The ABCs of molecular dynamics simulations on B-DNA, circa 2012

David L Beveridge[1],*, Thomas E Cheatham III[2] and Mihaly Mezei[3]

[1]*Department of Chemistry and Molecular Biophysics Program, Wesleyan University Middletown, CT 06459, USA*

[2]*Department of Medicinal Chemistry, University of Utah, Salt Lake City, UT 84112, USA*

[3]*Department of Structural and Chemical Biology, Mount Sinai School of Medicine, New York, NY 10029, USA*

**Corresponding author (Fax, 860-685-2211; Email, dbeveridge@wesleyan.edu)*

We dedicate this article to the memory of Prof Peter A Kollman (1944–2001), our colleague, mentor and friend.

This article provides a retrospective on the ABC initiative in the area of all-atom molecular dynamics (MD) simulations including explicit solvent on all tetranucleotide steps of duplex B-form DNA duplex, ca. 2012. The ABC consortium has completed two phases of simulations, the most current being a set of 50–100 trajectories based on the *AMBER ff99* force field together with the *parmbsc0* modification. Some general perspectives on the field of MD on DNA and sequence effects on DNA structure are provided, followed by an overview our MD results, including a detailed comparison of the *ff99/parmbsc0* results with crystal and NMR structures available for d(CGCGAATTCGCG). Some projects inspired by or related to the ABC initiative and database are also reviewed, including methods for the trajectory analyses, informatics of dealing with the large database of results, compressions of trajectories for efficacy of distribution, DNA solvation by water and ions, parameterization of coarse-grained models with applications and gene finding and genome annotation

[Beveridge DL, Cheatham TE III and Mezei M 2012 The ABCs of molecular dynamics simulations on B-DNA, circa 2012. *J. Biosci.* **37** 379–397]
DOI 10.1007/s12038-012-9222-6

1. Introduction

This article is a review of the progress made in the area of all-atom molecular dynamics (MD) computer simulations including solvent on duplex B-form DNA over the last 10 years by an international alliance of research groups known as the Ascona B-DNA Consortium (ABC). The original objective of this project was to produce a database of state-of-the-art MD on DNA in which all 136 unique tetranucleotide steps are represented, a sizable undertaking at the time (ca. 2001). The idea was to provide a comprehensive documentation of the performance of all-atom MD on DNA based on the best current nucleic acid force fields, and to obtain information for the comprehensive study and improved understanding of base pair sequence effects on structure and dynamics. Detailed knowledge of the effect of sequence on DNA structure is necessary for a general understanding of the sequence specificity of DNA–protein and DNA–ligand interactions as well as the role of DNA shape, flexibility, bending and bendability in mechanisms of action. In the course of this project, it has also been necessary to address critically some issues fundamental to the general application on MD to DNA, such as the parameterization of force fields, MD simulation protocols, the stability and convergence of MD simulations and methods for the analysis and validation of results. These topics are also reviewed.

The ABC initiative began in June 2001 at an informal lunchtime discussion during a workshop/conference on *Atomistic to Continuum Models for Long Molecules* in Ascona, Switzerland *(http://lcvmwww.epfl.ch/~lcvm/ascona2001/index.html),* and involved a number of those working on methods and applications of MD simulations to nucleic acids. There was specific interest in basic knowledge of the structure and motions of duplex B-form DNA, the predominant DNA

Keywords. ABC; DNA; molecular dynamics; sequence effects; tetranucleotide steps

Published online: 25 June 2012

conformation *in vivo* and the effect of base pair sequence on structure. Earlier studies of sequence effects in DNA had been carried out at the dinucleotide step level, but the influence of base pairs adjacent to dinucleotide steps was likely to be an important factor. In this case, at least a tetranucleotide model would be required to be able to fully understand sequence effects. Although the available experimental data for unbound DNA structures at the dinucleotide step level had been compiled (Berman *et al.* 1992) and analysed in detail (Olson *et al.* 1998; Zhurkin *et al.* 2005; Olson *et al.* 2006), sufficient statistical sampling of all tetranucleotide steps is not yet available.

At this point the idea was broached that the tetranucleotide database could be created using all-atom MD simulations. With the latest developments in nucleic acid force fields and access to high-performance computing, all-atom MD on DNA including explicit solvent was coming into its own and producing computational models of DNA for specific sequences that showed improved agreement with experiment (Beveridge and McConnell 2000; Cheatham 3rd and Kollman 2000). However, MD simulation projects described at this meeting typically dealt with one or just a few sequences, most at the dodecamer level of sequence length, with ~5 ns trajectories. Anticipating further advances in computer power, taking on a full set of tetranucleotide simulations at a meaningful level was estimated to require at least ~15 ns trajectories. At that time, a project of this magnitude was beyond the immediate capability of any one research group, but the idea that we could do this collectively with a 'divide and conquer' strategy was intriguing. The feasibility of such collaboration was influenced strongly by the ever-increasing capabilities of the Internet for ease of communication amongst participants and transferring data files. The proposed *modus operandi* was to parcel out the tetranucleotide sequences to various research groups willing and able to commit resources, and have each group perform the MD on their own computers using a commonly-agreed-upon force field and simulation protocols. Each of the potential participants was a user of the *AMBER* suite of programs for molecular simulations developed in the laboratory of our late colleague Peter A Kollman (Pearlman *et al.* 1995; Kollman *et al.* 2002; Case *et al.* 2005), and *AMBER* was thus adopted as our simulation engine. We recognized that not only getting the simulations done but the analysis of such a large database of results was going to be a considerable challenge in informatics. From the outset, our intention was to make the trajectories generally available to the field for further studies once they were completed.

