AN ABSTRACT OF THE THESIS OF

Beth E. Basham for the degree of Doctor of Philosophy in Biochemistry & Biophysics presented on March 11, 1998. Title: The Analysis and Prediction of DNA Structure.

Redacted for Privacy

Abstract approved: _____

P. Shing Ho

As genome sequencing projects begin to come to completion, the challenge becomes one of determining how to understand the information contained within the DNA. DNA is a polymorphic macromolecule; the A- B- and Z-DNA conformations have been observed by a variety of physical techniques. The magnitude of the energetic differences between these conformations suggests that these conformations may be important biologically and thus relevant in the analysis of genomes. A computer program, NASTE, was developed to evaluate the helical parameters of the set of Z-DNA crystal structures in order to determine the true conformation of Z-DNA and to understand the effects of various factors on the observed structure and stability. A thermodynamic method, elucidated in part with a genetic algorithm, was developed to predict the sequence-dependent propensity of DNA sequences for A- versus B-DNA in both the crystal and in natural DNA. Predictions from this method were tested by studying the conformation of short oligonucleotides using circular dichroism spectroscopy. Finally, the thermodynamic method was applied in an algorithm, AHUNT, to identify regions in genomic DNA with a high

propensity to form A-DNA. Significant amounts of A-DNA were identified in eukaryotic and archeabacterial genes. *E. coli* genes have less A-DNA than would be predicted from their (G+C) content. These results are discussed with respect to the intracellular environment of the genomes.

THE ANALYSIS AND PREDICTION OF DNA STRUCTURE

by

Beth E. Basham

A THESIS

submitted to

Oregon State University

in partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Completed March, 11 1998
Commencement June 1998

Doctor of Philosophy thesis of Beth E. Basham presented on March 11, 1998

APPROVED:

Redacted for Privacy

Major Professor, representing Biochemistry & Biophysics

Redacted for Privacy

Chair of Department of Biochemistry & Biophysics

Redacted for Privacy

Dean of Graduate School

I understand that my thesis will become part of the permanent collection of Oregon State University libraries. My signature below authorizes release of my thesis to any reader upon request.

Redacted for Privacy

Beth E. Basham, Author

ACKNOWLEDGMENT

The completion of this work would not have been possible without the help and support of many special people. I value the guidance, support and acute scientific insight of Dr. P. Shing Ho, who provided numerous wonderful opportunities and taught me how to ask the right questions and seek out the answers.

I must also acknowledge the past and present members of the Ho lab: Dr. Gary P. Schroth for the all advice and the critical evaluation of this work and my ideas, and fellow students Dr. Todd Kagawa, Dr. Blaine Mooers, Brandt Eichman and Jeff Vargason for their help and friendship.

The circular dichroism would not have been possible without the expert technical advice of Dr. W. Curtis Johnson and Jeannine Lawrence.

The contribution of the faculty at Oregon State University was also vital to this work, particularly the people involved in its evolution and evaluation: Dr. Michael Schimerlik, Dr. Philip McFadden, Dr. Victor Hsu and Dr. Mark Christensen. I would especially like to thank Dr. Victor Hsu for his helpful and patient discussions about DNA solution structure and Dr. Kensal van Holde for his interest in this work.

I also value the personal and professional friendships of Monika Ivancic, Debbie Mustacich, Indira Rajagopal, Kevin Ahern, Cyndi Thompson, and Laura Meek.

None of this would have been possible without the continuous support of my parents, Pat and John Etchells, my brother, Sean, and my grandparents, Ray and Edna Etzold. Finally, I am very thankful to Eric, my husband and my muse, for his unfaltering patience and inspiration.

CONTRIBUTION OF AUTHORS

Brandt Eichman assisted in the analysis of the Z-DNA crystal structures, in the compilation of the tables and in the preparation of the review (Chapter 2). He also analyzed the set of Z-DNA sequences for length effects. Dr. P. Shing Ho described the solvent structure of d(CGCGCG). Dr. Gary P. Schroth assisted in the design of the circular dichroism experiments and designed and collected the data for some of the oligonucleotide sequences presented here (Chapter 3, 4). Dr. P. Shing Ho guided all aspects of the projects described here.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

LIST OF FIGURES

LIST OF FIGURES (continued)

LIST OF FIGURES (continued)

LIST OF TABLES

LIST OF TABLES (continued)

DEDICATION

To Eric, for your support, understanding and encouragement.

Chapter 1

1. INTRODUCTION

Predicting the three dimensional structure and function of a macromolecule from its primary sequence of monomer units is one of the major goals in modern structural biology. For polymers of amino acids, this is the protein folding problem (Anfinsen, 1973). The complementary challenge for sequences of deoxyribonucleotides (DNA) is understanding how the sequence defines a three-dimensional structure and why that structure is important. This problem is intimately related to the emerging field of functional genomics which involves the systematic and global exploration of gene sequences to understand gene and genome function (Heiter and Boguski, 1997).

The human genome project and other sequencing projects have generated an enormous quantity of DNA sequence information (Table 1.1). It has been estimated that individual sequencing facilities can currently sequence 30 million bases per year (Rowen et al., 1997). While it is well known that DNA is the cell's instruction set for proteins, it is estimated that only 3% of the human genome actually codes for proteins (Voet and Voet, 1990). The challenge in understanding the rest of genome lies in

Table 1.1

Some genomes that have been or are currently being sequenced

Species	Genome size (million base pairs)
Aquifex aeolicus	1.5
Archaeoglobus fulgidus	2.2
Bacillus subtilis	4.2
Escherichia coli	4.6
Haemophilus influenzae	1.8
Helicobacter pylori	1.7
Methanobacterium thermautotrophicum	1.8
Methanococus jannaschii	1.8
Mycoplasma genitalium	0.6
Mycoplsama pneumoniae	0.8
Neisseria meningitidis	2.2
Pyrobaculum aerophilum	1.9
Pyrococcus horikoshii	2.0
Saccharomyces cerevisiase	16.0
Synechocystis sp. PCC6803	3.5
Treponema pallidum	1.1
Ureasplasma urealyticum	0.8
Homo sapiens	3000 (60 complete)

Compiled from (Gaasterland et al., 1997; Rowen et al., 1997)

developing new ways to understand and evaluate DNA sequence information (Hieter and Boguski, 1997) and in identifying the biological activities of noncoding DNA sequences. Presumably, the function of some noncoding regions is the regulation of cellular processes including DNA replication and transcription. This function could arise from the inherent dynamic nature of DNA structure or by recognition of specific DNA structures by proteins. In any case, predicting the three-dimensional structure of DNA from its sequence is an essential step in understanding genome function. The goal identified by functional genomics, that is using sequence information to assess gene function, may thus be addressed by considering the sequence-dependent structure of DNA. The question addressed in this thesis is: given a DNA sequence, what structure will it form and is that structure biologically important?

1.1 Biological functions of DNA that are defined by its structure

DNA is a molecule with biological activities defined by its structure. When Watson and Crick elucidated the structure of DNA in 1953, they could not help but notice that the base pairing in the structure suggested a "possible copying mechanism for genetic material" (Watson and Crick, 1953). Today many more activities for DNA are recognized, but the relationship to the structure is not as well defined. DNA encodes information about the sequence, folding pathway and ultimate destination of proteins in its primary sequence; but, admittedly, this activity does not relate to the

DNA's structure. However the sequence of DNA regulates the timing of gene expression (through the binding of transcription factors to promoters and enhancers) and tissue specific gene expression, by forming very specific interactions with proteins. These interactions are dependent on the structure of both the DNA and the protein (Luisi, 1995) and probably on the ability of both to adapt to one another (Grosschedl, 1995). DNA also has a positional activity. Stretches of DNA are present in promoters that keep regulatory elements in the promoter at the proper distance and in the correct orientation relative to one another. While the spacers do not bind proteins, only specific residues are tolerated at these positions, suggesting a relationship between sequence and structure (Barber et al., 1993; Warne and deHaseth, 1993; Werel et al., 1991). DNA also has the ability to bend or be otherwise flexible, and this has the potential to bring parts of genes into closer proximity with one another in three-dimensional space than would be expected from considering the DNA as a perfectly linear molecule. This bending appears to be important for gene regulation (Grosschedl, 1995). While all these biological functions of DNA structure are recognized, the detailed relationship between structure and function is not generally well understood.

1.2 DNA structure

DNA is a polymorphic molecule, and this is the basis for some of its regulatory functions. This polymorphism was established as early as 1953

when the structure of B-DNA (Watson and Crick, 1953) and the subsequent structure of A-DNA (Franklin and Gosling, 1953b) were established from analysis of diffraction patterns obtained from DNA fibers at different relative humidities. A- and B-DNA differ in such characteristics as helical repeat, rise and diameter. Fiber diffraction has subsequently revealed the helical repeat of other forms of regular double-stranded DNA structures (summarized in Saenger, 1984) (Table 1.2). These studies have established that DNA conformation is dependent on such factors as relative humidity, sequence and the nature of the stabilizing cation. B-DNA is the form found under conditions of high relative humidity, while the A, Z, C and D conformations are seen at lower relative humidities. Z-DNA (a left handed conformation with a helical twist of -12 residues per turn) was seen in alternating poly (dG-dC) sequences (Arnott et al., 1980) while the D-form was seen in alternating poly (dA-dT) sequence at low relative humidity (Arnott et al., 1974). The C-form was found at low relative humidity in the presence of lithium (Marvin et al., 1961). Finally, the A-form was seen at low relative humidity (Fuller et al., 1965). Fiber diffraction, then, does establish the polymorphic nature of DNA, and that sequence, hydration and cations affect its structure; but, it does not provide the molecular details of these structures.

Other physical techniques also offered some details of the relationship of nucleotide sequence and DNA structure. In particular, circular dichroism confirmed the results of the fiber diffraction studies and further

Table 1.2

DNA conformations studied using fiber diffraction (Saenger, 1984)

DNA Conformation	Helical Repeat (base pairs/turn)	Relative Humidity	Special requirements
Z	-12	43%	poly(dG-dC)
A	11	75%	-
B	10	92%	-
C	9	45-66%	Lithium salt
D	8	75%	poly(dA-dT)

elucidated the relationship between the structure and hydration for DNA conformations (Brahms and Mommaetrs, 1964; Ivanov et al., 1974; Ivanov et al., 1973; Sprecher et al., 1979). Most of these studies were confined to long polymers of DNA or DNA from natural sources. Therefore it was not straightforward to systematically study the effect of specific sequences of DNA on the structure.

In 1979, the advent of DNA synthesizers made it possible to synthesize large quantities of short oligonucleotides of specific sequences for crystallization. The first single-crystal structure of DNA was determined to 0.9 Å resolution by Wang, et al. (1979). This structure was that of the left handed form of Z-DNA. Shortly afterwards the single crystal structure of A-DNA was revealed by Shakked, et al. (1981) and that of B-DNA by Dickerson's group (Wing et al., 1980). This began an era of the systematic crystallizations of specific oligonucleotide sequences in an effort to understand the sequence, hydration and counterion dependence of not only the global helical structure of DNA, but also the fine, molecular details. To date, x-ray crystallography has revealed the molecular structures of the A-, B- and Z-DNA conformations. Thus attempts to correlate sequence and other effects with specific structures are feasible for these well characterized conformations.

The structures of A-, B- and Z-DNA that were determined from fiber diffraction and x-ray crystallography are shown in Figure 1.1. The global structure parameters of these conformations are summarized in Table 1.3.

A- and B-DNA are both right-handed helices. A-DNA is underwound relative to B-DNA; the helical repeat of A-DNA is 11 base pairs per turn as compared to 10 base pairs per turn for B-DNA. The helical rise (the distance between consecutive base pairs in the direction of the helix axis) of A-DNA (2.5 Å) is less than that of B-DNA (3.4 Å). Thus A-DNA is compressed along the helix axis and, concomitantly, a wider helical structure relative to B-DNA. This results in several distinct structural differences between A- and B-DNA that have the potential to be the basis for some of the structurally based activities of DNA. The base pairs in A-DNA are displaced away from the helix axis and are highly inclined relative to the axis. This is very different from B-DNA in which the helix axis runs through the base and the base pairs are perpendicular to the helix axis and therefore show no inclination. The structural consequences are that A-DNA has a wider, but shallower minor groove and a narrower, but deeper major groove as compared to B-DNA (Figure 1.1). In addition, A- and B-DNA differ in their sugar conformation, which is generally C3'-*endo* in A-DNA and C2'-*endo* in B-DNA. The glycosidic angle, χ, which describes the orientation of the base relative to the ribose sugar, is ~195° in A-DNA as compared to ~257° in B-DNA. Therefore, A- and B-DNA are structurally very distinct conformations.

Z-DNA is a left handed helix and is a dramatically different structure from both A- and B-DNA. Like A-DNA, Z-DNA is an underwound structure with a helical repeat of -12 base pairs per turn. However, Z-DNA is

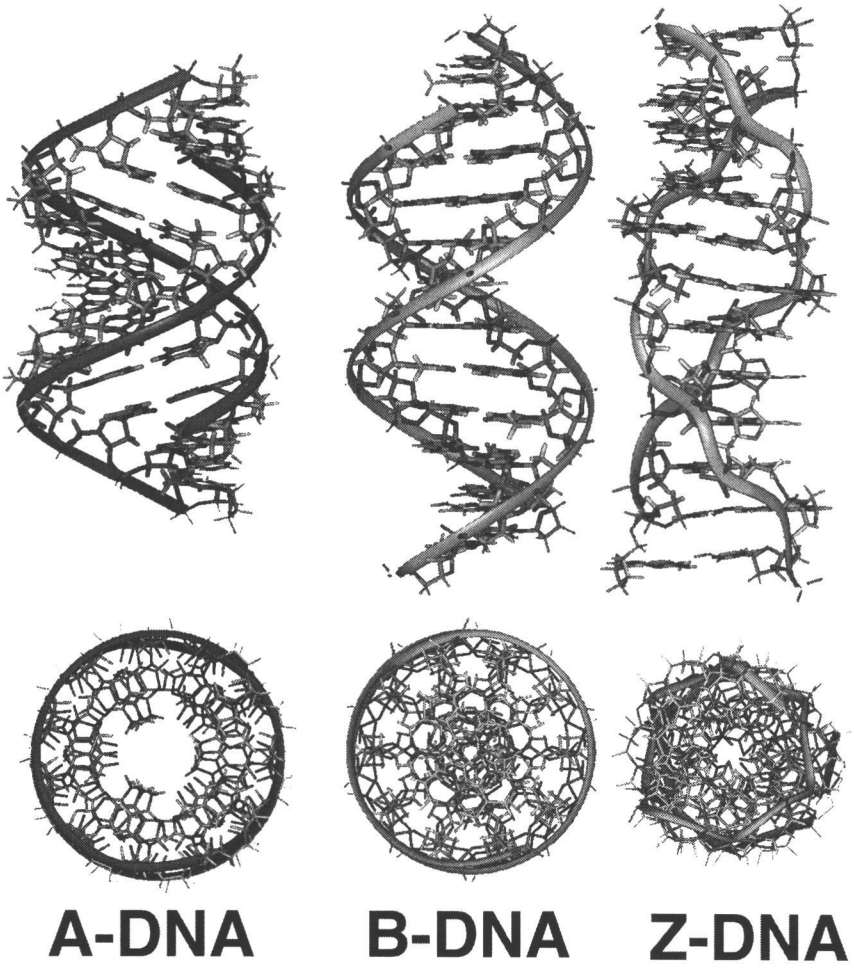


Figure 1.1

The A- B- and Z-conformations of DNA. Shown are models of 12 base pairs of A-, B- and Z-DNA built from standard models (Arnott and Hukins, 1972; Arnott et al., 1975b) viewed along the helix axis (top) and down the helix axis (bottom). Ribbons show the trace of the backbone of the helix.

Table 1.3

Helical parameters of A- B- and Z-DNA

	A-DNA	B-DNA	Z-DNA
Helical Twist (Ω)	33°	36°	-60°/dn
Helical Rise (D_z)	2.5 Å	3.4 Å	3.7 Å
Repeat Unit	nucleotide	nucleotide	dinunucleotide
Helical Repeat (base pairs/turn)	11	10-10.5	12
Pitch (length of one full turn of helix)	28 Å	34 Å	45 Å
Displacement of bases from helix axis	5 Å	0 Å	4 Å
Inclination	12°	2.4°	7°
Glycosidic Angle, - χ	195°	257°	208° (Py) 67° (Pu)
Sugar Pucker	C3'-*endo*	C2'-*endo*	C2'-*endo*(Py) C3'-*endo*(Pu)
Minor groove width	11 Å	5.7 Å	
Major groove width	2.7 Å	11.7 Å	
Minor groove depth	2.8 Å	7.5 Å	
Major groove depth	13.5 Å	8.5 Å	
Helical diameter	23 Å	19 Å	18 Å

dn - dinucleotide, Py - pyrimidine base, Pu - purine base

not simply a left handed version of A- or B-DNA. More importantly, the repeating unit for Z-DNA is not a mononucleotide as it is in A- and B-DNA, but rather a dinucleotide step. In general, sequences that form Z-DNA are those that contain a series of alternating pyrimidine and purine nucleotides. The pyrimidines have C2′-*endo* sugar puckers with the bases in the *anti* conformation whereas the purines have a C3′-*endo* sugar pucker with the bases in *syn* conformation (*anti* and *syn* refer to the orientation of the base relative to the sugar). The dinucleotide step in Z-DNA is composed of a pyrimidine and a purine. As in A-DNA, the bases in Z-DNA are displaced from the helix axis (by ~4 Å) and show moderate inclination relative to the helix axis. However, Z-DNA is a narrower and longer structure than B-DNA. The rise of Z-DNA is 3.7 Å as compared to 2.5 Å and 3.4 Å for A- and B-DNA respectively. The grooves in Z-DNA are also different from those of B-DNA. The minor groove is deep and narrow, and the major groove is not a groove at all, but rather a convex surface. Thus A-, B- and Z-DNA represent the well defined and distinct duplex structures of DNA.

While the dependence of the structure of DNA on the sequence has long been appreciated, only the rules for Z-DNA sequence-dependent conformation have been well characterized. Alternating pyrimidine-purine sequences favor Z-DNA, with the order of stability being d(m^5CpG) > d(CpG) > d(CpA)/d(TpG)> d(TpA) (Rich et al., 1984). For A-DNA the rules are less specific. It is known that long stretches of cytosines or guanines

favor A-DNA (Minchenkova et al., 1986; Peticolas et al., 1988), but there is no complete quantitative description of the sequence dependence of A-DNA formation. Additionally, the structural basis for the differences in stability due to sequence are not entirely understood.

The studies presented in this thesis provide an understanding of the relationship between the structure and sequence of DNA at two levels. In the first set of studies (Chapter 2), structures determined from x-ray crystallography are used to elucidate the intrinsic structural features of Z-DNA, the conformation for which the rules for stability have been best identified. This includes the effects of the crystal lattice on the detailed DNA structure, the effects of specific chemical modifications of the bases and the various cations on the structure and stability of Z-DNA. Chapter 3 explores the relationship between sequence and structure for the A-DNA conformation and aims to simply understand the relative stabilities of the different bases as A-DNA, since the rules for A-DNA formation have not been identified.

Once the rules for sequence-specific DNA structure have been established, it is of interest to identify the biological significance, if any, of specific structures and therefore of specific sequences. Since there are no reliable, general methods to probe for specific DNA structures *in vivo*, computational methods are an attractive way to search for sequences that have a high propensity to form non-B-DNA structures. For analyses of this kind, one searches for correlation between a sequence's potential to form a particular conformation and its physical location in the gene. This

approach was employed by Ho, et. al. (1986) to develop ZHUNT, an algorithm that uses the B-to-Z transition energies of the various dinucleotides to calculate the probability that a given sequence is in the Z conformation relative to average random sequences. An analysis of 137 human genes showed that Z-DNA forming sequences occur non-randomly and are more common near the 5′ end of genes. This suggested a putative role of Z-DNA in transcriptional regulation (Schroth et al., 1992). Chapter 4 of this work describes the development and application of an analogous search algorithm for A-DNA, AHUNT, which was developed to search genes for putative A-DNA regions, and to compare and contrast the propensity that A-DNA will form in regions of genes of different organisms.

In summary, these chapters explore aspects of sequence dependent DNA conformation. The thermodynamics and atomic-level structure of Z-DNA sequences are discussed in chapter 2. The remaining work focuses on the development of predictive rules for the propensities of specific sequences to form A-DNA and the application of these rules to exploring the biological significance of this structure.

Chapter 2

2. THE SINGLE-CRYSTAL STRUCTURES OF Z-DNA

Beth Basham, Brandt F. Eichman, and P. Shing Ho

Invited review, *The Oxford Handbook of Nucleic Acid Structure*, in press

2.1 Synopsis

In the nearly 20 years since it was first discovered, over 50 structures of left-handed Z-DNA have been studied in single crystals. These structures have provided insight into the physical properties of this unusual DNA duplex structure, particularly in terms of the types of sequences and base modifications that help to stabilize the conformation. Here, we review the structures of Z-DNA with an emphasis on how sequence affects the DNA structure and solvation of the DNA in trying to understand its stability relative to standard B-DNA.

2.2 Introduction

Z-DNA is a highly unique and unusual structure in biology. It is a left-handed double-helix, which at the time of its discovery in 1979 was dramatically different from any of the known forms of either DNA or RNA. Prior to the characterization of Z-DNA by crystallography, the fiber diffraction x-ray structures of naturally occurring and synthetic DNAs were all right-handed helices. B-DNA, first described by Watson and Crick (Watson and Crick, 1953), and A-DNA, described by Franklin and Gosling (Franklin and Gosling, 1953b) immediately afterwards, were the predominant models for the double-helix in solution and in the cell.

The discovery of Z-DNA itself was unusual in that it was the first detailed structure of any oligonucleotide to be determined by x-ray diffraction of single crystals. There had been prior spectroscopic evidence for a left-handed form of the synthetic sequence poly[d(GpC)] under high salt conditions (Pohl and Jovin, 1972), but the handedness of the structure could not be conclusively assigned until the crystal structure of d(CGCGCG) was determined by Wang, *et al.* (1979, 1981) in the laboratory of Dr. Alexander Rich, and confirmed by the structure of d(CGCG) by Drew, *et al.*, (1980) in Prof. Richard Dickerson's laboratory. The single crystal structures of A- and B-DNA were determined soon afterwards (Shakked et al., 1981; Wing et al., 1980). Z-DNA was an unusual structure when it was discovered in DNA crystals, since there had been no prior physical characterization (other than the circular dichroism results) or studies on the biology of this conformation.

The biology of Z-DNA is still widely debated. This problem of finding a function for a structure is difficult because it runs counter to the normal progression in biology, that is determining the structure responsible for a previously defined function. Since the crystal structures themselves cannot directly address this problem, we will not engage in this debate here.

Where the single crystal structures become very useful is in characterizing the physical properties of Z-DNA. This conformation has been studied extensively by not only crystallography, but also in solution by various spectroscopic and biochemical methods. Together, the results of

these studies have yielded a highly detailed description of what is required to induce and stabilize Z-DNA. From the early studies, it was suggested that this conformation can form only in alternating pyrimidine-purine (APP) sequences, specifically CG rich APP sequences, and in the presence of extremely high salt concentrations. It is now known that not only are d(T·A) base pairs accomodated by the Z-DNA structure, but nonalternating sequences can also adopt the Z-conformation (Table 2.1). Furthermore, Z-DNA can be induced to form under physiological conditions in the presence of cellular cations (e.g., polyamines) (Behe and Felsenfeld, 1981; Feuerstein et al., 1991), by negative supercoiling in closed circular plasmids (Thomas et al., 1991; Thomas and Thomas, 1994) or in the wake of a transcribing polymerase (Rahmouni and Wells, 1989). Many of the sequence rules for the formation of Z-DNA were determined from crystallographic studies in concert with solution studies. This review will focus on the more tangible issues of how the atomic structure of Z-DNA is affected by sequence and sequence modifications. For this discussion, the structural effects include both the conformation of the DNA and the interactions of the DNA with the solvent.

2.3 The prototypical Z-DNA structure of d(CGCGCG)

To date, there are over 50 single crystal structures of Z-DNA (Table 2.2), of sequences with lengths varying from 2 to 10 base pairs. This set

Table 2.1

Conditions that affect Z-DNA crystallization and stability

[a] B-to-Z transition free energies ($\Delta G°_T$) were determined in negatively supercoiled closed circular DNA.
[b] Modified nucleotide bases: dBr^5C = C5-bromodeoxycytosine; dm^5C = C5-methyldeoxycytosine; dEt^5C = C5-ethyldeoxycytosine; dI = deoxyinosine; dD = diaminodeoxypurine (2-aminodeoxyadenine); dU = deoxyuridine
[c] These references refer to studies of synthetic polymers, as opposed to single crystals.

Table 2.1

		Reference
Dinucleotide	$\Delta G°_T$ (kcal/mol-dn)[a]	
d(CpG)	0.66	(Peck and Wang, 1983)
d(CpA)/d(TpG)	1.34	(Vologodskii and Frank-Kamenetskii, 1984)
d(TpA)	≥2.4	(Ellison et al., 1986; McLean et al., 1988)
d(CpC)/d(GpG)	2.4	(Ellison et al., 1985)
d(TpC)/d(GpA)	2.5	(Ellison et al., 1985)
Substituent[b]		
$dBr^5C>dm^5C>dC>dEt^5C$		(Behe and Felsenfeld, 1981; Jovin et al., 1983; Moller et al., 1984; Sagi et al., 1991)[c]; (Chevrier et al., 1986; Fujii et al., 1982)
dG>dI		(Vorlickova and Sagi, 1991; Wang and Keiderling, 1993)[c]; (Ho et al., 1991)
dD>dA		(Coll et al., 1986)
dU>dT		(Schneider et al., 1992; Zhou and Ho, 1990)
Salts (Cations and Anions)		
$Co(NH_3)_6^{3+}>Ba^{2+}>Ca^{2+}>Mg^{2+}>Na^+>Li^+>NH_4^+$		(Behe and Felsenfeld, 1981; Kagawa et al., 1993)
$Spermine^{4+}>Spermidine^{3+}$		(Kagawa et al., 1993)
tetraalkylammonium ion^+		(McDonnell and Preisler, 1989)
tetraalkyl $carboxylate^-$		(McDonnell and Preisler, 1989)
Solvents		
Alcohols: Methanol, Ethanol, Propanol		(Preisler et al., 1995)
Polyols: MPD, Glycerol to stachyose		(Ho et al., 1991; Preisler et al., 1995; Tereshko and Milinina, 1990)

includes structures with standard and modified bases, with standard Watson-Crick and mismatched base pairs, with standard and modified deoxyribose backbones, and with various cations. The data set is dominated by self-complementary APP hexanucleotide sequences that start with a dC or its methylated analog, dm^5C at the 5'-end of each strand. In all these structures the overall conformation maintains the general features observed in the original Z-DNA structure of d(CGCGCG). Therefore, a logical starting point in this review is to discuss the conformation of d(CGCGCG) (Wang et al., 1979) as the prototypical Z-DNA structure to which all other structures will be compared. We will start by describing the gross morphology of the Z-DNA structure, which remains essentially identical in all structures of this conformation, followed by a detailed description of the helical parameters in which the external and internal influences on the fine structure are observed.

2.3.1 The structure of Z-DNA

The two obvious features of Z-DNA that distinguish it from both A- and B-DNA are that it is a left-handed double-helix and that its backbone has a characteristic zigzag pattern (Figure 2.1). It is the zig-zagged backbone that gives this form of DNA its name (Wang et al., 1979). The distinctive backbone pattern arises from an alternating conformation of the bases relative to the deoxyribose sugar, as defined by the rotation about the

Table 2.2

Catalog of Z-DNA crystal structures[a]

Sequence	NDB	PDB	Special Features/ Ions Present	Space Group	Resolution Å	R Factor %	*CS* Molar	Reference
d(CpG) family								
d(CpG)	zdb020	-		$P2_12_12_1$	0.8	13.6	-	(Ramakrishnan and Viswamitra, 1988)
d(CGCG)	zdd015	1zna	high salt, Cl^-	$C222_1$	1.5	19.9	-	(Drew et al., 1980)
d(CGCG)[b]	zdd022	-	spermine	$P6_5$	1.5	19.3	-	(Crawford et al., 1980)
d(CGCG)[b]	zdd023	-		$P6_5$	1.5	21.0	-	(Crawford et al., 1980)
d(CGCGCG)	zdf001	2dcg	Mg^{2+}, spermine	$P2_12_12_1$	0.9	14.0	2.19	(Wang et al., 1979)
d(CGCGCG)	zdf002	1dcg	Mg^{2+}	$P2_12_12_1$	1.0	17.5	2.00	(Gessner et al., 1989)
d(CGCGCG)	zdf029	1d48	spermine	$P2_12_12_1$	1.0	18.5	2.00	(Egli et al., 1991)
d(CGCGCG)	zdf035	131d	spermine, -110°C	$P2_12_12_1$	1.0	18.0	2.00	(Bancroft et al., 1994)
d(CGCGCG)	zdf052	293d	Mg^{2+}, spermidine	$P2_12_12_1$	1.0	19.1	-	(Ohishi et al., 1996)
d(CGCGCG)	zdf007	-	$Ru(NH_3)_6^{3+}$	$P2_12_12_1$	1.2	20.0	-	(Ho et al., 1987)
d(CGCGCG)	zdf019	-	Mg^{2+}, $Co(NH_3)_6^{3+}$	$P2_12_12_1$	1.3	18.5	-	(Gessner et al., 1985)
d(CGCGCG)	zdf044	-	Mg^{2+}, Co(II)Cl	$P2_12_12_1$	1.5	19.6	-	(Gao et al., 1993)
d(CGCGCG)	zdf045	-	spermine, Co(II)Cl	$P2_12_12_1$	1.5	23.7	-	(Gao et al., 1993)
d(GCGCGCG)/ d(CCGCGCG)	zdg054	-	5' overhang Mg^{2+}, $Co(NH_3)_6^{3+}$	$P2_12_12_1$	1.8	20.9	0.23	(Mooers et al.,1997)
d(GCGCGCG)/ d(TCGCGCG)	zdg056	-	5' overhang, Mg^{2+}, $Co(NH_3)_6^{3+}$	$P2_12_12_1$	1.9	19.1	0.23	(Mooers et al.,1997)
d(CGCGCGCG)[b]	zdh017	-		$P6_5$	1.6	19.0	-	(Fujii et al., 1985)
d(CGCGCGCGCG)[b]	zdj050	279d		$P6_522$	1.9	18.6	0.32	(Ban et al., 1996)
d(CCGCGG)	udf025	1d16		$C222_1$	1.9	18.5	0.14	(Malinina et al., 1994)

Table 2.2, continued

Covalent Modifications								
d(CGCGCG)	zdf028	1d39	Cu(II) Cl soaked	$P2_12_12_1$	1.2	19.8	2.00	(Kagawa et al., 1991)
d(Gm^{5}CGCGCG)	zdgb55	-	5' overhang, Mg^{2+}, $Co(NH_3)_6^{3+}$	$P2_12_12_1$	1.7	20.7	0.23	(Mooers et al., 1997)
d(m^{5}CGm^{5}CGm^{5}CG)	zdfb03	-		$P2_12_12_1$	1.3	15.6	0.57	(Fujii et al., 1982)
d(Br^{5}CGBr^{5}CGBr^{5}CG)	zdfb04	1dn4	291K	$P2_12_12_1$	1.6	13.3	-	(Chevrier et al., 1986)
d(Br^{5}CGBr^{5}CGBr^{5}CG)	zdfb05	1dn5	310K	$P2_12_12_1$	1.4	12.5	-	(Chevrier et al., 1986)
d(CACGTG)	zdf008	-		$P2_12_12_1$	2.5	22.9	3.30	(Coll et al., 1988)
d(CGCACG)/ d(CGTGCG)	zdf038	-		$P2_1$	2.5	16.1	-	(Sadsivan and Gautham, 1995)
d(CACGCG)/ d(CGCGTG)	zdf039	-		$P2_12_12_1$	1.6	19.9	0.60	(Sadsivan and Gautham, 1995)
d(m^{5}CGTAm^{5}CG)	zdfb06	-		$P2_12_12_1$	1.2	16.0	1.26	(Wang et al., 1984)
d(Br^{5}CGATBr^{5}CG)	zdfb09	-		$P2_12_12_1$	1.5	19.3	-	(Wang et al., 1985)
d(m^{5}CGUAm^{5}CG)	zdfb10	-	Cu(II)Cl soaked	$P2_12_12_1$	1.3	20.9	0.31	(Geierstanger et al., 1991)
d(CDCGTG)	zdfb11	-		$P2_12_12_1$	1.3	21.7	0.67	(Coll et al., 1986)
d(CGCm^{6}GCG)	zdfb21	1d24	O6 methylguanine	$P2_12_12_1$	1.9	19.0	0.71	(Ginell et al., 1990)
d(m^{5}CGUAm^{5}CG)	zdfb24	1d41		$P2_12_12_1$	1.3	20.8	0.31	(Zhou and Ho, 1990)
d(CGCGm^{4}CG)	zdfb25	1da2	N4 methoxycytosine	$P2_12_12_1$	1.7	18.1	0.88	(van Meervelt et al., 1990)
d(CGCUDCG)	zdfb31	1d76		$P2_12_12_1$	1.3	13.8	0.60	(Schneider et al., 1992)
d(CGCICG)	zdfb34	-		$P2_12_12_1$	1.7	14.8	0.35	(Kumar and Weber, 1993)
d(CGCGm^{5}CG)	zdfb36	133d	O4 methylcytosine	$P2_12_12_1$	1.8	18.9	2.29	(Cervi et al., 1993)
d(m^{5}CGGCm^{5}CG)	-	-	nonAPP	$P2_12_12_1$	1.6	20.8	1.26	(Eichman et al.,)
d(m^{5}CGGGm^{5}CG)/ d(m^{5}CGm^{5}CCm^{5}CG)	zdfb37	145d	non APP	$P2_12_12_1$	1.3	19.3	0.59	(Schroth et al., 1993)
d(CGCGm^{5}CG)	-	-	hemimethylated	$P2_12_12_1$	1.4	18.9	1.26	(Bononi, 1995)
d(CGTDCG)	zdfb41	210d		$P3_221$	1.4	17.4	1.35	(Parkinson et al., 1995)
d(CGTDC(Pt)G)	zdfb42	211d	Platinated Guanine	$P3_221$	1.6	17.0	1.60	(Parkinson et al., 1995)
d(CGCGOCG)	zdfb43	223d	oxydimethylene	$P2_12_12_1$	1.7	17.9	0.18	(Moore et al., 1995)

Table 2.2, continued

d(m^{5}CGm^{5}CGTG)	zdfb48	-	$BaCl_2$ soaked		1.3	19.7	-	(Gao et al., 1993)
d(CGCGBr^{5}CG)	zdfb51	242d	Hemibrominated	$P2_12_12_1$	1.6	17.0	-	(Peterson et al., 1996)
d(CGCATGCG)[b]	zdh016	-		$P6_5$	2.5	16.0	3.28	(Fujii et al., 1985)
d(CGCICICG)	zdh030	1d53		$P6_5$	1.5	22.5	0.35	(Kumar et al., 1992)
d(CGTACGTACG)	zdj018	1dn8	$Co(NH_3)_6^{3+}$	$P6_5$	1.5	25.0	-	(Brennan et al., 1986)
d(CGCGCG)	zdf040		racemic mixture	P1-	2.2	19.9	0.21	(Doi, 1993)
d(CGCGTG)	zdf046	-	Mg^{2+}, Co(II)Cl	$P2_12_12_1$	1.5	21.5	-	(Gao, 1993)
d(CGCGTG)	zdf047	-	Mg^{2+}, Cu(II)Cl	$P2_12_12_1$	1.5	18.0	-	(Gao, 1993)
Backbone Modifications								
a(C)d(G)a(C)d(G)a(C)d(G)	zdfs33	-		$P6_522$	1.3	28.7	0.61	(Zhang et al., 1992)
d(CG)r(CG)d(CG)	zhf026	-		$P2_12_12_1$	1.5	20.4	0.74	(Teng, et al., 1989)
d(CG)a(C)d(GCG)	zdfs27	-		$P2_12_12_1$	1.5	16.7	0.61	(Teng, et al., 1989)
Mismatches								
d(CGCGTG)	zdf013	-	G/T wobble base pair	$P2_12_12_1$	1.0	19.5	0.22	(Ho et al., 1985)
d(CGCGF^{5}UG)	zdfb12	1dnf	F^{5}U/G wobble base pair	$P2_12_12_1$	1.5	17.2	2.42	(Coll et al., 1989)
d(Br^{5}UGCGCG)	zdfb14	1da1	Br^{5}U/G base mismatch	$P2_12_12_1$	2.2	15.6	0.55	(Brown et al., 1986)

[a] NDB and PDB are the entry codes for the Nucleic Acid Database (Berman et al., 1992) and Protein Database (Bernstein et al., 1977), respectively. *CS* is the cationic strength of the solution used to crystallize the sequence, calculated as $CS = Z_i^2[M_i]$ (where Z is the charge and [M] is the concentration of each cation species *i* in the crystallization setup as described in (Ho et al., 1991)).

[b] Disordered structures

glycosidic bond (χ) (Figure 2.2). In the two right-handed DNA conformations, all the bases along the chains adopt an *anti* conformation (defined as χ between 90° and 270°, but typically with values of $\chi \sim 210°$). The bases are thus extended out and away from the phosphoribose backbone. In Z-DNA, the nucleotides alternate between the standard *anti* conformation ($<\chi> = 208°$) and the more compact *syn* conformation ($<\chi> = 67°$) (Table 2.3), with the base essentially sitting on top of the deoxyribose ring (Figure 2.1). The steric inhibition to pyrimidines adopting the *syn* conformation imposes the characteristic APP sequence motif commonly associated with Z-DNA. This alternating pattern of *anti-syn* nucleotide conformations is strictly adhered to in all the crystal structures of Z-DNA including those containing out-of-alternation sequences and non Watson-Crick base pairs. Thus, it is the alternation in the backbone and not the sequence that defines Z-DNA.

In an antiparallel DNA double helix, there is a major groove and a minor groove. B-DNA has a deep major groove and shallow minor groove. Z-DNA has a minor groove that is a deep narrow crevice, which brings the phosphate groups of opposite strands closer together than in A- or B-DNA. In contrast, the major groove of Z-DNA is more a convex surface than a true groove and, consequently, exposes more atoms to solvent than would be expect for B-DNA (Figure 2.1).

Figure 2.1

Structure of d(CGCGCG) as Z-DNA. A. The two stacked hexanucleotide duplexes in the crystal structure of d(CGCGCG) are shown as a stereodiagram. The upper duplex is shown as a CPK model using the van der Waal's radii of each atom to define spheres for each atom. The lower duplex is shown as a stick model, with the backbone phosphates traced with a ribbon to show the zigzag nature of Z-DNA. The nucleotides are numbered from the 5' to the 3' terminus of each strand, 1-6 for one strand and 7-12 for the comple-mentary strand. The d(CpG) dinucleotide in the *anti-p-syn* stacking arrangement (B) and the d(GpC) dinucleotide in the *syn-p-anti* stacking arrangement (C) are shown looking down the helix axis. Hydrogen bonds are shown as dashed lines connecting the bases of each base pair.

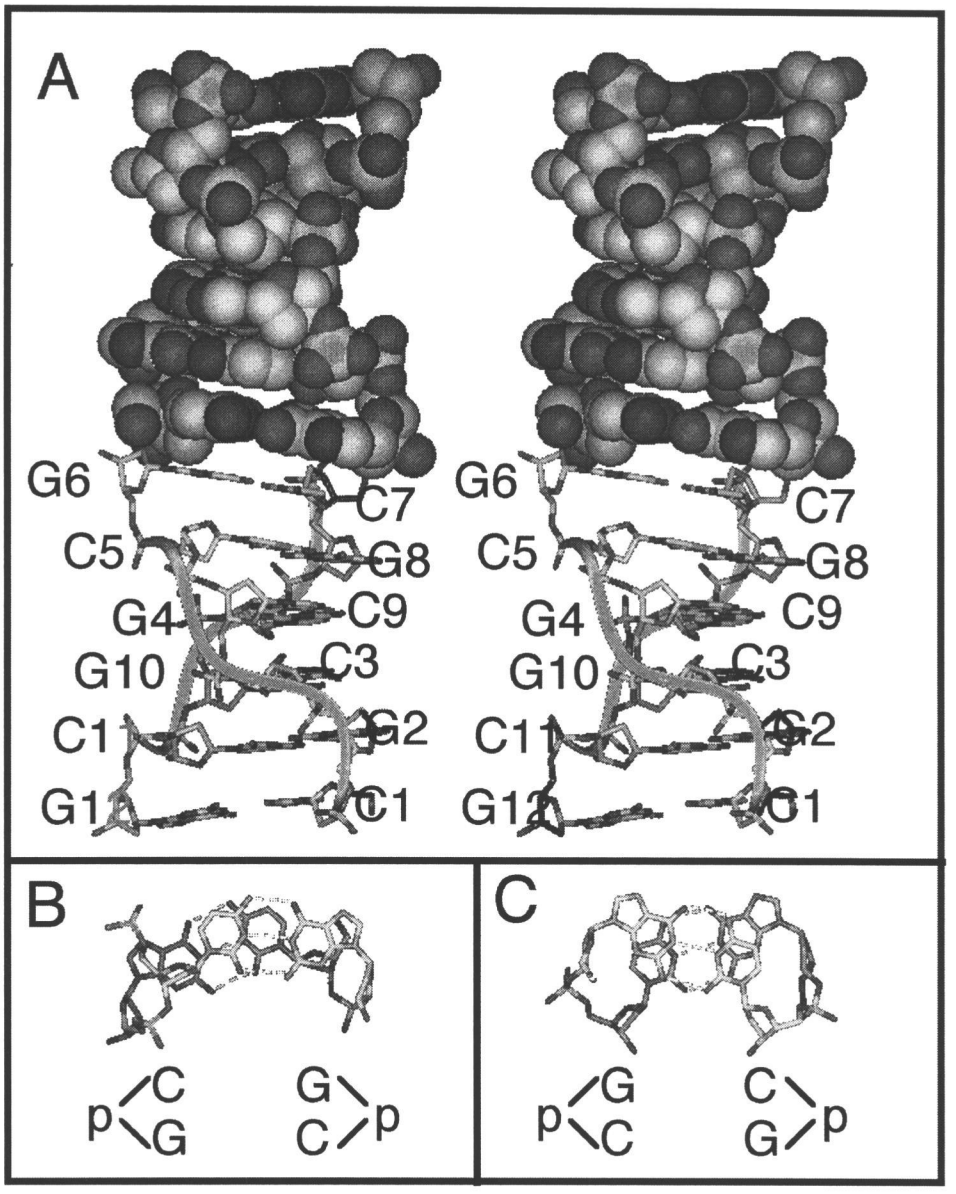

A
G6
C7
C5
G8
G4
C9
G10
C3
C1
G2
G12
C1
G6
C7
C5
G8
G4
C9
G10
C3
C11
G2
G12
C1
B
p
C
G
G
C
p
C
p
G
C
G
p

Figure 2.1

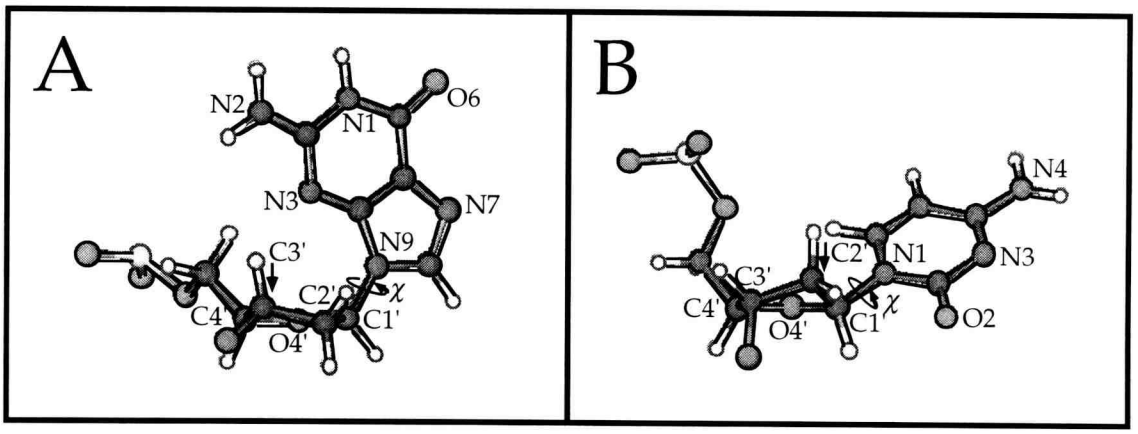


Figure 2.2

Comparison of the guanine nucleotide in the *syn* conformation (A) and cytosine in the *anti* conformation (B) of d(CGCGCG) as Z-DNA. The nitrogen and oxygens of each base, along with the atoms of the furanose ring of the 2'-deoxyribose sugar are labeled. The arrows show the carbon in the sugar rings that define the C3'-*endo* and C2'-*endo* sugar puckers for the guanine in *syn* (A) and cytosine in *anti* (B). The rotation about the glycosidic bond that defines the *syn* and *anti* conformations of each nucleotide are labeled as χ.

Table 2.3

Helical parameters of d(CGCGCG) crystallized in the presence of spermine^{4+} and Mg^{2+} (Wang et al., 1979)

Base	χ°	Sugar Pucker	Base	χ°	Sugar Pucker
C1	208.5	C2'-*endo*	G12	79.5	C2'-*endo*
G2	57.2	C3'-*endo*	C11	203.9	C2'-*endo*
C3	202.4	C2'-*endo*	G10	64.7	C3'-*endo*
G4	52.5	C3'-*endo*	C9	200.1	C2'-*endo*
C5	214.8	C2'-*endo*	G8	69.5	C3'-*endo*
G6	79.1	C2'-*endo*	C7	217.7	C2'-*endo*
$\langle\chi_{\text{Cytosine}}\rangle$	**207.9±7.1**		$\langle\chi_{\text{Guanine}}\rangle$	**67.1±11.1**	

d(CpG) step	Twist (Ω)	Rise (D_z)	Roll (ρ)	Tilt (τ)
(C1·G12)/(G2·C11)	-8.5	3.8	-3.0	6.9
(C3·G10)/(G4·C9)	-9.1	3.8	3.6	1.1
(C5·G8)/(G6·C7)	-10.6	4.3	-2.1	0.7
Average	**-9.4±1.1**	**4.0±0.3**	**-0.5±3.6**	**2.9±3.5**
d(GpC) step				
(G2·C11)/(C3·G10)	-48.8	3.7	-0.8	-0.6
(G4·C9)/(C5·G8)	-51.4	3.6	0.4	0.2
Average	**-50.1±1.8**	**3.7±0.1**	**-0.2±0.8**	**-0.2±0.6**

	Tip	Inclination	Buckle	Propeller	Displacement	
Base Pair	(θ)	(η)	(κ)	Twist (ω)	d_x	d_y
C1·G12	3.0	6.9	0.3	0.8	-3.3	2.5
G2·C11	2.1	7.5	4.8	2.1	-3.1	1.9
C3·G10	-1.5	6.4	2.8	5.6	-3.1	2.2
G4·C9	-1.1	6.6	5.9	3.4	-3.1	2.4
C5·G8	1.0	7.3	0.1	0.6	-3.5	2.0
G6·C7	0.9	7.7	4.4	3.2	-3.4	1.9
Average	**0.7±1.8**	**7.1±0.5**	**3.1±2.4**	**2.6±1.9**	**-3.3±0.2**	**2.2±0.3**

All parameters were calculated with NASTE. All values are in degrees, except rise (D_z) and displacement (d_x, d_y) which are in Å.

