Molecular Dynamics Simulations of the 136 Unique Tetranucleotide Sequences of DNA Oligonucleotides. I. Research Design and Results on d(C_pG) Steps

David L. Beveridge,* Gabriela Barreiro,* K. Suzie Byun,* David A. Case,† Thomas E. Cheatham III,‡ Surjit B. Dixit,* Emmanuel Giudice,§¶ Filip Lankas,‖** Richard Lavery,§ John H. Maddocks,** Roman Osman,¶ Eleanore Seibert,¶ Heinz Sklenar,†† Gautier Stoll,** Kelly M. Thayer,* Péter Varnai,§ and Matthew A. Young‡‡

*Chemistry Department, Molecular Biology & Biochemistry Department, and Molecular Biophysics Program, Wesleyan University, Middletown, Connecticut 06459 USA; †Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037 USA; ‡Departments of Medicinal Chemistry and of Pharmaceutics and Pharmaceutical Chemistry, University of Utah, Salt Lake City, Utah 84112-5820 USA; §Laboratoire de Biochimie Theorique, Institut de Biologie Physico-Chimique, Paris 75005, France; ¶Physiology and Biophysics, Mount Sinai School of Medicine, New York, New York 10029 USA; ‖J. Heyrovsky Institute and Center for Complex Molecular Systems and Biomolecules, 182 23 Prague, Czech Republic; **Institute of Mathematics B, Swiss Federal Institute of Technology, CH 1015 Lausanne, Switzerland; ††Theoretical Biophysics Group, Max Delbrück Center, D-13122 Berlin, Germany; and ‡‡Molecular and Cell Biology, University of California-Berkeley, Berkeley California 94720-3202 USA

ABSTRACT We describe herein a computationally intensive project aimed at carrying out molecular dynamics (MD) simulations including water and counterions on B-DNA oligomers containing all 136 unique tetranucleotide base sequences. This initiative was undertaken by an international collaborative effort involving nine research groups, the "Ascona B-DNA Consortium" (ABC). Calculations were carried out on the 136 cases imbedded in 39 DNA oligomers with repeating tetranucleotide sequences, capped on both ends by GC pairs and each having a total length of 15 nucleotide pairs. All MD simulations were carried out using a well-defined protocol, the AMBER suite of programs, and the parm94 force field. Phase I of the ABC project involves a total of ~0.6 μs of simulation for systems containing ~24,000 atoms. The resulting trajectories involve 600,000 coordinate sets and represent ~400 gigabytes of data. In this article, the research design, details of the simulation protocol, informatics issues, and the organization of the results into a web-accessible database are described. Preliminary results from 15-ns MD trajectories are presented for the d(CpG) step in its 10 unique sequence contexts, and issues of stability and convergence, the extent of quasiergodic problems, and the possibility of long-lived conformational substates are discussed.

INTRODUCTION

Basepair sequence effects on structure and dynamics are a key issue in understanding the biochemistry and biology of DNA at the molecular level. Most information on sequence effects to date has been limited to dinucleotide steps. However, recent more extensive considerations of the problem indicate that dinucleotide steps are sensitive to at least nearest-neighbor sequence contexts (Brukner et al., 1995b; Lankas et al., 2003; Packer et al., 2000b; Yanagi et al., 1991), and to even longer-range effects in the case of A-tracts (Burkhoff and Tullius, 1987) or allosteric effects (Kim et al., 1993). The minimum structural unit that reveals nearest-neighbor sequence context effects is the tetranucleotide step, of which there are 136 unique sequence permutations. At present, the experimental structural database of DNA tetranucleotide steps at atomic resolution, derived primarily from x-ray crystallography and emerging results from NMR spectroscopy, is quite sparse. However, the ability to model DNA structure in solution using all-atom molecular dynamics (MD) simulation has improved significantly in recent years (Beveridge and McConnell, 2000; Cheatham and Kollman, 2000; Giudice and Lavery, 2002; Miller et al., 1999; Orozco et al., 2003), and the study of sequence effects has become accessible to high-performance computing. In this article, we describe a project aimed at obtaining MD trajectories including water and counterions for all unique tetranucleotide base sequences. This project involves the participation of 17 investigators from nine independent research laboratories. This research was initiated in a workshop held at Ascona, Switzerland in June, 2001, and the participants in this project are henceforth referred to as the ''Ascona B-DNA Consortium'' (ABC).

Overall, we seek to obtain MD trajectories for the 136 unique DNA tetranucleotides imbedded in 39 DNA oligomers having repeating sequences. The oligomers are each 15 nucleotide pairs in length and are capped on both ends by GC pairs. All MD simulations were performed with a consensus protocol using the AMBER suite of programs (Case et al., 1999) and the parm94 force field of Cornell et al. (1995), which has been verified in test cases to produce good overall agreement between calculated and observed DNA structures in crystals and in solution (Arthanari et al., 2003; Bevan et al., 2000). MD trajectories of 15 ns have been obtained for each of the 39 cases. From these simulations, the MD protocol, convergence and stability issues, and quasiergodic problems due to substate sampling are assessed. In this

Submitted May 20, 2004, and accepted for publication August 3, 2004.

Address reprint requests to David L. Beveridge, E-mail: dbeveridge@wesleyan.edu.

© 2004 by the Biophysical Society

0006-3495/04/12/3799/15 $2.00

doi: 10.1529/biophysj.104.045252

article we present the research design of ABC, details of the simulation protocol, considerations on informatics and database issues, and results on the sequence context problem in d(CpG) steps. The development of a prototype web-accessible relational database for public dissemination of the results is described.

BACKGROUND

The first single crystal structure of the B-form DNA double helix (Wing et al., 1980) raised a number of fundamental questions about basepair sequence effects on structure, solvation, conformational stability, and axis bending (Dickerson and Drew, 1981; Drew and Dickerson, 1981; Drew et al., 1981). The idea that sequence-dependent structural deformations provide an analog code (indirect readout) for protein-DNA recognition that supplements the digital code, embodied in the pattern of noncovalent binding sites in the major and minor groove (direct readout), followed directly from this crystal structure (Dickerson, 1983). Many subsequent studies with implications regarding sequence effects on DNA structure have been carried out over the last 25 years (Neidle, 1999). A major line of investigation of DNA sequence effects on structure has been to try to understand oligomeric DNA structures in terms of sequence subunits. The minimum structural unit that carries information on the three-dimensional structure of DNA is the dinucleotide basepair step, 5′-dXpY-3′ where X and Y may be A, T, G, or C. The four alternatives lead to 16 XpY permutations, of which 10 are unique.

Questions have been raised since early studies of sequence effects as to whether dinucleotide step information is a sufficient basis for a description of sequence effects in oligomeric or polymeric DNA. The structure of any individual XpY step may clearly be subject to sequence context presented by the nearest neighboring basepairs. If such effects are important, the minimum monomeric unit necessary to describe the details of DNA structure would be tetranucleotide steps, of which there are 136 unique permutations. Evidence of higher-order cooperative behavior in DNA structure suggests that in some cases even tetranucleotide steps may be insufficient to fully characterize the system. However, for a systematic approach, the nature of the sequence-effect problem at the level of tetranucleotide steps needs to be fully examined.

The immediate issue is thus first-neighbor context effects on the structures of DNA dinucleotide steps, which requires knowledge of the structures of all 136 unique tetranucleotides. Crystal structures of DNA oligonucleotides serve as the primary source of data, the basis of a number of studies of the problem to date as described in research articles (El Hassan and Calladine, 1995; Olson et al., 1998; Suzuki et al., 1997), review articles (Neidle, 1999; Olson and Zhurkin, 1996), and texts (Calladine and Drew, 1997; Neidle, 2002; Saenger, 1984; Sinden, 1994). Even at the dinucleotide step level, the crystal structures present an uneven distribution of instances of each step, and are heavily biased toward cases with G's and C's. Issues with respect to the influence of packing effects and crystal imperfections have also been noted (Dickerson et al., 1987). In particular, sequence-dependent axis curvature of DNA is clearly sensitive to packing effects (DiGabriele et al., 1989; Johansson et al., 2000; Shakked et al., 1989). The determination of DNA structure in solution by NMR spectroscopy has been limited by the lack of tertiary contacts and the short-range nature of scalar couplings and NOE data. New NMR experiments based on residual dipolar coupling (RDC) hold the possibility of obtaining higher-resolution structures of oligonucleotides in solution (Vermulen et al., 2000) and may have sufficiently high resolution to accurately resolve DNA structure, but are just beginning to appear in the literature (Barbic et al., 2003; MacDonald and Lu, 2002; Tjandra et al., 2000). Another line of investigation has been to derive basepair step parameters empirically or semiempirically from experiment (Bolshoy et al., 1991; Liu and Beveridge, 2001). However, various dinucleotide step models give essentially similar accounts of the observed data within statistical uncertainty (Liu and Beveridge, 2001).

Two sets of structural indices based on trinucleotide steps have been derived from nucleosome positioning and DNase digestion (Brukner et al., 1995a,b; Satchwell et al., 1986). Both sets of results indicate significant context effects for dinucleotide steps, but the rankings do not correlate well with each other and likely index different aspects of sequence-dependent structural deformation and/or deformability. Kanhere and Bansal (2003) have reexamined this issue and indicate that trinucleotide scales do not lead to a good account of all the observed data on sequence-dependent curvature. At the tetranucleotide step level, the crystallographic database is still very sparse. Surveys of this data have raised the possibility of quite significant sequence effects. The most extensive theoretical consideration of the problem to date is due to Packer et al. (2000a,b), who presented detailed considerations based on the minimization of stacking energies for tetranucleotide steps as described by empirical energy functions.

Recently, all-atom molecular modeling of DNA structure via molecular dynamics simulations including explicit solvent (water molecules and mobile salt ions) and based on interactions described by empirical force fields, has reached a level at which accurate dynamical models of DNA structure in solution have been obtained. Various aspects of the field of MD simulations of DNA have been described in recent review articles (Beveridge and McConnell, 2000; Cheatham and Kollman, 2000; Giudice and Lavery, 2002; Norberg and Nilsson, 2002; Orozco et al., 2003). The simulation protocols employed by different groups are now reasonably uniform. The problem of long-range interactions is seemingly stabilized with the advent of the particle mesh Ewald (PME) method (Essmann et al., 1995) for periodic

boundary conditions, despite lingering concerns about long-range correlations (Hunenberger and McCammon, 1999; Smith and Pettitt, 1996). The energy functions incorporated in the suites of programs readily available for MD simulation such as AMBER (Case et al., 1999), CHARMM (Brooks et al., 1983), and GROMOS (Scott et al., 1999) each contain a full set of parameters for nucleic acids.

The AMBER parm94 nucleic acids force field as described by Cornell et al. (1995) is a reparameterization for MD with explicit solvent, and is termed ''second generation''. MD using AMBER and parm94 provided the first well-behaved MD trajectories of the DNA double helix (Cheatham et al., 1995, 1998; York et al., 1995; Young et al. 1997a,b). Known shortcomings in parm94 still include a sensitive problem in the coupling of base-sugar torsions and a systematic tendency toward slightly underwound structures. A modification known as parm99 has recently been proposed (Cheatham et al., 1999), which improves twist but appears less sensitive to changes in the environment (high salt, ethanol). The CHARMM force field for nucleic acids, as refined by MacKerell and co-workers (Foloppe and MacKerell, 2000; MacKerell and Banavali, 2000; MacKerell et al., 2000); and also the hybrid AMBER/CHARMM force field by Langley (1996, 1998), provide viable alternatives for MD on nucleic acids and also show good agreement with experiment. Comparative studies on force fields for nucleic acids have been described by Feig and Pettit (1998), Reddy et al. (2003), and Cheatham and Young (2001).

