pair B-form DNA duplex viewed into (left) the major groove and (right) the minor groove. Carbon atoms are black, oxygens are red, nitrogens are blue, hydrogens are white, and phosphorus atoms are orange.

groove (Figure 2). DNA sequences can be distinguished by the pattern of functional groups, e.g., hydrogen bond donors and acceptors, displayed on the edges of the base pairs. However, the sequence-dependent variation in conformation and counterion organization that distinguishes local DNA microstructure (Saenger, 1984) makes it difficult to design molecules with optimal shape and electrostatic complementarity to a particular DNA sequence.

Native DNA Recognition. Nature has selected for numerous DNA-binding proteins capable of specific sequence recognition based on ensembles of electrostatic and shape-selective interactions. Conversely, the enormous diversity in protein structure makes *de novo* protein design for the recognition of specific DNA sequences a challenging prospect. Although numerous structural motifs have been identified for protein-DNA recognition (Figure 3) (Love et al., 1995; Kim et al., 1993; Pavletich and Pabo, 1991; Ellenberger et al., 1992), a general recognition code correlating target DNA sequence with amino acid sequence composition has yet to be identified. On the other hand, small molecules are typically more limited in conformational flexibility than proteins, offering the chemist a more controllable platform for fine-tuning the shape complementarity necessary for DNA sequence discrimination. Nature has also provided a number of structurally diverse small molecules that recognize DNA by binding in the minor groove, intercalating between base pairs, or both (Figure 4) (Gao et al., 1992; Kamitori and Takusagawa, 1992; Paloma et al., 1994; Coll et al., 1987). The N-methylpyrrole carboxamide (Py) backbone of the antitumor antibiotics netropsin and distamycin A presents an attractive context for the design of polyamide ligands with altered DNA sequence selectivity.

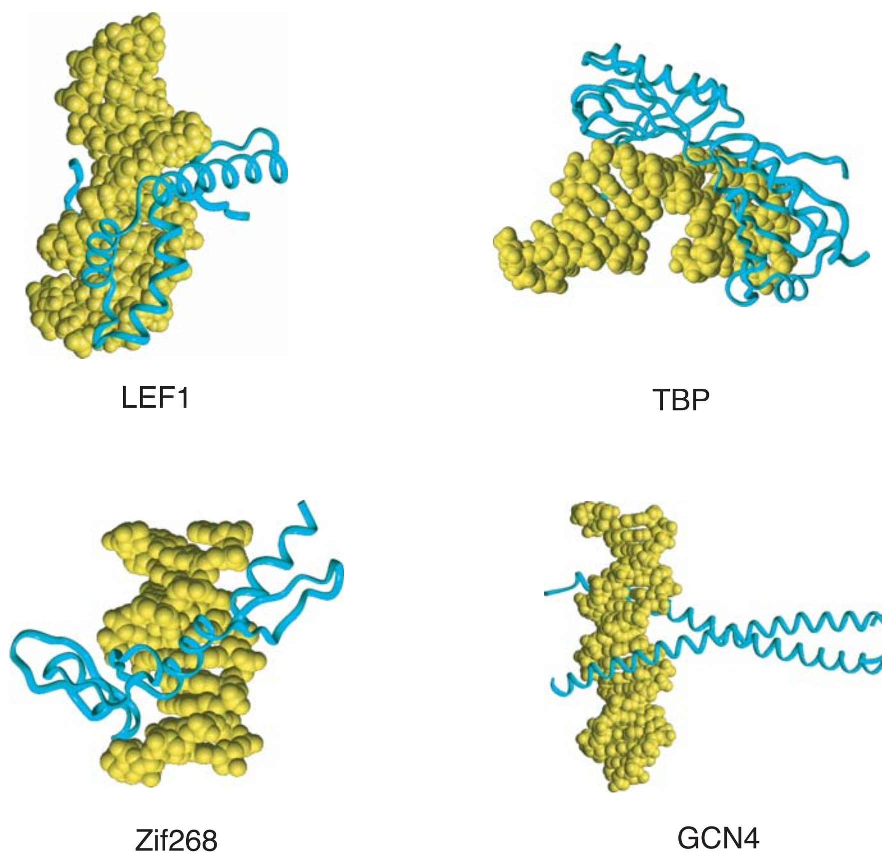

Figure 3 X-ray crystal structures of four protein-DNA complexes showing the diversity of structural motifs for protein-DNA recognition.