The conformations of the deoxyribose sugars along the phosphoribose backbone are strongly affected by the alternating *anti-syn* structure of Z-DNA. These sugar conformations are defined by how the furanose ring is distorted (or puckered) from planarity (Figure 2.2). In B-DNA, the sugars adopt the C2′-*endo* conformation in which the C2′ carbon sits above the plane (towards the base) formed by the C1', O4', and C4' atoms. In A-DNA, the sugar puckers are C3'-*endo*. The deoxyriboses of Z-DNA alternate between C3'-*endo* for the nucleotides in *syn* and C2'-*endo* for nucleotides in *anti* (Table 2.3). This alternating pattern is seen for all sequences, including nonAPP sequences that place pyrimidines in *syn*. In the crystal structure of d(CGCGCG), as well as other Z-DNA sequences, the 3'-terminal dG nucleotide has a C2'-*endo* sugar, even though the guanine is in *syn* (Table 2.3). This end effect and other exceptions to sugar pucker alternation most likely reflect distortions induced by the crystal lattice rather than any inherent sequence effect.

The phosphate backbone linking the sugars of each nucleotide shows two different conformations: Z_I and Z_{II}. The Z_I conformation is characterized by a pattern of alternating torsion angles along the backbone (α to ζ). The Z_{II} form shows exceptions to this alternating pattern (most prominently at α, β and γ), usually between the fourth and fifth base pairs of one strand of the hexamer duplex. The Z_{II} conformation rotates the phosphate out and away from the minor groove crevice at this nucleotide.

This differentiates one strand from the other in the crystal for most of the structures in which the asymmetric unit is the DNA duplex; however, the bases are not dramatically affected by these deviations. The Z_{II} pattern has been attributed to crystal packing effects, and has been suggested to be stabilized by a specific pattern of waters at the interface between Z-DNA duplexes (Schneider et al., 1992). The Z_I pattern is generally considered to be representative of the average structure of the Z-DNA backbone, while the existence of the Z_{II} pattern reflects the degree of flexibility in the backbone of an otherwise rigid structure.

This adherence to a characteristic zig-zagged pattern in the backbone, with the nucleotides always in the alternating *anti-syn* conformation even for non-APP and non-Watson-Crick base pairs, suggests that Z-DNA is very rigid in its conformity to a structural and not to a sequence pattern. The repeating unit of Z-DNA is therefore a dinucleotide with very well defined geometries. The zig-zag pattern of the backbone results in the stacking of the base pairs in two different arrangements. A d(CpG) dinucleotide places the pyrimidines in the *anti* conformation 5' to purines in *syn* along each chain (an *anti-p-syn* step), while the alternative d(GpC) dinucleotide has a purine in *syn* stacked over a pyrimidine in *anti* (a *syn-p-anti* step) (Figure 2.1B and C). In the *anti-p-syn* dinucleotide, the bases of the pyrimidines from opposite strands are actually stacked, while the purines stack over the deoxyriboses of the adjacent pyrimidines along the same strands. It has been suggested that this latter stacking is stabilized by a favorable

electrostatic interaction between the π electrons of the purine and the nonbonding electrons of the O4' oxygen of the sugar ring (Egli and Gessner, 1995). The *syn-p-anti* step places the six membered ring of the purine over the adjacent pyrimidine of the same strand. Thus, although an argument can be made that the *anti-p-syn* stacking arrangement is more stable, it is difficult to accurately compare the base stacking interactions because they are so different between the two stacking modes. There may be a difference imposed by the solvent interactions with Z-DNA, which will be discussed later. For now, we will treat the *anti-p-syn* dinucleotide as the repeating unit in Z-DNA. The sequence d(CGCGCG) can therefore be thought of as three stacked repeats of d(CpG) dinucleotides in the *anti-p-syn* conformation, with the interfaces being the *syn-anti* arrangements.

Thus, the overall shape of Z-DNA remains fairly consistent across all the Z-DNA crystal structures that have been determined. Factors such as sequence and solvent interactions affect the details of the structure, which are best described by the helical parameters.

2.3.2 The helix structure of d(CGCGCG)

We will compare the helical parameters of the various Z-DNA structures in order to elucidate the effect of any particular factor on the conformation. A set of standard definitions for helical parameters has previously been established by the EMBO Workshop (Diekmann, 1989);

however, some of these parameters have special meaning for Z-DNA. We therefore developed an algorithm, NASTE (Nucleic Acid Structure Evaluation), to calculate various helical parameters specifically for Z-DNA. The algorithm first transposes all base pairs to a common frame of reference, defined by the helix axis and the perpendicular from the helix axis to the base pair's long axis (Figure 2.3). The helical parameters of each base pair and each base step within the structures are calculated from this frame of reference (Figure 2.3). The effects of cations, sequence, base modifications and crystal packing forces on these parameters will be discussed in this review.

The helical parameters most useful in describing the *anti-p-syn* and *syn-p-anti* base steps of Z-DNA include the rise (D_Z) and helical twist (Ω) between each base pair. The single crystal structures of Z-DNA are long and narrow, with an average helical rise ($<D_Z>$) of 3.8 Å and a width of ~20 Å. By comparison, the $<D_Z>$ and width of B-DNA is 3.4 Å and ~24 Å, respectively. In Z-DNA , the average helical twist ($<\Omega>$) is -30° per base pair; thus, each base pair of Z-DNA is underwound on average by -66° relative to B-DNA ($<\Omega> = 36°$). For the remainder of this discussion we will only be comparing Z-DNA structures, and the terms "over-" and "underwound" will refer to more and less left-handed twists (negative Ω), respectively. The repeat unit of Z-DNA, however, is the dinucleotide and thus it is more accurate to compare Ω for the distinct dinucleotide repeats in

Figure 2.3

Helical parameters calculated with NASTE. Base pairs were transposed to a common frame of reference in which the base pair long axis was aligned on the y-axis and the base was perpendicular to the helix axis. The major groove points in the positive x direction. N is the N1 for pyrimidines and N9 for purines.

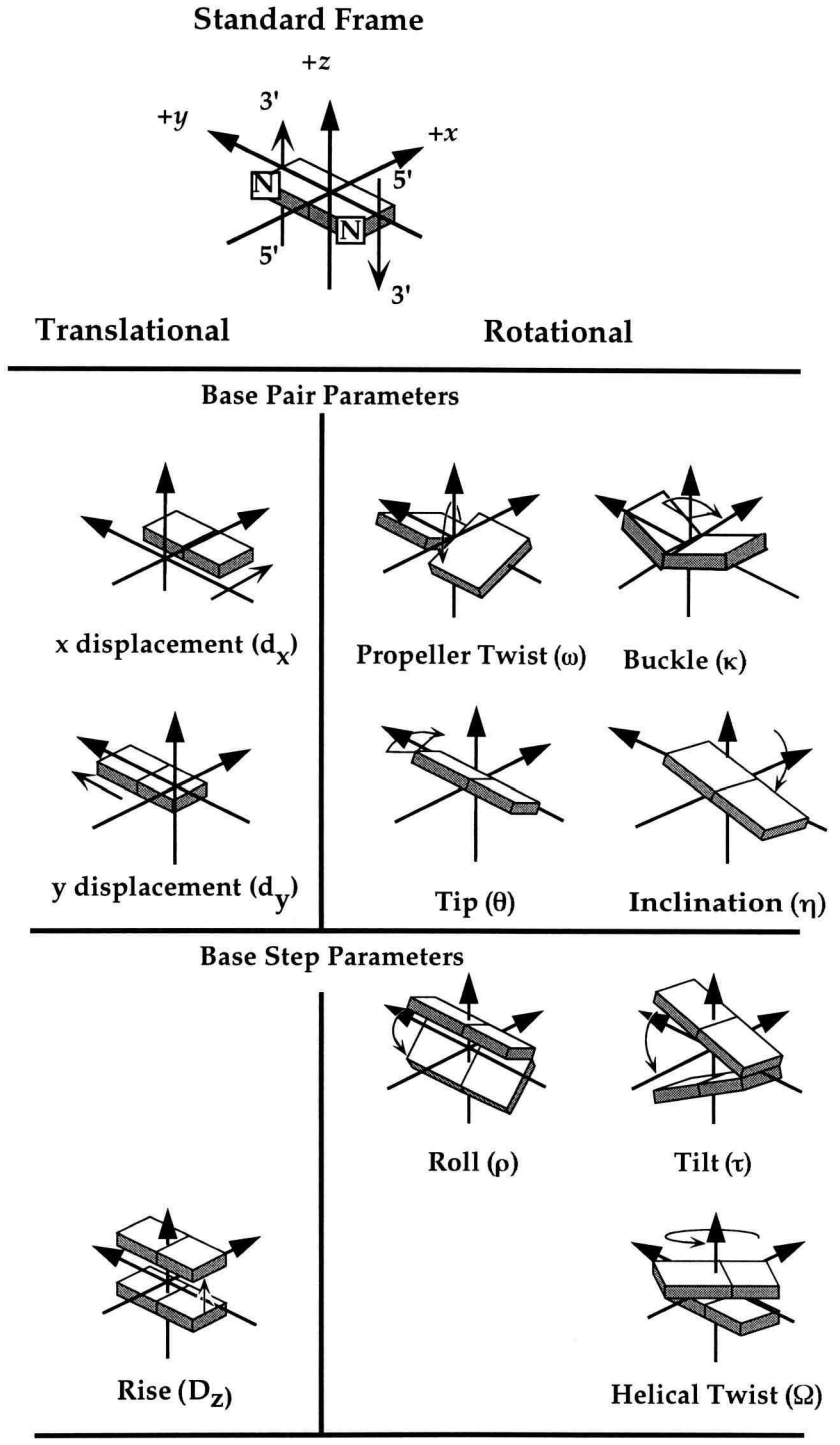

Standard Frame
+z
3'
+y
+x
5'
5'
3'
Translational
Rotational
Base Pair Parameters
x displacement (d_x)
Propeller Twist (ω)
Buckle (κ)
y displacement (d_y)
Tip (θ)
Inclination (η)
Base Step Parameters
Roll (ρ)
Tilt (τ)
Rise (D_z)
Helical Twist (Ω)

Figure 2.3

the structure. The d(CpG) step in d(CGCGCG) is characterized by <Ω> = -9.4° ± 1.1°, while for the d(GpC) step <Ω> = -50.1° ± 1.8° (Table 2.3), to give a total <Ω> = -59.5° for the sum of the dinucleotide steps (or <Ω> = 29.8° per base step).

Roll (ρ) and tilt (τ) describe the angles between adjacent base pairs along their long and short axes, respectively (Figure 2.3). Positive roll indicates that the bases open toward the major groove and positive tilt indicates that the angle opens toward the leading strand, which is defined as the strand containing the first nucleotide (Figure 2.1). NASTE's assignments of these paramters are consistent with these EMBO conventions (Diekmann, 1989). The d(CpG) steps in Z-DNA have a greater average roll (<ρ> = -0.5° ± 3.6°) than the d(GpC) steps (<ρ> = -0.2° ± 0.8°) (Table 2.3). The average roll for all steps is -0.4° ± 2.6°. Tilt does not differ between d(CpG) and d(GpC) steps. The average tilt is <τ> = 1.7° ± 2.9° for d(CGCGCG). By comparison, A-DNA tends to have a much larger roll (<ρ> = 6.3°) (Dickerson, 1992), and B-DNA tends to have only small degrees of roll (<ρ> = 0.6°) and tilt (<τ> = 0.0°) (Dickerson, 1992).

The structure of DNA duplexes is additionally described by how each base pair is rotated along its long and short axes (the rotational helical parameters) and translated along these axes (the displacement) relative to the helix axis. The rotational helical parameters (tip and inclination) are

calculated as the angle between the perpendicular to the base plane (the base normal) and the helix axis (Figure 2.3). Tip (θ) measures the rotation around the long axis of the base. In our comparisons, a positive value for tip indicates that the base pair is rotated toward the major groove. Inclination (η) measures the rotation around the short axis of the base pair, with a positive inclination reflecting a rotation toward the second strand. The average tip observed in Z-DNA (Table 2.3) is θ = 0.7° ± 1.8° and is greater than that observed for B-DNA (θ = 0°) (Dickerson, 1992), but much less than the average tip observed in A-DNA (θ = 11.0°) (Dickerson, 1992). Likewise, Z-DNA has more inclination (η = 7.1° ± 0.5°) (Table 2.3) than B-DNA (η = 2.4°) (Dickerson, 1992), but less than A-DNA (η = 12.0°) (Dickerson, 1992).

Like helical rise, displacement is a measure of translation, but in this case, translation of the position of the base pair relative to the helix axis. Displacement of the helix axis from the short axis at the center of the base is the x-displacement (d_x), while that along the long axis is the y-displacement (d_y). Positive values of d_x reflect a translation toward the major groove and positive d_y toward the leading strand (Figure 2.3). Base pairs in B-DNA are essentially centered on the helix axis and therefore show little or no displacement; however, in Z-DNA, the helix axis is displaced by ~4 Å into the minor groove. This can be separated into average values for d_x = -3.3 Å

± 0.2 Å and d_y = 2.2 Å ± 0.3 Å respectively for d(CGCGCG) (Table 2.3). By comparison, the helix axis of A-DNA is highly displaced toward the major groove (with d_x = 4.5 Å and d_y < 0.3 Å in the fiber structure).

Propeller twist (ω) and buckle (κ) describe the distortion about the short and the long axes, respectively, from planarity of the two bases within each base pair (Figure 2.3). In this analysis we report only the magnitude of these perturbations. The average propeller twist for Z-DNA (Table 2.3) (<ω> = 2.6° ± 1.9°) is less than that of B-DNA (<ω> = 11.0°) (Dickerson, 1992) and A-DNA (<ω> = 8.3°) (Dickerson, 1992). The average buckle is similar between Z-DNA (Table 2.3) (<κ> = 3.1° ± 2.4°) and A-DNA (<κ> = 2.4°) (Dickerson, 1992), but greater than in B-DNA (<κ> = 0.2°) (Dickerson, 1992).

In summary, Z-DNA is a long, narrow double helix in which the plane of the base pairs all lie essentially perpendicular to the helix axis, with the helix axis lying in the minor groove. The alternating helical twist angles reflect the distinct difference between the d(CpG) and d(GpC) dinucleotide steps.

2.3.3 The solvent structure of d(CGCGCG)

The arrangement of water molecules around the Z-DNA duplex is very important to the structure and stability of Z-DNA. Both the deep narrow minor groove crevice and the convex major groove surface are

important sites for Z-DNA interactions with solvent and with metal complexes. The most immediately obvious site of interaction at the major groove surface is the N7 nitrogen of the guanine bases. This is the most accessible nucleophilic group of the surface, and has been shown to form covalent adducts with transition metals (*e.g.*, copper (II) (Kagawa et al., 1991), (Geierstanger et al., 1991) and platinum (II) (Parkinson et al., 1995)). Perhaps more important in terms of their effect on the stability of Z-DNA, however, are the hydrogen bonding interactions. The potential hydrogen bonding groups in Z-DNA are basically the same as those present in B-DNA, with the exception that the N3 nitrogens of the adenine and guanine bases are not normally accessible in the minor groove crevice of Z-DNA. The hydrogen bonding groups interact with water molecules, with ligands of solvated magnesium and sodium complexes ($Mg(H_2O)_6^{2+}$ and $Na(H_2O)_n^{+}$, for n = 5-7 (reviewed in (Rich et al., 1984), (Jovin et al., 1987)), with the hexaammine complexes of cobalt and ruthenium ($Co(NH_3)_6^{3+}$ (Gessner et al., 1985) and $Ru(NH_3)_6^{3+}$ (Ho et al., 1987)) and with the polyamines spermine (Bancroft et al., 1994) and spermidine (Ohishi et al., 1996). We will focus first on the water structure and then on cation interactions and their effects on Z-DNA structure.

The solvent organization at the major groove surface and minor groove crevice of Z-DNA has been extensively studied (Gessner et al., 1994) for three crystal forms of d(CGCGCG) (the forms crystallized with only magnesium, only spermine, and mixed magnesium/spermine solutions).

The features that are common to these three crystal structures likely represent the typical organization of solvent around the d(C·G) base pairs of Z-DNA.

There are two conserved patterns of water interactions observed at the major groove surface (Figure 2.4). These regular solvent motifs connect cytosines to cytosines and guanines to guanines across the strands. In the first motif, two waters bridge adjacent cytosines on opposite strands of the *anti-p-syn* steps (the d(CpG) dinucleotides). This appears to be the more stable pattern of water organization. The bridging water molecules are very well ordered, as indicated by their low temperature factors (average = 16.8 $Å^2 \pm 5.1$ $Å^2$), and fall into very well defined geometries (with average water-to-cytosine hydrogen bond distances of 2.99 Å ± 0.15 Å, water-water distances of 2.94 Å ± 0.34 Å, and angles of 92.4° ± 6.5° for cytosine-to-water-to-water). The waters are not disrupted by either magnesium or spermine in the crystal, even though hydrated magnesium complexes are located in close proximity to these adjacent cytosines.

The second motif at the major groove surface is formed by single waters that directly connect two guanine bases on the opposite strands of the *syn-p-anti* steps. These are less regular in structure than those that bridge the cytosines (with average hydrogen bond distances of 3.05 Å ± 0.36 Å between the waters and the O6 oxygen of the guanines, and guanine-water-guanine angles of 69.5° ± 5.7°). They are also readily displaced by hydrated magnesium complexes and spermine. Thus, this set of bridging

Figure 2.4

Solvent interactions with d(CGCGCG) as Z-DNA. A. Stereodiagram comparing the solvent structures at the major groove surface and the minor groove crevice of Z-DNA. The waters that interact at the major groove surface are shown on the upper duplex. Hydrogen bonds between each water are shown as solid lines, while hydrogen bonds from each water to the DNA surface are shown as dotted lines. Waters that bridge the stacked cytosines of the d(CpG) dinucleotide (through the N4 amino groups of the bases) are shown as dark spheres (labeled W_{C1} and W_{C2}), while waters that bridge the stacked guanines of the d(GpC) dinucleotide (through the O6 oxygen of the bases) are shown as open circles (labeled W_{GA}). Waters that interact with the minor groove crevice are shown in the lower duplex. Dark spheres represent the spine of hydration that links the cytosines (through interactions with the O2 oxygen of the bases), while those that link the guanine N2 amino groups to the phosphoribose backbone are shown as open circles. The solvent interactions with the d(CpG) (B) and d(GpC) (C) dinucleotide steps are shown looking down the helix axis. In addition to the labels described above, W_{Gb} represents the waters that link the guanine bases in *syn* to the phosphoribose backbone. Waters that are hydrogen bonded to cytosines are shown as dark spheres, while those to guanine are shown as light spheres.

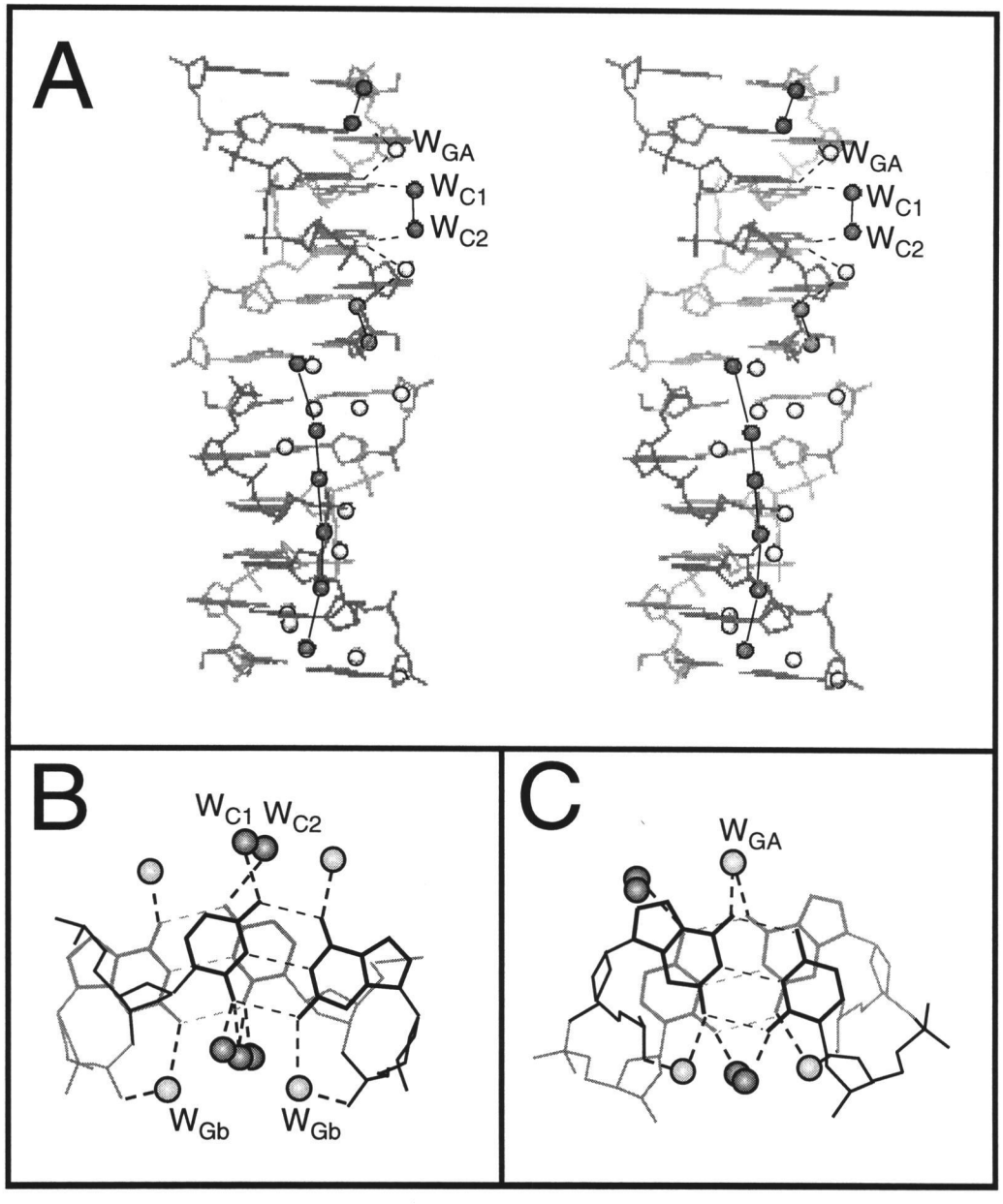


Figure 2.4

waters at the *syn-p-anti* steps are less regular and apparently less stable than those at the *anti-p-syn* steps.

The minor groove crevice is lined by a continuous network of well ordered water molecules. There are typically at least two water molecules lying in the plane of each d(C·G) base pair (Rich et al., 1984). These form hydrogen bonds to the O2 keto oxygen of the cytosine and the N2 amino group of the guanine bases. The interconnected waters bound to the O2 oxygens of the cytosine bases form a continuous network referred to as the spine of hydration. Similar spines of waters are observed in the minor grooves of B-DNA structures (Drew and Dickerson, 1981). The significance of regular networks of waters in the minor groove of DNA duplexes has previously been discussed for B-DNA (Berman, 1994). The basic conclusions from NMR studies on exchange between DNA-bound and bulk solvent were that these spines exists in solution in B-DNA (Kubinec and Wemmer, 1992; Liepinsh et al., 1992) and thus can be treated as an integral part of the DNA structure (Berman, 1994). These same concepts are likely to apply to the hydration spine in the more rigid Z-DNA structures.

The waters at the guanine bases are significant in that they bridge the N2 amino groups to the phosphate oxygens of the backbone (Figure 2.4B). This interaction may be important for stabilizing the *syn* conformation of the guanine bases. Any perturbation to the solvent interactions in the major groove surface and minor groove crevice caused by various base substituent groups will affect the stability of Z-DNA.

2.3.4 Cation effects on the structure of d(CGCGCG)

Z-DNA has been crystallized in the presence of several different types of cations including magnesium and the polyamines spermine and spermidine. The effect that cations have on the structure of Z-DNA is significant because the cations help to stabilize the left-handed structure in solution by screening, and thus shielding, the negatively charged phosphates. The phosphate-phosphate distances are closer in Z-DNA than in either A- or B-DNA because of the narrow minor groove crevice. The effect on the stability of Z-DNA is dependent on both the concentration and the charge of the cation, with higher charged ions being more effective at stabilizing this conformation. The stabilization of Z-DNA in solution follows the trend spermine^{4+} > spermidine^{3+} > Mg^{2+} > Na^{+} (Behe and Felsenfeld, 1981). In particular, the stability of different sequences as Z-DNA is dependent on the the cation strength of a solution ($CS = \sum Z_i^2[M_i]$, where Z_i is the charge and $[M_i]$ is the concentration of the cation type i). This relationship has been used as a quantitative method to predict the solutions for crystallizing different sequences as Z-DNA (Ho et al., 1991).

The polyamines have been extensively studied because they are known to aid in DNA condensation and to prevent thermal denaturation of the duplex (Morgan et al., 1986). Levels of polyamines are highly dependent on the cell cycle, and are perturbed in cancer cells [reviewed in (Tabor and Tabor, 1984)]. Therefore, polyamine binding has been of interest

not only to biologists, but also to crystallographers because of the analogy between crystallization and condensation.

Four crystal structures of d(CGCGCG) have been analyzed to determine the effect of polyamines and magnesium on the structure of Z-DNA. These structures include the magnesium only (MG) form (Gessner et al., 1989), the spermine only (SP) form (Egli et al., 1991), the mixed magnesium and spermine (MGSP) form (Wang et al., 1979), and the mixed magnesium and spermidine (MGSD) form (Ohishi et al., 1996). Although all these crystals were grown in the presence of sodium ions, it is the interactions of the multivalent cations (specifically Mg^{2+} versus the two polyamines) that will be discussed here.

The reference d(CGCGCG) structure to which we have been referring is the original MGSP form (Wang et al., 1979). The structures of the DNA in the MG, MGSP, and MGSD forms are nearly identical in all respects (Table 2.4), except for the ligand interactions. Thus, although the polyamines are more effective at stabilizing Z-DNA in solution, the crystal structures appear to be determined by the presence of magnesium. We will, therefore, treat the MG form as the reference for comparison, with the realization that the MGSP and MGSD forms are very similar to this.

The observed lattice of the SP crystal is different from that of the other d(CGCGCG) crystals, suggesting that this DNA structure is significantly different from the reference structure (Table 2.4). The DNA in the SP lattice is rotated by 70° around the helix axis, shifted by 3 Å along the helix

Table 2.4

Effect of cations on the Z-DNA structure of d(CGCGCG)

All values are in degrees, except rise (D_z) and displacement (d_x) which are in Å. MG refers to the crystal grown in the presence of Mg^{2+} only (Gessner et al., 1989), MGSP is the crystal grown with Mg^{2+} and spermine (Wang et al., 1979), MGSD is the crystal grown with Mg^{2+} and spermidine (Ohishi et al., 1996), and SP is the crystal grown with spermine only (Egli et al., 1991).

Table 2.4

	Crystal form			
	MG	**MGSP**	**MGSD**	**SP**
Twist (Ω)				
(C1·G12)/(G2·C11)	-9.2	-8.5	-9.0	-11.6
(G2·C11)/(C3·G10)	-48.9	-48.8	-48.8	-47.7
(C3·G10)/(G4·C9)	-9.4	-9.1	-8.7	-11.7
(G4·C9)/(C5·G8)	-50.8	-51.4	-51.6	-49.3
(C5·G8)/(G6·C7)	-12.2	-10.6	-11.6	-12.3
Average d(CpG)	**-10.3±1.7**	**-9.4±1.1**	**-9.8±1.6**	**-11.9±0.4**
Average d(GpC)	**-49.9±1.3**	**-50.1±1.8**	**-50.2±2.0**	**-48.5±1.1**
Rise (D_z)				
(C1·G12)/(G2·C11)	3.8	3.8	3.9	3.9
(G2·C11)/(C3·G10)	3.6	3.7	3.7	3.7
(C3·G10)/(G4·C9)	3.9	3.8	3.8	3.4
(G4·C9)/(C5·G8)	3.5	3.6	3.6	3.6
(C5·G8)/(G6·C7)	4.1	4.3	4.1	3.4
Average	**3.8±0.2**	**3.8±0.3**	**3.8±0.2**	**3.6±0.2**
Roll (ρ)				
(C1·G12)/(G2·C11)	-0.8	-3.0	-1.6	2.5
(G2·C11)/(C3·G10)	-1.5	-0.8	-0.9	-4.6
(C3·G10)/(G4·C9)	-1.1	3.6	-2.1	1.2
(G4·C9)/(C5·G8)	0.3	0.4	0.0	-1.9
(C5·G8)/(G6·C7)	3.6	-2.1	1.4	5.6
Average	**0.1±2.1**	**-0.4±2.6**	**-0.6±1.4**	**0.6±3.9**
Inclination (η)				
C1·G12	6.0	6.9	6.2	5.7
G2·C11	5.9	7.5	7.1	4.1
C3·G10	6.8	6.4	7.5	1.4
G4·C9	7.4	6.6	7.5	2.7
C5·G8	8.1	7.3	9.3	2.3
G6·C7	7.4	7.7	8.1	0.6
Average	**6.9±0.9**	**7.1±0.5**	**7.6±1.0**	**2.8±1.9**
x-displacement (d_x)				
C1·G12	-3.0	-3.3	-3.1	-4.8
G2·C11	-3.1	-3.1	-3.1	-4.8
C3·G10	-3.3	-3.1	-3.2	-3.9
G4·C9	-3.3	-3.1	-3.2	-3.4
C5·G8	-3.5	-3.5	-3.5	-3.8
G6·C7	-3.4	-3.4	-3.4	-4.2
Average	**-3.3±0.2**	**-3.3±0.2**	**-3.3±0.2**	**-4.2±0.6**

axis, and rotated around the intramolecular pseudo-2-fold axis as compared to the DNA in the MG and MGSP crystals (Egli et al., 1991). The most obvious difference between the Z-DNA structure of the SP form is the shorter $<D_Z>$ (3.6 Å) and larger $<d_X>$ (-4.2 Å) compared to the MG structure (3.8 Å and -3.3 Å respectively) (Table 2.4). As a result, the SP structure of Z-DNA is shorter and wider than the reference conformation. This is likely due to the binding of the spermine to the major and minor grooves. The $<\Omega>$ shows that the d(CpG) steps are overwound (by ~2°) while the d(GpC) steps are underwound (by ~1.0 °) in the SP structure relative to the reference MG structure (Table 2.4). The compensating under- and overwinding of the dinucleotide steps renders the overall Ω of the structure identical to that of the other Z-DNA structures. Finally, the SP form shows a slight increase in roll and a dramatic decrease in the inclination of the base pairs (Table 2.4) as compared to the reference structure.

In order to understand the effect of the cations on Z-DNA stability, we must first characterize the specific interactions that the cations and their ligands make with the DNA. Starting with the reference MG form, there are four unique hydrated magnesium clusters that were observed to bind the DNA duplex (Figure 2.5). This does not entirely neutralize the net -10 charge of the phosphoribose backbone in d(CGCGCG), requiring either one additional magnesium or two sodium ions that cannot be observed in the crystal structure. The observed ions, however, represent the specific

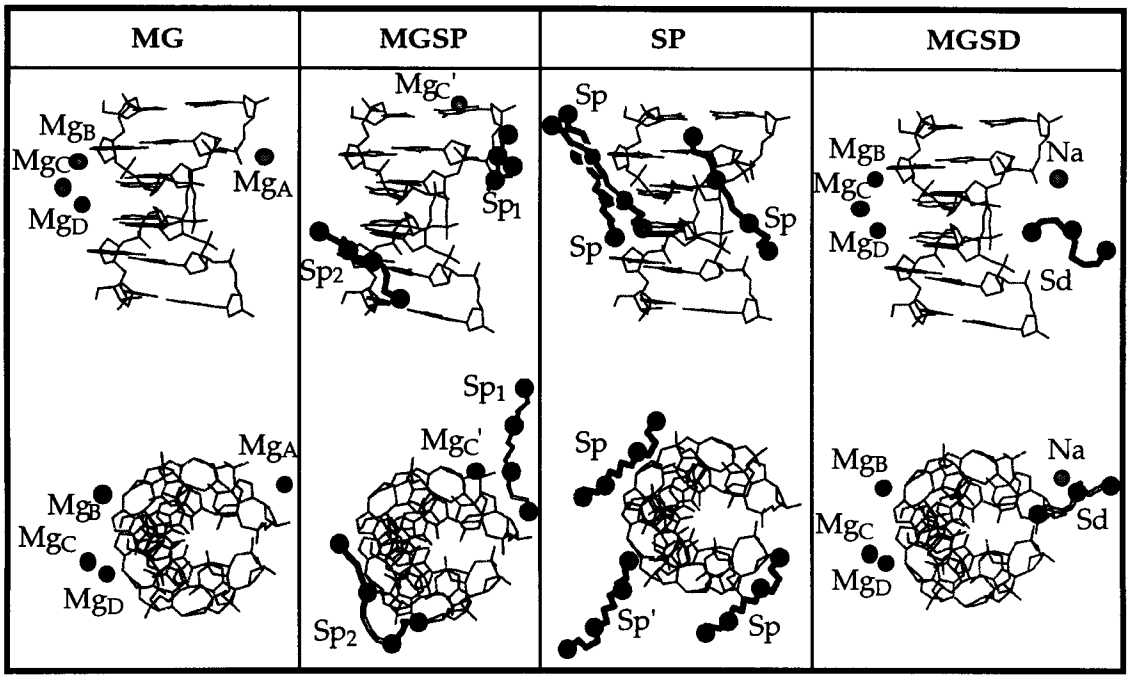


Figure 2.5

Comparison of the cation interactions between the magnesium only (MG), mixed magnesium/spermine (MGSP), spermine only (SP), and mixed magnesium/spermidine (MGSD) forms of d(CGCGCG). Shown are views perpendicular to (top) and down the helix axes (bottom) of each structure. In the structures of the polyamines (spermine and spermidine), the nitrogen atoms are shown as spheres. In the MG form of the structure, each unique magnesium ion (waters not shown) are labeled as Mg_A, Mg_B, Mg_C, and Mg_D. The two unique spermine molecules of the MGSP form are labeled as Sp_1 and Sp_2, while the single magnesium (which is symmetry related to Mg_C of the MG form) is labeled Mg_C'. Although there is only one unique spermine in the SP form, it makes three different interactions with each duplex. These three types of interactions are shown. Finally, the single unique spermidine (Sd), the three magnesiums (identical to Mg_B, Mg_C, and Mg_D of the MG form), and the cation identified as a sodium (labeled as Na, but is similar in position to Mg_A of the MG form) are shown for the MGSD form of d(CGCGCG).

interactions. The hexahydrated Mg_A (Table 2.5) makes six hydrogen bonding contacts with the DNA. It interacts with a phosphate oxygen of G8, C9, G10 and of the G6 of a neighboring duplex as well as with the N2 of G8 and the O4' of the G6 of the neighboring duplex. Mg_B (Table 2.5) is also hexahydrated and makes contacts with a phosphate oxygen of G6, G10 and C11, although the contact with the C11 oxygen is mediated by a water molecule. Three contacts are made with neighboring duplexes. These are with the O6 of G4, the phosphate oxygen of C5 and the N7 of G8. Mg^{2+} complexes Mg_C and Mg_D (Table 2.5) are linked together and share two water ligands. One of these ligands binds to N4 of C1 (a neighboring duplex) and additional water molecules mediate contacts with the N4 of C9 and the O6 of G6 on a neighboring duplex. Additionally, an Mg_C ligand makes a contact with the N7 of the same G6. Additional contacts of Mg_C ligands include water-mediated interactions with the phosphate oxygen of C5 the N4 of C1 in a neighboring duplex. Mg_D has additional interactions with the DNA, specifically with the O6 of G10 and a water-mediated contact with the C9 phosphate oxygen. It also makes contacts with two other duplexes, namely the O6 of G12 of one duplex and the O6 of G6 of another duplex. Thus, these Mg^{2+} complexes not only provide intramolecular stabilization of Z-DNA, but also stabilize the crystal through intermolecular interactions.

Table 2.5

Hydrogen bonding contacts of the four unique magnesium ions in the MG form of d(CGCGCG)

Cation	Residue	Atom
Mg_A	G6 (s)	PO, O4'
	G8	PO, N2
	C9	PO
	G10	PO
Mg_B	G4 (s)	O6
	C5 (s)	PO
	G6	PO
	G8 (s)	N7
	G10	PO
	C11	PO (w)
Mg_C	C1 (s)	N4 (w)
	C5	PO (w)
	G6 (s)	N7, O6 (w)
	C9 (s)	N4, PO (w)
		Mg_D (w)
Mg_D	C1(s)	N4
	G6	O6
	C9 (s)	N4, PO
	G10 (s)	O6, N7 (w)
	G12 (s)	O6
		Mg_C (w)

An interaction of the ion with adjacent, symmetry related residues is indicated by the designation (s), while (w) indicates that this contact is mediated through a coordinating water molecule.

In the MGSD form, there is a single unique spermidine per duplex that displaces Mg_A and the remaining three divalent cations are unperturbed (Figure 2.5). The spermidine itself interacts with the phosphoribose backbone of two adjacent duplexes in the crystal lattice. These are all mediated by water bridges between the amino nitrogens of the ligand and the oxygens of the phosphates. Specifically, these interactions (Table 2.6) are with phosphate oxygens of C3, G6 and G12. The C3 interaction is with a neighboring duplex. The interaction with G6 is equivalent to a contact made by Mg_A in the other structures. Interestingly, the amino groups of a truncated analogue (*N*-(2-amino-ethyl)-1,4-diamino-butane) of spermidine (spermindine is *N*-(2-amino-propyl)-1,4-diamino-butane) bind directly to the phosphates, and show direct interactions with the bases at the major groove surface (Ohishi et al., 1996).

Similarly, in the mixed MGSP form, one of the original Mg^{2+} clusters remain in place, but in this case the complexes Mg_A, Mg_B and Mg_D are displaced by the polyamines (Figure 2.5). There are two spermines per duplex in the asymmetric unit, each interacting with three DNA duplexes. $Spermine_1$ makes two contacts with the DNA (Table 2.6): one with the N7 of G4 and the other with the O6 of G8. The remainder of the interactions are with other duplexes. These interactions are with the phosphate oxygens of C5, G6 and C9 and the O4' of G6. There are also water-mediated contacts to phosphate oxygens G6, G8 and G10. The amino groups of $spermine_2$ also

Table 2.6

Hydrogen bonding contacts of polyamines and magnesium with the MGSP, MGSD, and SP forms of d(CGCGCG)[a]

[a] Sp and Sd denote the polyamines spermine^{4+} and spermidine^{3+}, respectively. PO denotes a phosphate oxygen. An interaction of the ion with an adjacent, symmetry related residue is indicated by the designation (s), while (w) indicates that this contact is mediated through a coordinating water molecule. "MG contact equivalent" refers to the analogous Mg^{2+} complex of the MG structure.
[b] MGSP refers to the crystal grown in the presence of Mg^{2+} and spermine.
[c] MGSD is the crystal structure of Z-DNA with Mg^{2+} and spermidine.
[d] SP is the structure that was crystallized only with spermine.

Table 2.6

Structure	Cation	Residue	Atom	MG contact equivalent
MGSP[b]	Mg_C	G6	N7	C (s)
	Sp_1[a]	G4	N7	-
		C5 (s)	PO	B, C
		G6 (s)	PO, PO (w),O4'	A,B
		G8	O6	-
		G8 (s)	PO (w)	A
		C9 (s)	PO	A,C,D
		G10 (s)	PO (w)	A,B
	Sp_2[a]	C1 (s)	5' OH	-
		G2	N7	-
		G2 (s)	N7, O6 (w)	-
		C3	O6, N4 (w)	-
		G10	O6, N7	D
		C11	N4	-
		C11 (s)	PO	-
		G12	O6	D
MGSD[c]	Mg_B			B
	Mg_C			C
	Mg_D			D
	Sd[a]	C3 (s)	PO (w)	-
		G6	PO (w)	A
		G12	PO (w)	-
SP[d]	Sp[a]	C3	PO	-
		G8	N7	B
		C9 (s)	PO	A,C,D
		G10 (s)	N7, O6	D
		C11(s)	PO	B
		G12 (s)	PO	-

make numerous contacts with the DNA and neighboring duplexes (Table 2.6). Direct interactions include those with the N7 of G2, the O6 of G10 and G12, and the N4 of C3 and C11 (although the interaction with C3 is water mediated). Interactions with other duplexes include the 5′ OH of C1, the phosphate oxygen of C11, the N7 of G2 and a water-mediated contact with the O6 of G2. The interactions of spermine with the DNA are similar to those observed in the MG structure and in fact many of the spermine contacts are equivalent to those seen with Mg^{2+} in the MG form (Table 2.6).

In the SP crystal, there is one spermine per duplex. Each spermine interacts with three different DNA molecules, and 3 spermines interact with each DNA molecule (Figure 2.5). This large number of interactions between the polyamine and the DNA is consistent with spermine's ability to condense DNA (Egli et al., 1991). The spermines interact with the DNA as follows (Figure 2.5): one binds in the major groove of the DNA, the second interacts with the phosphates along the minor groove and the third interacts with only the C9 and G10 of the DNA. Direct contacts are made between the phosphate oxygen of C3 and the N7 of G8. Interactions with neighboring duplexes include hydrogen bonds with the phosphate oxygens of C9, C11 and G12 as well as interactions with the N7 and O6 of G10 (Table 2.6). These interactions are common with the interactions observed between all four Mg^{2+} and the DNA in the MG structure (Table 2.6).

However, the interactions between spermine and the DNA duplex in the SP form versus the mixed cation MGSP form are not identical.

In summary, the cations make similar contacts with the DNA across the different structures, and display both intra- and interduplex interactions which often involve the coordination of bridging water molecules. When comparing the concentration of cations required to crystallize each of these forms of d(CGCGCG), it became evident that spermine had twice the effect expected relative to other cations. We therefore simply increased the effective *CS* for spermine on Z-DNA crystallization by a factor of 2 (this has already been incorporated into the *CS* values in Table 2.2). This is an empirical observation, but may be related to the base specific interactions of this polyamine with Z-DNA.

The binding of spermine to supercoil induced Z-DNA in closed circular DNA plasmids (Howell et al., 1996) appears to be consistent with that observed in the crystal. The association constant of spermine for Z-DNA (1.5×10^7 M^{-1} for d(CpG) and 1.2×10^8 M^{-1} for d(CpA/TpG) dinucleotides) is 100-fold greater than that for B-DNA (1.4×10^5 M^{-1}), consistent with the stabilizing effect that this polyamine has on the left-handed conformation. The size of the spermine binding site for Z-DNA was determined to be 10.4 d(C·G) base pairs. This is larger than that observed in the crystal structure of the SP form of d(CGCGCG) (1 spermine per duplex, or 6 bp/spermine). However, the crystal structure may exaggerate the number of ligands actually bound to the DNA. The

temperature factors for the spermines are about twice that observed for the DNA, even at -100°C. This suggests that the spermine is not fully occupied and, therefore, the number of ligands bound per duplex is likely to be significantly less than 1. This would give an overall binding size that is more consistent with the results from the solution studies.

2.3.5 Length effects on the structure of d(CpG) sequences as Z-DNA

The structures of alternating d(CpG) dinucleotides as Z-DNA have been determined for five different lengths of duplexes, from a single dinucleotide in the structure of d(CpG) to five dinucleotides in d(GCGCGCGCGC) (Table 2.2). A comparison of lengths shorter than that of the hexamer, d(CGCGCG), allows us to determine whether the conformation of the *anti-p-syn* stacking in d(CpG) is inherent to this dinucleotide in the absence of significant flanking base pairs. Comparisons of longer sequences address the questions of whether the structure of d(CGCGCG), or any hexanucleotide, can indeed be extrapolated to longer and even infinite lengths of Z-DNA, and whether the *anti-p-syn* dinucleotide of d(CpG) is the stable repeating unit of Z-DNA.

The overall conformations of the structures in this comparison are all very similar to one another and to d(CGCGCG), with just a few exceptions. One interesting feature that is common to all structures is the presence of the Z_{II} backbone conformation. The crystal lattice interactions

that are associated with this perturbation in the reference d(CGCGCG) structure are not identical across this set of structures. It is unclear then what specific crystal lattice interactions are directly responsible for this backbone conformation.