Although the present study is aimed at creating a well-defined computational vantage point on the problem of sequence effects on DNA structure, the project design allows us to address several important additional and timely methodological questions about MD on DNA oligonucleotides. Some of the principle concerns are: a), when is a simulation ''converged''; b), what length of trajectory is ''enough''; c), how sensitive are the results to the choice of initial configuration; and d), what are the meaningful ways and pitfalls in extracting ''structures'' from an MD trajectory and analyzing them? Questions a and b are in fact ''moving targets'' with no definitive answer. Convergence can never be unequivocally proved because there is no guarantee that the past behavior of a system in a simulation is predictive of the future; one may in principle always encounter new substates of a system with more extensive sampling or new modes of motion that have a slower relaxation time. One must deal with this pragmatically, running simulations for as long as possible and checking on the stability of diverse indices of dynamical structure as a function of time. Each property or structural index exhibits a characteristic time evolution in MD, and some have a shorter relaxation time and will be quicker to stabilize than others. Studies on a prototype B-form dodecamer (Ponomarev et al., 2004) indicate that DNA conformational and helicoidal parameters, have relaxation times of $<$500 ps. The rule of thumb is to sample for 10 times the relaxation time of all the indices of interest for a particular application (Haile, 1992). This indicates that 5-ns trajectories should be sufficient in the absence of substate problems (see below), and we are well in excess of that in the 15-ns trajectories carried out in phase I of this project. Observed diffusion constants indicate that motions of mobile counterions will be relatively slow to converge (Varnai and Zakrzewska, 2004). Ponomarev et al. (2004) have found in a prototype study that ion occupancies may take up to 100 ns to stabilize. However, in the same calculation, the DNA parameters were found to be well stabilized at 5 ns, and not sensitive to the fine details of ion convergence. The calculated DNA counterion radial distribution functions were found to be essentially unchanged after 3–5 ns, indicating that mean field effects of ions are dominant in DNA structure and that the excess sampling to get ion occupancies converged is a matter of granularity of the ion distributions.

What is referred to as the ''substate problem'' in macromolecular simulation is a quasiergodic issue. A flexible macromolecule has the potential for contributions from a manifold of thermally accessible substates, with DNA being particularly susceptible (McConnell et al., 1994; Poncin et al., 1992). Known examples of this are the BI-BII transitions (Hartmann et al., 1993) and α/γ-crankshaft motions (Varnai et al., 2002), and YpR hinge motions (Calladine and Drew, 1997). The latter have been noted to play an important role in structures of protein-bound DNA (Dickerson and Chiu, 1997) as well as DNA curvature (Beveridge et al., 2004). Indications from the crystallographic database are that certain basepair steps show high flexibility (El Hassan and Calladine, 1995) whereas those involved in A-tracts are relatively rigid (Young et al., 1995). The problem this poses to a simulation arises from the need to sample all thermally accessible substates adequately to obtain an ensemble of snapshots that properly represent the dynamical structure of the DNA. This requires additional sampling, which is numerically impeded when the paths between substates are narrow cols on the potential energy hypersurface and thus infrequent occurrences. However in examining this class of problems, computational modeling via MD has a unique vantage point, because a molecular level account of structure as a function of time can probably never be obtained experimentally. The substate issue calls attention to another problem in defining structure, because for a system with substates the idea of a single overall average structure of the system being representative of the dynamical structure of the system is challenged. For example for a system in a symmetric double minimum potential in which both states are thermally accessible, the average structure would have the least probability of occurrence in the ensemble. In this case the analysis should be based on the structures of substates and their respective statistical weights, i.e., the dynamical structure of the system.

In this article, the research design, details of the simulation protocol, informatics issues, and the organization of the

results into a relational database are described. Preliminary results concerning MD convergence issues and structural analyses after 15 ns of MD are presented, focusing on the d(CpG) step in its 10 unique sequence contexts. The d(CpG) step was chosen for preliminary analysis as a case in which x-ray structures indicate a potential for context-dependent substates (Calladine and Drew, 1997; El Hassan and Calladine, 1995). The extent to which any step has the potential for substates of any kind with differential stability sensitive to context effects will be an important issue in understanding DNA curvature and ligand-induced bending with substantial implications with respect to protein-DNA recognition.

METHODOLOGY

MD simulation is a computer ''experiment'' in which the atoms of a postulated system execute Newtonian dynamics on an assumed potential energy surface. The MD procedures specific for biological macromolecules have been well described by McCammon and Harvey (1986), Leach (1996), and Schlick (2002). The model system chosen for this study, the assumed potential energy surface (i.e., force field), and the simulation protocol are all operational variables in the calculation. An MD simulation on a DNA oligonucleotide begins with the choice of an initial configuration and an arbitrary arrangement of solvent (in this case, water) molecules and counterions. The initial configuration of the system is then subjected to energy minimization to relieve any major stresses, followed by a period in which the particle velocities (heating) are increased to reach the temperature of interest. The MD simulation then proceeds via Newtonian dynamics to locate a thermally bounded state of interest (equilibration) and subsequently to sample it (production). Analysis of the results is then based on the ensemble of structures that comprise the production segment of the simulation, providing sampling is sufficiently long and assuming approximate ergodicity within a Boltzmann distribution. What is ''sufficient'' for DNA simulations is one of the major problems addressed in this study.

All simulations have been carried out using the AMBER 6 or AMBER 7 suite of programs (Case et al., 1999) and the parm94 force field (Cornell et al., 1995). The simulations cover 39 double-stranded DNA oligomers, each being 15 basepairs in length. The sequences of these oligomers are discussed below. A consensus protocol was adopted for simulation in which the solute molecule is a 15-basepair oligonucleotide with 28 potassium ions added to achieve system electroneutrality. The DNA-ion complex is simulated in a truncated octahedral box having a face-to-face dimension of ~70 Å, which allows for a solvent shell extending for at least 10 Å around the DNA. The starting configuration has the oligomer in a canonical B form. The ions are randomly placed around the oligomer, and located at least 5 Å from any atom of the solute and at least 3.5 Å from one another in the initial structure. Ion interaction with other atoms are based on the potentials developed by Aqvist (1990). The neutral ion-oligomer complex is solvated with a layer of TIP3P water molecules (Jorgensen, 1981). Simulations are performed with periodic boundary conditions in which the central cell box contains ~8000 water molecules. Considering the DNA, counterions, and solvent water, the total system consists of ~24,000 atoms.

The preparations for MD simulations consists of an initial minimization followed by slow heating to 300 K at constant volume over a period of 100 ps using harmonic restraints of 25 kcal/mol/Å^2 on the solute atoms. These restraints are slowly relaxed from 5 to 1 kcal/mol/Å^2 during a series of five segments of 1000 steps of energy minimization and 50-ps equilibration using constant temperature (300 K) and pressure (1 bar) conditions via the Berendsen algorithm (Berendsen et al., 1984) with a coupling constant of 0.2 ps for both parameters. The final segments consists of 50-ps equilibration with a restraint of 0.5 kcal/mol/Å^2 and 50-ps unrestrained equilibration. The simulations were then continued for a total of 15 ns at constant temperature and pressure conditions, using the Berendsen algorithm (Berendsen et al., 1984) with a coupling constant of 5 ps for both parameters. Electrostatic interactions were treated using the particle mesh Ewald (PME) algorithm

TABLE 1 One-hundred thirty-six unique tetranucleotides (upper case), divided into 10 groups on the basis of their central dinucleotide step

GG	G	A	C	T
G	GGGG	GGGA	GGGC	GGGT
A	AGGG	AGGA	AGGC	AGGT
C	CGGG	CGGA	CGGC	CGGT
T	TGGG	TGGA	TGGC	TGGT
AA	G	A	C	T
G	GAAG	GAAA	GAAC	GAAT
A	AAAG	AAAA	AAAC	AAAT
C	CAAG	CAAA	CAAC	CAAT
T	TAAG	TAAA	TAAC	TAAT
GA	G	A	C	T
G	GGAG	GGAA	GGAC	GGAT
A	AGAG	AGAA	AGAC	AGAT
C	CGAG	CGAA	CGAC	CGAT
T	TGAG	TGAA	TGAC	TGAT
AG	G	A	C	T
G	GAGG	GAGA	GAGC	GAGT
A	AAGG	AAGA	AAGC	AAGT
C	CAGG	CAGA	CAGC	CAGT
T	TAGG	TAGA	TAGC	TAGT
GT	G	A	C	T
G	GGTG	GGTA	GGTC	GGTT
A	AGTG	AGTA	AGTC	AGTT
C	CGTG	CGTA	CGTC	CGTT
T	TGTG	TGTA	TGTC	TGTT
TG	G	A	C	T
G	GTGG	GTGA	GTGC	GTGT
A	ATGG	ATGA	ATGC	ATGT
C	CTGG	CTGA	CTGC	CTGT
T	TTGG	TTGA	TTGC	TTGT
GC	G	A	C	T
G	GGCC	GGCA	GGCC	ggct
A	AGCG	AGCA	AGCC	AGCT
C	CGCG	CGCA	cgcc	cgct
T	tgcg	TGCA	tgcc	tgct
CG	G	A	C	T
G	GCGG	GCGA	GCGC	gcgt
A	ACGG	ACGA	ACGC	ACGT
C	CCGG	CCGA	ccgc	ccgt
T	tcgg	TCGA	tcgc	tcgt
AT	G	A	C	T
G	GATG	GATA	GATC	gatt
A	AATG	AATA	AATC	AATT
C	CATG	CATA	catc	catt
T	tatg	TATA	tatc	tatt
TA	G	A	C	T
G	GTAG	GTAA	GTAC	gtat
A	ATAG	ATAA	ATAC	ATAT
C	CTAG	CTAA	ctac	ctat
T	ttag	TTAA	ttac	ttat

Lower-case entries correspond to redundant tetranucleotides whose complementary sequences are already present in the table.

(Essmann et al., 1995) with a real space cutoff of 9 Å, cubic B-spline interpolation onto the charge grid with a spacing of ~1 Å. SHAKE constraints (Ryckaert et al., 1977) were applied to all bonds involving hydrogen atoms. The integration time step was 2 fs. Center-of-mass translational motion was removed every 5000 MD steps to avoid the methodological problems described by Harvey et al. (1998). The trajectories were extended, as noted above, to 15 ns for each oligomer and conformations of the system were saved every 1 ps for further analysis.

In an effort to be objective about the convergence of MD simulations on DNA, it was agreed that first generation ABC study would involve 15 ns of simulation for each of the 39 oligomers, pooling the results and performing detailed analysis. This is at the high end of typical run lengths used in current published research. Each group was assigned responsibility for dealing with MD on four or five oligomers. Although no special resources were requested for carrying out these simulations, the entire data set was obtained in roughly three months on a heterogeneous mix of high-performance supercomputers and PC clusters. It should be stressed that this represents a considerable computational task, corresponding to a total of ~0.6 μs of simulation for systems containing ~24,000 atoms. The resulting trajectories involve 600,000 coordinate sets and represent roughly 400 gigabytes of data.