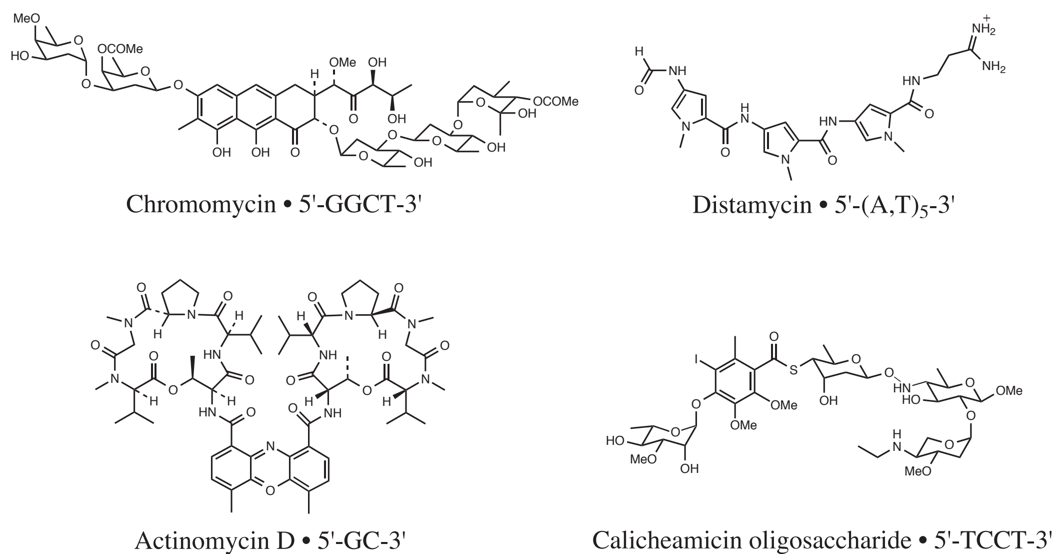

Figure 4 Chemical structures of several naturally occurring, DNA-binding small molecules with their optimal target DNA sequences indicated.

Minor Groove Recognition by Designed Ligands. The minor groove of DNA can be characterized by a somewhat uniform display of chemical functionality. A•T base pairs present relatively symmetric hydrogen bond acceptors, N3 of A and O2 of T. G•C base pairs present similar groups, N3 of G and O2 of C, in addition to the hydrogen bond-donating 2-amino group of guanine (G-NH_2) (Figure 5). The minor groove of A,T-tracts is both narrow due to propeller twisting of the base pairs and relatively deep due to lack of the protruding G-NH_2 (Fratini et al., 1982).

Understanding the sequence-dependent microstructure of DNA is of key importance for the study of ligand•DNA interactions. Analysis of numerous B-DNA single crystal x-ray structures reveals that certain base-base steps are more deformable than others (Dickerson, 2001). In particular, purine-purine steps such as A-A and G-A are inclined to be more rigid structures with a narrow minor groove and large negative propeller twist. It has been suggested that optimal base stacking is the primary factor governing this feature (Hunter, 1993).

