Structure and Dynamics of Okazaki Fragment Models

by

David E. Konerding

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA SAN FRANCISCO

[redacted]

Date University Librarian

Degree Conferred: ....

Dedication

This thesis is dedicated to Peter Andrew Kollman, whose early passing affects all of us. Peter was not just a accomplished scholar and academician, he was an excellent advisor and companion to his students. Many of the questions posed and answered in this thesis were suggested by Peter, and it is only right that this thesis is dedicated to him.

Preface

"I have made this letter longer than usual, because I lack the time to make it short."

— Blaise Pascal

Every attempt to employ mathematical methods in the study of chemical questions must be considered profoundly irrational and contrary to the spirit of chemistry. If mathematical analysis should ever hold a prominent place in chemistry - an aberration which is happily almost impossible - it would occasion a rapid and widespread degeneration of that science.

— A. Comte, Philosophie Positive, 1838

This dissertation could not have been completed without the assistance of a number of people. First and foremost, my wife Martha, for her dedication and companionship, which helped sustain me through the thesis writing process. Second, I would like to thank various members of the Graduate Group in Biophysics; in particular, Program Administrator Julie Ransom. Without her, I never would have managed to complete all the paperwork required to graduate on time. Third, I would like to thank my thesis advisor, Dr. Thomas L. James, for his support of my research, as well as many stimulating discussions. Fourth, I would like to thank all the other professors, postdoctoral scholars, and graduate students who have advised, aided, and abetted me for the last 6 years.

Chapter 2 of this thesis is a reprint of the material as it appears in Biochemistry (in press). The co-authors, Thomas L. James and William H. Gmeiner directed and supervised the research that forms the basis for chapter 2. David Konerding completed the bulk of effort required to publish this paper, and the chapter is comparable to a standard thesis chapter.

Chapter 4 of this thesis is a reprint of the material as it appears in Journal of Biomolecular NMR, 1999 Feb, 13(2):119-31. The co-authors, Peter A. Kollman and Thomas L. James directed and supervised the research that forms the basis for chapter 4. David Konerding completed the bulk of effort required to publish this paper, and the chapter is comparable to a standard thesis chapter.

Chapter 5 of this thesis is a reprint of the material as it appears in the Proceedings of the Pacific Symposium in Biocomputing, 2000.

Abstract

Structure and Dynamics of Okazaki Fragment Models by David Konerding

Three-dimensional structures of anticancer drugs bound to, and incorporated within, nucleic acids have proved beneficial in rationalizing how these drugs provide their anti-neoplastic functionality. We describe the NMR solution structure of an Okazaki fragment model containing the nucleoside analogs gemcitabine. The analog structures are compared with the reference structure to analyze the structural effect of nucleoside analog incorporation on the Okazaki fragment, and rationalize why these analogs lead to premature chain termination. Recent developments in molecular dynamics force field technology have enabled increasingly realistic simulations of biomolecular structure and dynamics. Increasing computer speeds, coupled with decreasing costs, enables larger systems to be studied in greater detail, with longer sampling intervals, and greater diversity of starting structures. The Okazaki fragment model as well as an analog-containing Okazaki fragment model were submitted to unrestrained molecular dynamics using the Cornell et al. force field, demonstrating the unsuitability of this force field in accurately simulating RNA-containing nucleic acid duplexes. Four simulations for ten nanoseconds each were performed on an all-RNA Okazaki fragment model, using the Cornell et al. and the Wang et al. force fields to determine the cause of the force field's shortcomings.

Th L. James

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 The Chemical Basis of Inheritance

The prehistory of the scientific study of nucleic acids begins with Darwin 1859, who proposed a mechanism for adaptive evolution via random mutation and selection pressure. Shortly after, trait inheritance was carefully quantified in peas by Gregor Mendel in 1866. Real chemical study of nucleic acids begin in 1869 when Johann Frederich Meischer purified "nuclein" from pus cells. Nuclein is the remnants of the cell after cell wall, cytoplasm and nuclear envelope have been extracted. His goal was only to identify the components of the nuclein; he apparently did not make any realization regarding the connection between nuclein and heredity. Miescher determined that nuclein was base soluble, contained anomalously high levels of phosphorous (only a trace element in proteins), and appeared to be of low molecular weight (a side effect due to the harsh extraction conditions used by organic chemists of the day). Later chemists purified nuclein completely free of protein using gentler techniques and accurately determined the ratios of carbon, hydrogen, nitrogen, oxygen and phosphorous in extremely high molecular weight polymers (DNA). By 1880, the stainable material of the nucleus, chromatin was considered the "hereditary particle", and there was clearly recognition of a link between nuclein, protein, and heredity within the nucleus, but the exact relationship was yet to be realized. There was at this point a recognition that long polymers were suited to the storage of heredity information but the identity of the heredity molecule was not yet recognized.

Genetics work by Morgan on flies in the early 20th century verified Mendel's hypothesis and was even capable of determining the relative chromosomal locations of hereditary genetic components. Genetic work rapidly moved from relatively large flies and plants down to some of the simplest living organisms (bacteria) and even primitive pseudo-organisms (viruses) for convenience. Before the middle of the 20th century, solid experimental evidence by Oswald Avery in 1949 [2] implicated DNA (the purified form of nuclein) as the "transforming principle"; when killed lethal virus was mixed with live nonlethal virus and injected into rats, it produced a live lethal virus. This was not a completely convincing argument to some scientists, but by then, most scientists were willing to entertain the hypothesis that the nucleic acid component of viruses contained the heredity function. By 1950, Hershey and Chase [43] produced convincing evidence using radioactively labelled viruses that DNA, not protein, was indeed the transforming principle. At this point it was considered merely a matter of time before the physical nature of molecular heredity was understood. In 1953, Watson and Crick [110], in the now famous paper "A Structure for Deoxyribonucleic Acid", produced a theoretical structure based on fibre diffraction data, molecular modelling, and the relative frequencies of the nucleotide bases. In this proposed structure, two polymers of DNA with complementary sequences formed an antiparallel helix. The structure made good chemical sense, with reasonable bonds lengths, bond angles, dihedrals, and nonbonded contacts. Further, the antiparallel complementary sequence relationship represented a simple, yet systematic method of molecular recognition. Watson and Crick concluded: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." With a plausible physical basis for molecular heredity, immense scientific inquiry begin into the details of the DNA molecule, how it encoded heredity, and the exact mechanism by which cells reproduce.

Enzymology by Kornberg [55] and others rapidly led to cell-free synthesis of complementary DNA strands using polymerase. Exquisite genetic and biochemical experiments led to the elucidation of the mapping between DNA sequences and protein sequences based on a triplet code. The triplet code in itself, a redundant mapping of the 64 unique triplet nucleobase sequences coding for 20 different amino acids, a start signal, and a stop signal. This coding convention is shared between nearly all forms of life, with the only exceptions being mitochondria and a few primitive bacteria.

1.2 Structural Studies of Nucleic Acids

Our knowledge of the structure of nucleic acids in solution has been greatly enhanced by quantitative statistical biology (biophysics). X-ray crystallography and solution phase NMR provide enough direct observations of nonbonded chemical structure to generate precisely determined accurate models of nucleic acid structures. The literature is rich and expanding ever more rapidly with detailed expositions of structures. In regular solvent conditions, the structure of a typical sequence of duplex DNA is close to canonical B-form, while duplex RNA is close to canonical A-form. Within these structural families there is still a great deal of allowed conformational variability. Each sequence exhibits its own characteristic fine structure based on subtle interactions of the individual bases, and some sequences exhibit larger than average intrinsic directed bends, greater flexibility and/or structural plasticity when bound to protein. With advanced structural and thermodynamic methods, this fine structure has been recognized as playing an important role in molecular recognition. For example, DNA-binding proteins have amazing specificity, binding to a tiny family of sequences tightly (nanomolar

K_d) and the rest of sequence space loosely or not at all (greater than micromolar K_d). This selectivity forms an important component in cellular regulation as well as reproduction and is an active topic of inquiry within diverse fields of study.

1.3 Nucleic Acid Analogs As Chemotherapeutics

One of the dominant aspects of modern nucleic acid science is the interest in using nucleic acids, rather than proteins or small-molecule ligands, in the development of effective drugs. One situation in which genome-targeted nucleic acid drugs have been particularly effective is in the treatment of cancer. Normally organismal cells regulate their own growth and replication. However, when the normal safeguards against uncontrolled growth are removed, cells grow rapidly, aggressively acquiring resources at an elevated rate. Many of the most successful cancer therapies target rapidly dividing cells; one approach provides analog nucleotides which are incorporated into the newly synthesized strand but lead to premature termination of replication. This method is not without side-effects, as there are rapidly dividing cells which are non-cancerous (such as epidermal tissue). One analog nucleotide (dFdCTP) is not only incorporated into the freshly synthesized replica of the genome of a dividing cell, but down-regulates the dCTP synthase enzyme. This has the effect of greatly increasing the analog's uptake during the cell replication cycle by depleting levels of the natural nucleoside. Understanding the physicochemical basis for the anticancer effects of these drugs is of critical interest.

1.4 Modern Structural Analysis of Nucleic Acid Chemotherapeutics

Structural analysis is critical to the understanding of how analog nucleotides perform their antineoplastic functionality. Typical antineoplastic analog nucleotides are incorporated into nascently replicated DNA strands during cell replication. The nucleotides interfere with the normal strand replication, although the specific details of the interference differ for each analog. Once an agent is known to be antineoplastic by observation of cell replication failure, closer inspection will be performed to determine the nature of the failure. In the drugs we inspect below (gemcitabine and cytarabine), termination of DNA replication is the typical outcome which leads to cell replication failure. Gemcitabine and cytarabine are incorporated into the newly synthesized DNA strand, and at some later point, the polymerase pauses and/or falls off the strand. It is very challenging to observe the polymerase "in situ" with high enough resolution to gain useful knowledge of why replication terminates prematurely. Instead, a model system is developed which represents the analog nucleotide in the context of newly replicated DNA, and the model system is subjected to various low- and high-resolution structural analysis techniques with the intent of determining the physical cause of replication termination.

Replication of genomic DNA begins at "origin" sites on the genome which contain sequence signals that are substrates for initiation of replication enzymes. Once replication has initiated, the two parental DNA strands are unwound by helicase enzymes, and replication occurs in a semiconservative manner (each strand serves as a template to form a new duplex). Continuous replication of DNA occurs on the "leading" strand by incorporation of complementary nucleotides by DNA polymerase in a 5' → 3' direction. Because no 3' → 5' polymerase exists, replication of the "lagging" strand occurs in discontinuous segments of 5' → 3' synthesis starting at short RNA primers.

Following synthesis, the RNA residues are removed and replaced by DNA residues. These discontinuous segments on the lagging strand are called "Okazaki fragments" and are actually RNA:DNA hybrid duplexes covalently linked to duplex DNA. We have designed an "Okazaki fragment model sequence" which is a minimal representation of the Okazaki fragment as it exists within a replicating genome. The sequence for the fragment model was chosen from the sequence at the origin of replication site of Simian Virus 40 (SV40); the replication cycle of SV40 is well-understood.

5' dC dA dA dA dG dA dT dT dC dC dT dC 3'
3' dG dT dT dT $\boxed{\text{dC}}$ dT dA rA rG rG rA rG 5'

We have determined the structure of the Okazaki fragment model, and using this as a basis, have determined two more structures of the model: with the analog nucleoside gemcitabine and with arabinocytosine. In both cases the analog nucleoside was incorporated at position 20 (the 7th base on the second strand, enclosed in a box in the sequence above). This position follows the RNA primer and represents the second nucleotide to be added by the 5' → 3' polymerase. Both of these drugs act as antineoplastic agents; this function is caused by incorporation of the analog nucleosides into the newly replicated strands, leading to premature replication termination.

1.5 Unrestrained PME Molecular Dynamics Simulations of the Okazaki Fragment Model

The Okazaki fragment model has been submitted to extensive simulation by fully solvated molecular unrestrained dynamics using the Cornell et al. and Wang et al. force fields with particle mesh Ewald electrostatics to determine why these modern force fields are incapable of correctly describing the solution structure of the Okazaki fragment model hybrid duplex.

1.6 Scope of the Following Chapters

The following chapters focus on the motiviation, methodology, and analysis of simulations of nucleic acid duplexes. Chapters 2 and 3 focus on the Okazaki fragment model sequence. Chapter 2 details the solution of an NMR structure of the Okazaki fragment model containing a gemcitabine residue while Chapter 3 describes extensive unrestrained molecular dynamics simulations of the Okazaki fragment model but containing all RNA residues. Chapter 4 describes free and restrained simulations run on decamer and trisdecamer DNA duplexes to determine whether the Cornell et al. force field is capable of reproducing NMR-determined structural features without NMR restraints, and whether structure refinement using PME electrostatics provides any benefits over unsolvated, distance-dependent dielectric electrostatics. Chapters 5 and 6 cover the software and system development which was necessary to produce, process and analyze the results in Chapter 3 conveniently and within a reasonable time period.

Chapter 2

NMR Structure of a Gemcitabine-Substituted Model Okazaki Fragment

The work presented in this chapter is a collaboration between myself, William H. Gmeiner, and Thomas L. James. Eric Trump and Ana Maria Soto performed the experimental data collection and William H. Gmeiner oversaw the project.

Reprinted from Biochemistry with permission. Originally published as: David Konerding, Thomas L. James, Eric Trump, Ana Maria Soto, Luis A. Marky and William H. Gmeiner. "NMR Structure of a Gemcitabine-Substituted Model Okazaki Fragment" Biochemistry (in press).

2.1 Authors

David Konerding[1], Thomas L. James[2], Eric Trump[3], Ana Maria Soto[4], Luis A. Marky[4] and William H. Gmeiner[5]*

1 Graduate Group in Biophysics, University of California at San Francisco, San Francisco, CA

2 Department of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, CA

3 Department of Chemistry, Emporia State University, Emporia, KS

4 Department of Pharmaceutical Sciences, University of Nebraska Medical Center, Omaha, NE 68198

5 Department of Biochemistry, Wake Forest University School of Medicine, Winston-Salem, NC 27157

Supported by NIH NCI CA60612 (WHG), NIH RR01081 and NIH GM39247 (DEK, TLJ) NIH GM42223 (LAM).

*Author to whom correspondence is addressed:

Phone: (336) 716-6216

Fax: (336) 716-7671

bgmeiner@wfubmc.edu

2.2 Abstract

Gemcitabine (2'-deoxy-2',2'-difluorodeoxycytidine; dFdC) is a potent anticancer drug that exerts cytotoxic activity, in part, through incorporation of the nucleotide triphosphate dFdCTP into DNA and perturbations to DNA-mediated processes. The structure of a model Okazaki fragment containing a single dFdC substitution, [GEM], was determined using NMR spectroscopy and restrained molecular dynamics to understand structural distortions that may be induced in replicating DNA resulting from dFdC substitution. The electrostatic surface of [GEM] was also computed to determine how the geminal difluoro group of dFdC perturbs DNA electrostatics. The stability of [GEM] was investigated using temperature-dependent UV spectroscopy. dFdC adopted a C3'-endo conformation in [GEM], and decreased the melting temperature of the duplex by 4.3 °C. dFdC substitution did not decrease helical stacking among adjacent purines in the DNA duplex region. dFdC substitution substantially altered the electrostatic properties of the model Okazaki fragment, with increased electron density in the vicinity of the geminal difluoro group. The results are consistent with dFdC substitution altering the structural, electrostatic and thermodynamic properties of DNA and interfering in DNA-mediated processes. Interference in DNA-mediated processes due to dFdC substitution likely contributes to the anticancer activity of dFdC.

2.3 Introduction

Gemcitabine (2'-deoxy-2',2'-difluorodeoxycytidine; dFdC) is one of the most widely used and efficacious anticancer drugs in current use. dFdC, either as a single agent or in combination chemotherapy, has demonstrated efficacy towards several of the most prevalent human cancers, including breast [96], ovarian [37] and lung cancers [50]. dFdC is also one of the most widely used anticancer drugs for treatment of the most devastating human cancers for which no adequate treatment options are available. For example, dFdC is used in the treatment of pancreatic cancer [49] and in the treatment of mesothelioma [27]. dFdC is active as a single agent [97], and also has been shown to possess palliative effects in cancer patients [40]. dFdC is most frequently used, however, in combination chemotherapy regimens [75]. dFdC is frequently combined effectively with cisplatin [52, 23], and the extent of platinum:DNA adduct formation positively correlates with the level of incorporation of dFdC into DNA [70]. dFdC is also used effectively in combination with topotecan [37], a potent inhibitor of DNA topoisomerase I (top1). The effectiveness of chemotherapeutic combinations combining dFdC with a drug that damages DNA, or interferes with DNA-mediated processes, suggests that the anticancer activity of dFdC is substantially due to the misincorporation of dFdCTP into DNA. In this paper, the effects of dFdC substitution on the structure, stability and electrostatics of a model Okazaki fragment are described (Figure 2.1).

dFdC is a deoxycytidine (dC) analog that is a member of the antimetabolite class of anticancer drugs [75]. dFdC requires metabolic activation to the nucleotide diphosphate (dFdCDP) and triphosphate (dFdCTP) forms to exert its cytotoxic effects [77, 41]. The cellular targets of dFdCDP and dFdCTP include DNA polymerases, ribonucleotide reductase and CTP synthase. Biochemical studies demonstrate that dFdCTP is incorporated predominantly into DNA, with little incorporation into RNA [76]. dFdCTP is a poor substrate for human DNA polymerases, however, with an efficiency of incorporation only 5% that of dCTP [48]. A strong correlation has been shown between incorporation of dFdCTP into DNA and the loss of viability of cells, as determined using a clonogenic assay [76]. Once dFdCTP is incorporated into DNA, the polymerase adds one additional nucleotide after which the polymerase pauses [48]. The effects of dFdCTP incorporation into DNA on further chain elongation are complex, and not currently understood in physicochemical detail. dFdC is found predominantly in internucleotide linkages in DNA [48]. dFdC is also a potent sensitizer to radiation, an activity that requires metabolism to dFdCTP and incorporation into DNA [62].

In addition to the DNA-directed effects of dFdC, dFdC also extensively modulates CTP and dCTP metabolism [42]. dFdC reduces cellular concentrations of CTP and dCTP 5% and 50%, respectively, apparently by blocking the activity of CTP synthase. Cellular depletion of CTP and dCTP inhibits DNA replication and RNA synthesis, respectively in cells exposed to dFdC. Reduction in dCTP pools also enhances the misincorporation of dFdCTP into DNA by increasing the dFdCTP:dCTP ratio. In this respect, the activity of the drug is self-potentiating with the non-DNA directed effects serving to enhance misincorporation of dFdCTP into DNA, thus enhancing the DNA-mediated effects of the drug [75]. dFdCTP also inhibits dCMP deaminase leading to decreased catabolism of dFdC [77]. This unique, self-potentiating activity of dFdC is thought to be responsible, in part, for the superior antitumor activity of dFdC relative to other nucleoside analogs used for the treatment of cancer.

Perturbations to DNA structure appear to play a significant role in the anticancer activity of nucleoside analogs such as arabinofuranosyl cytosine (cytarabine) and 5-fluorouracil (5-FU). For example, cytarabine, a dC analog that has potent anti-leukemic activity, has been shown to interfere with top1-mediated cleavage of duplex DNA [78]. Cytarabine misincorporation has also been shown to inhibit lagging strand DNA replication, and to cause DNA strand breaks to occur at sites distal relative to the site of misincorporation [58, 87]. The deoxythymidine triphosphate (dTTP) analog FdUTP, a metabolite of the widely used anticancer drug 5-FU, may be misincorporated into DNA, and the substituted DNA is a substrate for the DNA mismatch repair machinery, suggesting fluorpyrimidine substitution perturbs DNA structure [68]. Our laboratory has investigated the structural consequences of nucleotide analog substitution in DNA using NMR spectroscopy [34]. We recently investigated the structural basis for alterations in lagging strand replication due to cytarabine misincorporation by determining the NMR structure of a cytarabine-substituted model Okazaki fragment [36, 35]. These studies revealed that cytarabine substitution into the DNA duplex region of the model Okazaki fragment resulted in local distortion of the DNA duplex and increased global curvature of the model Okazaki fragment. We have also shown that FdU substitution perturbs the structure and stability of duplex DNA [90, 91]. These results are consistent with direct perturbations to DNA duplex structure arising from nucleotide analog misincorporation as being responsible for the DNA-mediated effects of the anticancer drugs cytarabine and 5-FU.

The anticancer activity of dFdC, like cytarabine and 5-FU, may result, in part, from perturbation of DNA-mediated processes that occur due to changes in the structure of the DNA as a consequence of dFdCTP misincorporation. The effects of dFdCTP misincorporation to the structure of duplex DNA have not been as extensively studied as for cytarbine and 5-FU, anticancer drugs that have been in clinical use for considerably longer. The geminal difluoro group of dFdC likely perturbs the electrostatic surface of duplex DNA into which dFdCTP has been misincorporated in a non-sequence selective manner. These electrostatic effects might alter cognate protein binding, and may be a significant factor in the altered kinetics of dFdCTP misincorporation (relative to dCTP), and for the pause in polymerase activity following dFdCTP misincorporation [48]. Alterations to the structure of duplex DNA due to dFdCTP misincorporation likely also contribute to perturbations in DNA processing. At present, few structural data are available concerning the effects of dFdC substitution on DNA structure. We have determined the NMR structure of a dFdC-substituted model Okazaki fragment ([GEM]; Figure 2.1) in order to evaluate the perturbation to duplex DNA structure resulting from dFdC substitution. The model Okazaki fragment studied had the same sequence as that previously used to investigate the effects of cytarbine misincorporation to DNA structure [36, 35]. Use of the same sequence facilitates direct comparison between the structural perturbations caused by the two dC analogs. The conformation of dFdC in [GEM] is markedly different from either cytarabine or dC in the same model Okazaki fragment sequence. Unlike cytarabine, however, the distorted nucleoside structure of dFdC is accommodated well into the same global solution structure as the native model Okazaki fragment. The results are consistent with perturbations to DNA-mediating processes due to dFdCTP misincorporation resulting both from altered local structure to the DNA duplex and changes in the electronic configuration of the nucleotide.

2.4 Materials and Methods

2.4.1 Preparation of Gemcitabine Phosphoramidite

The phosphoramidite of dFdC (5'-O-(4,4'-dimethoxytrityl)-3'-O-(2-cyanoethyl-N,N-diisopropyl-phosphoramidite-N4-benzoyl-2'-deoxy-2,2'-difluorocytidine) was synthesized in a manner similar to that previously described [85]. All reactions were carried out under argon in flame-dried glassware. Pyridine was distilled over calcium hydride and under argon prior to use. Gemcitabine was a gift (to W.H.G) from Eli Lilly. All other reagents were obtained from Aldrich. The intermediate amino-protected compound N4-benzoyl-2'-deoxy-2',2'-difluorocytidine, was prepared from the hydrochloride salt of gemcitabine by reaction with benzoyl chloride in pyridine with transient protection of the deoxyribose hydroxyls with trimethylsilyl chloride. The tritylated intermediate (5'-O-(4,4'-dimethoxytrityl)-N4-benzoyl-2'-deoxy-2',2'-difluorocytidine) was prepared by subsequent reaction with 4,4'dimethoxytritylchloride using standard methods. Gemcitabine phosphoramidite was then prepared by reaction with 2-cyanoethyl-N,N'-diisopropylchlorophosphoramidite in anhydrous tetrahyrofuran with diisopropylethylamine as base. The resulting 5'-O-dimethoxytrityl-3'-O-phosphoramidite was purified by column chromatography, and incorporated in place of one deoxycytidine (dFdC20; see Figure 2.1) during the chemical synthesis of the RNA/DNA hybrid strand of [GEM].

The oligonucleotides for the model Okazaki fragment containing dFdC were synthesized using an ABI 394 synthesizer in the Molecular Biology Core Laboratory at the Eppley/UNMC Cancer Center. The sequences for the DNA strand and the hybrid DNA/RNA strand, which included the site of dFdC substitution, are shown in Figure 2.1. The sequences selected were identical, other than the site of dFdC substitution, to the model Okazaki sequences previously investigated by our laboratory [36, 35]. Oligonucleotide synthesis was similar to that previously described for the cytarbine-substituted model Okazaki fragment except 4,5-dicyanoimidazole was used as activator and TOM-protected RNA amidites were used with a 6 min coupling time. Deprotection of RNA and DNA was accomplished using 2 mL of 40% aqueous methylamine and 2 mL of 33% ethanolic methylamine for each 10 μmol scale synthesis, followed by incubation at room temperature for 6-8 h. The solution was evaporated to dryness, and 4 mL of tetrabutylammonium fluoride were added and the mixture was heated to 50 oC until the oligonucleotide dissolved followed by slow cooling. The oligonucleotides were HPLC purified using a Waters DeltaPrep 4000 with a Hamilton PRP-1 polystyrene semiprep column. A 30 min linear gradient with initial conditions 96% buffer A (0.1M TEAA) 4% buffer B (80% aqueous acetonitrile) and final conditions 80% A, 20% B, was used for oligonucleotide purification.

2.4.2 Temperature-Dependent UV Spectroscopy

Absorbance versus temperature profiles (melting curves) were measured with a thermoelectrically controlled Aviv 14-DS spectrophotometer. The absorbance was monitored at 260 nm and the temperature was scanned at a heating rate of 0.5 °C/min. Analysis of the resulting melting curves, using an all or none approximation, allowed measurement of transition temperatures, T_m, which are the midpoint temperatures of the helix-coil transition of the duplex and van't Hoff enthalpies, ΔH_{vh}. The ΔH_{vh}'s were obtained in two different ways, from shape analysis of the melting curves and from the T_m-dependence on strand concentration, using the following equation: $1/T_m = R/\Delta H_{vh} ln(C_T/4) + \Delta S/\Delta H_{vh}$, where C_T is the total concentration of strands and R is the gas constant [67]. The concentration of each oligomer was determined at 260 nm and 80 °C, using extinction coefficients of 120.6 mM^{-1} cm^{-1} and 119.3 mM^{-1} cm^{-1}, for [OKA] and [GEM], respectively. These values were calculated by extrapolating the tabulated values of dimer and monomer bases at 25 °C[11] to high temperatures, using procedures reported earlier [66]. ΔS and ΔH were calculated from the latter equation while (G was calculated from the Gibbs equation: $\Delta G = \Delta H - T\Delta S$. All experiments were conducted in buffers containing 10 mM Sodium Cacodylate, 100 mM Sodium Chloride, adjusted to pH 7.

2.4.3 NMR Spectroscopy

NMR experiments were performed using a Varian UNITY 500 NMR spectrometer at the Eppley/UNMC Cancer Center. Samples for NMR analysis were prepared by mixing equimolar amounts of the DNA strand and the hybrid DNA/RNA strand. Sample concentration was approximately 1 mM duplex in 600 μL of 2 mM sodium cacodylate (pH 7.3), 100 mM NaCl, and 0.2 mM disodium EDTA. ^{1}H spectra were acquired using a ^{1}H{^{13}C,^{15}N} 5 mm PFG probe (Varian, Inc.), and were referenced to HDO at 4.76 ppm at 26 °C. 1D ^{1}H NMR spectra in 95% H_2O (5% D_2O) at 26, 29, 32, 35 and 38 °C were obtained using a 1-3-3-1 binomial pulse for water suppression [44]. 2D NOESY spectra in 95% H_2O solution were acquired using a 1-1 echo pulse sequence for water suppression. All spectra in H_2O were acquired with an 11 kHz spectral window centered about the ^{1}HDO resonance. NOESY spectra in D_2O were acquired for mixing times of 100, 150 and 200 ms using the standard three-pulse sequence using States' method of phase cycling for pure absorption spectra [101]. Four hundred free induction decays, 16 scans each, with alternating block acquisition, were collected in the t1 dimension. 2048 data points over a 5000 Hz spectral window were collected in the t2 dimension with the carrier frequency set at the ^{1}HDO resonance. A relaxation delay of 8 s was included between scans to allow sufficient relaxation for crosspeak quantitation. TOCSY spectra were acquired with 60 and 100 ms mix times using parameters similar to those used for described for NOESY spectra, except the relaxation delay was shortened to 3 s. ECOSY spectra were collected using 32 scans per increment with a 3 s relaxation delay and 4096 points in t2 [38]. All data were initially processed with VNMR v. 5.3B from Varian and then imported into SPARKY (UCSF) for analysis. The spectra were apodized using shifted Gaussian filter functions. After zero-filling in the t1 dimension, the final matrices were 2048 x 2048 points, except for ECOSY spectra that were 4096 x 2048 points. T1 and T2 data were obtained from inversion/recovery and CPMG experiments, respectively, and were fit to single-exponential functions using VNMR 5.3B. 1D 19F and 2D ^{1}H-19F HOESY spectra were acquired using a Nalorac 5 mm ^{1}H/19F probe. ^{1}H-19F HOESY spectra were acquired with using 5000 Hz spectral windows for both ^{1}H and 19F with 1024 points in t2 and 256 points in t1 [102].

2.4.4 Experimental Constraints

Constraints on interproton distances were determined from NOESY data sets in D_2O solution. Interproton distances were calculated from NOESY crosspeak intensities using MARDIGRAS [7]. Volume integrals were evaluated from NOESY experiments acquired with 100, 150 and 200 ms mix times, respectively. NOESY intensities evaluated for crosspeaks on both sides of the diagonal were averaged. A complete relaxation matrix was created for [GEM] using intensities evaluated experimentally for 364 interproton interactions, and estimated intensities from the geometry of the starting structure for those interproton interactions that could not be evaluated from the experimental data. The diagonal and off-diagonal terms were compared iteratively until the sum of the residual errors was minimized. MARDIGRAS calculations for [GEM] were carried out with three experimental data sets (100, 150, 200 ms), two geometries for the initial structure (A- and B-form double helices), and three values for the isotropic correlation time (τ_c = 1.1, 1.4 and 1.7 ns). Estimates of interproton distances associated with NOE crosspeak intensities resulting from each of the 18 MARDIGRAS calculations for each duplex were averaged. The average distance and standard deviation were then used to set the flat portion of the potential well for each distance constraint. Estimates of $^3J_{HH}$ from ECOSY spectra for the deoxyribonucleotides in [GEM] were the same, within experimental uncertainty, to those measured previously for [OKA]. Thus the torsion angle restraints used previously in the refinement of [OKA] were used in the refinement of [GEM].

2.4.5 Molecular Refinement

Molecular graphics images were produced using the Chimera package from the Computer Graphics Laboratory, University of California, San Francisco [47]. The electrostatic charge parameters for the GEM residue were determined as follows. The GEM residue was built as a model starting from dC with the LEaP module of AMBER [13]. The H2'1 and H2'2 atoms were replaced with F2'1 and F2'2 atoms, and the resulting model was input into Gaussian for optimization and electrostatic potential calculation [29]. The esp data from Gaussian were used to generate point charges for the entire residue using the program ANTECHAMBER [109]. The point charges were very similar to the dC residue. Fluorine charges were treated as equivalent during the RESP fit, and were calculated to be -0.2403 (H2'1 and H2'2 are 0.0718). The GEM residue was built with LEaP and then minimized using the fluorine charges and the fluorine bond parameters (F-CT-CT, 50.0 kcal/mol bond constant, optimal angle 109°).

The structure refinement of [GEM] (Figure 2.1) was performed as follows. Two initial models of [GEM] were created: one in canonical A and the other in canonical B form. The initial models were charge-neutralized with hydrated sodium ions [99] using the LEaP module of the AMBER molecular modeling suite. LEaP places the ions sequentially in the electrostatic energy minimum, until the charge of the system is neutral; 22 ions are necessary for a duplex composed of 24 residues in total. Following ion placement, the ions were subjected to 100 steps of steepest descent minimization and 400 steps of conjugate gradient minimization with the duplex harmonically restrained to its initial coordinates by a force of 10 kcal/mole. Following ion minimization, the neutralized duplex was submitted to 1000 steps of steepest descent and 4000 steps of conjugate gradient minimization. At this point, the initial model structures were considered equilibrated with respect to ion placement, and production refinement commenced. Independently, the A and B models were subjected to 30ps of simulated annealing. The SHAKE algorithm was applied to maintain idealized hydrogen bond and angle values [89]. The parameter schedule applied during the simulated annealing process is included in the Supplementary Information.

To increase conformational sampling, three runs per model structure were performed, with varying random number seeds. Following the simulated annealing, the final 5ps of each of the 6 runs were averaged to form 6 average structures. Each of the 6 average structures was minimized for 200 steps of steepest descent and 800 steps of conjugate gradient minimization, with no restraints applied. Following minimization, all three minimized A structures were averaged together and all three B structures were averaged together, and the average structures were minimized without restraints for 100 steps of steepest descent and 400 steps of conjugate gradient minimization. Following minimization, the two structures were averaged together and minimized further, without restraints, for 500 steps of steepest descent and 2000 steps of conjugate gradient minimization. Following minimization, the average structure was refined with 20 ps of restrained molecular dynamics at room temperature. The parameter schedule used is included in supplementary information. The final 5ps of the room temperature rMD were averaged, minimized for 10 steps to repair averaging artifacts, and submitted to a final restrained minimization of 1000 steps of steepest descent and 4000 steps of conjugate gradient minimization. The final averaged, minimized structure was submitted to structure analysis using the DIALS and WINDOWS application which computes CURVES parameters [81, 59]. Intermediate and final structures were analyzed for compliance with the experimental constraints using CORMA [6]. The sixth root compliance values were 0.0702, 0.0585 and 0.0576 for the NOE crosspeak intensities from the 100, 150 and 200 ms NOESY experiments, reflecting excellent compliance of the final structure with the experimental data.

2.4.6 Electrostatics Computations

The electrostatic contribution of the dC20 and dFdC20 residues were computed using Delphi [72]. The calculation included only the partial charges from the dC20 and dFdC20 residues, and no other charges from the molecules. Standard atomic radii were used for all charged atoms. The Delphi parameters used were: 100 iterations of the QDIFFX routine. The dielectric constant of the helix interior was 1 and the dielectric constant of the helix exterior was 80. The surface was calculated using MSMS [94], and the electrostatic potential was mapped to the surface using trilinear interpolation. The result was visualized using the "DelphiViewer" extension of Chimera (UCSF).

2.5 Results

2.5.1 NMR Analysis of Exchangeable ^{1}H

The exchangeable ^{1}H resonances of duplex DNA, RNA, or hybrid duplexes are informative concerning the formation and stability of specific base pairs in the duplex. Assignments for the imino and amino ^{1}H resonances of the G-C and A-T base pairs of [GEM] were obtained from analyses of 2D NOESY spectra in 95% H_2O solution (Supplementary Information). The chemical shift assignments for the exchangeable ^{1}H resonances of [GEM], summarized in Table 2.1, are similar to the corresponding position of [OKA], the model Okazaki fragment consisting of all native nucleotides [36]. In particular, the imino resonance for G5, the base pairing partner of dFdC20 in [GEM] is shifted only slightly relative to the chemical shift of G5 base paired to dC in [OKA] (12.68 vs 12.55 ppm; 24). Changes in chemical shift for other exchangeable ^{1}H resonances of [GEM], relative to [OKA], were of similar magnitude, suggesting the duplexes had similar overall structure. The intensities for T21 and T22 imino ^{1}H resonances of [GEM] were reduced, however, relative to the intensities observed for these resonances in [OKA] at identical temperatures, consistent with the DNA duplex region

(DDR) of [GEM] undergoing more frequent base pair opening than occurs for [OKA]. These results are consistent with thermodynamic measurements that reveal decreased global stability for [GEM] relative to [OKA] (Table 2.2).

2.5.2 Thermodynamic Measurements

The effects of dFdC substitution on the stability of the model Okazaki fragment [GEM] were investigated using temperature-dependent UV spectroscopy. The results are summarized in Table 2.2. The T_m for the dFdC-substituted duplex was 41.2 °C at a concentration of 13.3 x 10^{-6} M, a value 4.3 °C lower than for [OKA], the model Okazaki fragment consisting of only native nucleotides. The decreased stability caused by dFdC substitution was similar to that caused by cytarabine substitution [36]. The free energies for formation of each duplex were the same, within experimental error (-14 kcal/mol). The relative enthalpic and entropic contributions to the free energy differed between the duplexes, however. The relative stabilizing contribution from the enthalpic term was greater for [GEM] than [OKA] as measured from the concentration dependence of the T_m, but slightly greater for [OKA] than [GEM] measured from the shape of the melting curve (van't Hoff enthalpy; Table 2.2). It is worth noticing that the enthalpic contributions for [OKA] obtained from the concentration dependence of T_m and from the shape of the melting curves are in good agreement, suggesting that this transition occured in an all-or-none fashion. In the case of [GEM], the values obtained from these two approaches differ significantly from each other, which may indicate that the transition takes place through intermediate steps. The presence of a non two-state transition makes it difficult to elucidate the source of the destabilization caused by dFdC-substitution. However, the relative destabilizing contribution from the entropic term measured from the concentration dependence of the T_m was larger for [GEM] than [OKA]. This indicates that the substitution of dFdC decreases duplex stability of the model Okazaki fragment, inducing better base-pair stacking interactions, and at the same time causing a local ordering of ions and water molecules at the substitution site. This is consistent with the observations in the solution structure of [GEM].

2.5.3 NMR of Non-Exchangeable ^{1}H

The assignments for the non-exchangeable ^{1}H resonances of [GEM] were made using a similar approach to that described previously for the assignment of similar resonances for [OKA] [35]. In particular, the ^{1}H resonance assignments for the each of the 18 deoxyribose sugars were made using TOCSY and COSY experiments (Supplementary Information). Stereochemical assignments for H2'/2" were made on the basis of the relative intensities of the H1'-H2'/H2" crosspeaks for NOESY spectra acquired with 100 ms mix time. H2" resonated downfield of H2' for all deoxyribose sugars except G5. Sequential connectivity between the deoxyribose spin systems was established by the observation of (n)H8/H6-(n)H1' and (n)H8/H6-(n-1)H1' and/or (n)H8/H6-(n)H2'/H2" and (n)H8/H6-(n-1)H2'/H2" crosspeaks in the NOESY spectra. The ribose spin systems were assigned on the basis of (n)H1'-(n)H2' and (n)-H8/H6-(n-1)H2' crosspeaks in NOESY spectra and H3'-H4' crosspeaks in TOCSY and COSY spectra. Sequential connectivities for both types of nucleotides were also apparent from (n-1)H8/H6-(n)H5/M crosspeaks in NOESY spectra. Assignment of adenosine H2 resonances were made based on analysis of 2D NOESY data in both H_2O and D_2O solution. Each of the adenosine H2 resonances showed NOE crosspeaks either to the H1' resonance of the 3' neighbor (intrastrand) or to the H2 resonance of an adjacent adenosine, as well as to the imino ^{1}H of the base pairing partner dT.

2.5.4 Structure of dFdC20

The NMR assignments for dFdC20 were established based on NOESY crosspeaks between T21 H6 and dFdC20 H1', dFdC20 H6 and T19 H1', T19 H6 and dFdC20 H5, and T21 M to dFdC20 H6 (Supplementary Information). The assignment of the geminal difluoro 19F resonances of dFdC20 and additional interesidue assignments between dFdC20 and adjacent nucleotides were made from 2D HOESY spectra (Figure 2.2). The strongest HOESY crosspeaks were observed between the geminal difluoro resonances and dFdC20 H1'. Additional ^{1}H-19F HOESY crosspeaks were also observed between dFdC20 H6, H3' and H4' and the geminal difluoro resonances of dFdC20 and between T19 M and the geminal difluoro group suggesting efficient stacking of T19 and dFdC20. The HOE crosspeaks between the geminal difluoro group and dFdC20 H4' was slightly more intense than the corresponding crosspeak to dFdC20 H3', a result consistent with a short interatomic F2" to H4' distance and with a small pseudorotation angle for the dFdC sugar ($P < 40\,^\circ C$). The HOESY information was used only qualitatively in making resonance assignments and was not used in determining distance or angular restraints that were used in the restrained molecular dynamics refinement procedure. Refinement of the structure using the ^{1}H NMR data in conjunction with restrained molecular dynamics using the AMBER forcefield resulted in dFdC20 adopting an A-form sugar pucker.

Substitution of dFdC for dC in duplex DNA, or a model Okazaki fragment, is expected to alter the biophysical properties of the nucleic acid both in terms of the structure and the electrostatic potential in the vicinity of the site of substitution. The altered structural properties of the nucleic acid are expected to arise as a consequence of dFdC adopting an alternative conformation relative to the C2'-endo sugar pucker ordinarily adopted by dC in duplex DNA. Analysis of the final refined structure for [GEM] reveals that dFdC does adopt an atypical C3'-endo conformation (Figure 2.1). In addition to altering the structure of [GEM], the geminal difluoro group of dFdC substantially alters the electrostatic surface at the site of substitution. The electrostatic surface of dFdC, calculated using Delphi, is shown in Figure 2.3. The corresponding surface of dC is also shown for comparison in Figure 2.3. Fluorine substitution slightly increases the dimensions of the van der Waals surface of dFdC, relative to dC, however, the electronegative fluorine atoms result in a substantially altered electrostatic potential over this surface (Figure 2.3). The altered electrostatic potential will, in turn, affect the composition of electrolytes in proximity to the site of substitution, and potentially affect the affinity for dFdC-substituted DNA both or small molecules and cognate proteins.

2.5.5 Structure of [GEM]

The structure of [GEM] resulting from restrained molecular dynamics refinement based on the experimental NMR data is shown in Figure 2.4. Also shown in Figure 2.4 is the structure of [OKA], the model Okazaki fragment consisting of all native nucleosides [35]. The overall RMSD between the structures is 2.18 A; thus the overall similarity of the dFdC-substituted model Okazaki fragment is substantially more similar to [OKA] than was the case for the cytarabine-substituted model Okazaki fragment which had an overall RMSD of 4.1 A [35]. The structure of [GEM] was regionally similar to that previously reported for [OKA]. The structure consisted of three regions: a duplex DNA region (DDR; nucleotides 1-6 and 18-24), a hybrid duplex region (HDR; nucleotides 9-12 and 13-16) and a junction region (JR; nucleotides 7,8,17,and 18) [35]. The most substantial differences in structure between [GEM] and [OKA] occurred at the site of substitution (dFdC20) and in the JR.

The structure of the DDR of [GEM] was highly similar to that of the DDR of [OKA] with a pairwise regional RMSD of 1.26 A. The structure of the DDR of [GEM] is shown in Figure 2.5. The largest difference in this region between the two duplexes occurred at dFdC20 which had a sugar pucker of 20° compared to 130° for dC20 in [OKA]. The value of χ was also anomalously low for dFdC20 (205°) relative to dC20 in [OKA] (230°), but this is reasonable since chi and pucker are generally correlated.. Plots of all torsion angles for the entire duplex are included in the Supplementary Information. These altered values for torsion angles in dFdC20, relative to dC20, were accommodated into the structure of the DDR with relatively minor changes to the conformations of other nucleotides. In particular, base stacking among adjacent purines in the DDR of [GEM] was not disrupted by the substitution of dFdC20. The DDR of [GEM] contains three consecutive A-T base pairs with the three adenosines consecutive in the DNA strand. These three adenosines (A2, A3 and A4) were efficiently base stacked in [OKA], as was apparent from NOE crosspeaks between H2 of A3 and H2 of both A4 and A2 [35]. This efficient stacking of the purines was disrupted in the cytarabine-substituted model Okazaki fragment [ARAC], and the putative NOE crosspeak between A2 H2 and A3 H2 was not observed in the [ARAC] model Okazaki fragment [35]. Efficient base stacking was apparent based on observation of NOE crosspeaks for A3 H2 of [GEM] with both A2 H2 and A4 H2. Thus, the local conformation of dFdC20 did not disrupt helical stacking of contiguous purines in the DDR. In fact, dFdC20 substitution promoted helical stacking in the DDR as was evident by observation of an NOE crosspeak between A3 H8 and A2 H8 in [GEM] that was absent in [OKA] (supplementary information). Efficient base stacking among A2, A3 and A4 was observed in the NMR structure of [GEM], in accordance with the experimental observations (Figure 2.6).

Although base stacking and overall geometry were similar for the DDR of [GEM] and [OKA], significant differences in stacking geometry between adjacent adenosines occurred in the junction region (JR) of [GEM] compared to the corresponding region of [OKA]. In particular, the NOE crosspeak observed between A18 H2 and A6 H2 in both [OKA] and [ARAC] was absent in [GEM], and the H2 resonances of A18 and A17 of [GEM] differed considerably in chemical shift from the corresponding values in both [ARAC] and [OKA]. The structural differences in the JR of [GEM] relative to [OKA] may contribute to the observed destabilization of the duplex, with the decrease in melting point due to dFdC20 substitution being slightly greater than was observed for the cytarabine substituted model Okazaki fragment [ARAC], relative to [OKA] (2.2; [36]). Interestingly, although both dFdC and cytarabine substitution resulted in subtle changes in duplex structure near the site of substitution, they affected different regions of the duplex to a greater extent. Cytarabine, which actually adopted a more B-form sugar pucker in [ARAC] than was observed for dC in [OKA], affected base stacking in the DDR. dFdC, which adopted a more A-form sugar pucker in [GEM] than was observed for dC in [OKA], mainly affected geometry and base stacking in the JR of the model Okazaki fragment.

2.5.6 Electrostatics of [GEM]

The effects of nucleoside analog substitution to the biophysical properties of the nucleic acid result from both conformational effects and changes in the electrostatic surface. As shown in Figure 2.3, dFdC presents a much more highly electronegative surface than dC as a consequence of the highly electronegative geminal difluoro group. The electrostatic surface of the entire model Okazaki fragment is altered near the site of dFdC substitution as shown in Figure 2.4. While the phosphate groups and electronegative atoms that line the minor groove of duplex nucleic acids create a highly electronegative surface for [OKA], the substitution of dFdC20 in [GEM] substantially increases the electronegativity of the surface. Thus, while the inherent malleability of nucleic acids might mute the relatively subtle structural differences arising from dFdC substitution in vivo, the substantial differences in electronegativity arising from dFdC substitution may be important in protein binding and other important nucleic acid-mediated processes.

2.6 Discussion

Clinical experience with dFdC continues to reveal that this dC analog is an extremely potent anticancer drug with activity towards diverse tumor types [75]. Importantly, combination chemotherapy regimens including dFdC have been developed that have demonstrated activity for the treatment of pancreatic cancer [73], non-small cell lung cancer [57], malignant pleural mesothelioma [10], and other malignancies for which no adequate chemotherapeutic regimens have previously been developed [1]. Although combination chemotherapy with dFdC and other agents is rarely curative against these devastating malignancies, consistent observation of partial responses raises the possibility that modified chemotherapeutic regimens including dFdC will result in the predictable occurrence of complete responses to treatment. In order to increase the efficacy of chemotherapeutic regimens including dFdC, full elucidation of the mechanistic basis for the observed chemotherapeutic activity is required. In the present paper, we have described the effects of dFdC substitution to the structure, stability and electrostatics of a model Okazaki fragment. These results enhance our understanding of the physicochemical basis for the alteration of DNA-mediated properties due to dFdC substitution that contribute to the anticancer activity of this drug.

Although the mechanism of antitumor activity of dFdC is complex, the nucleoside must be phosphorylated in order to disrupt critical biochemical processes in tumor cells. In this regard, a novel polymeric form of dFdCMP has been developed by our laboratory that circumvents the requirement for the initial phosphorylation step by deoxycytidine kinase [56]. The activity of dFdC is most strongly correlated with the extent of intracellular dFdCTP formation and with the subsequent misincorporation of dFdCTP into DNA. Misincorporation of dFdCTP into DNA disrupts DNA synthesis and DNA-mediated processes (52-54). dFdCTP is a poor substrate for human DNA polymerases, however, with an efficiency of incorporation only 5% that of dCTP [48]. The physicochemical basis for the poor DNA polymerase substrate properties of dFdCTP have not been elucidated, however, the present studies indicate substantial differences between dFdC and dC with regards to the preferred nucleoside conformation, the relative charge distribution of the nucleoside, and the relative size of the nucleoside. Although the present studies do not permit delineation of which of these factors are most responsible for the poor substrate properties of dFdCTP, or for the pausing of the polymerase following dFdC misincorporation, they demonstrate features of the DNA polymerase:dFdCTP complex, and of the nascent DNA following dFdCTP misincorporation, that likely contribute to reduced rates of dFdCTP misincorporation into DNA and for DNA polymerase pausing. For example, the NMR structure of [GEM] revealed that dFdC adopted a C3'-endo sugar pucker rather than the C2'-endo sugar pucker characteristic of dC in B-form DNA. Thus, presuming dFdCTP also preferred a C3'-endo sugar pucker, an energetic penalty would be associated with adoption of a C2'-endo sugar pucker by dFdCTP in the DNA polymerase:dFdCTP complex to permit appropriate stereochemical alignment of reactive groups to allow phosphodiester bond formation to proceed. Likewise, assuming dFdC also adopted a C3'-endo sugar pucker as the terminal nucleotide in the nascent DNA, an energetic penalty would be assessed for conformational re-arrangement to a C2'-endo sugar pucker to permit proper alignment of reactive groups for phosphodiester bond formation for addition of the nucleotide following dFdCTP misincorporation.

In addition to altered conformational preferences affecting the rate of dFdCTP misincorporation and polymerase pausing, the electron-withdrawing geminal difluoro group significantly alters the electrostatic surface of the nucleoside. The altered electrostatic surface for dFdC, relative to dC, likely contributes to the reduced rate of misincorporation of dFdCTP into DNA and polymerase pausing. Altered electrostatics may affect complex formation between dFdCTP and DNA polymerase, possibly affecting the rate of dFdCTP misincorporation. Altered electrostatics may also effect adoption of the preferred orientation for the complex between nascent DNA and DNA polymerase. The electron withdrawing effects of fluorine also decrease the electron density of the triphosphate and 3'-hydroxyl groups of dFdCTP, thus reducing the reactivity of these sites during phosphodiester bond formation and decreasing the reaction rate. In particular, the decreased electron density at the 3'-hydroxyl of dFdC in the nascent DNA chain is likely a substantial reason for the pausing of DNA polymerase following misincorporation of dFdCTP. In addition to conformational and electrostatic effects, fluorine substitution also alters the size of dFdCTP relative to dC, and steric clashes may also contribute to the pausing of DNA polymerase following dFdCTP misincorporation.

