AN ABSTRACT OF THE DISSERTATION OF

Patricia A. Khuu for the degree of Doctor of Philosophy in Biochemistry and Biophysics presented on June 4, 2008.

Title: Analyses of Noncanonical DNA Structures.

Abstract approved:_____
Pui Shing Ho

The structural polymorphic nature of DNA has been long been recognized since the model of the right-handed B-DNA duplex proposed by James Watson and Francis Crick was accepted. Its malleability is thought to be critical in protein recognition and manipulation. In particular, it can form Holliday junctions, four-way DNA structural intermediates of homologous recombination mediated events in which four nucleic acid strands complex to give rise to double-helical arms extending from a central point. The variable recognition specificities of resolvases, a group of proteins that dissociate Holliday junctions into discrete duplexes, are addressed in this dissertation as related to the different structural conformations that can be adopted by junctions.

We also describe an unusual base pair observed in the crystal structure of a Holliday junction. The wobbled adenine-thymine base pair can be best ascribed to the assumption of a rare tautomeric form by one of the bases. Such observation provides unique and physiologically relevant evidence for a rare

nucleotide base tautomerization, with profound biological and chemical implications that deserve further investigation.

Finally, the evolutionary emergence of another known DNA structure, the left-handed Z-DNA, is also studied along with three other GC-rich genomic elements. The possible functions of Z-DNA have only been recently elucidated though it was the first single-crystal X-ray structure of a DNA double helix. In our phylogenomic analysis of the genomes of sixteen organisms, ranging from cyanobacteria to complex eukaryotes, we offer new insights into its evolutionary emergence near the transcription start site. A model is derived which correlates its emergence with other transcriptional regulatory sequences and evolutionary gain of function. Thus, the studies presented in this thesis expand our knowledge of noncanonical DNA structures and provoke questions about their functions in the cell.

Analyses of Noncanonical DNA Structures

by

Patricia A. Khuu

A DISSERTATION

submitted to

Oregon State University

In partial fulfillment of
the requirements for the
degree of

Doctor of Philosophy

Presented June 4, 2008
Commencement June 2009

Doctor of Philosophy dissertation of Patricia A. Khuu presented on June 4, 2008.

APPROVED:

Major Professor, representing Biochemistry and Biophysics

Chair of the Department of Biochemistry and Biophysics

Dean of the Graduate School

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

Patricia A. Khuu, Author

ACKNOWLEDGMENTS

In the laboratory of P. Shing Ho, I thank my mentor, Dr. P. Shing Ho, for his support and guidance and, particularly, for fostering leaps of the imagination while providing well-grounded restraint in scientific thinking. I thank Dr. Andrea R. Voth for her continued support and many fun discussions under the artificial light. Thank you to Dr. Franklin A. Hays for his enthusiasm, his invaluable patience and generosity with his computer knowledge and for sharing his love of science.

Special thanks Dr. H. Richard Faber for sharing his knowledge about crystallography and perspectives on life and science. Thank you to my committee members for their comments and doorway discussions. Very special thanks to Olga Golonzhka and Jeffery Monette for their friendship and cheers throughout the years. With great fondness to the staff, faculty and student members of the Department of Biochemistry and Biophysics, thank you for their ongoing encouragement and support.

CONTRIBUTION OF AUTHORS

P. Shing Ho was involved in the design, analysis and writing of each manuscript. Andrea R. Voth was involved in the writing and analysis of Chapter 2. Franklin A. Hays was involved in the analysis of Chapter 2, and he designed and determined the initial crystallization conditions for the oligonucleotide sequences used in Chapter 4. Maurice Sandor and Jennifer DeYoung were involved in the modifications to the Zhunt program in Chapter 3.

TABLE OF CONTENTS

Page

TABLE OF CONTENTS (continued)

Page

LIST OF FIGURES

Figure Page

LIST OF FIGURES (continued)

Figure Page

LIST OF TABLES

Analyses of Noncanonical DNA Structures

Chapter 1

Introduction

DNA, the genetic material that contains the information necessary for organismal development and cellular activities, is fairly simple, a polymer comprising of four standard nucleic acid units. Two are purines, adenine (A) and guanine (G), and two are pyrimidines, cytosine (C) and thymine (T). As discovered by Erwin Chagraff, the amount of A is proportionate to that of T, while the amount of C equals that of G. The observation, along with early x-ray diffraction patterns, led to the development of the DNA double helix model by James Watson and Francis Crick, who proposed that DNA exists as a helical secondary structure comprised of two complementary, antiparallel strands stabilized by interstrand hydrogen bonds. The hydrogen bonds are formed between the bases, where A is paired with T, and C with G. The model concurred with the available data while providing an elegantly simple structure that allowed for a mechanism of genetic information transmission, intrinsic within the complementary strands (Watson and Crick 1953).

Since the Watson and Crick (W-C) structure, denoted as B-DNA or the B form, has become part of the scientific canon, DNA is discovered to be highly

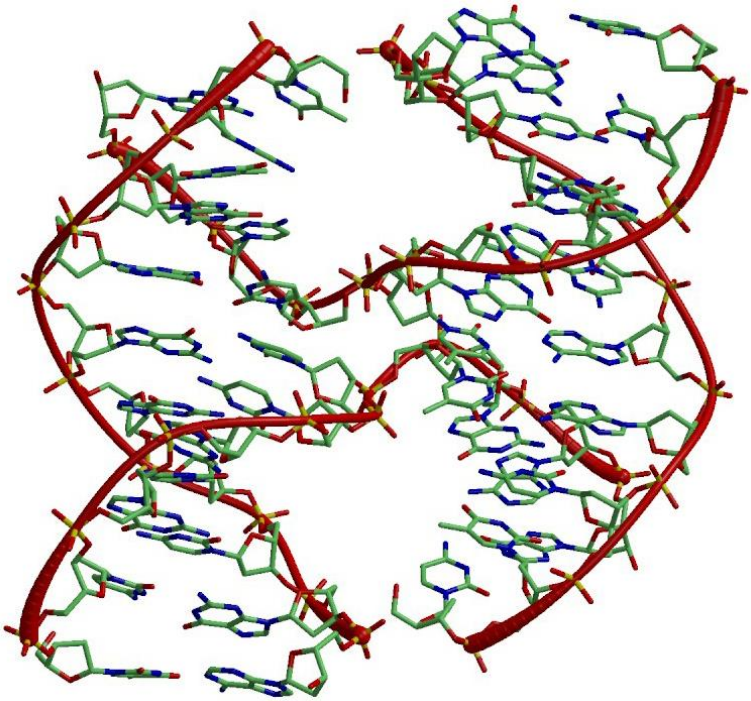
polymorphic, capable of assuming secondary structures of unexpected complexity. The DNA double helix is found to adopt other duplex conformations, such as the less hydrated A-DNA and the left-handed Z-DNA. Compared to B-DNA, the broader A-DNA, or A form, is also right-handed, but is observed under conditions of low humidity and possesses bases that have shorter rises and larger inclinations with respect to the helical axis (Arnott 1999; Arnott 2006). Z-DNA, conversely, is left-handed and narrower, with alternating *anti* pyrimidines and *syn* purine sequences (Rich and Zhang 2003). Cruciforms and Holliday junctions can arise from inverted repeat sequences to create four-way junctions that have B-DNA arms and are reflective of recombination intermediates (Eichman *et al.* 2002). Moreover, though the proposed W-C base pairings and B-DNA double helix structure remain the standard, studies reveal that the bases are capable of forming non-W-C hydrogen bonds that facilitate the formation of structures to include triplexes and tetraplexes. Triplex structures (triple helices, H-DNA) involve a third strand that is appended to a W-C duplex through Hoogsteen hydrogen bonds. Tetraplexes are observed in repeating tracts of guanines which can self-associate as tetrads, also through Hoogsteen hydrogen bonds (Arnott 1999).

The polymorphism of DNA secondary structures reflects a plasticity that is observed at the local level. Segments of DNA are differentially polymorphic, with many simple DNA repeat sequences capable of adopting at least two conformations. Most mainly exist as canonical B-DNA, but can transiently adopt other conformations.

Thus, while B-DNA is the predominant conformation, the capacity of DNA to adopt other conformations is provocative. It posits whether local non-B-DNA conformations are evolutionary artifacts or products of evolutionary selection, how primary sequence and secondary structure correlate, how non-B-DNA conformations may be involved in cellular function, and if they may have deleterious impacts. Continuing studies of nucleic acid structures have provided some elucidation into their possible functions. Some structures are thought to be important for protein recognition, demarcation of genic regions and induction of recombination (Sekharudu *et al.* 1993; Kono and Sarai 1999; Mito *et al.* 2005; Wang and Vasquez 2007). Nonetheless, given the nebulous grasp of how sequence dictates structural behavior and the persistent inability to accurately predict structure from sequence, the functional and physiological relevance of the structural polymorphism of DNA remains largely enigmatic and is derived from indirect evidence.

One approach to determine the biological significance of non-B-DBA structures has been the search for proteins specific to non-B-DNA structures. Identification of these proteins has provided evidence for the existence of these structures *in vivo.* One such structure is the Holliday junction. The four-way DNA junction was first proposed by Robin Holliday in 1964 as the central structural intermediate of recombination events (Holliday 1964). It is comprised of exchanging strands that form a branch point for four helices, allowing for two possible junction conformations, the mobile open-X and the locked stacked-X.

A.



B.

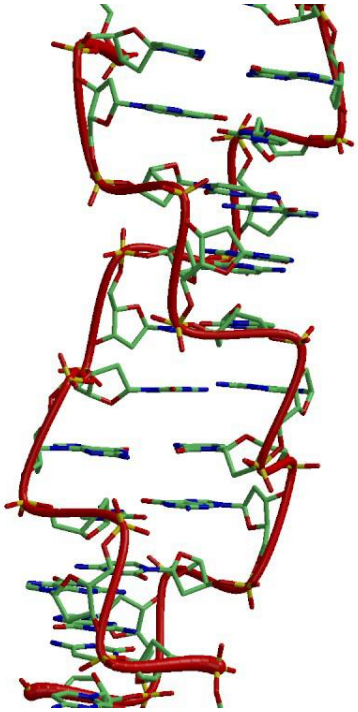


Figure 1.1. Molecular structure of (A) a stacked-X Holliday junction and (B) a Z-DNA duplex.

Conformational transition between the open- and stacked-X is observed in free solution (Lilley 2000).

The open-X conformation allows for branch migration of the junction as helices are splayed in an extended planar geometry under low cationic conditions. In the presence of higher cationic concentrations, two pairs of helical arms coaxially stack to form two duplexes (stacked-X) joined by exchanging strands and with no branch migration capability. The stacked-X conformation was first observed in electrophoretic (Duckett *et al.* 1988) and fluorescence resonance energy transfer (FRET) studies (Murchie *et al.* 1989), and later crystallized by our lab and Ortiz-Lombardı´ *et al.* (Ortiz-Lombardia *et al.* 1999; Eichman *et al.* 2000).

Two stacked-X junction isoforms are possible, defined by stacking partners. Nonetheless, a conformation bias for one isoform is observed that can be correlated with sequence, stacking interactions and electrostatic interactions (Liu *et al.* 2004; Liu *et al.* 2005). Subsequent single molecule FRET studies reveal that junctions can experience rapid exchange between the alternative stacking conformers, transiting through an open-X intermediate. Conformation bias influences the rate of exchange and dwell time in each conformation (McKinney *et al.* 2003).

As would be expected, proteins of diverse functions that were found to be highly selective for Holliday junctions have been implicated in pathways that include homologous recombination, DNA repair and resumption of arrested replication forks. These junction-specific proteins, complexed with other proteins,

rely on homologous recombination mediated mechanisms for the maintenance of genomic stability. Some examples include the prokaryotic RecA and its eukaryotic homolog, Rad51, which have been shown to facilitate homologous pairing and strand invasion (Henning and Sturzbecher 2003). BLM, a member of the RecQ family of DNA helicase, promotes branch migration of junctions to suppress improper DNA recombination events (Karow *et al.* 2000). Moreover, integration and excision of viral DNA into host genomes have been shown to require junction-specific enzymes, such as the bacteriophage λ-integrase and the related Cre recombinase, which catalyze site-specific DNA recombination via the formation of a Holliday junction (Chen *et al.* 2000; Subramaniam *et al.* 2003). Thus, the plethora of enzymes specific to Holliday junctions provides clear evidence for its existence *in* vivo and its biological relevance.

Of special interest to us are junction resolvases, which have been identified in diverse organisms (Aravind *et al.* 2000; Lilley and White 2001; West 2003; Khuu *et al.* 2006). These enzymes are critical for the dissociation of the linked junction duplexes, introducing cleavages near the point of strand exchange. The nicked strands in the dissociated duplexes are subsequently repaired by DNA ligases.

Resolvases bind specifically and tightly to junctions, with dissociation constants nearing K_d = 1 nM (Declais and Lilley 2008). Upon binding, biochemical studies suggest that the junction substrates become variably distorted by the resolvases despite their common enzymatic function. Three co-crystal structures of different resolvase-DNA complexes substantiate those findings to reveal very

diverse structural distortions imposed by the individual resolvases, RusA, T4 endonuclease VII and T7 endonuclease I. The crystal structure of RusA is bound to a duplex, and a symmetrical generated second duplex indicates that the bound junction substrate would be in the open-X conformation (Macmaster *et al.* 2006). Comparatively, the crystal structure of T4 endonuclease VII with a junction shows a junction substrate in a relatively planar stacked-X conformation (Biertumpfel *et al.* 2007). The crystal structure of T7 endonuclease I complexed with a junction also reveals a junction substrate in the stacked-X conformation, but the stacked-X conformation is highly distorted compared to a free junction and the substrate of T4 endonuclease VII (Hadden *et al.* 2007).

Nonetheless, while these co-crystal structures have provided invaluable new insights into molecular mechanism of resolvases, the conformations of the junction substrates, open- or stacked-X, initially recognized by the individual resolvases remain uncertain as the crystallized complexes may only reflect structural distortions induced after binding (Declais and Lilley 2008). Accordingly, in Chapter 2, we surveyed the structures of free and bound junction resolvases and present a model which correlates resolvase structure and its degree of sequence specificity with the initially recognized junction conformation**.** We predict that resolvases which recognize four-fold open-X structures would be tetrameric and exhibit some degree of sequence specificity, given that the open-X structure allows for branch migration to a specific sequence. Resolvases that

target the two-fold stacked-X structure would be dimeric and have no or minimal sequence specificity as the stacked-X conformation is already locked.

In Chapter 3, we further characterized the Holliday junction in a context that correlates with previous studies. Junctions previously studied in solution were sequence-locked, in that they were comprised of four different, nonhomologous sequences that were specifically complementary to form an immobile junction; otherwise, the junction would migrate out to the duplex ends. Crystal structures, conversely, involved junctions comprised of a singular, self-complementary sequence. We sought to bridge the disparity by crystallizing an immobile junction of four different sequences, similar to those of solution studies. The resultant junction crystal structure yielded unexpected insight into the nature of structure and sequence, with a potential significance to genomic integrity. We observe an anomalous A·T base-pairing in the purine-rich arm of the junction. The structure compellingly suggests that tautomerization of one of the participating bases produced the unusual base pair.

According to the canonical W-C model, component nucleotides adopt standard tautomeric forms that promote selective base-pairing of keto thymines with amino adenines and keto guanines with amino cytosines. The stability of these predominant tautomeric forms and their resulting canonical base pairs allows for generational transmissions of vital genetic information. The “rare tautomer hypothesis,” originally proposed by Watson and Crick in their seminal paper in 1953 and further developed by Topal and Fresco in 1976, however, offers

a mechanism that may undermine this fundamental biological scheme. The hypothesis postulates that bases could adopt other tautomeric forms to allow for formation of alternative, non-W-C mispairs which are not mismatches as they still retaining the geometry of W-C base pairs. These mispairs thus would be undetected and transmitted to subsequent generations as substitution point mutations (Fig. 1.2) (Watson and Crick 1953; Topal and Fresco 1976).

Nonetheless, the idea remains largely a hypothesis as only meager evidence, garnered under physiologically dubious conditions, exists for standard tautomers. Because their equilibrium frequency in solution is estimated to be on the order of 10^{-5}, special experimental conditions are necessary to generate quantitatively detectable rare tautomers for study of their biological relevance. Some experimental studies require extreme conditional induction, such as the gas phase, to observe unusual tautomeric forms. Shift of the equilibrium towards minor tautomeric forms are observed under high temperatures or low-dielectric medium. Unnatural (or modified) bases, like 2'-deozyisoguanosine (iG) and the nucleoside analogue dP (6-(2-deoxy-â-D-ribofuranosyl)-3,4-dihydro-8H-pyrimido[4,5-c][1,2]-oxazin-7-one), are used in NMR studies, because they can more easily be induced to structurally mimic rarely tautomers (Goodman and Ratliff 1983; Cheng *et al.* 2005; Podolyan *et al.* 2005). Other studies involve induction of tautomerization of the natural bases by metal ions. Rare tautomers observed in crystal structures have been induced by associated metals. Depending on type, proximal metal ions are thought to

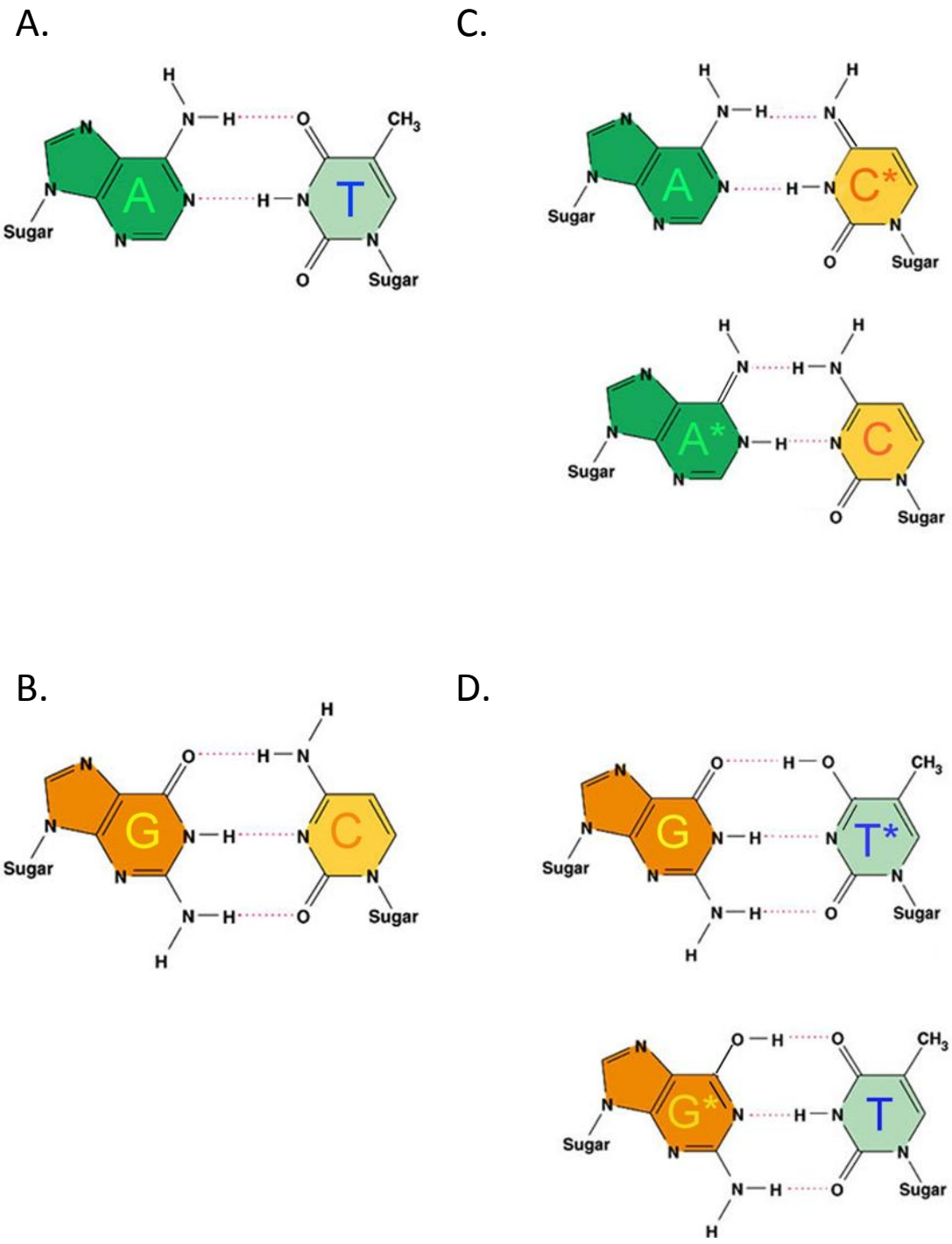


Figure 1.2. Standard Watson-Crick base pairs and mispairs. Canonical (A) adenine-thymine base pair and (B) guanine-cytosine base pair. The W-C geometry is maintained by alternative mispairs involving rare tautomeric forms *of (C) adenine and cytosine, and of (D) guanine and thymine.

destabilize the electron distribution of the base and change the tautomeric equilibria. Metal binding to specific sites of the base may cause shifting of the acidic proton to another site on the base, altering pK_a values and H-bonding donor and acceptor sites of the bases. Moreover, the metal ion may sterically impede complementary Watson-Crick base-pairing to generate alternative hydrogen-bonding patterns between complementary bases (Zamora *et al.* 1997).

None of these known factors that can induce tautomerization was observed in our structure. The structure was crystallized at room temperature, neutral pH and with only standard, natural bases. Thus, our structure of the rare tautomer may provide unique evidence for spontaneous tautomerization of a base, having profound mutagenic implications. In Chapter 3, we present a hypothesis that would account for its induction and stabilization by the primary sequence.

To further address the *in vivo* relevance of another non-B-DNA structure, in Chapter 4, we performed a phylogenomic analysis of four genomic elements across 17 prokaryotic and eukaryotic genomes: G+C content, CpG islands, NFI transcription factor binding sites and Z-DNA. Primarily, we were interested in Z-DNA, the left-handed helix which is a non-B-DNA structural element that has been extensively characterized and can be predicted from disparate sequences. The other three elements, in contrast, are defined by identifiable and quantifiable sequences.

Z-DNA is a left-handed double helix with a sugar phosphate backbone in a zigzag arrangement, hence the name (Fig. 1.1). It is formed by alternating sequences of purine residues adopting the *syn* conformation and pyrimidine residues remaining in the standard *anti* conformation. Though the most favoured sequence includes (dCdG)n, many other sequences can also transiently adopt the Z conformation, depending on conditions (Rich *et al.* 1984). Promoter regions of some transcriptionally active genes have been shown to readily form Z-DNA, only to revert back to B-DNA upon transcriptional down-regulation (Wittig *et al.* 1992). Formation of Z-DNA is further postulated to remove the negative supercoiling in the wake of a transcribing RNA polymerase (Liu and Wang 1987). Proteins specific to Z-DNA have been identified to include ADAR1, a nuclear-RNA-editing enzyme, (Herbert *et al.* 1995), DLM1, a protein linked to tumour tissues and interferon response (Fu *et al.* 1999), and E3L, a protein implicated in viral pathogenicity (Schwartz *et al.* 2001).

