Title
Realistic simulation of nucleic acids in solution

Author
Cheatham, Thomas Edward, III

Publication Date
1997

Peer reviewed|Thesis/dissertation

Realistic simulation of nucleic acids in solution

by

Thomas Edward Cheatham, III

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmaceutical Chemistry

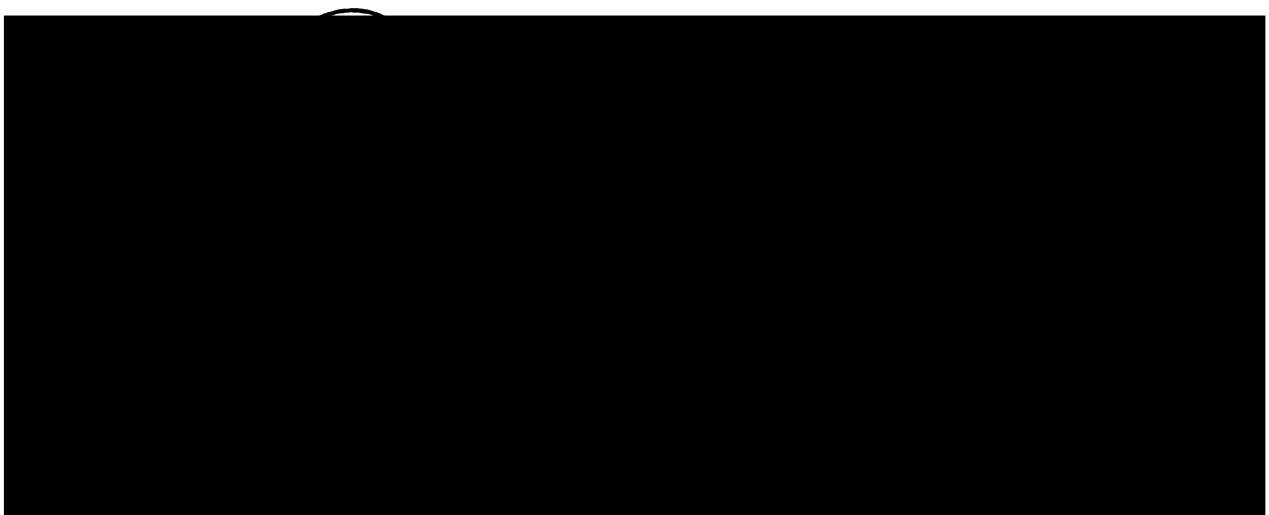
in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA

San Francisco



Date

University Librarian

Degree Conferred: . . . . . .

dedication

...to my parents and family for support all these years.

preface

One would think that writing the acknowledgement should be the easiest part of the thesis, but in fact, it is difficult since during six plus years of graduate school there are a large number of people who influence your thoughts, give support and encouragement, and help you hopefully grow into a better scientist and person. With this in mind, I cannot possibly list everyone who helped, but instead attempt to thank the people who come to mind as I close this large chapter in my life.

A large part of this work would not have been possible without the contributions and advice from outside collaborators; they include Roberto Gomperts and Michael Schlenkrich of SGI for shared memory parallelizing the AMBER code, Mark Berger and David Zirl for supporting the AMBER development through the loan of SGI hardware which sat on my desk, J. Kottalam of Cray Research for optimization and access to the CRI computers, and various people at the Pittsburgh Supercomputing Center, including David Deerfield, for their interest and help in parallelizing and optimizing AMBER. A particularly large thanks goes to Tom Darden from the National Institute of Environmental Health Sciences for giving me access to the particle mesh Ewald (PME) code, for inspiring and humorous conversations, for helping to teach me about PME, and for actively collaborating and being interested in my work. An additional partner in the PME collaboration who also deserves special thanks is Michael Crowley from PSC who parallelized the PME code for the T3D and later the T3E. My work would not have been possible without the help of these two individuals. Ken Merz and his talented graduate student Jim Vincent also deserve special mention for making an MPI version of the AMBER code available for general release; this code served as the foundation for the development and optimization of all the message passing code within AMBER 4.1. I

would also like to thank Bernie Brooks for letting me use the resources available in his lab to finish up this thesis.

At and around UCSF, I would like to thanks the AMBER developers for conspiring to let me into their cabal; in particular Dave Case for his wisdom and optimization/parallelization of the code, Dave Pearlman for good insight, quick and good natured response to my questions, excellent humor and group gossip; Bill Ross for helping me learn the obfuscations of AMBER, for maintaining the WWW page, for helping me to learn how to better handle code integration issues, and for the taking on the unglorious job of handling the day to day bug fixing, *etc*; and Jim Caldwell for his sagely advice about nuances in the code and group history. Dave Ferguson also deserves mention for making me realize that the Kollman group was the group I should join, despite that tainted wine and exploding champagne incident... I would also like to thank Jennifer Miller, Randy Radmer, Carlos Simmerling and Yax Sun for helpful discussions. A particularly strong acknowledgement goes to Dennis Benjamin, my classmate, not only for drinking fine beer, but for helping to push me to a new level scientifically. I would also like to thank all the faculty and staff who have befriended me over the years and made graduate school that much more pleasant; in particular Willa Crowell for just being so darn nice, helping with paper formatting and UCSF bureaucracy, for her work over the years planning the annual departmental retreat at Asilomar and helping me maintain a rather large budget for "entertainment" at that event. In addition I would like to thank Tack Kuntz for his insight and encouragement. Outside of "science", I would like to also thank my family and friends for their encouragement, particularly my good friend Hayden Foell for being there to watch me evolve over the last decade. Thanks also to John Colt for befriending Carol and I and the many good times we've shared.

In terms of funding and resources, I would like to acknowledge the NIH Biotechnology Training Grant (GM08388) and Professor Tony Hunt, the UCSF

Chancellors Fellowship, the UCSF Computer Graphics Laboratory (RR-1081) and Dr. Tom Ferrin, the Metacenter (MCA93S017P) and PSC for significant computational resources, Silicon Graphics Incorporated for the loan of machines to our group and me over the years and Cray Research for access to the Eagan site for testing and development of AMBER.

Some of the text in this thesis is based on published material, therefore I would like to acknowledge the references:

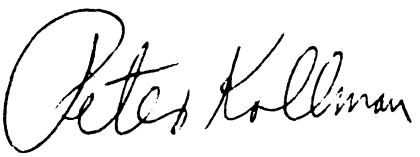
- Cheatham, T. E., III, Miller, J. L., Fox, T., Darden, T. A. & Kollman, P. A. (1995). Molecular dynamics simulations on solvated biomolecular systems - The particle mesh Ewald method leads to stable trajectories of DNA, RNA and proteins. *J. Amer. Chem. Soc.* **117**(14), 4193-4194.
- Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E., III, Debolt, S., Ferguson, D., Seibel, G. & Kollman, P. (1995). AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structure and energetic properties of molecules. *Comp. Phys. Comm.* **91**(1-3), 1-41.
- Cheatham, T. E., III & Kollman, P. A. (1996). Observation of the A-DNA to B-DNA transition during unrestrained molecular dynamics in aqueous solution. *J. Mol. Biol.* **259**(3), 434-44.
- Cheatham, T. E., III & Kollman, P. A. (1997). Molecular dynamics simulations can reasonably represent the structural differences in DNA:DNA, RNA:RNA and DNA:RNA hybrid duplexes. *J. Amer. Chem. Soc.* (1997) [in press].

...and finally, a special thanks to Peter Kollman for his constant encouragement, enthusiastic interest in my studies, access to computers and disks, and overall for helping me grow into an independent investigator; and to Carol Lim, my wife, for suddenly sitting next to me during our first year of graduate studies and her constant love and support ever since. Despite the hardships, this six years has been one of the most fullfilling times in my life.

abstract

Realistic simulation of nucleic acids in solution

Thomas Edward Cheatham, III



Advances in computer power, empirical force field representations, and the development of more reasonable means to handle the long ranged electrostatic interactions now allow routine "stable" nanosecond length unrestrained molecular dynamics simulations of nucleic acids in explicit water. This has allowed the observation of spontaneous A-DNA to B-DNA transitions in water, representation of the structure of A-RNA and DNA-RNA hybrids, and the stabilization of A-DNA in mixed water/ethanol solutions. The basics and history of the methods are discussed along with significantly more in-depth versions of published papers investigating the simulation of nucleic acids in various environments.

table of contents

list of tables

list of figures and illustrations

introduction

Not until the pioneering fiber diffraction studies of Franklin & Gosling (Franklin & Gosling, 1953), the proposition of the double helix by Watson & Crick (Watson & Crick, 1953), and subsequent crystallization experiments (Dickerson *et al.*, 1982) did a glimpse into the structure of nucleic acids emerge. Since this time, advances in X-ray crystallographic and NMR spectroscopic methods, coupled with the development of recombinant DNA technology, has led to an ever expanding body of information about nucleic acid structure. In this thesis, a series of papers and discussion are presented which suggest that molecular dynamics simulations with an empirical force field can serve as a complement to experimental studies and give insight into nucleic acid structure and dynamics. This is in large part thanks to increases in computer power, advances in force field representation and the application of methods for properly treating the long ranged electrostatic interactions. When I first came to UCSF and the Department of Pharmaceutical Chemistry in 1990, the simulation of nucleic acids was limited to *in vacuo* calculations or short simulations in explicit solvent. Although the first simulations of DNA in explicit water were reported in 1985 (Siebel *et al.*, 1985), even five years later, reasonable simulations with explicit solvent and full periodic boundary conditions were extremely time consuming, requiring run time on supercomputers such as the Cray Y-MP, and were generally limited to less than 200 picoseconds. Now, as I finish up this work in the winter of 1996-1997, multi-nanosecond simulations are routine and require only days on parallel computers such as the Cray T3E.

The work presented in this thesis took place mostly between 1994 and 1996 in the laboratory of Dr. Peter A. Kollman. The goal of this work, in the early stages, was to enable the simulation of small nucleic acids in explicit solvent on a nanosecond time scale and to test the latest force field for nucleic acids (Cornell *et al.*, 1995). Specifically, 10-mer

(and a few longer and shorter) duplexes of varying sequence were extensively studied. These solvated systems were studied since I believed that it was important to first test the methods and discover the limitations on these small representative systems prior to performing larger scale simulations on protein-nucleic acid complexes or higher order nucleic acid structures. Moreover, these small systems still offer a rich diversity in nucleic acid structure, such as the effect of sequence and the environment on the structure. Along with evaluation of the methods and force field, the hope was to demonstrate that molecular dynamics simulations can lead to useful insight into nucleic acid structure and dynamics. The larger challenge is to accurately represent this rich diversity in sequence dependent structure and the effect of the environment on structure; this is something which the methods are just now beginning to realize. The overall conclusion from all the data presented herein is that molecular dynamics simulations can, in some cases, give useful insight and a surprisingly good representation of nucleic acid structure. Moreover, in some cases the methods can reasonably represent the effect of the environment on the structure of nucleic acids. In other cases, the omnipresent "conformational sampling" problem becomes an issue and definitive conclusions become a bit more specious.

In this introduction, a discussion about nucleic acid structure and its relationship to function is presented along with some discussion of the methods and history of the atomic simulation of nucleic acids. This is followed by a summary of my involvement in the development and integration of enabling technology for studying nucleic acid structure and dynamics. The calculations presented herein represent large scale simulations for the time period; hence the use of high powered workstations and parallel supercomputers was necessitated. To support this goal, the molecular dynamics simulations code AMBER needed fairly general parallelism and overall optimization; this involved collaborations with scientists at Silicon Graphics, Inc., Pennsylvania State University, the Pittsburgh Supercomputing Center (PSC), Cray Research and other AMBER developers. An

additional enabling technology was the application of methods to properly treat the long ranged electrostatic interactions which involved a collaboration with Tom Darden at the National Institute of Environmental Health Sciences (NIEHS) and Michael Crowley (PSC). After the introduction, each chapter except for the final chapter represents a paper that is either published, in press, or submitted for review. A brief preface is presented before each paper which describes in a bit more detail the rational (at that time) for pursuing the research. The final chapter presents issues and problems encountered with the methods and is concluded with a discussion of molecular dynamics trajectory analysis and the software I developed for this purpose.

nucleic acids

structure and nomenclature

The nomenclature, chemical composition, and structure of nucleic acids has been reviewed extensively in a variety of biochemistry texts (for example see Stryer's Biochemistry (Stryer, 1988)), and more extensively in the "bible" of nucleic acid structure, Saenger's Principles of Nucleic Acid Structure (Saenger, 1984) hence will not be extensively reviewed herein. In general within the text nucleic acids are referred to by their principal bases; adenine (ADE, A), thymine (THY, T), guanine (GUA, G), cytosine (CYT, C) and uracil (URA, U) and the sequence is always written from the 5' to 3' end. Shown in Figure 1 are the principal bases and connectivity (in the 5' to 3' direction) of DNA. In the upper right of Figure 1 is shown the deoxyribose sugar. Changes in the conformation of the sugar are a large determinant of nucleic acid structure. These changes in the conformation of the sugar are often characterized by the sugar pucker pseudorotation phase and sugar pucker amplitude (Altona & Sundaralingam, 1972). The puckering is also referred to by common names, with puckers above the plane of the ring towards the 5' atoms called *endo* and puckers below the plane *exo*. Characteristic of B-DNA is a C2'-

endo sugar pucker (where the C2' atom is above the plane) which has a sugar pucker pseudorotation value in the ~144° to 180° range. Note that throughout the text "sugar pucker" and "pucker" are used synonymously with the sugar pucker pseudorotation value specified in degrees or by the common names (*i.e.* C3'-*endo*, C2'-*endo*, *etc.*).

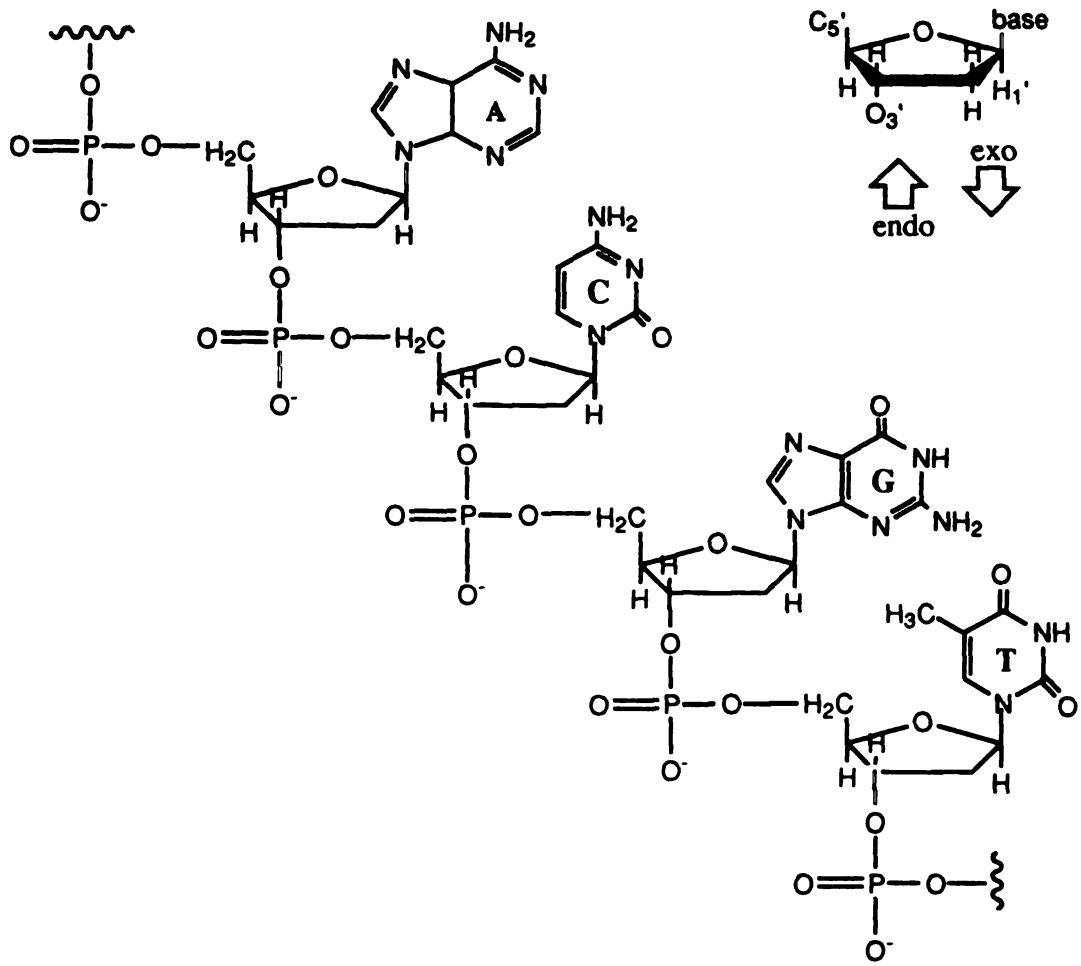


Figure 1: Atom names and connectivity in DNA

The double stranded structures are, in general, created using standard nucleotides, Watson-Crick base pairing and structures based on canonical geometries or model structures from X-ray crystallography or NMR methods. In all the simulations, except where otherwise

mentioned, the terminal nucleosides lack a terminal phosphate group. The model building is discussed in more detail in the methods section of each chapter and in more detail in Chapter 5. Shown in Figure 3 (at the end of this section) are stereoviews of canonical A, canonical B and the crystal (Prive *et al.*, 1991) structures of d[CCAACGTTGG]$_2$ which are studied extensively in this thesis. Helicoidal parameters used to describe the structure of the nucleic acid conform to the Cambridge convention (Dickerson, 1989); the calculation of these is discussed in more detail in the published discussion of "Curves" (Lavery & Sklenar, 1988; Lavery *et al.*, 1995) and "Dials and Windows" (Ravishanker *et al.*, 1989). Values for some of the helicoidal parameters which largely distinguish canonical A-DNA and canonical B-DNA, compared to the crystal structure of d[CCAACGTTGG]$_2$ (referred to by the pdb designation 5dnb) are presented in Table 1.

	A-DNA	B-DNA	5dnb
rise (Å)	2.56	3.38	3.43
x-disp (Å)	-5.43	-0.71	1.17
twist, Ω (°)	32.7	36.0	35.3
inclination (°)	19.1	-5.9	-2.3
pucker (°)	13 (C3'-*endo*)	192 (C2'-*endo*)	146

Table 1: Helicoidal values distinguishing A- and B-form geometries

In Table 1, the following trends are observed. The rise (along the helical axis) is lower in A-form than B-form structures, as is the base pair displacement from the helical axis (in the "x-direction", or x-disp). The helical twist in B-DNA represents 10 base pairs per

complete turn; A-DNA has a significantly lower twist. The inclination of the base pairs is significantly positive in A-DNA whereas it is slightly negative or near zero in B-DNA structures. A-DNA is characterized by C3'-*endo* sugar puckers and B-DNA by a mixture of puckers with C2'-*endo* puckers dominating. Note that throughout the text the words "conformation", "structure" and "geometry" are used synonymously and refer to the relative three dimensional atomic positions of all the atoms in the molecule. Also in the text the backbone angles are referred to by common names (α, β, *etc.*) defined below in Figure 2.

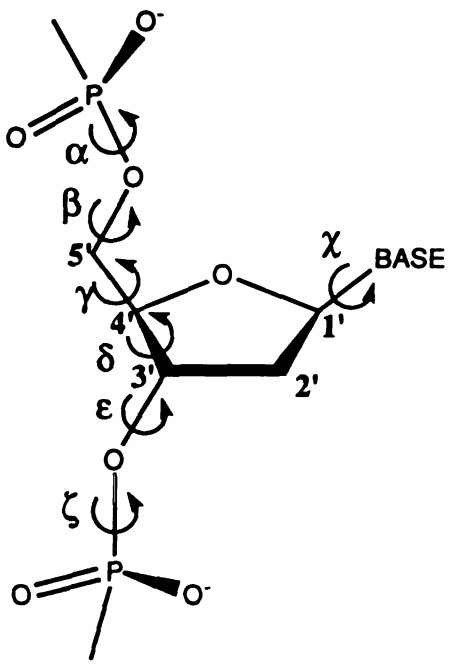


Figure 2: Nucleic acid backbone angles

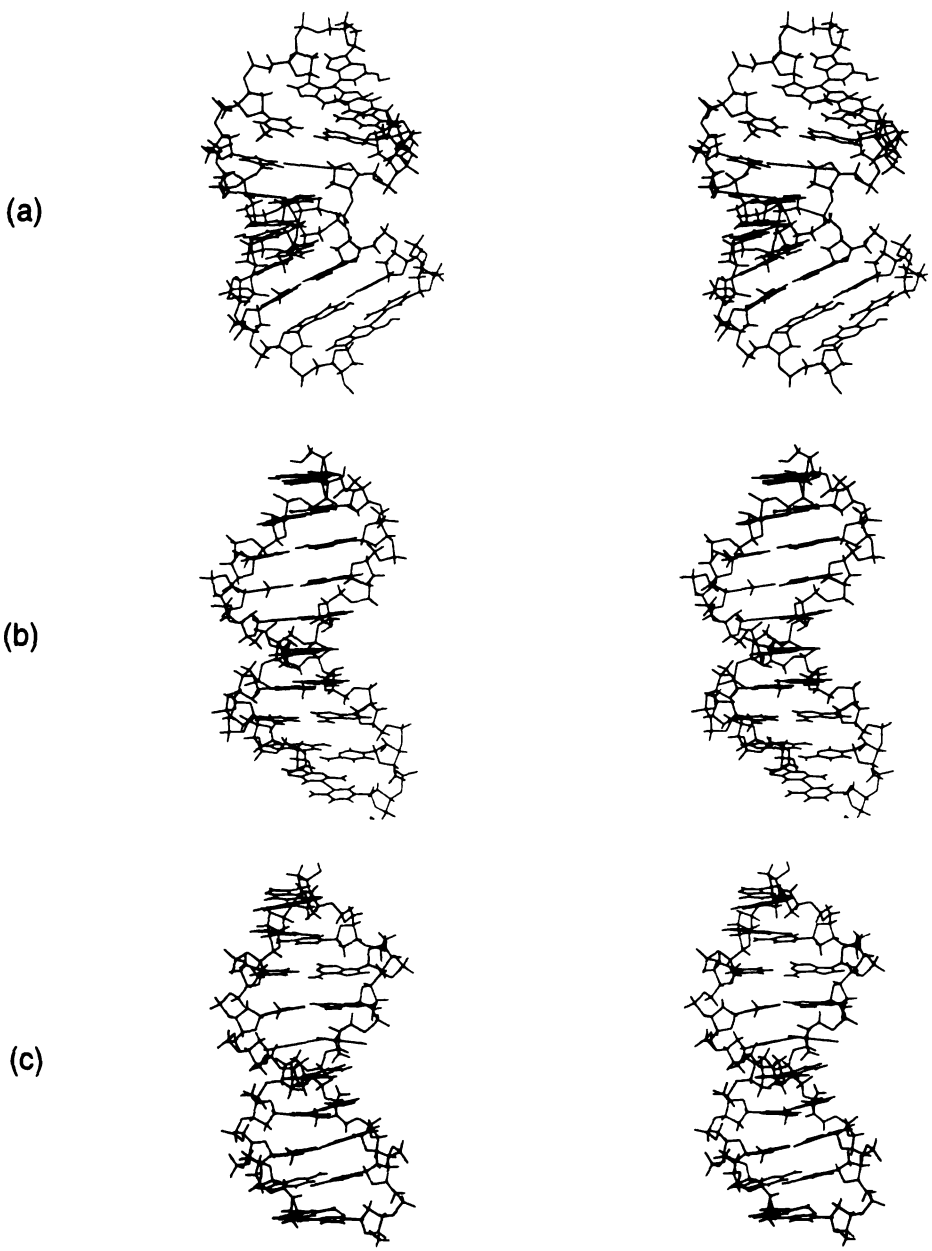


Figure 3: Stereo views of canonical A-DNA, canonical B-DNA and the crystal structure of d[CCAACGTTGG].

structure and function

Knowledge of the structure of nucleic acids gives insight to function. For example, differences in the structure between DNA:DNA, RNA:RNA and DNA:RNA hybrid duplexes help explain why the RNAse-H domain from the HIV-1 reverse transcriptase (Gotte *et al.*, 1995) preferentially degrades the RNA strand in DNA:RNA hybrids over RNA:RNA duplexes. Structural deformations of the nucleic acid also play a functional role. The widened minor groove characteristic of catabolite activating protein (CAP) induced bends (Schultz *et al.*, 1991) is the preferred site for DNase I and IN integrase activity (Muller & Varmus, 1994). For some enzymes, distortion of the DNA conformation is essential for catalysis (Steitz, 1990). Further examples include the unwinding of the DNA helix by zinc finger proteins (Shi & Berg, 1996) or the enhanced binding specificity of *Eco* R1 DNA methyltransferase to DNA due to disruption of base stacking (Allan & Reich, 1996).

In addition to structural deformations, structural *transitions* in nucleic acids also have an important function, such as the compaction of supercoiled DNA. The compaction occurs due to the induction of local segments of A-DNA or Z-DNA structure in an otherwise B-DNA coil (Levin-Zaidman *et al.*, 1996). Structural transitions to A-DNA or Z-DNA can be induced by the binding of poly-cationic ions such as spermine or neomycin (Robinson & Wang, 1996) or inhibited by other ligands, such as the minor groove binders netropsin and distamycin A (Burckhardt *et al.*, 1996). Another intriguing example of a structural transition which is functionally useful is the binding of small acid soluble spore proteins (SASPs) to DNA in gram positive bacteria which significantly increases the resistance of the bacteria to UV radiation damage (Mohr *et al.*, 1991). This occurs since the binding of the SASPs induces a transition from B-DNA to the more rigid and photoresistant A-DNA conformation. Although the inherent deformability of nucleic acid

structure is functionally important, clearly the flexibility of the nucleic acid also plays an important role as is suggested by the photoresistance of the "rigid" A-DNA structure. Changes in flexibility and alteration of the bending patterns are both suggested to play a role in both the activation and repression of RNA polymerase by the transcription factor MerR (although alteration of the bending patterns is more consistent with the data) (Ansari *et al.*, 1995). Differences in canonical structure, deformations of structure, transitions between structures, and the inherent flexibility of nucleic acids all play a functional role.

All of these structural transitions, deformations and changes in flexibility are possible since the conformation of a nucleic acid is strongly dependent on the environment. In fact, it is this strong dependence on the environment and inherent flexibility that makes structure determination of nucleic acids difficult. The more flexible the nucleic acid, the more difficult it is to obtain a high resolution structure. Although much has been learned about nucleic acid structure from X-ray crystallographic and NMR spectroscopic methods, both techniques have their specific limitations. Crystallography is limited by the need to obtain crystals which diffract to high resolution and the constraints imposed by crystal packing. Crystal packing can deform the structure as seen in the experiments where the same sequence crystallizes into different structures depending on the geometry of the crystal lattice (Dickerson *et al.*, 1987; Dickerson *et al.*, 1994; Shakked *et al.*, 1989) and the crystallization conditions. NMR methods, on the other hand, suffer from the difficulty in assigning all the proton resonance's and the fact that the information obtained is short ranged, representing less than 5 Å or three connected bonds in distance. Since the information derived is short ranged, this tends to question the reliability of NMR derived nucleic acid structures for analyzing longer ranged structural features, such as nucleic acid bending. Additionally, in some cases the refinement of NMR derived structures shows a strong dependence on the methods used; it has been demonstrated that refinement by simulated annealing with energy representations with and without explicit solvent converge

to different structures (Leijon *et al.*, 1995). Both X-ray crystallography and NMR methods have specific limitations which question their complete reliability in defining the effect of environment and sequence on the structure of nucleic acids. Clearly large advances in the understanding of nucleic acid structure have emerged, however it may be to early to determine the reliability of the statistical analyses of NMR derived structures (Ulyanov & James, 1995) or X-ray derived structures (Gorin *et al.*, 1995; Young *et al.*, 1995a). Therefore the development of theoretical methods which can accurately describe the structure and dynamics of nucleic acids is a clear advance. In the next section, an introduction to the simulation methods applied in this thesis is presented.

simulation methods

All of the simulations discussed herein were performed using AMBER (Pearlman *et al.*, 1995). AMBER-- "assisted model building with energy refinement"-- in common usage represents both a suite of programs for the simulation of biomolecules and a series of empirically derived molecular mechanical force field parameters (Cornell *et al.*, 1995; Weiner *et al.*, 1984). The force field is in general use in a variety of simulation programs and the tools provided and methods employed in AMBER are very similar to a number of other programs, including CHARMM (Brooks *et al.*, 1983), GROMOS (van Gunsteren & Berendsen, 1987), XPLOR, CEDAR, DISCOVER, SYBYL and others. AMBER and the Weiner *et al.* and Cornell *et al.* force fields are specifically tailored for the simulation of proteins and nucleic acids both *in vacuo* and in solution. An article by Pearlman *et al.* describes the development and history of AMBER (Pearlman *et al.*, 1995); this section elaborates somewhat on the discussion presented in that paper.

The potential energy function, force fields and molecular mechanics

The heart of all molecular mechanical force field methods is the potential energy representation; it is the accuracy of this representation which allows any connection to

physical reality. In AMBER, this is a rather simple and differentiable pairwise potential energy function (E_{AMBER}) shown in its simplest form in Figure 4.

$$E_{AMBER} = \sum_{bonds} k_r (r - r_{eq})^2 + \sum_{angles} k_\theta (\theta - \theta_{eq})^2$$
$$+ \sum_{dihedrals} \sum_{\eta} \frac{V_\eta}{2} [1 + \cos(\eta\phi + \gamma)]$$
$$+ \sum_{j=1}^{atoms} \sum_{i>j}^{atoms} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\varepsilon r_{ij}} \right]$$

Figure 4: The AMBER energy function

This function is appropriate for representing the classical interactions between pairs of atoms. It is not appropriate for simulations involving bond forming or breaking which require the application of a Quantum Mechanical treatment, nor are any non-additive or polarization effects represented, although in principle they can be added at an additional cost. Also not shown in Figure 4 are additional terms which may be added to restrain the system and scaling terms for the 1-4 nonbonded interactions. In the form above, the potential energy representation can be thought of as a sum over all the bonds (2 covalently bonded atoms), angles (the angle between two connected bonds), dihedrals (represented by 4 atoms and the rotation about the central bond or alternatively planarity constraints among 4 atoms with bonds to a common central atom) and all the "nonbonded" terms.

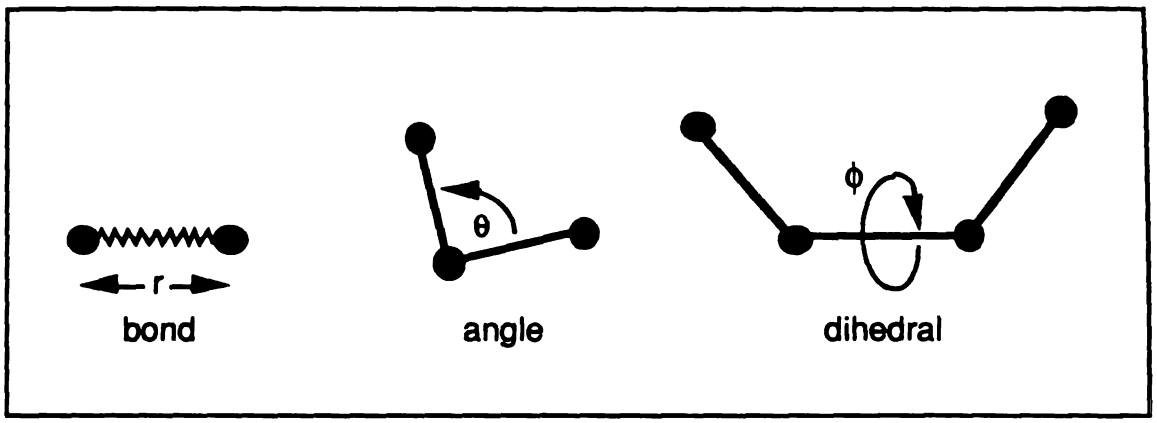


Figure 5: bonds, angles and dihedrals

The nonbonded terms represent the Lennard-Jones dispersion attraction and core repulsion and Coulombic electrostatic interactions between all pairs of atoms.

The philosophy of AMBER is the application of a fairly general (and transferable) force field. To this end, as few atoms types are chosen as possible to reasonably represent the system. For polymeric systems such as nucleic acids and proteins, consistent parameters are developed for a series of residues, typically at the nucleotide or amino acid level. It is then hoped that the parameters for the residues will be appropriate for larger macromolecules when the residues are connected together. The covalent terms (*i.e.* equilibrium distances and force constants) for the potential energy representation are developed from experiment and quantum mechanical calculations, where appropriate for a series of generic atom types. Nonbonded parameters for the dispersion attraction and core repulsion are also developed for these generic atom types, most recently based on liquid simulations, such as with the OPLS (optimized potential for liquid simulation) force field (Jorgensen & Pranata, 1990). The specific atom types are typically chosen based on the covalent connectivity, *i.e.* sp^3 carbon atom *vs.* a generic sp^2 carbon *vs.* sp^3 nitrogen, *etc.* with the goal of supplying transferability; for special chemical environments new atom types may be specified, such as an aromatic sp^2 carbon or sp^2 aromatic carbon in a five membered ring between two nitrogen's. The equilibrium values for the bond (r_{eq} in Å) and

angle (θ_{eq} in °) represent the ideal values and the instantaneous values are r and θ, respectively. The force constants for a particular bond and angle are k_{eq} (kcal/mol-Å^2) and k_θ (kcal/mol-deg^2), respectively. The representation for the dihedral angle terms is slightly more complicated due to complete rotation about the bond, hence the dihedral angle is modeled by a Fourier series where the periodicity is specified by η and phase by γ for a given angle ϕ. The force constant for each term in the series is $V_\eta/2$. The atomic repulsion and dispersion attraction are modeled by $1/r_{ij}^{12}$ and $1/r_{ij}^{6}$ terms, respectively, where r_{ij} represents the distance between atoms i and j. The parameters for individual atoms types, σ_i and ε_i represent the effective radius and equilibrium well depth for each atom. In AMBER, Lorentz-Berthelot combining rules are applied and used to represent the interactions between different atoms. All possible A_{ij} and B_{ij} are precalculated for each possible pair interaction based on the atomic radius (r^*) and equilibrium well depth (ε) as follows:

$$\varepsilon_{ij} = \sqrt{\varepsilon_i \varepsilon_j}$$
$$r_{ij}^* = \left(r_i^* + r_j^*\right)/2$$
$$A_{ij} = \varepsilon_{ij}\left(r_{ij}^*\right)^{12}$$
$$B_{ij} = 2\varepsilon_{ij}\left(r_{ij}\right)^6$$

Electrostatic interactions are represented, in the most simple form, by a Coulombic term which is proportional to the charges and inverse distance between atoms. Unlike the other parameters, the atomic charges are not as transferable and are strongly dependent on the chemical environment. To this end, each atom in a particular residue has a different charge (except for some atoms, such as methyl hydrogen atoms which may be equivalenced) which is calculated by performing a fit to an electrostatic potential generated from quantum mechanical calculations; this is discussed in much more detail in the respective force field papers. The form of the Coulombic term is rather simple; $q_i q_j/\varepsilon r_{ij}$, where q_i and q_j are the

charges on each of the atoms, r_{ij} is the distance between atoms and ε is the effective dielectric constant[1].

The most computationally demanding part of the potential energy evaluation is the calculation of the nonbonded terms. Since they involve all pairs of atoms, the calculation can become prohibitive for large numbers of atoms. Therefore, typically the interactions are truncated outside some cutoff distance, r_{cut}. This approximation, as is discussed in Chapter 1, can have drastic effects on the simulations. There are a variety of methods to cut off the longer ranged nonbonded interactions, ranging from the simple truncation of all interactions outside r_{cut} to more complex smoothing or switching of the discontinuities in the force or energy at the cutoff. An excellent description of cutoff methods and associated problems is presented by Steinbach & Brooks (Steinbach & Brooks, 1994). Typically, the interactions are cut off somewhere in the 9-15 Å range. Within the `sander` module of AMBER, which is used in the simulations presented in this thesis, little effort was spent to add proper methods to smooth the discontinuity at the cutoff, instead effort was spent incorporating code to properly treat the long ranged electrostatic interactions through the use of Ewald methods (as will be discussed in a later section of this introduction).

Throughout this text and within AMBER, all the distances are in angstroms (Å), angles in degrees (°), and energies in kcal/mol. For more information about particular force fields and the development of the parameter sets see the AMBER manuals, the published reports (Cornell *et al.*, 1995; Pearlman *et al.*, 1995; Weiner *et al.*, 1984) or the AMBER "world wide web" page originally developed by Bill Ross and myself at "http://www.amber.ucsf.edu". All of the calculations discussed in this thesis utilize the Cornell *et al.* (1995) force field, unless otherwise mentioned.

[1] Not shown is the constant $4\pi\varepsilon_0$ (where ε_0 is the permittivity of free space) in the denominator necessary to balance the units.

beyond "single point" energy evaluations

Given the molecular mechanics potential energy representation, various methods are used to study the energetic landscape beyond simple single point energy evaluations. These methods are discussed in much greater detail elsewhere[2]. Only a brief discussion is presented here in order to introduce some of the issues which will be developed further in later chapters. An obvious goal is the search for the lowest "energy" conformation, or "global" minimum, since it is expected that this will be the most representative structure. To this end, one can "minimize" the potential energy function of a molecule in a particular conformation. However, due to the complexity and high dimensionality of the potential energy hypersurface, minimization will generally only move the conformation of the molecule to a "local" minimum. Without complete sampling of the set of available conformations, it is impossible to determine if the minimum energy conformation is indeed the "global" minimum or even whether it is at all representative of what might be expected at room temperature[3].

Although it is meaningless to compare the molecular mechanical energies of two different molecules since these energies are not "absolute" and correspond to different

[2] The methods are reviewed to some degree in the AMBER paper (Pearlman et al., 1995) and in texts by McCammon & Harvey (McCammon & Harvey, 1987) and Brooks, Karplus & Pettitt (Brooks *et al.*, 1988). An excellent review of molecular dynamics and issues surrounding its use is presented by van Gunsteren and Berendsen (van Gunsteren & Berendsen, 1990). The "must have" technical reference is the seminal work by Allen & Tildesley (Allen & Tildesley, 1987).

[3] The expected distribution of conformations is that set of conformations that maximizes the entropy; this is the thermodynamic ensemble or Boltzmann distribution. In the text above, "energy" was used when in fact it is the conformation of lowest "free energy" or lowest energy and maximium degeneracy which is most favored. In general, not only is the lowest free energy conformation found or lowest free energy state populated, but conformations within a few kT are also populated where k is the Boltzmann constant and T is the temperature; at room temperature kT is approximately 0.6 kcal and it is not unreasonable to expect to find conformations within a few kT of each other. However, in molecular dynamics calculations it is highly unlikely during nanosecond length simulations to overcome barriers of more than a few kT; clearly surpassing a barrier to interconversion of 10-20 kcal/mol, such as *cis/trans* isomerization of proline, is highly unlikely, however smaller barriers, such as sugar repuckering or various backbone angle transitions in DNA should be readily be traversed at room temperature during a molecular dynamics simulation.

scales, it is possible to compare the energy of two different conformations of the same molecule *in vacuo*. This can give an indication of the relative difference in energy and may correlate with stability, however the relative energy difference is not a free energy since no entropic effects are included. For *in vacuo* calculations, an estimate of the vibrational free energy can be obtained by using a harmonic approximation to the normal modes to estimate the vibrational entropy. This is valid as long as both conformations represent true minima (which implies that the first six normal modes should be zero). It is this limitation which precludes the use of this approach in simulations with explicit solvent since it is intractable to "minimize" the positions of all the water, moreover, solvation is likely an ensemble property and not well represented by considering a small set of minima. In addition to giving an estimate of the vibrational entropy, the calculation of the normal mode frequencies can give insight into the "motion" of a molecule. Based on equipartition arguments, each normal mode frequency should be equally populated. However, low frequency modes (< 100 cm^{-1}) lead to most of the motion (Tidor *et al.*, 1983). Simulations on proteins suggest that the first 3-8 non zero normal mode frequencies account for 70% of the motion of the molecule (Levitt *et al.*, 1985). Analysis of these low frequency vibrational modes can give insight into which collective motions dominate. In Chapter 3, this is used to gain insight into the relative flexibility of A- and B-form geometries. It should be noted that the harmonic approximation to the normal mode frequencies may not be completely valid since proteins and nucleic acids have significant anharmonic character (McCammon & Harvey, 1987). However there are techniques for extracting information about the anharmonic modes from molecular dynamics trajectories (Levy *et al.*, 1984).

A drawback of minimization and normal mode calculations is the need for investigating a *representative* minimum energy conformation of the molecule; in other words, the insight gained is only as good as the choice of the initial conformation. Therefore, calculations of this type are generally limited to cases where the structure chosen

is reasonable, such as an experimentally derived structure, and where it is assumed that the force field will properly treat the structure as a minima. In the case of nucleic acids and minimizations *in vacuo*, the latter assumption is not always valid; without some representation of solvent in the simulations, the structure of a nucleic acid in solution will tend to distort from experimental values. Since minimization gets trapped into local minima, methods are desired which overcome this difficulty and reasonably represent an "ensemble" of energetically reasonable conformations. The difficulty in finding all the representative structures, for a given potential, is often termed the "local minimum" or "conformational sampling" problem. Given the expense of minimization calculations and the combinatorial explosion due to the independence of the degrees of freedom, exhaustive searching of all possible (or even reasonable) conformations is not possible. Therefore, other methods are such as Monte Carlo calculations and molecular dynamics methods are typically applied. In Monte Carlo calculations, random moves are made to a new conformation and the move is always retained if the new conformation is more favorable energetically. If it is less favorable energetically, the move is retained (probabilistically) some of the time. Molecular dynamics methods, on the other hand, assign random velocities about a mean temperature to each atom in the system, and propagate the dynamics of the atoms by integrating Newton's second law of motion. In Cartesian coordinates (r_i), where m_i is the mass of each atom (and t is time), the force is equal to the mass times the acceleration:

$$F_i(t) = m_i \frac{\partial^2 r_i(t)}{\partial t^2}.$$

With the Lagrangian defined in terms of the kinetic ($\frac{1}{2}\sum m_i v_i^2$ where v_i is the velocity on each atom) and potential energies (the E_{AMBER} as defined in Figure 4), the force is simply the negative gradient of the potential energy (which is calculated analytically):

$$F_i(t) = -\frac{\partial}{\partial r_i} E_{AMBER}(r_1,\ldots,r_N).$$

In AMBER, the propagation of the dynamics (or integration) is performed using the simple first order Leap-frog algorithm which is derived based on Taylor expansions; this is shown below for the velocities:

$$v_i\left(t_n+\frac{\Delta t}{2}\right) = v_i(t_n)+\left.\frac{\partial v_i}{\partial t}\right|_{t_n}\frac{\Delta t}{2}+\left.\frac{\partial^2 v_i}{\partial t^2}\right|_{t_n}\left(\frac{\Delta t}{2}\right)^2\frac{1}{2!}+\cdots$$

$$v_i\left(t_n-\frac{\Delta t}{2}\right) = v_i(t_n)-\left.\frac{\partial v_i}{\partial t}\right|_{t_n}\frac{\Delta t}{2}+\left.\frac{\partial^2 v_i}{\partial t^2}\right|_{t_n}\left(\frac{\Delta t}{2}\right)^2\frac{1}{2!}-\cdots$$

Subtracting, re-arranging terms and substitution of $\frac{\partial v_i(t)}{\partial t} = \frac{F_i}{m_i}$ leads to a representation of the "half step" velocities based on the previous half step velocities and the forces, where Δt is the time step:

$$v_i\left(t_n+\frac{\Delta t}{2}\right) = v_i\left(t_n-\frac{\Delta t}{2}\right)+\frac{F_i}{m_i}\Delta t$$

An analogous derivation leads to the on step positions.

$$r_i(t_n+\Delta t) = r_i(t_n)+v_i\left(t_n+\frac{\Delta t}{2}\right)\Delta t.$$

In order to properly represent the high frequency bond stretching motion, the equations of motion are integrated using a 1-2 femptosecond time step. This implies that a nanosecond simulation requires on the order of a million time steps.

Molecular dynamics has the added benefit of providing some estimate of the dynamics of the system and effective configurational entropy. However, given that the kinetic energy added to the system is finite– generally the dynamics are simulated at approximately room temperature or 300 K– the probability of overcoming large energetic

barriers is very small. Therefore, molecular dynamics calculations can get "stuck" in the vicinity of a local minima; this is exemplified nicely in Chapter 3 where it is observed that double stranded RNA duplex gets trapped in A-form and B-form structures at 300 K during the course of multi-nanosecond simulations depending on where the calculation is started. Monte Carlo methods also have some drawbacks. Not only is there no implicit time evolution to the system, and therefore no straightforward estimate of the dynamics (and no implicit entropic effects), it is often difficult to chose a set of moves that avoids trivial sampling (such as small atomic moves which change the energy and conformation only slightly) yet at the same time avoids unacceptable moves to high energy structures. The latter is particularly troublesome in simulations of flexible macromolecules with explicit solvent. In the limit of infinite sampling, both methods will ultimately converge to a Boltzmann distribution; however given that molecular dynamics simulations are currently restricted to the nanosecond time scale and Monte-Carlo calculations restricted to the evaluation of millions of moves (which both correspond to a roughly equivalent amount of sampling), the conformational sampling problem is still a major issue. There are a number of methods which have emerged which may prove useful for enhancing the effective sampling, such as Elber's locally enhanced sampling (Roitberg & Elber, 1991) method and Still's jumping between wells (Senderowitz *et al.*, 1995) adaptive Monte Carlo methods. However, it is still unclear whether these will prove useful in the simulation of solvated nucleic acids.

representation of solvent

In their biologically relevant forms proteins and nucleic acids exist in a solution of water, ions, and a rich diversity of other molecules. Even when extensively dehydrated, DNA still has associated water. The water is an integral part of the structure. At close range, serving both as a hydrogen bond donor and acceptor, water can specifically hydrogen bond to other molecules. As a bulk solvent, water screens charge interactions due to its high

dielectric constant (ε=78) and inhibits motion due to viscous damping forces. On a microscopic level, this screening arises from the polarizability of the medium. The total polarizability results from the reorientation of permanent dipoles (or orientational polarizability) and from electronic polarizability or creation of induced dipoles in the presence of an electric field (Gilson, 1995; Harvey, 1989; Sharp & Honig, 1990). In addition to the specific structural effects and polarizability, another important interaction of the solvent derives from non-specific entropic effects. Non-polar molecules, without hydrogen bonding capability or charges to interact with water, attempt to minimize their exposure to water and tend to associate; this hydrophobic effect is a large driving force in protein folding (Dill, 1990; Spolar & Record, 1994) and appears to be a larger driving force in DNA-ligand association than electrostatic effects (Misra & Honig, 1995).

Clearly some representation of water in molecular simulations is desirable. There are several ways to include solvent ranging from completely ignoring the water, to implicit models which represent water indirectly or macroscopically as a continuum, to explicitly including water with atomic detail.

in vacuo calculations

The calculation of the Coulombic interactions includes a parameter to specify the effective dielectric constant (ε). For gas phase calculations, ε=1. However, without explicit waters to screen the charges, *in vacuo* simulations tend to overemphasize the charge interactions. If the molecule of interest is truly in the gas phase, this is not a problem. However, with water solvated molecules such as proteins and nucleic acids simulated *in vacuo*, this is not realistic. Setting ε=1 in simulations of nucleic acids *in vacuo* leads to too much phosphate repulsion which distorts the structure. In the early simulations of nucleic acids *in vacuo* the structure fell apart unless the charges were removed (Levitt, 1983). Similarly with proteins, the lack of solvent on the surface provides no compensating interactions for the

polar groups which leads to excessive intramolecular hydrogen bonding and the formation of over stable ion pairs which tend to compact and distort the structure. A simple way to overcome this difficulty is to simply reduce the charges on ionic groups. However, ionic interactions are clearly important and simply "turning them off" is not the correct approach. Alternatively, the charges could be screened by raising the effective dielectric in the simulation; this however will screen the charges too much at short range and will tend to destabilize short ranged charge interactions such as hydrogen bonds.

Ideally, we desire a method that allows the full charges at short range (*i.e.* ε=1) and bulk solvent screening at longer range (*i.e.* ε=78 at ~20 Å). The simplest way to do this is through the application of modified dielectric functions or "effective" dielectric constants (Davis & McCammon, 1990). The most common and simplest form is the distance dependent dielectric constant where $\varepsilon=r_{ij}$ or $4r_{ij}$. This effectively screens the electrostatic interactions with distance and as an added bonus speeds up the calculation by removing the square root in the distance evaluation. A drawback is that this function screens too drastically at short range. A better treatment is the application of a slightly more complex, sigmoidal dielectric function (Daggett *et al.*, 1991; Hingerty *et al.*, 1985; Ramstein & Lavery, 1988) which tapers the short ranged electrostatic interactions more slowly, yet still effectively screens the longer ranged interactions as follows (where S is a parameter and D is the dielectric):

$$\varepsilon = D - \frac{D}{2} e^{-r_{ij}S}\left[r_{ij}S^2 + 2r_{ij}S + 2\right].$$

Alternatively, a dielectric that increases exponentially with distance (consistent with Debye-Hückel theory) has been applied in the simulation of nucleic acids (Sarai *et al.*, 1988; von Kitzing & Diekmann, 1987). A drawback of the distance based "effective" dielectric functions is the uniformity in the screening regardless of the proximity to solvent. This is a poor approximation for a macromolecule which tends to have a lower effective dielectric in

the interior of the molecule compared to bulk water. Moreover, these functions tend to cause the molecules to compact during molecular dynamic simulations and suppress motion (Harvey, 1989; Steinbach & Brooks, 1994).

implicit solvation

The simple distance based effective dielectric treatments completely neglect any representation of the hydrophobic effect. This can be included in an effective way based on the observation that the free energy of solvation for the saturated hydrocarbons in water is linearly related to the solvent accessible surface area (SASA). Eisenberg & MacLachlan developed a series of effective atomic solvation parameters (σ_i) based on the water/octanol transfer free energy for a variety of amino acid analogues (Eisenberg & McLachlan, 1986). In these, the free energy of interaction of a solute with water, $\Delta G_{residue}$, is related to the change in solvent accessible surface area (ΔA_i) as follows:

$$\Delta G_{residue} = \sum_{atoms\ i} \sigma_i \Delta A_i$$

Scheraga and coworkers also developed this method instead using vapor/water transfer free energies (Kang *et al.*, 1988; Ooi *et al.*, 1987; Williams *et al.*, 1992).

To use the method in molecular dynamics simulations, derivatives of the solvent accessible surface area are required. This was implemented into both CHARMM (Wesson & Eisenberg, 1992) and AMBER (Schiffer *et al.*, 1993) and in the later, dynamics cost roughly four times *in vacuo* simulations. The speed of the calculations can probably be improved thanks to the development of faster methods to calculate the SASA and its derivatives. A problem with this method is that the results are very sensitive to the parameterization and also that solvent screening is not included into the calculation; typically effective dielectric treatments are used in concert with this method. Moreover,

although protein folding may reasonably be represented by some representation of hydrophobic effects in a SASA term (Chiche *et al.*, 1990; Eisenberg & McLachlan, 1986; Novotny *et al.*, 1988) it is not clear if it will prove useful in simulations investigating the conformational preferences of DNA. The stability and conformational preferences of nucleic acids are largely due to base pair stacking, hydrogen bonding between the base pairs and with solvating water, and electrostatic interactions primarily related to the charged phosphate backbone and associated ions and solvent. Base stacking is clearly a major driving force in the stabilization of nucleic acid structure. This is supported by theoretical calculations (Friedman & Honig, 1995) and the observation that isolated bases stack in aqueous solution (Broom *et al.*, 1967; Ts'o *et al.*, 1962). Arguably, the base stacking interactions are largely hydrophobic and therefore may be modeled by a simple SASA term. However, despite fairly distinct structural differences between canonical A- and B-form geometries, the difference in solvent accessible surface area is actually slight (Alden & Kim, 1979); therefore it is unlikely that a simple SASA term will be sufficient to discriminate between the various DNA conformations.

In addition to including some representation of the hydrophobic effect, a better treatment of solvent polarization than is represented by the distance dependent dielectric is desired. A simple means to represent ionic solvation involves the inclusion of terms based on the Born model (Born, 1920). This is a solution to the Poisson equation for a charge q_i within a spherical cavity of radius r_i within a dielectric medium with dielectric constant ε_s and the free energy of solvation is as follows:

$$\Delta G_{solvation} = -166\frac{q_i^2}{r_i}\left(1-\frac{1}{\varepsilon_s}\right)$$

In the original usage, crystallographic radii were utilized which led to significant error in the calculated solvation free energies. Since this time, various adjustments to the radii have been proposed to give better results. In general, the results are very sensitive to the choice

of atomic radii and also the charges. This sensitivity both in the SASA "hydrophobic" term and in the representation of the solvent electrostatics or polarizability becomes a larger issue in the more advanced or derived treatments, such as the GBSA and Poisson-Boltzmann methods (discussed briefly below), where special care or parameterization is utilized to give a consistent "parameter" set (Sitkoff *et al.*, 1994).

The Born model requires a spherical cavity, hence the method is not appropriate for calculating the solvation free energy of general molecules. To provide a more general treatment, the model was extended by Still and coworkers and combined with a surface area treatment to create the "generalized Born surface area" (GBSA) approach (Still *et al.*, 1990). The key addition is the generalized Born term, $G_{polarization}$, which along with a SASA term ($G_{cavity} + G_{vdW}$), is added to the molecular mechanical energy. This polarization term is defined as follows:

$$G_{polarization} = -166\left(1-\frac{1}{\varepsilon}\right)\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{q_i q_j}{f_{GB}}, f_{GB} = \sqrt{r_{ij}^2 + \alpha_{ij}^2 e^{-D}}, \alpha_{ij} = \sqrt{\alpha_i \alpha_j}, D = \frac{r_{ij}^2}{2\alpha_{ij}^2}$$

The α_i or Born radii are calculated numerically for each charged atom in the solute and the values change as the calculation proceeds, as implemented in the program MacroModel, and other parameterizations are possible. The form of the functional above gives the Born energy for superimposed charges, the Onsager reaction field energy for a dipole in a spherical cavity at short range and at longer range the Born plus Coulomb dielectric polarization energy for two spheres. For a series of small molecules, this method gives reasonable free energies of solvation at modest computational cost (Still *et al.*, 1990).

A more complete description of the electrostatics can be obtained through continuum models, such as Poisson-Boltzmann (PB) electrostatic treatments. In the PB model, the solvent is treated completely macroscopically. The "solute" or molecule of interest is thought of as a region of a given dielectric bounded by a molecular surface with

embedded charges (at the atom centers) typically drawn from molecular mechanical force fields. Outside the molecular surface, the solute is surrounded by region of uniform "high" dielectric representing the solvent and possibly dissolved electrolyte. For a complete description of the relevant equations, see any of a number of reviews on the topic (Gilson, 1995; Harvey, 1989; Sharp & Honig, 1990). Except for some very simple cases, such as the Tanford-Kirkwood model which requires a spherical solute and independent ionizable groups (Tanford & Kirkwood, 1957) or other simple shapes such as cylinders, analytical solutions are not possible. Instead, often finite difference or other numerical methods are applied. There are a variety of programs for solving the Poisson-Boltzmann equation, such as the programs Delphi (Gilson *et al.*, 1987), UHBD (Madura *et al.*, 1995), MEAD (Bashford & Karplus, 1990) and others. Issues include the sensitivity to the approximating grid, radii for specifying the molecular surface and representation of the hydrophobic effect (typically through the addition of a surface area term). Although calculation of free energies of solvation are rapid compared to explicit free energy perturbation calculations in explicit solvent, the calculations are still moderately time consuming. Therefore, although the methods have been incorporated into molecular dynamics simulations (Gilson *et al.*, 1993; Gilson & Honig, 1991; Sharp, 1991; Zauhar, 1991), the complexity of the simulations has limited calculations to short time scales and few reports of its use in molecular dynamics simulations of proteins or nucleic acids have been published. Moreover, the models contain no microscopic description of the solvent. The results are also strongly dependent on the choice of dielectric, not only for the solvent which is fairly well defined, but for the solute of interest. In the case of proteins and nucleic acids controversy surrounds what precisely is the numerical value of the effective dielectric constant (Harvey, 1989; Warshel & Aqvist, 1991). Despite this, the PB model has proved useful in a variety of applications, such as the salt dependence of DNA-ligand interactions (Misra *et al.*, 1994), electrostatic contributions to the B-DNA to Z-DNA transition (Misra & Honig, 1996), base stacking (Friedman & Honig, 1995), salt bridges

in protein stability (Hendsch & Tidor, 1994), among other applications. A number of the more recent applications are reviewed by Gilson (Gilson, 1995).

An alternative method for representing the polarizability of the solvent which is intermediate between a continuum model and fully explicit water is the Langevin dipole and protein dipole-Langevin dipole (PDLD) methods developed by Warshel and coworkers (Warshel & Aqvist, 1991; Warshel & Levitt, 1976; Warshel & Russell, 1984). In this model, the polarizability due to reorientation of solvent dipoles is represented by fixed dipoles on a grid. This method has been useful for investigating solvation free energies and electrostatic components in free enzyme catalysis (Warshel & Aqvist, 1991) and appears to be fairly generally useful, although the general simulation community has not extensively adopted its use.

While these methods can give useful insight to the energetics and relative stability, the lack of a microscopic understanding limits its utility, particularly for investigating water's role in the structure, dynamics and relative flexibility of nucleic acids. This has recently been demonstrated to be an issue in simulations on DNA triplexes (Cheng & Pettitt, 1995) (as will be discussed in the section describing the history of nucleic acid simulations).

explicit solvent

Although much work has been done to represent explicit water, the models most frequently employed are the rigid three point waters, such as TIP3P (Jorgensen *et al.*, 1983) and SPC/E (Berendsen *et al.*, 1987). The three point water models are really very simple and characterized by three charges (q_O and q_H) in a fixed geometry (r_{OH}=0.9572 Å and $>_{HOH}$=104.52° for TIP3P and r_{OH}=1.0 Å and $>_{HOH}$=109.47° for SPC/E) , a van der Waals radius on the oxygen atom (r^*_O, ε_O) and optionally a van der Waals radius on the hydrogen. Four point models also exist where a fourth off center change is added on the OH bisector

0.15 Å from the oxygen, such as the TIP4P (Jorgensen *et al.*, 1983) model. The parameters for these models are summarized below.

	TIP3P	SPC/E	TIP4P
q_O	-0.834	-0.8476	-1.04
q_H	0.417	0.4238	0.52
r^*_O	1.7683	1.7766	1.7699
ε_O	0.152	0.1554	0.155

Table 2: rigid water models

During minimization, as long as atoms do not penetrate the oxygen van der Waals radius, and during Monte Carlo simulations (assuming individual atom moves in the water are not part of the move set), water is implicitly rigid. This is not true during molecular dynamics and therefore some method to keep the water models rigid must be applied. Typically, SHAKE (Ryckaert *et al.*, 1977) or the analytic version Settle (Miyamoto & Kollman, 1992) is applied to fix the positions of the hydrogens. The application of SHAKE has the added benefit of fixing the highest frequency motions in the system, typically the hydrogen bond stretching, which allows the simulations to be run with a slightly larger that one femptosecond time step (*i.e.* 1.5-2 fs). These rigid three point water models reasonably represent the density, interaction energies and first peak of the radial distribution function of bulk water. However, deficiencies include an underestimation of the compressibility, an absence of much structure beyond the first peak in the radial distribution function, and less tetrahedrality than expected. Additionally, TIP3P water diffuses too rapidly or

approximately twice what is seen experimentally. The fast diffusion could have implications in the molecular dynamics simulations where, as presented in Chapters 2 and 3, arguments are made about the flexibility and dynamics of the nucleic acids. However, given that simulations with SPC/E water show similar fluctuations and equivalent structures and hydration patterns (unpublished observations) to TIP3P, yet SPC/E diffuses at the expected rate, the arguments about flexibility with TIP3P are presumably still valid. This suggests that the choice of three point water model is not so critical.

However, recent work does show dependence on the treatment of the electrostatic interactions in the simulations of water solvated membranes (Feller *et al.*, 1996). Specifically, near the electrostatic cutoff (12-14 Å) "structure" appears in the radial distribution function where it should be flat. This leads to ~50% errors in surface tension and ~100% errors in the surface potentials compared to simulations applying an Ewald treatment (Ewald, 1921) (discussed below). Despite the better behavior when the long ranged electrostatic interactions are fully represented through the application of an Ewald treatment, it is not clear if the models, as parameterized, are fully appropriate since differences in some of the calculated water properties are observed. The inadequate representation of the tetrahedrality of the solvent may also have implications when trying to mimic the specific hydration of biomolecules.

Not withstanding the issues related to the long ranged electrostatic interactions, some of the deficiencies in the rigid three point water models can be corrected through the application of rigid four point models, such as TIP4P (Jorgensen *et al.*, 1983), or flexible water models (Ferguson, 1995; Levitt *et al.*, 1995; Mizan *et al.*, 1994). However, despite the availability of these additional water models, TIP3P and SPC/E water are typically the models of choice in the simulation of biomolecules even today. Part of this standard usage is probably historical. Moreover, the methods are typically optimized, and the force fields parameterized, for the rigid three point waters. In the case of the Cornell *et al.* force field,

the nucleic acid parameters are well "balanced" with respect to TIP3P. The issue of balance in the force field is important. Since we are not including electronic polarization effects explicitly (since induced dipoles cannot be formed in these simple models), it is worthwhile to utilize a water model that is "pre-polarized", such as TIP3P and SPC/E. These models mimic the polarization effects solely through dipole reorientation.

This issue of force field balance is a critical one. One problem with the previous "AMBER" force field (Weiner *et al.*) was that it tended to over-emphasize hydrogen bonds to water. This leads to break-up of the Watson-Crick base pairing unless extra restraining potentials are added. Part of the impetus for the development of the Cornell *et al.* force field was to fix this and other deficiencies of the earlier force field. The "balance" is in part restored by using a more consistent method for calculating the charges– the RESP methodology coupled with multi-conformation and multi-molecule fitting (Bayly *et al.*, 1993; Cieplak *et al.*, 1995) and a higher level "pre-polarized" basis set (6-31 G* instead of STO/3G) for the calculation of the electrostatic potential– and more consistent van der Waals parameters.

molecular dynamics with explicit solvent

Adding in explicit water tremendously increases the computational burden; therefore it is prudent to be cautious and add only as much explicit solvent as necessary. How much is necessary? Ideally enough water will be present to not only completely "solvate" the macromolecule, but also represent "bulk" water properties away from the macromolecule. This is an issue related not only to the quantity of water but to the boundary conditions applied. Additionally, to make the calculations more tractable, the pairwise interactions are typically cut off. While a cutoff in the 9-15 Å range is likely reasonable for Lennard-Jones interactions, it clearly is not appropriate for Coulombic electrostatic interactions which are

much longer ranged. In the next two sections, methods for handling the water boundary conditions and long ranged electrostatic interactions are discussed.

boundary conditions

The first simulation of a nucleic acid in explicit water involved a small duplex of DNA surrounded by a shell of 806 TIP3P water molecules (Siebel *et al.*, 1985). This droplet of water was in turn surrounded by a vacuum interface. This serves to solvate the DNA, however unless the droplet is very large, the "bulk" water properties are clearly misrepresented. The vacuum interface will alter the dipole reorientation at the surface and this in turn will change the polarizing effects of the solvent. An additional difficulty is that waters can "vaporize" or leave the droplet of solvent molecules. To circumvent the vaporization, a boundary potential can be added to restrain the waters within the droplet. However, even in the absence of a restraining boundary potential, the largest problem with this method for simulating the solvent is that the surface tension of the waters at the vacuum-water interface leads to high pressure in the system and reduced fluctuations in the system (Fox & Kollman, 1996; Steinbach & Brooks, 1994). Although the deficiencies can be minimized by adding more and more solvent, this makes the calculations much more expensive.

Various methods have been developed which circumvent the surface tension issue by "breaking up" the surface ordering of the water. Essex & Jorgensen demonstrate that a complicated restraining term at the surface, which in part destroys the ordering, can give better free energies than a harmonic potential (Essex & Jorgensen, 1995). As an alternative, stochastic boundary conditions can be applied which utilize the Langevin equation to add in frictional and random forces to the surface waters (Berkowitz & McCammon, 1982). This method has been applied with reasonable success in the simulation of proteins (Brooks *et al.*, 1985; Brooks *et al.*, 1988). A more complicated

treatment has been proposed by Roux and coworkers, where the sphere of waters is surrounded by a dielectric continuum (Beglov & Roux, 1994; Beglov & Roux, 1995). The serves to not only implicitly resolve some of the surface ordering effects, but also may add back in some representation of the solvent polarization due to bulk water. While this works reasonably well for small systems, for large systems such as nucleic acids the calculations become quite expensive since the molecule needs to be surrounded by a spherical shell of water and all interactions within the sphere calculated. An additional problem with these methods is how to represent pressure. To my knowledge, none of these models mentioned above which "break up" the surface ordering of water have been applied in simulations of nucleic acids, except recent work by Norberg & Nillson applying a stochastic boundary treatment with a large hydration sphere (~20 Å) to investigate guanylyl-3'-5'-uridine in solution (Norberg & Nilsson, 1995) and the glass transition temperature in DNA (Norberg & Nilsson, 1996).

Another means to rid the surface effect is to apply toroidal periodic boundary conditions (Allen & Tildesley, 1987; Born & Von Karman, 1912). This involves surrounding the simulation cell by virtual copies, replicating the simulation infinitely in each direction. This implies that a molecule leaving one side of the simulation cell immediately comes back in the opposite side.

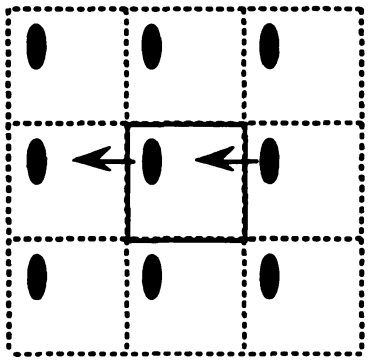


Figure 6: periodic boundary conditions

This avoids the vacuum interface at the expense of imposing a lattice around the molecule. Of course, as will be discussed later, imposing a lattice or crystal boundary conditions and "true" or infinite periodicity (where all interactions within the simulation cell and all copies of the simulation cell are represented) may have its own difficulties. However, when used in cutoff simulations, the explicit ordering is not a problem since a cutoff radius is chosen to be less than half the simulation cell (box) size and therefore the "image" molecules (*i.e.* copies of the molecule in the neighboring simulations cells and beyond) should not directly influence the structure since they are not within the cutoff sphere. Moreover, ordering of the entire lattice is inhibited by random thermal fluctuations from molecules which "suddenly" cross in or leave a given atoms sphere of influence. An alternative cutoff scheme in periodic boundary simulations is the minimum image condition, where only interactions within a single copy of the unit cell, centered on each atom, are considered. Although this also avoids seeing image molecules directly, it is problematic for non-spherical periodic boundary conditions since the "corner" interactions are over-represented. Moreover, reorientational motion is inhibited (Roberts & Schnitker, 1995). For this reason, they are typically not included in molecular dynamics simulations of biomolecules in explicit water.

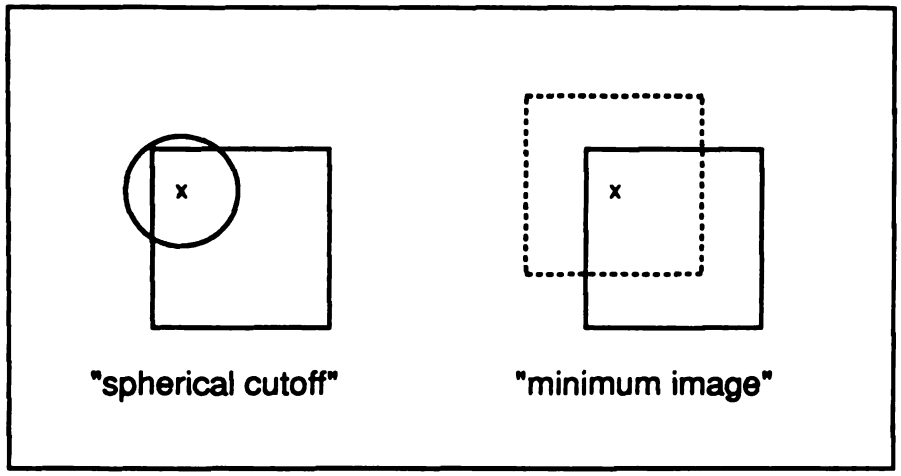


Figure 7: cutoff representations

An additional issue in simulations with periodic boundary conditions is how to handle fluctuations in the size or shape of the simulation cell. If the box size is held fixed, constant volume simulations are performed. When running a constant volume simulation, care must be taken to make sure the density is reasonable. If the volume is too large, as the simulation proceeds the solvent will relax to the equilibrium density and vacuum "bubbles" may appear. If the volume is too small, the density will be too high and fluctuations may be inhibited. An alternative method is the application of constant pressure. In this case, the box size is allowed to fluctuate and relax to a constant pressure; this can be done by a number of methods and in the simplest incarnation, Berendsen pressure coupling is often applied (Berendsen *et al.*, 1984). In the simulation of nucleic acids reported herein, constant pressure simulations were typically applied since this would not tend to inhibit conformational transitions which may tend to change the exposed volume of the macromolecule. Moreover, not only does this remove worries about vacuum bubbles and incorrect densities but it is more consistent with standard experimental conditions.

long ranged electrostatics

In order to limit the number of pairwise interactions in the evaluation of the molecular mechanical energy and forces, cutoffs on the nonbonded interactions are typically applied. Since the electrostatic interactions are long ranged, cutoffs can lead to drastic errors, particularly with highly charged systems. This has been shown in a variety of simulations. For example, in simulations with chloride ions (Cl^-) in water applying a cutoff, unexpected ion pairing was observed (Dang & Pettitt, 1987). Similarly, the potential of mean force profile representing the free energy of separating ferrous and ferric ions in water (Fe^{2+}-Fe^{3+}) with a 0.5 Å spline smoothed cutoff at 11.73 Å showed a well at ~6-7 Å and an attractive potential beyond 7 Å (Bader & Chandler, 1992). Control simulations run with the ions separated by 9 Å showed the ions to move closer together. These rather surprising results were shown to be an artifact of the truncation of the electrostatic forces. When the cutoff was removed through the use of "true" periodicity and application of an Ewald treatment (Allen & Tildesley, 1987; Ewald, 1921), chloride ion pairs were no longer observed (Hummer *et al.*, 1993) and the "well" in the ferric-ferrous ion PMF profile disappeared (Bader & Chandler, 1992). Cutoff simulations with less significantly charged biomolecules also show the anomalous properties of the cutoff. A nice visual example of the anomalies of the cutoff appears in a paper by Smith & Pettitt investigating the Coulombic electrostatic potential energy around a pentapeptide in explicit solvent with salt ions, where in contrast to the smooth isoenergy surface obtained with an Ewald treatment, rough and jagged isoenergy surfaces are seen with a cutoff with a shift (Smith & Pettitt, 1991).