Oligonucleotide sequences

A key element of our research design is that, rather than performing calculations on all 136 tetranucleotides using 136 different oligomers (for example, placing each tetranucleotide within a longer duplex, surrounded with some standard sequence), we carried out the calculations on oligomers with repeating tetranucleotide sequences (ABCDABCDABCD...); cf. Table 1. In this way, each oligomer can contain up to four distinct tetranucleotides. Thus moving a four-base ''reading frame'' along the oligomer, we locate successively ABCD, BCDA, CDAB, and DABC tetranucleotides. As shown in Table 2, this strategy enables all 136 tetranucleotides to be studied using only 39 oligomers. As concerns the length of the oligomers, 15 basepairs was chosen as a compromise between the necessity to avoid end effects and the computational expense of the simulations. Based on prior experience, it was also decided to cap the ends of each oligomer with a single GC pair to avoid fraying. This implies that a given 15-basepair oligomer contains three tetranucleotide repeats 5′-G-D-ABCD-ABCD-ABCD-G-3′. This choice means that if we decide to ignore two basepairs at either end of the oligomer, to avoid potential artifacts from end effects, there will still be two distinct copies of each unique

TABLE 2 Thirty-nine repeating-sequence oligonucleotides containing the 136 unique tetranucleotides

Oligonucleotide	1st tetranucleotide	2nd tetranucleotide	3rd tetranucleotide	4th tetranucleotide
GGGGGGGGGGGGG	GGGG	–	–	–
AAAAAAAAAAAAA	AAAA	–	–	–
CGCGCGCGCGCGC	GCGC	CGCG	–	–
TATATATATATAT	ATAT	TATA	–	–
AGAGAGAGAGAGA	GAGA	AGAG	–	–
TGTGTGTGTGTGT	GTGT	TGTG	–	–
AGGGAGGGAGGGA	GGGA	GGAG	GAGG	AGGG
CGGGCGGGCGGGC	GGGC	GGCG	GCGG	CGGG
TGGGTGGGTGGGT	GGGT	GGTG	GTGG	TGGG
GAAAGAAAGAAAG	AAAG	AAGA	AGAA	GAAA
CAAACAAACAAAC	AAAC	aaca/TGTT	acaa/TTGT	CAAA
TAAATAAATAAAT	AAAT	AATA	ATAA	TAAA
CGGCCGGCCGGCC	GGCC	–	CCGG	CGGC
AGGAAGGAAGGAA	GGAA	GAAG	AAGG	AGGA
TGGTTGGTTGGTT	GGTT	gttg/CAAC	TTGG	AGGA
TAATTAATTAATT	AATT	–	TTAA	TAAT
CGGACGGACGGAC	GGAC	gacg/CGTC	ACGG	CGGA
AGGCAGGCAGGCA	GGCA	gcag/CTGC	CAGG	AGGC
AGGTAGGTAGGTA	GGTA	GTAG	TAGG	AGGT
TGGATGGATGGAT	GGAT	GATG	ATGG	TGGA
CGGTCGGTCGGTC	GGTC	gtcg/CGAC	tcgg/CCGA	CGGT
TGGCTGGCTGGCT	ggct/AGCC	gctg/CAGC	CTGG	TGGC
CAAGCAAGCAAGC	AAGC	AGCA	gcaa/TTGC	CAAG
GAACGAACGAACG	aacg/CGTT	ACGA	CGAA	GAAC
TAACTAACTAACT	aact/AGTT	acta/TAGT	CTAA	TAAC
CAATCAATCAATC	AATC	atca/TGAT	tcaa/TTGA	CAAT
TAAGTAAGTAAGT	AAGT	AGTA	GTAA	TAAG
GAATGAATGAATG	AATG	ATGA	TGAA	GAAT
TGAGTGAGTGAGT	GAGT	AGTG	GTGA	TGAG
CGAGCGAGCGAGC	GAGC	AGCG	GCGA	CGAG
TGCGTGCGTGCGT	gcgt/ACGC	CGTG	GTGC	tgcg/CGCA
TAGATAGATAGAT	AGAT	GATA	ATAG	TAGA
GACAGACAGACAG	acag/CTGT	CAGA	AGAC	gaca/TGTC
TACATACATACAT	acat/ATGT	CATA	ATAC	taca/TGTA
AGCTAGCTAGCTA	gcta/TAGC	CTAG	–	AGCT
TGCATGCATGCAT	gcat/ATGC	CATG	–	TGCA
CGATCGATCGATC	GATC	atcg/CGAT	TCGA	–
TGACTGACTGACT	gact/AGTC	actg/CAGT	CTGA	TGAC
CGTACGTACGTAC	GTAC	tacg/CGTA	ACGT	–

The flanking dinucleotide caps have been removed. The upper- and lower-case conventions are as in Table 1.

tetranucleotide (ABCD, BCDA, CDAB, DABC) within the remaining 11-basepair fragment. Thus with MD on 39 of these oligonucleotides, we will be able to compare the properties of two copies of each tetranucleotide as a further convergence test of the simulations. To ensure that all participating groups were using strictly identical simulation protocols, standard scripts were made available via the group website. A standard naming convention was also adopted for all files generated during the simulations, to facilitate the exchange of data between the participating groups and to simplify setting up an overall database of the results.

Informatics and the ABC database

The net file size of the trajectories generated by the ABC simulations are in the range of several hundreds of gigabytes of data, making the distribution and handling of these files a difficult informatics task. A two-tier approach is being adopted to handle this data dissemination task. Apart from the data comprising the Cartesian coordinates of the molecular trajectory, the set of intra- and interbasepair helicoidal parameters, together with the conformational parameters of the sugar and phosphate backbone of DNA as calculated by CURVES (Lavery and Sklenar, 1996) provides a smaller but complete set of descriptors to define the fine structural details of the nucleic acid segments in each of the frames in these trajectories. A comparison of the various methods for calculating DNA structural parameters has been recently provided by Lu and Olson (Lu et al., 1999; Lu and Olson, 1999).

In the course of this project, a relational database was developed that simplifies and speeds up the task of querying this common repository for trajectory information about any of the simulations, or for comparing the parameters from the different simulations for characteristics of various subsets of the nucleic acid segments. The information extracted from the CURVES analyses are stored in the database as tables for the complete trajectory indexed with various identifications, defined on the basis of the simulation, the nucleotide position, the time step in the simulation, etc. Using the processing power of a structured query language (SQL) allows complex queries into the database. Thus a component of this project is a bioinformatics initiative aimed at the structured storage and handling of the results from large and numerous molecular simulations, which simplifies many of the technical complexities associated with the management and analysis of such large quantities of information.

The ABC database will be made accessible to the interested research community outside the consortium through the internet as soon as it is complete and fully tested. This interface provides a dynamic access to the simulation results harnessing the strength of SQL, limited only by the html interface. This system will permit queries executed from the web to extract the average helicoidal parameter values and their standard deviations over different time periods of the various trajectories, and enable comparisons between the various simulations on different sequences or in different user-defined conditions. The results can be either viewed as tables or displayed graphically in the web browser. The nature of queries that can be carried out from the web interface include viewing the mean and standard deviation profile of all the helicoidal parameters as a function of sequence in each of the trajectories and more elaborate searches such as comparing the statistical properties of any of the DNA structural parameter for a central basepair in all its relevant tetranucleotide combinations.

RESULTS AND DISCUSSION

The results described below were extracted from 15-mer DNA sequences composing the first round of ABC trajectories. We limit this preliminary discussion to results concerning the CpG step and its variability as a function of the flanking bases. Because this step exhibits inversion symmetry, 10 sets of flanking bases cover all the possible choices for the nearest neighbors:

RCGR: GCGG (≡CCGC), ACGG (≡CCGT),
ACGA (≡TCGT), GCGA (≡TCGC)
RCGY: GCGC, ACGC (≡GCGT), ACGT
YCGR: CCGG, CCGA (≡TCGG), TCGA

grouping the various CpG steps into three classes depending on whether they are flanked by purines (R) or pyrimidines (Y). There are four members of the RCGR class (which, by inversion symmetry, also includes the YCGY tetranucleotides) and three members of the RCGY and YCGR classes. These tetranucleotides can be found within 10 of the 39 oligomers studied by ABC.

By design, each of the ABC oligomers contains a minimum of two copies of each unique tetranucleotide, placed at least two basepairs from the ends of the oligomer. This gives us the chance of making internal comparisons of the structural and dynamic characteristics. It should be noted that these two copies are not necessarily symmetrically placed with respect to the center of the oligomer (e.g., G-CGTACG**T**ACGTAC-G, where the bold **T** marks the central basepair and the two copies of the ACGT tetranucleotides are underlined).

We begin by looking at the overall stability of the conformation of the oligomers used in this study. The typical low-resolution test is to examine the evolution of the structure during the trajectory by measuring the root-mean-square deviation (RMSD) with respect to a canonical B-form initial structure, an A-form reference state, or to the average structure reached at the end of the trajectory. Fig. 1 illustrates

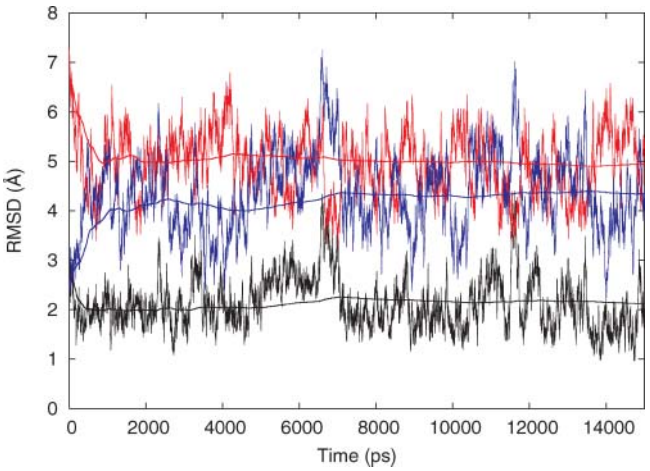

FIGURE 1 Evolution of the RMSD (Å; calculated for nonhydrogen atoms only) between the instantaneous conformation of the oligomer ACGT, a canonical B-DNA conformation (*blue*), a canonical A-DNA conformation (*red*), and the average conformation calculated from the last nanosecond of the trajectory (*black*). The smooth curves denote the corresponding running averages.

this test for the oligomer containing the ACGT tetranucleotide cited in the previous paragraph (hereafter referred to as the ''ACGT oligomer''). Comparisons with B-DNA, A-DNA, and the average conformation of the oligomer calculated over the last 5 ns of the trajectory all suggest that the overall structure of the oligomer rapidly stabilizes to a putatively stable state, and remains there, in RMSD fluctuations of ~1 Å, until the termination of the trajectory. This conformation is seen to lie at ~4 Å from the canonical B conformation and ~5 Å from canonical A-DNA. This global conformation is situated between the B and A forms as evidenced by high roll, negative slide, and lower rise and twist values than in canonical B-DNA. Examination of the helical and the backbone parameters provides more details about the dynamical structure. We will begin with the helical parameters and, in particular, with the interbasepair parameters relating to the CpG step. Fig. 2 shows time series for two important parameters, basepair rise and twist of the $C_{10}pG_{11}$ step within the ACGT oligomer. In each case, there are important oscillations on timescales ranging from a few picoseconds to several nanoseconds. Instantaneous values of both rise (1.6 Å → 5.8 Å) and twist (8° → 52°) cover ranges that are considerably greater than those seen in the crystallographic structures of B-DNA oligomers, even if CpG steps were originally classified as particularly flexible on the basis of such experimental data. The average values measured over the last 10 ns of the trajectory are 3.5 Å for rise and 32.6° for twist, with mean ± SD of 0.6 Å and 5.6°, respectively. Again, as in the case of the global helical conformation, stability as a function of time seems to have been achieved.