Results of our phylogenomic analysis are consistent with previous findings of the general occurrence of Z-DNA near the eukaryotic transcription start site. We offer new insights by following its evolutionary emergence in prokaryotic genomes through to early and more complex eukaryotic genomes. Our study reveals an evolutionary selection for Z-DNA sequences with increasing organismal complexity as it may have evolved more functional relevance. Moreover, we detect that it emerged at two distinct, though proximal, loci near the transcription start, possibly intimating an additional gain of function.

Our analyses of Holliday junction-specific resolvases, a wobbled base pair and the genomic occurrence of Z-DNA provide further evidence and new insights into how noncanonical structures may have evolved, perhaps, initially, as artifacts of cellular processes, to become functional elements in intricate, multi-tiered regulatory pathways.

Chapter 2

The Stacked-X DNA Holliday Junction and Protein Recognition

Patricia A. Khuu, Andrea Regier Voth, Franklin A. Hays and P. Shing Ho

Published in *J. Mol. Recognit.*

2006; 19: 234–242

2.1 Introduction

Homologous recombination is involved in a variety of cellular processes. Originally described as a means to generate genetic diversity by creating new gene combinations (Holliday 1964, 1974), homologous recombination is now recognized to be important for viral integration (Subramaniam *et al.* 2003), for maintaining genome stability (Flores-Rozas and Kolodner 2000) through recombination dependent repair of DNA lesions (Kreuzer 2004; Smith 2004) and restart of stalled replication forks (Cox *et al.* 2000; Cox 2001), and for proper segregation of homologous chromosomes during meiosis (McKim *et al.* 2002; Morrison *et al.* 2003; Kreuzer 2004; McKee 2004; Sherratt *et al.* 2004). Loss of recombination functions results in increased mutagenesis, mitotic and meiotic aneuploidy (MacDonald *et al.* 1994; Kamstra *et al.* 1999; Kwan*et al.* 2003), and DNA instability, which has been related to various diseases, including fragile-X syndrome (Bowater and Wells 2001; Fleming *et al.* 2003), colon cancer (Grady 2004), and aging (Lombard *et al.* 2005; Rodier *et al.* 2005). In addition, methods are currently being developed to apply recombination strategies to promote genetic therapy (see, for example, Urnov *et al.* 2005). The central intermediate in homologous recombination is the four-stranded DNA complex known as the Holliday junction. Thus, it is important to characterize the Holliday junction intermediate and how this DNA structure is recognized by recombinases, repair enzymes, and other junction binding proteins to fully understand the basic

mechanism of recombination. In this review, we will focus primarily on the detailed structure and structural determinants of the Holliday junction in free DNA, and speculate on how the DNA structure itself may be involved in how proteins recognize and bind to such junctions.

2.2 Structure of the DNA Holliday Junction

2.2.1 Early Models of junctions

The structure and conformation of the DNA Holliday junction (Fig. 2.1) have been of interest since it was first proposed by R. Holliday in 1964 (Holliday 1964). The basic structural features of DNA junctions in solution were elucidated in the 1980's by several groups using asymmetric sequence constructs that prevent migration and resolution of the junction off the ends (Kallenbach *et al.* 1983; Seeman and Kallenbach 1983; Seeman *et al.* 1985; Cooper and Hagerman 1987, 1989; Duckett *et al.* 1988; Lilley 1999, 2000). The general structure was found to be dependent on both the type and concentration of cations present in the solution, with the DNA junction adopting either a low salt extended-X form (Fig. 2.1b) or a high-salt compact stacked-X form of the junction (Fig. 2.1c, d) (reviewed in Lilley 1999, 2000). At low salt, the negatively charged phosphates remain largely unshielded and, thus, the arms are extended away from each other in an approximate 4-fold symmetric structure. At higher salt concentrations, condensation of cations around these phosphates allow formation of a more

compact structure in which the four arms pair and coaxially stack into two nearly continuous double-helices that are interrupted only by the crossing of strands. The model for the stacked-X junction relates the stacked duplexes by a positive (right-handed) rotation of ~60^{o} (Duckett *et al.* 1988). The strands of this latter stacked-X junction were proposed to be aligned antiparallel to each other, thereby forcing the two cross-over strands that link the stacked duplex arms to form a sharp U-turn. Notably, this antiparallel form of the junction would be topologically incapable of migrating along the DNA strands, while both the parallel stacked-X (as initially proposed by Holliday, Fig. 1a, Holliday 1964) and the extended open-X forms would be free to migrate along the duplex arms of homologous sequences.

How the arms of these asymmetric junctions pair defines different conformational isomer forms of the stacked-X junction. These conformational isomers are determined by the nucleotide sequences immediately around the junction cross-over. The interconversion between isomeric forms has been shown to be, again, cation dependent. Furthermore, recent single-molecule studies (McKinney *et al.* 2003) show that the interconversion between isomeric forms goes through the extended open-X structure. In homologous sequences, this conversion to the open-X form results in migration and subsequent resolution of the junction into discrete B-DNA duplexes (Lushnikov *et al.* 2003; McKinney *et al.* 2003).

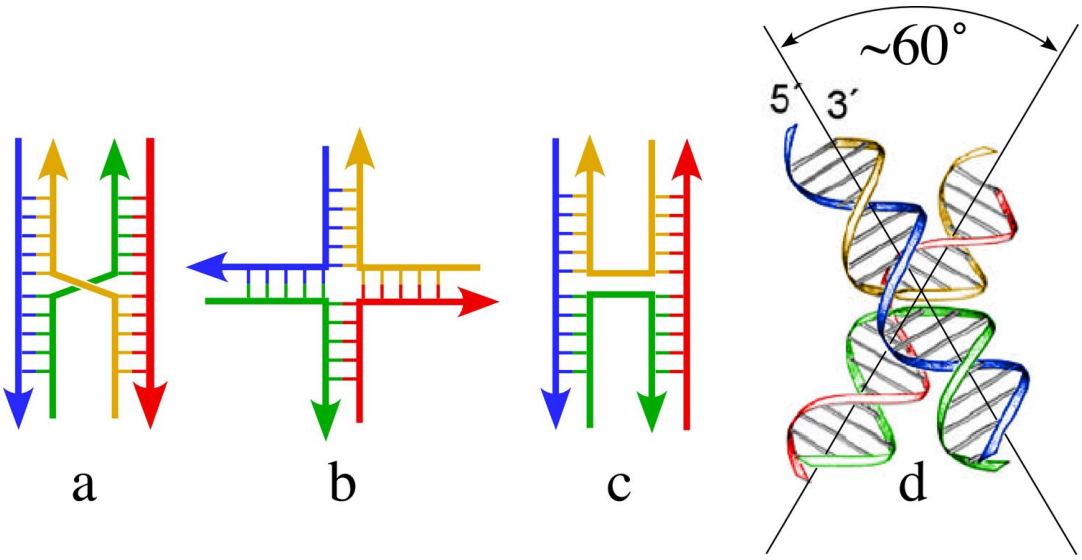
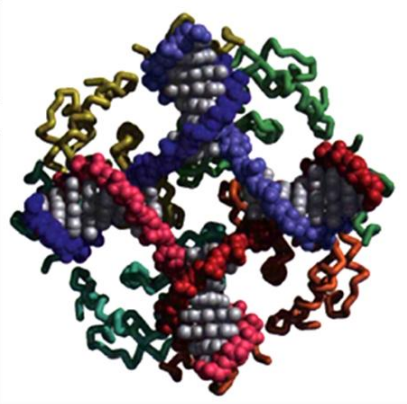


Figure 2.1. Structural forms of DNA Holliday junctions. a. The parallel stacked-X junction initially proposed by Holliday as the recombination intermediate (Holliday, 1964). b. The extended open-X form of a DNA junction. c. The antiparallel stacked-X junction does not allow for migration of the junction along the DNA strands. d. Model of the antiparallel stacked-X junction proposed from solution studies (Duckett *et al.* 1988).

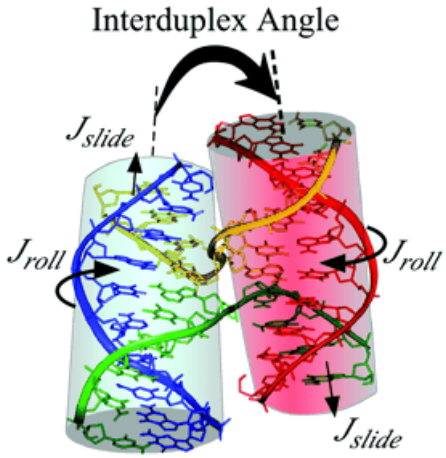
2.2.2 First single-crystal structures of junction

Despite this wealth of physical data in solution, obtaining the single-crystal structure of the Holliday junction had long been considered the 'Holy Grail' in DNA crystallography. The detailed molecular structure of the junction was first elucidated in the mid-1990's as complexes of DNA with various repair and recombination proteins, including the RuvA DNA repair protein (Hargreaves *et al.* 1998; Roe *et al.* 1998), RuvC (Bennett and West 1995a), Cre recombinase (Guo *et al.* 1997), andFlp recombinase (Chen *et al.* 2000). In all of these protein-bound structures, the DNA junction adopts some version of the extended open-X form, presumably to allow for migration of the junction cross-over along the DNA strands. The crystal structure of a four-way junction in the absence of protein was first seen in an RNA/DNAzyme complex (Nowakowski *et al.* 1999, 2000), while the structures of the DNA junction in its native state were finally solved nearly simultaneously by two different laboratories (Fig. 2.2) (Ortiz-Lombardı´a *et al.* 1999; Eichman *et al.* 2000). Both of these DNA junction structures were obtained serendipitously—the first was from a sequence that was designed to study the structure of adjacent G·A mismatched base pairs (Ortiz-Lombardı´a *et al.* 1999), while the second was intended to study the structure induced by interstrand thymine-thymine cross-links by the drug psoralen (Eichman *et al.* 2000, 2001). In the former structure, the sequence was an inverted repeat interrupted by G·A mismatches (5'-CCGG***GA***CCGG-3'), while in the latter case, the junction formed in a true inverted-repeat (IR) sequence with all standard Watson-Crick type base

a



b



c

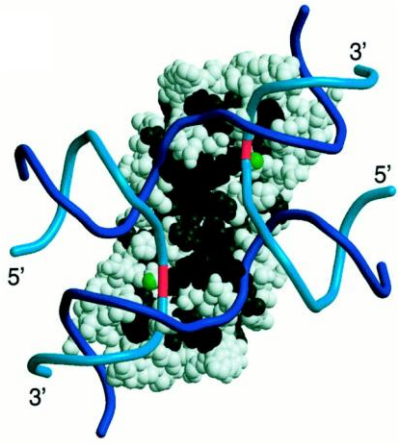


Figure 2.2. Single-crystal structures of Holliday junctions. a. The open-X junction in complex with the DNA repair protein RuvA (Hargreaves *et al.*, 1998). b. Antiparallel stacked-X junction in the sequence CCGGTACCGG (Eichman *et al.*, 2000). c. Model of the junction-resolving enzyme Hjc from Sulfolobus solfataricus bound to a stacked-X junction (Middleton *et al.*, 2004).

pairs (5'-CCGGT***ACC***GG-3'). The general features of both crystal structures were surprisingly similar to the molecular model proposed from solution work in 1988 (Duckett *et al.* 1988), with the junctions adopting the antiparallel stacked-X form, and the stacked duplexes related by a right-handed, albeit slightly less twisted (at ~40° rather than the 60° rotation relating the two pairs of stacked helical arms, Fig. 1) (reviewed in Ho and Eichman 2001; Hays *et al.* 2003b).

2.2.3 Effect of sequence and sequence-dependent interactions on formation and conformation

Comparison of the two junction forming DNA sequences to other similar sequences that had, to that point, been crystallized as standard B-DNA double-helices implicated the ACC trinucleotide at the $N_6N_7N_8$ nucleotide positions within the sequence motif CCnnn$N_6N_7N_8$GG as a common motif (a junction core) that stabilized the four-stranded structure in crystals (Eichman *et al.* 2000; Ho 2001; Hays *et al.* 2003b). This hypothesis was supported by the observation that these nucleotides are found at the cross-over of the junction and that the cytosine base at cytosine C_8 formed direct hydrogen bonds to the phosphate at the U-turn of the crossing strands. Interestingly, none of the divalent cations present in the crystallization solutions were identified in either crystal structure (Ortiz-Lombardı´a *et al.* 1999; Eichman *et al.* 2000). Thus, contrary to our expectations, there is a strong sequence-dependence for the formation and stabilization of DNA junctions in inverted-repeats, which had always been considered to be freely

migrating. This raises the question of how sequence, substituent groups, and cations affect the formation and conformation of Holliday junctions.

Crystal structures of Holliday junctions have now been determined from several IR sequences that contain the ACC trinucleotide motif, and show that the junction can form *(i)* with terminal C·G base pairs replaced by T·A (Thorpe *et al.* 2003), *(ii)* with the base at cytosine C_8 methylated (Vargason and Ho 2002), *(iii)* with the central thymine bases photocross-linked by psoralen (Eichman *et al.* 2001), *(iv)* with inosine 2-aminopurine (Hays *et al.* 2004) and brominated (Hays *et al.* 2003a) base analogues at the $N_6N_7N_8$ positions, and *(v)* in the presence of various divalent cations including magnesium (Eichman *et al.* 2000), calcium (Hays *et al.* 2003a), and strontium (Thorpe *et al.* 2003). The structures confirm that *(i)* the ACC core triplet is important for formation of junctions, *(ii)* the interaction between the N_4 amino at the major groove surface of cytosine C_8 with the cross-over phosphate is important not only for the formation of the junction but also in defining its conformational geometry, and *(iii)* divalent cations can be localized along the stacked DNA duplexes and that these cations can affect the local and global geometry of the junction. It should be noted that although the cytosine C_8 to phosphate hydrogen bond is seen as an intramolecular interaction in the junction, similar cytosine-phosphate hydrogen bonds were seen to provide sequence-dependent intermolecular interactions that 'locked' two B-DNA double-helices together (Timsit *et al.* 1989). Finally, an interesting variation on this interaction is that this aminophosphate hydrogen bond can be replaced, to some

degree, by a halogen bond, an underappreciated interaction between a polarizable halogen (in this case a bromine of a 5-bromouracil) and a Lewis base (oxygen, nitrogen, sulfur,etc.) (Auffinger *et al.* 2004).

We have recently applied a crystallographic screen of the general IR sequence CCnnn$N_6N_7N_8$GG (where $N_6N_7N_8$ is a combination of any of the four standard nucleotides, and nnn are trinucleotides that maintain the overall inverted repeat pattern in the sequence) to search for junction forming sequences that do not contain the ACC-motif. At this point, 63 of the 64 possible sequence combinations have been crystallized and the structures of 29 of these have been determined (Hays *et al.* 2005). The screen to date has identified a set of sequences that relate the formation of junctions to B-DNA duplexes, junctions to A-DNA, and A-DNA to B-DNA (Fig. 2.3). Among the structures that resulted from the screen are three new junction-forming sequences, all of which, in contrast to the ACC-core sequence, are identified as amphimorphic (capable of adopting both junctions and double-helical structures). The sequence GCC (these sequences are referred to according to their unique $N_6N_7N_8$ trinucleotides) was crystallized as a junction from Ca^{2+} solutions (Aymami *et al.* 2002; Hays *et al.* 2003a), but as B- DNA duplexes from Mg^{2+} solutions (Heinemann *et al.* 1992). ATC, however, forms both a junction and B-DNA in Ca^{2+}, with high salt favoring the stacked-X junction. We propose that, in this case, the lower cation concentration allows for unstacking of the junction into the open-X form, which subsequently allows for migration and consequently the resolution of the junction into individual B-DNA

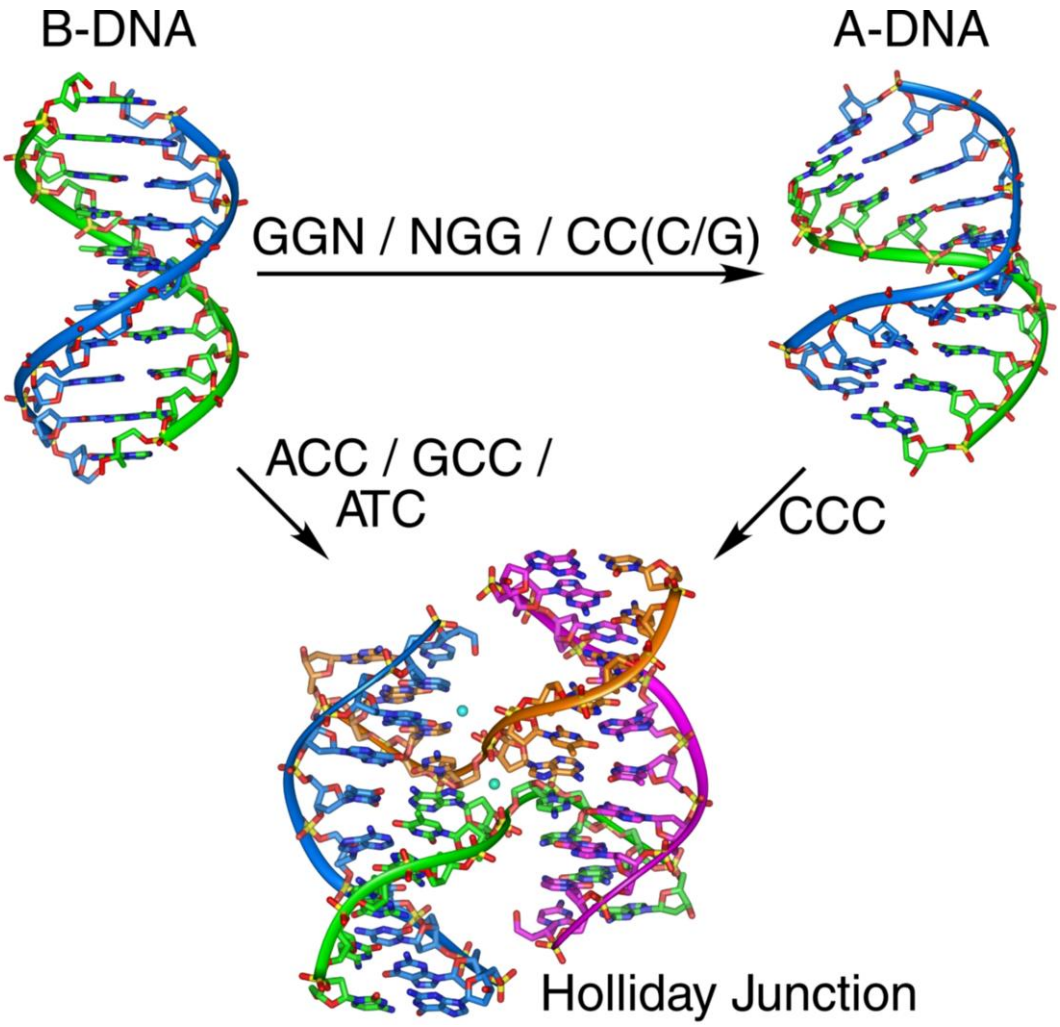


Figure 2.3. Sequence effects on B-DNA, A-DNA and Holliday junctions. The sequence dependent stabilization of DNA structures has been determined using a crystallographic screen of the decanucleotide sequence CCnnn $N_6N_7N_8$GG, where each position in $N_6N_7N_8$ is allowed to be any of the four standard nucleotides, and the trinucleotide nnn is specified to maintain the inverted repeat symmetry of the motif (Hays *et al.*, 2005). The $N_6N_7N_8$ trinucleotides that lead to formation of junctions from B-DNA,A-DNA from B-DNA, and junctions from A-DNA are indicated.

duplexes. Finally, CCC was seen in Ca^{2+} solutions to form a junction at high salt and the altered A-DNA duplex at lower salt (Hays *et al.* 2005). With these amphimorphic sequences spanning the junction and duplex forms of the DNA, Holliday junctions could be related to B-DNA and A-DNA through a structural map constructed from 32 unique single-crystal structures from 29 different sequences (Hays *et al.* 2005). The structures from this map show *(i)* that the C_8 to phosphate interaction is essential but not sufficient for formation of the junction, *(ii)* that the C_7 position is favored by pyrimidines (C>T) because of electrostatic interactions from either the N_4 amino of the cytosine or C_5-methyl of the thymine base to a phosphate of an opposing arm, and *(iii)* that the N_6 position shows the preference A>G>C, although the molecular rationale for the series has yet to be established (Fig. 2.4). In addition, the structural map supports the model that A-DNA is favored explicitly by GGN, NGG, and CC(C/G) trinucleotides within this sequence context, as expected. Thus, this structural map allows four-stranded junctions to be related to both B-DNA and A-DNA explicitly through sequence context. Finally, this sequence motif was seen to adopt structures and conformations relatively independent of crystal lattice and crystallization solutions effects, as evident from the broad range of crystal forms observed in the screen and the relatively limited crystallization solutions required to obtain this large variety of structural and crystal forms. We anticipate that additional junction forming sequences will be identified with the completion of the screen.

With the increasing number of crystal structures of DNA junctions, it became necessary to develop a set of definitions to accurately describe geometries of the four-stranded complexes relative to a set of defined planes (Vargason and Ho 2002; Watson *et al.* 2004) in order to quantitatively analyze and compare the effects of various factors, including sequence, salt and drugs on their conformations. An analysis of the currently available DNA structures shows that stacked-X type junctions in the crystal exhibit a more shallow J_{twist} (the angle relating the stacked duplex arms across the junction) as compared to the model derived from solution studies. In addition, the arms can be translated along their helix axes (characterized as J_{slide}) and rotated about their helical axes (measured by J_{roll}) to either bury or expose the major groove surfaces of the junction. These geometric perturbations are associated with the effect that sequence, salt, and substituents have on the intramolecular interactions at the junction core (Vargason and Ho 2002; Hays *et al.* 2003a; Watson *et al.* 2004).