We began the ABC initiative with limited expectations about whether it would actually get done, especially since everyone already had a full platter of other projects and commitments to various funding agencies to fulfill. The issue of obtaining funding for ABC was discussed, but we decided that in such a rapidly developing field it was better to just work on the project any way we could rather that commit the time (1–2 years, even if successful) to obtain a grant and then go from there. In summary, the consortium members found a quite viable way to work together, and simulations were set up and carried out over the course of the next ~2 years. To discuss preliminary results, informal gatherings of as many of the consortium members as possible were convened in conjunction with scheduled international symposia and CECAM Workshops. MD trajectories of 15 ns that comprise what we now call 'phase I' (ABC I) were completed on in a reasonable time frame, and are described in papers published in 2004 (Beveridge *et al.* 2004) and 2005 (Dixit *et al.* 2005). While interesting results were obtained, there turned out to be an ergodic problem in the MD force field that only materialized in longer trajectories beginning with canonical B-DNA. Specifically, transitions in the DNA backbone of the γ backbone torsion angle to *trans*, transitions that should be reversible (Varnai *et al.* 2002), were found to accumulate and eventually degrade the duplex structure. Considerable effort was required to correct this problem properly (Perez *et al.* 2007a, b; Svozil *et al.* 2008) and the whole set of simulations was rerun, this time at the level of 50 ns trajectories. This is referred to as ABC II and is described in an article that appeared in 2010 (Lavery *et al.* 2010). An ABC III initiative pushing simulations to the microsecond timescale is currently in progress, and future problems involving MD on DNA that require cooperation of laboratories and sharing of resources are being discussed amongst the active participants.

This article provides a retrospective on the ABC initiative circa 2012. A full list of those who contributed to the various phases of the ABC is provided in the papers published so far (Beveridge *et al.* 2004; Dixit *et al.* 2005; Lavery *et al.* 2009). Some projects that came about all or in part as a consequence of ABC or because of general public access to the database are also reviewed. Specifically, the focus in this article is (a) limited to the ABC project, (b) concerns only B-form DNA and (c) is *AMBER*-centric. A number of reviews with a broader purview of the field of MD on DNA during the time frame of ABC I and II are available in the recent literature (Beveridge and McConnell 2000; Cheatham 3rd and Young 2001; Giudice and Lavery 2002; Reddy *et al.* 2003; Mackerell Jr and Nilsson 2008; Orozco *et al.* 2008; Laughton and Harris 2011; Perez *et al.* 2012)

2. Background

JD Watson and Francis Crick in 1953 reported the discovery of the structure of DNA as a double helix of intertwined single-stranded polynucleotides interacting via complementary hydrogen bonds between nucleotide bases (Watson and Crick 1953). Some of the data they worked from were obtained from diffraction experiments on fibres (Franklin

and Gosling 1953), which led to a duplex structure now known as the right-handed B-form of DNA. From this structure, models for how DNA self-replicates and codes for proteins were deduced. Subsequently, higher-resolution fibre diffraction structures were obtained that have served to define all-atom models of the canonical form of B-DNA (Arnott *et al.* 1976). The first X-ray crystal structure of B-form DNA at molecular resolution was reported in 1980 by Dickerson and coworkers (Wing *et al.* 1980) and was generally consistent with the structure proposed by Watson and Crick. However, the higher resolution revealed sequence-dependent helix bending, conformational irregularities and hydration patterns. Dickerson (Dickerson 1983) recognized that a code for the sequence-specific interactions may be found not only in the patterns of donor and acceptor sites in the major and minor grooves but also in sequence-dependent shape and structure of DNA. The role of dynamics as well as structure should also be considered (Pastor and Weinstein 2001). These ideas have generally stood the test of time, and have become essential elements of core knowledge in the field.

Since these pioneering studies, the techniques for obtaining high-resolution models based on X-ray diffraction of DNA in crystals and more recently NMR spectroscopy on DNA in solution have been considerably refined. The field of purely computational modelling of DNA using molecular simulation techniques has developed in parallel (Beveridge and McConnell 2000; Cheatham 3rd *et al.* 2001; Cheatham 3rd and Young 2001; Perez *et al.* 2012). The method of choice in the ABC initiative is MD, which produces a time series of individual structures that define the MD model structure and motions. Molecular structures of DNA can be analysed in terms of the derived parameters shown in figure 1. The first MD on DNA was reported in 1983 (Levitt 1983; Tidor *et al.* 1983), following the first MD on a protein by several years (McCammon *et al.* 1977). The time step for numerical integration in MD is typically 1–2 fs, and early simulations were carried out on a picosecond time frame. In these early simulations, solvent was not considered explicitly, but introduced via a distance-dependent dielectric screening function. However, the early MD on DNA produced highly distorted and unreasonable conformations when the simulations were extended without unphysical approximations, i.e. neglecting electrostatic repulsions and other adjustments or constraints. MD on DNA was expected to present challenging problems to simulation due to the polyelectrolyte character of the system. Also, with solvent water and ions already well known to be an integral part of DNA stability (Saenger 1984), a molecular model for water was likely to be required for accurate simulations.