Within this set of nine $d(CpG)_n$ Z-DNA sequences (Table 2.2), the heptamers d(GCGCGCG)/d(CCGCGCG) and d(GCGCGCG)/d(TCGCGCG) do not have blunt ends (Mooers et al., 1997). These structures are essentially that of d(CGCGCG) as Z-DNA, with nucleotides dangling from the 5'-ends of each strand. These orphaned nucleotides pair between adjacent duplexes to form reverse Watson-Crick d(G·C) and reverse wobble d(G·T) base pairs that are sandwiched between two stacked d(CGCGCG) Z-DNA duplexes. These can therefore be treated as variations on the reference d(CGCGCG) structure in which the Z-DNA pattern is disrupted at the ends, serving as a true indicator of end effects. In this case, the overall structure of the Z-DNA duplexed region is remarkably similar to that of the reference d(CGCGCG) in all respects (Table 2.7). The terminal base pairs of the duplex region (C1·G12 and G6·C7, where the nucleotides are numbered according to the duplex Z-DNA regions only, ignoring the 5′ overhangs) show a larger buckle than found in any of the d(C·G) base pairs of d(CGCGCG), but otherwise all helical parameters are reproduced, including the average helical twist at each dinucleotide step and even the C3'-*endo* sugar pucker of the 3'-terminal guanine that breaks the alternating sugar conformation along each strand. The notable exceptions are the shorter rise and greater

Table 2.7

Helical base step and base pair parameters of $d(CG)_n$ sequences that crystallize as Z-DNA[a]

	d(CG)	d(CGCG)[b]	d(CCGCGG)	d(CGCGCG) MGSP	d(GCGCGCG)/ d(TCGCGCG)	d(GCGCGCG)/ d(CCGCGCG)	d(GCGCGCGCGC)[c]
d(CpG) steps							
Twist (Ω)							
(C1•G12)/(G2•C11)	-7.4	-12.3	-7.8	-8.5	-7.4	-6.5	-9.7
(C3•G10)/(G4•C9)		-13.6	-9.2	-9.1	-10.3	-10.5	
(C5•G8)/(G6•C7)				-10.6	-10.6	-9.8	
Average	**-7.4**	**-13.0±0.9**	**-8.5±1.0**	**-9.4±1.1**	**-9.4±1.8**	**-8.9±2.1**	**-9.7**
Rise (D_z)							
(C1•G12)/(G2•C11)	3.2	3.8	3.8	3.8	3.4	3.4	3.9
(C3•G10)/(G4•C9)		3.7	3.8	3.8	4.1	3.9	
(C5•G8)/(G6•C7)				4.3	3.8	3.9	
Average	**3.2**	**3.8±0.1**	**3.8**	**4.0±0.3**	**3.8±0.4**	**3.7±0.3**	**3.9**
Roll (ρ)							
(C1•G12)/(G2•C11)	-13.7	-7.1	-9.0	-3	0.6	0.9	3.2
(C3•G10)/(G4•C9)		-0.7	-4.0	3.6	4.7	-3.7	
(C5•G8)/(G6•C7)				-2.1	2.2	2	
Average[d]	**13.7**	**3.9±4.5**	**6.5±3.5**	**2.9±0.8**	**2.5±2.1**	**2.2±1.4**	**3.2**

Table 2.7 continued

Tilt (τ)							
(C1•G12)/(G2•C11)	-5.8	-10.1	-8.4	6.9	-5.8	-4.4	0.0
(C3•G10)/(G4•C9)		-3.4	0.7	1.1	0.2	0.6	
(C5•G8)/(G6•C7)				0.7	-2.7	-0.1	
Average[d]	**5.8**	**6.8±4.7**	**4.6±5.4**	**2.9±3.5**	**2.9±2.8**	**1.7±2.4**	**0.0**
d(GpC) steps							
Twist (Ω)							
(G2•C11)/(C3•G10)		-45.8	-44.0	-48.8	-47.5	-47.5	-50.3
(G4•C9)/(C5•G8)				-51.4	-47.1	-48	
Average		**-45.8**	**-44.0**	**-50.1±1.8**	**-47.3±0.3**	**-47.8±0.4**	**-50.3**
Rise (D_z)							
(G2•C11)/(C3•G10)		3.7	3.7	3.7	3.6	3.7	3.2
(G4•C9)/(C5•G8)				3.6	3.7	3.6	
Average		**3.7**	**3.7**	**3.7±0.1**	**3.7±0.1**	**3.7±0.1**	**3.2**
Roll (ρ)							
(G2•C11)/(C3•G10)		-4.2	-6.1	-0.8	-1.8	-1.8	-3.2
(G4•C9)/(C5•G8)				0.4	-3.7	0.1	
Average[d]		**4.2**	**-6.1**	**0.6±0.8**	**2.8±1.3**	**0.9±1.3**	**3.2**
Tilt (τ)							
(G2•C11)/(C3•G10)		4.3	-1.5	-0.6	1.9	1.1	0.0
(G4•C9)/(C5•G8)				0.2	1	2.6	
Average[d]		**4.3**	**-1.5**	**0.4±0.6**	**1.5±0.6**	**1.9±1.1**	**0.0**

Table 2.7 continued

Base Pairs							
Tip (θ)							
C1•G12	13.7	7.1	9.0	3	-0.6	-0.9	-1.6
G2•C11	1.7	2.8	3.0	2.1	1.2	0.9	1.6
C3•G10		2.1	-1.0	-1.5	-3.6	-2.8	
G4•C9		-7.4	-6.2	-1.1	0.1	-2.9	
C5•G8				1	-2.1	-0.9	
G6•C7				0.9	4.6	5.5	
Average[d]	**7.7±8.5**	**4.9±2.8**	**4.8±3.5**	**1.6±0.8**	**2.0±1.8**	**2.3±1.8**	**1.6**
Inclination (η)							
C1•G12	5.8	10.1	8.4	6.9	5.8	4.4	-2.7
G2•C11	2.2	5.9	7.0	7.5	3.8	3.2	-2.7
C3•G10		9.2	6.2	6.4	4	3.8	
G4•C9		4.9	7.4	6.6	5	6.4	
C5•G8				7.3	7.7	6.4	
G6•C7				7.7	5.4	2.6	
Average[d]	**4±2.5**	**7.5±2.5**	**7.3±0.9**	**7.1±0.5**	**5.3±1.4**	**4.5±1.6**	**2.7**
Propeller Twist (ω)							
C1•G12	10.6	1.9	3.2	0.8	-0.7	6.3	1.5
G2•C11	-15.7	2	-7.7	2.1	2.2	0.1	1.5
C3•G10		3.6	3.7	5.6	-0.9	-0.7	
G4•C9		4.2	0.9	3.4	1.2	-0.9	
C5•G8				0.6	-3.6	2.2	
G6•C7				3.2	5.4	0.1	
Average[d]	**13.1±3.6**	**2.9±1.2**	**3.9±2.8**	**2.6±1.9**	**2.3±1.8**	**1.7±2.4**	**1.5**

Table 2.7 continued

Buckle (κ)							
C1•G12	4.6	8.5	2.2	0.3	10.4	10	-10.5
G2•C11	1.2	-2.8	-6.6	-4.8	-5.9	-3.5	10.5
C3•G10		8.9	4.9	2.8	0.1	0	
G4•C9		-2	-1.5	-5.9	-0.7	-2.1	
C5•G8				0.1	8.2	3.8	
G6•C7				4.4	-8.3	-1.3	
Average[d]	**2.9±2.4**	**5.6±3.7**	**3.8±2.4**	**3.1±2.4**	**5.6±4.3**	**3.5±3.5**	**10.5**

[a] Underlined sequences denote the d(CpG) dinucleotides in the standard Z-DNA duplex. The numbering of residues refer only to those nucleotides in the duplex. All values are in degrees, except rise (D_z) and displacement (d_x) which are in Å. MGSP refers to the form of d(CGCGCG) crystallized in the presence of magnesium and spermine.

[b] d(CGCG) refers to the high-salt orthorhombic form (Drew et al., 1980).

[c] Only one value for each type of base step in the d(GCGCGCGCGC) decamer is shown. These values are repeated throughout the decamer because of the dinucleotide asymmetric unit.

[d] Averages of the magnitudes of roll, tilt, tip, inclination, propeller twist, and buckle are shown, and were calculated by Ave = $(\Sigma|q_i|) / i$, where q_i is the value of that parameter at base step *i*.

buckle in the corresponding terminal d(C·G) base pairs in the heptamer structures. Thus, these appear to be true end distortions associated with the lack of a Z-DNA-like d(GpC) step between duplexes in the crystal lattice.

Unlike d(CGCGCG), the short duplexes of d(CG) and d(CGCG) do not stack end-to-end to form pseudo-continuous helices. There are no internal d(CpG) steps in either structure, and only a single internal d(GpC) step in the tetramer structure. We would expect therefore that these structures are essentially "all ends". The dimer has a twist, rise and tilt comparable to the first d(CpG) step of the heptamer, making it less left-handed and shorter than comparable steps in d(CGCGCG) (Table 2.7). Thus the structure of d(CpG) is that of unconstrained Z-DNA ends. There are additional distortions to the dimer, such as the significant roll and propeller twist between and within the base pairs. These, however, may be related to the unusual ammonium cation present in this crystal that is not present in other Z-DNA structures.

The d(CpG) steps in the tetramer structure of d(CGCG) are the most overwound of all the structures, with $<\Omega> = -13°$. The single d(GpC) step, however, is significantly underwound, compensating for the overwound d(CpG) and resulting in an $< \Omega > = -29.4°$ per base step that is almost identical to that of the d(CGCGCG) structures (Table 2.7). It should be noted that only the high salt, orthorhombic form of d(CGCG) (Drew and Dickerson , 1981) was available for this comparison. However, the structure of

d(CCGCGG) has the 5'-terminal cytosine nucleotide flipped out to an extrahelical conformation and, thus, can be treated as a tetramer of four central Z-DNA base pairs, analogous to the treatment of the heptamers as six Z-DNA duplex base pairs with unusual ends. With respect to Ω, the tetramer within d(CCGCGG) is more similar to d(CGCGCG), particularly the MGSP form, than to that of d(CGCG). The similarity between the tetramer structures lies in the high negative roll of both the d(CpG) and d(GpC) base steps, a high negative tilt in the d(CpG) steps, large variations in the tip and inclination at each base pair, and large variations in propeller twist and buckle within each base pair (Table 2.7). These distortions are evidently associated with this short length of the duplex and, again, may reflect the structure of Z-DNA ends, as opposed to internal dinucleotides that one would expect in longer sequences.

There are a number of octanucleotide Z-DNA structures that have been determined, including that of d(CGCGCGCG), but they are all in disordered lattices. The only reliable parameter that we can determine from this structure is the average helical rise per base pair (3.6 Å/bp), which was calculated from the length of the helical axis (the crystallographic *c*-axis) of 43.6 Å for six base pairs (Fujii et al., 1985). This is shorter than the average for the alternating d(CG) tetramer and hexamer sequences.

The longest Z-DNA duplex crystal structure solved to date is that of the decamer d(GCGCGCGCGC). This sequence is unusual in that it starts

with a guanine nucleotide and thus there are more d(GpC) steps (5) than d(CpG) steps (4). Shorter alternating d(GpC) sequences that have been solved crystallographically (d(Gm^{5}CGCGC) and d(Gm^{5}CGm^{5}CGCGC)) were in the A-form (Mooers et al., 1995). The unmethylated versions of these hexamer and octamer sequences crystallize, but are highly disordered, with the octamer showing a strong Bragg reflection at 3.4 Å resolution suggesting that it is likely in the B-form. Thus, it appears that Z-DNA is not the preferred form in alternating $d(GC)_n$ sequences until $n \geq 5$ dinucleotides. This is consistent with the solution studies of Quadrafoglio, et al. (1984) which showed that oligonucleotides of $d(GC)_n$ are left-handed only in longer sequences ($n > 7$ dinucleotides), while shorter sequences ($3 < n < 7$ dinucleotides) remain right-handed even under dehydrating conditions. Thus it does appear that the d(CpG) step is the significant determinant for Z-DNA formation, and in oligonucleotides where the d(GpC) steps would be dominant, the left-handed conformation is not stable. In longer sequences, the number of destabilizing d(GpC) and stabilizing d(CpG) dinucleotides become equalized, allowing Z-DNA to form.

Unfortunately, the structure of d(GCGCGCGCGC) shows positional disorder and therefore end-effects could not be distinguished from the remainder of the structure. Still, the average values for the helical parameters can be compared to the averages for the shorter DNA lengths. For the most part, this decamer is very similar to the hexamers. The most interesting deviation is that the rise at the d(GpC) step is significantly

shorter (by ~0.5 Å) than in the hexamer or tetramer structures, to give an $<D_Z>$ = 3.6 Å (Table 2.7). When compared to the $<D_Z>$ of the octamer structure, which is intermediate between the shorter (tetramer and hexamer sequences) and this longer sequence, the compressed rise appears to be length dependent and suggests that the shorter sequences have an elongated d(GpC) step. Alternatively, the shorter $<D_Z>$ of the octa- and decanucleotide structures may be related to the crystal lattice since both are in disordered hexagonal space groups. In support of this, the disordered d(CGCG) tetramers crystallize in hexagonal space groups and show shorter rises (3.61 to 3.67 Å) when determined from the lengths of the helix axes (Crawford et al., 1980).

When comparing all these lengths of alternating d(CG) sequences as Z-DNA, the crystal structures of the hexanucleotides appear to indeed be a reasonable model for long, and perhaps even infinite lengths of Z-DNA. The helical twist for Z-DNA is very consistent at ~-30° per base step for all lengths. The rise at the Z-DNA stabilizing d(CpG) steps is ~3.8 Å, but may be slightly exaggerated in the shorter sequences at the d(GpC) step. The base pairs are all nearly perpendicular to the helix axis, with very little distortion to the base pair plane (as would be expected for this rigid structure). It is also clear that the base pairs at the ends of a Z-DNA stretch (as typified by the dimer and tetramer structures, and the ends of the heptamer structures) are more variable in structure.

2.4 Sequence and substituent effects on the structure and stability of Z-DNA

The tendency of dinucleotides to form Z-DNA is as follows: $d(m^5CpG) > d(CpG) > d(CpA)/d(TpG) > d(TpA)$ (Rich et al., 1984). In order to understand the structural basis behind these trends, DNA containing many different sequences and base modifications have been crystallized as Z-DNA. We will focus the discussion here on base modifications that both stabilize and destabilize the Z-conformation in terms of the substituent groups that are added, deleted, or replaced in the standard bases of cytosine, thymine, guanine and adenine (Figure 2.6).

In this analysis of sequences that have been crystallized as Z-DNA, we compare hexanucleotide sequences that have been crystallized in the same crystal lattice to determine which structural features are sequence versus crystal packing effects. The impact that sequence has on the stability of Z-DNA will be addressed by considering two measures of its stability relative to B-DNA. These are the solvent free energies (SFEs) and the cationic strength (*CS*) of the crystallization solution. SFEs are estimated from the solvent accessible surfaces (SAS) calculated for the DNA molecule and, therefore, reflect the energy associated with the DNA in an aqueous environment. The difference in SFE for a sequence in the Z-form versus the B form ($\Delta SFE_{Z\text{-}B}$) is indicative of the sequence's stability as Z-DNA (Kagawa et al., 1989). The other measure of Z-DNA stability that is relevant to the sequences in a single crystal relies on the recognition that the amount

Figure 2.6

Definitions and structures of variations in d(C·G) (top) and d(T·A) (bottom) type base pairs. Substituents at the C5 carbon of the pyrimidine bases are labeled R_A, while those at the C2 carbon of the purine bases are R_B. In the d(C·G) type base pairs, substituents at the C5 carbon of the cytosine base form 5-methylcytosine and 5-bromocytosine. Replacement of the amino group at the C2 carbon of guanine forms the unusual inosine nucleotide. In the d(T·A) type base pairs, removing the methyl group at the C5 carbon of thymine forms the unusual deoxynucleotide uridine, while adding an amino group to the C2 carbon of adenine forms the unusual 2-aminoadenine (or diaminoadenine) nucleotide.

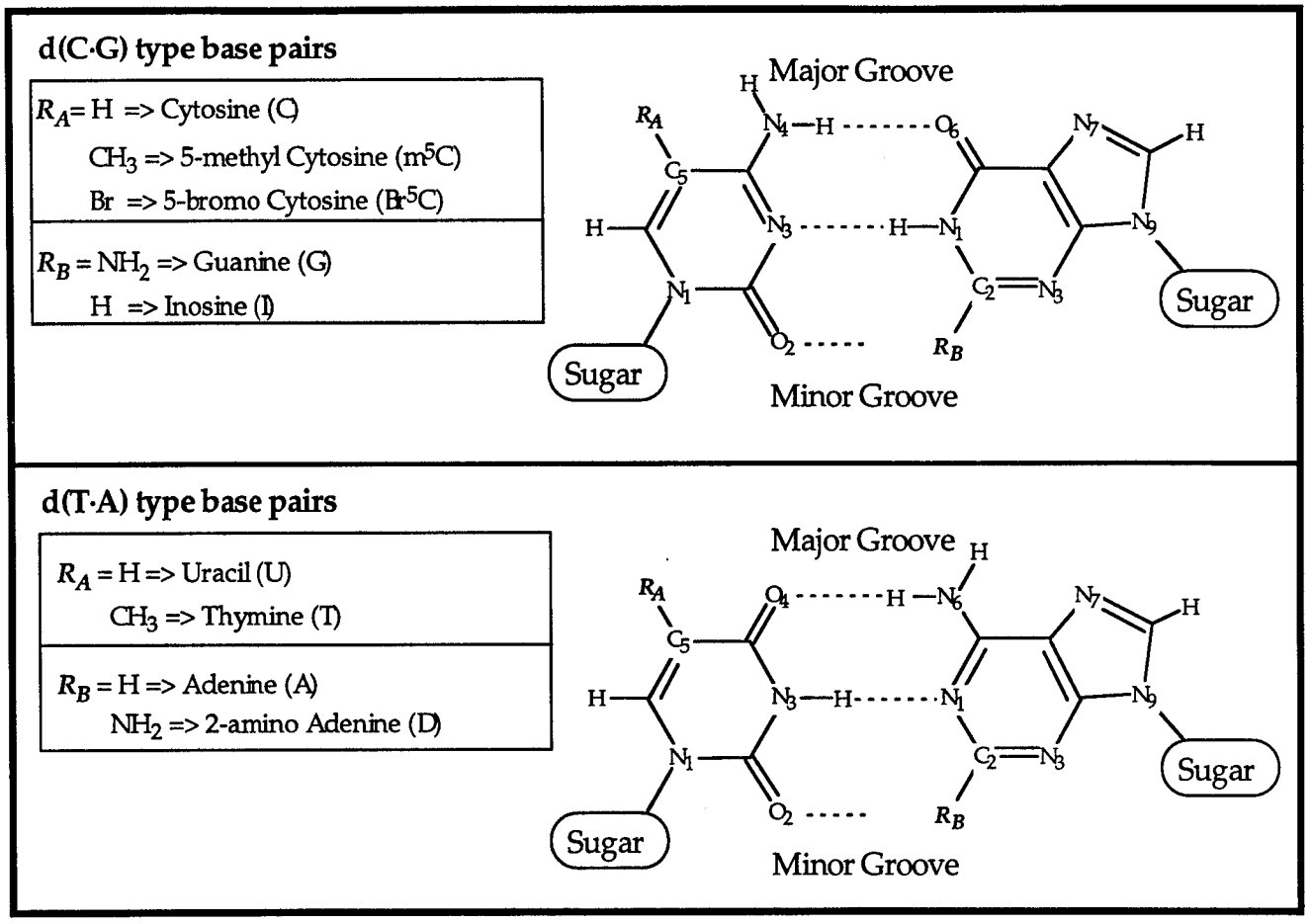

d(C·G) type base pairs
R_A= H => Cytosine (C)
CH_3 => 5-methyl Cytosine (m^5C)
Br => 5-bromo Cytosine (Br^5C)
R_B = NH_2 => Guanine (G)
H => Inosine (I)
Major Groove
Sugar
Sugar
Minor Groove
d(T·A) type base pairs
R_A = H => Uracil (U)
CH_3 => Thymine (T)
R_B = H => Adenine (A)
NH_2 => 2-amino Adenine (D)
Major Groove
Sugar
Sugar
Minor Groove

Figure 2.6

of salt required to convert a sequence to Z-DNA from B-DNA depends on the sequence's inherent stability as Z-DNA relative to B-DNA. Indeed, the quantity of salt (particularly the cations, as defined as the cation strength, or *CS*) required to crystallize sequences as Z-DNA was observed to be related to the relative stability of that sequence as Z-DNA (as estimated from ΔSFE_{Z-B}), and can be used to quantitatively predict the conditions for obtaining these crystals (Ho et al., 1991). This relationship can be attributed to the requirement that sequences undergo a transition from B- to Z-DNA along the crystallization pathway. Thus, in this analysis of sequence effects on DNA structure and stability, we will focus on the effect that various substituent groups have on the structure (both the DNA conformation and the solvent structure), ΔSFE_{Z-B} and the crystallization conditions for the various sequences that have been crystallized as Z-DNA.

2.4.1 Effects of cytosine methylation on Z-DNA structure

Methylation of cytosine at the C5 carbon of the base (m^5C) (Figure 2.6) has been studied extensively because of its effect on DNA transcription (Futscher et al., 199X). The effect that methylation has on Z-DNA has been studied both in solution (Behe and Felsenfeld, 1981) and in various crystal structures. These studies have shown that methylation stabilizes the Z-DNA conformation relative to the B-form. Using circular dichrosim spectroscopy to monitor salt and alcohol titrations, Behe and Felsenfeld

(1981) showed that poly[d(m^5CpG)] requires less salt or alcohol to convert to Z-DNA than the unmethylated poly[d(CpG)]. This stabilization is associated with the effect that the methyl group has on the hydrophobicity of Z- and B-DNA, as reflected in the ability of cations of the Hofmeister series to induce Z-DNA in poly[d(CpG)] and poly[d(m^5CpG)] (Kagawa et al., 1993). In the Hofmeister series, the cations would be expected to follow the trend $Mg^{2+}>Li^{+}>Na^{+}>K^{+}>NH_4^{+}$ in affecting the transition if indeed the hydrophobic effect is significant (Melander and Horvath, 1977).

The crystallization of methylated and unmethylated d(CGCGCG) reflects the stabilizing effect of cytosine methylation on Z-DNA. Methylated sequences require less salt to crystallize than unmethylated sequences. The sequence d(CGCGCG) was crystallized from a solution with a *CS* = 2.0 M whereas the methylated sequence d(m^5Cm^5CGm^5CG) required *CS* = 0.57 M cations (Table 2.2).

The crystal structure of d(m^5CGm^5CGm^5CG) (Fujii et al., 1982) showed that the methyl groups reside in protected and recessed pockets at the major groove surface formed by the base and sugar of the adjacent guanine nucleotide. Thus, by burying the methyl into a hydrophobic pocket, this group forms a hydrophobic patch that is less accessible to solvent in Z-DNA versus B-DNA. In addition, the methyl group is involved in favorable contacts with the base and sugar (Fujii et al., 1982).

The methyl group, however, should not be viewed as simply a substituent added to the d(CGCGCG) structure. It also affects the structure of Z-DNA, as is evident from the analysis of sequences having different degrees of cytosine methylation. In this analysis, we compare the sequence d(CGCGCG), which contains the fully unmethylated d(C·G) base pairs, to the sequences d(m^5CGm^5CGm^5CG), and d(CGCGm^5CG), in which the d(C·G) base pairs are fully and hemimethylated. Each d(CpG) and d(GpC) dinucleotide step was analyzed for the helical twist, rise, roll, tilt, propeller twist, buckle and x-displacement (Table 2.8).

The most significant effect of methylation on d(CpG) steps is on the helical twist (Table 2.8). In its fully unmethylated form, $\langle\Omega\rangle = -10.3° \pm 1.7°$ whereas the fully methylated dinucleotide, d(m^5CpG/m^5CpG), is overwound by ~5° with an $\langle \Omega \rangle = -15.0° \pm 1.0°$. This has been attributed to unfavorable steric contacts between the methyl group and the C2' carbon of the deoxyribose of the neighboring guanine (Fujii et al., 1982). In the d(GpC) steps, methylation affects both the twist and the roll independent of the dinucleotide's location in the sequence. In the unmethylated d(GpC/GpC) dinucleotide, $\langle \Omega \rangle = -49.9° \pm 1.3$, whereas the fully methylated d(Gpm^5C/Gpm^5C) is underwound by 5.8° with $\langle \Omega \rangle = -44.1° \pm 0.6$. Thus, there is a compensating over- and underwinding of the d(CpG) and d(GpC) steps so that the overall structure of methylated Z-DNA remains relatively unperturbed ($\langle\Omega\rangle = -30.3°$ per base step in the unmethylated and

Table 2.8

Comparisons of helical parameters for modified d(CpG) dinucleotides d(m^5CpG), d(Br^5CpG), and d(CpI)

	d(CGCGCG)	d(m^5CGm^5CGm^5CG)	d(CGCGm^5CG)	d(Br^5CGBr^5CGBr^5CG)	d(CGCGBr^5CG)	d(CGCICG)[a]
d(CpG) steps						
Twist (Ω)						
(C1·G12)/(G2·C11)	-9.2	-14.4	-10.8	-14.5	-8.9	-11.8
(C3·G10)/(G4·C9)	-9.4	-14.5	-12.1	-11.8	-11.6	-11.9
(C5·G8)/(G6·C7)	-12.2	-16.1	-14.6	-14.8	-13.9	-12.3
Average	**-10.3±1.7**	**-15.0±1.0**	**-12.5±1.9**	**-13.7±1.7**	**-11.5±2.5**	**-12.0±0.2**
Rise (D_z)						
(C1·G12)/(G2·C11)	3.8	3.9	3.9	3.9	3.8	3.6
(C3·G10)/(G4·C9)	3.9	3.7	3.8	3.5	4.0	2.1
(C5·G8)/(G6·C7)	4.1	3.9	3.9	3.9	3.8	3.4
Average	**3.9±0.2**	**3.8±0.1**	**3.9±0.1**	**3.8±0.2**	**3.9±0.1**	**3.0±0.8**
Roll (ρ)						
(C1·G12)/(G2·C11)	-0.8	0.1	1.5	0.0	-2.2	-4.6
(C3·G10)/(G4·C9)	-1.1	-2.6	-2.9	-2.1	2.1	-5.8
(C5·G8)/(G6·C7)	3.6	1.5	2.3	3.9	1.6	3.4
Average	**0.6±2.6**	**-0.3±2.1**	**0.3±2.8**	**0.6±3.0**	**0.5±2.4**	**-2.3±5.0**

Table 2.8, continued

Tilt (τ)						
(C1·G12)/(G2·C11)	-6.0	-7.1	-7.4	-9.2	-6.4	1.1
(C3·G10)/(G4·C9)	0.9	-0.1	0.7	0.6	-0.3	-1.7
(C5·G8)/(G6·C7)	-0.8	-1.0	1.5	0.3	-2.3	0.0
Average	**-2.0±3.6**	**-2.7±3.8**	**-1.7±4.9**	**-2.8±5.6**	**-3.0±3.1**	**-0.2±1.4**
d(GpC) steps						
Twist (Ω)						
(G2·C11)/(C3·G10)	-48.9	-43.6	-48.4	-45.4	-47.9	-49.2
(G4·C9)/(C5·G8)	-50.8	-44.5	-47.0	-46.0	-48.6	-48.2
Average	**-49.9±1.3**	**-44.1±0.6**	**-47.7±1.0**	**-45.7±0.4**	**-48.3±0.5**	**-48.7±0.7**
Rise (D_z)						
(G2·C11)/(C3·G10)	3.6	3.8	3.6	3.4	3.6	3.3
(G4·C9)/(C5·G8)	3.5	3.8	3.7	3.9	3.6	2.3
Average	**3.6±0.1**	**3.8±0.0**	**3.7±0.1**	**3.7±0.4**	**3.6±0.0**	**2.8±0.7**
Roll (ρ)						
(G2·C11)/(C3·G10)	-1.5	-4.6	-0.1	-4.3	-1.4	4.7
(G4·C9)/(C5·G8)	0.3	-2.4	1.1	0.0	0.3	4.4
Average	**-0.6±1.3**	**-3.5±1.6**	**0.5±0.8**	**-2.2±3.0**	**-0.6±1.2**	**4.6±0.2**
Tilt (τ)						
(G2·C11)/(C3·G10)	0.1	0.8	2.3	5.9	-0.5	-0.5
(G4·C9)/(C5·G8)	0.6	0.3	0.3	1.3	0.1	-0.6
Average	**0.4±0.4**	**0.6±0.4**	**1.3±1.4**	**3.6±3.3**	**-0.2±0.4**	**-0.6±0.0**

Table 2.8, continued

Base Pairs

Propeller Twist (ω)						
C1·G12	1.1	2.0	0.6	6.6	2.2	1.5
G2·C11	3.2	3.4	0.9	6.5	0.4	1.7
C3·G10	0.9	4.8	1.2	4.4	0.5	5.0
G4·C9	1.5	1.2	0.4	0.7	5.7	1.0
C5·G8	0.5	0.3	2.6	5.9	0.2	1.9
G6·C7	2.7	2.1	2.1	0.7	0.4	2.0
Average	**1.7±1.1**	**2.3±1.6**	**1.3±0.9**	**4.1±2.8**	**1.6±2.2**	**2.2±1.4**
Buckle (κ)						
C1·G12	1.9	6.2	4.4	11.6	2.2	5.3
G2·C11	3.5	4.8	0.8	3.5	3.5	8.7
C3·G10	3.0	2.1	1.5	2.2	0.6	8.7
G4·C9	2.4	5.7	3.4	10.3	0.8	8.3
C5·G8	2.0	5.3	4.1	4.2	4.1	3.6
G6·C7	0.0	3.8	3.5	0.6	4.0	0.7
Average	**2.1±1.2**	**4.7±1.5**	**3.0±1.5**	**5.4±4.5**	**2.5±1.6**	**5.9±3.3**

Table 2.8, continued

x-displacement (d_x)						
C1·G12	-3.0	-3.6	-3.1	-3.7	-3.2	-3.8
G2·C11	-3.1	-3.6	-3.2	-3.6	-3.3	-3.3
C3·G10	-3.3	-3.4	-3.1	-3.4	-3.3	-2.1
G4·C9	-3.3	-3.5	-3.2	-3.4	-3.3	-2.3
C5·G8	-3.5	-3.8	-3.8	-3.7	-3.9	-3.4
G6·C7	-3.4	-3.6	-3.5	-3.7	-3.6	-3.4
Average	**-3.3±0.2**	**-3.6±0.1**	**-3.3±0.3**	**-3.6±0.1**	**-3.4±0.3**	**-3.0±0.7**

All values are in degrees, except rise (D_z) and displacement (d_x) which are in Å.
[a] Parameters for d(CGCICG) were taken directly from the tables in (Kumar and Weber, 1993).

$< \Omega >$ = -29.8° per base step for the fully methylated d(CpG) sequences). This suggests, once again, that the primary determinant of Z-DNA structure is the d(CpG) step, with the d(GpC) steps acting to compensate for perturbations to the structure.

At the base pair level, buckle and x-displacement are the only parameters that are significantly affected by methylation. In this case, d(C·G) base pairs have an average buckle of 2.1° ± 1.2° and an average x-displacement of -3.3 Å ± 0.2 Å. whereas d(m^5C·G) base pairs have an average buckle of 4.7° ± 1.5° and an average x-displacement of -3.6 Å ± 0.1 Å (Table 2.8). This again is associated with steric interactions between the substituent and the neighboring guanine nucleotide.

Studies on the hemimethylated dinucleotides d(m^5CpG/CpG) and d(Gpm^5C/GpC) show that each methyl group acts independently to affect the structure and stability of Z-DNA. The structures of the two hemimethylated d(m^5CpG/CpG) and d(Gpm^5C/GpC) steps in the sequence d(CGCGm^5CG) are intermediate between those observed for the corresponding unmethylated and fully methylated dinucleotide steps (Table 2.8). This suggests that the hemimethylated form is a true conformational intermediate. This is evident when comparing, for example, the helical twist at the two d(m^5CpG/CpG) to the average of the corresponding dinucleotides at each position. The helical twist between base pairs d(C1·G12) and d(G2·m^5C11) is -10.8°, while the average twist for the

corresponding base pairs in the unmethylated and methylated structures is -11.8°. Likewise, the base step d(m^5C5·G8)/d(G6·C7) at the opposite end of the duplex has Ω = -14.6° as compared to $< \Omega >$ = -14.2° for the unmethylated and methylated analogues. The position effects are most likely due to crystal packing forces of the lattice. Similarly, $< \Omega >$ for the d(Gpm^5C/GpC) steps are intermediate between the analogous unmethylated and methylated structures.

We would expect hemimethylation of cytosines to have an intermediate effect on the stability of Z-DNA. Indeed, this is the case. The sequence d(CGCGm^5CG) was crystallized from a solution with *CS* = 1.26 M, which is intermediate between that required for unmethylated d(CGCGCG) (2.0 M) and the fully methylated d(m^5CGm^5CGm^5CG) sequence (0.57 M). The *CS* we predict for crystallizing this hemimethylated sequence is 1.52 M, assuming equal contributions from each methyl group. A salt titration of 24 alternating base pairs of d(C·G) showed that the midpoint for the transition from B- to Z-DNA in the unmethylated sequence occurred at ~1.25 M $MgCl_2$, while the fully methylated sequence was predominantly Z-DNA in the absence of added salt (Figure 2.7). This salt induced-transition in the sequence d(m^5CGCGm^5CGCGm^5CGCGCGCGCGCGCG), which has six true hemimethylated dinucleotide steps, had a midpoint for the transition at ~1 M $MgCl_2$ (Bononi, 1995). Again, we would predict a midpoint at 0.9 M for equal contributions from each methyl group. Thus, the

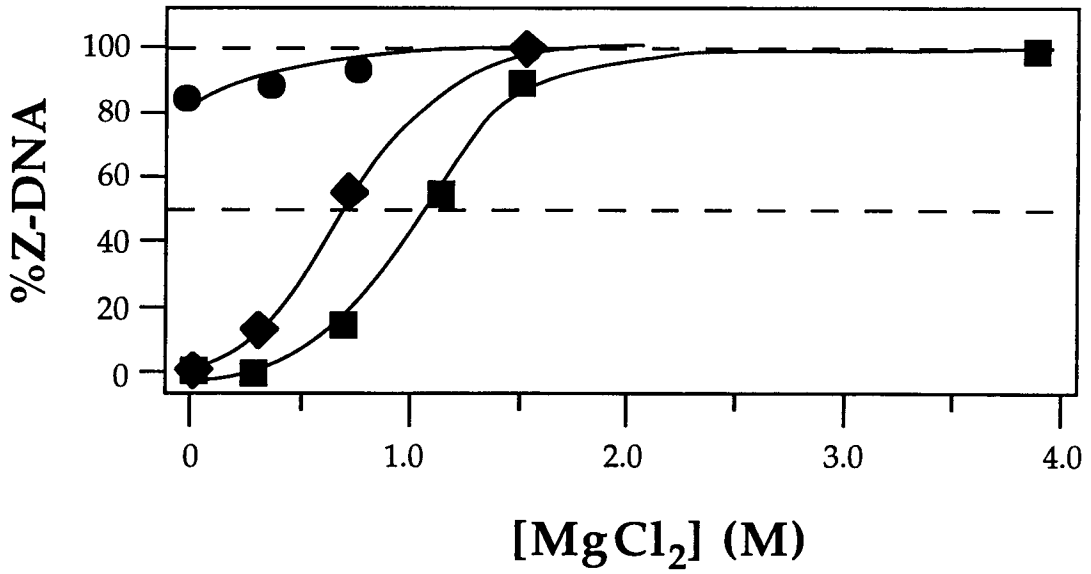


Figure 2.7

Titration of unmethylated, methylated, and hemimethylated d(CpG) dinucleotides with $MgCl_2$ to induce the formation of Z-DNA (Bononi, 1995). The unmethylated sequence $d(CG)_{12}$ (squares), fully methylated sequence $d(m^5CG)_{12}$ (circles), and hemimethylated sequence $d(m^5CGCG)_3(CGCG)_3$ (diamonds)were titrated with 0.0 to 4.0 M $MgCl_2$. The formation of Z-DNA was monitored by following the ratio of light absorbed at 260 nm versus 290 nm (A_{260}/A_{290} ratio). The conformations of the DNA at the beginning and end of the titration were confirmed to be that of B-DNA and Z-DNA, respectively, by circular dichroism spectroscopy.

hemimethylated dinucleotides represent a true intermediate, both structurally and thermodynamically, between fully unmethylated and fully methylated dinucleotides.

The Z-DNA stabilizing effect from cytosine methylation is likely associated with the hydration of the DNA structure. The only notable effect of cytosine methylation on the solvent structure in the crystal, however, is that the water which is hydrogen bonded to the cytosine N4 nitrogen at the major groove surface is slightly displaced away from the methyl group (Fujii et al., 1982). Otherwise, the arrangement of waters around the Z-DNA structure remains unperturbed. This is not entirely surprising because the methyl group actually sits recessed in a pocket of the major groove surface and thus is largely inaccessible to solvent.

The solvent free energies show that methylation makes the Z-DNA surface more hydrophobic; however, the ΔSFE_{Z-B} indicates that methylation works to stabilize Z-DNA primarily by destabilizing B-DNA (by making its surface even more hydrophobic, (Table 2.9)). Methylating the cytosines of Z-DNA increases the SFE_Z by 1.3 kcal/mol/bp and thus we would expect this to destabilize the left-handed conformation. In contrast, there is an even greater destabilization of B-DNA for these sequences (ΔSFE_{Z-B} = 0.29 kcal/mol/dn for the unmethylated sequence, but is -0.87 kcal/mol/dn for the methylated analogue). The SFE_Z and ΔSFE_{Z-B}, for the hemimethylated

Table 2.9

Solvent free energies of various dinucleotides as Z- and B-DNA

Dinucleotide *anti-p-syn*	SFE_Z (kcal/mol/dn)	SFE_B (kcal/mol/dn)	ΔSFE_{Z-B} (kcal/mol/dn)
d(CpG)	-12.97	-13.26	0.29
d(m^5CpG)/(CG)	-12.58	-12.08	-0.50
d(m^5CpG)	-10.28	-9.41	-0.87
d(TpA)	-9.90	-8.53	1.35
d(UpA)	-11.50	-10.64	0.86
d(ApT)[a]	-8.17	-9.45	1.28
d(GpC)/d(Gpm5C)[a]	-12.08	-11.62	-0.46
d(CpC)/d(GpG)[a]	-12.26	-12.90	0.64

SFE_Z and SFE_B are the solvent free energies for the dinucleotide in the Z and B conformations, respectively. SFE_Z was calculated from the crystal structure containing that dinucleotide step and SFE_B was calculated from idealized B-DNA models. ΔSFE_{Z-B} is the free energy difference for the dinucleotide step in the Z form versus the B form.

[a] Dinucleotide out-of-alternation (e.g., the first base pair of the step is *anti*, followed by *syn*).

dinucleotides in d(CGCGm^{5}CG) are again intermediate between the corresponding values for d(CGCGCG) and d(m^{5}CGm^{5}CGm^{5}CG) (Table 2.9).

2.4.2 *Effects of cytosine bromination on Z-DNA structure*

The effective radius (~2 Å) and hydrophobicity of a bromine atom is very similar to that of a methyl group so we would expect bromination of cytosines to have a similar effect in stabilizing Z-DNA, and on the structure of Z-DNA. The two Z-DNA sequences that have been crystallized which contain a brominated C5 of cytosine (Figure 2.6) are the fully brominated sequence d(Br^{5}CGBr^{5}CGBr^{5}CG) (Chevrier et al., 1986) and the hemibrominated sequence d(CGCGBr^{5}CG) (Peterson et al., 1996).

The effect of cytosine bromination on the stability of Z-DNA is equivalent to or greater than that of cytosine methylation. Poly(Br^{5}CpG) is constituitively in the Z-form even in the absence of alcohols and high concentrations of added salts (Moller et al., 1984). Like d(m^{5}CGm^{5}CGm^{5}CG), d(Br^{5}CGBr^{5}CGBr^{5}CG) required very little salt to crystallize. The enhanced stability of brominated Z-DNA compared to even the methylated form may result from the smaller perturbation of the structure.

Comparison of base dinucleotide parameters reveals trends similar to those seen for methylation (Table 2.8). Specifically, the d(CpG·CpG)

dinucleotide has an average twist of -10.3° ± 1.7°, while the d(Br^{5}CpG·Br^{5}CpG) dinucleotide is overwound by 3.4° to -13.7° ± 1.7°. Bromination, therefore, has an effect similar, but less dramatic, on Z-DNA structure than methylation. This smaller perturbation on the structure may be due to differences in the interactions with adjacent nucleotides between the spherically shaped bromine atom as opposed to a tetrahedral methyl group. As with the methylation effect on twist, the overwinding of the *anti-p-syn* step in brominated steps is compensated at the d(GpC) step to give no net difference in the helical twist of the hexanucleotide structures. The $< \Omega >$ = -30.3° per base step in d(CGCGCG) and the brominated structure has $< \Omega >$ = -30.2° per base step. Unlike other Z-DNA structures, the helical twist of d(Br^{5}CGBr^{5}CGBr^{5}CG) is not position dependent (Table 2.8), suggesting that the conformation of the fully brominated sequence is less affected by the crystal lattice. Additionally, only the Z_I backbone conformation is present in this fully brominated structure. Finally, bromination does not appear to have any effect on the base parameters of the tip, inclination, propeller twist, buckle, and x-displacement.

Unlike methylation, it is not clear if hemibromination represents an intermediate between fully brominated and fully unbrominated dinucleotides (Table 2.8). The helical twist (Ω = -8.9°) for the hemibrominated dinucleotide at one end is similar to that of d(CpG/CpG) (-9.2°). However, Ω = -13.9° for this same hemibrominated dinucleotide at the opposite end,

and is intermediate between Ω = -12.2° and -14.8° observed for the dinucleotides d(CpG/CpG) and d(Br^{5}CpG/Br^{5}CpG), respectively. Additionally, the $< \Omega >$ = -48.3° for the hemibrominated d(GpBr^{5}C/GpC) steps is identical to the average for d(GpC/GpC) and d(GpBr^{5}C/GpBr^{5}C). These observations are consistent with the hemibrominated sequence representing an intermediate conformation except at one terminal dinucleotide.

2.4.3 *Effects of the N2-amino of guanine on the structure and stability of Z-DNA*

Removing the amino group at the N2 position of guanine (to form inosine, dI) (Figure 2.6) would be expected to destabilize Z-DNA. This would eliminate one hydrogen bond within the base pair but, perhaps more importantly, would affect the spine of hydration in the minor groove of the Z-DNA duplex. The minor groove water that bridges this N2 amino group to the phosphate oxygens of the backbone is thought to be important for stabilizing the *syn* conformation of the guanine bases (Rich et al., 1984). The two published structures of inosine-containing Z-DNA are for the octamer sequence d(CGCICICG) (Kumar et al., 1992) and the hexamer d(CGCICG) (Kumar and Weber, 1993). Neither of these structures' coordinates were available for analysis by our program, but some helical

parameters could be gleaned from the data presented in the published papers.

The structure of d(CGCICICG) was disordered in the crystal and thus specific parameters for the d(CpI) and d(IpC) steps could not be distinguished from those of the d(CpG) and d(GpC) steps. The values reported were therefore averages for the respective dinucleotides. The average rise (3.6 Å for the d(CpG(I)) and 3.7 Å for d(G(I)pC)) and helical twist (16.5 ° for the d(CpG(I)) and 43.5° for d(G(I)pC)) of this structure are more similar to the tetramer d(CGCG) and the disordered octanucleotide d(CGCGCGCG) than to the parent hexanucleotide structures. This, again, may be related more to the hexagonal space group of these crystals rather than to any intrinsic structural property of Z-DNA.

The structure of d(CGCICG) shows a crystal lattice and conformation that is similar to the SP form of d(CGCGCG). The minor groove of the duplex is 0.6 Å narrower than the standard MGSP structure, but it was not clear whether this is primarily localized at the d(C·I) base pairs, or averaged over the structure. The water structure of d(CGCICG) was said to be similar to that of the spermine form of d(CGCGCG), including the continuous spine connecting the O2 oxygens of the cytosines along the minor groove crevice. This suggests that the N2 amino group is not absolutely essential to ordering the waters in the crevice. Still, the bridge from the purine to the phosphate cannot be made. The SFEs calculated suggest that d(C·I) base pairs are less stable as Z-DNA by 0.30 kcal/mol/bp compared to d(T·A) base

pairs. The sequence d(CICGCG) required a *CS* = 4.2 to crystallize as Z-DNA (Ho et al., 1991), the highest salt concentration required for any APP sequence.

2.4.4 The structure and stability of d(TpA) dinucleotides in Z-DNA

The observation that d(TpA) dinucleotides can be incorporated into the structure of Z-DNA extends the range of sequences that can adopt the left-handed conformation. Although this is an APP dinucleotide, it does not promote the formation of Z-DNA, and must be flanked by methylated d(m^5CpG) dinucleotides to crystallize, as in the sequence d(m^5CGTAm^5CG). The structure of d(m^5CGm^5CGm^5CG) therefore serves as the reference when analyzing this d(TpA) containing structure. The destabilization of Z-DNA in the crystals by the d(TpA) dinucleotide is reflected in the *CS* for the crystallization of d(m^5CGTAm^5CG) (1.3 M) as compared to that of d(m^5CGm^5CGm^5CG) (0.57 M) (Table 2.2).

The overall structure of d(m^5CGTAm^5CG) is indeed more similar to d(m^5CGm^5CGm^5CG) than to d(CGCGCG) in all respects (Tables 2.8 and 2.10). Differences in the structural details are attributed to replacing the central d(m^5CpG) dinucleotide with d(TpA). The helical twist is reduced by 1.7°, approaching that of d(CGCGCG). This is associated with a sliding of

Table 2.10

Helical parameters for d(A), d(T), d(U) and d(D)-containing sequences

	d(m^{5}CGTAm^{5}CG)	d(m^{5}CGUAm^{5}CG)	d(CGTDCG)	d(CDCGTG)	d(CDUDCG)
d(CpG) steps					
Twist (Ω)					
(C1·G12)/(G2·C11)	-16.1	-14.8	-13.5	-7.5	-13.5
(C3·G10)/(G4·C9)	-12.8	-13.8	-7.3	-12.4	-7.3
(C5·G8)/(G6·C7)	-14.9	-17.0	-13.5	-11.9	-13.5
Average	**-14.6±1.7**	**-15.2±1.6**	**-11.4±3.6**	**-10.6±2.7**	**-11.4±3.6**
Rise (D_z)					
(C1·G12)/(G2·C11)	3.9	3.8	4.2	4.0	4.2
(C3·G10)/(G4·C9)	3.3	3.4	3.6	3.9	3.6
(C5·G8)/(G6·C7)	3.9	3.8	4.3	3.8	4.3
Average	**3.7±0.3**	**3.7±0.2**	**4.0±0.4**	**3.9±0.1**	**4.0±0.4**
Roll (ρ)					
(C1·G12)/(G2·C11)	0.1	0.6	-9.4	0.5	-9.4
(C3·G10)/(G4·C9)	-1.4	-0.3	-6.6	2.6	-6.6
(C5·G8)/(G6·C7)	1.1	1.4	-5.8	0.3	-5.8
Average	**-0.1±1.3**	**0.6±0.9**	**-7.3±1.9**	**1.1±1.3**	**-7.3±1.9**

Table 2.10, continued

Tilt (τ)					
(C1·G12)/(G2·C11)	7.3	6.2	-9.7	3.9	-9.7
(C3·G10)/(G4·C9)	-2.4	3.8	0.1	-0.1	0.1
(C5·G8)/(G6·C7)	-1.1	-1.7	1.7	1.7	1.7
Average	**1.3±5.3**	**2.8±4.1**	**-2.6±6.2**	**1.8±2.0**	**-2.6±6.2**
d(GpC) steps					
Twist (Ω)					
(G2·C11)/(C3·G10)	-44.9	-44.7	-42.0	-49.7	-42.0
(G4·C9)/(C5·G8)	-44.2	-45.8	-42.0	-47.5	-42.0
Average	**-44.6±0.5**	**-45.3±0.8**	**-42.0±0.0**	**-48.6±1.6**	**-42.0±0.0**
Rise (D_z)					
(G2·C11)/(C3·G10)	3.9	3.9	3.6	3.6	3.6
(G4·C9)/(C5·G8)	3.8	3.6	3.6	3.7	3.6
Average	**3.9±0.1**	**3.8±0.2**	**3.6±0.0**	**3.7±0.1**	**3.6±0.0**
Roll (ρ)					
(G2·C11)/(C3·G10)	-5.3	4.0	-3.1	-0.6	-3.1
(G4·C9)/(C5·G8)	-5.2	6.5	-1.9	-3.1	-1.9
Average	**-5.3±0.1**	**5.3±1.8**	**-2.5±0.8**	**-1.9±1.8**	**-2.5±0.8**
Tilt (τ)					
(G2·C11)/(C3·G10)	-0.6	-1.1	-1.8	1.8	-1.8
(G4·C9)/(C5·G8)	0.6	0.7	0.2	0.3	0.2
Average	**0.0±0.8**	**-0.2±1.3**	**-0.8±1.4**	**1.1±1.1**	**-0.8±1.4**

Table 2.10, continued

Base Pairs					
Propeller Twist (ω)					
C1·G12	0.7	2.8	5.7	4.2	5.7
G2·C11	4.1	7.9	0.8	0.8	0.8
C3·G10	2.6	4.0	0.2	3.0	0.2
G4·C9	2.0	1.3	0.2	2.9	0.2
C5·G8	1.6	0.4	0.7	1.1	0.7
G6·C7	2.2	2.0	5.7	2.6	5.7
Average	**2.2±1.1**	**3.1±2.7**	**2.2±2.7**	**2.4±1.3**	**2.2±2.7**
Buckle (κ)					
C1·G12	5.5	7.7	2.9	0.9	2.9
G2·C11	7.5	7.2	1.2	1.8	1.2
C3·G10	5.1	2.0	6.9	1.0	6.9
G4·C9	8.8	4.3	6.9	1.8	6.9
C5·G8	2.1	0.3	1.2	5.2	1.2
G6·C7	2.8	2.5	2.8	6.3	2.8
Average	**5.3±2.6**	**4.0±3.0**	**3.7±2.6**	**2.8±2.3**	**3.7±2.6**

Table 2.10, continued

x-displacement (d_x)					
C1·G12	-3.8	-3.9	-3.7	-3.1	-3.7
G2·C11	-3.7	-3.6	-3.1	-3.1	-3.1
C3·G10	-3.4	-3.3	-2.8	-3.3	-2.8
G4·C9	-3.4	-3.2	-2.9	-3.3	-2.9
C5·G8	-3.8	-3.9	-3.3	-3.7	-3.3
G6·C7	-3.8	-3.9	-4.0	-3.4	-4.0
Average	**-3.7±0.2**	**-3.6±0.3**	**-3.3±0.5**	**-3.3±0.2**	**-3.3±0.5**

All values are in degrees, except rise (D_z) and displacement (d_x) which are in Å. Parameters for the reference sequences d(CGCGCG) and d(m^{5}CGm^{5}CGm^{5}CG) are shown in Table 2.8.

the d(T·A) base pairs towards each other. This sliding is localized, however, to only the d(TpA) dinucleotide, since the d(GpT) and d(Apm^{5}C) steps show increases in $< \Omega > = 0.5°$ each to compensate. The d(TpA) dinucleotide is also significantly compressed (D_Z = 3.3 Å) as compared to any d(CpG) dinucleotide.