The perhaps most dramatic evidence of anomalous behavior with a cutoff is the simulations of Schreiber & Steinhauser investigating a 17-residue alpha helical peptide in solution at various cutoffs compared to an Ewald treatment (Schreiber & Steinhauser, 1992). With a short cutoff (6 Å), the helix was unstable. With a 10 Å cutoff and with an

Ewald treatment, the helix was stable. Most interestingly, when an even longer cutoff was applied (14 Å), the helix was again unstable. This demonstrates that the errors introduced by truncation are not monotonically related to distance and suggest that even truncations in the 15-20 Å range may have problems. What is particularly interesting in these simulations is the fortuitously good agreement when an intermediate cutoff of 10 Å was used. This fortuitous agreement between Ewald and spherical cutoff methods (thanks to cancellation of errors) has been seen in other investigations (Roberts & Schnitker, 1995) and help explain why it was not until more recently that problems with the cutoff methods came to light. The observation that an even larger cutoff does not make the simulation "better" is especially sobering and points to the importance of properly treating the long ranged electrostatic interactions.

In our simulations, we were interested in using explicit water and counterions in the simulation of nucleic acids. Because of the issues with non-periodic boundary conditions (how to represent pressure, surface tension/ordering, lack of "bulk" water properties and the need for large numbers of water which slows the calculation), we opted to utilize periodic boundary conditions in all of the solvent simulations reported in this thesis. There are a variety of methods to properly represent the long ranged electrostatic interactions for simulations imposing periodic boundary conditions. These include the reaction field, Ewald, and fast multipole methods. In the end, thanks to available code and its appropriateness for systems in the ~10,000 atom range, we ended up using fast Ewald methods, in particular the particle mesh Ewald (Darden *et al.*, 1993; Essmann *et al.*, 1995) (PME) method.

reaction field

The reaction field method represents the field on a set of dipoles and includes a short range part (within a cutoff) representing the interaction of all the dipoles within a spherical cavity and a longer ranged part which is assumed to be a dielectric continuum (Onsager, 1936).

The calculation of the reaction field only adds a small computational cost over standard truncation. In spite of this, the method has not been applied extensively in the simulation of solvated biomolecules. This is likely due to the requirement that the spherical cavity be larger than the longest correlation length within it. This is likely on the order of the solute size in simulations of proteins or nucleic acids (~20-50 Å) which would require a periodic box much larger than standard periodic boxes and require the calculation of all pairwise interactions within the sphere. Additional issues with this method include the difficulty in conserving energy (due to discontinuities when molecules cross out of the cutoff sphere cavity) and the need to specify a fixed external dielectric constant. Realistically, the reaction field method is most appropriate for liquid simulations due to the implicit long range isotropy and shorter correlation lengths.

fast multipole methods

The fast multipole methods represent a set of methods which reduce the cost of computing the long ranged electrostatic interactions to linear time. This is a substantial savings and moreover the method is fairly easily parallelized. The basic idea is to recursively group distant atoms into multipoles based on hierarchical trees (Greengard, 1988; Greengard & Rokhlin, 1989). The method was extended to periodic systems by Schmidt & Lee (Schmidt & Lee, 1991). Modifications of the fast mulipole method, called the cell multipole method, by Ding *et al.* (1992) are considerably faster at the expense of accuracy. Details of the methods are presented in the review by Essmann and Darden (Essmann & Darden, 1996). There has been considerable controversy as to the relative performance of FMM methods to the fast Ewald methods (which are discussed in the next section) (Petersen, 1995; Solvason *et al.*, 1995). To obtain a similar level of accuracy to the Ewald methods, high order multipole (*i.e.* 8th order) expansions need be utilized. Moreover, although the complexity scales linearly with the number of atoms, the cost increases tremendously with higher order multipole expansions. The current belief is that the break

even point for the FMM and fast Ewald methods is in the 50,000 atom range or higher. Some issues with the FMM method are that energy conservation is difficult to obtain and errors appear to grow with system size.

Ewald[4]

Applying "true" periodic boundary conditions, as mentioned above, considers all pairwise interactions within the simulation cell and its infinitely replicated images in each direction. This requires a modification to the Coulombic term from its standard form with charges q_{atom} and coordinates r_{atom} representing the sum over all N atoms (and remembering that we are omitting the $4\pi\varepsilon_0$ term),

$$E_{electrostatic} = \frac{1}{2}\sum_{i,j=1}^{N}\frac{q_i q_j}{|r_i - r_j|}$$

to add in an additional summation over the lattice vectors, "n", where $n = n_x X + n_y Y + n_z Z$, as follows:

$$E_{electrostatic} = \frac{1}{2}\sum_{n}{}'\sum_{i,j=1}^{N}\frac{q_i q_j}{|r_i - r_j + n|}$$

The prime in the summation over lattice vectors implies that when n=0 the i=j terms are omitted. This series is conditionally convergent when the simulation cell is neutral and diverges if the simulation cell is not neutral. Ewald (Ewald, 1921) converted this summation over lattice vectors into a sum of two absolutely converging series. Naively this can be thought of as a short ranged part, or the direct sum, and a long ranged part in reciprocal space, called the reciprocal sum. The standard means for describing how the Ewald summation works is through the example presented in Allen & Tildesley (Allen & Tildesley, 1987), where we consider subtracting a screening charge distribution to make

[4] As mentioned in the preface, a large help in developing my understanding of Ewald methods and of other long ranged electrostatic treatments came from discussions with and papers by Tom Darden. In particular, in developing this section and the discussions on Poisson-Boltzmann electrostatics, great help came from a review chapter preprint entitled "Long-range electrostatic effects" by U. Essmann and T. A. Darden, sent to me by Darden.

the summation over charges "short ranged" (the direct sum) and then later add it back in (the reciprocal sum). This is shown schematically in Figure 8 for a series of point charges represented by arrows in (a) which is equivalent to the charges with the screening charge distributions added in (b) and subtracted back out in (c).

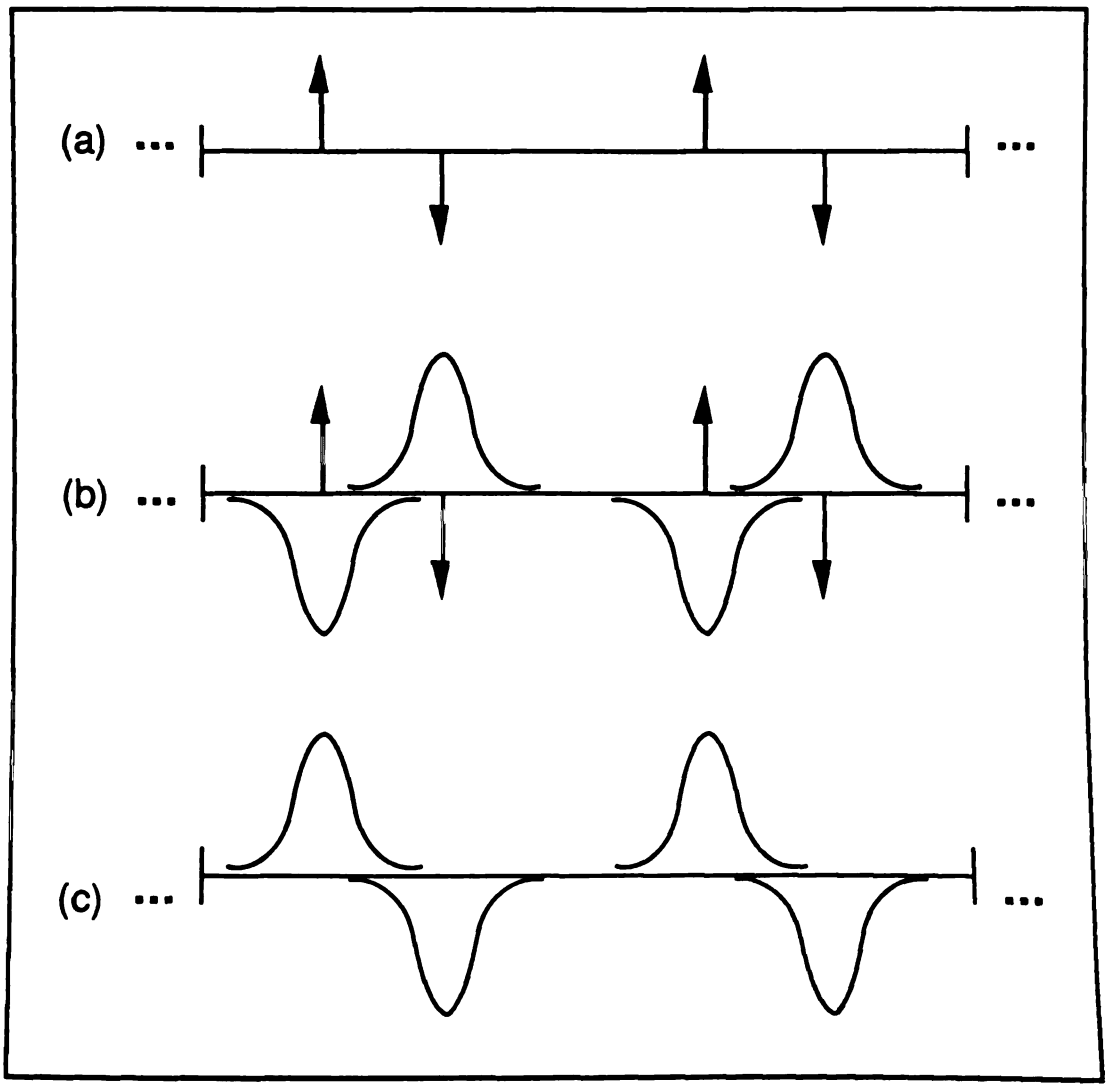


Figure 8: charges with screening Gaussians

For convenience, the form of this screening charge distribution is normally taken to be a Gaussian. In reality, although the reciprocal sum does in large part represent the long

ranged part, it still includes some short ranged interactions due to the finite tails of the Gaussian.

As mentioned above, the summation will diverge if the unit cell is not neutral. To overcome this difficulty for charged systems, either the simulation cell can be neutralized by subtracting off the net average charge smeared equally over every atom or a net neutralizing "plasma" can be added in to neutralize the system. The net neutralizing plasma need not be included explicitly (although it can be) since although the energies diverge, the forces are reasonable. In the version of the code used in the calculations presented in this thesis, for charged systems, the net charge was removed by subtracting it off each atom; in other words if a total charge of -18.0 existed in the simulation cell representing 10,000 atoms, -18.0/10,000 was added to each atom. This does not effect the charges very much as long as there are a sufficient number of atoms in the simulation cell; moreover in most of the simulations, explicit counterions were added to net neutralize the system.

The issue of conditional convergence is more complicated. Conditional convergence implies that the result depends on the order the terms in the summation are added up. As pointed out by Essmann and Darden in their excellent review on the subject (Essmann & Darden, 1996), the energies and forces will be different depending on whether the lattice is added up as slabs or as concentric spheres around the central simulation cell. As mentioned previously, the Ewald summation method formally converts the conditionally converging sum into a sum of two absolutely converging series, the direct sum (b) and the reciprocal sum (c); this is not formally correct since the conditionally converging part of the sum is missing. DeLeeuw *et al.* (1980) treat this issue explicitly, and show that the standard Ewald sum needs to be modified through the addition of a "surface" correction term which relates to the dipole of the simulation cell and contains the conditionally convergent part. This is developed by considering the summation over lattices as proceeding by adding in more and more concentric spheres until convergence. At the (infinite) boundary, the system can be considered to be surrounded by a vacuum or infinite

dielectric. In most implementations, the boundary is considered to be an infinite dielectric; these so-called conducting or "tin foil" boundary conditions make the surface correction term equal to zero. Roberts & Schnitker (1995) have analyzed the implications of applying non-conducting or vacuum (with the dipolar correction term) *versus* tin-foil boundary simulations. The basic conclusion is that since any disordered system of macroscopic size should not have any surface structure, small unit cells (which are characteristic of the simulation cells we are studying) should also not impose surface structure. This means that the dipolar correction term should vanish (*i.e.* we should apply tin-foil boundary conditions) since without surface ordering, surface polarization effects should vanish. In other words, the Ewald method is appropriate for simulations of liquids or "solution" phase simulations.

Applying the Ewald treatment sketched out above, following the notation of Essmann & Darden (Essmann *et al.*, 1995), we arrive at the following equation, assuming a Gaussian screening function of the form $\rho(r) = \beta^3 \exp(-\beta^2 r^2/\pi^{3/2})$ where β represents the width of the Gaussian:

$$E(r_1, r_2, \ldots, r_N) = \frac{1}{2}\sum_{n}{}'\sum_{i}\sum_{j}\frac{q_i q_j \cdot \mathrm{erfc}(\beta \cdot |r_i - r_j + n|)}{|r_i - r_j + n|}$$
$$+\frac{1}{2\pi L^3}\sum_{m\neq 0}\frac{\exp(-\pi^2 m^2/\beta^2)}{m^2}S(m)S(-m)$$
$$-\frac{\beta}{\sqrt{\pi}}\sum_i q_i^2 + \frac{2\pi}{3L^3}\left(\sum_i q_i r_i\right)^2$$
$$\text{where } S(m) = \sum_j q_j \cdot \exp(2\pi i\, m \bullet r_j)$$

In the above equation, the first term is the direct sum, the second term is the reciprocal sum (in fractional coordinates, m, and where L is the length of the simulation cell), the third term is the self energy (to avoid double counting) and the final term is the dipolar correction term (which we assume to be zero since we are applying tin-foil boundary conditions). The "erfc()" function above represents the complementary error function, defined as follows:

$$\mathrm{erfc}(x) = 1 - \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt$$

Since there is no analytical solution, this must be pre-computed numerically; a nice discussion of methods to table the values of the erfc is presented in the appendix to a paper by Feller *et al.* (1996). The erfc() function serves to screen the range of the direct sum; if the width of the Gaussian function is chosen to be small enough, the direct sum converges more quickly at the expense of a longer convergence for the reciprocal sum. For optimal efficiency in the standard Ewald, a value of β is chosen to be large enough such that the direct sum term is negligible beyond half the box length (L/2).

Although the Ewald treatment clearly gives a proper treatment of the long ranged electrostatic interactions, its use in the simulation of solvated biomolecules was limited due to the computational complexity compared to standard cutoff methods. Moreover, the real problems with the cutoff of the electrostatic interactions in the simulation of biomolecules did not become apparent until relatively recently when longer simulations became possible. However, thanks to recent advances in the methodology, a number of formally N•log(N) Ewald treatments have been developed, including the particle mesh Ewald (PME) (Darden *et al.*, 1993; Essmann *et al.*, 1995; Petersen, 1995), particle-particle particle-mesh Ewald (P3ME) (Hockney & Eastwood, 1981; Luty *et al.*, 1994; Luty *et al.*, 1995; Shimada *et al.*, 1993) and fast Fourier Poisson (York & Yang, 1994) methods. Without going into the details of each method, the basic idea is to convert the reciprocal sum into a form that can be evaluated quickly through the use of fast Fourier transform methods. In the case of the fast Fourier Poisson methods, Poisson's equation in periodic boundary conditions with Gaussian charge densities as sources (the reciprocal sum) is solved directly. This method is the most accurate, although more costly than the PME and P3ME methods. The PME and P3ME methods are very similar both algorithmically and in terms of computational cost. Both methods involve interpolating the charges onto a grid and then using a fast

Fourier transform to solve the discretized Poisson's equation. The differences come down to how the charges are interpolated.

In the PME method implemented by Darden, the charges were interpolated originally using a Lagrangian interpolation (Darden *et al.*, 1993) and more recently using a cardinal B-spline (Essmann *et al.*, 1995). The interpolation is performed such that the energies can be computed to the desired accuracy (depending on the interpolation order and grid spacing) and energy is conserved. Since the B-splines are differentiable, the forces are then calculated analytically from the charge grid. [With the Lagrangian interpolation, the forces also had to be interpolated (Darden *et al.*, 1993).] Since the forces are calculated analytically from the "approximating" charge grid, momentum is not strictly conserved (as is discussed in more detail in Chapter 1). The method can be switched to interpolate the forces directly at the expense of less accurate energies; however at the time it was thought that energy conservation was more important.

The P3ME methods use a simpler quadratic interpolation scheme (the "triangle charge" weighting scheme). To increase the accuracy, a least squares approach is used to pre-compute an "optimal influence function" (Hockney & Eastwood, 1981). Because of the use of the optimal influence function, they are able to obtain similar accuracy to the PME methods with a cubic interpolation order by using a quadratic interpolation order, therefore the P3ME method is slightly faster (neglecting the start up costs associated with calculating the optimal influence function). However, the values of the optimal influence function depend on box size. Therefore, although it can be adapted to run under constant pressure (perhaps by creating a series of optimal influence functions for various box sizes and interpolating between them) it is unclear how to use this method in cases where the box shape changes.

The PME method is more general since it can be tuned to arbitrary precision by increasing the interpolation order; not only is this useful for applications which require significant accuracy and stability (such as hybrid QM/MM methods) but it may be important

for speed since the direct sum part becomes rate limiting. As the time required to evaluate the direct sum is lessened (by increasing β and therefore decreasing the effective cutoff) more work ends up in the high frequency structure factors which require greater accuracy to evaluate correctly. An additional benefit of the PME method is that it generalizes to any inverse "r" potential (Essmann *et al.*, 1995), such as the Lennard-Jones dispersion attraction (r^{-6}). This is important for "true" periodic boundary simulations, since although the infinite sum for inverse third or greater terms is convergent, the interactions can be fairly long reaching. An easy way to visualize this is that as more and more concentric image shells around the simulation cell are added in, more of the attractive dispersion interaction is added in (albeit in small amounts until convergence). Cutting off this attractive potential at 7-12 Å could have deleterious effects, such as densities that are a little too low. Little has been discussed on this issue in the literature, however preliminary simulations of protein crystals by Darden suggest its importance (Darden, 1996). An additional benefit of including the long range dispersion attraction is that even shorter cutoffs may be used in the simulations, or ~5-7 Å, which (due to the constant speed up the reciprocal sum) speeds up the calculation. Unfortunately this is problematic with the AMBER force fields due to the use of Lorentz-Berthelot combining rules which requires the evaluation of seven reciprocal sums for the calculation of the long ranged attractive dispersion and therefore this method has not been used in practice. This issue can be overcome by either using geometric combining rules for the long ranged dispersion or alternatively updating the long ranged dispersion less frequently, using the methods employed with twin ranged cutoffs (Biesiadecki & Skeel, 1993). For more information about the methods, see the detailed review by Essmann & Darden (1996).

From the discussion above, it is clearly important to include some representation of the long ranged electrostatic interactions. The fast Ewald methods are a natural choice for simulations of solvated proteins and nucleic acids in the 5000 to 50,000 atom range. However, the imposition of "true" periodic boundary conditions comes at a cost,

specifically the possibility of inhibiting motion, inducing correlations, and other possible deleterious effects from the interactions of neighboring images. A simple means to explain the issue is with two simple examples; a freely rotating dipole and two charges in a periodic box (Smith & Pettitt, 1991; Valleau & Whittington, 1977).

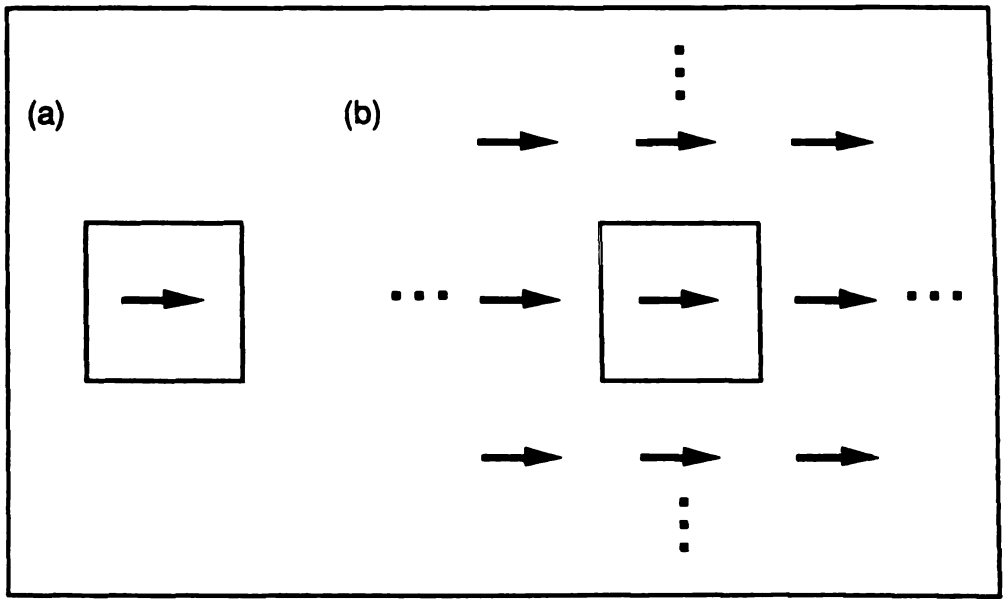


Figure 9: dipole in a periodic box

Given a dipole in a periodic box (such as in the figure above), it should be able to freely reorient. However, under true periodic boundary conditions (Figure 9b), the free reorientation may be inhibited by "image" dipoles.

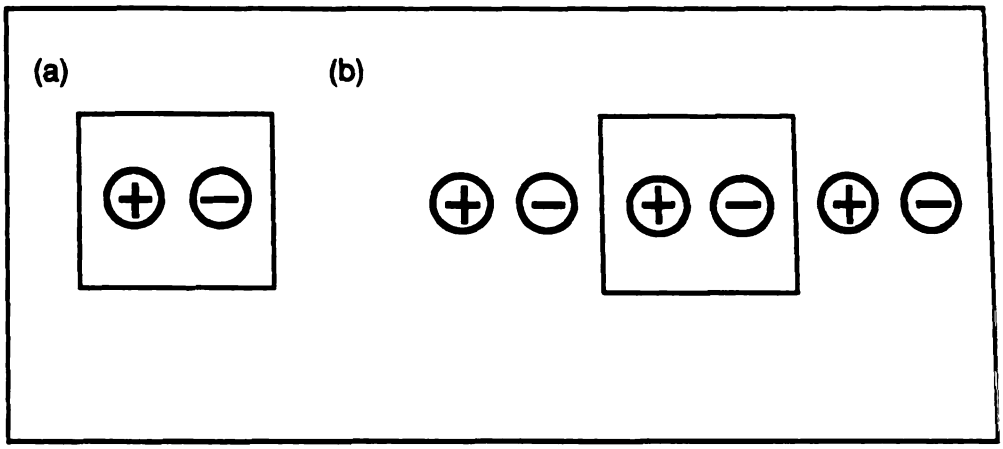


Figure 10: charges in a periodic box

In a similar manner, two like charges in a box should be attracted to each other and two opposite charges should repel each other. However, if we assume the charges are spaced at half the box length in a periodic box (shown in two dimensions in Figure 10b), no net force will be felt on the particles under true periodic boundary conditions. These effects could in principle lead to reduced fluctuations and induced correlations which in turn could have deleterious effects on the simulation. Fortunately, since water has such a high dielectric constant, this should not be too large an issue in the simulation of solvated biomolecules. However, *a priori*, it was not clear what effect the imposition of true periodicity might have. Therefore, in the first simulations (Chapter 1 and Chapter 2), considerable effort was placed on demonstrating that fluctuations were not inhibited and that conformational sampling was "reasonable". Recently a paper was published by Smith & Pettitt (1996) that suggests that the reorientation of a dipole or small protein in water is not hindered significantly. While there is a small effect observed for a dipole or quadrupole, the energy change is well below k_BT and according to the author's "may be safely ignored" for simulations in explicit water or other solvents with a high dielectric. However problems may arise in simulations with a lower effective dielectric, such as ethanol or hydrocarbon solvents, as was seen initially in the simulations of DNA in ethanol presented in Chapter 4.

An additional issue with imposing true periodicity is the problem of a molecule "seeing" its images in adjacent cells. To avoid artifacts, ideally enough water will be solvating the molecule of interest, or solute, that "bulk" properties can be maintained a distance away from the solute and images will not interact. This is a practical issue and suggests building a sufficiently large box. However, at the same time, it is desirable to have as small a box as reasonable to speed the calculations. In general, in the simulations presented herein, ~10 Å of water surrounded the solute in each direction initially. However, given that we used rectangular boxes in most of the simulations, rotation of the molecule can bring the "ends" of the nucleic acid duplexes closer together (as shown in Figure 11).

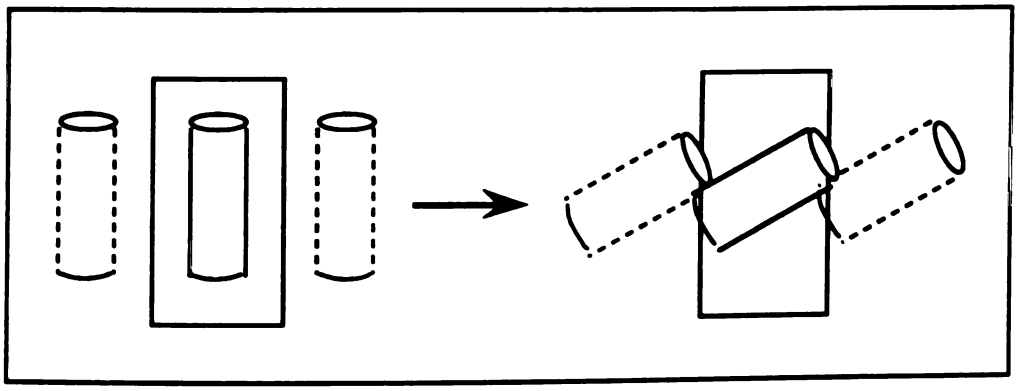


Figure 11: rotation in a rectangular periodic box

In the simulations discussed in this thesis, the anisotropy in the box dimensions is not too extreme (with the box sizes ~55-60 Å by ~40 Å by ~40 Å), however it is sufficient to bring the ends rather close together. Since the rotational correlation time is on the order of a nanosecond, it is expected that molecules will reorient during nanosecond length trajectories. It is expected that this problem will become much more of an issue when simulating larger nucleic acid sequences in rectangular boxes. Obviously this is not an issue in crystal simulations; however in general we were interested in mimicking "solution" conditions with an absence of crystal packing influences.

The final issue in using Ewald methods, or applying true periodicity, relates to a subtle force field dependence. Differences in the calculated properties are observed in liquid water simulations depending on whether a cutoff or Ewald treatment is applied to the electrostatic interactions.

	expt.	PME	8 Å	10 Å	12 Å	14 Å	SPC/E (cut)	SPC/E (pme)
ρ (g/cm^3)	0.997	0.937	0.954	0.963	0.974	0.978	1.0	0.99
E_i (kcal/mol)	-9.92	-9.05	-9.22	-9.27	-9.34	-9.37	-11.3	-11.1
D ($\times 10^{-5}$ cm^2/s)	2.3	~5.1	5.3				2.5	2.3

Table 3: Ewald and cutoff effects on water properties

As the data in Table 3 show, when applying an Ewald treatment, the density (ρ), average water-water interaction energy (E_i) and diffusion constant (D) are all underestimated with respect to simulations with a cutoff. Also apparent from the data is that as a longer cutoff is applied, the values (except for the diffusion constant) get closer to experiment; the values shown here were not corrected for the cutoff. All of the data was from simulations of 1000 water molecules, except for the diffusion constants and all the SPC/E values which were taken from Essmann *et al.* (1995); the values computed from independent water trajectories (with TIP3P and SPC/E models) using the mean square displacements to estimate the diffusion show the same trend. Although using an Ewald treatment only has a small effect on the calculated water parameters, it is a significant effect. The lowering of the diffusion constant (plus early fears stemming from a later retracted paper entitled "Ewald summation retards translational motion in molecular dynamics simulation of water" (Roberts & Schnitker, 1995; Teleman & Wallqvist, 1990)) led us to worry that application of the

Ewald method would lead to reduced fluctuations during the molecular dynamics simulations. Fortunately, these issues did not seem to be a large problem in the simulation of nucleic acids in water, especially compared to the larger problems which occur when long ranged electrostatic interactions are ignored.

enabling technologies

Much of this work in this thesis would not have been possible, as previously mentioned, without the help and extensive work performed by a series of collaborators. Of particular importance was the integration of Darden's particle mesh Ewald code into AMBER and the its subsequent parallelization. The following two sections outline this history and my role in the whole process. Overall, I did not serve as the innovator or primary developer of the work that follows; instead my role was more as of an integrator. I worked with the various parties to integrate the code into the general AMBER release, applied and tested the code, helped to ferret out limitations and fix bugs, and ultimately in the end helped it to mature (I hope!). Both of these efforts coincidentally came to my attention during the Computational Chemistry Gordon Conference in the summer of 1994. At this conference I met Dr. Thomas Darden from the National Institute of Environmental Health Sciences who introduced me to particle mesh Ewald methods, David Deerfield from the Pittsburgh Supercomputing Center (PSC) who set up collaborations within PSC to help optimize and develop AMBER, and Rob Stanton (who was then a graduate student in Dr. Ken Merz's group at Pennsylvania State University and) who told me of Jim Vincent's work (another graduate student in Merz's group) parallelizing AMBER under MPI. Darden and I began the collaboration to integrate his particle mesh code into the general AMBER release. Deerfield introduced us to Michael Crowley, who not only worked on fixing up the AMBER release for the Cray T3D, but parallelized the particle mesh Ewald code for the Cray T3D and later T3E on which a large part of these calculations were run. Independent of this effort, when I returned from the conference, I contacted Vincent who started a collaboration to put his code into the general AMBER release. All collaborations require some work to maintain confidences, retain openness, and give proper acknowledgment to the work done by each individual; in this respect I was fortunate that most collaborations worked without problems and still continue to this day. I learned a lot about politics

throughout this process, but in the end I think the entire AMBER simulation community has benefited from this work.

particle mesh Ewald within AMBER

Integration of Darden's code into version 4.1 of AMBER was relatively straightforward given working code by Darden. The largest issues were adding it in cleanly and gracefully into the source handling and compiling machinery, setting up the input and output, and testing it out. Discussing the complete details of the code integration is beyond the scope of this introduction; however a brief discussion follows. The code integrated was a modified version of the "public domain" water code distributed by Darden (Essmann *et al.*, 1995) and utilized B-splines instead of Lagrange interpolation for improved accuracy and analytical gradients. Integration was facilitated by simply bypassing the AMBER nonbonded routines to call PME routines instead. Small modifications to the force routine were necessary to support the additional energy terms and modification to the dynamics (`runmd`) to call PME pressure scaling routines instead. Rather than utilizing the AMBER pairlist, the PME routines build an atom based pairlist using a gridding procedure for extra speed. This differs from the standard "residue" or charge-group based cutoff within standard AMBER. However, given that the long ranged electrostatic interactions are explicitly included, and the "cutoff" refers to the Lennard-Jones interactions, an atom based cutoff is appropriate. Since the pairlist is already built for the Lennard-Jones interactions, it is useful to utilize this for the direct sum part of the Ewald summation since logistically, the direct sum is equivalent to the standard Coulombic term multiplied by the erfc() function. To make the range of the direct sum die out at the cutoff, the width of the Gaussian screening function is tailored to make the value of the direct sum at the cutoff be less than some tolerance; in the input file this is specified as DSUM_TOL and is generally equal to 0.000001 to 0.00001. The size of the reciprocal charge grid and B-spline interpolation order are specified in the input file. In general, the simulations are run with a cubic B-

spline and a charge grid that is approximately equal to the box size (implying ~1 Å grid spacing) but also a product of powers of 2, 3 or 5 (for FFT speed). The overall accuracy obtained by the program relates to a combination of the direct sum tolerance, the interpolation order and the grid size.

With the implementation in version 4.1 of AMBER, the interpolation on the charge grid in the reciprocal sum was set up to calculate the energies, and the forces were obtained analytically from the energy charge grid. This approximation, coupled with finite errors due to the grid size, interpolation order and direct sum tolerance in the PME method leads to incomplete conservation of force. Energy, on the other hand, is conserved. The lack of force conservation led to problems initially where a stochastic force component appears and gradually dumps translational kinetic energy into the system. Energy drains in the standard dynamics due to not updating the pairlist every step, SHAKE tolerances that are too high and the application of Berendsen pressure and temperature coupling lead to drops in the temperature. Therefore, the velocities are uniformly scaled up. This lead to a slow growth in the center of mass motion. As this motion increased, less kinetic energy was available to internal degrees of freedom, and effectively the motion of the molecule slowed down. Eventually, essentially all of the kinetic energy was in this one (translational) mode, the center of mass motion. This became known as the "flying block of ice" or "block of ice flying through space" problem. To circumvent this problem of force conservation with the PME method, the net center of force was subsequently removed at every step. At the time we instituted this fix, it hide the problem such that we thought it had been completely resolved; however as will be discussed at the end of Chapter 1, the problem came back to haunt me and others.

A final issue that is worth mentioning with respect to the PME method is "imaging" difficulties. When running a truly periodic simulation, the energies and forces are invariant to which simulation cell the molecule is in. In the case of DNA (as shown in Figure 12

below) as long as the two real strands "image" to be the same molecule, having two strands of DNA in the same cell or the strands split between multiple cells is irrelevant as far as the code is concerned.

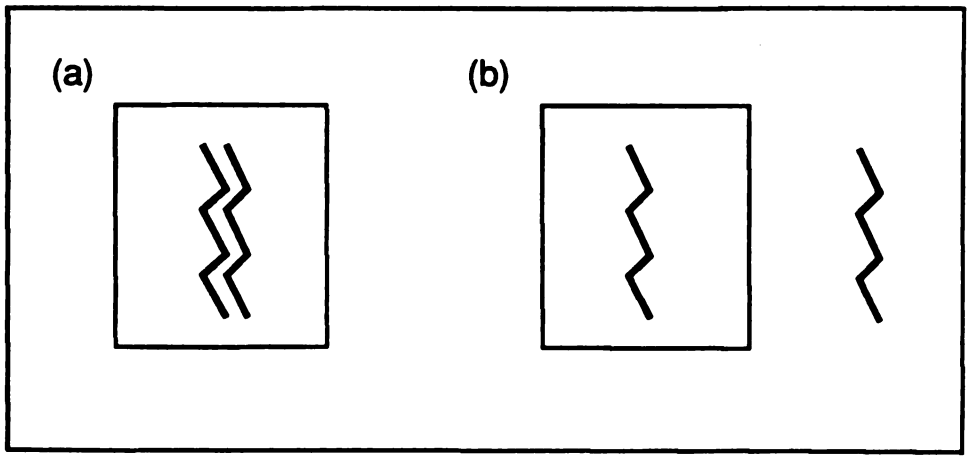


Figure 12: imaging in periodic boxes

This can cause problems when analyzing the trajectory from a simulation since molecules and/or residues will rotate and translate and eventually cross the box boundaries. Unless imaging is done consistently, the molecules or residues may get split across box boundaries. This is most easily seen by looking at the RMSd versus time for one of these trajectories where all of a sudden, the molecule appears to move ~1/2 the box in one step. Providing a general fix into the code to image consistently is not straightforward unless extra information is supplied by the user specifying what molecules should be imaged together. A way around this problem is to simply remove center of mass motion frequently for the solute since this will prevent it from ever crossing box boundaries. However, removal of this motion makes calculation of diffusion constants and other properties from the trajectory difficult unless information is retained about the removed motion.

To remedy this problem in the simulations reported, a supplementary program was written to "image" back the trajectory to the canonical form. Since residues, strands and/or various subparts of the molecule may have moved out of the box, it is not clear how to perform the imaging in a general fashion. To look at a movie, typically one wants the solute at the center of the simulation cell and the "box" of water shown around this. Therefore, the simplest way to image is to move the solute to the center of the box and then "image" the counterions or water; in other words, if a water or counterion molecule is outside the box centered on the solute, the water or counterion is translated an integral number of boxes to bring it back inside the box. This works well if there is only a single solute. However, in the case of duplex nucleic acids and a scenario as presented in Figure 12b, if the center of mass of the duplex is placed in the center of the box, both strands will be within the box and hence will not image. Moreover, the strands will still be separated. To fix this case, the first strand is centered and imaging applied. This brings the two strands together. Then both strands are centered and imaging takes place again.

Throughout the course of the simulations, various strategies were applied to remedy the contrived imaging cases as new cases appeared and broke the imaging procedure. In retrospect, the PME code should have been fixed such that this problem would not occur. However, it is not clear what the best way to do this is (except perhaps not to image at all within the PME code!). To overcome the imaging problem, I developed a utility to process trajectories that allowed flexibility in the imaging. This is discussed in more detail in Chapter 5. A static imaging procedure, as discussed in the previous paragraph, was also implemented into `carnal`.

AMBER parallelization[5]

Molecular dynamics simulations in the nanosecond range on small nucleic acids with explicit solvent are extremely time consuming, requiring significant computational resources. To allow AMBER to run on the emerging shared memory and message passing parallel machines, AMBER needed fairly general parallelization. Work on the parallelization of AMBER began with the version 3A, most notably the independent work of George Siebel and Tom Darden (NIEHS) on shared memory machines and Steve DeBolt's message passing parallelization on the nCUBE2/ hypercube (DeBolt & Kollman, 1993; DeBolt *et al.*, 1994). The first distributed parallel version of AMBER was with version 4.0 whence began a larger push with work continuing at UCSF by Steve DeBolt and later David Case on message passing versions and work by Silicon Graphics Incorporated (SGI) and myself at UCSF developing a shared memory parallel implementation for SGI multiprocessors[6]. In addition to the work at UCSF, other groups were also independently involved in parallelizing AMBER including Terry Lybrand and Eric Swanson at the University of Washington (master/slave PVM, eventually distributed with AMBER 4.0), Ken Merz and Jim Vincent at Pennsylvania State University (MPI), Tom Darden at the NIEHS, Thomas Huber at Ludwig Maximilian Universitaet in Muenchen Germany (KSR, TCGMSG), and various vendors (KSR, IBM, TMC, Cray), among others. When planning for the release of version 4.1 of AMBER, it was decided

[5] Note that a large part of this discussion is based on the Computer Physics Communication article entitled "AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules" by Pearlman *et al.* (Pearlman et al., 1995).

[6] Most of the work at shared memory parallelizing AMBER for SGI machines was performed by Roberto Gomperts (SGI) who worked closely with myself to integrate this into the general code distribution. Michael Schlenkrich (SGI) provided better scratch memory handling, parallelization of the pairlist generation and SHAKE. Marc Berger (the Chemistry Applications Specialist) and later David Zirl (who took over Marc's job) helped coordinate the collaboration and helped its development through the loan of shared memory hardware; these loaner machines provided the horse power for performing the initial equilibration, some production runs and the analysis of the results presented in this thesis.

early on that parallelism should be supported within the general code distribution for the computationally intensive `sander` (simulated annealing with NMR derived energy restraints; the standard molecular dynamics and minimization code) and `gibbs` (free energy code) modules. Given the tremendous amount of work done by various people, deciding what exactly to put into the version was not an easy task and led to difficult decisions regarding exactly what level of parallelism to support, how to make the version portable and clean, and how to choose and merge the excellent work contributed by the various groups. Initially it was decided to support both a shared memory parallel implementation for SGI multiprocessors and a general message passing version. The goal was to engineer a version of `sander` and `gibbs` that had minimal code modifications, complete integration into the standard source code release, and moderately scalability on a variety of hardware.

shared memory implementation

Version 4.1 of `sander` and `gibbs` supports shared memory parallelization on SGI multiprocessors for most of the time consuming parts of a typical molecular dynamics run. Specifically, this involves parallelization of the nonbond forces and energies, the creation of the residue based pairlist, SHAKE (Ryckaert *et al.*, 1977), and also the bonds, angles and dihedrals. This work was performed by Roberto Gomperts (Boston) and Michael Schlenkrich (Switzerland) of SGI based on version 4.0 and modified and integrated into the version 4.0 and 4.1 distributions by myself.

Generally, parallelism is supported through the use of `C$DOACROSS` comments in the source code which instruct the compiler to insert code to spawn multiple threads to run the next `DO` loop in parallel. In AMBER, an extra `DO iproc=1, numproc` loop is added around calls to various time consuming routines so that each of the `numproc` threads does one iteration of the loop. This provides a general and very simple means of implementing parallelism. The only difficulties are in figuring out for the parallel threads what variables

need to be local or shared, making sure that each thread has its own scratch arrays, load balancing the time spent in each parallel thread, and accumulating the results at the end of the loop. After execution in each parallel region, the results are accumulated into the standard AMBER data structures, implying that no special handling is needed for regions that have yet to be parallelized. The first thread uses the standard AMBER scratch arrays, whereas the other threads utilize a portion of integer and real memory allocated once, as needed. This is facilitated through the use of (non standard) FORTRAN pointers, supported in the SGI FORTRAN-77 release. The FORTRAN pointers associate the address of a variable with the variable name, thereby allowing routines such as `malloc()` to associate space with the variable. In the implementation, the pointers and current sizes are stored in a common block. This method of handling the space allocation is preferred over the standard `C$DOACCROSS` method of declaring each scratch array `LOCAL`, since it prevents needless reallocation of the space on subsequent `C$DOACROSS` calls in different scratch memory array contexts or in other words, one integer and one real scratch array can be allocated to fulfill all the scratch space needs. All of the SGI shared memory code is wrapped by the CPP define `SGI_MP` or `SHARED_MEMORY`. This version was ported by engineers at Cray Research (Jeyapandian Kottalam & Mike Page) to be compatible with Cray shared memory hardware. Overall this provides an efficient implementation with fairly minimal code modification and should be readily portable to other shared memory computers. As the code matures and prepares for version 4.2, it is worthwhile to question the utility of supporting two parallel versions. Despite the minimal footprint in the code, supporting two versions does significantly add to the "mess" in the code without adding much functionality. Moreover, vendor optimized MPI libraries are becoming available and it is not clear if the shared memory code in AMBER offers any performance advantage over the message passing version. In particular, an early version of 4.1 with a public domain MPI implementation (an early unoptimized version of mpich) outperformed the shared memory code on 4 processors of an R4400 SGI Challenge machine. However, until the

MPI version is made fully functional (*i.e.* at least supports polarization), the shared memory version should be retained. Moreover, a hybrid version with more local parallel (such as the shared memory version) support and message passing for less local communication may be ideal for emerging architectures, such as connected shared memory machines (*i.e.* SGI Origin 2000 or networks of Origin 200's) or networks of 2 or 4 processor Intel P6 processors. With respect to this, perhaps it is time to rebuild AMBER using a language that is more inherently parallelizable, such as High Performance Fortran or Fortran-90, or alternatively rebuild the data structures and memory handling to better support parallelization.

message passing implementation

After careful review of the available message passing implementations in version 4.0, the developers decided to integrate the Message Passing Interface Standard 1.0 (MPI) version developed independently under version 4.0 (`minmd`), and ported to version 4.1 (`sander` and `gibbs`), by Jim Vincent and Ken Merz from the Department of Chemistry at Pennsylvania State University. The primary reason this version was chosen was that the implementation was very elegant, the emerging MPI standard very portable, it was a SPMD (single program multiple data) implementation, and the code modifications required were actually very small. Integration and modification of the original AMBER/MPI work was performed in an active collaboration among Jim Vincent, myself and David Case[7].

The implementation in `sander` and `gibbs` provides for parallelization of the nonbond forces and energies, the creation of the residue-based pairlist, and the bond,

[7] In addition to the group coordinating the integration and optimization, special mention is also appropriate for the following people: (1) Michael Crowley and David Deerfield from the Pittsburgh Supercomputing Center for fixing bugs in the portable namelist for the Cray T3D, access to the T3D resources and for continuing a collaboration to port the particle mesh Ewald. (2) Jeyapandian Kottalam (Cray), Asiri Nanayakkara (PSC), and Ravi (PSC, Cray) for optimization. (3) Thomas Huber (Ludwig Maximilian Universitaet, Muenchen) for evaluation of the MPI and addition of TCGMSG calls (4) Steve Chen (IBM) for development of an IBM SP2 version similar in spirit to the current MPI version and (5) Bill Ross for help communicating the issues to the various parties.

angles and dihedrals (force, energy). For a more thorough discussion of the parallelization, see the discussion by Vincent & Merz (1995). The MPI calls are fully compliant with the FORTRAN interface of MPI 1.0. In order to provide portability to systems where MPI currently is not supported or is less efficient that native libraries, the MPI calls were implemented as subroutine wrappers of the MPI calls to PVM version 3.2 and higher, TCGMSG, Cray T3D shmem and recently Cray T3E shmem (the latter work by Michael Crowley, PSC), IBM SP1 and SP2, the Intel Paragon and the Convex Exemplar and can be easily adapted to run on any message passing computer.

The MPI version is very similar to native AMBER and all of the code is CPP wrapped with the CPP define `MPI`. One node is chosen as the master node and performs the initial startup code and all the I/O. After startup is complete, the master broadcasts (`MPI_BCAST`) the necessary common blocks, including all the necessary control data and initial coordinates, to each of the nodes. Within AMBER, most of the time consuming routines are driven by loops over residues. Load balancing is accomplished in general by having each node work on a subset of residues. Therefore, each node creates a portion of the total pairlist, calculates the nonbond energies and forces for this pairlist, and does a subset of the bonds, angles and dihedrals; then the partial forces, energies, and virial are combined on each node (`MPI_ALLREDUCE`). Each node then performs an update of the coordinates. For routines that are not parallelized yet, each node performs redundant work. The only communications necessary beyond the initial broadcast and combining of the forces, energies, and virial are for dynamic load balancing of the non-bond pairlist and for the routine which implements the fast 3-point water analytical SHAKE. The polarization code is not yet parallelized and therefore not supported in the `sander` or `gibbs` version compiled for MPI since these routines expect a complete copy of the pairlist on each node, rather than the distributed pairlist implemented in the AMBER/MPI version. Overall the implementation seems reasonably scaleable and eminently portable.

In the original version, load balancing of the pairlist is performed by manually redistributing the nonbonded pairs in a "brute force" fashion to balance the work done by each processor (as discussed in the article by Vincent and Merz (1995)). This requires an all-to-all broadcast to distribute the count of the nonbonded pairs from each processor to all the others. Then explicit communication (send/receive) is done to send pairs from the processor with the most pairs to the processor with the least, the processor with the second most pairs to the processor with the second least pairs, and so on. The leads to a fair bit of communication, however completely balances the pairlist. On machines with very fast communication, such as the Cray T3D and T3E, this does not incur a significant computational cost. However, on machines with slow communication, such as networks of workstations running PVM or the IBM SP-2, it is wise to minimize the communication.

The need for load balancing the list of pair interactions is because of the simple residue based distribution scheme. When the pairlist is first created, the distribution of residues is done in a modulo fashion, *i.e.* processor one gets residues $1 + i * npes$ and processor two gets residues $2 + i * npes$, *etc.* where *npes* is the number of processors and i is an integer from zero up until all the residues have been specified. This is done since for a given residue j, pair interactions only need to be calculated with residues $j+1$ and greater; therefore the bulk of the pair interactions are concentrated with the lower atom numbers. Because of the monotonically decreasing set of interactions between the pairs, this simple distribution scheme is not optimal and hence load balancing can be applied to redistribute the pairs. However, since the list of pairs interacting each time the pairlist is updated does not change much on subsequent updates, information from the previous pairlist update can be used to guide the distribution on a subsequent update. This has been implemented into an undistributed version of `sander` where the count of the pairs per residue is saved from previous pairlist updates and used to guide the distribution on subsequent steps. This reduces the communication cost.

Additional optimizations to the code included using tabled square root functionality on the Cray T3D (supplied by engineers at Cray Research, Inc.), cache-based optimizations and replacement of divides in the code by multiplications, wherever possible. All of these changes have helped to optimize the code, however more work needs to be done to support this goal, especially optimization for cache performance. Optimization for cache performance is very tricky however since small changes can have a profound effect on runtime performance, such as moving a variable in a common block and adding/removing code statements which make this process tedious. Moreover, changes appropriate for one machine may not be appropriate for another. Although considerable effort was spent in an attempt to optimize for the cache, little performance gain was realized. This optimization would be facilitated by a restructuring of the code to allow more flexible layout of memory to avoid cache contention. More flexible layout could be facilitated by allowing the ability to interchange the positions of arrays (in the large integer and real memory heaps) and provide padding of the array sizes to better match cache line sizes.

history of nucleic acid simulations

simulations of nucleic acids *in vacuo*

The earliest molecular dynamics simulations involving nucleic acids were published independently by Levitt (Levitt, 1983) and Karplus (Tidor *et al.*, 1983) who studied small double stranded nucleic acids *in vacuo*. Although Levitt's treatment completely ignored the electrostatic interactions, the DNA molecule remained intact and appeared quite dynamic, displaying large scale bending and twisting. In fact, an early indicator of the importance of electrostatics came from these simulations, since when the full charges were utilized, the DNA fell apart. The treatment by Tidor *et al.* (1983) applied a distance dependent dielectric to mimic solvent screening and lowered the charges on the phosphate ions to mimic counterion screening. They compared the dynamics expected based on harmonic (normal mode) dynamics to MD and showed the trend that B-form structures are more flexible than

Z-form DNA structures. Moreover, a nice discussion of how the molecular dynamics results can be compared to experimental results was also presented. These short calculations gave promise to the field and suggested that simulations of this type may give useful insight not only into DNA structure, but dynamics as well. In fact, these early results from molecular mechanics and molecular dynamics simulations suggested that nucleic acid duplexes are not represented by a unique structure but really an ensemble of similar structures. However, as pointed out by Poncin, Hartmann & Lavery (Poncin *et al.*, 1992) the energy calculations are not able to distinguish whether the nucleic acid duplex in solutions better represented by an equilibrium of conformational substates or if one dominant conformation is preferred. As more and more evidence accumulates, it appears that on a nanosecond time scale, the nucleic acid duplex is likely best represented as an ensemble of conformations. These early calculations (Olson, 1982a; Olson & Sussman, 1982) and analysis of the crystal structures (Sundaralingam & Westhof, 1981) also suggested that the sugar pucker is central in determining nucleic acid structure and flexibility.

After the first molecular dynamics simulations, there was a small flood of publications regarding the simulation of nucleic acids *in vacuo* which also included a few brief forays into simulation in explicit solvent. For a more thorough description of the early history, see the reviews by Beveridge *et al.* (Beveridge & Ravishanker, 1994; Beveridge *et al.*, 1993). The basic lesson learned is that the details of the calculation are very important. Not only is there a profound force field dependence– which remains true to this day– but there are definite issues with respect to the treatment of the electrostatic interactions, solvent, and counterions. From these simulations, it became clear that some representation of solvent screening was necessary. In these simulations, solvent screening was typically handled by an effective dielectric function, such as the distance dependent or sigmoidal dielectric functions discussed previously. Also apparent from these simulations

was the need to add some screening of the phosphate charge, either through the use of modified charges (as in the simulations of Levitt and Karplus discussed above) or through the use of explicit, net neutralizing, counter-ions. When using explicit counterions within *in vacuo* simulations, it is crucial to utilize softer and larger counter-ions which effectively represent the effect of the ion water solvation shell (Singh *et al.*, 1985), rather than the small and hard counter-ions used with explicit water. This is a lesson I learned early on in my graduate student career when I used the later type of counterion in DNA simulations without explicit solvent and saw that the counterion remained "locked" to the phosphate throughout the simulation. Interestingly, during the development of the calculations presented in Chapter 4, this "locking" of the ions to the phosphates was seen again in the simulation of DNA and Na^+ ions in pure ethanol. The lower dielectric of ethanol was not enough to solvate and screen the counterion charges causing them to remain locked to the phosphates (in their initial positions). This observation led me to start mixed solvent simulations with "pre-hydrated" ions and DNA. This avoids the problem with the ions since they are solvated. Moreover, pre-hydrating the DNA avoids the difficulty of hydrating the DNA during equilibration in a mixed ethanol/water box which is an extreme example of a conformational sampling problem. The DNA in simulations in ethanol are discussed in Chapter 4.

Overall, the published simulations of nucleic acids *in vacuo* give insight into nucleic acid structure and still serve as the method of choice for refining the structure of nucleic acids subject to experimental restraints, such as the X-PLOR refinement of crystal structures or the simulated annealing refinement of NMR structures[8]. However, on their own, the *in vacuo* simulations methods do not give much insight into the environmentally dependent structure of nucleic acids. However, they have given useful insight to a variety

[8] Interestingly, a recent paper demonstrates that inclusion of solvent in NMR structure refinment may be important; Leijon *et* al. (Leijon et al., 1995) show that a different structure is obtained for d(GCAAAAACG) depending on whether the restrained dynamics are run with explicit water or not.

of structural and dynamic issues. Prior to the first molecular dynamics simulations, a variety of groups were applying theoretical methods to better understand nucleic acid structure using tools ranging from simple "mechanical" treatments investigating the geometry (Calladine, 1982), to more detailed internal coordinate treatments (Tung & Harvey, 1984; Ulyanov & Zhurkin, 1984; Zhurkin *et al.*, 1982; Zhurkin *et al.*, 1980), to the atomic level molecular mechanics treatments adopted by our group. A review of some of the early theoretical treatments is presented by Olson (1982b). Zhurkin and co-workers, using an internal coordinate molecular mechanics treatment with a slightly screened dielectric (ε=4) (Zhurkin *et al.*, 1980), extensively studied the conformational dependence and flexibility of DNA duplexes. These studies clearly show the dependence of the helical twist both on the sugar pucker and the effective charge on the phosphates (Zhurkin *et al.*, 1982). As the sugar pucker decreases from 162° to 108°, the helical twist decreases from 40° to 32° and when the phosphates are less neutralized (from 100% to 50% neutralization) the helix unwinds by ~2°. In the limit of no counterions, the helix is likely unstable and expected to have a twist of ~30°. The helical twist also relates to the groove width since a wider groove (such as in A-DNA) implies a less wound helix. Further simulations (Gorin *et al.*, 1990) demonstrate multiple minima in the pyrimidine sugar pucker (with a minima appearing in the O1'-*endo* range) and higher average pucker value. These issues are of relevance to the "converged" average structures calculated from the simulations presented in this thesis. We observed sugar pucker values consistent with this analysis but also observed a lower than expected helical twist. The low twist lead us to perform a series of calculations where the charge on the phosphate was changed, barriers to sugar repuckering were changed or puckers were held fixed as are discussed briefly in Chapter 4.

The studies by Zhurkin and co-workers also investigated flexibility where fluctuations in the helical twist in the 2.4° to 3.0° range with less fluctuations in poly(dA)-poly(dT) than poly(dG)-poly(dC) were estimated. These fluctuations were slightly lower

than what is expected experimentally and significantly lower than what we observed in simulations, as is discussed in much greater detail in Chapter 2. Additionally, they show that B-DNA more easily bends into the grooves and in perpendicular directions with purine-pyrimidine dimers bending more easily into the minor groove and pyrimidine-purine more easily into the major groove where an equilibrium bend of ~6-12° is observed. This is consistent with an analysis of NMR structures performed recently where TpG steps are seen to bend into the major groove (Ulyanov & James, 1995) as is observed in the simulations presented in Chapter 2. The nucleic acid duplexes are surprisingly accommodating for bends in the helix axis; the helical axis can bend by 20° yet dihedrals vary by no more than 15° (Ulyanov & Zhurkin, 1984).

Molecular mechanical studies with an all-atom potential energy representation also suggested the relationships between sugar pucker and twisting MM (Kollman *et al.*, 1982) and also were useful to analyze the sequence dependence of stacking energies in canonical A-DNA and B-DNA models (Tilton *et al.*, 1983). These studies also helped to perhaps understand and resolve the paradox between the "rigid" DNA seen in hydrodynamic studies to the "flexible" DNA seen spectroscopically by suggesting that the backbone is inherently flexible and the bases relatively fixed (Keepers *et al.*, 1982). These studies were performed by modifying dihedral angles and moving bases in an attempt to mimic base pair opening. Surprisingly the backbone angles could adopt changes in angles (among g^+, t, g^-) in B-DNA models without a large energy penalty. This suggested that the backbone undergoes rapid motion between local torsional minima on a nanosecond time scale; this is confirmed in the more recent molecular dynamics simulations as is discussed in more detail in Chapter 2. Subsequent studies explain why Hoogstein base pairing is not favored in isolated A+T-rich DNA and how intercalators change this picture (Singh *et al.*, 1986) and suggested that right handed DNA was more stable than left handed on the basis of energy calculations (Pattabiraman *et al.*, 1987). All of these were performed applying a distance

dependent dielectric constant. In fact, the lack of a representation of solvent and complete salt effects makes the simulation suggesting that right handed duplexes are more stable than left handed duplexes incomplete since stable left handed helices are observed in Z-DNA at high salt. This is explained rather nicely through recent non-linear Poisson Boltzmann calculations. Although the more closely spaced phosphates destabilize Z-DNA compared to B-DNA, water and salt stabilize Z-DNA such that the total electrostatic free energy favors a B-DNA to Z-DNA transition in aqueous solvent (Misra & Honig, 1996). The fact that B-DNA is stable at low to medium salt suggests that non-electrostatic forces must be responsible for the relative stability of B-DNA in solution and that conformational entropy (*i.e.* the flexibility of B-DNA compared to Z-DNA) helps favor B-DNA.

As mentioned previously, the molecular dynamics studies without explicit solvent did not demonstrate tremendous stability. With the Weiner *et al.* force field, simulations on B-DNA lead to a structure intermediate between A- and B-form DNA whereas with the force field of Nilsson & Karplus (1984) a relatively rigid DNA and no sugar repuckering was seen, and the stable DNA had only 9.1 base pairs per turn (Beveridge *et al.*, 1991). As longer simulations became possible, it became clear that some representation of solvent was critical and many Monte Carlo and molecular dynamics studies on solvated nucleic acids were published. However, before going into these simulations, it should be mentioned that the various molecular mechanical treatments employing some continuum electrostatic method to represent solvent have and continue to give very useful insights. JUMNA calculations (Lavery *et al.*, 1995) by Lavery and co-workers which use a sigmoidal dielectric representation suggest that a B-DNA duplex is characterized by a set of conformational "substates" (Lavery & Hartmann, 1994; Poncin *et al.*, 1992). More recently, calculations of this type were able to reproduce the strong curvature in $(A_4T_4NN)_n$ sequences compared to $(T_4A_4NN)_n$ sequences and the proper trends in the minor groove widths (Sanghani *et al.*, 1996); this is something we are attempting to do with molecular

dynamics calculations in explicit solvent. An even better continuum representation of solvent is obtained from Poisson-Boltzmann electrostatic treatments. In addition to a better understanding of the relative stability of B-DNA and Z-DNA (Misra & Honig, 1996) and stacking (Friedman & Honig, 1995) as discussed previously, the Poisson-Boltzmann calculations have helped explain DNA-ligand interactions, shown the correct preference for binding mode and interestingly suggested that it is largely non-polar interactions that stabilize ligand binding (Misra & Honig, 1995; Misra *et al.*, 1994; Zakrzewska *et al.*, 1996). In spite of this, the methods lack a molecular level picture of the hydration since water is not explicitly included in the calculations. Recently this has been shown to be an issue in simulations of DNA triplexes where it was demonstrated that a distance dependent continuum solvent model and a Poisson-Boltzmann model predicted an antiparallel third strand with reverse-Hoogsteen base pairing to be more stable than a parallel third strand with Hoogsteen base pairing in d[CG-G]$_7$ triplexes which is opposite to the trend seen in simulations with explicit solvent (Cheng & Pettitt, 1992; Cheng & Pettitt, 1995). This implies that a minimal representation of the explicit water (perhaps only the first two hydration cells with a continuum treatment outside) is necessary. In spite of the expected need for water, short molecular dynamics simulations with explicit water do not automatically imply better results, as was seen in simulations by Fritsch *et al.* (Fritsch *et al.*, 1993) where dynamics with a distance dependent dielectric treatment agreed better with the experimental data. Part of the reason for this is that on a longer time scale, the force fields and methods were unable to reasonably represent the structure; for this reason and until more recently, Monte Carlo simulations with fixed nucleic acids arguably gave better insight into hydration than the molecular dynamics simulations. Adding explicit water brings additional complexity to the simulation. As mentioned previously, issues include what water model to utilize, how much water to add, what type of boundary to impose, how to make the calculation more tractable, and whether the nucleic acid force field

employed is balanced with respect to the water model. In the next section some results from the early and more recent simulations of nucleic acids in explicit solvent are presented.

simulations of nucleic acids in explicit solvent

The first reported molecular dynamics simulation of a nucleic acid in explicit water was reported in 1985 where a nucleic acid duplex, d[CGCGA], was surrounded by 8 Na^+ counterions and a spherical shell of 806 TIP3P water molecules (Siebel *et al.*, 1985). This simulation, run for a reported ~100 ps, required ~20 hours on a Cray X-MP 4/8; one of the larger supercomputers of the time. Compared to the *in vacuo* simulations done previously, there were not too many significant differences when solvent was added into the simulation except there was more twist and tilt in the central base pairs and a number of deviations from the classic C2'-*endo* sugar pucker conformation. The simulation did however give insight into the hydration, however extensive analysis of the water structure was not performed. Later molecular dynamics simulations (with a droplet of water) by Chuprima *et al.* (1991) did more extensively analyze this and conclude that the hydration of the minor groove is largely dependent on the width; however these simulations were very short (~25 ps). In the calculations of Seibel *et al.*, although the structure was more "B-like" than the *in vacuo* simulations of Levitt (1983), 100 ps was not sufficiently long enough for the instabilities due to the cutoff of the electrostatic interactions to become apparent. Also, by the end of the simulation the water droplet distorts to such a degree that one end of the helix is nearly exposed to the "vacuum." Of course, any distortions of the DNA could be prevented by holding the conformation of the DNA fixed. This is what was done in Monte Carlo simulations aimed at investigating the hydration of DNA which were published prior to the first molecular dynamics of DNA *in vacuo* in the detailed and often hard to follow papers of Clementi and coworkers (Clementi & Corongiu, 1981). To reduce the computational burden, small repeating units (such as base pairs or dinucleotide base pairs) were simulated with helical periodic boundary conditions. Later, entire duplex structures

were solvated and subject to hexagonal periodic boundary conditions. A series of Monte Carlo simulations by Beveridge and co-workers investigated the crystal structure of d[CGCGAATTCGCG] with 1777 waters in a hexagonal cubic cell under periodic boundary conditions (Subramanian & Beveridge, 1989; Subramanian *et al.*, 1988; Subramanian *et al.*, 1990b). These simulations found 10.4 waters per nucleotide comprising the first coordination sphere which was surrounded by a second hydration sphere of waters primarily in the minor and major groove which brought the hydration level to 17.4 water per nucleotide. This is close to the 20 water per nucleotide found in the early experiments (Falk *et al.*, 1963). Both GC and AT regions can support ordered water structure and the "spine of hydration" seen in crystal structures (Drew & Dickerson, 1981) was reproduced. Additionally, the expected "cone of hydration" or triad of waters around the phosphates (Pullman & Pullman, 1975) was found. Results on other systems also demonstrate reasonable agreement with the crystal structures (Eisenhaber *et al.*, 1990a; Subramanian & Beveridge, 1993; Subramanian *et al.*, 1990a) and demonstrate sequence specific hydration patterns (Eisenhaber *et al.*, 1990b; Eisenhaber *et al.*, 1989). As such, these models reasonably represent the hydration, however they completely neglect the dynamic flexibility seen in the nucleic acid structure. Molecular dynamics simulations, with the nucleic acid held fixed, also gave insight to the hydration and counterion association and moreover solvent molecule mobility's (Forester & McDonald, 1991). These simulations of nucleic acids were some of the first to include an Ewald treatment of the electrostatics.

After the publication of the solvated DNA simulation by Seibel *et al.*, there was a small flurry of related publications, most of which applied periodic boundary conditions and used standard models which neglected the long ranged electrostatic interactions (Auffinger *et al.*, 1995; Laughton *et al.*, 1995; Laughton & Neidle, 1992; McConnell *et al.*, 1994; Miaskiewicz *et al.*, 1993; Swaminathan *et al.*, 1991; van Gunsteren *et al.*, 1986). In

these simulations, as the length of the simulation got longer and longer distortions in the structure became more apparent. This is nicely shown in the simulations of McConnell *et al.* (1994) where during an ~1 ns simulation, the DNA moves more than 7 Å from the starting structure. Better results came from simulations which applied an Ewald treatment (Cheatham *et al.*, 1995; Weerasinghe *et al.*, 1995a; Yang *et al.*, 1995; York *et al.*, 1995; Zichi, 1995). Most of the simulations applying an Ewald treatment have applied constant volume conditions (most likely because of the difficulty implementing constant pressure with Ewald methods) and in the case of the simulations by Darden and coworkers, crystal boundary conditions where the unit cell as determined crystallographically is simulated. Constant volume as well as crystal packing may inhibit fluctuations.

This section on the simulation of nucleic acids both *in vacuo* and in solution has not been exhaustive; a few examples from the literature were presented to give a portrayal of the strengths and weaknesses of the methods and to serve as introduction to the simulations presented in this thesis. Moreover it should be noted that this discussion only focused on simulations of nucleic acids; simulations on proteins both *in vacuo* and with explicit solvent pre-date similar simulations on nucleic acids. Some of what was reviewed here will be reviewed again in the respective chapters. For more information, see reviews by Beveridge and coworkers (Beveridge & Ravishanker, 1994; Beveridge *et al.*, 1991; Beveridge *et al.*, 1993) and the recent review by Louise-May *et al.* (Louise-May *et al.*, 1996). To briefly summarize, a constant issue in all these simulations is the treatment of the electrostatic interactions. Additionally, issues arise as to the difference between "crystal" boundary conditions and simulations in "solution"; this is a particularly important issue in the simulation of nucleic acids since the structure is profoundly influenced by the environment and in some cases altered by crystal packing. Before the simulations in Chapter 1 were published, some of the longest simulations of nucleic acids applying methods with a cutoff lead to poor results (McConnell *et al.*, 1994), while those employing a proper treatment of

the long ranged electrostatics lead to more reasonable results (Weerasinghe *et al.*, 1995a; York *et al.*, 1995; Zichi, 1995).

Stable simulation of nucleic acids in solution

Initial successes...

After the summer of 1994, I worked with Tom Darden to integrate his particle mesh Ewald code into version 4.1 of AMBER and to ready the code for general release within AMBER. Michael Crowley joined the collaboration and turned his expertise to not only help make standard AMBER to run on the Cray T3D, but started his tremendous work aimed at getting a fairly general parallel 3D FFT running on the T3D which was integral to the particle mesh Ewald code (in the absence of a vendor supplied 3D FFT library which did not arrive until later). We were aided tremendously by a collaboration with Jim Vincent and Ken Merz of Pennsylvania State University which provided the base MPI version of AMBER 4.1 from which the parallel particle mesh Ewald version emanated. My role was an essential, but certainly not glorious, one serving as an integrator and tester. In fact, the strong credit really should go to Vincent for the MPI parallel version, Darden for the PME code and Crowley for parallelizing the PME code for message passing architectures. Dave Case was also involved optimizing the parallel code, as were a number of other AMBER researchers, including Thomas Huber, *etc*., all as were mentioned in a previous section. During the integration of the PME code into 4.1, we encountered a few difficulties as were mentioned in the introduction. Specifically, the "flying block of ice" and difficulties overcoming imaging problems and model building DNA. Particularly annoying was the "flying block of ice" problem since at first, Darden and I thought the problem was solved by simply zeroing the net force after each step. This occured while running some of the simulations presented in this chapter, which later had to be re-run or continued from the point where the temperature drop became problematic. Unfortunately, the problem developed again, particularly in longer trajectories (> 1 ns), due to a build up of center of mass motion as was discussed in the introduction. This is discussed in more detail at the end of this chapter. A drawback of having very fast computers at your disposal is that

trajectories that previously used to take months now take a week to days. Given multiple machines to run on, a tremendous amount of data can be generated in a relatively short time. The bottleneck in the simulation becomes not running the simulation, but archiving the trajectories, analyzing the results, and developing the software to analyze the results. Therefore, sometimes trajectories were not looked at as diligently as they should have been, and the "flying block of ice" crept in and turned a number of supposedly long molecular dynamics simulations into elaborate minimizations (*i.e.* a glorious temperature quenching protocol).