The obvious question is whether the properties of junctions seen in crystals have any relationship to the structure in solution. For example, all of the symmetric junctions in IR sequences have a more shallow twist angle relating the stacked duplex arms (40^o-45^o) compared to that determined for asymmetric junctions in non-IR sequence constructs (~60^o). Studies using atomic force microscopy and hydroxyl radical foot-printing to probe the geometry of symmetric junctions that contain the ACC-core (Sha *et al.* 2002), however, show that this angle is ~40^o and that the junction crosses between duplexes exactly between the

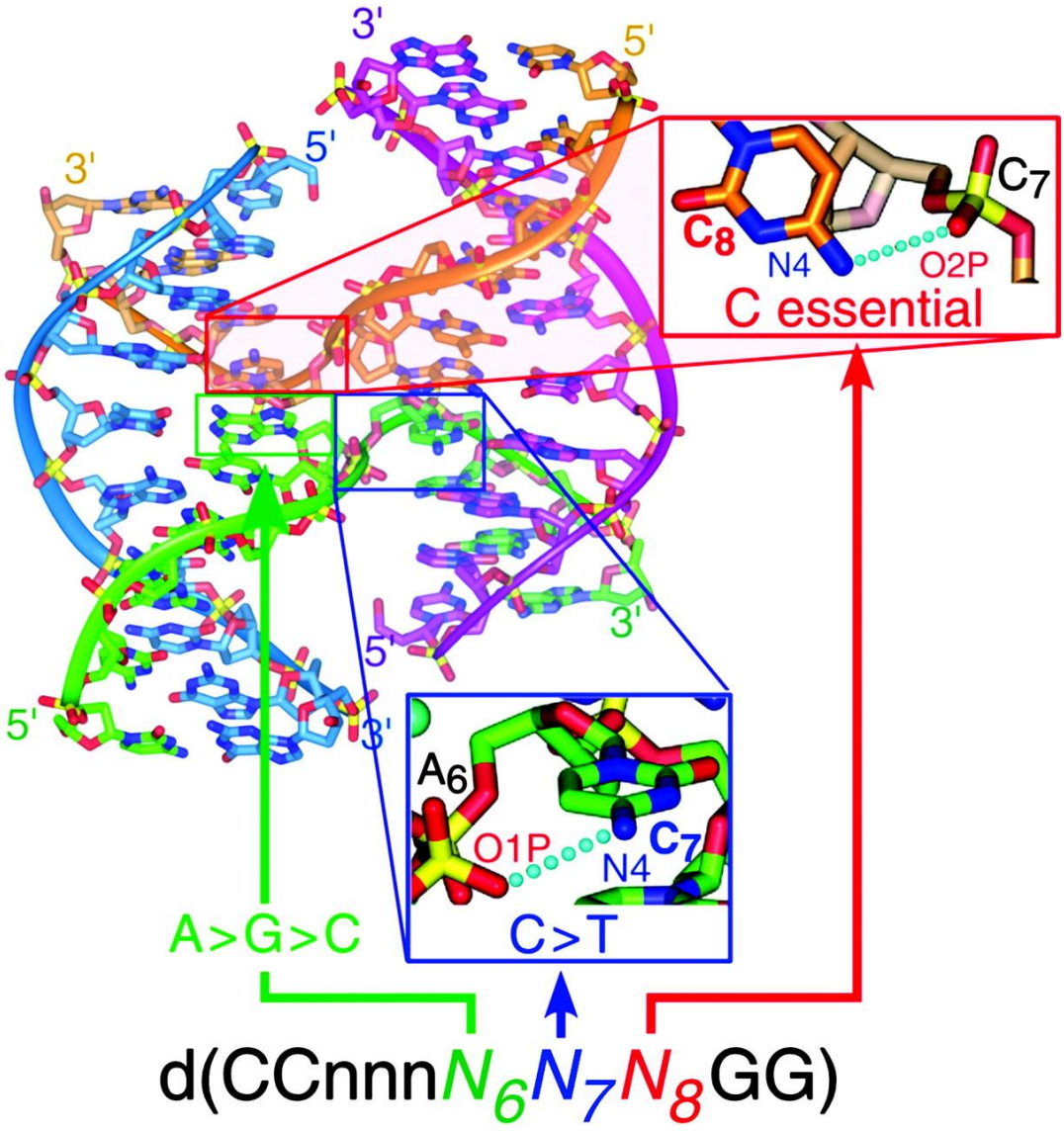


Figure 2.4. Stabilizing interactions in the Holliday junction. The formation of four-stranded DNA junction in the inverted-repeat sequence type d(CCnnn$N_6N_7N_8$GG) is dependent on the sequence and sequence dependent interactions at the $N_6N_7N_8$ trinucleotide. A hydrogen bond from the N_4 amino of a cytosine at N_8 to the phosphate at the junction cross-over is seen to be essential, but not sufficient, to specify formation of the junction. Cytosine is favored over thymine at the N_7 position (when N_6= A and N_8= C) because the electrostatic interaction from the phosphate oxygens of N_6 to the N_4 amino nitrogen of the cytosine base is stronger (distances range from 3.1 to 3.6A°) as compared to the C5 methyl group of the thymine base (distances range from 4.2 to 4.5A°). The preference is A>G>C (with N7 and N_8= C) for the nucleotide at N_6.

A and C nucleotides of the core triplet, as seen in the crystal structures (Hays *et al.* 2003b). These results indicate that specific interactions identified within the ACC core are responsible for specifying the geometry of the DNA junction even outside the environment of the crystal lattice.

Does this ACC core, however, help to stabilize the junction in solution? To address this question, we recently studied the parent (5'-CCGGT***ACC***GG-3')$_4$ junction by analytical ultracentrifugation and determined the dissociation constant for the junction to duplex equilibrium of 100–200 μM (Hays *et al.* 2006). Analytical ultracentrifugation studies showed that the similar sequence 5'-CCGCT***AGC***GG-3' (which does not crystallize as a junction, but as B-DNA duplexes) exists only as double-helices in solution. Thus, the ACC core is seen to contribute ~5 kcal/mol of stabilization to the tetrameric junction. Moreover, the junction is dissociated at low Ca^{2+} concentrations even at high DNA concentrations, consistent with the general understanding that the stacked-X junction is stabilized by high concentrations of divalent cations (Duckett *et al.* 1990). It is clear, therefore, that the stacked-X DNA junction in solution is well described by the single-crystal structures, including their sequence dependent formation and the intramolecular interactions associated with their formation. One must ask, however, whether the compact stacked-X structure seen in isolated DNA constructs is at all relevant to the biological mechanisms of recombination where the DNA does not exist in isolation, but in the context of protein complexes.

2.3 Junction Binding Proteins

There are currently numerous junction binding proteins known, consistent with the variety of cellular mechanisms associated with homologous recombination (Aravind *et al.* 2000; Lilley and White 2001); however, only a handful have been characterized in complex with their DNA substrates (Sharples 2001). Although the DNA junctions seen in crystals structures of all current protein-DNA complexes are in the open-X form, a large number of junction binding proteins show high affinity for the stacked-X junction, including the BLM protein associated with Bloom's syndrome (Karow *et al.* 2000). The general forms of the DNA in complexes with several dimeric resolvases, enzymes that make symmetric cuts at the point of strand exchange in four-stranded junctions, (White and Lilley 2001), including T7 endonuclease I (Declais *et al.* 2003) and Hjc resolvase (Fig. 2.2) (Fogg *et al.* 2001; Middleton *et al.* 2004), have been characterized biochemically by gel electrophoresis and fluorescence resonance energy transfer (reviewed in Lilley 2000). The dimeric junction resolving enzyme RusA has been shown to bind stacked-X junctions and specifically cut homologous sequences at the phosphodiester bond 5' of CC dinucleotides (Chan *et al.* 1997; Giraud-Panis and Lilley 1998). Although, as with many resolvases, RusA distorts the junction upon binding, this sequence specificity, along with the observations that NCC trinucleotides favor junction formation in the absence of proteins both in

crystals and in solution, suggests that the sequence context for junction formation may play a role in recognition.

Classifying the junction binding proteins according to whether they recognize the open-X or stacked-X form of the junction (those for which the structure of the DNA can be experimentally assigned, Table 2.1) provides some interesting insights into their structural specificity for the DNA substrate. Enzymes that recognize the two-fold symmetric stacked-X junction are all dimeric, while nearly all tetrameric proteins bind to the approximate four-fold symmetric open-X structure. Moreover, proteins that recognize open-X junctions also have some degree of sequence specificity (either a specific DNA sequence or damaged base pairs), while those that recognize stacked-X junctions are relatively non-specific at the sequence level (the lone exception in Table 2.1 is topoisomerase I from vaccinia, Liao *et al.* 2004). It seems reasonable, therefore, to think of the tetrameric proteins as a group that binds to the open-X form to allow for migration of the junction and provide a means for the protein to seek-out its target sequence.

One group of dimeric proteins that initially appears to violate these general trends includes RuvC (Bennett and West 1995b; Fogg *et al.* 2001), and the resolvases Cce1 from *S.cerevisiae* (White and Lilley 1997) and Ydc2 from *S. pombe* (White and Lilley 1998): these are all homodimers, but the DNA substrates in the complexes are seen to adopt the open-X structure. A more detailed analysis of this group suggests that these proteins are not so much exceptions to the

Table 2.1 Holliday Junction Binding Enzymes

Enzyme	Organism	Oligomeric State	Sequence Specificity[2]	References
Open-X Type DNA Junction Substrate				
Cre1	Bacteriophage P1	Tetramer	LoxP sequence	(Guo *et al.*, 1997)
RuvAB[1]	*E. coli*	Tetramer	Damaged DNA	(Hargreaves *et al.*, 1998; Roe *et al.*, 1998)
Flp[1]	*S.cerevisiae*	Tetramer	Flp Recombination target (FRT)[3]	(Chen *et al.* 2000)
λ Integrase[1]	Bacteriophage λ	Tetramer	TNNNTTNNTNN NANNAANNNG	(Biswas *et al.* 2005)
RecU	*B. subtilis*	Dimer	(G/t)G↓C(A/C)	(McGregor *et al.* 2005)
Induced Open-X Junction Substrate[4]				
RuvC[1]	*E. coli*	Dimer	(A/T)TT(G/C)	(Bennett and West, 1995b; Fogg *et al.* 2001)
Cce1	*S. cerevisiae*	Dimer	ACTA	(White and Lilley, 1997)
Ydc2	*S. pombe*	Dimer	CT and/or TT	(White and Lilley, 1998)
Stacked-X Type DNA Junction Substrate				
T4 nuclease VII	Bacteriophage T4	Dimer	None	(White and Lilley, 1997; Raaijmakers *et al.*, 1999)
T7 Endonuclease I	Bacteriophage T7	Dimer	None, (C/T)↓(C/T)	(Declais *et al.* 2003)
Hjc	*P. furiosu*	Dimer	None	(Middleton *et al.* 2004)
Hjc	*S. solfataricus*	Dimer	None	(Fogg *et al.* 2001)
Hje	*S. solfataricus*	Dimer	None	(Middleton *et al.* 2004)
Vtopo I	*Vaccinia*	**Dimer**	CCCTT↓N	(Liao *et al.* 2004)
Tetrahedral DNA Junction Substrate				
RusA	*E. coli*	**Dimer**	↓CC	(Chan *et al.*, 1997; Giraud-Panis, 1998)

[1]Crystallized DNA-complex.

[2]Binding or cutting sites (cut site specificity indicated by vertical arrow).

[3]Natural Flp Recombination target is an A/T-rich 48 bp sequence (Chen *et al*. 2000).

[4]Binds stacked-X, but induces an open-X structure in complex.

Enzymes are categorized according to conformation of the DNA substrate (open-X, stacked-X, or tetrahedral forms) as determined from crystal structures with junctions1, or inferred by biochemical data and/or molecular modeling (references are to studies that define the form of the DNA substrate).

general trends described above, but serve to bridge the dimeric stacked-X binding proteins with the tetrameric open-X binding proteins. These RuvC-related proteins are thought to initially recognize and bind to stacked-X junctions, but then to induce the DNA to adopt the more extended open-X form (Fogg *et al.* 2001). The induced structural perturbations to the DNA are associated with the dimerization of the protein—the monomers do not induce the open-X structure. Thus, the dimeric proteins all induce some structural perturbation to the stacked-X structure (from minor opening of the junction center and rotations of the stacked arms, to more dramatic changes to the open-X or even a possible tetrahedral form) presumably to allow the enzymes to gain access to the scissile bond or to stabilize a transition state (Sharples 2001).We suggest here that the protein-induced open-X structure may also allow the junction to migrate and the protein to seek-out its specific recognition site, even if it is the immobile stacked-X form that is initially recognized.

What then is the role of the stacked-X junction? We propose here a model in which the sequence dependent formation of this compact structure provides dimeric proteins, including nonspecific resolvases, with some degree of sequence specificity through an 'indirect-readout' mechanism (Dickerson 1983; Otwinowski *et al.* 1988; Olson *et al.* 1998; Lu *et al.* 2000; Arauzo-Bravo *et al.* 2005), as opposed to direct recognition of base pair identity (Fig. 2.5). In this model, inverted-repeat sequences that incorporate the ACC and, to lesser extents, the amphimorphic GCC, ATC and CCC trinucleotides help pause or fix junctions at specific sites along

a genome. The evidence for sequence specific pausing during junction migration was first seen in immobilized symmetric junctions by Seeman's group (Sun *et al.* 1998). It is then this stabilized junction that serves as the substrate for protein binding. In a classic example of an induced-fit model for enzymes, the protein subsequently induces distortions away from the intrinsic structure of the DNA junction in the course of its function. This perturbation can maintain the general topology and symmetry of the stacked-X junction, as with Hjc, or may induce the open-X structure if there is a need for the junction to migrate as the protein searches for a target sequence, as in the cases of Cce1 and Ydc2. Thus, in this model, sequence specificity is not conferred at the initial point of recognition and binding, but through indirect sequence dependent stabilization of the stacked-X junction rather than by direct read-out of the base pairs that define the DNA sequence.

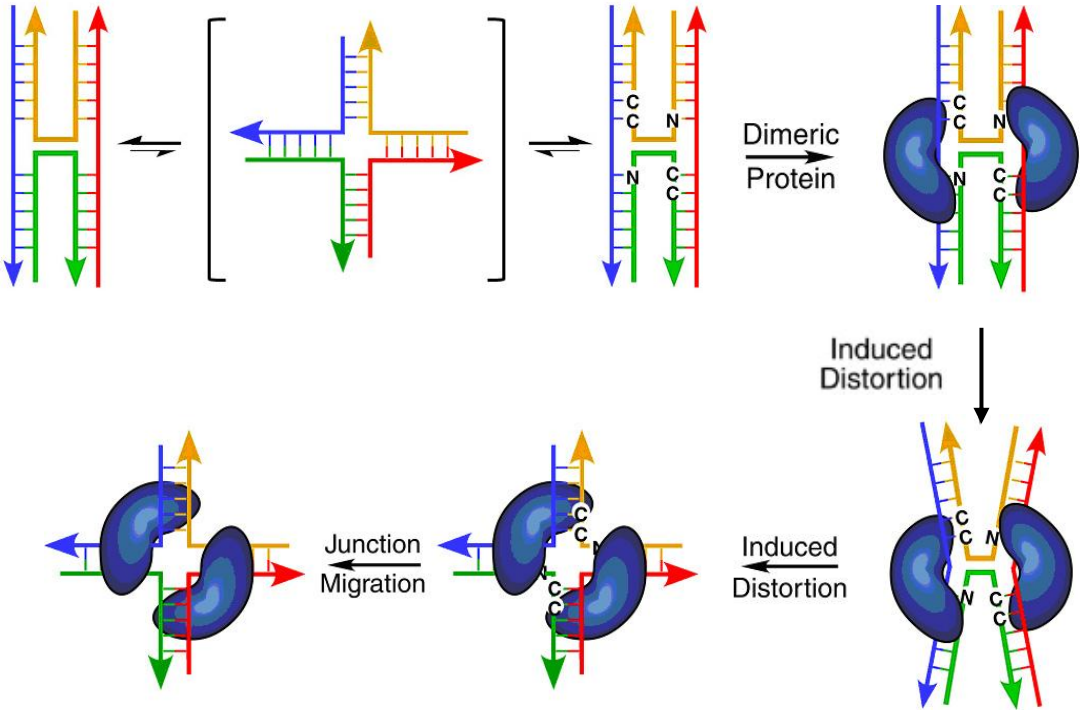


Figure 2.5. Proposed model for indirect sequence recognition of stacked-X junctions. In this model, migration of the Holliday junction in free DNA requires a transition to the open-X form. This is consistent with models proposed from single-molecule studies on junction isomerization (McKinney *et al.*, 2003) and translocation (Lushnikov *et al.*, 2003). However, certain sequences such as the ACC trinucleotide in an inverted repeat help to stabilize and stall the junction, thereby presenting a defined structure for recognition by a dimeric junction binding protein. In the complex, the protein induces a structural perturbation that either maintains the topology and symmetry of the stacked-X junction, or induces an open-X junction that can then migrate to a protein specific target site.

2.4 Summary and Perspectives

We now have a very detailed understanding of the DNA Holliday junction alone in the compact stacked-X form, at least within the context of inverted-repeat DNA sequences. Specific intramolecular interactions are seen to direct the formation and conformation of the DNA junction both in solution and in crystals. We suggest, therefore, that sequences that favor junction formation may provide a stable substrate for recognition and binding by proteins involved in recombination and DNA integration processes. This seems to be particularly important for dimeric enzymes that typically are not highly specific for a particular DNA sequence. Thus, specificity may be conferred by the ability of certain DNA sequences, particularly in inverted repeat sequences, to fix the junction and, thereby, indirectly confer sequence specificity through structure specificity. We recognize that the model proposed here is based on a small subset of known junction binding proteins, and the two general observations on which the model is based may need to be revised as the structures of the DNA substrates that are recognized and bound become characterized for additional proteins. Where do we go from here in terms of the structure of the Holliday junction? For the DNA itself, we are challenged to determine the structures of the stacked-X junction in nonsymmetric sequences to help bridge the conceptual gap between junctions in crystals and the wealth of information on their behavior in solution (as reviewed by Lilley 2000). In addition, it would be informative to determine the structure of

the open-X junction in the absence of protein (in order to understand how, by comparison, the protein affects this form) and perhaps other possible junction forms, including a potential tetrahedral four-stranded junction and three-way junctions. For junction binding proteins, the challenge has been to determine the structure of a resolvase in complex with a stacked-X substrate, particularly one that does not bind to or induce an open-X junction. Together, structural studies on junction binding proteins and their DNA substrates will provide us with an understanding for how sequence directed conformations contribute to 'indirect read-out' of genomic information, particularly for the ever growing class of biological functions that rely on the mechanism of genetic exchange first elucidated by R. Holliday over 40 years ago.

2.5 Acknowledgments

Work in the lab of PSH was funded by grants from the National Institutes of Health (R1GM62957) and the National Science Foundation (MCB0090615). PSH would like to thank the Franco-American Fulbright Commission and Prof. Eric Westhof at the CNRS at the Universite´ Louis Pasteur in Strasbourg, France for support and help in the writing of this manuscript.

Chapter 3

A wobble A·T base pair in the structure of an asymmetric Holliday junction: Evidence for a rare nucleotide base tautomer in a biological context

Patricia Khuu and P. Shing Ho

3.1 Summary

An asymmetric, locked DNA Holliday junction, comprised of four different sequences, in a crystal structure of 1.9 Å resolution exhibits a wobbled A·T base pair. Analysis of the base pair geometry suggests for rare tautomerization of one of the participating bases. As no wobbled base pair has been observed in other junction structures crystallized under comparable conditions, the tautomerization cannot be an artifact of crystallization conditions. The observation provides unique and more physiologically relevant evidence for a rare nucleotide base tautomerization, with profound biological and chemical implications.

3.2 Introduction

A four-stranded DNA complex was first proposed in 1964 by Robin Holliday to be the central intermediate in homologous recombination (Holliday 1964). The structure appears identical to two B-DNA duplexes, except they are joined by two strands that exchange across the duplexes to form what is now known as the Holliday junction. If the duplexes are homologous in sequence, the nucleotides form standard Watson-Crick base pairs, allowing for the isoenergetic migration of the cross-over along the DNA strands. The Holliday junction has now been

implicated in a variety of pathways that involve homologous recombination mediated mechanisms, including DNA repair and replication, resumption of stalled replication forks and viral genome integration (Nunes-Duby *et al.* 1987; Cox *et al.* 2000; Haber and Heyer 2001; Dickman *et al.* 2002; Declais *et al.* 2003; Subramaniam *et al.* 2003) and, consequently, the atomic details of its structure has been of great interest for the past four decades.

The first detailed structural models for the junction were derived from biochemical and biophysical studies of junctions assembled from four unique strands of DNA into a single junction complex. In these constructs, the junction is locked, in which migration is prohibited by the asymmetry of the sequences and would generate unfavorable base pair mismatches (Carlstrom and Chazin 1996; Miick, Fee *et al.* 1997; Grainger, Murchie *et al.* 1998; Sha, Liu *et al.* 2002; Cooper and Hagerman 1987; Duckett, Murchie *et al.* 1990; Clegg, Murchie *et al.* 1992; Clegg, Murchie *et al.* 1994). The resulting models include two conformations: an extended, “open-X” form and a compact, “stacked-X” form. The open-X structure, with a pseudo four-fold symmetry, exists under low salt conditions, having four duplex arms splayed out and away from the central crossover region, while the stacked-X structure is observed at high salt, with paired duplex arms stacking to form nearly continuous duplexes that are interrupted only by the crossing of the junctions. The stacked-X structure differs from Holliday’s original model in that the duplexes are not parallel with overlaying crossover strands. Rather, the sharp U-turns of crossing strands result in an antiparallel orientation of the duplexes.

The first crystal structures of DNA Holliday junctions were determined nearly a decade ago by two different laboratories. In both studies, the constructs were not of four different DNA strands, but assembled from the inverted repeat sequence d(CCGGTACCGG) (Eichman *et al.* 2000) or near inverted repeat sequence d(CCGGGACCGG) (Ortiz-Lombardia *et al.* 1999). In the former case, all base pairs form standard Watson-Crick base pairs, while in the latter, the central GA nucleotides form tandem G·A mismatches. Although these junctions are symmetric in that the sequences across the junctions are identical or near identical, they recapitulated the general features of the stacked-X model for asymmetric junctions derived in solution, including the antiparallel orientation of the crossing strands, and the ~40°-60° rotation of the stacked arms relative to each other. The symmetric junctions of crystallographic studies are locked not by the sequence complementarity, but by sequence specific hydrogen bonds that span across the tight U-turn of the crossing strands (Hays, *et al.* 2005). Still, there remain questions of whether the atomic details of the stacked-X junction constructed from four asymmetric sequences would conform to those of the symmetric junction structures.

To bridge the solution studies of asymmetric junctions with the crystallographic studies of symmetric junctions, we present the single crystal structure of an asymmetric Holliday junction construct assembled from four unique DNA strands. In general, the observed structure is nearly identical to the structures of the current symmetric constructs, with one major exception—an A·T

wobble base pair is observed in one arm of the junction, suggestive of the formation of rare nucleotide base tautomer (either an imino-adenine or the enol-thymine form). This is the first observation of a stable rare tautomer in a biologically relevant context, and lends credence to the rare tautomer hypothesis as a mechanism to effect spontaneous base substitution mutations proposed by James Watson and Francis Crick in their seminal 1953 paper (Watson and Crick 1953).

3.3 Materials and Methods

3.3.1 Crystallization, x-ray data collection and structure refinement.

DNA oligonucleotides were synthesized by Midland Oligos with the trityl-protecting group attached and subsequently purified by HPLC followed by size exclusion chromatography on a Sephadex G-25 column after detritylation with 3% acetic acid. The DNA was crystallized by the sitting-drop vapor diffusion method from solutions of 0.8 mM DNA and 25 mM sodium cacodylate (pH 7.0) buffer with 100 mM calcium chloride, and equilibrated against a reservoir of 35% aqueous 2-methyl-2,4-dimethylpentanediol. The first phase of growth involved formation of many “crystalline balls” in a droplet, with each “ball” consisting of highly dense, very fine crystal slivers growing from a central point. Expedited by agitation of the setup, one to a few of these splinter-like slivers, depending on conditions, grew to

become very stable, large diamond-shaped crystals with concurrent diminishing of the surrounding crystalline balls.