As a further test of convergence, we examine whether the two ACGT steps within the ACGT oligomer (i.e., A-C_6pG_7-T and A-$C_{10}pG_{11}$-T (which both satisfy the condition of being at least two basepairs from the ends of the oligomer) behave in a similar way during the trajectory. In Fig. 3, the six interbasepair parameters (shift, slide, rise, roll, tilt, twist) for these two steps are plotted for the last 10 ns of the trajectory. Plots of rise, tilt, and roll are virtually identical for the two steps. The plots of shift, slide, and twist indicate only relatively minor differences in form and in average values, with the differences in the latter being limited to the order of 0.2 Å for the translational parameters and to roughly 1° in twist. It is important to note that some plots show deviations from a normal distribution. Deviations from the normal distribution is diagnostic of either an underlying potential surface that is anharmonic with respect to this motion, or that the results reflect a superposition of thermally accessible

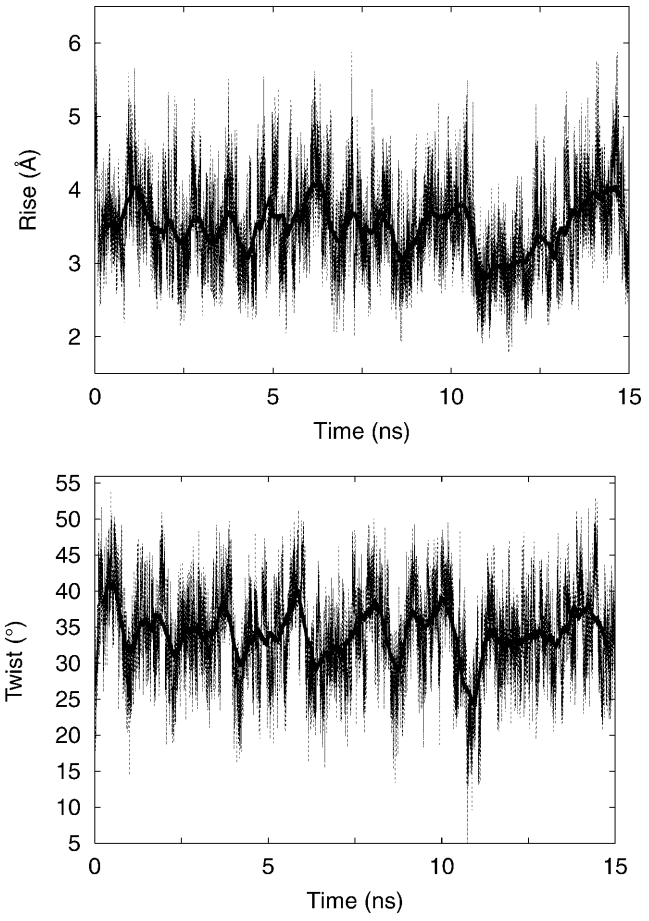

FIGURE 2 Time series for (*top*) rise (Å) and (*bottom*) twist (°) for the $C_{10}pG_{11}$ step within the ACGT oligomer. The thick line shows the same data smoothed using a 100-ps-wide sliding window.

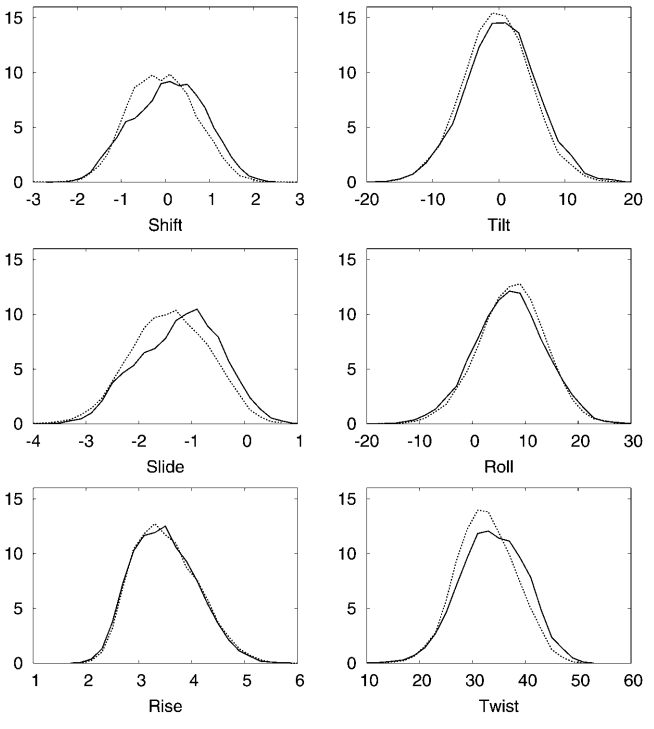

FIGURE 3 Frequency distributions of the six interbasepair parameters for the CpG steps within the ACGT oligomer; C_6pG_7 (*solid*) and $C_{10}pG_{11}$ (*dotted*). Values have been accumulated over the last 10 ns of the trajectory. Translations are given in angstroms and rotations are in degrees.

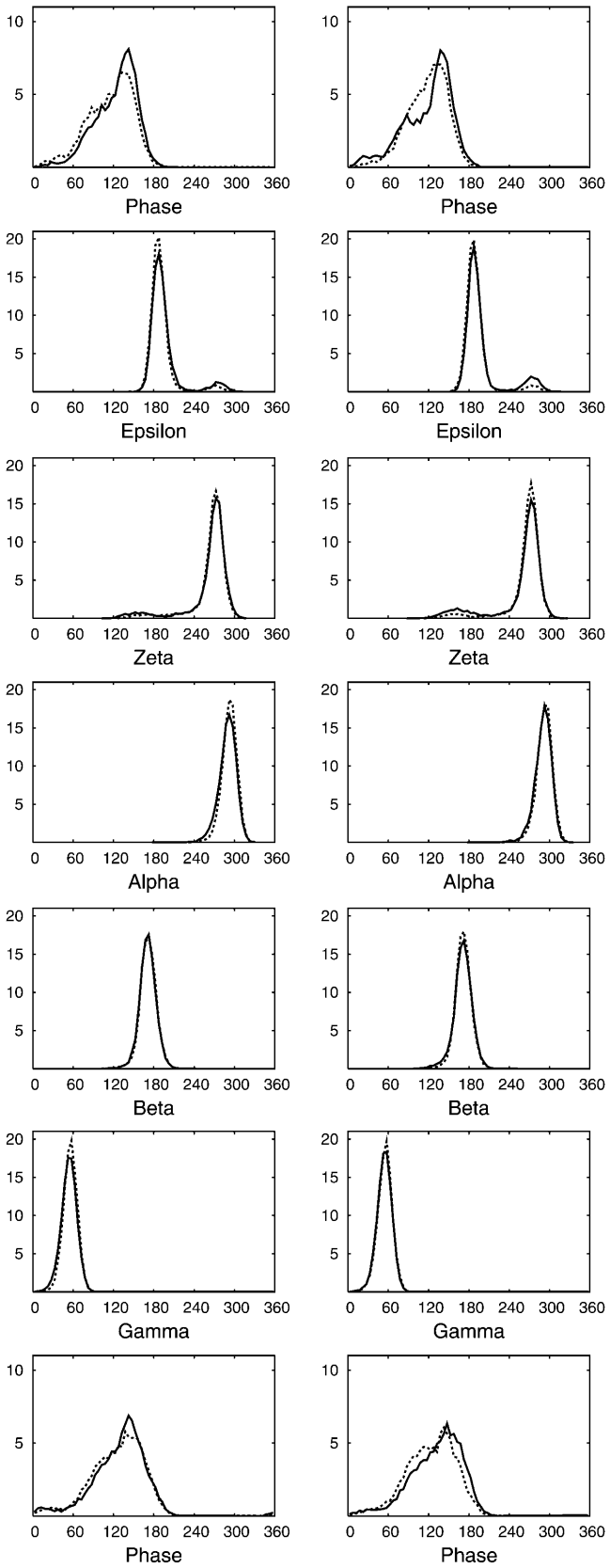

FIGURE 4 Frequency distribution of sugar puckers and the five internucleotide backbone angles for the two backbones of the CpG steps within the ACGT oligomer. (*Left*) First strand, C_6pG_7 (*solid*) and $C_{10}pG_{11}$ (*dotted*); (*right*) second strand, $C_{24}pG_{25}$ (*solid*) and $C_{20}pG_{21}$ (*dotted*). Parameters are ordered in the $5' \rightarrow 3'$ direction for each backbone. Values have been accumulated over the last 10 ns of the trajectory. All values are in degrees.

substates; either could be influenced by context effects. As noted above, there is precedent for the idea that substates corresponding to the open-hinge state (positive roll, negative slide) and closed-hinge (roll, slide both close to zero) for the basepair step exist. The analysis of slide in the 83 cases of CpG steps available in the database of crystal structures with resolution lower than 2.6 Å reveals a clear two-state distribution, a low-slide state close to −2 Å and a much more densely populated high-slide state ~0.5 Å (Packer et al., 2000b). Of the 83 cases, <15 of these CpG are flanked on the 5′ and 3′ ends, the rest being from terminal basepair steps in the DNA sequence (El Hassan and Calladine, 1996). The distribution of roll values in Fig. 3 is fairly broad and has at least a hint of asymmetry. Proceeding with the analysis of the conformational dynamics of ACGT, we examine the sugar-phosphate backbone parameters and sugar puckering calculated for the CpG steps. These data are shown in Fig. 4 for both strands of the C_6pG_7 and $C_{10}pG_{11}$ steps. The results again suggest that the two ACGT steps within the oligomer behave in a very similar way. Note that although most backbone dihedrals show single peaks in their probability distributions, ε (C3′-O3′) and ζ (O3′-P) have secondary populations in g^- and t states, respectively, within both backbones, corresponding to a small percentage of time spent in the B_{II} state.