The computer power to perform MD on DNA including explicit solvent became available circa 1990, but the force fields were still not up to the task. Breakthroughs in the field of MD on DNA using *AMBER* came with the development of the 'second generation' force field *ff94* (originally referred to as *parm94*) designed specifically for simulations including explicit solvent (Cornell *et al.* 1995), the availability of high-performance parallel computing compatible versions of the code, and implementation of fast Ewald methods (Essmann *et al.* 1995). All-atom MD on DNA including explicit solvent using the *ff94* force field produced the first MD models of stable B-form DNA in solution on nanosecond timescales (McConnell *et al.* 1994; Cheatham 3rd *et al.* 1995; Young *et al.* 1997a, b). Detailed comparisons of MD on DNA were carried out on crystal structures (Young et al. 1995) and on NMR structures, some comparing calculated

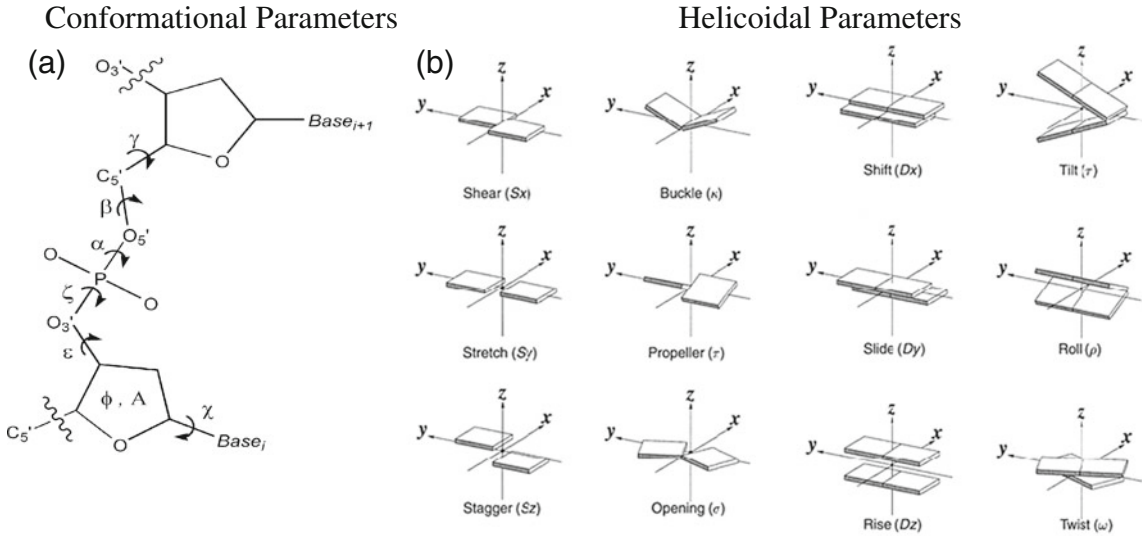

Figure 1. Definition of (**a**) conformational and (**b**) helicoidal parameters describing DNA structures (Lu and Olson 2003).

and observed NOESY peak volumes (Arthanari *et al.* 2003). Even though the MDs were ~5 ns and thus short by the standards of 2012, stable results were obtained and showed overall reasonable agreement between MD calculated and observed results, especially considering the complexity of the problem. However, some specific differences were also noted, particularly a general tendency in the MD models based on *ff94* to form under-twisted structures (Cheatham 3rd and Kollman 1997a, b, 1998; Cheatham 3rd *et al.* 1999) and an ergodic problem with the *ff94/ff98/ff99* force fields, discussed in detail below. A limitation of all-atom MD then and now is that only the relatively fast motions of a system are accessible to study, whereas many phenomena of biological interest occur on the milliseconds to seconds time frame and beyond. Thus, there has been and will be a continuing emphasis on extending MD on DNA to longer time scales to gain direct access to biological problems.

By the time of the 2001 Ascona meeting, successful MD simulations on DNA were being carried out in several laboratories with generally encouraging results (Miller *et al.* 1999; Cheatham 3rd and Kollman 2000), including a theoretical account of the hydration dependent A- to B-DNA transition (Cheatham 3rd and Kollman 1996, 1997a,b; Sprous *et al.* 1998). This system has also been the subject of a number of subsequent studies (Knee *et al.* 2008). The calculated distribution functions for mobile counterions provided a description of the ion atmosphere of DNA in remarkable agreement with Manning's counterion condensation theory (Manning 1978; Young *et al.* 1997a, b), and introduced the provocative idea that fractional occupation of ions in the grooves of DNA, i.e. that the minor groove spine of hydration reported in crystal structures may not all be water (Young *et al.* 1997a, b; McConnell and Beveridge 2000; Hamelberg *et al.* 2001). This point has been addressed further in a number of subsequent papers (Ponomarev *et al.* 2004; Rueda *et al.* 2004). The consensus MD view is now that the ions, as expected, preferentially sample electronegative sites around the DNA, but direct ion association with nucleotide bases in the grooves are found in only $\leq 10\%$ of the time (Rueda *et al.* 2004). In any case, with a reasonable MD model of DNA available, a number of questions about sequence effects on structure were addressed over the next few years, including the structure of A-tracts (~ straight helix axis) (Sherer *et al.* 1999; McConnell and Beveridge 2001; Madhumalar and Bansal 2003; Lankas *et al.* 2010), an account of augmented DNA bending in sequences with A-tracts phased by a full helix turn (bending locus often at CpG steps) (Lefebvre *et al.* 1995; Young and Beveridge 1998; Lankas *et al.* 2010) and an explanation of the relative bending in sequence of phased A_4T_4 (bent) compared with phased T_4A_4 (~straight) (Sprous *et al.* 1999). More details on the dynamics of DNA bending are available in recent reviews (Beveridge *et al.* 2004; Zhurkin *et al.* 2005).