The destabilization of Z-DNA by d(TpA) dinucleotides appears to be associated with the presence of the methyl group at the major groove surface of the thymine base and the absence of an N2 amino group from the minor groove crevice of the adenine base (Figure 2.6), both of which perturb the solvation around the d(T·A) base pairs (Kagawa et al., 1989). The cytosines of d(CpG) dinucleotides are bridged by a well defined pattern of waters at the major groove surface (Gessner et al., 1994). In contrast, the solvent structure at the d(TpA) dinucleotide major groove surface can best be described as a set of disordered waters and/or cation complexes, with no specific hydrogen bonding pattern to the thymine bases (Wang et al., 1984).

In comparison, the structure of d(m^{5}CGUAm^{5}CG) helps to pinpoint the role of the thymine methyl group in the instability of d(TpA) dinucleotides as Z-DNA. In this deoxyuridine-(Figure 2.6) containing structure of Z-DNA, the twist angle between the central d(U·A) base pairs approaches the values of Ω for the d(m^{5}CpG) dinucleotides (Table 2.8 and 2.10). This appears to result from the coupling of the two stacked uridine bases by a $Mg(H_2O)_6^{2+}$complex. This complex is analogous to the waters

that bridge the stacked cytosines at the major groove surface of the d(CpG) dinucleotides. The thymine methyls in the structure of d(m^5CGTAm^5CG) evidently disrupt these interactions. This was suggested by the lower $\Delta SFE_{Z\text{-}B}$ calculated for the d(UpA) as compared to the d(TpA) dinucleotides (Table 2.9).

Interestingly, the solvent in the minor groove crevice is also perturbed by the C5 methyl of the thymines in d(m^5CGTAm^5CG). The two well ordered waters typically observed at each d(C·G) base pair in Z-DNA (Figure 2.4) could not be located at either d(T·A) base pair (Wang et al., 1984). Thus, the spine of hydration in the minor groove crevice is disrupted at each d(T·A) base pair. This may contribute to the reduced stability of d(TpA) dinucleotides as Z-DNA. The water network in the minor groove of B-DNA is continuous even at the d(T·A) base pairs (Drew and Dickerson, 1981). In this case, the waters are hydrogen bonded to the N3 nitrogen of the purine ring, which is largely inaccessible in Z-DNA. Thus, there are no waters that bridge the N2 amino group of the purine to the phosphate backbone to stabilize the *syn* conformation, as was observed with the d(CpG) dinucleotides in Z-DNA.

The ordered hydration in the minor groove, however, is restored to the d(CpG)-like spine at d(U·A) base pairs of d(m^5CGUAm^5CG) (Zhou and Ho, 1990). This apparently results from a widening of the minor groove caused by the coupled binding of the uridine bases by the magnesium-water

complex at the major groove surface. There are two waters at each d(U·A) base pair. One water is directly hydrogen bonded to the O2 of the uridine base, while the second connects this water to the phosphoribose backbone of the opposite strand. Thus, although no water directly connects the adenine base to the backbone, there may still be a degree of stabilization of the *syn* conformation conferred by the pyrimidine-water-water-phosphate bridge. This would suggest that d(UpA) dinucleotides are more stable as Z-DNA than d(TpA) dinucleotides. Indeed, the sequence d(m^5CGUAm^5CG) was crystallized in solutions having *CS* = 0.31 M, which is less than half of that required to crystallize d(m^5CGTAm^5CG).

The magnesium complex of d(m^5CGUAm^5CG) can be displaced from the major groove by binding copper ions to the purines (Geierstanger et al., 1991). The result is that the minor groove crevice of the d(UpA) dinucleotide becomes narrower, although not as narrow as that of the d(TpA) dinucleotide. The effect of this on the spine of hydration is that the four waters at the d(UpA) dinucleotide are perturbed, but not displaced. One water remains hydrogen bonded to the O2 and in the plane of the uridine base. The second water for each base pair, however, is pushed out-of-plane and, therefore, cannot form the pyrimidine-water-water-phosphoribose bridge of the native d(UpA) structure. This displacement effectively isolates the cluster of four waters at the d(UpA) dinucleotides from those of the neighboring d(CpG) dinucleotides. Thus, although the

number of waters in the spine remain unchanged, its continuity along the minor groove and across the helix becomes disrupted by removing the magnesium complex at the major groove surface.

To see how perturbations to the major groove surface affect the stacking of the bases and the water structure of Z-DNA, we start with the d(UpA) dinucleotide of the copper soaked structure, which has a minor groove crevice that is intermediate in width (Figure 2.8). Introducing a magnesium complex at the major groove surface slides the base pairs to provide a wider crevice that can accommodate the four water molecules in the plane of the d(U·A) base pairs. Methylating the uridine bases, on the other hand, prevents the binding of this magnesium complex and slides the base pairs in the opposite direction to narrow the crevice. The narrower crevice prevents the waters from forming a well ordered network at the d(T·A) base pairs. Thus, the major and minor grooves of Z-DNA cannot be treated as two isolated domains of the structure. Perturbations to one side are transmitted through the double-helix to the other side of the duplex.

The other substituent that affects the stability of d(TpA) dinucleotides as Z-DNA is the N2 amino, or more precisely, the lack of this group on the adenine bases. The unusual base 2-aminoadenine (or diamminopurine, d(D)) (Figure 2.6) has been used to probe the effect of this group on Z-DNA structure and stability. Introducing this additional amino group to adenines apparently stabilizes Z-DNA. The sequence d(CGTDCG) was crystallized from solutions with *CS* = 1.1 M (Parkinson et al., 1995). These

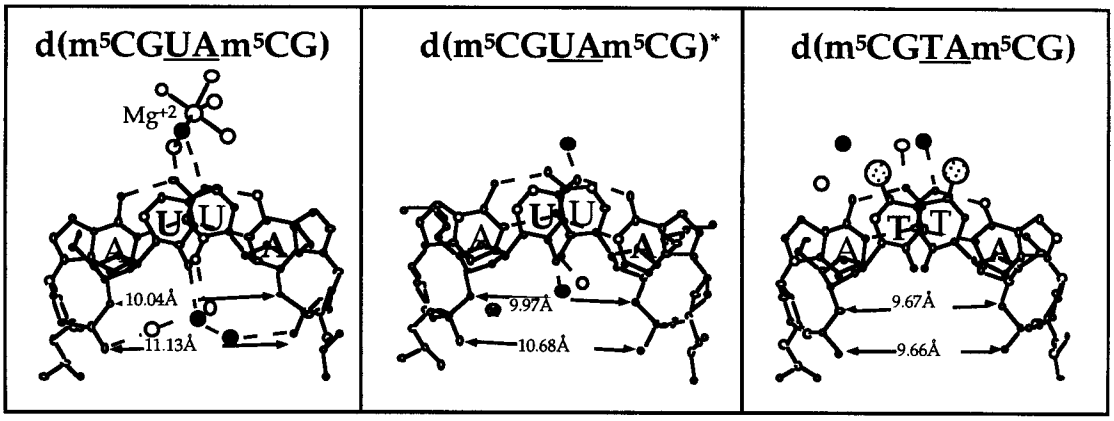


Figure 2.8

Comparison of the solvent structures and widths of the minor groove crevice of d(UpA) dinucleotides in Z-DNA. Shown are the structures of d(m^{5}CGUAm^{5}CG) (Zhou and Ho, 1990), d(m^{5}CGUAm^{5}CG) soaked with copper ((Geierstanger, et al., 1991), d(m^{5}CGUAm^{5}CG)*), and d(m^{5}CGTAm^{5}CG) (Wang, et al.,1984). The top base pair of each dinucleotide is shown with thick bonds and labeled in bold, while the lower base pairs are shown as thin bonds and labeled in standard type. Waters that interact with the top base pairs are shown as filled circles, while those interacting with the lower base pairs are open circles (the circle with a cross in the structure of d(m^{5}CGUAm^{5}CG)* sits between the two base pairs). Widths of the minor groove crevice are measured between the O3' oxygens, and between the closest oxygens of the phosphate group of the dinucleotides. The methyl group of the thymines in d(m^{5}CGTAm^{5}CG) are stippled.

conditions are comparable to that of d(m^{5}CGTAm^{5}CG), even though the cytosines are not methylated. There are two potential means by which dD stabilizes Z-DNA. The first effect would be the introduction of an additional hydrogen bond to the base pair, making d(T·D) more akin to d(C·G) base pairs. Since Z-DNA is a more rigid helix than B-DNA (Jovin et al., 1987; Rich et al., 1984), this would affect the difference in conformational entropy between the two DNA forms for the modified base pair. The second effect would be to place an additional hydrogen bonding function into the minor groove crevice to accommodate the waters of the hydration spine. This has been more extensively studied, and thus will be the focus of this discussion on d(TpD) dinucleotides.

The structure of the sequence d(CGTDCG) was solved in an unusual space group for Z-DNA, *P3$_2$21* (Parkinson et al., 1995). Although in a completely different lattice arrangement from other Z-DNA hexanucleotides, its structure shows many of the same features as standard Z-DNA (Table 2.8 and 2.10). It is, however, slightly underwound (the average helical twist is ~8° more positive) compared to d(CGCGCG), with most of this distortion associated with the d(GpC) steps (being approximately 7° to 9° less negative than comparable steps of d(CGCGCG)) and at one of the terminal d(CpG) steps (in this case 4.7° overwound in the left-handed direction). The minor groove is narrower as a result of appreciable negative roll at nearly all dinucleotide steps of the helix. These distortions may arise from the crystal lattice in that the terminal base pairs are not stacked end-to-

end to form essentially continuous strands of Z-DNA as in the "standard" hexamer crystals. The duplexes pack perpendicular to and against the major groove surface of the neighboring duplex. This general lattice is similar to A-DNA packing modes, except that the ends of the duplexes pack against the minor groove in the crystals of A-DNA hexanucleotides (Mooers et al., 1995). Thus, this structure may show more "end-effects" than would normally be observed. There are, however, some sequence dependent features.

Despite these distortions, the first hydration shell is again nearly identical to that of d(CGCGCG), if the waters at the interface between helices are ignored. The narrower minor groove crevice shifts the spine of hydration, but does not apparently "squeeze" any water out as in the d(TpA) dinucleotides. Thus, it is clear that the N2 amino group of the purine does play a significant role in defining the regular pattern of this water network. Both the crystallization conditions and salt titrations followed by circular dichroism spectroscopy show that d(TpD) dinucleotides are more stable as Z-DNA than are d(TpA), but less so than d(CpG) (Coll et al., 1986). Under dehydrating conditions, however, the hexamer d(TDTDTD) forms A-DNA instead of Z-DNA, as measured by circular dichroism. The flanking d(CpG) dinucleotides in d(CGTDCG) are required to induce d(TpD) to form Z-DNA, although the cytosine bases do not need to be methylated.

All this taken together suggests that demethylating the thymine and adding an amino group to the adenine (as in a d(UpD) dinucleotide) would

greatly enhance the stability of Z-DNA relative to the standard d(TpA) dinucleotide to the point where it should behave more like a d(CpG) base pair. Indeed, the structure of d(CGUDCG) (Schneider et al., 1992) most closely resembles that of the MG and MGSP forms of d(CGCGCG) in terms of the DNA conformation and the solvent interactions at the major groove surface and minor groove crevice. The *CS* for crystallization of this sequence as Z-DNA was identical to that of d(CGCGCG). It would be interesting to extrapolate from this to determine whether d(UDUDUD), as opposed to d(TDTDTD), would form Z-DNA in solution or in a crystal.

2.4.5 d(CpA)/d(TpG) dinucleotides in Z-DNA

One of the most prevalent simple repeating sequences found in eukaryotic genomes is the alternating pattern of d(CpA)/d(TpG) dinucleotides (Hamada and Kakunaga, 1982; Hamada et al., 1982; Schroth et al., 1992). These APP sequences are thought to form Z-DNA. Studies on Z-DNA formed in negatively supercoiled plasmids indicate that the order of stability for APP dinucleotides is d(CpG) > d(CpA)/d(TpG) > d(TpA) (Jovin et al., 1987; Rich et al., 1984). The thermodynamic propensity of a d(CpA)/d(TpG) dinucleotide to form Z-DNA is not, however, simply an average of the d(CpG) and the d(TpA) dinucleotides. The first conversion of a d(C·G) base pair in the standard d(CpG) dinucleotide to a d(T·A) base

pair is not as destabilizing as the second. Is this reflected in the crystal structure?

The single crystal structure of the sequence d(CACGTG) has been solved to ~2.5 Å resolution (Coll et al., 1988), which is one of the lowest resolution structures of Z-DNA. The structure shows two features that may contribute to the lower propensity of d(CpA)/d(TpG) dinucleotides to form Z-DNA. One is that the lack of an N2 amino group on the adenine base reduces the stacking surface and thus results in poorer stacking interactions at the d(ApC) steps as opposed to d(GpC) steps. This cannot be the major contributor, since only the d(ApC) step at the A8/C9 positions shows this poorer stacking. The d(ApC) step at A2/C3 compensates by placing the phosphate of C3 in the Z_{II} conformation. This displaces the A2 purine so that its six membered ring lies directly on top of the cytosine base.

The other effect is observed in the solvent structure of the minor groove. Although the minor groove crevice of this sequence is identical in width to that of d(CGCGCG), there were no ordered solvent molecules located at or near the adenine bases in the groove (Coll et al., 1988). The suggestion here was that the N2 amino group that is missing from the adenine base contributes to the disruption of the spine of hydration. As with the d(TpA) dinucleotide, the bridge from the purine base to the phosphoribose backbone, which appears to be important for stabilizing the purine in the *syn* conformation, is lost. In support of this proposition, the structure of d(CDCGTG) (Coll et al., 1986) shows the same organization of

water molecules in the minor groove as does d(CGCGCG). In addition, the structure of d(CDCGTG) is identical to the MGSP form of d(CGCGCG) in all respects (Table 2.8 and 2.10). This would contribute to the lower stability of d(CpA)/d(TpG) dinucleotides.

We had argued above with the structure of d(m^5CGUAm^5CG), however, that a wide minor groove, even in the absence of the N2 amino group on the purine, allows waters to organize into the well ordered spine in Z-DNA. If the widths of the minor groove crevice of d(CpA)/d(TpG) are identical to those of d(CpG) and d(CpD)/d(TpG), why are no ordered waters located near the adenines? It may be that the waters are less populated and thus could not be observed at the lower resolution of this structure. The structure of d(CACGCG)/d(CGCGTG) has been solved to 1.6 Å resolution (Sadsivan and Gautham, 1995), where one could expect to observe less populated solvent molecules. However, this asymmetric sequence shows orientational disorder about the dyad axis of the duplex; therefore, it would be difficult to definitively assign solvent structure at the d(T·A) base pairs. These base pairs effectively overlap in the electron density maps. The question therefore remains unanswered. If a higher resolution structure of d(CpA)/d(TpG) dinucleotide does indeed show the same type of pyrimidine-water-water-phosphoribose bridge as was observed with the d(UpA) step, then we can start to understand why introducing the first d(T·A) base pair into a dinucleotide is not as destabilizing to Z-DNA as the second.

2.4.6 Out-of-alternation structures

Z-DNA can tolerate dinucleotides that do not follow the APP rule for its formation (that is, they are out-of-alternation, and place pyrimidine bases in the disfavored *syn* conformation). The crystal structures of d(Br^{5}CGATBr^{5}CG) and d(m^{5}CGATm^{5}CG) were the first to indicate that the APP rule could be violated (Wang et al., 1985), and the structure of the brominated sequence was the one reported. In the structure of d(Br^{5}CGATBr^{5}CG), both thymine bases of the central dinucleotide adopt the *syn* conformation while the complementary adenines are *anti*. Still, the backbone conformation is remarkably similar to that of d(CGCGCG) (Table 2. 8 and 2.11). The twist angle (Ω) for the *anti-p-syn* step of the d(ApT) dinucleotide is -9°, while all the *syn-p-anti* steps are -49°. All nucleotides in the *anti* conformation have C2'-*endo* sugar puckers, while a majority of those in *syn* have C3'-*endo* puckers. Exceptions to this rule were at the guanines at the 3'-end of each strand. Thus, the alternating sugar conformations remain even when the pyrimidines are in *syn*.

The significant effects of the out-of-alternation base pairs on the structure of Z-DNA are seen in the stacking of the bases (Figure 2.9). The purine bases nearly completely overlap in the *anti-p-syn* stack, even more so than the pyrimidine bases of the standard APP sequences. The thymine bases, however, are completely unstacked in both the d(ApT) and d(GpA)

Table 2.11

The effects of out-of-alternation base steps on the helical structure of Z-DNA[a]

	d(m^{5}CGm^{5}CGm^{5}CG)	d(m^{5}CGGCm^{5}CG)	d(m^{5}CGGGm^{5}CG)/ d(m^{5}CGCCm^{5}CG)	d(m^{5}CGGGm^{5}CG)/ d(m^{5}CGC m^{5}CCG)	d(Br^{5}CGATBr^{5}CG)[b]
d(CpG) steps					
Twist (Ω)					
(C1•G12)/(G2•C11)	-14.4	-13.6	-13.6	-13.2	-13
(C3•G10)/(G4•C9)	-14.5	-11.4	-12.4	-12.2	-9
(C5•G8)/(G6•C7)	-16.1	-14.8	-14.7	-14.7	-12
Average	**-15.0±1.0**	**-13.3±1.7**	**-13.6±1.2**	**-13.4±1.2**	**-11±2.0**
Rise (D_z)					
(C1•G12)/(G2•C11)	3.9	4.0	3.9	4.0	
(C3•G10)/(G4•C9)	3.7	3.6	3.6	3.8	
(C5•G8)/(G6•C7)	3.9	3.6	3.8	4.0	
Average	**3.8±0.1**	**3.7±0.2**	**3.8±0.2**	**3.9±0.1**	
Roll (ρ)					
(C1•G12)/(G2•C11)	0.1	4.3	2.0	2.2	
(C3•G10)/(G4•C9)	-2.6	0.4	-0.9	0.2	
(C5•G8)/(G6•C7)	1.5	-0.3	-0.3	-2.0	
Average	**-0.3±2.1**	**1.5±2.5**	**0.3±1.5**	**0.2±2.1**	

Table 2.11, continued

Tilt (τ)					
(C1•G12)/(G2•C11)	-7.1	3.4	-1.9	5.1	
(C3•G10)/(G4•C9)	-0.1	5.1	2.0	1.8	
(C5•G8)/(G6•C7)	-1.0	-3.9	-1.7	2.5	
Average	**-2.7±3.8**	**1.5±4.8**	**-0.5±2.2**	**3.2±1.7**	
d(GpC) steps					
Twist (Ω)					
(G2•C11)/(C3•G10)	-43.6	-46.6	-46.8	-47.0	
(G4•C9)/(C5•G8)	-44.5	-46.8	-46.8	-47.3	
Average	**-44.1±0.6**	**-46.7±0.1**	**-46.8**	**-47.2±0.2**	**-49**
Rise (D_z)					
(G2•C11)/(C3•G10)	3.8	3.7	3.6	3.6	
(G4•C9)/(C5•G8)	3.8	3.7	3.8	3.8	
Average	**3.8**	**3.7**	**3.7±0.1**	**3.7±0.1**	
Roll (ρ)					
(G2•C11)/(C3•G10)	-4.6	5.6	0.2	1.3	
(G4•C9)/(C5•G8)	-2.4	1.3	-0.1	2.1	
Average	**-3.5±1.6**	**3.5±3.0**	**0.0±0.2**	**1.7±0.6**	
Tilt (τ)					
(G2•C11)/(C3•G10)	0.8	-1.4	-1.3	-0.1	
(G4•C9)/(C5•G8)	0.3	0.9	-0.1	-0.8	
Average	**0.6±0.4**	**-0.3±1.6**	**-0.7±0.8**	**-0.4±0.5**	

Table 2.11, continued

Base Pairs				
Propeller Twist (ω)				
C1•G12	2.0	5.0	0.9	2.2
G2•C11	3.4	5.8	1.8	1.7
C3•G10	4.8	2.4	3.0	4.5
G4•C9	1.2	3.1	2.2	1.7
C5•G8	0.3	1.1	2.3	1.0
G6•C7	2.1	1.3	2.8	0.8
Average	**2.3±1.6**	**3.1±1.9**	**2.2±0.7**	**2.0±1.3**
Buckle (κ)				
C1•G12	6.2	4.7	2.6	1.5
G2•C11	-4.8	-0.5	-0.8	-3.1
C3•G10	2.1	14.1	14.8	13.9
G4•C9	-5.7	-12.6	-5.4	-4.0
C5•G8	5.3	0.6	2.3	1.9
G6•C7	-3.8	-3.5	-3.2	-1.0
Average[c]	**4.7±1.5**	**6.0±5.9**	**4.8±5.1**	**4.2±4.9**

Table 2.11, continued

x-displacement (d_x)				
C1•G12	-3.6	-3.8	-3.7	-3.5
G2•C11	-3.6	-3.7	-3.7	-3.4
C3•G10	-3.4	-3.3	-3.3	-3.2
G4•C9	-3.5	-3.4	-3.4	-3.4
C5•G8	-3.8	-4.0	-3.9	-3.9
G6•C7	-3.6	-4.0	-3.7	-3.6
Average	**-3.6±0.1**	**-3.7±0.3**	**-3.6±0.2**	**-3.5±0.2**

[a] Base step, and base pair parameters are shown for crystallized Z-DNA structures containing out-of-alternation base pairs (underlined). All values are in degrees, except rise (D_z) and displacement (d_x) which are in Å.
[b] Values shown for d(Br^{5}CGATBr^{5}CG) are from (Wang et al., 1985).
[c] Averages for base pair buckle were calculated from the magnitudes of the values listed ($\langle\kappa\rangle = (\sum|\kappa_i|) / i$, where κ_i is the buckle at base pair *i*.

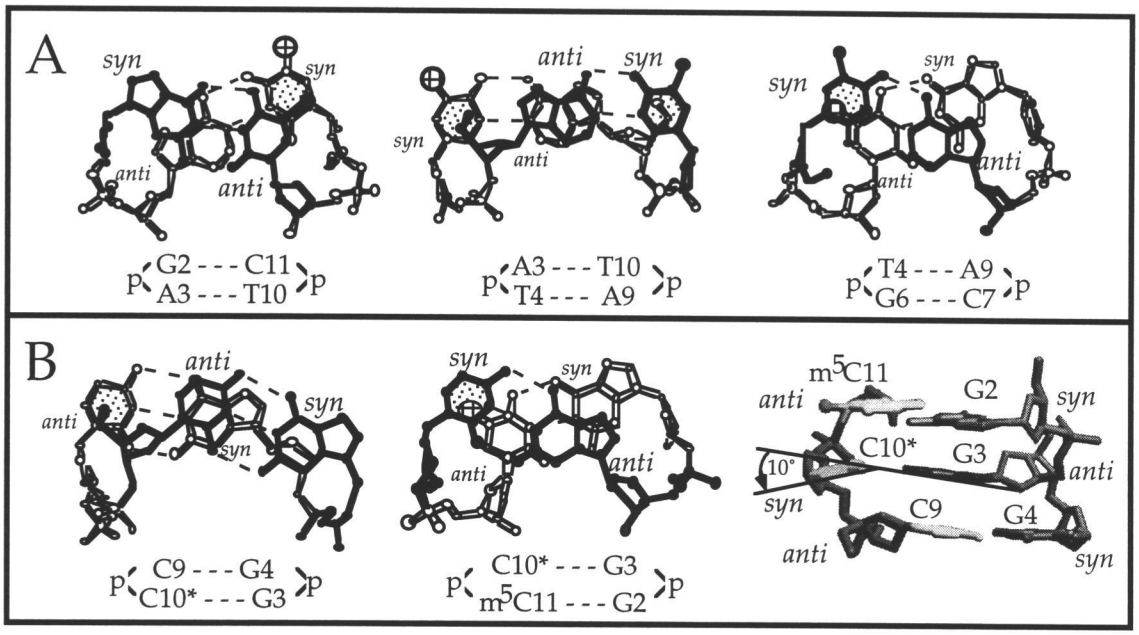


Figure 2.9

Comparison of the out-of-alternation bases in the structures of d(m^{5}CGATm^{5}CG) (A, (Wang et al., 1985)) and d(m^{5}CGGGm^{5}CG)/d(m^{5}CGm^{5}CCm^{5}CG) (B, (Schroth et al., 1993)). Shown are the dinucleotide stacks of the out-of-alternation base pairs. The pyrimidine bases that are in the disfavored *syn*-conformation are highlighted by the stippled rings. The top base pairs of the stacks are shown as solid atoms and bonds, while the bottom base pairs are in open atoms and bonds. A. Shown are views down the helix axis of the *syn-p-anti*, *anti-p-syn*, and *syn-p-anti* arrangements of the out-of-alternation base pairs in d(m^{5}CGATm^{5}CG). The structure shows that the thymine in *syn* is unstacked and protrudes away from the major groove surface for the d(G2·C11)/d(A3·T10), d(A3·T10)/d(T4·A9), and d(T4·A9)/d(G6·C7) stacked base pairs. The guanines at the two out-of-alternation base pairs stacked on-top of each other in the d(A3·T10)/d(T4·A9) stack. B. The *anti-p-syn* and *syn-p-anti* stacking of d(C·G) base pairs are shown down the helix axis and perpendicular to the axis of the d(m^{5}CGGGm5CG)/d(m^{5}CGCm^{5}Cm^{5}CG) structure. In the views down the axis, the single cytosine in *syn* is shown to also be unstacked and protruding away from the major groove surface.

steps and, therefore, protrude out from the major groove surface and into the solvent.

The organization of solvent in the minor groove is different from that of the APP d(m^{5}CGTAm^{5}CG) structure. In this latter case, no ordered waters were observed at the d(T·A) base pairs. The d(A·T) base pairs of the out-of-alternation structure do support ordered waters, but in a slightly different arrangement than in d(CGCGCG). In this case, the N3 nitrogen of adenine is accessible, as it is in B-DNA.

There are several questions that were left unanswered by this structure. Why are pyrimidines in *syn* unstable in Z-DNA? The supercoil induced B-Z transition free energy (ΔG°_T) for the APP dinucleotide d(CpA)/d(TpG) is 1.3 kcal/mol (Vologodskii and Frank-Kamenetskii, 1984), while that for the nonAPP dinucleotide d(TpC)/d(GpA) is 2.5 kcal/mol (Ellison et al., 1985) (Table 2.1). Thus placing a single thymine in *syn* requires 1.2 kcal/mol. The original explanation was that pyrimidines are sterically inhibited from adopting the *syn* conformation because of collisions between the base and the deoxyribose (Davies, 1978; Haschmeyer and Rich, 1967). The intramolecular distances from the thymine to the sugar in the d(Br^{5}CGATBr^{5}CG) structure, however, are only slightly shorter than those of guanines in *syn* to their sugars. It is unclear as to whether the stacking of bases accounts for this destabilizing effect since, although the thymines are poorly stacked, the adenines show better stacking interactions.

More likely, the protrusion of the out-of-alternation thymines into the solvent makes the difference. This will be discussed in greater detail later.

The other questions remaining are whether a single base pair that is out-of-alternation is more or less stable than two adjacent out-of-alternation base pairs in a nonAPP dinucleotide. Finally, are out-of-alternation d(T·A) base pairs more or less stable than d(G·C)? These can potentially be addressed by studying the structures of nonAPP d(GpC) sequences.

Only recently have structures of Z-DNA hexanucleotides been solved that place d(C·G) base pairs out-of-alternation. The first was the non-selfcomplementary sequence d(m^5CGGGm^5CG)-d(m^5CGCCm^5CG) (Schroth et al., 1993), which has a single cytosine base (underlined) in *syn*. Like the d(ApT) containing structure, this cytosine protrudes into the major groove, but the base pair is significantly buckled (Figure 2.9). This distortion to the base pair, which relieves the steric strain of placing the cytosine in *syn*, appears to be induced by the methyl group of an adjacent cytosine. We had proposed that in the absence of methylation of the flanking d(CpG) dinucleotides, the pyrimidine base would slide away from the ribose to relieve the steric strain, much like the thymines do in the d(Br^{5}mCGATBr^{5}CG) structure (Wang et al., 1985). In the refined structure, the steric energy was calculated to be essentially identical between this out-of-alternation structure and the standard structure of d(m^5CGm^5CGm^5CG).

The structure of d(m^5CGGGm^{5}CG)-d(m^5CGCm^{5}CCG), however, shows the out-of-alternation d(C·G) base pair with essentially the same high buckle, even in the absence of the methyl group of the adjacent cytosine. Similarly, both base pairs that are out-of-alternation in the structure of d(m^5CGGCm^{5}CG) show this same buckling (Table 2.11). Thus this distortion to the base plane is inherent to out-of-alternation base pairs, regardless of the flanking base pairs. It may simply be that the pyrimidine base in *syn* is not sandwiched by the base and deoxyriboses of the two flanking base pairs, as is the standard guanine base.

The cytosine in *syn* affects the solvent structure at both the major groove surface and minor groove crevice. In the minor groove of the out-of-alternation d(C·G) base pair, a water is hydrogen bonded to the N2 amino group of the guanine base and no waters are observed bound to the now inaccessible O2 oxygen of the cytosine, as in the nonAPP d(ApT) dinucleotides. In addition, there is no pattern of ordered waters around this d(C·G) base pair. This may, however, be associated with the orientational disorder of this non-selfcomplementary sequence.

The difference in stability between a standard d(C·G) base pair and an out-of-alternation d(G·C) base pair was estimated from supercoiled ccDNA studies to be 1.7 kcal/mol-bp (ΔG°_T for an APP d(CpG) dinucleotide (dn) is 0.7 kcal/mol-dn (Peck and Wang, 1983), while that for a d(CpC)·d(GpG) dinucleotide is 2.4 kcal/mol-dn (Ellison et al., 1985)). We believe that these solvent rearrangements play a role in this destabilization of Z-DNA.

Perhaps the two most dramatic examples of the out-of-alternation structures are the sequences d(CCCGGG) and d(m^5CGGCm^5CG). Both resemble sequences that one might expect to form A-DNA instead of Z-DNA. Indeed the reverse of the latter sequence, as in the hexamers d(GCCGGC) and its methylated analogue d(Gm^5CCGGC), have been crystallized as A-DNA (Mooers et al., 1995) . The structure of d(CCCGGG) has not been published in detail and thus we can not discuss it in this review (Malinina et al., 1994). We have recently completed the structure of d(m^5CGGCm^5CG) and find it to be nearly identical to the structure of d(CGCGCG) at the level of the DNA (Table 2.11). The one major exception is in the high buckle of the base pairs that are out-of-alternation (as discussed above). The other important structural perturbation is found in the solvent structure. At the major groove surface, the waters that bridge each cytosine are not as apparent, even at the flanking *anti-p-syn* d(CpG) dinucleotides. At the central out-of-alternation *anti-p-syn* d(GpC) dinucleotide, the two stacked guanines, however, show analogous solvent structures to those of the standard stacked cytosines in d(CGCGCG). For the flanking *anti-p-syn* d(CpG) dinucleotides in the minor groove, the waters that link the guanine N2 amino groups to the phosphoribose backbone were still observed and thus help to stabilize these in the *syn* conformation. The spine of hydration that links the cytosines in the minor groove, however, is no longer present. At the two central out-of-alternation d(G·C)

base pairs, the cytosines in *syn* are not at all accessible to solvent in the minor groove crevice. The two stacked guanines, however, are bridged by two waters that are analogous to the waters that normally form the spine that bridges the central cytosine bases in d(CGCGCG). This may help to increase the stability of the two base pairs that are out-of-alternation if they occur adjacent to each other as opposed to being separated in a sequence. Thus, although the DNA structure is not dramatically affected in this very unlikely Z-DNA sequence, the water interactions are.

2.5 Summary--Sequence effects on the structure and stability of Z-DNA

The nucleotide sequence affects not only the structure, but also the stability of Z-DNA. We have concentrated on how the major and minor grooves are affected, as well as the related solvent rearrangments at these surfaces because these are the classical explanations given for whether a DNA duplex conformation is stable or not. The characterization of Z-DNA sequences that contain d(m^5C·G), d(C·G), d(C·I), d(U·A), d(T·A), and d(T·D) base pairs in various combinations suggests that there are several distinct factors important for Z-DNA stability. Amidation of the purine base at the C2 helps to stabilize Z-DNA. Removing the N2 amino group from guanine destabilizes Z-DNA in d(C·G) sequences, while adding this group to adenine helps to stabilize the structure in d(T·A) containing sequences. Methylation at the C5 position of pyrimidine bases has both a stabilizing and

destabilizing effect on Z-DNA. Z-DNA is stabilized by methylation of cytosines, as in d(m^5C·G), and also when thymines are demethylated to form deoxyuridine. This apparent contradictory effect of methylation depends on its position relative to the amino and keto groups of the base pairs in the major groove.

We should stress, however, that comparisons of Z-DNA structures alone cannot provide an accurate account for the factors that stabilize a sequence in this form. These same parameters must be compared with the reference B-DNA structures of these sequences. Even then, however, it is not entirely clear how all these various factors contribute to the ability or inability of certain sequences to adopt the left-handed form of the duplex. For example, if one simply compares the spine of hydration in the narrow minor groove across the various Z-DNA structures, we see that this spine is disrupted by narrow minor grooves, the lack of an amino group contributed by the purine, and base pairs that violate the alternating pyrimidine-purine sequence motif for the *anti-p-syn* dinucleotide stacking. It is also clear that solvent interactions at the major groove (e.g., cation complexes that bridge stacked adjacent bases) will also affect the structure of the minor groove and its hydration spine. Whether this facilitates or hinders the formation of Z-DNA depends on your point of view. One can argue that solvent interactions are stabilizing since waters can form a direct bridge from the N2 amino group (if present) of the purine base to the DNA backbone, which would help to hold the base in the *syn* conformation. Furthermore, all this

says nothing about the effect of this amino group on the spine of hydration in the minor groove of B-DNA. However, a well structured water network can been argued to be destabilizing to either B- or Z-DNA from the perspective of the reduced entropy of the solvent structure (Schneider et al., 1992). Thus, although the large data set of single crystal structures for different sequences and substituent groups as Z-DNA provides a wealth of structural information, the details may not tell us much about the stability of this unusual conformation if we are confined to these qualitative comparisons.

One approach that does utilize the crystal structures to study and predict the effects of sequence of substituent groups on Z-DNA stability is to calculate solvent free energies (SFEs) from the structures, and to compare these to SFEs for the same sequences as B-DNA ($\Delta SFE_{Z\text{-}B}$). In this case, the reference B-DNA state is treated explicitly. Unfortunately, not all the various substituent modifications are well represented in B-DNA crystal structures; however, the SFEs calculated from B-DNA models constructed using idealized parameters appear to represent accurately the free energy for hydrating this form, even when compared to the conformations of sequences in single crystals (Basham et al., 1995; Kagawa et al., 1989).

For the standard APP sequences, we can derive a thermodynamic cycle (Kagawa et al., 1993) to elucidate how each base substituent affects the hydration and stability of Z-DNA. For example, deamidation of the guanine in d(C·G) base pairs to form d(C·I) has an energetic cost of +1.6

kcal/mol, whereas amidation of d(T·A) to form d(T·D) favors Z-DNA by -1.4 kcal/mol (Figure 2.10). This underscores the importance of the amino group in the minor groove and is consistent with its role in coordinating water molecules to form the spine of water molecules that traverse the minor goove (Wang et al., 1984). It also explains the apparent contradictory effect of methylation on the stability of Z-DNA, with methylation of cytosines favoring the left-handed form and the thymine methyl disfavoring this form. This is not intuitive, but when the SAS of each surface type are compared for B- and Z-DNA, they become more apparent (Table 2.12) . Methylation of cytosine does have the effect of increasing the overall exposed hydrophobic surface for the d(CpG) dinucleotides; however, this increase is significantly greater for B-DNA than Z-DNA and thus increases the relative stability of the left-handed form. It is now becoming evident that cytosine methylation destabilizes B-DNA, allowing the formation of A-DNA in crystals (Mooers et al., 1995), and increasing the frequency for cytosine deamindation in solution (Zhang and Mathews, 1994).

We can also make some predictions concerning the base pairs that are out-of-alternation. From the SFE calculations (Table 2.9), we can see that the d(ApT) as an out-of-alternation *anti*-p-*syn* dinucleotide is predicted to be less stable as Z-DNA as compared to the analogous d(GpC) out-of-alternation dinucleotide. This appears to be associated primarily not with

Table 2.12

Solvent accessible surface areas (Å^2) of dinucleotides steps as B- and Z-DNA

Conf.		Base Atoms					Ribose Atoms			
(B/Z)	Dinucleotide	C	CH_3(C5)	O	N	N2	C'	O'	P	Total
B	d(TpA)	43.6	44.8	32.6	55.6	-	182.8	51.4	132.8	543.6
Z	d(TpA)	46.0	46.2	27.0	50.8	-	188.2	42.0	133.6	533.8
B	d(TpD)	28.8	44.8	29.8	46.8	26.0	183.8	43.8	132.8	536.6
Z	d(TpD)	33.3	46.3	27.4	49.7	21.5	170.9	41.8	133.2	524.1
B	d(CpG)	49.4	-	31.0	59.1	23.6	185.5	47.2	132.6	528.4
Z	d(CpG)	56.3	-	44.2	48.7	19.6	184.0	47.4	132.1	532.3
B	d(CpG)	64.8	-	38.0	65.2	-	197.7	52.3	132.6	550.6
Z	d(CpG)	71.4	-	47.2	48.8	-	199.4	46.8	133.0	546.6
B	d(UpA)	63.2	-	39.2	57.6	-	190.0	51.4	133.6	535.0
Z	d(UpA)	68.6	-	37.8	57.0	-	194.6	42.0	133.8	533.8
B	d(m^5CG)	26.9	48.3	31.1	46.6	30.3	195.7	42.3	127.1	548.2
Z	d(m^5CG)	31.4	50.5	37.7	43.8	20.2	180.8	40.6	141.9	547.0

Table 2.12, continued

B	d(ApT)[a]	33.0	55.1	35.1	58.4	-	185.4	48.0	127.7	542.7
Z	d(ApT)[a]	55.4	80.3	24.6	37.2	-	171.8	48.9	131.0	549.2
B	d(GpC)[a]	34.3	-	28.3	57.9	28.1	187.8	40.3	128.8	505.6
Z	d(GpC)[a]	75.3	-	21.6	70.5	20.8	177.2	47.1	132.5	544.9
B	d(GpG)/ d(CpC)[a]	32.8	-	28.5	60.0	27.7	191.3	41.5	128.3	510.1
Z	d(GpG)/ d(CpC)[a]	67.2	-	25.5	62.0	20.6	183.6	40.9	135.9	535.7

[a] Out of alternation dinucleotide step

Figure 2.10

Effect of substituent groups on the differences in solvent free energies (ΔSFE) and the stability ($\Delta\Delta G^{\circ}_{T}$) of dinucleotides in Z-DNA versus B-DNA. A thermodynamic cycle is shown for the addition, removal, or replacement of various substituent groups, starting with the most stable dinucleotide as Z-DNA d(m^5CpG) to the least stable (d(CpI) and d(TpA)) and back to d(m^5CpG). The ΔSFE are shown for each dinucleotide, while the effect of the change in the substituent on the stability of Z-DNA ($\Delta\Delta G^{\circ}_{T}$) are shown for each modification step.

Figure 2.11

The relationship between the effective cation concentration of the crystallization solutions and the difference in solvent free energy between Z-DNA and B-DNA ($\Delta\Delta SFE_{Z\text{-}B}$) for sequences crystallized as Z-DNA. The effective cation concentration is calculated as the log of the cationic strength (log*CS*). The log*CS* that could be determined (Table 2.2) for all sequences (that contain base pairs of the type defined in Figure 2.6) are plotted relative to the $\Delta\Delta SFE_{Z\text{-}B}$ calculated (Table 2.9) for these sequences. The open circle represents the sequence d(CGCICG), which was crystallized by the hanging drop method of vapor diffusion. The line represents the best linear fit of the data for sequences with ΔSFE between -0.4 and +0.4 kcal/mol/dn (slope = 1.36, y-intercept = 0.05, R = 0.93). The plot asymptotes at both high and low values for log*CS*. At the high end, the salt concentrations reach the point of saturation in the crystallization solutions, while the low end represents the minimum amount of cations required to crystallize the DNAs (approximately equal to the concentration of mononucleotide equivalents in the DNA).

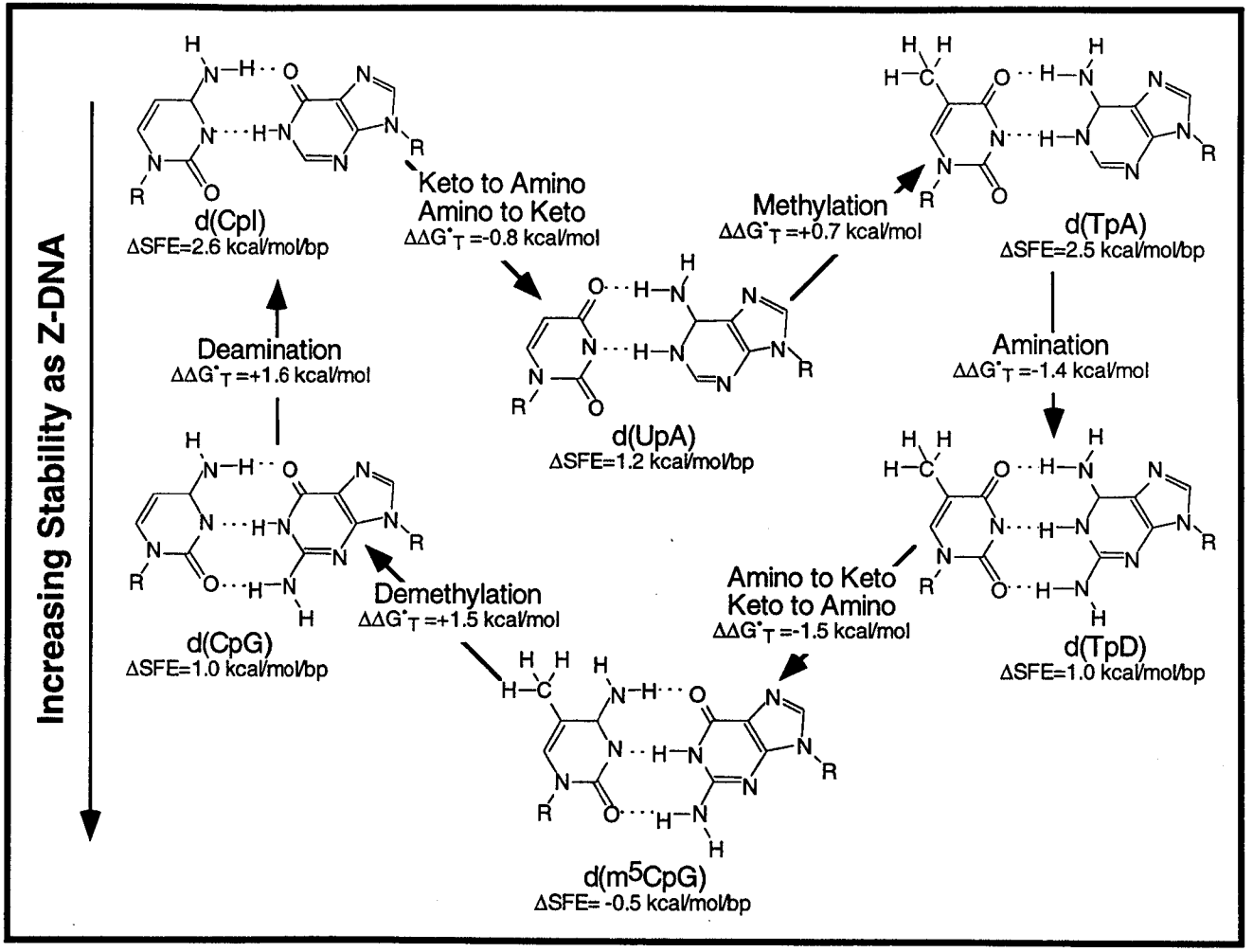

Increasing Stability as Z-DNA
d(CpI)
ΔSFE=2.6 kcal/mol/bp
Keto to Amino
Amino to Keto
ΔΔG°T =-0.8 kcal/mol
Methylation
ΔΔG°T =+0.7 kcal/mol
d(TpA)
ΔSFE=2.5 kcal/mol/bp
Deamination
ΔΔG°T =+1.6 kcal/mol
d(UpA)
ΔSFE=1.2 kcal/mol/bp
Amination
ΔΔG°T =-1.4 kcal/mol
d(CpG)
ΔSFE=1.0 kcal/mol/bp
Demethylation
ΔΔG°T =+1.5 kcal/mol
Amino to Keto
Keto to Amino
ΔΔG°T =-1.5 kcal/mol
d(TpD)
ΔSFE=1.0 kcal/mol/bp
d(m5CpG)
ΔSFE= -0.5 kcal/mol/bp

Figure 2.10

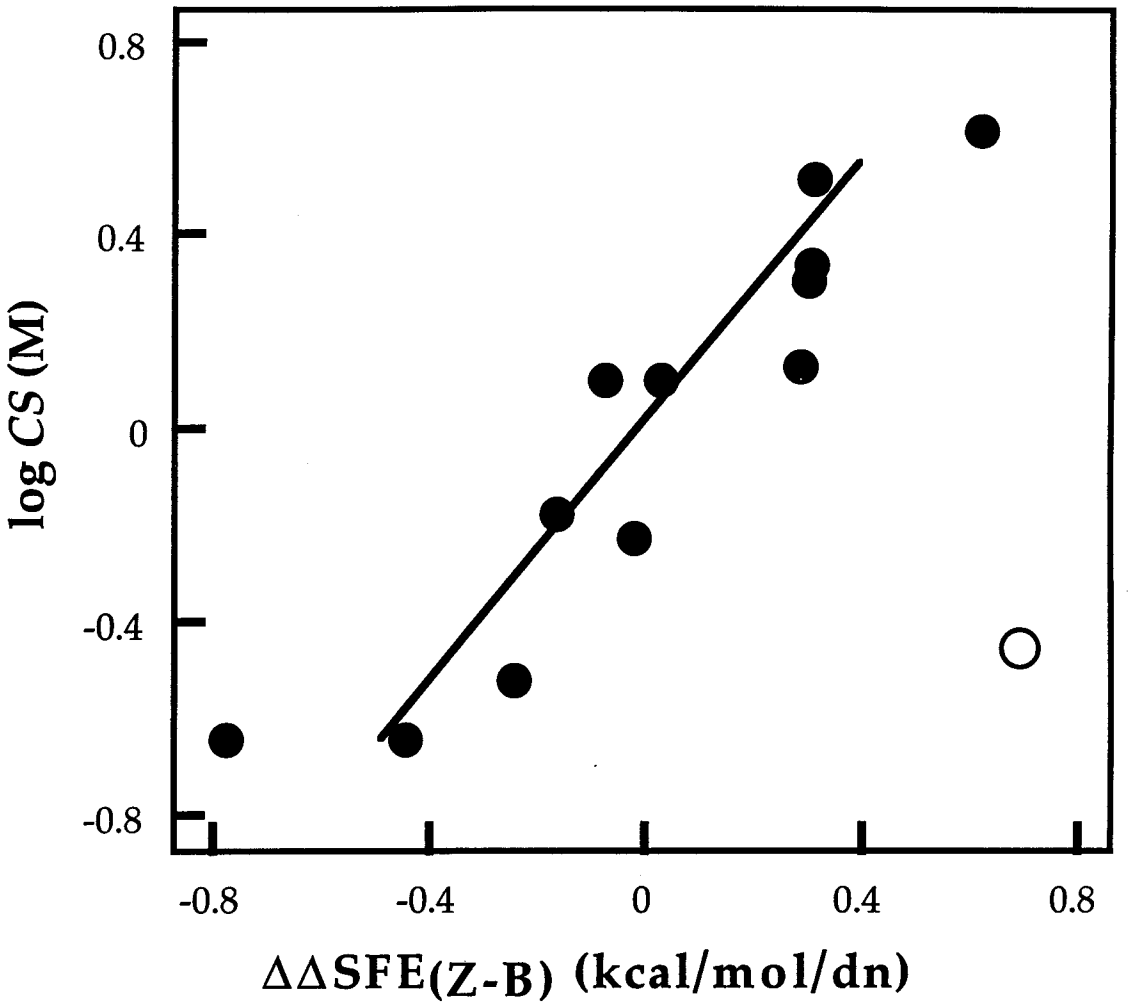


Figure 2.11

the out-of-alternation steps themselves (in this case, the d(ApT) step is actually more stable), but with how each out-of-alternation base pair affects the flanking base pairs. Finally, a single d(C·G) that is out-of-alternation is predicted to be only slightly less destabilized as Z-DNA as compared to the d(GpC) dinucleotide. Thus, we would expect that placing the cytosines of two adjacent base pairs in *syn* is more favorable than having them separated.