The cytotoxicity of dFdC is highly correlated with misincorporation of dFdCTP into DNA. Although dFdC misincorporation affects DNA polymerase processivity, dFdC is found predominantly in internucleotide linkages in DNA [48]. Thus the anticancer activity of dFdC appears to arise mainly as a consequence of alterations in DNA-mediated processes following dFdCTP misincorporation, and not as a result of alterations in DNA polymerase activity. The present study demonstrates that dFdC substitution alters the local structure of the DNA duplex region of a model Okazaki fragment. These results are consistent with conformational changes in DNA resulting from dFdCTP misincorporation as being responsible, in part, for the DNA-mediated effects of dFdC. In this regard, the dC analog cytarabine, a potent anti-leukemic agent, interferes with top1 mediated cleavage of DNA [78]. The same properties described above as contributing to reduced misincorporation of dFdCTP into DNA and DNA polymerase pausing also likely contribute to alterations in DNA-mediated processes due to dFdCTP misincorporation. At present, it is not possible to delineate the separate effects of altered structure, decreased stability, altered electrostatics and changes in size to alterations in DNA-mediated processes. Nonetheless, the present results are consistent with alterations to the biophysical properties of dFdC-substituted DNA contributing to the anticancer activity of this drug.

Recent studies have demonstrated that other nucleoside analogs that have potent anticancer activity exert their biological effects, in part, by altering the structural and/or electrostatic properties of DNA following their misincorporation. Thus, the dC analog cytarabine inhibits normal processing of supercoiled DNA by top1 [78]. Misincorporation of the 5-FU metabolite FdUTP renders the DNA a substrate for the mismatch repair machinery of the cell [68]. Additional biological studies will be required to identify the specific DNA-mediated processes that dFdC substitution interferes with. Knowledge of the biophysical alterations induced in DNA by dFdC-substitution may be useful in modulating these interactions in a manner that leads to more effective use of dFdC in combination chemotherapy. Alternatively, this information may prove useful in the design of alternative nucleoside analogs, or other drugs, that modulate DNA-mediated processes specifically in tumor cells.

2.7 Acknowledgements

The authors thank Junmei Wang (UCSF) for assistance in preparing the point charges and determining the fluorine bond parameters for dFdC.

2.8 Tables

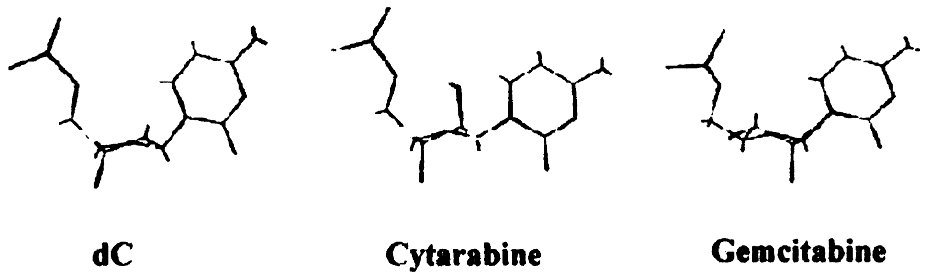
Base Pair	AH2/GNH2	T/G NH	C NH1/NH2
A3-T22	7.21	13.87	
A4-T21	7.54	13.93	
G5-GemC20	6.10	12.68	8.10/7.15
A6-T19	7.72	13.43	
T7-A18	7.87	13.44	
T8-A17	7.30	14.16	
C9-G16	5.86	12.12	8.25/6.75
C10-G15	6.04	12.70	8.36/6.80
T11-A14	7.68	13.76	

Table 2.1: Assignment of Exchangeable 1H Resonances for GEM

2.9 Figures

Oligomer	T_m(degrees C)^a	$\Delta G_{vH}^{CD}(kcal/mol)$	ΔH_{vH}^{CD}(kcal/mol)	ΔS_{vH}^{CD}(kcal/mol)
GEM	41.2	-14	-94 (-76)^b	-80
OKA	45.5	-14	-87 (-80)^b	-73

Table 2.2: Thermodynamic Parameters for the Formation of GEM and OKA at 20 degrees C.^{a} TM extrapolated for a total strand concentration of 13.3 μM. ^{b} Comparison of van't Hoff Enthalpies from obtained from analysis of the shape of melting curves and from TM dependence on strand concentration



5' CAAAGATTCCTC **[GEM]; X = dFdC**

3' GTTTXTAaggag

Figure 2.1: Structure of 2'-deoxycytidine (dC), cytarabine, and gemcitabine (dFdC). The nucleotides are shown in the same conformation each adopted in the final, refined NMR structures of OKA (ref 24), ARAC (ref 24), and GEM. The sequence of the model Okazaki fragment investigated in these studies is shown at the bottom of the Figure. Deoxyribonucleotides are indicated in upper case and ribonucleotides in lower case. The position of dFdC substitution in GEM is indicated by an "X".

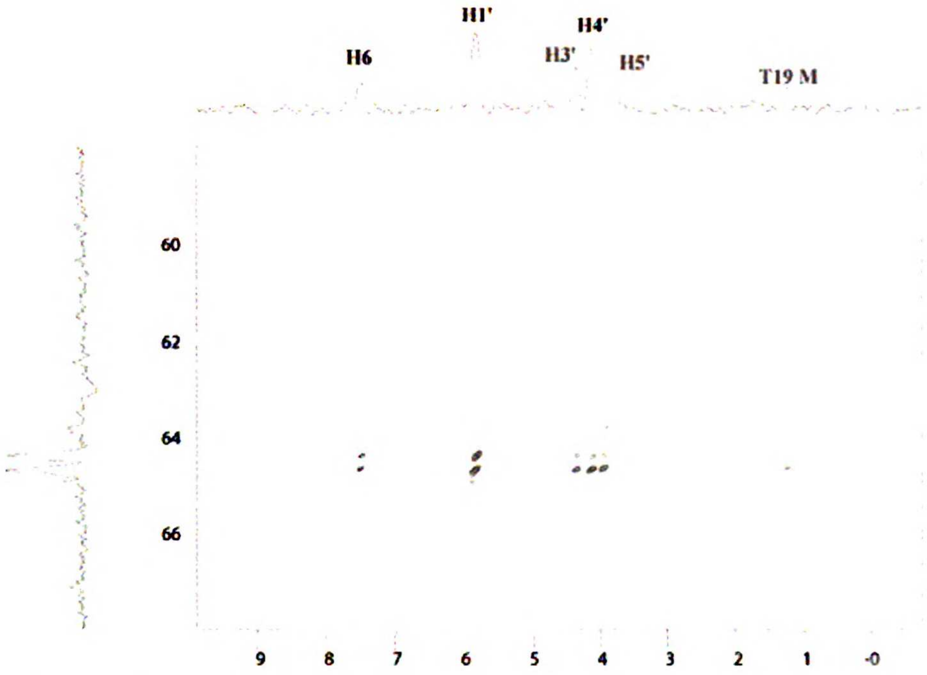


Figure 2.2: 2D ^{1}H-19F HOESY of GEM. Crosspeaks are indicated for the intraresidue interaction of the geminal difluoro group of dFdC20 with H6, H1', H3', H4' and H5'. Crosspeaks are also indicated for the interesidue interaction of F2'/F2" with the methyl group of T19.

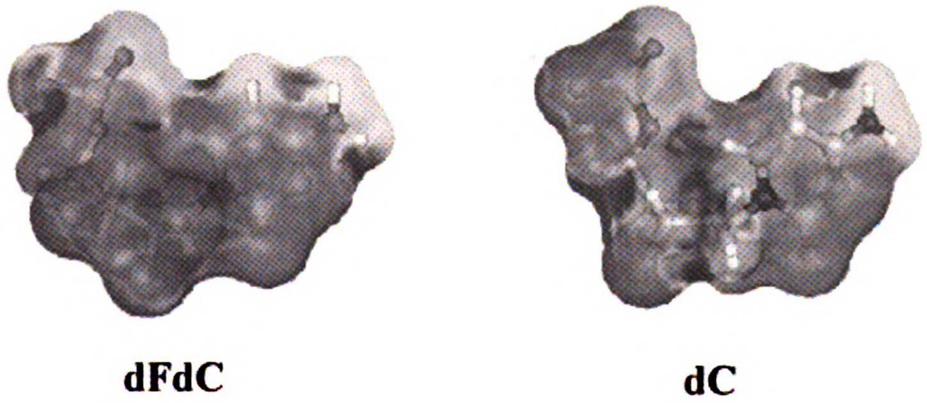


Figure 2.3: The electrostatic surface potential for dFdC and dC. Coordinates for each nucleotide were extracted from the NMR structures of GEM and OKA (ref 24), respectively. Electrostatic potential was calculated using Delphi.

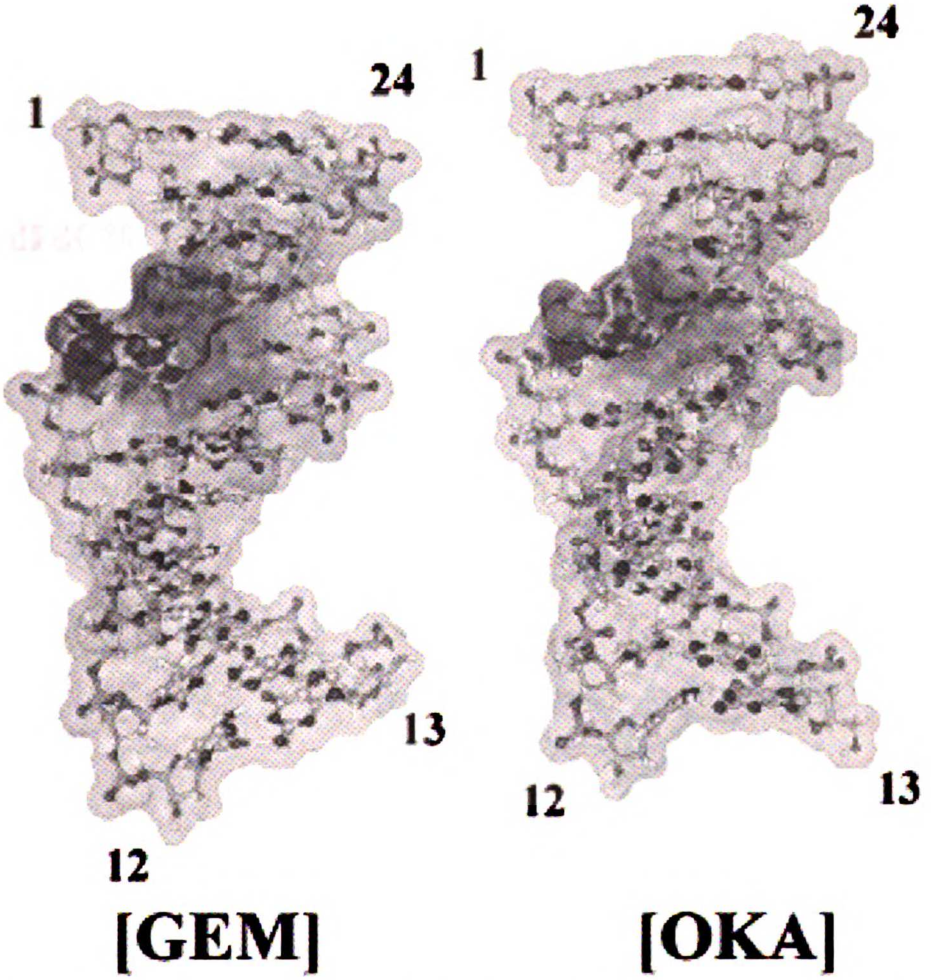


Figure 2.4: NMR structures of the dFdC-substituted model Okazaki fragment GEM (left) and the native Okazaki fragment sequence OKA (ref 24; right). All nucleotides except dFdC in GEM and dC in OKA are indicated as ball and stick figures. The electrostatic surface potential for dFdC and dC is indicacted in each structure, respectively.

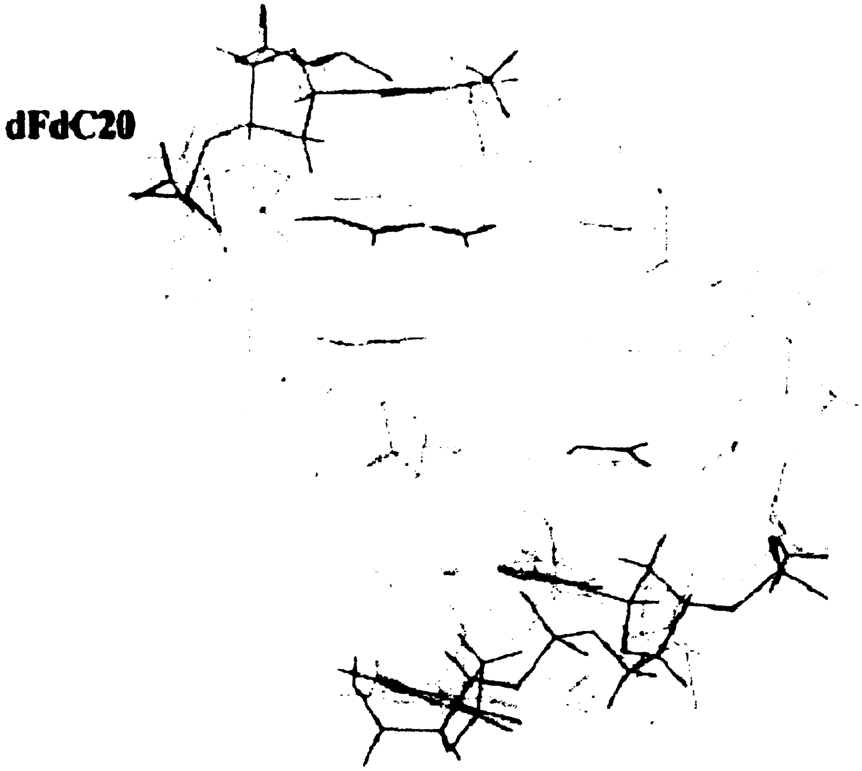


Figure 2.5: Superimposition of the DNA duplex region (DDR) for GEM (red except dFdC20 in blue) and OKA (green). Overall RMSD between the DDR of the two structures is 1.26 A.

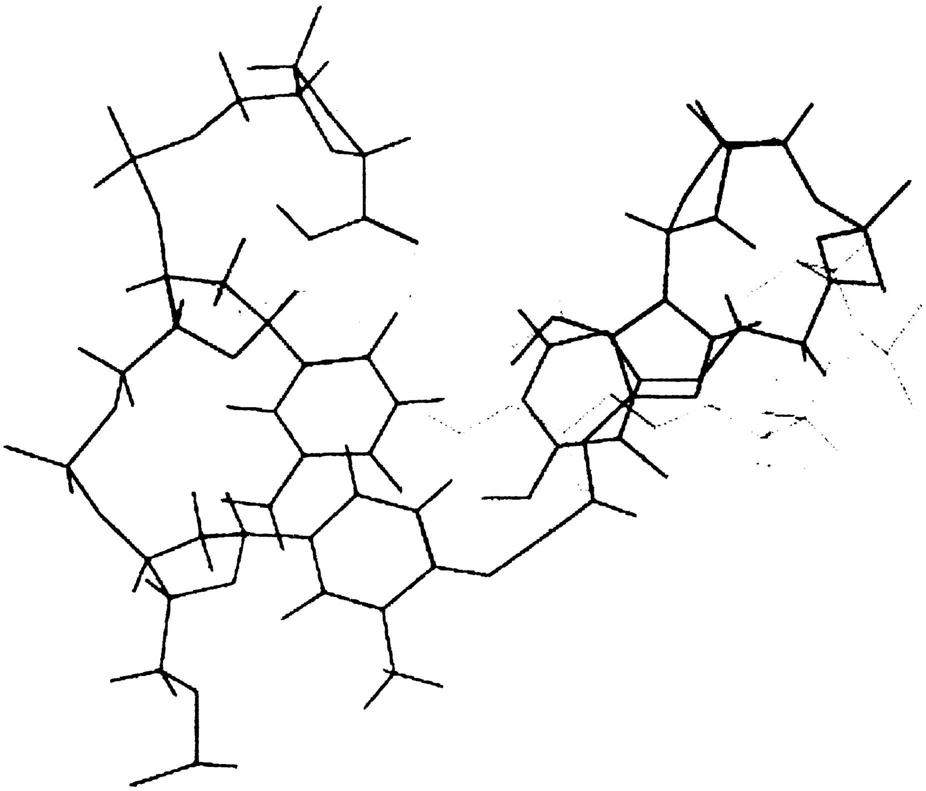


Figure 2.6: View along the dyad axis of GEM showing base stacking between the three consecutive adenosines A2, A3 and A4. Similar base stacking was observed in OKA, the model Okazaki fragment consisting of all native nucleosides, but was disrupted in ARAC, the model Okzaki fragment with cytarabine substituted at the same location as dFdC20 in GEM (ref. 24).

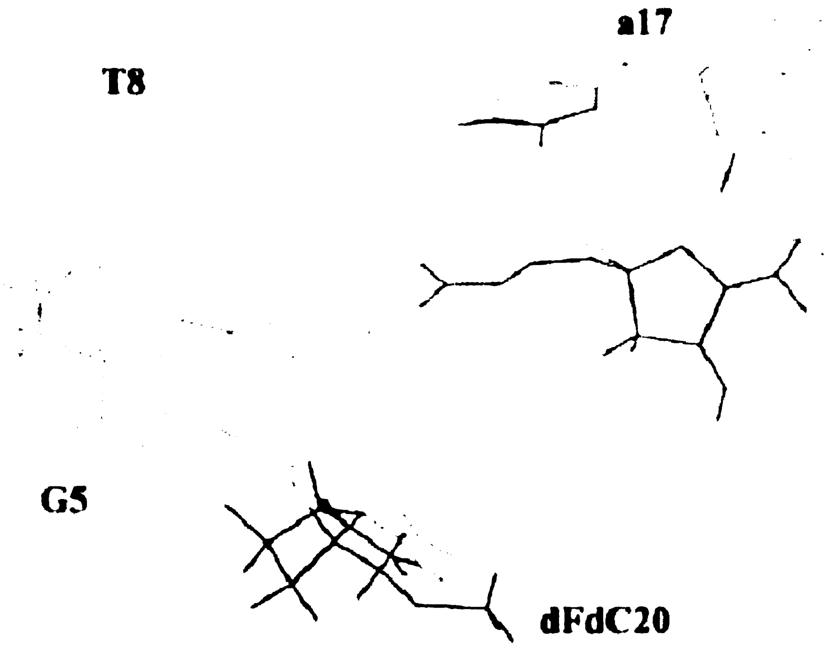


Figure 2.7: Structure of GEM in the junction region and near the site of dFdC substitution. dFdC20 is shown in blue while its base pairing partner G5 is shown in green. The three adenosines A6, A18 and a17 are shown in red and their base pairing partners T7, T8 andT19 are shown in yellow. Considerable base-base overlap occurs between the adjacent adenosine a17 and A18 as was observed in OKA, however A6 and A18 which are the purine components of adjacent base pairs, but on opposite strands of the duplex, do not overlap. Partial base overlap was observed between these adenosines in OKA.

Chapter 3

Unrestrained Molecular Dynamics of Okazaki Fragment Models

3.1 Introduction

Recent advances in molecular dynamics modeling methodology, coupled with dramatic increases in computing capacity, have led a number of researchers to investigate whether unrestrained force fields are capable of reproducing experimentally determined structural details of nucleic acids. An example of a recent "success" is the observation of an A-form to B-form transition in a DNA duplex [18]. The critical components for stable, accurate simulations of DNA are:

1. Carefully determined parameters which fit small molecule data well and are designed to be transferable to larger biomolecules

2. Explicitly hydrated/solvated system using TIP3P waters

3. Accurate long-range electrostatics using PPPM (particle-particle, particle-mesh) techniques such as PME

From crystallographic and NMR studies the DNA duplex is known to strongly favor the canonical B form (in a hydrated solution). When a solvated A-form DNA duplex is simulated using the AMBER [12] molecular dynamics application and the Cornell et al. force field, it rapidly (500ps) converts to B-form, as is expected. Further work on DNA duplexes showed that the Cornell et al. force field was capable of reproducing some aspects of experimentally determined sequence-specific structure of DNA duplexes, but that the twist parameter was systematically lower than expected (31-32 degrees) [54]. The exact value of the twist of arbitrary sequence DNA in solution is still a matter of debate, as both NMR [105] and gel analysis [83] point to an average twist of 33-35 degrees, while crystallography appears to prefer a larger value of 35-36 degrees [79]. The crystallography data is skewed by crystal packing effects which appear to stabilize the higher twist, while NMR lacks the long-range accuracy to determine twist over an entire molecule, and small local errors can lead to large global ones. The gel data lacks atomic resolution but should provide a very accurate measure of the twist of DNA adsorbed to a surface.

Recently, increasing interest in RNA tertiary structure has led to simulations of RNA to determine how the Cornell et al. force field performs on this class of nucleic acids. A priori, RNA can be expected to be a more challenging structure prediction target because the RNA sugar has a higher intrinsic repuckering barrier and because RNA sequences tend to form more complex tertiary structures than DNA. From the perspective of the AMBER force field, the change of the H2'2 atom to O2' and the added proton HO2' adds several bond, angle and dihedral terms (including terms which determine the chi and sugar pucker angles), and nonbonded interactions. Initial simulations of RNA [19, 69] have focused on duplexes and hairpins with known experimental structures. The simulations will typically start in a conformation which is not the experimental structure (for example, with a "wrong" conformation for a hairpin loop, or in the B-form family). The goal is to obtain a trajectory that moves from the "wrong" conformation to the "right" one. If the "wrong" conformation has a higher energy than the "right" one and the transition barrier is low enough, then the trajectory should spontaneously move to the "right" conformation. These simulations have demonstrated that the Cornell et al. force field does not handle RNA as well as DNA, in part because the simulations (one to two nanoseconds in length) have been too short to adequately surpass the higher energy barriers intrinsic to RNA. In the hairpin simulations, the "wrong" conformation was maintained for 1-2ns (showing a high transition barrier to the "right" structure), but when the loop sugars were converted to deoxyriboses, the loop rapidly found the "correct" conformation. It is not certain what the physical cause of the reduced conformational sampling of ribose residues is; however, it is known that the addition of the 2' hydroxyl alters the torsional barriers of the dihedrals of the sugar as well as chi as well as a limited effect on the backbone. Although this has a significant effect on the repuckering barrier, we will present data that suggests that the reason that B-RNA does not automatically convert to A-RNA in the 10ns timeframe is instead due to favorable hydrogen bonding between the 2' hydroxyl and the RNA backbone.

In conjunction with NMR studies of an Okazaki fragment model (chapter 2 and[31, 33, 32]) we carried out moderate-length simulations based on the Okazaki fragment model starting in both the A and B conformations (table 3.1). Unrestrained molecular dynamics of nucleic acid duplexes is a useful probe of structural dynamics (provided the simulated structure is close to the experimental structure and the force field is accurate), and can provide valuable insight in situations where the NMR data is affected by underlying molecular motion. The candidate structures from molecular dynamics can be used to generate structural ensembles with associated probabilities that fit the NMR data significantly better than a single structure, in situations where the structure is flexible or dynamic [107].

The Okazaki fragment model is a twelve base pair duplex with the following sequence:

5′ dC dA dA dA dG dA dT dT dC dC dT dC 3′

3′ dG dT dT dT dC dT dA rA rG rG rA rG 5′

Based on earlier results which suggested that unrestrained molecular dynamics using the Cornell et al. force field [54] was able to reproduce some aspects of two DNA duplex NMR structures, we applied this force field to the Okazaki fragment model. This is a hybrid-chimeric duplex, as positions 13-17 (first five residues of second strand) are RNA. However, it quickly became clear that while the simulation A-start form kept the RNA residues in C3'-endo sugar pucker, the simulation B-start was incapable of converting the RNA residues to C3'-endo. The DNA residues converged to the same values in the two free MD simulations, while RNA showed a complete lack of convergence. From this data we decided that solving the problem of poor convergence of RNA in the Cornell et al. force field was necessary before DNA:RNA hybrid simulations could produce useful results. To avoid complexity associated with DNA-RNA interactions all further simulations were performed on pure RNA:RNA duplexes. When pure RNA:RNA duplexes can be modeled successfully, the DNA:RNA simulations should be revisited.

Very recently, a new force field, Wang et al. [53], derives from and improves upon the Cornell et al. force field. Between Cornell et al. and Wang et al. two other force fields, parm96 and Cheatham et al [17], also known as parm98, were developed. parm96 was designed to improve certain aspects of the protein backbone parameters, while parm98 was developed with a specific goal of modifying the sugar pucker torsions to improve the predictive ability of the Cornell et al. force field for DNA, which underestimated the sugar pucker phase and helical repeat length (twist). Between Cornell et al. (parm94) and parm99, several terms that affect the sugar pucker were changed: CT-OS-CT-N* (C4' to O4' to C1' to base N) had extra barriers added, OS-CT-CT-OS (O5' to C5' to C4' to O4') barrier was increased, and OS-CT-N*-CK (chi) barrier was reduced. In Wang et al. (parm99), the sugar pucker was extensively modified to have much more sophisticated torsional parameterization, with effectively all torsions around the pucker being modified to either change barrier heights or add extra terms. These improvements were designed to address known issues with the behavior of the force field for both DNA and RNA.

In total, four simulations of 10ns each were carried out using the cluster system described in chapter 6. Additional trajectories (not discussed within this document) brought the total amount of time to 80ns. Each 10ns run took three months of computing time using four processors each, but the overall run time was less than 12 months because two or more simulations were run simultaneously.

3.2 Methods

3.2.1 Okazaki fragment model: initial model construction and simulation preparation

Okazaki fragment models were initially constructed as PDB files. The initial Okazaki fragment models were constructed as canonical A or B form DNA using the AMBER application "nucgen" with the given sequence. "nucgen" cannot generate hybrid/chimeric molecules nor can it make B-form RNA, so duplexes with all DNA residues were created, and the atom name H2'1 in all RNA residues was renamed to O2'. Following coordinate generation the AMBER application "LEaP" was used to convert the PDB files to AMBER "prmtop" and "prmcrd" files. LEaP added any hydrogens (such as RNA's HO2' attached to O2') that were missing from the PDB file. Before converting to "prmtop" format the model was solvated using the LEaP command "solvateBox". The solvating waters were taken from the LEaP WATBOX216 which is a box of 216 waters equilibrated and minimized using Monte Carlo (ref).

The model was solvated so that at least 10Å of water surrounded the model in each direction. Following solvation the system was neutralized by adding sodium ions at the global electrostatic minimum sequentially until neutrality was reached. Although this procedure is not guaranteed to produce the lowest possible electrostatic energy, equilibration and dynamics cause the ions to form a "cloud" of ions rather than maintaining the initial positioning. Following solvation and neutralization the solvent was minimized: 500 steps of steepest descent followed by 500 steps of conjugate gradient minimization (duplex restrained by positional restraints of 50kcal/mol). This minimization reduces unfavorable contacts between water molecules and the solute as well as between water molecules. During this minimization the duplex was held rigid to maintain the initial A-form or B-form conformation. Next, the solvent box was equilibrated to a density of one atmosphere by constant pressure dynamics (10ps at 100K followed by 10ps at 300K; duplex restrained by positional restraints of 50kcal/mol). The initial density was approximately 0.7atm due to poor interfacial contacts between the water and the solute (the LEaP water placement algorithm deletes water atoms and at the edge of the box that are too close to the solute or neighboring waters, leaving "density holes"). Before equilibration the box size of the A-start simulations was 46Å by 47Å by 65Å with a density of 0.76 atm; following equilibration the box size was 42Å by 43Å by 59Å with a density of 1.06atm. For the B-start simulations the box size changed from 47Å by 48Å by 70Å (.7atm) to 43Å by 43Å by 63Å (1.06atm). The B-RNA form is somewhat longer than A-RNA form causing the Z dimension of the box to be larger. After equilibration, extensive minimization of the solvent and the solute was performed. The minimization occurred in six steps, with 500 steps of steepest descent and 500 steps of conjugate gradient. The initial positional restraint on the duplex was 25kcal/mol, and was reduced by 5kcal/mol for each group; the final minimization applied 0kcal/mol positional restraints and thus allowed the duplex to relax. Following minimization the water was re-equilibrated to account for the minimized structure by unrestrained dynamics for 10ps at 300K. In this final step the unrestrained duplex moved approximately 1Å away from its initial structure. Finally, the system was considered ready for production dynamics. Identical conditions were used for all simulations, with the only exception being the nucleic acid starting model sequence and conformation, and the length of the simulation. Multiple runs were performed simultaneously, with four CPUs assigned to each run, to gain the maximum possible throughput.

Following production dynamics, all simulations were post processed in preparation for analysis. All trajectories were stripped of waters, imaged into the central periodic box, centered, and RMS-aligned to the initial snapshot using the PTRAJ application of the AMBER suite. Following processing by PTRAJ, the final 1ns of each simulation was averaged and minimized to produce a "final" structure. The Dials and Windows parameters, sugar pucker and backbone torsional angles were measured for all snapshots in the trajectory. Sugar pucker was calculated using the CARNAL application of the AMBER suite. The AMBER energy according to the Cornell et al. and Wang et al. force fields (vacuum system, distance dependent dielectric to simulate solvent screening, and no electrostatic cutoff) was calculated for each snapshot. All parameters and energies were also determined in the same manner for canonical A- and B-conformation starting structures. All processing and analysis were performed using the programs indicated (Dials and Windows, CARNAL, and AMBER), and were automated using Python scripts based on the Ensemble/Legacy of chapter 5.

3.3 Results

3.3.1 Okazaki fragment model: Analysis of Okazaki fragment (Hybrid-chimeric RNA:DNA) simulations

The initial simulations of the Okazaki fragment model (figures 3.1 through 3.8) demonstrate that while the DNA portions of the molecule are flexible and converge to values near that for the NMR structure (mainly C2'-endo for DNA residues), the RNA portions of the molecule appear to maintain the same values as the initial conformation for the entire 2-3ns simulation. For example, the RNA residues 13-17 in the canonical A start simulation using Cornell et al. remain in the C3'-endo sugar pucker (zero transitions), but the simulation starting with canonical B has more C2'-endo-like sugar puckers. There is an increased number of sugar pucker transitions between C2'-endo and C3'-endo. It appears the B-start simulations have not converged after 3ns, nor are they fully equilibrated with the force field. The lack of convergence demonstrates that the force field is not sufficient for unrestrained molecular dynamics to produce qualitatively correct predictions of conformation and dynamics of RNA residues within the timeframe simulated. Even terminal residues which are allowed greater flexibility than interior residues do not converge to A-RNA. We do not expect that the force field would support a transition from either Z-DNA (large transition barrier between right-handed and left-handed helices) or separated strands (extensive search of conformation space to find Watson-Crick base pairing), but the B- to A-form transition remains entirely within one conformational subfamily and thus should have a smaller transition barrier and should not need as extensive a search of conformational space. The specific height of the transition barrier between B-RNA and A-RNA is not known, nor is it known whether other conformational subfamilies have smaller transition barriers or deeper AMBER energy wells than B-RNA and A-RNA. Because we cannot get adequate convergence from B-RNA form, we cannot be confident that the simulation trajectory will consistently move to the "correct" conformation, limiting the utility of unrestrained molecular dynamics to provide information on the dynamics of known-structure systems. One approach to this problem is to start many simulations from the same conformation but with different random number seeds. This approach greatly improves the chances of seeing a transition to the "right" structure within the expected time frame.

3.3.2 Okazaki fragment model: Analysis Okazaki fragment (RNA:RNA) simulations

Average minimized structures from the last 1ns of the Okazaki fragment (RNA:RNA) simulations

The minimized average structures are presented in figures 3.17 through 3.24. As can be seen in the axial and side views, the A-start simulations (independent of the force field) remain in the A form with significant X-displacement from the helical axis, while the B-start simulations (also independent of the force field) have a structure which is neither A-form nor B-form but equidistant to both.

Pairwise RMS deviation of the Okazaki fragment (RNA:RNA) average structures with canonical A- and B-form RNA

By inspection of specific pairwise RMS between the average minimized structures in tables 3.2 and 3.3, it is clear that the A-start simulations remain very close to the canonical A-form (approximately 1.8Å) and distant from the canonical B-form (approximately 6Å). The time course of RMS (figure 3.13 to 3.16) shows that the A-start simulation actually started about 1.5-2Å from the true canonical A-form structure, even though the initial model was pure canonical A-form. The reason for this is that the equilibration procedure used to minimize the initial model, as described in the Methods section above, changed the structure of the model significantly, because the initial conformation of the molecule was unfavorable in the AMBER force field.

RMS deviation of the Okazaki fragment (RNA:RNA) simulations from canonical A- and B-form RNA

While the A-start simulations clearly maintain A-form for 10ns, the B-start simulations do not maintain the initial B-form (consistent with the experimental data since B-RNA is considered a highly unfavorable conformation) but they do not reach the A-form within the 10ns simulations (figures 3.13 to 3.13). The B-start simulations start about 2Å from canonical B (for the same reasons as the canonical-A start simulations above), and 5-6Å from canonical A. During the 10ns for both force fields, the B-RNA start simulation never gets any closer to canonical A than 3.5Å and spends most of its time 5-6Å away, with one excursion to 8Å before returning to 4Å in the Cornell et al. simulation. By the end of the 10ns the B-start simulations are both 4.5-5Å away from B and 5.4Å away from A. It is significant that the simulations are the same distance from canonical B as from canonical A, because this suggests that the simulation may indeed be moving along a transition pathway to canonical A and that the sampling time of the simulation is simply too short to observe the transition.

Helical parameters and sugar pucker/chi angles from Okazaki fragment (RNA:RNA) simulations

Coordinate RMS can be a misleading measure of difference between structures. While a low RMS clearly demonstrates that two structures have similar structures, a high RMS (>2.5Å) does not mean the two structures are dissimilar. This is because small local differences in a global variable such as helical twist can accumulate over the length of the molecule. These accumulations lead to large coordinate RMS even if the two global parts of the molecule are similar overall. For this reason it is instructive to compare structures using internal coordinates, such as helical parameters. Because helical parameters are effectively non-linear functions of the atomic coordinates, helical parameters determined for average, minimized structures lose some information compared to helical parameters which are computed for snapshots and then averaged. For the following analysis, the helical parameters were determined for the final 1ns and then averaged, instead of being computed for the average structures (figures 3.5 to 3.12).

The most important helical parameters for distinguishing A form from B form are base inclination, X displacement from the helical axis, and twist. Canonical A form has high base pair inclination (approximately 15 degrees), high negative X displacement (approximately -5 Å), and lower twist (33 degrees) while Canonical B has negligible inclination and X displacement and slightly higher twist (36 degrees, although see comments above regarding the actual twist value of B-DNA in solution). Sugar pucker, while not actually a helical parameter, is also a useful parameter to inspect, because canonical A-form duplexes typically have C3'-endo sugars while canonical B-form duplexes have C2'-endo sugars. This sugar pucker convention is adhered to in canonical form duplexes but many examples of non-C2'-endo sugars in B-form helices and non-C3'-endo sugars in A-form helices exist in the structural literature. However, it appears that a pure B-form helix is not compatible with C3'-endo sugars and that pure A-form helix is not compatible with C2'-endo sugars: in both cases, significant rearrangement of related parameters such as chi and the backbone torsion angles (and the helical parameters as a result) is required to accommodate the opposing sugar conformation. The chi residue in duplex DNA and RNA is anti (approximately 180) except in abnormal situations or fraying terminal residues. Single residues within the duplex may transiently take on alternative sugar pucker conformations, as observed within these simulations.

In all the canonical A-start simulations, the residues maintain canonical A-like chi (200 degrees) and pucker (near C3'-endo) except for the terminal residues, which break their Watson-Crick hydrogen bonds, de-stack and sample many chi angles and sugar puckers throughout the 10ns simulation. Inclination for the A-start simulations shows an interesting trend where the terminal residues are strongly positive and the value smoothly drops toward the canonical B values in the middle of the molecule. This smoothness is likely due to the difficulty of changing the inclination rapidly, since rapidly varying values of inclination disrupts favorable base-stacking. X displacement is solidly between A-RNA and B-RNA forms, but slightly closer to A-RNA form. All the helical parameters and pucker angles are effectively identical between the Cornell et al. and Wang et al. simulations. In the canonical B-start simulations, the results are much less consistent, showing much greater deviation from the starting B-RNA and target A-RNA conformations than the A-start simulations, and also much greater deviation between the Cornell et al. and Wang et al. force field. While the chi residues are all within a few degrees of the expected values during the A-start simulations, they are highly variable in the B-start simulations, mirroring the variance of pucker seen in the B-start simulations. In fact, chi and pucker are highly correlated, and inspection of the plots of these two parameters demonstrates this fact. It appears (from close inspection of sugar pucker transitions in the trajectories) that pucker primarily determines chi, in that the chi value adjusts to maximize base-pairing and base-stacking while being consistent with the pucker. This is not surprising given that the torsional barrier to chi rotation is much lower than that of sugar puckering in the AMBER force field. The puckers in the B-start simulations are highly dynamic during the course of the simulation. X-displacement stays strongly B-like, and for the internal four residues the inclination is B-like. The terminal residues approach A-like inclination values, but in the B-start simulations, the terminal residues are significantly more dynamic, with extensive de-stacking and Watson-Crick h-bond breaking. While it is possible that this de-stacking is a valid transition pathway which is required for the conversion between B-form and A-form RNA, since it allows for greater conformational sampling than a rigid duplex, it must require more than 10ns for this conversion to propagate to the stable interior of the duplex.

AMBER Energies from Okazaki fragment (RNA:RNA) simulations

Inspection of the AMBER energies of the canonical B-form and A-form RNA (as well as DNA), compared to the energies of the MD structures, is enlightening because it demonstrates the complexity of the potential surface of duplex nucleic acids (table 3.4). The initial, unminimized forms of canonical A and B DNA and RNA all have unfavorable total energies. All of the internal (bond, angle, dihedral) energies are unfavorable, with the canonical A conformation of DNA and the canonical B conformation of RNA showing highly unfavorable bond lengths. van der Waals is highly unfavorable, while electrostatics appear favorable, although not greatly so, for all the unminimized structures. The minimized structures of A and B DNA and RNA are more useful measures of intrinsic energies, with the caveat that these energies are those of the closest local minima to the starting conformation, rather than the deepest potential well of that conformational subfamily. Molecular dynamics searches out nonlocal potential wells more effectively than minimization. In the minimized structures, the bond, angle, and dihedral terms are significantly improved, as are the vdW and electrostatic energies. Only in the case of A-DNA does minimization not improve the electrostatic energy (it drops slightly, from -657 to -625, most likely due to increased phosphate repulsion). However, this is more than compensated for by greatly improved internal and vdW energies. The most stable minimized conformation by far is B-DNA, near -1300, while B-RNA and A-DNA are approximately equal (-960). A-RNA is the least stable conformation (-800). Of course, these values only represent the conformational energies of the minimized starting structures. Without the sampling of MD, potentially more stable structures within the same conformational subfamily will not be found. Further, these energies only show the enthalpic contribution to the free energy. The configurational entropy, translational/rotational entropy, and desolvation entropy cannot be directly measured using this method. A promising technique, MM-PBSA, has been applied to other RNA and DNA duplexes [100], to estimate the global free energy difference between various conformations. MM-PBSA estimates the free energy of the system by summing the following energies: average conformational energies over a number of snapshots within a subfamily energy well, conformational entropy determined by normal mode analysis, and a desolvation energy term based on the Poisson-Boltzmann equation solved at the surface of the molecule. In the duplex r(CCAACGUUGG)$_2$, B-RNA electrostatic energies are significantly more stable than A-RNA, while internal and van der Waals are about the same. It is only by adding in the solvation energy that the two conformations are seen to have roughly the same free energy. It is still not clear why the B-RNA is considered so stable according to the AMBER force field and the Poisson-Boltzmann solvation term, given that the structure is not observed in solution.

For the dynamic simulations, the average energies from the "start" and "end" of the simulations are far more useful because these values represent structures which are minimally and extensively (respectively) equilibrated with the force field. In the case of the "start" energies, the structures have already undergone minimization and dynamics in the presence of solvent, which allows the duplex to relax significantly more than the initial minimized canonical form. In the case of the "end" energies the molecule has relaxed significantly, undergone significant motion away from the initial model, and has had time to adjust to the solvent conditions. The trend for the starting structures is that the total energy of the B-form is better than the A form for both the Cornell et al. (-927 for B vs. -545 for A) and the Wang et al. force fields (-785 vs. -423). By the end of the simulation, the trajectories continue to maintain this trend for both the Cornell (-778 for B vs. -596 for A) and Wang (-653 vs. -447) force fields. By parsing out the various energy components, it appears that at the beginning, the internal terms (bond, angle, dihedral) are almost identical, but that electrostatics are more favorable for the

B-form (-1045 vs. -625 for Cornell and -995 vs. -576 for Wang). This is consistent with the observation that in the A-form the phosphates are significantly closer to each other and therefore have a greater repulsive force. The A-form is more favorable in the van der Waals term, however this difference is not nearly as large as the difference in electrostatics (-373 vs. -408 for Cornell and -364 vs. -408 for Wang). These trends are maintained throughout the simulation, because by the last 10ps, the electrostatic energies are still significantly more favorable for the B-start than the A-starts (even though the B-start simulations are not longer in the canonical B conformation). Again, it is clear that by ignoring the entropic contributions to the free energy we lack enough information to determine unambiguously why the B-start form or the final structures from the B-start simulations are so stable, given the lack of B-RNA in nature and the apparent decreased stability due to the steric repulsion of the 2' hydroxyl of RNA in the B-form.

Time course of helical parameters in Okazaki fragment (RNA:RNA) simulations

Because the structures are dynamic during the course of the simulation it is helpful to inspect the time course of the helical parameters and backbone angles in addition to just the averages (figures 3.25 through 3.40). In the case of the A-start simulations, the backbone and sugar puckers are relatively constant throughout the simulation. For non-terminal residues, there are only three sugar pucker transitions from C3'-endo to C2'-endo during the entire 10ns Cornell et al. A-start simulation, and two for the Wang et al. simulation, although the terminal residues repucker frequently. Wherever a sugar pucker transition is seen, the chi angle immediately adjusts to the new sugar pucker.

The helical parameters for the A-start simulation are also quiescent, with a gradual drop in inclination over the course of the simulation from about 20 to 0, the effect is lessened for the the terminal residues, which maintain their higher A-RNA-like structure. X-displacement stays well-fixed within the A-RNA regime.

In the B-RNA simulations the results are completely different. This information strongly demonstrates that while there is a significant barrier to repuckering within the the context of the internal residues of the A-form duplex, terminal residues in both subfamilies have a lowered barrier; therefore, the actual barrier to repuckering intrinsic to the torsional parameters and charges in the force field are not very high. Sugars repucker frequently in the B-start simulations; for both the Cornell et al. and the Wang et al. force field, each residue repuckers from C2'-endo to C3'-endo at least once, and some residues (including non-terminals) have long residence times (3ns) in C3'-endo. From inspection of movies of the simulation, it actually appears that the helix is attempting to undergo a transition, possibly to all C3'-endo A-form, but that the barrier to interconversion in the internal residues is too high to be achieved, while at the termini, which are less constrained, the transition barrier is more easily overcome. If this is true, it may only be necessary to sample for a longer time to see convergence between the A-start and B-start simulations. In the Cornell et al. simulation, the plasticity in the sugar pucker and chi torsion are mirrored by the greater (compared to the A-start simulations) dynamics of the inclination parameter. After 3ns, the inclination of nearly every residue undergoes a large change, first dropping and then returning to B-RNA form (near zero), and in some cases, jumping up to the A-RNA range. This dynamic behavior is not reproduced by the X displacement parameter, however, which stays strongly B-RNA like for the course of the simulations. In the Wang et al. simulation, however, the inclination does not undergo any large changes until the very end of the simulation. There is a general increase, over the first 3ns, to near the A-RNA regime, even in the inner residues, which is then followed by slow decline over the next 6ns back to B-RNA levels. This pattern is most noticeable in the terminal residues on one end of the duplex (C10-G15 through C12-G13), but propagates as far as the inner residues. This behavior is actually quite interesting, because it suggests that large, whole-molecule duplex transitions, stimulated by the increased flexibility and conformational freedom at the termini, may be a driving factor for the transition between the unstable B-RNA and the stable A-RNA forms. Alternatively, it may only be that the enhanced flexibility of the termini ("end-effects") are destabilizing the duplex and allowing greater conformational freedom without a concomitant subfamily transition.

Formation of hydrogen bonds stabilizes B-form RNA

The data above present an interesting picture with respect to the potential transition pathway and associated barrier(s) between B-form and A-form. The AMBER energies show that the B-form is quite stable within the Cornell et al. and Wang et al. force fields, even more stable (ignoring conformational and solvation entropic effects) than A-form. But, during the simulation the A-form is completely stable, moving very little from the initial conformation, while the B-form is very dynamic, rapidly moving away from the initial conformation to a form which is as close to A-form as it is to B-form. This suggests that conformational entropy within the potential well of the B-form is higher than the A-form, which could explain why the B-form is so stable in the AMBER force field. While sugar pucker rigidity is usually invoked to explain the large energy barrier to transition between B-form RNA to A-form RNA, the plasticity of the puckers in the B-start simulations suggests that another factor is inhibiting the transition. We inspected movies of the simulations and have determined a plausible explanation for the enhanced barrier between B-form and A-form DNA based on hydrogen bonds which are only formed within the B-form subfamily (figures 3.41 to 3.44. Because the AMBER force field does not have explicit hydrogen bond terms, the only contribution to the formation of these bonds is the favorable electrostatic interaction between oppositely charged atoms. There is no directional form of hydrogen bonds in the AMBER force field (it is purely an electrostatic interaction) which disallows or disfavors highly angular hydrogen formations. The magnitude of the individual partial atomic charges in the AMBER force field is very large, so electrostatic interactions are very strong, and without the hydrogen-bond directionality term, nonphysical hydrogen bond arrangements are greatly overemphasized by the force field.

In the canonical A-form structures used as the seeds for the A-start simulations, the hydrogen bonding opportunities are limited, mainly to intraresidue HO'2 to O3'. This is because the backbone is rotated away from the sugar in the A-form due to the C3'-endo pucker. During the course of the simulation, this pattern is mostly maintained, and although some HO'2 to O4' interresidue bonds appear, this interaction is only a small tail on the distribution which is skewed to larger (non-hydrogen-bond) distances. The HO'2 to O5' interresidue bond also appears, but it also is only the tail of a distribution which peaks in the non-hydrogen-bonded distance range. The HO'2 to O3' intraresidue hydrogen bond appears to be strongly correlated with the HO'2 to O2P and O1P interresidue bonds, because every time the HO'2 to O3' distance increases to break the bond, the O1P and O2P distances also increase (from 4.5 to 5.5Å). This occurs because a coupled transition between the backbone and the sugar repuckering moves the O3' out of the way in a concerted manner with the phosphate group. The HO'2 spends noticeably more time interacting with solvent than specifically bound to the duplex atoms in the A-form. Solvent bonding would affect the solvation around the site significantly; rather than forming a hydrophobic exclusion cage around the site, the waters would require less loss of entropy due to the favorable polar interactions. This could help explain why A-form RNA is more stable than B-form RNA in nature when the gas-phase AMBER energies show that the B-form is more stable.

The picture is very different for the B-start simulations, which start with a formed HO'2 to O5' interresidue hydrogen bond, and show significantly more dynamics as well as specific correlated transitions between hydrogen bonding states. From the movies of the B-start simulations it is clear that H-bond forming and breaking is tightly correlated to sugar pucker transitions and backbone readjustment. During the course of the B-start Cornell et al. simulation, the initial state of HO'2 (residue 5) to O5'(residue

5) interresidue hydrogen bond is maintained for 2.5ns before a significant transition, at which time a simultaneous sugar repuckering and phosphate/backbone readjustment occurs. Each time a significant rearrangement of hydrogen bonding occurs within a specific HO2' sugar, the sugar repuckers between C2'-endo and C3'endo, or at least moves to a nearby location on the pseudorotation cycle. This demonstrates that, at least within the B-RNA subfamily, the sugar pucker barrier is not particularly high, but that the HO2' spends significant time hydrogen bonded to the duplex rather than to the solvent, and that this stabilizes the B-RNA conformation (within the AMBER force field). Internal h-bonding should destabilize the duplex because solvent hydrogen bonding would have a more favorable free energy due to the reduced need for "water cages" around the mostly aliphatic sugar. Our observation is consistent with the behavior noted in [69] where removing the 2' hydroxyls was sufficient to allow the wrong conformation of a RNA hairpin's tetraloop to move quickly to the right one. Unfortunately, the hypothesis that 2' hydroxyl hydrogen bonds stabilize B-RNA is a difficult one to test using NMR because the 2' hydroxyl peaks are typically reduced due to solvent exchange and because no solvent conditions have been found to stabilize the B-RNA form.

3.4 Conclusions

Dramatic improvements in force field, molecular dynamics and computer technology have led to greatly improved prospects for modeling biomolecular structure and dynamics within the past few years. Initial success with convergence of DNA in the A- and B- forms to a common B-form lent credit to the idea that RNA could be modeled using molecular dynamics. However, RNA modeling has been found to be significantly less successful, due to the greater complexity of the RNA potential energy landscape. Since the only two differences between RNA and DNA are the addition of a 2' hydroxyl to the sugar and a removal of the methyl from thymine to make uracil, it is clear that any structural differences must be isolated to those two features. The thymine methyl should incur some effect on the conformation because the greater volume of the hydrophobic methyl group would force water in that area to form a cage structure (such an ordering requires a decrease in entropy which would cost stability and the molecule would adjust its conformation in response). However, since this change is limited to only one base, it is unlikely to contain nearly the same potency to change duplex conformation as the 2' hydroxyl. The 2' hydroxyl effect is very complicated however. First, the hydroxyl affects the torsional preferences within the sugar ring, making the C3'-endo conformation more stable than the C2'-endo conformation. Second, the added polarity in that region affects the formation of water structure around the grooves. Third, the hydroxyl, at least within the C2'-endo conformation, can form hydrogen bonds to the backbone/phosphate atoms, potentially stabilizing that conformation.