As the crystal structures of more and more oligonucleotides became available, information on patterns in DNA sequence effects accumulated (El Hassan and Calladine 1995; Suzuki *et al.* 1997; Olson *et al.* 1998). Classification of steps as either purine-pyrimidine (RY), purine-purine (RR) or pyrimidine-purine (YR) is useful to classify many of the observed conformational variations. YR steps (CA/TG, CG, TA) tend to have positive values of roll and higher than average twist, RY steps (AC/ GT, AT, GC) tend to have negative roll and lower than average twist, and RR steps (AA/TT, AG/ CT, GA/TC and GG/CC) are intermediate between these two groups. Dinucleotide steps also display sequence-dependent flexibility and deformability (bendability) (Ulyanov and Zhurkin 1984). Further evidence for the flexibility of CA/TG steps was noted early on from studies using empirical energy functions (Zhurkin 1985; Olson *et al.* 1998) A pattern with three rigid steps, AA/TT, AT and GA/TC, was noted (El Hassan and Calladine 1997), with the remaining more flexible steps further subdivided into bistable (all homogenous GpC steps) and flexible (CA/TG and TA). Mining the protein–DNA crystal structures further elucidated trends in sequence effects and correlations in the structural parameters (Olson *et al.* 1998, 2006), and provided further evidence that YR steps show the greatest flexibility, especially CA/TG. The subject of intrinsic flexibility of DNA and the ligand-induced bendability has been a subject of active interest due to the role these properties play in binding mechanisms (Olson and Zhurkin 2000; Lankas *et al.* 2004; Noy *et al.* 2004).

Ideas about the origins of sequence effects on DNA were first based on the steric clashes between purine and pyrimidine bases (Calladine 1982) and the variation in twist (Ulyanov and Zhurkin 1984). Further development of the steric effects view of sequence on structure was described by Suzuki *et al.* (1997) An alternative view based on base pair stacking interactions was subsequently advanced (Packer *et al.* 2000; Farwer *et al.* 2006). Normal mode analysis of oligonucleotide DNA using knowledge-based potentials obtained from mining the crystal structures successfully accounted for the bending persistence length and stretching modulus of DNA and indicated a sensitivity of twisting force constants to the base pair sequence (Matsumoto and Olson 2002). An MD study of two 18-base-pair DNA oligomers was reported in which all 10 unique dinucleotide base pair steps are represented (Lankas *et al.* 2003) and showed a trend in relative flexibility in roll, YR>RR>RY. The YR steps were also found to be the most flexible in tilt and partially in twist, supporting previous results. Slide–rise, twist–roll and twist–slide elastic couplings of various degrees were observed and a correlation of motions on a length scale of 2–3 base pairs was noted, which falls in the neighbourhood of first neighbour context effects. A number of subsequent studies of MD on DNA flexibility have extended and refined this account. (Lankas *et al.* 2000; Olson and Zhurkin 2000; Lankas *et al.* 2003, 2004; Noy *et al.* 2004).

With respect to the ABC initiative, it had by this point become clear that the issue of sequence effects on DNA the tetranucleotide level was only one of questions of interest. More fundamental was the issue of convergence and validation of the simulations – what length of trajectories were required to obtain stable results and how accurate are they over a broad range of sequences? One can envisage future use of good-quality computational models of DNA in diverse ways in problems ranging from structural biology to systems biology. Other questions had to do with the identification of thermally accessible sub-states, the nature of correlations within and between the conformational and helicoidal parameter sets, principal component and flexibility analysis of results, and sequence effects on DNA solvation. Another of the original motivations for doing the ABC project was to use ABC trajectories as a basis for choosing parameters for coarse-grained energy functions that could be used in calculations that predict structures for longer oligonucleotide sequences and for statistical mechanics partition functions. The progress in each of these areas related to the ABC simulations is also reviewed below.

3. Calculations

The effect of base pair sequence on structure can be conveniently expressed in terms of parameters derived from the Cartesian coordinates of model DNA structures (figure 1). The minimum structural unit that carries information on the three-dimensional structure of DNA is the dinucleotide base pair step, 5′-dXY-3′, where X and Y may be A, T, G or C. The four alternatives lead to 16 permutations, of which 10 are unique. The tetranucleotide problem arises since dinucleotide steps structures may depend on their nearest neighbours. In this case, the minimum structural unit necessary to study would be 4 base pair steps, of which there are 136 unique permutations.