X-ray diffraction data were collected from one of the single crystals at liquid nitrogen temperatures using CuKα radiation from a Rigaku (Tokyo, Japan) RU-H3R rotating anode generator with a Raxis-IV image plate detector. Diffraction images were processed and reflections merged and scaled using DENZO and SCALEPACK from the HKL suite of programs (Otwinowski 1997).

The crystal diffracted to 1.9 Å and was indexed in the *P1* space group, with the volume of the unit cell consistent with the four unique strands of one Holliday junction defining the asymmetric unit of the crystal. The *P1* unit cell volume is twice that of previous structures of symmetric junctions crystallized in the *C2* space group in which the asymmetric unit was composed of only two strands of the junction (one outer and one crossover), with the complete four-stranded complex generated by the crystallographic two-fold symmetry. This suggested that the asymmetric unit of the crystal was very similar to the stacked-X junction seen in previous crystals.

The structure was solved by molecular replacement using EPMR (Kissinger *et al.* 1999) with a complete four-stranded assembly constructed from the previous symmetric junction d(CCGGTACCGG) as the starting model. A solution with correlation coefficients of 79.8% and R_{cryst} of 39.8% was obtained with a complete four-stranded junction in the asymmetric unit. Subsequent refinement of the initial model in CNS (Brunger, Adams *et al.* 1998) through rigid body

refinement, simulated annealing, positional refinement, and addition of solvent molecules resulted in a final model with R_{cryst} of 22.43% and an R_{free} of 26.48% (Table 3.1).

3.3.2 Structure analysis

The DNA parameters of the refined structure were calculated with the programs CURVES (Lavery 1988) and 3DNA (Lu 2003).

3.4 Results

We had designed a DNA construct of four unique strands (Fig. 3.1A) that would anneal to generate a sequence-locked Holliday junction, typical of constructs used in biochemical and biophysical studies, such as those of recent single-molecule studies (McKinney, Declais *et al.* 2003; McKinney, Freeman *et al.* 2005). The design of the construct took into account the general crystal lattice of the symmetric junctions that had previously been successfully crystallized from inverted repeat sequences, allowing for T·A over A·T and C·G over G·C base pair stacking. In addition, the junction structure was designed so that the junction crossover would interrupt the coaxially stacked duplexes to result in a 4 base pairs over 6 base pairs stacking of the arms. Thus, the construct helps to bridge the biochemical studies of asymmetric junctions with the atomic details of the single-crystal structures of symmetric junctions. What we had not anticipated was that a

Table 3.1. Data collection and refinement statistics

Space group	*P1*
Unit cell constants (Å)	a = 22.344 b = 34.554 c = 37.631
	α = 109.921 β = 90.028 γ = 109.016
Wavelength (Å)	1.542
Resolution (Å)	50-1.90
Total (unique) no. of reflections	6797 (3048)
R_{sym} [a]	6.8 (28.9)
Completeness (%)[a]	89.9 (62.9)
<I/σI> [a]	15.7 (2.8)
Refinement	
R_{cryst}	22.43
R_{free} [b]	26.48
no. of DNA atoms	809
no. of solvent atoms	120

[a] Value in parentheses refer to the highest resolution shell.

[b] R_{free} is R_{cryst} for 10% of the reflections not used in refinement.

non-Watson-Crick wobble base pair would form, suggesting the occurrence of a rare tautomer in this DNA context.

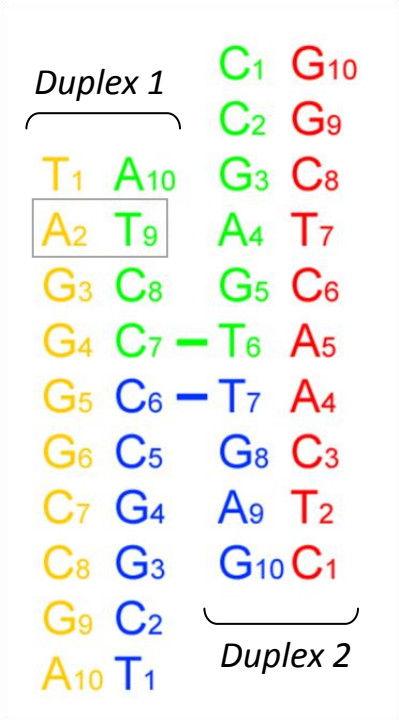
3.4.1 Junction structure

The crystal structure of the asymmetric junction (Fig. 3.1b) is in the antiparallel, stacked-X form, similar to that of previous structures of symmetric junctions. Each strand of the junction can be uniquely identified in the asymmetric unit, indicating that there is no averaging of the structure around the pseudo-two-fold axis that runs through the center of the junction—this is consistent with the *P1* space group of the crystals. The junction crossovers occur exactly as was designed in the construct, with the tight U-turns occurring at the phosphates that link the sixth and seventh nucleotides of the crossing strands.

The overall geometry of stacked-X junctions can be characterized by the relative orientations and positions of the coaxially stacked duplex arms, as defined by the angle relating the helical axes of the duplex arms (J_{twist}), the relative sliding of these helical axes (J_{slide}), and the angle that the centers of these duplexes form relative to the center of the junction (J_{roll}). The geometric parameters for this asymmetric junction when compared to those of symmetric junctions shows that the asymmetric junction is analogous to the symmetric junctions, though having a greater J_{twist} and J_{slide} (Table 3.2).

In the symmetric junctions, the geometries are linked by intrastrand molecular interactions that help to stabilize the U-turn and, consequently, lock the

A.



B.

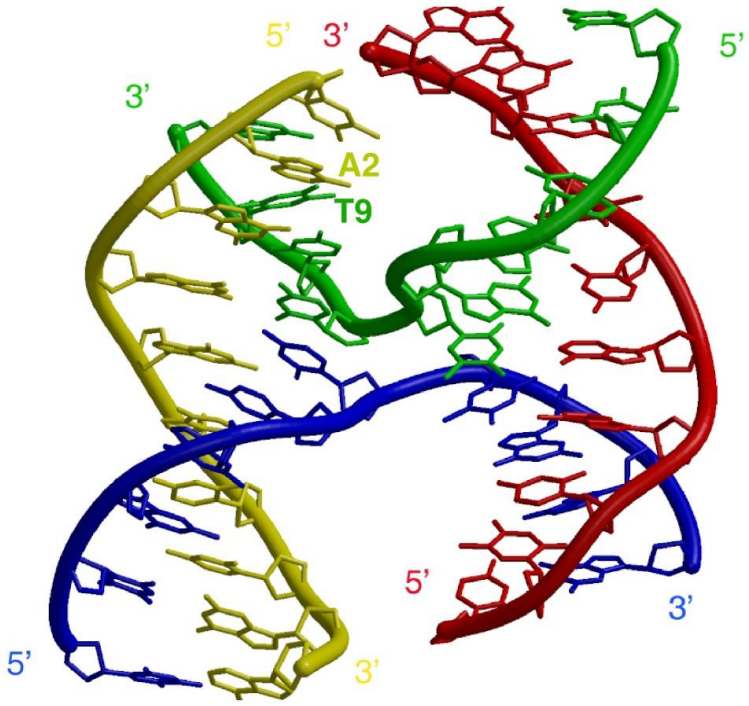


Figure 3.1. (A) Sequence assignment of sequence-locked asymmetric Holliday junction with position of wobble A·T base pair of Duplex 1 denoted *(box).* (B) Crystal structure of the asymmetric Holliday junction.

Table 3.2. Junction parameters

Structure	Sequence	NDB Accession Number	IDA (º)	J_{twist} (º)	J_{slide} (Å)	J_{roll} (º)
Asymmetric Junction	d(CCGAGTCCTA) d(CTCAACTCGG) d(TAGGGGCCGA) d(TCGGCCTGAG)		70.4	56.5	.84	134.5
Symmetric Junction						
ACC-4Na	d(CCGGTACCGG)	UD0008	65.9	37.8	.15	139.9
ACC-2Na	d(CCGGTACCGG)	UD0015	68.3	39.5	-.36	145.3
tACC-4Ca	d(TCGGTACCGA)	UD0018	66.6	39.3	.34	142.1
tACC-2Sr1	d(TCGGTACCGA)	UD0021	70.2	44.5	.57	150.1
tACC-2Ca	d(TCGGTACCGA)	UD0023	67.4	40.6	1.06	146.6
ACC-2Ca1	d(CCGGTACCGG)	UD0024	67.2	38.5	.15	142.2
ACC-2Ca2	d(CCGGTACCGG)	UD0025	66.7	37.7	.11	142.7
tACC-2Sr2	d(TCGGTACCGA)	UD0026	70.7	45.5	.46	150.3
Average			*67.9*	*40.4*	*0.31*	*144.9*
Standard Deviation			*1.6*	*2.8*	*0.4*	*3.6*

The geometric parameters J_{twist}, J_{slide}, J_{roll} and IDA have been calculated for the single-crystal DNA-only Holliday junction structures listed with global helix axes from CURVES.

junction in place. These include an essential hydrogen bond from the N4 amino group of the cytosine base at the C_8 nucleotide to the oxygen of the phosphate at sixth nucleotide, and a weaker electrostatic interaction from the seventh nucleotide , which could be a hydrogen bond from a cytosine (Eichman, Vargason *et al.* 2000; Hays, Vargason *et al.* 2003), a charge interaction from the methyl group of a thymine, or a halogen bond from a halogenated base (Voth, Hays *et al.* 2007) , to the oxygen of the phosphate at the sixth nucleotide position along the crossing strand. Of these core intrastrand interactions, only one is observed in the asymmetric junction: an electrostatic interaction between methyl group of T7 and the phosphate oxygen of C6 of a crossover strand (Fig. 3.1 and 3.2; *blue*). No intrastrand interaction is detected between atoms the T6 and C7 (*green*) of the other crossover strand. Moreover, a water molecule is observed at the central cavity of the crossover formed by the phosphates of the central C6 (*blue)* and T6 (*green)*. A water mediated interaction between the oxygen (O5) of the terminal C1 (*red*) and phosphate oxygen of G9 (*yellow*) is also detected.

As the asymmetric junction crystallized in a stacked-X isoform that precludes a purine at the sixth position of either crossover strand, the importance of a purine at this core position, involved in two of three stabilizing interactions in symmetric junctions, is not observed in this asymmetric junction. Rather, intrastrand interaction between two core pyrimidines, the methyl group of a thymine and phosphate oxygen of a cytosine, compensates in one crossover strand. The other strand does not exhibit any apparent stabilizing intrastrand core

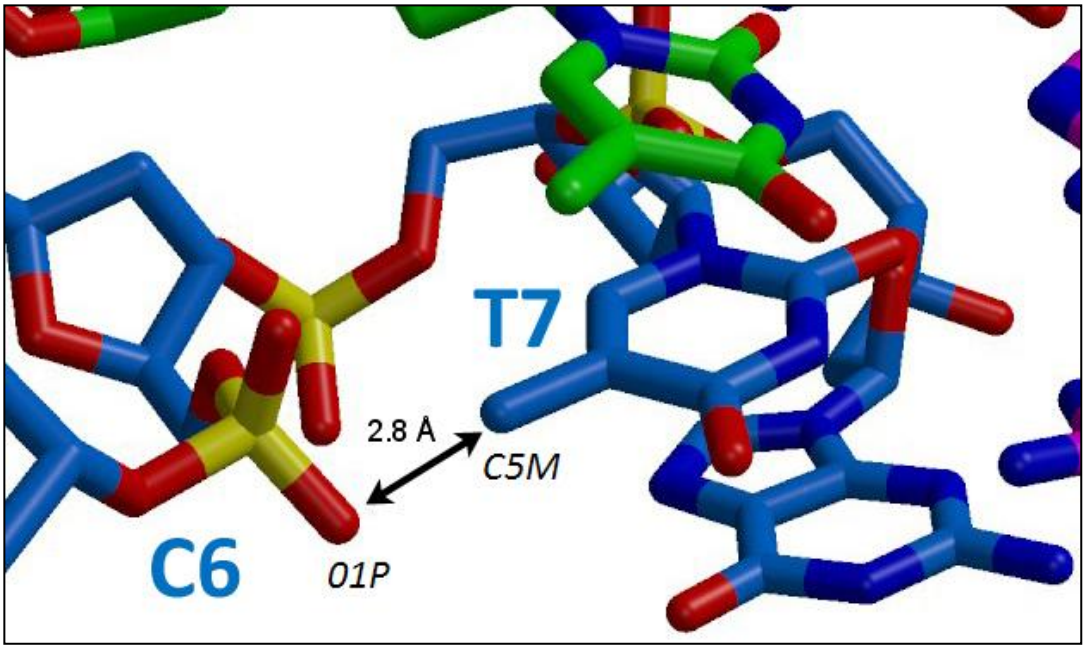


Figure 3.2. Intrastrand interaction between the methyl group of a thymine and phosphate oxygen of a cytosine of one crossover strand in the asymmetric junction (see Fig. 3.1; *blue*). The other crossover strand does not exhibit an analogous interaction.

interaction. Fewer compensatory stabilizing intrastrand interactions may explain the increased J_{twist} and J_{slide} of the asymmetric junction.

3.4.2 Wobble A·T base pair

Other than the discontinuity resulting from the junction crossover, the coaxially stacked arms form near continuous B-DNA type duplexes, with the base pairs near the junction and at the duplex ends all forming canonical Watson-Crick G·C or A·T type base pairs. The lone exception is the A2·T9 base pair of duplex 1 (Figure 3.1A). At these nucleotide positions, the N1 imino nitrogen of the adenine is closest to the O4 extracyclic oxygen (N···O distance of <2.8 Å) rather than to the N3 nitrogen of the thymine ring (N···N distance ≈ 3.1 Å), as would be expected for a standard A·T base pair, while the N6 extracyclic amino group of the adenine base is >3.3 Å from the O4 oxygen of the thymine (Fig. 3.3 A and B). The geometry indicates that the bases are paired by an N1···O4 hydrogen bond. This is supported by the C6 - N1···O4 angle of 110.73° and the N1···O4 – C4 angle of 109.87°, both of which approach the 120° for an ideal hydrogen bond. The A2·T9 is, therefore, in a wobble configuration, with the purine base sheared ~1.4 Å towards the major groove of the DNA duplex.

The apparent shearing of the base pair is not an artifact of how the structure was determined. Although this is a molecular replacement solution, the starting model had placed the A2 and T9 nucleotides of duplex 1 in the standard Watson-Crick configuration, the shearing persisted even after the base pairs were

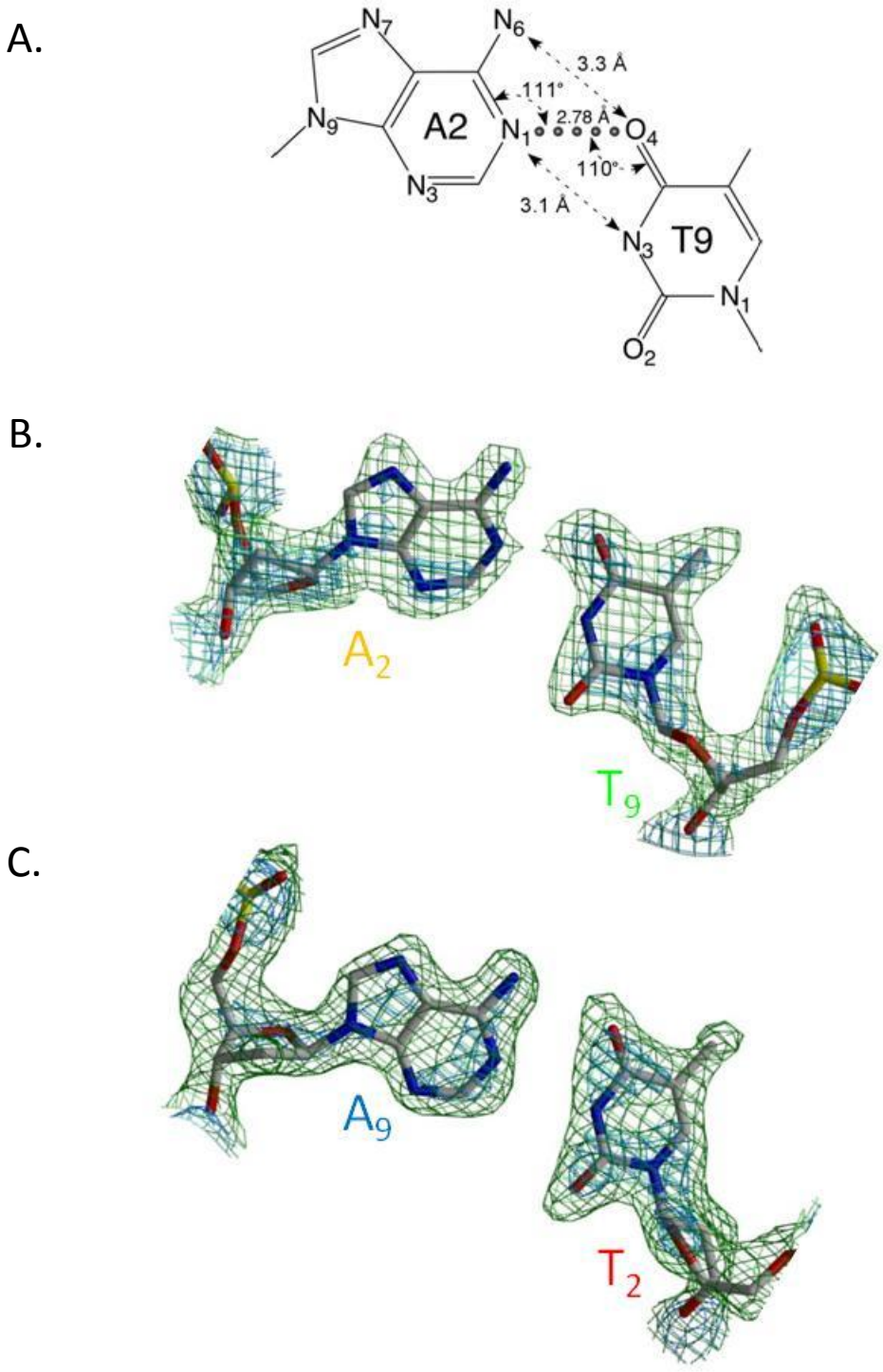


Figure 3.3. (A) Geometry of wobble A·T base pair. Dotted line indicates most probable hydrogen bonding distance and correlated angles. Straight dashed lines specify distances between hydrogen bond donors and acceptors if both bases were in the standard tautomeric form. Electron density map of (B) wobbled base pair and (C) symmetrically related standard T·A base pair located on the other four-base arm. $2F_o$-F_c map with 1.0σ (*green*) and 2.0σ (*blue*) contours.

forced back into the standard hydrogen-bonded form. We had also refined the structure in the absence of the bases of the nucleotides, allowing just the backbone to refine—in which case, the base pair was again sheared when the bases were added back. This indicated that the positions and orientations of the bases were defined not only by the electron density map but also the stereochemical constraints of the backbone. To force the adenine back to the standard configuration would require forcing the adenine back towards the minor groove by 1.4 Å, or a concerted shearing of the A and T by ~0.7 Å each, well beyond the ~0.3 Å coordinate error expected for a 1.9 Å resolution structure. The only source of error, therefore, would be for the DNA backbone in this region to be modeled incorrectly. Comparison of the electron density maps of the phosphodeoxyribose of A2 and T9 with the model reveals that the model fits the maps very well (Fig. 3.3B). Additionally, we did not observe any positive or negative difference density around the backbone or the bases to indicate that the atomic position of the final model was incorrect.

Average B-factors of the DNA backbone at these nucleotides were well within the mean for the structure as a whole (Fig. 3.4), indicating that these atoms are as well structured as the remainder of the DNA junction. Finally, the significant base shearing is seen only at this base pair, and not the two flanking base pair—if there had been a significant error in the backbone conformation at this penultimate base pair, we would have expected this error to propagate to the flanking base pairs as well, particularly to the terminal T1·A10 base pair of duplex

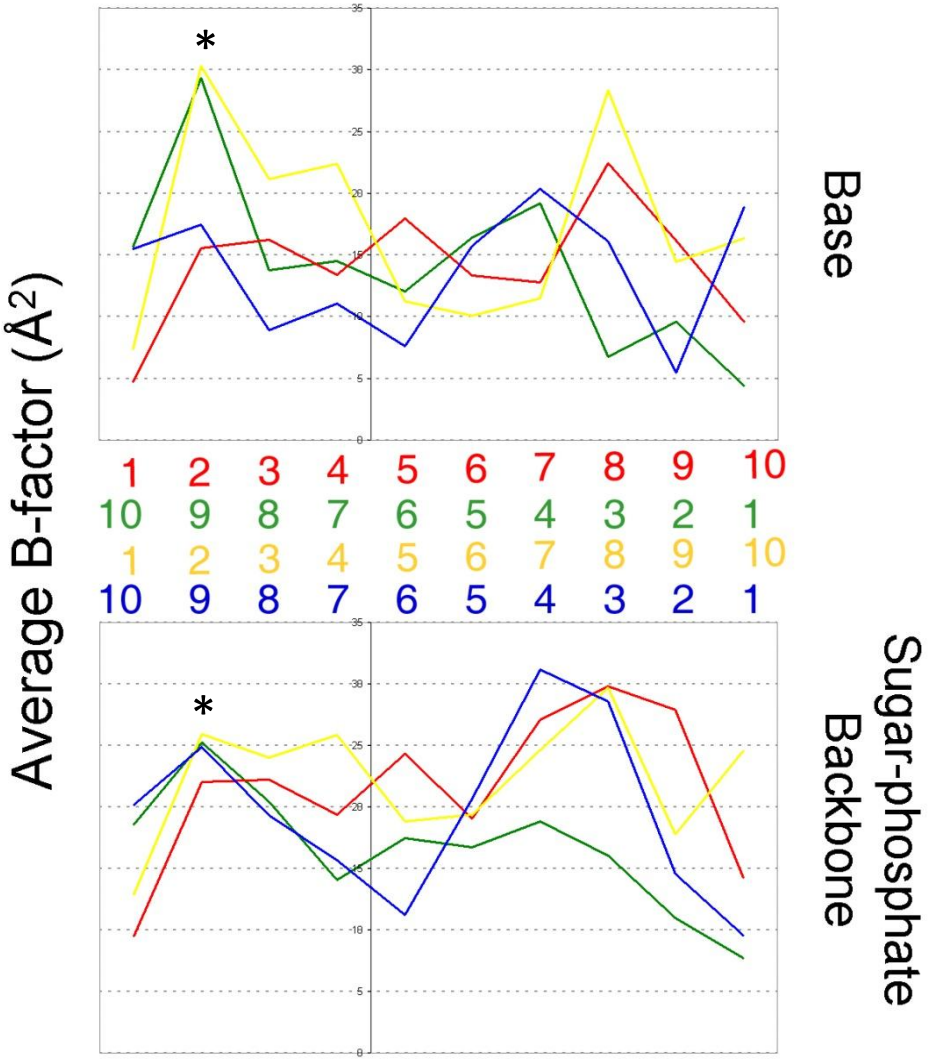


Figure 3.4. Average B-Factor values for atoms of (*top*) individual bases and (*bottom*) corresponding sugar-phosphate backbone. Colors are correlated to strand of sequence assignment (see Figure 3.1A). Values for atoms of wobble A·T are indicated by asterisks (*).