While all this development work was going on, we were itching to try out the methods. As soon as a reasonably stable version appeared, a number of us took to investigating not only the PME method and it applicability, but attempted to evaluate the (at the time) soon to be released Cornell *et al.* (1995) force field. This involved work by Jennifer Miller on an NMR-derived r(UUCG) RNA hairpin loop and stem structure, work by Thomas Fox on the crystal structure of ubiquitin and myself investigating the DNA crystal structure of the duplex d[CCAACGTTGG]$_2$. The results from these simulations were encouraging and led to the publication of the following communication in the *Journal of the American Chemical Society* (Cheatham *et al.*, 1995). Although Darden and co-workers had earlier investigated a number of systems with the PME method (York *et al.*, 1993a; York *et al.*, 1994; York *et al.*, 1995), all of these were crystal simulations employing constant volume and packed unit cell conditions. A potential criticism is that the packed unit cells and constant volume may inhibit fluctuations and be characterized by high pressures. We were more interested in "solution" conditions, and therefore constant pressure simulations in the absence of crystal packing. Few simulations employing Ewald treatments on biomolecules had appeared and to our knowledge, none on nucleic acids in "solution". Around the same time as our communication appeared, a number of articles applying an Ewald treatment to nucleic acids appeared. Specifically, Zichi simulated the

RNA hairpin loop GGGCGCAAGCUU (Zichi, 1995) and Pettitt and coworkers a model triple helix (Weerasinghe *et al.*, 1995a), although both of these were constant volume simulations.

In my simulations, I always applied constant pressure since this allows the box to relax to an equilibrium volume, density and pressure. Not only does this provide an extra check on the "balance" in the force field, but a worry with constant volume simulations is that the constraint on volume could in principle inhibit conformational transitions that may increase the exposed volume of the macromolecule. A clear example is tightly a packed crystal which cannot be expected to move too much in a constant volume simulation, hence tend to give low RMSd values to the starting structures. In fact, as a test of this hypothesis, I have always wanted immerse a crystal unit cell into a larger volume of water and then run dynamics to see if crystal packing is lost and crystal packing defects are ameliorated. Likewise, the converse, slowly turning on crystal packing during a simulation, would be useful to better understand the environmental effects on nucleic acid structure.

In a similar vein, all of my simulations were run with constant temperature with Berendsen scaling on all the atoms (Berendsen *et al.*, 1984). An issue that had crept into simulations of proteins in the past was termed the "hot solvent cold solute" problem, where during the simulations with constant temperature and due to poor coupling between the solvent and solute atoms, the kinetic energy would become disparingly partitioned between the solute and solvent degrees of freedom. To get around this problem, dual temperature coupling is applied, where the solvent and solute atoms are temperature scaled independently. A drawback of this method is that since there are generally far fewer atoms in the solute, the coupling to the temperature bath is effectively "tighter". In other words, the fluctuations in the solute temperature are effectively lowered. Given that the rate of overcoming conformational barriers is preportional to the temperature, less fluctuation in

temperature implies less excess kinetic energy available to overcome the barriers. While this can be remedied by changing the time constant for the temperature coupling to the solvent, I preferred instead to run simulations with a single coupling to the temperature bath. The origin is the hot solvent cold solute problem is likely related to the truncation of the Coloumbic interactions and therefore should not be an issue with Ewald simulations. This problem is not an issue in simulations of nucleic acids with the particle mesh Ewald, as can be seen in Figure 13 where the temperatures are plotted as a function of time for a simulation of the NMR derived structure of d[ACGTTGCCTTGCG]-d[CTCAAGGCAAGCT] (Mujeeb *et al.*, 1993) using the same protocols and simulation conditions as applied in Chapters 2 and 3.

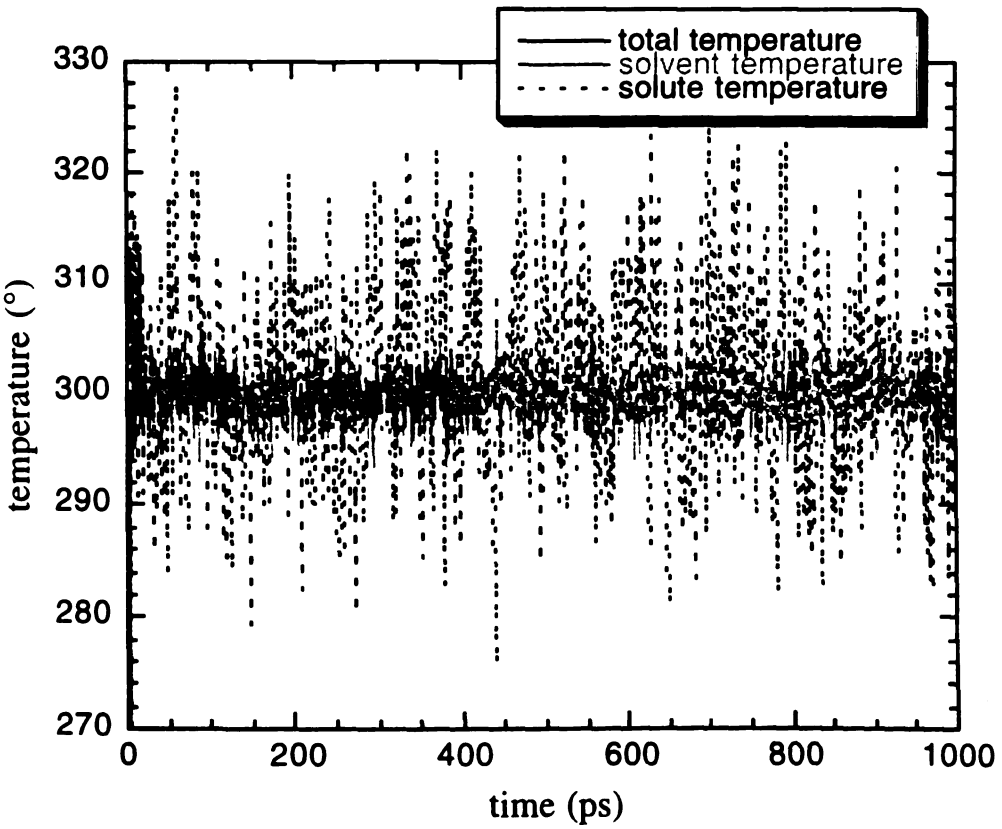


Figure 13: Temperature versus time in MD trajectories applying the PME methods and Berendsen temperature coupling

The following appeared the *Journal of the Americal Chemical Society*, **117**, pages 4193-4194 (1995).

Molecular Dynamics Simulations on Solvated Biomolecular Systems: The Particle Mesh Ewald Method Leads to Stable Trajectories of DNA, RNA and Proteins.

T. E. Cheatham, III
J. L. Miller
T. Fox
T. A. Darden [*]
and
P. A. Kollman

Department of Pharmaceutical Chemistry
University of California
San Francisco, California 94143-0446

[*] National Institute of Environmental Health Sciences
Research Triangle Park
North Carolina 27709

introduction

This communication presents results from molecular dynamics (MD) simulations with AMBER 4.1 (Pearlman *et al.*, 1995) and the Cornell *et al.* (1995) force field, of three different, fully solvated, fully charged, macromolecular structures: X-ray derived structures of d[CCAACGTTGG]$_2$ DNA (Prive *et al.*, 1991)[9] and ubiquitin (Vijay-Kumar *et al.*, 1987)[10] and an NMR derived structure of an r[UUCG] RNA hairpin loop and stem structure (Cheong *et al.*, 1990; Varani *et al.*, 1991). We compare the use of the Particle Mesh Ewald (PME) method (Darden *et al.*, 1993) for the treatment of long range electrostatic interactions to standard charge group based truncation cutoff (CUT) methods used in simulations with periodic boundary conditions.

An accurate representation of long range electrostatic interactions in MD simulations is extremely important in order to properly represent the structure, dynamics and energetics of biomolecular systems (Schreiber & Steinhauser, 1992; Steinbach & Brooks, 1994; York *et al.*, 1993b). This is particularly true for highly charged systems, such as DNA and RNA, where it has been difficult to obtain stable trajectories for a fully solvated system without imposing added restraints or artificially modifying the charges on the phosphate (McConnell *et al.*, 1994). Stable trajectories, ideally without the addition of restraining forces, are necessary for applications such as accurately predicting the binding free energy of ligand-macromolecule interactions (Miyamoto & Kollman, 1993).

The results on these three solvated biomolecular systems clearly demonstrate that with a modest computational burden, the PME method is not only generally applicable, but superior to standard cutoff methods. This is demonstrated by less RMS deviation from the

[9] PDB entry 5dnd.

[10] PDB entry 1ubq

experimentally observed structures by the PME method while maintaining reasonable atomic positional fluctuations. In the case of the DNA and RNA simulations, both CUT simulations led to distortion and breakup of the structures. The PME results presented on RNA, to our knowledge, are the first demonstration of stable solvated RNA MD trajectories.

A number of studies to date have applied MD simulations to study macromolecular crystals (York *et al.*, 1993a; York *et al.*, 1994). In these simulations, and more generally, with standard solvated periodic boundary systems (e.g. not true experimentally derived crystal unit cells), it has been found that it is critical to properly treat the electrostatics, preferably through the use of methods which account for the periodicity of the unit cell, such as PME, but minimally through the use of large cutoffs and/or methods which smooth out the potential and forces resulting from a truncated cutoff (Schreiber & Steinhauser, 1992). To our knowledge, the Ewald (Ewald, 1921) method has not been directly applied in standard solvated periodic boundary systems, except for in the work of Smith *et al.* (Smith & Pettitt, 1991) studying a small zwitterionic peptide and Schreiber & Steinhauser (1992) studying helix forming peptides. Both found that better treatment of the electrostatics was obtained using the Ewald method, in contrast to standard cutoff or switching function techniques, even though the Ewald method is known to introduce long range correlation of fluctuations (Valleau & Whittington, 1977).

In each of the macromolecular systems investigated, MD simulations were run[1] using the PME method, a charge group based truncation cutoff (CUT), and a group based

[1] DNA/RNA simulations: A rectangular box was constructed with counterions placed by the AMBER EDIT module. Simulations were run with SHAKE on hydrogens, a 2 fs timestep, a temperature of 300K with Berendsen temperature coupling, a 9 Å cutoff (also applied to Lennard-Jones interactions in PME), and constant pressure. The nonbonded pairlist was updated every 10 steps. No extra restraints were placed on the DNA/RNA systems following the equilibration period. Results were analyzed with the CARNAL module of AMBER 4.1. The PME charge grid spacing was approximately 1.0 Å and the charge grid was interpolated with a cubic B-spline. The DNA box size was 64.0 Å by 45.0 Å by 45.0 Å and the RNA box size was 47.0 Å by 44.0 Å by 39.0 Å. Equilibration was performed by placing 25 kcal/mol restraints on all solute atoms, minimizing the water for 1000 steps, followed by 3 ps of MD which allowed the water to relax around the solute. This was followed by 5 rounds of 600 step minimizations, reducing the solute restraints by 5 kcal/mol during each round. Finally,

truncation cutoff coupled with complete evaluation of all the solute-solute interactions (CUTSS). Over the course of the dynamics in each of these cases, the PME structures remained strictly closer to the experimentally observed structures [Figure 14], yet demonstrated significant positional fluctuations.

simulations on DNA: d[CCAACGTTGG)$_2$.

In the case of a 1 nanosecond PME DNA simulation, the structure remained in a strictly B-DNA form with an average RMS deviation 3.2 Å away from the crystal structure (2.9 Å away from canonical B-DNA (Arnott & Hukins, 1972)). This simulation, run with a fully charged DNA and explicit counterions, converged after roughly 200 ps to a structure which deviated from the average structure over 200 ps to 1 ns by an average RMS deviation of 1.4 Å (0.2 Å standard deviation) and did not demonstrate any persistent substates[12]. In contrast, 300 ps CUT and 300 ps CUTSS simulations demonstrated significant bending and structural distortion of the DNA, leading to a RMS deviation from the crystal structure of over 6 Å (Figure 14a). The distortion was particularly apparent in the CUTSS simulations (in both a rectangular and cubic box simulation) where the initial bending was so severe it led to breaking of the 3 terminal base pairs.

---

the system was heated to 300K over 10ps of MD and production runs initiated. Ubiquitin simulation: A rectangular box of 4385 TIP3P waters was constructed. Simulations were run with SHAKE on hydrogens, a 1.5 fs timestep, a temperature of 300K with Berendsen coupling, a 8 Å cutoff, and constant pressure. The nonbonded pairlist was updated every 10 steps. The system was equilibrated as follows: after initial minimization of the solvent box down to a RMS gradient of 0.5 kcal/mole-Å, 12 ps of dynamics on the solvent only was performed. This was followed by 6 minimizations of the whole system with the backbone atoms of the proteins restrained to the crystal positions by force constants of 1000, 100, 50, 15, 2, and 0 kcal/mol. Then two 3 ps MD at 100 K and 200K were performed, followed by production runs.

[12] This compares to the 1 ns switched cutoff simulations of McConnell *et al.* (1994) of d(CGCGAATTCGCG)$_2$, run with no explicit counterions, a reduced (to -0.24 au) phosphate charge and an extra hydrogen bonding potential to maintain Watson-Crick base pairing which converged to two structures approximately 4.5 and 7.5 Å away from canonical B-DNA.

simulations on RNA: 5'GGAC(UUCG)GUCC.

MD on the RNA system was performed using the CUT and CUTSS methods for 300 ps each, while the PME simulation was run for 1 ns. The two non-PME simulations diverged significantly (>4 Å) from the solution structure within 300 ps (Figure 14b). This deviation is primarily due to disruptions of the stem structure. As was seen in the DNA simulations, the CUTSS method demonstrated significantly worse behavior than the CUT method. The PME simulation stayed close to the solution structure throughout the entire 1 ns simulation (~1.5 Å RMS deviation). More importantly, the PME method correctly maintained the A-form stem structure. The largest deviations seen in the stem are in the closing base pair of the loop (C4:G9) where significant buckling occurred. Analysis of the loop structure showed that the major structural features, as reported by Cheong *et al.* (1990), were also maintained. The stability and accuracy of the PME method gives us confidence that these simulations can be used to examine some of the unresolved questions concerning the unique stability of this tetraloop. The results of the PME simulation will be discussed more fully elsewhere.

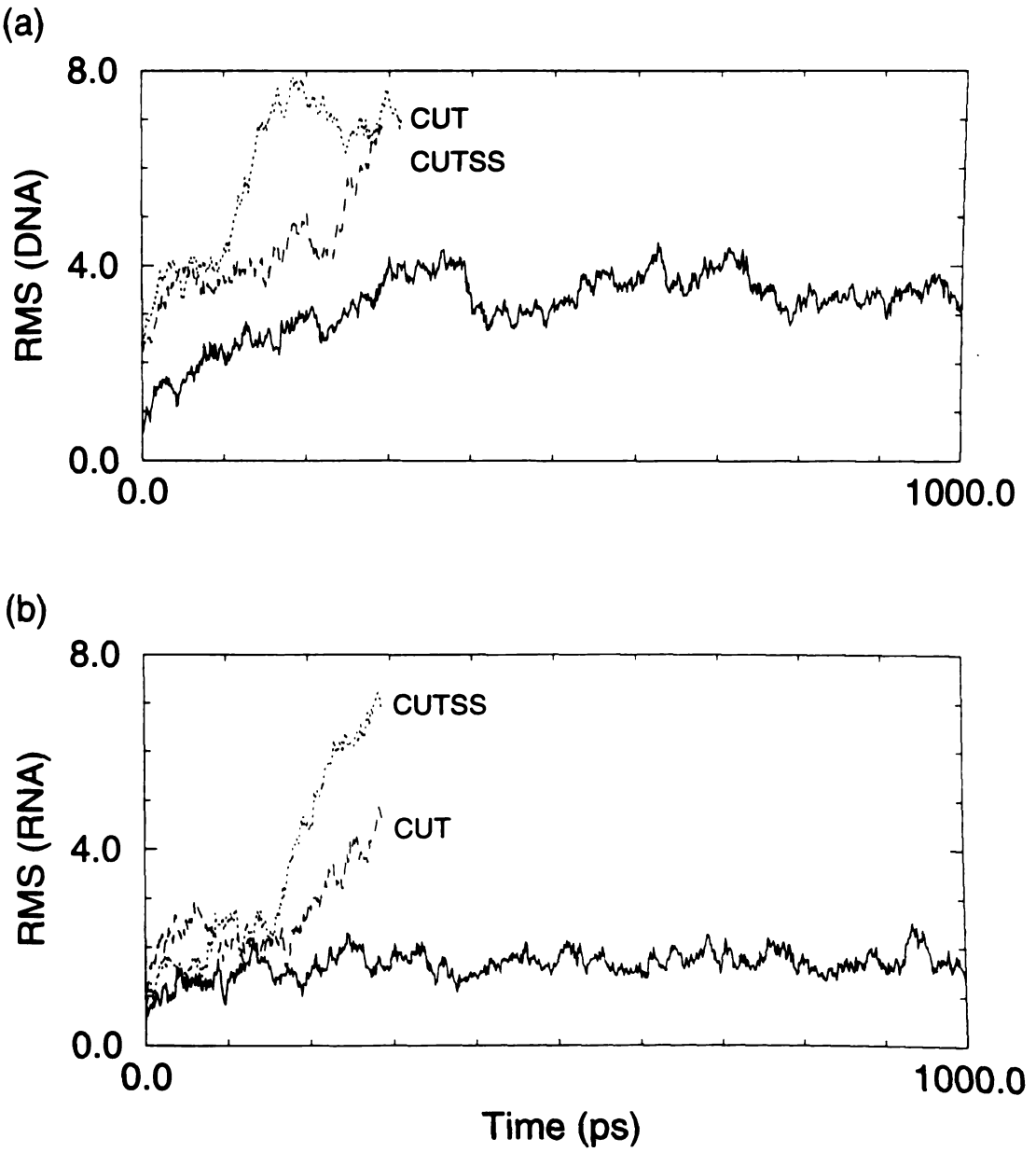


Figure 14: Time evolution of the all–atom rms deviation from the initial structure for (top) d[CCAACGTTGG]2 and (bottom) 5'GGAC(UUCG)GUCC). The solid line represents the PME, and the labeled dotted lines represent the CUT and CUTSS simulations.

simulations on the protein ubiquitin.

In contrast to the highly charged RNA and DNA systems, all the simulations of ubiquitin stay close to the initial structure. Without PME, the RMS deviation from the crystal structure rises for all non-hydrogen atoms to 1.8 Å and for the backbone atoms to 1.2 Å after 300 ps. With PME, the corresponding values are 1.2 Å and 0.9 Å. The low RMS deviation observed in the PME simulation is not due to damped motion since the calculated thermal factors are of similar magnitude to the crystal thermal factors [Figure 15]. In the solvent exposed surface residues the fluctuations are, not surprisingly, higher than that observed for the crystal (Vijay-Kumar *et al.*, 1987).

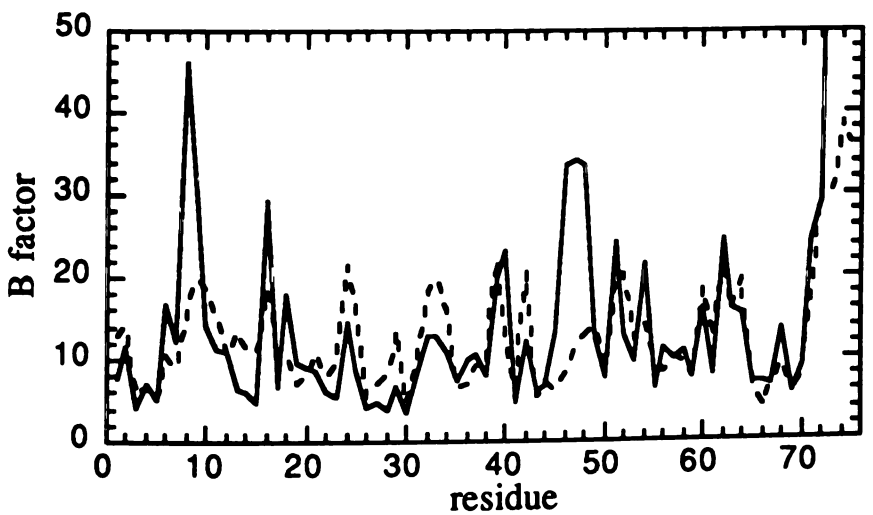


Figure 15: Experimental (dotted line) and PME trajectory derived (solid line) thermal B factors for ubiquitin, averaged over residues. The calculated values from the PME trajectory were obtained from the atomic fluctuations $\langle \Delta r^2 \rangle^{\frac{1}{2}}$ *by* $B = \frac{8\pi^2}{3}\langle \Delta r^2 \rangle$.

conclusion

In conclusion, it is demonstrated that PME is a powerful method to study MD of macromolecules in solution, particularly for highly charged systems, and superior to standard cutoff schemes with a modest computational burden (40-50% additional cost). Other studies have analyzed various approaches to electrostatic cutoffs for proteins and have concluded that "shifted" potentials seem to be the most robust (Daggett & Levitt, 1993; Steinbach & Brooks, 1994). It is especially encouraging that the PME demonstrated stable, unrestrained, nucleic acid trajectories. Although accurate agreement with a crystal structure is not expected, since we are not mimicing experimental crystal conditions, it may be significant that the NMR derived RNA structure remains closer to the initial structure than does the X-ray derived DNA structure. The low ubiquitin RMS backbone deviation is consistent with an available NMR study (Di Stefano & Wand, 1987) which suggests a hydrogen bonding pattern in solution that is consistent with the crystal structure. Furthermore, it is also consistent with prior findings on BPTI, which show a 0.8 Å backbone deviation between the crystal structures and the NMR solution structure (Berndt *et al.*, 1992), comparable to the 0.9 Å found here for the comparison between the MD and X-ray ubiquitin backbone structures. More in depth analysis of these simulation results will be published elsewhere[13].

[13] This is the end of the *J. Amer. Chem. Soc.* communication.

initial failures...

As mentioned briefly during the introduction and discussed in more detail in Chapter 5, there were a number of problems encountered when starting up these simulations. These included problems with the model building which led H1' atoms in cytosines to flip to the wrong side of the ring, difficulties imaging, and most notably the "block of ice flying through space" problem. It was thought that zeroing the net center of force at each step during the dynamics would rid the accumulation of net center of mass translational velocity. However, it was noticed that towards the end of the trajectories, the atomic positional fluctuations became significantly damped. Shown in Figure 16 are the atomic positional fluctuations over 100 picosecond intervals from a trajectory of d[CCAACGTTGG]$_2$ started in a canonical B geometry (using the protocol described in the next chapter). Along the x-axis in this plot is each atom in increasing order with the corresponding atomic positional fluctuations on the y-axis. Over the beginning part of the trajectory, the fluctuations are very similar. The highest fluctuations are seen at the termini and in the phosphate groups. At the end of the trajectory, all of the fluctuations have dropped below 0.5 Å except for the thymine methyl groups which are spinning very fast (which leads to the peaks of ~1.0 Å seen). The fluctuations have dropped since most of the kinetic energy is represented by the translation of the center of mass. The program `sander` prints out the center of mass translation kinetic energy (EKCMT) and this can be used to catch the problem at an earlier stage. Likewise, as the kinetic energy is drained from the internal degrees of freedom, the structure will tend to drop into a local minimum and the potential energy will also drop; this however occurs at a slower rate. This is equivalent to a quenched molecule dynamics run where the temperature is slowly reduced. This can be seen by taking a series of restart files from the dynamics, removing the center of mass motion, then calculating the temperature. This is shown in Figure 17 which shows the effective temperature from an effected trajectory of d[CCAACGTTGG]-r[CCAACGUUGG] in a canonical B conformation.

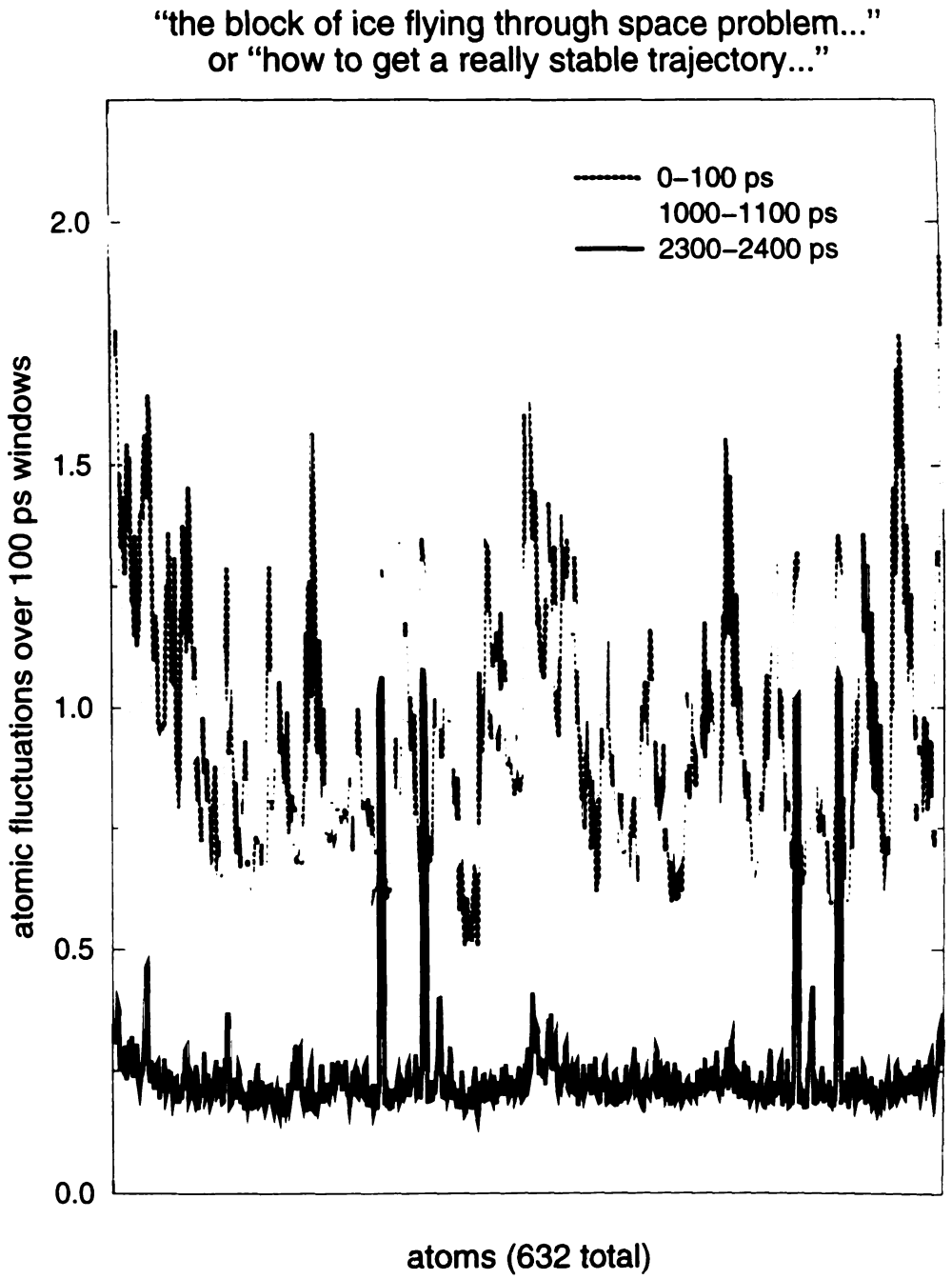


Figure 16: Atomic positional fluctuations from a trajectory of d[CCAACGTTGG]2 started in a canonical B–DNA geometry

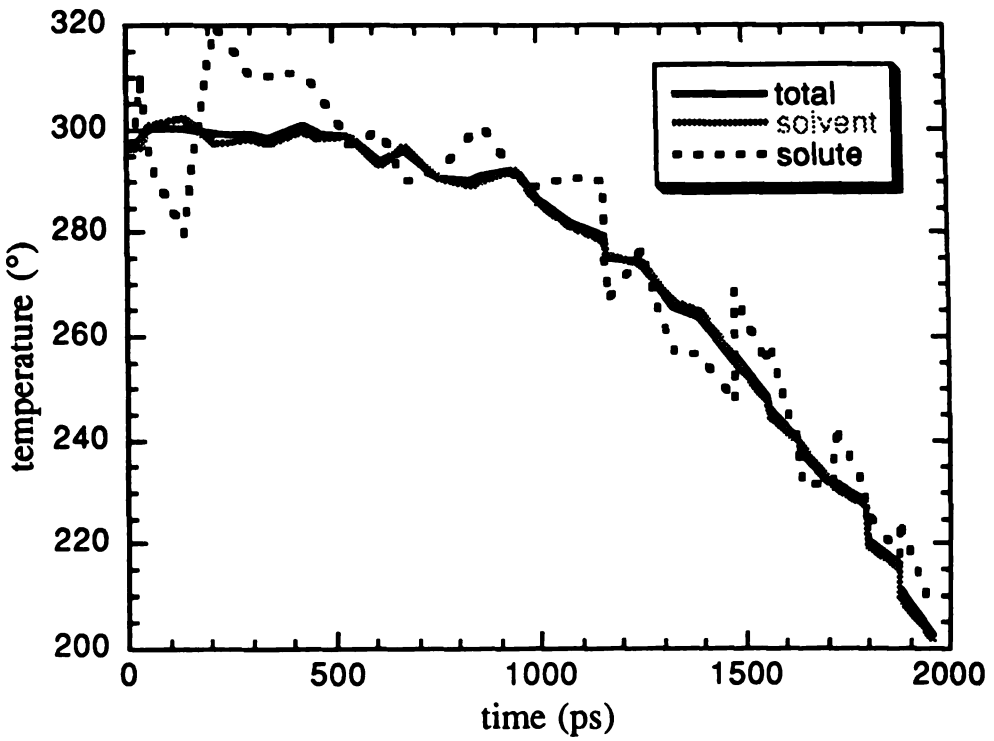


Figure 17: effective temperature in a trajectory suffering from the "flying block of ice" problem. This is the simulation of d[CCAACGTTGG]-r[CCAACGUUGG] in a canonical B conformation.

A simple way to remove the problem is to either not utilize a temperature coupling method which uniformly scales velocities (such as Anderson coupling) or to remove the center of mass motion of the entire simulation cell. In the simulations discussed in this thesis, the center of mass motion was removed each time the dynamics was restarted (roughly every 1-100 ps).

The growth in the center of mass motion is not a problem related to the use of the PME method. It is an inherent limitation of any method which controls temperature by

uniformly scaling the velocities when used in concert with methods that do not rigorously conserve energy or alternatively drain energy into low frequency modes. It also is not limited to simulation of solvated macromolecules, as it can be commonly observed in gas phase calculations and even in pure boxes of water. It was not found in earlier solvated simulations with a periodic box and a cutoff applied to the electrostatic interactions because of compensating heating effects due to truncation of the electrostatic forces and energies and random coupling from collisions of molecules moving into the cutoff sphere. These serve to hide the problem. It is highly likely that a simulation of an uncharged gas (i.e. neon) in a periodic box would succumb to the same problem. The growth of the center of mass motion is simplest to understand in terms of energy drains during the molecular dynamics. As the energy drops, the temperature drops and therefore when applying Berendsen temperature coupling the velocities of all the atoms are scaled up uniformly. This scales up the center of mass translational motion. For a molecule in the gas phase or an entire simulation cell in a periodic box, the potential energy is invariant to the translational motion. Since the periodic box or molecule in the gas phase cannot crash into other objects (*i.e.* other atoms or physical boundaries such as a wall), the kinetic energy in the translational motion of the center of mass cannot couple back into other modes. Therefore, as the temperature keeps falling, the center of mass motion keeps growing.

Where does the initial center of mass velocity originate in the PME calculations? The primary component is likely from the initial random velocity assignment. Typically the starting velocities of all the atoms are drawn from a Gaussian distribution about a given temperature which will likely lead to a finite center of mass velocity. Other sources of this are from lack of complete force conservation (which was compensated for in the PME calculations by removing the net force each step), small violations to Newton's third law (regarding equal and opposite forces) in the PME because of the analytic forces from the interpolated energy grid, perhaps drains of kinetic energy into low frequency modes and

other stochastic errors. Although all these sources of center of mass velocity are stochastic (rather than directed), any instantaneous net center of mass translation can be scaled up. Even without scaling up the motion, it is unlikely that it will disappear because of the low probability of another equal and opposite stochastic force appearing to cancel out the motion. Fortunately, these stochastic sources except for the initial velocity assignment on lead to very small center of mass translations. Therefore, the buildup of this motion is rather slow and essentially negligible on the 100 ps to 1 ns time scale as long as the energy drains are not too severe. This is not true of the initial velocity assignment which can lead to significant center of mass velocity hence it is fairly crucial to remove this, as I later learned after observing the "box of ice flying through space" in a number of trajectories.

Where do the energy drains in `sander` come from? When running the PME method with an appropriate grid size, interpolation order and direct sum tolerance the electrostatic energy is reasonably conserved, in principle. In practice however there are a few common sources of small energy drains.

- *pairlist update*: Unless the pairlist is updated every step or methods employed to build a slightly larger list with automatic updates when necessary, energy will drain.
- *SHAKE*: Unless the tolerance for the iterative procedure to fix the positions of the hydrogens is set very low (*i.e.* tolerance < ~0.0000001 Å) energy drains from the system.
- *integration time step*: In order to speed up the calculation, larger time steps are often applied. This leads to less accurate integration and a lack of energy conservation. Time steps less than 0.2 fs are likely necessary even when applying SHAKE.
- *Berendsen pressure scaling*: Scaling the positions of the molecules to perform pressure scaling leads to small violations in force conservation (equal and opposite forces), particularly if aniostropic scaling is applied. Moreover, the Berendsen pressure scaling

method leads to fairly large energy drains. When the system is at negative pressure and the box size scaled up to compensate, PdV is positive. Likewise, when the system is at pressures larger than the target pressure and the box size scaled down, PdV is also positive. This implies that whenever the pressure is negative or greater than the target presssure (P > 1 atm), energy is drained from the system during scaling of the box size. Since the fluctuations in the pressure are typically large (approximately +/-500 atm), energy is often drained[14].

In addition, any transfer of energy into a mode which cannot couple back into other internal degrees of freedom (such as drains into translational modes) can also lead to a small energy drain. Therefore it is prudent to monitor energy conservation and also center of mass kinetic energy.

In the calculations presented herein, the SHAKE tolerances were rather high (0.0005), the pairlist was not updated every step (typically updated every 10-25 steps), constant pressure was applied and the integration time step was a little too high (at 0.2 fs). Despite the small energy drains resulting from this, as long as the center of mass translational motion was removed every 10,000 to 1,000,000 time steps the center of mass motion did not grow tremendously. Temperature coupling plugs any energy drains by pumping more energy back into the system as necessary. The overall goal of these calculations was not to apply a "perfect" molecular dynamics simulation with ideal energy conservation, *etc.*, but to investigate the structure and dynamics of nucleic acids in solution and attempt to correlate the observation back to experiment. Use of larger time steps allows a longer time scale to be represented. Given the numerous errors in the force field representation, the lack of explicit polarization and a wide array of other errors, a slight lack

[14] I wish to thank Bernie Brooks for pointing this out.

of energy conservation from larger time steps and other sources was deemed a small problem once the "block of ice flying though space" problem was conquered!

The A-DNA to B-DNA transition

The results observed in the previous chapter were quite gratifying since they demonstrated the ability to stably simulate nucleic acids in solution when the long ranged electrostatic interactions are properly represented. This dismissed the need to lower charges on the phosphates or add extra base pair restraints to prevent the nucleic acids from falling apart and suggested the suitability of the Cornell *et al.* (1995) force field. The molecular dynamics trajectories also demonstrated stability on the nanosecond time scale. This is in the range of the longer simulations on nucleic acids at the time and no unsurmountable difficulties were apparent with the methods and simulations.

Of course it could be argued that the reason the results were less favorable when the electrostatic interactions was truncated at 9 Å was because this cutoff was too short or since the energies or forces were not smoothly switched, or alternatively since the equilibration protocol was not sufficient. To investigate the cutoff issue, several shorter simulations applying larger cutoffs were run with cutoffs of 12 Å, 15 Å and 20 Å. In all of these, distortions of the DNA structure appeared within the first 100 picoseconds (unpublished results). Interestingly the distortions were not equivalent; in the simulations with a cutoff of 12 Å of less, considerable bending was observed with break up of the terminal base pairs and in the 15 Å cutoff simulation, the DNA extended and middle base pairs broke apart. This is similar to the behavior seen in simulations of small helices in solution (Shreiber & Steinhauser, 1992). Simulations with larger cutoffs are extremely time consuming. Shown in Table 4 are timings for a 10-mer duplex for poly(A)-poly(T) and counterions with and without water. From this table, it is apparent that the PME simulations (with a 9 Å cutoff) are less computationally demanding than simulations

without PME and a 12 Å cutoff; the timings are comparable to simulations employing a dual cutoff at 9 Å and 15 Å.

in vacuo	*in vacuo*	solvated	solvated	solvated	solvated	solvated
656 atoms	656 atoms	9245 atoms	9245 atoms	9245 atoms	9245 atoms	9245 atoms
9 Å cutoff	no cutoff	9 Å cutoff	PME, 9 Å	dual 9/15 Å	12 Å cutoff	15 Å cutoff
64 s/ps	119 s/ps	1014 s/ps	1427 s/ps	1424 s/ps	1802 s/ps	3188 s/ps

Table 4: relative timings of sander with and without PME. Times were obtained on 1 processor of an SGI R8000.

In addition to issues related to the cutoff, it could also be argued that perhaps the equilibration protocol was not sufficient. Incomplete equilibration might lead to a situation where poorly equilibrated water may "crash" into the DNA causing structural distortions. This is unlikely to have occured in all the cutoff simulations yet never in the Ewald simulations unless perhaps the PME method overstabilizes the DNA. Moreover, care was taken to make sure that the solvent was well equilibrated by monitoring the pressure, temperature, volume, density, and potential energy of the system. During the initial "equilibration" all the values equilibrated to the expected range[15].

When we saw the results, we were rather surprised to learn that including the long ranged electrostatic interactions had such a profound effect. It seemed almost magical and therefore unbelievable. This led to lots of questions. Perhaps there was a bug? Or perhaps sampling is strongly inhibited when PME is turned on which prevents the nucleic

[15] For more information about the equilibration protocol beyond the discussion presented in this thesis, see the WWW page developed to help understand DNA simulations with AMBER at "http://www.amber.ucsf.edu/amber/equilibration/polyA-polyT.html".

acids from moving? Clearly movement was not inhibited, as shown in the previous chapter which displayed reasonable atomic positional fluctuations for ubiquitin. Nucleic acids also showed reasonable atomic positional fluctuations (data not shown). Given our concern, and consideration (of later retracted) statements that Ewald methods slow translational motion (Teleman & Wallqvist, 1990), I was worried (still) about sampling. Therefore I decided to try a larger experiment with conformational sampling. If sampling was not strongly inhibited, we should be able to start a simulation in various different geometries and converge to a common structure. In other words, canonical B-DNA and the crystal structure should converge to the same "B-DNA" structure. Likewise, since A-DNA is not favored in aquous solution at low salt, a simulation started in a canonical A geometry should transition to a B-DNA geometry and converge to the same structure as above. This is not only an issue with respect to sampling in the simulations, but also an issue of force field. In order to observe the transition, A-DNA must be less favorable and the barrier to the B->A transition must not be too high in order to surmount it during short (~1 ns) molecular dynamics simulations. Therefore, the simulations presented in the following paper were started (2 A-DNA and 2 B-DNA simulations). An additional simulation starting from the crystal structure was also run, however this was not discussed in the following paper. All in all, in 6 different simulations we saw convergence to a common structure. The issue of the force field is an important one as has been demonstrated in recent calculations. While the following paper was in press, Yang & Pettitt published results from the CHARMM 23 all hydrogen parameter set that showed a B-DNA to A-DNA transition during a 3.5 ns molecular dynamics simulation of $d(CGCGAATTCGCG)_2$ employing an Ewald method (Yang & Pettitt, 1996).

The following appeared in the *Journal of Molecular Biology* (1996) **259**, 434-444.

Observation of the A DNA to B DNA transition during unrestrained molecular dynamics in aqueous solution.

T. E. Cheatham, III and P. A. Kollman
Department of Pharmaceutical Chemistry
University of California
San Francisco, California 94143-0446

Summary:

A large challenge in molecular dynamics (MD) simulations of proteins and nucleic acids is to find the correct "experimental" geometry when a simulation is started a significant distance away from it. In this study, we have carried out four unrestrained ~1 ns length MD trajectories in aqueous solution on the DNA duplex d[CCAACGTTGG]$_2$, two beginning in a canonical A DNA structure and two beginning in a canonical B DNA structure. As judged by root mean squared coordinate deviations, average structures computed from all four of the trajectories converge to within ~0.8 - 1.6 Å (all atoms) of each other, which is 1.3 - 1.7 Å (all atoms of the central 6 residues from each strand) and 3.1 - 3.6 Å (all atoms) away from the B-DNA like X-ray structure reported for this sequence. To our knowledge, this is the first example of multiple nanosecond molecular dynamics trajectories with full representation of DNA charges, solvent and long range electrostatics that demonstrates both internal consistency (two different starting structures and four different trajectories lead to a consistent average structure) and considerable agreement with the X-ray crystal structure of this sequence and NMR data on duplex DNA in aqueous solution. This internal consistency of structure for a given sequence suggests that one can now begin to realistically examine sequence dependent structural effects in DNA duplexes using molecular dynamics.

Keywords:

molecular dynamics, nucleic acids, conformational transitions, A DNA, B DNA.

Given the critical biological role of DNA, it is essential that its structure and how it interacts with proteins be understood. Proteins recognize and bind to specific sequences of DNA, both in regulating gene expression and in repairing damaged DNA. A critical question is: do these proteins recognize the hydrogen bond patterns in the grooves of the DNA (Seeman *et al.*, 1976), the sequence dependent structure (Dickerson, 1982), or some combination of both. It is clear that sequence strongly influences structure, as is evident from sequence dependent DNA bending (Hagerman, 1992; Harrington & Winicov, 1994; Young *et al.*, 1995a). However, it is also clear that DNA structure is very strongly influenced by the environment. The effect of the environment ranges from global changes based on the solvent and ionic concentration leading to conversions between the distinct A-, B- and Z-DNA structural families to more local structural effects such as the radical 90° degree bend present in the DNA bound to CAP (catabolite activating protein; Shultz *et al.*, 1991) or differences in local helix parameters resulting from crystallization in trigonal *versus* monooclinic space groups (Lipanov *et al.*, 1993). Therefore, assessing the possible role of sequence dependent structual differences within the biologically relevant B-DNA structural family is a challenge. X-ray crystallography has given many insights, but suffers from limitations in the number of structures that can be crystallized and diffract to precise resolution, as well as interference from crystal packing effects (Dickerson *et al.*, 1987; Dickerson *et al.*, 1994; Lipanov *et al.*, 1993). NMR is becoming increasingly powerful in deriving duplex structures to high resolution (James, 1995), but the number of sequences that have been determined precisely is limited and the information derived is short ranged (representing < 5 Å in distance or three bonds in torsion restraints). Various theoretical models, such as those using empirical force fields and molecular mechanics/dynamics, have emerged, but to this point these have not been capable of sufficient realism and reliability to definitively address sequence dependent structure in DNA. Below, we suggest that a major step forward has been taken in molecular dynamics simulations of DNA, such that now one is in a position to use this method to begin to address sequence

dependent structures of DNA duplexes and perhaps modified and damaged DNA as well. Specifically, we have carried out four unrestrained nanosecond length simulations in aqueous solution on the DNA duplex d[CCAACGTTGG]$_2$, with the Cornell *et al.* (1995) force field, two beginning in a canonical A DNA structure and two beginning in a canonical B DNA structure. As judged by root mean squared coordinate deviations (RMSd), average structures computed from all four of the trajectories converge to within ~0.8 - 1.6 Å (all atoms) of each other, and are 1.3 - 1.7 Å (all atoms of the central 6 residues from each strand) and 3.1 - 3.6 Å (all atoms) away from the B-DNA like X-ray structure reported for this sequence (Prive *et al.*, 1991). The A DNA to B DNA transition, which represents an ~6 Å RMSd, occurs in these simulations on a roughly half nanosecond time scale. To our knowledge, this is the first demonstration of a free A DNA to B DNA transition occuring during an unrestrained molecular dynamics simulation in explicit solvent. Although the observed RMSd to canonical B DNA and the 5dnb crystal structure is fairly large, we do demonstrate stable convergence to the same B-DNA structure in all four trajectories reported herein and also from one trajectory starting from the 5dnb crystal structure reported elsewhere (Cheatham *et al.*, 1995). Moreover, we see expected sequence dependent bending. During the dynamics, we observe significant backbone fluctuations, such as B_I to B_{II} (ε,ζ: t, g- to g-, t) transitions, sugar repuckering, and we observe somewhat higher than expected fluctuations in the roll(ρ),twist(Ω) and tilt(τ). These simulations demonstrate a fairly wide radius of convergence for right handed DNA.

Historically, the first structural families of DNA emerged from fiber diffraction studies of DNA under varying conditions (Dickerson *et al*, 1982). These are the canonical A and B forms, which differ by a 5.7 Å RMSd between the A and B models of d[CCAACGTTGG]$_2$. The major differences between these forms relate to differences in the sugar pucker, parameters relating the base pairs to the helical axis, the rise between base pairs, and the width of the major groove. The preferred sugar pucker pseudorotation phase (pucker) value (Altona & Sundaralingam, 1972) is C2'-*endo* (~162°) for B DNA and C3'-

endo (~18°) for A DNA. The helicoidal parameters (Lavery & Sklenar, 1988) which show the largest differences between the canonical A DNA and B DNA are the displacement of the nucleotide base pair from the helical axis (x-displacement) which is in the range of -1.0 to 0.0 Å in B DNA and in the range of -6.0 to -5.0 Å for A DNA and the inclination of the base pairs which is ~-6° in B-DNA and greater than 19° in A DNA. The rise between base pairs is much less for A DNA which leads to an end-to-end length of ~23 Å for a canonical A DNA decamer compared to ~30 Å for B-DNA. Another distinguishing characteristic is the minor groove width, measured by close inter-strand phosphate distances, which is much larger in A DNA than B DNA. It is expected that in low salt and high humidity conditions, as are represented in these simulations, B DNA structures will be more stable than A DNA (Saenger, 1984).

Theoretical simulations of DNA structure have a rich history, as has been discussed in recent reviews by Beveridge (Beveridge *et al.*, 1993; Beveridge & Ravishankar, 1994). These studies clearly demonstate the importance of including some representation of the solvent. To date, very few realistic molecular dynamics simulations beyond 100 picoseconds with explicit solvent, counterions and full periodic boundary conditions have been reported on double-helical DNA. These past simulations, using a variety of different force fields, tended to drift significantly from the starting structures and in some cases displayed anomolous structure, such as base pair fraying. An early simulation of this type on d[CGCAACGC]• d[GCGTTGCG] (Van Gunsteren *et al.*, 1986) drifted rather quickly to a structure (time averaged over 50-80ps from the trajectory) that was 2.2 Å from canonical B-DNA and 3.5 Å from canonical A-DNA. Even with better treatments of the long range electrostatics through the use of reasonable 4 Å switching functions, a recent nanosecond simulation on d[CGCGAATTCGCG]$_2$ with periodic boundary conditions and explicit counterions, employing the GROMOS force field (Van Gunsteren & Berendsen, 1986) supplemented by Watson-Crick base pair restraints and reduced phosphate charges

to maintain base pairing, still showed considerable deviation (> 4.5 Å) from the starting structure (McConnell *et al.*, 1994).

Recently, a series of papers by Darden and Pedersen (Lee *et al*, 1995a, Lee *et al*, 1995b, York *et al*, 1995) have been published investigating nucleic acid crystal structures with an early version of the particle mesh Ewald code in AMBER. These studies demonstrate that with a proper treatment of the long range electrostatics, one can stay impressively close to the crystal structure. However, those simulations imposed constant volume and crystal symmetry, whereas the simulation we described previously (Cheatham *et al*, 1995) and herein apply constant pressure and attempt to mimic solution conditions. The above results and those of many other authors (Board *et al.*, 1992; Ding *et al.*, 1992; Luty *et al.*, 1994; Schreiber & Steinhauser, 1992; Smith & Pettitt, 1991; Steinbach & Brooks, 1994), suggest that in addition to an explicit representation of solvent, it is also critical to properly represent the long range electrostatics interactions, particularly for highly charged systems such as nucleic acids, in order to properly represent the structure, dynamics and energetics. For simulations in the ~10,000 atom range, the particle mesh Ewald [PME] method is considered to be among the most efficient and accurate ways to fully represent the long range electrostatic effects (Essmann *et al*, 1995; Petersen, 1995). However, one potential drawback of the Ewald method (Ewald, 1921) in its application to solvated biomolecular simulation is the effect of imposing true "infinite" periodic boundary conditions on the system. While imposing periodicity is clearly superior to applying a cutoff or surrounding the simulation cell by a vacuum interface, the simulation cell could be unduly influenced by the image cells (Luty *et al.*, 1994; Smith & Pettitt, 1991). This influence could lead to overstability of the simulated structure and/or misrepresented dynamics. The study presented here demonstrates that in the simulation of DNA, a complex conformational change is not inhibited and metastable structures do not appear to be overly stabilized.

Demonstration of "convergence" to a B-DNA-like structure for four independent molecular dynamics trajectories, two starting from canonical A DNA and two from canonical B DNA

Four separate molecular dynamics simulations were run, two starting from canonical A DNA (A1 and A2) and two starting from canonical B DNA (B1 and B2). In Table 5, a summary of the methods is presented along with the RMSd values between the starting structures and average structures computed from the trajectories. A comparison of these average structures demonstrates surprising agreement among them. All four simulations converge to average structures with RMSd values of ~0.8 - 1.6 Å among each other. It is also notable that the inner $d[AACGTT]_2$ core is within 1.3 - 1.7 Å from the crystal structure and closer to it than to B DNA (1.6 - 1.9 Å) or A DNA (2.3 - 2.9 Å). When the entire sequence is considered, the RMSd to the crystal (3.1 - 3.6 Å), canonical B DNA (2.9 - 3.3 Å) or canonical A DNA(3.4-4.1A) is higher. This is primarily due to bending towards the major groove at the TpG step (steps 2 and 8 which have a base pair roll, $\rho > \sim 10°$), which is not seen in the crystal (which has a minor groove roll at these steps, $\rho = \sim -11°$) or the canonical B DNA structure ($\rho = 0°$). Encouragingly this positive roll at TpG is consistent with NMR data (Ulyanov & James, 1995). Similar to what was seen in the early simulations of Van Gunsteren (1986), the RMSd shows that the structure is near midway between canonical A DNA and B DNA values, but still closer to B DNA. This serves to point out that high root-mean-square deviations are not good indicators of lack of similarity between DNA helices, since small changes in the helicoidal parameters can lead to rather large RMSd values. However, the excellent agreement between the average structures, especially considering we are including all atoms in the root mean square deviation calculation, is encouraging.

	$A1_{avg}$	$A2_{avg}$	$B1_{avg}$	$B2_{avg}$	A	B	5dnb
$A1_{avg}$	-	1.32	1.01	1.03	3.69	3.18	3.46
$A2_{avg}$	1.27	-	1.55	1.22	4.14	3.12	3.22
$B1_{avg}$	1.01	1.35	-	0.86	3.39	3.31	3.55
$B2_{avg}$	1.11	1.18	0.77	-	3.87	2.90	3.14
A	2.47	2.85	2.30	2.61	-	5.70	5.96
B	1.90	1.94	1.92	1.56	3.23	-	1.71
5dnb	1.71	1.67	1.60	1.28	3.15	1.17	-

Table 5: Summary of methods and RMSD values between the starting and average structures[16]*. RMS coordinate deviation (Å) between the average structures computed from the four trajectories, the Prive* et al. *(1991) crystal structure (5dnb) and canonical A-DNA (A) and B-DNA (B) models. The upper triangle is a comparison of all DNA atoms. The lower triangle represents the internal d[AACGTT]$_2$ DNA core or residues 3-8 and 13-18.*

[16] **Creating average structures**: Average structures from the trajectories were calculated using the carnal module of AMBER 4.1 (Pearlman et al., 1995) to coordinate average the RMS coordinate fit frames (over all DNA atoms) taken at 1 ps intervals. These structures were then minimized to a difference in the norm gradient of the energy of 1 kcal/mol-Å which served to relax unrealistic bonds and angles (leading to an overall RMS deviation change near 0.3 Å during the minimization). The $B1_{avg}$ and $B2_{avg}$ structures were calculated by averaging over their entire trajectories, respectively, from 0 to 1400ps. The A structures were created from the latter parts of the trajectories after the observation of the A DNA to B DNA transition. The $A1_{avg}$ structure was taken over 600 to 1456 ps from the A1 trajectory, the $A2_{avg}$ structure over 903 to 1253 ps from the A2 trajectory. **Creating the starting models**: Canonical A and B DNA starting structures (Arnott & Hukins, 1972) were created using the nucgen module of AMBER 4.1. **Simulation conditions**: A rectangular box was constructed with counter-ions placed by the AMBER EDIT module. Simulations were run with the sander module of AMBER 4.1 with SHAKE on the hydrogens, a 2 fs timestep, a temperature of 300K with Berendsen temperature coupling, a 9 Å cutoff applied to the Lennard-Jones interactions, and constant pressure with isotropic scaling. The nonbonded pairlist was updated every 10 steps. No extra restraints were placed on the DNA following the equilibration period. The PME charge grid spacing was approximately 1.0 Å and the charge grid was interpolated on a cubic grid with the direct sum tolerance set to 10^{-6}. The DNA box sizes after equilibration were approximately 60.0 Å by 40.0 Å by 40.0 Å for the A simulations (9248 atoms in A1 and A2) and 55.0 Å by 41.0 Å by 41.0 Å for the B simulations (9269 atoms in B). The A1 simulation was run for 1456 ps, the A2 for 1253 ps, and the two B trajectories for 1400 ps. **Equilibration**: non-PME equilibration was performed initially by first holding the DNA fixed and letting the water and counter-ions minimize for 10,000 steps, followed by dynamics for 25 ps where the temperature was raised from 100K to 300K in 2 ps. PME was then applied in all remaining simulations. Equilibration was continued with 25 kcal/mol restraints placed on all solute atoms, minimizing for 1000 steps, followed by 3 ps of MD which allowed the water to relax around the solute. Production runs of A2 were started from this point. In simulations A1, B1 and B2 this equilibration was followed by 5 rounds of 600 step minimizations, reducing the solute restraints by 5 kcal/mol during each round. Finally, the system was heated from 100K to 300K over 2 ps and then production runs were initiated.

In figure 18, stereo views of the structures used and generated in this study are presented. These are the 5dnb crystal structure (Prive *et al.*, 1991), the four overlapped converged average structures, canonical B DNA and canonical A DNA structures. The view shown for the overlapped structures (b) highlights part of the minor groove (top) and major groove (bottom) and shows a number of notable features in the structure. In particular, the base pairs appear flat and rather ideal. During the simulation, instantaneous structures show considerable deviation from this average picture and do not appear as regular or ideal. It is also readily apparent that the helix is underwound and bent at the central CpG and step 2 and 8 TpG steps. The deviations from the average structures appear mostly in the backbone regions. Overall, the average structures appear more "regular" than the crystal and B_{II} backbone conformations are not present.

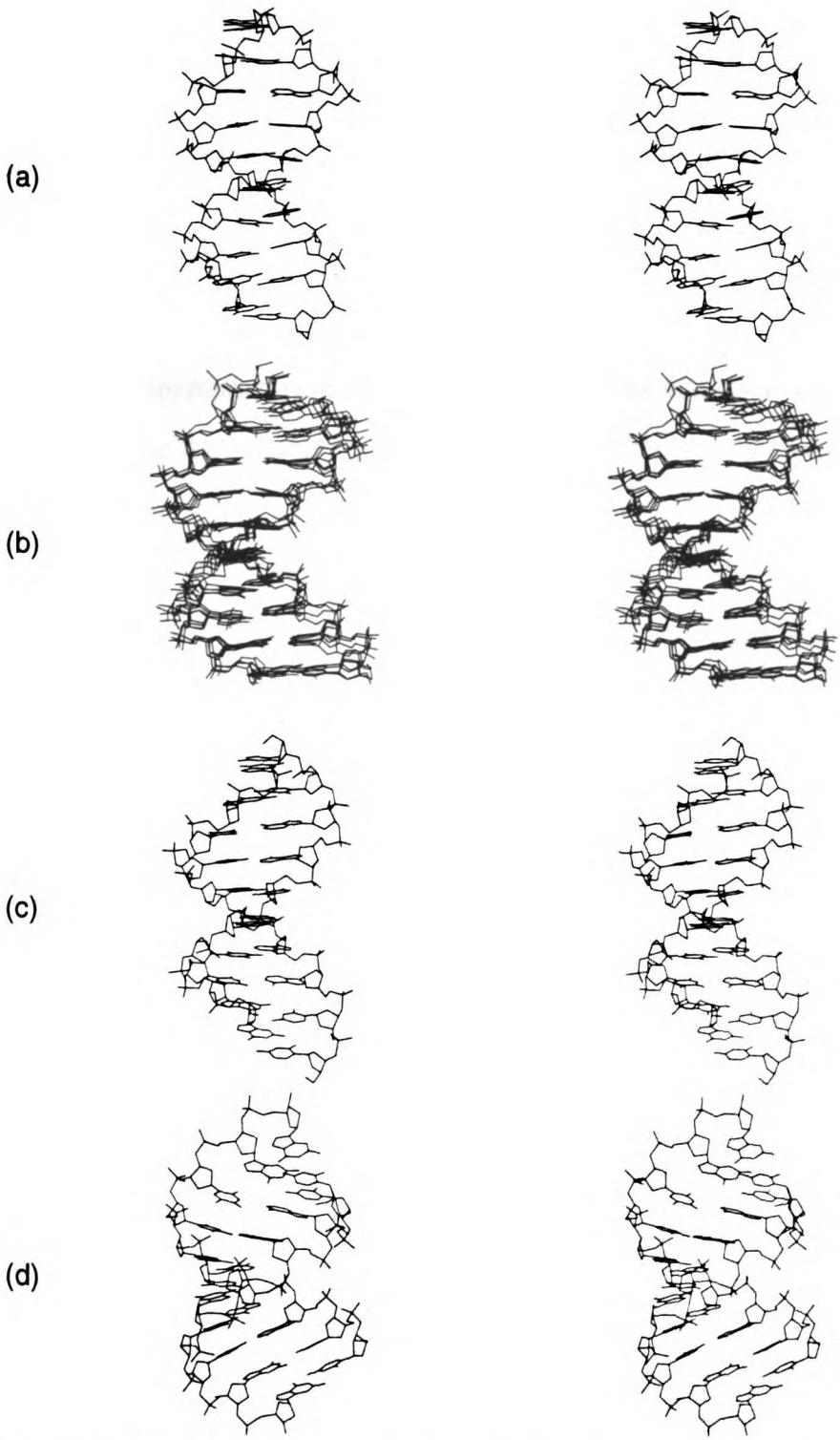


Figure 18: Stereo views of the structures used and generated in thie study. (a) The 5dnb structure (Prive et al., 1991). (b) An root-mean-square best fit of all the average structures computed herein, $A1_{avg}$, $A2_{avg}$, $B1_{avg}$ and $B2_{avg}$. (c) Canonical B-DNA. (d) Canonical A-DNA. All the structures (except the 5dnb crystal) were generated as described in the footnote to Table 6. The structures were all best fit (RMSd) in order to line them up within a common reference frame. All the atoms except the hydrogen atoms are displayed.

In Table 6, the average backbone angles, sugar pucker pseudorotation values and amplitudes, and helicoidal parameters with respect to the global helical axis over the 600-1456 ps range from the A1 trajectory calculated using the "Dials and Windows" (Ravishankar *et al.*, 1989) interface to "Curves" (Lavery & Sklenar, 1989) are presented. In addition to the mean values over this range, the standard deviations of the values (or fluctuations) and the maximum range observed for a given residue, base pair, or base pair step (where appropriate) are listed and compared to the values calculated for the 5dnb structure. The A1 trajectory over this range was chosen since the average structure generated from it ($A1_{avg}$) was the canonical A start furthest from the crystal structure. The values for the other average structures are very similar. In order to better understand the range of motion in the trajectory over the 600-1456 ps (or $A1_{avg}$) range, we calculated the RMSd of the average $A1_{avg}$ structure to each 1 ps frame making up the average. We observe that this value stays within 1.9 Å (all atom RMSd), or within 1.5 Å if three regions in the trajectory with lifetimes in the range of a few picoseconds to 40 ps are omitted. This is rather remarkable considering the fact that the two symmetric halves of this structure (obtained by rotating the duplex by 180° about any axis perpendicular to the helical axis and best fitting) differ from each other by a RMSd of 1.4 Å. This suggests that we have not sampled over a long enough time scale or over a sufficient number of conformations for full "convergence" to a symmetric average structure, as one would expect in solution in the absence of crystal packing effects. It is also apparent that the values reported show a wide range. However, the extremes are fairly isolated events since they are well outside the range of the fluctuations. Although one might expect the terminal bases to display the widest range of motions, this is not always the case.

	5dnb	mean	stddev	min	max	
alpha (°)	296.6	288.9	12.5	187.9	331.5	(G9)
beta (°)	166.3	168.7	12.6	96.2	213.8	(G9)
gamma (°)	49.2	54.1	10.4	8.0	92.9	(A3)
delta (°)	128.7	117.6	18.6	58.1	158.7	(C1)
epsilon (°)	204.9	199.7	20.3	151.4	329.0	(G6)
zeta (°)	239.7	256.1	25.9	77.4	303.4	(G9)
chi (°)	257.7	236.0	17.4	189.8	313.6	(A4)
pucker (°)	145.5	123.5	28.9	-5.9	207.2	(G10)
amplitude (°)	39.1	41.4	5.9	8.4	60.4	(A4)
twist (°)	35.3	30.2	5.1	1.7	40.6	(T8-G9)
roll (°)	4.0	7.1	8.4	-24.1	32.5	(C1-C2)
tilt (°)	0.0	0.1	5.2	-15.0	22.5	(T7-T8)
propellor (°)	-16.4	-10.8	12.4	-61.6	61.3	(C1-G20)
rise (Å)	3.43	3.48	0.5	1.31	5.04	(T8-G9)
x-disp (Å)	1.17	-1.27	0.6	-3.46	0.52	(C1-G20)
y-disp (Å)	0.00	0.01	0.5	-1.16	2.55	(C1-G20)
tip (°)	0.0	0.2	5.7	-27.5	13.7	(C1-G20)
buckle (°)	0.0	0.0	12.9	-42.2	36.6	(T7-T8)
opening (°)	4.3	2.3	5.4	-20.1	24.8	(T7-T8)
inclination (°)	-2.3	-6.2	5.9	-29.1	15.8	(T8-G9)
shear (Å)	0.0	0.0	0.4	-1.17	1.85	(C1-G20)
stretch (Å)	-0.1	0.2	0.2	-1.21	1.11	(C1-G20)
stagger (Å)	0.0	-0.2	0.5	-2.22	1.54	(C1-G20)
shift (Å)	0.0	0.0	0.6	-1.85	1.47	(A3-A4)
slide (Å)	-0.1	-0.2	0.4	-1.81	1.82	(C5-G9)
P_5-P_{20} (Å)	12.62	13.92	1.51	8.35	18.69	
P_6-P_{19} (Å)	12.45	12.18	1.26	9.54	15.71	
P_7-P_{18} (Å)	10.24	12.97	1.80	9.40	16.52	
P_8-P_{17} (Å)	10.24	11.72	1.27	7.95	15.32	
P_9-P_{16} (Å)	12.45	11.88	1.34	8.69	16.41	
P_{10}-P_{15} (Å)	12.62	13.17	1.78	8.78	17.70	

Table 6: Fluctuations, mean value and ranges for the backbone angles, sugar pucker pseudorotation angle and amplitudes, selected helicoidal parameters, and average minor groove distances[17].

[17] The data presented under the column heading **5dnb** represents values calculated from the 5dnb pdb entry (Prive et al., 1991). The other reported values are from the **A1** trajectory taken at 1 ps intervals over the range of 600 ps until the end of the run (1456 ps); this is the same range used to produce the $A1_{avg}$ structure reported in table I. After 600 ps, the

The average values reported in Table 6 are generally close to the crystal structure with a few exceptions. The relatively fixed angles (α, β and γ) are all within 8° of the crystal values. In contrast, the angles relating to the dynamic sugar conformation, specifically the χ glycosidic torsion, the sugar pucker and the backbone angle δ, as well as the backbone ε and ζ angles show higher deviation from the crystal structure, large ranges and significant fluctuations. These fluctuations are a favorable property of these simulations since they show motions on a time scale that is essentially averaged out in structures determined by X-ray crystallography. For example, rather than showing static B_{II} backbone conformations at the P_3/P_{13} and P_9/P_{19} phosphates as observed in the crystal, we see transient B_I to B_{II} (ε,ζ: t, g- to g-, t) transitions on a time scale ranging from 5 ps to 450 ps throughout the entire sequence. Similarly, we observe multiple sugar pucker transitions throughout the sequence, as is expected based on NMR results (Ulyanov *et al.*, 1995). However, we systematically underestimate the χ and δ angles and show an average pucker closer to C1'-*exo* than what is expected for "average DNA" as derived from NMR solution data (χ = 247° and pucker = 138°) (Ulyanov & James, 1995). Although the pucker phase is underestimated, we do show the preference for pyrimidine repuckering more frequently than purine repuckering and the tendency for purine to adopt higher phase values (Poncin *et al.*, 1992; Ulyanov *et al.*, 1995; Zhurkin *et al.*, 1991) and observe that

---

structure has definitely converted to B DNA. The $\mathbf{A1}_{\mathbf{avg}}$ mean values (**mean**) and fluctuations or standard deviations (**stddev**) reported here are averages of the individual residues, base pairs, or base pair steps (where appropriate). The maximum (**max**) and minimum (**min**) values reported here were chosen to represent the residue, base pair, or base pair step which showed the largest range over this interval and the appropriate residue, base pair or base pair step showing this range is noted in parenthesis. Note that the maximum and minimum values reported here are not strictly the largest or smallest value observed. The inter-strand phosphate distances reported here were chosen for convenience to match those chosen by Prive et al., 1991. They are labeled with the residue number of the phosphate. Note that over this interval, these distances do not necessarily represent the shortest interstrand phosphate distances. The sugar pucker, backbone torsion and helicoidal parameters were calculated using the Dials and Windows (Ravishankar et al., 1989) interface to Curves (Lavery & Sklenar, 1989). The overall averages of the helicoidal parameters reported are similar (data not shown) to values calculated with respect to a local axis prior to averaging.

the terminal nucleotides repucker at a much higher rate than the internal nucleotides. Furthermore, our average pucker value and frequent repuckering to O4' *endo* is consistent with the Raman results of Brahms *et al.* (1992). The fluctuations we see over 1 ns in ε and ζ appear larger than those observed in the double helix portion of a recent 1 ns simulation on a triplex $d(CG\cdot G)_7$ that also used an Ewald (Ewald, 1921) treatment (Weerasinghe *et al.*, 1995a).

Perhaps the most notable deviation from the crystal structure is the considerable unwinding of the double helix by ~4-5° (Ω = ~30.0° to 30.4°). For DNA in solution, various experiments suggest the following results: (1) electrophoretic band shifts due to the change in the superhelicity resulting from DNA inserts into plasmids suggest a twist Ω = 34.3° (Peck & Wang, 1981); (2) the periodicity of cleavage of surface absorbed DNA suggests a twist Ω = 34.0° (Rhodes & Klug, 1980) and (3) "average DNA" results from NMR suggest a twist Ω = ~35° (Ulyanov & James, 1995). In fact, twists nearly as low as seen here have been observed in solution NMR structures; for example, the average twist = 31.5° for the structure of $d[GCCGTTAACGGC]_2$ (Kim & Reid, 1992). The general lowering of the twist in solution with respect those found in crystal structures, coupled with the observation that small counter-ions and low ionic strength lead to decreases in the overall twist (Anderson & Bauer, 1978; Wang, 1969), suggest our twist may not be as poorly represented as it first appears. This is because our simulation with "net-neutralizing" sodium counter-ions essentially represents the limit of no added salt. However, the low ionic strength limit probably only represents an ~2° decrease over the solution values leading to an expected twist of Ω = 32° for B DNA at zero ionic strength, suggesting we are still underestimating the twist. The low twist we observe could be related to our overestimating the percent of sugars near O4'-*endo* or also due to overemphasis of the interstrand repulsion (Zhurkin *et al.*, 1982). Greater interstrand phosphate repulsion could, in principle, lead to widening of the minor groove which

correlates with the data shown in Table 6. However, an average structure computed over the 300-900ps range for the B1 trajectory has a greater twist of 31.4°, yet has the short minor groove width distances that are closer to the crystal values (B_{avg}, P_7-P_{18} = 10.24 Å, P_8-P_{17} = 11.97 Å), which tends to argue against this explanation. All of the average structures show the tendency to narrow in the minor groove. This narrowing seems to occur in the absence of crystal packing effects or the requirement of B_{II} backbone conformations at P_3/P_{13} and P_9/P_{19} as observed in the crystal (Prive *et al.*, 1991).

The time course of the A DNA to B DNA transition

This analysis focuses on those geometric variables that differ most significantly between A DNA and B DNA in order to clearly illustrate the time course of the A to B transition (Figure 19). Specifically the sugar pucker, end-to-end length, minor groove width, and the base pair inclination and x-displacement to the global helical axis are plotted. These variables have converged to within the B DNA range on the order of half a nanosecond.