The next question to be asked is whether the helicoidal and conformational parameters for other CpG steps behave in the same way as those in the ACGT oligomer. As an example, we have chosen the GCGC steps (within the perfectly alternating G-CGCGCGCGCGCGC-G oligomer, hereafter termed simply the ''GCGC oligomer''). Fig. 5 shows histograms for the interbasepair parameters of the C_6pG_7 and the $C_{10}pG_{11}$ steps analogous to those for the ACGT oligomer displayed in Fig. 3. In this instance, because of the regularly alternating dinucleotide sequence of the GCGC oligomer, we can actually extract information on two other GCGC tetranucleotides, namely those centered on the C_4pG_5 and C_8pG_9 steps. Note that both these tetranucleotides also satisfy our criteria that they should be placed at least two basepairs away from the ends of the oligomer. The results in Fig. 5 show significant differences with respect to those in Fig. 3. Although tilt and roll show almost identical histograms for all the CpG steps, the other interbasepair parameters show sharp disparities. Although the C_6pG_7 and $C_{10}pG_{11}$ steps behave like one another, they are generally centered at very different values than the C_4pG_5 and C_8pG_9 steps. In fact, it is the former pair of steps that occur at unusual values for B-DNA. This is particularly clear from the rise and twist histograms that show that the C_6pG_7 and

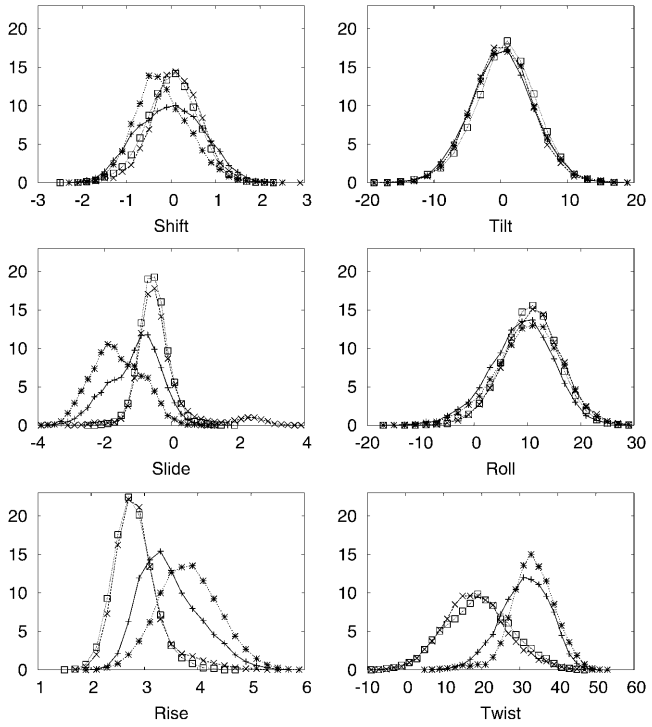

FIGURE 5 Frequency distributions of the six interbasepair parameters for the CpG steps within the GCGC oligomer. C_4pG_5 (+), C_6pG_7 (×), C_8pG_9 (*), and $C_{10}pG_{11}$ (□). Values have been accumulated over the last 10 ns of the trajectory. Translations are given in angstroms and rotations are in degrees.

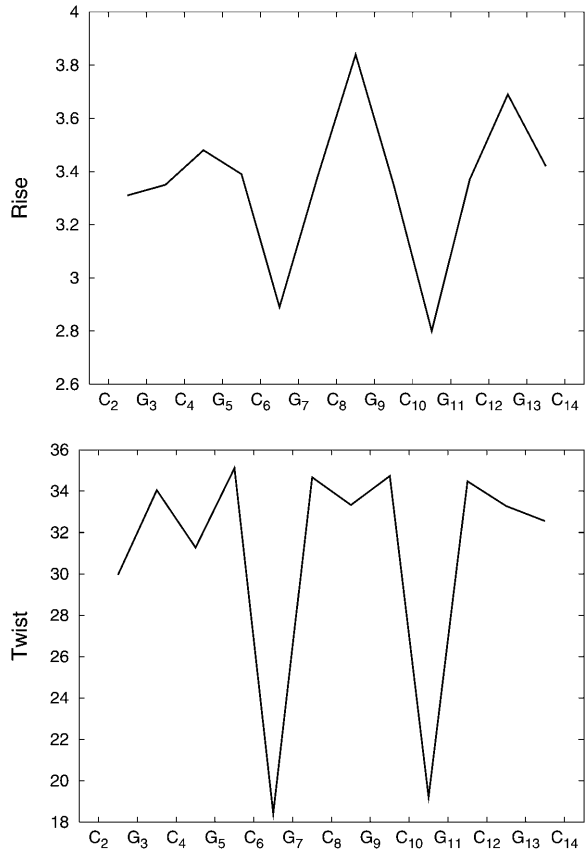

FIGURE 6 Average values of (*top*) rise (Å) and (*bottom*) twist (°) for the CpG step, measured over the last 10 ns of the trajectory of the GCGC oligomer.

$C_{10}pG_{11}$ steps are both compressed and strongly underwound. This difference becomes even more striking if we plot the average values of rise and twist along the GCGC oligomer (see Fig. 6). This figure also suggests that the unusual C_6pG_7 and $C_{10}pG_{11}$ steps, which are placed almost symmetrically with respect to the center of the GCGC oligomer, have a significant effect on the surrounding steps. This effect can be seen clearly in Fig. 5, as differences in the translational parameters for the C_4pG_5 and the C_8pG_9 steps (shown with *plus sign* and *asterisk*, respectively), the first of which lies near the end of the oligomer, whereas the second lies between the two perturbed steps. To summarize, we see that nominally equivalent CpG steps within the GCGC oligomer can exhibit average values of structural parameters that vary by ~1 Å in translation and ~20° in twist. These differences are an order of magnitude greater than those seen within the ACGT oligomer. How can this striking difference be explained?

The answer turns out to lie in the backbone geometry, and not in the linkages of the CpG steps we have analyzed. This is confirmed by the results in Fig. 7 that show histograms of the sugar puckers and phosphodiester dihedrals of both strands of the C_6pG_7 and $C_{10}pG_{11}$ steps. These values are not only similar for both steps, but also close to the values for the CpG steps of the ACGT oligomer shown in Fig. 4, despite the changes in helical parameters we have just described. The only visible change involves somewhat narrower peaks for the sugar-pucker distributions of the 3′ guanosines. The important changes actually occur in the 3′ neighboring GpC steps in the first strand (G_7pC_8 and $G_{11}pC_{12}$). As shown in Fig. 8, both the α (P-O5′) and γ (C5′-C4′) dihedral distributions are unusual, with α spending most of the trajectory in the g^+ state, rather than the usual g^- state and γ being *t* rather than the usual g^+. These changes also affect the other backbone angles of the junction, leading to a shift of β (O5′-C5′) from *t* toward g^- and a broad ζ distribution centered around 110°. Note that there are no such changes in the 3′-GpC steps of the second strand ($G_{25}pC_{26}$ and $G_{21}pC_{22}$) of the oligomer, although there is a shift in the $\varepsilon\zeta$ equilibrium to almost equally populated B_I and B_{II} states and a broadening of the β-distribution.

Analysis of time series of the relevant variables for the $C_{10}pG_{11}pC_{12}$ fragment of the backbone, Fig. 9, confirms that an $\alpha\gamma$-flip from the normal g^-g^+ state to a g^+t state indeed affects the rise and twist of the adjacent 5′ step. The results indicate that the $\alpha\gamma$-flip on the 3′ side of $C_{10}pG_{11}$ occurs after almost 4 ns of simulation and that the rise and twist of this step then drops sharply after a delay of roughly 1 ns. Similar coupling has been observed for other CpG steps that

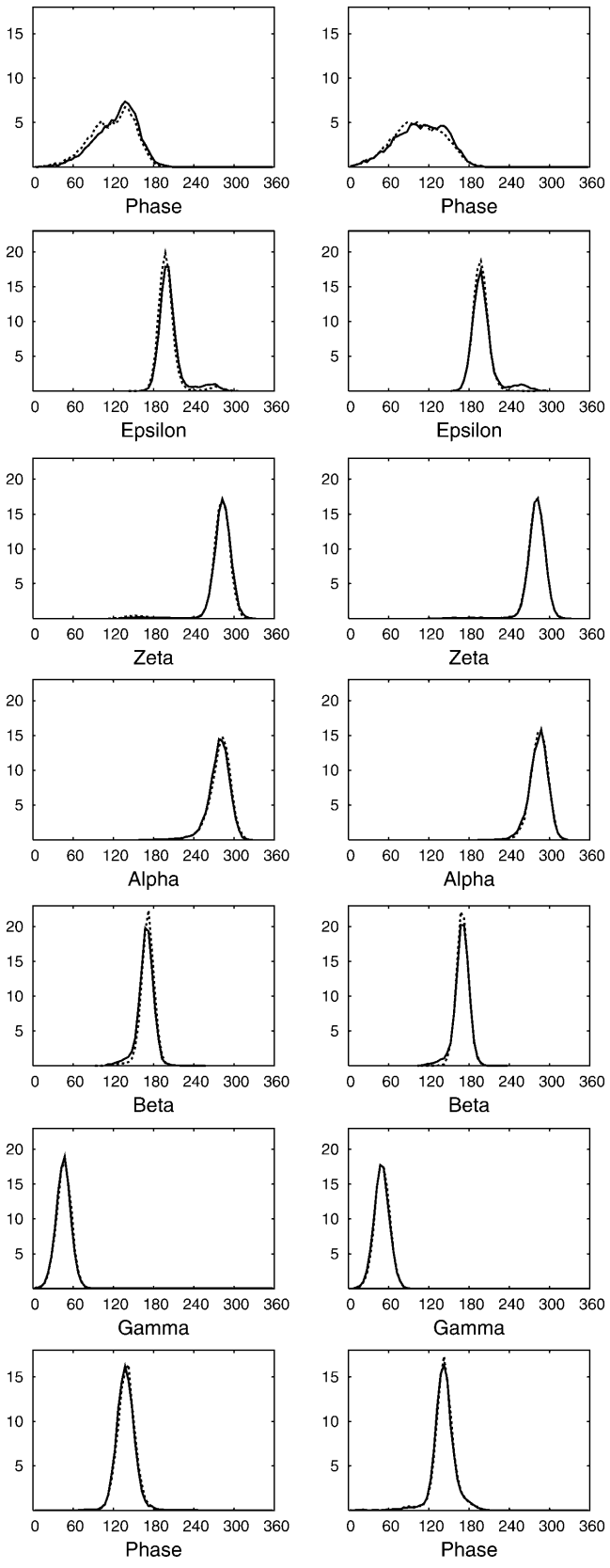

FIGURE 7 Frequency distributions of the sugar puckers and the five internucleotide backbone angles for the two backbones of the CpG steps within the GCGC oligomer. (*Left*) First strand, C_6pG_7 (*solid*) and $C_{10}pG_{11}$ (*dotted*); (*right*) second strand, $C_{24}pG_{25}$ (*solid*) and $C_{20}pG_{21}$ (*dotted*). Parameters are ordered in the $5' \rightarrow 3'$ direction of each backbone. Values have been accumulated over the last 10 ns of the trajectory. All values are in degrees.

exhibit exceptionally low rise and twist values in the series of oligomers studied here.