1. Each of the oligonucleotides was also analyzed by mass spectrometry to confirm the correct sequence for each DNA strand. Finally, this A2·T9 wobble was observed in the structure of the same junction construct, except with one of the cytosine bases brominated in order to define the orientation of the junction arms within the crystal unit cell. Thus, the experimental evidence indicates that A2·T9 is indeed an unusual wobbled base pair.

With only a single hydrogen bond between the wobble A2·T9 pair, the bases at these positions are not expected to be as well-ordered as other standard base pairs. The higher than average B-factors for the two wobbled nucleotide bases reflect the increased disorder, even though their respective backbone atoms remain well ordered. The helical parameters of the base pairs in and around the wobble indicates that the rise between the A2·T9 pair and the flanking nucleotides is extended (3.48 Å and 3.55 Å as compared to an average 3.40 Å for the other base pairs). In addition, there is a very large propeller twist between the two bases (-33.17° compared to the average -14.53° for the standard A·T base pairs in the structure). The wobble base pair also exhibit a large incline of 18.42°. Otherwise, the duplex appears relatively normal, suggesting that the standard B-DNA duplex can readily accommodate this unusual wobbled base pair.

3.5 Discussion

We present the single crystal structure of an asymmetric, sequence-locked Holliday junction constructed by the annealing of four unique DNA strands. This construct is similar to those that had previously been studied biochemically and biophysically and, most recently, by single-molecule methods. The similarity between structures of this asymmetric junction and the symmetric junctions indicates that the atomic details are not dependent on whether the Holliday junction is fixed by the sequence or by intrastrand interactions. Thus, the details for the effects of sequence (Carlstrom and Chazin 1996; Miick, Fee *et al.* 1997; Grainger, Murchie *et al.* 1998; Sha, Liu *et al.* 2002), cations (Cooper and Hagerman 1987; Duckett, Murchie *et al.* 1990; Clegg, Murchie *et al.* 1992; Clegg, Murchie *et al.* 1994), and phosphate charge (Liu, Declais *et al.* 2004; Liu, Declais *et al.* 2005) on the conformation, conformational isomerization, and dynamics of the recombination intermediate in solution can be related directly back to the atomic details seen in single crystals.

As with previous crystal studies on the junctions, the current structure also presented us with an unexpected result: an A·T base pair in the pseudo-continuous stacked B-DNA duplex arms that adopts a wobble configuration, where the adenine base is sheared by ~1.4 Å towards the major groove. The peculiar geometry can only be accommodated by a hydrogen bond between the N1 imino nitrogen of the adenine and the extracyclic O4 oxygen of the thymine bases.

The standard tautomer forms that define the hydrogen bond donor and acceptor groups of adenine and thymine would not account for structure of the A2·T9 wobble base pair. An N1···O4 hydrogen bond of this pair requires that one of the bases adopts a rare tautomer form, either the imino-tautomer of the adenine (with an imino extracyclic nitrogen at N6, and an N1 nitrogen that is changed from a hydrogen bond acceptor to a donor) or the enol-thymine (with a hydrogen shifted from the N3 nitrogen to the O4 and changing the oxygen from an acceptor to a donor). The experimental data cannot distinguish among the two possibilities, but the direction of the shearing may.

In the wobble configurations seen in the structures of G·T or A·C mismatches, it is the pyrimidine base that is sheared towards the major groove. A cobalt hexamine induced A·T wobble had previously been observed at the terminal base pair of a crystal structure of Z-DNA (Fig. 3.5) (Thiyagarajan, Rajan *et al.* 2004). In this case, the wobble was seen to be similar to the G·T and C·A mismatches, with the pyrimidine sheared towards the major groove. In this latter structure, the adenine was assumed to adopt the imino-tautomeric form, first because this was the form seen in mercury(II) induced tautomer of an A·U type base pair (Zamora, Kunsman *et al.* 1997) and, second the imino-adenine can form two hydrogen bonds between the A and T bases, similar to the canonical A·T base pair, with the thymine sheared towards the major groove. Finally, *ab initio* calculations on isolated bases suggest that the imino-tautomer of adenine is more stable than the enol-form of a cytosine base.

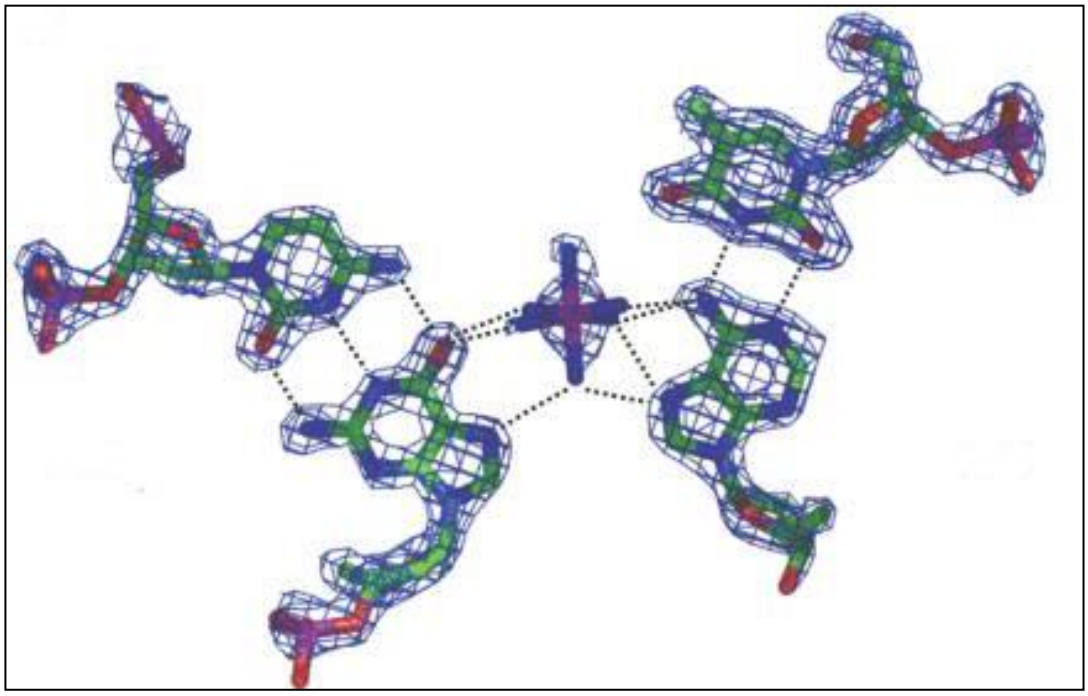


Figure 3.5. A cobalt hexamine induced A•T wobble (*right*) at the terminal base pair of a crystal structure of Z-DNA. Figure from Thiyagarajan*et al.* (2004) *Nucleic Acids Res*¸ **32**(19), 5945-53.

The observed A2·T9 wobble of the current structure shears the adenine towards the major groove, and has a single N1···O4 hydrogen bond. If it were an imino-adenine tautomer, we would have expected that the base pair would have wobbled in the opposite direction, with the thymine towards the major groove, which would have allowed for the formation of two interstrand hydrogen bonds. The opposite direction of the wobble would suggest that, in contrast to the metal induced tautomerizations, that it is the thymine that has adopted the rare enol-form.

The tautomerization of the thymine is not directly induced by the structure of the Holliday junction as the wobble A2·T9 base pair of duplex 1 is two steps away from the junction cross-over where structural perturbations would be more likely. In addition, an analogous T·A base pair at the opposite four-base pair arm of the current asymmetric junction adopts a canonical Watson-Crick base pair (Fig. 3.1A and 3.3C). Finally, no such wobbled A·T or G·C base pair has been reported in any crystal structure of junctions.

We thus propose that the A2·T9 wobble base pair in this structure can be attributed to sequence effects that stabilize a rare enol-tautomer of the thymine base or, less likely, the imino-tautomer of an adenine along a B-DNA duplex. With the A·T wobble at the end of 4 C·G base pairs (TCCCC·GGGGA), this is the only crystal structure of a B-type helix with this sequence motif; other sequences with this long stretch of contiguous C·G's adopt the alternative A-conformation

(McCall, Brown *et al.* 1985; Gao, Robinson *et al.* 1999; Ng, Kopka *et al.* 2000; Hays, Teegarden *et al.* 2005).

This proposal would suggest that a wobbled A·T base pair in B-DNA in specific sequence contexts would be highly mutagenic. In particular, an enol-thymine, upon replication, would readily undergo a transition mutation from an A·T to G·C base pair, consistent with the findings that this is one of the more common mutations observed in nature (Fig. 3.6). Indeed, A·C misincorporation has been shown to be dependent on the sequence of the flanking nucleotides, with misincorporation being on order of magnitude higher with polyG template or primers compared poly(A/T) (Topal, DiGuiseppi *et al.* 1980; Watanabe and Goodman 1981). These dramatic sequence dependent rates of misincoporation had previously been attributed to base-stacking effects at the DNA ends, but may now be seen to involve sequence influencing tautomer preference of the incoming or the template nucleotides. Indeed, the crystal structure of the high-fidelity DNA polymerase from *B. stearothermophilus* indicates that a rare protonated-tautomer of cytosine is paired to a mutagenic O6-methylguanine base (Warren, Forsberg *et al.* 2006), and that the rate of misincorporation of a T for a C is highly sequence dependent (Singer, Chavez *et al.* 1989).

The current structure, therefore, lends credence to the original “rare tautomer hypothesis” (Watson and Crick 1953) that uncommon tautomeric forms of the standard bases could result in nucleotide misincorporation by polymerases, inducing substitution mutation. In this case, we suggest that particular sequences

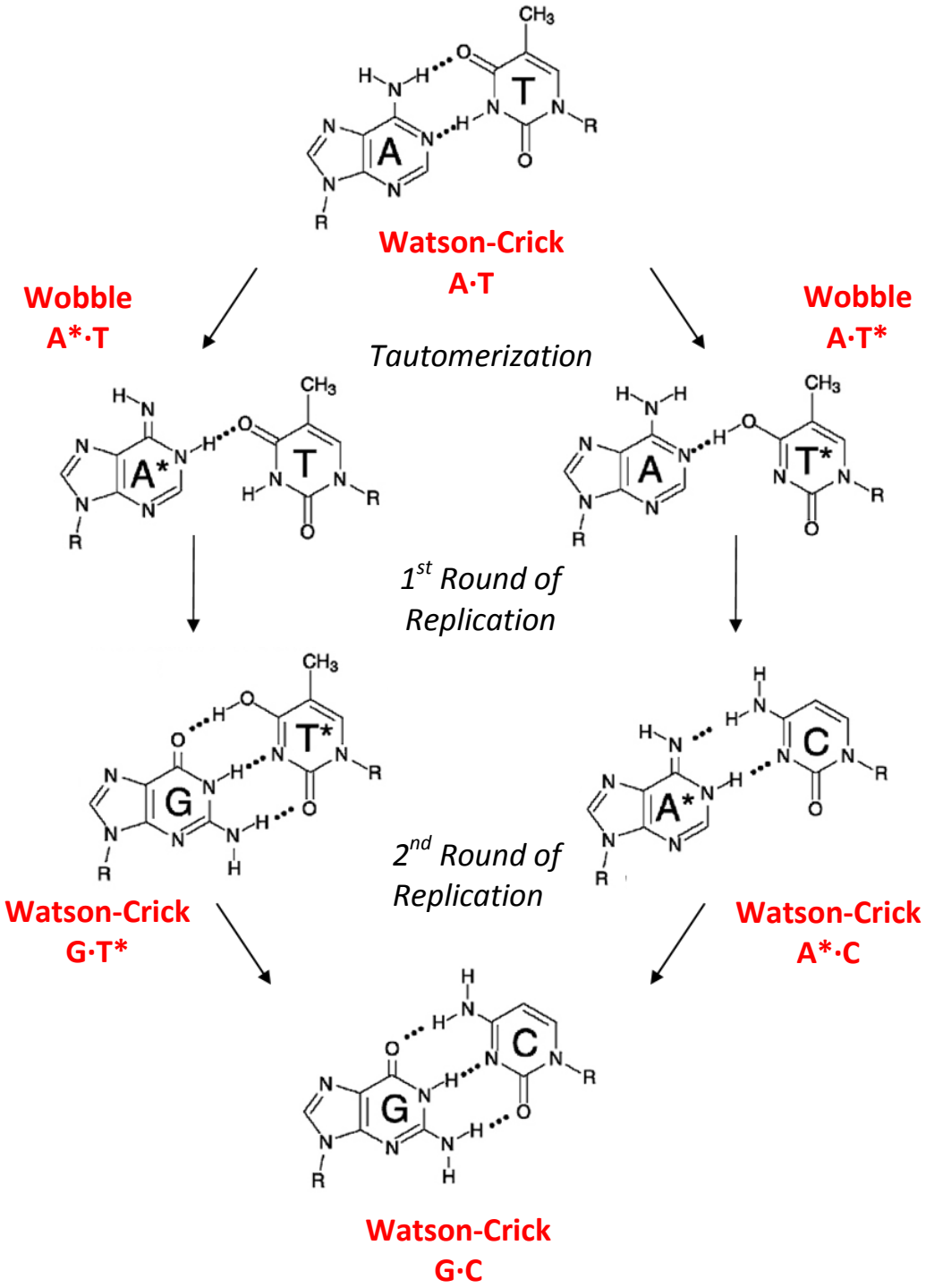


Figure 3.6. Transition mutation from A·T to G·C induced by tautomerization of one of the bases in the wobbled base pair. A* indicates an imino-adenine and T* indicates an enol-thymine.

help to stabilize rare tautomers in DNA. Particularly, long stretches of homo-CG base pairs promote either an enol-thymine or imino-adenine tautomer to contribute to the relatively common occurrence of transition mutations from A·T to G·C base pairs.

Chapter 4

Phylogenomic analysis of the emergence of GC-rich transcription elements

Patricia Khuu, Maurice Sandor, Jennifer DeYoung, and P. Shing Ho

Published in *Proc. Natl. Acad. Sci. USA,*

2007, **104** (42), 16528–16533

4.1 Summary

We have applied a comparative phylogenomic analysis to study the evolutionary relationships between GC content, CpG-dinucleotide content (CpGs), potential nuclear factor I (NFI) binding sites, and potential Z-DNA forming regions (ZDRs) as representative structural and functional GC-rich genomic elements. Our analysis indicates that CpG and NFI sites emerged with a general accretion of GC-rich sequences downstream of the eukaryotic transcription start site (TSS). Two distinct classes of ZDRs are observed at different locations proximal to the eukaryotic TSS. A robust CA/TG class of ZDRs was seen to emerge upstream of the TSS and independently of GC content, CpGs, and NFIs, whereas a second, weaker CG type appears to have evolved along with these downstream GC-rich elements. Taken together, the results provide a model for how GC-rich structural and functional eukaryotic markers emerge relative to each other, and indicate two distinct transition points for their occurrence: the first at the pro/eukaryotic boundary, and the second at or near the amniotic boundary.

4.2 Introduction

GC-rich regions of genomic DNA sequences are located at or near eukaryotic genes, serving as structural and functional "punctuation marks" for transcription (Zhang *et al*. 2004). Analysis of the prevalence and locations of GC-rich elements

across a large number of prokaryotic and eukaryotic genomes allows us to now trace their initial emergence and continued evolution in the eukaryotic genome and decipher the phylogenomic relationships between various transcription-related elements.

The GC content of a genome varies locally and regionally (Eyre-Walker and Hurst 2001; Zhang *et al*. 2004). Enrichment of GC-rich regions has been implicated in mutational bias, gene conversion bias, increased thermostability of the DNA duplex in thermophilic prokaryotes and warm-blooded eukaryotes, and structural plasticity associated with transcription (Bernardi 2000; Hurst and Williams 2000; Galtier 2003; Vinogradov 2003; Basak and Ghosh 2005). High GC content has also been correlated with short introns and elevated levels of gene transcription and recombination, whereas low GC content has been correlated with, among other things, tissue specificity and chromatin condensation (Montoya-Burgos *et al*. 2003; Versteeg *et al*. 2003; Vinogradov 2003; Vinogradov 2005). Increased GC content of sequences at and around the transcriptional start sites (TSSs) of genes suggests a functional relevance for GC-rich elements in higher eukaryotes (Zhang *et al*. 2004). Indeed, GC-rich mammalian genes exhibit up to 100-fold greater transcription rates than orthologous GC-poor genes (Kudla *et al*. 2006). Variations in GC content distribution may be a general property of the genes or may be associated with the emergence of GC-rich structural and functional transcriptional elements that contribute to the increased GC content. For example, CpG islands, defined as regions with GC content >50% and observed ratios of CpG dinucleotides

60% (Gardiner-Garden and Frommer 1987), have been shown to accumulate coincidentally with GC enrichment at the TSS of human genes and are used to predict genes in higher eukaryotes (Bird 1987; Antequera 2003). Other examples of GC-rich functional and structural elements include the CAAT-box sequence recognized by the nuclear factor I (NFI) transcription factor (Roulet *et al*. 2000) and CG-rich alternating pyrimidine-purine sequence regions with the potential to form left-handed Z-DNA (ZDRs) (Rich and Zhang 2003). Z-DNA has been implicated in several biological functions (Rich and Zhang 2003), including gene activation (Liu *et al*. 2001) and chromatin remodeling (Liu *et al*. 2006), and large-scale deletions in mammalian cells (Wang *et al*. 2006). Distributions of both NFI binding sites and ZDRs are correlated with the distribution of known and predicted genes across human chromosome 22 (Champ *et al*. 2004), accumulating around the TSS of human genes in a manner generally similar to those of GC content and CpG islands (Schroth *et al*. 1992; Champ *et al*. 2004). Here, we survey the patterns of occurrence of four GC-rich elements (GC content, potential CpG islands as reflected in the CpG-dinucleotide content, potential NFI bindings sites, and ZDRs) across a broad representation of genomic sequences to establish their phylogenomic relationships. Because CpG islands are not expected in prokaryotes, we do not directly count their occurrence, *per se*, but analyze for potential CpG islands by monitoring the percent of CpG dinucleotides across sequences. The patterns of distribution observed for these functional and structural elements

result in models for their emergence through divergence from a common GC-rich element and/or convergence of disparate elements.

4.3 Materials and Methods

4.3.1 Genome analyses

Sequences and annotations of prokaryotic genomes were accessed from the National Center for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov/genomes/lproks.cgi) and eukaryotic genomes from the Ensembl database (www.ensembl.org) (Birney *et al.* 2006) as their December 2005 releases. The current analyses include the genomes from 16 organisms (Table 4.1), representing four eubacteria, three archaea, yeast, worm, mosquito, two fish, chicken, and three mammals (rat, dog, and human). The particular eukaryotic genomes were chosen for analysis because of the consistency in their annotations in the Ensembl database and methylation of their genomes.

GC contents were calculated as the percent of G+C within 40-bp bins. Transcription start and stop sites for eukaryotic genes were as annotated in the Ensembl database (according to experimental transcripts). The near-identical distributions of GC content around the TSS and stop sites for human genes seen here, and as reported by Zhang *et al.* (2004), indicate that the annotations for these transcriptional markers are consistent with previous

Table 4.1. Analyses of prokaryotic and eukaryotic genomes for GC-rich transcriptional elements (percentage GC content, percentage CpG dinucleotides, number of NFI binding sites, and number of Z-DNA sequences)

	Genome size, Mbp (no. of chromosomes)	No. of genes	GC content, %	CpG, %	Total no. of NFIs	Total no. of ZDRs
Prokaryotes (Complete Microbial Genomes, www.ncbi.nlm.nih.gov/genomes/lproks.cgi)						
Eubacteria						
Synechocystis sp. PCC6083 (Syn)	3.57	3,218	47.72	10.05	15,982	118
B. subtilis subsp. *subtilis* str.168 (Bs)	4.21	4,226	43.52	10.92	7,827	2,143
E. coli K12 (Ec)	4.64	4,915	50.79	15.75	12,472	10,424
H. pylori J99 (Hp)	1.64	1,496	39.19	9.54	3,077	1,282
Archaea						
M. thermoautotrophicus ΔH (Mt)	1.75	1,921	49.54	7.80	2,791	206
A. fulgidus DSM 4304 (Af)	2.18	2,487	48.58	10.60	4,019	292
A. pernix K1 (Ap)	1.67	1,894	56.53	13.17	4,154	603

Eukaryotes (Ensembl v36-Dec 2005, www.ensembl.org/index.html)						
S. cerevisiae (Sc)	12 (16)	6,652[†]	38.42	6.73	20,729	1902
C. elegans (Ce)	100 (6)	19,723[†]	35.47	6.48	165,940	39,734
D. melanogaster (Dm)[‡]	118 (6)	13,733[§]	41.30	9.47	165,940	148,116
A. gambiae (Ag)	278 (5)	12,500[§]	44.73	11.06	388,446	693,596
T. nigroviridis (Tn)	402 (21)	15,357[†]	45.99	8.81	363,806	475,219
D. rerio (Dr)	1,688 (25)	18,009[§]	36.42	5.63	1,715,635	1,285,246
G. gallus (Gg)	1,054 (30)	15,348[§]	44.47	7.02	1,644,258	231,151
R. norvegicus (Rn)	2,719 (21)	21,939[§]	42.25	5.13	4,468,986	2,526,023
C. familiaris (Cf)	2,385 (39)	17,861[§]	40.90	5.14	4,555,797	1,061,843
H. sapiens (Hs)	3,272 (24[¶])	20,121[†]	41.53	5.44	5,693,028	1,065,255

[†]Number of annotated known RNA polymerase II (Pol II) transcribed genes.
[‡]Ensembl v42-Dec 2006.
[§]Number of annotated known and novel RNA Pol II transcribed genes.
[¶]Includes both the X and Y chromosomes.

analyses. Potential CpG islands (reflected in the CpG content) were analyzed similarly. The accepted definition for CpG islands (Gardiner-Garden and Frommer 1987) are long stretches of sequence (200 bp) with observed CpG occurrence versus expected occurrence (1/16 = 12.5%) 0.6; therefore, an actual CpG island would be five or more contiguous 40-bp bins having CpG contents 0.6, as calculated here. Because no CpG islands are expected in prokaryotic genomes, we calculated the potential for this element for all genomes (prokaryotic and eukaryotic) as the percent of CpG dinucleotides within a 40-bp window, rather than actual CpG islands, *per se*. NFI binding sites were analyzed according to affinity constants reported by Roulet *et al.* (Roulet *et al*. 2000), with a binding score of 65 (of a possible 100) representing a nonrandom and potentially functional binding site (enhancing gene expression *in vitro*) as the criteria for identifying an NFI site for this study, as described in Champ *et al.* (2004). With the threshold set at a lower binding score, the distributions around the TSSs approach the distributions of random sequences (i.e., become flat), whereas, at higher scores, the distributions become noisier (because of smaller numbers), but retain the general shapes reported here. ZDRs were defined as contiguous stretches of DNA with a P_Z score 500 bp, as defined in Champ *et al. (*2004), using the program Z-hunt (Ho *et al*. 1986; Schroth *et al*. 1992).