A novel research design for ABC was proposed by Richard Lavery (Beveridge *et al.* 2004) and adopted for all phases of the project to date. Instead of running 136 different MD trajectories with one tetrad per oligomer embedded as the central 4 base pairs of a dodecamer, Lavery's idea was to pack repeating tetranucleotide sequences (WXYZWXYZWXYZ…), where W, X, Y and Z are any of the four nucleotide bases, A, T, G or C, within a single longer sequence. In this way, each oligomer can contain up to four distinct tetranucleotides, WXYZ, XYZW, YZWX and ZWXY. This strategy enables all 136 tetranucleotides to be studied using only 39 oligomers. Oligomers of 13 base pairs were used in ABC I and of 18 base pairs in ABC II. The ends of each oligomer were capped with a single GC pair (13-mers) and two GC pairs (18-mers) to avoid fraying. Thus, a given 18-base-pair oligomer contains 3 tetranucleotide repeats 5′-GC- YZ- WXYZ-WXYZ-WXYZ-GC-3′ and a four base pair repeating sequence, WXYZ, which occurs 3.5 times. Having multiple copies of each sequence is advantageous for statistics, and comparison of several different examples of any given sequence serves as a test for convergence.

Each oligomer was constructed as a canonical B-form DNA to begin with. All MD simulations were carried out with periodic boundary conditions on a truncated octahedral cell (figure 2) using the *AMBER* suite of programs (Case *et al.* 2006) using *ff94* (Cornell *et al.* 1995) in ABC phase I and *ff99* (Cheatham 3rd *et al.* 1999) with the modifications of *parmbsc0* (Perez *et al.* 2007a, b) in ABC phase II. The water model used was SPC/E (Berendsen *et al.* 1987) ion parameters were those of Aqvist (Aqvist 1990) in ABC I and Smith and Dang (Dang 1995) in ABC II. (Note: improved parameters for ions are now available) (Joung and Cheatham 3rd 2009). Long-range electrostatic interactions were treated using the particle mesh Ewald method (Cheatham *et al.* 1995). A standard protocol of minimization, heating and equilibration was applied. Production simulations were carried out using an NPT ensemble with bonds to hydrogen atoms were restrained, allowing for stable simulations with a 2 fs time step. Centre-of-mass motion was removed every 5000 steps to avoid kinetic energy building up in translational motion (Harvey *et al.* 1998) and to keep the solute centered in the simulation cell.

4. Results

4.1 *ABC Phase I (Beveridge et al. 2004; Dixit et al. 2005)*

The MD simulations in ABC I provided a database of the trajectories and structural parameters for all tetranucleotide steps based on 15 ns trajectories. The dinucleotide steps present clear differences with respect to YR, RY and RR/YY steps. The YR and GG steps present large positive roll towards the major groove. The average twist values for the YR steps are in general lower than the RR and RY steps, and the difference is even more pronounced. The YR steps present relatively large and positive roll values, i.e. a local bending towards the major groove. The GC and GA steps exhibit the highest average twist values in our MD simulations, and the CG, GG, and AG the lowest. This result is in good accord with the high-twist profile (HTP) and low-twist profiles (LTP) classification in a previous analysis of crystal structures (Yanagi *et al.* 1991). The range of MD calculated values for rise, slide and shift are relatively narrow: 0.4 Å for rise, 0.7 Å in slide and 0.2 Å in shift, with standard deviations in the range of 0.1–0.2 Å. However, average values in the database indicate anti-correlated changes in rise and slide values trending as YR<RY<RR for rise and RR<RY<YR for slide.

Some conformational transitions to structures distinct from the canonical B-form such as BI/ BII and α/γ flips (see below) were found. There are strong correlations between the

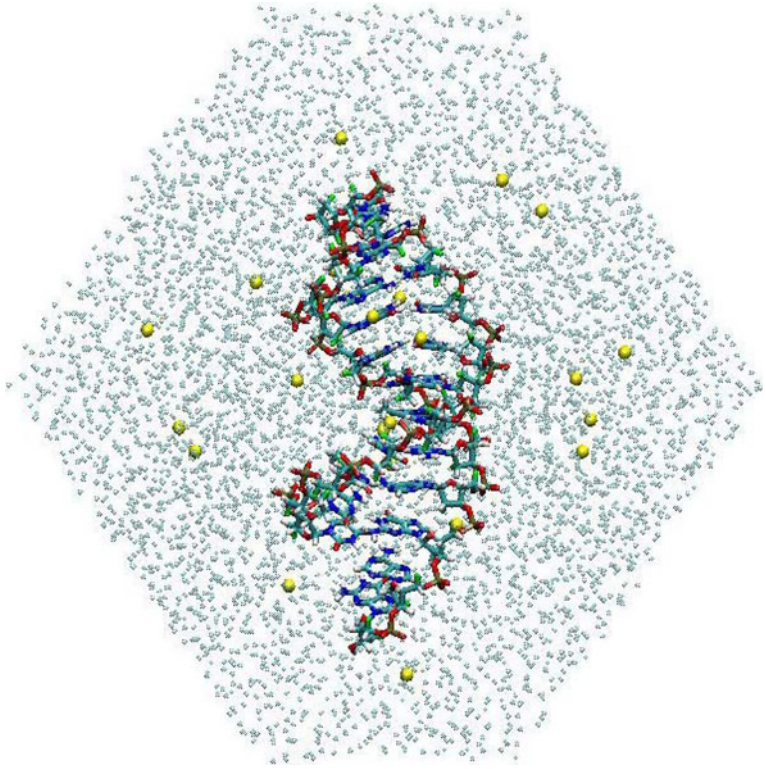

Figure 2. An image of the truncated octahedral cell used in MD simulations on DNA under periodic boundary conditions in the ABC project (Lavery *et al.* 2010). The environment of the DNA includes water molecules represented as small blue spheres and ions represented by the larger yellow spheres.

backbone conformational angles and the helicoidal properties such as twist, rise and slide. The calculated mean lifetimes of BI and BII in ABC I are 918 ps and 180 ps, respectively. Although YR steps are intrinsically flexible, they are also least affected by the neighbouring base pairs. Conversely, these steps have a significant structural impact when adjacent to a RR or RY step, which are intrinsically more rigid than YR.