Although the A-start simulations are extremely stable on the 10ns time frame, suggesting that the A-RNA structure represents a strong potential well within the AMBER force field. B-RNA is not nearly as stable nor does it converge to the A-RNA structure. There are several possibilities compatible with this observation. First, it is possible that B-RNA is unstable within the AMBER force field and that the barrier is simply too high to reach A-RNA in 10ns. Second, it is possible that B-RNA is unstable, but that there is a low-barrier pathway to another conformation equally distant from B- as from A-form, but with a higher barrier to A-RNA. Third, it is possible that A-RNA and B-RNA are both less stable than the observed mixed structure, but that the A-RNA barrier is much higher than the B-RNA barrier. Because the simulation was only carried out for 10ns, and the energy barriers between the various forms are too high to overcome, a longer simulation, perhaps 25 to 50ns, might be sufficient to overcome the barrier. To overcome a transition free energy of 5kcal/mol should take approximately 1ns, 6kcal/mol should take approximately 3nsec, and 7kcal/mol 20ns. So a 50ns simulation should be more than sufficient for B-RNA to convert to A-RNA assuming a 7kcal barrier, and that there are no deep potential wells other than B-RNA or A-RNA. Unfortunately, if the energy barrier is 8kcal/mol the amount of time required is on the order of 100ns. The three proposals above can all be tested by running longer simulations, and the answer can be determined by observing whether A-RNA remains in A-RNA form or converts to the mixed form, and whether B-RNA remains in the mixed form or converts to A-RNA. The primary challenge is in estimating the transitional energy barrier between A-RNA and B-RNA and possibly between A-RNA, B-RNA and mixed RNA, to determine whether the calculation can be made in a reasonable amount of time. This barrier can be approximated or calculated using various computational techniques although no general technique exists to find the lowest energy pathway between two subfamilies. Currently, runs using dnaMiniCarlo [103] are being carried out to determine a low-energy pathway, and initial investigation into PEDC [30] as a possible technique is being carried out as well. If such a pathway is generated, it can be submitted to MM-PBSA analysis to determine the free energy difference between the B-RNA and highest energy structure in the pathway. Several estimates in the literature have been made to identify the contributors to the energy barrier, but unfortunately, since this barrier has both significant enthalpic and entropic sources, and because the barrier could be cooperative (movement within one base leading to a lower energy barrier in an adjacent base), a simple energy-additive calculation would not be sufficient to produce an accurate estimate. Alternatively, another method to estimate the barrier would be to run many simulations of B-RNA, each with a distinct random number seed, and find the average amount of time for the B-RNA to A-RNA conversion. This would only work if the barrier is in a region reasonable given the amount of computational resources available (there is no guarantee the transition will ever occur), but if it is, the barrier can be estimated fairly confidently, although the estimate will be based on specific aspects of the force field.

3.5 Acknowledgements

D.E.K. would like to acknowledge the Computer Graphics Lab at the University of California, San Francisco, for use of their visualization and server resources. ‘Molecular graphics images were produced using the MidasPlus program from the UCSF Computer Graphics Laboratory, University of California, San Francisco (supported by NIH P41-RR01081).

3.6 Tables

Name	Forcefield	Composition	Starting Conformation	Length in ns
ALLDNA_96	Cornell	D	B	1.06
OKAZAKI_A_96	Cornell	D/RD	A	1.98
OKAZAKI_B_96	Cornell	D/RD	B	1.99
CYTARABINE_96	Cornell	D/RD	B	3.35
OKBstart_ALLRNA_parm96	Cornell	R/R	B	10.07
OKBstart_ALLRNA_parm99	Wang	R/R	B	10.10
OKAstart_ALLRNA_parm96	Cornell	R/R	B	10.09
OKAstart_ALLRNA_parm99	Wang	R/R	B	10.07

Table 3.1: List of simulations performed. ALLDNA_96 refers to simulation of Okazaki fragment sequence with all DNA residues. OKAZAKI_A_96 refers to simulation of Okazaki fragment sequence started in canonical A conformation. OKAZAKI_B_96 refers to simulation of Okazaki fragment sequence started in canonical B conformation. CYTARABINE_96 refers to simulation of Okazaki fragment sequence with cytarabine at residue 20. OKBstart_ALLRNA_parm96 refers to Okazaki fragment sequence with all RNA residues started in the canonical B conformation with the Cornell et al force field. OKAstart_ALLRNA_parm96 refers to Okazaki fragment sequence with all RNA residues started in the canonical A conformation with the Cornell et al force field. OKBstart_ALLRNA_parm99 refers to Okazaki fragment sequence with all RNA residues started in the canonical B conformation with the Wang et al force field. OKAstart_ALLRNA_parm99 refers to Okazaki fragment sequence with all RNA residues started in the canonical A conformation with the Wang et al force field.

	BDNA	ARNA	Astart parm96	Astart parm99	Bstart parm96	Bstart parm99
BDNA		5.942	6.553	5.94	4.616	5.326
ARNA	3.164		1.882	1.857	5.442	5.411
Astart parm96	3.159	1.182		1.249	5.663	5.486
Astart parm99	2.815	1.494	0.703		4.981	4.774
Bstart parm96	2.165	2.385	1.96	1.825		3.25
Bstart parm99	2.31	1.988	1.573	1.457	1.278	

Table 3.2: Pairwise RMS between static structures. BDNA and ADNA are canonical B and canonical A (Arnott) structures generated by the "nucgen" program of the AMBER suite. Astart_parm96 through Bstart_parm99 are average structures of the final 1ns of the "ALLRNA" simulations listed in Table 3.1. RMS is in Å, computed for all atoms, mass-weighted. The upper right diagonal matrix contains RMS computed using all residues while the lower left contains RMS computed only for the internal hexamer.

	BDNA	ARNA	Astart parm96(1st)	Astart parm99(1st)	Bstart parm96(1st)	Bstart parm99(1st)
BDNA		5.942	6.436	6.508	1.872	1.696
ARNA	3.164		1.679	1.551	5.086	5.911
Astart parm96(1st)	3.409	1.071		1.588	5.549	6.394
Astart parm99(1st)	3.097	0.851	1.112		5.622	6.351
Bstart parm96(1st)	1.464	3.187	3.318	3.008		1.992
Bstart parm99(1st)	1.186	3.447	3.605	3.253	1.214	

Table 3.3: Pairwise RMS between static structures. BDNA and ADNA are canonical B and canonical A (Arnott) structures generated by the "nucgen" program of the AMBER suite. Astart_parm96 through Bstart_parm99 are the first snapshot of the "ALLRNA" simulations listed in Table 3.1. RMS is in Å, computed for all atoms, mass-weighted. The upper right diagonal matrix contains RMS computed using all residues while the lower left contains RMS computed only for the internal hexamer.

Name	Etot	Ebond	Eangle	Evdw	E14vdw	Eeel	E14eel	Edihed
BDNA(nomin)	882	940	641	169	307	-860	-769	452
BRNA(nomin)	10320	1393	640	8623	514	-648	-727	525
ARNA(nomin)	1534	931	764	4	406	-307	-705	441
ADNA(nomin)	3891	1318	1076	2255	257	-657	-742	384
BDNA(min)	-1298	28	136	-257	205	-990	-829	409
BRNA(min)	-957	29	231	-247	202	-774	-851	453
ARNA(min)	-799	25	124	-363	186	-377	-815	421
ADNA(min)	-965	29	156	-285	191	-625	-852	421
OKBstart parm96 start	-927 ± 51	192 ± 24	440 ± 8	-373 ± 8	208 ± 5	-1045 ± 37	-826 ± 11	477 ± 5
OKBstart parm99 start	-785 ± 28	196 ± 14	439 ± 21	-364 ± 9	205 ± 2	-995 ± 13	-829 ± 5	563 ± 5
OKAstart parm96 start	-545 ± 55	195 ± 21	386 ± 4	-408 ± 4	211 ± 7	-625 ± 32	-813 ± 5	509 ± 10
OKAstart parm99 start	-423 ± 63	194 ± 8	389 ± 15	-408 ± 11	216 ± 6	-576 ± 30	-814 ± 9	576 ± 15
OKBstart parm96 end	-778 ± 64	203 ± 10	435 ± 12	-383 ± 9	208 ± 7	-906 ± 46	-820 ± 9	485 ± 6
OKBstart parm99 end	-653 ± 42	191 ± 7	428 ± 7	-384 ± 3	212 ± 7	-861 ± 34	-823 ± 5	584 ± 7
OKAstart parm96 end	-596 ± 25	201 ± 10	406 ± 15	-407 ± 7	213 ± 4	-715 ± 28	-809 ± 14	515 ± 5
OKAstart parm99 end	-447 ± 38	197 ± 9	390 ± 11	-405 ± 6	212 ± 8	-617 ± 31	-811 ± 15	587 ± 12

Table 3.4: Average and standard deviation of AMBER energy for static structures and simulations. Energy is calculated using the Cornell et al force field (parm96) unless noted (Wang et al is "parm99"). "nomin" means the energy was computed for the unminimized canonical form generated by "nucgen" while "min" means the structure was minimized for 10000 steps (5000 steepest descent followed by 5000 conjugate gradient). The minimization typically terminated before reaching 10000 steps showing that a local energy minimum was found. "start" means that the first 10 snapshots from a simulation were used while "end" means the last 10 were used.

3.7 Figures

3.7.1 Figure Legends

Figure 3.1 to Figure 3.12 plot the average Dials and Windows parameters for several static structures and several dynamic simulations.. "Okazaki_Astart" refers to the simulation labelled "OKAZAKI_A_96" and 'Okazaki_Bstart" refers to the simulation labelled "OKAZAKI_B_96" (Table 3.1), both of which contain RNA residues at positions 13-17. "OKAstart_ALLRNA_parm96.9000.10000" refers to the "OKAstart_ALLRNA_parm96" simulation in Table 3.1 containing all RNA residues. ARNA and BDNA refer to canonical structures generated using the "nucgen" tool of the AMBER suite. nucgen generates Arnott (REF) canonical A and B conformations. XDP stands for "x displacement" and "INC" stands for "inclination" defined by Dials and Windows (REF). All time averages were computed over the last 1ns of the trajectory, and no average was computed for the static structures.

Figure 3.25 to Figure 3.40 plots the time course of Dials and Windows parameters for the simulations containing all RNA residues. Okazaki fragment free MD refers to the 10ns Okazaki simulations of pure RNA duplexes. OKAstart-parm96 refers to the simulation started in the canonical A form with the parm96 (Cornell et al) force field. OKAstart-parm99 started in canonical A with the parm99 (Wang et al) force field. OKBstart-parm96 started in canonical B with parm96 and OKBstart-parm99 started in canonical B with parm99. For parameters specific to a base, strand one is black, and strand two is gray. For parameters specific to a base pair or pair of base pairs, only black is used. In the histogram plots, the empty bars represent the first strand and the filled bars represent the second strand.

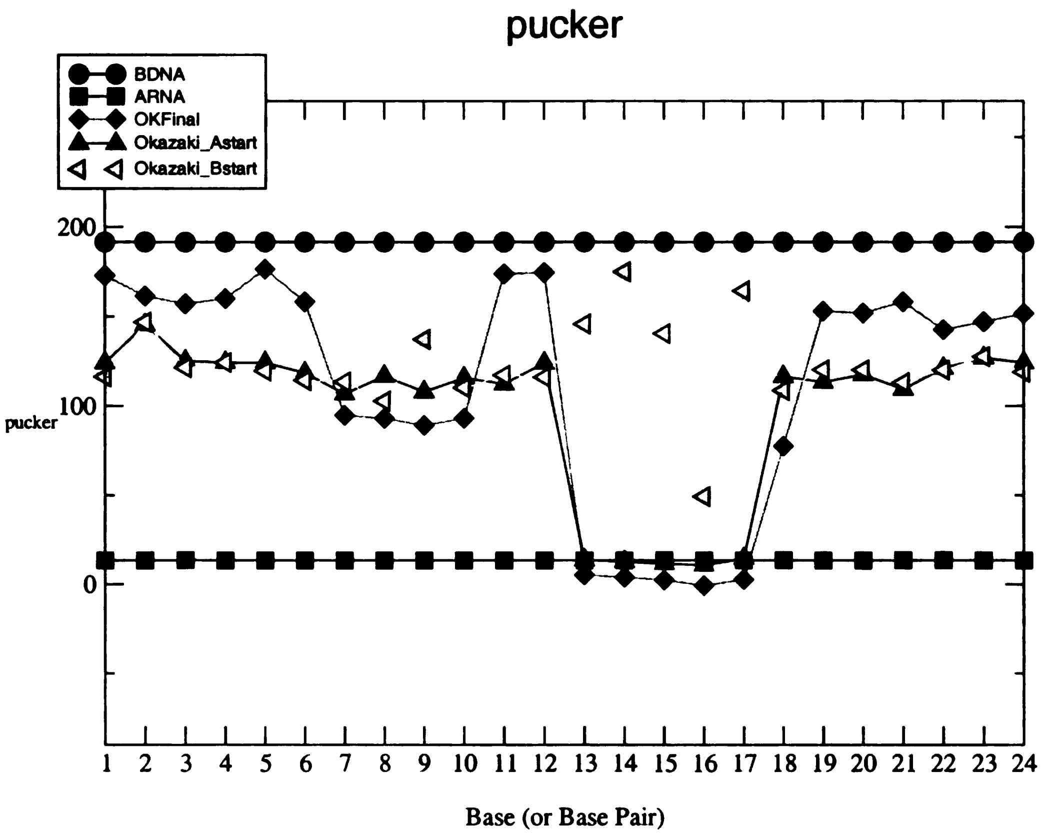


Figure 3.1: Time average from time=9ns to time=10ns of pucker in Canonical A, Canonical B and Okazaki fragment free MD simulation (NMR-v-freeMD)

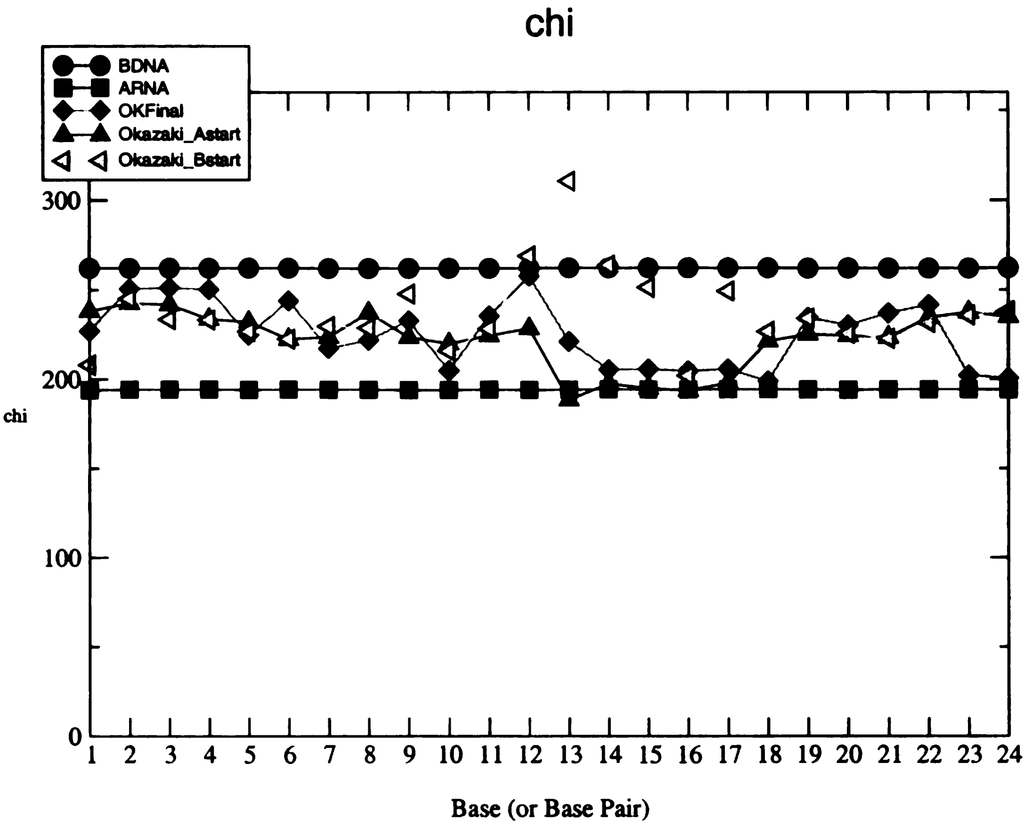


Figure 3.2: Time average from time=9ns to time=10ns of chi in Canonical A, Canonical B and Okazaki fragment free MD simulation (NMR-v-freeMD)

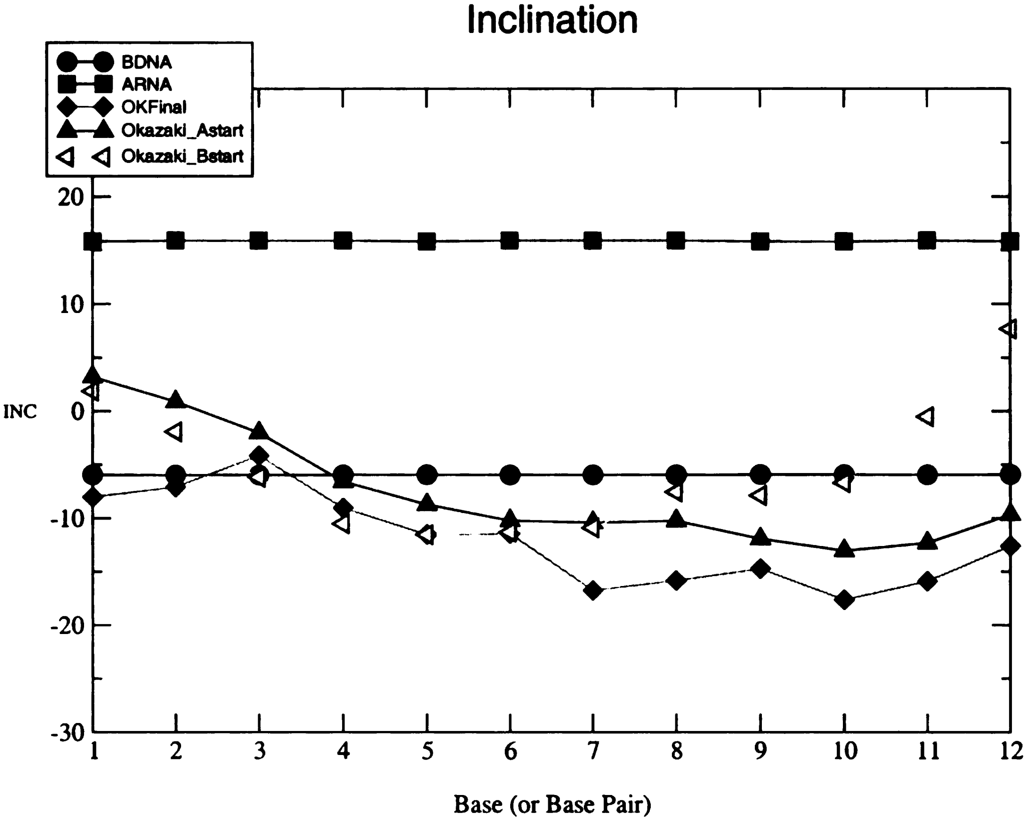


Figure 3.3: Time average from time=9ns to time=10ns of INC in Canonical A, Canonical B and Okazaki fragment free MD simulation (NMR-v-freeMD)

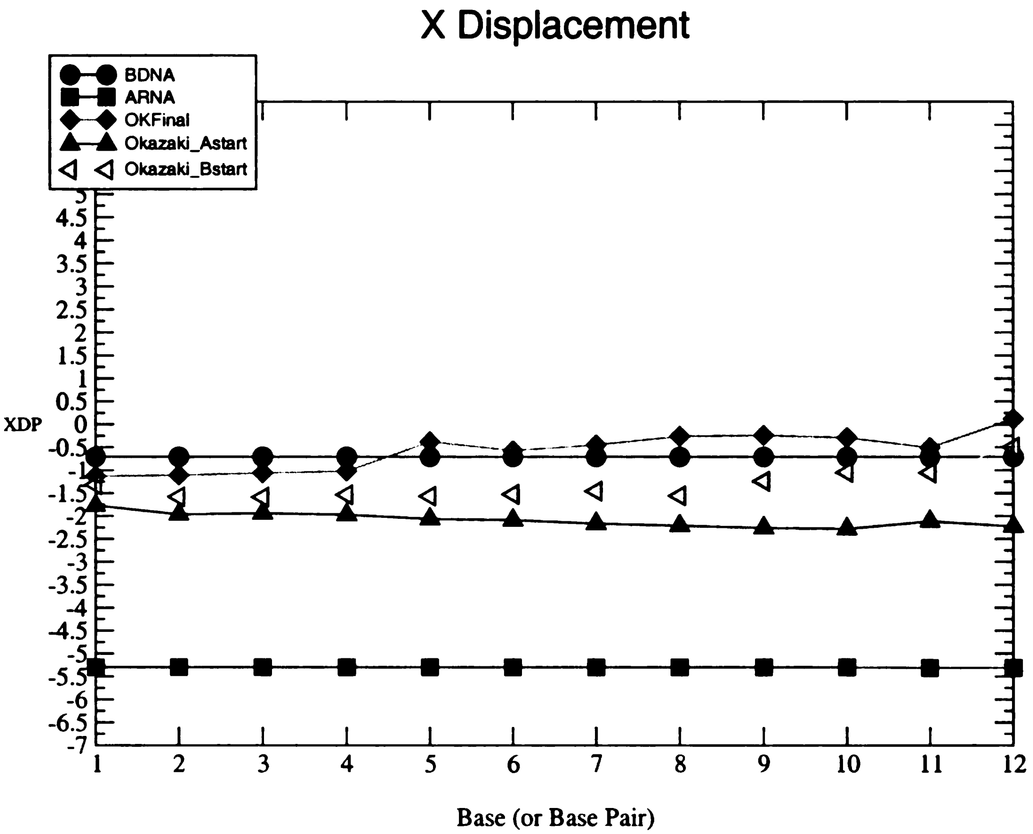


Figure 3.4: Time average from time=9ns to time=10ns of XDP in Canonical A, Canonical B and Okazaki fragment free MD simulation (NMR-v-freeMD)

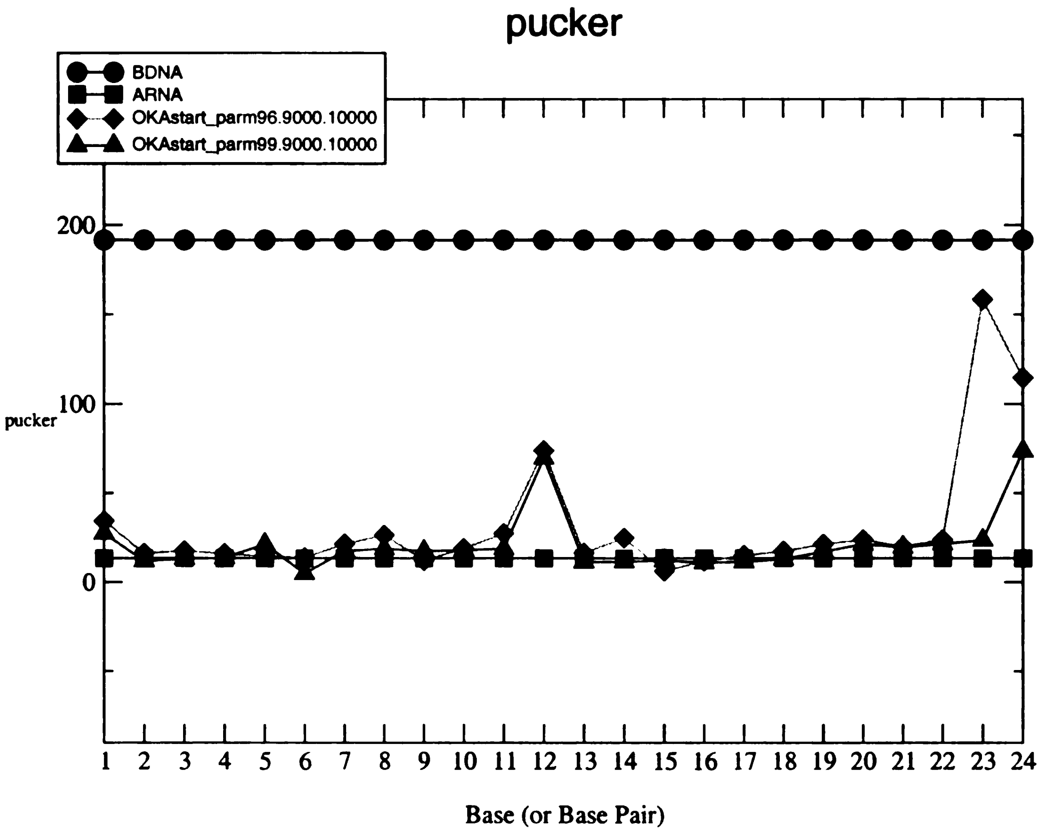


Figure 3.5: Time average from time=9ns to time=10ns of pucker in Canonical A, Canonical B and Okazaki fragment free MD simulation (Aform96v99)

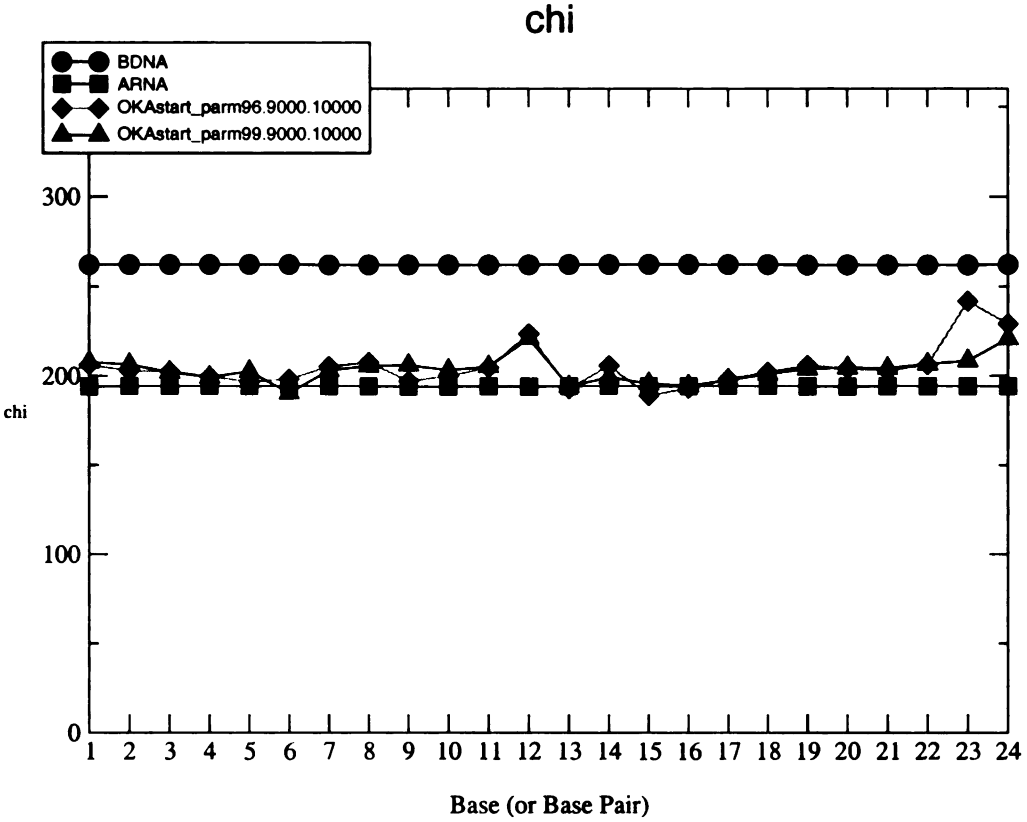


Figure 3.6: Time average from time=9ns to time=10ns of chi in Canonical A, Canonical B and Okazaki fragment free MD simulation (Aform96v99)

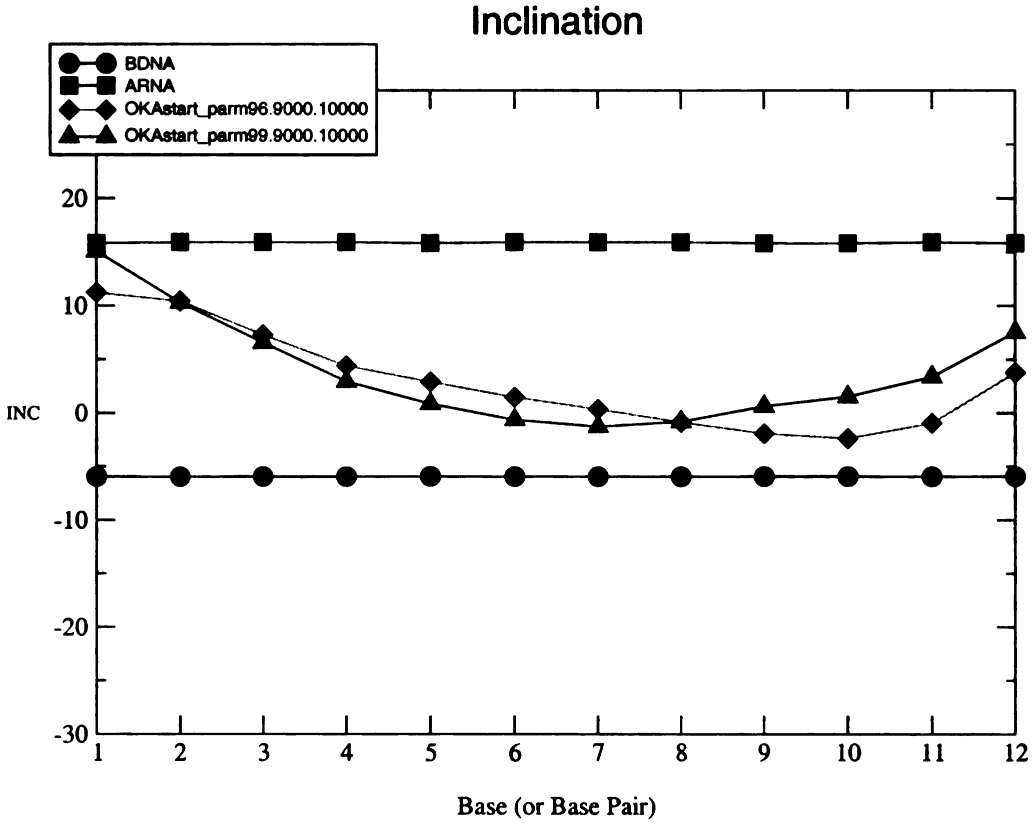


Figure 3.7: Time average from time=9ns to time=10ns of INC in Canonical A, Canonical B and Okazaki fragment free MD simulation (Aform96v99)

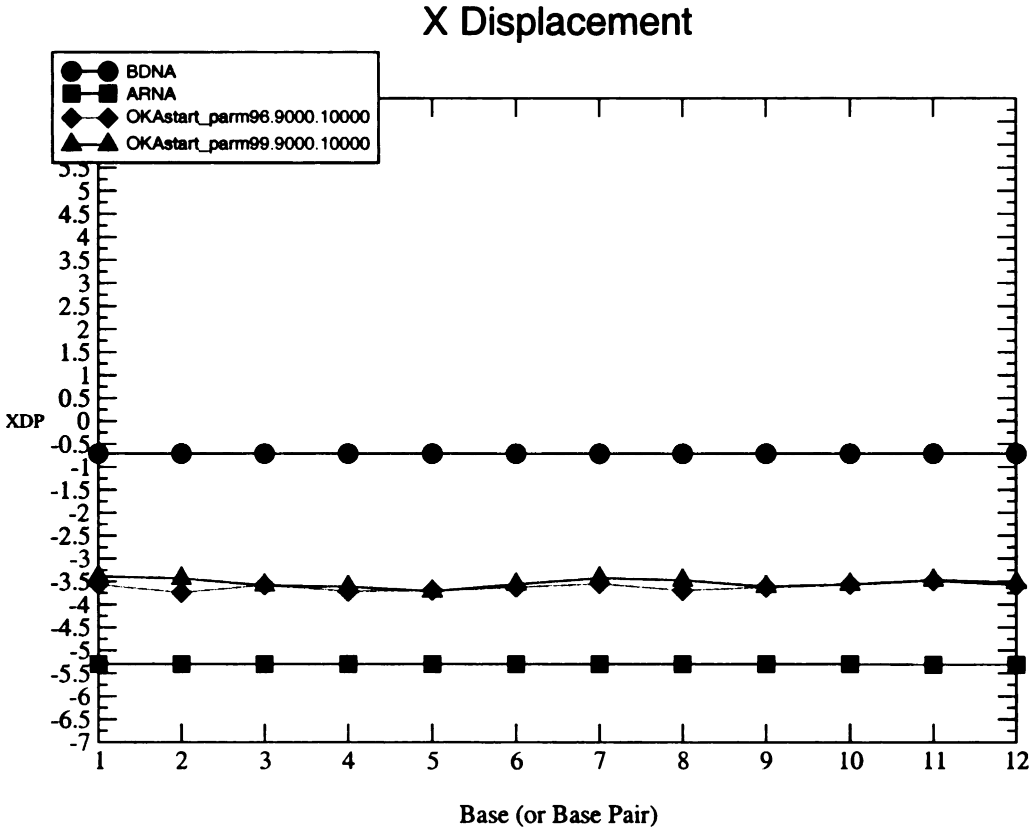


Figure 3.8: Time average from time=9ns to time=10ns of XDP in Canonical A, Canonical B and Okazaki fragment free MD simulation (Aform96v99)

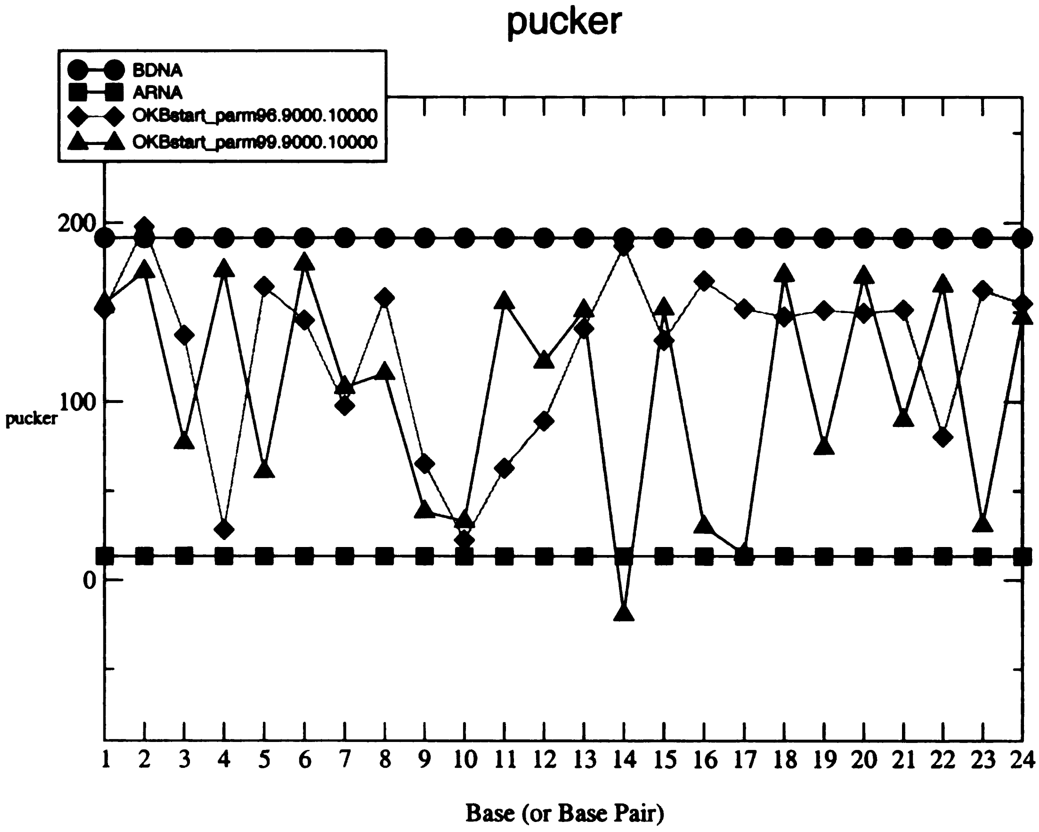


Figure 3.9: Time average from time=9ns to time=10ns of pucker in Canonical A, Canonical B and Okazaki fragment free MD simulation (Bform96v99)

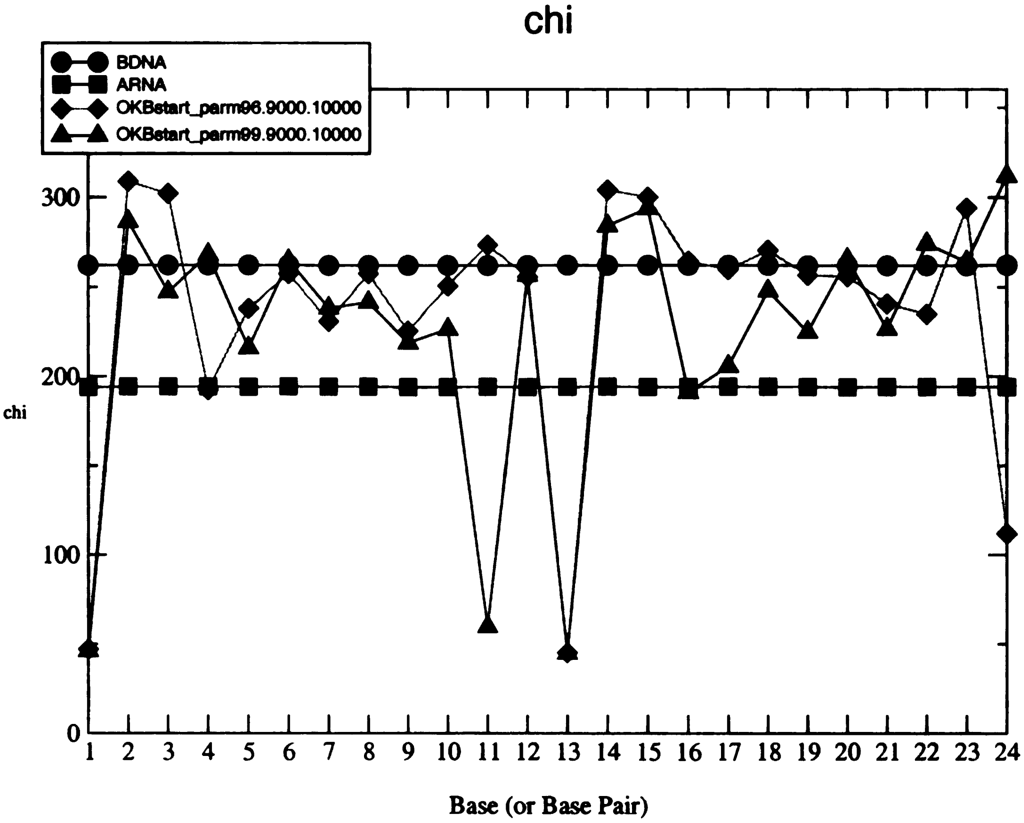


Figure 3.10: Time average from time=9ns to time=10ns of chi in Canonical A, Canonical B and Okazaki fragment free MD simulation (Bform96v99)

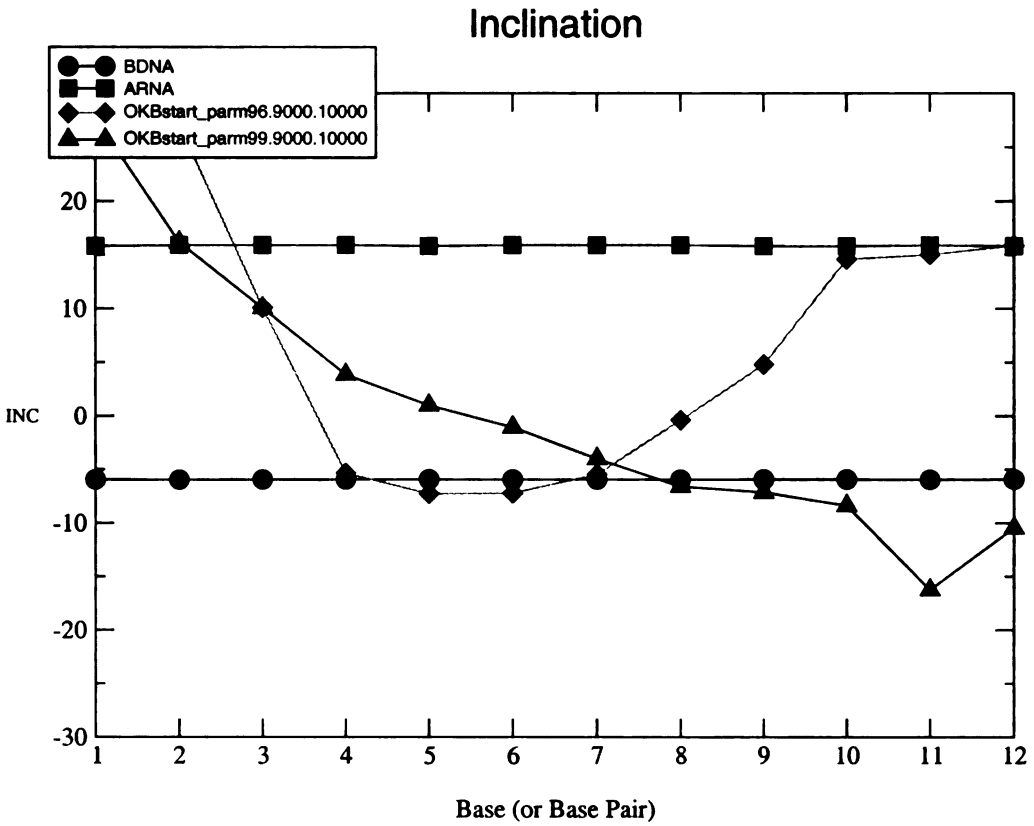


Figure 3.11: Time average from time=9ns to time=10ns of INC in Canonical A, Canonical B and Okazaki fragment free MD simulation (Bform96v99)

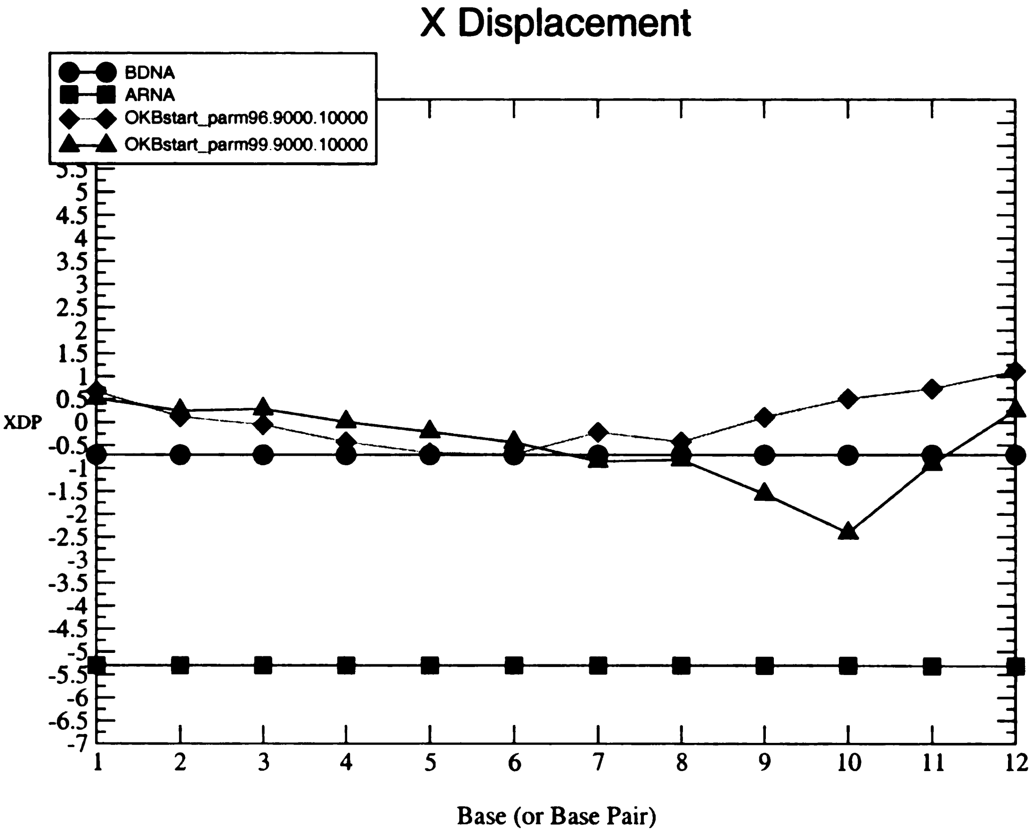


Figure 3.12: Time average from time=9ns to time=10ns of XDP in Canonical A, Canonical B and Okazaki fragment free MD simulation (Bform96v99)

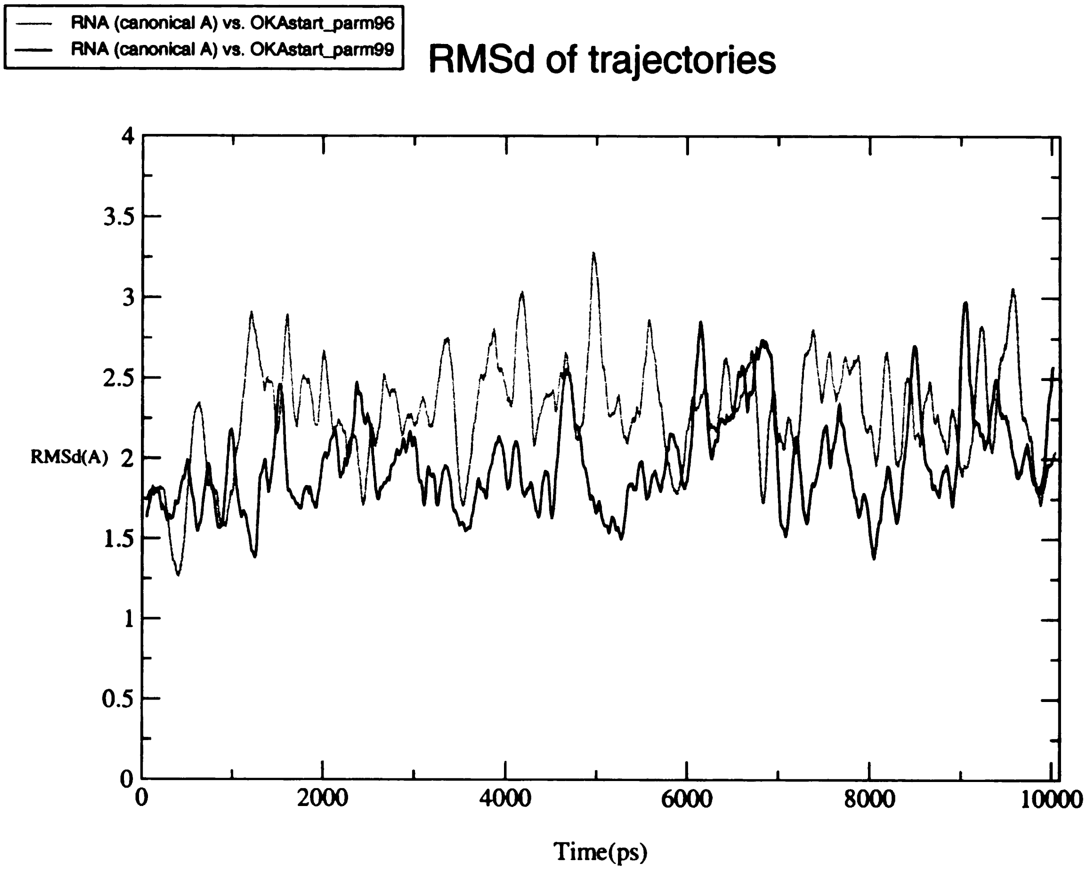


Figure 3.13: Time course of RMS of Okazaki sequence RNA simulation starting in canonical A form against canonical A

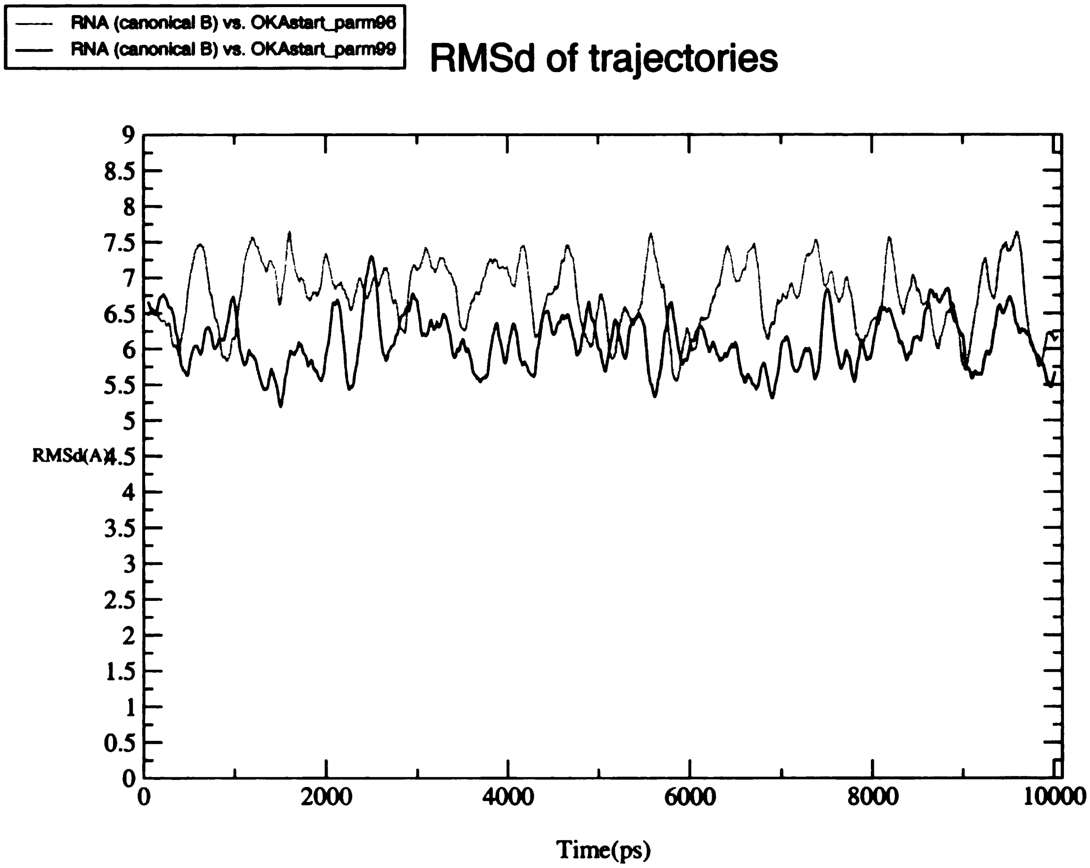


Figure 3.14: Time course of RMS of Okazaki sequence RNA simulation starting in canonical A form against canonical B