Upon putting all of this together in the context of the crystallography of Z-DNA, it became evident that the SFE calculations are useful as an analytical tool for predicting the target salt concentrations for obtaining crystals of this conformation (Ho et al., 1991). A comparison of the CS for crystallization of the current Z-DNA sequences (Table 2.2) shows a strong corelation to the $\Delta SFE_{Z\text{-}B}$ for these sequences (Figure 2.11). This relationship apparently arises for the stabilization of the Z- versus the B-form as both the salt and alcohol concentrations in the crystallization setups are increased. The pathway for crystallization, therefore, directs the DNA to the left-handed form in solution, while avoiding various amorphous precipitant forms along the way.

The shortcoming of the SFE approach to studying stability is that we do not utilize any of the detailed information on solvent interactions gleaned from the high resolution single-crystal structures. The general hydration parameters from the SFE calculations should somehow be related to these specific patterns of water structure. This is perhaps where Z-DNA

may play its most significant role in physical biochemistry. The accumulated structural and thermodynamic data for all these various sequences can provide a benchmark for the development of molecular forcefields. It serves much the same function as the hydrogen atom to physical chemists. The properties of Z-DNA are now very well understood; now we need to develop the theories to explain the properties. Once developed, these same principles should be generally applicable to the study and prediction of all classes of biological macromolecules.

2.6 Acknowledgments

This work has been supported by grants from the National Science Foundation (MCB9304467), the National Institutes of Health (R05GM54538A), and the Environmental Health Sciences Center at Oregon State University (NIEHS ES00210). We would like to thank Mason Kwong and Christine Nguyen for their help with this project.

Chapter 3

3. AN A-DNA TRIPLET-CODE: THERMODYNAMIC RULES FOR PREDICTING A- AND B-DNA

Beth Basham, Gary P. Schroth and P. Shing Ho

Published in the *Proceedings of the National Academy of Sciences of the United States of America*

The National Academy of Sciences, Washington D.C.

1995, 92: 6464-6468

3.1 Synopsis

The ability to predict macromolecular conformations from sequence and thermodynamic principles has long been coveted, but has generally not been achieved. We show that differences in the hydration of DNA surfaces can be used to distinguish between sequences that form A- and B-DNA. From this, a 'triplet-code' of A-DNA propensities was derived as energetic rules for predicting A-DNA formation. This code correctly predicted >90% of A- and B-DNA sequences in crystals, and correlates with A-DNA formation in solution. Thus, with our previous studies on Z-DNA, we now have a single method to predict the relative stability of sequences in the three standard DNA duplex conformations.

3.2 Introduction

A long held precept in biochemistry is that the conformation of a biomolecule is defined by, and thus can be predicted from its sequence. This is the "protein folding problem" for polypeptides (Anfinsen, 1973). An apparently simpler, although no less trivial problem lies in predicting the structure of DNA, a highly polymorphic molecule. Despite the large volume of data on DNA structures in solution, in fibers, and in single crystals, there are currently no general rules to accurately predict the ability

of a sequence to adopt one of the standard duplex conformations of A-, B-, and left-handed Z-DNA. This is the analogous "DNA-folding problem". We present here a set of thermodynamic rules, based on hydration of DNA surfaces, to distinguish between A- and B-DNA forming sequences in crystals and in solution.

Soon after Watson and Crick described the structure of B-DNA (Watson and Crick, 1953), the alternative A-DNA conformation was reported by Franklin and Gosling (Franklin and Gosling, 1953b). A-DNA is a shorter, broader helix as compared to B-DNA, and is characterized by base pairs that are highly inclined and displaced away from the helix axis, and a helical repeat of 11 base pairs per turn (bp/turn) as compared to 10 - 10.5 bp/turn for B-DNA (Saenger, 1984). Although the conformation of DNA in fibers depends on its water content and sequence (Franklin and Gosling, 1953a,b), this relationship has not generally been exploited to predict sequences that form A-DNA.

Predicting DNA secondary structure has traditionally relied on simple sequence rules. For example, Z-DNA is often assigned to alternating pyrimidine-purine sequences that are d(C/G) rich (Rich et al., 1984; Jovin et al., 1987; Peticolas et al., 1988), while non-alternating d(G/C) rich sequences that contain d(CC/GG) steps favor A-DNA (Peticolas et al., 1988). These rules, although successful in predicting simple A-DNA (Ivanov and Krylov, 1992) and Z-DNA (Jovin et al., 1987) sequences, are not general for the large number of oligonucleotides that have been crystallized as A- and

B-DNA. For example, d(CCAGGCCTGG) and d(ACCGGCCGGT) have the same base compositions and number of d(CC/GG) steps, but crystallize as B- and A-DNA respectively (Frederick et al., 1989; Heinemann and Alings, 1989). Furthermore, A-DNA has been crystallized in sequences that are 50% d(A/T), and at least one B-DNA sequence contains only d(C/G) base pairs (Dickerson, 1992).

We had previously shown that the sequence dependent stability of Z-DNA is related to the free energy required to hydrate exposed DNA surfaces (the solvent free energies, or SFEs). Z-DNA was found to be more hydrophobic than B-DNA (Kagawa et al., 1989). This difference becomes greater for sequences that are less stable as Z-DNA (Kagawa et al., 1993). This is consistent with Z-DNA being stabilized by dehydrating conditions (Rich et al., 1984; Jovin et al., 1987).

A-DNA is also dehydrated compared to B-DNA (Dickerson, 1990). The positions of waters around a DNA structure is correlated with its conformation (Schneider et al., 1992). Therefore A-DNA stability should also be dependent on the hydration of the DNA surface (Alden and Kim, 1979). Here, we use SFEs to derive a set of thermodynamic rules to predict sequences that form A- or B-DNA. First, the differences in SFEs of A- and B-DNA models (ΔSFE_{A-B}) were determined for sequences that have been crystallized as A-DNA and B-DNA. From this, we derive a set of thermodynamic rules to describe the context dependent A-DNA propensities for individual base pairs. These rules accurately predict the conformations of

oligonucleotides that have been crystallized as A- and B-DNA and predict the behaviors of well defined oligonucleotide sequences in solution.

3.3 Materials and Methods

3.3.1 SFE calculations:

For this work, we assembled a data set of A- and B-DNA crystal structures (those solved to 2.7 Å or better) that contain only standard Watson-Crick base pairs, have standard bases and phosphoribose backbones, and that did not contain any drugs or other known ligands. This includes 17 unique B-DNA sequences and 17 A-DNA sequences (Table 3.1), as listed in the Nucleic Acids Data Base (Berman et al., 1992) or the Brookhaven Protein Data Base (Bernstein et al., 1977). A- and B-DNA models were constructed for each sequence using the program HyperChem (AutoDesk, Inc.). Although the "best" model for each sequence may be its crystal structure, no sequence in this data set has been crystallized in both conformations. To treat all sequences equally and consistently, the models were generated using standard helical parameters (Arnott and Hukins, 1972; Arnott et al., 1975b).

The SFE of each DNA model was calculated as previously described (Kagawa et al., 1989). In short, the solvent accessible surface (SAS) of the

DNA was determined by rolling a probe (1.45 Å in radius) over its structure (Connolly, 1983). SFEs were calculated by applying an atomic solvation parameter (ASP) to the accessible surfaces (Kagawa et al., 1989; 1993), (SFE = $\sum SAS_i \times ASP_i$ for each atom type i). The terminal base pairs were excluded from each SFE calculation because of our inability to accurately model the hydration of the ends (Kagawa et al., 1989).

We had previously shown (Kagawa et al., 1989), and confirm in this study, that the SFEs of B-DNA crystal structures are variable at the base pair level. When averaged over the entire sequence, however, the SFEs of the crystal and model structures are nearly identical ($\Delta SFE = 0.02 \pm 0.29$ kcal/mol/bp), indicating that the hydration of B-DNA in the crystal is accurately represented by the SFE of the model structures. The crystal and model structures of A-DNA hexanucleotides are very similar (Mooers et al., 1995), as were their respective SFEs. Not surprisingly, however, the SFEs of A-DNA octanucleotides, which are greatly distorted by crystal lattice effects (Timsit and Moras, 1992), are very different. This was further impetus for using the idealized models to calculate SFEs, since it is difficult to accurately account for crystal packing effects and difficult to generate A-DNA models equivalent to a crystal structure for sequences that have only been crystallized as B-DNA.

3.3.2 Solution studies:

Oligonucleotides for solution studies were synthesized on an Applied Biosystems DNA synthesizer in the Center for Gene Research and Biotechnology at Oregon State University. These were passed over a Sephadex G-25 column and subsequently annealed in 3.33 mM Tris, pH 8.1 buffer containing 0.03 mM EDTA and 4 mM NaCl. The sequences were diluted to 0.4 absorbance units in a 10 mM Tris, pH 8.1 buffer containing 0.02 mM EDTA, and titrated with trifluoroethanol (TFE, from Sigma). The solubility of the sequences defined the maximum concentration of TFE added in the titrations. Circular dichroism (CD) spectra of each sample were recorded on a JASCO J-720 spectrometer.

3.4 Results

3.4.1 A- and B-DNA data sets

In this study, we use a data base of A- and B-DNA crystal structures to define the sequence determinants for A-DNA formation. Is the ability to crystallize a DNA sequence in a particular conformation a good indicator of its stability in that conformation? It has been suggested that the DNA conformation in a crystal is strongly influenced by the crystallization solutions and by crystal lattice forces (Timsit and Moras, 1992). To address this first

point, we compared the concentration of salts and alcohol precipitants reported for crystallizing A- and B-DNA (Dickerson, 1992). We had previously observed that the cation strength ($CS = \Sigma[Z_i^2 \times C_i]$, where Z_i is the charge and C_i the concentration of each cation type i) to crystallize hexanucleotides as Z-DNA is related to the stability of these sequences as Z-DNA (Kagawa et al., 1993; Ho et al., 1991). For A- and B-DNA, this relationship does not hold. On average, $CS = 0.2 \pm 0.3$ M per millimole phosphate for A-DNA sequences and 0.2 ± 0.2 M per millimole phosphate for B-DNA (crystallization conditions from Dickerson, 1992). Only cobalt hexaamine, which induces A-DNA in solution (Xu et al., 1993a,b), directly affects the conformation of the DNA in the crystal. In addition, neither the temperature nor precipitant concentrations were correlated with the crystallization of either conformation. Although most oligonucleotides require 2-methyl-2,4-dimethylpentanediol (MPD) or some other alcohol precipitant for crystallization, several A-DNA crystals have been obtained in the absence of any alcohol (Timsit and Moras, 1992). Whether a sequence crystallizes as A- or B-DNA is therefore not generally defined by the crystallization conditions, with the caveat that these solutions are normally more hydrophobic than standard aqueous buffers.

Some have also suggested that the conformations of DNA in crystals are defined primarily by lattice packing forces, and that sequence length is an important determinant of structure. Dickerson et al. (1994), however, have argued against this 'tyranny of the lattice'. For instance, decamer and

dodecamer sequences have been crystallized as both A- and B-DNA. In addition, hexanucleotides have been crystallized as B-DNA (Cruse et al., 1986), Z-DNA (Rich et al., 1984), and most recently as A-DNA (Mooers et al., 1995), indicating that hexamers "fit" into the crystal lattices of all three DNA conformations. Therefore, the conformation in the crystal is not strictly defined by the DNA length. Here, we use idealized models for A- and B-DNA to avoid any potential influence of crystal packing effects on our results.

3.4.2 Distinguishing A- and B-DNA by SFEs

We would expect that A-forming sequences will be more hydrophilic as A-DNA than as B-DNA, and *vice versa*. Thus, the difference in SFE for A- versus B-DNA (ΔSFE_{A-B}) of an A-forming sequence should be lower than that of a B-forming sequence. This is indeed the case. The average ΔSFE_{A-B} for A-DNA sequences is 0.53 kcal/mol/bp lower than those of the B-DNA sequences (Table 3.1). The sequences fall into two distinct populations, with a 99.8% confidence limit as defined by a standard t-test (Figure 3.1). Thus ΔSFE_{A-B} can discriminate between sequences that crystallize as A- or B-DNA.

A ΔSFE_{A-B} = 0.50 kcal/mol/bp (~1 standard deviation from the mean of each population) was used as the criteria to distinguish A- from B-DNA sequences. This being a positive value suggests that most of sequences, including those crystallized as A-DNA, are primarily B-DNA in aqueous

Figure 3.1

Distributions of ΔSFE_{A-B} for A-DNA (top) and B-DNA (bottom) sequences. Values for ΔSFE_{A-B} (see Table 3.1) were rounded to the nearest 0.1 kcal/mol/bp. The curves represent the standard gaussian distributions calculated from the means and standard deviations of the ΔSFE_{A-B} for each population (Table 3.1). The shaded area represents 1 standard deviation from the mean of each population.

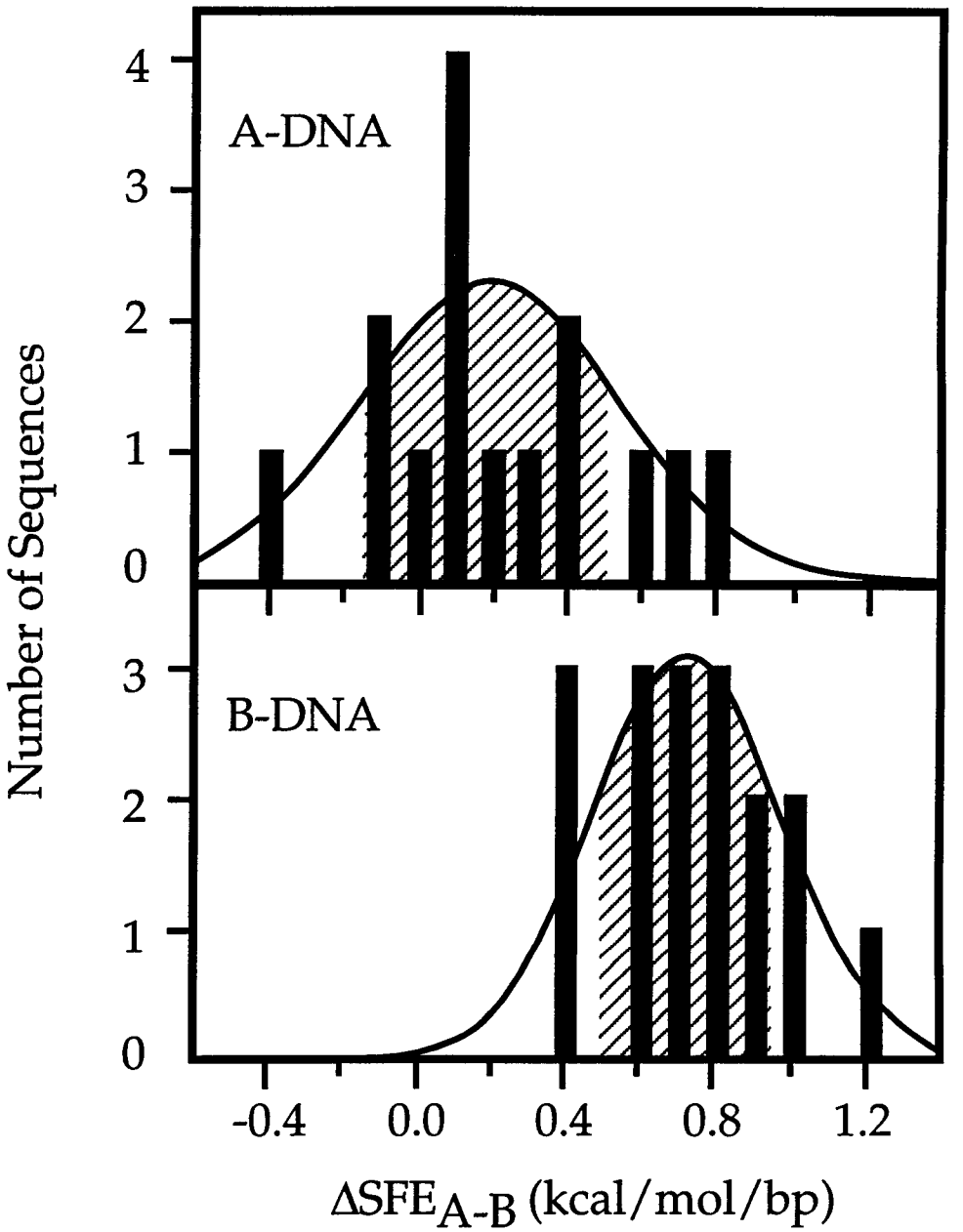

A-DNA
B-DNA
Number of Sequences
4
3
2
1
0
3
2
1
0
-0.4
0.0
0.4
0.8
1.2
ΔSFE A-B (kcal/mol/bp)

Figure 3.1

Table 3.1

Solvent free energies (SFEs) of A- and B-DNA sequences modeled from standard helical parameters.

B-DNA sequences	SFE_A	SFE_B	ΔSFE_{A-B}	A-DNA Sequence	SFE_A	SFE_B	ΔSFE_{A-B}
d(CCAGGCCTGG)†	-4.89	-5.47	0.58	d(GCCGGC)†	-6.32	-5.91	-0.41
d(CCAACGTTGG)†	-4.66	-5.03	0.37	d(GGGGCCCC)	-5.8	-5.97	0.17
d(CGATCGATCG)	-4.44	-4.99	0.55	d(GGGATCCC)	-4.99	-5.29	0.30
d(CGATTAATCG)	-3.83	-4.55	0.72	d(GCCCGGGC)	-5.97	-5.85	-0.12
d(CGATATATCG)	-3.72	-4.56	0.84	d(CCCCGGGG)	-5.93	-5.84	-0.09
d(CCGGCGCCGG)	-5.24	-5.94	0.70	d(GTACGTAC)	-4.52	-4.57	0.05
d(CATGGCCATG)	-4.52	-4.89	0.37	d(CTCTAGAG)	-4.62	-4.58	-0.04
d(CCAAGCTTGG)†	-4.69	-5.08	0.39	d(GTGTACAC)	-4.32	-4.44	0.12
d(CCATTAATGG)	-3.84	-4.50	0.67	d(GGGCGCCC)	-5.79	-5.93	0.14
d(CGTGAATTCACG)	-3.99	-4.87	0.88	d(GGGTACCC)	-4.9	-5.29	0.39
d(CGTAGATCTACG)	-4.02	-4.83	0.81	d(GTCTAGAC)	-4.56	-4.60	0.04
d(CGCGAATTCGCG)	-4.26	-5.43	1.17	d(ACCGGCCGGT)	-5.29	-6.11	0.82
d(CGCATATATGCG)	-4.03	-5.01	0.98	d(GCGGGCCCGC)	-5.25	-5.96	0.71
d(CGCAAATTTGCG)	-4.17	-4.79	0.62	d(CCCGGCCGGG)	-5.27	-5.89	0.63
d(CGCAAAAAAGCG)†	-4.04	-4.90	0.87	d(CCCCCGCGGGGG)	-4.90	-5.29	0.39
d(CGCGAAAAAACG)†	-4.12	-4.93	0.81	d(CCGTACGTACGG)	-4.18	-5.11	0.93
d(CGCAAAAATGCG)	-4.17	-5.14	0.97	d(GCGTACGTACGC)	-4.18	-5.13	0.95
B-DNA Average (Standard Deviation)			0.72 (0.23)	A-DNA Average (Standard Deviation)			0.19 (0.33)

†These sequences were not used in the derivation of APEs in Figure 3.2.

solution. The ΔSFE_{A-B} values presented here apply primarily to the more hydrophobic crystallization solutions. We would expect, however, that sequences with $\Delta SFE_{A-B} \leq 0.5$ kcal/mol/bp should have some A-DNA characteristics, even though the oligonucleotides would be predominantly B-DNA in aqueous solution. From this, we derived an A-propensity energy, or APE = ΔSFE_{A-B} - 0.50 kcal/mol/bp, as the thermodynamic propensity of a sequence to adopt the A-conformation. The uncertainty in discriminating between A- and B-DNA according to APEs is 0.03 kcal/mol/bp (the overlap at 1 S.D. from the mean of each population). Thus an APE $\leq$ -0.02 kcal/mol/bp should favor A-DNA, while an APE $\geq$ +0.02 favors B-DNA.

Of the 17 B-DNA sequences, 14 have APEs > 0 (predicted to form B-DNA), while 3 have APEs < 0 (predicted to be A-DNA). Of the 17 A-DNA sequences, 12 have negative APEs, while 5 have positive APEs. The latter five sequences include the three decamers and two of the three dodecamers. The two dodecanucleotides were crystallized with cobalt hexaamine, which stabilizes A-DNA in solution (Xu et al., 1993b), and, therefore, were induced to adopt the A-conformation. As such, these sequences were excluded from the remainder of the study. No other sequence in the data set had been crystallized with cobalt hexaamine. The APEs therefore properly assigned 12/15 (80%) of the sequences crystallized as A-DNA.

3.4.3 A triplet code to predict A-DNA formation

To be useful as predictive tool, the APEs were redefined at the base pair level. The solvation of a base pair in A- or B-DNA must be considered in the context of all possible neighboring 5′ and 3′ base pairs, that is as a trinucleotide. A sequence is a linear combination of trinucleotides and its hydration free energy is the average hydration across these triplets. For example, d(GGGGCCCC) is composed of 4 d(GGG/CCC) and 2 d(GGC/GCC) triplets, and its APE (-0.35 kcal/mol/bp) is simply the weighted average of the APEs of these trinucleotides. Using the method of singular-value decomposition, we can thus determine the contribution of each trinucleotide to the average APEs of the sequences in our data set.

There are potentially 32 unique trinucleotide combinations in duplex DNA. The APEs of 25 unique trinucleotides were determined from our data set to derive a "triplet code" for A-DNA stability (Figure 3.2). This was limited by the size of our data set. In this triplet code, negative APEs indicate that the central nucleotide favors A-DNA, while positive values favor B-DNA. Of the 25 unique triplets in Figure 3.2, 7 are strongly A-forming (APE $\leq$ -0.5 kcal/mol/bp), 8 are strongly B-forming (APE $\geq$ 0.5 kcal/mol/bp), and 10 do not appear to strongly favor either form. Not surprisingly, d(CCC/GGG) is strongly A-forming, while triplets that contain only d(T/A) base pairs are all very strong B-formers. All strong A-triplets have at least one d(C/G) nucleotide, which fits the general rule that d(C/G)

N_{i-1}	N_i								N_{i+1}
	C		G		A		T		
C	-0.59	(13)	0.71	(11)	-1.76	(2)	-1.97	(1)	C
	0.29	(11)	0.29	(11)	nd		nd		G
	0.04	(2)	-0.33	(5)	-2.12	(2)	0.15	(3)	A
	nd		2.49	(7)	-0.10	(4)	nd		T
G	-0.49	(8)	-0.49	(8)	0.69	(1)	0.69	(1)	C
	0.71	(11)	-0.59	(13)	-1.97	(1)	-1.76	(2)	G
	-0.74	(3)	-0.48	(1)	1.56	(2)	-1.74	(9)	A
	nd		1.52	(2)	0.41	(6)	nd		T
A	1.52	(2)	nd		nd		0.41	(6)	C
	2.49	(7)	nd		nd		-0.10	(5)	G
	nd		0.11	(3)	2.06	(3)	0.57	(5)	A
	nd		nd		0.58	(6)	0.58	(5)	T
T	-0.48	(1)	-0.74	(3)	-1.74	(8)	1.56	(2)	C
	-0.33	(5)	0.04	(2)	0.15	(3)	-2.12	(1)	G
	-1.06	(1)	-1.06	(1)	0.10	(2)	0.10	(2)	A
	0.11	(3)	nd		0.57	(5)	2.06	(2)	T

Figure 3.2

The A-DNA triplet code of A-DNA propensity energies (APEs). APEs, in kcal/mol/bp, are for trinucleotides consisting of a central base pair (N_i) and the 5'-flanking (N_{i-1}) and 3'-flanking (N_{i+1}) base pairs. The number of times each unique triplet is represented in the data set are shown in parentheses. An APE $\leq$ -0.02 kcal/mol/bp favors A-DNA while an APE $\geq$ 0.02 kcal/mol/bp favors B-DNA. Triplets labeled not determined (n.d.) were not represented in the curent data set.

favors A-DNA. There were some notable exceptions, however. The alternating triplet d(CGC/GCG) is a strong B-former. Also, while d(GCC/GGC) is strongly A-forming, d(CCG/CGG) is B-forming. The triplet d(GTA/TAC) is predicted to be a strong A-former (APE = -1.74 kcal/mol-bp), even though it is centered around a d(T/A) base pair, and is >50% d(T/A) in composition. We should note that additional crystal structures will help to improve the reliability of the APEs and to fill the gaps remaining in Figure 3.2.

3.4.4 APE predictions for A- and B-DNA in crystals

How well does this triplet code predict the conformations of the sequences in the data set? Of the 26 sequences used to derive the APEs, 24 were predicted correctly as either A- or B-DNA and 2 incorrectly (Table 3.2). For 6 sequences that were not used to derive the triplet code, the conformations of 5 were predicted correctly and unambiguously, while one was incorrectly predicted (Table 3.3). Thus for the 32 sequences that have been crystallized to date and for which we can derive APE values from Figure 3.2, the APEs predict >90% (29/32) correctly and unambiguously, and 9% (3/32) incorrectly. We suspect that the conformations of these three latter cases are strongly influenced by the crystal lattice.

3.4.5 Conformations of oligonucleotides in solution

To determine whether the APE triplet code is useful for predicting A-DNA in solution, we monitored the CD spectra of four dodecanucleotides titrated with TFE, which induces A-DNA in solution (Sprecher et al., 1979). Two of the sequences are predicted by the APEs to strongly favor A-DNA and two to be stable as B-DNA (Table 3.4).

The CD spectra of A-DNA are characterized by, among other things, a strong positive band at 270 nm, a weaker negative band at 240 nm, and a strong negative band at about 210 nm (Ivanov et al., 1973). B-DNA spectra are distinguishable from A-DNA spectra in that the 270 nm band is less intense and the 240 nm band is more intensely negative (Brahms and Mommaetrs, 1964). An increase in the intensity of the bands at 270 nm and 210 nm is indicative of a B- to A-DNA transition.

A typical TFE induced B- to A-DNA transition is observed for d(GGCGGCGGCGGC) (Figure 3.3a). This sequence (APE = 0.10 kcal/mol/bp) shows a typical B-DNA CD spectrum at low TFE (<60%). Upon titration to 71% TFE, the spectrum shifts to that of A-DNA, as evidenced by the increase in intensity at 210 nm and 270 nm, and precipitates at even higher TFE concentrations.

The sequence d(GCGCGCGCGCGC) (APE = 0.71 kcal/mol/bp) also shows a characteristic B-DNA spectrum at TFE concentrations < 68%. In 75% TFE the CD spectrum is characteristic of Z-DNA, with positive bands at

Table 3.2

Conformations of A- and B-DNA sequences as predicted from the APEs

Sequence	APE	ΔSFE_{A-B}	Residual	Predicted Conformation
A-DNA Sequences				
d(GGGGCCCC)	-0.56	-0.35	0.21	A
d(GGGATCCC)	-0.22	-0.22	0.00	A
d(GCCCGGGC)	-0.26	-0.64	-0.38	A
d(CCCCGGGG)	-0.30	-0.61	0.31	A
d(GTACGTAC)	-0.33	-0.47	0.14	A
d(CTCTAGAG)	-0.56	-0.56	0.00	A
d(GGGCGCCC)	-0.12	-0.38	0.26	A
d(GGGTACCC)	-0.27	-0.13	-0.14	A
d(GTCTAGAC)	-0.48	-0.48	0.00	A
d(GTGCGCAC)	-0.60	-0.60	0.00	A
d(ACCGGCCGGT)	0.40	0.30	0.10	B
d(GCGGGCCCGC)	-0.02	0.19	-0.21	A
d(CCCGGCCGGG)	-0.13	0.11	-0.24	A
d(CCCCCGCGGGGG)	-0.15	-0.13	-0.02	A
d(CCGTACGTACGG)	0.36	0.41	-0.05	(B)†
d(GCGTACGTACGC)	0.44	0.43	0.01	(B)†
B-DNA Sequences				
d(CGATCGATCG)	0.04	0.03	0.01	B
d(CGATTAATCG)	0.19	0.20	-0.0.1	B
d(CGATATATCG)	0.31	0.32	-0.01	B
d(CCGGCGCCGG)	0.20	0.18	0.02	B
d(CATGGCCATG)	-0.16	-0.15	-0.01	A
d(CCATTAATGG)	0.16	0.15	0.01	B
d(CGCATATATGCG)	0.47	0.46	0.01	B
d(CGCAAAAATGCG)	0.45	0.45	0.00	B
d(CGTGAATTCACG)	0.36	0.36	0.00	B
d(CGCAAATTTGCG)	0.10	0.10	0.00	B
d(CGTAGATCTACG)	0.29	0.29	0.00	B
d(CGCGAATTCGCG)	0.65	0.65	0.00	B

The comparable ΔSFE_{A-B} values (with 0.5 kcal/mol/bp subtracted from the values in Table 3.1) and the residual difference between the average APE (kcal/mol/bp) and ΔSFE_{A-B} (kcal/mol/bp) are listed.

Table 3.3

Conformations predicted and observed in A-DNA crystals.

		Conformation	
Sequence	**APE**	**Crystal**	**Predicted**
d(GGATGGGAG)/d(CTCCCATCC)	-0.45	A	A
d(GGCCGGCC)	-0.23	A	A
d(GGTATACC)	0.12	A	B
d(GGGCGCCC)	-0.12	A	A
d(ATGCGCAT)	-0.04	A	A
d(GCCGGC)	-0.10	A	A

The conformations of six sequences not included in the data set for deriving the APEs (kcal/mol/bp) are predicted using the values in Figure 3.2.

Table 3.4

Conformations of dodecanucleotides determined by CD spectroscopy.

		DNA Conformation		
Sequence	**APE**	**Predicted**	**0% TFE**	**High TFE**
d(GGCGGCGGCGGC)	0.10	B	B	A (71%)
d(GCGCGCGCGCGC)	0.71	B	B	Z (75%)
d(CCCCCGCGGGGG)	-0.15	A	A-like	A (68%)
d(CCCCGTACGGGG)	-0.33	A	A-like	A (75%)

Spectra for these sequences are show in Figure 3.3. The APE (kcal/mol/bp) for the sequence was calculated from the APEs in Figure 3.2. The percent TFE of the solution in which the conformation was observed is shown in parentheses.

220 nm and 270 nm, and a strong negative band at ~190 nm (Figure 3.3b) (Riazance et al., 1987). The previously calculated ΔSFE of 0.15 kcal/mol/bp for Z- versus B-DNA (Kagawa et al., 1993) suggests that the order of conformation stability for alternating d(GC) sequences is B > Z > A. Therefore, the sequence would be expected to form Z- rather than A-DNA under dehydrating conditions.

The sequence d(CCCCCGCGGGGG) (APE = -0.15 kcal/mol/bp) has been crystallized as A-DNA (Verdaguer et al., 1991). The CD spectrum of this sequence in 0% TFE shows an intense positive band at 260 nm (Figure 3.3c), which is more characteristic of A-DNA than of B-DNA. The addition of TFE does not dramatically change the overall shape of the spectrum, and is clearly that of A-DNA at 68% TFE. Continued titration of this sequence to 83% TFE resulted in a transition to another unidentifiable conformation. Is this sequence predominantly A-DNA even in aqueous solution? This sequence is a self-complementary oligonucleotide analog of poly(dG)•poly(dC). Poly(dG)•poly(dC) shows only an A-DNA fiber diffraction pattern (Arnott et al., 1975a). More recent CD and NMR studies have established that it is A-DNA in aqueous solution (Sarma et al., 1986), and this has been confirmed using chiral probes (Mei and Barton, 1988). In addition, linear dichroism studies in neutral aqueous buffer show that the bases of poly(dG)•poly(dC) are more dramatically inclined than observed for B-DNA, and this is characteristic of A-DNA (Kang and Johnson, 1994).

Figure 3.3

CD spectra of DNA dodecanucleotides titrated with TFE. All spectra were recorded in buffers containing 10 mM Tris, pH 8.1, 0.02 mM EDTA, 4 mM NaCl and TFE (concentrations are labeled with each spectrum). DNA concentrations were adjusted to an absorbance of 0.4 at 260 nm. Spectra are shown for sequences (A) d(GGCGGCGGCGGC), (B) d(GCGCGCGCGCGC), (C) d(CCCCCGCGGGGG), and (D) d(CCCCGTACGGGG).

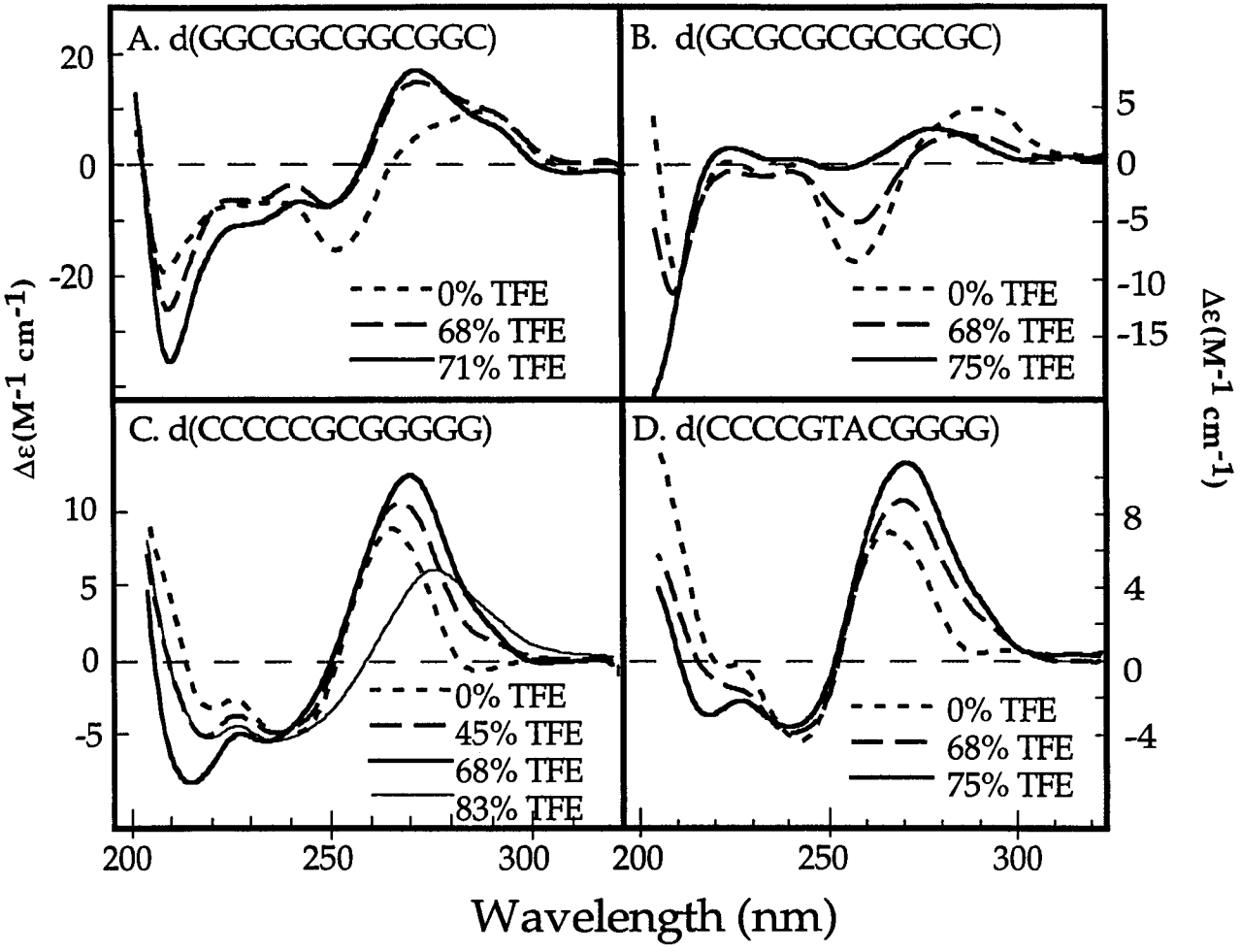

A. d(GGCGGCGGCGGC)
B. d(GCGCGCGCGCGC)
C. d(CCCCCGCGGGGG)
D. d(CCCCGTACGGGG)
0% TFE
68% TFE
71% TFE
75% TFE
45% TFE
83% TFE
Δε(M-1 cm-1)
Wavelength (nm)
200
250
300

Figure 3.3

The CD of d(CCCCCGCGGGGG) in buffer is similar to that of poly(dG)•poly(dC) in buffer (Gray and Bollum, 1974). Since the polymer is A-DNA in buffer, we conclude that our sequence is in the A-form as well. The sequence d(CCCCGTACGGGG), a somewhat weaker A-forming sequence (APE = -0.03 kcal/mol/bp), has a CD spectrum Figure 3.3d) in 0% TFE which is nearly identical to that of d(CCCCCGCGGGGG) in 0% TFE. The titration to A-DNA is analogous to d(CCCCCGCGGGGG). Thus, even though the APE for this sequence is only slightly negative, its behavior in solution is very similar to that of the standard A-DNA forming sequence d(CCCCCGCGGGGG). These results show that oligonucleotides can form A-DNA in the absence of organic precipitants, including MPD (Timsit and Moras, 1992).

3.5 Discussion

In this study, we show that sequences crystallized as A-DNA can be distinguished from those that crystallize as B-DNA by comparing the free energy required to hydrate the exposed DNA surfaces (as measured by A-DNA propensity energies, or APEs). From this, we derived an A-DNA triplet code that is better than 90% accurate in predicting sequences that crystallize as A- and B-DNA. Finally, we observed that sequences with positive APEs have typical B-DNA CD spectra in aqueous solution, while sequences with negative APEs showed characteristic A-DNA spectra, even

in the absence of TFE. The APEs, therefore, provide an accurate method to predict the crystal and solution conformations of short DNA sequences.

Can the APEs be useful for predicting A-DNA formation in a cell? Recently, A-DNA has been suggested to be involved in promoter recognition by *Escherichia coli* RNA polymerase (Warne and deHaseth, 1993), in proper phasing of CRP-binding sites (Barber et al., 1993), and in binding small acid-soluble proteins in dormant spores of *Bacillus* (Stetlow, 1992). A broader role may be found for A-DNA, but this form is difficult to detect within longer stretches of B-DNA (Ivanov and Krylov, 1992). Thus predictions based on thermodynamic rules may be highly useful, as we have seen with Z-DNA (Schroth et al., 1992). In order to apply the APE values derived here to predicting A-DNA in genomic sequences, we must include the free energy for forming junctions between A- and B-DNA (1.2 to 1.5 kcal/mol) (Ivanov and Krylov, 1992; Ivanov et al., 1974; 1985).

We had previously shown that SFE calculations can account for the sequence dependent stability of left-handed Z-DNA. Here, this same method is shown to accurately predict the relative stabilities of the two right-handed forms of DNA. Thus we can now begin to consider a more complex three-state equilibrium between A-, B- and Z-DNA.

3.6 Acknowledgments

This work was supported by grants from the American Cancer Society (NP-740C) and the National Science Foundation (MCB 9304467). GPS has been supported by a postdoctoral fellowship from the American Cancer Society (PF-3749). We wish to thank Dr. C. Robert for his input, and Prof. W. C. Johnson for help in recording and interpreting CD spectra.

Chapter 4

4. THE IDENTIFICATION OF A-DNA IN GENOMIC DNA SEQUENCES

Beth Basham and P. Shing Ho

Formatted for submission

4.1 Synopsis

As the amount of information in DNA sequences databases increases, the need for methods to evaluate the sequence information and identify nonstandard DNA structures becomes more important. One such tool is reported here: an algorithm, AHUNT, which uses a set of A-DNA propensity energies (APEs) to predict which sequences have a high propensity to form A-DNA. The solvent free energy differences for 41 sequences in the crystal structure database modeled as A-DNA versus B-DNA and information from a UV photofootprinting analysis of the 5S ribosomal RNA gene (Becker and Wang, 1989) were used as parameters in a genetic algorithm to derive a complete set of APEs. The APEs predict the conformation of the DNA in the crystal structures (78%), show an 86% correlation with the photofootprinting analysis and correlate with the titration behavior of short oligonucleotides in solution. AHUNT was used to evaluate genes from 7 species, showing that *E. coli* genes tend to have significantly less A-DNA than would be expected as compared to random DNA and eukaryotic genes. This type of comparison will be useful in exploring the biological significance of A-DNA and understanding the structurally dynamic nature of DNA.

4.2 Introduction

DNA forms many different structures *in vitro*. These include the standard double-stranded right-handed A- and B-DNA conformations as well as the left-handed structure of Z-DNA. Triple-stranded H-DNA, quadruple-stranded guanine quartets and cruciforms add to DNA's repertoire of structures (Rich, 1993). The importance of many of these structures biologically is not well understood in part because the ability to identify and predict non-B-DNA structures in biological systems is limited. The challenge lies in trying to define the energetic and environmental requirements of sequence-specific DNA conformations.

For some non-B-DNA structures this goal has been accomplished. A computer algorithm, ZHUNT, was developed to analyze sequences for stretches of DNA that have a high propensity to form the left-handed conformation of DNA, Z-DNA (Ho et al., 1986). This program has been used to evaluate genomic sequences (Futscher et al., 199X; Ho et al., 1986; Schroth et al., 1992) for Z-DNA. This type of analysis has shown that Z-DNA forming sequences are conspicuously absent from prokaryotic genes (Gross and Garrad, 1986) and are disproportionately located near the 5′ end of human genes (Schroth et al., 1992) suggesting a role for this conformation in eukaryotic transcriptional regulation. A similar analysis identified potential cruciforms and triple-stranded H-DNA in human genes, but not

in the *E. coli* genome, again, suggesting a functional role for these conformations in eukaryotes (Schroth and Ho, 1995). However, these represent only a small set of non-B-DNA structures that may be functionally important.

Another well studied conformation of DNA is A-DNA. A-DNA is a right-handed double-helical conformation of DNA that is structurally distinct from B-DNA. A-DNA is a shorter and wider helical structure than B-DNA (Figure 1.1) with a helical repeat of 11 base pairs per turn as compared to B-DNA's 10 to 10.5 base pairs per turn. It is, therefore, an underwound structure relative to B-DNA with an average helical twist of 33° versus 36° for B-DNA. Additionally, the distance between consecutive stacked base pairs is 0.7 Å shorter in A-DNA (2.7 Å) as compared to B-DNA (3.4 Å). The base pairs are displaced from the helix axis by about 5 Å in A-DNA, whereas in B-DNA they show little displacement. The bases are also highly inclined in A-DNA relative to B-DNA where there is no inclination of the bases. The ribose sugars are in the C3′ *endo* conformation in A-DNA crystal structures whereas in B-DNA they are C2′ *endo*. All of these parameters define A-DNA as an underwound structure with distorted bases and a wider circumference than B-DNA.

The energetic difference between A- and B-DNA has been estimated at 1.2 to 1.5 kcal/mol (Ivanov and Krylov, 1992; Ivanov et al., 1974; Ivanov et al., 1985) and subtle changes in the relative humidity or ion concentrations in the cell or supercoiling (Krylov et al., 1990) could induce a B-to-A

transition. This low energetic barrier suggests A-DNA could be a source of biological regulation at the DNA level. For example, a B-to-A transition could change the shape of and the distance between important protein binding sites in a gene.

Several putative biological activities for A-DNA have previously been identified. It has been suggested that A-DNA is involved in promoter recognition by *E. coli* RNA polymerase (Warne and deHaseth, 1993), in the phasing of CRP-binding sites (Barber et al., 1993; Ivanov et al., 1995), and in the binding of small proteins to DNA as a mechanism of dormancy in *Bacillus* (Stetlow, 1992). Crystallographic and spectroscopic analysis of DNA-ligand complexes and DNA-protein complexes provide several examples in which the A-DNA conformation is stabilized. In the complex of the TATA binding protein with DNA, the DNA in the TATA box has a structure that is more A-DNA like than B-DNA like (Guzikevich-Guerstein and Shakked, 1996). The complex of the SRY protein (sex-determining region Y, a HMG protein) with DNA in solution shows that the DNA assumes a conformation that has many characteristics of A-DNA (Werner et al., 1995). In the crystal structure of the antitumor drug, cisplatin, with DNA, the DNA on one side of the induced bend is in the A-form (Takahara et al., 1995). There is also evidence that A-DNA may be a functional link between the cellular environment (hydrophobicity, ionic strength, pH and polyamine concentrations) and DNA packaging (Reich et al., 1991). While A-DNA has been studied in detail spectroscopically and by

x-ray diffraction, there is as yet no straightforward method to detect A-DNA *in vivo*. This has limited efforts to monitor the occurrence of A-DNA in genomic DNA and hence identify its potential functions in the cell.

The conformation that DNA adopts is dependent on its nucleotide sequence and to some extent its environment. A-DNA was originally reported to be found under conditions of low water activity (low relative humidity) (Franklin and Gosling, 1953a,b) which immediately suggested that hydrophobicity plays a role in its stability. The base sequence rules that are most obvious are that poly(dG) and poly (dC) sequences favor A-DNA while d(A)/d(T) rich sequences favor B-DNA. However, the (G+C) content of a sequence is too simplistic a rule to identify potential A-DNA regions in genomic DNA. The overall ability of the (G+C) content rules to predict the crystal structure conformations is only 61% (Table 4.1). A method based on the free energy difference between the A- and B-conformations for different base stacks in a dinucleotide has been proposed by Ivanov's group (Ivanov et al., 1983; Minchenkova et al., 1986). In this method, the stability of a base step as A- or B-DNA is reflected in the relative humidity at the midpoint of the B-to-A transition (RH_{mid}) for the dinucleotide: CC/GG and AA/TT stacks have RH_{mid} of 88.3% and 66% respectively and all other stacks have RH_{mid} = 81%. However, the ability of this method to predict the formation of A-DNA in crystals is no better than from (G+C) content alone (Table 4.1).

Table 4.1

Comparison of methods to predict the sequence dependence of A- and B-DNA crystal structures.