Notably, a parallel study of all tetranucleotide steps based on 136 separate 10 ns MD simulations in dodecamer sequences was reported by Fujii *et al.* (2007), also based the *AMBER ff99* force field. Their results support that WATX tetramers show the most rigidity and the WYRX steps show the largest flexibility. Here, MD results were compared with data compiled from 239 protein–DNA complexes and agree with the experimental data quite well (Olson *et al.* 2006). The sequence dependent deformability was analysed on the basis of conformational entropy to study indirect readout and nucleosome positioning.

4.2 *The α/γ flip problem*

The AMBER force fields *ff94* (Cornell *et al.* 1995), *ff98* (Cheatham 3rd *et al.* 1999) and *ff99* (Wang *et al.* 2000) were parameterized when 'state-of-the-art' simulations were on the 1 ns timescale and *ab initio* quantum mechanical calculations were limited to small model systems and to moderate levels of theory. Starting MD from a canonical B or a B-form crystal structure, both of these force fields performed well in simulations on DNA in the 5–10 ns range, a typical trajectory length circa 2004. However, a then recently published MD on DNA extended to 50 ns showed a number of irreversible α/γ transitions from the canonical to a non-canonical sub-state, accompanied by severe distortions of the structure (Varnai and Zakrzewska 2004). Likewise, persistent α/γ transitions to this non-canonical state began to appear in ABC I. One would expect such transitions, but they should only occur infrequently, i.e. they should be short-lived and reversible. This turned out to be a general sequence-independent ergodic problem in the ff94 and ff99 force fields and was causing an unphysical artifact to appear in long simulation times.

In response, a re-examination of the α/γ torsional behaviour in *ff99* was carried out based on refined quantum chemical studies of the relevant part of the DNA backbone, resulting in the modification to *ff99* now known as *parmbsc0* (Perez *et al.* 2007a, b; Svozil *et al.* 2008). This was a non-trivial undertaking, since not only the behaviour of the α/γ torsional needs to be corrected but extensive testing is required to

make sure other errors were not inadvertently introduced. The *ff99/parmbsc0* force field for DNA has now been shown to produce stable MD on DNA on a ~1 μs timescale. Notably, the α/γ flip problem has not yet been seen in shorter (10 ns or less) simulations starting from canonical B-form DNA, and so this does not necessarily invalidate earlier all-atom MD studies prior to ABC. However, this artefact is expected to be more common in longer MD runs, and the ABC participants decided to rerun the entire set of 39 sequences using the *parmbsc0* force field modification. By this time, the gold standard for MD trajectories had increased to ~100 ns (Ponomarev *et al.* 2004; Varnai and Zakrzewska 2004). Thus, an ABC phase II round of simulations was set in motion with a target of generating MD on DNA for all tetranucleotide steps at the 50–100 ns level. This phase of the ABC initiative was carried out over the next 4 years

4.3 *ABC Phase II (Lavery et al. 2010) – Sequence averaged results*

Mean values averaged over all sequences considered describe a model that clearly belongs to the B-DNA family. Base pairs show relatively small average deformations, excepting a propeller which has an average deviation of 11°. There is a weak positive average inclination to the helical axis (6.8°) and moderate shift towards the major groove (1.4 Å). The inter-base pair parameters show an average rise of 3.32 Å and a twist of 32.6°. The calculated value for average twist is improved over that found with the *AMBER ff94* or *ff99* force fields without the recent bsc0 modifications to the backbone parameters. Shift and tilt are on the average quite small, but there is an overall tendency to negative slide (0.44 Å) and positive roll (3.6°). Rise and twist show large ranges (~ 4.5 Å and ~76°, respectively), reflecting large fluctuations in base pair steps where local axis bending can reach 20°. The axis bending averages 20° with a 12° standard deviation, although fluctuations up to 40–50° are not uncommon.

The MD calculated groove widths show values consistent with B-form DNA, with a narrow minor groove (6.4 Å) and a wide major groove (12.3 Å), with average depths of 4.7 Å and 6.2 Å respectively. The groove widths have considerably large thermal dispersion. The major groove depth fluctuates twice as much as that of the minor groove, with standard deviations of 2 Å and 0.8 Å, respectively. Large fluctuations in groove geometries result in backbone distortions over several base pairs. The thermal dispersion of both grooves covers a range from completely closed to 2.5 times their normal widths. This has significant implications with respect to ligand-induced changes in DNA structure, since apparent sequence-dependent changes in groove widths may not be statistically significant.