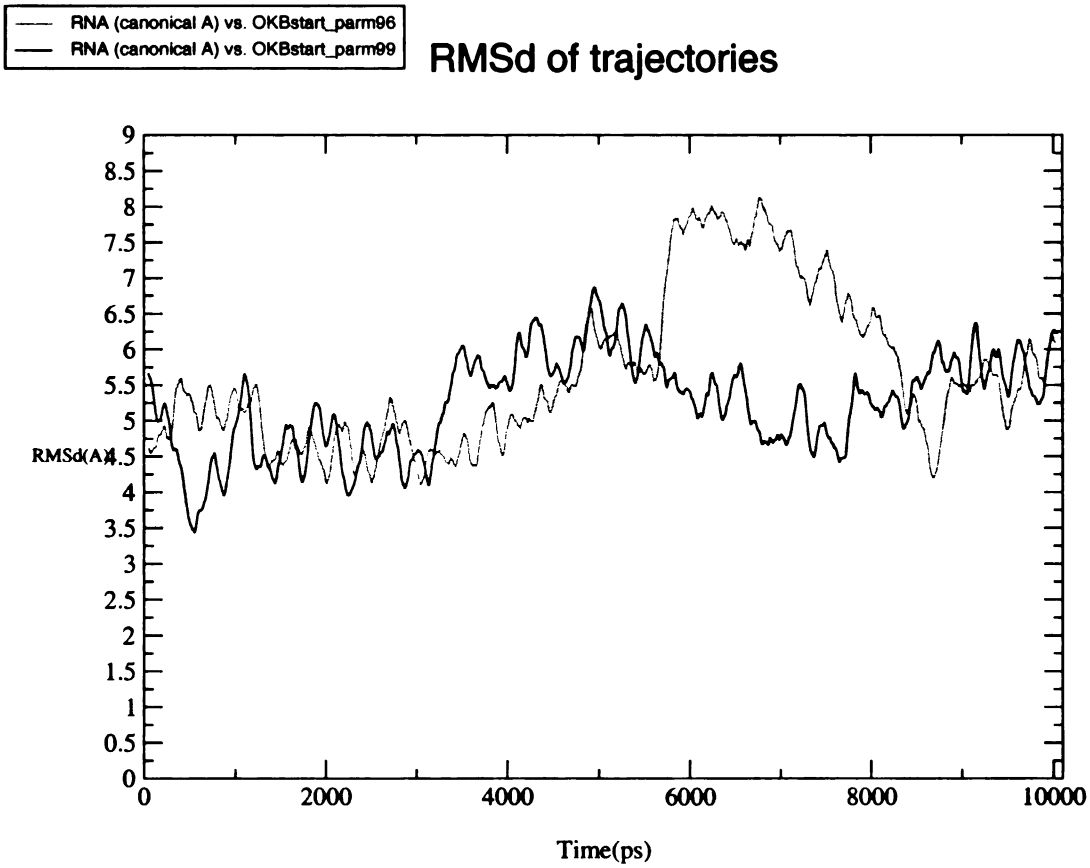


Figure 3.15: Time course of RMS of Okazaki sequence RNA simulation starting in canonical B form against canonical A

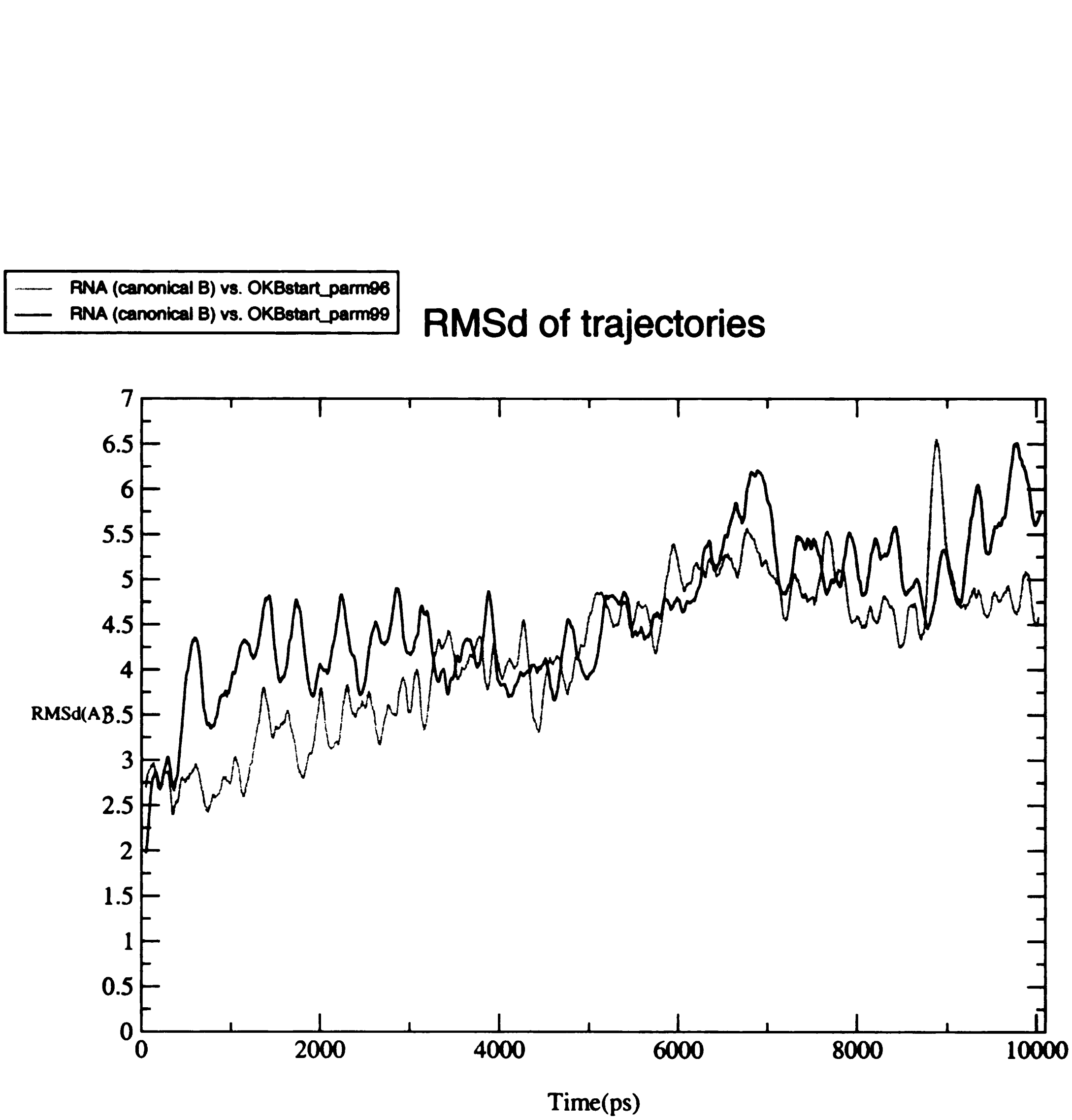


Figure 3.16: Time course of RMS of Okazaki sequence RNA simulation starting in canonical B form against canonical B

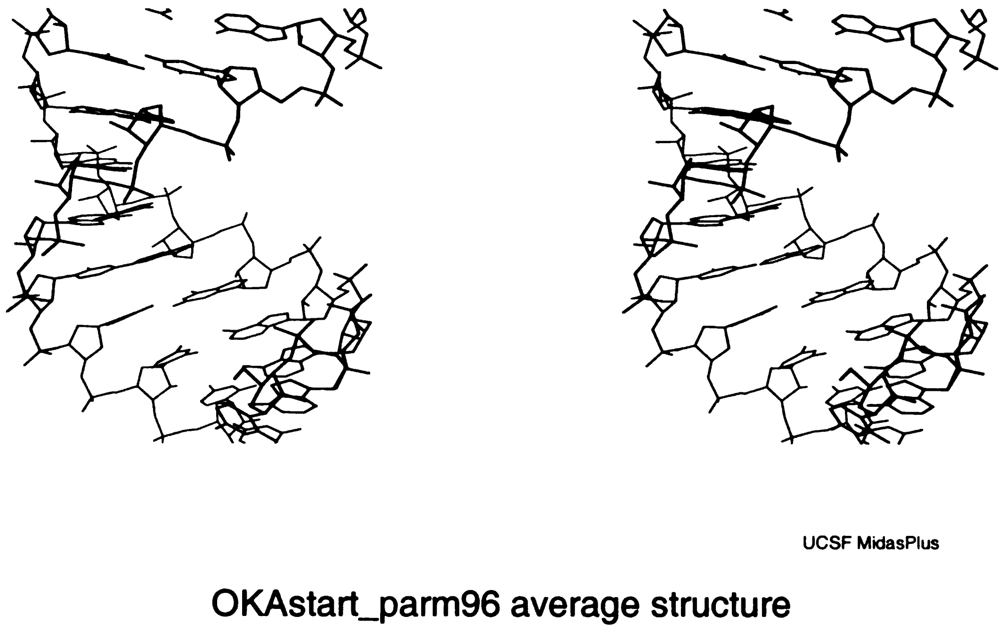


Figure 3.17: View of OKAstart_parm96 structure (minimized average structure of last 1ns)

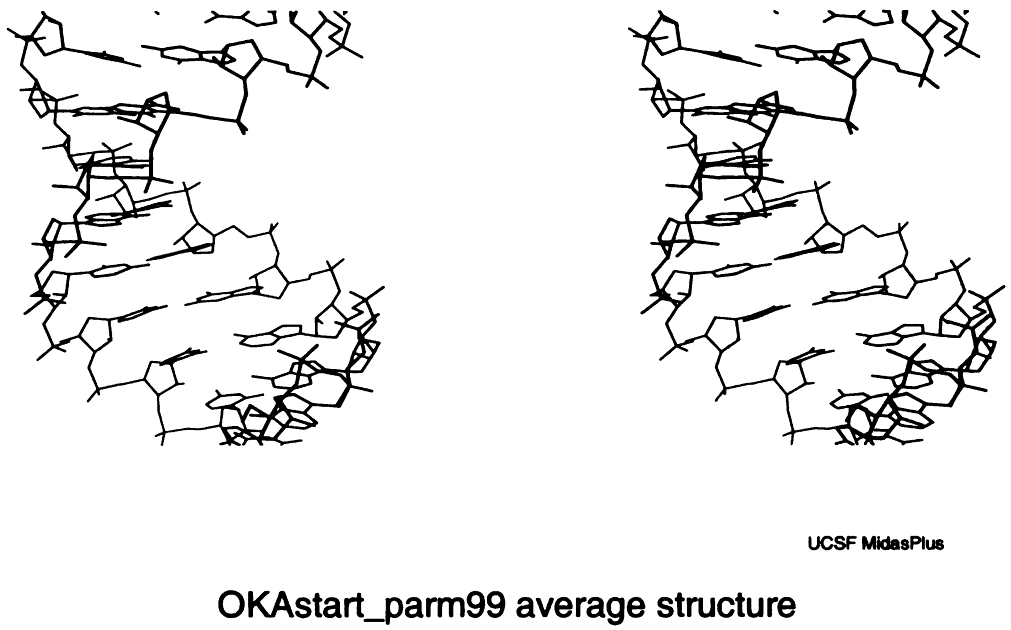


Figure 3.18: View of OKAstart_parm99 structure (minimized average structure of last 1ns)

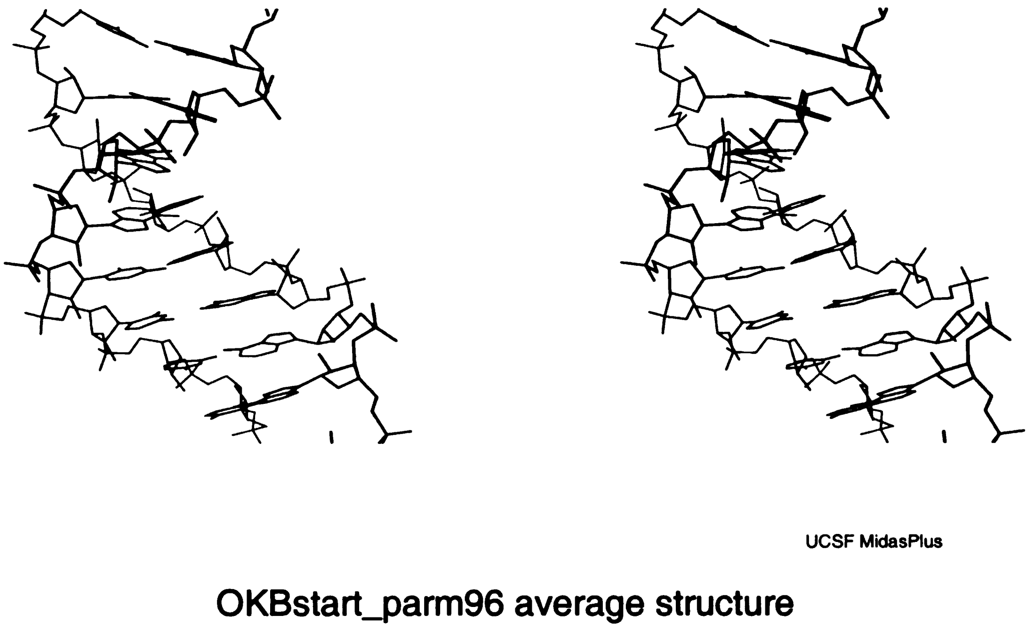


Figure 3.19: View of OKBstart_parm96 structure (minimized average structure of last 1ns)

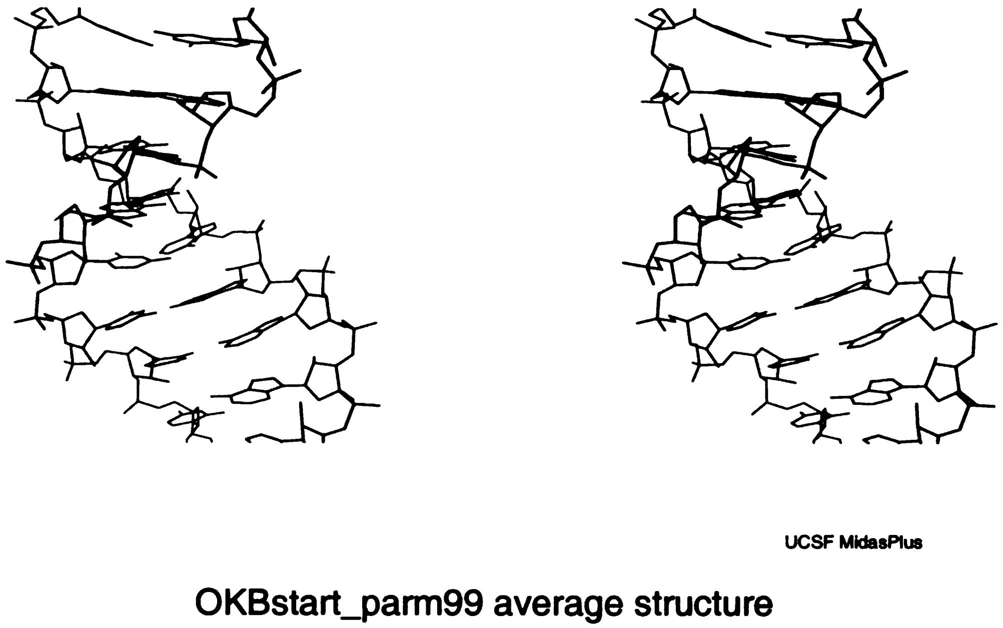


Figure 3.20: View of OKBstart_parm99 structure (minimized average structure of last 1ns)

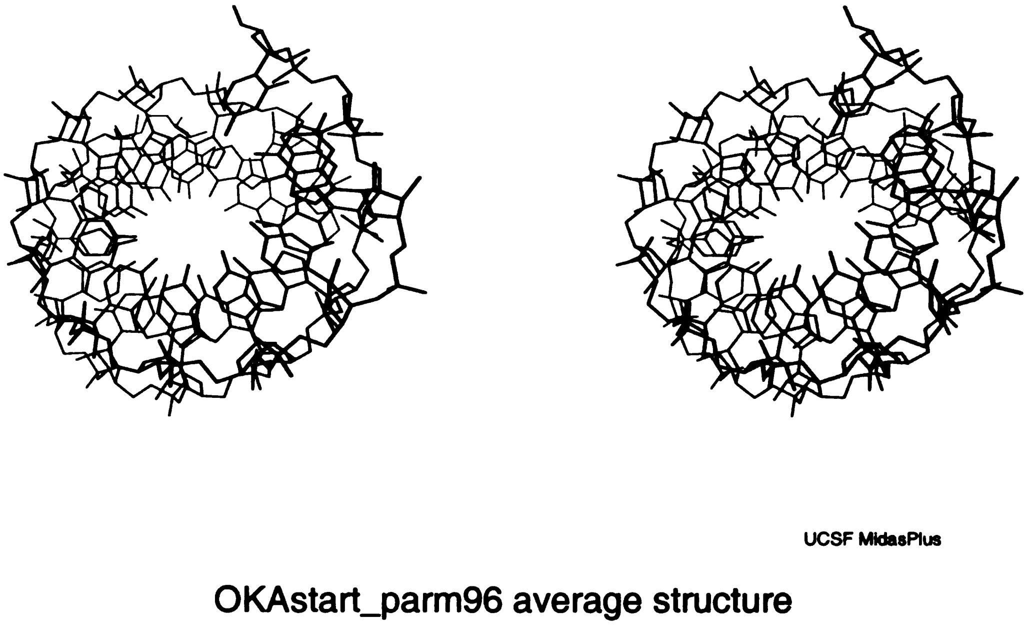


Figure 3.21: Axis view of OKAstart_parm96 structure (minimized average structure of last 1ns)

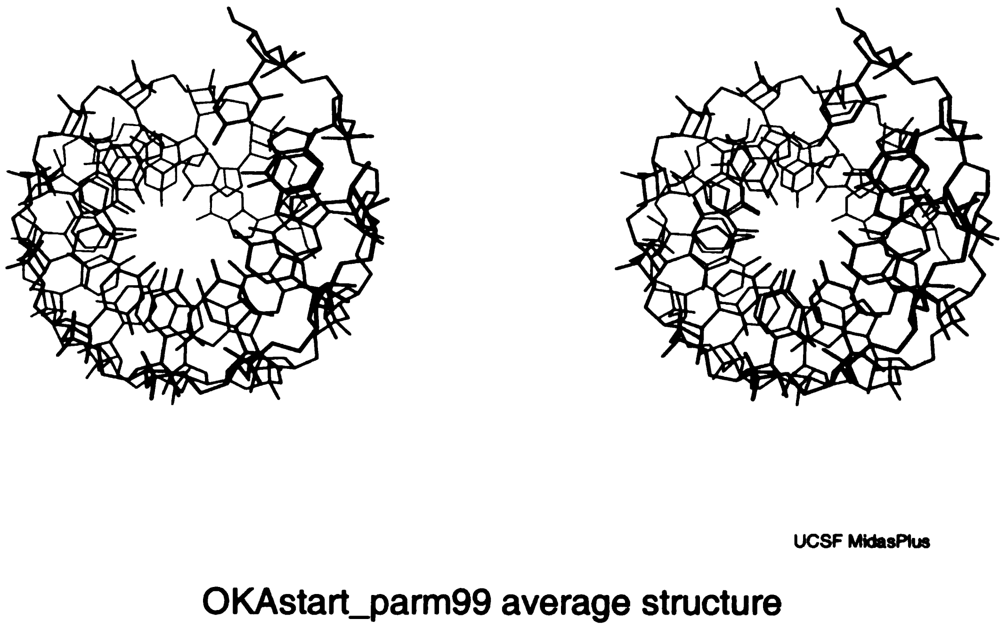


Figure 3.22: Axis view of OKAstart_parm99 structure (minimized average structure of last 1ns)

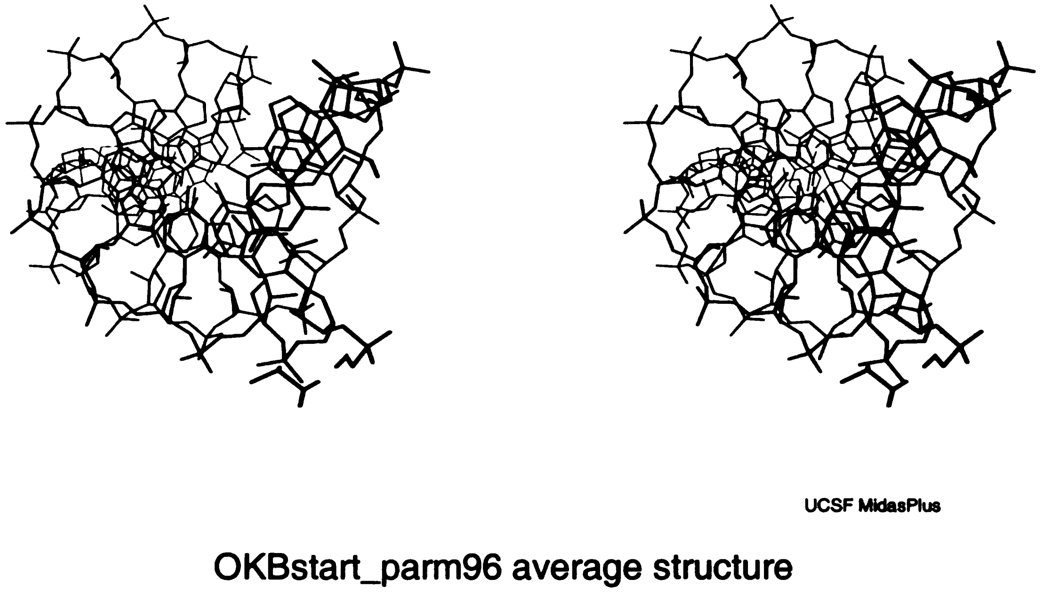


Figure 3.23: Axis view of OKBstart_parm96 structure (minimized average structure of last 1ns)

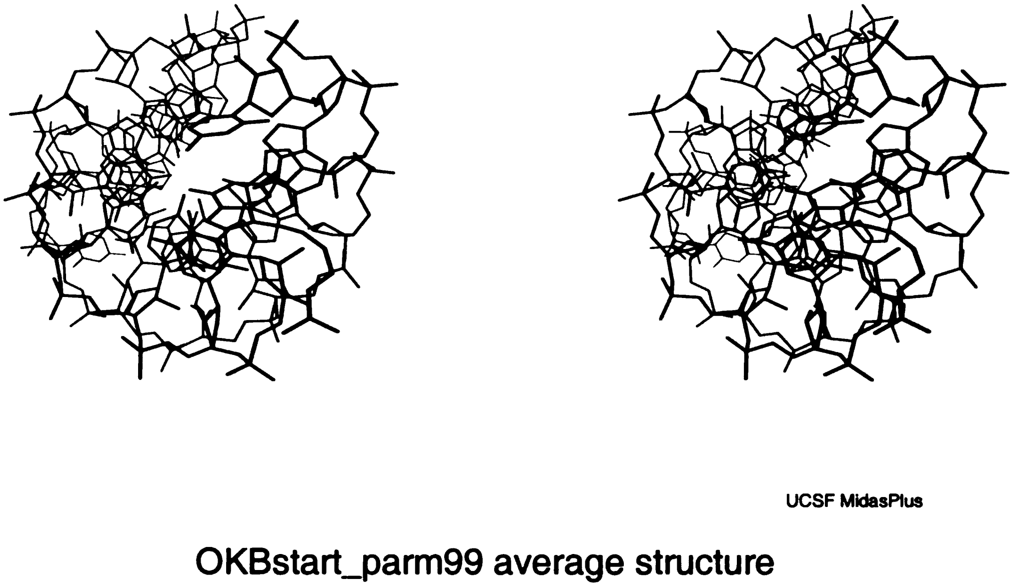


Figure 3.24: Axis view of OKBstart_parm99 structure (minimized average structure of last 1ns)

OKAstart_parm96, parameter pucker

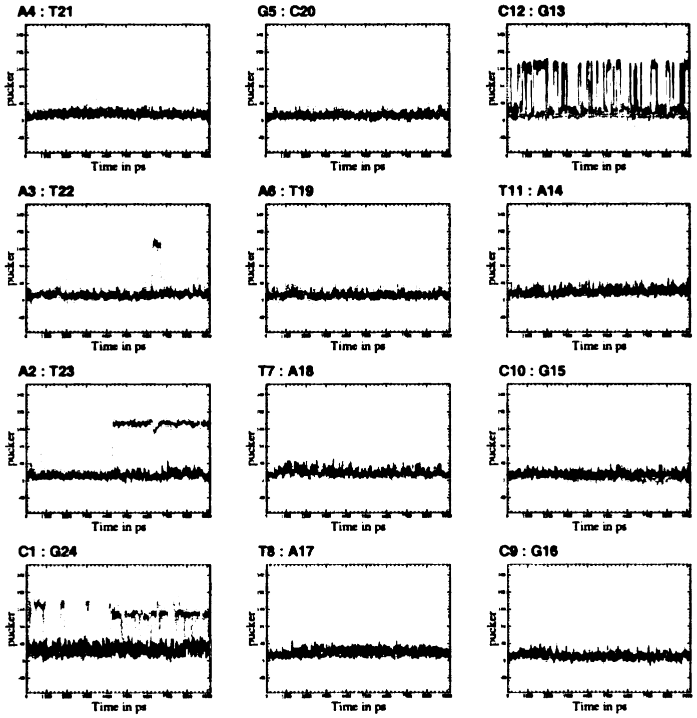


Figure 3.25: Time course of pucker in Okazaki fragment free MD (OKAstart-parm96)

OKAstart_parm96, parameter chi

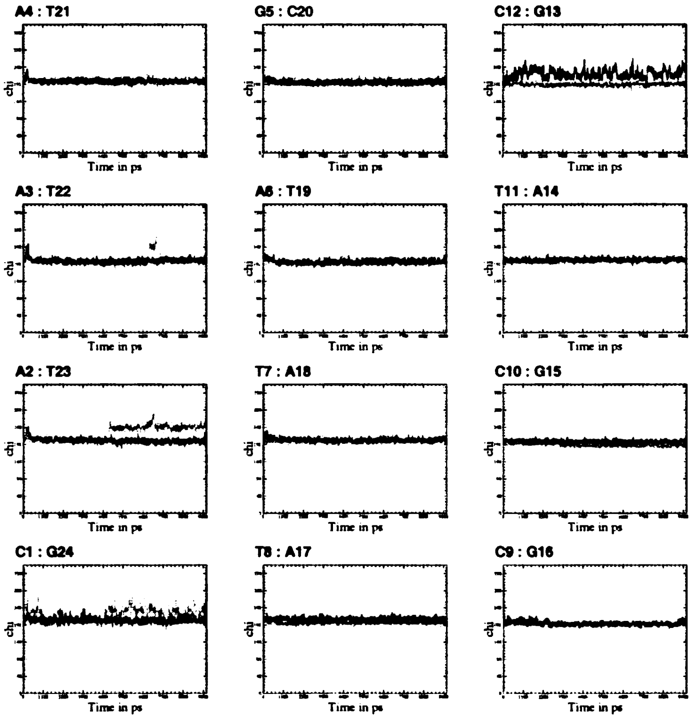


Figure 3.26: Time course of chi in Okazaki fragment free MD (OKAstart-parm96)

OKAstart_parm96, parameter INC

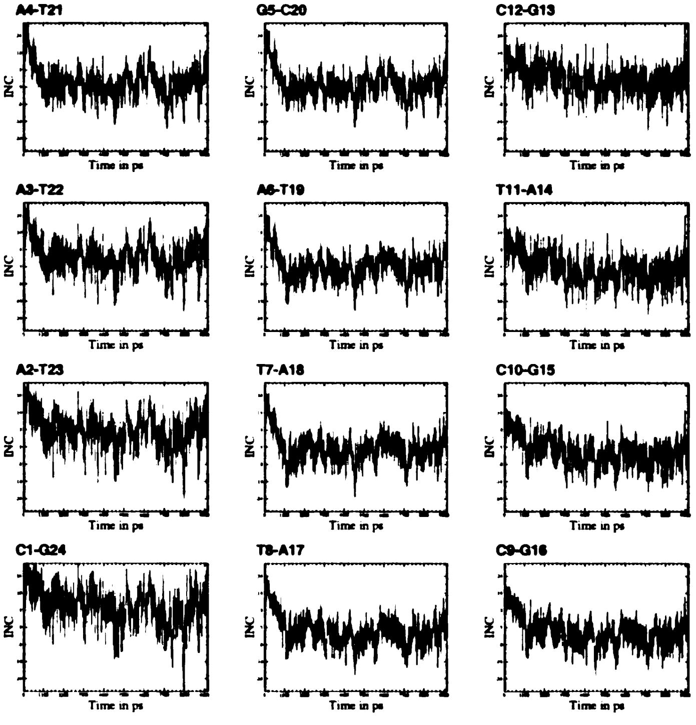


Figure 3.27: Time course of INC in Okazaki fragment free MD (OKAstart-parm96)

OKAstart_parm96, parameter XDP

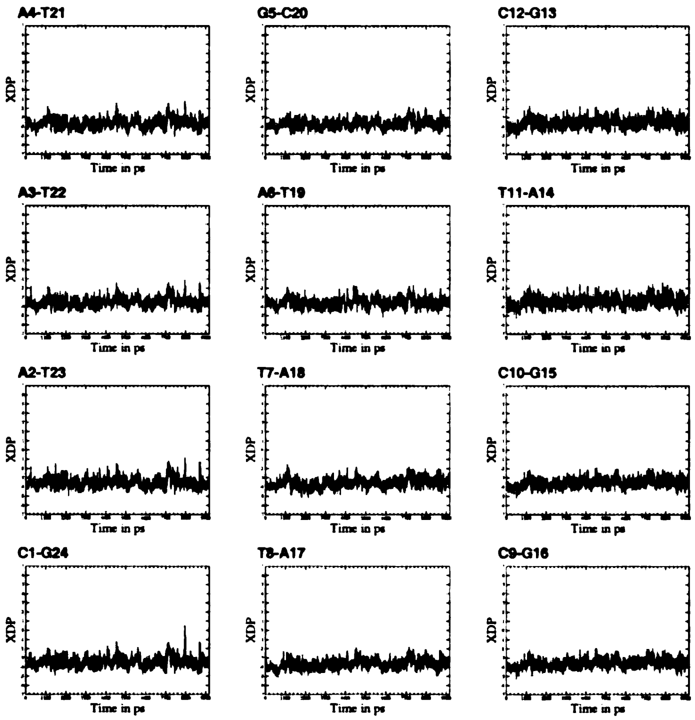


Figure 3.28: Time course of XDP in Okazaki fragment free MD (OKAstart-parm96)

OKAstart_parm99, parameter pucker

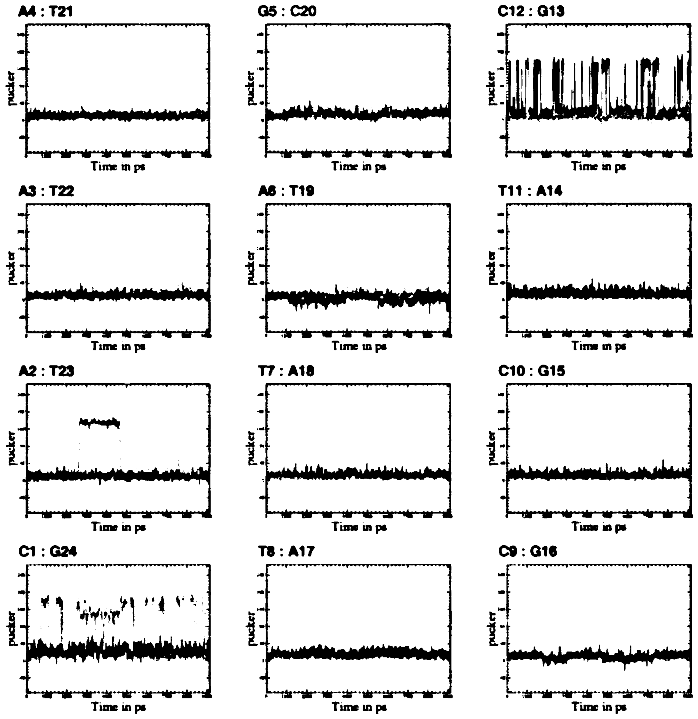


Figure 3.29: Time course of pucker in Okazaki fragment free MD (OKAstart-parm99)

OKAstart_parm99, parameter chi

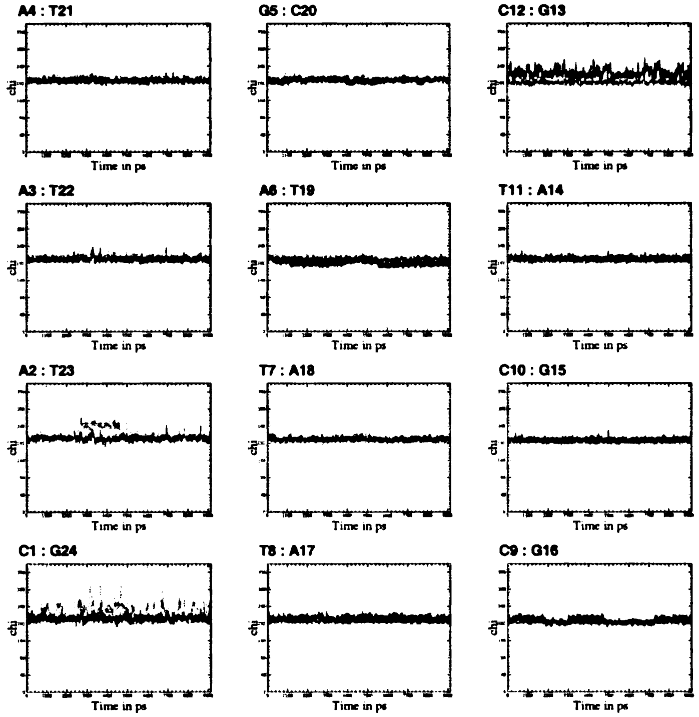


Figure 3.30: Time course of chi in Okazaki fragment free MD (OKAstart-parm99)

OKAstart_parm99, parameter INC

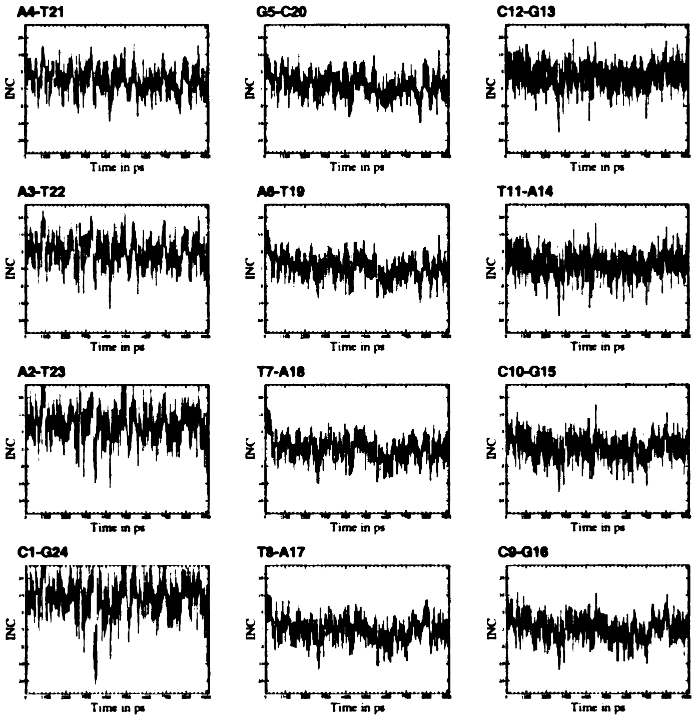


Figure 3.31: Time course of INC in Okazaki fragment free MD (OKAstart-parm99)

OKAstart_parm99, parameter XDP

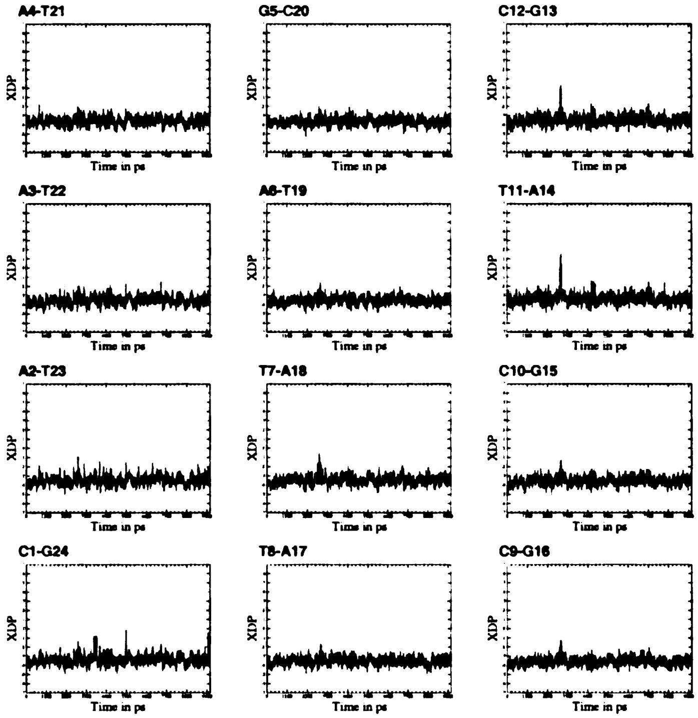


Figure 3.32: Time course of XDP in Okazaki fragment free MD (OKAstart-parm99)

OKBstart_parm96, parameter pucker

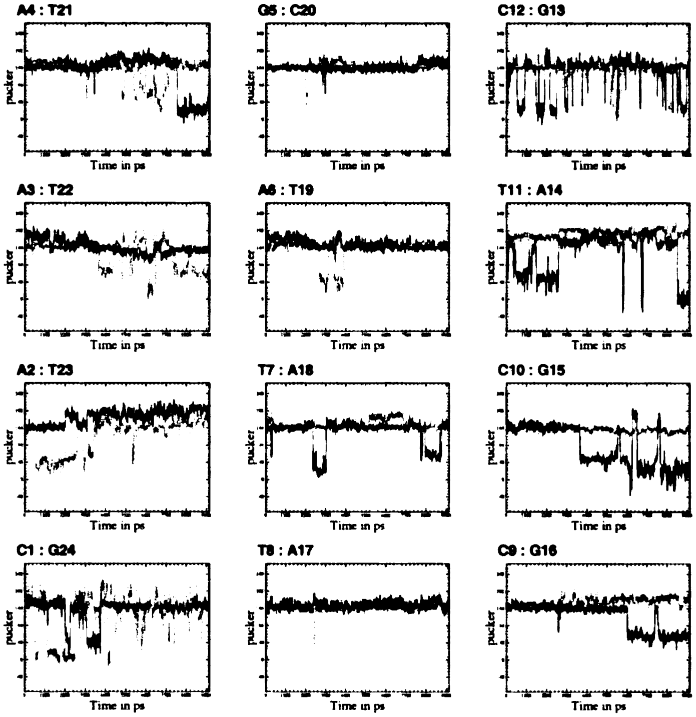


Figure 3.33: Time course of pucker in Okazaki fragment free MD (OKBstart-parm96)

OKBstart_parm96, parameter chi

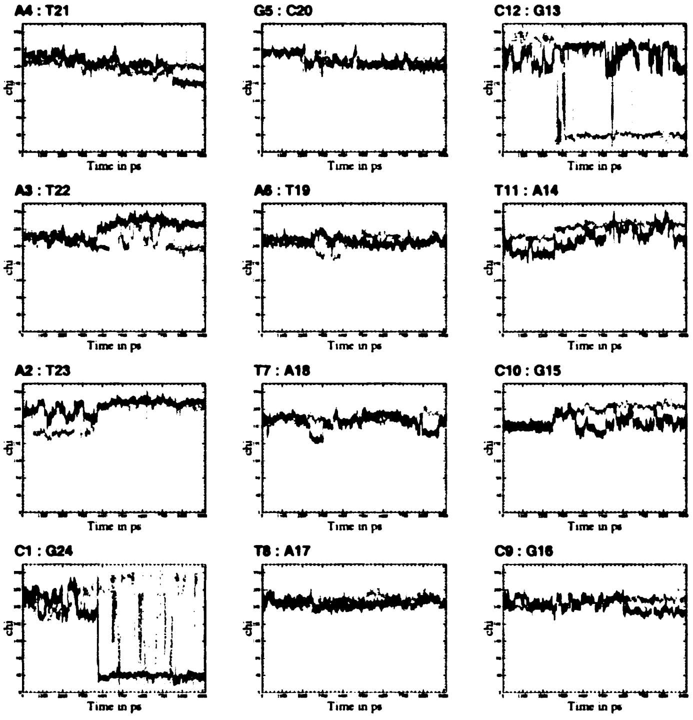


Figure 3.34: Time course of chi in Okazaki fragment free MD (OKBstart-parm96)

OKBstart_parm96, parameter INC

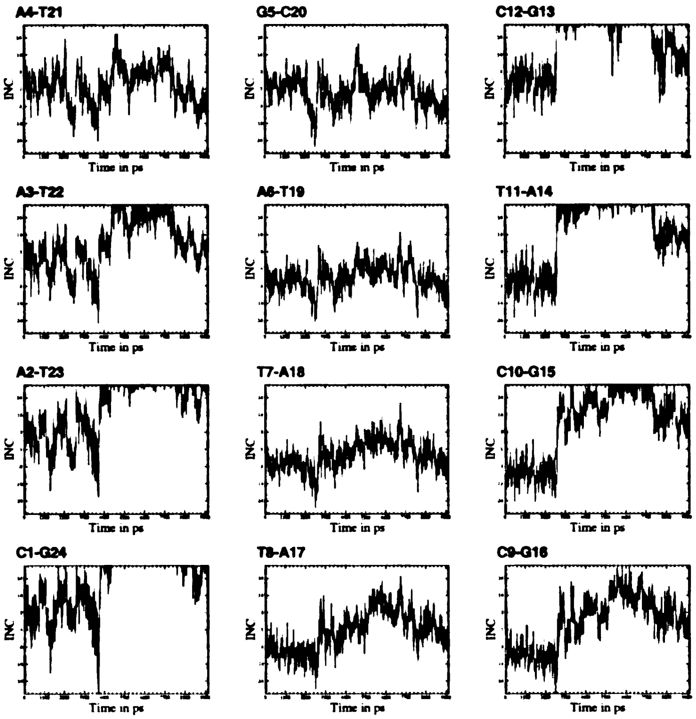


Figure 3.35: Time course of INC in Okazaki fragment free MD (OKBstart-parm96)

OKBstart_parm96, parameter XDP

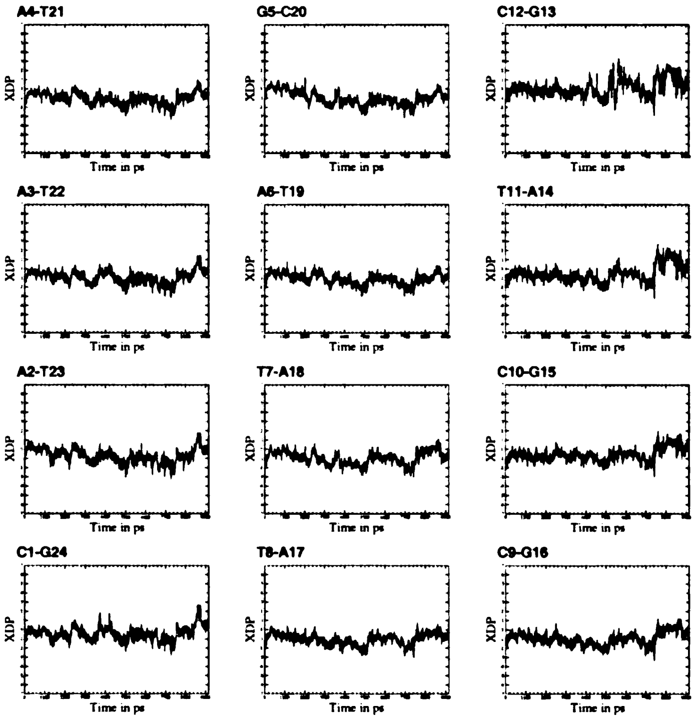


Figure 3.36: Time course of XDP in Okazaki fragment free MD (OKBstart-parm96)

OKBstart_parm99, parameter pucker

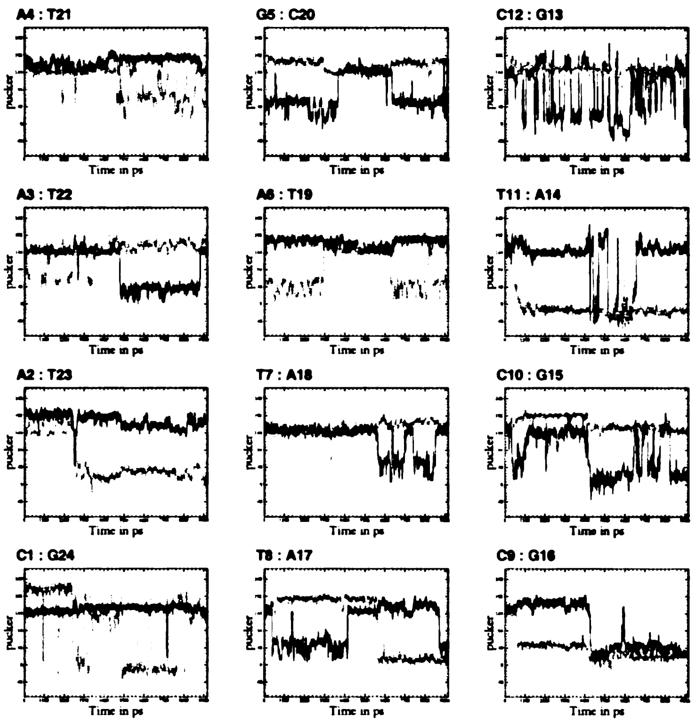


Figure 3.37: Time course of pucker in Okazaki fragment free MD (OKBstart-parm99)

OKBstart_parm99, parameter chi

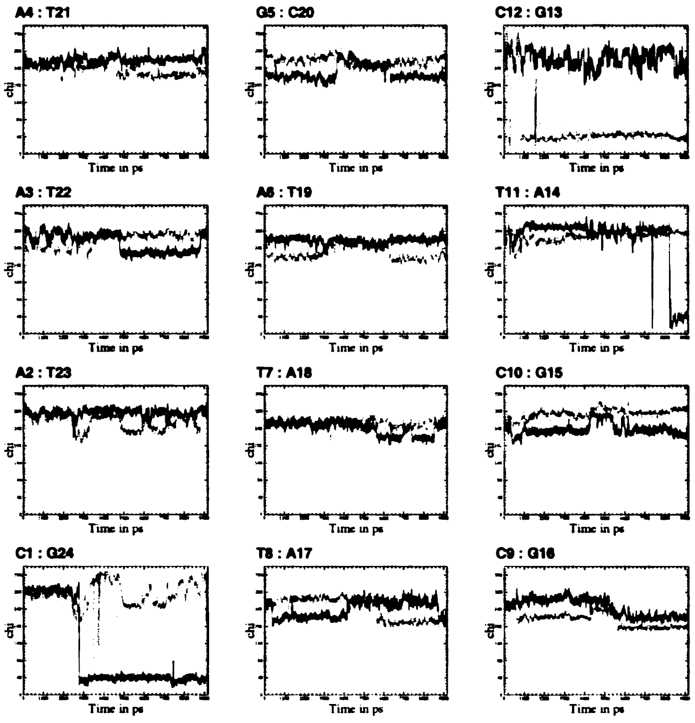


Figure 3.38: Time course of chi in Okazaki fragment free MD (OKBstart-parm99)

OKBstart_parm99, parameter INC

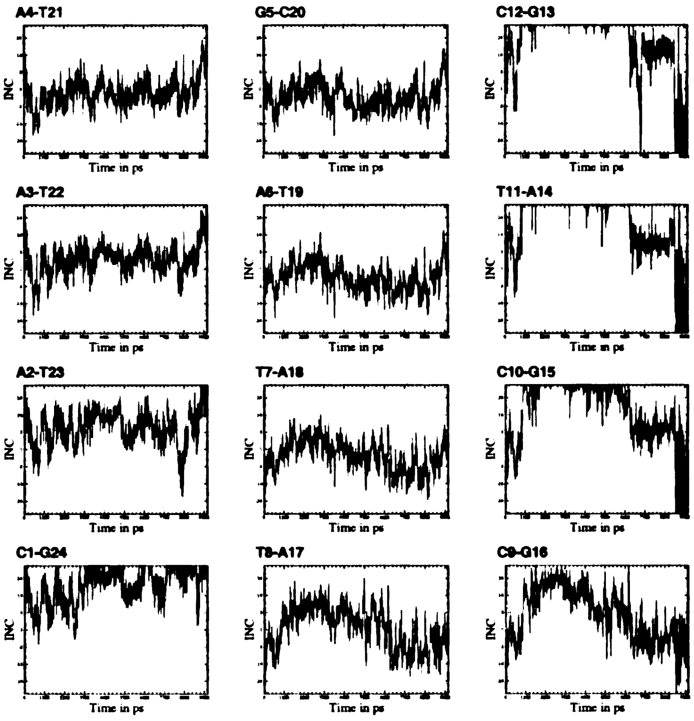


Figure 3.39: Time course of INC in Okazaki fragment free MD (OKBstart-parm99)

OKBstart_parm99, parameter XDP

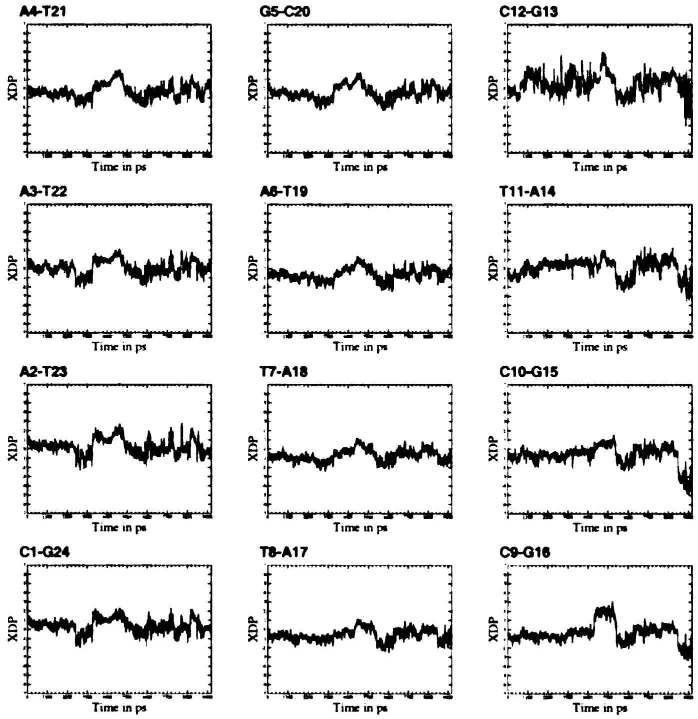


Figure 3.40: Time course of XDP in Okazaki fragment free MD (OKBstart-parm99)