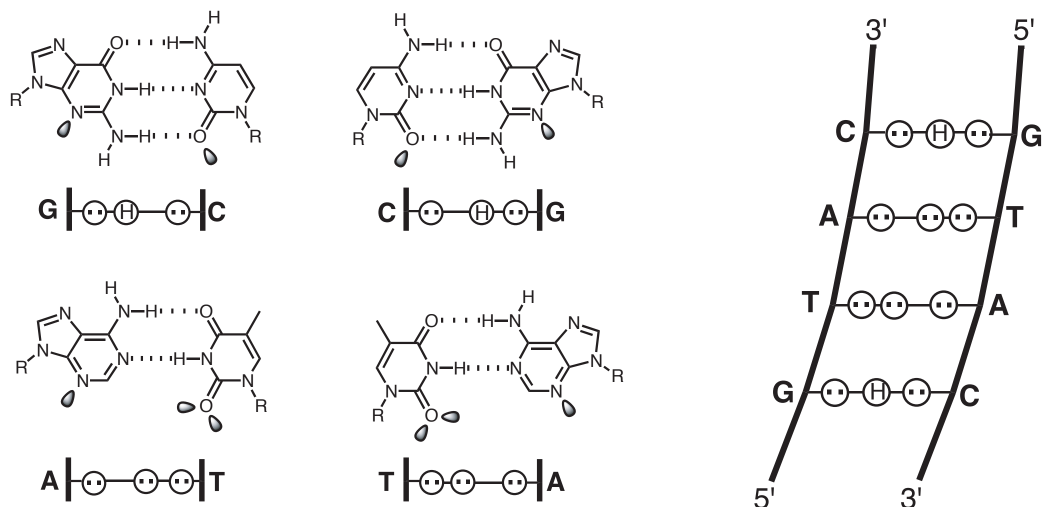

Figure 5 Minor groove hydrogen bonding patterns of the four Watson-Crick base pairs. Circles with dots represent lone pairs on purine N3 or pyrimidine O2, and circles with an H represent the 2-amino group of guanine (G-NH_2).

Distamycin and netropsin bind in the minor groove of A,T-tract DNA (Zimmer and Wahnert, 1986). Upon binding, these ligands displace the spine of hydration in a multidentate fashion, forming hydrogen bonds to proximal purine N3 and pyrimidine O2 atoms (Kopka et al., 1985; Coll et al., 1987), which provides a large entropic driving force for binding (Chalikian et al., 1994). Moreover, the bound ligands fit snugly, making extensive van der Waals contacts to the walls of the minor groove. The x-ray structure of netropsin bound as a 1:1 complex to DNA (Figure 6, left) inspired the lexitropsin model, where it was predicted that replacing one or both Py residues in netropsin with N-methylimidazole carboxamide (Im) would confer G•C recognition by simultaneously alleviating a steric interaction with the C3-H of Py and forming a hydrogen bond from G-NH_2 to Im-N3 (Kopka et al., 1985). Subsequent footprinting experiments revealed that Im-Py polyamides tolerate G•C base pairs but show little sequence-specificity (Lown et al., 1986). Remarkably, the structural basis for the lexitropsin model, as envisioned in a 1:1 complex, was never verified structurally.

Wemmer and coworkers made the unanticipated observation that distamycin can bind A•T tracts of DNA in an antiparallel 2:1 fashion (Figure 6, right), even at low ligand:DNA stoichiometries (Pelton and Wemmer, 1989). Therefore, Py/Py pairs, as well as Py, prefer A,T over G,C. It was subsequently discovered by Dervan and coworkers at Caltech that the unsymmetrical ring pair Im/Py can distinguish G•C from C•G and both from A•T and T•A base pairs (Dervan, 2001). Further invention of the new ring pair Hp/Py (Hp = 3-hydroxypyrrole, Hp/Py specifies T•A) completed a recognition code to target all four Watson-Crick base pairs in the minor groove of DNA (Figure 7).