In contrast to $B_I \rightarrow B_{II}$ transitions, which are both relatively rare and short lived for the tetranucleotides presently investigated, it appears that $\alpha\gamma$-flips can persist for at least the 15-ns trajectory that has been carried out in the first round of ABC simulations. This is consistent with the larger barriers that separate the g^-g^+ and g^+t states revealed by Varnai et al. (2002) in recent free-energy simulations. Such long-lived conformational substates clearly cannot be correctly sampled on the timescales of these calculations, and provide an explanation of the surprisingly irregular helical parameters obtained in the case of the GCGC oligomer. In fact, similar problems occur with a number of other CpG containing oligomers (GCGG, TCGG, TCGA) whose sequences also favor long-lived $\alpha\gamma$-flips. The question of whether this is a viable component of a dynamical model or a force-field artifact will need to be investigated further with longer trajectories, noting that $\alpha\gamma$-flips are observed in DNA sequence complexed to proteins but in few cases otherwise. If we now want to look provisionally at sequence effects on the CpG step, we have to deal with perturbations caused by $3'$ flanking $\alpha\gamma$-flips. Using these results, the only alternative is to filter out those parts of the trajectories associated with the unusual $\alpha\gamma$-states as artifact and analyze the remainder of the trajectory. Naturally, this means that the statistical quality of the data for some tetranucleotides may be reduced, or even that, in some cases, we may actually have no data left to analyze. This filtering has been applied in the results on d(CpG) contained in Table 3 and affects the GCGG, TCGG, and TCGA tetranucleotides. Data for GCGC can, however, be recovered by using the steps present in this dinucleotide repeat oligomer that are not affected by the $\alpha\gamma$-flips. Table 3 contains the mean value and the standard deviation of the six interbasepair helical parameters for the 10 possible tetranucleotide environments of the CpG step, averaged over the last 10 ns of the trajectories.

What can we read from the remaining data? A first remark is that, despite the $3'$ $\alpha\gamma$-filtering, there are still individual parameters that show visible differences between the two nominally identical steps in the given oligomers. This is clearly the case for TCGA, where there are clear discrepancies in slide, roll, and twist, and for GCGC, which shows a discrepancy in slide. For the remaining steps the differences between the two tetranucleotide copies are generally <0.3 Å for translational parameters and $<2°$ for rotational parameters. If we assume that such variations limit the precision of the parameters resulting from these

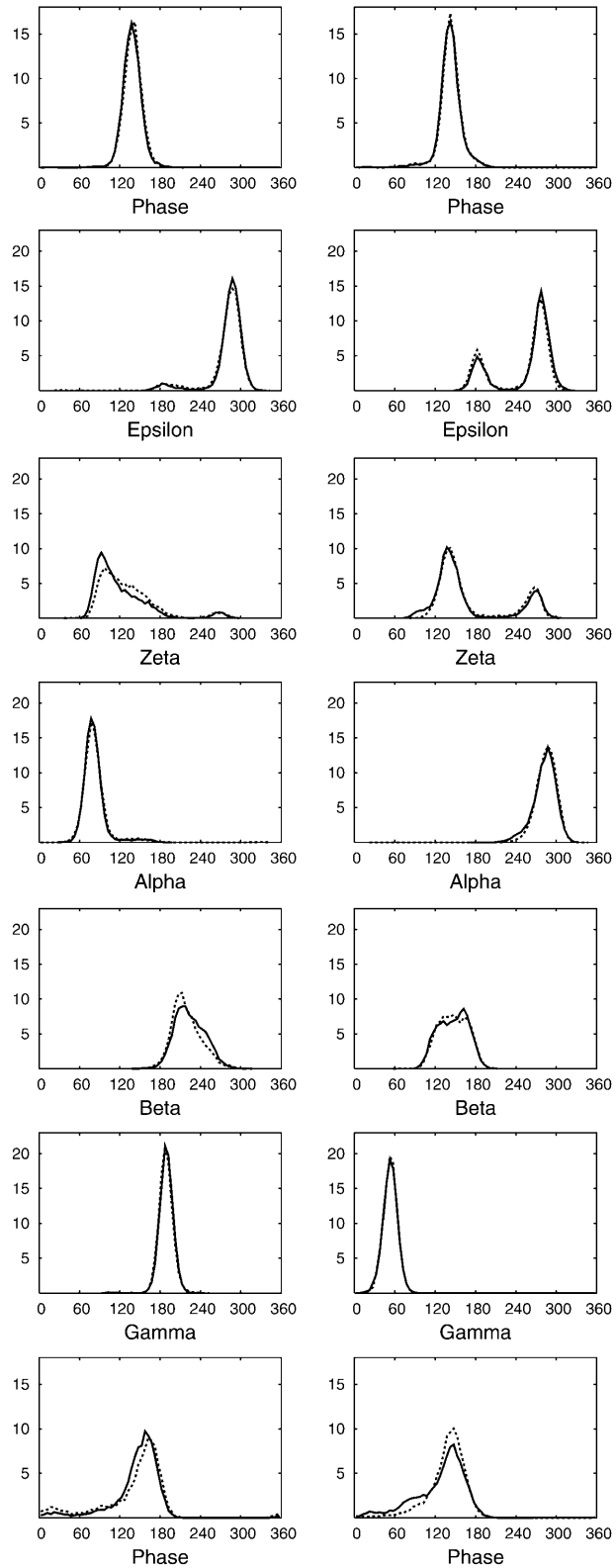

FIGURE 8 Frequency distributions of the sugar puckers and the five internucleotide backbone angles for the two backbones of the GpC steps within the GCGC oligomer. (*Left*) First strand, G_7pC_8 (*solid*) and $G_{11}pC_{12}$ (*dotted*); (*right*) second strand, $G_{25}pC_{26}$ (*solid*) and $G_{21}pC_{22}$ (*dotted*). Parameters are ordered in the $5' \rightarrow 3'$ direction of each backbone. Values have been accumulated over the last 10 ns of the trajectory. All values are in degrees.

simulations, then there are surprisingly few visible context effects on the six interbasepair parameters for the CpG. Shift and tilt need not be analyzed because their values are very small for all the CpG steps. For the remaining translational parameters, we can cite low rise for the CCGG and GCGG steps (although data on only one nucleotide are available in the latter case), higher rise in the RCGN steps, and somewhat lesser slide for GCGA. For rotational parameters, low twists occur for GCGG (one data point), GCGA, CCGG, and TCGG, whereas the RCGY steps (ACGT, GCGC, GCGT) have rather higher values. If we now look at the standard deviations of the interbasepair parameters, the situation can be rapidly summarized by saying that only minor context effects are visible in the MD modeling of the d(CpG) step. The limited crystal structure data that are currently available do not help in confirming or refuting this observation, although the idea of sequence-directed structural properties relies on the existence of such context effects. More detailed analysis of the available data is required to decisively conclude on the capabilities of the present level of approximations employed in molecular dynamics simulations of DNA to address this issue.

CONCLUSIONS

This article describes the approach adopted by the ABC consortium to perform molecular dynamics simulation on all tetranucleotide basepair steps in DNA. A single dinucleotide step, CpG, in all its possible neighboring sequence contexts, is analyzed here. By bringing together a small number of interested laboratories to work toward a common goal, the ABC project has been able to attack a problem that was beyond the computational possibilities of any single laboratory. The results obtained from this series of MD trajectories form a coherent database that can, in the future, be used as a reference for further studies of MD on DNA and for comparison with experiment. In providing data on all tetranucleotide sequence contexts, ABC has been able to achieve a goal that is, at present, unattainable by experiment. The results are surprising in several respects. First, although many structural and dynamic features of the oligomers studied have converged to stable values, the results indicate that slow backbone transitions prevent a complete sampling of the conformation space of B-DNA in the MD. For the same reason it is not yet possible to characterize all the consequences of such backbone transitions, which can occur independently or be coupled together, and which can influence the structural and dynamic behavior beyond the junction where the transition occurs. If we filter out such effects, the remaining conformational sampling appears to be reasonably balanced, but, surprisingly, suggests that the

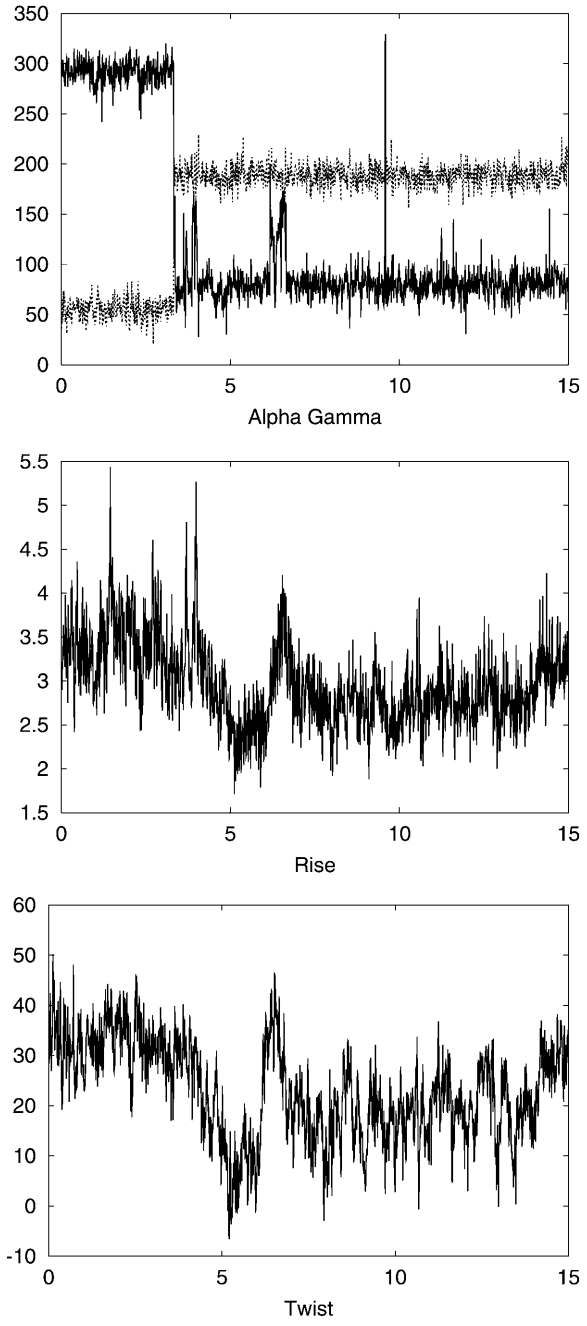

FIGURE 9 Time series plots of the $\alpha\gamma$-configuration of the $G_{11}pC_{12}$ step: α (*solid*), γ (*dotted*), and of the rise (Å) and twist (°) of the $C_{10}pG_{11}$ step within the GCGC oligomer over 15 ns of MD.