HO'2 Distances for residue 5

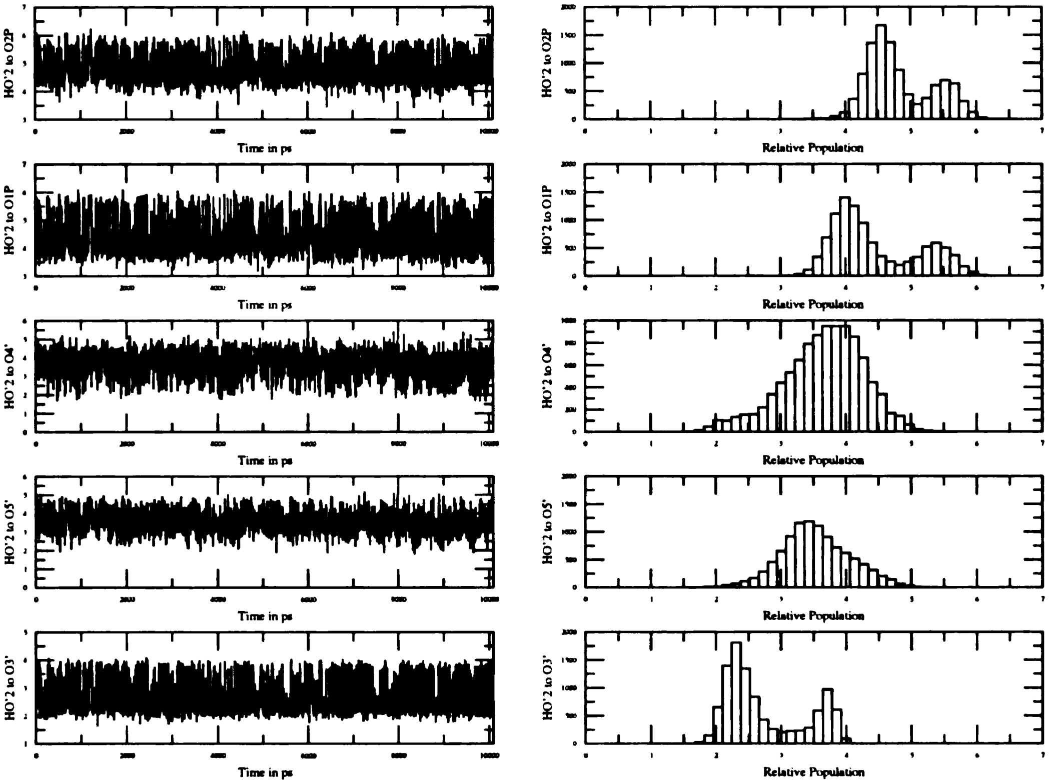


Figure 3.41: Time course of hydrogen bonds of Okazaki sequence RNA simulation starting in canonical A using parm96

HO'2 Distances for residue 5

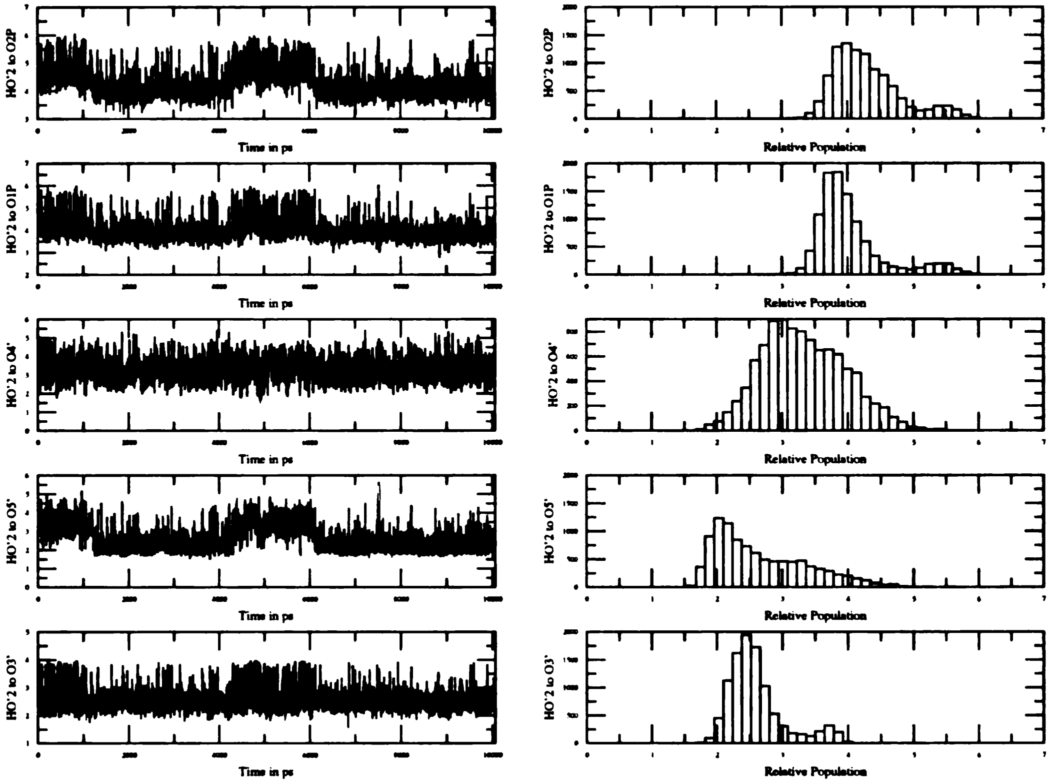


Figure 3.42: Time course of hydrogen bonds of Okazaki sequence RNA simulation starting in canonical A using parm99

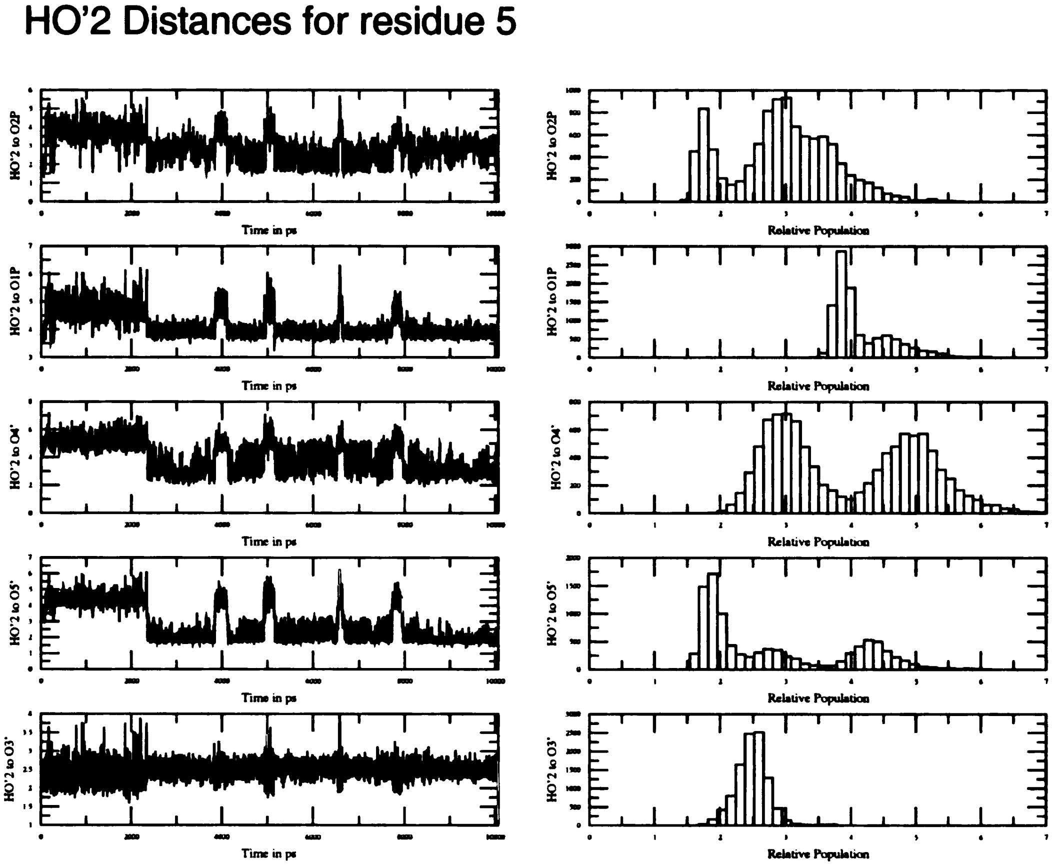


Figure 3.43: Time course of hydrogen bonds of Okazaki sequence RNA simulation starting in canonical B using parm96

HO'2 Distances for residue 5

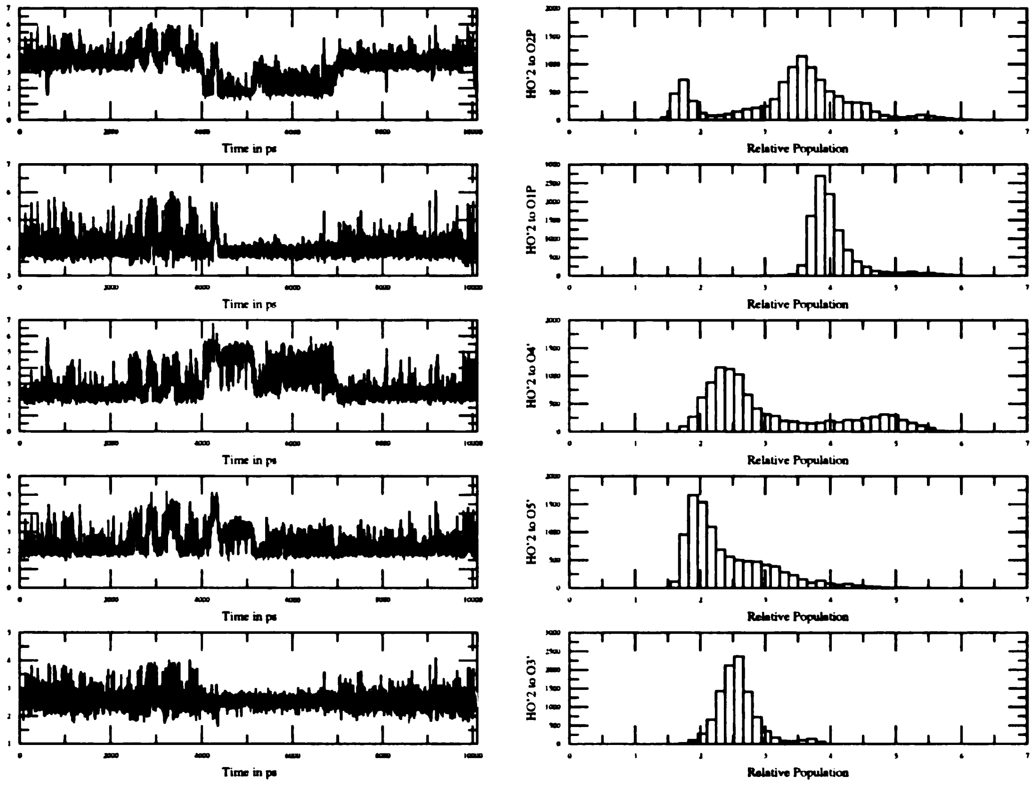


Figure 3.44: Time course of hydrogen bonds of Okazaki sequence RNA simulation starting in canonical B using parm99

Chapter 4

Restrained molecular dynamics of solvated duplex DNA using the particle mesh Ewald method

The work presented in this chapter is a collaboration between myself, Thomas E. Cheatham III, Peter A. Kollman and Thomas L. James. Thomas E. Cheatham III performed some simulations, and and Thomas L. James oversaw the project.

Reprinted from the Journal of Biomolecular NMR with permission. Originally published as: David E. Konerding, Thomas E. Cheatham III, Peter A. Kollman and Thomas L. James. "Restrained molecular dynamics of solvated duplex DNA using the particle mesh Ewald method" Journal of Biomolecular NMR (1999) 13(2):119-31.

4.1 Authors

David E. Konerding[1], Thomas E. Cheatham III[2], Peter A. Kollman[3] and Thomas L. James[4]

1 Graduate Group in Biophysics, Box 0446, University of California, San Francisco, CA 94143, U.S.A.

2 Laboratory of Biophysical Chemistry 12A-2041, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, MD 20892-5626, U.S.A.

3 Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143, U.S.A.

4.2 Abstract

Restrained and unrestrained aqueous solution molecular dynamics simulations applying the particle mesh Ewald (PME) method to DNA duplex structures previously determined via in vacuo restrained molecular dynamics with NMR-derived restraints are reported. Without experimental restraints, the DNA decamer, dCATTTGCATC · dGATGCAAATG and trisdecamer, dAGCTTGCCTTGAG · dCTCAAGGCAAGCT, structures are stable on the nanosecond time scale and adopt conformations in the B-DNA family. These free DNA simulations exhibit behavior characteristic of PME simulations previously performed on DNA sequences, including a low helical twist, frequent sugar pucker transitions, B_I-B_{II} ($\epsilon - \zeta$) transitions and coupled crankshaft ($\alpha - \gamma$) motion. Refinement protocols similar to the original in vacuo restrained molecular dynamics (RMD) refinements but in aqueous solution using the Cornell et al. force field [20] and a particle mesh Ewald treatment produce structures which fit the restraints very well and are very similar to the original in vacuo NMR structure, except for a

significant difference in the average helical twist. Figures of merit for the average structure found in the RMD PME decamer simulations in solution are equivalent to the original in vacuo NMR structure while the figures of merit for the free MD simulations are significantly worse. The free MD simulations with the PME method, however, lead to some sequence-dependent structural features in common with the NMR structures, unlike free MD calculations with earlier force fields and protocols. There is some suggestion that the improved handling of electrostatics by PME improves long-range structural aspects which are not well defined by the short-range nature of NMR restraints.

Abbreviations: NOE, nuclear Overhauser effect; MD, molecular dynamics; RMD, restrained MD; PME, Particle Mesh Ewald; rmsd, root-mean-square deviation.

4.3 Introduction

Improving the resolution of the structure of DNA in solution is a major challenge which requires both experimental data and theoretical modeling. It has long been recognized that solvent conditions profoundly influence the structure of DNA [28] and that the specific sequence can also play a role in the structure and deformability of nucleic acids in solution [108]. Static structure and dynamic deformability properties of DNA have important implications for recognition by proteins in biological processes such as transcription, as well as packaging and damage repair and may have applications in drug design and gene therapy. Therefore, correctly representing the structure of DNA in solution is paramount if understanding of these processes at a molecular level is to be achieved.

Defining the atomic level structure of DNA in solution to adequate resolution has been difficult. NMR is the principal method for determining high resolution solution structures of DNA. However, NOE intensities can only be observed for interproton pair distances up to 5-6Å, limiting our ability to define global DNA structure accurately beyond the base pair or base step level. Additionally, overlap of protons in the NOESY spectrum and poor signal-to-noise ratio can lead to the absence of data. Of the approximately 2000 interproton pairs with distances less than 6Å in a canonical B-DNA decamer, we can expect to observe only about 400-500, with a precision of about 0.25-1Å. In addition, due to molecular motion in solvated DNA, NOESY intensities are subject to averaging which complicates the process of determining accurate distances [106]. For these reasons, structures based on NMR data must include explicit a priori knowledge of chemical structure. Normally, this information is present in the form of a force field, including explicit parameters for the bond lengths, bond angles and dihedral angles, as well as atomic charges and van der Waals parameters for the system being studied. Typically, a published DNA structure results from refinement of a starting model by in vacuo restrained molecular dynamics (RMD) using a particular force field. In this case, the chemical structure is maintained by the force field, while the specific tertiary conformation of the molecule is achieved by means of interproton distance and torsion angle restraints (derived from NOE intensities and COSY coupling constants, respectively) added to the force field [95]. Because DNA is a highly charged biomolecule, accurate unrestrained molecular simulations of duplexes have been very hard to generate. Consistent improvement of methodology [5] has improved the quality of simulations, but obtaining stable trajectories for long time periods (> 1 ns) has been very difficult. Recently [22, 113], a promising technique, particle mesh Ewald, was developed. The PME method computes a full representation of the electrostatic interactions for a periodic lattice using screened real space sums and Fourier transforms to evaluate the reciprocal space interactions, providing an alternate method for computing long-range electrostatics in nucleic acid simulations. Its application has led to stable nucleic acid dynamics trajectories of 1 ns and longer [16, 65] without need for artificial restraints. Use of PME with the Cornell et al. force field has been validated by impressive results reproducing many of the known conformational features of solvated

DNA. Simulations using unrestrained PME starting from canonical B and canonical A form DNA converge to structures very similar to experimental crystal structures, reproducing sequence-specific properties such as roll and tilt [14]. To determine whether these advances in force field and simulation methodologies could increase the quality of DNA structures determined by NMR, we examined the effect of PME molecular dynamics simulations on a DNA decamer, dCATTTGCATC · dGATGCAAATG, and a trisdecamer, dACGTTGCCTTGAG · dCTCAAGGCAACGT, both free and restrained by NMR data, and compared the results with the in vacuo RMD simulations used in the originally refined structures [71, 111, 112]. These sequences were chosen since they represent some of the highest resolution NMR structures refined in our laboratory. Unrestrained and restrained simulations have been carried out, and the resulting ensembles obtained from both types of simulations have been analyzed to determine whether more accurate structures are found using the improved methodology.

4.4 Methods

4.4.1 Decamer setup and equilibration

A total of five different MD simulations using the Cornell et al. force field with PME have been performed on the DNA decamer using the AMBER 4.1 and AMBER 5.0 suite of programs. One long unrestrained MD run was carried out for 2 ns. Following this, restrained PME simulations starting from four different initial structures were carried out for 100 ps each, using the distance and torsion angle restraints of the original in vacuo NMR refinements.

4.4.2 DNA decamer solvation and equilibration

A rectangular periodic box containing TIP3P water and 18 sodium counterions was constructed using the AMBER EDIT module to solvate and neutralize the originally refined NMR DNA structure. The water box extended approximately 10Å away from any solute atom, yielding approximately 3000 water molecules with a box size of 52Å by 44Å by 45Å, giving an approximate concentration of 16 mM DNA and 270 mM sodium. PME simulations were run with SHAKE on hydrogens (tolerance = 0.0005Å), a 1 fs time step, a temperature of 300 K with Berendsen temperature coupling [3] with solvent and solute coupled to a common bath, a 9Å cutoff applied to Lennard-Jones interactions in PME, and constant pressure with pressure scaling (τ_T = 0.2 ps, τ_P = 0.2 ps). The nonbonded list was updated every 10 steps. The PME charge grid spacing was approximately 1.0Å, and the charge grid was interpolated with a cubic B-spline [25]. Water and solute equilibration was performed by minimizing the water and counterions for 2500 steps while holding the DNA fixed to its initial atomic coordinates (the original in vacuo NMR structure). Next, 25 ps of non-PME dynamics were run, raising the temperature of the system from 100 K to 300 K while holding the DNA fixed to its atomic coordinates. Then, 1000 steps of minimization were applied, allowing the water and counterions to move freely while the DNA was restrained to its atomic coordinates using a harmonic potential of 25 kcal/mol. Following this, 10 ps of non-PME dynamics were run, allowing the water and counterions to move freely while restraining the DNA with a 1000 kcal/mol harmonic potential. Next, five consecutive 2000-step minimizations were performed, decreasing the harmonic potential from 20 to 0 kcal/mol in 5 kcal/mol steps. As a final equilibration step, a 3 ps PME dynamics run with no restraints on DNA, counterions or water, warmed the system from 100 K to 300 K. At this point the system was considered to be equilibrated, and production runs at 300K were initiated. As a control for the PME force field, a long (2 ns) free PME simulation was run. The free PME simulation, referred to as fPME in this paper, remained stable and exhibited structural behavior similar to previously published DNA free PME simulations [14, 15, 114]. The rmsd (all atom, mass-weighted) of the average structure to the initial in vacuo refined NMR structure is 3.23Å (Table 4.1), the average twist of the molecule is 29°, roll is positive, inclination is slightly negative, and has an average pseudorotation value near 120° (Table 4.2).

4.4.3 Original DNA decamer refinement in vacuo

The original in vacuo NMR refinement is described in detail in [112]. Following 20 ps of RMD refinement with a 1 fs timestep at 300 K, the restraint-minimized structure was submitted to 100 ps of in vacuo RMD, generating the trajectory referred to as ivRMD in this paper.

4.4.4 DNA decamer refinement using PME

For the DNA decamer, four independent 100 ps restrained PME molecular dynamics runs were performed. The restraint force field utilized was k_{dist} = 20 kcal/(mol·Å^2) for distance restraints and k_{tors} = 60 kcal/(mol·rad^2) for torsion angle restraints. The restraint set used was identical to the original NMR restraint set: 100 torsion angle restraints, 398 distance restraints, and 48 additional Watson-Crick hydrogen bond restraints to maintain base pairing. The hydrogen bond flat angle restraints were 10 kcal/(mol·rad^2) and the distance restraints were 18 kcal/(mol·Å^2). These hydrogen bond restraints were maintained for consistency with the initial refinement. The four starting structures for the runs corresponded to the initial and final frames of the unrestrained PME simulation, canonical A-DNA and canonical B-DNA. These simulations are referred to as RMDI, RMDF, RMDA and RMDB. The starting structures from the free PME simulation were already equilibrated in solvent, while the canonical A and B forms were solvated and equilibrated using the method described above for equilibrating the original NMR structure. Velocities were assigned from a Maxwellian distribution to give a temperature of 300 K. The restraints were identical to those used in the original in vacuo NMR refinement. The restraint protocol is as follows: after an initial unrestrained PME dynamics equilibration period of 10 ps, the distance restraint forces were ramped from k_{dist} = 1 kcal/(mol·Å^2) to 20 kcal/(mol·Å^2) and torsion angle restraint forces were ramped from k_{tors} = 3 kcal/(mol·rad^2) to 60 kcal/(mol·rad^2). The target temperature remained constant at 300 K throughout the restrained portion of the simulation. This protocol is similar to the original NMR refinements but had a duration of 100 ps instead of 20 ps.

4.4.5 Analysis of decamer simulation trajectories

For each simulation, a representative ensemble comprising the final 50 ps at 1 ps intervals of the simulation, and the corresponding average structure over this range, were chosen to represent the structure and allow direct comparison between the simulations. These subsets were used to eliminate bias from the early sections of the trajectories, which were close to the starting structure, and to select the same number of samples from each simulation.

To describe the conformational space sampled in the simulations, the backbone torsion angles and helical parameters of the DNA structures were calculated using the Dials and Windows [82] interface to Curves [60]. The parameters were computed for each structure in an ensemble, then arithmetically averaged. The results of the average of the parameters is more informative than parameters calculated from the average structure, because coordinate averaging is subject to motion artifacts and, with the ensemble, a statistical distribution of values is obtained rather than a single value. The quality of fit of structures compared to the experimental data was calculated using a sixth-root weighted R-factor, Rx [51]. The calculation of the Rx factor used the experimental NOESY intensities with a mixing time of 140 ms and a correlation time of 2 ns. The experimental NOESY intensities were available for the decamer but not the trisdecamer, so Rx has not been calculated for the trisdecamer simulations. For the calculation of structural energy and figures of merit, the sampling ensembles were used. Energies of structures from the simulations were computed using AMBER 5.0 with the

Cornell et al. force field applied to water- and counterion-stripped sample frames. No electrostatic cutoff was applied, and a distance-dependent dielectric with a dielectric constant of 4 was used to represent crudely the dielectric screening by bulk solvent. Energies of the individual members of the sampling ensembles were computed, then arithmetically averaged. The coordinate-average structures were computed by averaging the trajectory subsets described above. Following this averaging, the rmsd of all atoms with mass weighting was computed for pairs of structures. The CARNAL module of AMBER 4.1 was used for coordinate averaging and computing rmsd.

4.4.6 Trisdecamer setup and equilibration

A series of different MD simulations using the Cornell et al. force field and PME [16] have been performed on the DNA trisdecamer using the AMBER 4.1 and AMBER 5.0 suite of programs. One long unrestrained MD simulation using the PME force field was carried out for 1 ns, and will be referred to as fPME. As with the decamer fPME simulation, the trisdecamer simulation exhibited the same structural properties as previously published free PME simulations of duplex DNA. rmsd (all-atom, mass-weighted) to the initial in vacuo refined NMR structure is 3.17Å (Table 4.3), the average twist of the molecule is 30°, roll is positive, the molecule has a slightly negative inclination and an average pseudorotation value near 120° (Table 4.4). Following this, restrained PME simulations starting from four different starting structures (two canonical A-DNA and two canonical B-DNA form DNAs), each with a different random number seed, were carried out for 250 ps each using the distance and torsion angle restraints of the original in vacuo NMR refinements. These simulations will be referred to as RMDA1, RMDA2, RMDB1 and RMDB2. The PME equilibration was essentially the same as the decamer simulation, except in the case of the restrained simulations, the first 25 ps involved non-PME dynamics and were followed by 25 ps of PME dynamics with the DNA held fixed in both cases before PME production dynamics. In the production dynamics, the restraint force constants were k_{dist} = 20.0 kcal/(mol·Å^2) and k_{tors} = 60.0 kcal/(mol·rad^2). The restraints were ramped up smoothly to 1/4 strength during the first 2-10 ps, followed by ramping to full strength over the next 10 ps. The time step was 2 fs. Similarly, in the absence of artificial hydrogen bond restraints, it is necessary to ramp up the restraint force constants slowly since otherwise the structure may rapidly distort (e.g., by breaking base pairs) in order to instantaneously satisfy the restraints.

Analysis of trisdecamer simulation trajectories In each case, a representative ensemble comprising the final 50 ps at 1 ps intervals of the simulation and the corresponding average structure over this range was chosen to represent the structure and allow direct comparison between the simulations. The conformational properties of the trisdecamer ensembles were calculated using Dials and Windows, and the energetics and quality of fit factors were computed using AMBER 5.0, in the same manner as for the decamer ensembles.

4.5 Simulation details

The decamer and trisdecamer simulations were run using the Sander module of AMBER 4.1 and 5.0 on the Cray T3D and T3E at the San Diego Supercomputer Center, Pittsburgh Supercomputer Center and local computers.

4.6 Results

Unrestrained and restrained PME simulations were run for both the DNA decamer, dCATTTGCATC · dGATGCAAATG, and trisdecamer, dAGCTTGCCTTGAG · dCTCAAGGCAAGCT. These sequences have been well defined by NMR data, with a high number of restraints per residue. The decamer has nearly 20 distance restraints per residue, which are derived from the cross-relaxation matrix MARDIGRAS analysis of the NOE intensities. The MARDIGRAS bounds are extremely tight; the average flatwell width over all distance restraints was 0.25Å. The tightness of the bounds is due to the MARDIGRAS technique. A newer method for determining distance bounds from NOE intensities, RANDMARDI [63], would produce wider bounds which better represented the inherent imprecision of the intensities. However, the original distance bounds were used to maintain consistency with the original refinement. The trisdecamer has approximately 10 distance restraints per residue. The results of both the decamer and trisdecamer simulations demonstrate similar general trends between the free and restrained simulations; for brevity, we will focus first on the results of the decamer simulations in depth, then compare these with the trisdecamer simulations.

4.6.1 Decamer simulations: Comparison of energies and goodness of fit between unrestrained PME, restrained PME, and restrained in vacuo structures

As described in the Methods, the structural energy and goodness of fit were computed for sampling periods from the MD simulations. Instead of computing the structural energy and fit of the average structures, we have computed the average structural energy and fit from the sample structures making up the ensemble. There is a distinct advantage in computing the average energy of these samples from an ensemble in that the conformational energy is not artificially high due to the coordinate averaging process. Dynamic processes in the simulation, such as backbone torsion angle fluctuations, helix axis flexibility, and methyl rotation lead to anomalously high bond and angle energies upon straight coordinate averaging. Minimization is necessary to eliminate these artifacts but will only move the structure to a nearby local energy minimum. The local energy minimum may not correspond to the global minimum sought by RMD simulation. Therefore it is not very informative to compare the structural energies of minimized average structures generated from different ensembles. Even in the best case, where the dynamics represent fluctuations around a single mean rather than transitions between conformationally accessible substates, these energies are not particularly informative. Average energies computed from the simulations are illuminating. The free PME simulation has a slightly lower conformational energy (439 kcal/mol vs. 540-580 kcal/mol, where conformation energy is the sum of bond length, angle, dihedral and nonbonded terms) (Table 4.5) than the restrained simulations, which makes sense because the restraints tend to move the structure simulations away from the idealized, lower conformational energy structure preferred by the force field in an effort to minimize the artificial restraint energy. It is quite reasonable that the constraint energy of the free PME structure is significantly higher than the constraint energy of the restrained runs (2396 kcal/mol vs. about 280-290 kcal/mol), and the Rx, 0.13, is much higher than for the original in vacuo RMD structure (0.06). An Rx of 0.13 is approximately the same as canonical A- (0.16) or B-DNA (approximately 0.11), essentially a non-fit to the data. Comparing the in vacuo RMD energies with the RMD PME energies, we see that virtually the same Eamber and Econst values, as well as AVDB and Rx, are found for both types of restrained simulations. No matter which type of fit we use to compare the free MD with the RMD runs, a clear trend exists: the free MD structures do not fit the data nearly as well as the RMD runs. What is surprising is that all the RMD runs, in vacuo or solvent/PME, have the same quality of fit, independent of the force field and whether explicit solvent is included. Given the dissimilarity of the magnitude of twist (vide infra) between the in vacuo run and the restrained PME runs, this demonstrates that the quality of fit is not adequate to distinguish the absolute value of twist between the two structures. However, the restraints are still sufficient to determine the relative value of twists between base steps. This is rationalized by the fact that NOEs are short range, never giving direct information beyond the base pair or base step. Nevertheless, if short-range distances are determined with sufficient precision and combined with a sophisticated force field, which accurately represents the long-range features of the molecule, the global structure of the DNA should be accurately defined.

4.6.2 Decamer simulations: Restrained PME molecular dynamics compared to restrained in vacuo molecular dynamics: Comparison of time-averaged helical and pseudorotation parameters

The restrained PME simulations agree well with the restrained in vacuo simulation when helical parameters are compared (Figure 4.1). This demonstrates that the restraints act independently of the force field to determine sequence-specific variations in helical parameters. While all the different RMD PME runs converge to identical structures, ranging between 0.25 and 0.95Å (all atoms, mass-weighted pairwise rmsd between average structures from the sampling ensembles) as shown in Table 4.1, the RMD PME runs do not find precisely the same structure as the original in vacuo RMD simulation structure (rmsd ranging between 1.3 and 1.6Å).

Despite the agreement in helical parameters, we are not confident that 100 ps simulations using this protocol represent adequate sampling; for example, the unusual $\alpha - \gamma$ conformation at the T5pG6 step seen in the free PME simulation is reproduced only in the RMD simulations which started from the PME or in vacuo structures which had that conformation initially. The RMD simulations started from canonical A-DNA and B-DNA, which do not have that $\alpha - \gamma$ conformation, do not converge to the unusual $\alpha - \gamma$ conformation in the 100 ps refinement. Moreover, careful inspection of backbone angle and helical parameters shows RMD simulations starting from A and B tend to cluster together and simulations starting from the in vacuo or free PME cluster together. The lack of complete agreement suggests incomplete sampling. This clustering effect is most noticeable in the angle at the G6 base and may represent a correlation between the $\alpha - \gamma$ conformation and the angle at that step. Admittedly, the backbone of DNA is really well defined by NMR restraints, as few torsions are measured by COSY. Moreover, the barriers to rotation around these dihedral angles are likely large enough to preclude observing transitions during short simulations, so this probably represents a sampling problem with MD which is exacerbated by restraints which further inhibit sampling and by the presence of explicit water which slows the dynamics. All of these observations suggest the need for longer restrained simulations or a methodology which increases conformational sampling by reducing energy barriers. In spite of the lack of complete convergence, magnitude and sequence-specific variations of roll and tilt, as well as pseudorotation, are nearly identical between the PME and in vacuo RMD simulations. X-displacement, inclination and other helical parameters are in excellent agreement in both magnitude and sequence-specific variation as well (Figure 4.1), although some solvated simulations do not converge as well as others. This lack of convergence also suggests that 100 ps RMD simulations are not long enough to adequately allow the transition to the final RMD structure. In free PME, the A to B DNA transition takes approximately 250 to 500 ps; whereas in the restrained simulations there is no repuckering and limited sampling, yet an A to B transition is enforced in 50 ps. Even with restraints, the energy barrier between the starting structure and the correct solution structure may be high enough that longer restrained simulations will be necessary to find the global minimum energy. While sequence-specific variations in twist are maintained between PME and in vacuo restrained structures, the magnitude differs. It appears to be 'stuck' at an average of 33° (Table 4.2), which is between the value favored by the free PME structure (average = 29°) and the restrained in vacuo structure (average = 36°). This difference in magnitude of twist is the primary contributor to the surprisingly high atomic rmsd (1.3-1.6Å) between RMD PME and in vacuo RMD structures. Naturally the question is raised why tilt and roll converge to values so close to the original NMR structure while twist mirrors the sequence-specific variations but not the actual magnitude. If one considers the influences of NMR restraints and the force field on a DNA structure, the logical explanation for this derives from the combination of the lack of explicit twist-defining NMR data and the underestimated twist with the Cornell et al. force field. 'Local' (base pair step) twist is poorly defined by the NMR data, and the 'global' twist of the entire molecule is primarily determined by the force field. Roll and tilt components of base pair steps are very well defined by NMR data, primarily due to the spatial arrangement of NOESY distances between base pairs in a step [104]. This is an important aspect of DNA solution structure that has not been fully addressed in the literature to date. The equilibrium twist in the restrained PME run is balanced between the value predicted by the free PME simulation and the value predicted by the original restrained in vacuo simulation. Thus, the force field used does play a significant role in defining the global helical parameters, even when experimental distance restraints are added. We note that the distance restraints, which never extend beyond a single base pair or base step, are not necessarily violated when twist, tilt, and roll are modulated. There are two reasons why this is the case. First, NOE restraints are imprecise, i.e., upper and lower bounds define acceptable values, and so small changes in twist may not cause the NOE restraints to be violated. Second, compensations for large changes in twist by other parameters such as slide and displacement can avoid strong violations of the NOEs [104]. Thus, it is not unreasonable that we can expect the force field to exert a significant influence on the average twist magnitude. In spite of this, NMR data still have an important role in determining sequence-specific variations from the mean. It is hoped that with optimization of the Cornell et al. force field to mend the low-twist bias, a more reasonable representation of the twist in these structures can be determined. It is important to note that the RMD PME simulations do have an average twist which is close to the helical periodicity (10.6 bases per turn, twist = 34°) measured in solution using an independent enzymatic method [84]. Additionally, we note that re-inspection of the original decamer NMR spectra reveals the absence of a cross peak between methyl hydrogens of bases T5 and T13. In the original in vacuo RMD structure, the methyl hydrogens are close enough (<6Å) that a small peak should be observable. In the free PME and RMD PME structures, the methyl hydrogen distances are increased, due to the decreased bending at the

T5-A8 steps, in better agreement with the spectra. While absence of a peak does not prove there is no intermolecular contact, it suggests that the original model may not be as accurate as the PME RMD models. These two observations represent important validations of the Cornell et al. force field with the PME method for use in NMR structure refinement.

4.6.3 Trisdecamer results: Comparison of in vacuo RMD with free PME and RMD PME

In contrast to the straightforward RMD PME refinements of the decamer, restrained simulations of the trisdecamer from different starting structures did not converge as readily to a common structure (Figure 4.2, Tables 4.3 and 4.6). In part, this is due to the quality of the restraints: there are fewer restraints per residue for the trisdecamer, and bounds are not as precise. Additionally, we found it necessary to add Watson-Crick hydrogen bond restraints to maintain base pairing during the simulations starting from A-DNA, since without these, the base pairs broke to instantaneously satisfy the 'B-DNA' restraints while in an 'A-DNA' geometry. The simulations were run for a longer period (250 ps instead of 100 ps) and still had trouble converging: the A-DNA simulations still have high inclination (approximately 20°), although the x-displacement and pseudorotation angles are compatible with the B-DNA simulations. The average structures from the simulations starting from A-DNA are 2.76 to 3.3 Å from the original NMR structure (Table 4.3), and the simulations starting from B-DNA are 1.7 to 2.1Å from the original NMR structure, while the simulations starting from A-DNA are 2.4 to 3.2Å from the simulations starting from B-DNA. The difficulty of converging from A-DNA to B-DNA, despite the longer simulation, is likely due to difficulties in conformational sampling coupled with lower quality restraints. Adding explicit water slows conformational transitions, and spontaneous A-DNA to B-DNA transitions in free PME require ca. 500 ps [14]. This time scale, coupled with the inhibited sampling observed with the restraints, suggests that longer simulation times are necessary or alternative methods need to be applied to enhance sampling. Mirroring the trouble we had in obtaining converged structures (vide supra), the goodness of fit of the RMD PME structures starting from A-DNA does not fit the NMR data as well as the B-DNA start simulations or the originally determined in vacuo structure. The ensemble average Econst for the A-DNA starting structure simulations were above 700 kcal/mol, while the B-DNA starting structure simulations were around 230 kcal/mol and the original in vacuo simulation around 330 kcal/mol (Table 4.6).

4.7 Discussion and conclusions

We have shown that unrestrained MD simulations of DNA sequences, which have previously been characterized by NMR, are stable on the nanosecond time scale using the Cornell et al. force field with PME. However, the sampled structures are not fully consistent with the NMR data. In contrast, the restrained PME simulations converge to a common structure even when different starting structures are used, and the sequence-specific properties observed in the original in vacuo RMD simulations are reproduced fairly well, with the exception of helical twist. Unrestrained simulations using PME with the Cornell et al. force field have already been observed to show a lower twist than experimental data imply, so the lower twist in the RMD simulations is not surprising. Moreover, the PME RMD simulations, which converge to a structure with much lower average helical twist from the original in vacuo refinements, manifest nearly identical structural energies and figures of merit. This shows that NOE intensities, even in well-determined systems, cannot accurately define the magnitude of helical twist. Dependence on an accurate force field is therefore necessary in DNA structure refinement using NMR data. Commonly, NMR refinements will use restrained MD with several different starting structures. Convergence from several different starting structures to a single common structure (usually measured by atomic rmsd) is treated as a measure of precision, in that the NMR data is sufficient to guide an RMD run to a single structure which satisfies the restraints. In this work, we used canonical A- and B-DNA forms as starting structures and measured whether convergence was reached. Although excellent convergence was readily achieved using the decamer data set and A-DNA and B-DNA forms, the trisdecamer data set was not sufficient to drive the A-DNA and B-DNA forms to the same final structure under the simulation conditions employed. In particular, the trisdecamer RMD simulation that started from A-DNA still maintained significant A-DNA conformational features such as x-displacement and inclination. Additionally, the and torsion angles at step T5pG6 of the decamer do not converge from the simulations started from the original NMR structure conformation or the final PME conformation to the more typical values seen for the rest of the sequence and those seen in the canonical A-DNA and B-DNA simulations. Normally, RMD refinement utilizes a simulated annealing approach, where the temperature of the simulation is raised to high values along with the force constants of the experimental restraints, followed by a cooling period where temperature and restraints are dropped significantly. This should help guide initial structures over large energy barriers to a region near the correct solution structure. Simulated annealing is challenging to implement when full solvation and periodic boundary conditions are used, since the commonly used water models were not parameterized for use at high temperature, and high temperature leads to lower water density (which could disrupt the structure) or higher pressure (which may inhibit sampling further). However, judicious modifications to the simulation protocol, such as constant volume instead of constant pressure as well as shorter time steps, should allow higher temperatures during RMD runs. In the meantime, other methods of passing over the energy barrier of A-DNA to B-DNA interconversion need to be used. In the current work, it was necessary to use H-bond restraints and longer (250 ps vs. 100 ps) simulations for the trisdecamer in canonical A-DNA and B-DNA conformation to approach the ‘correct’ final structure without distortion of terminal base pairs. Since we know that free PME simulations readily interconvert from A-DNA to B-DNA on a 500 ps-1 ns timescale, this suggests that longer simulations, at least 500 ps-1 ns, may be necessary when using PME RMD with explicit water. A simple way to overcome some of these problems is to continue with the standard, rapid and efficient in vacuo refinement to generate a set of structures compatible with the data followed by submitting these structures to 50-500 ps of RMD in explicit solvent with PME and a reasonable nucleic acid force field. Alternatively, enhanced sampling methodologies, such as locally enhanced sampling, may be applied to effectively reduce barriers to conformational transition [86, 98]. It is believed that with better nucleic acid force fields, modern simulation techniques and inclusion of explicit solvent, more reliable refinement of NMR models can be performed to produce more realistic structures.

4.8 Acknowledgements

D.E.K. would like to acknowledge the Computer Graphics Lab at the University of California, San Francisco, for their generous donation of computer time on the Cray T3D and T3E at the San Diego Supercomputer Center. We would also like to thank the Pittsburgh Supercomputer Center for significant computational resources (MCA93S017P and CHE880090P). P.A.K. would like to acknowledge the National Institutes of Health for research support (CA-25644). T.L.J. would like to acknowledge the National Institutes of Health for research support (GM-39247).

4.9 Tables

	fPME[a]	RMDI	RMDF	RMDA	RMDB	ivRMD	Adna	Bdna
fPME		1.78	1.75	1.89	2.01	2.89	2.56	3.65
RMDI	2.04		0.24	0.61	0.78	1.23	2.75	2.79
RMDF	2.02	0.25		0.60	0.76	1.28	2.73	2.82
RMDA	2.19	0.62	0.60		0.86	1.41	2.73	2.86
RMDB	2.27	0.88	0.83	0.95		1.39	2.91	2.59
ivRMD	3.23	1.35	1.39	1.46	1.58		3.18	2.68
Adna	2.87	3.23	3.22	3.25	3.40	3.73		4.73
Bdna	4.23	3.14	3.16	3.24	2.96	2.96	5.59	

Table 4.1: RMS deviations between pairs of average structures from the free PME, restrained PME, and in vacuo decamer simulations. Rms deviations between pairs of coordinate averaged structures of the decamer. Lower left is all-atom mass-weighted rms, upper right is the inner octamer mass-weighted rms.

[a] Acronyms used in Tables 4.1, 4.2 and 4.5: fPME: 2 nsec free PME simulation of decamer DNA; RMDI: RMD PME starting from the initial free PME conformation; RMDF: RMD PME starting from the final free PME conformation; RMDA: RMD PME starting from canonical A form; RMDB: RMD PME starting from canonical B form; IvRMD: RMD in vacuo (original NMR simulation); Adna: A-DNA; Bdna: B-DNA.

	fPME[a]	RMDI	RMDF	RMDA	RMDB	ivRMD
Shear	0.0(0.2)	0.1(0.1)	0.0(0.1)	0.0(0.1)	0.0(0.1)	0.1(0.1)
Stretch	0.2(0.1)	-0.1(0.1)	-0.1(0.1)	-0.1(0.1)	-0.1(0.0)	-0.3(0.1)
Stagger	-0.2(0.2)	-0.1(0.1)	-0.1(0.1)	-0.1(0.2)	-0.1(0.1)	0.0(0.1)
Buckle	-0.8(4.5)	-2.1(3.3)	-2.2(2.4)	-1.7(2.9)	-2.1(2.7)	-5.0(3.0)
Propeller	-11.8(3.4)	-12.4(3.2)	-11.9(3.1)	-12.5(2.7)	-11.4(2.8)	-18.4(3.2)
Opening	3.8(2.3)	-0.5(1.7)	-0.7(1.5)	-0.2(1.6)	-2.1(1.5)	-5.9(1.7)
Shift	-0.0(0.1)	-0.1(0.1)	-0.1(0.1)	-0.1(0.1)	-0.1(0.1)	-0.1(0.0)
Slide	-0.2(0.1)	-0.2(0.1)	-0.2(0.1)	-0.2(0.1)	-0.3(0.0)	-0.2(0.0)
Rise	3.3(0.2)	3.2(0.1)	3.2(0.1)	3.1(0.1)	3.2(0.1)	3.1(0.0)
Tilt	0.5(1.9)	1.3(1.2)	0.8(1.2)	1.2(1.1)	0.8(1.0)	1.0(0.9)
Roll	7.4(2.7)	1.9(1.5)	2.0(1.6)	1.0(1.5)	3.8(1.3)	1.0(1.4)
Twist	29.0(1.2)	33.1(0.5)	32.8(0.6)	32.9(0.6)	33.0(0.5)	36.3(0.6)
X Disp.	-1.9(0.8)	-1.9(0.3)	-1.8(0.3)	-2.0(0.3)	-1.3(0.2)	-1.6(0.2)
Y Disp.	-0.0(0.5)	-0.4(0.2)	-0.4(0.3)	-0.4(0.2)	-0.5(0.2)	-0.4(0.2)
Inclination	-4.8(6.2)	4.1(2.6)	4.0(2.4)	5.7(2.6)	-0.0(2.2)	6.3(2.3)
Tip	-0.6(3.8)	1.4(2.1)	1.5(2.4)	1.5(2.0)	0.5(1.9)	1.2(1.7)
Axis X Disp.	-0.0(0.1)	-0.0(0.0)	-0.0(0.0)	-0.0(0.0)	0.0(0.0)	-0.0(0.0)
Axis Y Disp.	-0.1(0.0)	-0.1(0.0)	-0.1(0.0)	-0.1(0.0)	-0.1(0.0)	-0.1(0.0)
Axis inc.	0.5(1.3)	0.8(0.8)	0.7(0.9)	0.9(0.8)	-0.2(0.7)	0.4(0.7)
Axis tip	6.4(2.2)	1.9(1.0)	2.1(1.2)	1.5(1.1)	3.4(0.8)	1.5(1.0)
delta	110.7(3.3)	114.1(1.4)	113.7(1.5)	113.3(1.6)	113.2(1.9)	117.6(1.6)
epsilon	178.8(3.0)	172.3(1.6)	172.8(1.7)	172.0(1.5)	172.4(1.8)	171.3(1.5)
zeta	254.9(3.7)	256.1(1.7)	255.9(1.6)	256.4(1.6)	256.4(1.9)	255.4(1.9)
alpha	264.9(1.8)	267.0(2.5)	267.5(2.2)	277.9(1.7)	277.3(2.1)	262.8(2.9)
beta	162.5(2.7)	164.3(2.0)	164.0(1.8)	163.0(1.7)	163.2(1.5)	165.1(1.5)
gamma	61.2(2.1)	57.0(2.0)	56.8(1.7)	48.5(1.7)	48.4(1.5)	61.9(2.4)
chi	229.9(2.7)	236.8(1.8)	236.8(1.7)	237.3(2.0)	237.0(1.6)	237.3(1.6)
Pucker	114.8(9.5)	129.6(1.4)	129.3(1.4)	129.1(1.9)	128.8(1.7)	131.9(1.6)
Amplitude	40.6(1.4)	32.2(0.7)	32.2(0.9)	32.3(0.9)	32.4(0.6)	32.3(0.7)

Table 4.2: Average helical parameters and backbone angles of decamer simulation ensembles. Standard angle and helical values averaged over residues, base pairs, or base pair steps (where appropriate) for the decamer structures specified. Average values were calculated by arithmetically averaging the values calculated for the individual structures within each sampling ensemble. Standard deviations are parenthesized.

[a] Acronyms are explained in the footnote to Table 4.1.

	fPME[a]	ivRMD	RMDA1	RMDA2	RMDB1	RMDB2
fPME		2.83	3.28	3.03	2.10	1.78
ivRMD	3.17		2.90	2.42	1.48	1.85
RMDA1	3.43	3.30		0.79	2.79	3.15
RMDA2	3.15	2.76	0.92		2.37	2.79
RMDB1	2.21	1.70	2.94	2.43		0.68
RMDB2	1.89	2.10	3.26	2.84	0.72	

Table 4.3: Rms deviations between pairs of average structures from the free PME, restrained PME, and in vacuo trisdecamer simulations. Lower left is all-atom mass-weighted, upper right is the inner decamer.

[a] Acronyms used in Tables 4.3, 4.4 and 4.6: fPME: 1 nsec free PME simulation of trisde- camer DNA; ivRMD: Original NMR structure; RMDA1: RMD PME starting from A-DNA conformation; RMDA2: RMD PME starting from A-DNA conformation; RMDB1: RMD PME starting from B-DNA conformation; RMDB2: RMD PME starting from B-DNA conformation.

	fPME[a]	ivRMD	RMDA1	RMDA2	RMDB1	RMDB2
Shear	-0.0(0.1)	-0.0(0.0)	-0.0(0.1)	-0.1(0.1)	-0.1(0.1)	-0.2(0.1)
Stretch	0.1(0.1)	-0.2(0.0)	-0.0(0.2)	0.1(0.1)	0.0(0.1)	0.0(0.1)
Stagger	-0.2(0.1)	0.1(0.0)	-0.1(0.1)	-0.1(0.2)	0.0(0.1)	-0.1(0.2)
Buckle	-0.1(3.8)	0.6(0.0)	-9.4(2.4)	-9.4(3.3)	-2.4(3.1)	-0.3(2.8)
Propeller	-5.8(3.8)	-18.3(0.0)	-3.2(3.8)	-5.8(2.7)	-10.6(3.2)	-10.7(3.5)
Opening	2.1(1.9)	-4.4(0.0)	-1.6(2.3)	-0.8(1.8)	-2.7(1.7)	-2.2(1.6)
Shift	-0.1(0.1)	-0.0(0.0)	0.2(0.1)	0.1(0.1)	0.0(0.1)	0.0(0.1)
Slide	-0.2(0.1)	0.0(0.0)	-0.1(0.1)	-0.1(0.1)	-0.1(0.1)	-0.1(0.1)
Rise	3.6(0.1)	3.1(0.0)	3.1(0.1)	3.1(0.1)	3.3(0.1)	3.3(0.1)
Tilt	0.6(1.8)	1.4(0.0)	-1.5(1.6)	0.2(1.2)	0.4(1.3)	0.4(1.3)
Roll	8.5(1.7)	3.0(0.0)	4.0(2.3)	2.7(2.0)	4.2(1.4)	4.4(1.4)
Twist	30.2(0.6)	34.3(0.0)	32.7(0.9)	33.6(0.7)	33.4(0.6)	32.8(0.6)
X Disp.	-1.2(0.4)	-2.4(0.0)	-3.1(0.4)	-3.2(0.4)	-1.9(0.3)	-1.8(0.3)
Y Disp.	-0.1(0.4)	-0.4(0.0)	0.5(0.5)	0.1(0.3)	-0.2(0.3)	-0.2(0.3)
Inclination	-9.3(3.1)	6.3(0.0)	20.1(4.0)	21.0(3.6)	3.9(2.7)	1.0(2.9)
Tip	1.8(2.7)	0.6(0.0)	-6.9(3.6)	-3.7(3.0)	-2.2(2.4)	-2.2(2.0)
Axis X Disp.	-0.0(0.0)	0.0(0.0)	0.0(0.0)	0.0(0.0)	0.0(0.0)	-0.0(0.0)
Axis Y Disp.	-0.1(0.0)	-0.0(0.0)	-0.1(0.0)	-0.1(0.0)	-0.1(0.0)	-0.1(0.0)
Axis inc.	-0.6(1.1)	0.3(0.0)	-1.8(1.0)	-0.5(1.0)	-0.3(1.0)	-0.2(1.0)
Axis tip	7.2(1.3)	2.2(0.0)	5.2(1.9)	3.5(1.6)	3.9(1.0)	3.8(1.1)
delta	114.2(2.7)	117.4(0.0)	118.7(1.3)	119.2(1.5)	119.0(1.3)	118.6(1.7)
epsilon	198.4(2.5)	178.5(0.0)	198.5(1.9)	199.2(2.0)	186.5(1.6)	185.7(1.7)
zeta	259.1(3.4)	272.1(0.0)	254.9(2.1)	256.3(1.8)	267.1(1.7)	267.3(1.9)
alpha	280.8(1.7)	291.7(0.0)	286.7(7.7)	291.2(1.9)	292.0(1.8)	294.0(2.3)
beta	171.2(2.1)	180.8(0.0)	170.3(3.3)	172.0(1.6)	176.4(1.5)	177.3(1.6)
gamma	62.3(1.7)	54.0(0.0)	43.4(2.9)	44.0(2.8)	50.7(1.6)	49.2(2.0)
chi	231.9(2.9)	244.7(0.0)	257.2(1.9)	255.3(1.8)	245.1(2.0)	244.2(2.3)
Pucker	117.8(5.8)	138.5(0.0)	141.9(1.8)	142.4(2.0)	141.7(1.5)	141.4(1.5)
Amplitude	40.4(1.3)	29.1(0.0)	28.6(0.8)	28.5(0.9)	28.8(0.8)	28.9(0.8)

Table 4.4: Average helical parameters and backbone angles of trisdecamer simulation ensembles. Standard angle and helical values averaged over residues, base pairs, or base pair steps (where appropriate) for the trisdecamer structures specified. Average values were calculated by arithmetically averaging the values calculated for the individual structures within each sampling ensemble. Standard deviations are parenthesized. The ivRMD standard deviations are zero because it is only a single structure.