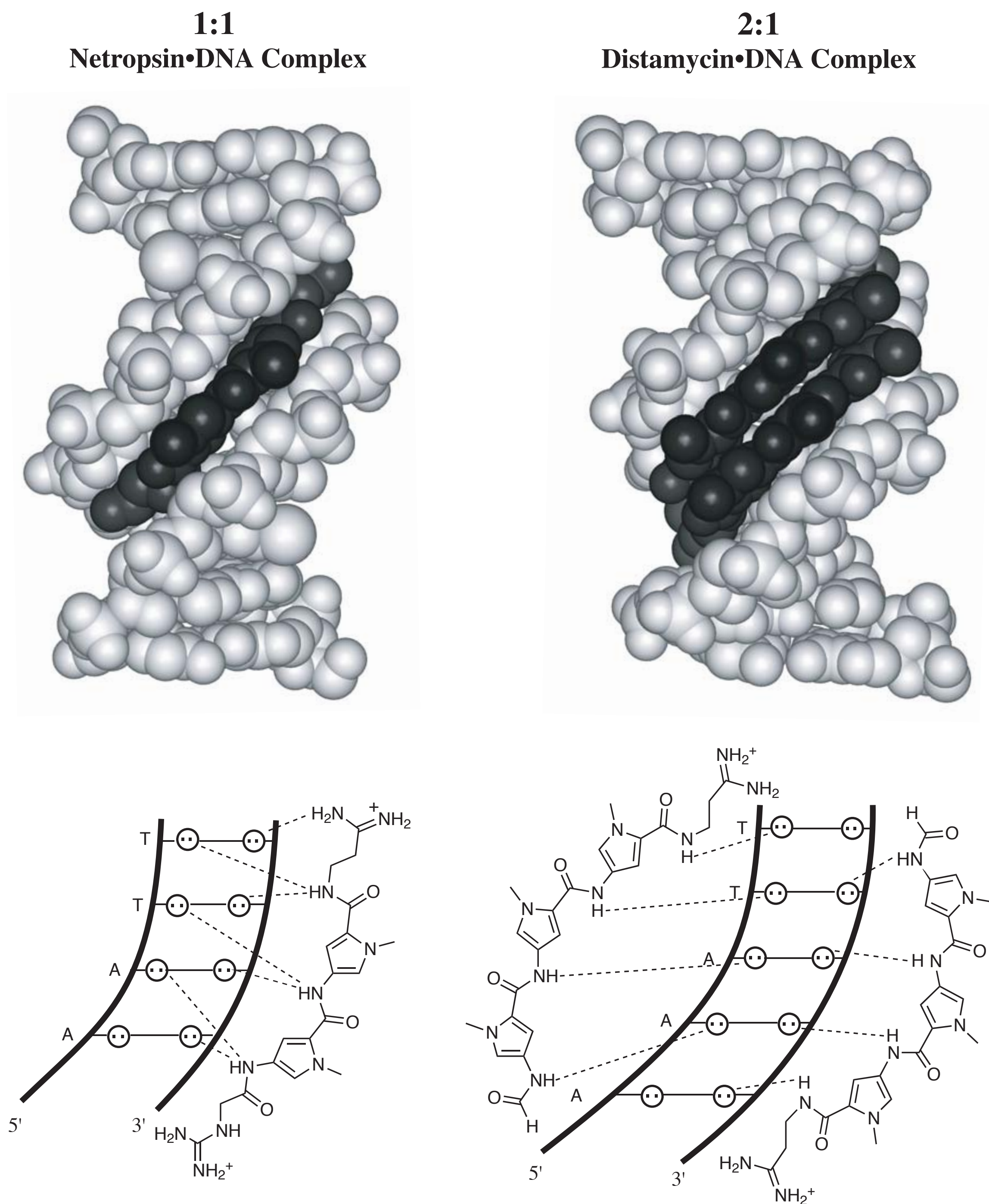

Figure 6 (left) X-ray crystal structure of the 1:1 netropsin•DNA complex; (right) NMR structure of the 2:1 distamycin•DNA complex. Binding models of each complex are shown below with hydrogen bonds indicated as dashed lines.