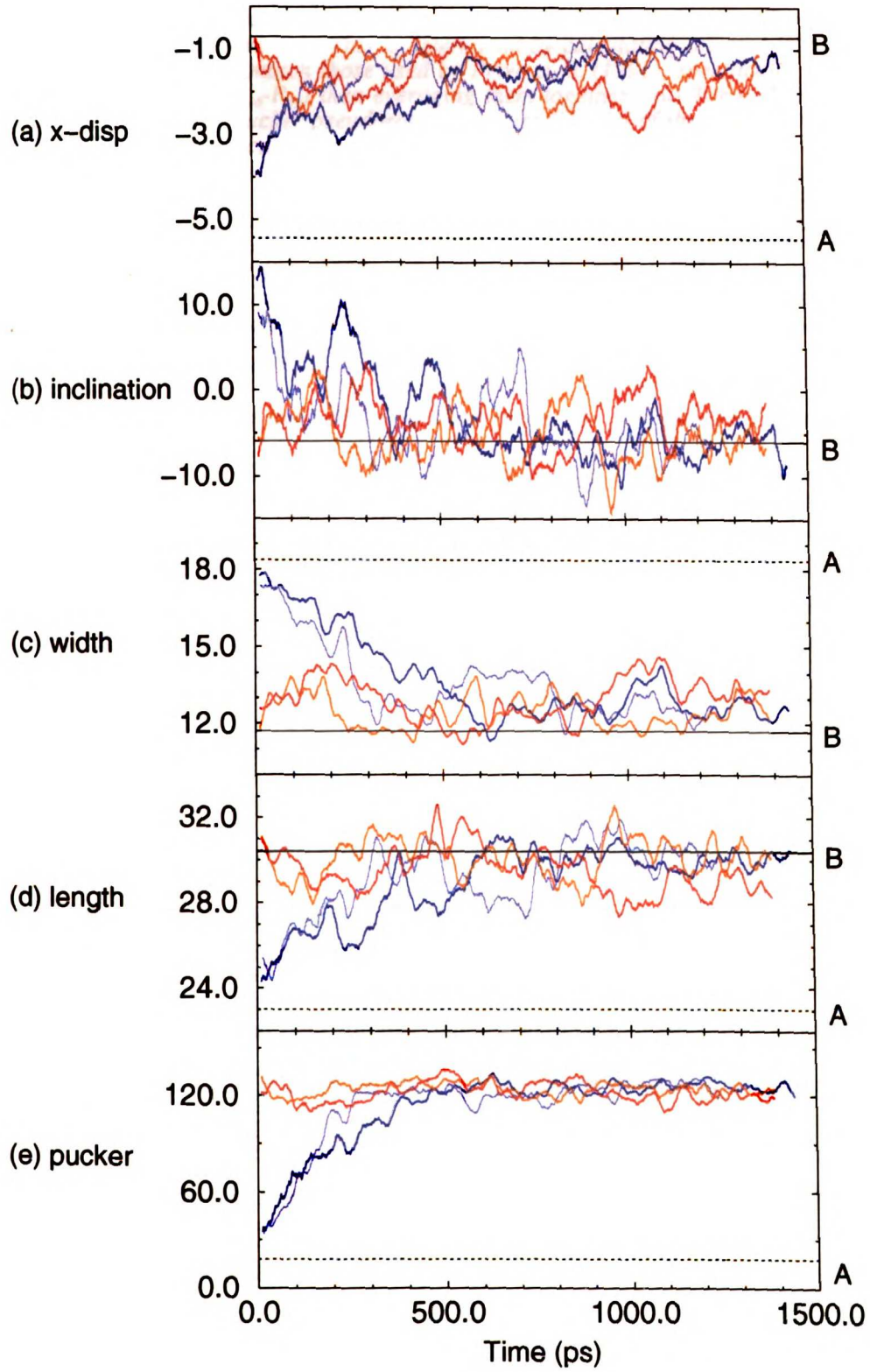

(a) x–disp
−1.0
−3.0
−5.0
B
A
(b) inclination
10.0
0.0
−10.0
B
B
(c) width
18.0
15.0
12.0
A
B
(d) length
32.0
28.0
24.0
B
A
(e) pucker
120.0
60.0
0.0
A
0.0
500.0
1000.0
1500.0
Time (ps)

Figure 19: (previous page). Time course for the A-DNA to B-DNA transition. Results from the A-DNA trajectories (A1 and A2) are printed in blue (A1 is dark blue and A2 is light blue) and for the B-DNA trajectories in red (B1) and orange (B2). The plots are as follows. (a) The average x-displacement (Å). (b) The average base-pair inclination (°). (c) The average width of the minor groove (Å). This is determined by calculating selected interstrand phosphate distances (to match those used by Prive et al. *(1991), specifically P_5-P_{20}, P_6-P_{19}, P_7-P_{18}, P_8-P_{17}, P_9-P_{16} and P_{10}-P_{15}) then everaging them together. (d) The end-to-end length (Å). (e) The everage sugar pucker pseudorotation values (°). All of the plots represent an average of the individual values calculated from the trajectories at 1 ps intervals from all the residues, base-pairs, or base pair steps (where appropriate). Furthermore, the data have been smoothed by performing a running average in time over 25 ps prior to plotting. Additionally, when the canonical values are within the range displayed on the plots, that are marked with a continuous line (B-DNA) or dotted line (A-DNA). The canonical values out of range are the typical B-DNA pucker in the ~160° range or greater and the A-DNA inclunation which is > 19°. All of the values, with the exception of the groove widths, were calculated using the Curves (Lavery & Sklenar, 1988) interface from Dials and Windows (Ravishankar* et al., *1989).*

Considering the average over all the residues presented in Figure 19, it does not appear that any one of these variables "drives" the A to B transition. The data for the A trajectories (light and dark blue) shows that initially the inclination spikes rapidly downward and the length and x-displacement spike upward rapidly moving to near B DNA values in about 200 ps. In the range of 200 - 450 ps, the individual sugars repucker away from C3'-*endo*. The slowest converging parameter is the rise in the x-displacement from -2.5 to -1.0 Å which takes roughly twice as long as the sugar repucker. In all of the trajectories, the x-displacement spans a large range and spends a significant amount of time in the -3.0 to -1.0 Å range. Considerable fluctuation in the individual parameters is evident. Considering the data presented in Figure 19, and given that we have only observed two transitions in one direction, there is not sufficient data to derive a single mechanism for the A DNA to B DNA transition. However, our data suggest that we observe a generalized motion in all the parameters towards the B form and that one parameter alone, such as the sugar repuckering of individual residues, does not necessarily drive the transition. Further analyses are underway to attempt to further seek a cause and effect mechanism for this transition.

A closer observation of the data presented in Figure 19 shows that the end-to-end length and x-displacement in the B1 trajectory start moving towards more A-like values between ~900 and 1200 ps. In all of our trajectories, the x-displacement displays a wide

range and spends a considerable amount of time in the -1.0 to -3.0A range. These low values in the B1 trajectory are the result of significant bending towards the major groove, most notable at the TpG steps, with bending also at the central CpG and terminal GpG steps. All the average structures show similar bending patterns; however, the total bend is less at the terminal residues and is partially compensated by small bending into the minor groove at the GpT and TpT steps. This is in marked contrast to the crystal structure, which shows bending into the minor groove at the TpG steps compensated by bending into the major groove at the TpT steps. As mentioned previously, the observed bending alone could lead to significant RMSd values and partially explains why the inner core is closer to the crystal than the entire sequence. Since our simulation is set up to mimic solution conditions and does not impose the experimentally derived unit cell packing and crystal boundary conditions, we are pleased to see consistency with results obtained by NMR methods (Ulyanov & James, 1995) which suggest preferential bending into the major groove at TpG and CpG steps and general consistency with analysis of crystal structure bending patterns (Young *et al.*, 1995a) rather than precise agreement with the 5dnb crystal structure (Prive *et al.*, 1991).

Analysis of the fluctuations in the "fundamental parameters"— twist, roll and tilt: how significant are they?

The twist (Ω) and roll (ρ) [and to a lesser extent the tilt (τ)] have recently been described as "fundamental" parameters which determine the long range structure in nucleic acids (Gorin *et al*, 1995). The fluctuations we observe show the expected trend that rolling is preferred over tilting ($\Delta\rho$ = 7.5° - 8.7°, $\Delta\tau$ = 4.6° - 5.7°). Our fluctuations in twist ($\Delta\Omega$ = 4.9° to 5.3°) appear in the range of the higher estimates and measurements to date. Time correlated single photon counting of intercalated ethidium bromide on pBR_{322} DNA gives a $\Delta\Omega$ of 5° degrees (Selvin *et al.*, 1992). Analysis of fluorescence anisotropy decay using

an elastic chain model predicts fluctuations in twist in the range of $\Delta\Omega$ = ~3.3° to 5.1° in the nanosecond time range (Barkley & Zimm, 1979). Our observations, and the results cited above, are larger than those found in previous theoretical treatments which span the range of $\Delta\Omega$ = 1.2° to 4° (Cognet *et al.*, 1995; Levitt, 1978; Zhurkin *et al.*, 1982), but are in the range of estimates based on the analysis of crystal data ($\Delta\Omega$ = ~5° - 6°) (Gorin *et al.*, 1995; Olson *et al.*, 1995). Typically the observed fluctuations are related back to experimental data *via* the measurement of the persistence length. The persistence length relates to the limit of how widely spaced two nucleotides in a chain have to be before their directions (based on an internal reference frame) are no longer correlated. As a polymer becomes more rigid, the persistence length increases. Experimentally, this length is ~500 Å in a 0.1 N Na^+ solution (Eisenburg, 1987; Hagerman, 1988; Sobel & Harbst, 1991). For short chains at the low salt limit, such as ours, the persistence length is probably longer (Hagerman, 1981). Given a persistence length in this range, the expected fluctuations in roll and tilt are on the order of $\Delta\tau$ = ~3° and $\Delta\rho$ = ~4° (Olson *et al*, 1993). If we use the equation for persistence length based on the model of anisotropic bending in the appendix of Olson *et al.* (1993), our calculated length is ~223 Å. This value is clearly too low and suggests that either we are overemphasizing the dynamics (as has been seen in low temperature studies of myoglobin by Loncharich & Brooks, 1990) and/or that the fluctuations in roll, tilt and twist are not independent. Given that we are looking at a small piece of DNA, the fluctuations in roll, tilt, and twist may be correlated, hence the formula of Olson *et al.* (1993) may not be appropriate to apply. However, this does not exclude the possibility that we may be overemphasizing the motions. Given that we are imposing infinite periodic boundary conditions and were initially concerned that we may be overstabilizing certain structures and/or inhibiting the dynamics, the higher than expected fluctuations are in some sense gratifying.

Conclusions

We have shown that four molecular dynamics trajectories (each of length >1 nsec) on the DNA duplex d[CCAACGTTGG]$_2$, two starting from canonical A DNA and two starting from canonical B DNA converge to a common structure, which is in rather good agreement with available data from X-ray crystallography and NMR. The fact that independent geometries converge to a common structure that is in reasonably consistent with experiment, using a molecular dynamics model that uses a state of the art force field, simulation protocol and representation of long rang electrostatics is encouraging. This suggests that such simulation approaches may now be able to join X-ray crystallography and NMR in usefully analyzing sequence dependent structural effects in DNA, something that is of the utmost importance in understanding sequence dependent protein recognition of DNA. Our studies also show a reasonable representation of dynamical aspects of DNA (e.g. sugar and other torsional changes and fluctuations in twist, roll and tilt). The largest discrepancy with experiment in the simulations appears to be a rather small average twist (~30°) compared to solution measurements (31° to 35°). It is clear that more work is necessary to better understand the DNA dynamics and to precisely map out the role of the microenvironment (water and counter-ions) around the DNA. In addition, further study is necessary to determine how well one can represent sequence specific bending and twisting patterns. Preliminary work on a variety of sequences (T. E. Cheatham, unpublished results) suggests that we do, in general, accurately represent sequence specific bending patterns (Ulyanov & James, 1995; Young *et al.*, 1995a), such as a small bend into the major groove in poly(A)-poly(T) sequences (Young *et al.*, 1995b) but poorly represent sequence specific twisting. One further possible criticism of this work is that we do not know whether the force field employed here would tend to overstabilize B DNA structures, although the relative energy of sugar repuckering of the nucleosides with this force field seems reasonably consistent with experiment (Cornell *et al.*, 1995). Preliminary work on

RNA/DNA hybrids (T.E. Cheatham, unpublished results) and on DNA helices containing phosphoramidates (P. Cieplak, unpublished results) would suggest that we are not overstabilizing B DNA. Current work on RNA duplexes with the same sequence (r[CCAACGTTGG]$_2$) lends further support since we find that the Cornell *et al.* (1995) force field does support "stable" canonical A RNA. (RMSd 0.8 Å for a 600 psec trajectory: T.E. Cheatham, unpublished results). The results in this study are very encouraging since they display a radius of convergence on the order of 6 Å for right-handed double-stranded DNA. This leads to the hope that future molecular dynamics simulations will lead to a thorough understanding of sequence specific structure and dynamics in nucleic acids and the role of these structural and dynamic effects in protein recognition.

Acknowledgements:

P.A.K. is grateful to acknowledge research support from the NIH through grant CA-25644. We would like to acknowledge Nikolai Ulyanov, Uli Schmitz, Carlos Simmerling, Thomas Fox and Jennifer Miller for helpful discussions, Silicon Graphics, Inc., and the Pittsburgh Supercomputing Center (PSC) (MCA93S017P) for computational resources, the UCSF Computer Graphics Laboratory (RR-1081), Michael Crowley of PSC for parallelizing the PME on the Cray T3D, and Tom Darden for helpful discussions and for making the PME code available. T.E.C. was a recipient of an NIH Biotechnology Training Grant (GM08388) and is a UCSF Chancellor's Graduate Research Fellow.

RNA:RNA and DNA:RNA hybrid duplexes

The observation of a spontaneous A-DNA to B-DNA transition in two separate simulations of d[CCAACGTTGG]$_2$ suggests that conformational sampling is not seriously inhibited in simulations of duplex DNA. This further suggests that our worries about true periodicty and poor sampling in PME simulations of water solvated DNA are unfounded. Clearly the DNA duplex is also very flexible. During the dynamics, higher than expected fluctuations in the roll, tilt and twist are observed as is sugar repuckering and backbone transitions (such as B_I to B_{II} correlated angle flips) thoughout the duplex. Despite the tremendous mobility, convergence to a common *average* structure was seen in six separate simulations (two B-DNA, two A-DNA and two crystal structure starts). Instantaneously, however, the nucleic acid is characterized by small structural distortions ~1.5-2.0 Å away from this idealized average structure.

The observation of the A-DNA to B-DNA transition also in part validates the nucleic acid force field parameters (Cornell *et al.*, 1995) since B-DNA is stabilized over A-DNA which is expected for solvated DNA in low salt conditions. However, there did appear to be some deficiencies, most notably the low twist (~30°) and slightly lower than expected average sugar pucker value for the average structure. Despite the deficiencies, clearly a B-DNA structure is well represented by the force field. Even better (and certainly unexpected) was the observation of sequence specific bends into the minor groove at the TpG and CpG steps, in constrast to the bend in the minor groove at the TpG steps seen in the crystal structure (Prive *et al.*, 1991). These bending patterns are consistent with results seen in NMR experiments (Ulyanov & James, 1995). Since the simulations were set up to mimic solution conditions rather than the packed crystal, the observation of these "solution" characteristic bends was encouraging.

All these results were very exciting and suggested that we were now at a point to seriously investigate sequence dependent structure and attempt to model the effect of the environment on the structure of nucleic acids in solution. This led the research direction to bifuricate into a number of separate, but related, paths; this was made possible by the availability of tremendous allocations of computer time on the Cray T3D and most recently the Cray T3E. Additionally, many calculations were run on a 2-processor R8000 Challenge machine loaned by SGI for development of parallel AMBER. Rather than jumping into larger simulations of protein-DNA complexes or larger heterogenous nucleic acid structures, I continued to investigate small nucleic acid duplexes in solution. A variety of different DNA sequences were studied in order to investigate if the A-DNA to B-DNA transition was a fluke related to the $d[CCAACGTTGG]_2$ structure and to investigate sequence dependent structure (poly(A)-poly(T), poly(G)-poly(C), $d[ATATATATAT]_2$), to investigate A-tract bending($d[AAAAATTTTT]_2$, $d[TTTTTAAAAA]_2$, $d[AAAACGTTTT]_2$, $d[TTTTCGAAAA]_2$), and to see if we could remain close to a well refined NMR structure (d[CTCAAGGCAAGCT] (Mujeeb *et al.*, 1993)). In addition, many different simulations were run using the same 10-mer sequence $d[CCAACGTTGG]_2$ to perform a "sensitivity" analysis. What happens if we run without salt or with 1 M Na^+ Cl^-? What happens if the charges are reduced on the phosphate groups or water? What happens if the pressure is reduced? What happens if sugar puckers are held fixed at C3'-*endo* or C2'-*endo* or if the sugar pucker parameters are changed? All of these simulations were started to see how sensitive the simulation protocol and force field was to these small changes and to see if we might be able to "fix" the deficiencies such as low twist and represent the effect of the environment (*i.e.* salt effects).

Of course an even simpler way to investigate the sensitivity of the nucleic acid force field parameters is to compare the results of simulations on DNA with corresponding RNA duplex or DNA:RNA hybrid duplex structures. RNA duplexes are generally observed in a

canonical A geometry and "B-RNA" has never been observed experimentally. DNA:RNA hybrid duplexes are found in a conformation intermediate between canonical A and B geometries. The results from simulations on RNA:RNA and DNA:RNA hybrid duplexes of the same 10-mer sequence (r[CCAACGUUGG]$_2$ and d[CCAACGTTGG]-r[CCAACGUUGG]) are presented in the following paper that was recently accepted (pending revision) to the *Journal of the American Chemical Society*. The suggested revisions are a considerable shortening of the text; as one of the anonymous reviewers remarked "the article reads too much like a PhD thesis." Rather than include the revised text, the originally submitted text is included herein since it contains much more background, more figures and more analysis. From the outset, the goal was to see if the molecular dynamics simulations when started from canonical A and canonical B geometries could distinguish the structural features and flexibility between DNA:DNA, RNA:RNA and DNA:RNA hybrids. In some cases the simulations can, and in other cases conformational sampling is limited and "expected" transitions (B-RNA to A-RNA) are not observed.

Accepted pending revision to the *Journal of the American Chemical Society* (1997).

Molecular dynamics simulations can reasonably represent the structural differences in DNA:DNA, RNA:RNA and DNA:RNA hybrid duplexes.

Thomas E. Cheatham, III and Peter A. Kollman*

University of California San Francisco

Department of Pharmaceutical Chemistry

San Francisco, CA 94143-0446

Keywords: Molecular dynamics, particle mesh Ewald, nucleic acid, hybrid, hydration, flexibility, counterions, A-DNA, B-DNA, A-RNA, B-RNA.

Running Tittle: MD on DNA, RNA and hybrid duplexes

Abstract: Nanosecond length simulations applying the particle mesh Ewald method within AMBER 4.1 on canonical A-form and B-form geometries of d[CCAACGTTGG]$_2$, r[CCAACGUUGG]$_2$ and d[CCAACGTTGG]-r[CCAACGUUGG] duplexes are reported. DNA duplexes only adopt a stable B-DNA geometry, in contrast to RNA duplexes which adopt both a stable A-RNA and "B-RNA" geometry. The "B-RNA" can be converted to A-RNA by forcing a concerted flip in the sugar puckers from C2'-*endo* to C3'-*endo*. The A-RNA structure displays features similar to A-form crystal structures, specifically interstrand purine stacking at the central pyrimidine-purine step is observed. When started in a canonical A-form geometry, DNA:RNA hybrid duplexes converge to a structure that is characteristic of experimental solution structures; specifically, a minor groove width intermediate between A-form and B-form geometries, the RNA strand in an A-form geometry, a mixture of C2'-*endo* and C3'-*endo* sugar puckers in the DNA strand, expected distribution of backbone angles and reasonable agreement with the helicoidal parameters is observed. In all of the simulations reported, A-form geometries appear to be less flexible than B-form geometries. There are also significant differences in the patterns of hydration and counterion association between A-form and B-form duplexes. In A-RNA, sodium counterions tend to associate into "pockets" in the major groove whereas these counterions tend to associate into the minor groove in B-form structures.

Introduction:

In order to better understand biological information transfer, molecular interactions of nucleic acids, and the polymorphic character of nucleic acid conformation, it is important to understand the structure, dynamics and relative flexibility of DNA:DNA, RNA:RNA and DNA:RNA duplexes. A better understanding of the differences in sequence specific structure and dynamics can provide insight into protein-nucleic acid interactions, such as why the HIV-1 virus-encoded reverse transcriptase RNase H domain degrades the RNA strand of DNA:RNA hybrids faster than RNA:RNA duplexes (Gotte *et al.*, 1995) and what structural change in DNA:RNA hybrids, compared to duplex DNA, lead to the affinity change of the RNA polymerase core enzyme for the σ subunit (Hansen & McClure, 1980). Flexibility is clearly important in protein-nucleic acid recognition; rigidifying critical residues in the unbound protein can reduce the entropic cost of induced fit, as shown with the interaction of methionyl tRNA synthetase and $tRNA^{Met}$ (Ribas de Pouplana *et al.*, 1996). DNA:RNA complementary hybridization is important in a variety of biological processes including DNA replication (Ogawa & Okazaki, 1980), normal and reverse transcription (Varmus, 1988), and recombination (Daniels & Lieber, 1995). In addition, a better understanding of DNA:RNA hybrid structure is important for antisense drug development, since the potential drug-mRNA complex needs to be recognized by RNase H to allow the RNA to be degraded and the drug to have potent inhibitory activity (Uhlmann & Peyman, 1990).

To date, most of our understanding of nucleic acid structure has come from X-ray crystallographic, and NMR, CD and Raman spectroscopic studies. Theoretical calculations have been of some use; however, earlier simulations employing molecular dynamics methods with an explicit representation of solvent and counterions [see reviews by Beveridge *et al.* (Beveridge & Ravishanker, 1994; Beveridge *et al.*, 1993)] were limited to a short time scale (~100 ps) and during the simulation typically displayed anomalous structure (such as base pair fraying). More recent simulations of nucleic acids with explicit

water on a longer time scale (~1 ns) suggest the importance of properly treating the long ranged electrostatic interactions (Cheatham *et al.*, 1995; Louise-May *et al.*, 1996; Weerasinghe *et al.*, 1995a; York *et al.*, 1995; Zichi, 1995). In addition, there is a dependence of the results on the molecular mechanical force field applied. For example, Yang & Pettitt observed a B-DNA to A-DNA transition (Yang & Pettitt, 1996) when the CHARMM-23 (Brooks *et al.*, 1983) all hydrogen parameter set (Mackerell *et al.*, 1995) was applied with an Ewald treatment on the dodecamer d[CGCGAATTCGCG]$_2$ which suggests that the A form of this structure is more stable. In contrast, B-DNA is more stable than A-DNA when the force field described by Cornell *et al.* (1995) is applied in molecular dynamics simulations with the particle mesh Ewald method (Essmann *et al.*, 1995) within AMBER 4.1 (Pearlman *et al.*, 1995) to a variety of DNA sequences (Cieplak *et al.*, 1996; Duan *et al.*, 1996; Young *et al.*, 1996), including the above dodecamer (Cheatham & Kollman, 1996b).

In this study, comparable simulations with RNA:RNA (r[CCAACGUUGG]$_2$) and DNA:RNA (d[CCAACGTTGG]-r[CCAACGTTGG]) duplexes were performed to determine if we might be able to properly represent the various differences in structure and dynamics among these models. In solution, DNA is expected to be within the larger B-type domain of right handed duplex conformations. Crystallographic and NMR studies clearly demonstrate the heterogeneity in the B-DNA "family" of structures, most notably from sequence specific structure (bending, twisting), various accessible backbone conformations (B_I and B_{II}) and inherent flexibility resulting from sugar repuckering. This flexibility is manifest, not only by noting how easily the DNA can be deformed by crystal packing forces (Dickerson *et al.*, 1987; Dickerson *et al.*, 1994; Ramakrishnan & Sundaralingam, 1993), but since slightly different structures appear when DNA is crystallized into different space groups (Lipanov *et al.*, 1993; Shakked *et al.*, 1989) in contrast to RNA (Portmann *et al.*, 1995). Moreover, B-form structures tend to diffract to lower resolution than A-form structures and B-form fibers have lower crystallinity and

lesser order than A-form fibers (Thomas *et al.*, 1995). The flexibility of B-DNA is further confirmed in NMR experiments which suggest a large range of possible conformations (Fujiwara & Shindo, 1985; Shindo *et al.*, 1985), sugar repuckering (Ulyanov *et al.*, 1995) and (α,γ) backbone "crankshaft" transitions (Ravishanker *et al.*, 1989; Schmitz *et al.*, 1993; Weisz *et al.*, 1994). A comparison of J coupling constants measured by NMR show the enhanced flexibility of DNA duplexes compared to the more rigid RNA duplexes. Theoretical calculations also suggest an inherent flexibility in B-DNA, such as the "substates" of B-DNA conformations suggested by Lavery (Lavery & Hartmann, 1994; Poncin *et al.*, 1992) to the frequent repuckering and (ε,ζ: t, g- to g-, t) backbone transitions observed during molecular dynamics simulations (Cheatham & Kollman, 1996b). RNA duplexes, which are known to adopt a fairly small range of conformations within the A family (A, A'), are generally more rigid than corresponding DNA duplexes, as can be seen in the ^{31}P NMR experiments (Fujiwara & Shindo, 1985; Shindo *et al.*, 1985) and indirectly via crystallography (Portmann *et al.*, 1995). On the other hand, an analysis of crystal structures suggests that double stranded DNA and RNA have a similar level of vibrational motion and sampling of conformational substates (Holbrook & Kim, 1984). However, the simulations reported herein support the idea that A-form structures are more rigid than B-form structures.

Hybrid duplexes with one strand RNA and the complementary strand DNA tend to crystallize in the A-form (Egli *et al.*, 1992; Wang *et al.*, 1982), but in solution are found in a conformation intermediate between an A- and B- form geometry. This is based on fiber diffraction data which suggest a different conformation than true an A or B form geometry for hybrids at high relative humidity (Arnott *et al.*, 1986; Zimmerman & Pheiffer, 1981) with the DNA strands adopting C2'-*endo* and the RNA strands adopting C3'-*endo* sugar puckers. The CD data confirm this picture and further suggest that the overall helix is more A-like than B-like, with positive base pair inclination to the helical axis, small positive roll into the major groove, small positive buckle, negative propeller, and negative x-

displacement from the helical axis (Gray & Ratliff, 1975; Hall & McLaughlin, 1991; Roberts & Crothers, 1992; Steely *et al.*, 1986). Further support comes from the NMR data which clearly show that the RNA strand remains in an A-form geometry with C3'-*endo* puckers throughout, while the DNA strand is in a near B-form geometry with some controversy as to whether the pucker is O4'-*endo* or a mixture of C2'-*endo* and C3'-*endo* (Chou *et al.*, 1989; Fedoroff *et al.*, 1993; Fujiwara & Shindo, 1985; Gao & Jeffs, 1994; Gonzalez *et al.*, 1994; Katahira *et al.*, 1990; Lane *et al.*, 1993; Salazar *et al.*, 1993). The latter is more consistent with the J coupling and dynamics data (Gonzalez *et al.*, 1995). The NMR data also suggest differences in the expected distribution of backbone angles between the DNA and RNA strands in DNA:RNA hybrid duplexes. Beyond the angles directly correlated with the sugar pucker, specifically δ and χ which should be lower in the RNA strand, it is generally observed that α is typically lower, and ε, ζ and γ are slightly higher in the RNA strand than in the DNA strand. All of the experimental data also suggest that the minor groove width in DNA:RNA hybrids is intermediate between A- and B-form duplexes. JUNMA (Lavery *et al.*, 1995) minimizations (Sanghani & Lavery, 1994) and *in vacuo* molecular dynamics simulations (Fritsch & Wolf, 1994) support these observations.

In our simulations, we see the expected structural and dynamic trends. Specifically, we observe that A-RNA duplex structures are stable and within the canonical A family of structures, and moreover display sequence specific features that are consistent with the crystal data, specifically at the central CpG step which has a large rise and low helical twist value. DNA:RNA hybrids also show the expected structural trends with a DNA strand that repuckers between C2'-*endo* and C3'-*endo* sugar puckers, groove widths intermediate to A-form RNA:RNA and B-form DNA:DNA duplexes, expected distributions of backbone angles and reasonable agreement with the helicoidal parameters.

Methods:

The creation of the initial structures, equilibration and dynamics were performed as described in our previous paper (Cheatham & Kollman, 1996b). The starting canonical A- and B- form duplex structures (Arnott & Hukins, 1972) of d[CCAACGTTGG]$_2$, r[CCAACGUUGG]$_2$ and d[CCAACGTTGG]-r[CCAACGUUGG] were generated using the NUCGEN module of AMBER 4.1 (Pearlman *et al.*, 1995). Hydrogens were added with the EDIT module with AMBER 4.1 and the initial hydrogen positions were minimized (*in vacuo*) while holding all non-hydrogen atoms fixed. Care was taken to insure the hydrogens were added with the proper stereochemistry. Explicit net-neutralizing sodium counterions were placed at the phosphates of these models by the EDIT module of AMBER 4.1 and the nucleic acid and 18 counterions were surrounded by a periodic box of TIP3P waters which extended approximately 10 Å (in each direction) from the nucleic acid atoms. This leads to a periodic box size of ~55 Å by ~42 Å by ~42 Å for the B-form structures and ~59 Å by ~40 Å by ~40 Å for the A-form structures. The parameters described by Cornell *et al.* (1995) [see also http://www.amber.ucsf.edu] were used in all of the simulations. All simulations were run using the sander module of AMBER 4.1 with SHAKE (Ryckaert *et al.*, 1977) or SETTLE (Miyamoto & Kollman, 1992) (tolerance = 0.0005 angstroms) on the hydrogens, a 2 fs time step, a temperature of 300K with Berendsen temperature coupling (Berendsen *et al.*, 1984) and a time constant of 0.2 ps, a 9 Å cutoff applied to the Lennard-Jones interactions, and constant pressure with isotropic molecule based scaling (Berendsen *et al.*, 1984) with a time constant of 0.2 ps. The nonbonded list was updated every 10 steps. Equilibration was performed by first holding the positions of the DNA fixed and running 1000 steps of minimization followed by dynamics for 25 ps with a cutoff of 9 Å on all interactions. In order to avoid shifting of the two DNA strand molecules during constant pressure equilibration (when the DNA was held fixed), both strands were treated as if they were a single molecule. After this initial equilibration, all subsequent simulations were run using the particle mesh Ewald method (PME) (Essmann *et al.*, 1995)

within AMBER 4.1 using a cubic B-spline interpolation order and a 10^{-5} tolerance for the direct space sum cutoff. To speed up the fast Fourier transform in the calculation of the reciprocal sum, the size of the PME charge grid is chosen to be a product of powers of 2, 3, and 5 and to be slightly larger than the size of the periodic box. This leads to a grid spacing of ~1 Å or less. Equilibration was continued with 25 kcal/mol-Å^2 restraints placed on all solute atoms, minimization for 1000 steps, followed by 3 ps of MD which allowed the water to relax around the solute. This equilibration was followed by 5 rounds of 600 step minimization where the solute restraints were reduced by 5 kcal/mol during each round. Finally, the system was heated from 100K to 300K over 2 ps and then production runs were initiated. It should be noted that the main goal of the equilibration protocol outlined above is to first let the counterions and water equilibrate, then secondarily let the DNA slowly relax away from the starting geometry to avoid bad contacts, relieve poor bond, angle and dihedral deviations in the model structure, yet help it remain "close" to the initial structure. The most important step in this regard is the initial water and counterion equilibration. To determine if 25 ps is enough time to relax the solvent, the pressure, volume and density are typically monitored. Although not shown, these indicators are easily equilibrated within the 25 ps of water/counterion equilibration. For more discussion about the equilibration protocol used herein, see the presentation available on the AMBER world wide web page at "http://www.amber.ucsf.edu/amber/tutorial/polyA-polyT/".

Since the pairlist is not updated every step, the SHAKE tolerance used is rather modest (0.0005), and constant pressure is utilized, some small energy drain during the simulations can occur. Since uniform scaling of velocities by Berendsen coupling was utilized to bring the very slowly dropping temperature back up to 300K, the center of mass velocity can slowly grow. Therefore, periodically in the simulation (at every restart or every ~80-100 ps) this center of mass velocity was removed during the production dynamics.

Simulations were run on RNA:RNA duplexes starting from canonical A (2030 ps, referred to as A-RNA) and canonical B (2370 ps, referred to as B-RNA) duplexes. After ~1.5 ns with the canonical B start, it was realized that the trajectory had converged to a "B-RNA" conformation that is remarkably close to the average B-DNA structure observed in the corresponding DNA:DNA simulations of $d[CCAACGTTGG]_2$. Therefore, in an attempt to push the "B-RNA" structure away from the B-DNA average structure and perhaps initiate a B to A RNA transition, a simulation was run for 540 ps restarting the trajectory from 1565ps with the temperature increased to 400K. An additional simulation was also started from 1565 ps and run for 1070 ps where a concerted flip in the sugar puckers was forced by applying restraints on the C1'-C2'-C3'-C4' torsion of each nucleotide for a limited time as described below. To determine how to best restrain the C1'-C2'-C3'-C4' torsion to give a particular sugar pucker pseudorotation value, calculations on adenine nucleotides *in vacuo* and in solution were run with various force constants on the restraints. In order to maintain reasonable distributions of the sugar pucker about the mean pucker pseudorotation value, flatwell restraints were applied. To constrain the pucker to C3'-*endo*, flatwell restraints are applied with no penalty between 30° and 40°, parabolic penalties from 30° to 20° and 40° to 50°, and linear penalties outside this range. To constraint the pucker to C2'-*endo* the flatwell restraint is applied between -40° to -38°, parabolic penalties from -40° to -44° and -38° to -34°, and linear penalties outside this range. The typical force constant necessary to "restrain" the pucker to the appropriate range is on the order of 30 kcal/mol-radian^2. However, to force a concerted flip larger restraints were necessary; the goal was to allow a quick concerted flip in the pucker such that the biasing restraints could then be turned off. In this simulation, a concerted flip from the "B-RNA" C2'-*endo* puckers to C3'-*endo* was forced by gradually increasing the restraint penalty force constant from 0 to 300.0 kcal/mol-radian^2 over 5 ps, then gradually reducing this restraint from 300.0 kcal/mol-radian^2 to 30.0 kcal/mol-radian^2 over the next 45 ps. After this time, all the restraints were turned off, and free dynamics

were continued. Note that a simulation where the force constant on the restraints were held at 30.0 kcal/mol-radian^2 for 25 ps initially, then removed, was not sufficient to completely force a concerted flip in the puckers.

Simulations were also run both on canonical A (2005 ps, referred to as A-hybrid) and canonical B (2045 ps, referred to as B-hybrid) forms of DNA:RNA hybrid d[CCAACGTTGG]-r[CCAACGUUGG] duplexes. The canonical B-form DNA:RNA hybrid simulation was continued from 2045 ps for ~400 ps at 400K. Simulations were also run, and some of the results previously reported (Cheatham & Kollman, 1996b) on canonical A and canonical B (1400 ps, referred to as A-DNA and B-DNA, respectively) models of d[CCAACGTTGG]$_2$ duplexes.

All of the results were analyzed using the carnal, anal, nmode, and mdanal modules of AMBER 4.1, the Dials and Windows (Ravishanker *et al.*, 1989) interface to Curves (Lavery & Sklenar, 1988), a more recent version of Curves, version 5.1 dated June 1996, or some adapted trajectory analysis software (rdparm). Standard angle (α, β, γ, δ, ε, ζ, χ) (Saenger, 1984) and helicoidal parameter (Dickerson, 1989) names and definitions are presented in the analysis. Sugar pucker pseudorotation values and sugar pucker amplitudes were calculated based on the Altona & Sundaralingam conventions (Altona & Sundaralingam, 1972); in the text "sugar pucker" or "pucker" will be used synonymously with "sugar pucker pseudorotation phase". Nucleic acid residue names are referred to in the text as one letter codes. Where necessary, a subscript for the residue number is also presented; the residue number is in the 5' to 3' direction with the first strand numbered 1-10 and the second strand 11-20. To avoid confusion between base pairs and base pair steps in the text, base pair steps are denoted with a "p", *i.e.* TpG steps in contrast to TG base pairs.

Average structures from the trajectories were calculated using the carnal module of AMBER to coordinate average the RMS coordinate fit frames (over all DNA atoms) taken at 1 ps intervals. No extra processing of these average coordinates (*i.e.* minimization) was

performed. Since these structures were not minimized post averaging, they may contain some anomalous structural features, such as is exemplified with DNA thymine methyls which will on average display each hydrogen co-linear with the C-C bond. This may lead to higher calculated RMSd values and slight differences in the calculated helicoidal parameters. However, minimization to fix up the structure is tricky since it is impractical to include water in the calculation. Without water to balance the interactions, *in vacuo* minimization will distort the structure significantly from what is observed during the dynamics with explicit water. In order to investigate the effect of minimization on the average structures, two short minimizations of the B-form RNA:RNA duplex average structure over 1370-2370 ps, one with a constant dielectric constant and the other using a distance dependent dielectric function with a dielectric constant of 4, were run where the energy was minimized until the RSMd in energy between steps changed by less than 1.0 kcal/mol. These short minimizations led to structures that were only 0.37 Å and 0.22 Å away from the average structure, respectively. Despite the rather small root-mean-square deviation between the structures, there are small, but significant, differences in the backbone angles (less then 5°) between the average and minimized structures and a higher average sugar pucker amplitude (43.0°) in the minimized structure. This is not surprising since during the minimization process, the dihedral angle values will tend to move towards minimizing their deviation from the equilibrium force field values. Although the helicoidal parameters are very sensitive to the base atom positions, the differences in the helicoidal parameters between the minimized and non-minimized structures are significantly smaller than the standard deviations of the values over the trajectory. Given these small differences between the minimized and average structures, and considering the fairly low all atom self symmetric RMSd values (ssRMSd, A-RNA is ~0.34 Å and B-RNA is ~1.0 Å) of the average structures, the non-minimized average structures calculated will be used in the analysis presented herein. The self symmetric RMSd (ssRMSd) values are defined herein for duplexes in which both strands have the same sequence; the ssRMSd is the RMSd of

the structure to the symmetric structure obtained by rotating the duplex to match the second strand to the first.

Diffusion constants were calculated using the Einstein relationship (or the slope of the mean square displacements in angstroms versus time in picoseconds multiplied by 10.0/6.0 which leads to units of 10^{-5} cm^2/s) using the software rdparm. The average self diffusion constants of TIP3P water and Na+ counterions in these simulations (A-RNA, B-RNA, A-hybrid, B-hybrid and B-DNA) over the course of a nanosecond are 4.7-4.9 x 10^{-5} cm^2/s and 1.2-2.1 x 10^{-5} cm^2/s respectively. This is slightly lower than the calculated values for pure TIP3P as is expected due to some condensation of the ions and water with the nucleic acids. This data is presented to here to show that diffusion is not seriously inhibited by the imposition of true periodicity [see Essman *et al.* (1995) for a more thorough analysis of water diffusion with and without Ewald]. Atomic positional fluctuations were determined over nanosecond portions of the trajectory using the mdanal module from AMBER 4.1. Normal mode calculations were performed using the nmode module from AMBER 4.1.

Solvent and counterion distributions were calculated by binning atom positions from RMS coordinate fit frames over all DNA atoms at 1 ps intervals into 0.5 $Å^3$ grids over 1 nanosecond portions from the trajectories. In other words, the value of each grid element represents the number of times the coordinates of the center of a particular atom of interest (*i.e.* water oxygen) were within the 0.5 $Å^3$ represented by that particular grid element. These grids can then be contoured using the density delegate of UCSF MidasPlus (Ferrin *et al.*, 1988). For 1000 frames, the expected number of waters per grid element, assuming bulk water density, is 4.18. In the graphics of the water and counterion hydration presented, the contouring of the water/counterion density was typically performed at 12.0 or 15.0 hits per 0.5 $Å^3$, or approximately three times expected bulk water density. In the text, the contouring level will be referred to as "*x* hits per 0.5 $Å^3$" which represents *x* visits to each 0.5 $Å^3$ grid from 1000 frames of the trajectory taken at 1 ps intervals.

All the molecular graphics images herein were produced using the MidasPlus software available from the Computer Graphics Laboratory, University of California, San Francisco. All the molecular dynamics calculations were either run on an SGI R8000 at UCSF or 16 processors of the Cray T3D at the Pittsburgh Supercomputing Center using a modified version of the sander module of AMBER 4.1. The Cray T3D parallel version was adapted from the MPI version of sander originally developed by Vincent & Merz (Vincent & Merz, 1995) and incorporated into AMBER 4.1. Parallelization of the particle mesh Ewald code specifically for the Cray T3D and also more generally under MPI was performed by Michael Crowley of the Pittsburgh Supercomputing Center. Approximately 1 week on 16 processors of the Cray T3D or 17.5 days on 1 processor of the 75 MHz SGI R8000 is required to simulate each system for 1 nanosecond.

Results and Discussion:

RNA maintains a stable A-RNA structure with features similar to A-form nucleic acid crystal structures:

In the simulations we find that when the RNA duplex is started in a canonical A geometry it remains in a canonical A geometry. The A-form RNA:RNA duplex (A-RNA) adopts a "stable" average structure over the last nanosecond of a ~2 ns simulation that is within ~2 Å of canonical A RNA. The convergence to this average structure is rather good, as judged by comparing the "self" symmetric root mean square deviation (ssRMSd) which is only 0.34 Å. The data in Table 7 shows that the A-RNA remains very close to canonical A. Moreover, the data in Table 8 show that the structure is characteristic of the canonical A family of RNA structures (Saenger, 1984).

	A	B	B-DNA	A-RNA	B-RNA	A-hybrd	B-hybrd
A	*	3.81 3.79 5.80	3.23 3.10 3.86	1.74 1.88 2.09	3.14 3.50 3.82	2.15 1.72	3.41 3.35
B	2.68 2.70 3.29	*	1.74 1.94 3.02	3.39 3.44 5.41	2.03 2.18 3.47	2.65 3.32	1.71 1.93
B-DNA	2.28 2.25 2.61	1.13 1.29 1.69	*	2.36 2.35 2.97	0.81 0.85 0.95	1.46 2.34	0.80 0.90
A-RNA	1.19 1.14 1.64	2.49 2.58 3.18	2.00 1.96 2.17	*	2.23 2.45 2.72	1.03	2.44
B-RNA	2.37 2.46 2.80	1.23 1.51 2.04	0.61 0.70 0.76	2.00 2.01 2.13	*	2.64	1.12
A-hybrd	1.55 0.96	1.85 2.37	1.17 1.84	0.79	2.03	*	1.83 2.39 2.56
B-hybrd	2.20 2.43	1.19 1.17	0.36 0.75	2.08	0.86	1.09 1.99 1.75	*

Table 7: Root mean square deviations (RMSd) of all atoms (not mass weighted) in various structures in angstroms. The upper triangle is over all residues, the lower triangle is the internal 6 residues from each strand. For each comparison, three numbers are presented, where applicable. The top number is the RMSd of the first strand (residues 1-10 in the upper triangle or residues 3-8 in the lower triangle), the middle number is the second strand (residues 11-20 in the upper triangle and residues 13-18 in the lower triangle), and the bottom number is the RMSd of both strands. In the case of the DNA:RNA hybrids, the DNA strand is always the first strand. The canonical A and canonical B models were built into both DNA:DNA and RNA:RNA duplexes as discussed in the methods. The average structures are generated by a straight average of the RMS fit coordinate streams taken at 1 ps intervals as discussed in the methods; no minimization of these structures was performed. The average DNA model, B-DNA, is described in our previous paper (Cheatham & Kollman, 1996b) and represents the B2 trajectory; the average is over 0-1400 ps. A-RNA and B-RNA represent the canonical A and canonical B starts of the RNA:RNA duplex over 1030-2030 ps and 1370-2370 ps, respectively. The A-hybrd and B-hybrd average structures are from the canonical A and canonical B DNA:RNA hybrid duplex simulations over 1005-2005 ps and 1045-2045 ps, respectively.

	A-RNA 1030-2030ps	B-RNA 1370-2370ps	A-hybrid 1005-2005ps	B-hybrid 1045-2045ps	B-DNA 400-1400ps
α (°)	277.0 (10.5)	286.3 (12.5)	279.9 (10.6)	288.1 (15.0)	290.4 (11.6)
β (°)	175.6 (9.5)	173.3 (12.8)	173.8 (9.9)	171.6 (12.6)	168.4 (12.6)
γ (°)	69.8 (8.9)	52.7 (11.9)	66.0 (9.4)	55.4 (13.9)	54.3 (10.6)
δ (°)	79.3 (8.2)	127.7 (15.1)	95.2 (13.2)	120.2 (16.9)	116.6 (18.0)
ε (°)	201.6 (10.1)	200.3 (18.4)	194.5 (9.7)	194.7 (19.9)	197.0 (18.5)
ζ (°)	291.4 (8.6)	247.1 (26.0)	281.8 (11.3)	257.9 (26.2)	258.0 (25.5)
χ (°)	201.7 (9.1)	248.7 (17.5)	214.9 (12.5)	241.4 (18.6)	234.2 (16.7)
pucker (°)	22.6 (16.5)	137.0 (24.4)	66.8 (22.2)	127.6 (27.4)	122.8 (28.1)
amplitude (°)	38.9 (5.0)	40.2 (5.2)	38.0 (5.4)	37.8 (5.8)	37.1 (6.1)
propeller (°)	-12.6 (12.0)	-10.4 (13.6)	-13.2 (11.4)	-13.0 (13.5)	-10.4 (12.3)
buckle (°)	-0.8 (10.6)	1.2 (13.5)	4.3 (10.7)	-1.7 (13.7)	0.4 (11.4)
opening (°)	3.3 (6.0)	4.5 (7.5)	1.9 (5.9)	4.0 (8.2)	2.0 (5.6)
rise (Å)	2.7 (0.6)	3.2 (0.6)	3.0 (0.5)	3.2 (0.6)	3.3 (0.5)
tilt (°)	-0.5 (5.4)	-0.7 (6.7)	0.0 (5.4)	-1.1 (6.8)	0.0 (5.4)
roll (°)	2.4 (7.7)	2.2 (9.5)	1.6 (7.9)	2.6 (8.9)	1.3 (8.7)
twist (°)	30.9 (4.3)	30.6 (5.4)	30.5 (4.2)	30.6 (6.5)	30.9 (5.1)
x-disp (Å)	-5.2 (0.8)	-3.1 (0.9)	-4.5 (0.8)	-2.7 (1.2)	-3.0 (0.7)
y-disp (Å)	0.1 (0.7)	0.5 (0.7)	-0.3 (0.6)	-0.1 (1.3)	0.0 (0.5)
inc (°)	15.0 (9.3)	9.8 (10.8)	11.1 (9.2)	6.1 (13.5)	4.9 (7.3)
tip (°)	-1.9 (6.7)	-2.4 (7.9)	4.6 (6.1)	-1.8 (13.5)	0.4 (5.8)
shear (Å)	0.0 (0.5)	0.1 (0.5)	0.0 (0.4)	-0.1 (0.5)	0.0 (0.4)
stretch (Å)	0.4 (0.4)	0.3 (0.3)	0.2 (0.3)	0.2 (0.3)	0.1 (0.2)
stagger (Å)	-0.2 (0.5)	-0.2 (0.6)	-0.2 (0.5)	-0.2 (0.5)	-0.2 (0.5)
shift (Å)	0.0 (0.6)	0.0 (0.7)	0.0 (0.6)	0.1 (0.7)	0.0 (0.6)
slide (Å)	-0.1 (0.4)	-0.1 (0.5)	-0.1 (0.4)	-0.1 (0.5)	-0.1 (0.4)

Table 8: Standard angle and helicoidal values and standard deviations (in parenthesis) averaged over all the residues, base pairs or base pair steps (where appropriate) for the various duplex structures specified[18].

[18] The average values were calculated by determining the values for each average structure. The average structure is calculated by best fitting the RMSd over all atoms (mass weighted) at 1 ps intervals and performing a straight coordinate average. Each average structure represents a nanosecond portion of the trajectory (as specified). The values in parenthesis represent the standard deviation of the values calculated for each 1 ps frame during the nanosecond portion of the trajectory specified. The "A" or "B" above denotes the starting conformation, which is canonical A or canonical B respectively. All the values were calculated using the dials and windows (Ravishanker et al., 1989) interface to Curves (Lavery & Sklenar, 1988) or a more recent version of Curves, version 5.1, dated June 1996.

This average A-RNA structure has an average helical inclination of 15.0°, x-displacement from the helical axis of -5.2 Å, C3'-*endo* sugar puckers, and the rise between the base pairs of 2.7 Å. Interestingly, the A-RNA average structure does not have the sequence specific bending patterns, such as the notable roll into the major groove at the TpG and CpG steps, seen in comparable B-DNA simulations (Cheatham & Kollman, 1996b). Instead, as shown by the dark black lines in Figure 20, we see a generalized base pair roll into the major groove and relatively more uniform base pair propeller and buckle. Although the average structure appears to be close to canonical A, there are some notable deviations. In particular, a recent analysis of individual dimer steps from DNA crystal structures clearly demonstrate that A-form and B-form structures show little overlap in the respective helicoidal parameters, particularly in slide and roll (Gorin *et al.*, 1995). In this analysis, Gorin *et al.* show that A-form structures typically display a more negative average slide and higher average roll. Our slide versus roll values for the A-RNA simulation tend to fall into the "B-DNA" region, or on the boundary of the B-DNA region, depending on exactly how the values are calculated (discussed in more detail below). Overall, our slide values are of a slightly lower magnitude than has been observed in various A-DNA (Haran *et al.*, 1987; Ramakrishnan & Sundaralingam, 1993) or A-RNA (Portmann *et al.*, 1995) crystal structures. The significance of these small deviations is difficult to judge, since we are comparing a model of a "solution" structure of A-RNA to data derived from the analysis of A-DNA crystal structures.

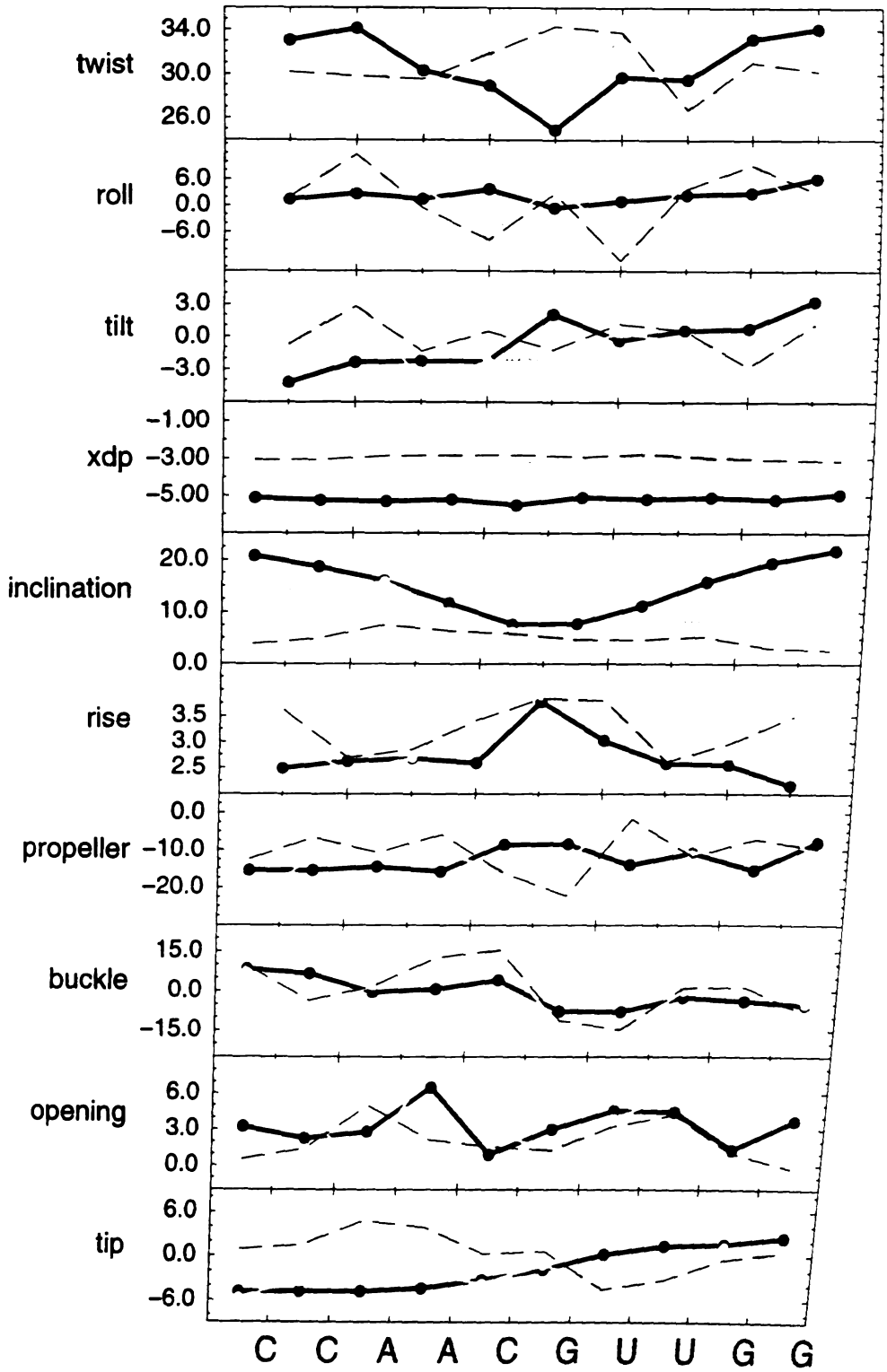

twist
34.0
30.0
26.0
roll
6.0
0.0
-6.0
tilt
3.0
0.0
-3.0
xdp
-1.00
-3.00
-5.00
inclination
20.0
10.0
0.0
rise
3.5
3.0
2.5
propeller
0.0
-10.0
-20.0
buckle
15.0
0.0
-15.0
opening
6.0
3.0
0.0
tip
6.0
0.0
-6.0
C C A A C G U U G G

Figure 20 (previous page): Helicoidal parameters calculated with the dials and windows (Ravishanker et al., 1989) interface to Curves (Lavery & Sklenar, 1988) for average structures from the trajectory. The canonical A-start RNA:RNA duplex average structure from 1030-2030 ps (A-RNA) is represented in black, the canonical B-start RNA:RNA duplex average structure from 1370-2370 ps (B-RNA) in gray and the canonical B start DNA:DNA duplex average structure from 400-1400 ps (B-DNA) is shown as a dotted black line[19].

Beyond the general observations, the most notable sequence specific deviations from canonical A-RNA can be seen by examining the dark black lines in Figure 20. In particular, the central CpG step shows an anomalous rise of 3.77 Å and a low helical twist of 25.0°, with correspondingly lower inclination values of ~8° and relatively less propeller twisting of ~ -8°, at the central CG and GC base pairs, and a negative cup (or difference between the CG and GC base pair buckle) equal to -11.7°. Neglecting for a moment the lower than expected inclination and the absence of a more positive roll at this step (instead a lower roll of -0.5° is observed), the "anomalies" noted above are characteristic of a "low twist profile" base pair step which is expected for CpG steps (Yanagi *et al.*, 1991). Moreover, an analysis of canonical A-forms does clearly show the negative correlation between rise and inclination (Babcock & Olson, 1994). This is expected since A-form structures compensate for the lower rise between base pairs by inclining the base pairs (Yanagi *et al.*, 1991). The only unexpected observation in these values is the low roll (into the minor groove) at this step, since roll is inversely correlated with twist.

[19] The average structures were created by averaging all the coordinates of the nucleic acid from RMS fit frames taken at 1 ps intervals. The twist, roll, tilt, inclination, propeller twist, buckle, opening and tip are all represented in degrees and the x-displacement from the helical axis and rise are represented in angstroms. The values are presented traversing the helix from left to right representing the 5' to 3' direction. Note that although the x-axis legend specifies a "U" for uracil, thymine nucleotides were used in the DNA strands.

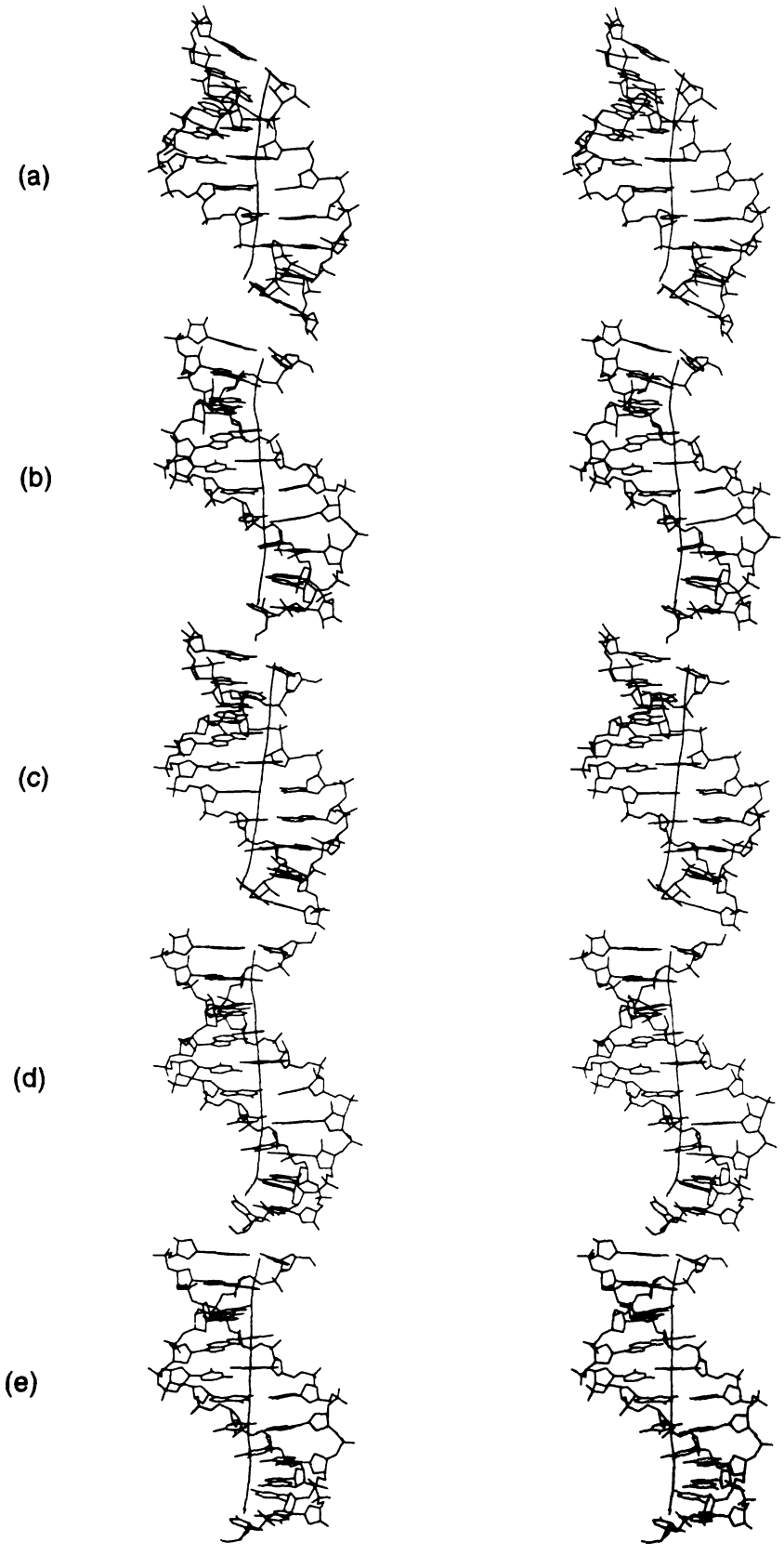

(a)
(b)
(c)
(d)
(e)

Figure 21 (on previous page): Average structures and the global helical axis (calculated from Curves 5.1 (Lavery & Sklenar, 1988)) for all of the average structures represented in the calculations described herein are plotted in stereo. Each structure represents the final nanosecond from their respective trajectories and are calculated from a straight coordinate average over all nucleic acid atoms from RMS fit coordinate frames taken at 1 ps intervals. All of the nucleic acid atoms, except the hydrogens, are displayed. All the plots were created using MidasPlus (Ferrin et al., 1988). (a) A-RNA: canonical A start of the RNA:RNA duplex over 1030-2030 ps. (b) B-RNA: canonical B start of the RNA:RNA duplex over 1370-2370 ps. (c) A-hybrid: canonical A start of the DNA:RNA duplex over 1005-2005 ps. (d) B-hybrid: canonical B start of the DNA:RNA duplex over 1045-2045 ps. (e) B-DNA: canonical B start of the DNA:DNA duplex over 400-1400 ps.

Figure 21a displays a stereoview picture of the A-RNA average structure and the calculated global helical axis. Traversing the helical axis down the sequence from top to bottom, it appears to curve one direction, until the CpG step is reached (in the middle), where the direction changes, then it curves off in another direction at the end. Based on the Curves analysis, the largest deviation from true helicity occurs at this central CpG step. Looking at the base pair stacking, it appears that the cytosines of the CG base pairs are relatively unstacked and the two guanines from opposite strands have positively shifted and negatively slid so as to partially stack on top of each other. This result is very similar to what has been seen in the A-DNA crystal structures of $d[CCCCGGGG]_2$ (Haran *et al.*, 1987), $d[GGGCGCCC]_2$ (Shakked *et al.*, 1989) and other octamers (Eisenstein & Shakked, 1995; Ramakrishnan & Sundaralingam, 1993) where interstrand stacking of the central guanines from opposite strands is observed and low twist and more slide (both of which improve the interstrand stacking), low propeller twist (which reduces interstrand steric clashes (Calladine, 1982)) and low roll angle (which compensates for the slide) are observed. The helicoidal structure at the CpG step is locally perturbed in order to more favorably stack at the intrinsically weak 5' to 3' pyrimidine on purine step. It is also very interesting that we observed a spontaneous crankshaft transition in the α and γ backbone angles (see Figure 22) from the more common *gauche-*, *gauche+* state to *trans*, *trans* values at ~872 ps between the CpG step of the first strand which serves to increase the base pair separation (Olson, 1982b) and may serve to improve the cross-strand overlap of the

guanine bases (Haran *et al.*, 1987). The all trans conformation of the P-O5'-C5'-C4' bonds, low twist, negative slide and interstrand guanine stacking at the central CpG steps has been observed in a variety of A-DNA tetragonal crystal structures (Eisenstein & Shakked, 1995; Haran *et al.*, 1987) and in one strand of the rhombohedral crystal structure of $r[CCCCGGGG]_2$ (Portmann *et al.*, 1995).

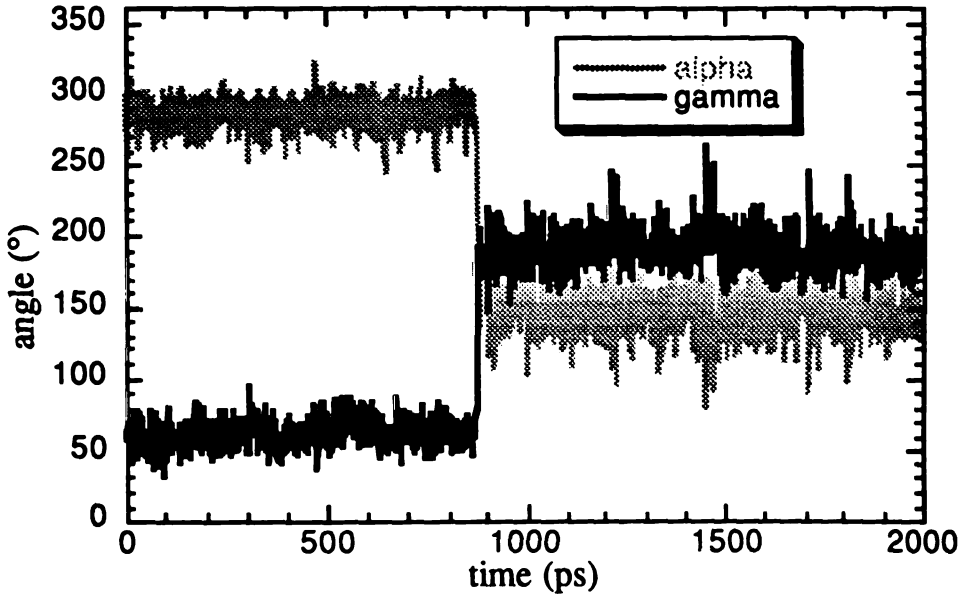


Figure 22: Plot of the alpha (α, O3'-P-O5'-C5', in black) and gamma (γ, C5'-C4'-C3'-O3', in gray) backbone torsions angles as a function of time for the CpG step from first strand in the A-RNA (canonical A start of the RNA:RNA duplex) trajectory.

It is significant that the interstrand guanine stacking, and single crankshaft (α,γ) transition, is observed spontaneously at the central CpG step in solution phase simulations of a decamer. All of the tetragonal octamer A-DNA crystal structures analyzed by Ramakrishnan & Sundaralingam (Ramakrishnan & Sundaralingam, 1993) have a pyrimidine-purine, and most commonly a CpG, base pair step in the center of the duplex; sequences with other types of step at the center do not crystallize into the tetragonal form. Moreover, each of these tetragonal structures has a low twist at the central pyrimidine-purine step. In addition to the above mentioned features, there is also postulated a particular hydration pattern in the tetragonal crystal structures; specifically a chain of water molecules across the CpG step that links the minor groove to the backbone of other duplexes in the unit cell is observed which may contribute to the stability of the tetragonal geometry (Eisenstein & Shakked, 1995). It is not clear if other sequences avoid crystallizing in the tetragonal group because they do not have the deformable weak stacking interaction found in pyrimidine-purine steps and the water stabilizing crystal contacts which allow for facile crystallization or alternatively if the interstrand stacking of the CpG step is an inherent contextual sequence specific structure. Support for the hypothesis that it is the deformability of this step, induced by crystal packing, that leads to the low twist and interstrand purine stacking comes from analysis of A-DNA crystal structures, which demonstrate that the low twist occurs only at pyrimidine-purine steps at the center of the helix and not at other pyrimidine-purine steps (Ramakrishnan & Sundaralingam, 1993). However, the spontaneous observation of low twist and interstrand guanine stacking at the CpG steps without the characteristic crystal packing and hydration patterns (discussed later) in our simulations demonstrate that the crystal packing forces alone do not induce the central CpG step behavior. Instead, the interstrand stacking is a real contextual sequence dependent structural effect.

The central pyrimidine-3'-5'-purine step is clearly more deformable, as results of crystallizing the r[CCCCGGGG]$_2$ (Portmann *et al.*, 1995) and d[GTGTACAC] (Jain & Sundaralingam, 1989) sequences in two distinct lattices attest. The CpG steps in our simulations show somewhat enhanced flexibility which suggests a greater deformability. In the first strand (where the crankshaft transition to *trans*, *trans* is observed in the α and γ angles), the standard deviations in α (15.2°), β (10.5°) and γ (13.1°) are significantly larger than the average fluctuations at these angles (Table 8) and the fluctuations in the helical twist and rise at this step are more than 10% above the average. Given the above, the question still remains as to why this behavior is only observed at the central, and not flanking, pyrimidine-purine steps.

Analysis of average structures vs. averages of the analysis of snapshots from the trajectory:

Interpreting and judging the validity of correlations found in the helicoidal parameters is often difficult. While some correlations in the helicoidal parameters are expected and structurally significant, some may result from the methods used to calculate the helicoidal parameters (Babcock & Olson, 1994), some may be an artifact of insufficient sampling in the simulation leading to a misrepresentative average structure, and some may be an artifact of the force field representation. One means of checking the consistency of the average structure is by comparing the average of all the values calculated from analysis of each individual snapshot in the trajectory to those values calculated from the one average structure. To avoid confusion in the description, the former will be referred to as the "mean of the snapshots" and the latter as the "average". When comparing the mean of the snapshots to the average, all the backbone angles and sugar puckers are all respectively within ~1° of each other. The agreement seen here is even better than the agreement between the average structure and the average structure after minimization, as discussed in

the methods section. However, the sugar pucker amplitude appears to be slightly lower in the average than is expected based on averaging the amplitudes from each snapshot; this observation is fairly easy to rationalize based on the flattening of the ring during coordinate averaging of the various sugar puckers. In the case of sugar amplitude, the mean of the snapshots seems more appropriate.

A more significant difference between the mean of the snapshots and the average is found for the helicoidal parameters. Specifically, the average x-displacement is 1.52 Å higher, the average inclination ~10.6° lower, the average roll ~5.6° larger and the average rise between base pairs 0.3 Å higher when calculated using the mean of the snapshots from the A-RNA simulation over 1030-2030 ps. These differences are actually quite large and in the case of the x-displacement and inclination, greater than the standard deviations. The differences in rise and inclination alone correspond to the difference between A-RNA and A'-RNA (Saenger, 1984). These differences are not restricted to the A-RNA simulation, but are seen in all of the simulations run. It is not clear what the significance of this is, however it is worth pointing out since analysis of the snapshots (*i.e.* as with dials and windows) is actually quite common. These differences may suggest that we have not sampled long enough for the time average to converge. Alternatively, there could be a systematic difference in the two types of analysis (such as with the averaging of the sugar pucker amplitudes). Either way, the differences we observe demonstrate the sensitivity in the calculation of the helicoidal parameters to the structure. It should also be noted that although the values are shifted, depending on the way the data is analyzed, the trends or relative values for various base pairs or base pair steps are still maintained. In other words, the anomalous rise and lower inclination at the CpG step in the A-RNA simulation is still apparent whether the mean of the snapshots or the average structure is analyzed, it is just the actual value that is shifted. This implies that either the average or the mean of the snapshots is appropriate for analyzing sequence dependencies as long as the analysis is internally consistent. In Table 8, we present the analysis of the average structure since this

nanosecond time averaged structure is perhaps more representative of what is observed experimentally by NMR or crystallography. The standard deviations, on the other hand, can only be estimated by analyzing a series of individual frames.