[a] Acronyms are explained in the footnote to Table 4.3.

	ivRMD[a]	RMDI	RMDF	RMDA	RMDB	free PME
Eamber	582.90	543.79	554.59	551.68	544.93	439.76
Econst	284.87	283.03	299.69	284.66	283.87	2396.33
AVDB	0.15	0.14	0.15	0.14	0.14	0.39
Rx	0.06	0.06	0.06	0.06	0.06	0.13

Table 4.5: Energies and statistics of fit for decamer simulation structures.
[a] Acronyms are explained in the footnote to Table 4.1.

	ivRMD[a]	A1	A2	B1	B2	fPME
Eamber	169.80	914.06	895.66	669.73	655.69	565.77
Econst	337.78	820.28	735.69	230.12	232.11	3553.34
AVDB	0.16	0.28	0.26	0.14	0.14	0.57

Table 4.6: Energies and statistics of fit for trisdecamer simulation structures.
[a] Acronyms are explained in the footnote to Table 4.3.

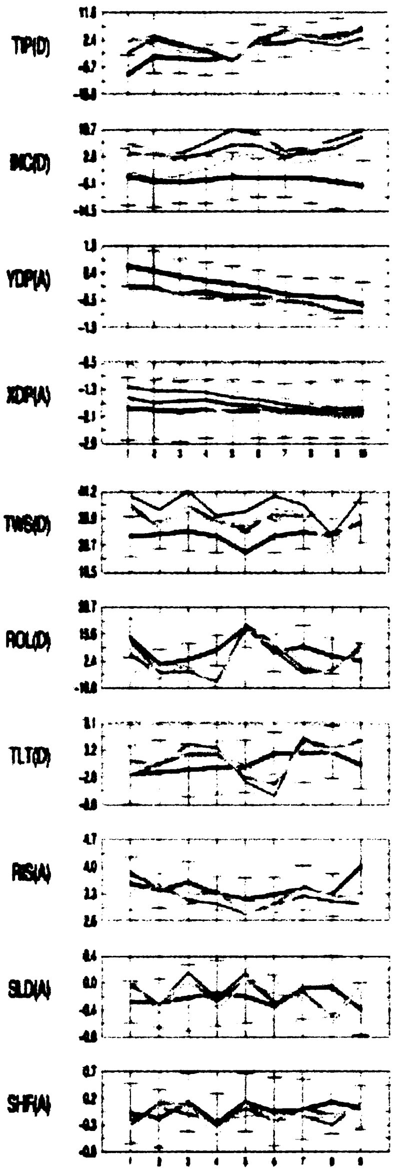


Figure 4.1: Average of helical parameters and backbone angles over the ensemble structures from the original in vacuo RMD, free PME and PME RMD simulations calculated with the Dials and Windows interface to Curves. Parameters were calculated for individual structures taken from the sampling ensembles of the trajectories, and then arithmetically averaged. The x-axis represents the base position in the sequence and the y-axis is the parameter value. Parameters in Åare marked (A) and parameters in degrees are marked (D). The lines are colored as follows: Free PME (black), ivRMD (red), RMDI (green), RMDF (blue), RMDA (yellow), RMDB (brown). Vertical bars represent the standard deviation for the free PME simulation.

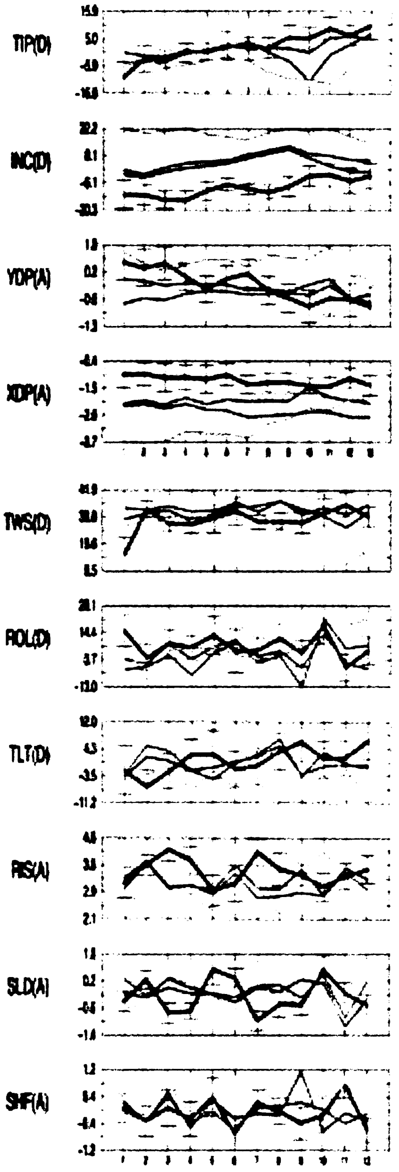


Figure 4.2: Average of helical parameters and backbone angles over the ensemble structures from the original in vacuo RMD structure, free PME simulation and PME RMD simulations calculated with the Dials and Windows interface to Curves. Parameters were calculated for individual structures taken from the sampling ensembles of the trajectories, and then arithmetically averaged. The x-axis represents the base position in the sequence and the y-axis is the parameter value. Parameters in Å are marked (A) and parameters in degrees are marked (D). The lines are colored as follows: Free PME (black), RMDB4 (red), RMDA4 (green), ivRMD (blue). Vertical bars represent the standard deviation for the free PME simulation.

Chapter 5

The Ensemble/Legacy Chimera Extension: Standardized User and Programmer Interface to Molecular Ensemble Data and Legacy Modeling Programs

Reprinted from the Proceedings of the Pacific Symposium in Biocomputing with permission. Originally published as: David E. Konerding. "The Ensemble/Legacy Chimera Extension: Standardized User and Programmer Interface to Molecular Ensemble Data and Legacy Modeling Programs". Proceedings of the Pacific Symposium in Biocomputing (1999).

5.1 Author

David E. Konerding,

Graduate Group in Biophysics, Box 0446 UCSF,

San Francisco, CA 94143-0446

5.2 Abstract

Ensemble/Legacy is a toolkit extension of the Object Technology Framework (OTF)[46] that exposes an object oriented interface for accessing and manipulating ensembles (collections of molecular conformations that share a common chemical topology) and driving Legacy programs (such as MSMS[93], AMBER[74], X-PLOR[8], CORMA[64], MARDIGRAS[64], Dials and Windows[80], and CURVES[61]). Ensemble/Legacy provides a natural programming interface for running Legacy programs on ensembles of molecules and accessing the resulting data. Using the OTF reduces the time cost of developing a new library to store and manipulate molecular data and also allows Ensemble/Legacy to integrate into the Chimera[45] visualization program. The extension to Chimera exposes the Legacy functionality using a graphical user interface that greatly simplifies the process of modeling and analyzing conformational ensembles. Furthermore, all the C++ functionality of the Ensemble/Legacy toolkit is "wrapped" for use in the Python[88] programming language.

5.3 Introduction

5.3.1 Flexibility and Dynamics of Biomolecules

Biophysical analysis of molecules in solution reveals that proteins and nucleic acids are flexible and dynamic. This flexibility plays a critical role in many biological situations, such as cellular regulation, genomic replication, and environmental interaction.

In some cases, molecules are rigid and the only flexibility is small positional fluctuation around average atomic positions, while in other cases there may be large conformational rearrangements from one conformational subfamily to another. Being able to extract and visualize relevant dynamic information from experimental data and theoretical predictions is a major challenge. Management and reduction of data to produce useful information is often hampered by the sheer volume of the data to be analyzed, and by the difficulty of converting the information into a form required by analysis programs.

5.3.2 The Ensemble/Legacy Toolkit

To address the issue of modeling and visualizing molecular ensemble data we have developed the Ensemble/Legacy toolkit. Ensemble/Legacy provides the biomolecular scientist with both a GUI (graphical user interface) and a programming library to analyze and visualize ensembles of structures. The Ensemble/Legacy library is built on top of Chimera and the OTF (see figure 5.1), eliminating the need to develop a new molecular visualization tool designed specifically for ensembles. This design decision allows programmers using the Ensemble/Legacy extension to focus on implementing code that performs the required functionality (such as an interface to a pre-existing Legacy program) and allows users to learn only one molecular visualization interface.

5.3.3 OTF: a Molecular Applications Framework

Because the Ensemble/Legacy Toolkit depends heavily on the OTF, we will describe the design philosophy and interface structure of the OTF first. The Object Technology Framework is freely available software (http://www.cgl.ucsf.edu/otf) developed by the Computer Graphics Lab at UCSF to automate the process of generating C++ classes to facilitate rapid biomolecular application development. The OTF stores molecular data using C++ classes known as otf::Molecule, otf::Residue, and otf::Atom. The otf::Molecule, otf::Residue and otf::Atom classes represent a molecule by storing its atom type information (carbon, oxygen, nitrogen, etc), topological structure (bonds), and atomic coordinates. Multiple structures are accommodated by storing multiple coordinate sets for a given molecule. Logically, an instance of otf::Molecule contains otf::Residue instances (one for each residue in the molecule), each of which contains otf::Atom instances (one for each atom in the residue). This demonstrates the straightfoward link between reality, where "molecules are composed of residues which are composed of atoms", and the class structure of the OTF.

5.4 Design Structure of the Ensemble/Legacy Toolkit

5.4.1 Ensemble::Molecule class

The Ensemble::Molecule C++ class is the primary data object used by Ensemble/Legacy to store molecular ensembles. This class inherits from the otf::Molecule class, and mostly duplicates its behavior. The primary difference in the Ensemble::Molecule class is that it overrides the normal otf::Molecule behavior when coordinate data is read. Otf::Molecule reads all the coordinate sets from a molecular structure at once; however, Ensemble::Molecule reads one frame of coordinates in at a time (on demand) and caches the coordinate sets for later use. This design decision was made to reduce the start-up time cost and memory footprint when reading large ensembles of data but still allow high performance for both random and sequential access of ensemble structures. Typical users of the otf::Molecule class will be loading a structure with a single coordinate set; consuming a modest amount of memory (10000 atoms * 3 coordinates/atom * 8 bytes/float <= 1MB) while users of the Ensemble::Molecule class may be reading up to 1000 or more coordinates sets (230MB). Since a typical workstation may well have less physical memory than this, it is clear that there is a significant advantage to loading only frames that are going to be used.

5.4.2 Ensemble::Legacy class

Ensemble/Legacy also has an Ensemble::Legacy class that forms an abstract representation of running a legacy program as a subprocess. Legacy programs are well-established as useful tools, but are not easily automated or have other deficiencies which limit their usability in an automated modelling process. Legacy programs often require complicated input files which are challenging to build correctly by hand, and they frequently require customized versions of PDB[4] files (or other molecular data formats) which do not conform to the file format standard. The Ensemble/Legacy toolkit knows what each different legacy program expects in terms of input files, and automatically generates the appropriately formatted input files. Support for individual Legacy programs such as MSMS (a program to calculate molecular surfaces), AMBER (a suite of molecular modelling tools), X- PLOR (a package for crystallographic and NMR structure refinement), CORMA (a program to compute the R factor of an NMR structure), Dials and Windows and CURVES (programs which determine the helical structure of a DNA or RNA duplex) exist in subclasses of the Ensemble::Legacy class which provide customized code specific to the Legacy application.

5.4.3 The Chimera Movie GUI

The Movie GUI is a plug-in tool for the Chimera visualization program which allows the user to visualize an ensemble as a movie. Each frame of the movie maps one-to-one with a coordinate set in the ensemble. Frames can be single-stepped in forward or reverse direction or run as a movie where each frame is displayed rapidly in sequence. The movie feature is useful for qualitatively assessing a molecular dynamics or Monte Carlo simulation, while the single-step feature allows detailed scrutiny of individual structures in an ensemble. Furthermore, the Movie GUI 'supports all of the legacy interfaces, exposing their features through an intuitive graphical interface. At every step of the movie, the user can interactively run a legacy program on the current frame and immediately see the results.

The Movie GUI has evolved through several incarnations. The original Movie GUI was written as a delegate for the Midas display system[26]. The delegate mechanism allowed Midas to be extended with new features not anticipated by the original authors without the need for modification of the Midas source code. However, the delegate system was awkward, and allowed for only limited extension of Midas. The second Movie GUI was written as a standalone program with its own molecular visualization program. While this allowed for detailed control by the Movie GUI, developing a molecular visualization program is a major undertaking in itself; this version of Movie was primarily used as a feature testbed. For the third incarnation of Movie, we have chosen to use Chimera. Chimera is a fully extensible molecular visualization program and exposes much of its functionality to C++ and Python programs. This allows the Movie GUI extension to have detailed control over Chimera without the need to modify the Chimera source code, and allows efficient communication of molecular structure data between the Movie GUI and Chimera. This efficiency and control are absolutely necessary for Movie to be a useful and effective program.

5.5 The Ensemble/Legacy Toolkit Python/C++ Interface

5.5.1 “Wrapping” C++ classes for Python

As mentioned earlier, the Ensemble/Legacy toolkit is written primarily in C++. However, all of the functionality is “wrapped” so as to be available to the Python programming language as modules. Python is a high-level object oriented interpreted programming language that is easy to learn and use. In fact, Python code is so readable it is often called “executable pseudocode”. Python supports heterogeneous lists, hash arrays (called “dictionaries”), and other high-level data structures which are absent from the core C and C++ languages. Python provides a number of advantages over other high-level object-oriented interpreted programming languages. First and foremost, Python is is easy to learn and use, unlike C, or C++. Second, it is easy to understand Python programs written by others, unlike perl, because the Python syntax enforces readability. By using a language which is easy to learn and use, we make it more likely that Ensemble/Legacy will be adopted by users in the scientific community. Also, Python is open source, cross-platform (Unix, Windows, and Macintosh), and is being adopted to solve problems in many different application spaces. There is at least one other[92] published example using Python in molecular modelling and visualization.

One of the most powerful features of the OTF (and the Ensemble/Legacy library as a result) is the ability to automatically "wrap" C++ code as a Python extension. This allows programs written in Python to access OTF data and code directly from Python. The extension method of Python works as follows: a programmer writes code in C++ which declares the structure of the C++ data and code in a manner than Python can understand. This C++ code is compiled to a shared object module (also known as a "dynamic link library" in Windows terminology). A running Python script can "import" a C++ shared object module in the same way it can "import" a regular Python module. Methods within the C++ module which are called by the Python module execute as natively compiled code. When the native code finishes, control returns to the Python intepreter at the point immediately following the method invocation. From the perspective of the Python program, the method call was simply a call to another Python module. For this to work correctly, the C++ module code must be carefully written to use Python data structures to communicate with the calling Python code.

Although the extension mechanism is powerful, it requires a fair amount of work to "wrap" C++ library code which has already been written. For each public method and data item in the C++ code, a "wrapper" function which handles the Python to C++ (and back) translation must be written. If the C++ code is undergoing development, changes to the C++ interface must be reflected in the extension code. To simplify the process of “wrapping”, the OTF provides a powerful tool called “wrappy” [21]. Wrappy takes as input a C++ header file which defines the interface to a C++ class and outputs source code for the Python extension module. The extension has all the necessary support for accessing public methods and data of the C++ class. This automation greatly reduces the time necessary to build the Python/C++ interface and allows the programmer to focus on developing robust code. Ensemble/Legacy uses wrappy to wrap the AMBER molecular structure I/O code, which is written in C++ for maximum speed.

5.5.2 Example: the AMBER Legacy Interface

The most useful and powerful legacy interface is to the AMBER suite of programs. The AMBER legacy interface can convert a protein or nucleic acid stored in the Ensemble::Molecule format so that it can be used by AMBER for molecular mechanics/dynamics, or free energy perturbation simulation. Every AMBER option is exposed as a parameter in the AMBER legacy interface, thus allowing direct control over the course of the simulation. The Ensemble/Legacy toolkit legacy interfaces can automate an entire simulation methodology, from initial model construction to the final analysis stages. Furthermore, since the simulation methodology is stored as a Python program, it is very easy to change simulation parameters and analyze how the changes affect the simulation. The following example demonstrates how a programmer can use the AMBER legacy interface to run minimization and molecular dynamics on a structure from a previous trajectory.

```
##Construct an ``Ensemble.Molecule'' object from the
##our starting structure
m = Ensemble.Molecule(startingStructure)
##Create an ``AMBER Legacy Object'' from the Molecule
##object
a = Ensemble.Legacy.AMBER(m)
```

```
## Now minimize the molecule for 500 steps
## using the AMBER suite
minimizedStructure = a(type=MINIMIZE, steps=500)
## Now run molecular dynamics on the minimized structure
## for 1000 steps
m = Ensemble.Molecule(minimizedStructure)
a = Ensemble.Legacy.AMBER(m)
dynamicsStructures = a(type=DYNAMICS, steps=1000)
```

Note that in the case of a minimization, only a single structure is returned (the minimized structure) while in the case of dynamics, an ensemble of structures is returned (each structure is a step in the dynamics simulation). Since these structures are stored in the Ensemble::Molecule format, they can be submitted to Dials and Windows or CORMA for analysis as shown in the examples below.

5.5.3 Example: the Dials and Windows Legacy interface.

The “Dials and Windows” program requires a collection of single-structure PDB files representing the ensemble of molecular structures for which the helical parameters are to be computed. The Ensemble::Legacy::Dials class has functionality for writing ensembles (using any supported file format) to disk as PDB files using the otf::PDBio class. It is necessary for the Ensemble::Legacy::Dials class to ensure that all the residue names in the molecules written to the PDB file format conform to the file format expected by Dials. Dials requires that the nucleic acid residue names follow a particular convention (ADE, GUA, THY, CYT, URA). This can conflict with AMBER, which uses more descriptive nucleic acid residue names such as DG5 to represent a 5’ terminal deoxyguanosine. Ensemble::Legacy::Dials then generates a Dials input file and runs Dials as a child process. Dials computes the helical parameters for the ensemble of structures. When control returns to the Ensemble::Legacy::Dials module, it determines whether Dials executed successfully, and if so, parses the output file to read in the helical parameters. These data are stored in a collection of hash arrays, indexed by the type of helical parameter.

While the graphical user interface has been written to support common requests (such as running Dials And Windows on an ensemble of nucleic acid structures then plotting the results) it is intended that advanced users will use the programming interface to perform tasks which are not specifically supported by the GUI. To demonstrate the straightforward mapping between the GUI and the progrmaming interface, we will demonstrate using an example which runs Dials And Windows on an ensemble and then plots the results. Each program statement corresponds to a GUI window in the figures.

```
d = Ensemble.Legacy.Dials(dynamicStructures)
##Run Dials and Windows on the ensemble
data = d()
## See figure 2; only plot backbone data
paramType = BACKBONE
## See figure 3; Only plot the pucker and amplitude
parameters = (PUCKER,AMPLITUDE)
## See figure 4; only plot bases 1 and 2
bases = (1,2)
## See figure 5; now plot the data
Ensemble.Analysis.Dials.PlotData(data, bases,
parameters)
```

5.5.4 Example: the CORMA Legacy Interface

The CORMA program computes the R factor of a model structure given experimental NMR data. CORMA requires PDB files representing the molecular structure for which the NMR R factor is to be computed. CORMA requires the PDB files to contain experimental correlation times stored in the "temperature factor" field of the PDB file. The Ensemble::Legacy::Corma class has functionality for taking structures stored in the Ensemble::Molecule class and writing them to disk as PDB files using the otf::PDBio class with user-supplied correlation times. Since correlation times can either be constant for a whole molecule or may vary on a per- atom basis, Ensemble::Legacy::Corma allows the correlation time to be supplied as either a scalar or a vector. Ensemble::Legacy::Corma then generates a CORMA command input file and runs CORMA as a child process. CORMA computes the R factor of the PDB files written by Ensemble:Legacy::Corma. When control returns to Ensemble::Legacy::Corma, it determines if CORMA executed successfully and if so, parses the CORMA output file to determine the R factor value(s). Although the Corma legacy program is designed to produce an R factor for a single PDB file, the Ensemble::Legacy::Corma interface is designed to automate the process of calculating R factors for each structure in an ensemble of structures. In the case of a single structure, a scalar value is returned, while in the case of an ensemble of structures, a vector of R factors is returned. The CORMA application can also take a collection of PDB files and compute an "ensemble" R factor; Ensemble::Legacy::Corma supports this as well. This demonstrates how Ensemble/Legacy is able to handle "ensemble" data in a natural way: as scalar data when a single structure is considered, vector data when an ensemble of structures is considered, and possibly scalar data produced from a reduction of vector data (such as the mean R factor from an ensemble of structures). Here we use the CORMA legacy interface to compute the R factor for each of the structures generated by molecular dynamics above. Then, the Python module "Statistics" is used to compute the mean and standard deviation for the R factors computed by CORMA.

```
c = Ensemble.Legacy.Corma(dynamicStructures)
##Run CORMA on the ensemble
```

```
data = c()
##Average the R factors returned from CORMA
averageR = Statistics.average(data.r)
standardDeviation = Statistics.standardDeviation(data.r)
```

5.6 Conclusion

Modelling biomolecular structures is a fruitful but challenging endeavor. As experimental techniques make rapid advances in the ability to probe molecular flexibility and dynamics, greater demands are being placed on analysis programs to reduce this data to easily presented and understood information. Ensemble/Legacy was designed to address the issues associated with molecular ensembles; in particular, to handle various forms of ensembles in a consistent manner, and to provide services for analyzing ensemble data by providing an object oriented interface to legacy programs. We have chosen a design strategy which allows for the greatest flexibility for the user. While all the necessary tools are made available in the form of an extensible programming library, there is also a graphical user interface which provides this functionality in a user-friendly format. Furthermore, by building on top of the extensible architecture built into the OTF and Chimera, we are able to take advantage of work done by others rather than having to "reinvent the wheel again and again". This library is already being used by several users in the Computer Graphics Lab to analyze real experimental data and produce publication- quality plots, demonstrating the utility of the library in real-life situations.

5.7 Acknowledgements

David Konerding gratefully acknowledges financial support from his advisor (NIH grant GM39247, Thomas James, PI). Additionally, he would like to thank the UCSF

Computer Graphics Laboratory (NIH NCRR P41-RR01081, Thomas Ferrin, PI), for access to computational resources and technical information which made the development of Ensemble/Legacy possible. Furthermore, he would like to show his gratitude to Conrad Huang, Eric Petterson, Greg Couch and Thomas Ferrin of the Computer Graphics Lab for advice and assistance during the development of this program.

5.8 Figures

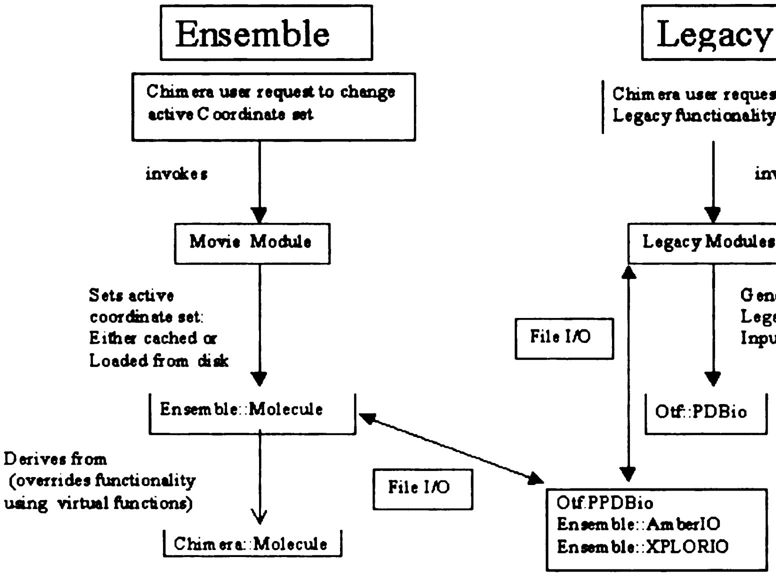

Ensemble
Chimera user request to change active Coordinate set
invokes
Movie Module
Sets active coordinate set: Either cached or Loaded from disk
Ensemble::Molecule
Derives from (overrides functionality using virtual functions)
Chimera::Molecule
File I/O
Legacy
Chimera user request for Legacy functionality
invokes
Legacy Modules
Generates Legacy Input
Otf::PDBio
File I/O
Otf.PPDBio
Ensemble::AmberIO
Ensemble::XPLORIO

Figure 1. Structure of the Ensemble/Legacy toolkit and its interaction with Chimera. The Ensemble interface replaces Chimera's normal Molecule class with a derived class "Ensemble::Molecule".

Figure 5.1: structure of the Ensemble/Legacy toolkit and its interaction with Chimera. The Ensemble interface replaces Chimera's normal Molecule class with a derived class "Ensemble::Molecule".

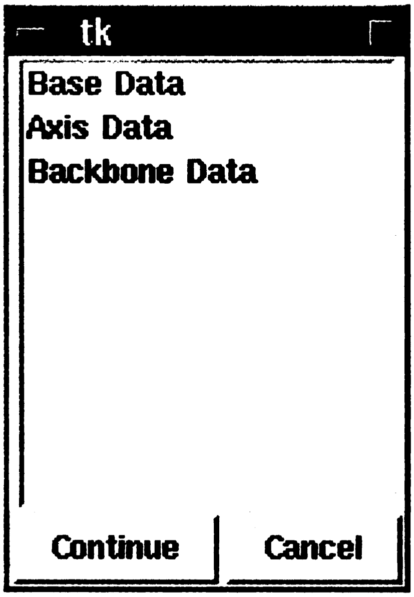


Figure 5.2: The Dials and Windows Parameter Type Selection Dialog allows the user to select which type of Dials and Windows parameter to plot. Dials and Windows computes three types of parameters: "base data", "axis data", and "backbone data". In this figure, the user has selected "backbone data", which presents further choices.

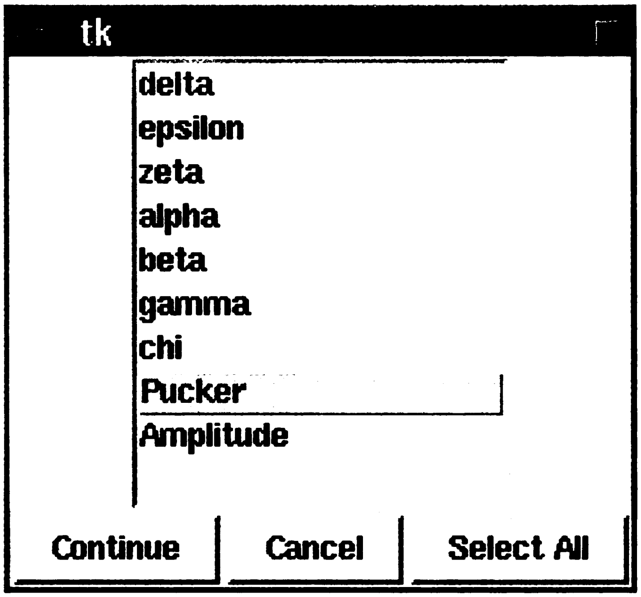


Figure 5.3: The Dials and Windows Backbone Parameter Selection Dialog allows the user to select which type of backbone data to plot. In this figure, the user has selected "Pucker" to plot the time course of the Pucker variable.

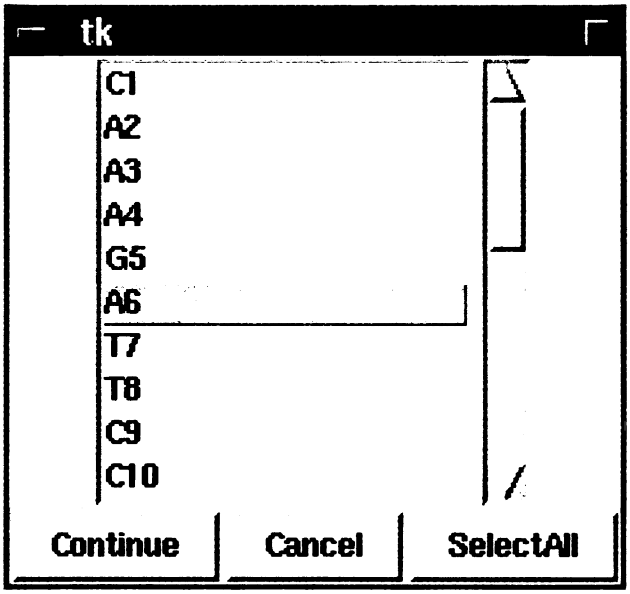


Figure 5.4: The Dials and Windows Base Selection Dialog allows the user to select which bases in the molecule to plot. This list is built at run-time depending on the constituents of the molecule. In this figure the user has selected base "A6" to plot.

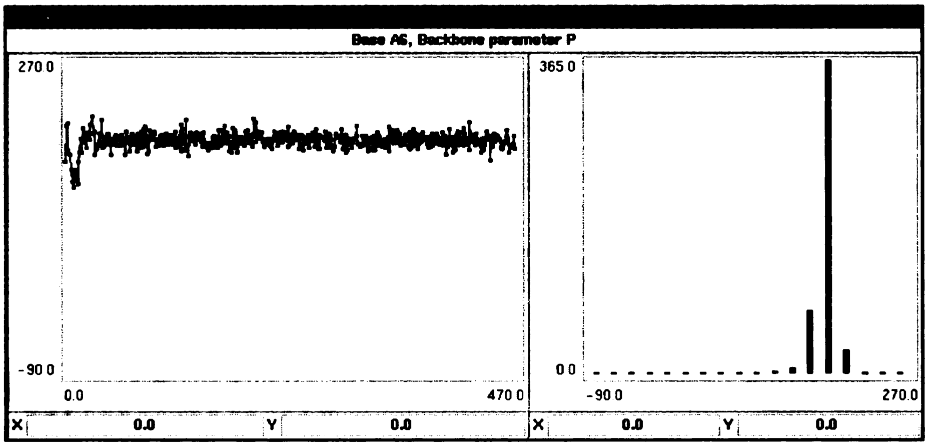


Figure 5.5: The Dials and Windows Data Plot Window plots the data requested by the user. Both the time course and a histogram are plotted. If the Movie GUI is running the user may "jump" to a particular frame by clicking on the corresponding data point in the time course plot.

Chapter 6

Inexpensive Computational Cluster Design for Large Molecular Dynamics Simulations

6.0.1 Cluster design and implementation

A major challenge in producing the long molecular simulations reported in this document was to provide computing resources which could attain the required performance with minimal capital investment and system administration. Molecular dynamics simulations have an enormous requirement for computer resources. The timescales of interesting molecular conformational transitions (a small protein folds to its native conformation in approximately 1 millisecond to 1 second) are 2-6 orders of magnitude longer than the fastest parallel supercomputers are capable of providing in 1 year (longer than most supercomputers would be dedicated to a single calculation). Chemists must be content, at least for now, with performing calculations on very small (500 residues or fewer) proteins or nucleic acids for short time scales (1 nanosecond to 1 microsecond). These times are sufficient to predict a number of properties exhibited by the systems under study, such as transient conformational transitions (kinetics), but not sufficient to determine the molecular partition function or predict native folds (thermodynamics).

We evaluated several choices when faced with the need to provide resources to produce our molecular simulations. First, we determined our requirements: the need to produce at least 2 nanoseconds per month of simulation, and to be able to run 4 multiple simulations simultaneously with no performance degradation. Second, the cost of the system needed to be very low, no more than $15,000. Third, the setup and administration of the system should take very little time.

After examining a number of solutions, including access to external supercomputers, fast RISC servers, and inexpensive personal computers (PCs), a solution based on off-the-self PCs connected by standard 100BaseT ethernet networking technology was selected. External supercomputers were not an acceptable solution because it was not possible to easily obtain enough service units to compute the requisite simulations. Fast but expensive RISC servers were eliminated because the base cost of the systems were too high to run many simulations simultaneously with no loss of performance. The PC solution had a number of advantages: reasonably high compute and network performance, and very low cost hardware due to the standard, off-the-shelf availability of components. The performance of a single PC was about 1/2 that of a high-end RISC server, but at less than 1/4 the cost. Although PCs would normally run the Microsoft Windows operating system, this OS is unsuitable for low-cost, diskless workstations for a number of reasons (lack of diskless support, difficult to administer remotely, expensive compilers). Instead, the system runs the Linux operating system, a zero-cost OS (in capital investment) with diskless support, powerful remote administration support, and a high-quality, high-performance free compiler (gcc). In practice, "zero cost" does not mean "without any monetary or time cost", because Linux, like any operating system, requires system administration.

The final specification of the system is as follows: 6 dual CPU Pentium-III 600MHz PCs with 256MB RAM, 7 Intel EtherPro 100BaseT network interface cards, 1 9GB SCSI hard drive and SCSI adapter, 1 8-port 100BaseT switch ($15,000 total).

One node was designed as the "master" node which contained the hard drive and two NICs ("external"/Internet and "internal"/cluster). Red Hat 6.2 Linux was installed on the master node, with all the necessary packages (NFS server, DHCP server, NTP server, TFTP server, Portable Batch Queuing system). Minimal "images" of the Red Hat 6.2 Linux OS were generated, one per "client" node, and stored on the master node's filesystem. Each image contained the minimum necessary files to boot the system as a diskless workstation. All shared filesystems (/home, /usr/local, and /opt) were mounted from the master node using NFS. Password and other configuration files are synchronized by simple scripts which copy the appropriate files to each of the client images when necessary. The PBS batch queuing system was used to manage the client nodes' compute resources using a straightforward method- no more than one job is assigned per CPU. Users can request a specific number of nodes or require nodes with specific properties to be selected when the job is run via comments in the batch command file.

The AMBER suite of programs was compiled and linked with the MPICH [39] implementation of the MPI libraries, providing support for parallel computation across cluster nodes. Simple PBS batch script codes were used to request multiple processors and MPICH codes were used to inform AMBER on which processors to run.

6.0.2 Cluster performance results

The cluster was tested for accuracy, reliability, and performance. Accuracy was tested by running comparisons against the AMBER reference results and was found to be acceptable. Single-CPU performance of AMBER was determined using the sander binary compiled using the gcc and g77 (the Linux C and Fortran compilers, respectively), tested and found to be as expected given the hardware configuration. Multiple-CPU performance of AMBER was determined using gcc/g77 and the MPICH library. Multiple CPU performance was measured against the single process reference (Figure 6.1). It was noted that scaling of AMBER beyond 3-4 CPUs did not provide any benefit but that multiple jobs running simultaneously scaled independently. The primary cause of the poor scaling is the high latency and low bandwidth of 100BaseT networking. The cluster performed reliably, with individual jobs running many days with no problems; the cluster nodes themselves ran for over 100 days with no hardware failures. After an extensive initial test period, the cluster entered production mode and has remained reliable ever since.

One may argue that this scaling is too poor to be considered for production work, and that specialized supercomputers with very fast interconnect or even a PC cluster with a faster interconnect (such as Myrinet, Giganet, or VIA) than 100BaseT would be a better choice. However, we argue that due to the economics of PC clusters, the choice of relatively slow (low bandwidth and high latency) interconnect is sensible when it meets the performance requirements. If a user requires a single job to finish quickly and a poorly scaling cluster cannot meet that requirement, a specialized system must be considered. However, if many jobs are run simultaneously (and this is often the case when the user wishes to get better sampling of configurational space, or when many users are using the same computing resource), then a poorly scaling cluster will be acceptable. The reason for this is that adding N more nodes to the cluster is a linear cost function (N nodes * cost per node) as long as the cost of the switching technology is linear. This is approximately true up to 32 nodes, and can be extended to even more nodes using a hierarchical switch model. The administrative overhead of maintaining a cluster of many nodes is not that great because each compute node requires only a minimal file system with a few modifications from the template- the network address is the only aspect which needs to be changed. Further, the batch queuing system can automatically remove down nodes from the list of compute servers, so that the system is tolerant of hardware failures on compute nodes. A poorly scaling cluster that meets the requirements will almost certainly have a better price/performance ratio than a specialized cluster.

6.0.3 Incremental Enhancements

The cluster was enhanced over time to accommodate greater numbers of users and jobs. Four dual Celeron 533MHz nodes and three dual Pentium III 667MHz nodes were added. To accommodate the additional network requirements the 8-port 100BaseT switch was replaced with a 16-port 100BaseT switch (Allied Telesyn FS-716). This switch contains enough internal bandwidth to allow full duplex communication between 8 pairs of systems simultaneously (1.6Gbytes/sec). The total cost of 7 new nodes and the switch was under $5000. The operating system was upgraded from Red Hat 6.2 to Red Hat 7.2, which contains the following enhancements: a new version of the Linux kernel (2.4) which supports a journaled filesystem (ext3), a newer version of the C library, newer versions of nearly all user programs and libraries NFS version 3, faster networking and disk I/O, and the associated user-space tools required to utilize these features. No difference in multiple CPU AMBER performance was detected after upgrading the cluster.

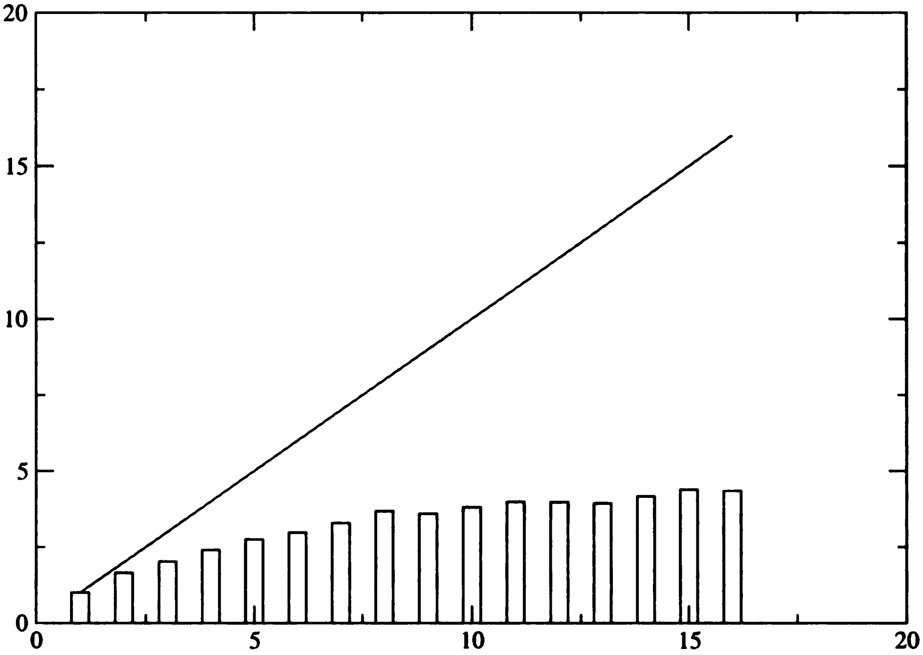


Figure 6.1: Plot of the deviation from scaling of the Linux cluster. Reference value of 1 CPU is 1819 seconds for 1ps of fully solvated molecular dynamics of the Okazaki fragment model using PME, a 2fs timestep, constant pressure and SHAKE on all hydrogen bond lengths and angles. At the reference rate with 1 CPU, 47 ps/day would be produced, 21 days/nanosecond, or 7 months for 10 nanoseconds. It would take 2.9 months for 10 nanoseconds using 4 CPUs or 1.7 months using 16 CPUs.

6.2 Tables

Number of processors	Time in seconds	Scaling factor
1	1819	1
2	1097	1.65
3	892	2.03
4	755	2.40
5	656	2.77
6	584	3.11
7	554	3.28
8	505	3.60
9	498	3.65
10	469	3.87
11	485	3.75
12	466	3.90
13	459	3.96
15	450	4.04
16	442	4.11

Table 6.1: Scaling of cluster

Chapter 7

Future Directions

It is clear from the unrestrained molecular dynamics simulations of RNA:DNA hybrid chimeras and RNA:RNA duplexes that more work is required before we can confidently simulate RNA using molecular dynamics and the AMBER force field. We have shown that even if the AMBER force field does favor the A-RNA conformation over B-RNA (as demonstrated by 10ns simulations which remain in A-RNA and 10ns simulations started in B-form which move away from the B-form to a structure equidistant from canonical A and B), our simulations were not able to complete transition B-form to A-form.

It appears that while the AMBER force field clear favors the A-RNA form, we did not sample conformational space adequately when starting from B-RNA to observe a transition. Because MM-PBSA studies have shown that A-RNA is slightly more favorable than B-RNA in the AMBER force field due mainly to entropic effects rather than enthalpic effects, it is not surprising that we do not see an immediate transition from B-RNA to A-RNA. However, the enhanced conformational sampling of the B-RNA simulation strongly suggests that the B-RNA form is highly unstable and that a 10ns timeframe is not sufficient for B-start simulations to reach a stable conformational region as seen in the A-start simulations. It is not clear that the new Wang et al force field provides any significant improvement over Cornell et al, because at least from the simulations presented here, there is not a large difference in the structures after 10ns (3.25Å), and the Wang et al structure is not any closer to A-RNA at the end of the simulation.

We propose to carry out further MD simulations with the goal of reaching a stable conformation starting from the B-RNA form. In the current study, we simulated an RNA:RNA duplex starting from A-RNA form and B-RNA form, using two force fields, for a total of 4 simulations. Based on the computer resources available, two and later three simulations were run simultaneously to obtain optimal throughput. Rather than running A-start simulations or comparing two force fields, we instead suggest that the simulations all start from the B-RNA form, and that each simulation uses a different random number seed. Even if the characteristic time scale to pass an energy barrier is on the order of 10ns, it is not likely that one simulation will pass that barrier in a 10ns simulation. However, if a large number of simulations are run simultaneously, the chances that at least one of them will overcome the barrier is greatly increased. In the limit of infinite simulations, the estimate of the transition barrier can be determined from the average time to transition. Although infinite simulations are impractical to run, it is possible to greatly enhance the number of simulations which can be run simultaneously for a minimal investment. Because CPU speeds have increased greatly, it is now reasonable to run 10ns simulations on a single CPU within the same time frame that previously required multiple CPUs running in parallel. Although the trend in the molecular dynamics community has been to run long single simulations [24], we argue that not only does the the multiple simulation technique require less expensive computer hardware (no need for fast interconnect) but that it produces statistically more reliable information. By recasting the problem as an embarrassingly parallel one, the same amount of information can be gained for a much lower cost, or, much more information can be gained for the same cost. Of course, if one requires that a simulation be of a specific length which cannot be obtained by a single CPU within a reasonable time, parallelization is still an option, especially small-scale "tight" parallelization (3-4 CPUs) combined with embarrassing parallelization.

7.0.1 Proposed Enhancements

We propose the following hardware configuration to address the problem of locating a transition between B-RNA and A-RNA. This hardware configuration aims to reduce total hardware cost while using the same software configuration previously developed in chapter 6

First, as 1000BaseT (gigabit) networking has become affordable it is worth considering as a replacement for 100BaseT. Gigabit networking does not provide significantly better latency than 100BaseT, but provides much larger bandwidth. For cluster applications which are bandwidth limited, gigabit networking would be a reasonable investment. In our situation, however, gigabit networking will not provide better scaling, since AMBER's parallelization design was tuned for very low latencies. Further, gigabit networking has such a high throughput that it requires highly tuned network drivers, and can use more throughput than is available on the 32bit/33MHz PCI bus. Older CPUs were not even capable of using the complete gigabit bandwidth available due to high CPU utilization but more recent CPUs and network drivers have led to greatly decreased CPU utilitization. It is unlikely that any significant difference would be seen in AMBER scaling using gigabit networking without extensive tuning. Second, although administration of the client nodes is not very challenging, it could be made simpler using a "Single System Image". In the current cluster configuration, each compute node has its own image of the operating system (the root filesystem, /, /etc, and /var). A single system image allows all nodes to share exactly the same instance of the root filesystem. Any node-specific configuration files can be managed using CDSLs

(Context-Dependent Symbolic Links). SSI is not a standard component of any Linux distribution yet, and the non-standard offerings are not yet mature enough to commit to. Third, since clusters tend to generate data much more quickly than single systems, a high performance, high capacity, reliable file system is a necessity. We propose that the cluster be upgraded to use an external SCSI-based RAID5 filesystem. RAID5 is tolerant to single disk failures and external RAID arrays provide significantly better performance and manageability than the software RAID functionality provided by the Linux kernel. Fourth, the client nodes are approaching obsolescence (typically defined as 3 years), so we propose to upgrade all client nodes:

15 Dual Athlon 1800, 256MB RAM, 100BaseT (15*$2,500)

1 16 port 100BaseT switch ($500).

Total cost: $38,000.

Each Athlon 1800 processor should be about 3 times the speed of the Intel 600MHz processor used in the existing cluster configuration, based on SpecFP [spec.org reference] which is an accurate predictor of AMBER performance. If a single CPU is not fast enough to allow the simulation to be run in a reasonable amount of time, the number of total simulations can be cut in half and each 2-CPU node used a small parallel computer. Dual CPUs typically scale better than two CPUs connected by 100BaseT due to the lower latency and higher bandwidth of the motherboard's bus compared to the networking technology. The increase in total computational power from 13 dual 600MHz nodes to 15 dual 1800MHz nodes is significant, approximately 3.5 times faster. By attaching these nodes to the master node already used in the cluster, the cluster will effectively simply be a larger compute resource than before, rather than requiring a large time investment in cluster redesign or monetary investment in the form of faster interconnect or commercial clustering software.

One of the primary issues with embarassingly parallel computing is that the amount of data is greatly increased. Although we propose to address the data capacity issue above with a large RAID filesystem, the other issue is one of practical data reduction. With the current Ensemble/Legacy library, much of the parsing and plotting is done in a very inefficient manner: data is re-parsed from textual output files from programs such as CARNAL and Dials and Windows and then written as textual input to the plotting program. We intend to modify the Ensemble/Legacy library so that the data is only parsed once and then stored in netcdf [9]. netcdf is a platform independent compact data storage format. Python has excellent support for netcdf, and the plotting program used reads netcdf files natively. This will greatly reduce the time required to go from raw trajectory data to readily analyzable plots and tables.

Bibliography

[1] A.A. Adjei, C. Erlichman, J.A. Sloan, J.M. Reid, H.C. Pitot, R.M. Goldberg, P. Peethambaram, A. Atherton, L.J. Hanson, S.R. Alberts, and J. Jett. *Journal of Clinical Oncology*, 18, 2000.

[2] O.T. Avery, C.M MacLeod, and M. McCarty. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. i. induction of transformation by a deoxyribonucleic acid fraction isolated from pneumococcus iii. *Journal of Experimental Medicine*, 79:137–58, 1944.

[3] H.J.C. Berendsen, J.P.M Postma, W.F. van Gunsteren, A. DiNola, and J.R. Haak. *Journal of Chemical Physics*, 81:3684–3690, 1984.

[4] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. E. J. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. *Journal of Molecular Biology*, 112, 1977.

[5] D.L. Beveridge and G. Ravishanker. *Current Opinion in Structural Biology*, 4:246–255, 1994.

[6] B.A. Borgias and T.L. James. *Journal of Magnetic Resonance*, 79, 1988.

[7] B.A. Borgias and T.L. James. *Journal Magnetic Resonance*, 87, 1990.

[8] A. T. Breunger, P. D. Adams, and L. M. Rice. *Current Opinion in Structural Biology*, 8(5):606–11, 1998.

[9] S.A. Brown, M. Folk, G. Goucher, and R. Rew. Software for portable scientific data management. *Computers in Physics*, 7(3):304–308, 1993.

[10] M.J. Byrne, J.A. Davidson, A.W. Musk, J. Dewar, G. van Hazel, M. Buck, N.H. de Klerk, and B.W.S. Robinson. *Journal of Clinical Oncology*, 17, 1999.

[11] C.R. Cantor, M.M. Warshow, and H. Shapiro. *Biopolymers*, 9, 1970.

[12] D. A. Case, D. A. Pearlman, J. W. Caldwell, T. E. Cheatham III, W. S. Ross, C. Simmerling, T. Darden, K. M. Merz, R. V. Stanton, A. Chen, J. J. Vincent, M. Crowley, V. Tsui, R. Radmer, Y. Duan, J. Pitera, I. Massova, G. L. Seibel, U. Chandra Singh, P. Weiner, and P. A. Kollman. *AMBER version 6.0*, 1999.

[13] D.A. Case, D.A. Pearlman, J.W. Caldwell, T.E. Cheatham III, W.S. Ross, C.L. Simmerling, T.A. Darden, K.M. Merz, R.V. Stanton, A.L. Cheng, J.J. Vincent, M. Crowley, D.M. Ferguson, R.J. Radmer, G.L. Seibel, U.C. Singh, P.K. Weiner, and P.A. Kollman. *University of California, San Francisco*, 2000.

[14] T.E. Cheatham and P.A. Kollman. *Journal of Molecular Biology*, 259:434–444, 1996.

[15] T.E. Cheatham and P.A. Kollman. *Structure, Motion, Interaction and Expression of Biological Macromolecules, Proceedings of the Tenth Conversation*, 1998.

[16] T.E. Cheatham, J.L. Miller, T. Fox, T.A. Darden, and P.A. Kollman. *Journal of the American Chemical Society*, 117:4193–4194, 1995.