TABLE 3 Context effects on the interbasepair parameters of the CpG step

Tetranucleotide	*Time*	⟨*Shift*⟩	σ_{Shift}	⟨*Slide*⟩	σ_{Slide}	⟨*Rise*⟩	σ_{Rise}
GCGG	0.0	–	–	–	–	–	–
	2.4	−0.2	0.6	−1.0	0.7	2.8	0.6
ACGG	10.0	0.0	0.7	−1.0	0.7	3.2	0.5
	10.0	−0.0	0.7	−1.1	0.7	3.1	0.6
ACGA	10.0	0.3	0.7	−1.5	0.9	3.7	0.8
	10.0	0.1	0.7	−1.2	0.8	3.4	0.7
GCGA	10.0	0.0	0.7	−0.8	0.5	3.0	0.4
	10.0	−0.0	0.7	−0.9	0.6	3.0	0.5
ACGT	10.0	0.1	0.8	−1.2	0.8	3.5	0.6
	10.0	−0.1	0.8	−1.4	0.7	3.5	0.6
GCGC	10.0	0.0	0.7	−1.1	0.8	3.5	0.6
	10.0	−0.2	0.6	−1.6	0.8	3.8	0.6
GCGT	10.0	0.2	0.7	−1.2	0.8	3.4	0.6
	10.0	0.1	0.7	−1.4	0.7	3.5	0.6
TCGG	0.0	–	–	–	–	–	–
	10.0	−0.0	0.7	−0.9	0.6	3.0	0.5
CCGG	10.0	0.0	0.6	−0.8	0.6	2.8	0.4
	10.0	0.0	0.6	−1.0	0.6	2.9	0.5
TCGA	7.3	0.0	0.8	−0.9	0.6	3.2	0.5
	5.6	−0.3	0.7	−1.5	0.6	3.6	0.6

Tetranucleotide	*Time*	⟨*Tilt*⟩	σ_{Tilt}	⟨*Roll*⟩	σ_{Roll}	⟨*Twist*⟩	σ_{Twist}
GCGG	0.0	–	–	–	–	–	–
	2.4	0.4	4.4	7.6	6.0	24.9	5.8
ACGG	10.0	−0.1	5.1	7.0	6.9	33.9	6.1
	10.0	0.1	4.7	8.1	6.5	30.2	5.4
ACGA	10.0	1.5	5.3	7.4	7.5	31.4	5.7
	10.0	0.9	5.1	8.9	6.6	29.1	6.9
GCGA	10.0	1.3	4.8	9.9	6.0	26.7	7.7
	10.0	0.8	5.3	11.2	6.7	25.1	9.9
ACGT	10.0	0.3	5.4	7.0	6.7	33.9	6.1
	10.0	−0.3	5.1	7.4	6.3	32.6	5.6
GCGC	10.0	0.1	4.7	8.8	6.1	31.3	6.4
	10.0	0.4	4.6	10.1	6.3	33.3	5.9
GCGT	10.0	1.0	5.1	8.6	6.2	32.4	6.6
	10.0	0.9	4.9	7.7	6.4	32.1	5.1
TCGG	0.0	–	–	–	–	–	–
	10.0	−0.9	4.6	10.0	6.3	25.6	6.8
CCGG	10.0	0.0	4.3	8.3	5.6	27.6	6.7
	10.0	−0.4	4.3	7.9	6.0	28.6	6.7
TCGA	7.3	−0.2	5.1	11.4	6.6	29.3	6.6
	5.6	0.6	5.0	8.3	6.5	26.1	6.5

The table contains the mean value and the standard deviation of each parameter for the two copies of each tetranucleotide in the oligomers studied. The time of sampling (ns) after filtering out unusual $\alpha\gamma$-configurations is indicated. Translations are in angstroms and rotations are in degrees.

surrounding sequence has a very small effect on the properties of the CpG step. This indicates that any difference in the underlying potential as a consequence of helix context is probably only a fraction of a kcal/mol. Note that the MD results do not necessarily preclude the possibility of substates in the dynamical structure of CpG, but indicate that the higher energy state is not thermally accessible at 300 K, the temperature of the simulation.

The preliminary analysis offered here for the CpG step anticipates the issues for other steps and at least some of the problems involved and issues to be considered. However, before drawing any general conclusions from this first phase of the ABC initiative and resulting database, it is clearly necessary to complete the analysis of all 136 unique tetranucleotides. The results, even in this preliminary state, indicate that the dynamics of DNA introduce significant effects that raise a cautionary flag with respect to studies based on MD on DNA with short trajectories. A fuller

knowledge and better understanding of this is of course one of the objectives of this project. The subsequent analysis step, in itself, poses a challenging informatics problem, given that the combined trajectories represent roughly 0.6 μs of simulation and require almost 0.5 terabytes of storage. At this point all simulations from the initial phase of ABC are completed and analysis is underway. The data obtained will hopefully allow us to obtain an increasingly clear view of context effects, to better understand the importance of such phenomena as conformational substates, and also to define how end effects and length effects can influence the behavior of DNA fragments.

From this study, we hope to provide a benchmark of what can be expected from MD on DNA based on the parm94 empirical force field, and place subsequent applications of MD on a well-characterized theoretical basis. The sensitivity of these results to choice of force field is likewise interesting in pointing the way to appropriate improvements in functional forms or parameters. The sensitivity of the results to ionic strength is an additional important question to consider. To accomplish our task, it is quite possible that longer trajectories will be required. Given the experience gained to date by the ABC collaboration, it is reasonable to think of extending this computational effort by an order of magnitude if necessary.

The ABC collaboration commenced at a workshop on ''Atomistic to Continuum Models for Long Molecules and Thin Films'' held at the Mte Verita Conference Centre in Ascona, Switzerland in July, 2001. Funding for this meeting was provided by the Center Stefano Franscini, the European Office of Aerospace Research and Development, Air Force Office of Scientific Research, United States Air Force Research Laboratory, United States Office of Naval Research (Europe), Compaq, the European Science Foundation-Programme SIMU, and the Ecole Polytechnique Federale de Lausanne (EPLF), Switzerland. The ABC project was advanced in a Centre Européen de Calcul Atomique et Moléculaire (CECAM) workshop in Lyon, France the next year, and a meeting, ''DNA and Beyond: Structure, Dynamics and Interactions,'' held at the EPFL in April, 2003, sponsored by the Bernoulli Center of the EPFL and Hewlett-Packard, Inc.

Generous support for both these meetings is gratefully acknowledged. Prof. Beveridge acknowledges support for this research from the National Institute of General Medical Sciences (NIGMS) (grant no. GM37909) and the Keck Center for Integrative Genomics at Wesleyan University. The participation of Kelly M. Thayer in this project was supported by an NIGMS training grant in Molecular Biophysics to Wesleyan University (grant no. GM 08271). Dr. Gabriela Barreiro acknowledges CNPq/Brazil (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for a postdoctoral fellowship. Supercomputer time was generously provided under the auspices of the Partnerships for Advanced Computational Infrastructure program on the facilities of the National Center for Supercomputing Applications (NCSA) at the University of Illinois at Champaign/Urbana. The contribution of Dr. Richard Lavery and coworkers was supported by grants from the Centre National de la Recherche Scientifique, France. Dr. Peter Varnai thanks the Wellcome Trust for an International Prize Travelling Research Postdoctoral Fellowship (grant reference 060078). Prof. Cheathem acknowledges computational time from a National Research Advisory Council allocation (MCA01S027) and friendly user time on computer hardware at the University of Kentucky, the Pittsburgh Supercomputing Center, and the NCSA. Prof Osman acknowledges support from National Cancer Institute grant CA 63317. Eleanore Seibert was supported by National Institutes of Health training grants GM08553 and CA78207. Prof. Cheatham also acknowledges support from the Center for High Performance Computing at the University of Utah and financial support from the National Science Foundation (CHE-0326027). Dr. Filip Lankas acknowledges the support provided by the Centre for Complex Molecular Systems and Biomolecules (LN00A032) financed by the Ministry of Education of the Czech Republic. Prof Maddocks acknowledges the support for this research provided by the Swiss National Science Foundation and via a research collaboration between the EPFL and Hewlett-Packard.

REFERENCES

Aqvist, J. 1990. Ion-water interaction potentials derived from free energy perturbation simulations. *J. Phys. Chem.* 94:8021–8024.

Arthanari, H., K. J. McConnell, R. Beger, M. A. Young, D. L. Beveridge, and P. H. Bolton. 2003. Assessment of the molecular dynamics structure of DNA in solution based on calculated and observed NMR NOESY volumes and dihedral angles from scalar coupling constants. *Biopolymers.* 68:3–15.

Barbic, A., D. P. Zimmer, and D. M. Crothers. 2003. Structural origins of adenine-tract bending. *Proc. Natl. Acad. Sci. USA.* 100:2369–2373.

Berendsen, H. J. C., J. P. M. Postma, W. F. van Gunsteren, and A. DiNola. 1984. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* 81:3684–3690.

Bevan, D. R., L. Li, L. G. Pedersen, and T. A. Darden. 2000. Molecular dynamics simulations of the d(CCAACGTTGG)2 decamer: influence of the crystal environment. *Biophys. J.* 78:668–682.

Beveridge, D. L., S. B. Dixit, G. Barreiro, and K. M. Thayer. 2004. Molecular dynamics simulations of DNA curvature and flexibility: dynamical aspects of helix phasing and premelting phenomena. *Biopolymers.* 73:380–403.

Beveridge, D. L., and K. J. McConnell. 2000. Nucleic acids: theory and computer simulation, Y2K. *Curr. Opin. Struct. Biol.* 10:182–196.

Bolshoy, A., P. McNamara, R. E. Harrington, and E. N. Trifonov. 1991. Curved DNA without A-A: experimental estimation of all 16 DNA wedge angles. *Proc. Natl. Acad. Sci. USA.* 88:2312–2316.

Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. 1983. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4:187–195.

Brukner, I., R. Sanchez, D. Suck, and S. Pongor. 1995a. Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J.* 14:1812–1818.

Brukner, I., R. Sanchez, D. Suck, and S. Pongor. 1995b. Trinucleotide models for DNA bending propensity: comparison of models based on DNaseI digestion and nucleosome packaging data. *J. Biomol. Struct. Dyn.* 13:309–317.

Burkhoff, A. M., and T. D. Tullius. 1987. The unusual conformation adopted by the adenine tracts in kinetoplast DNA. *Cell.* 48:935–943.

Calladine, C. R., and H. R. Drew. 1997. Understanding DNA: The Molecule and How It Works. Academic Press, San Diego, CA.

Case, D. A., D. A. Pearlman, J. W. Caldwell, T. E. Cheatham III, W. S. Ross, C. Simmerling, T. Darden, K. M. Merz, R. V. Stanton, and A. Cheng. and others. 1999. AMBER: Version 6. Version 6.0. San Francisco: University of California.

Cheatham III, T. E., P. Cieplak, and P. A. Kollman. 1999. A modified version of the Cornell et al. force field with improved sugar pucker phases and helical repeat. *J. Biomol. Struct. Dyn.* 16:845–862.

Cheatham III, T. E., and P. A. Kollman. 2000. Molecular dynamics simulation of nucleic acids. *Annu. Rev. Phys. Chem.* 51:435–471.

Cheatham III, T. E., J. L. Miller, T. Fox, T. A. Darden, and P. A. Kollman. 1995. Molecular dynamics simulations on solvated biomolecular

systems: the particle mesh Ewald method leads to stable trajectories of DNA, RNA, and proteins. *J. Am. Chem. Soc.* 117:4193–4194.

Cheatham III, T. E., J. L. Miller, T. I. Spector, P. Cieplak, and P. A. Kollman. 1998. Molecular dynamics simulations on nucleic acid systems using the Cornell et al force field and particle mesh Ewald electrostatics. *In* Molecular Modeling of Nucleic Acids. American Chemical Society, Washington, DC. 285–303.

Cheatham III, T. E., and M. A. Young. 2001. Molecular dynamics simulation of nucleic acids: successes, limitations, and promise. *Biopolymers.* 56:232–256.

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

Dickerson, R. E. 1983. The DNA helix and how it is read. *Sci. Am.* 249: 94–111.

Dickerson, R. E., and T. K. Chiu. 1997. Helix bending as a factor in protein/DNA recognition. *Biopolymers.* 44:361–403.

Dickerson, R. E., and H. R. Drew. 1981. Structure of a B DNA dodecamer II. Influence of base sequence on helix structure. *J. Mol. Biol.* 149: 761–786.