B-RNA is also stable on the 1-2 nanosecond time scale:

When the simulation is started in a canonical B form, we do not observe a spontaneous B to A transition in the RNA:RNA duplex simulations. Instead, snapshots from the trajectory remain in the B family for over 2 ns and move towards an average structure that is exceptionally close to the "average DNA" structure calculated for DNA:DNA duplexes of the same sequence (Cheatham & Kollman, 1996b). As shown in Table 7, the RMSd of the canonical B start RNA:RNA duplex over 1370 - 2370 ps to the canonical B start DNA:DNA duplex over 400 - 1400 ps is ~1 Å and the "B-RNA" self agreement (ssRMSd) is 1.0 Å. Sugar repuckering from C2'- *endo* to C3'-*endo* (and back) does occur in "B-RNA", but at a much lower rate than is observed in the DNA simulations. Helicoidal analysis of the average B-RNA structure displayed in Figure 20 in gray compared to the DNA:DNA average B duplex structure (dotted black line) shows remarkable agreement between the two structures. Except for the base pair inclination at the terminal base pairs, all of the helicoidal parameters for the B-RNA and B-DNA structures (Figure 20, Table 8) are in the expected range. Although the x-displacement is slightly lower (~ -3 Å) than is expected for B-form structures (0 to -2 Å), if the mean of the snapshots over the same nanosecond portion is used to calculate the value, the x-displacement moves into the expected range (~ -1.3 Å) (Cheatham & Kollman, 1996b).

The largest discrepancies between the B-RNA and B-DNA structures are in the base pair inclination and tip. Since the inclination at the terminal base pairs has risen above the average B-DNA values and moved closer to A-RNA values, the tip values more closely resemble the A-RNA values (black line) and since some of the puckers do display some

transient C3'-*endo* puckers, perhaps it is simply necessary to continue the simulation for a longer time in order to see the B to A transition. The fact that we can stably simulate a B-RNA structure for longer than 2 nanoseconds suggests that the B-RNA structure is certainly a minimum energy conformation with the Cornell *et al.*. (1995) force field. JUMNA (Lavery *et al.*, 1995) minimization's by Lavery (Lavery, 1996), with both the Cornell *et al.*. (1995) and the standard JUNMA force fields, support this observation. If, as suggested by Olson *et al.* (Olson & Sussman, 1982), the barrier to sugar repuckering is on the order of ~4 kcal/mol for RNA and ~2 kcal/mol for DNA, the time for an A to B or B to A transition should be $\sim e^{-\Delta\Delta E^{\ddagger}/RT}$ or approximately 20 times longer for RNA. This implies a time scale on the order of ~10 nanoseconds based on the observation of a A-DNA to B-DNA transitions in ~500ps (Cheatham & Kollman, 1996b).

Can we force a B-RNA to A-RNA transition?

The B-RNA simulation displays more frequent C2'-*endo* to C3'-*endo* sugar repuckering with longer lasting C3'-*endo* puckers compared to the infrequent and short-lived C2'-*endo* puckers observed during C3'-*endo* to C2'-*endo* repuckering in the A-RNA simulation. This suggests that the transition to C3'-*endo* from C2'-*endo* in B-RNA is easier than the C3'-*endo* to C2'-*endo* repuckering in A-RNA. Moreover, looking at the black line in Figure 23 which displays some of the common indicators of A vs. B form geometry as a function of time, transient spikes in the inclination and dips in the rise and x-displacement, such as just after 500 ps, are observed which indicate the structure becomes more "A-form" like during the B-RNA simulation. These observations indirectly suggest that the A-RNA structure is more stable, but that the barrier allowing the concerted repuckering necessary for a B-RNA to A-RNA transition is too high and cannot be surmounted in 1-2 nanosecond simulations. To investigate this, simulations were run at a higher temperature, 400 K, which should allow for more frequent repuckering. Interestingly, although the increase in

temperature to 400K does lead to a slight increase the rate of sugar repuckering, the repuckering is not nearly as pronounced as is observed in the DNA simulations. Although the temperature does not significantly increase the rate of sugar repuckering, it does lead to transient breaking of the Watson-Crick base pairs and a correspondingly large increase in the fluctuations within the helicoidal parameters. The largest disruption in the structure comes from terminal base pair fraying on one side of the helix where approximately halfway through the simulation, the terminal CG base pair breaks and the base pairs from across the strand pack on top of each other. Given the disruption in the structure, it appears that 400 K is too "hot" for these simulations. While it may be possible to run at an elevated temperature somewhere between 300 K and 400 K that still maintains the base pairing, it is not anticipated that this will provide for a significant enough rate enhancement in the sugar puckering to allow a concerted change to C3'-*endo* on a nanosecond time scale.

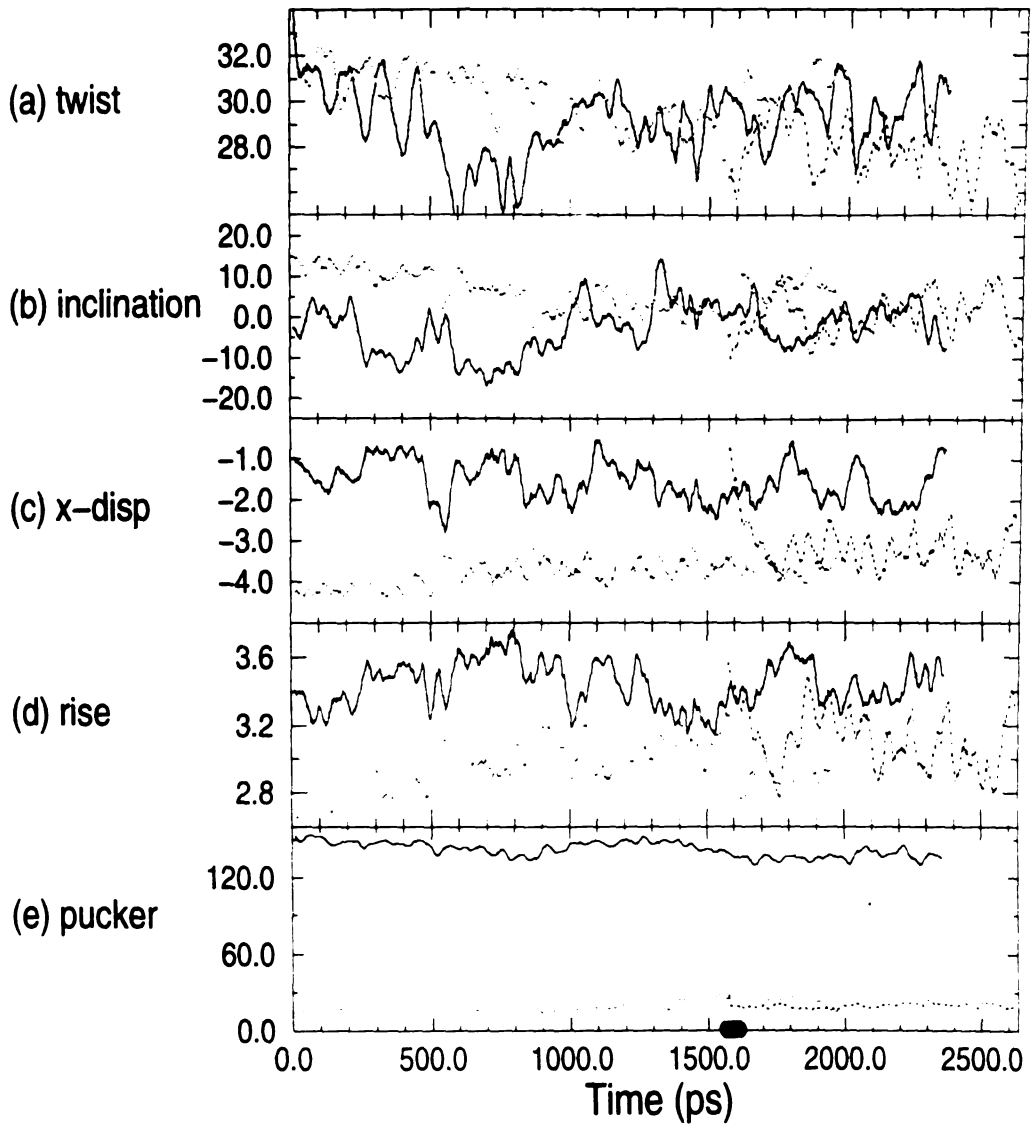


Figure 23: Plot of selected helicoidal parameters versus time for the B–RNA (black), A–RNA (gray), and the B–RNA simulation continued from 1565 ps where a concerted flip in the puckers was forced over 50 ps (dashed black). The helical twist, base pair inclination and sugar pucker pseudorotation phase (pucker) are in degrees and the rise between base pairs and x–displacement from the helical axis are in angstroms. The data represents an average over all nucleotides, base pairs, or base pair steps, as appropriate, and has been smoothed by performing a running average over 25 ps. The black oval along the x–axis of the pucker graph represents the time over which the restraints on the C1'–C2'–C3'–C4' torsions were applied to induce a concerted flip in the sugar puckers.

To see if we could observe a B-RNA to A-RNA transition more directly, a simulation was run where a concerted change in the pucker to C3'-*endo* was forced. As discussed in the methods section, this was accomplished by forcing the C1'-C2'-C3'-C4' torsion over a short period of time (~50ps), to values which lead to C3'-*endo* puckers. As Figure 23 shows, when the C3'-*endo* restraints are applied, the rise and x-displacement (dotted lines) move rather quickly (over approximately 250 ps) from the B-RNA values (black) to the A-RNA values (gray). The restraints were only applied for 50 ps (a time corresponding to the back oval on the x-axis of the pucker graph) and after the restraints were removed, all the puckers remained C3'-*endo* for the ~1 ns simulation (except for one short C3'-*endo* to C2'-*endo* repuckering event at one of the terminal guanines). An average structure calculated from the trajectory, starting after the first 70 ps and representing one nanosecond of simulation, converged to within 1.07 Å of the average structure calculated over the last nanosecond of the A-RNA trajectory. Convergence to this average structure was not as close (ssRMSd ~0.64 Å) as what was seen in the A-RNA simulation (which had a ssRMSd ~0.34 Å). A major difference in these average structures is the backbone conformations.

As discussed previously, one of the strands in the A-RNA simulation displayed a crankshaft (α, γ: g-,g+ to t,t) transition between the CpG step (Figure 22); it was claimed that this allows for better cross-strand overlap (Haran *et al.*, 1987) and increased separation of the base pairs which lead to better interstrand guanine overlap. In the simulation with the concerted pucker flip, instead of the (α, γ) crankshaft transition at CpG step, it is observed at the adjacent ApC step in both strands. Moreover, the second strand additionally displays a crankshaft transition at the UpG step and a *trans* γ at the first cytosine; the latter was also observed in the second strand of the A-RNA average structure. The crankshaft transition at the ApC step in both strands occurs within the first 3-4 ps as the sugar repuckers to C3'-*endo* under the influence of the applied restraints. The sudden transition in the sugar pucker to C3'-*endo* also effects the α, ε and ζ backbone angles at

this step. The most uncharacteristic backbone angle changes occur at the ApC step of the first strand. At this step, β is *gauche+* rather than *trans*, and remains *gauche+* throughout the one nanosecond simulation. As shown in Figure 24, transitions in ε and ζ occur simultaneously with α transitions from *trans* to *gauche-* . ζ remains in a *gauche+* or *gauche-* conformation and when it is in a *gauche-* conformation, ε moves away from *trans*. The characteristic B_I (ε,ζ:t,g-) backbone conformation is never observed at this step. The likely reason for the anomalous behavior at this step is since the backbone was in a B_{II} or (ε:ζ; g-, t) backbone conformation at the start of the flip in pucker to C3'-*endo*. Since these B_{II} backbone conformations are not observed in the A-RNA simulation, this suggests that they are unfavorable in A-form structures, which may explain the unexpected behavior. It is somewhat surprising that during one nanosecond of dynamics, the structure does not transition to the more characteristic B_I backbone conformation at this step. Moreover, it appears that "crankshaft" backbone transitions are rare events in RNA simulations.

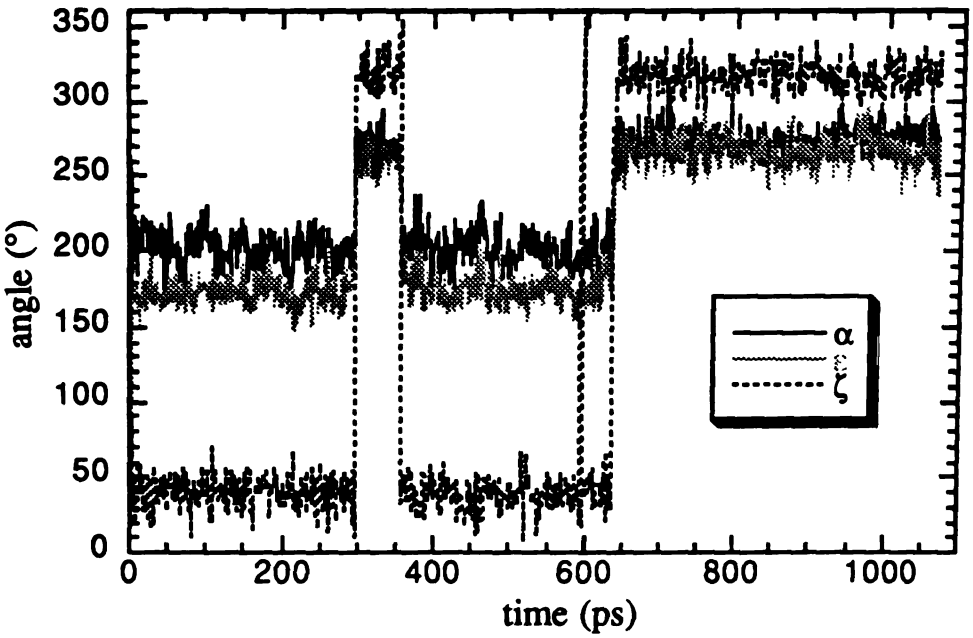


Figure 24: Plot of the alpha (α, O3'-P-O5'-C5', in black), epsilon (ε, C4'-C3'O3'-P, in gray) and zeta (ζ, C3'-O3'-P-O5', dashed black) backbone torsions angles (degrees) as a function of time (ps) for the ApC step from the first strand in the B-RNA simulation where a concerted flip in the sugar pucker pseudorotation phase from C2'-endo to C3'-endo was forced.

Note that with the exception of the anomalies mentioned above, all of the other backbone angles are in the expected range and we also still observe the distinctive rise and low inclination at the CpG step. However, instead of a low twist at the CpG step, the twist is above average at 36.4°. Interstrand stacking of the guanines is still allowed by compensating lower twists and slide values at the adjacent ApC (16.5°, -1.4 Å) and GpT (20.1°, -1.2 Å) steps. A slightly higher than average x-displacement (which is -5.0 Å) is also observed at the CG (-4.7 Å) and GC (-4.8 Å) base pairs. This backbone and helicoidal arrangement apparently leads to slightly better overlap of the guanines at the central step. This can be seen by examining the stereoview plot shown in Figure 25. This shows a view down the helical axis (shown in the upper center of the picture in black) with the C_5-G_{16} base pair stacking on top of the G_6-C_{15} and T_7-A_{15} base pairs. In black is

shown the nanosecond time averaged structure from the simulation with the concerted flip in the puckers. The displacement of the top guanine from the helical axis is less and it appears to better stack on the guanine below it than is seen in the average structure from the A-RNA simulation (gray). Analysis of the interaction energy of the two guanines *in vacuo*, using the coordinates of the average structures after they have been briefly minimized to an RMS energy gradient of 1.0 kcal/mol with a distance dependent dielectric function and dielectric constant of 4, applying the program ANAL from AMBER 4.1 (Pearlman *et al.*, 1995), suggests that the concerted pucker flip guanine stacking energy is indeed more favorable than that of the A-RNA. However, the better stacking of the guanine leaves the paired cytosines even less favorably stacked and more solvent exposed as can be seen in Figure 25.

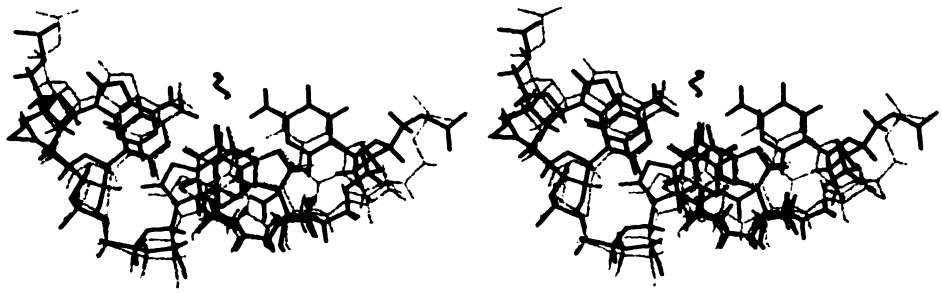


Figure 25: Stereo view plots, generated with MidasPlus (Ferrin et al., 1988), of three stacked base pairs from average structures calculated from the B-RNA simulation where a concerted flip in puckers was forced (70-1070 ps, in black) and the A-RNA simulations (1030-2030 ps,) in gray). The structures were RMS fit to these three base pairs and the view is looking down the helical axis (shown as the line visible in the top center of the figure in black) with the C_5-G_{16} base pair on top of the G_6-C_{15} and the T_7-A_{14} is on the bottom.

Looking back at Figure 20 and the data for the B-RNA simulation (in gray), the low twist at the CpG and adjacent steps is not seen. However, with the change in puckers to C3'-*endo* B-RNA converts to A-RNA and the low twist (albeit at the steps adjacent to the CpG and not the CpG step itself) and interstrand stacking of the guanines appears. This is interesting since the backbone compensates immediately to the pucker change at the steps adjacent to the CpG step. This observation also provides further evidence that the intrastrand stacking is a real contextual sequence dependent structural effect. The helicoidal parameters do take some time to convert to A-RNA values. In Figure 26, snapshots from the simulation where the concerted flip in pucker to C3'-*endo* are displayed. From these, it is clear that application of the restraints causes massive structural perturbation. When the pucker is converted to C3'-*endo*, this immediately decreases the intrastand phosphate distance leading to significant base pair buckling. The conversion to C3'-*endo* was not done smoothly, but abruptly as can be seen by looking at the snapshots in Figure 26. By 40 ps into the simulation, the terminal base pair is almost broken due to significant propeller twisting. However, by 70 ps into the simulation (or 20 ps after the termination of the restraints), the structure begins to settle down and display more reasonable helicoidal values. Figure 26 shows that the structure can react to the fairly drastic and quick transition from C2'-*endo* to C3'-*endo* puckers without completely breaking up. Short term effects on the helicoidal parameters are clearly evident. More importantly, backbone angles do get caught in specific conformations, such as the *gauche+* ApC β angle from the first strand which persists for longer than one nanosecond after the restraints are removed.

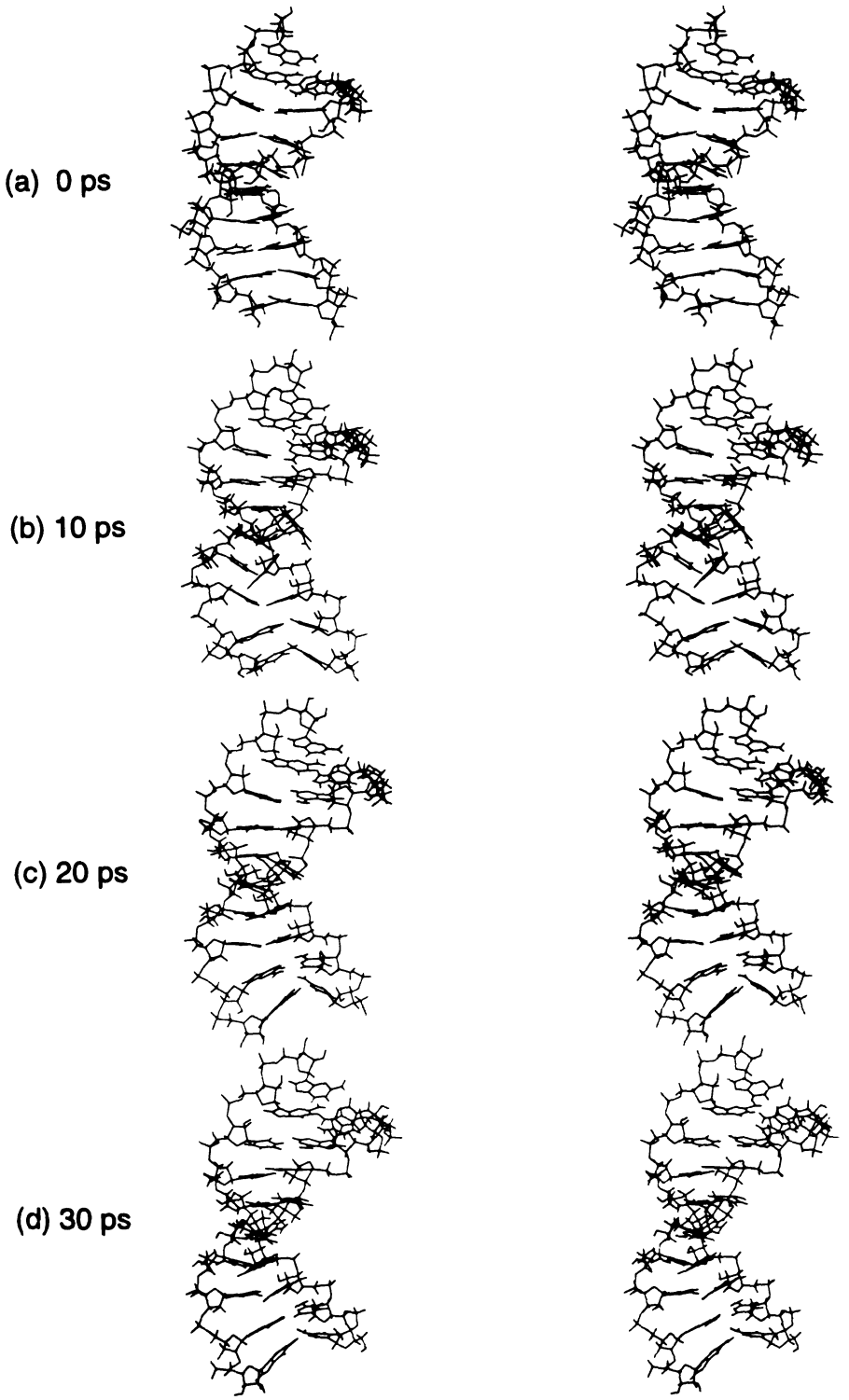


Figure 26: Snapshots from the simulation of B-RNA where a concerted flip in the puckers was forced, in stereo, are displayed in plots (a) through (h), representing the structure at each 10 ps interval. At 0 ps (a), no restraints have yet been applied and by 50 ps (f), all restraints have been removed, as is discussed in the methods section. The snapshots were all atom RMS fit to a common reference frame prior to display. *[continued on next page]*

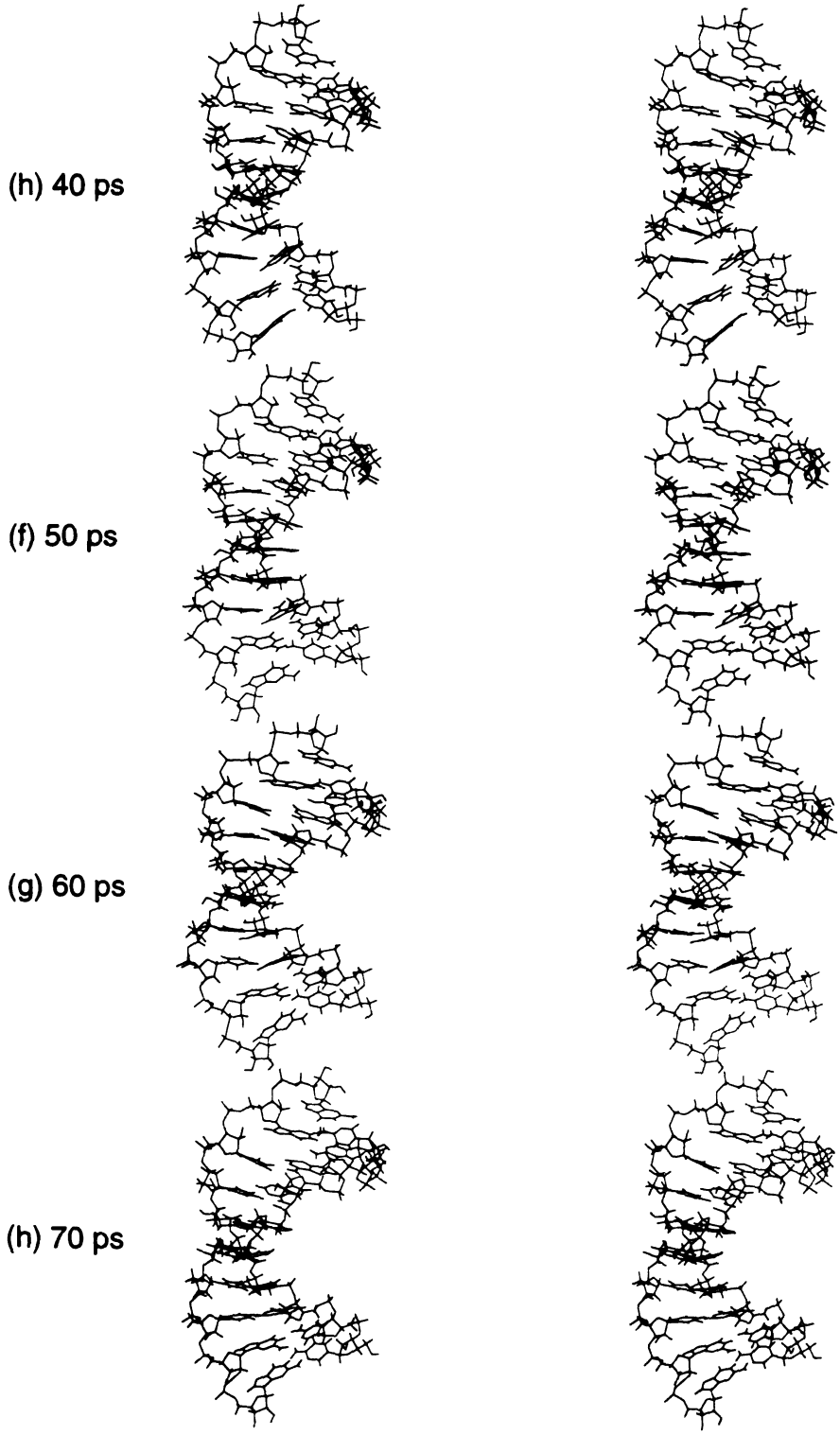


Figure 26, continued.

The A-RNA structure is more rigid than B-RNA or B-DNA:

Despite the observation of a crankshaft (α,γ: g-, g+ to t, t) transition at the CpG step in one of the strands of the A-RNA average structure (see Figure 22), the self convergence of the A-RNA average structure (ssRMSd = 0.34 Å) is much better than is observed in the B-RNA structure (1.0 Å). This high level of self convergence is not unexpected, since as mentioned in the introduction to this chapter, the A family of nucleic acid structures are expected to be more rigid than the B family. The rigidity of the A-RNA structure is readily apparent in the simulation. Little sugar repuckering and no correlated (ε,ζ: t, g- to g-, t) B_I to B_{II} backbone transitions, both of which are seen frequently in DNA simulations, are observed. The sugar repuckering that is seen in the A-RNA simulation is limited; the only events that are observed occur at the terminal guanine residues where G_{10} repuckers twice, once for ~10 ps and a second time for ~500 ps, and G_{20} repuckers once for ~100 ps. The overall fluctuations in the backbone torsion angles (see Table 8) are also reduced in the range of 10-50% compared to simulations of the RNA (or DNA) started in the B family, with the angles related to the pucker displaying the largest reduction in relative fluctuations (δ,γ,χ). Although the fluctuations in the backbone angles are reduced, we still observe higher than expected fluctuations in roll, tilt and twist as seen in previous simulations (Cheatham & Kollman, 1996b).

Interestingly, the B-RNA structure is not only more flexible than A-RNA, but also more flexible than the B-DNA simulations, as judged by looking at the standard deviations in backbone angles and helicoidal values presented in Table 8. All the values show enhanced fluctuation, except for the sugar pucker (and therefore the sugar pucker amplitude and δ backbone angle) and the ε and ζ backbone angles. Although the sugar repuckering is more frequent in the B-RNA simulation, where greater than 10 events longer than ~100 ps are observed, than in the A-RNA simulation (where as previously discussed only the terminal guanines repuckered) the repuckering is considerably less frequent than is

observed in the B-DNA simulations (*i.e.* roughly 10% of what is seen in the B-DNA simulations). This is presumably due to the larger barrier to repuckering in RNA (Olson & Sussman, 1982). The fluctuations in ε and ζ backbone angles are also less since the B_I to B_{II} backbone transitions are less frequent; however these B_I to B_{II} backbone transitions do occur in the B-RNA structure, which suggests that the lack of these transitions in A-RNA is not due to the presence of the O2' hydroxyl group, but due to the A-form geometry.

DNA:RNA hybrids-- flexible DNA and rigid RNA:

Hybrid duplexes which have one of the strands RNA and the other DNA, as previously discussed, are known to adopt a "mixed" form between a canonical A and B geometry. Simulations of DNA:RNA hybrid duplex structures (d[CCAACGTTGG]-r[CCAACGUUGG]) started in both canonical A and canonical B structures for ~2 ns each were performed to determine if molecular dynamics simulations with the Cornell *et al.* (1995) force field could accurately represent the structure of hybrid duplexes. In these simulations, as seen in the RNA:RNA simulations, the RNA strand remains in either a canonical A or canonical B geometry depending on the initial RNA conformation. The DNA strand on the other hand, as in the DNA:DNA simulations, undergoes an A-DNA to B-DNA transition regardless of its, or the RNA strands, starting structure.

As shown in Table 7, the DNA strand is closer to a canonical B geometry when the hybrid is started from a canonical B geometry (RMSd ~1.71 Å to canonical B-DNA and ~0.80 Å to the average B-DNA) than when it is started from a canonical A geometry (RMSd ~2.65 Å to canonical B-DNA and ~1.46 Å to average B-DNA). This is not surprising since, as seen in the RNA:RNA duplex B-RNA simulation, the RNA strand does not convert to a A geometry during 2+ ns of simulation. When the hybrid is started in a canonical A geometry, the DNA strand moves away from an A-form geometry but does not go quite all the way to a B-form geometry which demonstrates that it is clearly

influenced by the RNA strand's A-like conformation. This is of particular interest since it is this structure that is most consistent with and relevant to what has been seen experimentally.

In Figure 21c and 21d, average structures over the last nanosecond from the two trajectories are displayed to allow comparison to the RNA:RNA and DNA:DNA simulations. The average structure from the DNA:RNA hybrid duplex simulation started in a A geometry (A-hybrid) shown in Figure 21c has a surprisingly straight and regular helical axis compared to the other structures displayed. There is no kink at the TpG steps as is seen in the B-RNA and B-DNA average structures since there is not any significant bending into the major groove at this step. There is no change in direction of the helical axis at the central CpG step as is seen in the A-RNA structure, since this average structure does not have the low twist, high rise and interstrand guanine stacking seen in the A-RNA structure at the central CpG step. As will be discussed in more detail later, the A-hybrid structure is less inclined and less bent overall, with a narrower minor groove than is seen in the corresponding A-RNA structure.

The average structure calculated from the final nanosecond of the DNA:RNA hybrid duplex simulation started in a B geometry (B-hybrid) shown in Figure 21d is very similar, as is expected based on the low RMSd values reported in Table 7, to the B-DNA (Figure 21e) and B-RNA (Figure 21b) structures. Noteworthy is the kink in the helical axis and inclination of the cytosine at the bottom of the structure (Figure 21d). Looking at the RNA strand (on the right) it can be seen that the O2' hydroxyl of the first cytosine residue (bottom) is pointing away from the helical axis rather than up and along the helical axis as is seen in the other RNA nucleotides. This is because this cytosine repuckered to C3'-*endo* and remained C3'-*endo* during the final nanosecond of the simulation from which the average structure was created. The 3'-terminal guanine ribonucleotide also repuckers frequently during the simulation and some of the interior steps also have some persistent C3'-*endo* puckers during the final nanosecond of the simulation (C_{12} for ~50ps and U_{18} for

~500 ps). These repuckering events, which are more frequent and more persistent than was observed in the simulation of the B-RNA duplex suggests that perhaps a longer simulation may allow the RNA strand to convert to A geometry. However, simulations run at 400 K were not sufficient to drive the transition; instead, as was observed in the simulation of the B-RNA at 400 K, the structure became distorted and the terminal base pairs frayed and stacked upon each other.

The A-hybrid average structure displays many properties that are consistent with experimental results. In particular, the DNA strand has sugar puckers that are primarily C2'-*endo* whereas the RNA strand has nearly all C3'-*endo* sugar puckers throughout. As shown in Table 8, the helicoidal parameters are also consistent with what is known from experiment about DNA:RNA hybrid structure. Specifically, the A-hybrid average structure is positively inclined (11.1°), has a small positive roll in the major groove (1.6°), small positive buckle (4.3°), negative propeller twist (-13.2°) and a negative x-displacement from the helical axis (-4.5 Å). In Figure 27, the individual backbone angles for all the nucleotides are represented for the A-hybrid structure (gray), A-RNA (black) and B-DNA (dashed black). From this figure the (α, γ) crankshaft transition, discussed previously, at the CpG step in the first strand of the A-RNA is readily apparent. A similar crankshaft transition, at the UpU step can be seen in the RNA strand of the A-hybrid structure (top right of Figure 27). In general the A-hybrid angles of the DNA strand (gray, left side) tend to match the B-DNA angles (black dashed) and the angles of the RNA strand (gray, right side) tend to match the A-RNA angles (black). We see the expected trend in the A-hybrid structure, as discussed in the introduction to this paper, that the α angle is lower, and the ε, ζ and γ angles slightly higher in the RNA strand (gray, right) than in the DNA strand (gray, left). The average ε and ζ angles at each step for the B-RNA (not shown), A-RNA and A-hybrid are all similar, in contrast to the B-DNA simulations where peaks and troughs in ε and ζ are evident. These peaks and troughs result from relatively more frequent B_I to B_{II} backbone transitions which push the average ε closer to *gauche*- and the average ζ closer

to *trans* in the B-DNA and B-hybrid (data not shown) simulations. Despite the similarity of the average ε and ζ angles in the B-RNA and A-hybrid simulations, B_I to B_{II} backbone transitions are still observed in both the DNA and RNA strands, albeit at a lower rate than is seen in the B-DNA simulations. These transitions are more frequent and correlated in the DNA strand than in the RNA strand. During more than 2 nanoseconds of simulations of A-RNA, B_I to B_{II} backbone transitions were never observed. This provides further evidence that B_I to B_{II} transitions are easier in B-form structures and that the B_I state is preferred in A-form structures.

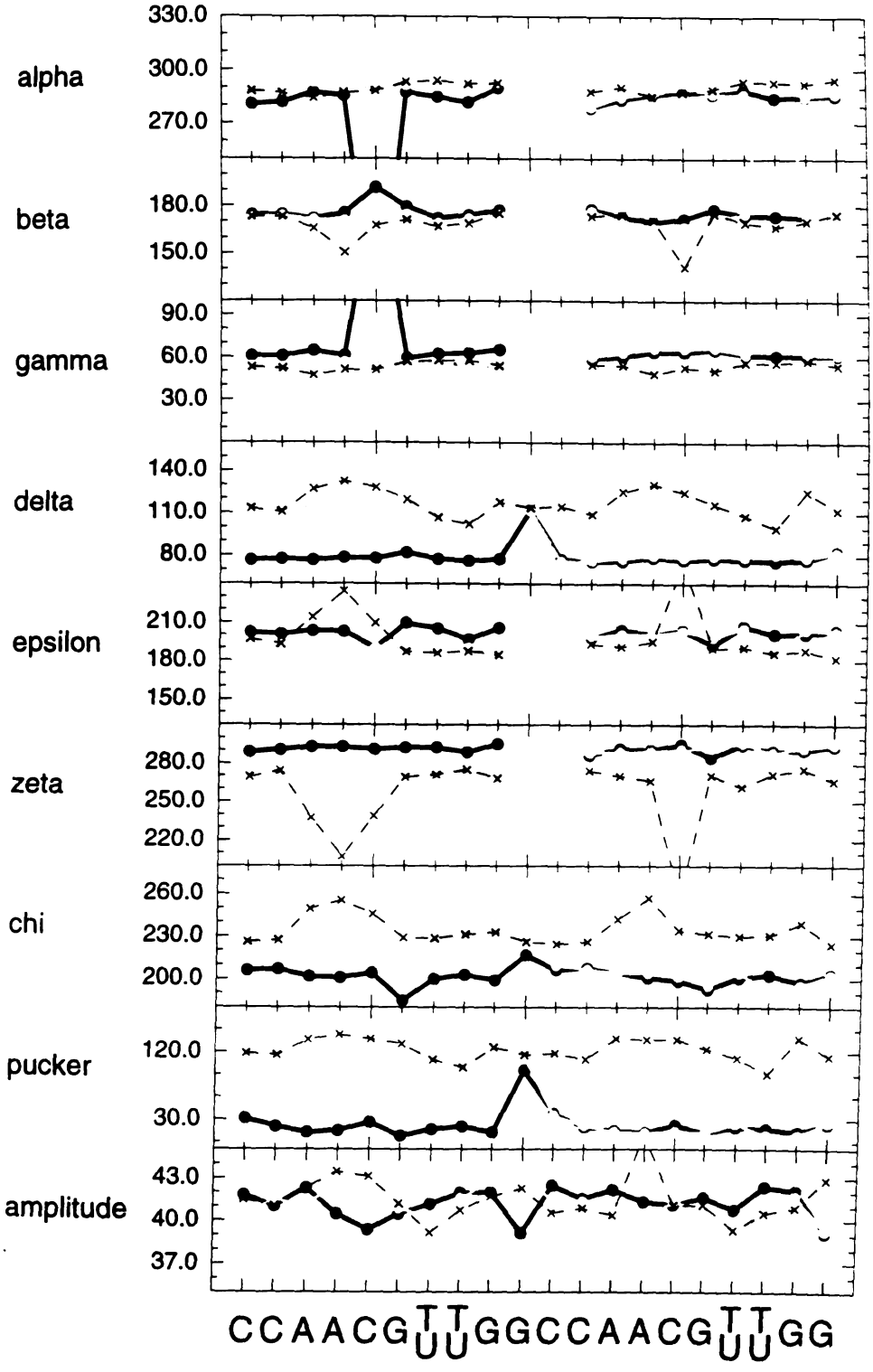


Figure 27: Average backbone angles along the sequence from 5' to 3' for the first strand followed by the second strand (from left to right) for average structures from the A–RNA (black, 1030–2030 ps), A–hybrid (gray, 1005–2005 ps), and B–DNA (dashed black, 400–1400 ps) trajectories. All the angles are listed in degrees.

The groove widths are also consistent with the experimental data on DNA:RNA hybrid duplex structures. Shown in Table 9 are average interstrand phosphate distances (and standard deviations in parenthesis) across the minor groove for each of the simulations.

	A-RNA 1030-2030 ps	B-RNA 1370-2370 ps	A-hybrid 1005-2005 ps	B-hybrid 1045-2045 ps	B-DNA 400-1400 ps
P_5-P_{20}	17.32 (0.69)	15.52 (1.58)	15.28 (0.90)	13.53 (1.46)	14.19 (1.13)
P_6-P_{19}	17.12 (0.68)	12.72 (1.31)	15.72 (0.78)	12.15 (1.18)	12.66 (1.53)
P_7-P_{18}	16.60 (0.54)	11.17 (0.97)	15.07 (0.81)	10.59 (0.87)	11.10 (1.07)
P_8-P_{17}	16.14 (0.53)	10.72 (1.00)	14.82 (0.80)	11.17 (1.31)	10.98 (1.21)
P_9-P_{16}	16.93 (0.58)	11.87 (1.67)	15.47 (0.88)	12.72 (1.61)	11.65 (1.48)
P_{10}-P_{15}	17.69 (0.71)	15.10 (1.90)	15.94 (0.98)	13.51 (1.58)	14.05 (1.14)

Table 9: Minor groove widths in the various models (as denoted by the row headings) represented by selected inter-strand phosphate distances (as specified in column 1) in angstroms. Distances were selected to match those chosen in analysis of the B-DNA crystal structure (Prive et al., 1991) and previously reported calculations (Cheatham & Kollman, 1996b). The distances are averages over 1 ns, in angstroms, and in parenthesis are the standard deviations.

The A-RNA (left) has the widest minor groove and little sequence specific narrowing. The B-form structures of DNA, RNA and the B-hybrid have narrow minor grooves and display sequence specific narrowing in the center of the helix similar to that observed in the crystal structure (Prive *et al.*, 1991). The A-hybrid structure has a minor groove width intermediate between the A-RNA and B-DNA structures. There is also no significant sequence specific narrowing at the center of this duplex. Overall the A-hybrid structure is closer to an A-form geometry than a B-form geometry, although the DNA strand has primarily C2'-*endo* sugar puckers.

C2'-*endo* sugar puckers tend to increase the intrastrand phosphate separation and therefore B-DNA structures have a larger rise between base pairs and are longer overall. This brings up an interesting question with respect to the structure of DNA:RNA hybrids of

longer sequence. As the helix becomes longer, how can the structure compensate for the larger discrepancy between the longer end to end length of the DNA strand and the shorter RNA stand; will the DNA strand "shrink" or perhaps present more C3'-*endo* puckers, in order to maintain good structure? Clearly hybrid structures can adopt a canonical A geometry (which would alleviate this problem) as has been seen in crystal structures (Egli *et al.*, 1992; Wang *et al.*, 1982). The current simulations do not address this question. However, these simulations do further support the observation that the B structure is more stable for DNA, as A to B transitions have been seen for the DNA strand in DNA:DNA and DNA:RNA hybrid duplexes. Moreover, it is clear that molecular dynamics simulations can reasonably represent the difference in structure between duplex A-RNA, DNA:RNA hybrids and B-DNA.

DNA:RNA hybrids and nucleic acid sugar repuckering:

The time course of the individual sugar puckers at 1 ps intervals and the histograms of the individual sugar puckers are shown in Figure 28. The DNA strand is represented in Figures 28a (pucker versus time) and 28b (histogram) with a graph for each nucleotide from the 5' (top) to the 3' end (bottom). The corresponding RNA base pair is shown alongside in Figure 28c (pucker versus time) and Figure 28d (histogram) for the RNA strand from the 3' (top) to the 5' end (bottom). The one letter code for each nucleotide is specified in the upper right of the histograms (Figures 28b and 28d). From the data in Figure 28a, it is clear that the DNA strand is repuckering throughout the simulation. Also from this figure, the time course of the transition from C3'-*endo* to C2'-*endo* puckers, indicative of the A-DNA to B-DNA transition in the left strand, is also evident. Most of the DNA sugars have repuckered within the first 100 ps and all have repuckered by ~300 ps. This is similar to the time course seen in the DNA:DNA duplex simulations. The histograms of the DNA strand (Figure 28b) show a mix of puckers with C2'-*endo* puckers

favored. The RNA strand (shown in Figures 28c and 28d), on the other hand, is clearly not repuckering, except for one repuckering event at the terminal guanine and C3'-*endo* puckers dominate (Figure 28d). Although the data is not shown, the number of sugar repuckering events seen in the RNA strand of the A-hybrid is consistent with the A-RNA duplex simulation. B-RNA, in both the B-hybrid and B-RNA simulations, does repucker more frequently than A-RNA, however at a significantly lower rate than is seen in the DNA simulations. In general, all of these simulations suggest that the RNA does not repucker too frequently on a nanosecond time scale. In the B-RNA simulations the repuckering from C2'-*endo* to C3'-*endo* is longer lived than is seen in the DNA simulations, which implies, as expected (Olson & Sussman, 1982), that the barrier to repuckering in RNA is higher than in DNA. The repuckering of the sugars in the DNA strand of the A-hybrid, despite the presence of the RNA strand, occurs at a similar rate and gives a similar distribution to that seen in the corresponding B-DNA duplex simulations.

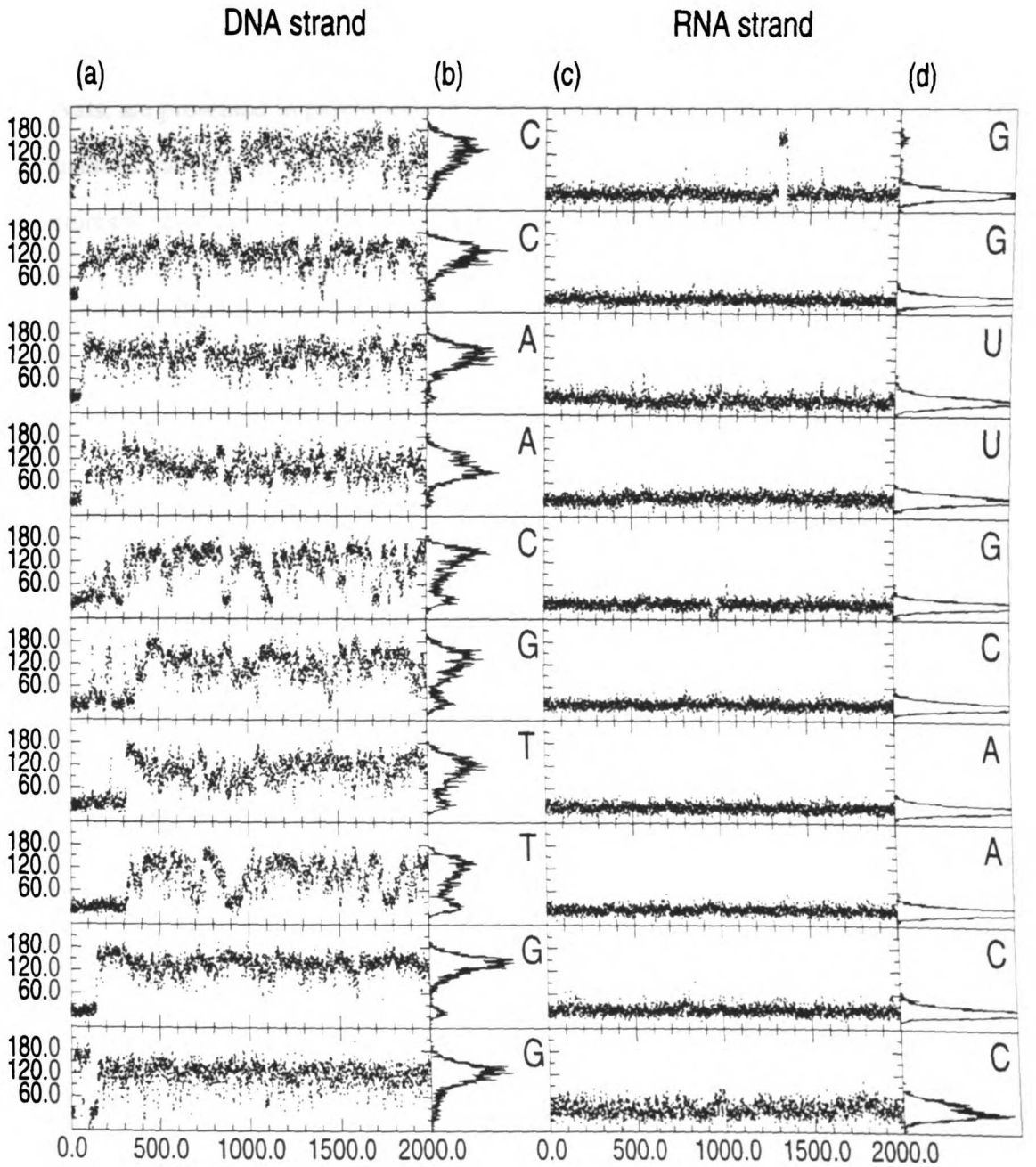


Figure 28: Sugar pucker pseudorotation phase (degrees) versus time (ps) and histograms for each individual nucleotide from the A–hybrid simulation. (a) The pucker versus time from top to bottom for the DNA strand from 5' to 3'. (b) The histogram of the pucker from top to bottom for the DNA strand from 5' to 3' along with 1–letter code labels for the DNA residues. Figures (c) and (d) are the same as (a) and (b) except for the DNA strand from 3' to 5'.

Inherent fluctuations during molecular dynamics:

The standard deviations, or fluctuations, in the backbone angles and helicoidal parameters, calculated based on analysis of 1 nanosecond portions of the trajectories at 1 picosecond intervals, are presented in parenthesis in Table 8. All of the B-form structures have larger fluctuations than are seen in the A-form structures. The greater flexibility of the B-form structures is consistent with experiment since, as discussed in the introduction to this paper, A-form structures are generally found to be more rigid than B-form structures. What does this flexibility, on a picosecond to nanosecond time scale, mean and where does it come from? The fluctuations reported in Table 8 relate to differences both in the simple anharmonic atomic motions and short time scale (ps) collective motions such as base pair propeller twisting, sliding, and helical twisting and bending. Sugar repuckering may also influence the flexibility; RNA, with a larger barrier to sugar repuckering might be expected to be more rigid than DNA (Olson & Sussman, 1982). Moreover, backbone transitions (B_I to B_{II}, (α,γ) crankshaft, *etc.*) are expected to also influence the dynamics. However, the data in Table 8 suggest the enhanced flexibility of B-form over A-form structures is not entirely due to more frequent repuckering or the presence of backbone transitions. Considerably less repuckering is seen in the RNA simulations, yet both the B-RNA and the B-hybrid structures have fluctuations in the backbone angles and helicoidal parameters that are higher than are seen in A-RNA. DNA simulations where the puckers are held fixed at C3'-*endo* and C2'-*endo* also confirm that the observed flexibility is not dependent on repuckering; when the puckers are forced to remain C3'-*endo* (and the structure moves closer to a A-form geometry), the fluctuations are significantly damped. When the pucker is held fixed at C2'-*endo*, the fluctuations are slightly enhanced with respect to simulations where the pucker is not held fixed (Cheatham & Kollman, 1996a). The flexibility also is not due solely to B_I to B_{II} transitions in the backbone since these do not happen as frequently in B-RNA as in the B-DNA, yet B-RNA shows a similar magnitude in the

fluctuations. However, the absence of B_I to B_{II} transitions in the A-RNA simulation may partially explain the rigidity. Although the B-hybrid structure does appear to be the most flexible, as can be seen by examining the fluctuations in Table 8, the high relative fluctuations are mostly the result of large fluctuations in the individual angles, base pairs and base pair steps at the 5'-end of the RNA strand. As mentioned previously, the 5'-terminal cytosine in the RNA strand of the B-hybrid repuckers to C3'-*endo*. This observation, coupled with the enhanced flexibility, indirectly suggests that the B-hybrid is in a higher energy state.

The high flexibility of all the B-form structures suggest that the free energy landscape around the minimum is flatter, and the B-form is a broader energy minimum allowing more picosecond time scale motions, such as base pair twisting, sliding and propeller twisting, among other collective motions (Tidor *et al.*, 1983). The A-form geometry of RNA on the other hand may represent a deeper and tighter minimum. This helps explain why is hard to find the A-RNA state during the B-RNA dynamics; it is difficult to transition from the broad flat energy minimum of the B-form geometry representing a high configurational entropy state to the more ordered A-state, not to mention the difficulty in overcoming the concerted barriers to sugar repuckering. A comparison of the atomic positional fluctuations over nanosecond portions of all the trajectories (data not shown) clearly demonstrate that the individual atomic motions in the B-form structures (B-DNA, B-RNA, B-hybrid) are uniformly higher than those observed in the A-form structures (A-RNA, A-hybrid). The average atomic positional fluctuations are roughly 20-30% higher in the B-form structures (B-RNA = 1.35 ± 0.40 Å, B-hybrid = 1.37 ± 0.41 Å, B-DNA 1.24 ± 0.36 Å) than the A-form structures (A-RNA = 1.09 ± 0.35 Å, A-hybrid = 1.03 ± 0.29 Å). The greater fluctuations in the B-form structures could result from "lower" low frequency vibrational modes. Low frequency vibrational modes (less than 100 cm^{-1}) contribute to most of the atomic motion (Tidor *et al.*, 1983) due to the inverse relationship between fluctuations and normal mode frequency. Simulations on

proteins suggest that the first 3-8 non zero normal mode frequencies account for ~70% of the motion (Levitt *et al.*, 1985). Therefore, to investigate if the reduced fluctuations in the A-RNA are due to less low frequency modes in the A-form structures than are present in the B-form structures, normal mode calculations were run on the A-RNA and B-RNA average structures *in vacuo* with a dielectric constant of 4 and a distance dependent dielectric function (to mimic solvent screening). To insure the starting structures are true minima, the first six normal mode frequencies should be zero. Therefore, the structures were minimized for 1000 steps of conjugate gradient minimization (with the program sander), followed by Newton-Raphson minimization (with the program nmode) until maximum in the energy gradient was less than 0.00071 kcal/mol and 0.00055 kcal/mol for the B-RNA and A-RNA average structures, respectively. Despite the stringent minimization, the sixth normal mode frequency was still ~0.05 cm^{-1}. As mentioned previously, *in vacuo* minimization will tend to significantly distort the structures, particularly in the absence of counterions and solvent to balance and screen the electrostatic interactions. The minimizations caused the A-RNA and B-RNA structures to move 1.9 Å and 2.4 Å from the starting structures, respectively. Therefore, caution should be taken in interpreting the results from these normal mode calculations since the structures may not be truly representative of the average structure in solution; despite this, the normal mode analyses on these minimum energy structures may provide insight into the relatively flexibility of A-RNA and B-RNA.

The normal mode calculations suggest that both structures have an equivalent number of normal modes less than 20 cm^{-1}, but that the B-RNA structure is slightly less entropically favored (B-RNA = 1696.4 cal/mol-K, A-RNA = 1706.1 cal/mol-K). Given that both structures have an equivalent number of modes less than 20 cm^{-1} it is expected that both structures should have roughly equivalent atomic positional fluctuations; this is somewhat contradicted by the higher entropy of the A-RNA. However, some of the first few non-zero normal mode frequencies are lower for B-RNA than A-RNA which may

partially explain the greater flexibility observed in the B-form structures. The first five non-zero normal modes for B-RNA are 1.916, 2.294, 3.951, 4.030 and 5.285 cm^{-1} compared to 2.106, 2.605, 3.359, 4.467 and 4.660 cm^{-1} for A-RNA. Considering the relative difference in the sum of the first few normal mode frequencies, which is generally less than 5%, the differences in the low frequency normal modes for B-RNA and A-RNA are not enough to explain the observed 20-30% difference in the atomic positional fluctuations. This approximation completely neglects the solvent and counterion environment which may also influence the flexibility as is suggested indirectly in the analysis of hydration and counterion association which follows. The differences in relative flexibility in the two states may represent a mechanism for recognition and distinction between RNA, DNA and hybrid duplexes beyond the obvious differences in structure.

Hydration: B-form structures and flexibility revisited.

The relative flexibility of B-form and A-form structures, as mentioned above, is clearly not due entirely to backbone transitions, sugar repuckering or a significant difference in the *in vacuo* low frequency normal modes. Perhaps hydration or the associated salts play a role in the relative rigidities? From the early fiber diffraction studies on oriented DNA fibers which show that relative humidity influences DNA structure (Franklin & Gosling, 1953), to gravimetric and spectroscopic studies which characterize bound water [for a review, see Texter (Texter, 1978)], to high resolution crystal structures which map out the precise positions of nucleic acid associated water (Egli *et al.*, 1996; Schneider & Berman, 1995), it is clear that water is an integral part of nucleic acid structure. Since water, and likely also salt, is an integral part of the structure, it is likely that both the water and salt influence the dynamics. Perhaps the A-form geometry is partially rigidified by more tightly bound or specifically associated water and counterions? This is a difficult question to answer directly. However, our simulations suggest that the more flexible structures have "less"

associated water and counterions. Additionally, the A-form structure seems to have more counterions specifically associated in the major groove, and more specific hydration in the grooves, than B-form structures which could lead to higher stability and less flexibility (discussed below).

Molecular dynamics simulations can shed useful insight into water and counterion behavior since one can track, as a function of time, the water and salt structure. However, tracking individual waters and/or bookkeeping individual hydrogen bonds is tedious due to the large number of waters; moreover, presentation of the data is difficult and hard to follow. Therefore, methods which look at the average water structure using radial pair distribution functions of the water around a particular group, such as a phosphate, or cylindrical pair distribution functions which characterize water around a helical axis (Weerasinghe *et al.*, 1995b), or more elaborate methods which associate waters with particular groups, such as proximity analysis (Mehrotra & Beveridge, 1980) are often applied. These methods can quantitate the number of waters around a particular group and easily show the relative population of the water at a given distance with respect to bulk water, but do not show the individual positions of the water. Instead, overlapped snapshots from the trajectory can be graphically displayed and the waters visualized (Subramanian *et al.*, 1988). However, with many snapshots, this picture becomes clouded and pulling out the most representative water interactions is difficult. A perhaps better way to see the influence is by visualizing the most probable, or average, water positions relative to the nucleic acid over the course of the simulation. This can be done by constructing a grid around the nucleic acid and counting the number of water or counterions that hit each particular grid element over the simulation. This data grid can then be contoured to give a picture of the hydration; this is similar to what has been done in the analysis of water in nucleic acid crystal structures (Schneider *et al.*, 1993; Umrania *et al.*, 1995), characterization of pharmacophores (Rosenfield *et al.*, 1984), and analysis of counterion density around DNA (Laughton *et al.*, 1995). In Figure 29, this type of analysis is applied

to visualize the water density around the average structures computed over 1 nanosecond portions of the trajectories. The purpose of this current analysis is not to precisely map out the locations of the waters in the simulations, but to provide a general picture of the overall hydration and to characterize the distinction between the hydration of B-DNA (Figure 29b), B-RNA (Figure 29c) and the B-hybrid (Figure 29d) structures contoured at an equivalent level (15.0 hits per 0.5 $Å^3$ grid element, or ~3.6 times the expected water density). It should be noted that the structures shown are the average structures from the trajectory; an average structure does not clearly show the relative range of motion of all the atoms in the duplex over the course of the simulation. In general, each individual 1 ps frame deviates from the average structure by on the order of ~1.3±0.3 Å, with B-DNA closest to the average (1.2 Å) and the B-hybrid furthest from the average (1.35 Å). Each individual frame never gets closer than 0.6 Å, nor further than 2.4 Å, from the average structure over the final nanosecond of each simulation.

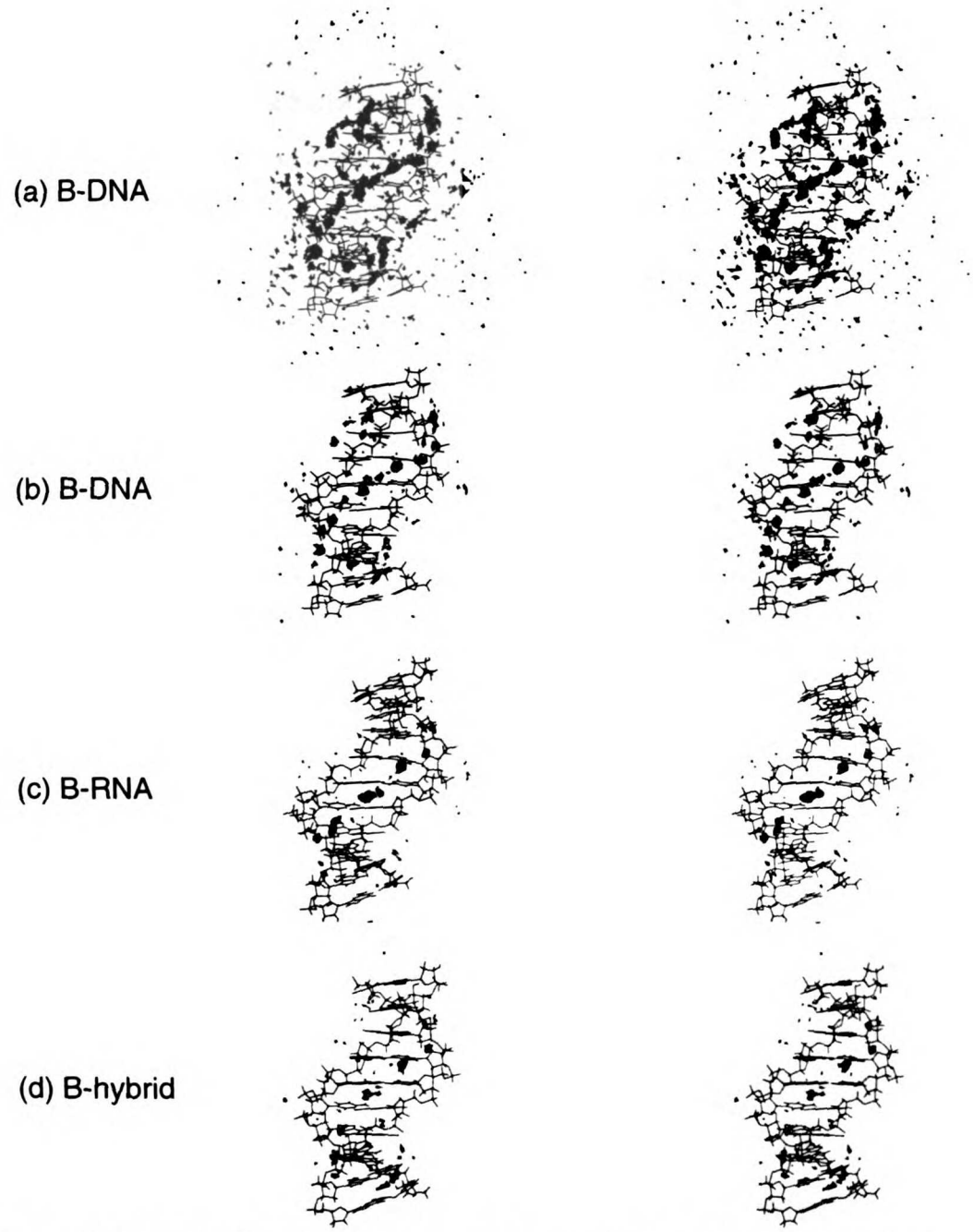


Figure 29: Hydration of the average structures. Stereo view picture of the average structures from various trajectories are presented along with contoured water oxygen atom density. The contours of the water oxygen density over 1 nanosecond from each trajectory, at 1 ps intervals, into 0.5 angstom grid elements over a 50 angstom cubed grid are displayed using the density delegate from MidasPlus. (a) B-DNA average structure over 400-1400 ps at a contour level of 12.0 hits per grid element. (b) B-DNA average structure at a contour level of 15.0 hits per grid element. (c) B-RNA average structure over 1370-2370 ps at a contour level of 15.0 hits per grid element. (d) B-hybrid average structure over 1045-2045 ps at a contour level of 15.0 hits per grid element.

Figure 29a displays the B-DNA average structure contoured at 12.0 hits per 0.5 $Å^3$ grid element or ~2.9 times bulk water density. The condensation of the water around the DNA is clearly evident. The highest density appears in the minor groove where waters which directly hydrogen bond to the nucleic acid, and secondary waters directly on top of those waters, can be visualized. The former sit deep in the minor groove and interact with the bases and the sugar O4' oxygens. At the lower contour levels, two to three waters per base pair step are visible, except in the narrow portion of the minor groove at the center of the duplex, where one water per base pair step deep in the groove is visible. The most obvious hydration pattern is this "spine of hydration" (Kopka *et al.*, 1983; Subramanian *et al.*, 1988) which extends out of the center of the duplex in the minor groove. Density is also clearly visibly associated with the backbone and major groove as well, especially at the lower contour levels (Figure 29a). The beginnings of a spine of hydration in the major groove is visible at both ends of the helix which twists the other direction around the helical axis compared to the minor groove and backbone spines of hydration. In Figure 29, very little preferential hydration of the phosphate groups is visible, however water density can be seen off the bisector of the phosphate oxygens. This is likely due to the diffuse nature of the "cone of hydration" (Pullman & Pullman, 1975; Subramanian & Beveridge, 1989) around the phosphates and the enhanced mobility of the phosphate atoms with respect to the other nucleic acid atoms. Based on this analysis, no specific hydration of the O2' hydroxyls by water oxygen atoms in the B-RNA is seen. This is most likely since the O2' hydroxyls spend a significant time hydrogen bonded to the O5' and/or O1P atoms of the following residue. The 3'-terminal guanines have no following residue to interact with, hence the C1'-C2'-O2'-HO2' torsion is essentially freely spinning.

Comparing the various B-form simulations (which all converged to the same structure) we generally see equivalent hydration patterns in the minor groove. However, at an equivalent water density contour, the water occupancy appears to be highest in the B-DNA and lowest in the B-hybrid structure. The trend here mimics the trends in relative

flexibility, with the more flexible B-hybrid appearing to have lower water occupancy. The trend does not result from the solvent diffusing more rapidly, on average, in the B-hybrid than the other simulations since the average diffusion of water in all of these simulations is comparable.

The lower density observed in the more flexible structures could be an artifact of the coordinate fitting and water oxygen atom gridding and visualization procedure. An example of artifactual behavior is readily apparent in Figure 29 where it appears that the ends of the duplex are less hydrated. In order to create the grid of water density, each snapshot is RMS fit to a common reference frame, which is this case was all the nucleic acid atoms, and then the grid is constructed. Since the ends of nucleic acids are more flexible, the density appears lower. This is similar to what is seen in the crystal structures. The higher mobility leads to relatively less and more irregular solvent density. In fact, it was not until the low temperature crystal structures (Drew & Dickerson, 1981) (which effectively reduce the thermal fluctuations) that explicit water density around the phosphates was visualized. This is the basic point. If water is in more regular positions, such as is seen in both the grooves of A-RNA (discussed below), it will be easily visualized and moreover may tend to rigidify the structure. Alternatively, if the structure is more flexible, less water density anchoring both sides of the grooves (for example) may be present. Of course, a causal relation is not apparent; while it is clear that a more rigid structure will lead to more well defined water positions, these simulations do not determine if water rigidifies the structure.

It should be noted that if the grid is built around a set of coordinates that are RMS fit to only the first base pair of the helix, specific water density is clearly visible at the end of the helix (data not shown). This implies that in order to specifically analyze the hydration each individual base pair, or base pair step, should be RMS fit and the water density independently calculated to remove these dynamic effects. When this is done, the

hydration results are in general agreement with the analysis of crystal structures presented to date (Schneider & Berman, 1995; Schneider *et al.*, 1993; Umrania *et al.*, 1995).

From the pictures presented in Figure 29, the minor groove in the B-form geometry is clearly preferentially hydrated and is probably directly involved in stabilizing the B-form geometry. This is in agreement with nucleic acid crystal structures of B-DNA which show preferential hydration in the minor groove (Kopka *et al.*, 1983) and in some cases, little hydration of the major groove (Edwards *et al.*, 1992). The observation of less hydration in the major groove is also in agreement with NMR studies on B-DNA which suggest that the water in the major groove is highly mobile and characterized by residence times less than 500 ps (Liepinsh *et al.*, 1992). Although the major groove is less hydrated, there is visible major groove hydration. This hydration is most visible in the B-DNA structure (Figure 29a,b) and resembles a "spine" that runs down the middle of the major groove between the bases. This hydration is also apparent in the B-hybrid structure where the major groove density appears darker at the bottom of the helix, where it interacts with 3' terminal guanines of the DNA strand, than on the top where it interacts with the 3' terminal guanines of the RNA strand. Not only is the density lower on the top part of the B-hybrid, density also appears lower in the B-RNA structure which suggests that the minor groove of RNA in a B-form geometry is less preferentially hydrated than in B-DNA.

Hydration of A-RNA:

The locations of bound water in A-form structures is distinctly different than B-form structures as has been seen in the analysis of crystal structures (Eisenstein & Shakked, 1995; Schneider *et al.*, 1993) and via fiber diffraction (Langan *et al.*, 1992). The deeper major groove of A-form structures, and rotation of the phosphate group into the major groove, leads to more well defined hydration in the major groove. The minor groove is also hydrated, however the more open minor groove of A-DNA is not as extensively

hydrated as B-DNA (Schneider *et al.*, 1993). A-RNA, on the other hand, has an extensively hydrated minor groove, likely due to the addition of the O2' hydroxyl groups which provide an anchor point for hydration traversing the minor groove, as is seen in a recent high resolution A-RNA crystal structure of r[CCCCGGGG]$_2$ (Egli *et al.*, 1996). The extensive hydration of A-RNA, and differences between A-RNA and B-RNA/B-DNA, are readily apparent in the simulations of A-RNA. In Figure 30a, a contour plot of the A-RNA hydration, contoured at 12.0 hits per 0.5 Å^3, is shown with a view into both grooves. The extensive hydration of the major and minor grooves is readily apparent, and the overall hydration patterns are very distinct from that seen in the B-form structures (Figure 29).