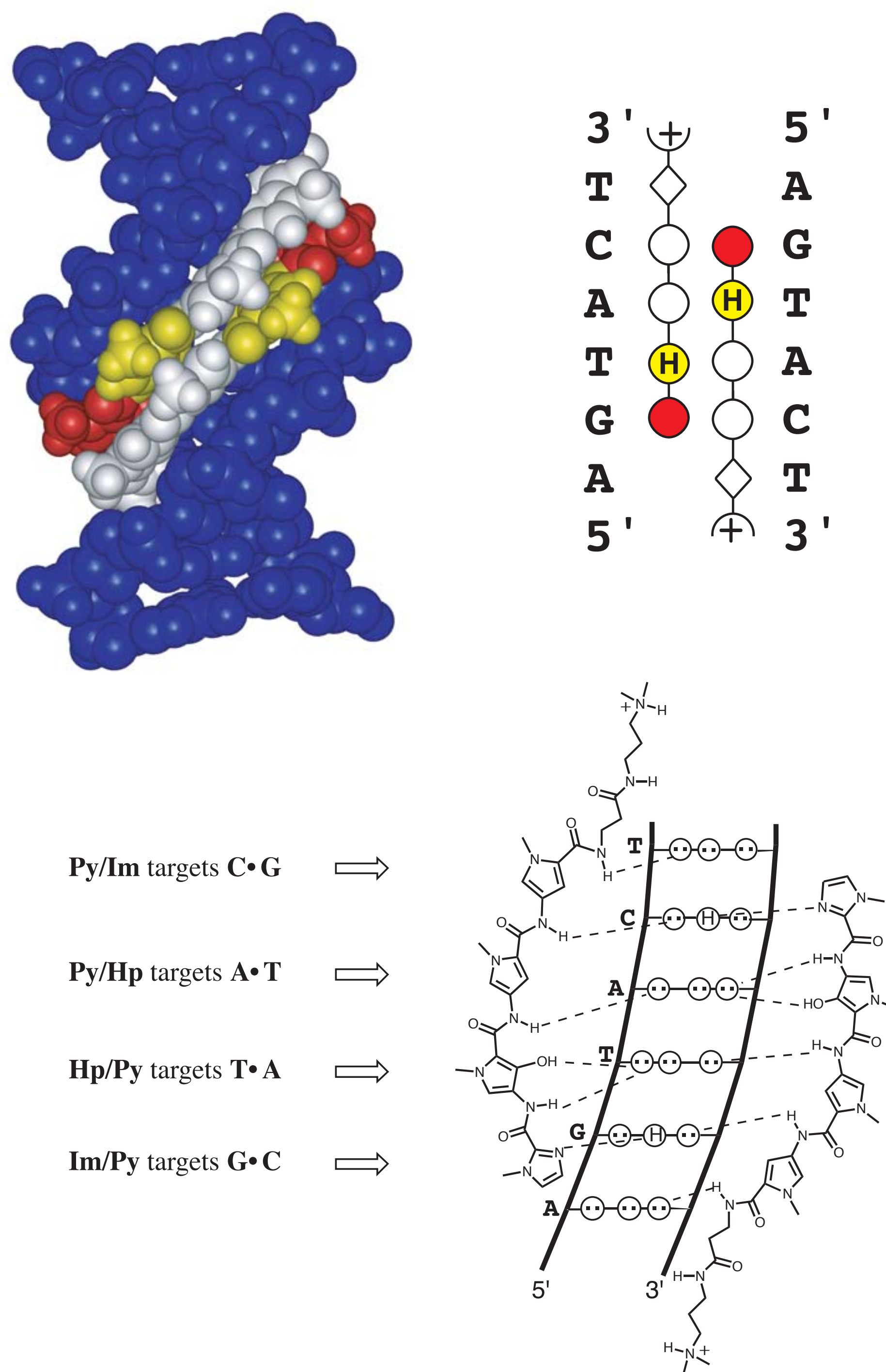

Figure 7 Pairing Rules. (top left) x-ray crystal structure of ImHpPyPy-β-Dp (Dp = dimethylaminopropylamide) bound in a 2:1 complex with its target DNA site, 5'-AGTACT-3' (Kielkopf et al., 1998b). A schematic is shown to the right, which represents Im residues as red circles, Hp as yellow circles containing an H, Py as white circles, and β as white diamonds. The hydrogen bonding schematic at the bottom shows the pairing rules for targeting all four Watson-Crick base pairs of DNA.