4.3.2 Quantitative analysis of distributions

The occurrence of each class of CpG, NFI, and ZDR elements were counted in 40-bp bins starting from –2 kb to +2 kb relative to the TSS of identified genes. Note that, for prokaryotic and lower eukaryotic genomes, this 4-kb window for analysis encompasses primarily coding sequences, whereas, for higher eukaryotes, it would include mostly noncoding sequences of introns, the 5'-untranslated regions, and intergenic DNA. In addition, for the compact prokaryotic and yeast genomes, there would be significant overlapping of transcriptional start and stop sites from multiple genes within this analysis window. However, these overlapping sites occur randomly across the windows, and any suppression or accumulation of elements is expected to contribute to the overall background levels within the windows and not contribute significantly to the patterns of positive and negative spikes identified at the reference TSS for the analysis of a particular genome. Indeed, the features observed for these compact genomes distribute over a very narrow range (200 bp). The magnitudes of each element within the 40-bp bins are normalized as the number of standard deviations from the mean of the average count for that organism, allowing us to directly compare the uniqueness of the GC-rich elements across all organisms regardless of the background levels. The center and boundaries of each peak in the normalized distributions were determined by calculating the first derivative of the distribution (APPENDIX). Centers were defined as points where the first derivative crosses zero, and boundaries were defined as the positive and negative peaks of the

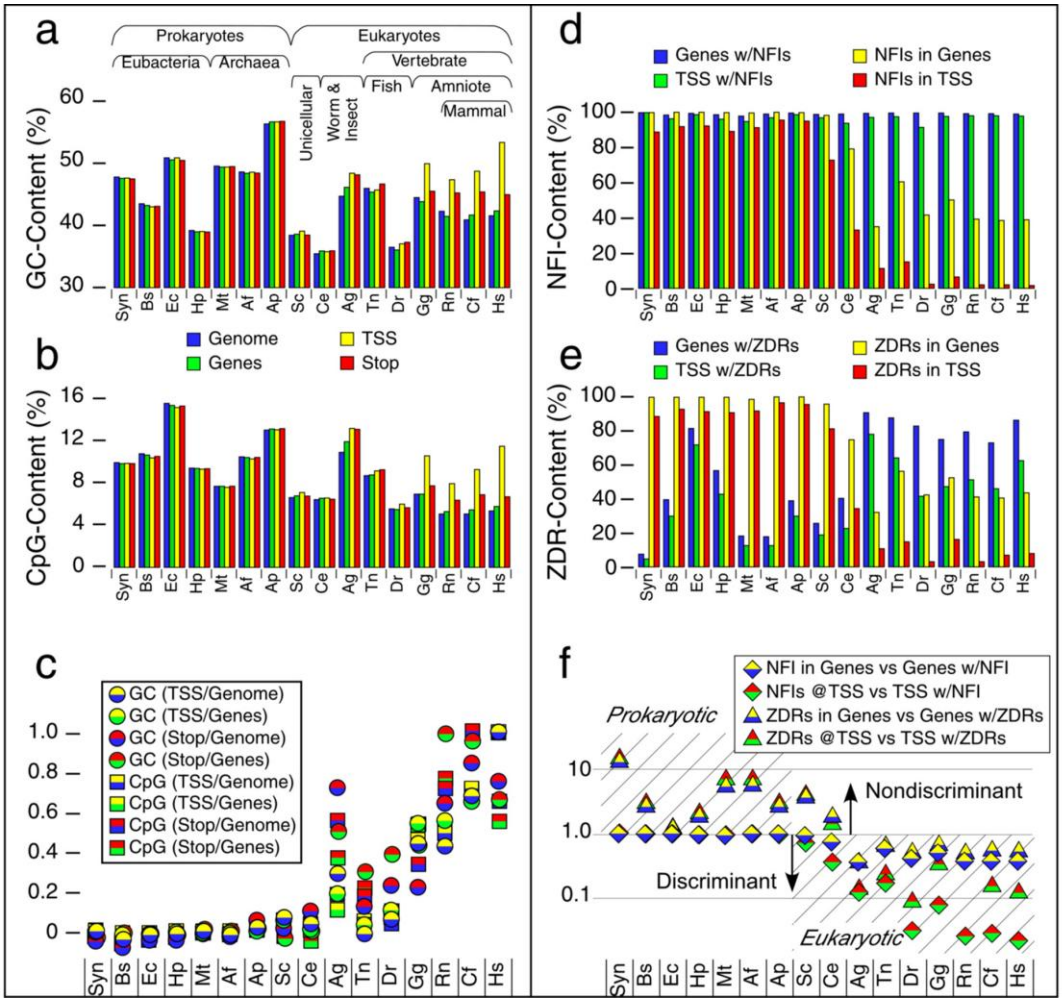
derivative. Although these boundaries do not account for the entire area within each peak, they provide a consistent definition that is independent of shifting baselines.

4.4 Results

In the current study, we survey the annotated genomes of nine eukaryotic and seven prokaryotic representative organisms. The genomes were placed in approximate order of increasing evolutionary complexity (Fig. 4.1A): cyanobacteria (generally recognized as one of the most primitive prokaryotic organisms (Taylor 1993)), other eubacteria, archaea, simple eukaryotes, and higher eukaryotes (including invertebrates and vertebrates, with chicken and mammals representing the amniotic organisms).

We initially compared the global GC and CpG content across the genomes, within genes, and at the TSS and termination (Stop) sites of genes (Fig. 4.1 A, B). This analysis shows that eukaryotic genomes are generally GC-poor (significantly less than 50%), as expected, whereas prokaryote genomes vary 50% (Basak and Ghosh 2005). Neither overall GC nor CpG content across genomes shows any particular pattern that correlates to the order of organismal complexity. Compared with genomic and genetic levels, however, one distinguishing feature is the significant enrichment of both GC content and CpG dinucleotides at the TSS and, to a lesser extent, at the stop sites, starting at the

Figure 4.1. Occurrence of GC-rich elements across organisms. (*A* and *B*) The abundance of GC-rich and CpG sites are shown for the genomes (blue) and genes (green) and at the transcription start (TSS, yellow) and termination (stop, red) sites of genes across prokaryotic and eukaryotic organisms. The TSS and stop sites are defined as the 40-bp bin that is centered at the respective gene marker. Gene sequences include the ORFs as well as the promoter and termination sequences both upstream of the TSS and downstream of the termination sequence, respectively. Representative genomes are arranged by approximate increasing complexity in the following order: cyanobacterium *Synechocystis* sp. (Syn); eubacteria *Bacillus subtilis* (Bs), *Escherichia coli* (Ec), and *Helicobacter pylori* (Hp); archaea *Methanothermobacter thermoautotrophicus* (Mt), *Archaeoglobus fulgidus* (Af), *Aeropyrum pernix* (Ap); unicellular eukaryote *Saccharomyces cerevisiae* (Sc); the invertebrate worm *Caenorhabditis elegans* (Ce); fruit fly *Drosophila melanogaster* (Dm); mosquito *Anopheles gambiae* (Ag); fugu fish *Tetraodon nigroviridis* (Tn), and zebrafish *Danio rerio* (Dr); chicken *Gallus gallus* (Gg); the mammals rat *Rattus norvegicus* (Rn), dog *Canis familiaris* (Cf) and human *Homo sapiens* (Hs). (*C*) Analysis of GC content (circles) and CpG content (squares) at the TSS and stop sites relative to those of the genome and in genes. The two colors in each point represent the ratio of content in the TSS versus the overall genome (yellow over blue), TSS versus genes (yellow over green), the stop site versus genome (red over blue), or stop versus genes (red over green). These ratios increase with increasing organismal complexity, indicating a general accumulation of these global elements at the beginning and end of genes. (*D* and *E*) The contents of NFI and ZDR sites were analyzed as the percent of genes (blue) and TSS (green) with these elements, and as the percent of NFI and ZDR sites within genes (yellow) and within TSS (red) sequences. (*F*) The ratio of NFI (diamonds) and ZDRs (triangles) in all genes (yellow) versus genes with either NFI or ZDR (blue) and at the TSS of all genes (red) versus TSS of genes with either NFI or ZDR (green) drops one order of magnitude with increasing complexity, reflecting an increasing discrimination in the location of these functional and structural elements. A ratio 1.0 was taken as evidence for no discrimination in localization of each element within genes and at the TSS of genes, but a ratio <1.0 indicates discrimination in localization of the elements.


a
Prokaryotes
Eukaryotes
Eubacteria
Archaea
Vertebrate
Unicellular
Worm & Insect
Fish
Amniote
Mammal
GC-Content (%)
60
50
40
30
Syn Bs Ec Hp Mt Af Ap Sc Ce Ag Tn Dr Gg Rn Cf Hs
b
Genome
Genes
TSS
Stop
CpG-Content (%)
16
12
8
4
0
c
GC (TSS/Genome)
GC (TSS/Genes)
GC (Stop/Genome)
GC (Stop/Genes)
CpG (TSS/Genome)
CpG (TSS/Genes)
CpG (Stop/Genome)
CpG (Stop/Genes)
1.0
0.8
0.6
0.4
0.2
0
d
Genes w/NFIs
TSS w/NFIs
NFIs in Genes
NFIs in TSS
NFI-Content (%)
100
80
60
40
20
0
e
Genes w/ZDRs
TSS w/ZDRs
ZDRs in Genes
ZDRs in TSS
ZDR-Content (%)
f
NFI in Genes vs Genes w/NFI
NFIs @TSS vs TSS w/NFI
ZDRs in Genes vs Genes w/ZDRs
ZDRs @TSS vs TSS w/ZDRs
Prokaryotic
10
1.0
0.1
Nondiscriminant
Discriminant
Eukaryotic

amniotic boundary between fish and birds. This enrichment is already evident in early eukaryotes (Fig. 4.1C), although it is not as striking.

This boundary becomes better defined from the analyses of the functional NFI and structural ZDR elements (Fig. 4.1D–F). Global distributions were analyzed for their general occurrences in genes (as opposed to intergenic regions) and at the TSS, whereas more specific distributions were determined from the number of genes and TSS that include these elements. Nearly all genes and TSS sequences across the genomes have at least one potential NFI binding sequence. However, the total numbers of overall NFI sites within genes and at the TSS decrease significantly starting at *S. cerevisiae*, and continue to drop precipitously with increased eukaryotic complexity. ZDR content shows parallel decreases within genes and TSS sequences, but a gradual increase is observed in the number of genes and TSS with ZDRs, consistent with the developing functionality of this structural element. Unlike the more general GC and CpG content that shows increasing accumulation at the TSS relative to their distributions in genes and across genomes (Fig. 4.1C), NFIs and ZDRs actually appear to become discriminated against in genes and at the TSS. Because NFI binding has no known function in prokaryotes, the sequences recognized by this eukaryotic activator should occur randomly (i.e., there would be no discrimination as to when and where NFIs might occur in prokaryotes). High ratios of NFIs in genes vs. genes with NFIs (Fig. 4.1F) indicate that there are many potential binding sites among the genes that have such sites and thus an indiscriminant distribution of such sites

across the gene. Low ratios, on the other hand, reflect a more discriminant localization with fewer sites within genes. For prokaryotes, this high ratio indicates very little discrimination for these potential binding sites in their genes. In eukaryotes, this ratio is reduced, reflecting an increase in discrimination for NFIs as a functional element within each gene. The same trends are observed for ZDRs in genes and for NFIs and ZDRs at the TSS of genes, suggesting that, as these elements assume specific functions, their localization along the genome becomes more explicit. The global analyses of these GC-rich elements indicate a distinct phylogenomic boundary at the lower eukaryotes (yeast and worms).

Detailed analyses of the distribution profiles of GC content, CpGs, and NFIs around the TSS (Fig. 4.2) suggest that these eukaryotic elements are closely related. The distributions show sharp dips immediately upstream of the TSS in eubacteria and archaea, which can be attributed to the localization of AT-rich promoters upstream of prokaryotic genes. In lower eukaryotes (*C. elegans*), the distributions show a broad negative peak upstream of the TSS, followed by a sharp spike immediately 5' of the TSS and a weak positive shoulder further downstream of the TSS. This general pattern is sustained, but broadened in *D. melanogaster*. Interestingly, this negative–positive pattern around the TSS mirrors that for H3 localization in *D. melanogaster* genes (Mito *et al*. 2005). In *A. gambiae*, the upstream negative distribution is lost, but the downstream distribution becomes a broad, highly asymmetric positive peak. In vertebrates (*D. rerio* and *H. sapiens*), the positive distribution becomes sharper, more symmetric, and

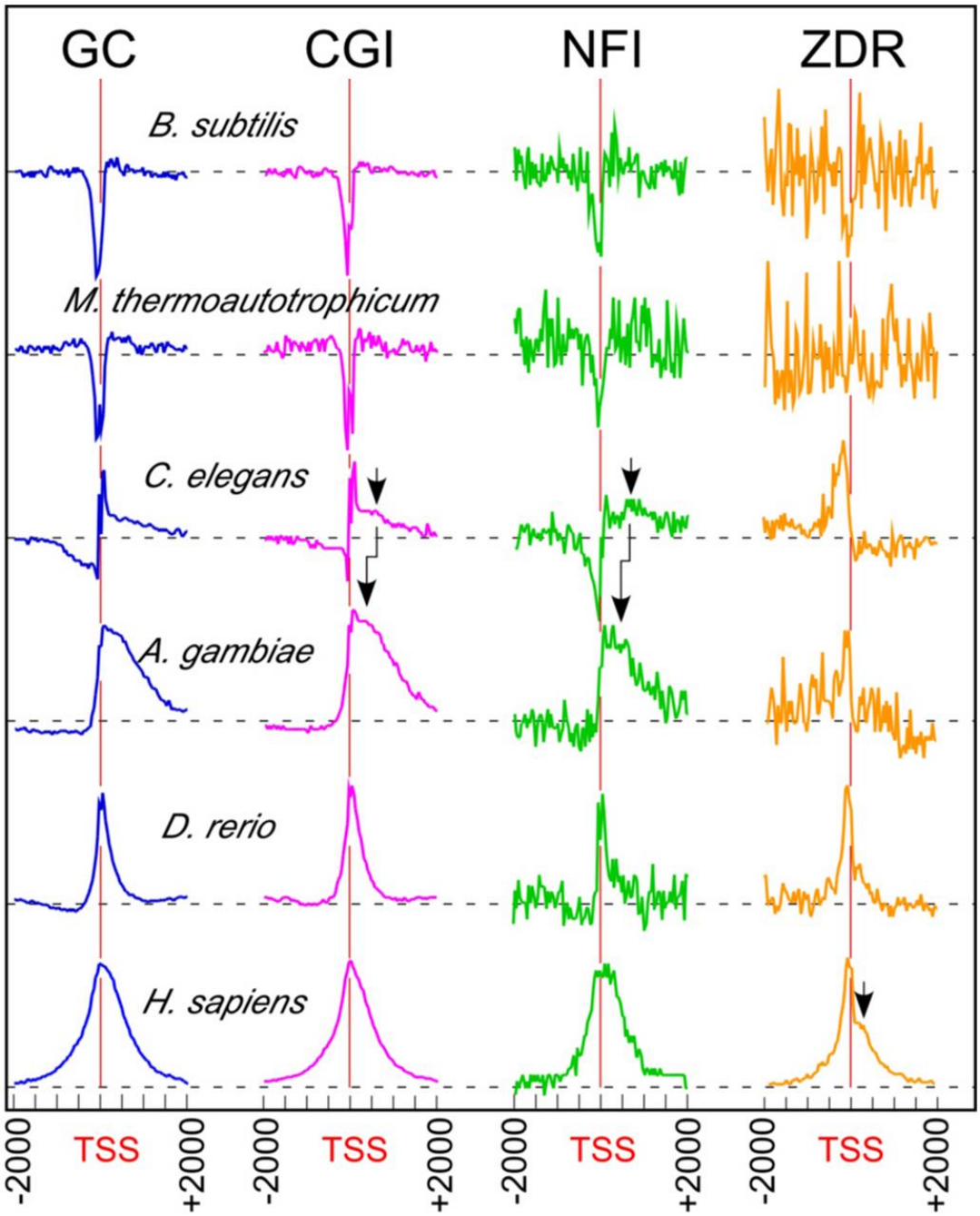


Figure 4.2. Distribution of GC-rich elements around the TSS of genes. The distributions (normalized for peak height) of GC content (GC) and CpG, NFI, and ZDR sites are plotted from –2,000 bp upstream to +2,000 bp downstream of the TSS for a set of representative genomes. Dashed horizontal lines indicate the average number of each element seen in the genomes. Arrows indicate positions of secondary (weaker) shoulders or peaks that are identified in these distributions.

centered at the TSS. Thus, distribution of GC content in eukaryotic genomes starts as asymmetric (being both positive and negative in lower eukaryotes) and becomes more symmetric and positive with increasing organismal complexity. The broad positive downstream shoulder seen in *C. elegans* becomes more distinct as a separate peak in the CpG and NFI distributions. This pattern suggests that the highly asymmetric positive downstream distributions in *A. gambiae* result from either migration toward or enhancement of this broad downstream peak around the TSS. A similar analysis of these distributions around termination sites showed no regular patterns across the phyla (data not shown). The phylogenomic pattern of ZDR distributions shows a weak suppression just upstream of the TSS of eubacteria, and no discernible pattern in archaea. A sharp positive peak is seen to emerge upstream of the TSS in *C. elegans*. A second, broader positive distribution is seen to emerge downstream of the TSS, weakly in *D. rerio*, but more distinctly in *H. sapiens*. Thus, the emergence of ZDRs appears to be distinct from that of GC content, CpGs, and NFIs. The distributions in Fig. 4.2 were quantified by first calculating their first derivative, which allowed us to identify the centers and the boundaries of all peaks in each distribution (APPENDIX). The intensity of each peak was calculated by summing the data between the peak boundaries. This analysis provides an overall picture across all organisms in the study of how each GC-rich element emerges and positions itself relative to the TSS. The patterns for GC content and CpGs are similar, both showing a negative peak 5' of the TSS in prokaryotes that evolutionarily migrates further upstream in lower eukaryotes

(Fig. 4.3 A and B). Within eukaryotes, the peaks are positive and start downstream of the TSS, but migrate toward the TSS with increasing organismal complexity. The CpGs, however, are centered at the TSS for all amniotic organisms (recapitulating the amniotic boundary seen in Fig. 4.1), whereas GC-content distributions are centered at the TSS only in *H. sapiens*, *S. cerevisae*, and *C. elegans* show both upstream suppression and downstream accumulation of GC content, indicative of the lower eukaryotes serving as a transitional boundary with features of both kingdoms. The NFI pattern displays features similar to GC content and CpGs, but with the pro/eukaryotic boundary extended into insects.

The phylogenomic pattern of ZDR distributions is weakly suppressed immediately upstream of the TSS of eubacterial genes, but shows no discernible pattern in archaea. A sharp positive peak is seen to emerge upstream of the TSS in lower eukaryotes and remains 100–200 bp upstream of the TSS in all eukaryotes. This class of strong, upstream Z-DNA elements (ZDR1) clearly arose independently of the other GC-rich elements. A second, weaker distribution of ZDRs (ZDR2) starts as a negative peak far downstream of the TSS in *C. elegans*, which becomes weakly positive in *R. norvegicus* and centered closer to the TSS in *H. sapiens*. Thus, there are apparently two distinct classes of ZDRs that emerge: the first is distinct from and the second appears correlated to GC enrichment and the emergence of CpGs and NFIs across the phylogenomic spectrum.

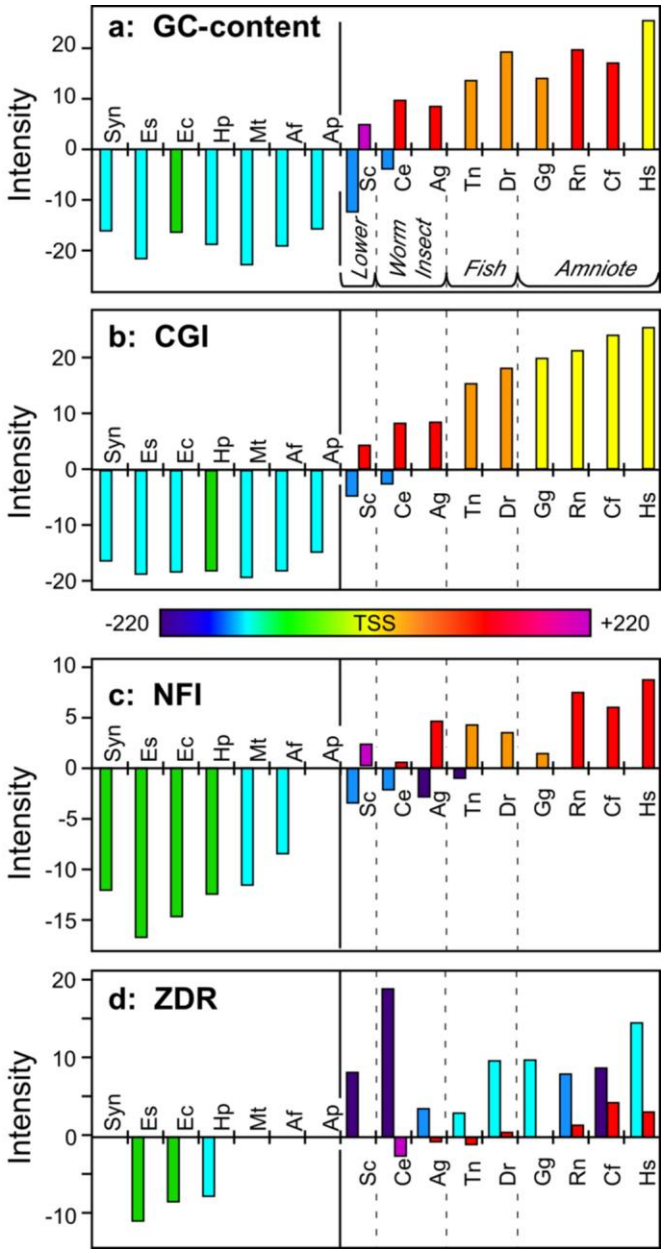


Figure 4.3. Phylogenomic patterns of enrichment or suppression of GC-rich transcriptional elements. The intensity of GC content (*A*), CpG (*B*), NFI (*C*), and ZDR (*D*) distributions are shown for various representative genomes. The positions of the centers of each distribution are shown relative to the TSS (red arrow in Fig. 4.2) of the genes in the genomes, with those centered increasingly upstream shown as green to blue to violet and those increasingly downstream shown as orange to red to magenta (the color scale for positions of centers is shown between the panels).

4.5 Discussion

When considered together, the phylogenomic patterns suggest a model for the emergence of GC-rich structural and functional elements for eukaryotic genes. It is clear from these patterns that CpG dinucleotides (and, by extension, CpG islands) and the GC-rich NFI transcriptional activator binding sites both emerged coincidentally with the increasing GC content just 3' of the eukaryotic TSS. The starting and ending positions for the centers of these distributions suggests that CpG islands evolved with a subset of high GC-content regions that are relatively close to the TSS, and NFI sequences started further downstream of the TSS. All three elements migrate toward the TSS in parallel, as one would expect for elements that become increasingly important for transcriptional regulation, but the migration stops once it reaches the TSS. An attractive alternative model is that, rather than the transcriptional elements migrating relative to the TSS, both TA- and GC-rich elements are relatively fixed across the various organisms, but the TSS migrates evolutionarily in the 3' direction (Fig. 4.4). This would be consistent with the dramatic increase in the size of the transcriptosome (the proteins of the transcription machinery) at the pro/eukaryotic boundary (the number of subunits of RNA polymerase triples from *E. coli* to yeast) and the increased numbers of transcriptional regulatory elements in the higher eukaryotes. The increase in size and complexity of the transcriptosome that accompanies evolutionary complexity would provide a physical rationale for the downstream migration of the TSS away

from the primordial TA-rich transcriptional elements.