Backbone angles show that conventional states dominate for α/γ = gauche$^-$/gauche$^+$, with ε/ζ = trans/gauche (BI) and 15% of ε/ζ = gauche$^-$/trans (BII). Averaging over the entire dataset shows only 1% of non-canonical α/γ states, which supports that the repair of the α/γ flip problem has been successful, if not over-corrected. The average sugar pucker φ has a phase of 137° (C1′-exo, but near to the line with C2′-endo and an amplitude of 40°. All parameters show occasional, large deviations from their average values, often connected with incipient base pair opening. Backbone torsions show generally very large and polymodal fluctuations, but canonical α/γ and ε/ζ sub-states dominate on average.

4.4 *Results averaged over flanking sequences*

The averages and standard deviations of base pair step parameters averaged over all flanking sequences are shown in figure 3. Sequence effects at this level are seen to be relatively small overall, with statistically significant variations limited to a few steps. In particular, the YR steps show low rise, low twist and high positive roll. These steps also show much lower proportion of BII states in either strand, whereas RR steps have significant amounts (25–50%) of BII in the Watson strand and RY steps have more BII in the Crick strand. Negative rolls (bending towards the minor groove) occur for GC, GT, AT and AA. The AA and GA steps have the largest values of twist, with averages of 35.3°. Thermal fluctuations in all of the base pair steps are also quite similar, although the more flexible twist and rise of YR steps stand out somewhat. However, the difference between average values of the inter-base pair parameters for the 10 base pair steps and sequence-averaged values is not statistically significant. A set of distribution functions for base pair step parameters is shown in figure 4. Most are close to normal, but slide and twist show bimodal distributions at several steps and there is some evidence of shoulders. Bimodal distributions in twist are found for all RY steps (TG, TA, and CG). Twist is broad in GG, and GA shows a shoulder. CG essentially favours a low twist in CCGA and a high twist in ACGT, whereas ACGA is in equilibrium between these two states. These differences are linked to BI/BII transitions of the 3′ nucleotides. A bimodal distribution of base pair slide was found for RR steps.

4.5 *Analysis of tetranucleotide steps*

First nearest-neighbour sequence effects on dinucleotide steps are revealed in the analysis of tetranucleotide fragments. Tilt and roll are hardly affected, but shift and slide both show variations of up to 1 Å, but the changes occur mainly for RR steps. Rise and twist show changes of up to 0.7 Å and 18° in YR steps, especially TG and CG. The

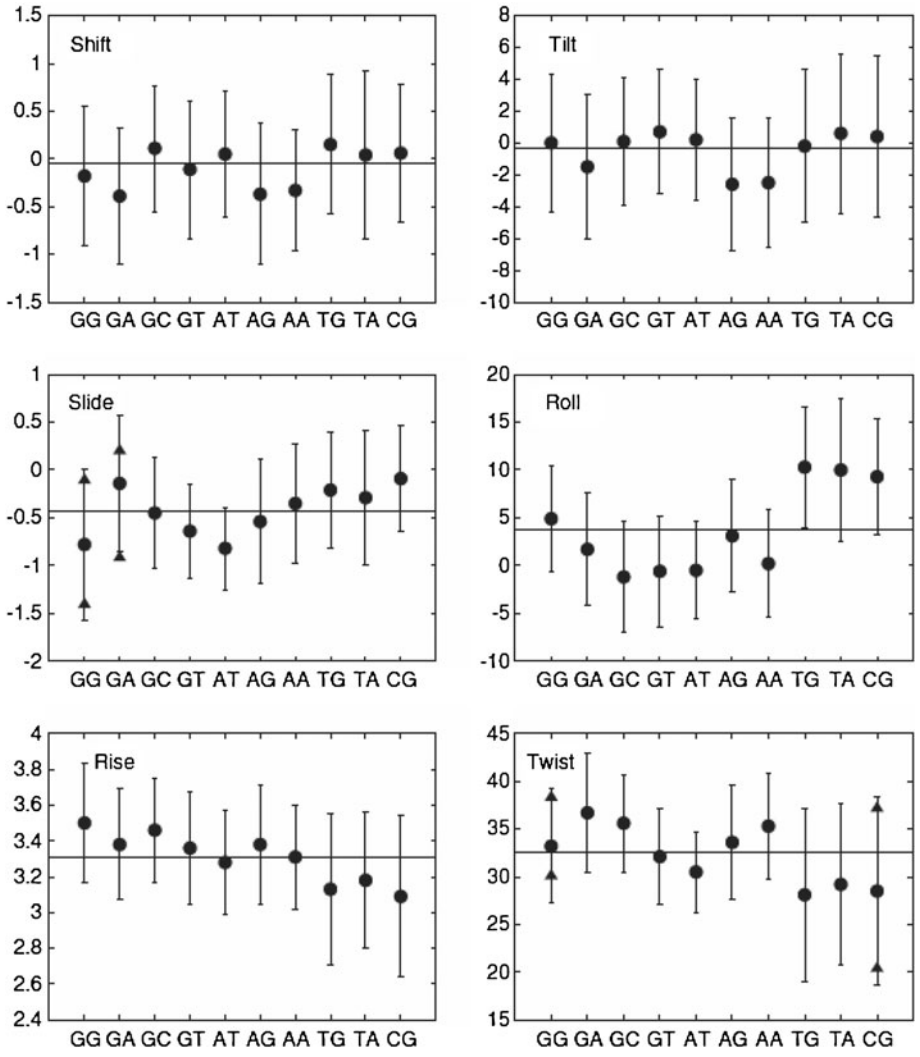

Figure 3. Average values and standard deviations for base pair step parameters from ABC II (Lavery *et al.* 2010).