	%(G+C)[d]	RH_{mid}[e] (%)	SFE[f] (kcal/mol/bp)
Criteria			
B_{range}[a]	57 ± 10	79.3 ± 2.9	0.23 ± 0.24
A_{range}[b]	76 ± 22	83.4 ± 2.7	-0.36 ± 0.28
Inconclusive range[c]	54 to 66	80.7 to 82.2	-0.01 to 0.08
Predictive Ability			
% Correct	61%	61%	78%
% Incorrect	8%	8%	15%
% Inconclusive	31%	31%	7%

[a]B_{range} is the average predicted value for sequences crystallized as B-DNA, ± 1 standard deviation.
[b]A_{range} is the average predicted value for sequences crystallized as A-DNA, ± 1 standard deviation.
[c]Inconclusive range is the values at which A_{range} and B_{range} overlap.
[d]%(G+C) refers to using (G+C) content only .
[e]RH_{mid} is the predicted relative humidity at the midpoint of the B-to-A transition and is calculated from the kinds of base stacks (Ivanov et al., 1983; Minchenkova et al., 1986).
[f]SFE is the solvent free energy method (Basham et al., 1995; Kagawa et al., 1993; 1989).

We have previously shown that a method based on solvent free energies (SFE), the approach that was successful in predicting Z-DNA (Kagawa et al., 1993; 1989), could also predict the sequence-dependent conformation of A-DNA seen in the crystal structures (Basham et al., 1995). The SFE method, which relies on the difference in solvent free energy for the sequence modeled as A versus B-DNA to distinguish between A- and B-DNA, shows a significant difference in the ΔSFE_{A-B} for sequences crystallized as A- versus B-DNA (Basham et al., 1995). The SFE method correctly predicts the conformation of 78% of the sequences using criteria in which the A-conformation is predicted for sequences with $\Delta SFE_{A-B} < -0.1$, and B-DNA is predicted for sequences with $\Delta SFE_{A-B} > 0.08$.

We had previously used this thermodynamic measurement based on hydrophobicity to develop a quantitative method to predict a sequence's stability as A-DNA (Basham et al., 1995). This is a set of A-DNA propensity energies (APEs) that represent the difference in solvent free energy (ΔSFE_{A-B}) for a particular base pair (in the context of its nearest 5′ and 3′ neighbors) in the A- versus the B-conformation (Basham et al., 1995). The APEs predict the sequence-dependent conformation of greater than 90% of the sequences that were crystallized and correlate well with the titration behavior of short oligonucleotides in solution. Unfortunately this set of A-DNA propensity energies is incomplete because not all possible triplets were represented in the crystal database at the time of its derivation.

Here we seek to develop a complete set of A-DNA propensity energies that can be extended to identify sequences with a high propensity to form A-DNA in genomic DNA. Our hypothesis is that if the APEs do indeed represent A-DNA stability then we should be able to correlate the SFE information with data from experiments that probe for A-DNA to develop a set of APEs that are generally useful for predicting the function of A-DNA. For the current work, we have re-derived the APEs from the $\Delta SFE_{A\text{-}B}$s for a larger set of crystallized sequences, and correlated these with quantitative results from a UV photofootprinting study of the 5S ribosomal RNA gene (Becker and Wang, 1989). In the UV photofootprinting study, 14 regions of the gene were shown to undergo the B-to-A transition independently of one another (Figure 4.1). This was determined by measuring the percent trifluoroethanol (TFE) at the midpoint of the titration from B-DNA to A-DNA (TFE_{mid}). These midpoints were not entirely correlated with the (G+C) content of the regions. We expect the APEs to better predict the observed transition.

The goal of the studies described here is to derive a set of APEs that incorporate information from both the DNA crystals and the natural DNA that can be generally applied to predicting A-DNA structural propensities under all circumstances. A genetic algorithm was developed to use the $\Delta SFE_{A\text{-}B}$s with the results of the UV photofootprinting analysis of the titration of the 5S gene to derive a set of APEs descriptive of both systems. This

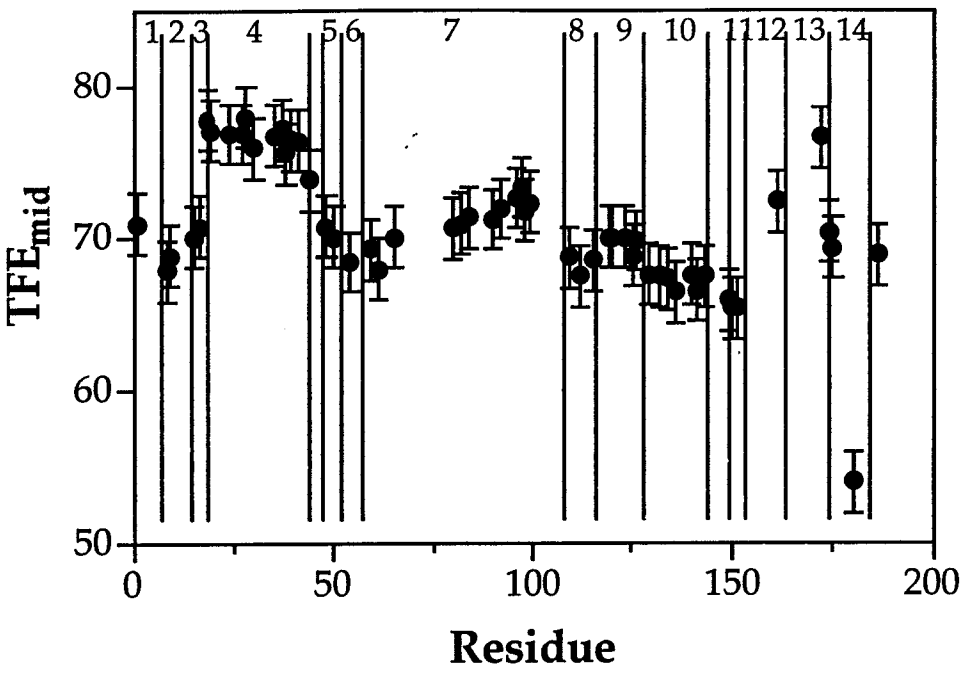


Figure 4.1

The percent TFE at the midpoint of the B-to-A transition (TFE_{mid}) for residues in the 5S rRNA gene as measured by UV photofootprinting (Becker and Wang, 1989). The midpoints ± the error are shown for the residues for which this measurement was done. Vertical lines indicate the 14 regions of the 5S rRNA gene that were identified as undergoing the B-to-A transition independently of one another by this technique.

algorithm yielded a complete set of APEs that predict the conformation of 78% of the sequences that have been crystallized as A- or B-DNA and are reasonably well correlated (86%) with the photofootprinting analysis of the 5S rRNA gene. We have developed an algorithm, AHUNT, which applies this set of APEs to search for regions with a high propensity to form A-DNA within genomes. Here we detail the derivation of the APEs and their application to locate A-DNA in DNA sequences from various eukaryotic and prokaryotic organisms.

4.3 Methods

4.3.1 Calculation of ΔSFE_{A-B}

The differences in solvent free energies between A- and B-DNA (ΔSFE_{A-B}) were calculated as previously described (Basham et al., 1995; Kagawa et al., 1989; 1993) for 41 DNA sequences in the Nucleic Acids Database (Berman et al., 1992). The A-DNA propensity energies (APEs) were derived by subtracting 0.5 kcal/mole/bp from the ΔSFE_{A-B} values. This is the energy at which A- and B-DNA are distinguishable by this method.

4.3.3 Genetic algorithm

The aim in deriving the APEs was to develop a complete and consistent solution set that predicts which sequences will crystallize as A-DNA and the midpoints of the B-to-A transition measured for the 5S rRNA gene by Becker and Wang. Specifically, the observations that were to be fit were the 41 $\Delta SFE_{A\text{-}B}$s and 11 TFE_{mid}s. This is a multivariable fitting problem in which 32 variables are to be fit to 52 equations. A genetic algorithm was used to solve for these variables. A genetic algorithm is an efficient, iterative, computational method of optimization which capitalizes on the principles of evolution to efficiently explore large areas of potential solution space (Cartwright, 1995; Coley, 1996; Holland, 1992). It is especially applicable to multivariable systems and to problems in which a straightforward or mathematically simple algorithm is not possible (Bohm, 1996). Genetic algorithms have been applied to many types of biological problems including RNA secondary structure prediction (Benedetti and Morosetti, 1995), the protein folding problem (Sun, 1993; Jones, 1994; Bowie and Eisenberg, 1994), and modeling protein/ligand docking (Jones et al., 1995).

The genetic algorithm was developed as a computational analogy to evolution. Thus the principles of evolution describe the genetic algorithms very well. According to the theory of evolution, genes are expressed in organisms and the ability of the organisms to survive in the environment is tested (survival of the fittest). The most fit organisms in a

population are most likely to pass their genes on to the next generation. The source of speed and efficiency in evolution is the recombination event, in which paired chromosomes exchange genetic material. This step is responsible for the rapid evolution of highly specialized organisms. In a genetic algorithm, solutions are represented as chromosomes, with each variable in the solution being a gene. The environment is the set of observations which the variables must describe. The expression of the genes is achieved by determining the fitness, which is a quantitative measurement of how well the genes model the observations to be fit.

The steps in a genetic algorithm are summarized in Figure 4.2. The first step is to randomly generate a set of potential solutions from random. The ability of each potential solution to model the observed data is then quantitatively evaluated and reported as the fitness. The best solutions are recombined with one another to generate new, and often dramatically more fit daughter solutions. The fitness of these daughter solutions are evaluated and then the population of solutions is reduced to its original size by retaining only the most fit solutions. This process is repeated until a solution that satisfies a minimum criteria is met.

4.3.2.1 Population

In the first step, a population of potential solutions (in this case, the APEs) to the problem are randomly generated. In this case, the population

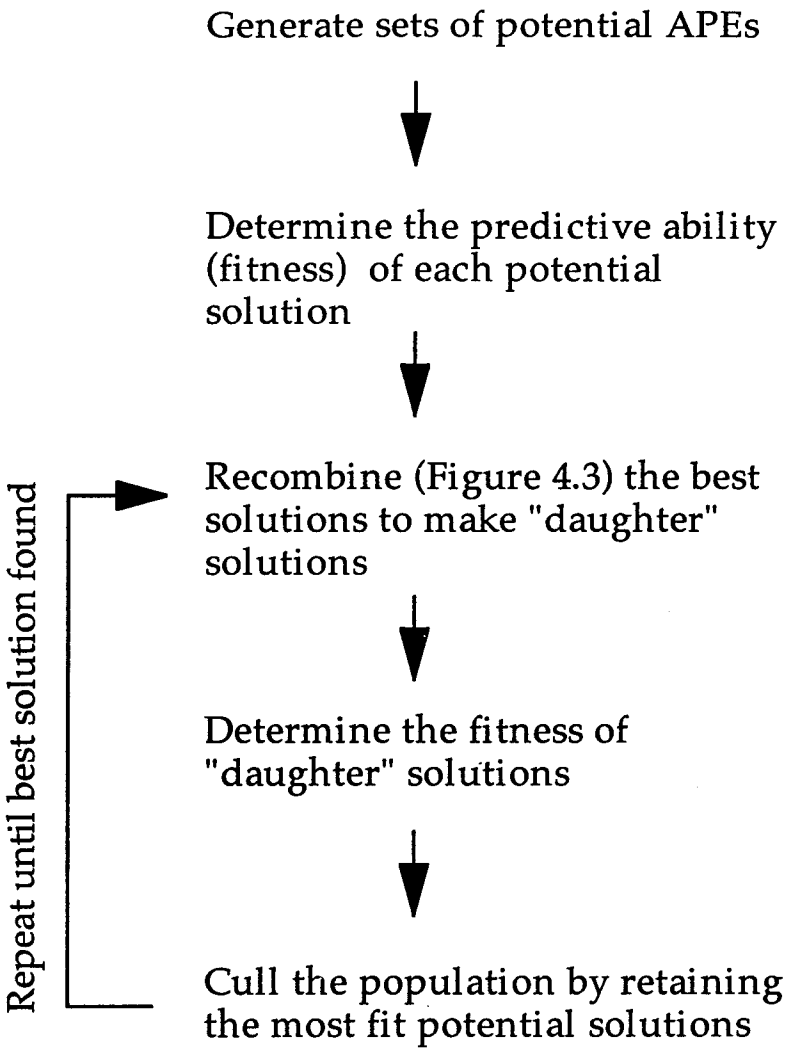


Figure 4.2

Schematic diagram of the genetic algorithm. This method was used to solve for a set of APEs consistent with both the ΔSFE_{A-B}s of the sequences that have been crystallized and the TFE_{mid} for the regions of the 5S gene.

Figure 4.3

Recombination produces new solutions. Each dot represents a variable (analogous to a gene) and potential solutions are shown as sets of dots (analogous to a chromosome). The number in parenthesis represents the fitness, of the predictive ability of the potential solution.

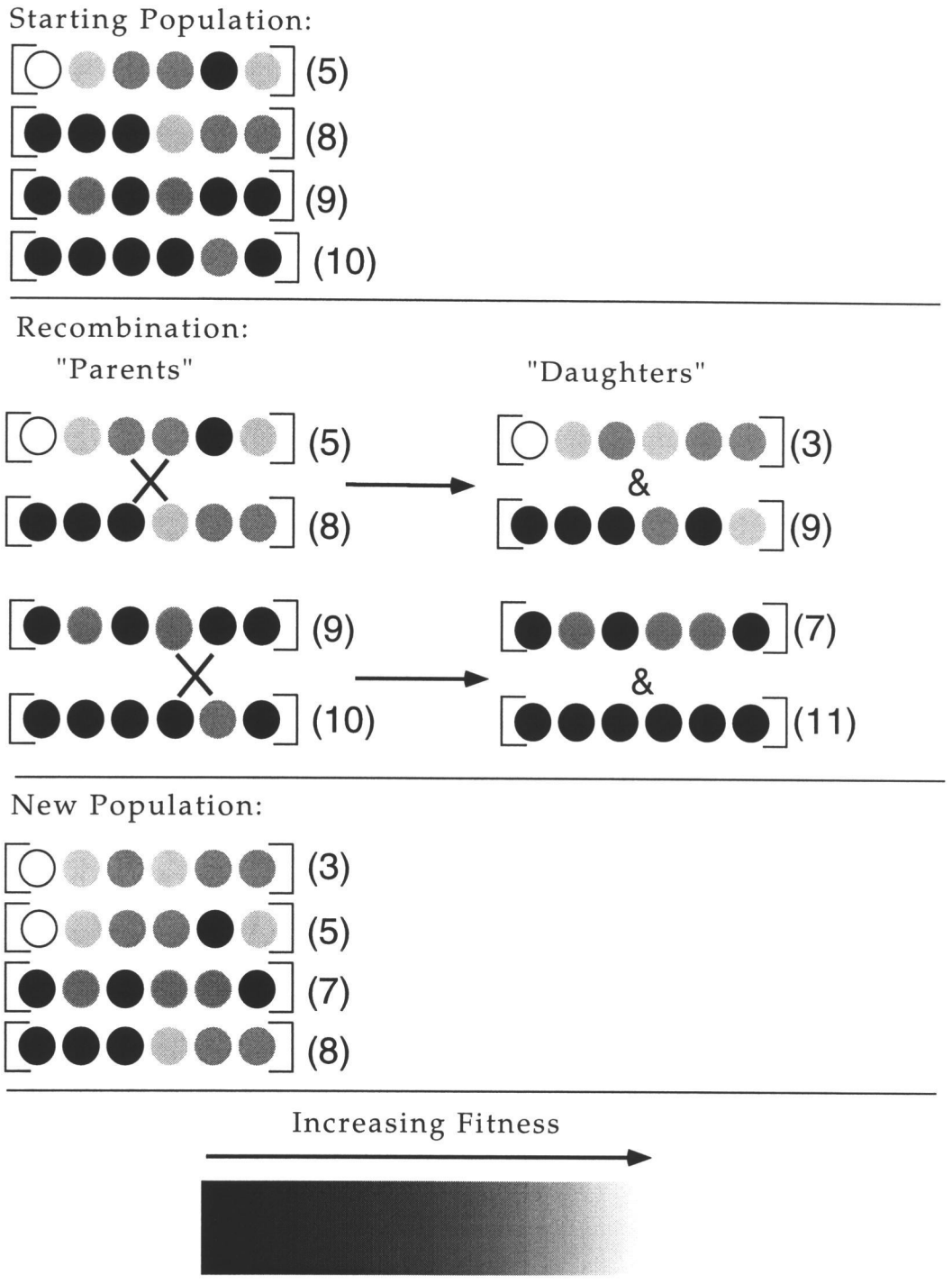

Starting Population:
(5)
(8)
(9)
(10)
Recombination:
"Parents"
"Daughters"
(5)
(8)
(3)
&
(9)
(9)
(10)
(7)
&
(11)
New Population:
(3)
(5)
(7)
(8)
Increasing Fitness

Figure 4.3

size was maintained at 5000 potential solutions. Each potential solution (which is analogous to a chromosome) was composed of 32 variables (or genes) which were the A-DNA propensity energies for each nucleotide in the context of all possible nearest 5′ and 3′ neighbors.

4.3.2.2 Fitness

Finding a solution with a genetic algorithm is highly dependent on the quality of the fitness function (analogous to the selective pressure in evolution) used to apply the solution to model the data. In the genetic algorithm applied to the 5S and SFE data, it was important to choose a function such that both data sets were appropriately weighted. Three factors were identified as important and were incorporated into the fitness function. First, the APEs had to model the ΔSFE_{A-B} of the sequences that have been crystallized as A- or B-DNA since we have previously shown that this value distinguishes between sequences that have been crystallized as A- versus B-DNA (Basham et al., 1995). To quantitatively determine how well a potential solution reflected the ΔSFE_{A-B} of the sequences in the crystal structure database, Fit_{SFE} was calculated as the mean absolute difference between the average APE (the prediction) and the ΔSFE_{A-B} (the observation) calculated for all 41 sequences in the data set.

$$Fit_{SFE} = (\Sigma\, Sqrt(\Delta SFE_{A-B} - APE)^2)/41$$

Second, since we had predicted that the APEs should be correlated with the %TFE at the midpoint of the titration to A-DNA (TFE_{mid}) of 11 regions of the 5S rRNA gene shown to undergo the B-to-A transition independently of one another (Becker and Wang, 1989), the fitness function must incorporate this correlation. Therefore, the fitness function includes a term that reflects the correlation of the APE of a region (APE_{region}) with the TFE_{mid} observed for the region. Specifically, the linear least squares best fit line was determined for the APE_{region} versus the TFE_{mid}:

$$APE_{region} = Slope^{*}(TFE_{mid}) + intercept.$$

and one minus the correlation coefficient for this line was used as the quantitative determinant of fitness for the 5S data, Fit_{5S}.

$$Fit_{5S} = 1\text{-}R^2$$

The final criteria in the fitness function is that the solution set of APEs that equally satisfies both the $SFE_{A\text{-}B}$ and the TFE_{mid} data. Therefore, the fitness function includes a term for the absolute difference between the two fits, Fit_{Diff}:

$$Fit_{Diff} = |\, Fit_{SFE} - Fit_{5S} \,|$$

In summary, the fitness function used in this genetic algorithm is:

$$Fitness = Fit_{5S} + Fit_{SFE} + Fit_{Diff}$$

4.3.2.3 Recombination

Recombination is the most important step in the genetic algorithm and is responsible for the dramatic increase in search speed and search space achieved with genetic algorithms. This step is analogous to recombination that occurs in natural DNA; that is parts of one potential solution are exchanged for parts of another potential solution. The most fit solutions (i.e. those with lowest fitness) have the highest probability of undergoing recombination with one another. Therefore, good parts of different potential solutions are brought together (Figure 4.3) to form solutions that may be better than either parent.

4.3.2.4 Mutation

Mutation is used in genetic algorithms to fine-tune the solutions. In this algorithm, the mutation rate was a function of the number of recombining potential solutions. Mutations were introduced by substituting a gene (APE) with a new, random value during the recombination step.

4.3.2.5 Culling the population

Borrowing the "survival of the fittest" concept from evolution, the population was maintained at 5000 potential solutions by retaining only

the most fit potential solutions after each recombination step. Culling the population ensured that the system was evolving toward better solutions and is necessary for maintaining the speed of the algorithm.

4.3.2.6 Termination

The algorithm was terminated when there was no significant change in the population fitness. This was after 1000 cycles.

4.3.3 TFE titrations of DNA oligonucleotides

As an independent test of the predictive ability of the APEs, the titration behavior in trifluoroethanol (TFE) of specific oligonucleotides was measured with circular dichroism. Circular dichroism is extremely sensitive to changes in the structure of DNA. The amount of TFE needed to induce the B-to-A transition in a specific sequence should be proportional to the A-DNA propensity energy of the sequence. Dodecameric oligonucleotides were titrated with TFE as previously described (Basham et al., 1995). Synthetic oligonucleotides were desalted over a Sephadex G-25 column and annealed to form duplexes in 3.33 mM Tris (pH 8.1) buffer with 0.03 mM EDTA and 50 mM NaCl. All sequences were diluted to 0.4 absorbance units in 10 mM Tris (pH 8.1), 0.02 mM EDTA and various amounts of TFE was added to bring the TFE concentration to five different

target concentrations ranging from 0% to 83% TFE. The maximum amount of TFE used in each titration was limited by the solubility of the DNA sequences. Circular dichroism spectra were recorded at each TFE concentration on a JASCO J-720 spectrometer at room temperature in a 1 mm pathlength cell.

4.3.4 AHUNT

In order to apply the APEs to locate regions of DNA in genomic sequences that have a high propensity to form A-DNA, AHUNT was developed. AHUNT calculates the average APE for a residue centered in an *n* residue window. The window incorporates the effects of neighboring bases on the residue being evaluated. The average APE is the average of the APEs of all the triplets in the window. In this case, a window size of seven residues was used which represented five triplets.

The average APEs were then evaluated to determine which residues form regions that are distinct from other regions of the sequence (Figure 4.4). This was done by calculating the rate of change of the APEs as a function of the residues. This is roughly the first derivative of the average APE as a function of the residue. Residues at the maximums or minimums of the derivative curve define boundaries between potential regions. Potential regions are then assigned conformations: strong A (if $<APE_{region}>$ is less than -0.2), strong B ($<APE_{region}>$ is greater than 0.2) or

intermediate (a/b) (<APE_{region}> is between -0.2 and 0.2). The cutoffs for A-DNA were derived from the energy needed to nucleate a B-to-A transition which has been estimated at 1.2 to 1.5 kcal/mol (Ivanov and Krylov, 1992; Ivanov et al., 1974; 1985). Therefore a seven residue sequence with an <APE> of -0.2 kcal/mol/bp has enough energy to overcome the nucleation barrier to the formation of the A-DNA conformation. Since AHUNT only assigns regions to sequences at least 5 residues long (with the ends of the regions excluded), an APE of -0.2 kcal/mol/bp provides enough energy for this transition. Regions that are predicted to be strong A- and B-formers will be categorized as A-DNA and B-DNA respectively while regions that do not strongly favor A- or B-DNA (i.e. those that have APEs close to 0 kcal/mol/bp) will be categorized as a/b-DNA. It must be noted that regions that are predicted to be a/b-DNA are presumably in the B-DNA conformation; however, these regions do not favor B-DNA as strongly as those assigned as B-DNA and these sequences should convert to A-DNA more readily than regions identified as strong B-DNA regions.

In order to compare sequences with different (G+C) contents and to compare genomic DNA to random DNA, the amount of A- B- and a/b-DNA was calculated for 30 random sequences, each 1500 residues long with (G+C) contents ranging from 30% to 75% in 5% increments. Individual genes were then compared to random sequences with comparable (G+C) content and the amount of enhancement or suppression of a particular conformation (A-, B- or a/b-DNA) in the gene was assessed relative to

Figure 4.4

The determination of A- and B-DNA regions using AHUNT. First derivatives are calculated as the average over 5 consecutive residues. Shaded regions are those predicted to be A-DNA ($APE_{region} < -0.2$ kcal/mol/bp) and hashed regions are those predicted to be B-DNA ($APE_{region} > 0.2$ kcal/mol/bp).

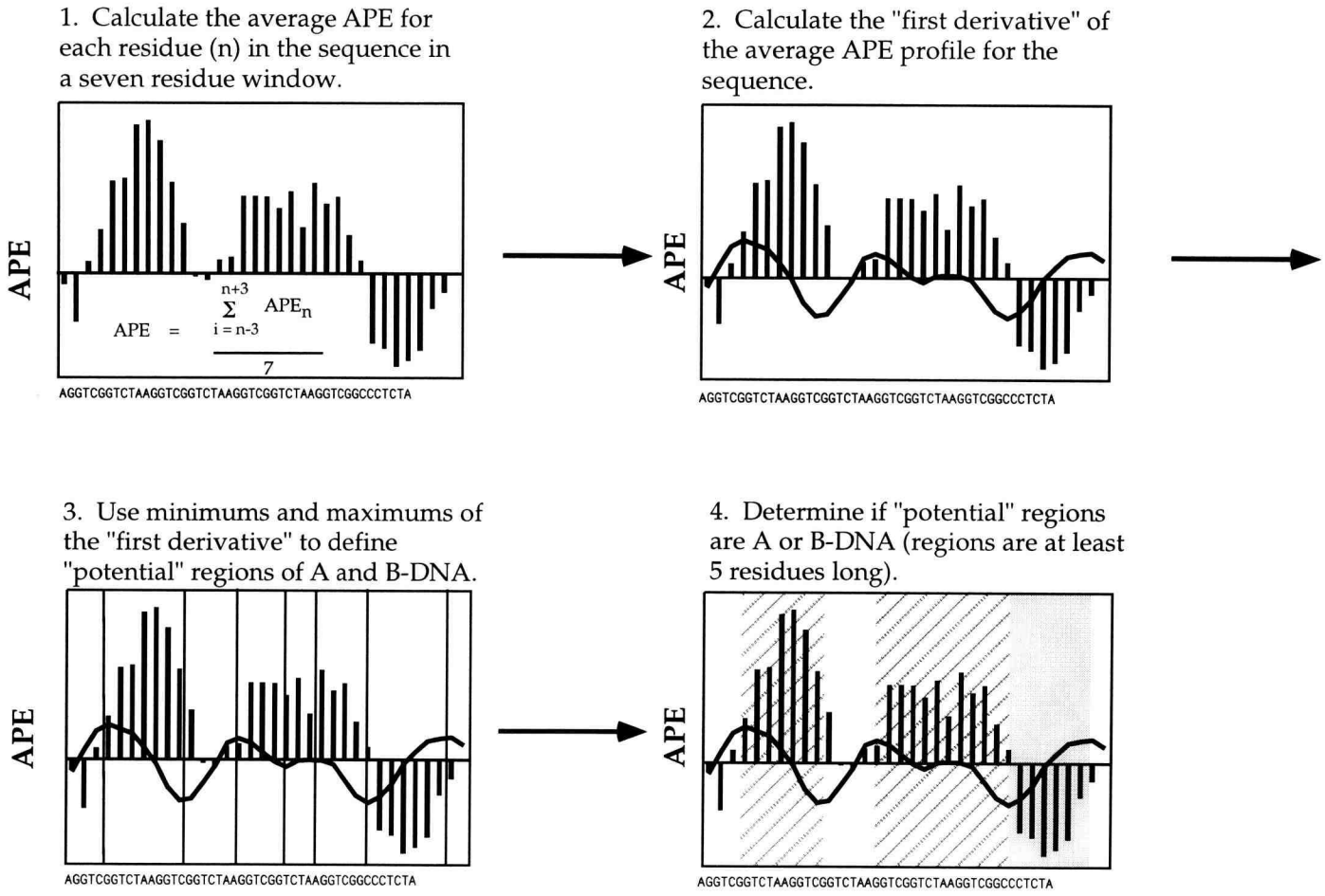

1. Calculate the average APE for each residue (n) in the sequence in a seven residue window.
APE
$APE = \frac{\sum_{i=n-3}^{n+3} APE_n}{7}$
AGGTCGGTCTAAGGTCGGTCTAAGGTCGGTCTAAGGTCGGCCCTCTA
2. Calculate the "first derivative" of the average APE profile for the sequence.
APE
AGGTCGGTCTAAGGTCGGTCTAAGGTCGGTCTAAGGTCGGCCCTCTA
3. Use minimums and maximums of the "first derivative" to define "potential" regions of A and B-DNA.
APE
AGGTCGGTCTAAGGTCGGTCTAAGGTCGGTCTAAGGTCGGCCCTCTA
4. Determine if "potential" regions are A or B-DNA (regions are at least 5 residues long).
APE
AGGTCGGTCTAAGGTCGGTCTAAGGTCGGTCTAAGGTCGGCCCTCTA

Figure 4.4

random DNA with the same (G+C) content. Thus while genes with higher (G+C) contents are more likely to have more A-DNA, this method identifies those genes that are predicted to have significantly more or less of a particular conformation as compared to random DNA of the same (G+C) content.

4.3.5 The application of AHUNT to analysis of genomic DNA

Sets of genes from humans, *E. coli, Arabadopsis thaliana, Saccharomyces cerevisiae, Sulfolobus solfataricus, Thermus aquaticus* and *Chlamydia* sp. (Table 4.2) were analyzed for the amount of the gene that is predicted to be in strong A-DNA, B-DNA or a/b-DNA regions. Only sequences determined from genomic DNA sequences were analyzed.

4.4 Results

We had previously developed a thermodynamic method based on hydration free energies that predicts the conformation of right-handed DNA sequences in crystals and correlates with the titration behavior of short oligonucleotides in solution (Basham et al., 1995). The thermodynamic method yielded a set of A-DNA propensity energies (APEs) which represent the stability as A-DNA of a base at a position *n* in the sequence in the context of its nearest 5′ (*n-1*) and 3′ (*n+1*) neighbors. The APEs were

Table 4.2

Species analyzed with AHUNT for the amount of A- B- and a/b-DNA

Species	# of genes	<length>[a]	<%(G+C)>[b]	<% in regions>[c]
Homo sapiens	154	6733	52	63
Arabadopsis thaliana	24	3424	39	75
Saccharomyces cerevisiae	35	2595	38	76
Escherichia coli	104	2153	51	67
Sulfolobus solfataricus	32	2136	37	75
Thermus aquaticus	22	1764	68	59
Chlamydia sp.	24	1661	42	72

[a]<length> is the average length (base pairs) for the set of genes analyzed.
[b]<%(G+C)> is the average (G+C) content for the set of genes analyzed.
[c]<% in regions> is the average percent of the gene that was assigned to A-B- or a/b-DNA regions.

derived from the sequence dependent difference in solvent free energy for sequences modeled as A-DNA versus B-DNA (ΔSFE_{A-B}). This value distinguishes between sequences which have been crystallized as A- or B-DNA (Basham et al., 1995).

Unfortunately this set of APEs is incomplete and some triplets are not well represented due to the composition of the nucleic acids database at the time of the original derivation. In order to complete the APE table and at the same time incorporate sequence-dependent structural effects observed in biological DNA into the model, we incorporated the work of Becker and Wang who reported a UV photofootprinting analysis of 185 nucleotides in the 5S ribosomal RNA gene from *Xenopus laevis*. These studies identified 14 regions of the gene that undergo the B-to-A transition independently of one another (Becker and Wang, 1989). This technique assumes that DNA residues in the A-DNA conformation are less susceptible to UV photoproduct formation than residues in the B-conformation, presumably because of the reduced distance and improper geometry between consecutive stacked residues in A-DNA. The 5S gene was photofootprinted at various concentrations of trifluoroethanol (TFE). TFE is known to induce the B-to-A transition (Sprecher et al., 1979). The percent TFE at the midpoint of the B-to-A transition (TFE_{mid}) was determined for about one third of the residues in the gene (Figure 4.1). TFE_{mid} for a residue should be correlated with the energy required to convert the region

containing that base pair to A-DNA, and hence with the A-DNA propensity of the region.

Becker and Wang reported the average TFE_{mid} for the residues in the 11 of the 14 regions identified that undergo the B-to-A transition independently of one another. For 8 of these regions, the average TFE_{mid} over the region is correlated with the %(G+C) in the region (Figure 4.6a). However, this relationship does not apply for 3 of the regions. Since (G+C) content in itself does not explain the sequence-dependent conformation observed in the crystal structure database or the UV photofootprinting results, we attempted to use the APEs to model the relationship between sequence and TFE_{mid}. TFE_{mid} should be proportional to the sequence's propensity to form A-DNA; that is, sequences with a higher propensity to form A-DNA (and thus more negative APEs) should require less TFE to titrate to A-DNA. Likewise sequences with more positive APEs, which should therefore favor B-DNA, should require more TFE. Unfortunately we could not directly extend the APEs to test this since the APEs were not complete for all possible base triplets.

However, the UV footprinting experiment does provide quantitative sequence-dependent information for the B-to-A transition and we sought to incorporate that information with the $\Delta SFE_{A\text{-}B}$s of the crystal structure data set. In order to derive a self-consistent set of APEs that should be applicable to both measurements, it was necessary to develop a means of determining all the APEs simultaneously. A genetic algorithm

was developed for this purpose. Our hypothesis in applying a genetic algorithm to this problem was that the two different data sets (the 5S data and ΔSFE_{A-B}s) would be modeled by the same set of APEs if both methods are indeed quantitative measurements of A-DNA stability. The genetic algorithm provided an efficient method to incorporate both sets of measurements in the refinement of the APEs (Figure 4.2).

4.4.1 The genetic algorithm identifies a set of APEs that reflect both the SFE and the UV footprinting information

A genetic algorithm is a nonlinear minimization technique that relies on the principles of evolution to search large areas of solution space efficiently. In the algorithm used here, an initial pool size of 5000 potential solutions converged after 1000 cycles (Figure 4.5) on the set of APEs shown in Figure 4.7. The final fitness of 0.3 is consistent with an estimated 15% error in the measurements of TFE_{mid} and the calculation of ΔSFE_{A-B} or a net prediction of 85% of the data. Three independent runs of the algorithm starting with randomly assigned starting values resulted in the same set of APEs. These APEs predict the ΔSFE_{A-B}s, and 5S titrations.

The APEs predict the conformation of 18/21 sequences that have been crystallized as A-DNA and 16/20 sequences that have been crystallized as B-DNA (Table 4.3). This is a total prediction rate of 83% for DNA conformation in crystals. Of the 7 sequences that are not unambiguously predicted, 3 have APEs close to 0 indicating that the sequence does not have

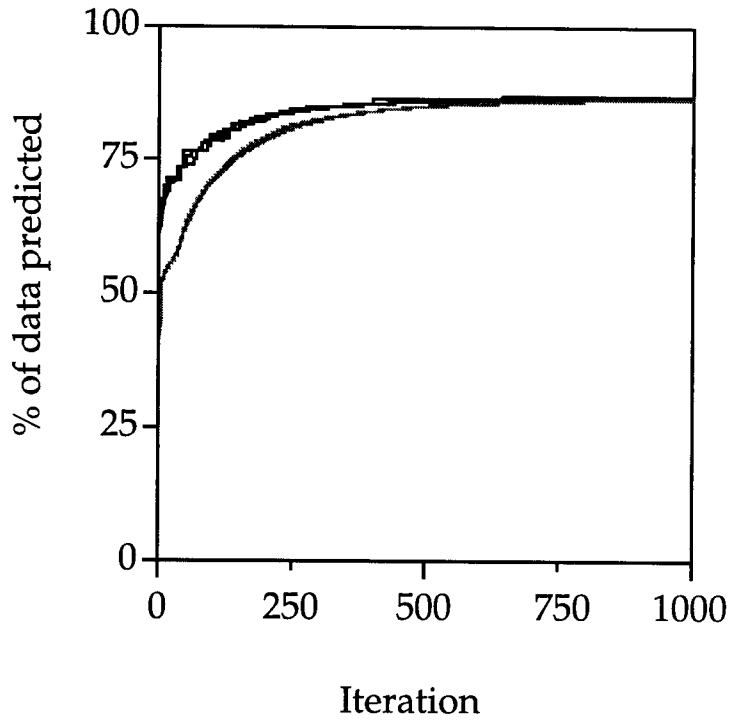


Figure 4.5

Summary of genetic algorithm runs. The % of the observed data that the APEs generated with the genetic algorithm predict as a function of the iteration cycle. Three separate runs are shown. The dark line is the fitness of the most fit solution at each step and light line is the average fitness for the whole population at the step. An initial population size of 5000 potential solutions was used in each run. The maximum prediction is 100% and the minimum is 0%.

Table 4.3

ΔSFE_{A-B}, <APE> and predicted conformations for DNA sequences used to derive the APEs

A-DNA Sequence	ΔSFE_{A-B}	<APE>	Predicted Conformation	B-DNA Sequence	ΔSFE_{A-B}	<APE>	Predicted Conformation
GCCGGC	-0.91	-0.21	A	CCAGGCCTGG	0.08	0.23	B
GGGGCCCC	-0.33	-0.44	A	CCAACGTTGG	-0.13	0.04	B
GGGATCCC	-0.2	-0.12	A	CGATCGATCG	0.05	0.23	B
GCCCGGGC	-0.62	-0.31	A	CGATTAATCG	0.22	0.53	B
GGTATACC	-0.14	0.15	B	CGATATATCG	0.34	0.62	B
CCCCGGGG	-0.59	-0.36	A	CCGGCGCCGG	0.20	0.01	A/B
CTCTAGAG	-0.54	-0.10	A	CATGGCCATG	-0.13	-0.01	A/B
GTACGTAC	-0.45	-0.29	A	CCAAGCTTGG	-0.11	-0.24	A
GGGCGCCC	-0.36	-0.09	A	CCATTAATGG	0.17	0.37	B
GGGTACCC	-0.11	-0.35	A	CTCTCGAGAG	-0.12	-0.33	A
ATGCGCAT	-0.6	-0.17	A	CCACTAGTGG	0.26	0.48	B
GTGTACAC	-0.38	0.00	A/B	CGCGAATTCGCG	0.67	0.48	B
GTCTAGAC	-0.46	-0.17	A	CGCAAAAAAGCG	0.37	0.28	B
GTGCGCAC	-0.58	-0.09	A	CGCATATATGCG	0.48	0.30	B
GAAGCTTC	-0.65	-0.18	A	CGCAAAAATGCG	0.47	0.27	B
GGCCGGCC	-0.58	-0.25	A	CGTGAATTCACG	0.38	0.51	B
ACGTACGT	-0.34	0.15	B	CGCAAATTTGCG	0.12	0.29	B
ACCGGCCGGT	-0.01	-0.13	A	CGTAGATCTACG	0.31	0.04	B
CCCGGCCGGG	0.13	-0.25	A	CGCGAAAAAACG	0.31	0.53	B
GCGGGCCCGC	0.21	-0.09	A	CGCGTTAACGCG	0.63	0.41	B
CCCCCGCGGGGG	-0.11	-0.17	A				

A-DNA and B-DNA sequences are the sequences in the NDB (Berman et al., 1992) that have been crystallized as A- or B-DNA. <APE> was calculated from the APEs in Figure 4.7.

Table 4.4

APEs (calculated from Figure 4.7) and predicted and observed conformations for sequences not used in the derivation of the APEs

Sequence	Crystal Conformation	APE (kcal/ mol/bp)	Predicted Conformation
GGCATGCC	A	-0.47	A
CCCTAGGG	A	0.22	B
GGATGGGAG	A	-0.22	A
CCGGGCCCGG	A	-0.25	A
GCACGCGTGC	A	-0.01	A/B
CTCGAG	B	-0.34	A
CGCTAGCG	B	0.20	B
CGCAATTGCG	B	0.19	B
CGCGATATCGCG	B	0.51	B
CGCTCTAGAGCG	B	-0.01	A/B

a strong thermodynamic preference for either conformation. For the sequences not used in the derivation of the APEs, the APEs predict the correct conformation of 6/10 of the sequences (Table 4.4). Two sequences have APEs that are ambiguous and the other two sequences are predicted to be the incorrect conformation. One of the sequences crystallized as B-DNA that was incorrectly predicted by the APEs to be A-form, d(CTCGAG), was observed to be a highly distorted B-DNA structure (Wahl et al., 1996).

The APEs calculated using the genetic algorithm are well correlated with the midpoint of the B-to-A transition over the region as determined by UV footprinting. Figure 4.6 shows a plot of TFE_{mid} determined from UV footprinting as a function of %(A+T) in the region and as a function of the average APE over the region ($\langle APE_{region} \rangle$). The correlation (reported as R^2) is 58% with %(A+T) whereas the APEs derived with the genetic algorithm yielded $\langle APE_{region} \rangle$ values that are correlated with an R^2 of 86% with TFE_{mid}.

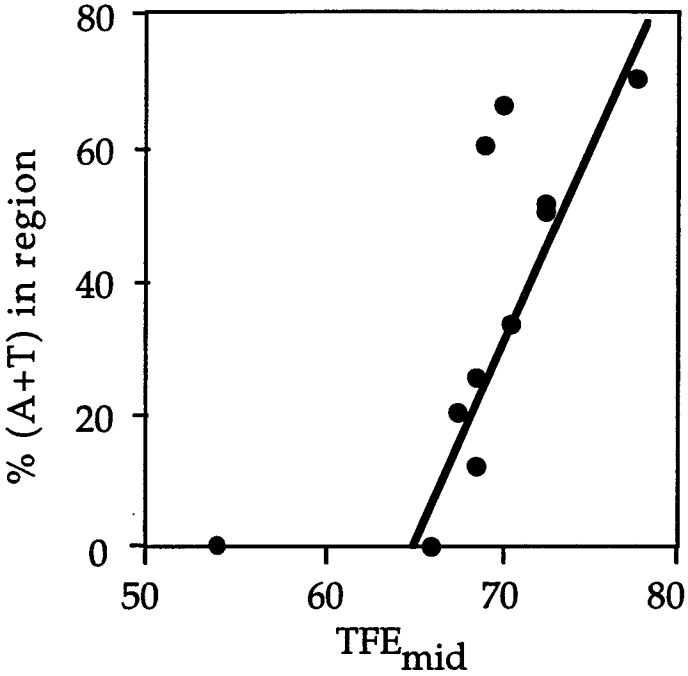
4.4.1.1 Description of APE trends

The APEs derived with the genetic algorithm (Figure 4.7) show many of the trends apparent in the original APEs (Basham et al., 1995). For example, the triplets CCC/GGG, GCA/TGC and TAC/GTA still have a strong propensity to form A-DNA, while the triplets CGC/GCG, GAT/ATC and AAA/TTT still favor B-DNA. Additionally, the trend that (G+C) rich

Figure 4.6

The APEs calculated with the genetic algorithm predict TFE_{mid} for 11 regions of the 5S rRNA gene. TFE_{mid} is the %TFE at the midpoint of the B-to-A transition. a. The correlation coefficient for (R^2) for the %(A+T) in the region and the TFE_{mid} measured for the region (Becker and Wang, 1989) is 58%. b. The correlation coefficient for (R^2) for the <APE> of the same regions and the TFE_{mid} measured for the regions (Becker and Wang, 1989) is 86%.

a.


80
60
40
20
0
% (A+T) in region
50
60
70
80
TFE$_{mid}$

b.

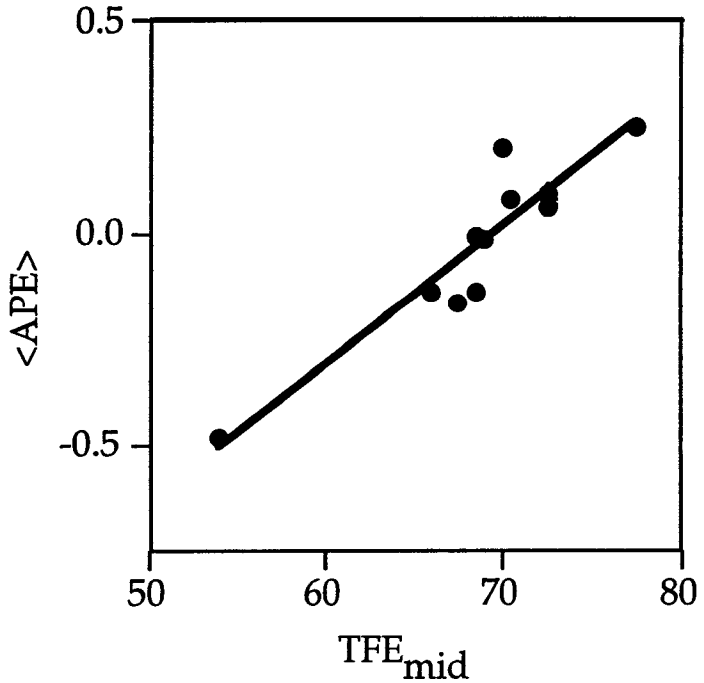

0.5
0.0
-0.5
<APE>
50
60
70
80
TFE$_{mid}$

Figure 4.6

N_{i-1}	N_i								N_{i+1}
	C		G		A		T		
C	-0.50	(42)	0.55	(31)	0.70	(10)	-0.68	(10)	C
	-0.09	(32)	-0.09	(32)	1.05	(6)	1.05	(6)	G
	-0.64	(17)	0.00	(18)	0.79	(13)	0.34	(13)	A
	0.82	(11)	0.24	(17)	0.46	(16)	-0.82	(11)	T
G	-0.33	(28)	-0.33	(28)	-0.87	(7)	-0.87	(7)	C
	0.55	(31)	-0.50	(42)	-0.68	(10)	0.70	(10)	G
	-1.53	(15)	-0.32	(8)	0.58	(11)	-0.55	(21)	A
	-0.29	(13)	0.00	(16)	0.46	(17)	-0.22	(9)	T
A	0.00	(16)	-0.29	(13)	-0.22	(9)	0.46	(17)	C
	0.24	(17)	0.82	(11)	-0.82	(11)	0.46	(16)	G
	-0.16	(4)	0.03	(11)	0.89	(16)	1.01	(16)	A
	1.50	(6)	1.50	(6)	0.74	(13)	0.74	(13)	T
T	-0.32	(8)	-1.53	(15)	-0.55	(21)	0.58	(11)	C
	0.00	(18)	-0.64	(17)	0.34	(13)	0.79	(13)	G
	0.28	(8)	0.28	(8)	0.92	(12)	0.92	(12)	A
	0.03	(11)	-0.16	(4)	1.01	(16)	0.89	(16)	T

Figure 4.7

The A-DNA triplet code of A-DNA propensity energies (APEs) (in kcal/mol per bp). Values for base pair triplets are for the central base pair (N_i) in the context of the 5′ flanking (N_{i-1}) and 3′ flanking (N_{i+1}) base pairs. The number of times each unique triplet is represented in the combined data sets is shown in parentheses. Triplets with an APE ≤ -0.2 strongly favor A-DNA, while triplets with an APE ≥ 0.2 strongly favor B-DNA.

triplets favor A-DNA and that (A+T) rich triplets favor B-DNA still holds (16/32 and 24/32). However, the APEs also predict that the (G+C) rich triplets CCT/AGG, CGC/GCG, CAC/GTG and CAG/CTG favor B-DNA while the (A+T) rich triplets AAG/CTT and TAC/GTA are more stable as A-DNA, confirming the observation that (G+C) content in itself does not adequately predict A-DNA stability.

4.4.1.2 The APEs correlate with the titration behavior of short oligonucleotides in solution

As an independent test of the APEs, we have compared the APEs with the titration behavior of short oligonucleotides in solution. We had previously compared the titration behavior of DNA oligonucleotides with TFE to the sequences' APEs (Basham et al., 1995). Sequences with negative APEs (which should favor A-DNA) have different titration profiles than sequences with positive APEs. Some sequences with negative APEs have CD spectra consistent with that of A-DNA throughout the titration because the normal 280 nm to 270 nm peak shift is not observed. With the exception of $d(AGC)_3$, other sequences with negative APEs required 75% or less TFE to convert to A-DNA. Sequences which were predicted to favor B-DNA do not appear to convert to A-DNA, and the sequence d(GCGCGCGCGCGC) which has a very positive APE (0.55 kcal/mol/bp), actually converted to Z-DNA in the presence of high amounts of TFE.