[17] T. E. Cheatham, III, P. Cieplak, and P. Kollman. A modified version of the cornell et al force field with improved sugar pucker phases and helical repeat. *Journal of Biomolecular Structure and Dynamics*, 1999.

[18] T.E. Cheatham, III and P.A. Kollman. Observation of the a-DNA to b-DNA transition during unrestrained molecular dynamics in aqueous solution. *JMB*, 1996.

[19] T.E. Cheatham, III and P.A. Kollman. Molecular dynamics simulations highlight the structural differences amount DNA:DNA, RNA:RNA, and DNA:RNA hybrid duplexes. *JACS*, 1997.

[20] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K. M. Merz, J., D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, and P.A. Kollman. *Journal of the American Chemical Society*, 117:5179–5197, 1995.

[21] G.S. Couch. Wrappy – a python wrapper generator for c++ classes. *OReilly Open Source Convention Python Conference Proceedings, 1999, http://conferences.oreilly.com*, 1999.

[22] T.A. Darden, D. York, and L.G. Pedersen. *Journal of Chemical Physics*, 98:10089–10092, 1993.

[23] H. Von der Maase. *European Journal Cancer*, 36, 2000.

[24] Y. Duan and P.A. Kollman. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science*, 282:740–4, 1998.

[25] U. Essman, L. Perera, M. Berkowitz, T. Darden, H. Lee, and L. Pedersen. *Journal of Chemical Physics*, 103:8577–8593, 1995.

[26] T. E. Ferrin, G. S. Couch, C. C. Huang, E. F. Pettersen, and R. Langridge. *Journal of Molecular Graphics*, 9(1):27–32, 1991.

[27] P.M. Fracasso, B.R. Tan Jr., M. Grieff, J.S. Stephenson Jr., H. Liapis, N.L. Umbeck, D.D. Von Hoff, and E.K. Rowinsky. *Journal of the National Cancer Institute*, 91, 1999.

[28] R.E. Franklin and R.G. Gosling. *Acta Crystallographica*, 6:673–677, 1953.

[29] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, V.G. Zakrzewski, J.A. Montgomery, R.E. Stratmann, J.C. Burant, S. Dapprich, J.M. Millam, A.D. Daniels, K.N. Kudin, M.C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi., R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G.A. Petersson, P.Y. Ayala, Q. Cui, K. Morokuma, D.K. Malick, A.D. Rabuck, K. Raghavachari, J.B. Foresman, J. Cioslowski, J.V. Ortiz, B.B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R.L. Martin, D.J. Fox, T. Keith, M.A. Al-Laham, C.Y. Peng, A. Nanayakkara, C. Gonzalez, M. Challacombe, P.M.W. Gill, B.G. Johnson, W. Chen, M.W. Wong, J.L. Andres, M. Head-Gordon, E.S. Replogle, and J.A. Pople. *Gaussian, Inc., Pittsburgh PA*, 1998.

[30] B. Gilquin, C. Guilbert, and D. Perahia. Unfolding of hen egg lysozyme by molecular dynamics simulations at 300k: insight into the role of the interdomain interface. *Proteins*, 41(1):58–74, 2000.

[31] W. H. Gmeiner, D. Konerding, and T. L. James. Effect of cytarabine on the NMR structure of a model okazaki fragment from the sv40 genome. *Biochemistry*, 1999.

[32] W. H. Gmeiner, D. Konerding, R. T. Pon, and T. L. James. NMR structure of a model okazaki fragment from the sv40 genome. *Biochemistry*.

[33] W. H. Gmeiner, A. Skradis, R. T. Pon, and J. Liu. Cytarabine-induced destabilization of a model okazaki fragment. *NAR*, 1998.

[34] W.H. Gmeiner. *Current Medicinal Chemistry*, 5, 1998.

[35] W.H. Gmeiner, D. Konerding, and T.L. James. *Biochemistry*, 38, 1999.

[36] W.H. Gmeiner, A. Skradis, R.T. Pon, and J-Q. Liu. *Nucleic Acids Research*, 26, 1998.

[37] S. Greggi, M.G. Salerno, G. DAgostino, G. Ferrandina, D. Lorusso, L. Manzione, S. Mancuso, and G. Scambia. *Oncology*, 60, 2000.

[38] C. Griesinger, O.W. Sorenson, and R.R. Ernst. *Journal of the American Chemical Society*, 107, 1985.

[39] W. Gropp, E. Lusk, N. Doss, and A. Skjellum. A high-performance, portable implementation of the MPI message passing interface standard. *Parallel Computing*, 22(6):789–828, sep 1996.

[40] V. Heinemann. *Oncology*, 60, 2001.

[41] V. Heinemann, L. Schulz, R.D. Issels, and W. Plunkett. *Seminars in Oncology*, 22 (suppl 11), 1995.

[42] V. Heinemann, L. Schulz, R.D. Issels, and W. Plunkett. *Seminars in Oncology*, 22, 1997.

[43] A.D. Hershey and M. Chase. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *Journal of General Physiology*, pages 39–56, 1956.

[44] P.J. Hore. *Journal of Magnetic Resonance*, 55, 1983.

[45] C. Huang, G. Couch, E. Petterson, and T. E. Ferrin. Chimera: An extensible molecular modeling application constructed using standard components. *Proceedings of the Pacific Symposium on Biocomputing*, 1996(1), 1996.

[46] C. C. Huang, G. S. Couch, E. F. Pettersen, T. E. Ferrin, A. E. Howard, and T. E. Klein. *Proceedings of the Pacific Symposium in Biocomputing*, pages 349–61, 1998.

[47] C.C. Huang, G.S. Couch, E.F. Pettersen, and T.E. Ferrin. *Proceedings of the Pacific Symposium on Biocomputing*, 1996.

[48] P. Huang, S. Chubb, L.W. Hertel, G.B. Grindley, and W. Plunkett. *Cancer Research*, 51, 1991.

[49] H.A. Burris III, M.J. Moore, J. Andersen, M.R. Green, M.L. Rothenberg, M.R. Modiano, M.C. Cripps, R.K. Portenoy, A.M. Storniolo, P. Tarassoff, R. Nelson, F.A. Dorr, C.D. Stephens, and D.D Von Hoff. *Journal Clinical Oncology*, 15, 1997.

[50] A. Illiano, E. Barletta, V. de Marino, C. Battiloro, M. Barzelloni, F. Scognamiglio, N. Rossi, G. Zampa, M. de Bellis, and C. Gridelli. *Anticancer Research*, 20, 2000.

[51] T.L. James. *Current Opinion in Structural Biology*, 1:1042–1053, 1991.

[52] P.A. Bunn Jr. and K. Kelly. *Clinical Cancer Research*, 5, 1998.

[53] Peter A. Kollman Junmei Wang, Piotr Cieplak. How well does a restrained electrostatic potential (resp) model perform in calculating conformational energies of organic and biological molecules. *Journal of Computational Chemistry*, 2000.

[54] D. E. Konerding, T. E. Cheatham, III, P. A. Kollman, and T. L. James. Restrained molecular dynamics of solvated duplex DNA using the particle mesh ewald method. *JBNMR*, 1999.

[55] A. Kornberg, I.R. Lehman, M.J. Bessman, and E.S. Simms. Enzymatic synthesis of deoxyribonucleic acid. *Biochimica et Biophysica Acta*, 21:197–98, 1956.

[56] R. Kotchetkov, B. Groeschel, W.H. Gmeiner, A.A. Krivtchik, E. Trump, M. Bitoova, J. Cinatl, B. Kornhuber, and J. Cinatl. *Anticancer Research*, 20, 2000.

[57] J.R. Kroep, G. Giaccone, C. Tolis, D.A. Voorn, W.J.P. Loves, C.J. van Groeningen, H.M. Pinedo, and G.J. Peters. *British Journal of Cancer*, 38, 2000.

[58] D.W. Kufe, D. Munroe, D. Herrick, E. Egan, and D. Spriggs. *Molecular Pharmacology*, 26, 1984.

[59] R. Lavery and H. Sklenar. *Journal of Biomolecular Structure and Dynamics*, 6., 1989.

[60] R. Lavery and H. Sklenar. *Journal of Biomolecular Structure and Dynamics*, 6:655–667, 1989.

[61] R. Lavery and H. Sklenar. *Journal of Biomolecular Structure and Dynamics*, 6(4):655–67, 1989.

[62] T.S. Lawrence, M.A. Davis, A. Hough, and A. Rehemtulla. *Clinical Cancer Research*, 7, 2001.

[63] H. Liu, H. Spielmann, N. Ulyanov, and D.E. Wemmer. *Journal of Biomolecular NMR*, 6:390–402, 1995.

[64] H. Liu, H. P. Spielmann, N. B. Ulyanov, D. E. Wemmer, and T. L. James. *Journal of Biomolecular NMR*, 6(4):390–402, 1995.

[65] S. Louise-May, P. Auffinger, and E. Westhof. *Current Opinion in Structural Biology*, 6:289–298, 1996.

[66] L.A. Marky, K.S. Blumenfeld, S. Kozlowski, and K.J. Breslauer. *Biopolymers*, 22, 1983.

[67] L.A. Marky and K.J. Breslauer. *Biopolymers*, 26, 1987.

[68] M. Meyers, M.W. Wagner, H-S. Hwang, T.J. Kinsella, and D.A. Boothman. *Cancer Research*, 61, 2001.

[69] Jennifer L. Miller and Peter A. Kollman. Theoretical studies of an exceptionally stable RNA tetraloop: Observation of convergence from an incorrect nmr structure to the correct one using unrestrained molecular dynamics. *Journal of Molecular Biology*, 270:436–450, 1998.

[70] C.J.A. Van Moorsel, H.M. Pinedo, G. Veerman, A.M. Bergman, C.M. Kuiper, J.B. Vermorken, W.J.F. Van der Vijgh, and G.J. Peters. *British Journal of Cancer*, 35, 1999.

[71] A. Mujeeb, S.M. Kerwin, G.L. Kenyon, and T.L. James. *Biochemistry*, 32:13419–13431, 1993.

[72] A. Nicholls and B. Honig. *Journal of Computational Chemistry*, 12, 1991.

[73] H. Oettle, D. Arnold, C. Hempel, and H. Rless. *Anticancer Drugs*, 11, 2000.

[74] D. A. Pearlman, D. A. Case, J. C. Caldwell, W. S. Ross, T. E. Cheatham III, D. N. Ferguson, G. L. Seibel, U. C. Singh, P. K. Weiner, and P. A. Kollman. *AMBER version 4.1*, 1995.

[75] G.J. Peters, C.L. van der Wilt, C.J.A. van Moorsel, J.R. Kroep, A.M. Bergmann, and S.P. Ackland. *Pharmaceutical Therapeutics*, 87, 2000.

[76] W. Plunkett, P. Huang, C.E. Searcy, and V. Gandhi. *Seminars in Oncology*, 23, 1996.

[77] W. Plunkett, P. Huang, Y.X. Xu, V. Heinemann, R. Grunewald, and V. Gandhi. *Seminars in Oncology*, 22 (suppl 11), 1995.

[78] P. Pourquier, Y. Takebayashi, Y. Urasaki, C. Gioffre, G. Kohlhagen, and Y. Pommier. *PNAS*, 97:1885–1890, 2000.

[79] G.G. Prive, K. Yanagi, and R.E. Dickerson. The structure of the b-DNA decamer C-C-A-A-C-G-T-T-G-G and comparison with the isomorphous decamers C-C-A-A-G-A-T-T-G-G and C-C-A-G-G-C-C-T-G-G. *Journal of Molecular Biology*, 217(177), 1991.

[80] G. Ravishanker, S. Swaminathan, D. L. Beveridge, R. Lavery, and H. Sklenar. *Journal of Biomolecular Structure and Dynamics*, 6(4), 1998.

[81] G. Ravishanker, S. Swaminathan, D.L. Beveridge, R. Lavery, and H. Sklenar. *Journal of Biomolecular Structure and Dynamics*, 6, 1989.

[82] G. Ravishanker, S. Swaminathan, D.L. Beveridge, R. Lavery, and H. Sklenar. *Journal of Biomolecular Structure and Dynamics*, 6:669–699, 1989.

[83] D. Rhodes and A. Klug. *Nature*, 286:573–578, 1980.

[84] D. Rhodes and A. Klug. *Nature*, 286:573–578, 1980.

[85] F.C. Richardson, K.K. Richardson, J.S. Kroin, and L.W. Hertel. *Nucleic Acids Research*, 20, 1992.

[86] A. Roitberg and R. Elber. *Journal of Chemical Physics*, 95:9277–9286, 1991.

[87] D.D. Ross, D.P. Cuddy, N. Cohen, and D.R. Hensley. *Cancer Chemotherapy and Pharmacology*, 31, 1992.

[88] G. Rossum. *http://www.python.org*, 1991.

[89] J.P. Ryckaert, G. Cicotti, and H.J.C Berendsen. *Journal of Computational Physics*, 23, 1977.

[90] P.V. Sahasrabudhe, R.T. Pon, and W.H. Gmeiner. *Nucleic Acids Research*, 23, 1995.

[91] P.V. Sahasrabudhe, R.T. Pon, and W.H. Gmeiner. *Biochemistry*, 35, 1996.

[92] M. F. Sanner, B. S. Duncan, C. J. Carillo, and A. J. Olson. *Proceedings of the Pacific Symposium on Biocomputing*, 1999.

[93] M. F. Sanner, J.-C. Spehner, and A.J. Olson. *Biopolymers*, 38(3):305–20, 1996.

[94] M.F. Sanner, A.J. Olson, and J-C. Spehner. *Proceedings of the 11th ACM Symposium on Computational Geometry*, 1995.

[95] U. Schmitz and T.L. James. *Methods in Enzymology*, 261:3–44, 1995.

[96] A.D. Seidman. *Oncology*, 60, 2001.

[97] A.D. Seidman. *Oncology*, 15, 2001.

[98] C. Simmerling, J. Miller, and P.A. Kollman. *Journal of the American Chemical Society*, 120:7149–7155, 1998.

[99] U.C. Singh, S.J. Weiner, and P.A. Kollman. *PNAS*, 82, 1985.

[100] Jayashree Srinivasan, Thomas E. Cheatham, Piotr Cieplak, Peter A. Kollman, and David A. Case. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. *Journal of the American Chemical Society*, 120(7):9401–9409, sep 1998.

[101] D.J. States, R.A. Haberkorn, and D.J. Ruben. *Journal of Magnetic Resonance*, 48, 1982.

[102] K. Stott and J. Keeler. *Magnetic Resonance in Chemistry*, 34, 1996.

[103] N. B. Ulyanov, U. Schmitz, and T. L. James. Metropolis monte carlo calculations of DNA structure using internal coordinates and NMR distance restraints: An alternative method for generating a high-resolution solution structure. *JBNMR*, 1993.

[104] N.B. Ulyanov, A.A. Gorin, V.B. Zhurkin, B.-C. Chen, M.H. Sarma, and R.H. Sarma. *Biochemistry*, 31:3918–3930, 1992.

[105] N.B. Ulyanov and T.L. James. Statistical analysius of DNA duplex structures in solution derived by high resolution NMR. *Applied Magnetic Resonance*, 1994.

[106] N.B. Ulyanov, U. Schmitz, A. Kumar, and T.L. James. *Biophysical Journal*, 68:13–24, 1995.

[107] Nikolai Ulyanov, Ulrich Schmitz, A. Kumar, and Thomas L. James. Probability assessment of conformational ensembles - sugar repuckering in a DNA duplex in solution. *Biophysical Journal*, 68(1):13–24, 1995.

[108] A.H. Wang, G.J. Quigley, F.J. Kolpak, J.L. Crawford, J.H. van Boom, G. van der Marel, and A. Rich. *Nature*, 283:680, 1979.

[109] J. Wang, W. Wang, and P.A. Kollman. *Journal Computational Chemistry*, submitted.

[110] J.D. Watson and F.H.C. Crick. A structure of deoxyribose nucleic acid. *Nature*, 171:737–38, 1953.

[111] K. Weisz, R.H. Shafer, W. Egan, and T.L. James. *Biochemistry*, 31:7477–7487, 1992.

[112] K. Weisz, R.H. Shafer, W. Egan, and T.L. James. *Biochemistry*, 33:354–356, 1994.

[113] D.M. York, T.A. Darden, and L.G. Pedersen. *Journal of Chemical Physics*, 99:8345–8348, 1993.

[114] M.A. Young, G. Ravishanker, and D.L. Beveridge. *Biophysical Journal*, 73:2313–2336, 1997.

Chapter 8

Appendices

8.1 Supplementary tables and figures for Gmeiner paper

Parameter Schedules used in rMD refinement of [GEM].

Parameter	Initial Step	Final Step	Initial Weight	Final Weight
TEMP0	0	1000	0.4	0.4
TEMP0	1001	5000	0.4	100.0
TEMP0	5001	15000	100.0	300.0
TEMP0	15001	18000	300.0	150.0
TEMP0	18001	30000	150.0	150.0
REST	0	1000	0.0	0.0
REST	1001	5000	0.0	1.0
REST	5001	30000	1.0	1.0

Table 8.1: Initial refinement of starting models.

Parameter	Initial Step	Final Step	Initial Weight	Final Weight
TEMP0	0	500	100	100
TEMP0	501	2000	100	300
TEMP0	2001	10000	300	300
TEMP0	10001	20000	300	100
REST	0	500	0	0.1
REST	501	1000	0.1	1
REST	1001	20000	1	1

Table 8.2: Room temperature final refinement of averaged structures.

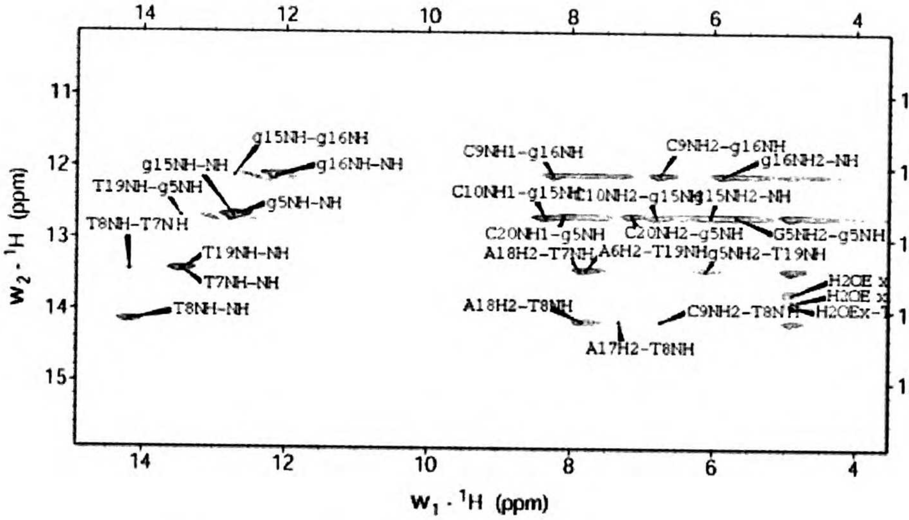


Figure 8.1: 2DNOESY of the exchangeable 1H region of GEM

2D TOCSY Spectra of [GEM].

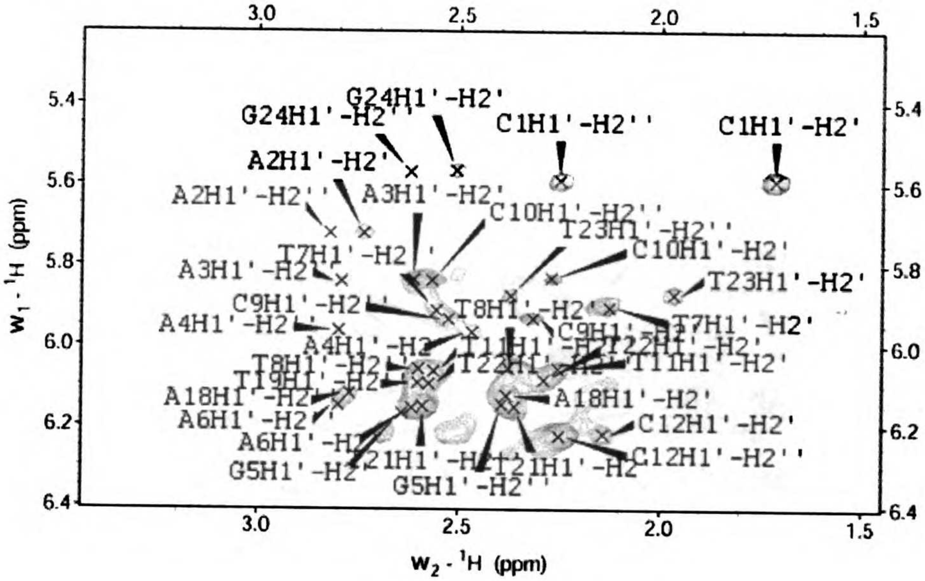


Figure 8.2: H1'-H2'/H2" Region.