The results of the phylogenomic analysis suggest that the stronger ZDR1-type structural elements emerged independently of GC and CpG content, even though Z-DNA is characteristic of alternating GC-rich sequences. ZDR1s are most likely alternating CA/TG-type Z-DNA sequences, as opposed to the prototypical alternating GC motif. ZDR1 sequences, however, are not simply repeats of CA/TG, as seen in the repetitive regions of eukaryotic chromosomes, but are similar to the CA/TG-rich sequences characteristic, for example, of the promoters in rat genes (Rothenburg *et al*. 2001). The convergence of ZDR1s toward the downstream GC-rich elements such as NFI may reflect the emergence of the more complex mechanism of structural/nuclear factor coactivation, as seen in the human CSF-1 promoter (Rich and Zhang 2003).

The lower intensity ZDR2 class follows the general trend of the GC-rich elements, suggesting that these are the prototypical GC-type Z-DNA sequences and they arose perhaps as a consequence of GC content and CpG islands rather than as a distinct element in itself. The emergence of GC-rich isochores has been proposed to be associated with Z-DNA, as well as thermal stability and helix bendability (Vinogradov 2003). The emergence of two distinct classes of ZDRs may reflect the plurality of functions now recognized for Z-DNA in various genomes (Rich and Zhang 2003).

When viewed as a whole, the phylogenomic relationships seen here

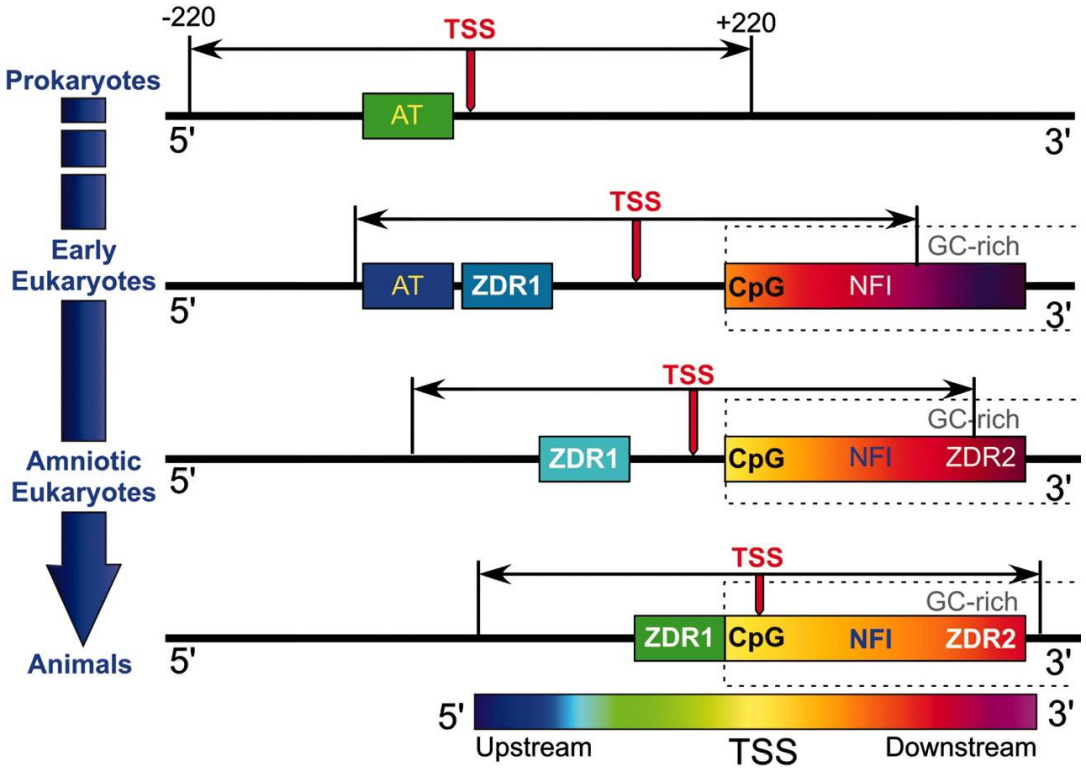


Figure 4.4. Model for the emergence of GC-rich transcriptional elements and migration of the transcription start site (TSS, red arrow) of genes from prokaryotes to early eukaryotes to amniotic eukaryotes, and, finally, to higher eukaryotes. In this model, the prokaryotic AT-rich promoters and the GC-rich eukaryotic elements are seen to be fixed, whereas the TSS and the analysis window (±220 bp relative to the TSS) migrate in the 3' direction as the size and complexity of the transcriptosome increases. In prokaryotes, the primary transcriptional control elements are AT-rich upstream promoters, which account for the strong suppression of GC-rich elements that are characteristic of eukaryotes. Early eukaryotes show both the persistence of localized upstream AT-rich promoters that are characteristic of prokaryotes, as well as the accumulation of the eukaryotic GC-rich elements (CpG and NFI elements) as the GC content (dashed boxes) downstream of the TSS increases. The first Z-DNA regions (ZDR1) also emerge at this point, independently of the other GC-rich elements. In progressing toward amniotic eukaryotes, the TSS migrates further in the 3' direction, followed by the ZDR1 elements. This is in concert with AT-rich promoters becoming less distinct (and, therefore, their locations are not specified in this figure), the emergence of a second class of CG-rich Z-DNAs (ZDR2) accumulated downstream of the TSS, and convergence of the upstream ZDR1 sites with the GC-rich elements.

suggest that GC-rich transcriptional elements evolved gradually rather than abruptly across organisms, but with two distinct boundaries. The lower eukaryotes can be perceived as the pro/eukaryotic transition, showing characteristics of both types, consistent with a continuity across this transition. The second interface is at or near the amniotic transition, where the GC content changes from a broad asymmetric to a sharper symmetric distribution, CpG dinucleotides have fully localized at the TSS, and ZDR2-type sequences are enriched rather than suppressed. Thus, these GC-rich elements are a means to decipher phylogenomic relationships at the gene level, even without knowing their specific functions. What remains unclear at this level of analysis is whether patterns of emergence of these punctuation elements are entirely organismal or related to the emergence of specific genes or gene functions in each class of organism.

4.6 Acknowledgments

We thank Drs. C. K. Mathews, G. Merrill, and M. Freitag at Oregon State University for reviewing the manuscript and providing comments before submission. This work was funded by National Institutes of Health Grant R01GM62957A and the Medical Research Foundation of Oregon.

Chapter 5

Discussion

Although B-DNA is predominant under physiological conditions, alternative structures may exist locally and transiently in different segments of the genome. To date, more than ten different non-B-DNA structures have been discovered, with some implicated in specific mechanisms relevant to genetic stability and diseases. These structures may function to influence chromatin structure, gene expression and gene regulation (Sinden 1994). Our studies focused on the Holliday junction, a rare tautomeric base pair and the left-handed Z-DNA. Until recently, these structures and their biological relevance remained poorly understood. Our findings provide unique insights that help to further elucidate their roles in biological processes.

Our crystallographic study of the asymmetric Holliday junction allowed us to reconcile and extend the relevance of the solution and crystallographic studies of junctions. It revealed that the findings associated with symmetric junctions are also valid with sequence-locked asymmetric junctions. Thus, the results would facilitate future studies of junctions and junction-specific proteins, such as the co-crystallization of junction complexed with resolvases.

The unexpected structure of tautomeric A·T base pair is particularly intriguing as it is unique in its formation. Known factors that affect proper base

pairing or induce tautomerization are apparently absent. With no proximal ion observed, flanked by standard W-C base pairs and distal to the junction cross-over where occurrences of distortions to standard base pairing geometries would be more likely, the wobble base pair adds insight into other factors that may promote and stabilize rare tautomers. Arguably, two obvious physiologically irrelevant factors may be plausible contributors to tautomerization: crystallization conditions and junction structure context. However, since our crystallization conditions and junction structure are consistent with those of other previously crystallized junctions and nucleic acid structures which exhibit no rare tautomers, we believe that the proposed tautomerization cannot be an artifact of crystallization conditions (Watson *et al.* 2004). Accordingly, since no other crystallized junction exhibits wobbled geometry of a W-C base pair particular to the B-DNA-like arm under comparable conditions, we also contend that junction structure context does not directly induce the evident wobble. The junction structure has only localized disruptions at the strand crossover; otherwise, its stacked duplex arms are nearly identical to continuous B-DNA duplexes, indicating that the observed wobble is not directly associated with the four-stranded junction itself.

We thus suggest that the observed wobble was induced by its sequence context. Four tandem guanines are adjacent to the wobble site (italicized): 5'-T*A*GGGGCCGA-3'. The conformation duality of polyguanine tracts has been suggested by multiple models which find that some sequences with polyguanine

tracts have minima in two conformational states: B-DNA and A-DNA. Regions of DNA with extended tracts of guanines undergo a dynamic equilibrium between the A-DNA-like and B-DNA-like conformations. Vertical shifts and local conformation adjustments allow consecutive base pairs to optimize base stacking energy, which predispose some sequences to adopt nonstandard conformations, such as A-DNA and Z-DNA (Watson *et al*. 2004; Pilet *et al*. 1975; Packer *et al*. 2000; Gardiner *et al*. 2003; Lankas *et al*. 2000; Stefl *et al*. 2001; Lankas *et al*. 2002). Moreover, other studies have shown that polyguanine tracts longer than two bases enhance structural bend and exhibit A-like behavior despite being in the B form (Milton *et al.* 1990; Lindqvist and Graslund 2001). Correlated to the current junction structure, these findings suggest that the properties particular to polyguanine tracts which enable them to adopt the A conformation may also promote and stabilize rare tautomers of flanking nucleotides. Such occurrences may induce spontaneous tautomerization of a base, with potentially mutagenic consequences. We have provided unique evidence of the structure of a rare tautomer that have perhaps added new perspective of sequence as a source of mutation.

Lastly, our phylogenomic studies of 16 genomes yielded valuable insights into the origin and propagation of Z-DNA forming sequences. The left-handed Z-DNA duplex was the first single-crystal structure of a DNA double helix, but its biological relevance has only been recently illuminated. Our findings further corroborate its biological role. The results show a prevalence of sequences near

the eukaryotic transcription start site that can adopt the Z-DNA conformation, suggestive of a fundamental function for Z-DNA. Moreover, two distinct populations of sequences with the potential to form Z-DNA intimate multiplicity of function.

Our method of analysis has provided a unique glimpse into the origin and evolution of a sequence motif and could be applicable for study of other motifs. In discovering the possible evolutionary origin of a sequence motif, we can also learn about its relationship to other sequence motifs and possibly reveal unnoticed relationships between protein families that co-evolved with them.

Bibliography

Antequera, F. (2003). Structure, function and evolution of CpG island promoters. *Cell Mol. Life. Sci.* **60**(8), 1647-58.

Arauzo-Bravo MJ, Fujii S, Kono H, Ahmad S, Sarai A. (2005) Sequence-dependent conformational energy of DNA derived from molecular dynamics simulations: toward understanding the indirect readout mechanism in protein-DNA recognition*. J. Am. Chem. Soc.* **127**(46), 16074–16089.

Aravind L, Makarova KS, Koonin EV. (2000) SURVEY AND SUMMARY: holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories. *Nucleic Acids Res.* **28**(18), 3417–3432.

Arnott S. (1999) Polynucleotide secondary structures: an historical perspective. *The Oxford Handbook of Nucleic Acid Structures.* Neidle S (ed.). Oxford University Press: New York, 1–38.

Arnott S. (2006) Historical article: DNA polymorphism and the early history of the double helix. *Trends in Biochem. Sci.* **31**, 349-54.

Auffinger P, Hays FA, Westhof E, Ho PS. (2004) Halogen bonds in biological molecules. *Proc. Natl. Acad. Sci.* **101**(48), 16789–16794.

Aymami J, Pous J, Lisgarten JN, Coll M. (2002) Crystallization and preliminary X-ray analysis of the DNA decamers d(CCGGATCCGG) and d(CCGGCGCCGG). *Acta Crystallogr. D. Biol. Crystallogr.* **58**(Pt 2), 310–311.

Basak S, Ghosh TC. (2005) On the origin of genomic adaptation at high temperature for prokaryotic organisms. *Biochem. Biophys. Res. Commun*. **330**(3), 629-32.

Bennett RJ, West SC. (1995a) Structural analysis of the RuvCHolliday junction complex reveals an unfolded junction. *J. Mol. Biol.* **252**, 213–226.

Bernardi G. (2000) Isochores and the evolutionary genomics of vertebrates. *Gene*. **241**(1), 3-17.

Biertumpfel C, Yang W and Suck D. (2007) Crystal structure of T4 endonuclease VII resolving a Holliday junction. *Nature.* **449**, 616-20.

Bird A. (1987) CpG islands as gene markers in the vertebrate nucleus. *Trends. Genet.* **3**, 342-348.

Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Flicek P, Gräf S, Hammond M, Herrero J, Howe K, Iyer V, Jekosch K, Kähäri A, Kasprzyk A, Keefe D, Kokocinski F, Kulesha E, London D, Longden I, Melsopp C, Meidl P, Overduin B, Parker A, Proctor G, Prlic A, Rae M, Rios D, Redmond S, Schuster M, Sealy I, Searle S, Severin J, Slater G, Smedley D, Smith J, Stabenau A, Stalker J, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodwark C, Hubbard TJ. (2006) Ensembl 2006. *Nucleic Acids. Res.* **34**, D556–D561.

Biswas T, Aihara H, Radman-Livaja M, Filman D, Landy A, Ellenberger T. (2005) A structural basis for allosteric control of DNA recombination by lambda integrase. *Nature.* **435**(7045), 1059–1066.

Bowater RP, Wells RD. (2001) The intrinsically unstable life of DNA triplet repeats associated with human hereditary disorders. *Prog. Nucleic Acid Res. Mol. Biol.* **66**, 159–202.

Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D Biol. Crystallogr*. **54**(Pt 5), 905-21.

Carlstrom G, Chazin WJ. (1996). Sequence dependence and direct measurement of crossover isomer distribution in model Holliday junctions using NMR spectroscopy. *Biochemistry*. **35**(11), 3534-44.

Champ PC, Maurice S, Vargason JM, Camp T, Ho PS. (2004) Distributions of Z-DNA and nuclear factor I in human chromosome 22: a model for coupled transcriptional regulation. *Nucleic Acids Res.* **32**(22), 6501-10.

Chan SN, Harris L, Bolt EL, Whitby MC, Lloyd RG. (1997) Sequence specificity and biochemical characterization of the RusA Holliday junction resolvase of Escherichia coli. *J. Biol. Chem.* **272**, 14873–14882.

Chen Y, Narendra U, Iype LE, Cox MM, Rice PA. (2000) Crystal structure of a Flp recombinase-Holliday junction complex: assembly of an active oligomer by helix swapping. *Mol. Cell.* **6**(4), 885–897.

Cheng X, Kelso C, Hornak V, de los Santos C, Grollman AP and Simmerling C. (2005) Dynamic behavior of DNA base pairs containing 8-oxoguanine. *J. Am. Chem. Soc.* **127**, 13906-18.

Clegg RM, Murchie AI, Lilley DM. (1994) The solution structure of the four-way DNA junction at low-salt conditions: a fluorescence resonance energy transfer analysis." *Biophys. J.* **66**(1), 99-109.

Clegg RM, Murchie AI, Zechel A, Carlberg C, Diekmann S, Lilley DM. (1992). Fluorescence resonance energy transfer analysis of the structure of the four-way DNA junction. *Biochemistry*. **31**(20), 4846-56.

Cooper JP, Hagerman PJ. (1987) Gel electrophoretic analysis of the geometry of a DNA four-way junction. *J. Mol. Biol*. **198**, 711–719.

Cooper JP, Hagerman PJ. (1989) Geometry of a branched DNA structure in solution. *Proc .Natl. Acad. Sci.* **86**, 7336–7340.

Cox M. (2001) Recombinational DNA repair of damaged replication forks in Escherichia coli: questions. *Annu. Rev. Genet.* **35**, 53–82.

Cox MM, Goodman MF, Kreuzer KN, Sherratt DJ, Sandler SJ, Marians KJ. (2000) The importance of repairing stalled replication forks. *Nature.* **404**, 37–41.

Declais AC and Lilley DM. (2008) New insight into the recognition of branched DNA structure by junction-resolving enzymes. *Curr. Opin. Struct. Biol.* **18**, 86-95.

Declais AC, Fogg JM, Freeman AD, Coste F, Hadden JM, Phillips SE, Lilley DM. (2003) The complex between a four-way DNA junction and T7 endonuclease I. *EMBO J.* **22**(6), 1398–1409.

Dickerson RE. (1983) The DNA helix and how it is read. *Sci. Am.* **249**(6), 94–98.

Duckett DR, Murchie AI, Lilley DM. (1990) The role of metal ions in the conformation of the four-way DNA junction. *EMBO J.* **9**(2), 583–590.

Duckett DR, Murchie AIH, Diekmann S, von Kitzing E, Kemper B, Lilley DMJ. (1988) The structure of the Holliday junction, and its resolution. *Cell.* **55**, 79–89.

Eichman BF, Mooers BHM, Alberti M, Hearst JE, Ho PS. (2001) The crystal structures of psoralen cross-linked DNAs: drug dependent formation of Holliday junctions. *J. Mol. Biol.* **301**, 15–26.

Eichman BF, Ortiz-Lombardia M, Aymami J, Coll M, Ho PS. (2002) The inherent properties of DNA four-way junctions: comparing the crystal structures of holliday junctions. *J. Mol. Biol*. **320**, 1037-51.

Eichman BF, Vargason JM, Mooers BHM, Ho PS. (2000) The Holliday junction in an inverted repeat sequence: sequence effects on the structure of four-way junctions. *Proc. Natl. Acad. Sci.* **97**, 3971–3976.

Eyre-Walker A, Hurst LD. (2001) The evolution of isochores. *Nat Rev Genet* **2**(7), 549-55.

Fleming K, Riser DK, Kumari D, Usdin K. (2003) Instability of the fragile X syndrome repeat in mice: the effect of age, diet and mutations in genes that affect DNA replication, recombination and repair proficiency. *Cytogenet. Genome Res.* **100**(1–4), 140–146.

Flores-Rozas H, Kolodner RD. (2000) Links between replication, recombination and genome instability in eukaryotes. *Trends Biochem. Sci.* **292**, 196–200.

Fogg JM, Kvaratskhelia M, White MF, Lilley DM. (2001) Distortion of DNA junctions imposed by the binding of resolving enzymes: a fluorescence study. *J. Mol. Biol.* **313**(4), 751–764.

Fu Y, Comella N, Tognazzi K, Brown LF, Dvorak HF, Kocher O. (1999) Cloning of DLM-1, a novel gene that is up-regulated in activated macrophages, using RNA differential display. *Gene.* **240**, 157-63.

Galtier N. (2003) Gene conversion drives GC content evolution in mammalian histones. *Trends. Genet.* **19**(2), 65-8.

Gao YG, Robinson H, Wang AH. (1999) High-resolution A-DNA crystal structures of d(AGGGGCCCCT). An A-DNA model of poly(dG) x poly(dC). *Eur. J. Biochem*. **261**(2), 413-20.

Gardiner EJ., Hunter CA, Packer MJ, Palmer DS, Willett P. (2003) Sequence-dependent DNA structure: a database of octamer structural parameters. *J. Mol. Biol.* **332**, 1025-35.

Gardiner-Garden M, Frommer M. (1987). CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**(2), 261-82.

Giraud-Panis MJ, Lilley DM. (1998) Structural recognition and distortion by the DNA junction-resolving enzyme RusA. *J. Mol. Biol.* **278**, 117–133.

Goodman MF, Ratliff RL. (1983) Evidence of 2-aminopurine-cytosine base mispairs involving two hydrogen bonds. *J. Biol. Chem.* **258**, 12842-6.

Grady WM. (2004) Genomic instability and colon cancer. *Cancer Metastasis Rev.* **23**(1–2), 11–27.

Grainger RJ, Murchie AI, Lilley DM. (1998). "Exchange between stacking conformers in a four-Way DNA junction." *Biochemistry*. **37**(1), 23-32.

Guo F, Gopaul DN, van Duyne GD. (1997) Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. *Nature*. **389**(6646), 40–46.

Hadden JM, Declais AC, Carr SB, Lilley DM, Phillips SE. (2007) The structural basis of Holliday junction resolution by T7 endonuclease I. *Nature*. **449**, 621-4.

Hargreaves D, Rice DW, Sedelnikova SE, Artymiuk PJ, Lloyd RG, Rafferty JB. (1998) Crystal structure of E. coli RuvA with bound DNA Holliday junction at 6 Å resolution. *Nat. Struct. Biol.* **5**, 441–446.

Hays FA, Jones ZJ, Ho PS. (2004) Influence of minor groove substituents on the structure of DNA Holliday junctions. *Biochemistry.* **43**(30), 9813–9822.

Hays FA, Schir V, Ho PS, Demeler B. (2006) Solution formation of Holliday juncions in inverted-repeat DNA sequences. *Biochemistry*. **45**(8), 2461–2471.

Hays FA, Teegarden A, Jones ZJ, Harms M, Raup D, Watson J, Cavaliere E, Ho PS. (2005) How sequence defines structure: a crystallographic map of DNA structure and conformation. *Proc. Natl. Acad. Sci.* **102**(20), 7157–7162.

Hays FA, Vargason JM, Ho PS. (2003a) Effect of sequence on the conformation of DNA holliday junctions. *Biochemistry*. **42**(32), 9586–9597.

Hays FA, Watson J, Ho PS. (2003b) Caution! DNA crossing: crystal structures of Holliday junctions. *J. Biol. Chem.* **278**(50), 49663–49666.

Heinemann U, Alings C, Bansal M. (1992) Double helix conformation, groove dimensions and ligand binding potential of a G/C stretch in B-DNA. *EMBO J.* **11**, 1931–1939.

Henning W, Sturzbecher HW. (2003) Homologous recombination and cell cycle checkpoints: Rad51 in tumour progression and therapy resistance. *Toxicology.* **193**, 91-109.

Herbert A, Lowenhaupt K, Spitzner J, Rich A. (1995) Chicken double-stranded RNA adenosine deaminase has apparent specificity for Z-DNA. *Proc. Natl. Acad. Sci.* **92**, 7550-4.

Ho PS, Eichman BF. (2001) The crystal structures of DNA Holliday junctions. *Current Opin. Struct. Biol.* **11**, 302–308.

Ho PS, Ellison MJ, Quigley GJ, Rich A. (1986) A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences. *EMBO. J.* **5**(10), 2737-44.