Table 4.5

Conformations of dodecanucleotides in aqueous solution and at high concentrations of TFE

Sequence	APE	Predicted Conformation	Conformation 0% TFE	Conformation High TFE
$d(G_{12})/d(C_{12})$	-0.50	A	A-like	A (68%)
$d(C_4G_4C_4)/d(G_4C_4G_4)$	-0.38	A	A-like	A (68%)
d(GGGGCCGGCCCC)	-0.35	A	B	A (75%)
$d(AGC_3)/d(TCG_3)$	-0.29	A	B	A (83%)
d(CCCCGTACGGGG)	-0.28	A	A-like	A (68%)
d(GGCCGGCCGGCC)	-0.23	A	B	A (75%)
d(CCCGTCGACGGG)	-0.22	A	A-like	A (75%)
d(CCCCCGCGGGGG)	-0.20	A	A-like	A (75%)
d(GGCGGCGGCGGC)	0.01	a/b	B	A (71%)
d(CGATACGTATCG)	0.23	B	B	B-like (83%)
d(CGCATATATGCG)	0.30	B	B	B-like (83%)
d(GCGCGCGCGCGC)	0.55	B	B	Z (75%)

Predicted conformations were determined from the APEs (Figure 4.7) and the conformations were determined by CD spectroscopy.

With the exception of the sequence d(GGCGGCGGCGGC) which has an APE of 0.01 kcal/mol/bp, and is therefore inconclusive, and the sequence d(AGCAGCAGCAGC), the APEs predict the titration behavior of the short oligonucleotides very well (Table 4.5). The overall predictive ability of the APEs for the titration behavior of the oligonucleotides is 10/12 or 83%.

4.4.2 AHUNT: an application of the APEs to predict gene structure

4.4.2.1 Testing and calibration of AHUNT with the 5S gene

How useful are the APEs in predicting the conformation of DNA in genomes? To answer this question, we developed the program, AHUNT, which uses the APEs to determine which parts of DNA gene sequences have a high propensity to form A-DNA regions. The APEs corroborate Becker and Wang's conclusion that the 5S rRNA gene contains regions that undergo the B-to-A transition independently of one another (Becker and Wang, 1989) (Figure 4.8). Specifically, AHUNT identifies 13 different regions in this gene and recognizes the independent regions identified by UV footprinting. AHUNT predicts more regions in the 5S rRNA gene than UV footprinting did because AHUNT has nucleotide-level resolution. In the UV photofootprinting assay, TFE_{mid} was measured for only one third of the residues in the sequence. AHUNT parameters were optimized by

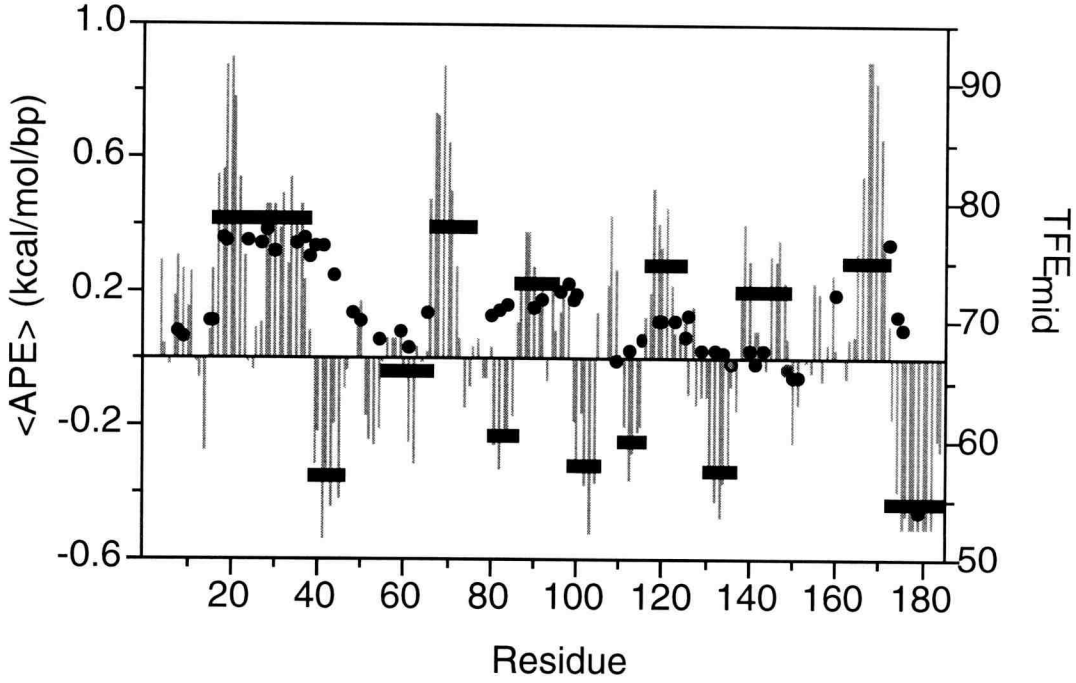


Figure 4.8

AHUNT's predictions correlate with the observed TFE_{mid} for the 5S rRNA gene. Solid vertical bars denote the <APE> (average APE) (kcal/mol/bp) calculated with AHUNT from the APEs (Figure 4.7) for the 5S ribosomal RNA gene using a window size of 7 nucleotides. • represents the % TFE at the B-to-A transition midpoint (TFE_{mid}) (Becker and Wang, 1989). Horizontal bars represent the separate regions that AHUNT identified. Regions with $<APE_{region}> > 0.2$ kcal/mol/bp are predicted to be strong B-DNA formers and regions with $<APE_{region}> < -0.2$ kcal/mol/bp are predicted to be strong A-DNA forming regions. Intermediate regions (a/b-DNA) have -0.2 kcal/mol/bp $< <APE_{region}> < 0.2$ kcal/mol/bp.

attempting to predict these regions using different window sizes. A window size of 7 residues was determined to be optimal. Shorter windows resulted in many more regions and a longer window predicted too few regions.

4.4.2.2 A-DNA is not localized to a specific part of human genes

In order to apply the APEs to identify biological activities for A-DNA, AHUNT was used to address two questions. First, is A-DNA localized to specific parts of the gene? If A-DNA is more frequent in part of the gene and less frequent in another, that would suggest a specific function. For example, analysis of the set of human genes with ZHUNT showed that Z-DNA forming sequences are disproportionately located near the 5′ ends of the genes. For this analysis, the genes analyzed were divided into four parts: promoter, intron, exon and 3′ flanking (3′ of the poly-A tail). The (G+C) content, the frequency of A-DNA regions and the percent of the gene that AHUNT identified as being in strong A-DNA regions were calculated for the distinct parts of 154 human genes (Figure 4.9). There are no significant differences in any of these measurements between the four different gene parts.

AHUNT does identify some human genes that have a significant amount of A-DNA (Table 4.6). The c-fos proto oncogene contains 2.3% more A-DNA in regions than a random sequence with comparable (G+C)

Figure 4.9

Distributions of characteristics of A-DNA regions in four sections of genes for the set of 154 human genes. 5′ is the sequence 5′ of the transcription start site. Intron and Exon refer to intron and exon sequences. 3′ is the region of the sequence after the putative poly A tail. Distributions are shown as box and whisker plots: The box represents 50 percent of the data, the center line is the median and the whiskers are the minimum and maximum values. a. the percent (G+C) for the different sections of the genes. b. the number of A-DNA regions per 100 residues. c. the percent of the gene section that AHUNT identifies as being a strong A-DNA forming region (%A-DNA).

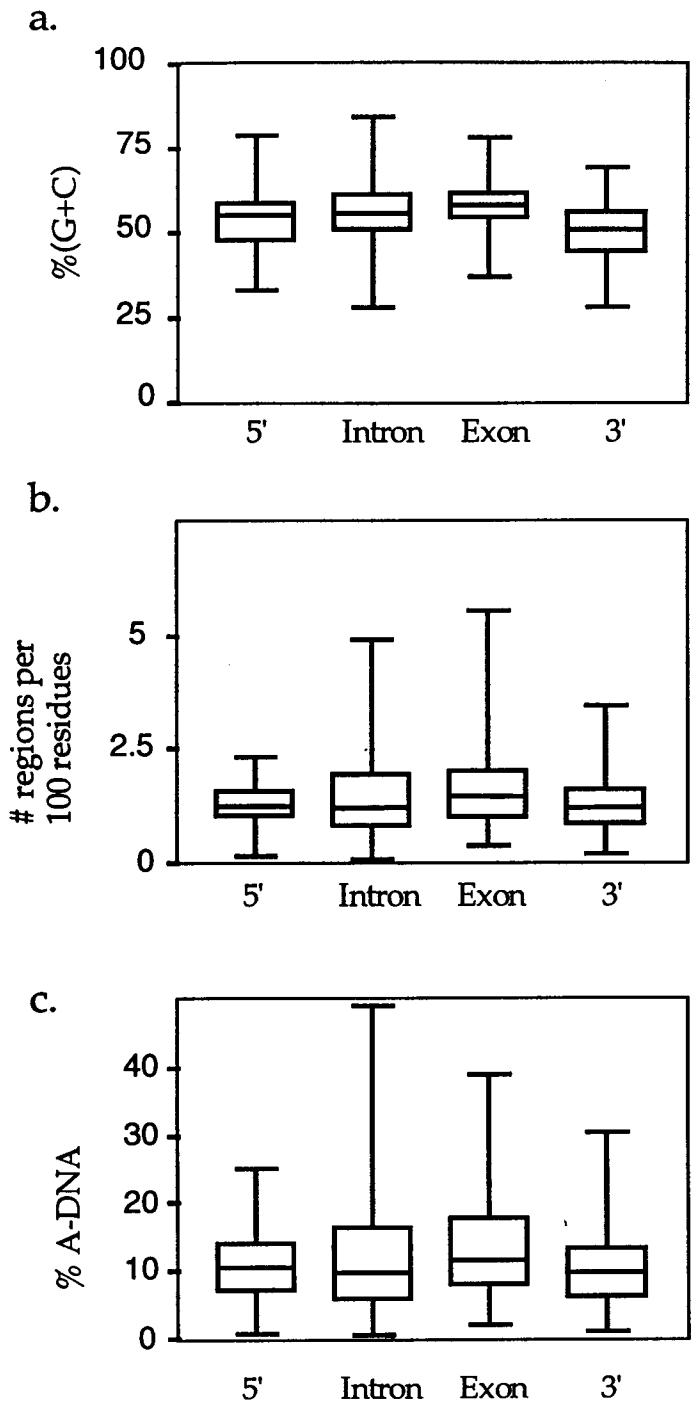

a.
100
75
50
25
0
%(G+C)
5'
Intron
Exon
3'
b.
5
2.5
0
regions per
100 residues
5'
Intron
Exon
3'
c.
40
30
20
10
0
% A-DNA
5'
Intron
Exon
3'

Figure 4.9

Table 4.6

Human genes with significantly more A-DNA than would be expected relative to random DNA

Gene	Gene sections with more A-DNA than random
Brain natriutetic protein	5′, transcribed
Atrial natriuretic factor	5′
c-fos proto-oncogene	5′, transcribed
Urokinase-plasminogen activator gene	5′
Tissue factor gene	5′, 3′
Histone H3 gene	transcribed

5′ and 3′ refer to the 5′ and 3′ untranscribed regions respectively.

content and many of these regions tend to be very long. The promoter region is predicted to contain 13 A-DNA forming regions, 5 of which are more than 10 residues long. The first and second introns contain A-DNA regions that are very long (50 and 28 residues respectively) and exon 4 contains 2 long A-DNA regions (15 and 17 residues). For comparison, a random sequences with comparable (G+C) content has A-DNA regions that have an average length of 9.0 ± 4.6 residues.

4.4.2.3 Analysis of genes from different species

AHUNT was also used to determine if there is a difference in the amount of A-DNA between prokaryotes and eukaryotes. However, it was very important to first determine what amount of A-DNA in a gene sequence is biologically significant. Additionally, since there is a weak correlation between the (G+C) content and A-DNA stability, it is also important to develop a method to compare genes with different (G+C) contents. Therefore, the effect of (G+C) content on the measurement of A- B- and a/b-DNA with AHUNT was determined. 30 random sequences 1500 residues long with specific (G+C) contents were constructed for (G+C) contents from 30% to 75% in 5% increments. A length of 1500 residues was used since that is the shortest length at which the mean amount of the gene predicted to be in A, B, and a/b-DNA regions does not differ significantly from longer sequences. However, sequences at this length have a

significant amount of variance associated with the measurement. Therefore, a length of 1500 provides a conservative description of the effect of (G+C) content on the measurement of the amount of A- B- and a/b-DNA in regions for random DNA.

The relationship between the percent of the sequence that is in A-, B- and a/b-DNA regions with respect to (G+C) content fits a second order polynomial with R^2 >99% (Figure 4.10) for all measurements. It is significant that these are not linear relationships since we have shown that (G+C) content is not a complete model for A-DNA. Additionally, relationships were derived for the mean ± 1 standard deviation for A- B- and a/b-DNA as a function of %(G+C). These separate fits were necessary because the variance associated with the mean is not constant for all (G+C) contents (Figure 4.10). Using these relationships, we may now determine which genes have significantly more or less A- B- or a/b-DNA than comparable random sequences.

Genes from seven species were analyzed for the amount of A-, B- and a/b-DNA in regions. Analysis of 154 human genes shows that these genes are predicted to have more A-DNA and less B-DNA in regions than random DNA of comparable (G+C) content (Figure 4.10). In general the eukaryotic genes tend to have more A-DNA than comparable random sequences and most have less B-DNA than random sequences (Table 4.7).

Figure 4.10

A-DNA, B-DNA and a/b-DNA content of 154 human genes versus the (G+C) content of the gene. The percent of the gene that is predicted to be in A-DNA, B-DNA and a/b-DNA regions was calculated with AHUNT. The dashed line represents the average corresponding % of the random sequence predicted to be in A- B- or a/b-DNA regions. The solid lines represent one standard deviation from this mean as a function of (G+C) content.

Figure 4.11

A-DNA, B-DNA and a/b-DNA content of 104 *E. coli* genes versus the (G+C) content of the gene. The percent of the gene that is predicted to be in A-DNA, B-DNA and a/b-DNA regions was calculated with AHUNT. The dashed line represents the average corresponding % of the random sequence predicted to be in A- B- or a/b-DNA regions. The solid lines represent one standard deviation from this mean as a function of (G+C) content.

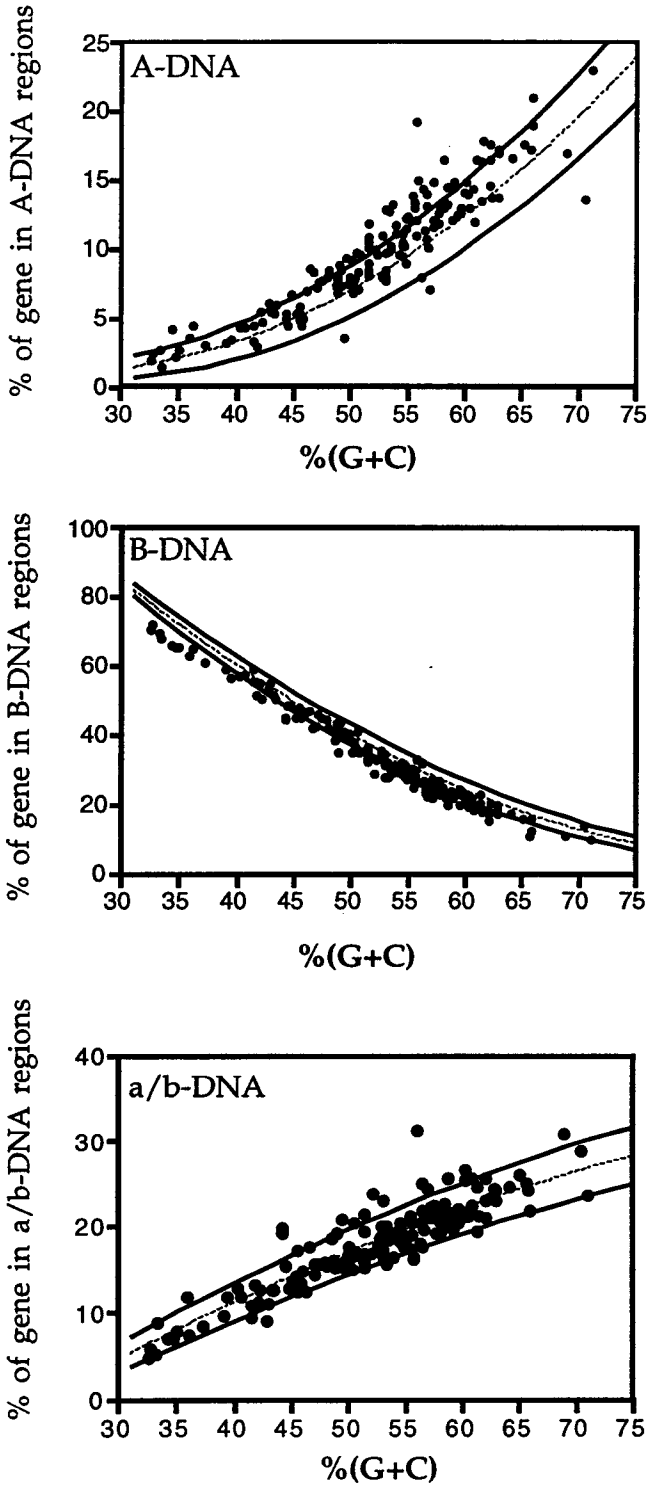

A-DNA
% of gene in A-DNA regions
%(G+C)
B-DNA
% of gene in B-DNA regions
%(G+C)
a/b-DNA
% of gene in a/b-DNA regions
%(G+C)

Figure 4.10

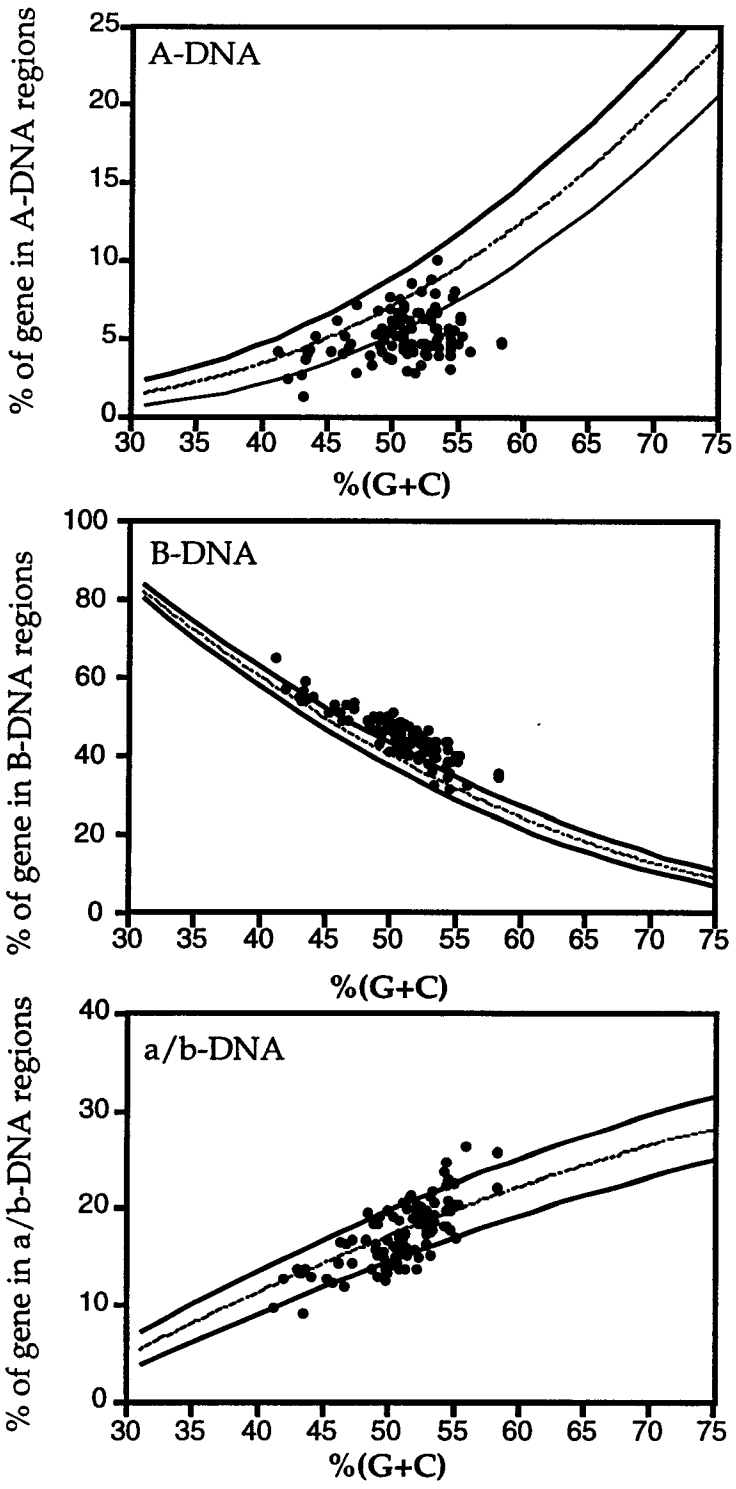

A-DNA
% of gene in A-DNA regions
%(G+C)
B-DNA
% of gene in B-DNA regions
%(G+C)
a/b-DNA
% of gene in a/b-DNA regions
%(G+C)

Figure 4.11

The set of 104 *E. coli* genes shows the most significant differences in A- B- and a/b-DNA propensity (Table 4.7). Specifically, 59% of the *E. coli* genes surveyed have less A-DNA and 75% have more B-DNA (Figure 4.11) than comparable random sequences. No *E. coli* gene analyzed has more A-DNA or less B-DNA than comparable random sequences.

The sets of genes of the two archea bacteria species analyzed are very different from each other and from *E. coli*. *S. sulfolobus*, a thermophilic bacteria that grows under conditions of low pH and has a low (G+C) content (37% for the complete genome), does not in general enhance for or suppress A-DNA or a/b-DNA. However, 52% of the genes analyzed have less B-DNA than random DNA of comparable (G+C) content (Table 4.7). In contrast, *T. aquaticus* (which has a high (G+C) content and grows at high temperatures) genes are enriched for A-DNA and a/b-DNA (55% and 78% of the genes have more A- and a/b-DNA respectively than comparable random sequences) (Table 4.7). 95% of the genes analyzed have significantly less B-DNA than random sequences (Table 4.7). Thus in general, these extremophiles tend to suppress B-DNA.

Finally, genes from *Chlamydia*, a virus that infects eukaryotes, were analyzed for A- B- and a/b-DNA. *Chlamydia sp.* shows trends more similar to the eukaryotes than to any of the other prokaryotes. Specifically, *Chlamydia* genes tend to have more A-DNA and less B-DNA than random sequences of comparable (G+C) content. 38% have more a/b-DNA than random and 13% have less (Table 4.7).

Table 4.7

Summary of the percent of genes with more or less DNA in A-, B- or a/b-DNA regions than random DNA.

Species	# of genes	A^+	B^+	a/b^+	A^-	B^-	a/b^-
Homo sapiens	154	38	0	12	2	52	6
Arabodopsis thaliana	24	42	4	21	4	75	8
Saccharomyces cerevisiae	35	23	3	14	3	49	3
Escherichia coli	104	0	78	13	59	0	16
Sulfolobus solfataricus	32	15	3	15	12	52	12
Thermus aquaticus	22	55	0	78	9	95	0
Chlamydia sp.	24	42	4	38	0	71	13

A^+, B^+ and a/b^+ are the percent of the genes analyzed that have significantly more A-DNA, B-DNA or a/b-DNA than the comparable random sequence.
A^-, B^- and a/b^- are the percent of the genes analyzed that have significantly less A-DNA, B-DNA or a/b-DNA than the comparable random sequence.

In summary, the eukaryotes seem to enhance for strong A-DNA forming sequences in their genomes and suppress strongly B-DNA forming sequences. The prokaryotes as a group do not share any single trend (Table 4.7). Most notably, though, *E. coli* seems to select against A-DNA and favor B-DNA. The thermophilic bacteria suppress strongly B-DNA forming regions, and *Chlamydia* DNA is more like that of eukaryotic DNA by this analysis.

4.5 Discussion

We report here an improved set of A-DNA propensity energies (APEs) that model the sequence dependence of DNA crystal structure conformation (78%) and correlate with the %TFE at the midpoint of the B-to-A transition (86%) for the regions in the 5S gene that have been identified as undergoing the B-to-A transition independently of one another (Becker and Wang, 1989). In an independent test, the APEs correlate with the titration behavior of short oligonucleotides in solution (83%). The APEs predict that most DNA is constiuitively in the B-form since the average value of all the APEs is 0.14 kcal/mol/bp. These APEs have been used in an algorithm, AHUNT, to predict regions in genomic DNA that have a strong propensity to form A- or B-DNA. Analysis of 104 genes from *E. coli* shows that these sequences have less A-DNA and more B-DNA than random sequences of comparable (G+C) content. Eukaryotic and *T.*

aquaticus genes tend to have more A-DNA than random. The archea-bacteria genes analyzed suppress B-DNA and Chlamydia, a bacteria that infects eukaryotes, shows trends similar to those observed for eukaryotic genes. Finally, AHUNT has identified long regions of A-DNA in the human c-fos proto-oncogene, although A-DNA does not in general appear to be localized to any specific part of human genes.

The APEs do not predict the conformation of all the sequences in the crystal structure database, however, this does not invalidate their usefulness or their application to the identification of A-DNA in genomes. Some of the incorrect predictions may relate to the fact that the APEs do not take crystal packing effects which may drive some sequences into the conformation not predicted by the APEs or the ΔSFE_{A-B} into account. It has been suggested that the length of the sequence determines the crystal lattice and that this in turn drives the conformation in the crystal (Timsit and Moras, 1992). While this rule may be valid for sequences that do not have a strong thermodynamic propensity toward A- or B-DNA, it is not generally true. Sequences have been crystallized as A- and B-DNA with lengths of 6, 8, 10 and 12 residues long. In fact, d(CGCTAGCG), which is the only octamer sequence that has been crystallized as a B-DNA, is predicted to have an APE of 0.2 and therefore should be stable as B-DNA. However, the effects of crystal packing on those sequences that do not have a strong propensity toward either A- or B-DNA should not be ignored.

The regions of the 5S gene that AHUNT identifies are not entirely identical to those identified by UV photofootprinting. However, the UV footprinting provided conformational information about only one third of the residues, whereas AHUNT provides nucleotide level resolution. For that reason, region assignments are slightly different.

The B-to-A transition is an energetically attractive mechanism for many types of genetic regulation. DNA binding proteins may prefer sequences that are more structurally malleable. For example, the sequence co-crystallized with the TATA binding protein, d(GGTATACC) has an APE of 0.15 kcal/mol/bp and therefore does not strongly favor either A- or B-DNA. In the crystal structure, the DNA is a distorted A-DNA structure (Guzikevich-Guerstein and Shakked, 1996). Sequences that strongly favor A- or B-DNA may be important as spacer DNA; which is responsible for the proper phasing of important regulatory sites. Finally, A-DNA may be important in phasing bends in DNA.

In order to investigate the biological significance of A-DNA, AHUNT was developed to identify regions within a sequences that have a high propensity to form A-DNA. The requirement that A-DNA regions have <APE> less than -0.2 kcal/mol/bp ensures that the nucleation energy for the B-to-A transition (estimated at 1.2 to 1.5 kcal/mol (Ivanov and Krylov, 1992; Ivanov et al., 1974; Ivanov et al., 1985)) is met. Finally, by identifying regions with a high propensity toward a specific structure rather

than specific residues, AHUNT identifies residues that can cooperatively convert to A-DNA.

For this analysis we have analyzed genes by calculating the percent of the gene sequence that is predicted to be in strong A-DNA or strong B-DNA regions or defined regions that favor neither A- or B-DNA, but meet AHUNT's criteria for a region (a/b-DNA). This latter classification includes sequences that are most likely B-DNA-like under standard conditions; however, these sequences should require less perturbation than the strong B-DNA sequences to convert to A-DNA. Additionally, we have calibrated AHUNT with a large set of random sequences to determine how (G+C) content affects the amount of A- B- or a/b-DNA reported by AHUNT for purely random sequences. This allows us to compare genes with different (G+C) contents in a meaningful and quantitative way.

Analysis of genes with AHUNT has shown that *E. coli* genes tend to have significantly less A-DNA and significantly more B-DNA, than their (G+C) content would predict. It is not entirely surprising that the genes analyzed from the extremely thermophilic bacteria do not show the same trends that the *E. coli* sequences do because the archea bacteria have evolved under very different conditions to fill very specific niches and are not even closely phylogenetically related to *E. coli*.

Eukaryotic genes tend to be enhanced for A-DNA and suppress B-DNA. The A-DNA regions do not map to any particular gene parts. It is likely that A-DNA may map to particular protein binding sites; however,

the sequence databases do not contain this information in such a manner that this may be systematically explored. The A-DNA in eukaryotic genes does not occur disproportionately in any part of the gene, so strong B-DNA regions are not a feature of all protein-coding portions of genes.

Why does *E. coli* have significantly less DNA in A-DNA forming regions and significantly more in B-DNA regions, whereas the other species considered show, in general, a bias toward A-DNA and against B-DNA? Perhaps, these trends reflect the competition between maintaining DNA that is structurally active (due to its polymorphic nature) and the effects of the local DNA environment. That is, certain types of environments are more conducive to allowing a structurally flexible molecule to exist, whereas other environments do not adequately shield the DNA from the forces in the cell that could induce structural changes under the wrong conditions or at the wrong time.

Eukaryotes have a nucleus which protects the DNA from cytoplasmic fluctuations that could induce the B-to-A transition. *E. coli* has a nucleoid structure (Drlica, 1987) which requires a high concentration of cations to maintain the integrity of the nucleoid (Griffith, 1976; Stonington and Pettijohn, 1971), but does not have a membrane. This structure does not protect the DNA from changes in the cytoplasm as well. Finally, *E. coli* DNA is highly negatively supercoiled (negative supercoiling induces the B-to-A transition (Krylov et al., 1990)).

The presence of a nucleus is not the only factor that can protect DNA's structural variability; the proteins that interact with the DNA also can affect the preferred conformations of the DNA. Comparison of the eukaryotic species shows that the higher eukaryotes, which have histones, tend to be more enriched for A-DNA than the *S. cerevisae* genes which does not have histone H1.

The thermophilic bacteria have a very different set of problems in maintaining DNA structural integrity, and therefore it is not surprising that their usage of A- B- and a/b-DNA is very different from that of *E. coli.* The thermophilic bacteria have evolved mechanisms to keep their DNA double stranded, since the temperatures under which they grow demand strand separation. In general the DNA binding proteins found in archea-bacteria tend to bind more like the eukaryotic histones, that is they condense the DNA and remain bound at relatively high ionic strength (Musgrave et al., 1992). *Sulfolobus* has a number of small basic proteins (the Sac and Sso families) that are associated with its DNA (Choli et al., 1988ab) and the NMR structure of the Sac7 protein suggests that this protein interacts with both the major and minor grooves of the DNA (Edmondson et al., 1995).

Thus different species may have adopted different strategies for maintaining structurally active, but structurally regulated DNA. *E. coli* may restrain its DNA structure to be more B-DNA like since the DNA is in an environment such that it is more susceptible to the changes in the

intracellular environment. In contrast, the eukaryotes, which have histones and a true compartmentalized nucleus, may allow more structural freedom to their DNA. The thermophilic eubacteria and archaebacteria have mechanisms to keep their DNA double stranded, and these may also serve to protect the structure of the DNA, and therefore allow their DNA sequences to be more structurally polymorphic.

Chapter 5

5. SUMMARY

The quest in structural biology is to understand the inherent biological function associated with specific sequences of monomeric units. The standard paradigm is that the primary sequence of monomers defines a macromolecule with an explicit three-dimensional structure and a specific biological function. A critical step in understanding a molecule's biological function is to accurately predict its three-dimensional structure. This work addresses this goal for sequences of deoxyribonucleotides, which form DNA, as it explores the question: given a DNA sequence, what structure will it form and is that structure biologically important.

This question is explored at two different levels. For Z-DNA sequences, in which the relationship between the primary sequence and the global structure is well understood, the set of single crystal structures was systematically analyzed to determine the intrinsic atomic-level structure of left-handed DNA, Z-DNA. Rules to explain the sequence-dependence of the right-handed A-DNA structure were not well established; therefore, chapter 4 details the development of a thermodynamic method to predict which sequences have a high propensity to form A-DNA. Chapter 5 explores the biological significance of A-DNA by analyzing genomic DNA sequences with the thermodynamic predictive model.

In addressing these questions, information about DNA structure and stability was obtained from many different structural biology techniques. Each method, however, operates within its own set of limitations and introduces or obscures structural parameters in the macromolecule that is being studied. The challenge in extracting useful information from these techniques is to resolve the actual relevant information. In addressing the goals outlined above, multiple sources of structural information were utilized to develop predictive, sequence-dependent models for DNA structure and function.

5.1 The systematic evaluation of Z-DNA structure

Z-DNA has been analyzed by a variety of methods in order to understand the sequence-dependence of its structure and stability. X-ray crystallography is the most frequently used technique to ascertain the high resolution three-dimensional structure of this conformation. The stability of specific sequences as Z-DNA is generally measured by supercoiling induction. Chapter 2 is a systematic analysis of the Z-DNA crystal structure database. The chapter details the use information from structure and stability measurements to explore how sequence, substituent modifications and the DNA's environment define the structure and stability of Z-DNA.

The computer program, NASTE, was developed specifically to calculate structural parameters for all the Z-DNA crystal structures in the

Nucleic Acid Database (Berman et al., 1992) in a standard frame of reference. Other structure analysis programs were not applicable to the analysis of Z-DNA crystal structures because they are not generalized enough to properly analyze a left handed helix. The parameters calculated with NASTE were used to systematically explore how Z-DNA structure and stability are related and how specific moieties on the bases, crystal packing effects and cationic environment account for differences in the three-dimensional structures. The specific aim was to discover the true conformation of Z-DNA (independent of artifacts of the crystal environment) and how different factors contribute to its stability. 56 different Z-DNA crystal structures were compared and contrasted. These structures differed from each other in factors such as space group (which defines the crystal lattice), length, sequence, modifications to bases and the cationic environment. The analysis of the Z-DNA crystal structure database was done by comparing crystal structures to their appropriate reference crystal structures. In this way, the effect of specific variables on the structure could be extracted from the other variables that might be present in a particular structure. For example, the structural parameters of the sequence d(m^5CGTAm^5CG) were compared to those of d(m^5CGm^5CGm^5CG) in order to understand the effect of a d(TpA) step on Z-DNA structure and stability, independent of the effects of cytosine methylation.

Several rules were established. First, the dinucleotide is the repeating unit (and not the mononucleotide as in A- and B-DNA) and this is the basis for the zigzag structure that defines Z-DNA. Second, the bases in the dinucleotide unit are always in the *anti-p-syn* conformation, even in the sequences where the sequence does not follow the alternating pyrimidine purine motif that generally defines Z-DNA stability. Third, the pattern of alternating sugar puckers (C3′-*endo* for nucleotides in *syn* and C2′-*endo* for nucleotides in *anti*) is generally conserved; however, it may be influenced by crystal packing effects at the ends. Finally, the *anti*-p-*syn* step is always severely underwound relative to the *syn*-p-*anti* step. While the twist angles of these two steps vary slightly from structure to structure, the net twist of the dinucleotide step is consistently very close to -60°.

Analysis of sequences of various lengths showed that hexamers are a good model for Z-DNA since they contain the features of structure common with all the other lengths. However, the crystal lattice environment does perturb the hexamers in a minor way. One would expect that d(CGCGCG) would have the same twist at each d(CpG) step; however, a position dependence is observed. This is attributed to the lattice in which the hexamers crystallize, $P2_12_12_1$. In this crystal lattice, a two-fold screw axis relates the motif to the adjacent duplexes. As a result, the duplexes are stacked end on end in the z direction, but are stacked with a dinucleotide displacement in the x and y directions. Thus the first dinucleotide step is in

a slightly different lattice environment than the other steps and a slight decrease in the twist is observed.

Cytosine methylation is by far the most understood substituent and is present in many of the crystallized Z-DNA sequences. Methylation has been thoroughly studied by both crystallography and circular dichroism spectroscopy (Behe and Felsenfeld, 1981) which have established that it stabilizes the Z-DNA conformation. Analysis of structures with methylated cytosines has revealed the structural basis for the increase in relative stability as Z-DNA of sequences with methylated cytosines and the importance of hydration. The d(m^5CpG) step is overwound in order to accommodate the methyl group and the d(Gpm^5C) step is underwound to maintain the standard dinucleotide conformation. Sequences that were brominated at the C5 of cytosine were compared to the methylated sequences. The structural parameters calculated with NASTE revealed that bromination is in many ways analogous to methylation of cytosine. Interestingly, this comparison also showed that brominated sequences do not exhibit the same position dependence of the twist that is seen in the other hexamers suggesting that the fully brominated structure is less affected by crystal lattice effects.

This systematic analysis of out-of-alternation Z-DNA structures also revealed a significant structural consequence associated with the incorporation of an out of alternation base step. The analysis of the crystal structures of d(m^5CG<u>G</u>Gm^5CG)-d(m^5CGC<u>C</u>m^5CG) and d(m^5CG<u>G</u>Gm^5CG)-

d(m^5CGC<u>C</u>m^5CG), in which the underlined cytosine is in *syn*, suggested that the high buckle seen at this cytosine in *syn*, was a steric consequence of its ribose interacting with the methyl of the flanking 3′ 5-methyl-cytosine. However, analysis of d(m^5CG<u>G</u>Gm^5CG)-d(m^5CGC<u>m^5C</u>CG) in which the flanking 3′ cytosine is not methylated revealed that the high buckle is characteristic of all cytosines in *syn* in Z-DNA.

The effect of the cationic environment on the structure of Z-DNA was also evaluated using NASTE. Structures with magnesium present were essentially identical even in the presence of spermine or spermidine. However, in the crystal structure in which there was no magnesium and spermine was the only cation, the orientation of the DNA in the crystal lattice and the DNA structure was changed. Since there are no other structures in the database that are oriented in the lattice in the same way, it is not possible to determine which structural perturbations in the DNA are due to its position in the lattice and which are due to the absence of magnesium. Analysis of other sequences crystallized in the presence of only spermine as the counterion may be able to resolve this.

In summary, chapter 2 establishes a framework for evaluating multiple structures of the same global conformation. By comparing structures to the appropriate reference structures and correlating structural information with stability measurements, it is possible to identify the generalities of the Z-DNA conformation and to understand how different

variables introduced during crystallization affect the conformation. This type of analysis was possible for the set of Z-DNA crystal structures because the database is dominated by sequences that represent systematic changes from the original d(CGCGCG) sequence. When there are enough related crystal structures for the other conformations of DNA, a similar analysis will be possible for those conformations and a more detailed analysis of DNA structure and crystal packing effects will be possible.

5.2 Development of a general predictive method for A-DNA sequence-dependent stability

The structure and stability rules are not as well defined for A-DNA as they are for Z-DNA. Chapters 3 and 4 explore the development, testing and application of a model of A-DNA sequence dependent stability. This work started with several observations. First, DNA sequences crystallize as either A- or B-DNA, and do so independent of the crystallization conditions. Second, (G+C) content is not a sufficient predictor of a sequence's propensity to form A-DNA, and third, hydrophobicity plays a role in A-DNA stability (Franklin and Gosling, 1953a,b).

Hydrophobicity is important for stabilizing the A-DNA conformation; therefore, we hypothesized that a thermodynamic model based on hydration could explain the sequence dependence of A- versus B-DNA stability. This method was successful in modeling the difference in Z- versus B-DNA stability (Kagawa et al, 1989; Ho et al., 1991). The

thermodynamic method involved calculating the solvent accessible surface of the DNA and determining the free energy of hydration associated with the exposed surface. Hydrophobic groups make the solvent free energy less favorable, whereas hydrophilic groups are stabilizing. Chapter 3 describes how differences in solvent free energy for sequences modeled as A- versus B-DNA (ΔSFE_{A-B}) were calculated in order to determine the difference in stability of a sequence in the A-DNA conformation versus the B-DNA conformation. Ideal models of A- and B-DNA (developed from the fiber diffraction studies) were used to model A- and B-DNA because this eliminated effects due to crystal packing from the models. The calculation of ΔSFE_{A-B} for the sequences in the crystal structure database showed that sequences that crystallized as A-DNA have significantly more negative ΔSFE_{A-B}s than sequences which crystallized as B-DNA. Therefore ΔSFE_{A-B} is a predictive parameter for A-DNA stability in the crystal.

Calculation of ΔSFE_{A-B} in this manner is not reasonably directly applicable to large numbers of sequences of DNA or to the very long sequences that compose genes and genomes. Therefore, the information from the ΔSFE_{A-B}s was used to develop a general method to predict A-DNA stability. The ΔSFE_{A-B}s were deconvoluted into a set of A-DNA propensity energies (APEs) which reflect a base pair's relative stability as A- versus B-DNA in the context of its nearest 5′ and 3′ neighbors. This triplet code predicted the conformation of 90% of the sequences in the crystal structure database at the time of the derivation.

The crystal structure database does not contain all the possible triplets, and the goal was to develop a set of energies that were predictive not only for DNA in crystals, but also for DNA sequences in genomes. Results from a UV photofootprinting study of the 5S ribosomal RNA gene from *Xenopus laevius* (Becker and Wang, 1989) provided correlated sequence and structure information about the B-to-A transition. In the UV photofootprinting analysis of DNA, photoproduct formation was measured for one third of the nucleotides in the gene (Becker and Wang, 1989). It is believed that the A-DNA conformation protects DNA from UV-induced photoproduct formation and therefore residues that are in the A-conformation are less reactive than residues in the B-DNA conformation (Becker and Wang, 1989). In Becker and Wang's experiment, the DNA was titrated with trifluoroethanol (TFE) to induce A-DNA, and the photoproduct formation was measured as a function of TFE concentration. This experiment, then, provided a quantitative measure of the sequence specificity of the B-to-A transition. The conclusion was that there are 11 regions of the 5S rRNA gene that undergo the B-to-A transition independently of one another.

If the $\Delta SFE_{A\text{-}B}$s and the UV photofootprinting experiment were both correlated with A-DNA stability, then a single model could represent both types of data. A genetic algorithm was developed to incorporate data from both measurements into the derivation of the APEs. A genetic algorithm is a method of optimization which harnesses the principles of evolution to rapidly and efficiently explore solution space. Genetic algorithms have

been applied to energy minimization problems (i.e. problems that are normally tackled with molecular dynamics or simulated annealing algorithms); however, in these algorithms the function that is minimized is essentially that calculated by traditional refinement protocols (LeGrand and Merz, 1994). The challenge in applying the genetic algorithm to this problem was understanding how the two different measurements related to one another and how to treat them appropriately in defining the APEs. The algorithm minimized the deviation of the model's parameters (the APEs) from the ΔSFE_{A-B} and UV footprinting measurements. It also minimized the difference in deviations between the two data sets in order to prevent the algorithm from converging on a solution that satisfied one data set at the expense of the other. Thus, the solution obtained does not represent the crystal structure data set or the UV photofootprinting experiment maximally, but the ultimate set of APEs reflects a compromise between both data sets. The complete set of APEs predicts the conformation of 78% of sequences in the crystal structure database and correlates (R^2 = 86%) with the % TFE at the midpoint of the B-to-A transition for the 11 regions of the 5S ribosomal RNA gene that were identified as undergoing the B-to-A transition independently of one another.

5.3 Testing the A-DNA propensity energies

As a test of the APEs that was independent of the systems used to derive them, the B-to-A transitions of 12 dodecameric sequences were monitored with circular dichroism spectroscopy (CD). The thermodynamic model predicts that the midpoint of the B-to-A transition for these sequences should be correlated with their APEs. In this experiment, the sequences were titrated with TFE and the CD spectrum was recorded at various TFE concentrations. The results of these experiments are correlated with the APEs, however, not in the manner anticipated. Sequences with positive APEs show one of two behaviors. Two sequences show spectra that are not consistent with a complete B-to-A transition over the titration and $d(GC)_6$ actually converts to Z-DNA in the presence of high TFE concentrations. This is not surprising since this sequence strongly favors Z-DNA. Sequences with negative APEs also show one of two behaviors. Sequences with at least three cytosines or guanines in a row in positions 1-3 have CD spectra consistent with that of A-DNA throughout the titration and other sequences with negative APEs require less than 83% TFE to titrate to A-DNA.

Our interpretation that the conformation of the sequence d(CCCCCGCGGGGG) is predominantly A-form in aqueous solution has been challenged (Vorlickova et al., 1996). Vorlickova, et. al. repeated the experiment using ethanol to induce the B-to-A transition in

d(CCCCCGCGGGGG) and showed that the change in the CD spectrum was not the same as that seen with TFE. There was no increase in intensity at 270 nm, but there was a shift of the peak similar to that seen in the presence of very high TFE in some of the dodecamers examined in this work. Additionally, when a titration was performed with TFE, Vorlickova, et. al. interpreted the increase in ellipticity at 269 nm as indicative of a reversible transition with fast kinetics with a high energy barrier between the two conformations (Vorlickova et al., 1996). The titrations with TFE and ethanol (Vorlickova et al., 1996) do not establish that the conformation of d(CCCCCGCGGGGG) in aqueous solution is not A-DNA, though. While these experiments have established that the spectra obtained with ethanol versus TFE are different, neither Vorlickova, et. al's titrations or the titrations presented in these chapters establish the conformation of d(CCCCCGCGGGGG) in aqueous solution. In short, CD does not reveal if the transition observed with TFE is an X-to-A or an A-to-X transition (where X is a non-A-DNA conformation).

Unfortunately, the CD experiments in these chapters did not provide transition midpoints that could be correlated with the APEs. There are several reasons why this did not occur. First, CD spectra are very sequence specific. That is the spectrum of one sequence is not identical to that of another sequence in the same conformation. Second, the CD signal observed is the average spectrum of all the molecules in all the conformations present in the solution. Since the spectra for the endpoints of the

titration (that is 100% A-DNA and 100% B-DNA) or the ratio of the different conformations in any given solution are not explicitly known, it is not straightforward to decompose any spectrum into its parent spectra. Additionally, since the spectrum is sequence dependent, it is not possible to use spectra from other sequences (in which one is confident of the conformation) to do this. These kinds of comparisons and a paradigm that DNA in aqueous solution is entirely B-DNA, while DNA in high concentrations of TFE or ethanol is predominantly A-DNA may have resulted in the misinterpretation of many spectra and titrations. Until there is a way to unambiguously determine the ratios of conformations in a solution, the results from TFE titrations such as those described in chapters 3 and 4 are strictly qualitative.

How would one determine the relative ratios of A and B-DNA at the endpoints of a titration or the spectra of pure A- or pure B-DNA for a given sequence and establish the aqueous conformation of d(CCCCCGCGGGGG)? NMR has the potential to provide information about the relative ratios of A- and B-DNA in solution and establish an equilibrium constant for the B-to-A transition for sequences of DNA. A-DNA should have a very different set of NOE signals from the base protons to the sugar protons than B-DNA (van de Ven and Hilbers, 1988). The limitation with this method applied to A- and B-DNA involves the rate of interconversion between A- and B-DNA. That is if the molecules are converting between the two conformations on the NMR timescale (50 to 500 milliseconds mixing time) then the

resulting spectra will be an average spectra and resolution will be lost. Additionally, for this type of analysis, it is important to determine the structure and ratios of all the structures in the solution. This is not necessarily straightforward (van de Ven and Hilbers, 1988). Finally, an NMR approach would only be feasible for low alcohol or aqueous conditions since at the concentrations of DNA required for NMR, the DNA is not soluble in high concentrations of alcohol.