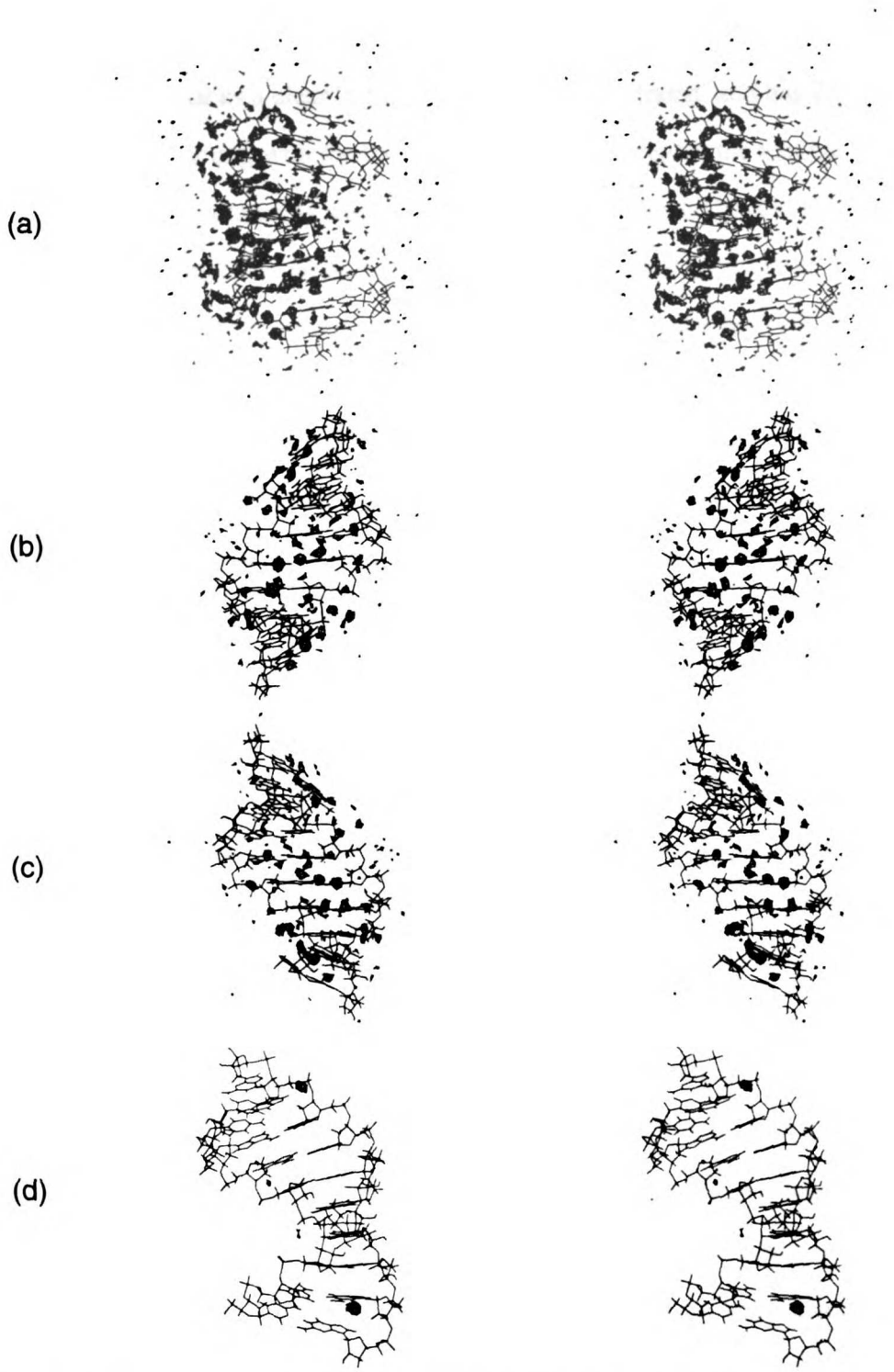


Figure 30: Average structure from the A-RNA trajectory (1030-2030 ps) in various view with contoured water oxygen density (a-c) and counterion density (d) at 1 ps intervals. In figure (a), water oxygen density at 12.0 hits per grid element is displayed with a view into both grooves. Water oxygen density at 15.0 hits per grid element is shown with a view into the minor groove (b) and major groove (c). (d) Sodium counterion density is displayed at a contour level of 12.0 hits per grid element

In contrast to the visible "spine of hydration" in the B-form structures, the minor groove in A-RNA lacks a clear "spine", appears more hydrated, and has 2-3 waters per base pair step interacting with the bases and backbone (except at the terminal residues which appear less hydrated as was seen and discussed in the analysis of the hydration of the B-form structures). In the high resolution A-RNA crystal structure of $r[CCCCGGGG]_2$, two common transversal hydration motifs are seen which link the O2' hydroxyl's across the minor groove (Egli *et al.*, 1996). Most of the major groove in this crystal structure is characterized by two waters between the RNA strands linking the O2' atoms from adjacent residues in base pair steps across the groove; this motif has the shortest distance between O2' atoms in different strands. In the average A-RNA structure from the molecular dynamics simulation, the average distance between the O2' atoms in adjacent base pair steps across the minor groove is slightly larger (8.74 Å) than is observed in the crystal structure (8.52 Å). The other transversal hydration motif was observed at two base pair steps in the crystal structure of $r[CCCCGGGG]_2$ where three waters linking the O2' atoms across base-paired nucleotides are seen; the distance across the base pair linking the O2' atoms is 11.30 Å in the crystal, compared to 10.93 Å, on average, in the A-RNA average structure. Figure 30b shows the minor groove hydration of the A-RNA average structure, contoured at a level of 15.0 hits per 0.5 $Å^3$ (or ~3.6 times the expected water density). From this figure, both transversal minor groove hydration motifs are simultaneously apparent, such as can be seen at the central CpG step. The behavior at the AU base pairs is slightly different, as might be expected since the AU base pair only has two hydrogen bond acceptors in the minor groove compared to the three present in GC base pairs. At the UpU steps, two waters deep in the groove are seen interacting with the adenine O2 and uracil N3 atoms. This leads to the observation of one and two waters, respectively, in the two O2' atom transversal hydration motifs discussed above. A third water interacts with the uracil O2' atom and the water bound to the uracil N3 atom. In addition to the transversal water linking the O2' atoms within the minor groove, hydration

of the O2' atoms out of minor groove and above the backbone can be seen (as displayed in Figure 30b). This water hydrating the O2' atom likely interacts with water directly solvating the backbone. Overall, the O2' atoms appear to be solvated more in A-RNA than was seen in the B-RNA simulation.

The backbone is also fairly extensively hydrated. The relatively short distance between intrastrand phosphates in A-form structures commonly leads to single water bridges between the adjacent O1P atoms (Saenger *et al.*, 1986; Westhof, 1988). In the A'-RNA structure of r[CCCCGGGG]$_2$ (Egli *et al.*, 1996), this motif is seen despite the slightly longer intrastrand phosphate distance in A'-RNA. In this crystal structure, the bridged water tends to be closer to the O5' side of the bridge. In the A-RNA average structure, although water oxygen density bridging the O1P atoms is generally observed, the tendency of the water to be closer to the O5' atoms is not reproduced. In some cases this water density is closer to the O5' atoms, however more often the water oxygen density is closer to the O3' side of the O1P-water-O1P bridge where the water is closer to other donors, such as the O4' and purine N7 atoms (where present). It is not clear if the simulation is incorrect here. The differences in hydration could relate to differences in crystal versus solution phase structures, differences between A'-RNA and A-RNA structures, sequence specific hydration patterns, or may relate to subtle deficiencies in the simple partial charge model's representation of hydrogen bond directionality. In addition to O1P-water-O1P bridges, water can be seen bridging the O2P atoms as well.

Although the major groove is not as well hydrated as the minor groove, judged by comparing the water oxygen density presented in Figure 30, there is still specific hydration and apparently more hydration, particularly in the center of the duplex, than is seen in the corresponding B-form structures (displayed in Figure 29). In figure 29c, at the CG base pair (or the fifth base pair down from the top) water can be seen traversing the major groove from the O1P of one strand, across the bases to the O1P of the other strand. Hydration is also present at the GpG step at both ends of the duplex where conserable

water oxygen density is present. As is discussed in the next section, this water is hydrating counterions in the GpG "pocket" interacting between the N7 atoms. At the water oxygen densities displayed, we do not observe the regular major groove hydration pattern seen in the crystal structure (Egli *et al.*, 1996). This could relate to sequence specific hydration patterns which might lead to differences in the hydration between the A-RNA studied here and the A'-RNA crystal structure or deficiencies in the model. This is currently being investigated in our lab in simulations on the r[CCCCGGGG]$_2$ duplex structure.

Overall, the A-RNA average structure seems more specifically hydrated than the B-form structures studied; this is probably since A-RNA is more rigid. The observation of different hydration patterns, not only between A-form and B-form structures but between A-RNA and A-DNA (Schneider *et al.*, 1993), and sequence dependently within A-RNA, suggests that specific hydration patterns may be important for recognition and distinction among various nucleic acids.

Counter-ions in the groove:

The phosphates pointing into the major groove in A-RNA lead to a more negative electrostatic potential in the major groove than is seen in B-form structures. Therefore, it is expected that counterions will preferentially move into the major groove rather than the more hydrophobic minor groove of A-RNA. Chemical acylation experiments contradict this hypothesis and suggest that the narrow and deep major groove of A-RNA may be inaccessible for specific recognition (Weeks & Crothers, 1993). Despite this, NMR and crystallography experiments suggests that ions (Ba^{2+} and $Co(NH_3)_6]^{3+}$) can specifically interact in the major groove, most notably interacting with guanine N7 atoms (Gao *et al.*, 1995). These experiments further suggest that smaller metal ions, such as Co^{2+} and Ni^{2+} with their intact hydration shells, cannot penetrate into the deep minor groove without leading to unfavorable contacts (Gao *et al.*, 1995; Gao *et al.*, 1993). Sodium ions, on the

other hand, have more deformable hydration shells since the water interacts with the ion electrostatically in contrast to the more covalent ion-water interactions of water with transitional metal ions. This suggests that Na^+ ions may penetrate the deep and narrow minor groove of A-RNA; this is seen in our simulations where sodium counterions move to distinct locations in the major groove of A-RNA. Symmetrically, at both ends of the duplex, high density is seen for counterions at the GpG steps, closely associated with the N7 atoms of the guanines, as can be seen in Figure 30d which shows the sodium atom density at 12.0 hits per 0.5 $Å^3$. Although the view is not exactly the same, in Figure 30c as mentioned previously, water is visible at the GpG steps which forms part of the coordination shell around the sodium atoms. Also visible in Figure 30d is more faint density more towards the center of the duplex associated with the adenines at the ApA step and also at the ApC step. This counterion density, at a contouring level equivalent to Figures 29a and 30a, shows that counterions specifically, and preferentially, associate with the major groove of A-RNA. It is not until the contour level is lowered to 4.0 hits per 0.5 $Å^3$ that counterion density is seen in the minor groove of A-RNA and this density is not deep in the groove, but relatively close to the backbones interacting with the minor groove waters and the backbone phosphates and O2' hydroxyls. The density seen in these simulations for the counterions is not large since only 18 counterions were added to neutralize the system and these counterions diffuse throughout the simulation box during the trajectory. However, it may be claimed that, even though greater than 2 nanoseconds of simulation were run for each of the models with RNA and only the final nanosecond was used in the analysis, the counterion positions could be biased by the initial starting coordinates. In other words, perhaps counterions appear in the major groove of A-RNA since they were started near the major groove. However, we think that this is not the case, in part because not only do the counterions diffuse at near the expected rate (1.2-2.1 x 10^{-5} cm^2) but the counterions in the major groove exchange during the simulation. In Figure 31 distances between key atoms which represent the major groove and the individual

counterions versus time for all the base pairs are shown. The distance to each individual counterion is represented by a different color. Comparing the A-RNA (Figure 31a) to the B-RNA (Figure 31b), counterions are clearly more often closer to atoms in the major groove in the A-RNA simulation than in the B-RNA simulation. The guanine N7 atom distances from the A-RNA simulation (top right and bottom left of Figure 31a), show that not just one ion interacts with the GpG step at both ends of the duplex, but at least two distinct ions interact at different times. A persistent interaction is also seen with the N7 atoms of first strand adenines. By following a specific color, such as the maroon, a particular ion can be followed moving from interactions with the center of the duplex up through ~1 ns where it moves to interact with the GpG step N7 atoms (at the top of the duplex) over the latter part of the trajectory. The B-RNA simulation clearly shows less specific, close and persistent interaction of sodium ions and groove atoms. Moreover, tracking an individual ion, such as the "blue" ion seen in the top of Figure 31b, the interaction is less specific to the GpG step; it moves from interactions with the guanine O6 atoms to interactions with the uracil O4 atoms.

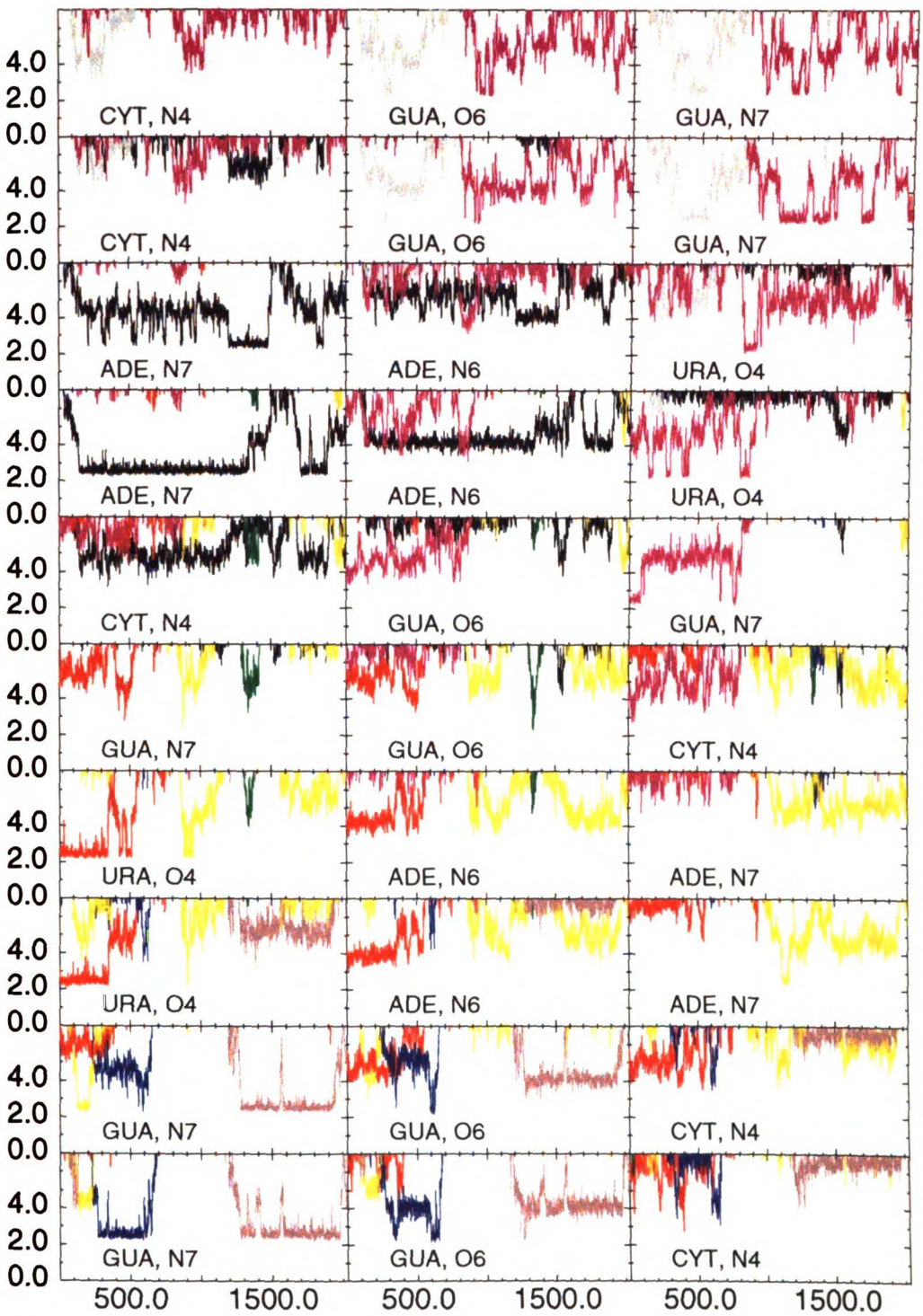


Figure 31: Distances from individual counterions to atoms along the duplex representing the major groove. Each individual graph represents the distances of each counterion, represented by a different color, to an atom (as labeled in each subgraph) as a function of time. Each base pair is represented from the 1st (top) to the 10th (bottom). The distances along the y–axis are in angstroms and the time in picoseconds. (a) Distances for the A–RNA simulation. (b) Distances for the B–RNA simulation (next page).

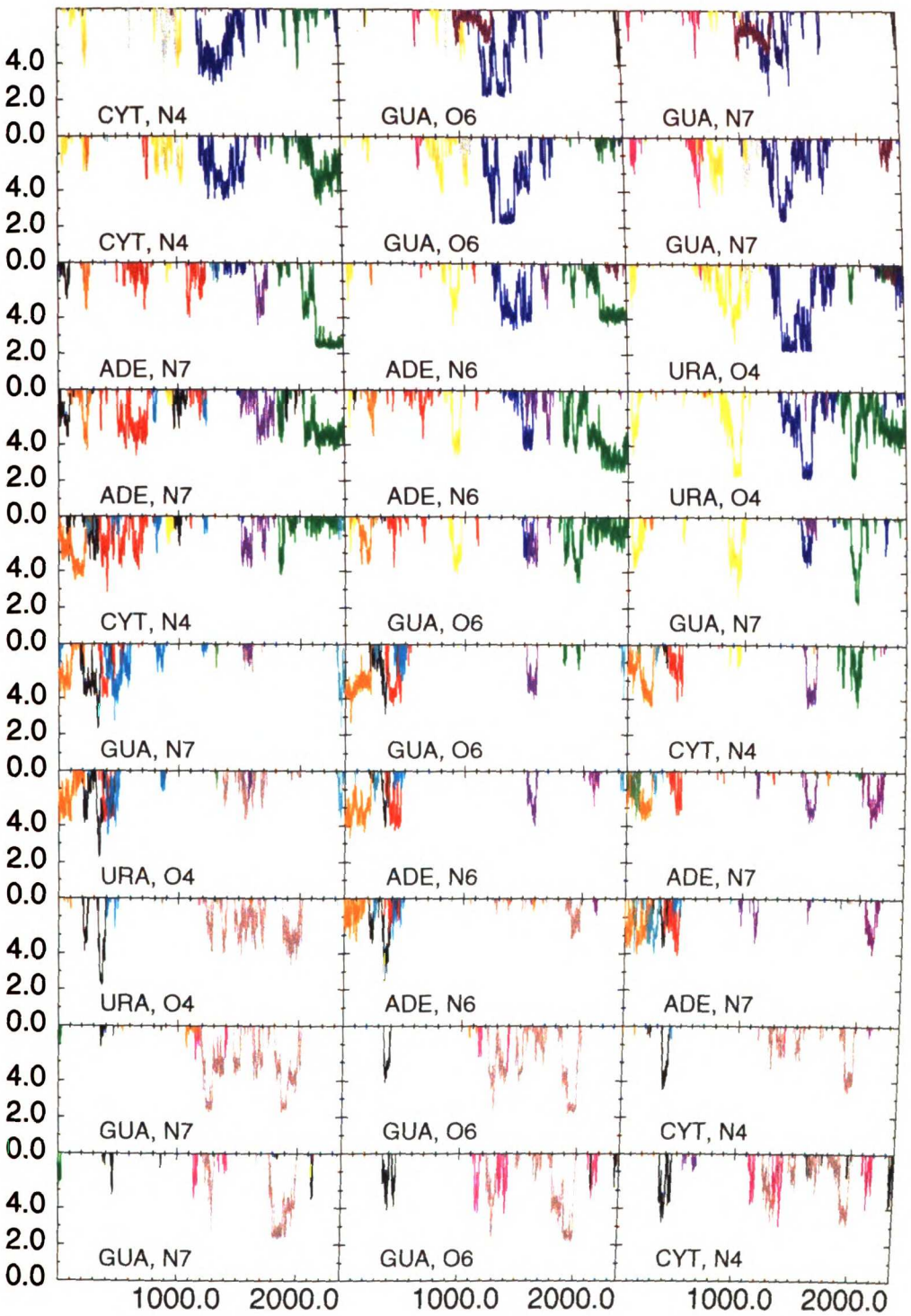


Figure 31: continued.

The density of counterions associated with the A-RNA structure at the 12.0 contour level is greater than is seen in any of the other trajectories (B-RNA, B-DNA, B-hybrid and A-hybrid). In fact, density in the B-RNA simulation does not appear at all until the contour level is dropped down to 8.0 hits per 0.5 $Å^3$, where a little density appears in the minor groove. As the contour level is lowered in the B-RNA, counterions appear associated in the minor groove, major groove, and along the backbone. The major groove counterion positions are not in the distinct "pockets" seen in the A-RNA simulation where the counterions tend to interact with the N7 atoms at purine-purine steps, but is more diffuse and resembles the inverse spine of hydration seen in the major groove of the B-DNA simulation. The fact the ions are seen in the same place as waters, and moreover provide diffuse density along a particular location, such as the major groove or minor groove, suggests that the "waters" often seen in crystallography experiments may be counterions since both will lead to a comparable scattering/density. A recent paper by Young and Beveridge makes the same point (Young *et al.*, 1996); in this paper simulations on the dodecamer $d[CGCGAATTCGCG]_2$ are analyzed, with various initial counterion positions, to investigate the interaction of counterions in the grooves. In their discussion of major groove counterion association, it is suggested that the most favorable "pocket" in the major groove of B-DNA will be intrastrand GpG "pockets" with ions interacting with the guanine O6, followed by ApA "pockets", where the ion would interact with the thymine O4 atoms. In figure 31b, representing the B-RNA simulation, this general trend is seen. The ions are closest to the guanine O6 or uracil O4 atoms when in GpG or ApA "pockets" respectively. However, since the occupancy of the ions in the major groove of B-RNA is lower than is seen in the A-RNA, not enough occupancy of the major groove by ions was observed by us to unequivocally support the conclusions of Young *et al.* (1996) on ions in grooves. However, our B-RNA simulations do support their observation that there is an overall lesser propensity for fractional occupation by mobile counterions in the major groove than the minor groove in B-form structures. The current simulations add to this observation of

Young *et al.* (1996) by pointing out that the propensity for ions in the grooves of A-form structures is distinct from B-form structures. Although intrastrand GpG and ApA pockets are still favored in the A-RNA major groove (see Figure 31b and Figure 30c), the ions tend to interact more strongly with the guanine or adenine N7 atoms rather than the guanine O6 or thymine/uracil O4 atoms. The general trend seen in these simulations is that the B-form structures tend to favor counterions in the minor groove, whereas the A-form structures favor having counterions in the major groove. Given that the counterions in the major groove of A-RNA tend towards specific locations, rather than the diffuse density seen in the major groove of B-form structures, suggests that the counterions may in part stabilize and rigidify the A-RNA structure. A good test of these hypotheses regarding counterion association would be to run simulations of both A- and B-form structures where a significant number of counterions were initially placed ~10 Å from the solute to see if the same localization of the counterions is observed, similar to the experiments performed Young *et al.* (1996).

Conclusion:

These calculations show that molecular dynamics simulations-- with a reasonable force field (Cornell *et al.*, 1995), proper treatment of the long ranged electrostatics (Cheatham *et al.*, 1995; Essmann *et al.*, 1995; York *et al.*, 1995) and representation of the solvent and counterions-- can accurately represent the differences in structure between A-RNA, B-DNA and DNA:RNA hybrid duplexes. Spontaneously in the simulation of A-RNA, a single (α,γ) crankshaft transition in one strand, along with the observation of low twist and interstrand guanine stacking, occured at the central CpG step. This feature, seen in a variety of A-RNA (Portmann *et al.*, 1995) and A-DNA (Eisenstein & Shakked, 1995; Haran *et al.*, 1987; Ramakrishnan & Sundaralingam, 1993) duplex crystal structures is thought to result from crystal packing. These results suggest that this feature is a context

dependent, sequence specific structure that can appear even in the absence of crystal packing or specific hydration patterns. It is clearly a context dependent effect since this behavior is not seen at other, non-central, pyrimidine-3'-5'-purine steps in these simulations or in crystal structures. The results also suggest that the A-RNA structure does not have the sequence specific narrowing at the center of the duplex as seen in the B-DNA crystal structure (Prive *et al.*, 1991), nor the sequence specific bending patterns seen in the crystal (where TpG bends into the minor groove) or in molecular dynamics simulations (where TpG and CpG bends into the major groove) (Cheatham & Kollman, 1996b). Instead, a generalized roll into the major groove is observed.

The A-hybrid simulations demonstrate that the DNA strand can undergo an A-DNA to B-DNA transition, despite the presence of the A-RNA strand, to converge to a structure that has features very similar to what has been seen by NMR. In particular, the DNA strand has sugar puckers that interconvert between C2'-*endo* and C3'-*endo*, while the RNA strand sugars remain in a C3'-*endo* conformation. Additionally, a minor groove width is intermediate between A-RNA and B-DNA is observed along with positive base pair inclination to the helical axis, negative propeller twist and negative x-displacement from the helical axis. Similar to the A-RNA, the A-hybrid structure displays a small positive roll into the major groove; however the low twist and interstrand guanine stacking at the central CpG step is not observed. Clearly the DNA strand is more deformable, as the properties of the overall duplex are more similar to an A-form geometry than to a B-form geometry, yet the DNA strand still adopts primarily C2'-*endo* sugar puckers.

Each of the structures, A-RNA, B-DNA and A-hybrid, have distinct structural features which allow for discrimination and help explain how enzymes can distinguish between the structures (Fedoroff *et al.*, 1993; Lane *et al.*, 1993). In addition to the structural differences, there are also differences in the relative flexibility between A-form and B-form structures, the latter being considerably more flexible. Although the data is not conclusive, it appears the solvent and counterions influence the flexibility of the structures.

The rather specific association of water and counterions into the major groove and transversal water bridging the O2' atoms across the minor groove of A-RNA may stabilize the structure. The B-form structures on the other hand are characterized by more diffuse counterion association to the backbone and in the grooves and a less specifically stabilizing "spine" of hydration in the minor groove. Overall the more flexible structures show less specific hydration (which may result out of the analysis) and the fractional occupancy of counterions in the major groove of B-RNA is considerably less than that seen in A-RNA.

The convergence to the "same" B-form geometry, when simulations are started from canonical A-DNA or B-DNA, canonical B-RNA, or a canonical B-form hybrid duplex is somewhat surprising. It was previously thought that the unacceptable stereochemistry of the O2' hydroxyl "bumping" into the following phosphate group, sugar ring and base in RNA would destabilize the B-form geometry and make B-RNA unfavorable (Dickerson, 1984). While the hydroxyl does point up towards the following nucleotide (as can be seen in Figure 21b), the interaction is not unfavorable; in fact, the hydroxyl group hydrogen bonds with one of the phosphate oxygens, the O5' of the backbone, or both from the following residue. This stable interaction of the O2' hydroxyl, coupled with the increased barrier to sugar repuckering (Olson & Sussman, 1982), helps explain why B-RNA is a stable conformation during more than two nanoseconds of simulation and why it is difficult to force the B-RNA to A-RNA transition. Based on the current data and since we have not observed a spontaneous B- to A- RNA transition, it is impossible to determine which structure is more stable. Simulations are currently underway in an attempt to directly calculate the relative free energies of the A-RNA and B-RNA models and the free energy barrier to interconversion (Cheatham *et al.*, 1996c). However, the observation that C3'-*endo* to C2'-*endo* repuckering occurs less frequently than the reverse during 2 ns of simulation on the RNA duplexes, coupled with the observation of the sugar pucker transition to C3'-*endo* in the terminal cytosine of the RNA

strand from the B-hybrid and the large fluctuations, indirectly suggest that A-RNA may be the more stable, consistent with observation.

The observation of stable B-RNA in the RNA and hybrid duplexes suggest that conformational sampling of RNA is clearly an issue. Not only do the structures get locked in B-RNA or A-RNA conformations, the backbone can get locked into conformations that are persistent for longer than a nanosecond time scale (*i.e.* the ApC step in the simulation with the concerted flip in puckers). Little repuckering occurs, no B_I to B_{II} backbone transitions are observed in A-RNA simulations and few in B-DNA, and few (α, γ) crankshaft transitions are observed in over 4 ns+ simulation of RNA duplexes. Moreover, short simulations may not be sufficient to observe backbone transitions, such as the (α, γ) crankshaft seen after ~1 ns of simulation in only one strand of the A-RNA simulation.

The rigidity of the RNA presents difficulties to modellers of RNA. Limits in computer power, and complexity of the calculations, currently restrict simulations to a nanosecond time scale which is not long enough to allow sampling between the various conformational states. Clearly simulations on RNA could benefit from application of some reasonable means to enhance conformational sampling, such as locally enhanced sampling (Roitberg & Elber, 1991). With the current methods, it is likely that a given RNA model, such as a tRNA crystal or model structure will likely remain close to the initial model during a simulation, despite the validity of the starting structure, for many nanoseconds. This apparent stability of the structure does not validate the force field, per se, since the RNA may be caught in a metastable state. This has been observed in nanosecond time scale simulations of an RNA hairpin (Miller & Kollman, 1996). These results suggest that fairly long simulations, *i.e.* tens of nanosecond, may be necessary to investigate RNA structures via unrestrained molecular dynamics. Since the barriers to conformational transition in DNA are clearly smaller and transitions such as the A-DNA to B-DNA transition can readily be observed (Cheatham & Kollman, 1996b), simulations in the 1-2 nanosecond time scale may be sufficient to properly represent right handed DNA duplexes.

Despite these caveats with respect to conformational sampling, nanosecond time scale simulations seem to be able to provide useful insights into the overall sequence specific structure and dynamics of nucleic acids.

Acknowledgments:

P.A.K. is grateful to acknowledge research support from the NIH through grant CA-25644. T.E.C. would like to acknowledge research support as a NIH Biotechnology Training Grant (GM08388) and UCSF Chancellor's Graduate Research Fellow. We would also like to acknowledge the Pittsburgh Supercomputing Center (PSC) (MCA93S017P) and Silicon Graphics Incorporated for significant computational resources; the UCSF Computer Graphics Laboratory (RR-1081); Michael Crowley of PSC for parallelizing the PME code on the Cray T3D; Tom Darden for helpful discussion and for making the PME code available; Jim Vincent and Ken Merz for the original MPI version of AMBER; Matt Young & David Beveridge and Yong Duan & John Rosenberg for making preprints of their work available; Jed Pitera and Randy Radmer for helpful discussions about the density grid generation and contouring; and Jennifer Miller for helpful discussions.

B-DNA to A-DNA transitions? The effect of the environment on nucleic acid structure.

The simulations discussed in Chapter 1 clearly show that it is possible to reliably simulate B-DNA structures in aqueous solution on a nanosecond time scale. The discussion in Chapter 2 enhances this observation by demonstrating a preference for B-DNA structures over A-DNA structures. Six different simulations all converge to a common structure and the dynamics in each of these simulations suggests that the DNA is very flexible. In addition to stably simulating B-DNA structures, A-RNA is also stable on a multi-nanosecond time scale as is discussed in Chapter 3. In addition, hybrid duplexes started in an A-form geometry converge to a structure different than the corresponding A-RNA or B-DNA duplexes; this structure is consistent with experiment. The simulations also show the expected differences in dynamics and flexibility; A-form structures are much more rigid than corresponding B-form structures (irrespective of whether the duplex is RNA or DNA). These simulations suggest that the force field can differentiate DNA and RNA structures. Additional simulations on a variety of different sequences by myself and others further suggest that the force field can also differentiate between various DNA sequences and converge to different sequence specific B-DNA structures.

Although it is gratifying to see that A-RNA is stable with the Cornell *et al.* (1995) force field, the simulations on RNA also found a stable "B-RNA" structure. Even when the B-RNA simulation is continued beyond the two nanoseconds reported (now up to 8.5 nanoseconds, unpublished results), B-RNA is still stable. We would like to believe that the simulation protocol and force field favor A-RNA structures. In Chapter 3 it was argued that since it was possible to force a B-RNA to A-RNA transition and since the fluctuations in the B-RNA structures are larger than B-DNA that A-RNA should be more stable than B-RNA. However, this is a very weak argument and in fact unsubstantiated. To prove that A-RNA is more stable than B-RNA, either spontaneous transitions from B-RNA to A-

RNA need to be observed or alternatively the relative free energy determined. Simulations are currently in progress (Cheatham *et al.*, 1996c) in an attempt to do the later by applying weighted histogram analysis methods (WHAM) (Boczko & Brooks, 1993; Boczko & Brooks, 1995; Kumar *et al.*, 1992; Kumar *et al.*, 1995) to a large series of calculations on B-RNA, A-RNA and intermediate structures.

Despite the assertion that B-RNA is less stable than A-RNA, B-RNA is clearly stable on a multi-nanosecond time frame in the molecular dynamics simulations. Yet, B-RNA has never been observed experimentally either by X-ray crystallography or NMR and other spectroscopic methods to my knowledge. No B-RNA structures appear in the PDB database or nucleic acid database (Berman *et al.*, 1992). Moreover, it was thought that steric repulsion from the 2' hydroxyl group would disfavor the B-geometry (Dickerson, 1984). Therefore, the observation of B-RNA is somewhat surprising. After discussing our results with Richard Lavery, he checked to see if JUMNA (Lavery *et al.*, 1995) calculations would also produce a stable B-RNA; they do under certain conditions (personal communication, R. Lavery). Although B-RNA may represent a stable minimum energy conformation, it may never be observed experimentally since it is significantly less stable than A-RNA and hence represents a very minor population. However, B-RNA as a minimum energy conformation could also be an artifact of the force field representation. Maybe the force field "overstabilizes" B-form structures? Therefore it is necessary to resolve this issue and either overcome the conformational sampling problem or demonstrate a stable A-DNA under conditions which favor A-DNA. Conformational sampling is a major issue and care should be taken in relying on the results in cases when the results cannot be correlated back to experiment, such as in the case of B-RNA structures. Miller found similar conformational sampling difficulties with RNA hairpin structures which converged to different structures (a correct and incorrect NMR structure) depending on the starting structures (Miller & Kollman, 1996).

To move to the next level, neglecting for the time being the conformational sampling problem, it is critical to show some influence of the environment on nucleic acid structure. As mentioned briefly in the introduction to this paper, the structure of nucleic acids is profoundly influenced by the environment. In the next section the effect of the environment on the structure of nucleic acids is discussed; this presentation is not meant to be exhaustive but to supplement and reintroduce what was presented in the introduction to this thesis with some analysis of the implications for molecular simulations. This discussion also mainly concerns the transition between A-DNA and B-DNA structures since these are within the realm of what can be studied realistically with molecular dynamics simulations. Spontaneous structural transitions from Z-DNA to B-DNA or those involving the binding of proteins are likely not amenable to simulation methods, at this time, due to difficulties in conformational sampling. After this, some of the "less than successful" simulations attempted to show some environmental dependence are presented. The chapter is ended with a preprint of work investigating the simulation of $d[CCAACGTTGG]_2$ in ethanol and ethanol/water solutions.

The environmental dependence of nucleic acid duplex structure: The transition between A-DNA and B-DNA

The structure of nucleic acids is profoundly influenced by hydration, salt effects and interaction with other molecules. The dependence of the structure on hydration was readily apparent in the early fiber diffraction studies which demonstrated that decreasing the relative humidity lead to a B-DNA to A-DNA transition with Na-DNA (Franklin & Gosling, 1953). Similarly, reducing the effective water concentration through the addition of solvents such as ethanol, isopropanol or dioxane can also lead to a B-DNA to A-DNA transition (Ivanov *et al.*, 1973); however the effect is subtle since methanol does not induce

a B-DNA to A-DNA transition. Moreover, ions also modulate the nature of this transition. Shown in Figure 32 is a schematic representing melting temperature as a function of ethanol concentration for DNA in the presence of sodium or potassium as determined in mechanochemical studies of highly oriented DNA fibers (this is adapted from Piskur & Rupprecht, 1995). Piskur & Rupprecht claim that the local maximum in the T_m and the narrow change in T_m over this range implies cooperativity and an increase in interhelical interactions.

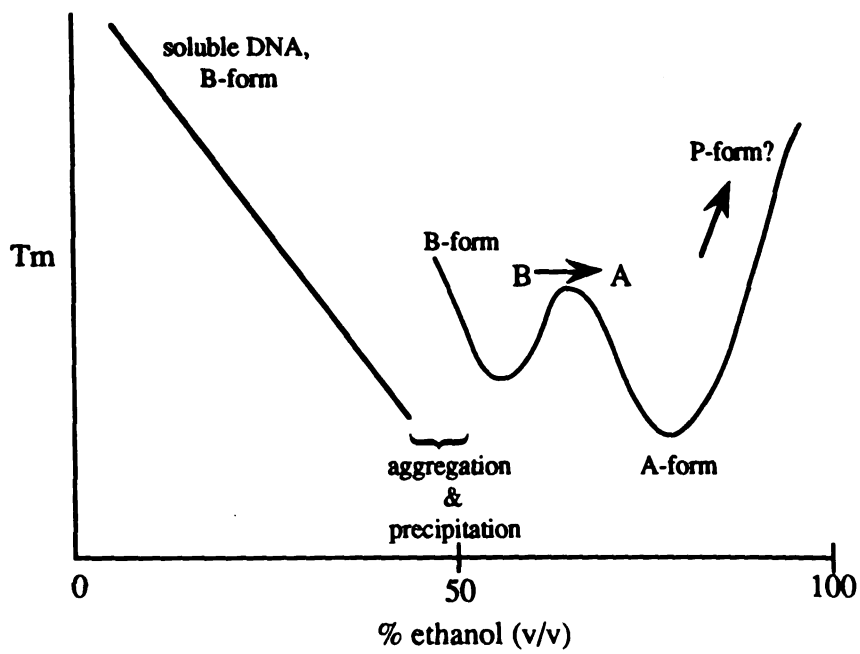


Figure 32: Schematic of melting temperature (T_m) as a function of ethanol concentration for Na^+ or K^+ DNA.

Increasing the ethanol concentration leads to aggregation and possible precipitation; this is a common laboratory technique for purifying DNA and care must be exercised to study DNA in solution at concentrations of ethanol above 50%. As the ethanol concentration gets very large, a very stable P-form DNA is formed. P-DNA is a highly aggregated form of DNA and characterized by a complete (and reversible) loss of base stacking (Zehfus & Johnson, 1984).

The above schematic is profoundly influenced by the nature of the counterion. In the presence of cesium, no local maximum in the melting temperature is seen at the B to A transition. In the presence of lithium or magnesium, no transition to A-DNA is observed and instead B-DNA to C-DNA is observed with lithium (Bokma *et al.*, 1987) and B-DNA to P-DNA with magnesium (Zehfus & Johnson, 1984). B-DNA is favored in the order $Cs^+ > Rb^+ > K^+ > Na^+$, which is likely due to direct coordination to minor groove base atoms and/or stabilization of the spine of hydration in the minor groove. The B-form is most stabilized by cesium atoms; crystallographically these ions are shown to bind deep into the minor groove (Bartenev *et al.*, 1983). AT-rich regions are stabilized more than GC-rich regions which helps explain the greater tendency of GC-rich sequences to undergo a B-DNA to A-DNA transition. This has been attributed to the lack of a distinct spine of hydration in the minor groove of GC base pairs (Drew & Dickerson, 1981). Lithium and magnesium, which are much more strongly hydrated, likely do not interact directly with the minor groove but through a water bridge (Bartenev *et al.*, 1983; Buckin *et al.*, 1994). Since these ions are more strongly hydrated, they are more effective at dehydrating the macromolecule which may partially explain why transitions to C-DNA or P-DNA are observed instead of B-DNA to A-DNA transitions.

A-DNA is stabilized in the opposite order; $Na^+ > K^+ > Cs^+$. This relates directly to relative strength of interhelical forces in A-DNA as measured by soft ($< 25\ cm^{-1}$) modes (Weidlich *et al.*, 1988; 1990). Aggregation and interhelical forces stabilize the A-DNA structure, however sedimentation studies clearly show that A-DNA can form in the absence of aggregation (Potaman *et al.*, 1980). In order for interhelical interactions and aggregation to stabilize A-DNA, something must overcome the repulsion from the negatively charged phosphate groups. As the concentration of ethanol increases, the effective dielectric constant drops and less water is available to screen the phosphate charges. This suggests that positively charged ions must somehow be involved. The same argument holds for P-

DNA as well, and in fact, aggregated Mg^{2+}-DNA does not dissolve on re-addition of water because of the strong interactions (Schultz *et al.*, 1994). Recently a paper was published by Robinson & Wang (1996) which showed that various polycationic ions could convert B-DNA to A-DNA and moreover that these shared common structural motifs. Specifically, hexaamminecobalt(III) or $Co(NH_3)_6^{3+}$, neomycin and spermine could induce a B-DNA to A-DNA transition in $d[A_2G_{15}C_{15}T_2]_2$. The led them to investigate the structure of a smaller sequence where it was found that $Co(NH_3)_6^{3+}$ bound to an A-form of $d[ACCCGCGGGT]_2$ in a 4:1 complex. Ions bind the A-DNA structure in two major modes; deep in the major groove interacting with the guanines and on top of the major groove bridging the phosphate groups from opposing strands (Gao *et al.*, 1995; Robinson & Wang, 1996). The binding deep in the major groove is particular; the square planar $Pt(NH_3)_4^{2+}$ presumably does not have the proper arrangement of the amine groups to interact with the guanines to stabilize the A-DNA conformation and a B-DNA to A-DNA transition is not observed. The interaction of the ions in the major groove and bridging strands also likely stabilizes Z-DNA structures. Binding of cations in the major groove is also a mechanism by which proteins interact with nucleic acids, such as the arginine rich HIV-1 Rev peptide interacting with RRE RNA (Battiste *et al.*, 1996). It may also be involved with SASP binding to DNA which induces a B-DNA to A-DNA transition, however no structural data has been reported yet in this system.

The general conclusion from these observations is that anything which stablizes the minor groove and minor groove hydration in B-form structures should stabilize B-DNA and anything which helps stabilize interhelical contacts, screens charge and/or counterions binding in the major groove should stabilize A-DNA. Consistent with this, B-DNA is stabilized, and an A-DNA or Z-DNA to B-DNA transition induced in ethanolic solutions by the addition of the minor groove binders netropsin and distamycin A (Burckhardt *et al.*,

1996). Also consistent is the B-DNA to A-DNA transition, discussed above, induced by the interaction of polycationic ligands with the major groove.

Preliminary simulations investigating the effect of small perturbations to the force field.

In order to investigate the effect of the environment, the simplest simulations to perform are those that modify the parameters or change the simulation protocol by including molecules which have been parameterized and tested with the Cornell *et al.* (1995) force field. This avoids the painful task of developing new parameters. Moreover homogeneous systems are much easier to set up than systems with multiple solvent or solute types. Additionally, the simulations should be sufficiently similar to spirit to those previously run so that the earlier simulations can serve as a control. To this end, a variety of simulations were run which tweaked the parameters and played around with the ionic strength. Part of the incentive for performing these calculations was as a kind of "sensitivity analysis"; in other words, it is important to understand how sensitive the results are to the particular parameters and ionic strength. Surprisingly little effect was seen in these simulations, except for simulations where the sugar puckers were held fixed at C3'-*endo* or C3'-*endo*. The ethanol simulations discussed at the end of the chapter were "on the horizon" and slated for future simulation since Thomas Fox in the group was near to completing his development of various solvent parameters at the time many of these simulations were run.

Before briefly presenting the results from these simulations, it should be mentioned that a large incentive for trying these out was to investigate some of the deficiencies found in earlier simulations. In Chapter 2, it was mentioned that the largest deficiencies in the

current representation of the B-DNA structure was a lower than expected helical twist (~30°). Experimentally, helical twists in the 34° to 35° range are expected (Peck & Wang, 1981; Rhodes & Klug, 1980; Ulyanov & James; 1995). The low twist observed in the simulations could be related to an oversampling of sugars in the O4'-*endo* range or also because of too much interstrand phosphate repulsion, as was seen in previous internal coordinate molecular mechanical treatments by Zhurkin and co-workers (Zhurkin *et al.*, 1982). Moreover, experimentally it is known that small counterions and low ionic strength lead to decreases in the overall twist (~1-2°) (Anderson & Bauer, 1978; Wang, 1969). To investigate this issue, a series of simulations were run with no salt, high salt, and reduced phosphate charges. In all of these, a protocol matching that used in Chapters 2 and 3 was applied to either the average structure from the A-DNA simulation in water (calculated from the **A1** trajectory after the transition to B-DNA or from 580 ps to 1456 ps, $\mathbf{A1}_{avg}$) or continued the B-DNA in water simulation from one nanosecond (**B2**) [see Chapter 2]. In most of the simulations little effect on the B-DNA structure was noticed and average structures calculated from the trajectories were generally within ~1.5 Å of the starting structure which is within the range of the fluctuations. The following simulations were run:

- **no salt** ($A1_{avg}$ for 1085 ps): No counterions were included in the simulation and the average net charge was removed by smearing the average net charge over every atom. No effect on the average structure was observed and the calculated average structure did not deviate significantly from the control (0.89 Å). The hope was to see a lowering of the helical twist or at least some distortion of the helix because of the increased phosphate repulsion, however none was seen. At the no salt limit, previous calculations by Zhurkin (1982) suggested the DNA would be very undertwisted and likely unstable.

- **high salt** ($A1_{avg}$ for 1350 ps): In addition to the 18 Na^+ counterions, 50 additional Na^+ and Cl^- atoms were added to the simulation, in hand picked randomly distributed positions (replacing water molecules). No effect on the average structure was observed and the average structure remains within 1.20 Å of the control average structure. The hope was to see an effective raising of the helical twist, however no effect was seen.

These two simulations with various salt concentrations when compared to the control simulations suggest that the amount of salt present is not critical. This is both good and bad. The "good news" is that including large amounts of salt is not necessary; this vastly simplifies the setup and equilibration. The "bad news" is that a salt effect is not seen. After these simulations were run, Young *et al.* (1996) sent us a preprint of their work which investigates the association of counterions in the grooves of DNA. This was discussed to some degree in Chapter 3. Consistent with the results of that paper, the high salt simulations run here show counterions entering the minor groove. More interesting is the rather close approach of Cl^- ions. It is expected that divalent salts (such as Mg^{2+}) will have a much more profound effect; discussions with Matt Young suggest that he has seen a larger effect with divalent counterions in his simulations.

- **high salt, poly(G)-poly(C)**: Although not previously mentioned, simulations were run on 10-mer poly(G)-poly(C) sequences in both canonical A and canonical B geometries. An A-DNA to B-DNA transition was observed within 300 ps. Since sequences with high GC content are more likely to crystallize into an A-DNA structure, high salt simulations were also run on poly(G)-poly(C) since it was expected that this sequence would more easily transition to A-DNA. However, no effect on the structure was seen in over one nanosecond of dynamics.

- **"mock water"** (B2 for 800 ps): The charges on the TIP3P water atoms were reduced by 10%. The average structure remains within 0.48 Å, however water properties are

significantly altered. The water diffusion constant is roughly doubled (11.27×10^{-5} cm^2/s). Although the fluctuations in the helicoidal parameters do not appear significantly larger than the control simulations, the plot of RMSd over time has spikes that show that the structure moves up to 3.5 Å away from the starting structure during the trajectory. Clearly the dynamic properties are altered. Note that in these simulations, the goal was to reduce the strength of hydration which may possibly allow an B-DNA to A-DNA transition. Reducing the water charges by 20% led the system to continuously expand or "blow up" until the simulation terminated. This can be prevented by running with constant volume conditions or constant pressure with large coupling times.

- **SPC/E water** (B2 for 1155 ps): TIP3P (Jorgensen *et al.*, 1983) and SPC/E (Berendsen *et al.*, 1987) are the common rigid three point watel models applied in the simulation of biomolecules. SPC/E has some more favorable properties, in particular the diffusion is closer to experiment (at the expense of the average water-water energy). The purpose of this simulation was to see if changing the water model had any effect on the structure of our B-DNA model. No significant change in the structure was observed (RMSd 1.0 Å).

- **reduced phosphate charges** (B2 for 1060 ps): Zhurkin previously demonstrated that interstrand phosphate repulsion could lead to a reduced helical twist (Zhurkin *et al.*, 1982). To goal therefore was to reduce the phosphate charge to see if the helical twist would go up. The charges on the phosphates were reduced by half. This was accomplished by adding 0.2 to the phosphate oxygens and 0.05 to the O5' and O3' oxygens. The change on the Na^+ counterions was reduced to 0.5 to balance the phosphate charge. No effect on the structure was observed (RMSd 0.69 Å).

The average helicoidal parameters for the last three simulations, compared to a control simulation, are plotted in Figure 33 to allow comparison between the structures.

From this plot it is evident that the structures are very similar; the differences are typically within the range of the fluctuations.

- **reduced base stacking** (B2 for 1000 ps): The idea here is that perhaps over-stacking of the base pairs is leading to undertwisting in the DNA structure. To test this, the dispersion attraction interactions were reduced for all the non-hydrogen base atoms. This is accomplished by reducing the well depth (ε) by 20% and increasing the radius (r^*) to keep the nonbonded repulsion roughly the same. This procedure results in a reduction in the strength of the dispersion attraction interactions by roughly 10%. No effect on the structure was seen (RMSd ~0.71 Å). This suggests that the structure is not tremendously sensitive to the van der Waals parameters. However visualizations of the dynamics ("movies") and the RMSd over time to the starting structure suggest that more movement is seen during the dynamics. This however does not greatly effect the fluctuations in the helicoidal parameters.

- **smaller periodic boxes** (1000 ps): Rather than surround the DNA initially by 10 Å of water in each direction, the DNA was surrounded by 5 Å. This leads to a significantly smaller box (~43 Å by ~36 Å by ~36 Å) and significantly fewer atoms (~5700 compared to ~9300 atoms). No effect on the calculated average structure was observed (RMSd ~0.69 Å). This method is not generally advisable however since rotation in the periodic box will bring the ends of the duplex very close together. If the rotation is inhibited this may prove very useful for NMR refinement since the computational complexity is lessened and the restraints should removing any biasing effect of the true periodicity.

- **low pressure** (B2 for 1000 ps): Similar in spirit to the "mock water" simulations, the goal was to reduce hydration, in this case by running the simulation with a negative target pressure. Interestingly, as the box initially expands, there is a transient move

closer to A-DNA, however this does not persist and no effect is seen on the average structure.

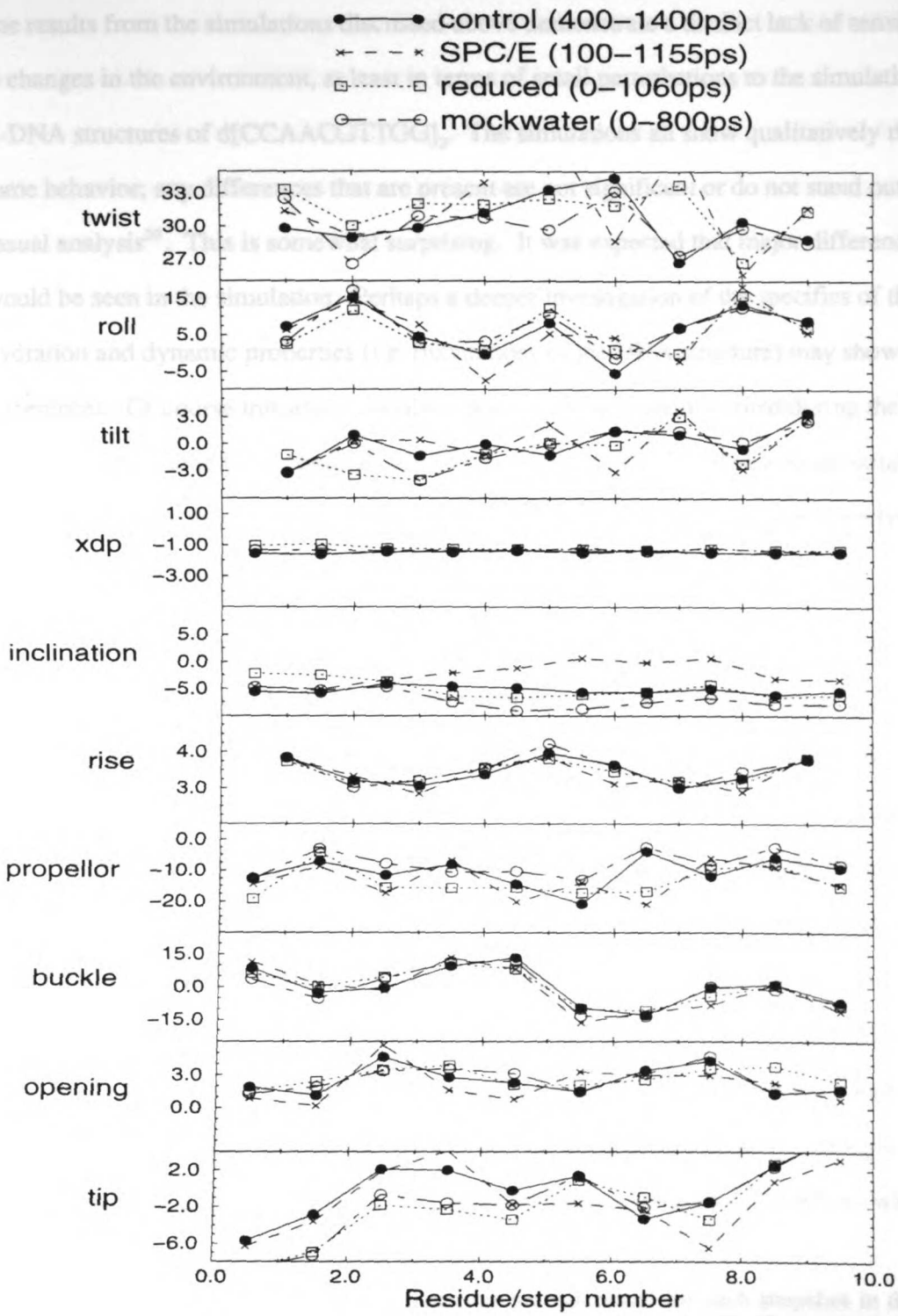


Figure 33: Helicoidal parameters from selected simulations. The twist, roll, tilt, inclination propellor twist, buckle, opening, and tip are all in degrees; the x–displacement from the helical axis and rise are in angstroms

The results from the simulations discussed above demonstrate a distinct lack of sensitivity to changes in the environment, at least in terms of small perturbations to the simulation of B-DNA structures of d[CCAACGTTGG]$_2$. The simulations all show qualitatively the same behavior; any differences that are present are not significant or do not stand out in the casual analysis[20]. This is somewhat surprising. It was expected that major differences would be seen in the simulation. Perhaps a deeper investigation of the specifics of the hydration and dynamic properties (*i.e.* fluctuations of the DNA structure) may show some differences. Of course this analysis only applies to conformations visited during the trajectory which tend to be conformations in the vicinity of the B-DNA average structure. It is not clear how these perturbations in the simulation protocol may change the relative stability of other conformations, such as the A-DNA.

Perhaps what is most distressing in this analysis is that the B-DNA structure of d[CCAACGTTGG]$_2$ we do find is different than the crystal structure (Prive *et al.*, 1991) and that small perturbations to the force field and protocol have little effect. The B-DNA structure we do find may be relevant, however there is no direct "solution" data to compare with. Therefore it is difficult to judge the "validity" of the structure. Clearly the structure has some validity as it appears to be close to the crystal when only the central residues are considered (Chapter 2, Table 5). Moreover the differences observed in the model structure are likely due to bends at the TpG and CpG steps which is consistent with the NMR data (Ulyanov & James, 1995). In some sense, we should be grateful that we are doing this well! What is somewhat worrisome to me is that the simulation protocol is strongly favoring a *certain* B-DNA structure. This is seen in the simulations of "B-RNA" which

[20] For every trajectory run, helicoidal parameters are calculated for each snapshot in the trajectory at 1 picosecond intervals. This data is processed to calculate the averages over the simulation and the standard deviations. The standard deviations represent the "fluctuations". Additionally, the helicoidal parameters are calculated for the average structure computed by a coordinate average of the all DNA atom RMS fit trajectory. From this analysis, all of the structures from the various simulations discussed above are very similar.

converge to a very similar structure. From the data we have so far, it is not possible to determine whether this B-DNA structure is completely representative of the actual conformation of this sequence in aqueous solution. It was hoped that small perturbations, such as performed above, would provide the means to "tweak" the simulations to give "more reasonable" results. In other words, these small perturbations would have a larger effect in modulating the DNA structure so that they could be used later to develop a better protocol and force field for simulating nucleic acid structure. Based on the "stability" of the B-DNA models studied, and the specific limitations (such as low twist), it is not clear what may eventually provide the means to "tweak" the protocols to better represent nucleic acid structure.

Sequence dependent structure and correlation with experiment.

The lack of sensitivity to the simulation protocol is rather surprising. Moreover, it is difficult to judge the validity of the "B-DNA" model of d[CCAACGTTGG]$_2$ since the only structures of this sequence solved experimentally have been using X-ray crystallography. Concurrent with the above work, simulations investigating DNA structures that have been determined by NMR methods were also initiated. These should be more representative of the "solution" conformation of B-DNA. Before briefly discussing these results, it should be noted that the structures determined by NMR are in some cases very sensitive to the refinement protocol (Leijon *et al.*, 1995). However, on average our simulated structure should hopefully be able to satisfy the NMR NOE and coupling data; this level of analysis has not be applied yet and will not be discussed in this section. Preliminary simulations have been run on the sequence d[ACGTTGCCTTGAG]-d[CTCAAGGCAAGCT] starting with the high resolution NMR structure (Mujeeb *et al.*, 1993) and run for one nanosecond (applying the simulation protocol outlined in Chapter 2). In Figure 34 is shown the

helicoidal parameters and in Figure 35 the average structure from the simulation compared to the NMR derived structure. Clearly there are distinct differences in the helicoidal parameters. In general, the simulation methods reproduce the experimental rolling patterns, but completely misrepresent the sequence specific twisting patterns. Overall the simulated structure is significantly under twisted. However the structure is not completely unreasonable as the RMSd of the NMR structure to an average structure computed over the last 800 ps of the trajectory is only 2.40 Å. The agreement is even better over eight residue (:1-8 represents the first eight base pairs) windows.

residues	:1-8	:2-9	:3-10	:4-11	:5-12	:6-13
RMSd (Å)	1.82	1.83	1.84	2.05	2.11	1.84

Table 10: Root mean square deviation over eight residue windows in the average structure from 200-1000 ps compared to the NMR structure (Mujeeb et al., 1993) of d[ACGTTGCCTTGAG].

Further analysis of this data, specifically to investigate the time averaged agreement of the NMR restraints, and other NMR structures is necessary to better validate the force field representation. In spite of this, particle mesh Ewald methods and explicit solvent may prove very useful in NMR structural refinement of nucleic acids.

In addition to the simulations reported above, simulations have also been run on various other 10-mer DNA duplexes, including poly(A)-poly(T), $d[TTTTTAAAAA]_2$, $d[AAAAATTTTT]_2$, $d[TTTTCGAAAA]_2$, $d[AAAACGTTTT]_2$, and $d[ATATATATAT]_2$. In each of the cases where the DNA duplex was started in a canonical A geometry, a transition to B-DNA was observed on a 500 ps time scale. Moreover, each of these simulations converges to a different structure and show expected sequence specific bends.

All show uncharacteristically low helical twist values, however poly(A)-poly(T) has the highest average twist (~32°). This is expected; experiments suggest a reduced periodicity in solution compared to random sequence DNA (Rhodes & Klug, 1980). The results will be presented in detail in an upcoming paper.

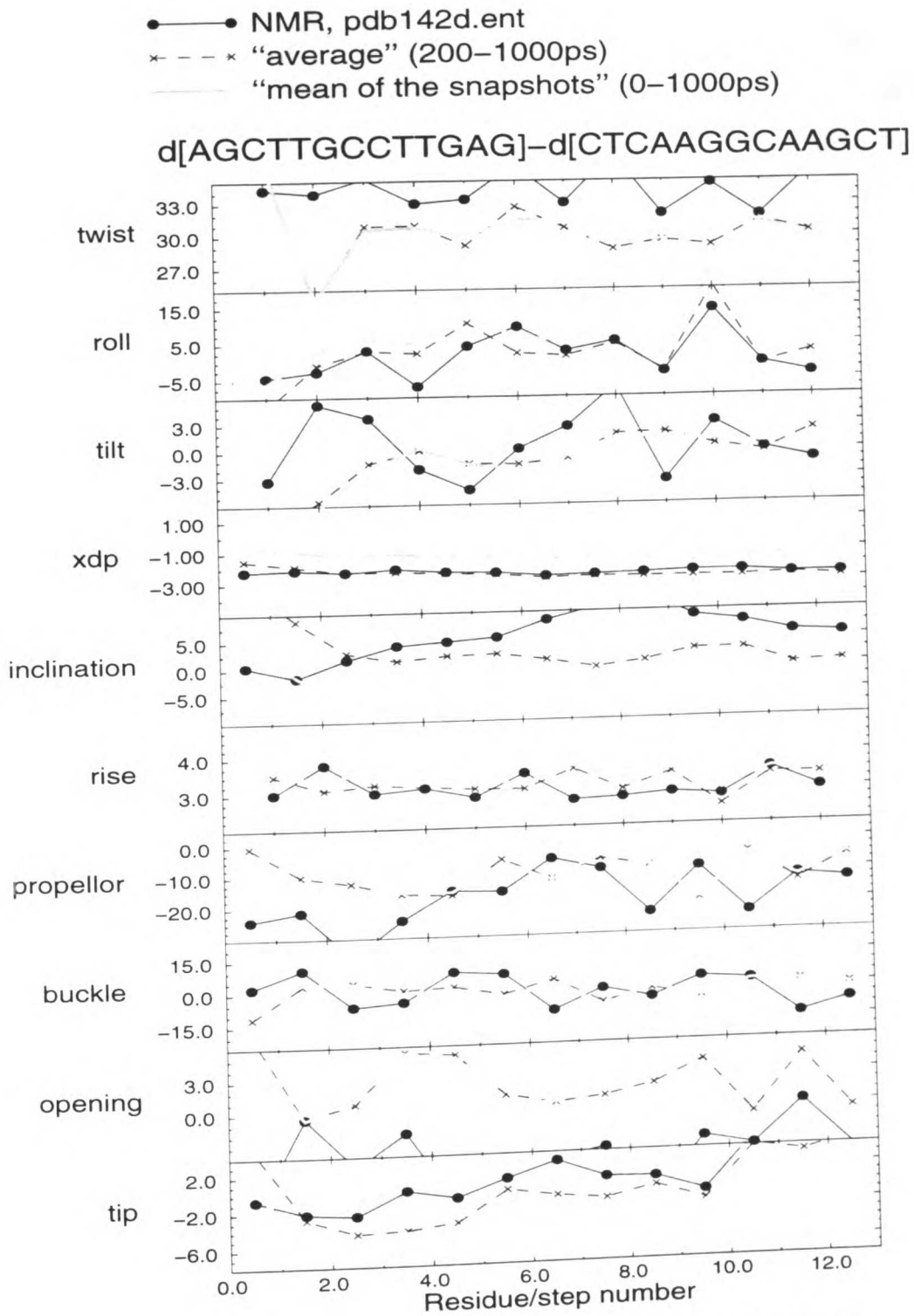


Figure 34: Helicoidal parameters calculated from the average structure and mean of the snapshots at 1 ps intervals (both over 200–1000 ps) compared to the NMR structure of d[AGCTTGCCTTGAG]–d[CTCAAGGCAAGCT] (Mujeeb et al., 1993).

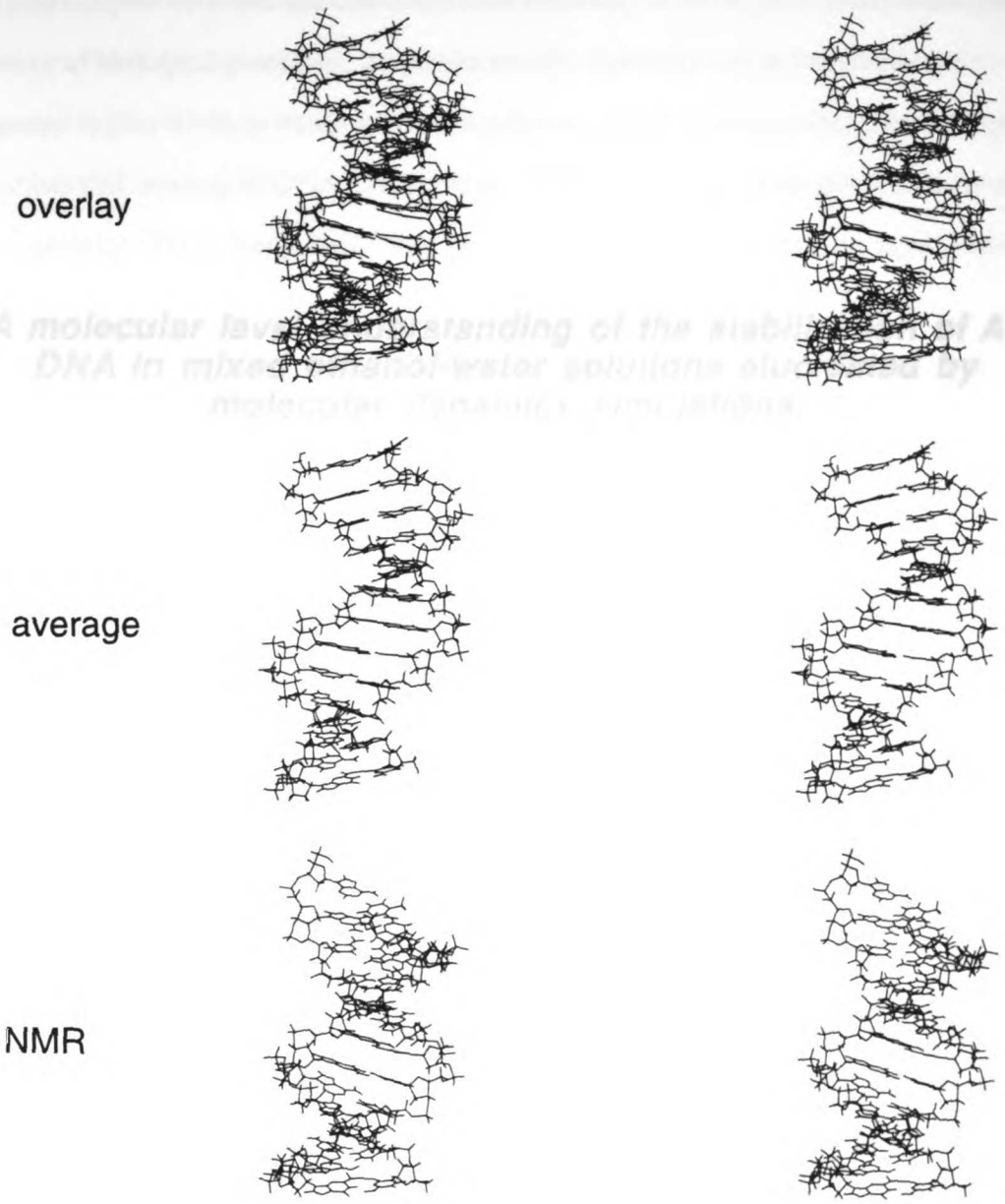


Figure 35: Stereo view plots of the average structure from 200-1000 ps compared to the NMR structure (Mujeeb et al., 1993) of d[AGCTTGCCTTGAG]-d[CTCAAGGCAAGCT]. Shown are an overlay of the structures and both structures separately; the structures were all atom RMS fit prior to the display.

A molecular level understanding of the stabilization of A-DNA in mixed ethanol-water solutions elucidated by molecular dynamics simulations.

Thomas E. Cheatham, III

Michael F. Crowley*

Thomas Fox

and Peter A. Kollman

University of California San Francisco

Department of Pharmaceutical Chemistry

San Francisco, CA 94143-0446

and

* Pittsburgh Supercomputing Center

4400 Fifth Avenue

Pittsburgh, PA 15213

The polymorphic structure and conformational flexibility of DNA plays an important role in a variety of biological processes. Sequence specific flexibility and deformability are suggested to play a role in transcriptional regulation; specific examples include transcription factor induced bending of DNA (Ansari *et al.*, 1995), unwinding of the DNA helix by zinc finger proteins (Shi & Berg, 1996), and the transportation of one duplex through another by type II DNA topoisomerases (Berger & Wang, 1996). In addition to structural deformations induced by ligand or protein binding, conformational transitions in DNA among the various canonical forms is also biologically relevant. A particularly intriguing example is the induction of an A-form geometry in an otherwise B-DNA duplex by α/β-type small acid-soluble spore proteins which serves to increase resistance to UV radiation damage in Gram-positive bacteria (Mohr *et al.*, 1991). Local transitions to Z-DNA in alternating purine-pyrimidine segments may also help promote and regulate DNA condensation (Levin-Zaidman *et al.*, 1996). These conformational transitions occur because the conformation of DNA is strongly dependent upon the environment. Variations in ionic strength and identity, sequence, water activity and ligand/protein binding all can modulate DNA structure[21]. An example is the B-DNA to A-DNA transition observed in

[21] Examples of environmentally induced changes in secondary structure include increasing ethanol concentrations leading to B-DNA to A-DNA (Ivanov et al., 1973), C-DNA (Bokma *et al.*, 1987) or P-DNA (Zehfus & Johnson, 1984) transitions depending on the nature of the counterion (Piskur & Rupprecht, 1995); binding of neomycin or spermine to induce a B-DNA to A-DNA transition (Robinson & Wang, 1996) or binding of distamycin A or netropsin to induce a A-DNA or Z-DNA to B-DNA transition (Burckhardt *et al.*, 1996); or binding of $Co(NH_3)_6^{3+}$ to induce a B-DNA to A-DNA or Z-DNA transition (Robinson & Wang, 1996). Common to each of these means to transition among the various structural forms seems to be both a subtle modulation of the hydration of DNA and the ionic association with the negatively charged phosphate backbone and grooves. Models such as the "groove binding model" (Bartenev *et al.*, 1983; (Ivanov et al., 1973) explain the stabilization of B-DNA by ions interacting directly with base atoms in the minor groove (Cs+>K+>Na+) or indirectly via water bridges (Li+). B-DNA is also stabilized by extensive solvation of the grooves and backbone (Falk *et al.*, 1963), such as the "spine of hydration" in the minor groove (Drew & Dickerson, 1981). A-DNA, on the other hand, is significantly less hydrated (Falk et al., 1963; (Wolf & Hanlon, 1975) and is stabilized by aggregation and interhelical contacts (Piskur & Rupprecht, 1995) and also by ions interacting primarily in the major groove (such as $Co(NH_3)_6^{3+}$) (Robinson & Wang, 1996) and with the phosphates. There is a profound dependence on the nature of the ionic species; for example, the square planer $Pt(NH_3)_4^{2+}$ ion and strongly hydrated Mg^{2+} and Li^+ ions, in contrast to the octahedral $Co(NH_3)_6^{3+}$ ion, do not induce a B-DNA to A-DNA transition (Robinson & Wang, 1996).