Dickerson, R. E., D. S. Goodsell, M. L. Kopka, and P. E. Pjura. 1987. The effect of crystal packing on oligonucleotide double helix structure. *J. Biomol. Struct. Dyn.* 5:557–579.

DiGabriele, A. D., M. R. Sanderson, and T. A. Steitz. 1989. Crystal lattice packing is important in determining the bend of a DNA dodecamer containing an adenine tract. *Proc. Natl. Acad. Sci. USA.* 86:1816–1820.

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

Drew, H. R., R. M. Wing, T. Takano, C. Broka, S. Tanaka, K. Itikura, and R. E. Dickerson. 1981. Structure of a B DNA dodecamer I: conformation and dynamics. *Proc. Natl. Acad. Sci. USA.* 78:2179–2183.

El Hassan, M. A., and C. R. Calladine. 1995. The assessment of the geometry of dinucleotide steps in double-helical DNA; a new local calculation scheme. *J. Mol. Biol.* 251:648–664.

El Hassan, M. A., and C. R. Calladine. 1996. Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *J. Mol. Biol.* 259:95–103.

Essmann, U., L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. 1995. A smooth particle mesh Ewald method. *J. Chem. Phys.* 103:8577–8593.

Feig, M., and B. M. Pettitt. 1998. Structural equilibrium of DNA represented with different force fields. *Biophys. J.* 75:134–149.

Foloppe, N., and A. D. MacKerell, Jr. 2000. All-atom empirical force field for nucleic acids. I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comp. Chem.* 21: 86–104.

Giudice, E., and R. Lavery. 2002. Simulations of nucleic acids and their complexes. *Acc. Chem. Res.* 35:350–357.

Haile, J. M. 1992. Molecular Dynamics Simulation: Elementary Methods. John Wiley and Sons, New York, NY.

Hartmann, B., D. Piazzola, and R. Lavery. 1993. BI-BII transitions in B-DNA. *Nucleic Acids Res.* 21:561–568.

Harvey, S. C., R. K. Z. Tan, and T. E. Cheatham, Iii. 1998. The flying ice cube: velocity rescaling in molecular dynamics leads to violation of energy equipartition. *J. Comput. Chem.* 19:726–740.

Hunenberger, P. H., and J. A. McCammon. 1999. Effect of artificial periodicity in simulations of biomolecules under Ewald boundary conditions: a continuum electrostatics study. *Biophys. Chem.* 78:69–88.

Johansson, E., G. Parkinson, and S. Neidle. 2000. A new crystal form for the dodecamer C-G-C-G-A-A-T-T-C-G-C-G: symmetry effects on sequence-dependent DNA structure. *J. Mol. Biol.* 300:551–561.

Jorgensen, W. L. 1981. Transferable intermolecular potential functions for water, alcohols and ethers. Application to liquid water. *J. Am. Chem. Soc.* 103:335–340.

Kanhere, A., and M. Bansal. 2003. An assessment of three dinucleotide parameters to predict DNA curvature by quantitative comparison with experimental data. *Nucleic Acids Res.* 31:2647–2658.

Kim, U. S., B. S. Fujimoto, C. E. Furlong, J. A. Sundstrom, R. Humbert, D. C. Teller, and J. M. Schurr. 1993. Dynamics and structures of DNA: long-range effects of a 16 base-pair (CG)8 sequence on secondary structure. *Biopolymers.* 33:1725–1745.

Langley, D. R. 1996. The BMS Nucleic Acid Force Field. Bristol-Myers Squibb, Wallingford, CT.

Langley, D. R. 1998. Molecular dynamic simulations of environment and sequence dependent DNA conformations: the development of the BMS nucleic acid force field and comparison with experimental results. *J. Biomol. Struct. Dyn.* 16:487–509.

Lankas, F., J. Sponer, J. Langowski, and T. E. Cheatham 3rd. 2003. DNA basepair step deformability inferred from molecular dynamics simulations. *Biophys. J.* 85:2872–2883.

Lavery, R., and H. Sklenar. 1996. Curves 5.1: Helical Analysis of Irregular Nucleic Acids. Institut de Biologie Physico-Chimique, Paris, France.

Leach, A. R. 1996. Molecular Modeling: Principles and Applications. Addison Wesley Longman, Essex, UK.

Liu, Y., and D. L. Beveridge. 2001. A refined prediction method for gel retardation of DNA oligonucleotides from dinucleotide step parameters: reconciliation of DNA bending models with crystal structure data. *J. Biomol. Struct. Dyn.* 18:505–526.

Lu, X.-J., M. Babcock, and W. Olson. 1999. Mathematical overview of nucleic acid analysis programs. *J. Biomol. Struct. Dyn.* 16:833–843.

Lu, X.-J., and W. K. Olson. 1999. Resolving the discrepancies among nucleic acid conformational analyses. *J. Mol. Biol.* 285:1563–1575.

MacDonald, D., and P. Lu. 2002. Residual dipolar couplings in nucleic acid structure determination. *Curr. Opin. Struct. Biol.* 12:337–343.

MacKerell, A. D., Jr., and N. Banavali. 2000. All-atom empirical force field for nucleic acids. II. Application to molecular dynamics simulations of DNA and RNA in solution. *J. Comput. Chem.* 21:105–120.

MacKerell, A. D., Jr., N. Banavali, and N. Foloppe. 2000. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers.* 56:257–265.

McCammon, A. J., and S. C. Harvey. 1986. Dynamics of Proteins and Nucleic Acids. Cambridge University Press, Cambridge, UK.

McConnell, K. M., R. Nirmala, M. A. Young, G. Ravishanker, and D. L. Beveridge. 1994. A nanosecond molecular dynamics trajectory for a B DNA double helix: evidence for substates. *J. Am. Chem. Soc.* 116: 4461–4462.

Miller, J. L., T. E. Cheatham III, and P. A. Kollman. 1999. Simulation of nucleic acid structure. *In* Oxford Handbook of Nucleic Acid Structure. S. Neidle, editor. Oxford University Press, Oxford, UK and New York, NY. 95–115.

Neidle, S. editor. 1999. Oxford Handbook of Nucleic Acid Structure. Oxford University Press, Oxford, UK and New York, NY.

Neidle, S. 2002. Nucleic Acid Structure and Recognition. Oxford University Press, Oxford, UK.

Norberg, J., and L. Nilsson. 2002. Molecular dynamics applied to nucleic acids. *Acc. Chem. Res.* 35:465–472.

Olson, W. K., A. A. Gorin, X. J. Lu, L. M. Hock, and V. B. Zhurkin. 1998. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA.* 95:11163–11168.

Olson, W. K., and V. B. Zhurkin. 1996. Twenty Years of DNA Bending. *In* Biological Structure and Dynamics. R. H. Sarma and M. H. Sarma, editors. Adenine Press, Albany, NY. 341–370.

Orozco, M., A. Perez, A. Noy, and F. J. Luque. 2003. Theoretical methods for the simulation of nucleic acids. *Chem. Soc. Rev.* 32:350–364.

Packer, M. J., M. P. Dauncey, and C. A. Hunter. 2000a. Sequence-dependent DNA structure: dinucleotide conformational maps. *J. Mol Biol.* 295:71–83.

Packer, M. J., M. P. Dauncey, and C. A. Hunter. 2000b. Sequence-dependent DNA Structure: tetranucleotide conformational maps. *J. Mol. Biol.* 295:85–103.

Poncin, M., B. Hartmann, and R. Lavery. 1992. Conformational sub-states in B-DNA. *J. Mol. Biol.* 226:775–794.

Ponomarev, S., K. M. Thayer, and D. L. Beveridge. 2004. Ion motions in molecular dynamics simulations on DNA. *Proc. Natl. Acad. Sci. USA.* 101:14771–14775.

Reddy, S. Y., F. Leclerc, and M. Karplus. 2003. DNA polymorphism: a comparison of force fields for nucleic acids. *Biophys. J.* 84:1421–1449.

Ryckaert, J. P., G. Ciccotti, and H. J. C. Berendsen. 1977. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* 23: 327–336.

Saenger, W. 1984. Principles of Nucleic Acid Structure. Springer Verlag, New York, NY.

Satchwell, S. C., H. R. Drew, and A. A. Travers. 1986. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.* 191: 659–675.

Schlick, T. 2002. Molecular Modeling and Simulation: An Interdisciplinary Guide. J. E. Marsden, L. Sirovich, S. Wiggins, and S. S. Antman, editors. Springer, New York, NY.

Scott, W. R. P., P. H. Hunenberger, I. G. Tironi, A. E. Mark, S. R. Billeter, J. Fennen, A. E. Torda, T. Huber, P. Kruger, and W. F. Gunsteren. 1999. The GROMOS biomolecular simulation program package. *J. Phys. Chem. A.* 103:3596–3607.

Shakked, Z., G. Guerstein Guzikevich, M. Eisenstein, F. Frolow, and D. Rabinovich. 1989. The conformation of the DNA double helix in the crystal is dependent on its environment. *Nature.* 342:456–460.

Sinden, R. R. 1994. DNA Structure and Function. Academic Press, San Diego, CA and London, UK.

Smith, P. E., and B. M. Pettitt. 1996. Ewald artifacts in liquid state molecular dynamics simulations. *J. Chem. Phys.* 105:4289–4293.

Suzuki, M., N. Amano, J. Kakinuma, and M. Tateno. 1997. Use of a 3D structure data base for understanding sequence-dependent conformational aspects of DNA. *J. Mol. Biol.* 274:421–435.

Tjandra, N., S.-I. Tate, A. Ono, M. Kainosho, and A. Bax. 2000. The NMR structure of a DNA dodecamer in an aqueous dilute liquid crystalline phase. *J. Am. Chem. Soc.* 122:6190–6200.

Varnai, P., D. Djuranovic, R. Lavery, and B. Hartmann. 2002. Alpha/gamma transitions in the B-DNA backbone. *Nucleic Acids Res.* 30: 5398–5406.

Varnai, P., and K. Zakrzewska. 2004. DNA and its counterions: a molecular dynamics study. *Nucleic Acids Res.* 32:4269–4280.

Vermulen, A., H. Zhou, and A. Pardi. 2000. Determining DNA global structure and DNA bending by application of NMR residual dipolar couplings. *J. Am. Chem. Soc.* 122:9638–9647.

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

Yanagi, K., G. G. Prive, and R. E. Dickerson. 1991. Analysis of local helix geometry in three B-DNA decamers and eight dodecamers. *J. Mol. Biol.* 217:201–214.

York, D. M., W. Yang, H. Lee, T. Darden, and L. G. Pedersen. 1995. Toward the accurate modeling of DNA: the importance of long- range electrostatics. *J. Am. Chem. Soc.* 117:5001–5002.

Young, M. A., B. Jayaram, and D. L. Beveridge. 1997a. Intrusion of counterions into the spine of hydration in the minor groove of B-DNA: fractional occupancy of electronegative pockets. *J. Am. Chem. Soc.* 119:59–69.

Young, M. A., G. Ravishanker, and D. L. Beveridge. 1997b. A 5-nanosecond molecular dynamics trajectory for B-DNA: analysis of structure, motions and solvation. *Biophys. J.* 73:2313–2336.

Young, M. A., G. Ravishanker, D. L. Beveridge, and H. M. Berman. 1995. Analysis of local helix bending in crystal structures of DNA oligonucleotides and DNA-protein complexes. *Biophys. J.* 68: 2454–2468.