A final difference between the CD study reported here and other CD studies of DNA is that most other CD studies were done on long polymers of DNA. The sequences used here were only 12 residues long, and fraying at the ends of these short sequences would have an influence on the spectra. In summary, these chapters suggest that CD studies of DNA should be interpreted with caution and in conjunction with evidence from other physical techniques.

5.4 The application of the A-DNA propensity energies to identifying A-DNA in genomes

Typically, the problem addressed by structural biology is one of defining a structure for a particular function. Although a role for A-DNA has not been identified, A-DNA is a very attractive candidate for some of the functions which are known for DNA. For example, the small energetic barrier between A- and B-DNA (1.2 to 1.5 kcals) (Ivanov and Krylov, 1992; Ivanov et al., 1974; Ivanov et al., 1985), suggests that A-DNA could be

important as a conformational switch. The B-to-A transition has been shown to be very sensitive to environment, and therefore it is easy to imagine a scenario in which changes in the environment could induce a change in the DNA structure which in turn could alter the binding affinity of a regulatory protein. However, the traditional physical methods used to identify A-DNA are not applicable to searching for this conformation in genomic DNA. In addition, it is difficult to probe for the A-DNA conformation *in vivo*. Therefore, a computation strategy based on the derived thermodynamic parameters was used to search for a biological activity for A-DNA. This allowed us to use the vast amount of sequence information available to try to correlate sequences with a high propensity to form A-DNA with known activities of DNA.

The program, AHUNT, is the first computational attempt to identify A-DNA in long sequences of DNA. Rather than identifying specific residues with a high propensity to form A-DNA, AHUNT identifies regions of a sequence that collectively have a high propensity for A-DNA formation. To do this, AHUNT uses the APEs, which are a thermodynamic measurement of a nucleotide's propensity to form A-DNA, to calculate a residue's propensity to adopt the A-DNA conformation in the context of its nearest three 5′ and 3′ residues. Regions of genes that have similar propensities are grouped and assigned a conformation according to the average APE of the region. Regions have several characteristics. They are at least 5 residues long and at least 80% of the residues have an average APE that

satisfies a minimum energetic requirement. AHUNT identifies 3 distinct classes of regions. Strong A-DNA-forming regions have an average APE less than -0.2 kcal/mol/bp. Strong B-DNA-forming regions have an average APE greater than 0.2 kcal/mole/bp. Regions assigned as a/b-DNA do not strongly favor either A- or B-DNA and are characterized by an average APE intermediate between -0.2 kcal/mol/bp and 0.2 kcal/mol/bp. These latter regions are most likely constitutively B-DNA, but should require less perturbation to undergo the B-to-A transition than regions assigned as B-DNA.

Several questions may be addressed with AHUNT in attempting to correlate the A-DNA structure with biological functions:

1. Do putative A-DNA forming regions correlate with known biological activities of DNA?
2. Are putative A-DNA forming regions more prevalent in specific parts of the gene?
3. Do different species (or kingdoms) have different amounts of A-DNA?

To address these questions, a method of comparing and contrasting genes had to be developed. Since genes are non-random sequences of DNA, differences between the predicted amount of A-DNA in genes and comparable random sequences are significant. (G+C) content was used as the standard to which all sequences were compared. (G+C) content was chosen because it is a characteristic of the gene that is well correlated with

the predicted A-DNA content of the gene. Significantly, this relationship is not linear, but at least second order. A linear relationship would suggest that (G+C) content is a sufficient model to account for A-DNA formation. This same analysis was performed for B- and a/b-DNA regions and similar second order correlations were observed relative to the (G+C) content. It is thus straightforward to identify sequences that deviate from random DNA with the same (G+C) content with respect to the amount of A-, B- or a/b-DNA in regions. This establishes criteria for identifying significant amounts of a specific conformation in a gene.

Analysis of the genes from several species showed that *E. coli* has significantly less A-DNA and more B-DNA than comparable random sequences. Genes analyzed from other species have more A-DNA and less B-DNA than random sequences of comparable (G+C) content. These include eukaryotic genes from humans, *Arabadopsis thaliana* and *Saccharomyces cerevisiae* and genes from the prokaryotes *Sulfolobus solfataricus, Thermus aquaticus* and *Chlamydia*. The analysis of human genes with AHUNT did not identify any specific parts of these genes that contain significantly more or less A-DNA than other parts; however, some human genes do contain some very long A-DNA regions.

In light of this analysis, one may ask whether A-DNA is a biologically significant conformation. If A-DNA is indeed biologically significant, we would expect that there would be either significant differences or significant similarities in its predicted occurrence between different species or king-

doms or in different parts of the gene. Significant differences would suggest a species-specific or kingdom-specific role. Common differences from random DNA or localization to specific parts of the gene could suggest a general role for A-DNA biologically.

At this point, there are several conclusions that may be drawn. First, AHUNT in its current form may not be a good tool for identifying A-DNA in genes. In this analysis it was assumed that the nucleation energy is the same for all sequences, however, this may not be true. This would affect the cutoffs used for A-, B- and a/b-DNA. Additionally, comparison of genes to random DNA with the same (G+C) content may not accurately emphasize A-DNA in genes since the random DNA has a uniform (G+C) content whereas the gene does not necessarily have a uniform (G+C) content. However, AHUNT does predict regions that correlate very well with the observations of the photofootprinting assay, so these factors may be a source of error, but most likely are not significant.

Second, the correlations that were attempted with AHUNT were limited due to the information available in the sequence files. The sequence databases do not yet contain consistent information about the identification of specific gene parts. Information about protein binding sites, spacer sequences which are responsible for phasing regulatory elements, and sequences that are responsible for bending the DNA is not regularly present in the databases. For complete genomes, the information is even less complete. While the genomes are analyzed for putative protein

binding sites, the sequence files do not contain complete mapped information about the genome that would make it usable in an analysis of this type. Fortunately, a current effort in genomics is to make the sequence databases more relational and multidimensional. This should make analyses like AHUNT more informative.

This work, however, does provide a testable hypothesis. This hypothesis, developed in chapter 4, suggests that the DNA intracellular environment affects the conformational freedom of the DNA. The intracellular environment can potentially effect changes in DNA structure through perturbations in water activity, salt or polyamine concentrations or the expression of special DNA binding proteins. If A-DNA, or any other specific structure of a sequence of DNA, is critical biologically, then the cell must ensure that the conditions under which the DNA can adopt that specific structure are well regulated. This may be done in one of two ways. First, the DNA structure may be constrained by physical structures in the cell. These include the nucleus, the nuclear matrix, histones and other DNA binding proteins. All of these factors are found associated with eukaryotic DNA. The other way to control the structure of the DNA is through the sequence. Specifically, the genome may be composed of sequences that are less likely to be perturbed by changes in the intracellular environment (i.e. sequences that are predicted to be strong A-DNA or strong B-DNA formers).

The prediction of this model is that species that have external mechanisms to constrain their DNA in a certain conformation should have more conformational freedom and hence more structurally dynamic sequences than species that use the sequence to maintain a specific structure. This was observed for the genes of the species analyzed with AHUNT. Species that have a nucleus had more A-DNA and less B-DNA than comparable random sequences. The trend was not as significant for yeast, which lacks histone H1, and therefore has a genome that is less constrained by physical structures in the cell. In contrast, in *E. coli*, the DNA is in a less protective nucleoid structure and is therefore more subject to potential B-to-A transition inducing factors. Thus, in *E. coli* there may be a mechanism to actively suppress A-DNA forming sequences in order to more tightly regulate its genome's ability to respond to environmental changes. The results from the analysis of *E. coli* genes with AHUNT reflect this: these genes had less A-DNA than comparable random sequences. The thermophilic bacteria have evolved a number of mechanisms to keep their DNA double-stranded, and these proteins and polyamines could serve the same role as the histones in the eukaryotes. Analysis of genes from thermophilic bacteria with AHUNT shows that these genes tend to have more A-DNA than comparable random sequences.

A prediction of this model is that species which do not have a nucleus to protect the DNA would have less A-DNA than species with a nucleus if it is necessary for the organism to regulate its DNA structure.

Investigation of other bacteria should show a suppression of A-DNA as observed for *E. coli* genes. Likewise, consideration of other species with different DNA environments should be informative.

While a specific biological function for A-DNA has not been identified here, this work does emphasize the importance of considering DNA and genomes as dynamic systems rather than as static B-DNA structures. This work has explored the inherent dynamic nature of DNA sequences from two aspects. First, it has identified the generalities of the Z-DNA conformation and the effect of sequence and environmental factors on the structure and stability of Z-DNA. Second, it has identified the thermodynamic requirements of the sequence-specific B-to-A transition and used these rules to explore the structure of DNA in genes. These studies are an understanding of the relationship between DNA sequence and structure and should prove valuable in understanding genome structure and function in the post-genome sequencing era.

BIBLIOGRAPHY

Alden, C., and Kim, S. 1979. Solvent-accessible surfaces of nucleic acids. *J. Mol. Biol.* **132**: 411-434.

Anfinsen, C. 1973. Principles that govern the folding of protein chains. *Science* **181**: 223-230.

Arnott, S., Chandrasekaran, R., Birdsall, D. L., Leslie, A. G. W., and Ratcliff, R. L. 1980. Left handed DNA helices. *Nature (London).* **283**: 743-745.

Arnott, S., Chandrasekaran, R., Hukins, D. W. L., Smith, P. J. C., and Watts, L. 1974. Structural details of a double-helix observed for DNA's containing alternating purine-pyrimidine sequences. *J. Mol. Biol.* **88**: 523-533.

Arnott, S., Chandrasekaran, R., and Selsing, E. 1975a. The variety of polynucleotide helicies. In *Structure and Conformation of Nucleic Acids and Protein-Nucleic Acid Interactions,* M. Sundaralingam and S. Rao, eds. (Baltimore: University Park Press), pp. 577-596.

Arnott, S., and Hukins, D. 1972. Optimised parameters for A-DNA and B-DNA. *Biochem. Biophys. Res. Comm.* **47**: 1504-1509.

Arnott, S., Smith, P., and Chandrasekaran, D. 1975b. Atomic coordinates and molecular conformations for DNA-DNA, RNA-RNA and DNA-RNA helices. In *Handbook of Biochemistry and Molecular Biology,* G. Fasman, ed. (Cleveland: CRC), pp. 411-422.

Ban, C., Ramakrishnan, B., and Sundaralingham, M. 1996. Crystal structure of the self-complementary 5'-purine start decamer d(GCGCGCGCGC) in the Z-DNA conformation - part I. *Biophysical J.* **71**: 1215-1221.

Bancroft, D., Williams, L. D., Rich, A., and Egli, M. 1994. The low-temperature crystal structure of the pure-spermine form of Z-DNA reveals binding of a spermine molecule in the minor groove. *Biochemistry* **33**: 1073-1086.

Barber, A. M., Zhurkin, V., and Adhya, S. 1993. CRP-binding sites: evidence for two structural classes with 6-bp and 8-bp spacers. *Gene* **130**: 1-8.

Basham, B., Schroth, G. P., and Ho, P. S. 1995. An A-DNA triplet code: thermodynamic rules for predicting A- and B-DNA. *Proc. Natl. Acad. Sci., USA* **92**: 6464-6468.

Becker, M. M., and Wang, Z. 1989. B-A transitions within a 5S ribosomal RNA gene are highly sequence specific. *J. Biol. Chem.* **264**: 4163-4167.

Behe, M., and Felsenfeld, G. 1981. Effects of methylation on a synthetic polynucleotide: the B-Z transition in poly(dGm^{5}dC)•poly(dG-m^{5}dC). *Proc. Natl. Acad. Sci., USA* **78**: 1619-1623.

Benedetti, G. And Morosetti, S. 1995. A genetic algorithm to search for optimal and suboptimal RNA secondary structures. *Biophys. Chem.* **55**: 253-259.

Berman, H. M. 1994. Hydration of DNA: take 2. *Curr. Opin. Struct. Biol.* **4**: 345-350.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.-H., Srinivasan, A. R., and Schneider, B. 1992. The Nucleic Acid Database: a comprehensive relational database of three-dimensional structures of nucleic acids. *Biophysical J.* **63**: 751-759.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodger, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**: 535-542.

Bohm, G. 1996. New approaches in molecular structure prediction. *Biophys. Chem.* **59**: 1-32.

Bononi, J. 1995. Effect of hemi-methylated CG dinucleotide on Z-DNA stability; crystallographic and solution studies. (MS thesis, Oregon State University).

Bowie, J. U. and Eisenberg, D. 1994. An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. *Proc. Natl. Acad. Sci, USA.* **91**: 4436-4440.

Brahms, J., and Mommaetrs, W. 1964. A study of conformation of nucleic acids in solution by means of circular dichroism. *J. Mol. Biol.* **10**: 73-88.

Brennan, R. G., Westhof, E., and Sundaralingam, M. 1986. Structure of a Z-DNA with two different backbone chain conformations. Stabilization of the decadeoxynucleotide d(CGTACGTACG) by $(Co(NH_3)_6)^{3+}$. *J Biomol. Struct. Dynam.* **3**: 649-665.

Brown, T., Kneale, G., Hunter, W. N., and Kennard, O. 1986. Structural characterization of the bromouracil-guanine base pair mismatch in a Z-DNA fragment. *Nucleic Acids Res.* **14**: 1801-1809.

Cartwright, H. M. 1995. The genetic algorithm in science. *Pest. Sci.* **45**: 171-178.

Cervi, A. R., Guy, A., Leonard, G. A., Téoule, and Hunter, W. N. 1993. The crystal structure of N^4-methylcytosine•guanosine base-pairs in the synthetic hexanucleotide d(CGCGm^4CG). *Nucleic Acids Res.* **21**: 5623-5629.

Chevrier, B., Dock, A. C., Hartmann, B., Leng, M., Moras, D., Thuong, M. T., and Westhof, E. 1986. Solvation of the left-handed hexamer d(^{5}BrC-G-^{5}BrC-G-^{5}BrC-G) in crystals grown at two temperatures. *J. Mol. Biol.* **188**: 707-719.

Choli, T., Henning, P., Wittmann-Liebold, B., and Reinhardt, R. 1988a. Isolation, characterization and microsequence analysis of a small basic methylated DNA-binding protein from the archaebacterium *Sulfolobus solfataricus*. *Biochem Biophys. Acta* **950**: 193-203.

Choli, T., Wittmann-Liebold, B., and Reinhardt, R. 1988b. Microsequence analysis of DNA-binding proteins 7a, 7b and 7e from the archeabacterium *Sulfolobus aciocaldarius*. *J. Biol. Chem.* **263**: 7087-7093.

Coley, D. A. 1996. Genetic Algorithms. *Contemp. Phys.* **37**: 145-154.

Coll, M., Fita, I., Lloveras, J., Subirana, J. A., Bardella, F., Huynh-Dinh, T., and Igolen, J. 1988. Structure of d(CACGTG), Z-DNA hexamer containing AT base pairs. *Nucleic Acids Res.* **16**: 8695-8705.

Coll, M., Saal, D., Frederick, C. A., Aymami, J., Rich, A., and Wang, A. H. J. 1989. Effects of 5-fluorouracil/guanine wobble base pairs in Z-DNA. Molecular and crystal structure of d(CGCGFG). *Nucleic Acids Res.* **17**: 911-923.

Coll, M., Wang, A. H. J., van der Marel, G. A., van Boom, J. H., and Rich, A. 1986. Crystal structure of a Z-DNA fragment containing thymine/2-aminoadenine base pairs. *J. Biomol. Struct. Dyn.* **4**: 157-172.

Connolly, M. L. 1983. Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221**: 709-713.

Crawford, J. L., Kolpak, F. J., Wang, A. H. J., Quigley, G. J., van Boom, J. H., van der Marel, G. A., and Rich, A. 1980. The tetramer d(CpGpCpG) crystallizes as a left-handed double helix. *Proc. Natl. Acad. Sci., USA* **77**: 4016-4020.

Cruse, W., Salisbury, S., Brown, T., Cosstick, R., Eckstein, F., and Kennard, O. 1986. Chiral phoshorothioate analogues of B-DNA. *J. Mol. Biol.* **192**: 891-905.

Davies, D. B. 1978. *Progress in NMR Spectroscopy.* Volume 12. (Pergamonn, Oxford).

Dickerson, R. 1990. *Structures and Methods: Vol. 3, DNA and RNA,* R. Sarma and M. Sarma, eds. (Albany, NY: Adenine Press).

Dickerson, R., Goodsell, D., and Neidle, S. 1994. ...the tyranny of the lattice... *Proc. Natl. Acad. Sci., USA* **91**: 3579-3583.

Dickerson, R. E. 1992. DNA structure from A to Z. *Methods Enzymol.* **211**: 67-111.

Diekmann, S. 1989. Definitions and nomenclature of nucleic acid structure parameters. *EMBO J.* **8**: 1-4.

Doi, M., Inoue, M., Tomoo, K., Ishida, T., Ueda, Y., Akagi, M., and Urata, H. 1993. Structural characteristics of enantiomorphic DNA: crystal analysis of racemates of the d(CGCGCG) duplex. *J. Am. Chem. Soc.* **115**: 10432-10433.

Drew, H. R., and Dickerson , R. E. 1981a. Conformation and dynamics in a Z-DNA tetramer. *J. Mol. Biol.* **152**: 723-736.

Drew, H. R., and Dickerson, R. E. 1981b. Structure of a B-DNA dodecamer: III geometry of hydration. *J. Mol. Biol.* **151**: 535-556.

Drew, H. R., Takano, T., Tanaka, S., Itakura, K., and Dickerson, R. E. 1980. High-salt d(CpGpCpG), a left-handed Z-DNA double helix. *Nature (London)* **286**: 755-756.

Drlica, K. 1987. The Nucleoid. In *Escherichia coli and Salmonella typhimurium - Cellular and Molecular Biology*, F. C. Neidhardt, ed. (Washington DC: Association of Microbiology), pp. 91-103.

Edmondson, S. P., Qui, L., and Shriver, J. W. 1995. Solution structure of the DNA-binding protein Sac7d from the hyperthermophile *Sulfolobus acidocaldarius*. *Biochemistry* **3**: 13289-13304.

Egli, M., and Gessner, R. V. 1995. Stereoelectronic effects of deoxyribose O4' on DNA conformation. *Proc. Natl. Acad. Sci. USA* **92**: 180-184.

Egli, M., Williams, L. D., Gao, Q., and Rich, A. 1991. Structure of pure-spermine form of Z-DNA (magnesium free) at 1 Å Resolution. *Biochemistry* **30**: 11388-11402.

Eichman, B. F., Basham, B., Schroth, G. P., and Ho, P. S. 199X. The Z-DNA structure of the out-of-alternation d(GpC) dinucleotide.

Ellison, M. J., Feigon, J., Kelleher, R. J., III, Wang, A. H. J., Habener, J. F., and Rich, A. 1986. An assessment of the Z-DNA forming potential of alternating dA-dT stretches in supercoiled plasmids. *Biochemistry* **25**: 3648-3655.

Ellison, M. J., Kelleher, R. J., III, Wang, A. H. J., Habener, J. F., and Rich, A. 1985. Sequence-dependent energetics of the B-Z transition in supercoiled DNA containing nonalternating purine-pyrimidine sequences. *Proc. Natl. Acad. Sci., USA* **82**: 8320-8325.

Feuerstein, B. G., Williams, L. D., Basu, H. S., and Marton, L. J. 1991. Implications and concepts of polyamine nucleic acid interactions. *J. Cell. Biochem.* **46**: 37-47.

Franklin, R. E., and Gosling, R. G. 1953a. Evidence for 2-chain helix in crystalline structure of sodium deoxyribonucleate. *Nature (London)* **172**: 156-157.

Franklin, R. E., and Gosling, R. G. 1953b. Molecular configuration in sodium thymonucleate. *Nature (London)* **171**: 740-741.

Frederick, C., Quigley, G., Teng, M., Coll, M., Van Der Marel, G., Van Boom, J., Rich, A., and Wang, A. 1989. Molecular structure of an A-DNA decamer d(ACCGGCCGGT). *Eur. J. Biochem.* **181**: 295-307.

Fujii, S., Wang, A. H. J., Quigley, G. J., Westerink, H., van der Meral, G., van Boom, J. H., and Rich, A. 1985. The octamers d(CGCGCGCG) and d(CGCATGCG) both crystallize as Z-DNA in the same hexagonal lattice. *Biopolymers* **24**: 243-250.

Fujii, S., Wang, A. H. J., van der Marel, G., van Boom, J. H., and Rich, A. 1982. Molecular structure of $d(m^5dC\text{-}dG)_3$: the role of the methyl group on 5-methyl cytosine in stabilizing Z-DNA. *Nucleic Acids Res.* **10**: 7879-7892.

Fuller, W., Wilkins, M. H. F., Wilson, H. R., and Hamilton, L. D. 1965. The molecular configuration of deoxyribonucleic acid. IV. X-ray diffraction study of the A-form. *J. Mol. Biol.* **12**: 60-80.

Futscher, B. W., Rice, J. C., Ho, P. S., and Dalton, W. S. 199X. Transcriptional activation of human MDR1 associated with methylation of its 5' CpG island. (submitted).

Gaasterland, T., Andersson, S., and Sensen, C. Magpie Genome Sequencing Project List. Internet WWW page at *<http://www.mcs.anl.gov/home/gaasterl/genomes.html>* (version current 12 Nov. 1997).

Gao, Y. G., Sriram, M., and Wang, A. H. J. 1993. Crystallographic studies of metal ion-DNA interactions: different binding modes of cobalt(II), copper(II) and barium(II) to N^7 of guanines in Z-DNA and a drug-DNA complex. *Nucleic Acids Res.* **21**: 4093-4101.

Geierstanger, B. H., Kagawa, T. F., Chen, S.-L., Quigley, G. J., and Ho, P. S. 1991. Base-specific binding of copper(II) to Z-DNA: the 1.3-Å single crystal structure of $d(m^5CGUAm^5CG)$ in the presence of $CuCl_2$. *J. Biol. Chem.* **266**: 20185-20191.

Gessner, R. V., Frederick, C. A., Quigley, G. J., Rich, A., and Wang, A. H.J. 1989. The molecular structure of the left-handed Z-DNA double helix at 1.0-Å atomic resolution: geometry, conformation, and ionic interactions of d(CGCGCG). *J. Biol. Chem.* **264**: 7921-7935.

Gessner, R. V., Quigley, G. J., and Egli, M. 1994. Comparative Studies of high resolution Z-DNA crystal structures. *J. Mol. Biol.* **236**: 1154-1168.

Gessner, R. V., Quigley, G. J., Wang, A. H.-J., van der Marel, G. A., van Boom, J. H., and Rich, A. 1985. Structural basis for stabilization of Z-DNA by cobalt hexaammine and magnesium cations. *Biochemistry* **24**: 237-240.

Ginell, S. L., Kuzmich, S., Jones, R. A., and Berman, H. M. 1990. Crystal and molecular structure of a DNA duplex containing the carcinogenic lesion O^6-methylguanine. *Biochemistry* **29**: 10461-10465.

Gray, D., and Bollum, F. 1974. A circular dichroism study of poly(dG), poly(dC) and poly(dG•dC). *Biopolymers* **13**: 2087-2102.

Griffith, J. D. 1976. Visualization of prokaryotic DNA in a regularly condensed chromatin-like fiber. *Proc. Natl. Acad. Sci. USA* **73**: 563-567.

Gross, D. S. And Garrad. 1986. The ubiquitous potential Z-forming sequence of eukaryotes, $(dT\text{-}dG)_n \bullet (dC\text{-}dA)_n$ is not detectable in the genomes of eubacteria, archaebacteria, or mitochondria. *Mol. Cell. Biol.* **6**: 3010-3013.

Grosschedl, R. 1995. Higher-order nucleoprotein complexes in transcription: analogues with site-specific recombination. *Curr. Opin. Cell Biol.* **7**: 362-370.

Guzikevich-Guerstein, G., and Shakked, Z. 1996. A novel form of the DNA double helix imposed on the TATA-box by the TATA-binding protein. *Nature Struct. Biol.* **3**: 32-37.

Hamada, H., and Kakunaga, T. 1982. Potential Z-DNA forming sequences are highly dispersed in the human genome. *J. Cell. Biochem.* **6**: 944.

Hamada, H., Petrino, M. G., and Kakunaga, T. 1982. A novel repeated element with Z-DNA forming potential is widely found in evolutionarily diverse eukaryotic genomes. *Proc. Natl. Acad. Sci., USA* **79**: 6465-6469.

Haschmeyer, A. E. V., and Rich, A. 1967. Nucleoside conformations: An analysis of steric barriers to rotation about the glycosidic bond. *J. Mol. Biol.* **27**: 369-384.

Heinemann, U., and Alings, C. 1989. Crystallographic study of one turn of G/C rich B-DNA. *J. Mol. Biol.* **210**: 369-381.

Hieter, P., and Boguski, M. 1997. Functional genomics: it's all how you read it. *Science* **278**: 601-602.

Ho, P. S., Ellison, M. J., Quigley, G. J., and Rich, A. 1986. A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences. *EMBO J.* **5**: 2737-2744.

Ho, P. S., Frederick, C. A., Quigley, G. J., van der Marel, G. A., van Boom, J. H., Wang, A. H. J., and Rich, A. 1985. GT wobble base-pairing in Z-DNA at 1.0 angstrom atomic resolution: the crystal structure of d(CGCGTG). *EMBO J.* **4**: 3617-3623.

Ho, P. S., Frederick, C. A., Saal, D., Wang, A. H.-J., and Rich, A. 1987. The interactions of ruthenium hexaammine with Z-DNA: crystal structure of a $Ru(NH_3)_6^{+3}$ salt of d(CGCGCG) at 1.2 Å resolution. *J. Biomol. Struct. Dynam.* **4**: 521-534.

Ho, P. S., Kagawa, T. F., Tseng, K., Schroth, G. P., and Zhou, G. 1991. Prediction of a crystallization pathway for Z-DNA hexanucleotides. *Science* **254**: 1003-1006.

Holland, J. H. 1992. Genetic algorithms. *Sci. Am.* 66-72.

Howell, M. L., Schroth, G. P., and Ho, P. S. 1996. Sequence-dependent effects of spermine on the thermodynamics of the B-DNA to Z-DNA transition. *Biochemistry* **35**: 15373-15382.

Ivanov, V., and Krylov, D. 1992. A-DNA in solution as studied by diverse approaches. *Methods Enzymol* **211**: 111-127.

Ivanov, V., Minchenkova, L., Minyat, E., Frank-Kamenetskii, F., and Schyolkina, A. 1974. The B to A transition of DNA in solution. *J. Mol. Biol.* **87**: 817-833.

Ivanov, V., Minchenkova, L., Schyolkina, A., and Poletayev, A. 1973. Different conformations of double-stranded nucleic acid in solution as revealed by circular dichroism. *Biopolymers* **12**: 89-110.

Ivanov, V., I, Minchenkova, L. E., Chernov, B. K., McPhie, P., Ryu, S., Garges, S., Barber, A. M., Zhurkin, V. B., and Adhya, S. 1995. CRP-DNA complexes: Inducing the A-like form in the binding sites with an extended central spacer. *J. Mol. Biol.* **245**: 228-240.

Ivanov, V. I., Krylov, D. Y., and Minyat, E. E. 1985. Three-state diagram for DNA. *J. Biomol. Struct. Dynam.* **3**: 43-55.

Ivanov, V. I., Krylov, D. Y., Minyat, E. E., and Minchenkova, L. E. 1983. B-A transition in DNA. *J. Biomol. Struct. Dynam.* **1**: 453-476.

Jones, D. T. 1994. De novo protein design using pairwise potentials and a genetic algorithm. *Protein Sci.* **3**: 567-574.

Jones, G., Willett, P., Glen, R. C. 1994. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **245**: 43-53.

Jovin, T. M., McIntosh, L. P., Arndt-Jovin, D., Zarling, D. A., Robert-Nicoud, M., van de Sande, J. H., and Jorgenson, K. F. 1983. Left-handed DNA: from synthetic polymers to chromosomes. *J. Biomol. Struct. Dynam.* **1**: 21-57.

Jovin, T. M., Soumpasis, D. M., and McIntosh, L. P. 1987. The transition between B-DNA and Z-DNA. *Annu. Rev. Phys. Chem.* **38**: 521-560.

Kagawa, T. F., Geierstanger, B. H., Wang, A. H.J., and Ho, P. S. 1991. covalent modification of guanine bases in double-stranded DNA: the 1.2-Å Z-DNA structure of d(CGCGCG) in the presence of $CuCl_2$. *J. Biol. Chem.* **266**: 20175-20184.

Kagawa, T. F., Howell, M. L., Tseng, K., and Ho, P. S. 1993. Effects of base substituents on the hydration of B- and Z-DNA: correlations to the B- to Z-DNA transition. *Nucleic Acids Res.* **21**: 5978-5986.

Kagawa, T. F., Stoddard, D., Zhou, G., and Ho, P. S. 1989. Quantitative analysis of DNA secondary structure from solvent-accessible surfaces: the B- to Z-DNA transition as a model. *Biochemistry* **28**: 6642-6651.

Kang, H., and Johnson, W. 1994. Infrared linear dichroism reveals that A-, B- and C-DNAs in films have bases highly inclined from perpendicular to the helix axis. *Biochemistry* **33**: 8330-8338.

Krylov, D. Y., Makarov, V. L., and Ivanov, V. I. 1990. The B-A transition in superhelical DNA. *Nucleic Acids Research* **18**: 759-761.

Kubinec, M. G., and Wemmer, D. E. 1992. NMR evidence for DNA bound water in solution. *J. Am. Chem. Soc.* **114**: 8739-8740.

Kumar, V. D., Harrison, R. W., Andrews, L. C., and Weber, I. T. 1992. Crystal structure at 1.5 Å Resolution of d(CGCICICG), an octanucleotide containing inosine, and its comparison with d(CGCG) and d(CGCGCG) structures. *Biochemistry* **31**: 1541-1550.

Kumar, V. D., and Weber, I. T. 1993. Crystal structure of a Z-DNA hexamer d(CGCICG) at 1.7Å resolution: inosine•cytidine base-pairing, and comparison with other Z-DNA structures. *Nucleic Acids Res.* **21**: 2201-2208.

LeGrand, S., and Merz, K. J. 1994. The genetic algorithm and protein tertiary structure prediction. In *The protein folding problem and tertiary structure prediction*, K. J. Merz and S. LeGrand, eds. (Boston: Birkhauser), pp. 109-124.

Liepinsh, E., Otting, G., and Wuthrich, K. 1992. NMR observation of individual molecules of hydration water bound to DNA duplexes: direct evidence for a spine of hydration water present in aqueous solution. *Nucleic Acids Res.* **20**: 6549-6553.

Luisi, B. 1995. DNA-protein interactions at high resolution. In *DNA-protein: structural interactions*, D. Lilley, ed. (New York: IRL Press), pp. 1-48.

Malinina, L., Urpi, L., Salas, X., Huynh-Dinh, T., and Subirana, J. A. 1994. Recombination-like structure of d(CCGCGG). *J. Mol. Biol.* **243**: 484-493.

Marvin, D. A., Spencer, M., Wilkins, M. H. F., and Hamilton, L. D. 1961. The molecular configuration of DNA. III. X-ray diffraction study of the C form of the lithium salt. *J. Mol. Biol.* **3**: 547-565.

McDonnell, N. B., and Preisler, R. S. 1989. Hydrophobic moieties in cations, anions and alcohols promote the B-to-Z transition in poly [d(G-C)] and poly [d(G-m^{5}C)]. *Biochem. Biophys. Res. Comm.* **164**: 426-433.

McLean, M. J., Lee, J. W., and Wells, R. D. 1988. Characteristics of Z-DNA helicies formed by imperfect (purine-pyrimidine) sequences in plasmids. *J. Biol. Chem.* **263**: 7378-7385.

Mei, H. Y., and Barton, J. K. 1988. Tris(tetramethylphenanthroline)ruthenium(II): a chiral probe that cleaves A-DNA conformations. *Proc. Natl. Acad. Sci., USA* **85**: 1339-1343.

Melander, W., and Horvath, C. 1977. Salt effects on hydrophobic interactions in precipitation and chromatography of proteins: An interpretation of the lyotropic series. *Arch. Biochem. Biophys.* **183**: 200-215.

Minchenkova, L. E., Schyolkina, A. K., Chernov, B. K., and Ivanov, V. I. 1986. CC/GG contacts facilitate the B to A transition of DNA in solution. *J. Biomol. Struct. Dynam.* **4**: 463-475.

Moller, A., Nordheim, A., Kozlowski, S. A., Patel, D., and Rich, A. 1984. Bromination stabilizes poly(dG-dC) in the Z-DNA form under low salt conditions. *Biochemistry* **23**: 54-62.

Mooers, B., Eichman, B., and Ho, P. S. 1997. The structures and relative stabilities of d(GG) reverse Hoogsteen, d(GT) reverse wobble, and d(GC) reverse Watson-Crick base pairs in DNA crystals. *J. Mol. Biol.* **269**: 796-810.

Mooers, B. H. M., Schroth, G. P., Baxter, W. W., and Ho, P. S. 1995. Alternating and non-alternating dG-dC hexanucleotides crystallize as canonical A-DNA. *J. Mol. Biol.* **249**: 772-784.

Moore, M. H., van Meervelt, L., Salisbury, S. A., Kong Thoo Lin, P., and Brown, D. M. 1995. Direct observation of two base-pairing modes of a cytosine-thymine analogue with guanine in a DNA Z-form duplex: significance of for base analogue mutagenesis. *J. Mol. Biol.* **251**: 665-673.

Morgan, J. E., Blankenship, J. W., and Matthews, H. R. 1986. Association constants for the interaction of double-stranded and single-stranded DNA with spermine, spermidine, putrescine, diaminopropane, N^1- and N^8-acetylspermidine, and magnesium: Determination from analysis of the broadening of thermal denaturation curves. *Arch. Biochem. Biophys.* **246**: 225-232.

Musgrave, D. R., Sandman, K. M., Stroup, D., and Reeve, J. N. 1992. DNA-binding proteins and genome topology in thermophilic prokaryotes. In *Biocatalysis at extreme temperatures - Enzyme systems near and above 100°C.*, M. W. W. Adams and R. M. Kelly, eds. (Washington, DC: American Chemical Society), pp. 174-188.

Ohishi, H., Nakanishi, I., Inubushi, K., van der Marel, G. A., van Boom, J. H., Rich, A., Wang, A. H. J., Hakoshima, T., and Tomita, K. 1996. Interaction between the left-handed Z-DNA and polyamine-2: the

crystal structure of $d(CG)_3$ and spermidine complex. *FEBS Lett.* **391**: 153-156.

Parkinson, G. N., Arvanitis, G. M., Lessinger, L., Ginell, S. L., Jones, R., Gaffney, B., and Berman, H. M. 1995. Crystal and molecular structure of a new Z-DNA crystal form: d[CGT(2-NH_2-A)CG] and its platinated derivative. *Biochemistry* **34**: 15487-15495.

Peck, L. J., and Wang, J. C. 1983. Energetics of B-to-Z transition in DNA. *Proc. Natl. Acad. Sci., USA* **80**: 6206-6210.

Peterson, M. R., Harrop, S. J., McSweeney, S. M., Leonard, G. A., Thompson, A. W., Hunter, W. N., and Helliwell, J. R. 1996. MAD phasing strategies explored with a brominated oligonucleotide crystal at 1.65 Å resolution. *J. Synch. Rad.* **3**: 24-34.

Peticolas, W., Wang, Y., and Thomas, G. 1988. Some rules for predicting the base-sequence dependence of DNA conformation. *Proc. Natl. Acad. Sci., USA* **85**: 2579-2583.

Pohl, R. M., and Jovin, T. M. 1972. Salt induced co-operative conformational change of a synthetic DNA: Equilibrium and kinetic studies with poly (dG-dC). *J. Mol. Biol.* **647**: 375-396.

Preisler, R. S., Chen, H. H., Colombo, M. F., Choe, Y., Short, B. J. J., and Rau, D. C. 1995. The B form to Z form transition of poly (dG-m^5dC) is sensitive to neutral solutes through an osmotic stress. *Biochemistry* **34**: 14400-14407.

Quadrifoglio, F., Manzini, G., and Yathindra, N. 1984. Short oligonucleotides with $d(C\text{-}C)_n$ sequences do not assume left-handed conformation in high salt conditions. *J. Mol. Biol.* **175**: 419-423.

Rahmouni, A. R., and Wells, R. D. 1989. Stabilization of Z-DNA *in vivo* by localized supercoiling. *Science* **246**: 358-363.

Ramakrishnan, B., and Viswamitra, M. A. 1988. Crystal and molecular structure of the ammonium salt of the dinucleoside monophosphate of d(CpG). *J. Biomol. Struct. Dynam.* **6**: 511-523.

Reich, Z., Ghirlando, R., and Minsky, A. 1991. Secondary conformational polymorphism of nucleic acids as a possible functional link between cellular parameters and DNA packaging processes. *Biochemistry* **30**: 7828-7836.

Riazance, J., Johnson, W., McIntosh, L., and Jovin, T. 1987. Vacuum UV circular dichroism is diagnostic for the left-handed Z-form of poly[d(A-C)•d(G-T)] and other polydeoxynucleotides. *Nucleic Acids Res.* **15**: 7627-7636.

Rich, A. 1993. DNA comes in many forms. *Gene.* **135**: 99-109.

Rich, A., Nordheim, A., and Wang, A. H. J. 1984. The chemistry and biology of left-handed Z-DNA. *Ann. Rev. Biochem.* **53**: 791-846.

Rouviere-Yaniv, J. 1978. Localization of the HU protein on the Escherichia coli nucleoid. In *Cold Spring Harbor Symp. Quant. Biol.* **42**: 439-447.

Rowen, L., Mahairas, G., and Hood, L. 1997. Sequencing the human genome. *Science* **278**: 605-607.

Sadsivan, C., and Gautham, N. 1995. Sequence-dependent microheterogeneity of Z-DNA: the crystal and molecular structures of d(CACGCG)•d(CGCGTG) and d(CGCACG)•d(CGTGCG). *J. Mol. Biol.* **248**: 918-930.

Saenger, W. 1984. Principles of Nucleic Acid Structure (New York, NY: Springer-Verlag).

Sagi, J., Szemzo, A., Otvos, L., Vorlikckova, M., and Kypr, J. 1991. Destabilization of the duplex and the high-salt Z-form of poly(dG-methyl^{5}dC) by substitution of ethyl for the 5-methyl group. *Int. J. Biol. Macromol.* **13**: 329-336.

Sarma, M. H., Gupta, G., and Sarma, R. H. 1986. 500-MHz ^{1}H NMR study of poly(dG)•poly(dC) in solution using one-dimensional nuclear overhauser effect. *Biochemistry* **25**: 3659-3665.

Schneider, B., Cohen, D., and Berman, H. 1992. Hydration of DNA bases: analysis of crystallographic data. *Biopolymers* **32**: 725-750.

Schneider, B., Ginell, S. L., Jones, R., Gaffney, B., and Berman, H. M. 1992. Crystal and molecular structure of a DNA fragment containing a 2-aminoadenine modification: the relationship between conformation, packing, and hydration in Z-DNA hexamers. *Biochemistry* **31**: 9622-9628.

Schroth, G. P., Chou, P. J., and Ho, P. S. 1992. Mapping Z-DNA in the human genome. *J. Biol. Chem.* **267**: 11846-11855.

Schroth, G. P. and Ho, P. S. 1995. Occurrence of potential cruciform and H-DNA forming sequences in genomic DNA. *Nucleic Acids Res.* **23**: 1977-1983.

Schroth, G. P., Kagawa, T. F., and Ho, P. S. 1993. Structure and thermodynamics of nonalternating C•G base pairs in Z-DNA: The 1.3Å crystal structure of the asymmetric hexanucleotide. *Biochemistry* **32**: 13381-13392.

Shakked, D., Rabinovich, D., Cruse, W. B. T., Egert, E., Kennard, O., Sals, G., Salisbury, S. A., and Viswamitra, M. A. 1981. Crystalline A-DNA: the x-ray analysis of the fragment d(G-G-T-A-T-A-C-C). *Proc. Roy. Soc.* **B 213**: 479.

Sprecher, C., Baase, W., and Johnson, W. 1979. Conformation and circular dichroism of DNA. *Biopolymers* **18**: 1009-1019.

Stetlow, P. 1992. DNA in dormant spores of Bacillus species is in an A-like conformation. *Molecular Microbiology* **6**: 563-567.

Stonington, O. G., and Pettijohn, D. E. 1971. The folded genome of *Eschericia coli*. *J. Mol. Biol.* **68**: 6-9.

Sun, S. 1993. Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Protein Sci.* **5**: 762-785.

Tabor, C. W., and Tabor, H. 1984. Polyamines. *Ann. Rev. Biochem.* **53**: 749-790.

Takahara, P., Rosenzweig, A., Frederick, C., and Lippard, S. 1995. Crystal structure of double-stranded DNA containing the major adduct of the anticancer drug cisplatin. *Nature (London)* **377**: 649-652.

Teng, M., Liaw, Y. C., van der Marel, G. A., van Boom, J. H., and Wang, A. H. J. 1989. Effects of the O2' hydroxyl group on Z-DNA conformation: structure of Z-RNA and (araC)-[Z-DNA]. *Biochemistry* **28**: 4923-4928.

Tereshko, V., and Milinina, L. 1990. Different forms of the double helix in the pCpGpCpGpCpG crystals. *J. Biomol. Struct. Dynam.* **7**: 827-836.

Thomas, T. J., Gunnia, U. B., and Thomas, T. 1991. Polyamine-induced B-DNA to Z-DNA conformational transition of a plasmid DNA with $(dG\text{-}dC)_n$ insert. *J. Biol. Chem.* **266**: 6137-6141.

Thomas, T. J., and Thomas, T. 1994. Polyamine-induced Z-DNA conformation in plasmids containing $(dA\text{-}dC)_n \bullet d(G\text{-}dT)_n$ inserts and increased binding of lupus autoantibodies to the Z-DNA form of plasmids. *Biochem. J.* **298**: 485-491.

Timsit, Y., and Moras, D. 1992. Crystallization of DNA. *Methods Enzymol.* **221**: 409-449.

van de Ven, F., and Hilbers, C. 1988. Nucleic acids and nuclear magnetic resonance. *Eur. J. Biochem.* **178**: 1-38.

van Meervelt, L., Moore, M. H., Lin, P. K. T., Brown, D. M., and Kennard, O. 1990. Molecular and crystal structure of d(CGCGmo^{4}CG): N^4-methoxycytosine-guanine base pairs in Z-DNA. *J. Mol. Biol.* **216**: 773-781.

Verdaguer, N., Aymami, J., Fernandez-Forner, D., Fita, I., Coll, M., Huynh-Dinh, T., Igolen, J., and Subirana, J. 1991. Molecular structure of a complete turn of A-DNA. *J. Mol. Biol.* **221**: 623-635.

Voet, D., and Voet, J. 1990. *Biochemistry* (New York: Wiley & Sons).

Vologodskii, A. V., and Frank-Kamenetskii, M. D. 1984. Left-handed Z form in superhelical DNA: A theoretical study. *J. Biomol. Struct. Dynam.* **1**: 1325-1333.

Vorlickova, M., and Sagi, J. 1991. Transitions of poly (dI-dC), poly (dI-methyl^{5}dC) and poly (dI-bromo^{5}dC) among and within the B-, Z-, A- and X-DNA families of conformations. *Nucleic Acids Res.* **21**: 2343-2347.

Vorlickova, M., Subirana, J. A., Chladkova, J., Tejralova, I., Huynh-Dinh, T., Arnold, L., and Kyper, J. 1996. Comparison of the solution and crystal conformations of (G+C) rich fragments of DNA. *Biophysical J.* **71**: 1530-1538.

Wahl, M. C., Tao, S. T., and Sundaralingam, M. 1996. Crystal structure of the B-DNA hexamer d(CTCGAG): model for an A-to-B transition. *Biophysical J.* **70**: 2857-2866.

Wang, A. H.-J., Hakoshima, T., van der Marel, G., van Boom, J. H., and Rich, A. 1984. AT base pairs are less stable than GC base pairs in Z-DNA: the crystal structure of d(m^{5}CGTAm^{5}CG). *Cell* **37**: 321-331.

Wang, A. H.-J., Quigley, G. J., Kolpak, F. J., Crawford, J. L., van Boom, J. H., van der Marel, G., and Rich, A. 1979. Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature (London)* **282**: 680-686.

Wang, A. H. J., Gesser, R. V., van der Marel, G. A., van Boom, J. H., and Rich, A. 1985. Crystal structure of Z-DNA without an alternating purine-pyrimidine sequence. *Proc. Natl. Acad. Sci., USA* **82**: 3611-3615.

Wang, A. H. J., Quigley, G. J., Kolpak, F. J., van der Marel, G., van Boom, J. H., and Rich, A. 1981. Left-handed double helical DNA: variations in the backbone conformation. *Science* **211**: 171-176.

Wang, L., and Keiderling, T. A. 1993. Helical nature of poly(dI-dC)•poly(dI-dC). Vibrational circular dichroism results. *Nucleic Acids Res.* **21**: 4127-4132.

Warne, S. E., and deHaseth, P. L. 1993. Promoter recognition by *Escherichia coli* RNA polymerase. Effects of single base pair deletions and insertions in the spacer DNA separating the -10 and -35 regions are dependent on spacer DNA sequence. *Biochemistry* **32**: 6134-6140.

Watson, J. D., and Crick, F. H. C. 1953. A structure for deoxyribose nucleic acid. *Nature (London)* **171**: 737.

Werel, W., Schickor, P., and Heumann, H. 1991 . Flexibility of the DNA enhances promoter affinity of *Escherichia coli* RNA polymerase. *EMBO J.* **10**: 2589-2594.

Werner, M., Huth, J., Gronenborn, A., and Clore, G. 1995. Molecular basis of human 46X,Y sex reversal revealed from the three-dimensional solution structure of the human SRY-DNA complex. *Cell* **81**: 5705-5714.

Wing, R. M., Drew, H. R., Takano, T., Broka, C., Tanaka, S., Itakura, K., and Dickerson, R. E. 1980. Crystal structure analysis of a complete turn of B-DNA. *Nature (London)* **287**: 755-758.

Xu, Q., Rao, S., Jampani, B., and Braunlin, W. 1993a. Rotational dynamics of hexaamminecobalt(II) bound to oligomeric DNA: correlation with cation-induced structural transitions. *Biochemistry* **32**: 11754-11760.

Xu, Q., Shoemaker, R. K., and Braunlin, W. H. 1993b. Induction of B-A transitions of deoxyoligonucleotides by multivalent cations in dilute aqueous solution. *Biophysical J.* **65**: 1039-1049.

Zhang, H., van der Marel, G., van Boom, J., and Wang, A. H. J. 1992. Conformational perturbation of the anticancer nucleoside arabinosylcytosine on Z-DNA: molecular structure of $(araC\text{-}dG)_3$ at 1.3 Å resolution. *Biopolymers* **32**: 1559-1569.

Zhang, X., and Mathews, C. K. 1994. Effect of DNA cytosine methylation upon deamination-induced mutagenesis in a natural target sequence in duplex DNA. *J. Biol. Chem.* **269**: 7066-7069.

Zhou, G., and Ho, P. S. 1990. Stabilization of Z-DNA by demethylation of thymine bases: 1.3-Å single-crystal structure of $d(m^5CGUAm^5CG)$. *Biochemistry* **29**: 7229-7236.