76%, 80% or 84% ethanol (v/v) solutions of DNA fibers in the presence of Na^+, K^+, or Cs^+ respectively (Piskur & Rupprecht, 1995). In contrast, the more strongly hydrated counterions (Mg^{2+}, Li^+) inhibit the transition to A-DNA and B-DNA to P-DNA (Zehfus & Johnson, 1984) or B-DNA to C-DNA (Bokma et al., 1987) transitions are observed instead. Although a number of theories have been advanced regarding the molecular mechanisms underlying the stabilization of A-DNA in mixed water/ethanol solutions, no one has solved a structure of an "isolated" A-DNA in a mixed ethanol/water solution. In this report, we demonstrate that molecular dynamics simulations can provide this molecular level description. In constrast to pure water, where an A-DNA to B-DNA transition is observed on a roughly 500 ps time scale (Cheatham & Kollman, 1996b), simulations of $d[CCAACGTTGG]_2$ in pure ethanol or mixed water ethanol show a profoundly different behavior. In pure ethanol, simulations starting from an A-DNA structure move away from the canonical geometry (~4.0 Å all atom root-mean-square deviation, or RMSd, by ~2 ns) and local distortions in the helicoidal parameters are evident. In contrast, A-DNA in a mixed water and ethanol solution remains in a canonical A geometry for at least three nanoseconds of molecular dynamics.

Molecular dynamics simulations on nucleic acids have a relatively rich history starting with the first simulations of Levitt (Levitt, 1983) and Karplus (Tidor *et al.*, 1983). The early simulations (see reviews by Beveridge (Beveridge & Ravishanker, 1994; Beveridge *et al.*, 1993)) were generally limited to short time scales (< 200 ps), typically displayed anomolous structure such as base pair fraying, and demonstrated the need for including some representation of solvent and a reasonable treatment of the highly charged phosphate backbone. Advances in computer power, empirical force field representations, and the development of more reasonable means to handle the long ranged electrostatic interactions now allow routine "stable" nanosecond length unrestrained molecular dynamics simulations of nucleic acids in water with explicit counterions (for a review see Louise-May *et al.* (Louise-May *et al.*, 1996)). From these studies, it has become clear that a well

balanced force field is necessary to properly represent the expected structural preferences and dynamics. However, the precise balance in the force field has not been fully put to the test. While our previous simulations demonstrate that B-DNA is favored over A-DNA in pure water (Cheatham & Kollman, 1996b), as is expected, "stable" B-DNA could have been an artifact of the force field representation. Therefore, simulations were run using the same $d[CCAACGTTGG]_2$ sequence, same force field (Cornell *et al.*, 1995), and an equivalent simulation protocol[22] applying the particle mesh Ewald method (Essmann *et al.*, 1995) implemented into AMBER 4.1 (Pearlman *et al.*, 1995) under conditions which are expected to stabilize A-DNA. The results suggest that the current DNA simulation protocols have reached a sufficient level of realism such that they can properly represent the effect of the environment on DNA structure.

When A-DNA is placed in pure ethanol (with or without counterions)[23], the structure moves away from the canonical A geometry, with all atom RMSd values rising

---

[22] All the simulations were performed applying the particle mesh Ewald method within AMBER 4.1 with constant pressure and temperature (300K) with Berendsen temperature coupling, SHAKE on hydrogens, a 2 fs time step and a 9 Å cutoff applied to the Lennard-Jones interactions. Differences from the previous study include the use of an all-atom ethanol model and slightly longer equilibration periods when the DNA was held fixed (100 ps in pure ethanol, and ~300 ps in the water/ethanol with the ions and water held fixed as well) in order to better equilibrate the ethanol prior to production simulations. An all-atom flexible ethanol model was developed using the standard nonbonded and intra-molecular parameters from the Cornell *et al.* force field and charges generated from a restrained electrostatic potential fit to a 6-31 G* wavefunction of the minimum geometry of ethanol (Bayly *et al.*, 1993). The charges and atom types for the CH_3CH_2OH molecule are as follows: CT, -0.0990; HC, 0.0345; CT, 0.3318; H1, -0.0294; OH, -0.6718 and HO, 0.4143. These charges correspond to a total dipole moment of 1.77 Debye which is about 5% higher than the gas phase dipole moment for ethanol of 1.68 Debye. Molecular dynamics simulations of a periodic box of 213 ethanol molecules under constant pressure conditions yielded, with this parameter set, a density of 0.781 g/cm^3 and a heat of vaproization of 9.989 kcal/mol; these compare well with the experimental values of 0.789 g/cm^3 (error -1.0%) and 9.67 kcal/mol (error 3.2%), respectively. When PME is applied, the density (0.800 g/cm^3) and heat of vaporization (10.21 kcal/mol) increase slightly.

[23] The simulation of A-DNA in pure ethanol with Na^+ counterions had 929 ethanol molecules and was placed in a box of ~58 Å by ~46 Å by ~46 Å and represented 9011 atoms. The simulation without counterions had 941 ethanol molecules in a box of roughly the same size and represented 9101 atoms; the net charge of -18.0 on the system was neutralized by smearing it over all atoms, or by subtracting -18.0/9101 from the charge on each atom. The latter simulation was run since in the simulation with counterions, the non-hydrated counterions never moved away from the phosphate atoms which could have lead to the distortions in the structure. However, when the counterions were removed, the simulation moved away from canonical A-DNA even more rapidly.

monotonically and approaching ~4 Å after ~2 ns of simulations. The structures visited during the dynamics are characterized by local distortions in the helicoidal parameters. These distortions in the structure are not surprising since no water is present in these simulations to solvate the DNA; even under extremely dehydrating conditions, A-DNA still has some tightly associated water (Wolf & Hanlon, 1975). Moreover, in solutions of greater than 90% ethanol, A-DNA to P-DNA transitions are typically encountered (Piskur & Rupprecht, 1995; Zehfus & Johnson, 1984). P-DNA is a highly aggregated form of DNA that appears to have lost all of its base stacking. In these simulations, since we are only simulating one structure in a periodic box, the nucleic acid cannot aggregate. However, the simulations in pure ethanol suggest that we are seeing some secondary structural collapse in the DNA (as is evident by low twist (~18°), high roll (~23°) and large rise (~4.8 Å) at some base pair steps), which is consistent with experiment, however the simulations have not been run long enough to definatively support this claim. Despite this, these results are presented since the behavior observed in pure ethanol is clearly different than is observed in water or in a mixed ethanol/water solution.

In mixed water and ethanol solutions with ~80% ethanol, it is expected that A-DNA should be stable. Therefore simulations were run in mixed water and ethanol solutions. To avoid a very long equilibration protocol to hydrate the DNA when the nucleic acid is placed into a pre-equilibrated box of water and ethanol, "hydrated" DNA was placed into a box of pure ethanol. In other words, a shapshot from a previous trajectory of A-DNA was taken[24] and stripped of all the waters except the 500 waters closest to any DNA or counterion atom. This leads to ~20 waters/nucleotide and ~6 waters/counterion which is equivalent to the hydration expected for B-DNA (Falk et al., 1963; Wolf & Hanlon, 1975).

[24] The A-DNA "snapshot" was generated and equilibrated from a canonical A-form geometry as described previously (Cheatham & Kollman, 1996b) except that that the pucker was forced to remain C3'-endo by the addition of a flatwell restraint on the C1'-C2'-C3'-C4' torsion to keep the angle between 30° and 40° as discussed in our previous work (Cheatham & Kollman, 1996a); this prevents the structure from undergoing an A-DNA to B-DNA transition. The snapshot represents the structure after 20 ps of production dynamics; at this point, the ions are still in rather close proximity to the phosphate groups of the nucleic acid.

A three nanosecond molecular dynamics simulation was performed on A-DNA in this mixed ethanol/water solution[25] In Figure 36 is presented the RMSd to A-DNA over the course of the simulations in pure ethanol and mixed water/ethanol. The structure in the mixed water ethanol simulation stays within ~2.0 Å of canonical A-DNA.

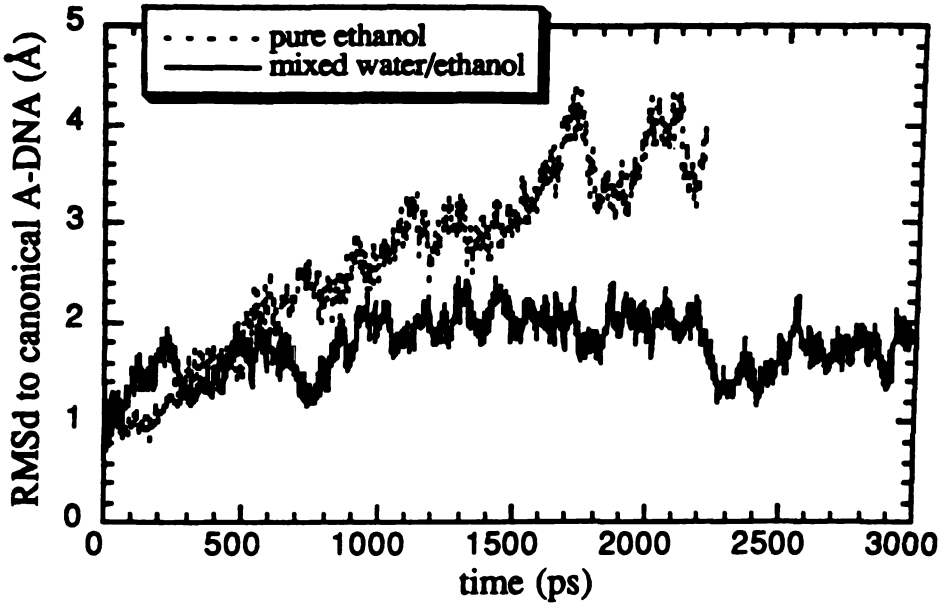


Figure 36: RMSd over the course of the A-DNA simulation. The root-mean-square deviation (RMSd) of all DNA atoms to canonical A-DNA is shown as a function of time for the simulation of A-DNA in mixed ethanol/water solution (black line) and in pure ethanol (dashed line) .

The helicoidal parameters presented in Table 11 confirm that the structure is canonical A and moreover even closer to canonical A than the A-RNA structure calculated previously for this sequence (Cheatham & Kollman, 1996a).

[25] The simulation contained 500 waters and 1240 ethanol molecules in addition to the DNA atoms and 16 Na^+ counterions. The equilibration protocol involved performing ~300 ps of dynamics with the DNA, Na^+ ions, and water held fixed to allow the ethanol to equilibrate. The box size was ~60 Å by 50 Å by 50 Å and the PME charge grid was 64 Å by 54 Å by 54 Å. Otherwise, the simulation conditions were equivalent to our previous work (Cheatham & Kollman, 1996a).

	A	B (in water)	A-RNA
twist (°)	32.6	30.9	30.9
rise (Å)	2.92	3.26	2.70
inclination (°)	15.8	4.9	15.0
propeller (°)	-12.6	-10.4	-12.6
x-disp (Å)	-4.0	-2.96	-5.19
RMS to A (Å)	1.66	3.85	2.09
pucker (°)	80.1	122.8	22.6
χ (°)	222.6	234.2	201.7
δ (°)	102.8	116.6	79.3

Table 11: Selected helicoidal and backbone angle values. Values of various helicoidal parameters, backbone angles, and sugar pucker values are represented for the A-DNA simulation in mixed water and ethanol solution (A, over 2000-3000 ps). For comparison, values of B-DNA in pure water (B, over 400-1400 ps) and A-RNA (A-RNA, over 1030-2030 ps) are also presented. All of the helicoidal values were calculated using the Dials and Windows interface (Ravishanker et al., 1989) to Curves (Lavery & Hartmann, 1994) from average structures over 1 nanosecond portions of the trajectories. The average structures were created by RMS fitting all DNA atoms at 1 ps intervals from the trajectories and coordinate averaging. The helicoidal values reported are averages over all base pairs or base pair steps where appropriate. The sugar pucker, χ and δ angles are 1 ps averages over 1 nanosecond portions of the dynamics, averaged over all nucleotides.

The structure is clearly very close to canonical A-DNA and does not display the intrastrand stacking at the central CpG step in corresponding A-RNA simulations (Cheatham & Kollman, 1996a) and in a number of A-DNA crystal structures (Ramakrishnan & Sundaralingam, 1993; Eisenstein & Shakked, 1995). The differences from canonical A-DNA are an average pucker value that is higher than expected which indicates significant sugar repuckering to C2'-*endo* and an axial rise that is slightly higher than expected; this leads to a structure that is perhaps closer to A'-DNA (Saenger, 1984).

The observation of a "stable" A-DNA structure, despite this higher average pucker and repuckering between C3'-*endo* and C2'-*endo*, suggests that it is not a difference in the relative stability of C3'-*endo* sugar puckers in mixed ethanol/water versus pure water that leads to the stabilization of A-DNA. In Figure 37 is presented the time course of the sugar puckers. Sugar repuckering occurs at every nucleotide with the terminal residues

displaying more C2'-*endo* puckering on average. The time course is shown to demonstrate that there is not a general trend towards C2'-*endo* puckers at later parts of the simulation, or in other words, we are not seeing a slow transition to B-DNA. Interestingly, the terminal groups, particularly the guanines, spend a significant time with C3'-*endo* puckers. It is as if the structure is trying to convert to a B-DNA structure but something is holding the middle residues in an A-DNA geometry; this is likely water and counterions in the major groove as is discussed later.

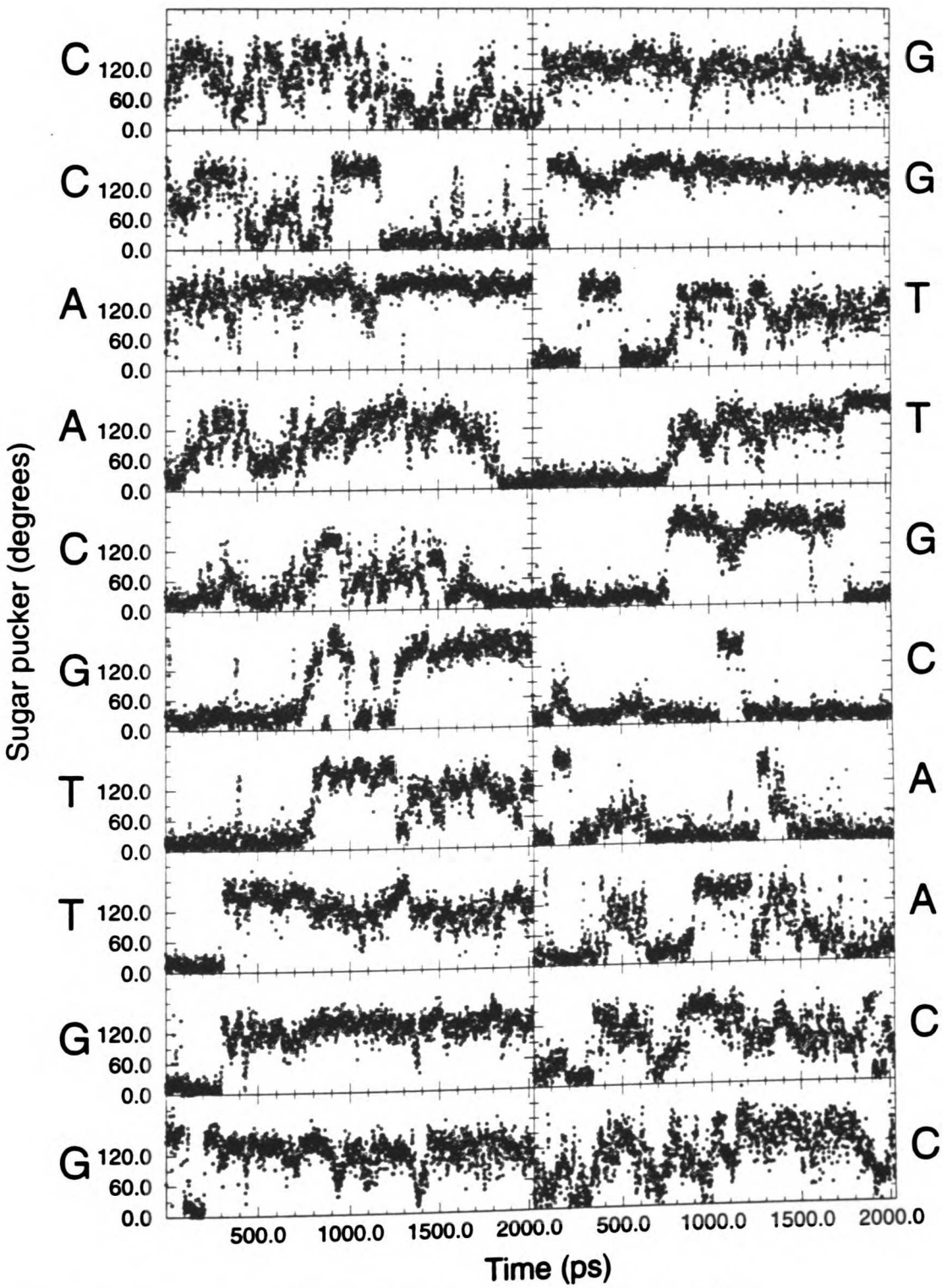


Figure 37: Time course of the individual sugar puckers from the simulation of A–DNA in a mixed water and ethanol solution

A-DNA is generally less hydrated than B-DNA; this is part of the reason why a B-DNA to A-DNA transition is observed experimentally in mixed ethanol/water solutions. In Figure 38 is shown the number of water and ethanol oxygens atoms within 3.4 Å of any DNA atom over the course of the simulation in mixed water and ethanol solution.

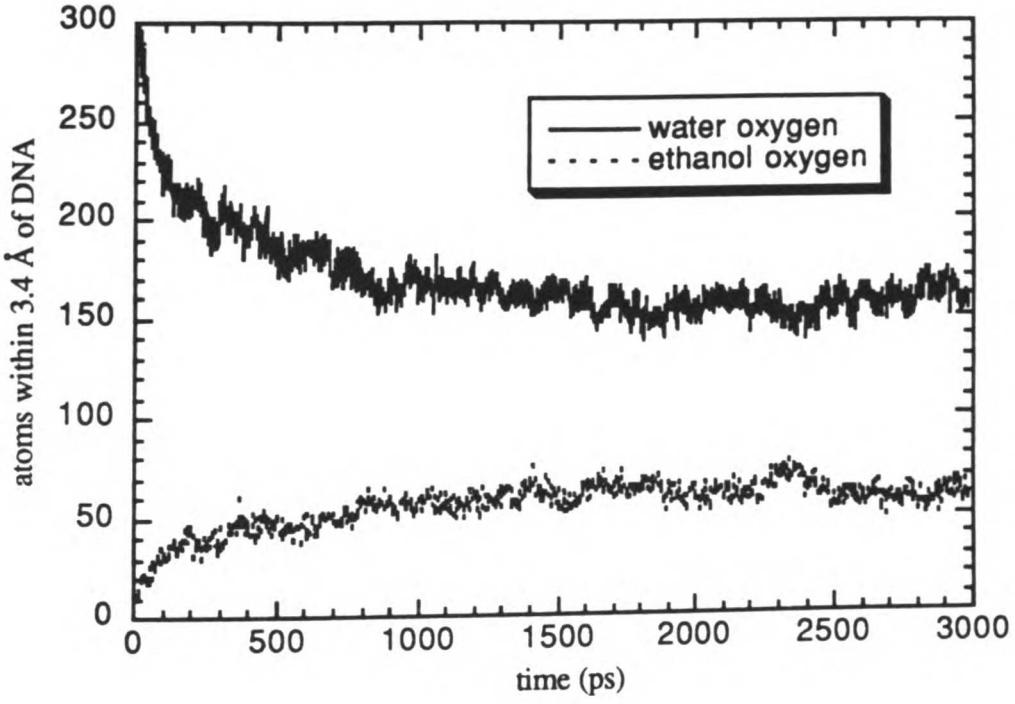


Figure 38: Closely associated oxygen atoms: The graph represents the number of water oxygens (solid line) and ethanol oxygens (dashed line) within 3.4 Å of any DNA atom over the course of the A-DNA simulation in mixed water/ethanol solution.

As the simulation proceeds, water is clearly being "pulled" away from the DNA. By ~2 nanoseconds, ~15 waters per base pair and ~6 ethanols per base pair are associated with the DNA. To better visualize the water and counterion association, iso-contours of the water oxygen, ethanol oxygen and Na^+ density over the final nanosecond of the trajectory is displayed as a stereo view in Figure 39. Despite less waters per base pair, the DNA is still extensively hydrated. Water density is extensive in the major groove with a large body of

closely associated water bridging the two strands in a bend across the major groove. The extensive hydration and counterion association likely overcomes the inter-strand phosphate repulsion and helps stabilize A-DNA. In addition to the extensive hydration transverse to the major groove, spines of hydration are also apparent. The minor groove is also extensively hydrated, however the density appears lower than is seen in corresponding simulations of A-RNA in water. Likewise, the major groove hydration is more extensive than was seen in corresponding A-RNA simulations and the hydration of the phosphate backbone greater than was seen in simulations of B-DNA in water (Cheatham & Kollman, 1996a). Ethanol also appears closely associated with the DNA, however this density appears (from Figure 39) to be most closely associated with the backbone and the terminal part of the minor groove. Less apparent hydration in the minor groove and extensive hydration and ion association in the major groove are consistent with the experimental observation of ions in the major groove of A-DNA. A-DNA is presumably stabilized not only due to the ions and hydration in the major groove, but likely absence of stabilizing counterion association and hydration in the minor groove consistent with the "groove binding model" (Bartenev et al., 1983; Ivanov et al., 1973).

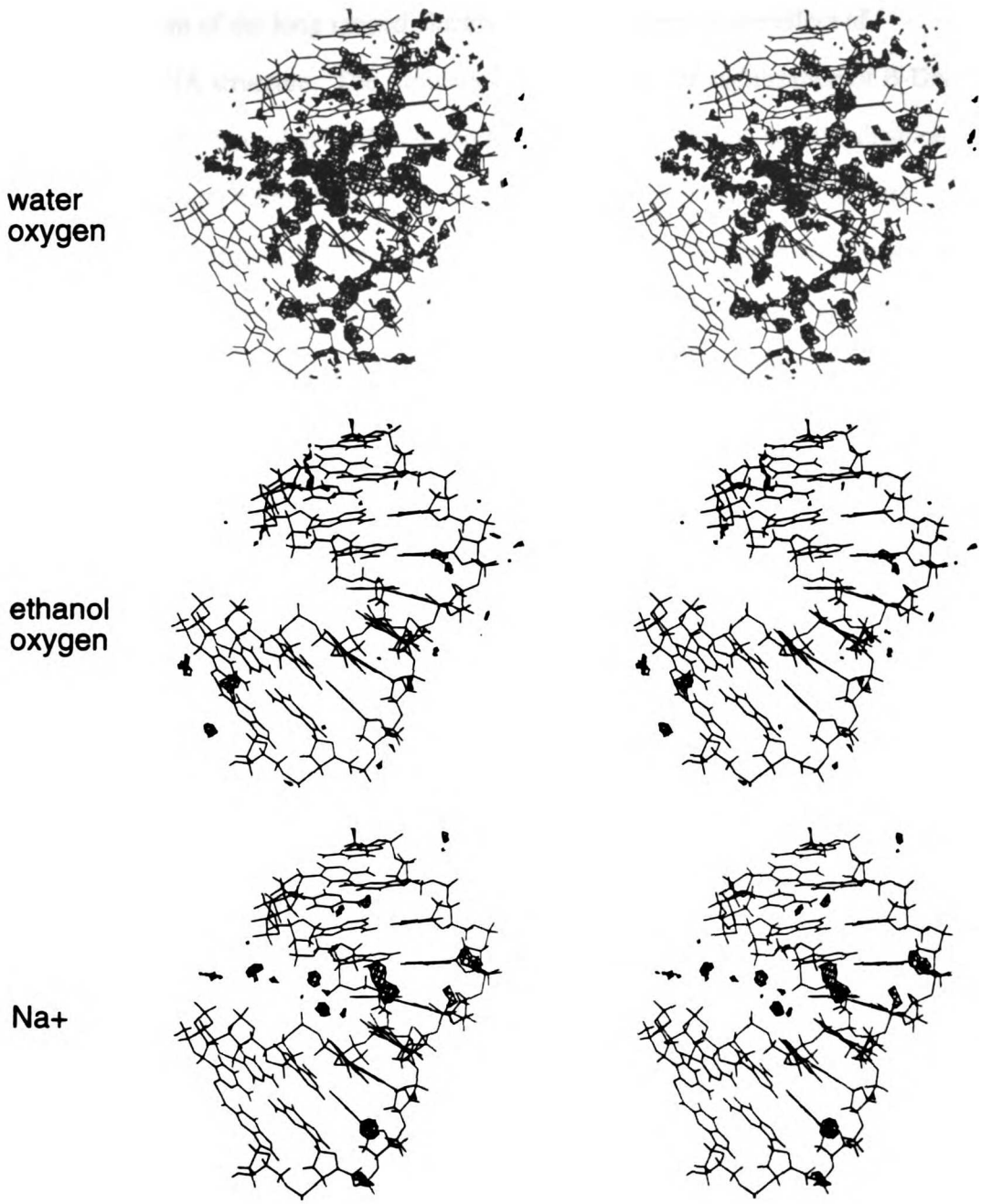


Figure 39: Stereo views of the average structure over 2000-3000 ps along with iso-contours of the water oxygen, ethanol oxygen and sodium ion density. This data is generated (Cheatham et al., 1996a) by RMS fitting all atom of the DNA to the first frame, at 1 ps intervals. Then a 50 angstrom grid is constructed around the DNA and over the trajectory the counts of each particular atom type in each grid element (0.5 angstroms cubed) are saved. This data is contoured at 12.0 hits per grid element or roughly ~3.0 times the bulk water density

These results suggest that molecular dynamics simulations, with a reasonable force field and proper treatment of the long ranged electrostatics, can represent the effect of environment on DNA structure. The simulations demonstrate the stabilization of B-DNA and A-RNA in water, "unstable" A-DNA and B-DNA in pure ethanol, and stabilization of A-DNA in mixed water ethanol solution. Consistent with experiment, these studies suggest that the relative stabilization of, and transition between, A- and B-form geometries involves subtle differences in the specific hydration and counterion association with the nucleic acid. Whereas B-DNA is stabilized by extensive hydration in the minor groove and by hydrated ions in the minor groove anchoring the two strands ($Na^+ < K^+ < Cs^+$), A-DNA is stabilized by major groove hydration and ion association, and also by ion-mediated interhelical bonds across the major groove and between duplexes ($Na^+ > K^+ > Cs^+$) (Piskur & Rupprecht, 1995). In fact, the early CD and Raman spectroscopic studies, as discussed by Potaman *et al.* (Potaman *et al.*, 1980), were criticized since it was difficult to distinguish whether aggregation (representing strong interhelical interactions), the presence of A-DNA, or both lead to the altered spectroscopic signature in solutions with high ethanol concentrations. However, sedimentation studies clearly show that A-DNA can form in the absence of aggregation (Potaman et al., 1980). Extensions of these simulations are currently underway to investigate B-DNA in mixed water and ethanol solutions.

issues in the simulation of nucleic acids

This section discusses briefly some of the problems I had performing all the calculations presented in this thesis. The hope is that someone reading this can take heed and avoid the same mistakes, or at least beware of the issues.

model building nucleic acids

An ideal starting structure for molecular dynamics studies is an experimentally derived structure. Generally these are deposited to the Brookhaven "PDB" (protein databank) depository (Sussman, 1996). When placed in the pdb, the structures will in general have a nomenclature different from what AMBER expects and often the structures lack hydrogen atoms. With care, users can modify the names in the pdb to match AMBER and then use some program to add in the missing hydrogens, such as `protonate` which comes with the standard distribution of AMBER. In the absence of an experimentally derived structure, when simulating nucleic acid duplexes, all hope is not lost since duplexes tend to adopt "canonical" geometries, such as the canonical A-DNA and B-DNA structures found from fiber diffraction (Arnott & Hukins, 1972) or the left-handed Z-DNA structure found crystallographically (Arnott *et al.*, 1980; Wang *et al.*, 1979). There are a variety of programs available for building canonical structures both within the AMBER distribution and available through other sources. One such program worth mentioning is the nucleic acid builder (NAB) which is more like a "language" for building arbitrary nucleic acid structures (Macke & Case, 1996).

To build the canonical A- and B-form structures used in the simulations herein, the program `nucgen` from the AMBER release was used. This builds an arbitrary sequenced DNA and/or RNA strands into a duplex geometry without hydrogen atoms. After fixing any discrepancy in naming conventions, the hydrogens are added using the `edit` module of AMBER. To add in hydrogen atoms, `edit` is guided by the standard database of residues (from `prep`). As long as this database is consistent, hydrogen placement will work as expected. However, due to some glitches in the geometry and hydrogen placement of a few of the residues, most notably cytosine, the hydrogens were initially placed with a poor geometry. During the initial minimization to "fix" up the structure, the H1' atoms on cytosine's flipped to the wrong stereochemistry. This did not come to my

attention until after running dynamics for a significant time period, at which time the behavior of the duplex was noticeably incorrect. This problem recurred due to carelessness multiple times. The lesson learned from this experience is to not trust the model building programs and to very carefully "check" the starting structure to make sure it represents what it should before beginning any very long calculations.

equilibration

The standard means for solvating a solute with AMBER, as discussed previously, is to surround the solute by either a droplet (cap or blob) or a periodic box of water. This is done by first overlaying the solute with enough pre-equilibrated boxes of water to span the range of interest, followed by removing any waters that overlap the solvent or are beyond the given range. This is shown schematically below for a periodic box:

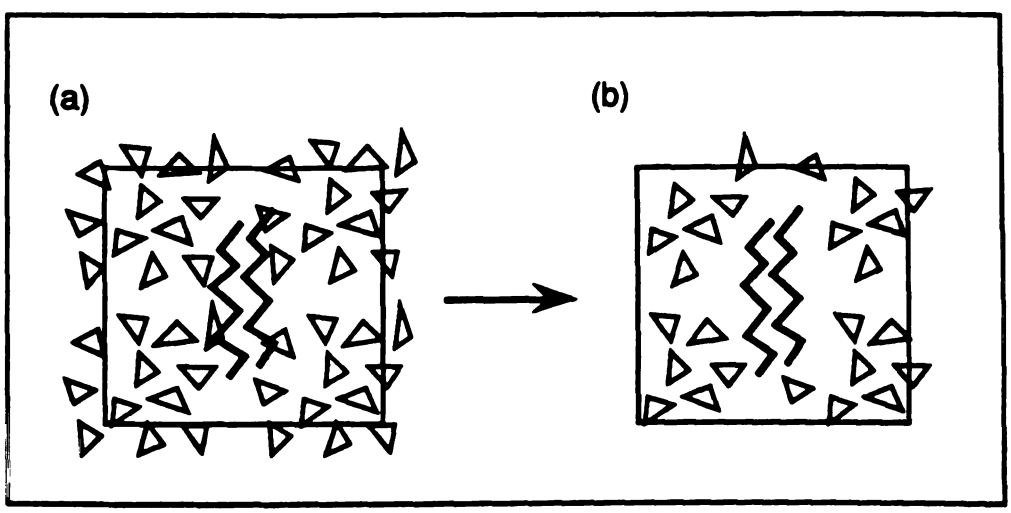


Figure 40: solvating a periodic box

The pre-equilibrated box of water supplied with AMBER is a cube of 216 TIP3P waters generated and equilibrated some time ago from Monte-Carlo simulations. When

surrounding a solute with more than ~33 $Å^3$ of water, more than one pre-equilibrated cube of waters is placed over the solute. Due to anisotropy in the cube of equilibrated waters, when the cubes are placed adjacent to each other (to overlap the solute), some gaps appear between the cubes (as can be seen in Figure 40). Likewise, when waters overlapping the solute are removed, space appears around the solute. Prior to using the newly created periodic box, it is advisable to equilibrate the solvent to let the water relax to properly interact with the solute, fill the gaps, and eventually equilibrate to the expected density (for that particular water model). As discussed previously, minimization will not relax the solvent sufficiently and either a molecular dynamics or Monte-Carlo type procedure are required. If a standard constant volume molecular dynamics (or Monte-Carlo) simulation is run for this purpose, the extra space in the periodic box will tend to collapse and coalesce into a vacuum "bubble", unless the initial volume of the box was chosen such that the box size matches the density. However, achieving this match is rather difficult based on the anisotropy of the box, the typically complex surface of the solute and the "mechanics" of the box building. Therefore it is fairly critical to equilibrate the system with constant pressure simulations which allow the box size to change and therefore collapse the vacuum bubble. In simulations where this is not done, the vacuum bubble has disastrous effects on the solute structure when it "hits" the solute, as was seen in some of my earlier simulations where I first equilibrated with constant volume.

Another issue with equilibration of the solvent is keeping the solute "fixed". Given that likely the solvent is being placed around a reasonable solute structure, it is important to maintain the solute structure while relaxing the less reasonable solvent structure. Although solvent clearly does influence the solute structure, it is important to be able to separate real effects of the solvent from ill-effects of the solvent during the equilibration phase. To this end, in the simulations herein, the solute was held fixed during water minimization (to fix bad contacts) and equilibration. After the initial equilibration, restraints were placed on the

solute atoms and cycles of minimization and dynamics (as discussed in the chapters that follow) were performed where the restraints were gradually reduced. Within AMBER, there are two ways to maintain the solute structure; this can be done either by applying harmonic restraints to the atoms in the solute or by effectively holding the solute fixed by zeroing the forces on the solute atoms (using the `belly` option). Both of these methods, within AMBER, have issues which must be confronted when running constant pressure simulations. When applying restraints, the force constants must be sufficiently high to maintain the structure; in spite of this there is a small drift in the solute atom positions from the reference coordinates. This is likely insignificant. However, during constant pressure simulations, the box size may change (and likely shrink in the current `edit` or `LEaP` solvation protocol which adds in extra space between the pre-equilibrated cubes to avoid bad contacts) and therefore the position of the solute atoms will change with respect to the reference coordinates. To compensate for this, the pressure scaling (or box size changes) are applied to the reference coordinates as well. However, at the end of the simulation, these modified reference coordinates are not saved. Therefore on subsequent dynamics runs using the original reference coordinates with the final coordinates from the previous dynamics steps (the `restrt` file) significant deviations and restraint violations will be apparent since the origins of the solutes are no longer consistent. In many cases, the dynamics will simply terminate because of SHAKE or other failures that result from the system virtually exploding due to high velocities imparted by the high restraint violations. An obvious means to overcome this difficulty is to modify the code to save the "scaled" reference coordinates. This solution has not been placed into the standard release. An alternate method around the problem is to instead of using the old reference coordinates, use the final restrained (and slightly different) coordinates as the new reference coordinates. This however leads to an accumulation of error. For this reason, and based on speed (since all the solute-solute pair interactions need not be calculated), I applied the `belly` method where the solute is held fixed by zeroing the forces on the solute atoms. This

option also has conflicts with constant pressure, specifically, the box is scaled independently of the belly. Within AMBER there are two methods to perform pressure scaling; these are molecule based or atom based scaling. In atom based scaling, the box size is increased or decreased by scaling the relative positions of all the atoms. With the belly option and the solute held fixed, if the box size shrinks, all the atoms in the solute get closer together. This distorts the structure. During standard dynamics this would not be an issue since the intermolecular forces would compensate for the small change in atomic positions (each step) to correct the structure. However, with these forces zeroed, the molecule will shrink. On the other hand, in molecule based scaling, the box size is changed by scaling the relative positions of each molecule's center of mass. This is the obvious method to apply when equilibrating the water around a solute. However, if as in the case of DNA, the solute represents more than one molecule (*i.e.* the two DNA strands in the duplex), the positions of the fixed solute molecules will be shifted with respect to each other. Although the code could be changed to not scale atoms that are fixed with the belly option, this is not practical since the belly is general (*i.e.* can be placed on any atoms including solvent) and omitting scaling on some molecules could lead to locally high or low pressures that could disrupt the pressure scaling process. The alternative method is to, for the purposes of equilibration, treat the separate solute molecules as a single molecule. This is the procedure used in the simulations herein. To perform this requires modification of the parameter/topology file[26]. In addition to modifying the box information, if absurdly

---

[26] At the bottom of the parameter/topology file (`prmtop`), assuming no perturbation is being performed, information necessary for the periodic box is stored. This gives information which specifies how many atoms are in each molecule. The format of this is

- `IPTRES, NSPM, NSPSOL` (format 12I6): The final residue which is part of the solvent, the total number of molecules and th first solvent molecule (respectively)
- `NSP` (format 12I6): The number of atoms in each molecule; a total of `NSPM` entries.
- `BETA, BOX(1), BOX(2), BOX(3)` (format 5E16.8): the angle of the box and the box dimensions, respectively.

To merge the first two molecules into a single molecule, `NSPM` and `NSPSOL` are decremented by one and the first two `NSP` entries added together and merged, then the entire `NSP` array shifted.

high pressures are encountered on the first step, the box size can increased (by editing the `prmtop` file).

For more information about the equilibration protocol used and equilibration issues in general, see the tutorial I developed for the "Methods and Applications of Molecular Mechanics and Dynamics to Molecules of Biological Interest" workshop held August 7-10, 1996 at the Pittsburgh Supercomputing Center and placed on the WWW at "http://www.amber.ucsf.edu/amber/tutorial/".

trajectory analysis

Molecular dynamics simulations provide a tremendous wealth of information in terms of the time sequence of the structures. Typically, during the dynamics snapshots of the instantaneous positions of the molecules are saved at regular intervals by dumping to a file. This *trajectory* is later analyzed by supplemental programs. [Although other programs, such as CHARMM (Brooks *et al.*, 1983), have extensive built-in trajectory analysis capability, AMBER does not.] With the AMBER distribution comes the program `carnal`; this was used to analyze root mean square deviations (RMSd) between the structures (which represents the similarity of best fit structures) and various values, such as dihedral angles, bond lengths and sugar pucker values. Helicoidal values were calculated using Lavery's `curves` program (Lavery & Sklenar, 1988) and versions of this program hooked into more general trajectory analysis software, `dials_and_windows` (Ravishanker *et al.*, 1989). To analyze the trajectories herein, often RMSd plots as a function of time for all atoms to starting structures or average structures were plotted, as were plots of selected dihedrals, sugar puckers, bond lengths and helicoidal parameters as a function of time. In addition to the time course, averages and standard deviations were also calculated. To simplify the task, perl scripts were written to automatically process the output from the analysis programs to write summary files and create graphs. A particularly useful program for plotting is the freely available `xmgr` or ACE/gr 2D graphing program;

this program was used to create most of the 2D graphs presented in this thesis. In addition to investigating the various parameters as a function of time, it is useful to utilize molecular graphics to visualize the structures. To this end, the program `midas` developed by the Computer Graphics Lab at UCSF was used extensively (Ferrin *et al.*, 1988), as was the public domain code `moil-view` developed by Carlos Simmering and Terry Lybrand's `MDdisplay` program for looking at "movies" of the trajectories. Most of the molecular graphics images presented in this thesis were generated by `midas`. Average structures were created from the trajectories, using `carnal`, by best fitting the coordinates to a common reference frame followed by a straight coordinate average. Minimization can be subsequently applied to fix distortions in the structure (such as seen with methyl group hydrogens). This, as discussed in Chapter 3, does not alter the structure or calculated properties too drastically. Helicoidal values are very sensitive to how they are calculated; this is discussed in significantly more detail in Chapter 3.

While I was at UCSF, a new graphical model builder (`LEaP`) was being developed and slated to replace the old style method of creating input files using the series of programs `prep`, `edit`, `link` and `parm`. Inevitably, a few bugs appeared in `LEaP`, such that the parameter/topology files generated were not consistent with the old tried and true methods. After being "burned" a few times and in order to help ferret out the bugs, I started to write a program called `rdparm` to decipher the data in the parameter/topology (or `prmtop`) file. This program allows users to print out all the parameters in a given topology file (*i.e.* bonds, angles, dihedrals, Lennard-Jones, etc). Gradually this was extended to allow simple modifications to the `prmtop` (such as modifying the `prmtop` to remove or add waters, change parameters, add/remove bonds/angles/dihedrals, add constraints, *etc.*). During the development of this program, the issues with imaging in Ewald trajectories became apparent, so I started to add in trajectory analysis capability. With this, I could "fix" the imaging problem. I also added capability to calculate RMSd

values and generate 2D RMS plots (discussed below) and perform very crude clustering based on the RMSd values. Given the "block of ice flying though space" problem (as is discussed at the end of Chapter 1) where molecules would move through the periodic box rather rapidly and the subsequent imaging problems, the simple algorithms for imaging to fix the problem broke down. Therefore, I reevaluated the trajectory analysis capability in `rdparm` and started over generating a new more general and extendible utility for processing the trajectories. This new utility, a part of `rdparm`, is called `newtransform`.

The idea was to develop a utility that was data driven, had flexible input and output formats and was readily extendible. Also, since the time consuming part of analyzing the trajectory is often uncompressing the files and performing the input and output, I wanted the ability to do multiple analyses at once and the ability to handle compressed files. To give a better example, here is a command file for `newtransform`:

```
newtransform
trajin traj1.Z   1  20 1
trajin traj2.Z   1 100 1
trajout traj.0-120ps
center :1-20 mass origin
image origin
rms first :1-20
strip :WAT
go
```

This processes two compressed files which has the coordinates initially centered to the origin, images, rms fits to the first frame using residues 1-20, strips of the water residues and then writes the processed coordinates to a trajectory file. The input file is processed to create a list of input trajectories (`trajin`) and output trajectories (`trajout`) where the trajectories can be in AMBER trajectory, AMBER restart or PDB format. Then everything else in the input file is placed into a list of "actions" which represent routines that are called

to process each particular coordinate frame. The final "go" tells the program to start the transformation. A state representing the `prmtop` (*i.e.* atom names, *etc.*) is initially created (which may be modified by the actions, such as after stripping waters) then each coordinate file is first opened and checked. After the initial check, each trajectory (in the order specified) is re-opened and the coordinates read in frame by frame. For each frame, each "action" is called in the order specified and the transformed coordinates from the current action get sent to the next action. If writing a processed trajectory (or a series of PDB files), after the final action the current transformed snapshot is written out. The "action" routines take a common form and can associate data which is later written out (such as RMSd values over time, *etc.*) after all the snapshots have been processed. The flow can be explained more easily in the following diagram:

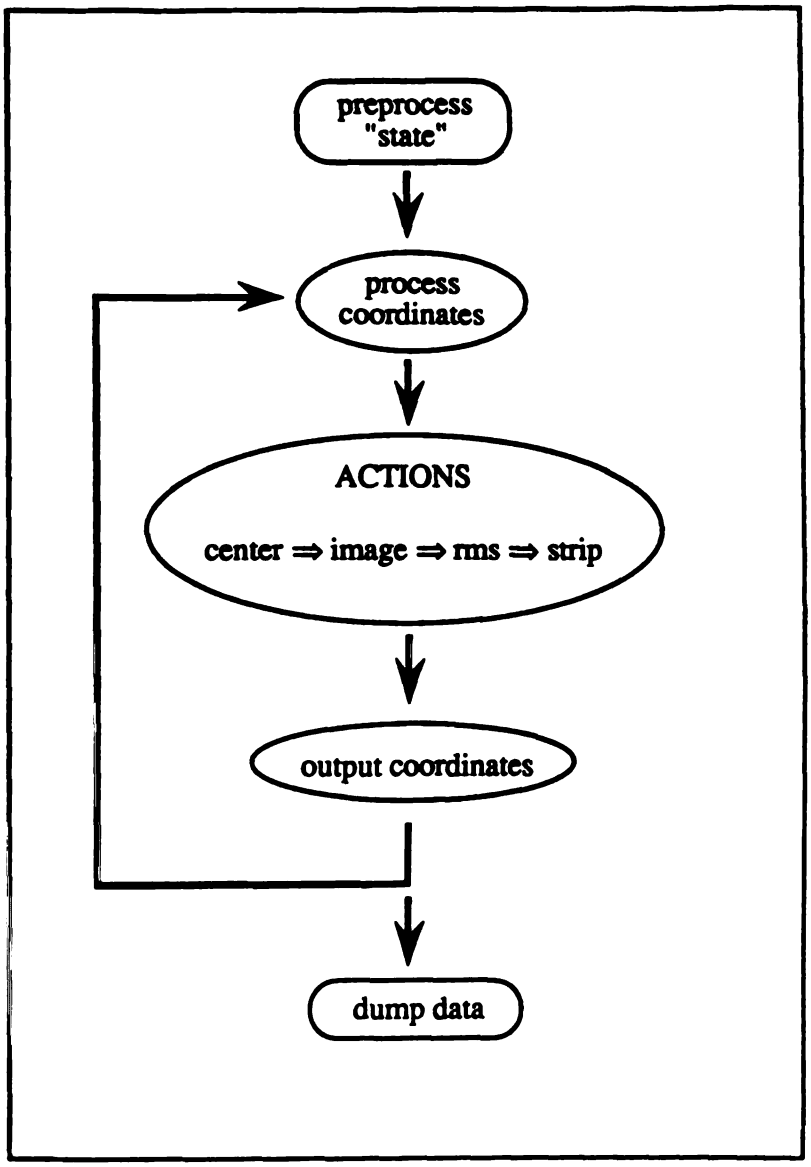


Figure 41: code flow in `newtransform`

The basic idea is to make the transformation procedure versatile and easily extendible. For more information, see the code which is written entirely in C. The next sections discuss some of the functionality added into rdparm and used to analyze the trajectories. All of it is rather straightforward and has been performed previously.

1D and 2D RMSd plots

Often in the presentation of results from molecular dynamics simulations, the "goodness" of the trajectory is often equated with how close the structure stays to the starting geometry, or how low the RMSd value is. This is not necessarily the best measure of how good a trajectory is for a number of reasons. A low RMSd value could mean that the force field representation is good, alternatively it may suggest that the simulation never went anywhere due to conformational sampling difficulties (high pressure, crystal packing). Likewise, a high RMSd value does not necessarily imply a significant structural difference since small changes in structure can lead to high RMSd values. However, on some level the RMSd values can give some indication of the progress during the trajectory, the stability, and an indication of the fluctuations. Presented in Figure 42 is the RMSd values versus time representing the A-DNA to B-DNA transition presented in Chapter 2.

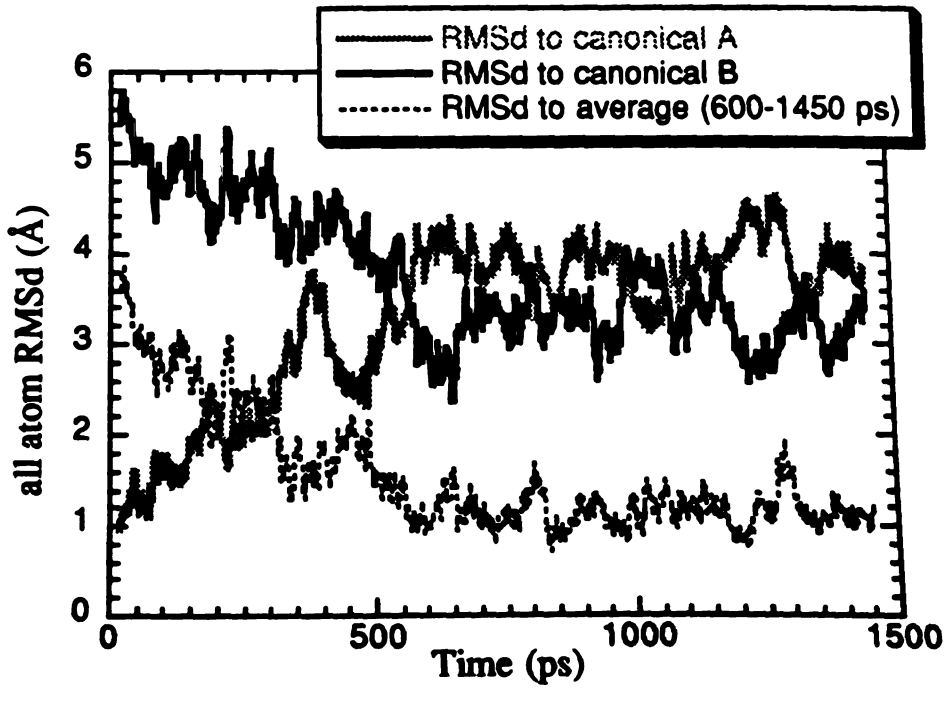


Figure 42: RMSd versus time for the A-DNA to B-DNA transition

From this plot, it is fairly clear that the structure is moving from a canonical A structure to be more B-DNA like. Moreover, the structure converges to an average structure and during the dynamics the instantaneous structures after 500 ps are within ~1.5 Å of this average structure. The fluctuations in the RMSd show that the structure is mobile, however it is not clear from this plot how these structures interrelate.

To get a better idea about how the structures are changing over time and how they all interrelate, a 2D RMS plot is very useful. In this plot, increasing time is represented on both the x- and y- axis and each point (x,y) represents the RMSd value between structures at time x and time y. This means of displaying the data highlights the similarity of structures and implicity clusters the trajectory. Low RMSd values (represented by whiter shades) will be easily seen. If the structural changes are diffuse or the structure remains near the starting geometry, little information can be gained from the plot. However, when distinct structural changes are observed, the plot can be quite informative. A 2D RMSd plot is presented in the discussion by McConnell *et al.* (1994) to display the conformational substates of DNA visited during a trajectory. A similar plot is shown in Figure 43. This 2D RMS plot represents the RMSd of the nucleic acid base atoms over 95 picoseconds taken from the later part (1775-1870 ps)[27] of a molecular dynamics trajectory (at 1 ps intervals) of $d(A_{10})$-$d(T_{10})$. The white represents an RMS of zero (such as along the diagonal axis where the structures are equivalent) and the black represents the maximum RMSd between the structures. Two whitish squares are visible, one in the lower corner and one in the upper corner. The darker blocks show that these two regions are dissimilar and represent slightly different conformations, *i.e.* approximately one third of the way

[27] The trajectory of $d(A_{10})$-$d(T_{10})$ was run starting in a canonical A geometry utilizing the particle mesh Ewald method and simulation conditions equivalent to those reported in Chapter 2. During the trajectory, the A-DNA structure converted to a B-DNA.

through this trajectory, the bases appear to move into a different conformation or "substate".

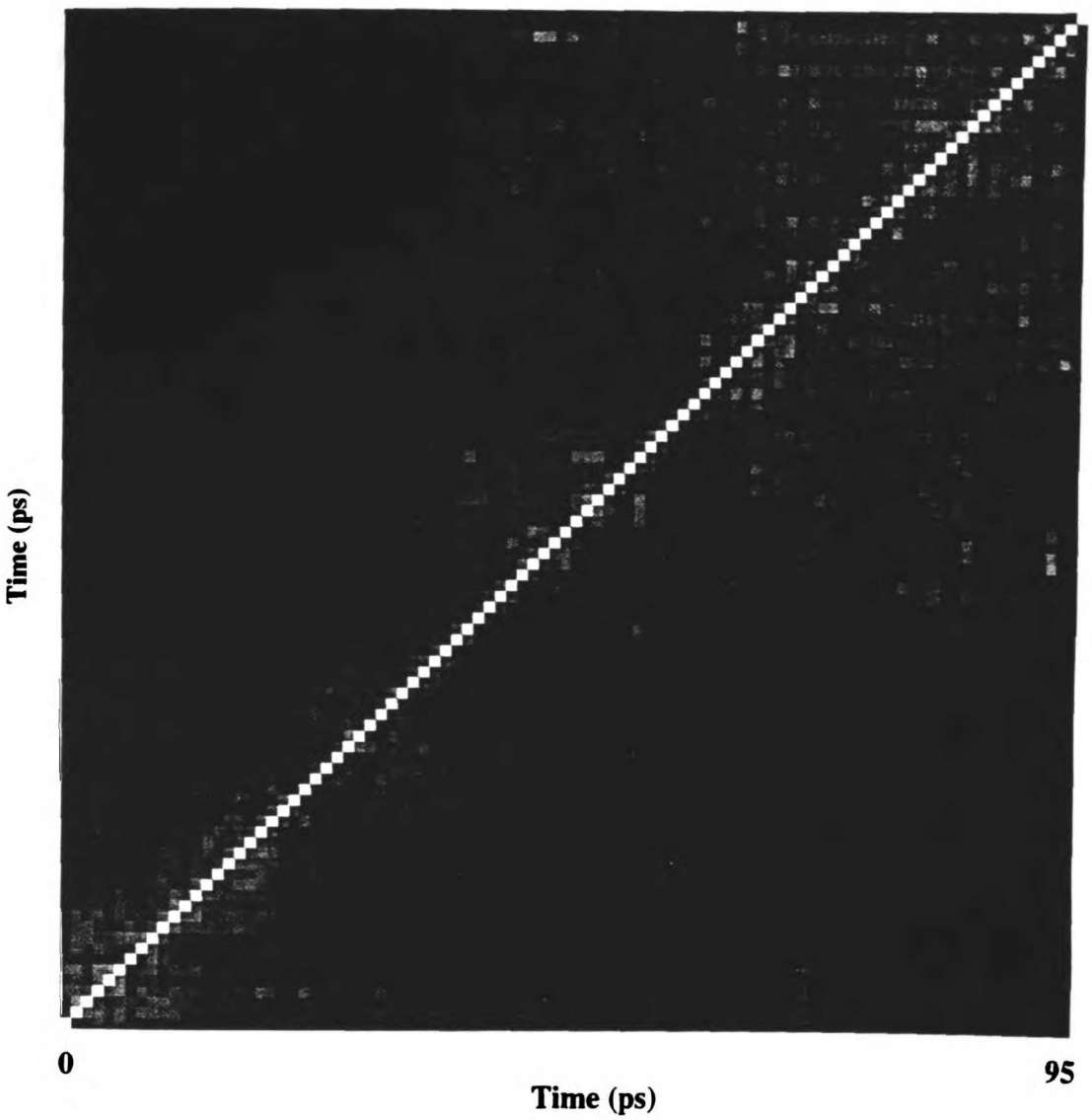


Figure 43: 2D RMSd plot

diffusion

There are two common ways to calculate diffusion constants (D) in molecular dynamics simulations. Either they are calculated directly using the Einstein relationship and the average mean square displacements or from the velocity autocorrelation function. The velocity autocorrelation function (in terms of the center of mass velocity at time t, $v_i(t)$) is defined as follows:

$$D = \frac{1}{3}\int_0^{\infty} dt \langle v_i(t) \circ v_i(t_0) \rangle$$

Although the diffusion constant is more accurately calculated using the velocity autocorrelation, the diffusion constants were calculated using the mean square displacements (from coordinates r at various times t during the simulation):

$$2D = \lim_{t \to \infty} \frac{\langle |r(t) - r(t_0)|^2 \rangle}{3(t - t_0)}$$

This is necessary since often the velocities are not archived along with the coordinates in order to save space and the velocity autocorrelation function is not calculated "on the fly" in `sander`. Despite less accuracy, the results calculated by both methods are comparable.

Calculating the diffusion constant from the mean squared displacements is complicated by the need to keep track of imaging and coordinate re-centering (which will mess up the average displacement). The `diffusion` command was added into the `newtransform` code to calculate the mean squared displacements from a trajectory, correctly taking into account imaging. This prints out data representing the mean squared displacements (Å^2) as a function of time (ps) which when plotted will give a slope proportional to the diffusion constant. The units typically reported in molecular simulations

are 10^{-9} m^2/s or 10^{-5} cm^2/s. To convert the slope ($Å^2/ps$) to the proper units (taking into account the three degrees of freedom represented by the three in the denominators above) it is multiplied by 10.0/6.0.

hydration

Molecular dynamics simulations in explicit water contain much information about the hydration and average water properties. Often presentation of this data is very tricky. On one level, all of the hydrogen bonds to water can be counted and specified and the data presented as a giant table. However, reading and understanding this table is often very difficult. In a similar manner, in the text each particularly interesting water interaction can be mentioned and described; however this also gets tedious and the data is hard to follow. Therefore it is desirable to simplify the data somehow by looking at average properties such as the counts of water molecules within a given range of a particular hydration site or by visualizing the average water density. Variations on both methods were implemented into the `newtransform` trajectory analysis code and are discussed below along with a few related methods from the published literature.

radial distribution functions, proximity, "counting" waters

The classic definition of a pair distribution function, as presented in Allen & Tildesley (1987), is "the probability of finding a pair of atoms a distance r apart, relative to the probability expected for a completely random distribution at the same density." This can be represented discretely as follows:

$$g(r) = g\left(x^* \Delta r + \frac{1}{2}\Delta r\right)_{x=0,1,\ldots} = \frac{histogram_{normalized}[x]}{histogram_{ideal}[x]}$$

For a given bin size (Δr) we compare as a function of distance r (or over bins, x) the normalized count of atoms ($histogram_{normalized}$) at that distance ($x^*\Delta r$) to what is expected ideally ($histogram_{ideal}$). This is often referred to as a radial distribution function (RDF) and

is calculated for various atoms in a molecule, such as the oxygen-oxygen and hydrogen-hydrogen RDF's for TIP3P water shown in Figure 44.

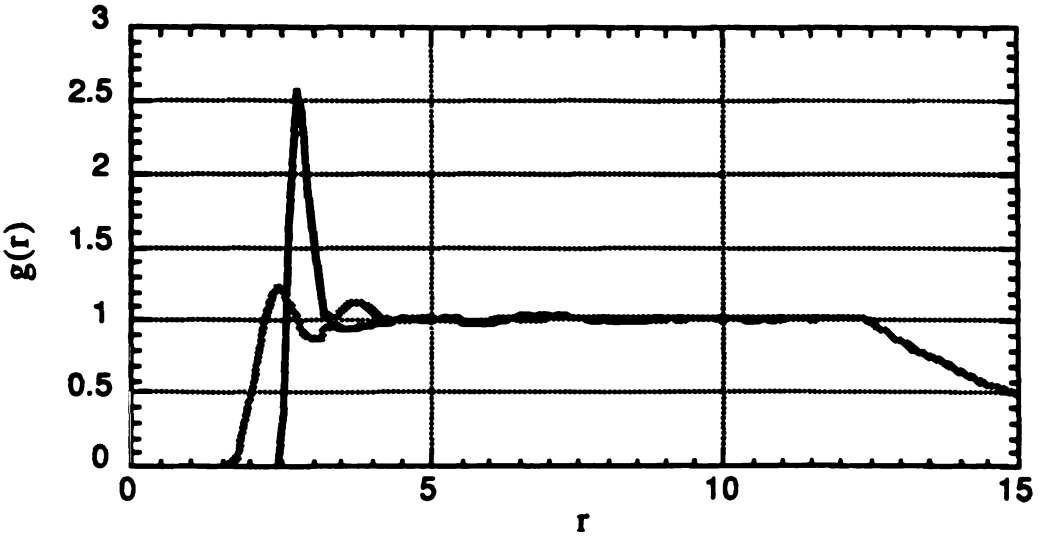


Figure 44: water-water radial distribution functions. In black is presented the oxygen-oxygen RDF and in gray the hydrogen-hydrogen (for the H1 atoms) RDF from a 100 ps simulation of 500 TIP3P waters in a periodic box.

The RDF above for oxygen (in black) shows a peak at ~3.5 Å which is ~2.6 times the expected density and which represents the first hydration sphere around a water. If the curve is radially integrated (*i.e.* it is necessary to multiply the value at each r by the radial volume of the subshell) to 3.5 Å, a water occupancy of ~5.2 waters is found in this first hydration sphere. After the first peak there is a slight dip (representing the excluded volume of the oxygens in the first hydration sphere) followed by a second peak (representing the second hydration sphere). After this, the RDF gradually flattens out implying that "bulk" density of the water is observed in this range. After 12.5 Å (in the figure above), the RDF values fall. This is an artifact of the periodic box size (~25 Å) since "image" waters are not included in the construction of the RDF.

The RDF can be measured experimentally in diffraction experiments (Narten & Levy, 1971; Thiessen & Narten, 1982). When compared to experiment, TIP3P has a first peak in the RDF that is too large (compared to the experimentally determined occupancy of 5.0 at 3.5 Å) and too little structure beyond the first peak (representing the second and third hydration spheres) (Jorgensen *et al.*, 1983).

To calculate the RDF, all the distances between pairs of atoms are binned. Then this histogram needs to be normalized by the number of frames in the trajectory and the number of "molecules". The number of "molecules" depends on the nature of the RDF:

- *solute to solute*: the radial distribution function represents the distribution of the solute to itself. This is a trivial RDF and the number of molecules needed to normalize is one.
- *solvent to solute*: the radial distribution function represents the distribution of solvent molecules around the solute molecule. This is a standard RDF used to investigate the number of waters around a given hydration site such as an ion; the number of molecules for the normalization is the number of solvent molecules.
- *solvent to solvent*: the radial distribution function represents the distribution of each solvent molecule around every other solvent molecule. This is the RDF used to investigate bulk water properties; the number of molecules for the normalization is $N_{solvent} * (N_{solvent} - 1) / 2$, where $N_{solvent}$ is the number of solvent molecules.

To calculate the "bulk" or expected number of molecules at a given distance we simply multiply the density (or inverse volume) by the volume of the radial subshell. In practice, two methods are commonly used to calculate the volume of the subshell; these are the approximate method where $dV=4\pi r^2\Delta r$ or the exact method where $dV=(4/3)\pi[(r+\Delta r)^3 - r^3]$. When implementing code to perform the RDF, the accuracy of the approximate method was checked.

spacing (A)	error (%)
1.0	9.8
0.5	4.9
0.1	1.0
0.05	0.5

Table 12: approximate versus *exact radial volume integration*

The overall integration error obtained by using the approximate method is rather high (see Table 12) unless the bin size is rather small. The error in the value of each radial volume element (dV) is presented in Figure 45. From this analysis, it was decided to use the exact method in the calculations of the RDF.

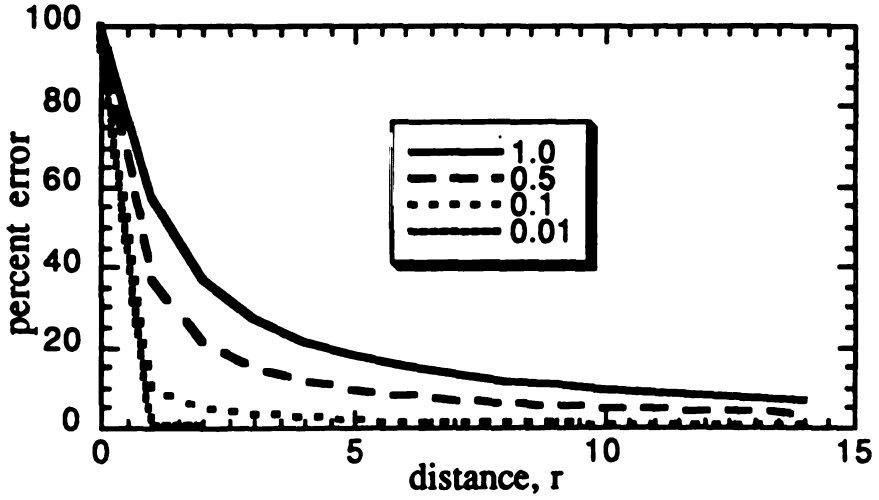


Figure 45: radial volume element integration errors. Using approximate methods ($4\pi r^2\Delta r$) to calculate the volume of a radial subshell leads to significant errors. Plotted as a function of distance is the percent error for various Δr values.

As with the calculation of diffusion constants, care needs to be taken to image the distances and note that distances beyond one box length are not considered (unless the normalization is changed to include the count of the "image" molecules). At large distance, the value of the RDF should go to unity. This is a good test of any implementation.

One problem with the RDF is how to handle the excluded volume of the solute when the solute is large, such as is the case with DNA and proteins. In this case, the bulk density will be altered. For example, consider a central phosphate group in a long helix compared to an isolated ion. The count of waters around the phosphate will be lower than around the ion since waters can only approach the phosphate on one side. While this is expected, the whole radial distribution function will be offset if the density is used to normalize it. Therefore, when used in analysis, rather than using the density reported in the simulation, the inverse volume (which represents the idealized density) is used to normalize since this will allow integration of the RDF to give reasonable values. It should also be mentioned that when using large non-spherical solute molecules, the RDF should only be calculated with respect to individuals groups (and can be subsequently averaged) since otherwise the RDF will otherwise not be very informative. To analyze nucleic acid simulations Pettitt's group has opted to use "cylindrical" distribution functions which calculate the pair distribution with respect to a global helical axis (Mohan *et al.*, 1993; Weerasinghe *et al.*, 1995b); this however has the drawback that occupancies can not be accurately calculated by direct integration.

An additional issue is how to determine which waters are associated with a specific site when two sites are close to each other, such as adjacent phosphates in a DNA duplex. To overcome this difficulty, Mezei and Beveridge (1986) use a proximity criterion and quasi-component distribution functions to break up an RDF into primary contributions from different solute environments as well as representing the primary, secondary, *etc.* contributions from the environment. While this can decompose the hydration, there is some subjectivity in terms of picking the boundaries of the hydration shells. Moreover, it is not clear if the more complex procedure gives any more insight into the hydration of nucleic acids. Better understanding can be obtained by visualizing the specific hydration.

"movies" and grid occupancies

The radial distribution function gives an idea of how much water surrounds a given hydration site, but does not give an indication of specific structure of the water. If only a limited number of snapshots from a trajectory are to be analyzed, a simple way to gather this information is by overlapping the solute and displaying close waters with molecular graphics. This was done in the early Monte Carlo simulations with fixed solutes (Subramanian & Beveridge, 1989; Subramanian *et al.*, 1990a; Subramanian *et al.*, 1988; Subramanian *et al.*, 1990b) and later in molecular dynamics simulations (Chuprina *et al.*, 1991). However, where there are many snapshots, this kind of picture becomes cluttered. To reduce the clutter, only the most favorable hydration waters can be displayed. To support this, code was added to `newtransform` to bin atom positions in a trajectory.

The binning procedure works by constructing a grid centered on the molecule of interest and, as the trajectory is processed, incrementing counters in each grid element if the atom of interest (*i.e.* a water oxygen) is within that particular grid element. To make the binning of atom positions meaningful, the molecule of interest should remain in a common reference frame. Therefore, the molecules of interest are RMS fit to a common reference frame and imaged prior to constructing the grid. After processing the trajectory, this grid can be contoured.

The `grid` command to `newtransform` creates a grid in XPLOR density format which can be visualized using the density delegate to MidasPlus (Ferrin *et al.*, 1988). The grids are usually created with 0.5 Å or 1.0 Å spacing and the counter for each grid element is only updated if the center of the molecule (or atom) is within that cell. Assuming a bulk water density of 1 g/mL and converting the units, this works out to ~0.03346 molecules/Å^3 or equivalently 1 water per ~29.9 Å^3. With a 0.5 Å grid spacing, the expected number of waters per box is ~0.00418 molecules per grid element. To determine the expected count in each grid element, this number is multiplied by the number of frames visited.

The gridding procedure was used extensively to analyze the hydration of DNA as is presented and discussed in Chapter 3 and Chapter 4. This type of analysis is nothing new and is similar to the reported analysis of crystal structures by the Berman (Schneider *et al.*, 1993) and Goodfellow (Umrania *et al.*, 1995) groups, the characterization of pharmacophores (Rosenfield *et al.*, 1984) and the analysis of counterion density around DNA (Laughton *et al.*, 1995). The density is not a true electron density such as is represented in the XPLOR density file since the data has not been Fourier transformed; instead the raw occupancies are binned and contoured.

A drawback of the method relates to the flexibility of the solute. If the solute is highly mobile and the grid rather large, then mobile parts of the solute (such as the backbone) will effectively reduce the water counts in the static grid despite the common reference frame. In other words, a water in close proximity to a terminal phosphate group may appear in the grid as a low occupancy water if the terminal phosphate is moving around even though the water is tightly associated to the moving phosphate group. To eliminate this, smaller grids centered on the mobile parts can be created.

bibliography

Alden, C. J. & Kim, S.-H. (1979). Solvent-accessible surfaces of nucleic acids. *J. Mol. Biol.* **132**, 411-434.