Holliday R. (1964) A mechanism for gene conversion in fungi. *Genet. Res.* **5**, 282–304.

Holliday R. (1974) Molecular aspects of genetic exchange and gene conversion. *Genetics*. **78**, 273–287.

Hurst LD, Williams EJ. (2000) Covariation of GC content and the silent site substitution rate in rodents: implications for methodology and for the evolution of isochores. *Gene*. **261**(1), 107-14.

Kallenbach NR, Ma RI, Wand AJ, Veeneman GH, van Boom JH, Seeman NC. (1983) Fourth rank immobile nucleic acid junctions. *J. Biomol. Struct. Dyn.* **1**(1): 159–168.

Kamstra SA, Kuipers AG, De Jeu MJ, Ramanna MS, Jacobsen E. (1999) The extent and position of homoeologous recombination in a distant hybrid of Alstroemeria: a molecular cytogeneti assessment of first generation backcross progenies. *Chromosoma.* **108**(1), 52–63.

Karow JK, Constantinou A, Li J-L, West SC, Hickson ID. (2000) The Bloom's syndrome gene product promotes branch migration of Holliday junctions. *Proc. Natl. Acad. Sci. USA.* **97**, 6504–6508.

Khuu P, Sandor M, DeYoung J, Ho PS. (2007) Phylogenomic analysis of the emergence of GC-rich transcription elements. *Proc. Natl. Acad. Sci. USA*. **104**(42), 16528-33.

Khuu PA, Voth AR, Hays FA, Ho PS. (2006) The stacked-X DNA Holliday junction and protein recognition. *J. Mol. Recognit.* **19**, 234-42.

Kono H, Sarai A. (1999) Structure-based prediction of DNA target sites by regulatory proteins. *Proteins*. **35**, 114-31.

Kreuzer KN. (2004) Interplay between DNA replication and recombination in prokaryotes. *Annu. Rev. Microbiol*. **59**, 43–67.

Kudla G, Lipinski L, Caffin F, Helwak A, Zylicz M. (2006) High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol* **4**(6), e180.

Kwan KY, Moens PB, Wang JC. (2003) Infertility and aneuploidy in mice lacking a type IA DNA topoisomerase III beta. *Proc. Natl. Acad. Sci. USA*. **100**(5), 2526–2531.

Lankas, F., J. Sponer, P. Hobza and J. Langowski (2000). "Sequence-dependent elastic properties of DNA." J Mol Biol 299, 695-709.

Lankas, F., T. E. Cheatham, 3rd, N. Spackova, P. Hobza, J. Langowski and J. Sponer (2002). "Critical effect of the N2 amino group on structure, dynamics, and elasticity of DNA polypurine tracts." Biophys J 82, 2592-609.

Liao S, Mao C, Birktoft JJ, Shuman S, Seeman NC. (2004) Resolution of undistorted symmetric immobile DNA junctions by vaccinia topoisomerase I. *Biochemistry.* **43**(6), 1520-1531.

Lilley DM, White MF. (2001) The junction-resolving enzymes. *Nat. Rev. Mol. Cell. Biol.* **2**(6), 433–443.

Lilley DMJ. (1999) Structures and interactions of helical junctions in nucleic acids. *Oxford Handbook of Nucleic Acid Structure*, Neidle S (ed.). Oxford University Press: New York. 471–498.

Lilley DMJ. (2000) Structures of helical junctions in nucleic acids. *Quart. Rev. Biochem.* **33**, 109–159.

Lindqvist M, Graslund A. (2001). An FTIR and CD Study of the Structural Effects of G-tract Length and Sequence Context on DNA Conformation in Solution." *J. Mol. Biol.* **314**, 423-432.

Liu H, Mulholland N, Fu H, Zhao K. (2006) Cooperative activity of BRG1 and Z-DNA formation in chromatin remodeling. *Mol. Cell. Biol.* **26**(7), 2550-9.

Liu J, Declais AC, Lilley DM. (2004) Electrostatic interactions and the folding of the four-way DNA junction: analysis by selective methyl phosphonate substitution. *J. Mol. Biol.* **343**, 851-64.

Liu J, Declais AC, McKinney SA, Ha T, Norman DG, Lilley DM. (2005) Stereospecific effects determine the structure of a four-way DNA junction. *Chem. Biol.* **12**, 217-28.

Liu LF, Wang JC. (1987) Supercoiling of the DNA template during transcription. *Proc. Natl. Acad. Sci. USA.* **84**, 7024-7.

Liu R, Liu H, Chen X, Kirby M, Brown PO, Zhao K. (2001) Regulation of CSF1 promoter by the SWI/SNF-like BAF complex. *Cell*. **106**(3), 309-18.

Lombard DB, Chua KF, Mostoslavsky R, Franco S, Gostissa M, Alt FW. (2005) DNA repair, genome stability, and aging. *Cell* **120**(4), 497–512.

Lu X-J, Shakked Z, Olson WK. (2000) A-form conformational motifs in ligand-bound DNA structures. *J. Mol. Biol.* **300**, 819–840.

Lushnikov AY, Bogdanov A, Lyubchenko YL. (2003) DNA recombination: holliday junctions dynamics and branch migration. *J. Biol. Chem.* **278**(44), 43130–43134.

MacDonald M, Hassold T, Harvey J, Wang LH, Morton NE, Jacobs P. (1994) The origin of 47,XXY and 47,XXX aneuploidy: heterogeneous mechanisms and role of aberrant recombination. *Hum. Mol. Genet.* **3**(8), 1365–1371.

Macmaster R, Sedelnikova S, Baker PJ, Bolt EL, Lloyd RG, Rafferty JB. (2006) RusA Holliday junction resolvase: DNA complex structure--insights into selectivity and specificity. *Nucleic Acids Res.* **34**, 5577-84.

McCall M, Brown T, Kennard O. (1985) The crystal structure of d(G-G-G-G-C-C-C-C). A model for poly(dG).poly(dC). *J. Mol. Biol*. **183**(3), 385-96.

McGregor N, Ayora S, Sedelnikova S, Carrasco B, Alonso JC, Thaw P, Rafferty J. (2005) The structure of Bacillus subtilis RecU Holliday junction resolvase and its role in substrate selection and sequence-specific cleavage. *Structure.* **13**(9), 1341–1351.

McKee BD. (2004) Homologous pairing and chromosome dynamics in meiosis and mitosis. *Biochim. Biophys. Acta*. **1677**(1–3), 165–180.

McKim KS, Jang JK, Manheim EA. (2002) Meiotic recombination and chromosome segregation in Drosophila females. *Annu. Rev. Genet.* **36**, 205–232.

McKinney SA, Declais AC, Lilley DM, Ha T. (2003) Structural dynamics of individual Holliday junctions. *Nat. Struct. Biol.* **10**(2): 93–97.

McKinney SA, Freeman AD, Lilley DM, Ha T. (2005) Observing spontaneous branch migration of Holliday junctions one step at a time. *Proc. Natl. Acad. Sci. USA*. **102**(16), 5715-20.

Middleton CL, Parker JL, Richard DJ, White MF, Bond CS. (2004) Substrate recognition and catalysis by the Holliday junction resolving enzyme Hje. *Nucleic Acids Res.* **32**(18), 5442–5451.

Miick SM, Fee RS, Millar DP, Chazin WJ. (1997) Crossover isomer bias is the primary sequence-dependent property of immobilized Holliday junctions. *Proc. Natl. Acad. Sci. USA.* **94**(17), 9080-4.

Milton DL, Casper ML, Wills NM, Gesteland RF. (1990) Guanine tracts enhance sequence directed DNA bends. *Nucleic Acids Res.* **18**, 817-20.

Mito Y, Henikoff JG, Henikoff S. (2005) Genome-scale profiling of histone H3.3 replacement patterns. *Nat. Genet*. **37**, 1090-7.

Montoya-Burgos JI, Boursot P, Galtier N. (2003) Recombination explains isochores in mammalian genomes. *Trends Genet.* **19**(3), 128-30.

Morrison C, Vagnarelli P, Sonoda E, Takeda S, Earnshaw WC. (2003) Sister chromatid cohesion and genome stability in vertebrate cells. *Biochem. Soc. Trans.* **31**(Pt 1), 263–265.

Murchie, AI, Clegg RM, von Kitzing E, Duckett DR, Diekmann S, Lilley DM. (1989) Fluorescence energy transfer shows that the four-way DNA junction is a right-handed cross of antiparallel molecules. *Nature.* **341**, 763-6.

Ng HL, Kopka ML, Dickerson RE. (2000) The structure of a stable intermediate in the A <--> B DNA helix transition. *Proc. Natl. Acad. Sci. USA*. **97**(5), 2035-9.

Nowakowski J, Shim PJ, Prasad GS, Stout CD, Joyce GF. (1999) Crystal structure of an 82-nucleotide RNA-DNA complex formed by the 10–23 DNA enzyme. *Nat. Struct. Biol.* **6**, 151–156.

Nowakowski J, Shim PJ, Stout CD, Joyce GF. (2000) Alternative conformations of a nucleic acid four-way junction. *J. Mol. Biol.* **300**, 93–102.

Olson WK, Gorin AA, Lu XJ, Hock LM, Zhurkin VB. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA.* **95**(19), 11163–11168.

Ortiz-Lombardıa M, Gonzalez A, Eritja R, Aymamı J, Azorın F,Coll M. (1999) Crystal structure of a DNA Holliday junction. *Nat. Struct. Biol.* **6**, 913–917.

Otwinowski Z, Schevitz RW, Zhang RG, Lawson CL, Joachimiak A, Marmorstein RQ, Luisi BF, Sigler PB. (1988) Crystal structure of trp repressor/operator complex at atomic resolution. *Nature*. **335**(6188), 321–329.

Packer MJ, Dauncey MP, Hunter CA. (2000) Sequence-dependent DNA structure: tetranucleotide conformational maps. *J. Mol. Biol.* **295**, 85-103.

Pilet J, Blicharski J, Brahms J. (1975) Conformations and structural transitions in polydeoxynucleotides. *Biochemistry.* **14**, 1869-76.

Podolyan Y, Gorb L, Leszczynski J. (2005) Rare tautomer hypothesis supported by theoretical studies: ab initio investigations of prototropic tautomerism in the N-methyl-p base. *J. Phys. Chem. A.* **109**, 10445-50.

Raaijmakers H, Vix O, Tor I, Golz S, Kemper B, Suck D. (1999) X-ray structure of T4 endonuclease VII: a DNA junction resolvase with a novel fold and unusual domain-swapped dimer architecture. *EMBO J.* **18**, 1447–1458.

Rich A, Nordheim A, Wang AH. (1984) The chemistry and biology of left-handed Z-DNA. *Annu. Rev. Biochem.* **53**, 791-846.

Rich A, Zhang S. (2003) Timeline: Z-DNA: the long road to biological function. *Nat Rev Genet.* **4**, 566-72.

Rodier F, Kim SH, Nijjar T, Yaswen P, Campisi J. (2005) Cancer and aging: the importance of telomeres in genome maintenance. *Int. J. Biochem. Cell. Biol.* **37**(5), 977-990.

Roe SM, Barlow T, Brown T, Oram M, Keeley A, Tsaneva IR, Pearl LH. (1998) Crystal structure of an octameric RuvA-Holliday junction complex. *Cell.* **2**, 361–372.

Rothenburg S, Koch-Nolte F, Rich A, Haag F. (2001) A polymorphic dinucleotide repeat in the rat nucleolin gene forms Z-DNA and inhibits promoter activity. *Proc. Natl. Acad. Sci. USA*. **98**(16), 8985-90.

Roulet E, Bucher P, Schneider R, Wingender E, Dusserre Y, Werner T, Mermod N. (2000) Experimental analysis and computer prediction of CTF/NFI transcription factor DNA binding sites. *J. Mol. Biol.* **297**(4), 833-48.

Schroth GP, Chou PJ, Ho PS. (1992) Mapping Z-DNA in the human genome. Computer-aided mapping reveals a nonrandom distribution of potential Z-DNA-forming sequences in human genes. *J. Biol. Chem.* **267**(17), 11846-55.

Schwartz T, Behlke J, Lowenhaupt K, Heinemann U, Rich A. (2001) Structure of the DLM-1-Z-DNA complex reveals a conserved family of Z-DNA-binding proteins. *Nat. Struct. Biol.* **8**, 761-5.

Seeman NC, Kallenbach NR. (1983) Design of immobile nucleic acid junctions. *Biophys. J.* **44**(2), 201–209.

Seeman NC, Maestre MF, Ma RI, Kallenbach NR. (1985) Physical characterization of a nucleic acid junction. *Prog. Clin. Biol. Res.* **172A**, 99–108.

Sekharudu CY, Yathindra N, Sundaralingam M. (1993) Molecular dynamics investigations of DNA triple helical models: unique features of the Watson-Crick duplex. *J Biomol Struct Dyn* **11**, 225-44.

Sha R, Liu F, Seeman NC. (2002) Atomic force microscopic measurement of the interdomain angle in symmetric Holliday junctions. *Biochemistry.* **41**(19), 5950–5955.

Sharples GJ. (2001) The X philes: structure-specific endonucleases that resolve Holliday junctions. *Mol. Microbiol.* **39**(4), 823–834.

Sherratt DJ, Soballe B, Barre FX, Filipe S, Lau I, Massey T, Yates J. (2004) Recombination and chromosome segregation. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **359**(1441), 61–69.

Sinden RR. (1994) DNA Structure and Function. Academic Press, New York.

Singer B, Chavez F, Goodman MF, Essigmann JM, Dosanjh MK. (1989) Effect of 3' flanking neighbors on kinetics of pairing of dCTP or dTTP opposite O6-methylguanine in a defined primed oligonucleotide when Escherichia coli DNA polymerase I is used. *Proc. Natl. Acad. Sci. USA.* **86**(21), 8271-4.

Smith KC. (2004) Recombinational DNA repair: the ignored repair systems. *Bioessays.* **26**(12), 1322–1326.

Stefl RL, Trantirek L, Vorlickova M, Koca J, Sklenar V, Kypr J. (2001) A-like guanine-guanine stacking in the aqueous DNA duplex of d(GGGGCCCC). *J. Mol. Biol.* **307**, 513-24.

Subramaniam S, Tewari AK, Nunes-Duby SE, Foster MP. (2003) Dynamics and DNA substrate recognition by the catalytic domain of lambda integrase. *J. Mol. Biol.* **329**(3): 423–439.

Sun W, Mao C, Liu F, Seeman NC. (1998) Sequence dependence of branch migratory minima. *J. Mol. Biol*. **282**, 59–70.

Taylor T. (1993). The Biology and Evolution of Fossil Plants Englewood Cliffs, NJ, Prentice Hall.

Thiyagarajan S, Rajan SS, Gautham N. (2004) Cobalt hexammine induced tautomeric shift in Z-DNA: the structure of d(CGCGCA)*d(TGCGCG) in two crystal forms. *Nucleic Acids Res.* **32**(19), 5945-53.

Thorpe JH, Gale BC, Teixeira SC, Cardin CJ. (2003) Conformational and Hydration Effects of Site-selective Sodium, Calcium and Strontium Ion Binding to the DNA Holliday Junction Structure d(TCGGTACCGA)(4). *J. Mol. Biol.* **327**(1): 97–109.

Timsit Y, Westhof E, Fuchs RPP, Moras D. (1989) Unusual helical packing in crystals of DNA bearing a mutation hot spot. *Nature*. **341**, 459–462.

Topal MD, DiGuiseppi SR, Sinha NK. (1980) Molecular basis for substitution mutations. Effect of primer terminal and template residues on nucleotide selection by phage T4 DNA polymerase in vitro. *J. Biol. Chem.* **255**(24), 11717-24.

Topal MD, Fresco JR. (1976) Complementary base pairing and the origin of substitution mutations. *Nature.* **263**, 285-9.

Urnov FD, Miller JC, Lee YL, Beausejour CM, Rock JM, Augustus S, Jamieson AC, Porteus MH, Gregory PD, Holmes MC. (2005) Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature.* **435**(7042), 646–651.

Vargason JM, Ho PS. (2002) The effect of cytosine methylation on the structure and geometry of the Holliday junction: the structure of d(CCGGTACm5CGG) at 1.5 A resolution. *J. Biol. Chem.* **277**(23), 21041–21049.

Versteeg R, van Schaik BD, van Batenburg MF, Roos M, Monajemi R, Caron H, Bussemaker HJ, van Kampen AH. (2003) The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res*. **13**(9), 1998-2004.

Vinogradov AE. (2003) DNA helix: the importance of being GC-rich. *Nucleic Acids Res*. **31**(7), 1838-44.

Vinogradov AE. (2003) Isochores and tissue-specificity. *Nucleic Acids Res.* **31**(17), 5212-20.

Vinogradov AE. (2005) Dualism of gene GC content and CpG pattern in regard to expression in the human genome: magnitude versus breadth. *Trends Genet*. **21**(12), 639-43.

Voth AR, Hays FA, Ho PS. (2007) Directing macromolecular conformation through halogen bonds. *Proc. Natl. Acad. Sci. USA*. **104**(15), 6188-93.

Wang G, Christensen LA, Vasquez KM. (2006) Z-DNA-forming sequences generate large-scale deletions in mammalian cells. *Proc. Natl. Acad. Sci. USA*. 103(8): 2677-82.

Wang G, Vasquez KM.(2007) Z-DNA, an active element in the genome. *Front Biosci.* **12**, 4424-38.

Warren JJ, Forsberg LJ, Beese LS. (2006) The structural basis for the mutagenicity of O(6)-methyl-guanine lesions. *Proc. Natl. Acad. Sci. USA.* **103**(52), 19701-6.

Watanabe SM, Goodman MF. (1981) On the molecular basis of transition mutations: frequencies of forming 2-aminopurine.cytosine and

adenine.cytosine base mispairs in vitro. *Proc. Natl. Acad. Sci. USA.* **78**(5): 2864-8.

Watson J, Hays FA, Ho PS. (2004) Definitions and analysis of DNA Holliday junction geometry. *Nucleic Acids Res.* **32**(10), 3017-3027.

Watson JD, Crick FH. (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*. **171**, 737-8.

West SC. (2003) Molecular views of recombination proteins and their control. *Nat. Rev. Mol. Cell Biol*. **4**, 435-45.

White MF, Giraud-Panis M-JE, Pohler JRG, Lilley DMJ. (1997) Recognition and manipulation of branched DNA structure by junction-resolving enzymes. *J. Mol. Biol.* **269**, 647–664.

White MF, Lilley DM. (1997) The resolving enzyme CCE1 of yeast opens the structure of the four-way DNA junction. *J. Mol. Biol.* **266**(1), 122–134.

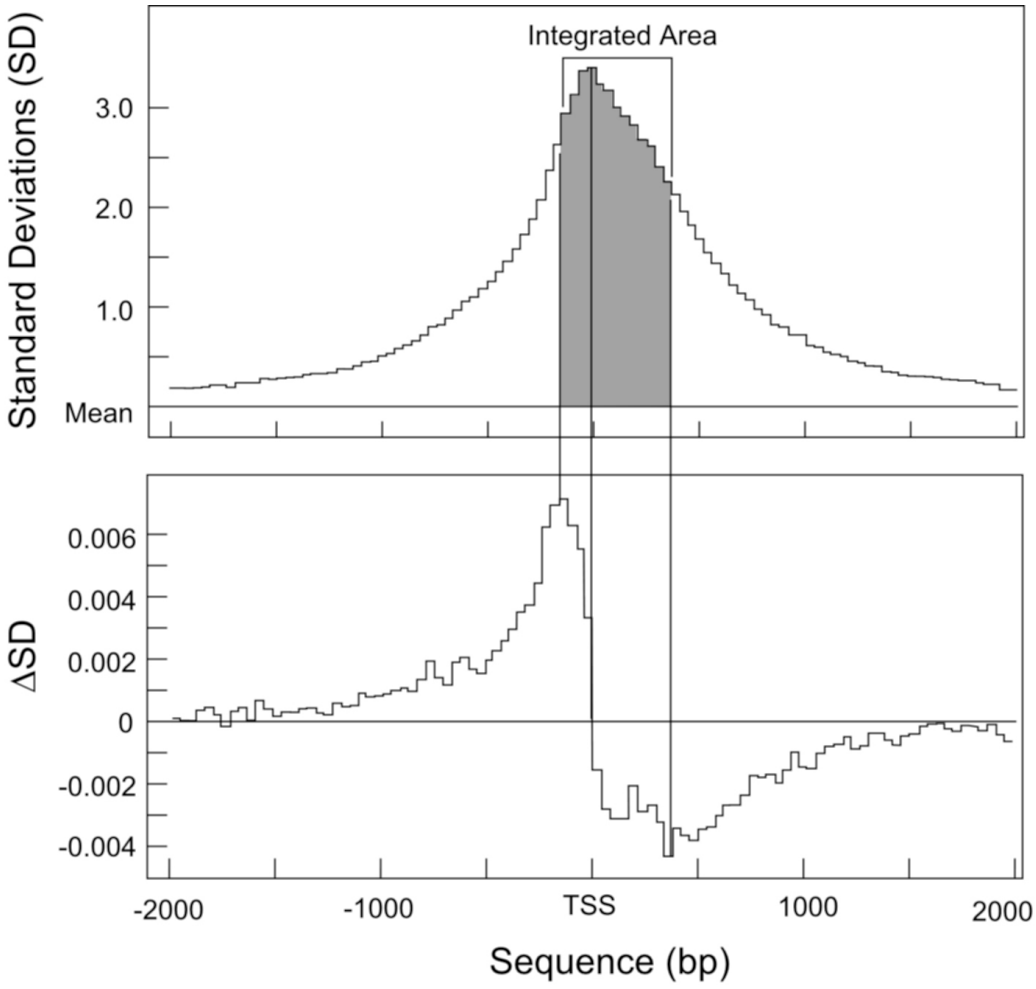
White MF, Lilley DM. (1998) Interaction of the resolving enzyme YDC2 with the four-way DNA junction. *Nucleic Acids Res.* **26**(24), 5609–5616.

Wittig B, Wolfl S, Dorbic T, Vahrson W, Rich A. (1992) Transcription of human c-myc in permeabilized nuclei is associated with formation of Z-DNA in three discrete regions of the gene. *EMBO J.* **11**, 4653-63.

Zamora F, Kunsman M, Sabat M, Lippert B. (1997) Metal-Stabilized Rare Tautomers of Nucleobases. 6. Imino Tautomer of Adenine in a Mixed-Nucleobase Complex of Mercury(II). *Inorg. Chem.* **36**, 1583-1587.

Zhang L, Kasif S, Cantor CR, Broude NE. (2004) GC/AT-content spikes as genomic punctuation marks. *Proc. Natl. Acad. Sci. USA*. **101**(48), 16855-60.

APPENDIX



Appendix. Distribution of CpG content around the transcription start sites (TSS) of human genes. (Upper) The percent of CpG dinucleotides in each 40-bp bin is represented as the number of standard deviations from the mean CpG content of the genome. (Lower) First derivative of the mean CpG content in Upper is numerically calculated as the change in the standard deviation (DSD) between bins. The point where the DSD crosses zero defines the position of the peak in Upper. The positive and negative peaks in DSD define the boundaries for the bins that are summed to define the intensity of the peak in Upper