Antiparallel polyamide subunits stacked in a 2:1 complex can be covalently linked, head-to-tail, to form hairpin polyamides with substantially increased binding affinity and sequence specificity (Mrksich et al., 1994). Polyamides composed of multiple contiguous Py and Im residues are overcurved with respect to the DNA helix (Kelly et al., 1996). However, Py residues can be substituted with the flexible, A,T-specific beta-alanine residue (β), in order to relax ligand curvature and restore ligand-DNA alignment, thereby restoring binding affinity (Trauger et al., 1996a; Turner et al., 1998; de Clairac et al., 1999) (Figure 8). An important example of affinity restoration via β substitution is the hairpin polyamide ImPyImPy-γ-ImPyImPy-β-Dp (Dp = dimethylaminopropylamide), which targets the core sequence 5′-GCGC-3′, according to the pairing rules, but with low affinity (Swalley et al., 1997). By judicious replacement of Py with β-alanine, i.e., Im-Py-Im → Im-β-Im, a second generation hairpin polyamide, Im-β-ImPy-γ-Im-β-ImPy-β-Dp, restores the dissociation constant (K_D) to subnanomolar (Figure 8) (Turner et al., 1998).

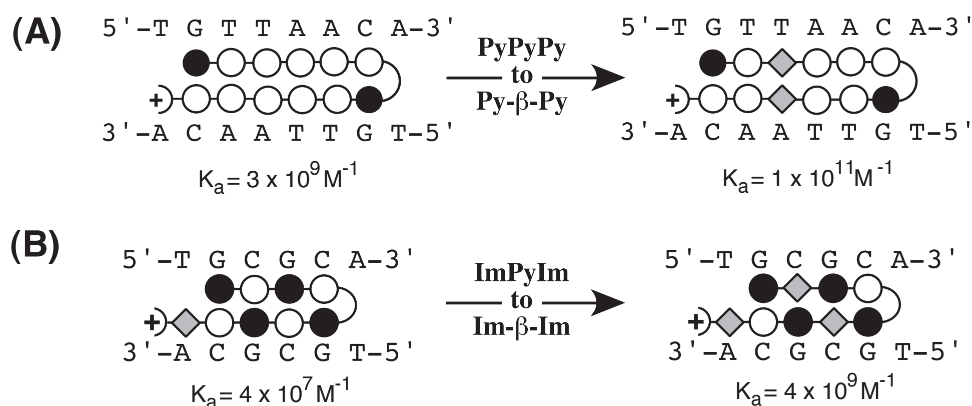

Figure 8 Restoration of binding affinity upon replacement of Py with β. Im and Py residues are illustrated as shaded and nonshaded circles, respectively; gray diamonds indicate β; and semicircles represent the γ-turn residue.

Limitations. Sequence specificity in a genomic context will require high fidelity targeting of long DNA sequences (16-20 base pairs). Employing the current state-of-the-art hairpin polyamides may be problematic because hairpin molecular weight increases dramatically with binding site size. This increase can be deleterious to cellular and nuclear uptake, which are essential for therapeutic applications. Therefore, the search for alternative binding modes remains an important direction. The "pairing rules" have proven useful for recognition of hundreds of predetermined DNA sequences by designed polyamides. However, sequence-dependent structural variation is thought to reduce binding affinity at numerous DNA sequences, most notably those containing 5'-GA-3' and 5'-GNG-3' steps (Swalley et al., 1997; Herman, 2001). The development of higher fidelity recognition rules, which will require a better understanding of sequence-dependent structural effects, is an important goal**.**

The 1:1 Motif. Although the structure and energetics of the 2:1 motif have been explored extensively (Dervan, 2001; Kielkopf et al., 2000), relatively little was known (prior to the thesis presented here) about the 1:1 motif beyond the initial work on netropsin and distamycin. In a recent breakthrough, Laemmli and coworkers reported the use of Py-Im polyamides to effect phenotypic changes in *Drosophila melanogaster* (Janssen et al., 2000a; 2000b). In their report, 5'-GAGAA-3' repeat sequences were targeted at very high affinity using β-rich polyamides that bind in a 1:1 fashion and in a sequence-dependent, single orientation (Figure 9). This new mode of 1:1 recognition offers the opportunity to target longer DNA sequences with considerably smaller molecules, due to the requirement for only a single subunit per binding site, which may further confer the benefit of increased cellular uptake. Additionally, the 5'-GAGAA-3' sequence