Allan, B. W. & Reich, N. O. (1996). Targeted base stacking disruption by the *Eco*R1 DNA methyltransferase. *Biochem.* **35**, 14757-14762.

Allen, M. P. & Tildesley, D. J. (1987). *Computer simulation of liquids*, Oxford University Press, Oxford.

Altona, C. & Sundaralingam, M. (1972). Conformational analysis of the sugar ring in nucleosides and nucleotides. A new description using the concept of pseudorotation. *J. Amer. Chem. Soc.* **94**(23), 8205-8212.

Anderson, P. & Bauer, W. (1978) Supercoiling in closed circular DNA: Dependence upon ion type and concentration. *Biochem.* **17**, 594-601.

Ansari, A. Z., Bradner, J. E. & O'Halloran, T. V. (1995). DNA-bend modulation in a repressor-to-activator switching mechanism. *Nature* **374**(6520), 371-375.

Arnott, S., Chandrasekaran, R., Birdsall, D. L., Leslie, A. G. & Ratliff, R. L. (1980). Left-handed DNA helices. *Nature* **283**(5749), 743-5.

Arnott, S. & Hukins, D. W. (1972). Optimised parameters for A-DNA and B-DNA. *Biochem. Biophys. Res. Comm.* **47**(6), 1504-9.

Auffinger, P., Louise-May, S. & Westhof, E. (1995). Multiple molecular dynamics simulations of the anticodon loop of $tRNA^{Asp}$ in aqueous solution with counterions. *J. Amer. Chem. Soc.* **117**, 6720-6726.

Babcock, M. S. & Olson, W. K. (1994). The effect of mathematics and coordinate system on comparability and "dependencies" of nucleic acid structure parameters. *J. Mol. Biol.* **237**(1), 98-124.

Bader, J. S. & Chandler, D. (1992). Computer simulation study of the mean forces between ferrous and ferric ions in water. *J. Phys. Chem.* **96**, 6423-6427.

Barkley, M. D. & Zimm, B. H. (1979). Theory of twisting and bending of chain macromolecules; analysis of the fluorescence depolarization of DNA. *J. Chem. Phys.* **70**, 2991-3007.

Bartenev, V. N., Golovanov, E. I., Kapitonova, K. A., Mokulskii, M. A., Volkova, L. I. & Skuratovskii, I. Y. (1983). Structure of the B DNA cationic shell as revealed by an x-ray diffraction study of CsDNA. Sequence specific cationic stabilization of B form DNA. *J. Mol. Biol.* **169**, 217-234.

Bashford, D. & Karplus, M. (1990). pKa's of ionizable groups in proteins: Atomic detail from a continuum electrostatic model. *Biochem.* **29**, 10219-10225.

Battiste, J. L., Hongyuan, M, Rao, N. S., Tan, R., Muhandiram, D. R., Kay, L. E., Frankel, A. D. & Williamson, J. R. (1996). α-helix-RNA major groove recognition in an HIV-1 Rev peptide-RRE DNA complex. *Science* 273, 1547-1551.

Bayly, C. I., Cieplak, P., Cornell, W. D. & Kollman, P. A. (1993). A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges- the RESP model. *J. Phys. Chem.* **97**(40), 10269-10280.

Beglov, D. & Roux, B. (1994). Finite representation of in infinite bulk system-- solvent boundary potential for computer simulations. *J. Chem. Phys.* **100**, 9050-9063.

Beglov, D. & Roux, B. (1995). Dominant solvation effects from the primary shell of hydration-- approximations for molecular dynamics simulations. *Biopol.* **35**, 171-178.

Berendsen, H. J. C., Grigera, J. R. & Straatsma, T. P. (1987). *J. Phys. Chem.* **91**, 6269-6274.

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *J. Comp. Phys.* **81**, 3684-3690.

Berger, J. M. & Wang, J. C. (1996). Recent developments in DNA topoisomerase II structure and mechanism. *Cur. Op. Struct. Biol.* **6**, 84-90.

Berkowitz, M. L. & McCammon, J. A. (1982). Molecular dynamics with stochastic boundary conditions. *Chem. Phys. Lett.* **90**, 215-217.

Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S. H., Srinivasan, A. R. & Schneider, B. (1992). The nucleic acid database- A comprehensive relational database of 3-dimensional structures of nucleic acids. *Biophys. J.* **63**(3), 751-759.

Berndt, K. D., Guntert, P., Orbons, L. P. & Wuthrich, K. (1992). Determination of a high-quality nuclear magnetic resonance solution structure of the bovine pancreatic trypsin inhibitor and comparison with three crystal structures. *J. Mol. Biol.* **227**(3), 757-75.

Beveridge, D. L. & Ravishanker, G. (1994). Molecular dynamics studies of DNA. *Curr. Opin. Struct. Biol.* **4**, 246-255.

Beveridge, D. L., Swaminathan, S., Ravishanker, G., Withka, J., Srinivasan, J., Prevost, C., Louise-May, S., DiCapua, F. M. & Bolton, P. H., Eds. (1991). Methodological considerations on molecular dynamics simulations of DNA oligonucleotides. *Advances in Biomolecular Simulation*; AIP conference proceedings **239**.

Beveridge, D. L., Swaminathan, S., Ravishanker, G., Withka, J. M., Srinivasan, J., Prevost, C., Louise-May, S., Langley, D. R., DiCapua, F. M. & Bolton, P. H. (1993). Molecular dynamics simulations on the hydration, structure and motions of DNA oligomers. In *Water and Biological Molecules* (Westhof, E., ed.), pp. 165-225. Macmillan Press.

Biesiadecki, J. J. & Skeel, R. D. (1993). Dangers of multiple time step methods. *J. Comp. Phys.* **109**(2), 318-328.

Board, J.A., Jr., Causey, J.W., Leathrum, J.F.,Jr., Windemuth, A. & Schulten, K. (1992). Accelerated molecular dynamics simulation with the parallel fast multipole algorithm. *Chem. Phys. Lett.* **198, 89-94.**

Boczko, E. M. & Brooks, C. L. I. (1993). Constant-temperature free energy surfaces for physical and chemical processes. *J. Phys. Chem.* **97**(17), 4509-4513.

Boczko, E. M. & Brooks, C. L. I. (1995). First-principles calculation of the folding free energy of a three-helix bundle protein. *Science* **269**(5222), 393-396.

Bokma, J. T., Johnson, W. C. J. & Blok, J. (1987). CD of the Li-salt of DNA in ethanol/water mixtures: Evidence for the B- to C-form transition in solution. *Biopolymers* **26**, 893-909.

Born, M. (1920). Volumen der Hydratationswärme der Ione. *Z. Phys. Chem.* **1**, 45-48.

Born, M. & Von Karman, T. (1912). Über schwingungen in Raumgittern. *Physik. Z.* **13**, 297-309.

Brahms, S., Fritsch, V., Brahms, J.G. & Westhof, E.(1992). Investigations on the dynamic structures of adenine- and thymine- containing DNA. *J. Mol. Biol.* **223**, 455-476

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D., J, Swaminathan, S. & Karplus, M. (1983). CHARMm: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.* **4**, 187-217.

Brooks, C. L. I., Brunger, A. & Karplus, M. (1985). Active site dynamics in protein molecules: a stochastic boundary-molecular dynamics approach. *Biopol.* **24**, 843-865.

Brooks, C. L. I., Karplus, M. & Pettitt, B. M. (1988). *Proteins. A theoretical perspective of dynamics, structure, and thermodynamics.* Advances in Chemical Physics (Prigogine, I. & Rice, S. A., Eds.), 71, John Wiley & Sons, New York.

Broom, H. D., Schweizer, M. P. & Ts'o, P. O. P. (1967). Interaction and association of bases and nucleosides in solutions. V. Studies of the association of purine nucleosides by vapor pressure osmometry and by proton magnetic resonance. *J. Amer. Chem. Soc.* **89**, 3612-3622.

Buckin, V. A., Kankiya, B.I., Rentzeperis, D. & Marky, L .A. (1994). Mg^{2+} recognizes the sequence of DNA through its hydration shell. *J. Amer. Chem. Soc.* **116**, 9423-9429.

Burckhardt, G., Walter, A. & Zimmer, C. (1996). Reversal of the Z- to B- conformation of poly(dA-dT)-poly(dA-dT) induced by netropsin and distamycin A. *J. Biomol. Struct. Dyn.* **13**(4), 671-676.

Calladine, C. R. (1982). Mechanics of sequence-dependent stacking of bases in B-DNA. *J. Mol. Biol.* **161**(2), 343-52.

Cheatham, T. E., III & Kollman, P. A. (1996a). Molecular dynamics simulations can reasonably represent the structural differences in DNA:DNA, RNA:RNA and DNA:RNA hybrid duplexes. *J. Amer. Chem. Soc.*, [accepted, pending revision].

Cheatham, T. E., III & Kollman, P. A. (1996b). Observation of the A-DNA to B-DNA transition during unrestrained molecular dynamics in aqueous solution. *J. Mol. Biol.* **259**(3), 434-44.

Cheatham, T. E. III., Crowley, M., Duan, Y., Rosenberg, J. M. & Kollman, P. A. (1996c). [work in progress].

Cheatham, T. E., III, Miller, J. L., Fox, T., Darden, T. A. & Kollman, P. A. (1995). Molecular dynamics simulations on solvated biomolecular systems - The particle mesh Ewald method leads to stable trajectories of DNA, RNA and proteins. *J. Amer. Chem. Soc.* **117**(14), 4193-4194.

Cheng, Y.-K. & Pettitt, B. M. (1992). Hoogsteen versus reverse-Hoogsteen base pairing-DNA triplex helices. *J. Amer. Chem. Soc.* **114**(12), 4465-4474.

Cheng, Y.-K. & Pettitt, B. M. (1995). Solvent effects on model d(CG-G)$_7$ and d(TA-T)$_7$ DNA triplex helices. *Biopol.* **35**, 457-473.

Cheong, C., Varani, G. & Tinoco, I. J. (1990). Solution structure of an unusually stable RNA hairpin, 5'GGAC(UUCG)GUCC. *Nature* **346**(6285), 680-682.

Chiche, L., Gregoret, L. M., Cohen, F. E. & Kollman, P. A. (1990). Protein model structure evaluation using the solvation free energy of folding. *Proc. Natl. Acad. Sci.* **87**, 3240-3243.

Chou, S. H., Flynn, P. & Reid, B. (1989). High-resolution NMR study of a synthetic DNA-RNA hybrid dodecamer containing the consensus pribnow promoter sequence: d(CGTTATAATGCG)-r(CGCAUUAUAACG). *Biochem.* **28**(6), 2435-43.

Chuprina, V. P., Heinemann, U., Nurislamov, A. A., Zielenkiewicz, P., Dickerson, R. E. & Saenger, W. (1991). Molecular dynamics simulation of the hydration shell of a B-DNA decamer reveals two main types of minor-groove hydration depending on groove width. *Proc. Natl. Acad. Sci.* **88**, 593-597.

Cieplak, P., Cheatham, T. E., III & Kollman, P. A. (1996). Controlling the B to A conformational transition in deoxyoligonucleotides. *J. Amer. Chem. Soc.* [submitted].

Cieplak, P., Cornell, W. D., Bayly, C. & Kollman, P. A. (1995). Application of the multimolecule and multiconformational RESP methodology to biopolymers- charge derivation for DNA, RNA, and proteins. *J. Comp. Chem.* **16**(11), 1357-1377.

Clementi, E. & Corongiu, G. (1981). Solvation of DNA at 300K: Counter-ion structure, base pair sequence recognition and conformational transitions. A computer experiment. In *Biomolecular Stereodynamics* (Sarma, R. H., ed.), Vol. 1, pp. 209-259. Adenine Press, New York.

Cognet, J. A. H., Boulard, Y. & Fazakerley, G. V. (1995). Helical parameters, fluctuations, alternative hydrogen bonding, and bending in oligonucleotides containing a mismatched base-pair by NOESY distance restrained and distance free molecular dynamics. *J. Mol. Biol.* **246**, 209-226.

Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. & Kollman, P. A. (1995). A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Amer. Chem. Soc.* **117**(19), 5179-5197.

Daggett, V., Kollman, P. A. & Kuntz, I. D. (1991). Molecular dynamics simulations of small peptides: Dependence on dielectric model and pH. *Biopol.* **31**, 285-304.

Daggett, V. & Levitt, M. (1993). Realistic simulations of native-protein dynamics in solution and beyond. *Annu. Rev. Biophys. Biomol. Struct.* **22**, 353-80.

Daniels, G. A. & Lieber, M. R. (1995). RNA:DNA complex formation upon transcription of immunoglobulin switch regions: implications for the mechanism and regulation of class switch recombination. *Nuc. Acids Res.* **23**(24), 5006-11.

Dang, L. X. & Pettitt, B. M. (1987). Chloride ion pairs in water. *J. Amer. Chem. Soc.* **109**, 5531-5532.

Darden, T. A. (1996). Simulations on BPTI crystals: inclusion of long ranged dispersion attraction [personal communication].

Darden, T. A., York, D. M. & Pedersen, L. G. (1993). Particle mesh Ewald - An N log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **98**(12), 10089-10092.

Davies, D. B. (1978). Conformations of nucleosides and nucleotides. *Prog. Nuc. Magn. Res. Spect.* **12**, 135-225.

Davis, M. E. & McCammon, J. A. (1990). Electrostatics in biomolecular structure and dynamics. *Chem. Rev.* **90**, 509-521.

DeBolt, S. E. & Kollman, P. A. (1993). AMBERCUBE MD, parallelization of Ambers molecular dynamics module for distributed-memory hypercube computers. *J. Comp. Chem.* **14**, 312.

DeBolt, S. E., Pearlman, D. A. & Kollman, P. A. (1994). Free energy perturbation calculations on parallel computers- Demonstrations of scalable linear speedup. *J. Comp. Chem.* **15**, 351-373.

DeLeeuw, S. W., Perram, J. M. & Smith, E. R. (1980). Simulation of electrostatic systems in periodic boundary conditions I: Lattice sums and dielectric constants. *Proc. Royal Soc. London* **A373**, 27-56.

Dickerson, R. E. (1984). The DNA helix and how it is read. *Sci. Amer.* 251, 94-112.

Dickerson, R. E. (1989). Definitions and nomenclature of nucleic acid structure components. *Nuc. Acids Res.* **17**(5), 1797-803.

Dickerson, R. E., Drew, H. R., Conner, B. N., Wing, R. M., Fratini, A. V. & Kopka, M. L. (1982). The anatomy of A-, B-, and Z-DNA. *Science* **216**(4545), 475-85.

Dickerson, R. E., Goodsell, D. S., Kopka, M. L. & Pjura, P. E. (1987). The effect of crystal packing on oligonucleotide double helix structure. *J. Biomol. Struct. Dyn.* **5**(3), 557-79.

Dickerson, R. E., Goodsell, D. S. & Neidle, S. (1994). "...the tyranny of the lattice...". *Proc. Natl. Acad. Sci.* **91**(9), 3579-83.

Dill, K. A. (1990). Dominant forces in protein folding. *Biochem.* **29**(31), 7133-7155.

Ding, H. Q., Karasawa, N. & Goddard, W. A. (1992). Atomic level simulations on a million particles- the cell multipole method for Coulomb and London nonbond interactions. *J. Chem. Phys.* **97**(6), 4309-4315.

Di Stefano, D. L. & Wand, A. J. (1987). Two-dimensional 1H NMR study of human ubiquitin: a main chain directed assignment and structure analysis. *Biochem.* **26**(23), 7272-81.

Drew, H. R. & Dickerson, R. E. (1981). Structure of a B-DNA dodecamer. III. Geometry of hydration. *J. Mol. Biol.* **151**(3), 535-56.

Duan, Y., Wilkosz, P., Crowley, M. & Rosenberg, J. M. (1996). The Ewald summation improves the fidelity of molecular dynamics simulations of the DNA dodecamer d(CGCGAATTCGCG). *J. Mol. Biol.*, [submitted].

Edwards, K. J., Brown, D. G., Spink, N., Skelly, J. V. & Neidle, S. (1992). Molecular Structure of the B-DNA dodecamer d(CGCAAATTTGCG)$_2$. An Examination of Propeller Twist and Minor-groove Water Structure at 2.2 Å Resolution. *J. Mol. Biol.* **226**, 1161-1173.

Egli, M., Portmann, S. & Usman, N. (1996). RNA hydration: a detailed look. *Biochem.* **35**(26), 8489-94.

Egli, M., Usman, N., Zhang, S. G. & Rich, A. (1992). Crystal structure of an Okazaki fragment at 2-A resolution. *Proc. Natl. Acad. Sci.* **89**(2), 534-8.

Eisenberg, D. & McLachlan, A. D. (1986). Solvation energy in protein folding and binding. *Nature* **319**, 199-203.

Eisenburg, H. (1987). DNA flexing, folding and function. *Accts. Chem. Res.* **20**, 276-282.

Eisenhaber, F., Mannik, J. H. & Tumanyan, V. G. (1990a). Structural principles of B-DNA grooves hydration in fibers as revealed by Monte Carlo simulations and X-ray diffraction. *Biopol.* **29**, 1453-1464.

Eisenhaber, F., Tumanyan, V. G. & Abagyan, R. A. (1990b). Structure of the hydration shells of oligo (dA-dT)- oligo (dA-dT) and oligo d(A)- oligo(dT) tracts in B-type conformation on the basis of Monte Carlo calculations. *Biopol.* **30**, 563-581.

Eisenhaber, F., Tumanyan, V. G., Eisenmenger, F. & Gunia, W. (1989). Hydration of B-DNA: Comparison between the water network around poly(dG)-poly(dC) and poly(dG-dC)-poly(dG-dC) on the basis of Monte Carlo computations. *Biopol.* **28**, 741-761.

Eisenstein, M. & Shakked, Z. (1995). Hydration patterns and intermolecular interactions in A-DNA crystal structures. Implications for DNA recognition. *J. Mol. Biol.* **248**, 662-678.

Essex, J. W. & Jorgensen, W. L. (1995). An empirical boundary potential for water droplet simulations. *J. Comp. Chem.* **16**, 951-972.

Essmann, U. & Darden, T. A. (1996). Long-ranged electrostatic effects [in review].

Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H. & Pedersen, L. G. (1995). A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **103**(19), 8577-8593.

Ewald, P. (1921). Investigations of crystals by means of Roentgen rays. *Ann. Phys. (Leipzig)* **64**, 253-264.

Falk, M., Hartman, K. A. & Lord, R. C. (1963). Hydration of Deoxyribonucleic acid. II. An infared study. *J. Amer. Chem. Soc.* **85**, 397-391.

Fedoroff, O., Salazar, M. & Reid, B. R. (1993). Structure of a DNA:RNA hybrid duplex. Why RNase H does not cleave pure RNA. *J. Mol. Biol.* **233**(3), 509-23.

Feller, S. E., Pastor, R. W., Rojnuckarin, A., Bogusz, S. & Brooks, B. R. (1996). Effect of electrostatic force truncation on interfacial and transport properties of water. *J. Phys. Chem.* **100**, 17011-17020.

Ferguson, D. M. (1995). Parameterization and evaluation of a flexible water model. *J. Comp. Chem.* **16**, 501-511.

Ferrin, T. E., Huang, C. C., Jarvis, L. E. & Langridge, R. (1988). The MIDAS display system. *J. Mol. Graphics* **6**, 13-27.

Forester, T. R. & McDonald, I. R. (1991). Molecular dynamics studies of the behaviour of water molecules and small ions in concentrated solutions of polymeric B-DNA. *Mol. Phys.* **72**(3), 643-660.

Fox, T. & Kollman, P. A. (1996). The application of different solvation and electrostatic models in molecular dynamics simulations of ubiquitin: How well is the X-ray structure "maintained". *Proteins* **25**, 315-334.

Franklin, R. E. & Gosling, R. G. (1953). The structure of sodium thymonucleate fibres. I. The influence of water content. *Acta. Cryst.* **6**, 673-677.

Friedman, R. A. & Honig, B. (1995). A free energy analysis of nucleic acid base stacking in aqueous solution. *Biophys. J.* **69**(4), 1528-1535.

Fritsch, V., Ravishanker, G., Beveridge, D. L. & Westhof, E. (1993). Molecular dynamics simulations of poly(dA)-poly(dT): Comparisons between implicit and explicit solvent representations. *Biopol.* **33**, 1537-1552.

Fritsch, V. & Wolf, R. M. (1994). Molecular Dynamics Simulations of a r($GA_{12}G$)-d($CT_{12}C$) Hybrid Duplex. *J. Biomol. Struct. Dyn.* **11**(6), 1161-1174.

Fujiwara, T. & Shindo, H. (1985). Phosphorus-31 nuclear magnetic resonance of highly oriented DNA fibers. 2. Molecular motions in hydrated DNA. *Biochem.* **24**(4), 896-902.

Gao, X. & Jeffs, P. W. (1994). Sequence-dependent conformational heterogeneity of a hybrid DNA-RNA dodecamer duplex. *J. Biomol. NMR* **4**(3), 367-84.

Gao, Y.-G., Robinson, H., van Boom, J. H. & Wang, A. H.-J. (1995). Influence of counter-ions on the crystal structures of DNA decamers: Binding of $[Co(NH_3)_6]^{3+}$ and Ba^{2+} to A-DNA. *Biophys. J.* **69**, 559-568.

Gao, Y.-G., Sriram, M. & Wang, A. H.-J. (1993). Crystallographic studies of metal ion - DNA interactions: different bindingmodes of cobalt(II), copper(II) and barium(II) to N^7 of guanines in Z-DNA and a drug - DNA complex. *Nuc. Acids Res.* **21**(17), 4093-4101.

Gilson, M. K. (1995). Theory of electrostatic interactions in macromolecules. *Cur. Op. Struct. Biol.* **5**, 216-223.

Gilson, M. K., Davis, M. E., Luty, B. A. & McCammon, J. A. (1993). Computation of electrostatic forces on solvated molecules using the Poisson-Boltzmann equation. *J. Phys. Chem.* **97**, 3591-3600.

Gilson, M. K. & Honig, B. (1991). The inclusion of electrostatic hydration energies in molecular mechanics calculations. *J. Comp. Aided Mol. Design* **5**, 5-20.

Gilson, M. K., Sharp, K. A. & Honig, B. H. (1987). Calculating the electrostatic potential of molecules in solution: Method and error assessment. *J. Comp. Chem.* **9**, 327-335.

Gonzalez, C., Stec, W., Kobylanska, A., Hogrefe, R. I., Reynolds, M. & James, T. L. (1994). Structural study of a DNA-RNA hybrid duplex with a chiral phosphorothioate moiety by NMR: extraction of distance and torsion angle constraints and imino proton exchange rates. *Biochem.* **33**(37), 11062-72.

Gonzalez, C., Stec, W., Reynolds, M. A. & James, T. L. (1995). Structure and dynamics of a DNA.RNA hybrid duplex with a chiral phosphorothioate moiety: NMR and molecular dynamics with conventional and time-averaged restraints. *Biochem.* **34**(15), 4969-82.

Gorin, A. A., Ulyanov, N. B. & Zhurkin, V. B. (1990). S-N transition of the sugar ring in B-form DNA. *Molekulyarnaya Biologiya* **24**(5), 1300-1313.

Gorin, A. A., Zhurkin, V. B. & Olson, W. K. (1995). B-DNA twisting correlates with base-pair morphology. *J. Mol. Biol.* **247**(1), 34-48.

Gotte, M., Fackler, S., Hermann, T., Perola, E., Cellai, L., Gross, H. J., Le Grice, S. F. & Heumann, H. (1995). HIV-1 reverse transcriptase-associated RNase H cleaves RNA/RNA in arrested complexes: implications for the mechanism by which RNase H discriminates between RNA/RNA and RNA/DNA. *EMBO J.* **14**(4), 833-41.

Gray, D. M. & Ratliff, R. L. (1975). Circular dichroism spectra of poly[d(AC):d(GT)], poly[r(AC):r(GU)], and hybrids poly[d(AC):r(GU)] and poly[r(AC):d(GT)] in the presence of ethanol. *Biopol.* **14**(3), 487-98.

Greengard, L. (1988). *The rapid evaluation of potential fields in particle systems*, MIT Press, Cambridge, MA.

Greengard, L. & Rokhlin, V. (1989). On the evaluation of electrostatic interactions in molecular modeling. *Chemica Scripta* **29A**, 139-144.

Hall, K. B. & McLaughlin, L. W. (1991). Thermodynamic and structural properties of pentamer DNA-DNA, RNA-RNA, and DNA-RNA duplexes of identical sequence. *Biochem.* **30**(44), 10606-13.

Hagerman, P. J. (1981). Investigation of the flexibility of DNA using transient electric birefringence. *Biopol.* **20**, 1503-1535.

Hagerman, P. J. (1988). Flexibility of DNA. *Ann. Rev. Biophys. Chem.* **17**, 265-286.

Hagerman, P.J. (1994). Straightening out the bends in curved DNA. *Biochim. Biophys. Acta.* **1131***(2), 125-132.*

Hansen, U. M. & McClure, W. R. (1980). Role of the sigma subunit of Escherichia coli RNA polymerase in initiation. I. Characterization of core enzyme open complexes. *J. Biol. Chem.* **255**(20), 9556-63.

Haran, T. E., Shakked, Z., Wang, A. H. & Rich, A. (1987). The crystal structure of d(CCCCGGGG): a new A-form variant with an extended backbone conformation. *J. Biomol. Struct. Dyn.* **5**(2), 199-217.

Harrington, R.E. & Winicov, I. (1994). New concepts in protein-DNA recognition: sequence directed DNA bending and flexibility. *Prog. Nucleic Acid Res. Mol. Biol.* **47**, 195-270.

Harvey, S. C. (1989). Treatment of electrostatic effects in macromolecular modeling. *Proteins* **5**, 78-92.

Hendsch, Z. S. & Tidor, B. (1994). Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Prot. Sci.* **3**, 211-226.

Hingerty, B. E., Ritchie, R. H., Ferrell, T. L. & Turner, J. E. (1985). Dielectric effects in biopolymers: The theory of ionic saturation revisited. *Biopol.* **24**, 427-439.

Hockney, R. W. & Eastwood, J. W. (1981). *Computer simulation using particles*, McGraw-Hill, New York.

Holbrook, S. R. & Kim, S.-H. (1984). Local mobility of nucleic acids as determined from crystallographic data. I. RNA and B form DNA. *J. Mol. Biol.* **173**, 361-388.

Hummer, G., Soumpasis, D. M. & Neumann, M. (1993). Computer simulations do not support Cl-Cl pairing in aqueous NaCl solution. *Mol. Phys.* **81**, 1155-1163.

Ivanov, V. I., Minchenkova, L. E., Schyolkina, A. K. & Poletayev, A. I. (1973). Different conformations of double-stranded nucleic acids in solution as revealed by circular dichroism. *Biopol.* **12**, 89-110.

Jain, S. & Sundaralingam, M. (1989). Effect of crystal packing environment on conformation of the DNA duplex. Molecular structure of the A-DNA octamer d(G-T-G-T-A-C-A-C) in two crystal forms. *J. Biol. Chem.* **264**(22), 12780-4.

James, T.L. (1995). Nuclear magnetic resonance and nucleic acids. *Methods in Enzymology* **261**.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926-935.

Jorgensen, W. L. & Pranata, J. (1990). Importance of secondary interactions in tripy hydrogen bonded complexes- guanine-cytosine vs. uracil-2,6-diaminopyridine. *J. Amer. Chem. Soc.* **112**, 2008-2010.

Kang, Y. K., Nemethy, G. & Scheraga, H. A. (1988). Free energies of hydration of solute molecules. 4. Revised treatment of the hydration shell model. *J. Chem. Phys.* **79**, 926-935.

Katahira, M., Lee, S. J., Kobayashi, Y., Sugeta, H., Kyogoku, Y., Iwai, S., Ohtsuka, E., Benevides, J. M. & Thomas, G. J. (1990). Raman spectral studies of nucleic acids. 36. Structure in solution of the RNA-DNA hybrid $(rA)_8$-$(dT)_8$ determined by NMR and Raman spectroscopy. *J. Amer. Chem. Soc.* **112**(11), 4508-4512.

Keepers, J. W., Kollman, P. A., Weiner, P. K. & James, T. L. (1982). Molecular mechanical studies of DNA flexibility: Coupled backbone torsion angles and base-pair openings. *Proc. Natl. Acad. Sci.* **79**, 5537-5541.

Kim, S.-G. & Reid, B. R. (1992). Solution structure of T_nA_n DNA duplex GCCGTTAACGGC containing the *HpaI* recognition site. *Biochem.* **31**, 12103-12116.

Kollman, P. A., Keepers, J. W. & Weiner, P. K. (1982). Molecular-mechanics studies on $d(CGCGAATTCGCG)_2$ and dA_{10}-dT_{10}: An illustration of the coupling between sugar repuckering and DNA twisting. *Biopol.* **1`**, 2345-2376.

Kopka, M. L., Fratini, A. V., Drew, H. R. & Dickerson, R. E. (1983). Ordered water structure around a B-DNA dodecamer. A quantitative study. *J. Mol. Biol.* **163**(1), 129-46.

Kumar, S., Bouzida, D., Swendsen, R. H., Kollman, P. A. & Rosenberg, J. M. (1992). The weighted histogram analysis method for free-energy calculations on biomolecules. 1. The method. *J . Comp. Chem.* **13**(8), 1011-1021.

Kumar, S., Rosenberg, J. M., Bouzida, D., Swendsen, R. H. & Kollman, P. A. (1995). Multidimensional free-energy calculations using the weighted histogram analysis method. *J. Comp. Chem.* **16**(11), 1339-1350.

Lane, A. N., Ebel, S. & Brown, T. (1993). NMR assignments and solution conformation of the DNA.RNA hybrid duplex d(GTGAACTT)-r(AAGUUCAC). *Eur. J. Biochem.* **215**(2), 297-306.

Langan, P., Forsyth, V. T., Mahendrasingam, A., Pigram, W. J., Mason, S. A. & Fuller, W. (1992). A high angle neutron fibre diffraction study of the hydration of the A conformation of the DNA double helix. *J. Biomol. Struct. Dyn.* **10**(3), 489-503.

Laughton, C. A., Luque, F. J. & Orozco, M. (1995). Counterion distribution around DNA studied by molecular dynamics and Quantum Mechanical simulations. *J. Phys. Chem.* **99**, 11591-11599.

Laughton, C. A. & Neidle, S. (1992). Molecular dynamics simulation of the DNA triplex $d(TC)_5$-$d(GA)_5$-$d(C^+T)_5$. *J. Mol. Biol.* **223**, 519-529.

Lavery, R. (1996). [personal communication]

Lavery, R. & Hartmann, B. (1994). Modelling DNA conformational mechanics. *Biophys. Chem.* **50**(1-2), 33-45.

Lavery, R. & Sklenar, H. (1988). The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dyn.* **6**(1), 63-91.

Lavery, R., Zakrzewska, K. & Sklenar, H. (1995). JUMNA (junction minimisation of nucleic acids). *Comp. Phys. Comm.* **91**, 135-158.

Lee, H. Darden, T.A. & Pedersen, L. (1995a). Molecular dynamics simulation studies of a high resolution Z DNA crystal. *J. Chem. Phys.* **102**, 3830-3834

Lee, H., Darden, T.A. & Pedersen. L. (1995b). Accurate crystal molecular dynamics simulations using particle mesh Ewald: RNA dinucleotides-ApU and GpC. *Chem. Phys. Lett.* **243**, 229-235

Leijon, M., Zdunek, J., Fritzsche, H., Sklenar, H. & Graslund, A. (1995). NMR studies and restrained molecular dynamics calculations of a long A+T-rich stretch in DNA. Effects of phosphate charge and solvent approximations. *Eur. J. Biochem.* **234**, 832-842.

Levin-Zaidman, S., Reich, Z., Wachtel, E. J. & Minsky, A. (1996). Flow of structural information between four DNA conformational levels. *Biochem.* **35**, 2985-2991.

Levitt, M. (1978). How many base-pairs per turn does DNA have in solution and in chromatin? Some theoretical calculations. *Proc. Nat. Acad. Sci.* **75**, 640-644.

Levitt, M. (1983). Computer simulation of DNA double-helix dynamics. *Cold Spring Harbor Symp. Quant. Biol.* **47**, 251-262.

Levitt, M., Hirshberg, M., Sharon, R. & Daggett, V. (1995). Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Comp. Phys. Comm.* **91**, 215-231.

Levitt, M., Sander, C. & Stern, P. S. (1985). Protein normal-mode dynamics: Trypsin inhibitor, Crambin, Ribonuclease and Lysozyme. *J. Mol. Biol.* **181**, 423-447.

Levy, R. M., Kushick, J., Perahia, D. & Karplus, M. (1984). Evaluation of configurational entropy for proteins: Application to molecular dynamics simulations of an α-helix. *Macromol.* **17**, 1370-1374.

Liepinsh, E., Otting, G. & Wuthrich, K. (1992). NMR Observation of Individual Molecules of Hydration Water Bound to DNA Duplexes: Direct Evidence for a Spine of Hydration Water Present in Aqueous Solution. *Nuc. Acid Res.* **20**(24), 6549-6553.

Lipanov, A., Kopka, M. L., Kaczor-Grzeskowiak, M., Quintana, J. & Dickerson, R. E. (1993). Structure of the B-DNA decamer C-C-A-A-C-I-T-T-G-G in two different space groups: conformational flexibility of B-DNA. *Biochem.* **32**(5), 1373-89.

Loncharich, R.J. & Brooks B. R. (1990). Temperature dependence of dynamics of hydrated myoglobin. Comparison of force field calculations with neutron scattering data. *J. Mol. Biol.* **215**, 439-455.

Louise-May, S., Auffinger, P. & Westhof, E. (1996). Calculations of nucleic acid conformations. *Cur. Op. Struct. Biol.* **6**(3), 289-298.

Luty, B., Davis, M. E., Tironi, I. G. & van Gunsteren, W. F. (1994). A comparison of particle-particle particle-mesh and Ewald methods for calculating electrostatic interactions. *Mol. Sim.* **14**, 11-20.

Luty, B. A., Tironi, I. G. & van Gunsteren, W. F. (1995). Lattice-sum methods for calculating electrostatic interactions in molecular simulations. *J. Chem. Phys.* **103**, 3014-3021.

Macke, T. J. & Case, D. A. (1996). NAB (Nucleic Acid Builder). The Scripps Research Institute, La Jolla, CA.

Mackerell, A. D., Wiorkiewiczkuczera, J. & Karplus, M. (1995). An all-atom empirical energy function for the simulation of nucleic acids. *J. Amer. Chem. Soc.* **117**(48), 11946-11975.

Madura, J. D., Briggs, J. M., Wade, R. C., Davis, M. E. & McCammon, J. A. (1995). Electrostatics and diffusion of molecules in solution- Simulations with the University of Houston brownian dynamics program. *Comp. Phys. Comm.* **91**(1-3), 57-95.

McCammon, J. A. & Harvey, S. C. (1987). *Dynamics of proteins and nucleic acids*, Cambridge University Press, Cambridge.

McConnell, K. J., Nirmala, R., Young, M. A., Ravishanker, G. & Beveridge, D. L. (1994). A nanosecond molecular dynamics trajectory for a B DNA double helix - Evidence for substates. *J. Amer. Chem. Soc.* **116**(10), 4461-4462.

Mehrotra, P. K. & Beveridge, D. L. (1980). Structural analysis of molecular solutions based on quasi-component distribution functions. Application to $[H_2C0]_{aq}$ at 25 C. *J. Amer. Chem. Soc.* **102**, 4287-4294.

Mezei, M. & Beveridge, D. L. (1986). Structural chemistry of biomolecular hydration via computer simulation: The proximity criterion. *Meth. Enzym.* **127**, 21-47.

Miaskiewicz, K., Osman, R. & Weinstein, H. (1993). Molecular dynamics simulation of the hydrated d(CGCGAATTCGCG)$_2$ dodecamer. *J. Amer. Chem. Soc.* **115**, 1526-1537.

Miller, J. L. & Kollman, P. A. (1996). Theoretical studies of an exceptionally stable RNA tetraloop: Observation of convergence from an incorrect NMR structure to the correct one using unrestrained molecular dynamics. *J. Mol. Biol.* [submitted].

Misra, V. K. & Honig, B. (1995). On the magnitude of the electrostatic contribution to ligand-DNA interactions. *Proc. Natl. Acad. Sci.* **92**, 4691-4695.

Misra, V. K. & Honig, B. (1996). The electrostatic contribution to the B to Z transition of DNA. *Biochem.* **35**(4), 1115-1124.

Misra, V. K., Sharp, K. A., Friedman, R. A. & Honig, B. (1994). Salt effects on ligand-DNA binding. Minor groove binding antibiotics. *J. Mol. Biol.* **238**, 245-263.

Miyamoto, S. & Kollman, P. A. (1992). SETTLE - an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J. Comp. Chem.* **13**(8), 952-962.

Mizan, T. I., Savage, P. E. & Ziff, R. M. (1994). Molecular dynamics of supercritical water using a flexible SPC model. *J. Phys. Chem.* **98**, 13067-13076.

Mohan, V., Smith, P. E. & Pettitt, B. M. (1993). Molecular dynamics simulation of ions and water around triplex DNA. *J. Phys. Chem.* **97**, 12984-12990.

Mohr, S. C., Sokolov, N. V. H. A., He, C. & Setlow, P. (1991). Binding of small acid-soluble spore proteins from *Bacillus subtilis* changes the conformation of DNA from B to A. *Proc. Nat. Acad. Sci.* **88**, 77-81.

Muller, H. P. & Varmus, H. E. (1994). DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. *EMBO. J.* **13**(19), 4704-4714.

Mujeeb, A., Kerwin, S. M., Kenyon, G. L. & James, T. L. (1993). Solution structure of a conserved DNA sequencfrom the HIV-1 genome: restrained molecular dynamics simulations with distance and torsion angle restraints derived from two-dimensional NMR spectra. *Biochem.* **32**(49), 13419-13431.

Narten, A. H. & Levy, H. A. (1971). *J. Chem. Phys.* **55**, 2263.

Nilsson, L. & Karplus, M. (1984). *J. Comp. Chem.* **1**, 591-616.

Norberg, J. & Nilsson, L. (1995). NMR relaxation times, dynamics, and hydration of a nucleic acid fragment from molecular dynamics simulations. *J. Phys. Chem.* **99**, 14876-14884.

Norberg, J. & Nilsson, L. (1996). Glass transition in DNA from molecular dynamics simulations. *Proc. Natl. Acad. Sci.* **93**(19), 10173-10176.

Novotny, J., Rashin, A. A. & Bruccoleri, R. (1988). Criteria that discriminate between native proteins and incorrectly folded models. *Proteins* **4**, 19-30.

Ogawa, T. & Okazaki, T. (1980). Discontinuous DNA replication. *Ann. Rev. Biochem.* **49**, 421-57.

Olson, W. (1982a). Theoretical studies of nucleic acid conformation: Potential energies, chain statistics, and model building. In *Topics in nucleic acid structure* (Neidle, S., ed.). Macmillan, London.

Olson, W. K. (1982b). Computational studies of polynucleotide flexibility. *Nuc. Acids Res.* **10**(3), 777-87.

Olson, W. K. & Sussman, J. L. (1982). How flexible is the furanose ring? 1. A comparison of experimental and theoretical studies. *J. Amer. Chem. Soc.* **104**, 270-278.

Olson, W. K., Babcock, M. S., Gorin, A., Liu, G., Marky, N. L., Martino, J. A., Pedersen, S. C., Srinivasan, A. R., Tobias, I., Westcott, T. P. & Zhang, P. (1995). Flexing and folding double helical DNA. *Biophys. Chem.* **55**, 7-29.

Olson, W. K., Marky, N. L., Jernigan, R. L. & Zhurkin, V. B. (1993). Influence of fluctuations on DNA curvature. A comparison of flexible and static wedge models of intrinsically bent DNA. *J. Mol. Biol.* **232**, 530-554.

Onsager, L. (1936). Electric moments of molecules in liquids. *J. Amer. Chem. Soc.* **58**, 1486-1493.

Ooi, I., Oobatake, M., Nemethy, G. & Scheraga, H. A. (1987). Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci.* **84**, 3086-3090.

Pattabiraman, N., Rao, S. N., Scott, K., Langridge, R. & Kollman, P. A. (1987). Molecular mechanical studies on left- and right- handed B-DNA. *Biopol.* **26**, 403-414.

Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. S., Cheatham, T. E., Debolt, S., Ferguson, D., Seibel, G. & Kollman, P. (1995). AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structure and energetic properties of molecules. *Comp. Phys. Comm.* **91**(1-3), 1-41.

Peck, L. J. & Wang, J. C. (1981). Sequence dependence of the helical repeat of DNA in solution. *Nature* **292**, 375-378.

Petersen, H. G. (1995). Accuracy and efficiency of the particle mesh Ewald method. *J. Chem. Phys.* **103**(9), 3668-3679.

Piskur, J. & Rupprecht, A. (1995). Aggregated DNA in ethanol solution. *FEBS Lett.* **375**, 174-178.

Poncin, M., Hartmann, B. & Lavery, R. (1992). Conformational sub-states in B-DNA. *J. Mol. Biol.* **226**(3), 775-94.

Portmann, S., Usman, N. & Egli, M. (1995). The crystal structure of r(CCCCGGGG) in two distinct lattices. *Biochem.* **34**(23), 7569-75.

Potaman, V. N., Bannikov, Y. A. & Shlyachtenko, L. S. (1980). Sedimentation of DNA in ethanol-water solutions within the interval of B to A transition. *Nuc. Acids Res.* **8**(3), 635-642.

Prive, G. G., Yanagi, K. & Dickerson, R. E. (1991). Structure of the B-DNA decamer C-C-A-A-C-G-T-T-G-G and comparison with isomorphous decamers C-C-A-A-G-A-T-T-G-G and C-C-A-G-G-C-C-T-G-G. *J. Mol. Biol.* **217**(1), 177-99.

Pullman, A. & Pullman, B. (1975). New paths in the molecular orbital approach to solvation of biological molecules. *Ann. Rev. Biophys.* **7**, 505-566.

Ramakrishnan, B. & Sundaralingam, M. (1993). Crystal packing effects on A-DNA helix parameters: a comparative study of the isoforms of the tetragonal & hexagonal family of octamers with differing base sequences. *J. Biomol. Struct. Dyn.* **11**(1), 11-26.

Ramstein, J. & Lavery, R. (1988). Energetic coupling between DNA bending and base pair opening. *Proc. Natl. Acad. Sci.* **85**, 7231-7235.

Ravishanker, G., Swaminathan, S., Beveridge, D. L., Lavery, R. & Sklenar, H. (1989). Conformational and helicoidal analysis of 30 ps of molecular dynamics on the d(CGCGAATTCGCG) double helix: "curves", dials and windows. *J. Biomol. Struct. Dyn.* **6**(4), 669-99.

Rhodes, D. & Klug, A. (1980). Helical periodicity of DNA determined by enzyme digestion. *Nature* **286**, 573-578.

Ribas de Pouplana, L., Auld, D. S., Kim, S. & Schimmerl, P. (1996). A mechanism for reducing entropic cost of induced fit in protein-RNA recognition. *Biochem.* **35**(25), 8095-8102.

Roberts, R. W. & Crothers, D. M. (1992). Stability and properties of double and triple helices: dramatic effects of RNA or DNA backbone composition. *Science* **258**(5087), 1463-6.

Roberts, J. E. & Schnitker, J. (1995). Boundary conditions in simulations of aqueous ionic solutions: A systematic study. *J. Phys. Chem.* **99**, 1322-1331.

Robinson, H. & Wang, A. H.-J. (1996). Neomycin, spermine and hexaamminecobalt(III) share common structural motifs in converting B- to A-DNA. *Nuc. Acids Res.* **24**(4), 676-682.

Roitberg, A. & Elber, R. (1991). Modeling side chains in peptides and proteins: Application of the locally enhanced sampling and the simulated annealing methods to find minimum energy conformations. *J. Chem. Phys.* **95**(12), 9277-9286.

Rosenfield, R. E. J., Swanson, S. M., Meyer, E. F. J., Carrell, H. L. & Murray-Rust, P. (1984). Mapping the atomic environment of functional groups: turning 3D scatter plots into pseudo-density contours. *J. Mol. Graphics* **2**, 43-46.

Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. (1977). Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comp. Phys.* **23**, 327-341.

Saenger, W. (1984). *Principles of Nucleic Acid Structure.* Springer Advanced Texts in Chemistry (Cantor, C. E., Ed.), Springer-Verlag, New York.

Saenger, W., Hunter, W. N. & Kennard, O. (1986). DNA conformation is determined by economics in the hydration of phosphate groups. *Nature* **324**, 385-388.

Salazar, M., Fedoroff, O. Y., Miller, J. M., Ribeiro, N. S. & Reid, B. R. (1993). The DNA strand in DNA.RNA hybrid duplexes is neither B-form nor A-form in solution. *Biochem.* **32**(16), 4207-15.

Sanghani, S. R., Zakrzewska, K., Harvey, S. C. & Lavery, R. (1996). Molecular modelling of $(A_4T_4NN)_n$ and $(T_4A_4NN)_n$: sequence elements responsible for curvature. *Nuc. Acids Res.* **24**(9), 1632-1637.

Sanghani, S. R. & Lavery, R. (1994). Theoretical Studies of DNA-RNA Hybrid Conformations. *Nuc. Acids Res.* **22**(8), 1444-1449.

Sarai, A., Mazur, J., Nussinov, R. & Jernigan, R. L. (1988). Origin of DNA helical structure and its sequence dependence. *Biochem.* **27**, 8498-8502.

Schiffer, C. A., Caldwell, J. W., Kollman, P. A. & Stroud, R. M. (1993). Protein structure prediction with a combined solvation free energy-molecular mechanics force field. *Mol. Sim.* **10**(2-6), 121.

Schmidt, K. E. & Lee, M. A. (1991). Implementing the fast multipole method in three dimensions. *J. Stat. Phys.* **63**, 1223-1235.

Schmitz, U., Ulyanov, N. B., Kumar, A. & James, T. L. (1993). Molecular dynamics with weighted time-averaged restraints for a DNA octamer. Dynamic interpretation of nuclear magnetic resonance data. *J. Mol. Biol.* **234**(2), 373-89.

Schneider, B. & Berman, H. M. (1995). Hydration of the DNA bases is local. *Biophys. J.* **69**(6), 2661-9.

Schneider, B., Cohen, D. M., Schleifer, L., Srinivasan, A. R., Olson, W. K. & Berman, H. M. (1993). A systematic method for studying the spatial distribution of water molecules around nucleic acid bases. *Biophys. J.* **65**(6), 2291-303.

Schreiber, H. & Steinhauser, O. (1992). Cutoff size does strongly influence molecular dynamics results on solvated polypeptides. *Biochem.* **31**(25), 5856-60.

Schultz, S. C., Shields, G. C. & Steitz, T. A. (1991). Crystal structure of a CAP-DNA complex- the DNA is bent by 90 degrees. *Science* **253**, 1001-1007.

Seeman, N.C., Rosenburg, J. M. & Rich, A. (1976). Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl. Acad. Sci.* **73**, 804-808.

Selvin, P. R., Cook, D. N., Pon, N. G., Bauer, W. R., Klein, M. P. & Hearst, J. P. (1992). Torsional rigidity of positively and negatively supercoiled DNA. *Science* **255**, 82-85.

Senderowitz, H., Guarnieri, F. & Still, W. C. (1995). A smart Monte Carlo technique for free energy simulation of multiconformational molecules, direct calculations of the conformational populations of organic molecules. *J. Amer. Chem. Soc.* **117**(31), 8211-8291.

Shakked, Z., Guerstein-Guzikevich, G., Eisenstein, M., Frolow, F. & Rabinovich, D. (1989). The conformation of the DNA double helix in the crystal is dependent on its environment. *Nature* **342**(6248), 456-60.

Sharp, K. A. (1991). Incorporating solvent and ion screening into molecular dynamics using the finite-difference Poisson-Boltzmann approach. *J. Comp. Chem.* **12**(4), 454-468.

Sharp, K. A. & Honig, B. (1990). Electrostatic interactions in macromolecules: theory and applications. *Annu. Rev. Biophys. Biophys. Chem.* **19**, 301-332.

Shi, Y. & Berg, J. M. (1996). DNA unwinding induced by zinc finger protein binding. *Biochem.* **35**, 3845-3848.

Shimada, J., Kaneko, H. & Takada, T. (1993). Efficient calculations of Coulombic interactions in biomolecular simulations with periodic boundary conditions. *J. Comp. Chem.* **14**, 867-868.

Shindo, H., Fujiwara, T.,.Akutsu, H., Matsumoto, U. & Kyogoku, Y. (1985). Phosphorus-31 nuclear magnetic resonance of highly oriented DNA fibers. 1. Static geometry of DNA double helices. *Biochem.* **24**(4), 887-95.

Schultz, J., Rupprecht, A., Song, Z., Piskur, J., Nordenskiold, L. & Lahajnar, G. (1994). A mechanochemical study of MgDNA fibers in ethanol-water solutions. *Biophys. J.* 66, 810-819.

Siebel, G. L., Singh, U. C. & Kollman, P. A. (1985). A molecular dynamics simulation of double-helical B-DNA including counterions and water. *Proc. Nat. Acad. Sci.* **82**, 6537-6540.

Singh, U. C., Pattabiraman, N., Langridge, R. & Kollman, P. A. (1986). Molecular mechanical studies of d(CGTACG)2: Complex of triostin A with the middle AT base pairs in either Hoogsteen or Watson-Crick pairing. *Proc. Natl. Acad. Sci.* **83**, 6402-6406.

Singh, U. C., Weiner, S. C. & Kollman, P. A. (1985). Molecular dynamics simulations of d(C-G-C-G-A)-d(T-C-G-C-G) with and without "hydrated" counterions. *Proc. Natl. Acad. Sci.* **82**, 755-759.

Sitkoff, D., Sharp, K. A. & Honig, B. (1994). Accurate calculation of hydration free energies using macroscopic solvent models. *J. Phys. Chem.* **98**, 1978-1988.

Smith, P. E. & Pettitt, B. M. (1991). Peptides in ionic solution- A comparison of the Ewald and switching function techniques. *J. Chem. Phys.* **95**(11), 8430-8441.

Smith, P. E. & Pettitt, B. M. (1996). Ewald artifacts in liquid state molecular dynamics simulations. *J. Chem. Phys.* **105**(10), 4289-4293.

Sobel, E. S. & Harbst, J. A. (1991). Effects of Na+ on the persistence length and excluded volume of T7 bacteriophage DNA. *Biopol.* **31**, 1559-1564.

Solvason, D., Kolafa, J., Petersen, H. G. & Perram, J. W. (1995). A rigorous comparison of the Ewald method and the fast multipole method in two dimensions. *Comp. Phys. Comm.* **87**, 307-318.

Spolar, R. S. & Record, M. T. (1994). Coupling of local folding to site-specific binding of proteins to DNA. *Science* **263**, 777-784.

Steely, H. T., Jr., Gray, D. M. & Ratliff, R. L. (1986). CD of homopolymer DNA-RNA hybrid duplexes and triplexes containing A-T or A-U base pairs. *Nuc. Acids Res.* **14**(24), 10071-90.

Steinbach, P. J. & Brooks, B. R. (1994). New spherical-cutoff methods for long-range forces in macromolecular simulation. *J. Comp. Chem.* **15**(7), 667-683.

Steitz, T. A. (1990). Structural studies of protein-nucleic acid interaction: the sources of sequence-specific binding. *Quart. Rev. Biophys.* **23**(3), 205-280.

Still, W. C., Tempczyk, A., Hawley, R. C. & Hendrickson, T. (1990). Semi analytical treatment of solvation for molecular mechanics and dynamics. *J. Amer. Chem. Soc.* **112**, 6127-6128.

Stryer, L. (1988). *Biochemistry*. 3rd ed., W.H. Freeman and Company, New York.

Subramanian, P. S. & Beveridge, D. L. (1989). A Theoretical Study of the Aqueous Hydration of Canonical B d(CGCGAATTCGCG): Monte Carlo Simulation and Comparison with Crystallographic Ordered Water Sites. *J. Biomol. Struct. Dyn.* **6**(6), 1093-1122.

Subramanian, P. S. & Beveridge, D. L. (1993). A Monte Carlo simulation study of the aqueous hydration of d(CGCGCG) in Z form. *Theor. Chim. Acta.* **85**, 3-15.

Subramanian, P. S., Pitchumani, S., Beveridge, D. L. & Berman, H. M. (1990a). A Monte Carlo simulation study of the aqueous hydration of r(GpC): Comparison with crystallographic ordered water sites. *Biopol.* **29**, 771-783.

Subramanian, P. S., Ravishanker, G. & Beveridge, D. L. (1988). Theoretical considerations on the "spine of hydration" in the minor groove of d(CGCGAATTCGCG).d(GCGCTTAAGCGC): Monte Carlo computer simulation. *Proc. Natl. Acad. Sci.* **85**(6), 1836-40.

Subramanian, P. S., Swaminathan, S. & Beveridge, D. L. (1990b). Theoretical account of the "spine of hydration" in the minor groove of d(CGCGAATTCGCG). *J. Biomol. Struct. Dyn.* **7**, 1161-1165.

Sundaralingam, M. & Westhof, E. (1981). The nature of the mobility of the sugar and its effects on the dynamics and functions of RNA and DNA. In *Biomolecular Stereodynamics* (Sarma, R. H., ed.), Vol. 1, pp. 301-326. Adenine Press, New York.

Sussman, J. (1996). Protein Data Bank. http://www.pdb.bnl.gov, Brookhaven, NY.

Swaminathan, S., Ravishanker, G. & Beveridge, D. L. (1991). Molecular dynamics of B-DNA including water and counterions: A 140 ps trajectory for d(CGCGAATTCGCG) based on the GROMOS force field. *J. Amer. Chem. Soc.* **113**, 5027-5040.

Tanford, C. & Kirkwood, J. G. (1957). Theory of protein titration curves. I. General equations for impenetrable spheres. *J. Amer. Chem. Soc.* **79**(20), 5333-5339.

Teleman, O. & Wallqvist, A. (1990). Ewald summation retards translational motion in molecular dynamics simulation of water. *Int. J. Quant. Chem.* **24**, 245-249.

Texter, J. (1978). Nucleic acid-water interactions. *Prog. Biophys. Mol. Biol.* **33**(1), 83-97.

Thiessen, W. E. & Narten, A. H. (1982). *J. Chem. Phys.* **77**, 2656.

Thomas, G. J., Jr., Benevides, J. M., Overman, S. A., Ueda, T., Ushizawa, K., Saitoh, M. & Tsuboi, M. (1995). Polarized Raman spectra of oriented fibers of A DNA and B DNA: anisotropic and isotropic local Raman tensors of base and backbone vibrations. *Biophys. J.* **68**(3), 1073-88.

Tidor, B., Irikura, K. K., Brooks, B. R. & Karplus, M. (1983). Dynamics of DNA oligomers. *J. Biomol. Struct. Dyn.* **1**, 231-252.

Tilton, R. F., Weiner, P. K. & Kollman, P. A. (1983). An analysis of sequence dependence of the structure and energy of A- and B-DNA models using molecular mechanics. *Biopol.* **22**, 969-1002.

Ts'o, P. O. P., Melvin, I. S. & Olson, A. C. (1962). Interaction and association of bases in aqueous solution. *J. Amer. Chem. Soc.* **85**, 1289-1296.

Tung, C.-S. & Harvey, S. C. (1984). A molecular mechanical model to predict helix twist angles of B-DNA. *Nuc. Acids Res.* **12**(7), 3343-3356.

Uhlmann, E. & Peyman, A. (1990). Antisense Oligonucleotides- A new therapeutic principle. *Chem. Rev.* **90**, 543-584.

Ulyanov, N. B. & James, T. L. (1995). Statistical analysis of DNA duplex structural features. *Meth. Enzym.* **261**, 90-120.

Ulyanov, N. B., Schmitz, U., Kumar, A. & James, T. L. (1995). Probability assessment of conformational ensembles: sugar repuckering in a DNA duplex in solution. *Biophys. J.* **68**(1), 13-24.

Ulyanov, N. B. & Zhurkin, V. B. (1984). Sequence-dependent anisotropic flexibility of B-DNA. A conformational study. *J. Biomol. Struct. Dyn.* **2**(2), 361-385.

Umrania, Y., Nikjoo, H. & Goodfellow, J. M. (1995). A knowledge-based model of DNA hydration. *Int. J. Radiat. Biol.* **67**(2), 145-52.

Valleau, J. P. & Whittington, S. G. (1977). A guide to Monte Carlo for Statistical Mechanics: 1. Highways. In *Statistical Mechanics, A. A Modern Theoretical Chemistry* (Berne, B. J., ed.), Vol. 5-6. Plenum Press, New York.

Van Gunsteren, W.F. & Berendsen, H.J.C. (1986). *GROMOS86: Groningen Molecular Simulation System*; University of Groningen: Groningen, The Netherlands.

van Gunsteren, W. F. & Berendsen, H. J. C. (1987). *Groningen molecular simulation (GROMOS) library manual.* BIOMOS, Nijenborgh, Groningen, The Netherlands.

van Gunsteren, W. F. & Berendsen, H. J. C. (1990). Computer Simulation of Molecular dynamics: Methodology, Applications, and Perspectives in Chemistry. *Angew. Chem. Int. Ed. Engl.* **29**, 992-1023.

van Gunsteren, W. F., Berendsen, H. J. C., Geurtsen, R. G. & Zwinderman, H. R. J. (1986). A molecular dynamics computer simulation of an eight-base-pair DNA fragment in aqueous solution: Comparison with experimental two-dimensional NMR data. In *Computer Simulation of Chemical and Biomolecular Systems*, pp. 287-303. NY Acad. Sci. 482, New York.

Varani, G., Cheong, C. & Tinoco, I. J. (1991). Structure of an unusually stable RNA hairpin. *Biochem.* **30**(13), 3280-3289.

Varmus, H. (1988). Retroviruses. *Science* **240**(4858), 1427-35.

Vijay-Kumar, S., Bugg, C. E. & Cook, W. J. (1987). Structure of ubiquitin refined at 1.8 Å resolution. *J. Mol. Biol.* **194**(3), 531-544.

Vincent, J. J. & Merz, K. M. (1995). A highly portable parallel implementation of Amber4 using the Message Passing Interface standard. *J. Comp. Chem.* **16**(11), 1420-1427.

von Kitzing, E. & Diekmann, S. (1987). Molecular mechanics calculations of dA_{12}-dT_{12} and the curved molecule d(GCTCGAAAAA)$_4$-d(TTTTTCGAGC)$_4$. *Eur. Biophys. J.* **15**, 13-26.

Wang, J. C. (1969). Variation of the average rotation angle of the DNA helix and the superhelical turns of covalently closed cyclic [bacteriophage] DNA. *J. Mol. Biol.* **43**, 25-39.

Wang, A. H., Quigley, G. J., Kolpak, F. J., Crawford, J. L., van Boom, J. H., van der Marel, G. & Rich, A. (1979). Molecular structure of a left-handed double helical DNA fragment at atomic resolution. *Nature* **283**, 743-745.

Wang, A. H., Fujii, S., van Boom, J. H., van der Marel, G. A., van Boeckel, S. A. & Rich, A. (1982). Molecular structure of r(GCG)d(TATACGC): a DNA--RNA hybrid helix joined to double helical DNA. *Nature* **299**(5884), 601-4.

Warshel, A. & Aqvist, J. (1991). Electrostatic energy and macromolecular function. *Ann. Rev. Biophys. Biophys. Chem.* **20**, 267-298.

Warshel, A. & Levitt, M. (1976). Theoretical studies of enzyme reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* **103**, 227-249.

Warshel, A. & Russell, S. T. (1984). Calculation of electrostatic interactions in biological systems and in solution. *Quart. Rev. Biophys.* **17**, 283-422.

Watson, J. D. & Crick, F. H. C. (1953). A structure for deoxyribonucleic acid. *Nature* **171**, 737-738.

Weeks, K. M. & Crothers, D. M. (1993). Major groove accessibility of RNA. *Science* **261**, 1574-1577.

Weerasinghe, S., Smith, P. E., Mohan, V., Cheng, Y. K. & Pettitt, B. M. (1995a). Nanosecond dynamics and structure of a model DNA triple helix in saltwater solution. *J. Amer. Chem. Soc.* **117**(8), 2147-2158.

Weerasinghe, S., Smith, P. E. & Pettitt, B. M. (1995b). Structure and stability of a model pyrimidine-purine-purine DNA triple helix with a GC-T mispatch by simulation. *Biochem.* **34**, 16269-16278.

Weidlich, T., Lindsay, S. M. & Rupprecht, A. (1988). Counterion effects on the structure and dynamics of solid DNA. *Phys. Rev. Lett.* **61**, 1674-1677.

Weidlich, T., Lindsay, S. M., Rui, Q., Rupprecht, A., Peticolas, W. L. & Thomas, G. A. (1990). A Raman study of low frequency intrahelical modes in A-, B- and C-DNA. *J. Biomol. Struc. Dyn.* **1**, 139-171.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G. S., Profeta, J. & Weiner, P. (1984). A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Amer. Chem. Soc.* **106**, 765-784.

Weisz, K., Shafer, R. H., Egan, W. & James, T. L. (1994). Solution structure of the octamer motif in immunoglobulin genes via restrained molecular dynamics calculations. *Biochem.* **33**(1), 354-66.

Wesson, L. & Eisenberg, D. (1992). Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Prot. Sci.* **1**, 227-235.

Westhof, E. (1988). Water: An integral part of nucleic acid structure. *Ann. Rev. Biophys. Biophys. Chem.* **17**, 125-144.

Williams, R. L., Vila, J., Perrot, G. & Scheraga, H. A. (1992). Empirical solvation models in the context of conformational energy searches: Application to Bovine Pancreatic Trypsin Inhibitor. *Proteins.* **14**, 110-119.

Wolf, B. & Hanlon, S. (1975). Structural transitions of deoxyribonucleic acid in aqueous electrolyte solutions. II. The role of hydration. *Biochem.* **14**(8), 1661-1670.

Yanagi, K., Prive, G. G. & Dickerson, R. E. (1991). Analysis of local helix geometry in three B-DNA decamers and eight dodecamers. *J. Mol. Biol.* **217**(1), 201-14.

Yang, L., Weerasinghe, S., Smith, P. E. & Pettitt, B. M. (1995). Dielectric response of triplex DNA in ionic solution from simulations. *Biophys. J.* **69**, 1519-1527.

York, D. M., Darden, T. A., Pedersen, J. G. & Anderson, M. W. (1993a). Molecular dynamics simulation of HIV-1 protease in a crystalline environment and in solution. *Biochem.* **32**(6), 1443-1453.

York, D. M., Darden, T. A. & Pedersen, L. G. (1993b). The effect of long-range electrostatic interactions in simulations of macromolecular crystals- A comparison of the Ewald and truncated list methods. *J. Chem. Phys.* **99**(10), 8345-8348.

York, D. M., Wlodawer, A., Pedersen, L. G. & Darden, T. A. (1994). Atomic-level accuracy in simulations of large protein crystals. *Proc. Natl. Acad. Sci.* **91**(18), 8715-8.

York, D. & Yang, W. (1994). The fast Fourier Poisson method for calculating Ewald sums. *J. Chem. Phys.* **101**(4), 3298-3300.

York, D. M., Yang, W., Lee, H., Darden, T. A. & Pedersen, L. (1995). Toward accurate modeling of DNA: the importance of long-range electrostatics. *J. Amer. Chem. Soc.* **117**, 5001-5002.

Young, M. A., Ravishanker, G., Beveridge, D. L. & Berman, H. M. (1995). Analysis of local helix bending in crystal structures of DNA oligonucleotides and DNA-protein complexes. *Biophys. J.* **68**(6), 2454-68.

Young, M.A., Srinivasan, J., Golier, I., Kumar, S., Beveridge, D.L. & Bolton, P.H. (1995b). Structure determination and analysis of local bending in an A-tract DNA duplex: comparison of results from crystallography, Nuclear Magnetic Resonance, and molecular dynamics simulation on d[CGCAAAAAGCG]$_2$. *Methods in Enzymology* **261**, 121-144.

Young, M. A., Jayaram, B. & Beveridge, D. L. (1996). Counterions may intrude into the spine of hydration in the minor groove of B-DNA: fractional occupancy of electronegative pockets. *J. Amer. Chem. Soc.* [in press].

Zakrzewska, K., Madami, A. & Lavery, R. (1996). Poisson-Boltzmann calculations for nucleic acids and nucleic acid complexes. *Chem. Phys.* **204**, 263-269.

Zauhar, R. J. (1991). The incorporation of hydration forces determined by continuum electrostatics into molecular mechanics simulations. *J. Comp. Chem.* **12**(5), 575-583.

Zehfus, M. H. & Johnson, W. C. J. (1984). Conformation of P-form DNA. *Biopolymers* **23**, 1269-1281.

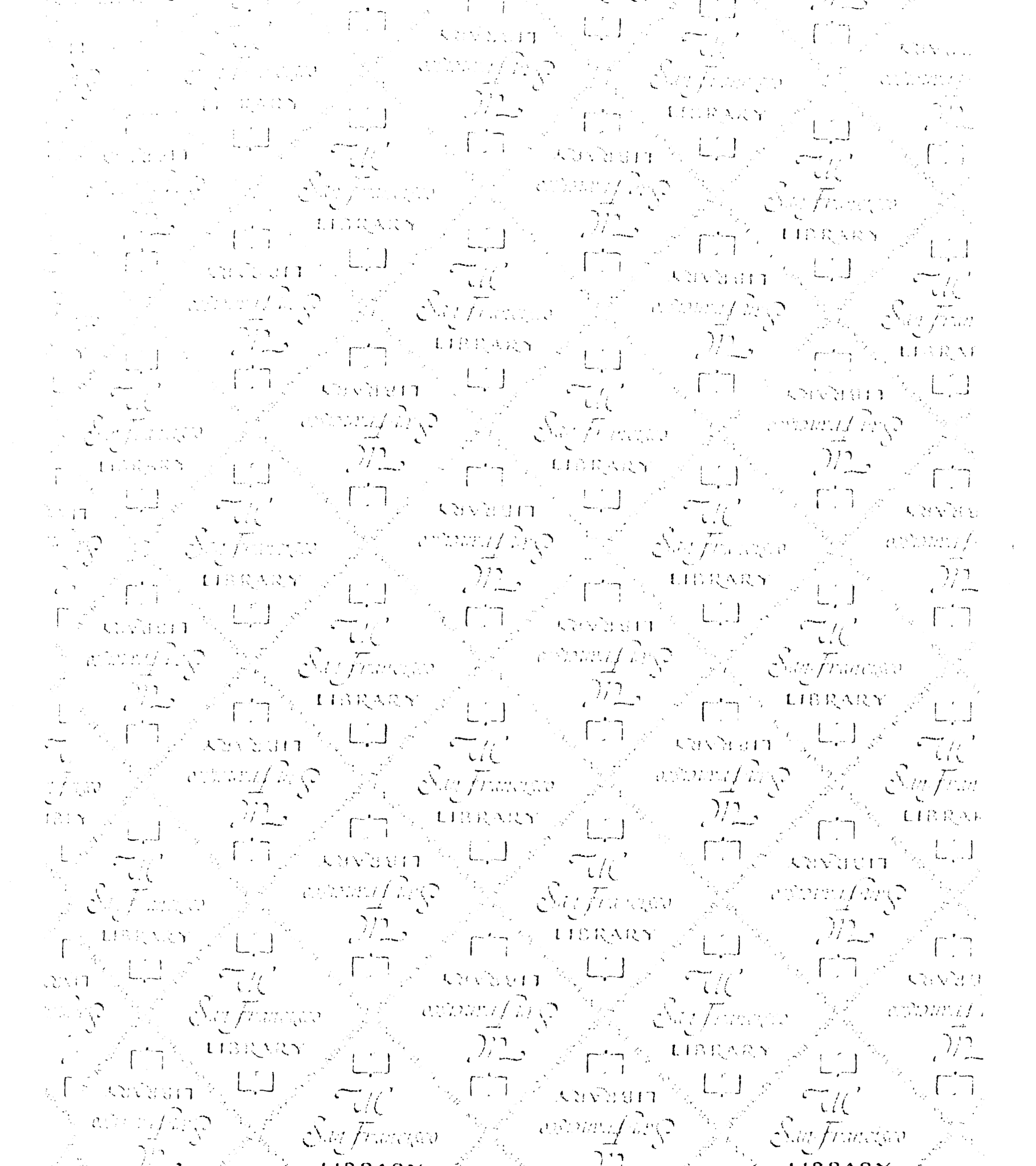
Zhurkin, V. B., Lysov, Y. P., Florent'ev, V. L. & Ivanov, Y. (1982). Torsional flexibility of B-DNA as revealed by conformational analysis. *Nuc. Acids Res.* **10**(5), 1811-1830.

Zhurkin, V. B., Poltev, V. I. & Florent'ev, V. L. (1980). Atom-atom potential functions for conformational calculations of nucleic acids. *Mol. Biol.* **14**, 1116-1130.

Zhurkin, V. B., Ulyanov, N. B., Gorin, A. A. & Jernigan, R. L. (1991). Static and statistical bending of DNA evaluated by Monte Carlo simulations. *Proc. Nat. Acad. Sci.* **88**, 7046-7050.

Zichi, D. A. (1995). Molecular dynamics of RNA with the OPLS force field. Aqueous simulation of a hairpin containing a tetranucleotide loop. *J. Amer. Chem. Soc.* **117**(11), 2957-2969.

Zimmerman, S. B. & Pheiffer, B. H. (1981). A RNA-DNA hybrid that can adopt two conformations: An x-ray diffraction study of poly(rA)-poly(dT) in concentrated solution or in fibers. *Proc. Natl. Acad. Sci.* **78**(1), 78-82.


San Francisco
LIBRARY