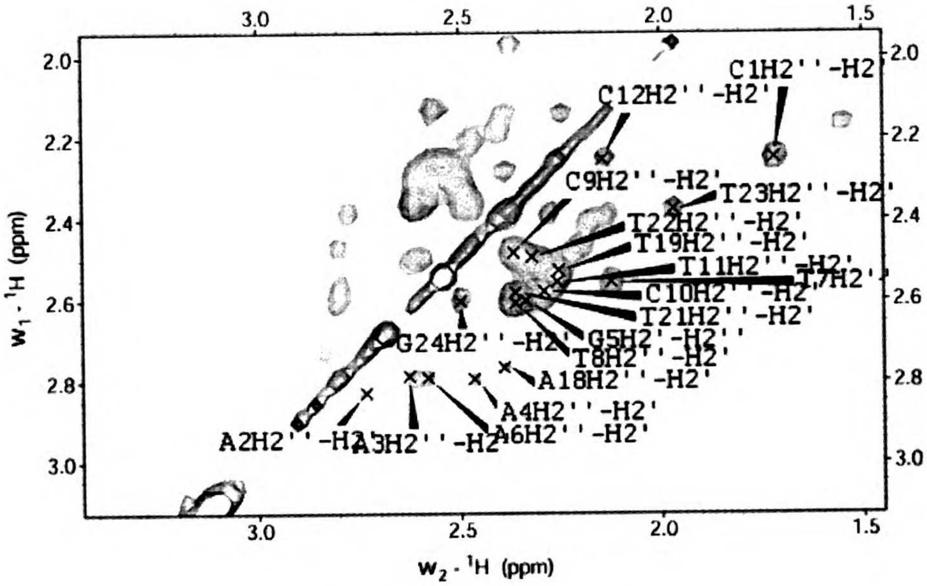


Figure 8.3: H2'/H2" Region of GEM

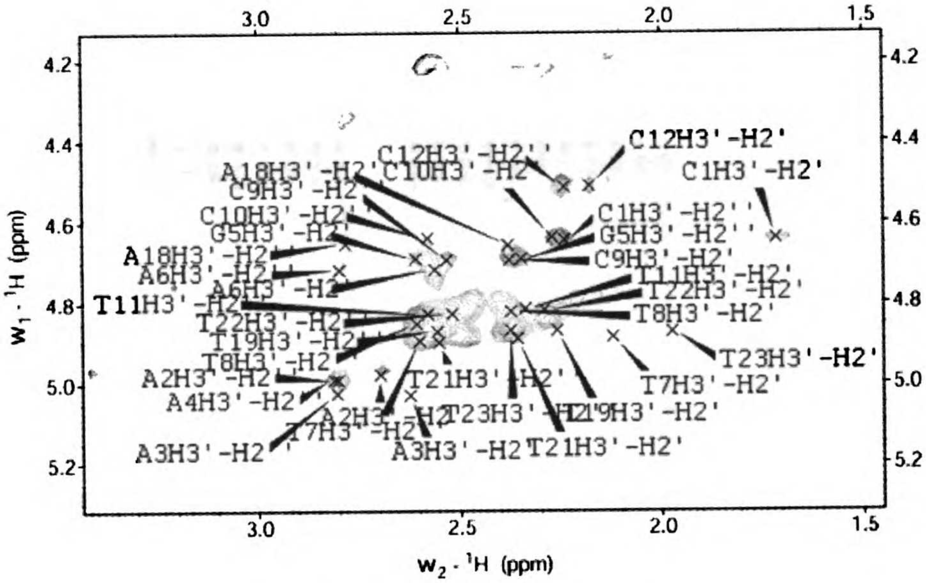


Figure 8.4: H3'-H2'/H2" Region of GEM

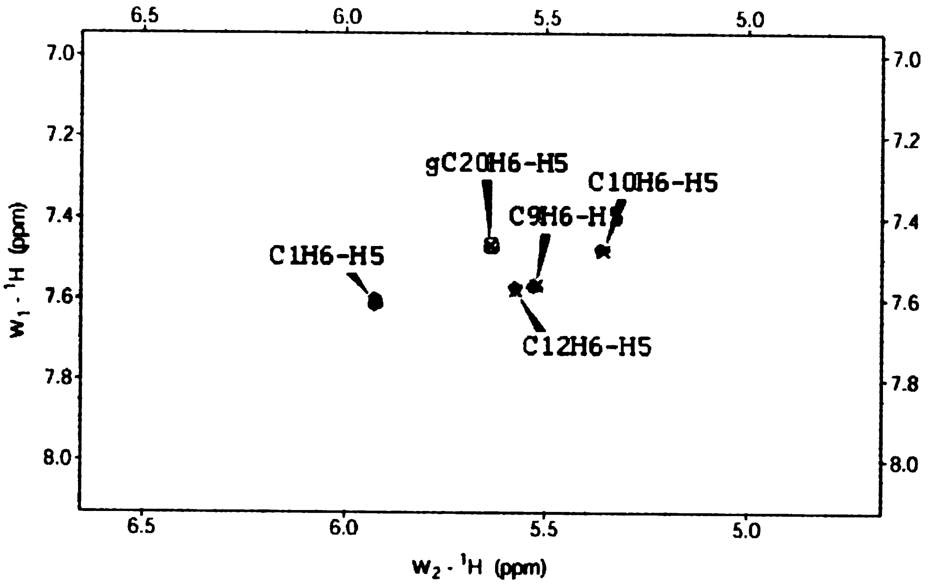


Figure 8.5: H6-H5 Region of GEM

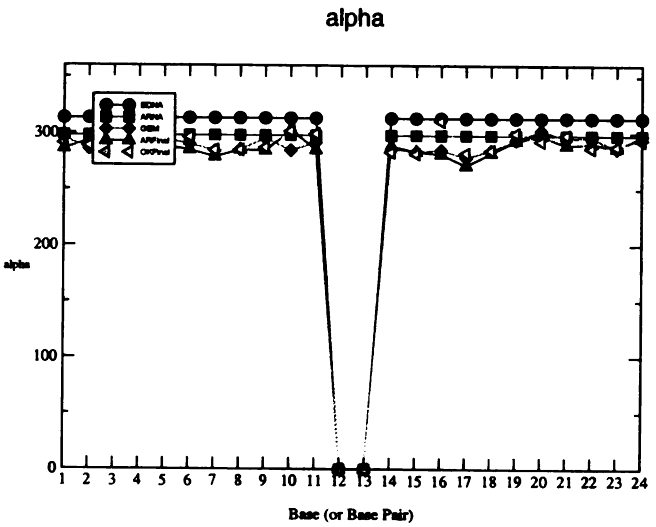


Figure 8.6: Dials plot of alpha parameter

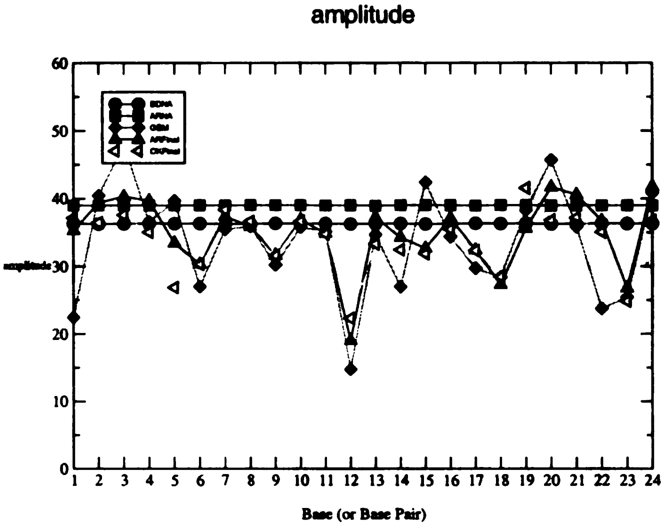


Figure 8.7: Dials plot of amplitude parameter

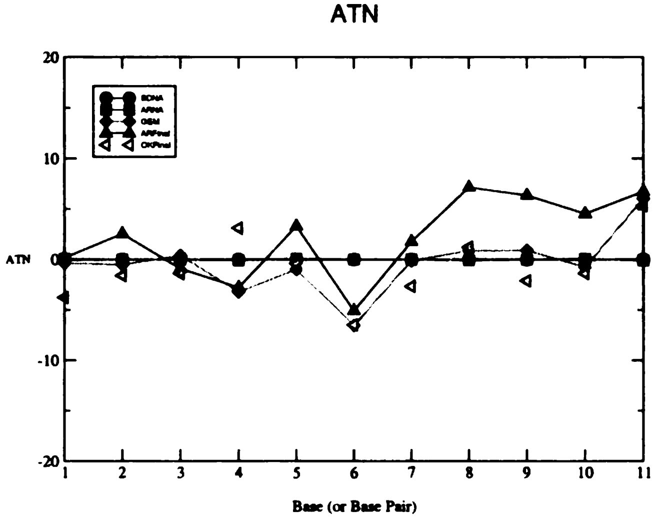


Figure 8.8: Dials plot of ATN parameter

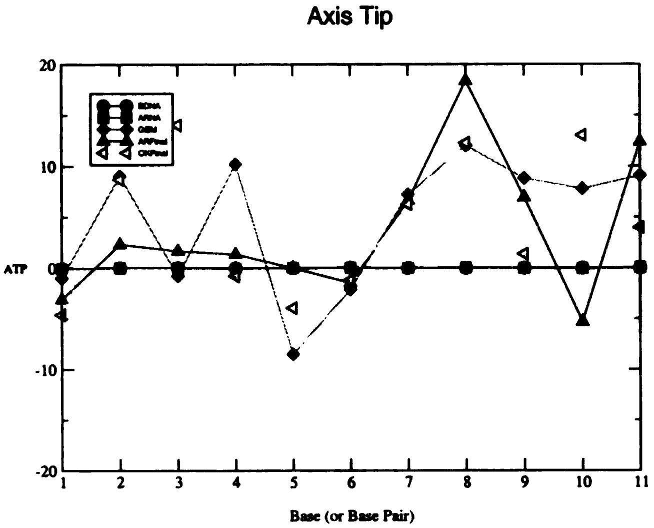


Figure 8.9: Dials plot of ATP parameter

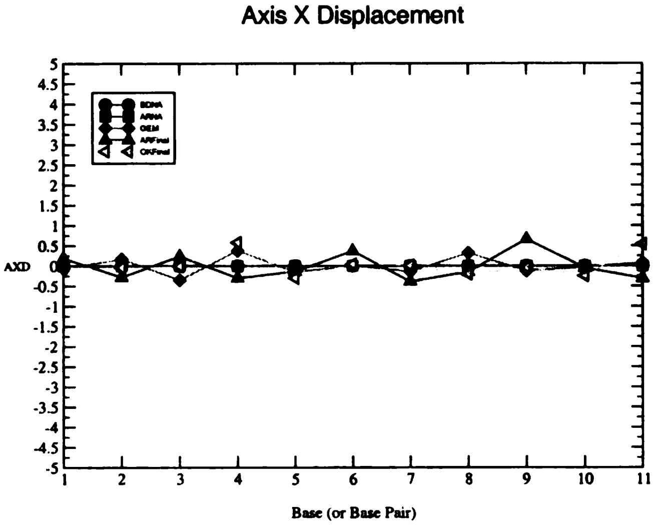


Figure 8.10: Dials plot of AXD parameter

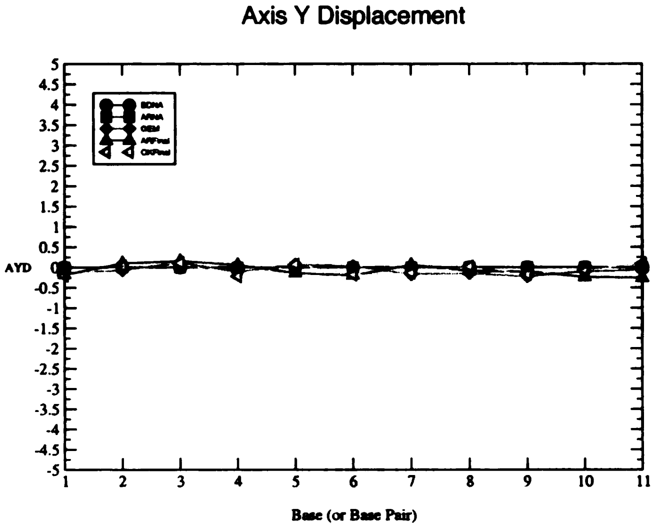


Figure 8.11: Dials plot of AYD parameter

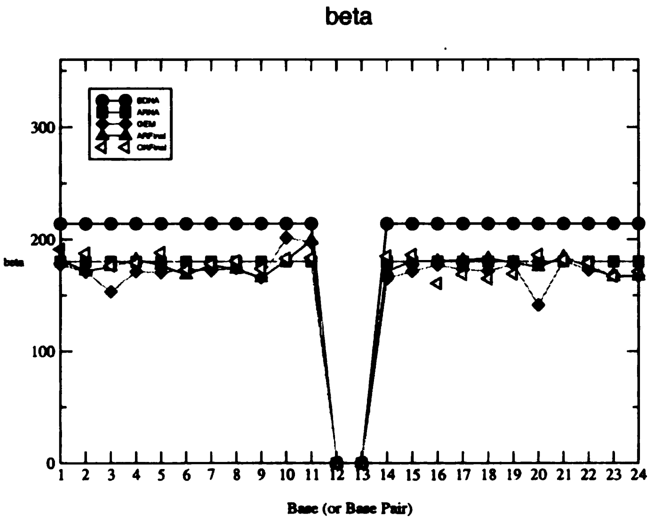


Figure 8.12: Dials plot of beta parameter

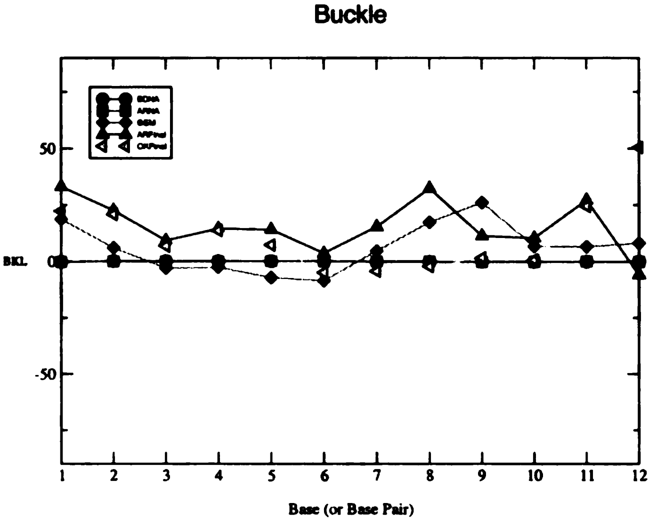


Figure 8.13: Dials plot of BKL parameter

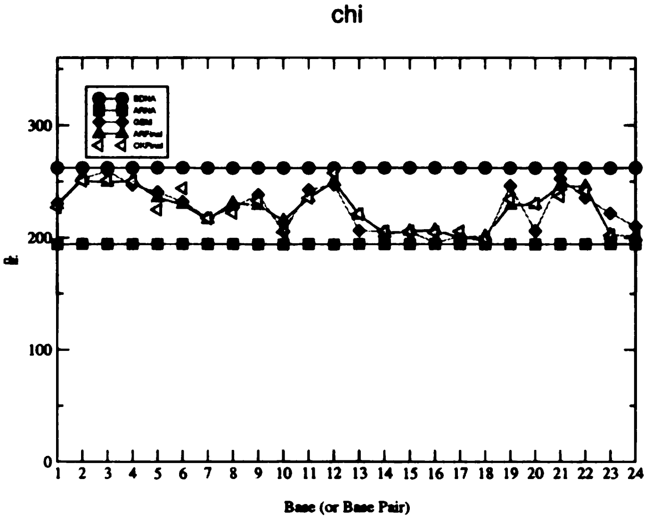


Figure 8.14: Dials plot of chi parameter

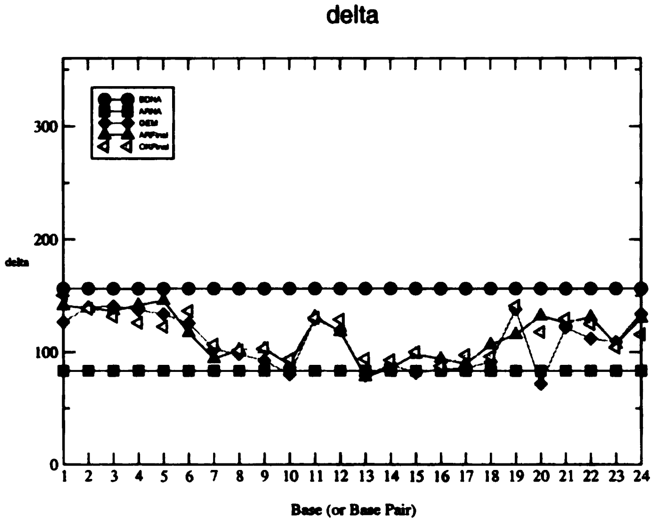


Figure 8.15: Dials plot of delta parameter

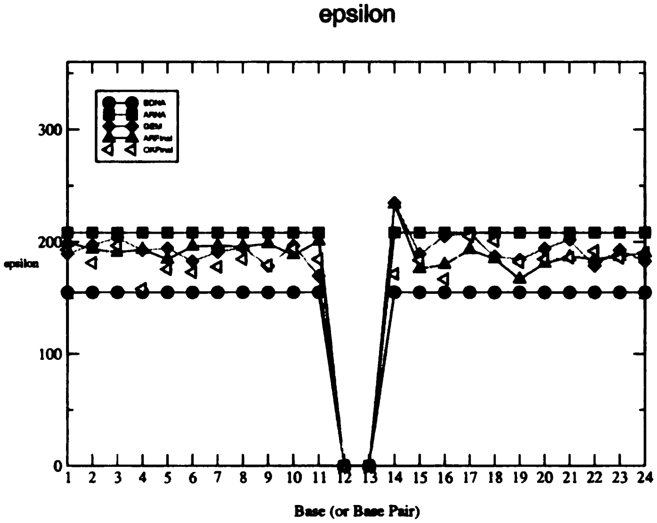


Figure 8.16: Dials plot of epsilon parameter

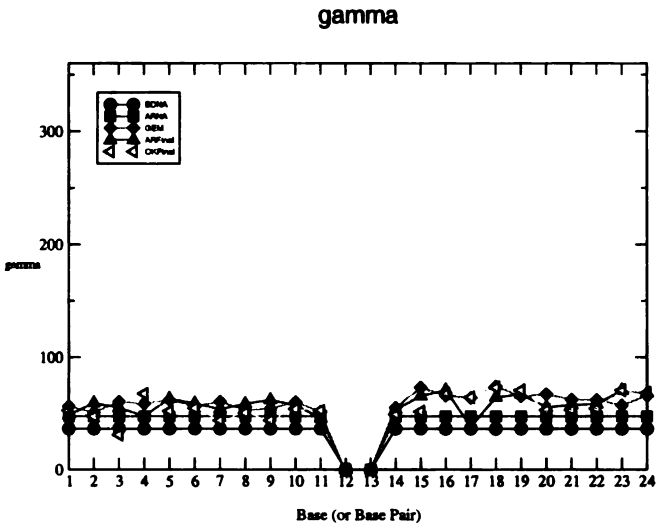


Figure 8.17: Dials plot of gamma parameter

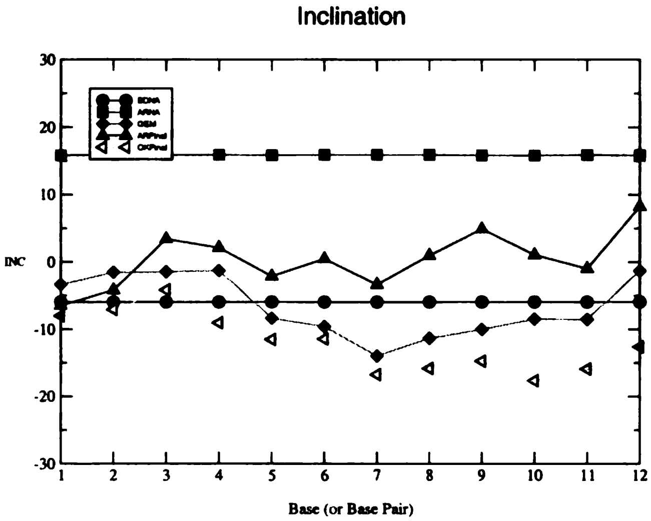


Figure 8.18: Dials plot of INC parameter

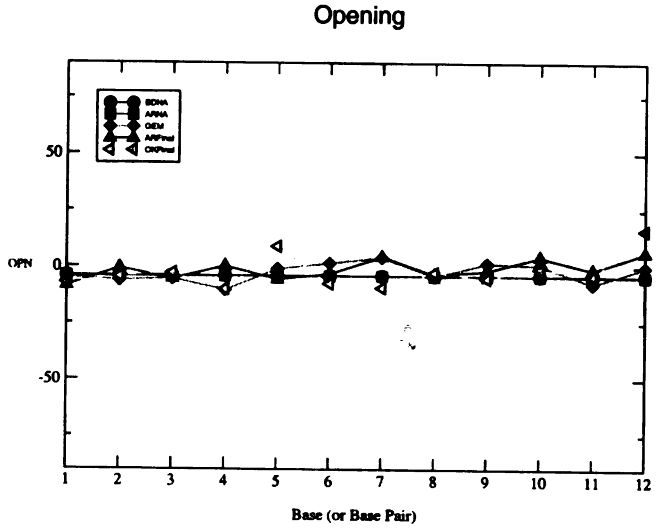


Figure 8.19: Dials plot of OPN parameter

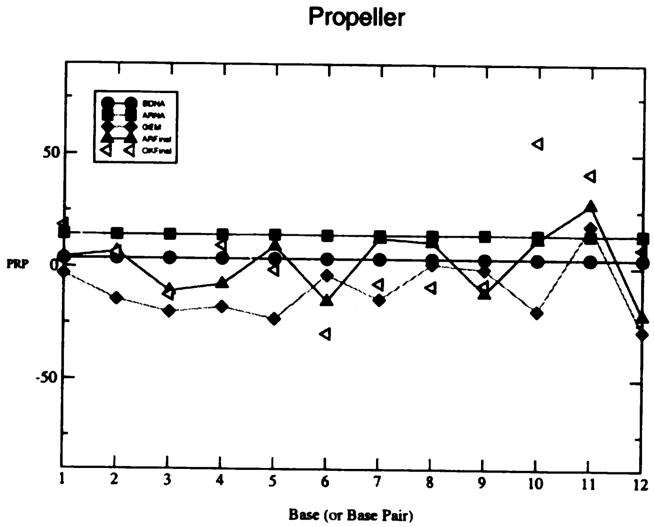


Figure 8.20: Dials plot of PRP parameter

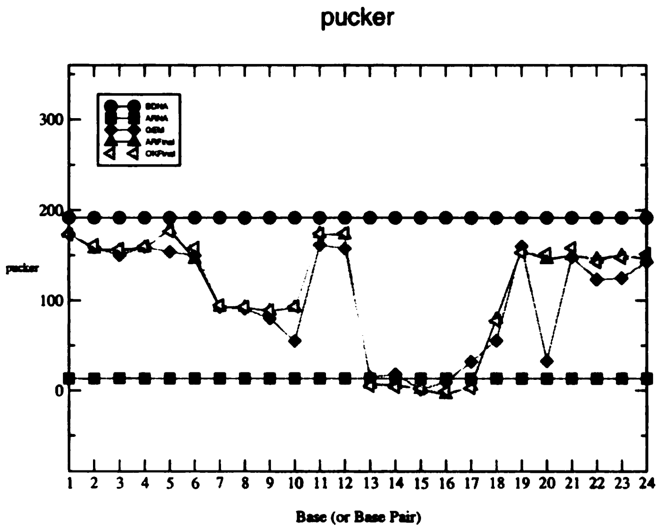


Figure 8.21: Dials plot of pucker parameter

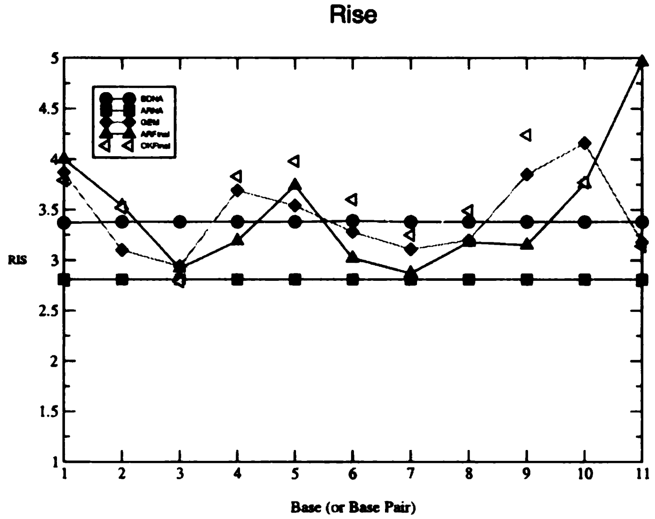


Figure 8.22: Dials plot of RIS parameter

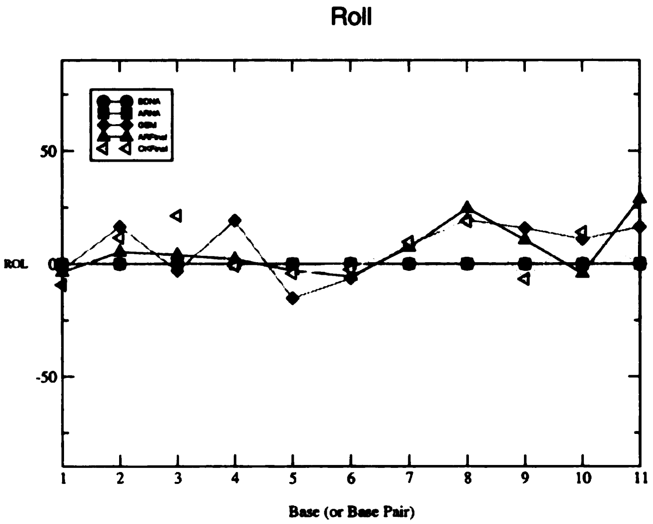


Figure 8.23: Dials plot of ROL parameter

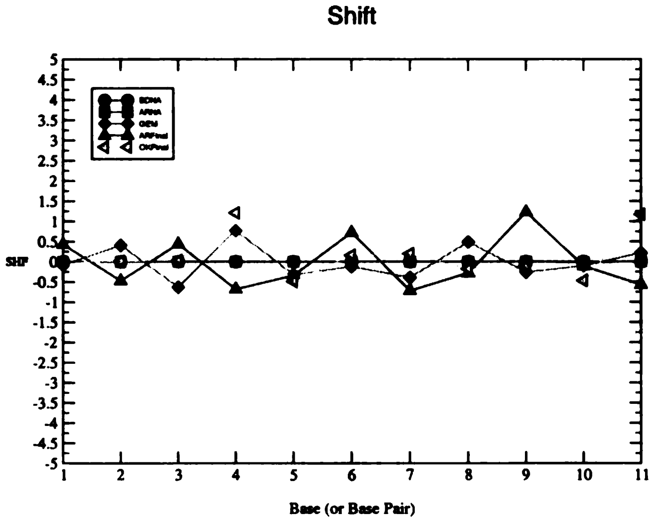


Figure 8.24: Dials plot of SHF parameter

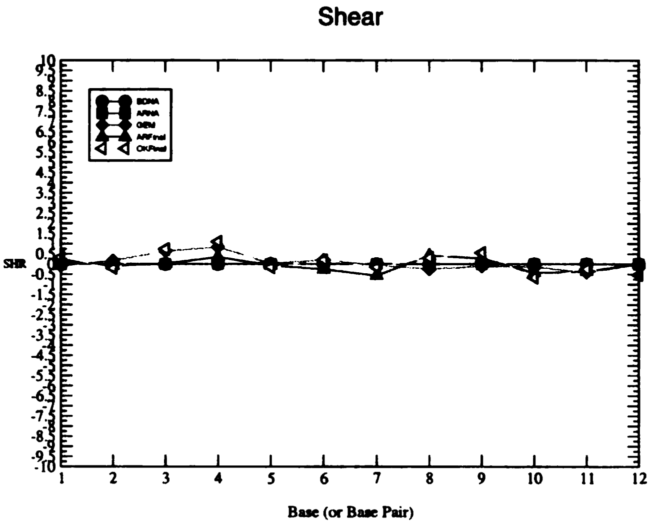


Figure 8.25: Dials plot of SHR parameter

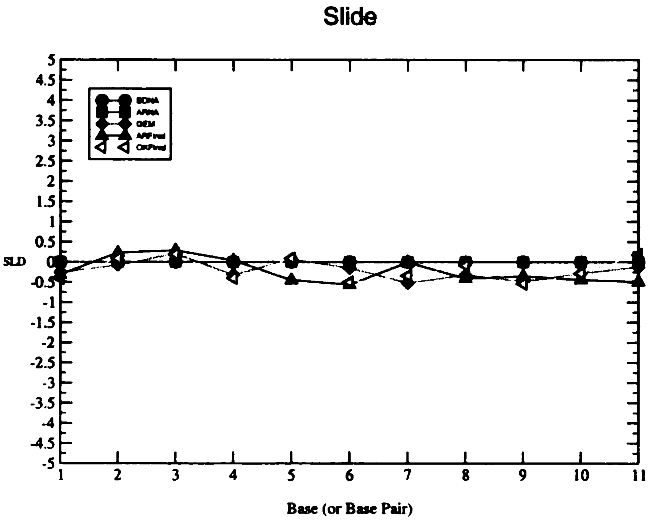


Figure 8.26: Dials plot of SLD parameter

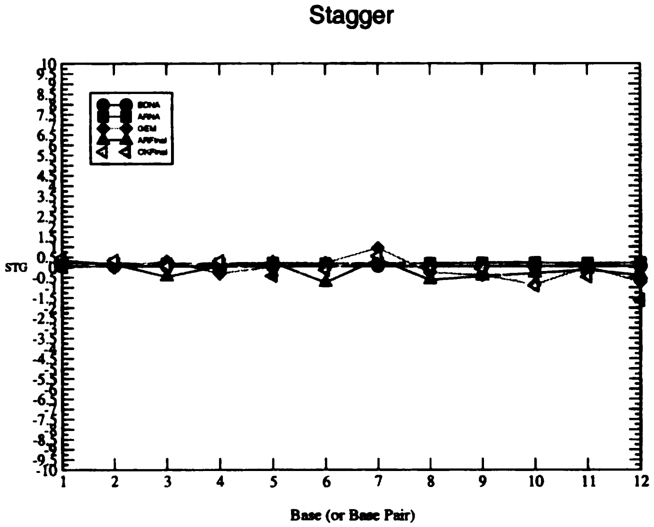


Figure 8.27: Dials plot of STG parameter

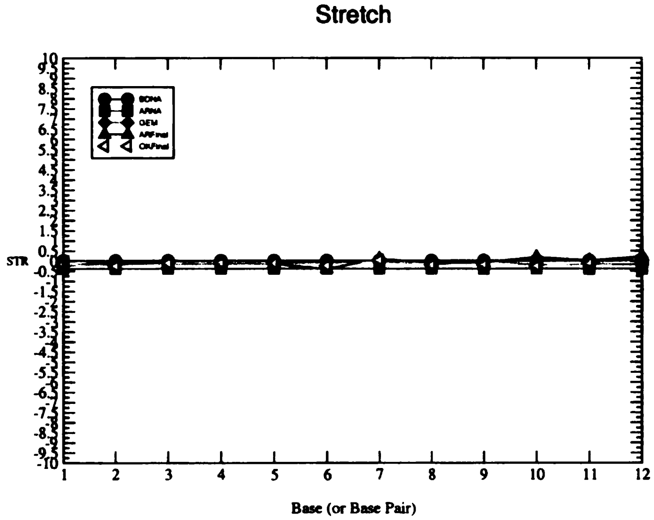


Figure 8.28: Dials plot of STR parameter

Tip

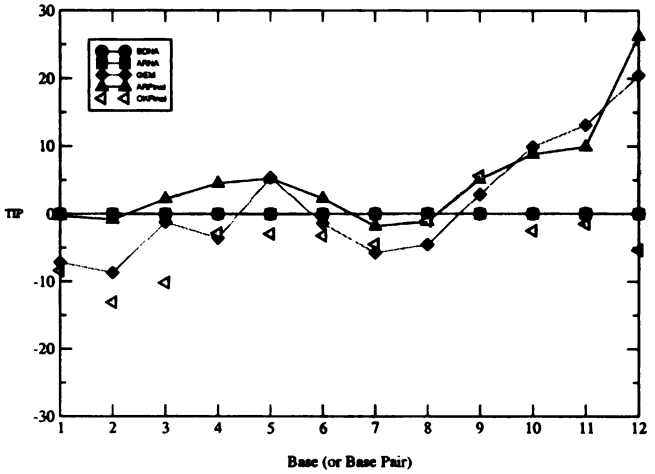


Figure 8.29: Dials plot of TIP parameter

Tilt

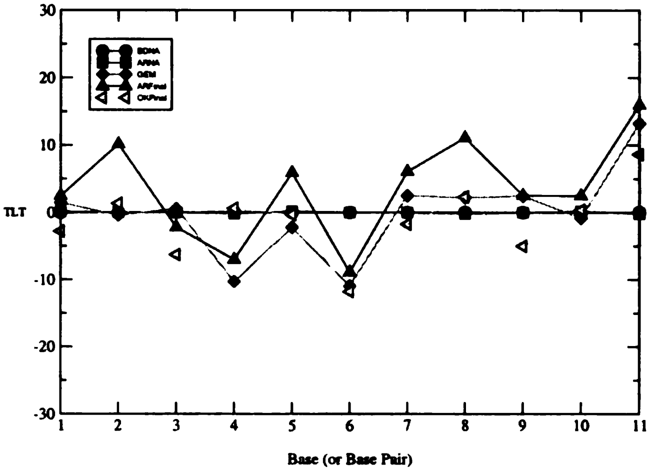


Figure 8.30: Dials plot of TLT parameter

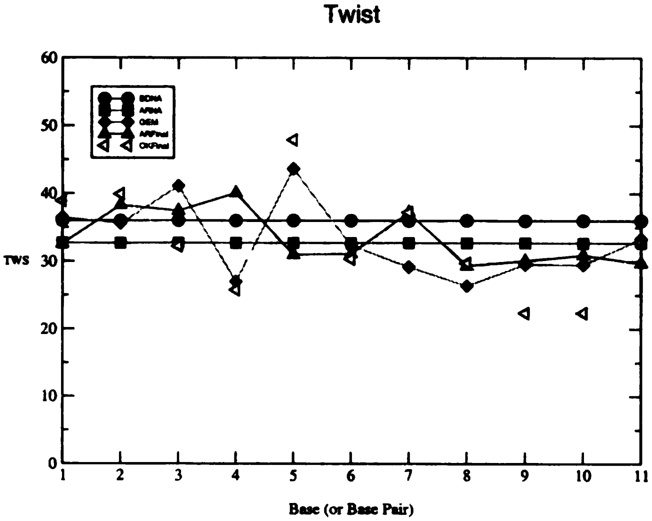


Figure 8.31: Dials plot of TWS parameter

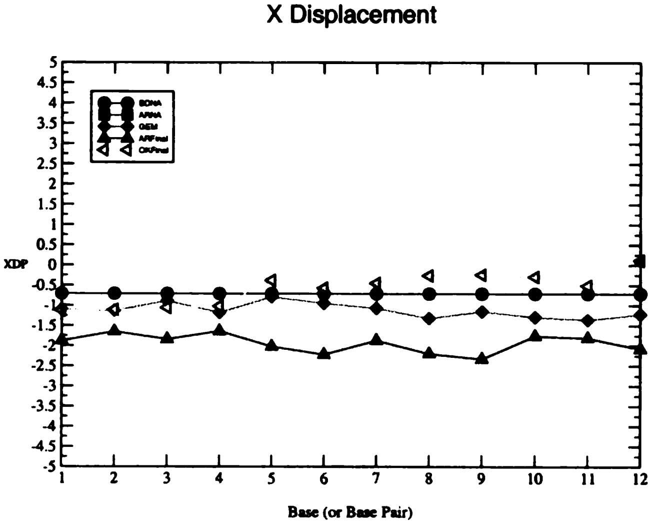


Figure 8.32: Dials plot of XDP parameter

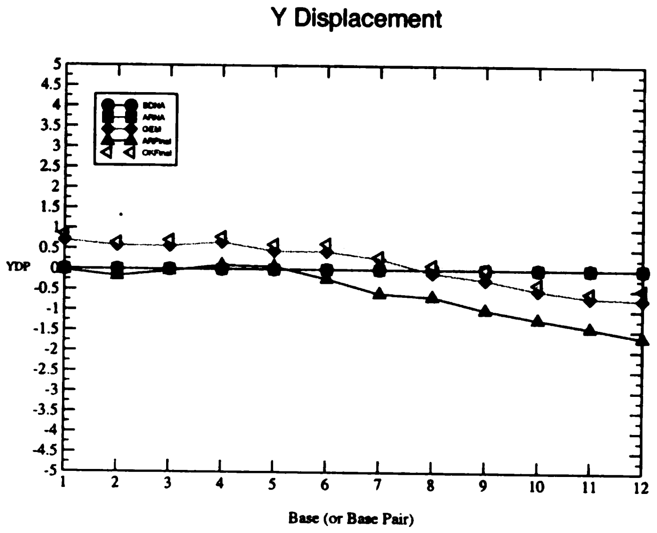


Figure 8.33: Dials plot of YDP parameter

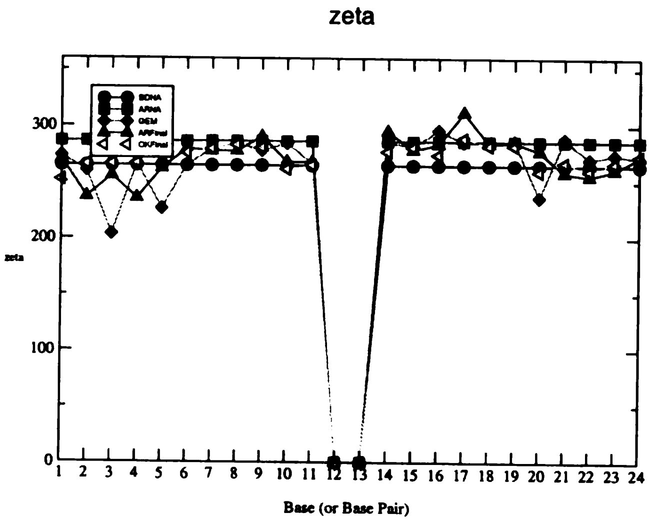


Figure 8.34: Dials plot of zeta parameter

8.2 Supplementary Figures for Free MD

Figure 8.36 to Figure 8.50 plot the histogram of Dials and Windows parameters for the RNA duplex simulations. Same definitions as previous, except histogram is calculated only for time=9ns to time=10ns. Strand one is filled gray, strand two is unfilled.

Figure 8.51 to Figure 8.54 plot the time course and histogram of AMBER energies (distance-dependent dielectric, no explicit solvent, no distance cutoff for nonbonded electrostatic and van der Waals). Same definitions as previous. Histograms are calculated for the entire simulation.

OKAstart_parm96, parameter pucker

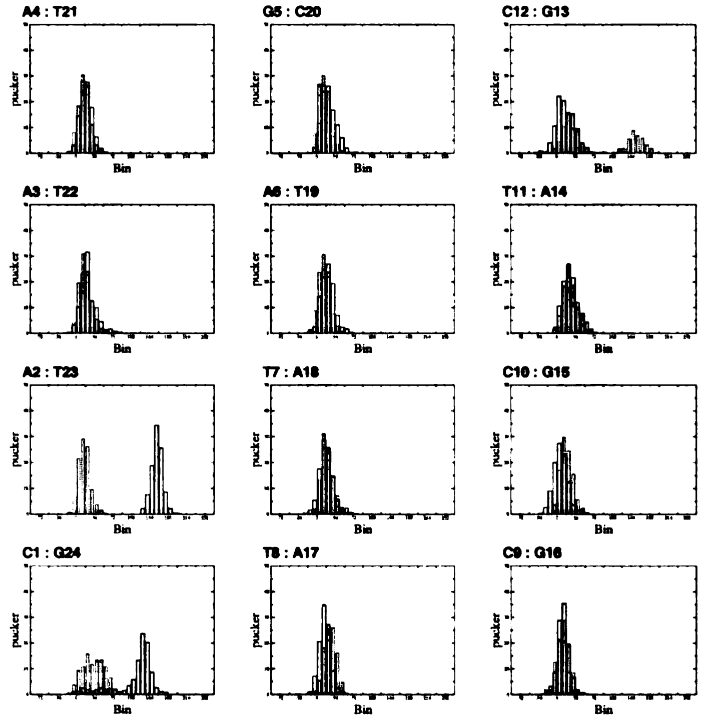


Figure 8.35: Histogram of pucker in Okazaki fragment free MD simulation (OKAstart-parm96 from time=9ns to time=10ns

OKAstart_parm96, parameter chi

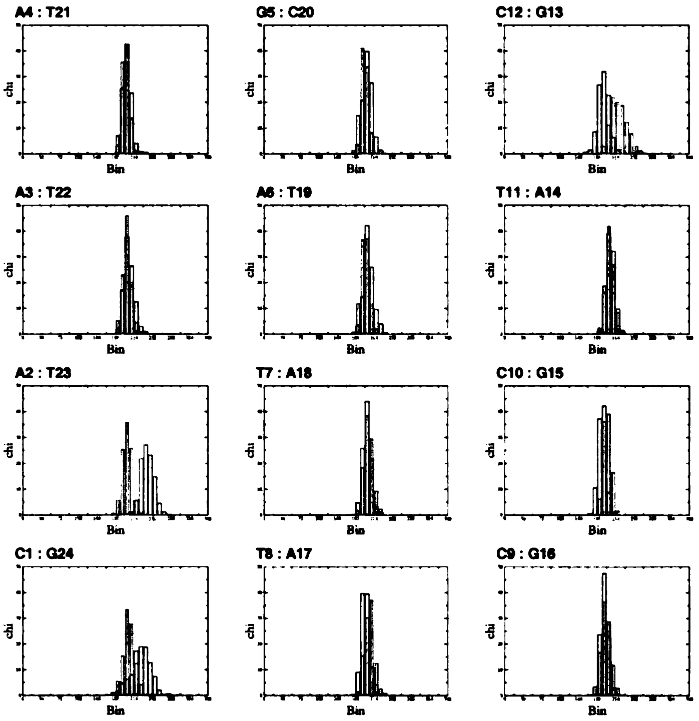


Figure 8.36: Histogram of chi in Okazaki fragment free MD simulation (OKAstart-parm96 from time=9ns to time=10ns

OKAstart_parm96, parameter INC

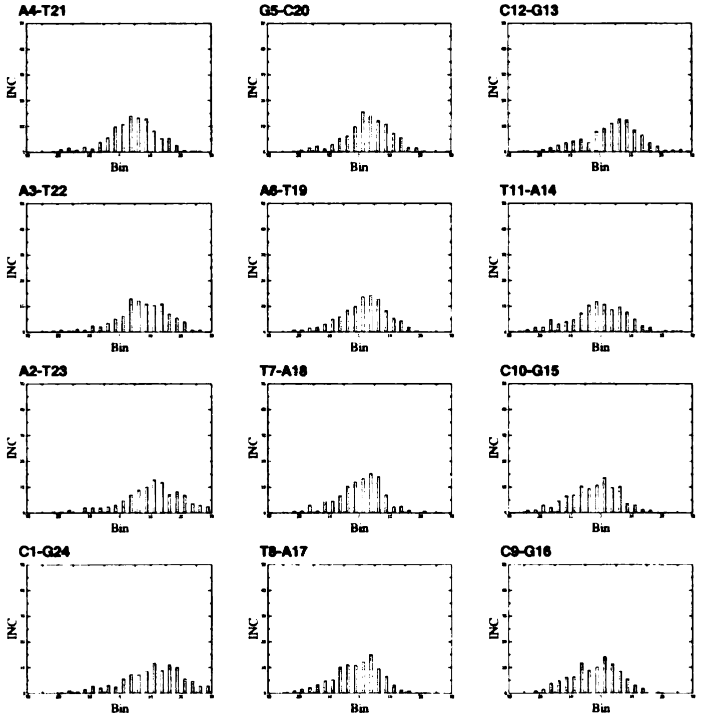


Figure 8.37: Histogram of INC in Okazaki fragment free MD simulation (OKAstart-parm96 from time=9ns to time=10ns

OKAstart_parm96, parameter XDP

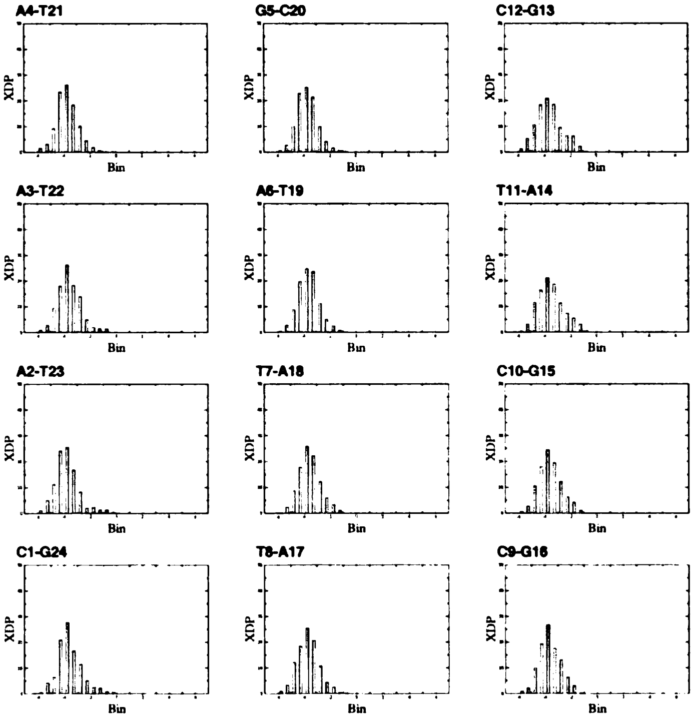


Figure 8.38: Histogram of XDP in Okazaki fragment free MD simulation (OKAstart-parm96 from time=9ns to time=10ns

OKAstart_parm99, parameter pucker

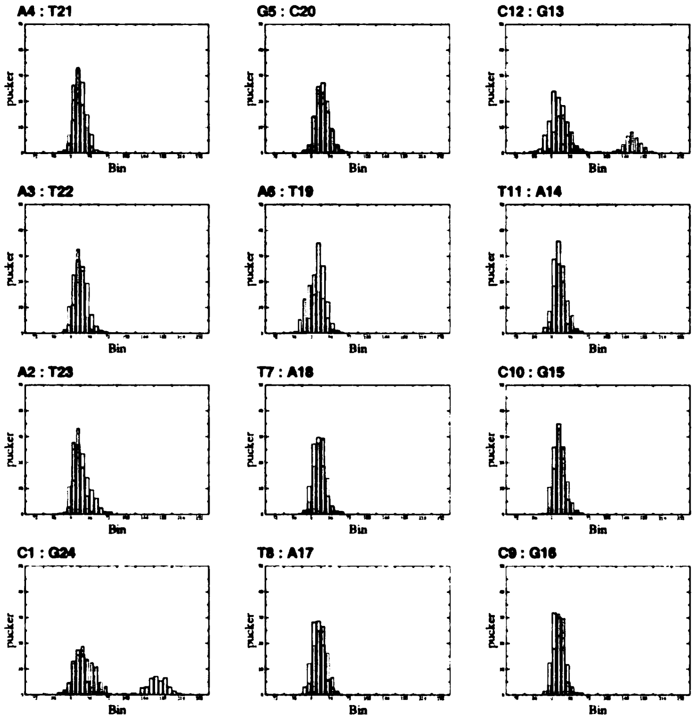


Figure 8.39: Histogram of pucker in Okazaki fragment free MD simulation (OKAstart-parm99 from time=9ns to time=10ns

OKAstart_parm99, parameter chi

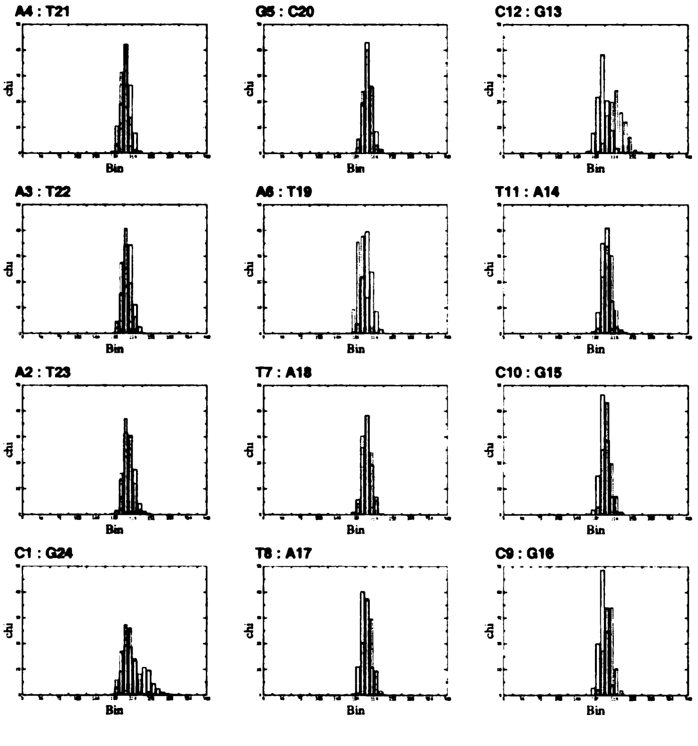


Figure 8.40: Histogram of chi in Okazaki fragment free MD simulation (OKAstart-parm99 from time=9ns to time=10ns

OKAstart_parm99, parameter INC

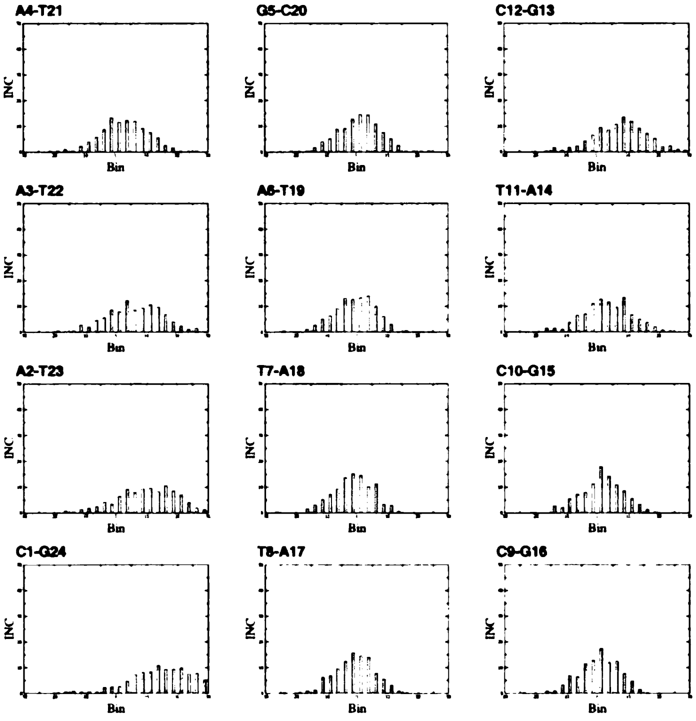


Figure 8.41: Histogram of INC in Okazaki fragment free MD simulation (OKAstart-parm99 from time=9ns to time=10ns

OKAstart_parm99, parameter XDP

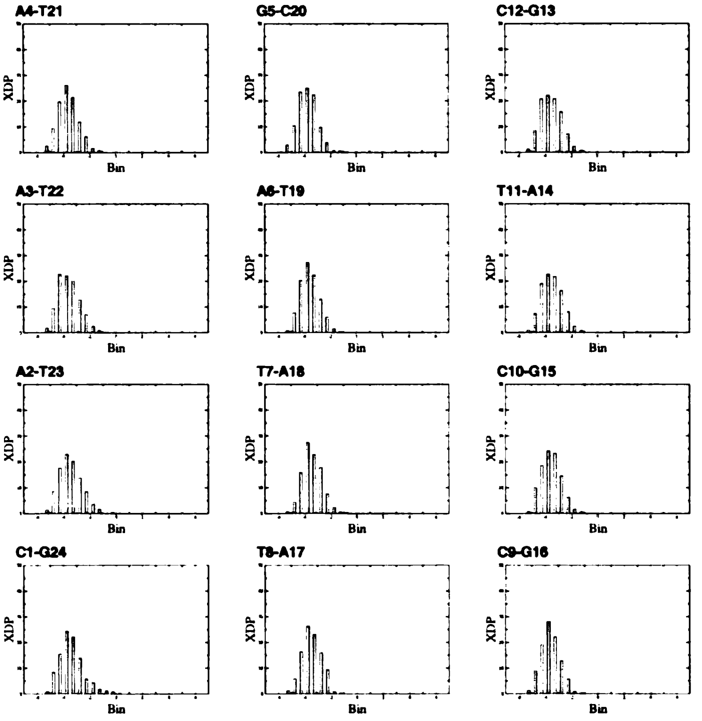


Figure 8.42: Histogram of XDP in Okazaki fragment free MD simulation (OKAstart-parm99 from time=9ns to time=10ns

OKBstart_parm96, parameter pucker

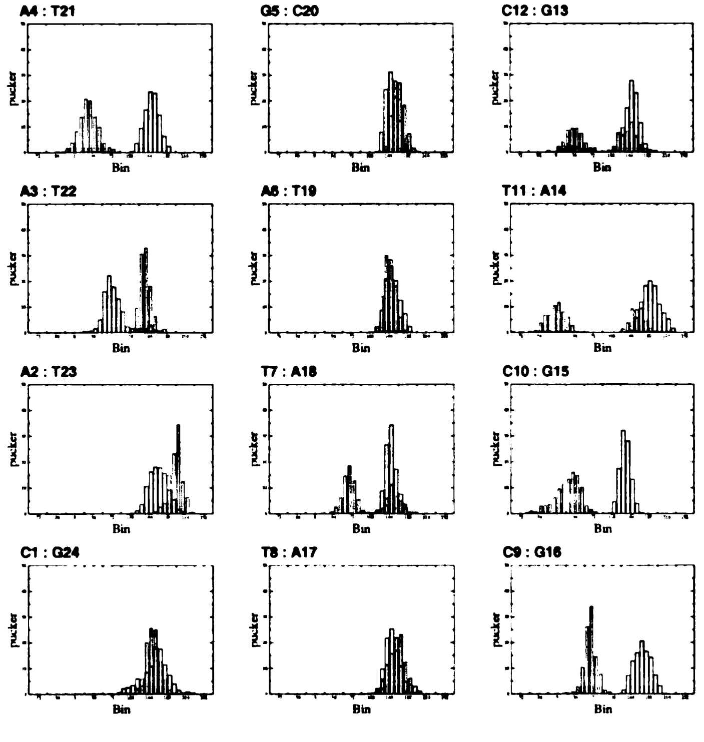


Figure 8.43: Histogram of pucker in Okazaki fragment free MD simulation (OKBstart-parm96 from time=9ns to time=10ns

OKBstart_parm96, parameter chi

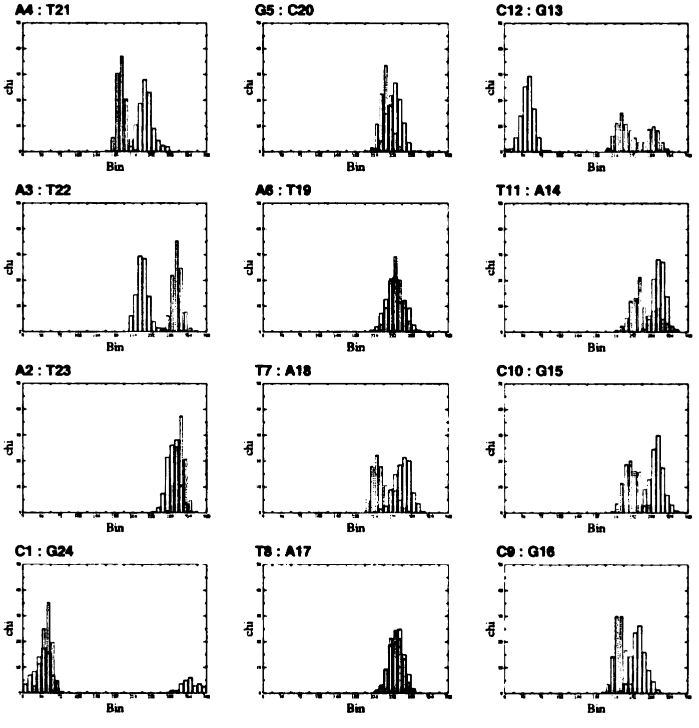


Figure 8.44: Histogram of chi in Okazaki fragment free MD simulation (OKBstart-parm96 from time=9ns to time=10ns

OKBstart_parm96, parameter INC

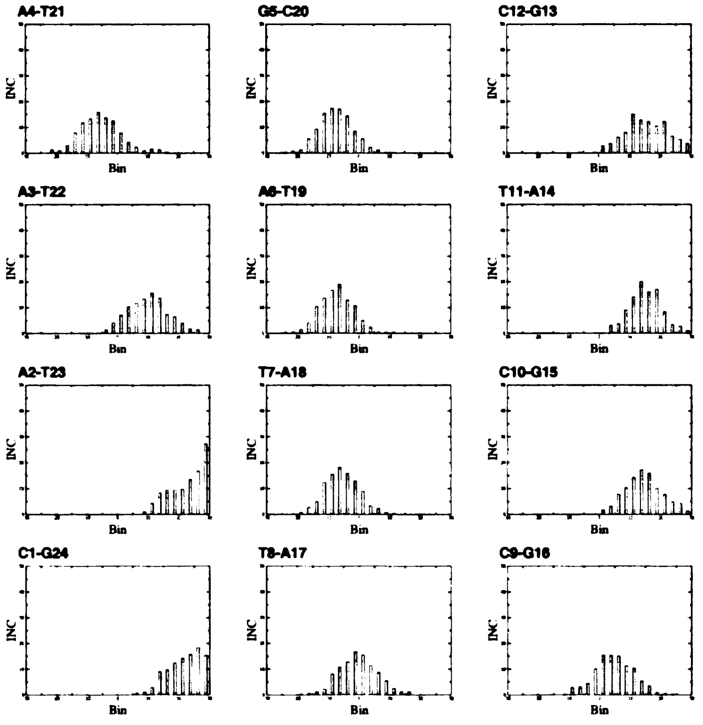


Figure 8.45: Histogram of INC in Okazaki fragment free MD simulation (OKBstart-parm96 from time=9ns to time=10ns

OKBstart_parm96, parameter XDP

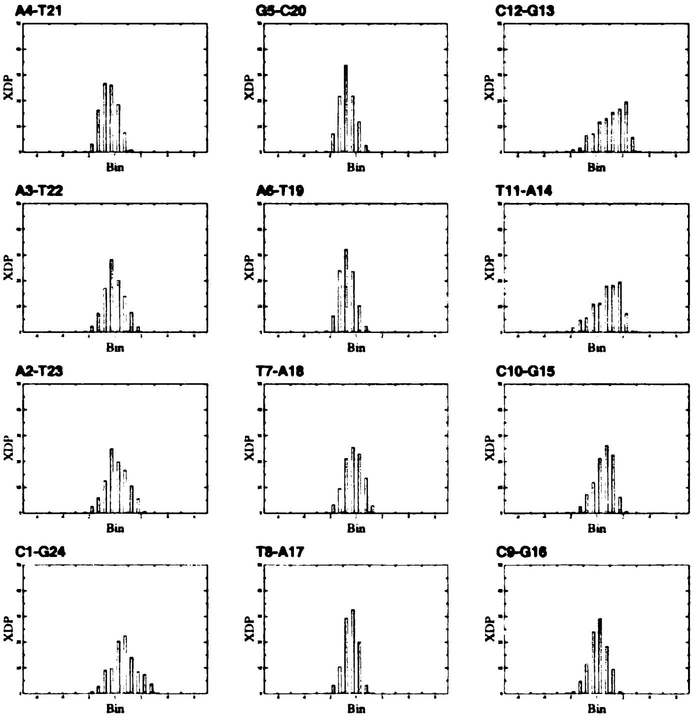


Figure 8.46: Histogram of XDP in Okazaki fragment free MD simulation (OKBstart-parm96 from time=9ns to time=10ns

OKBstart_parm99, parameter pucker

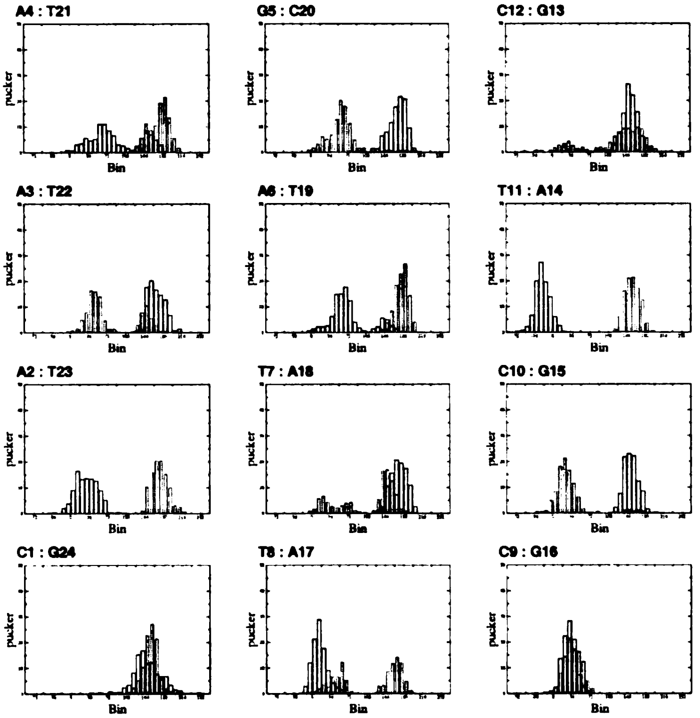


Figure 8.47: Histogram of pucker in Okazaki fragment free MD simulation (OKBstart-parm99 from time=9ns to time=10ns

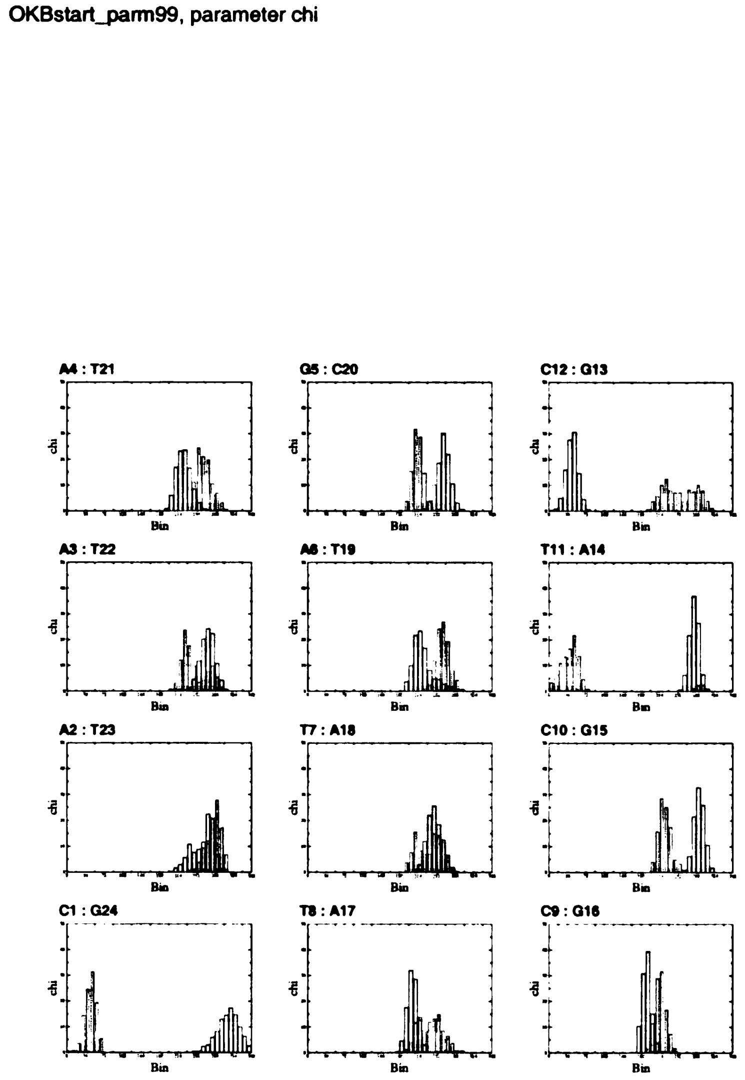


Figure 8.48: Histogram of chi in Okazaki fragment free MD simulation (OKBstart-parm99 from time=9ns to time=10ns

OKBstart_parm99, parameter INC

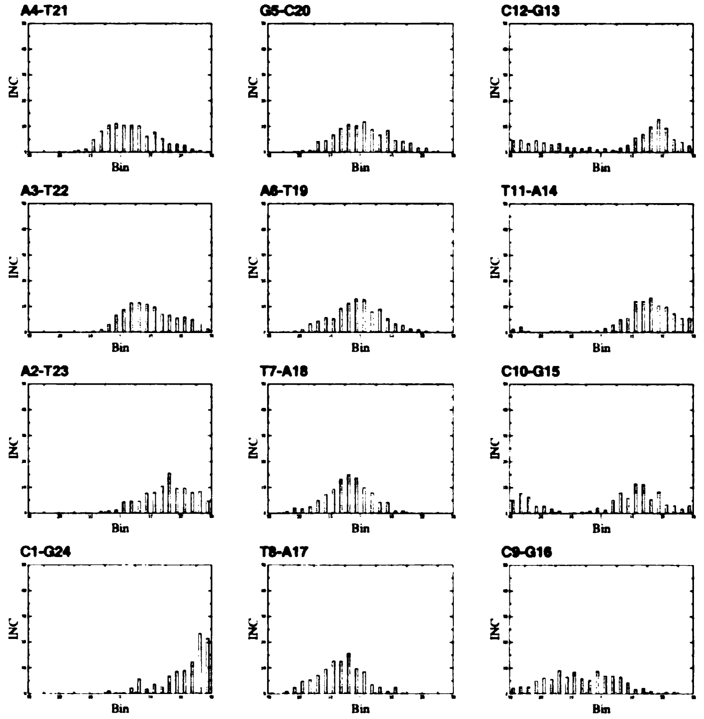


Figure 8.49: Histogram of INC in Okazaki fragment free MD simulation (OKBstart-parm99 from time=9ns to time=10ns

OKBstart_parm99, parameter XDP

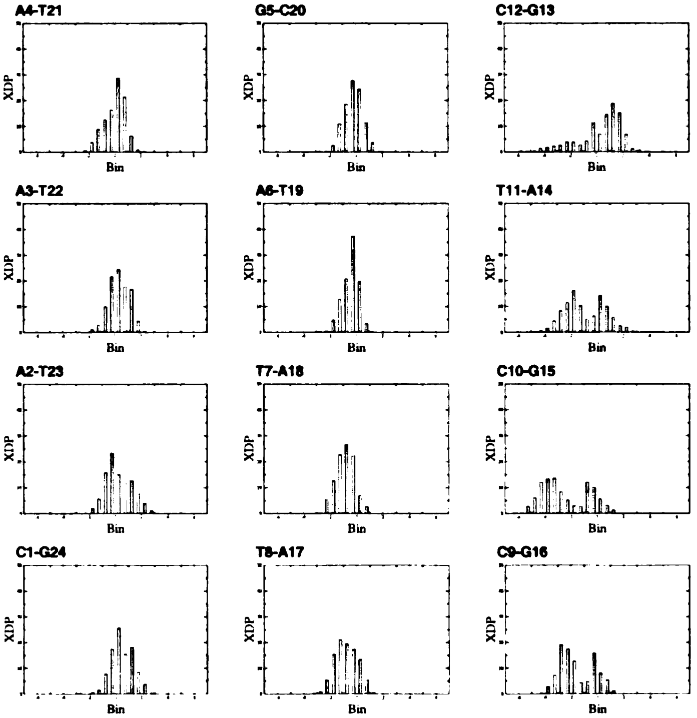


Figure 8.50: Histogram of XDP in Okazaki fragment free MD simulation (OKBstart-parm99 from time=9ns to time=10ns

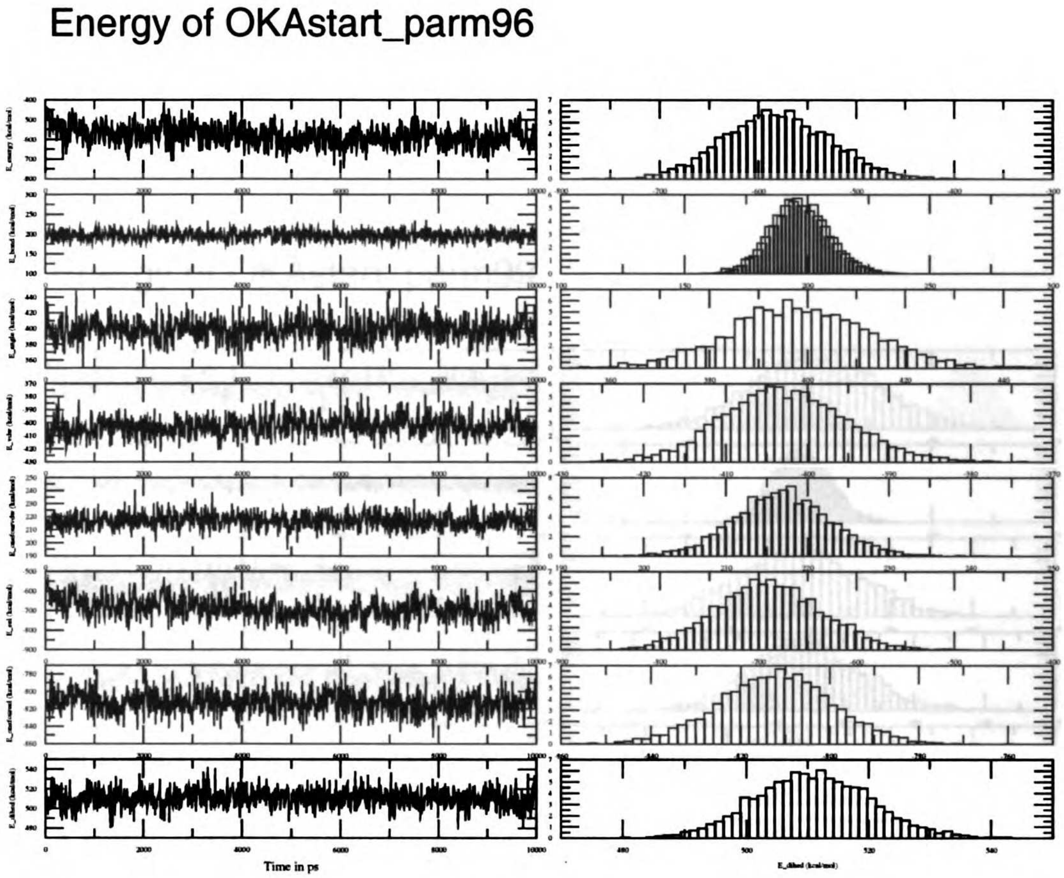


Figure 8.51: Time course of energies for RNA OKAstart-parm96

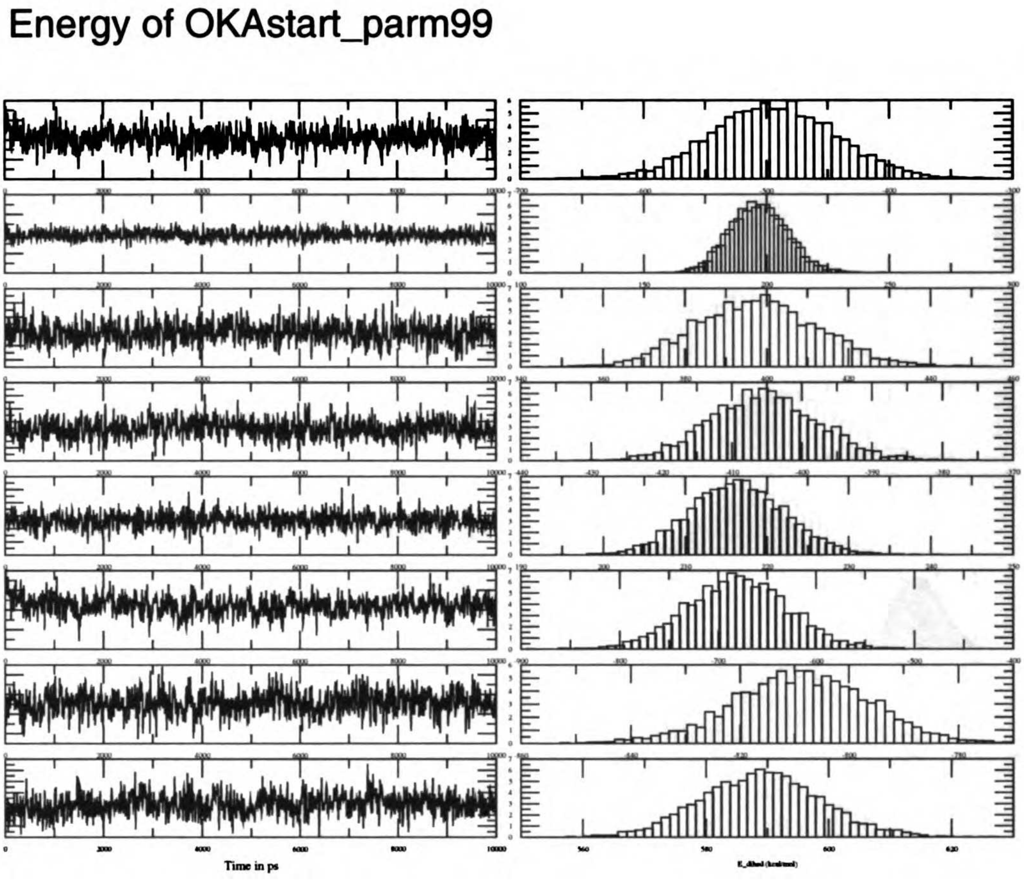


Figure 8.52: Time course of energies for RNA OKAstart-parm99

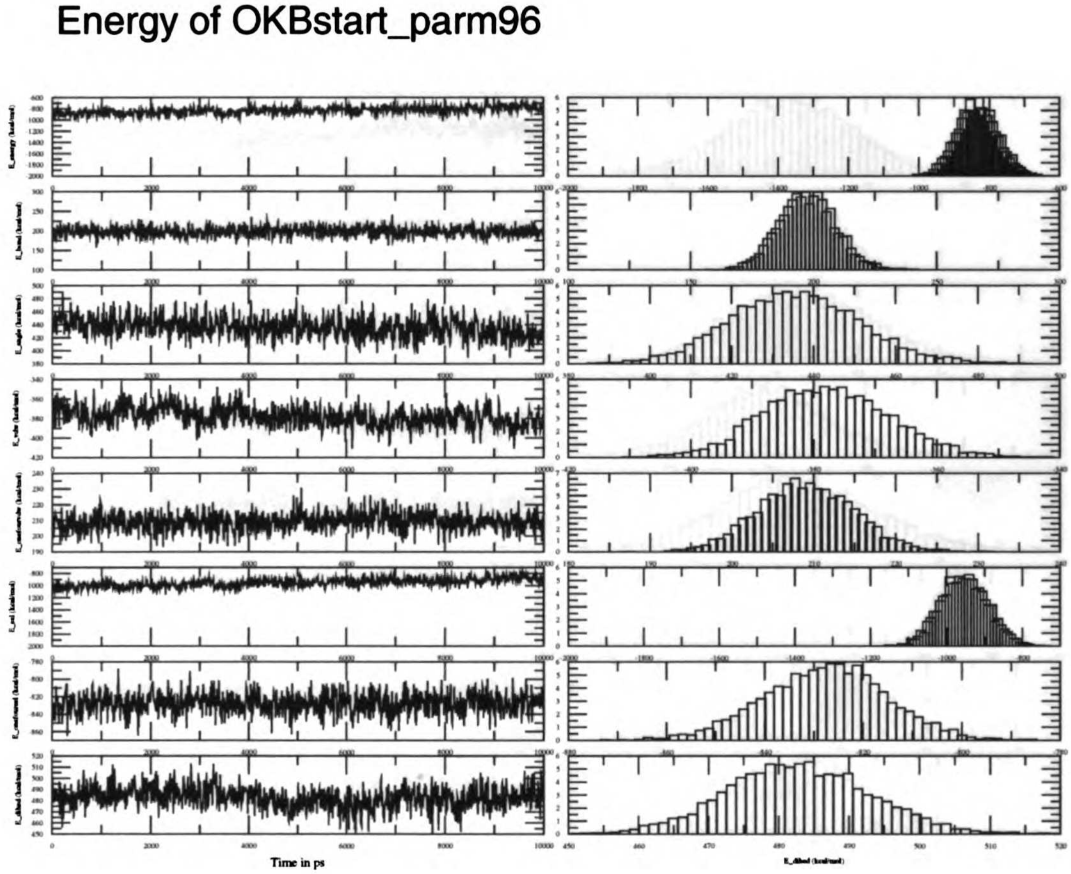


Figure 8.53: Time course of energies for RNA OKBstart-parm96

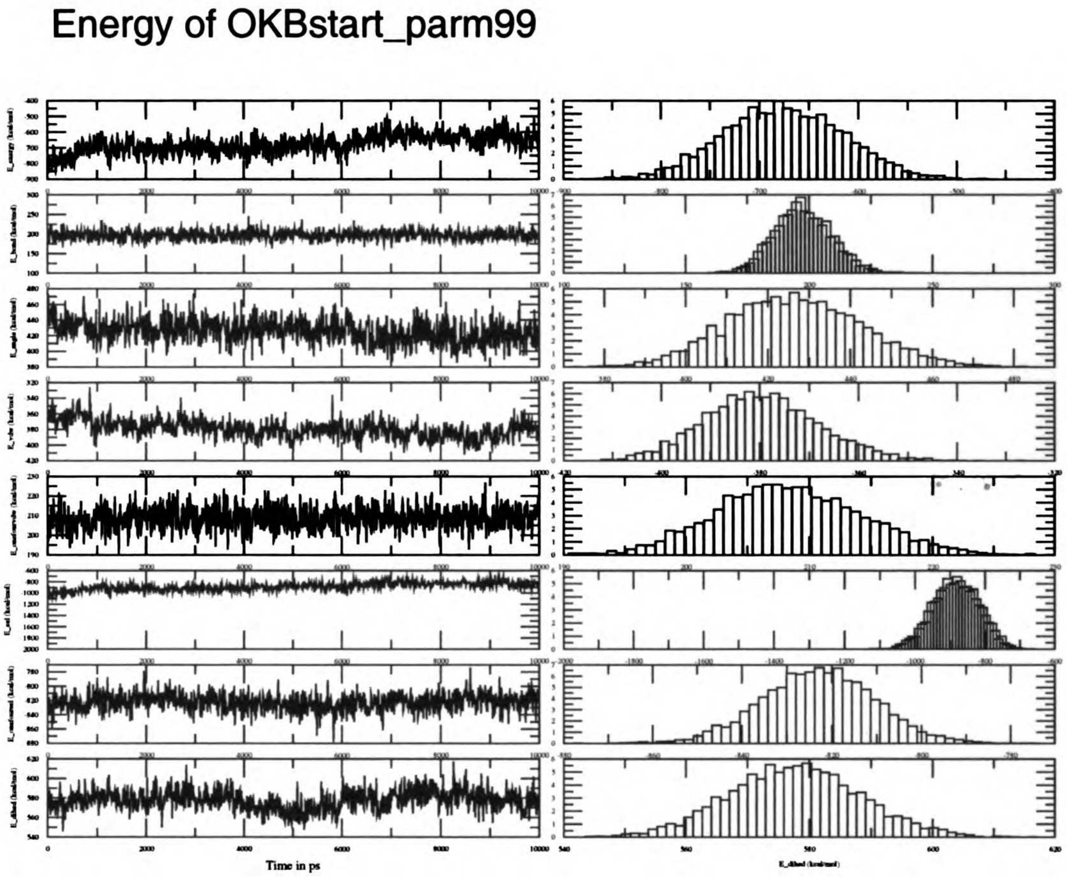


Figure 8.54: Time course of energies for RNA OKBstart-parm99