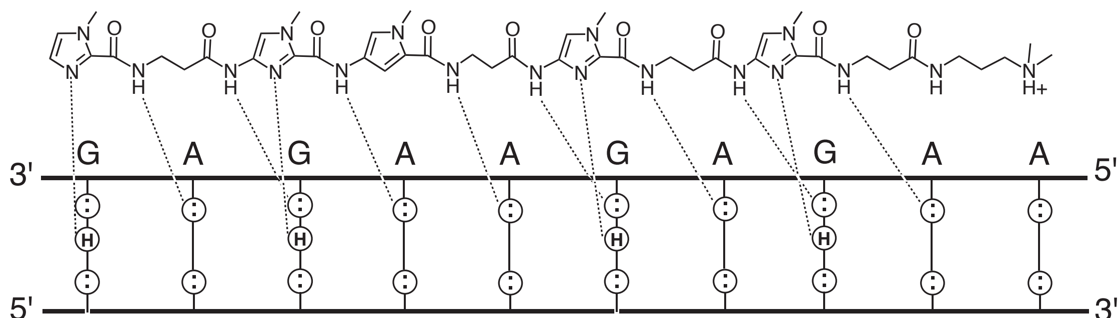

Figure 9 Laemmli model for 1:1 polyamide-DNA recognition using a β-linked Py-Im polyamide. Putative hydrogen bonds are indicated by dashed lines.

is almost entirely composed of 5'-GA-3' units, which are problematic for hairpin polyamides. Therefore, this class of molecules holds the potential to address numerous problems in DNA recognition and, hence, merits further exploration.

Description of this Work. The thesis presented here investigates Laemmli's 1:1 motif in order to further understand and exploit this novel mode of DNA recognition. First, the importance of β-alanine for 1:1 recognition was examined. In this study, Py residues within Im**Py**Py and Im**Py**Im contexts were replaced with β, and it was found using quantitative DNase I footprinting that the Im-β-Im subunit is important for high-affinity binding. This study also demonstrates the capability of a single ligand to bind very different DNA sequences depending on the stoichiometry of complexation, i.e., 1:1 or 2:1. Since the 1:1 and 2:1 motifs have inherently different rules for recognition, this finding poses the design problem of how to control the binding mode and therefore the DNA sequence target (Dervan and Urbach, 2001). Next, the possibility of developing a 1:1 recognition code was explored. By selectively mutating

polyamide residues and DNA base pairs, and comparing the association constants for the resulting complexes, it was found that Im residues tolerate all four Watson-Crick base pairs; Py and β residues are specific for A•T and T•A base pairs; and Hp specifies a single base pair, A•T, in the sequence context 5'-AAAGAGAAGAG-3' (Urbach and Dervan, 2001). Efforts to improve upon this recognition code using novel heterocyclic amino acids, such as furan, thiophene, thiazole, and hydroxythiophene, are presented (Marques et al., in preparation). Additionally, the sequence-dependence of ligand orientation and the effect of ligand size on binding affinity were explored. The intellectual core of this thesis is supported by the determination and analysis of the solution structure of a 1:1 polyamide-DNA complex using two-dimensional NMR methods. The high-resolution structure reveals B-form DNA with a narrow minor groove and large negative propeller twist, which are shown to be stabilized by bifurcated hydrogen bonds between polyamide NH groups and purine N3 and pyrimidine O2 atoms at each base step. The first direct evidence is provided for hydrogen bond formation between Im-N3 and G-NH_2 in the 1:1 motif, thus confirming the original lexitropsin model (Urbach et al., 2002). Finally, the ambiguity of sequence targeting depending on stoichiometry was addressed. It was discovered that hairpin and 1:1 binding modes, which are dependent on ligand conformation, can be effectively controlled by changing the linker between polyamide subunits. In the system examined, a β linkage specifies the 1:1 mode by >150-fold. Replacement of β with α-(*R*)-acetamido-γ-aminobutyric acid enforces an 82,000-fold reversal of specificity in preference for the hairpin motif (Urbach et al., *in preparation*).