absence of a significant fraction of BII states for YR steps holds in all sequence environments. However, the presence of BII for RR and RY steps depends strongly on the flanking bases. Thus, for example, AA steps have 50–80% BII in the Watson strand if the 5′-neighbour is a pyrimidine, but less than 20% if it is a purine. Similar patterns are seen for GG and AG, although in these cases a 5′ adenine suppresses BII despite a 3′ pyrimidine. For RY steps, GT shows 40–80% BII in the Crick strand if the 3′ neighbour is a purine and a comparable percentage in the Watson strand in CGCG. GT has even more specific effects, with significant BII in the Crick strand only for AGTA, AGTG, GGTA and GGTG sequences. With respect to fluctuations, the effects are relatively small for rise, tilt and roll, but more significant for shift and slide, where given environments can modify standard deviations by 60–70%. The standard deviations of twist values can double with respect to flanking sequence. A 5′-C and a 3′-A leads to high twist fluctuations for all RR and RY steps (except TG), while a 5′-T and a 3′-G leads to high twist fluctuations only for the AA step.

4.6 *Beyond tetranucleotide steps*

In ABC II, there is data on four pentanucleotide environments around all trinucleotide sequences, namely CXYZC, AXYZA, GXYZG and TXYZT. The results show that in each case the central nucleotide pair changes conformation significantly as a function of the next-nearest neighbours. A full description and understanding of this requires a more elaborate research design, but clearly next-nearest-neighbour effects on nucleotide pair conformations may be significant.

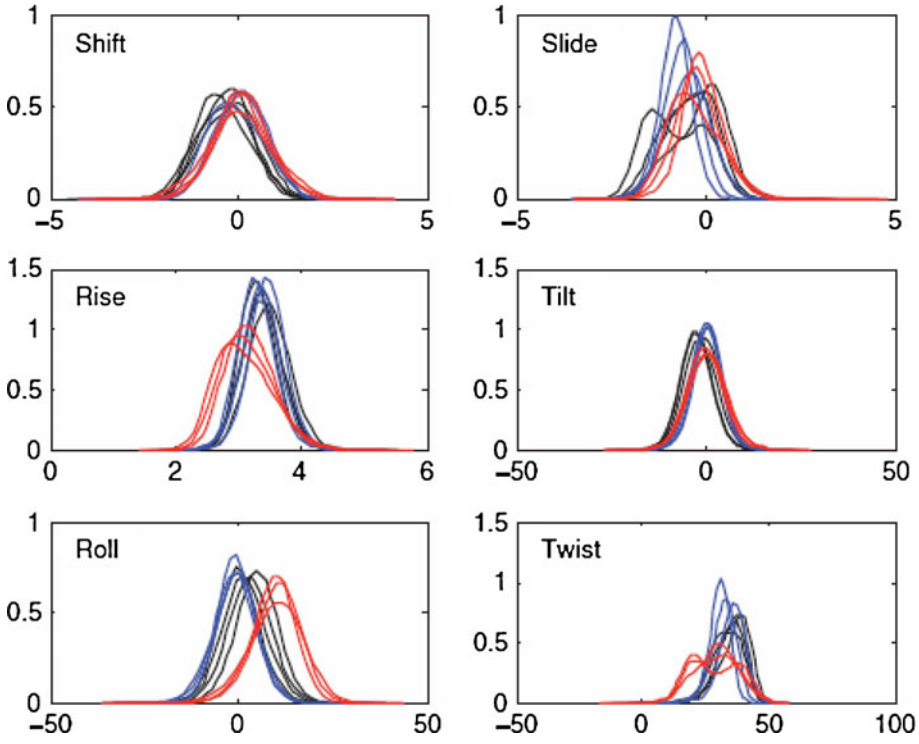

Figure 4. Distribution functions of base pair step parameters from ABC II for RR (black), YR (red) and RY (blue) (Lavery *et al.* 2010).

4.7 *Comparison with experimentally observed structures*

Direct comparison with experimental data is essential to validation of any MD on DNA. There is crystal structure and NMR data on some sequences containing tetrameric steps, but quantities directly comparable with ABC results are few and not sufficient for establishing a comprehensive vantage point (Olson *et al.* 2006). However, there are now refined X-ray and NMR structures for the Dickerson's sequence d (CGCGAATTCGCG), and all-atom MD trajectories on d (CGCGAATTCGCG) using same force field as used in ABC II are available for trajectories from 100 ns up to 1.2 μs. Selected comparisons of observed and calculated helicoidal parameters as a function of sequence are shown in figure 5. Here the experimentally observed and MD calculated *parmbsc0* results for the helical parameters roll, tilt, twist, shift slide and rise agree at >95% confidence level, which provides a good validation of the force field used in ABC II on this prototype case. Overlays of MD structures for d(CGCGAATTCGCG) shown in figure 6 convey the ensemble nature of the dynamical structure and flexibility of DNA even at the 12-mer level. Further direct comparison of ABC results and MD on DNA in general with experiment will need to be an objective of future studies in the field.

4.8 *ABC Phase III*

During the time frame of phase II of the ABC initiative, MD trajectories on d(CGCGAATTCGCG) in the range of 1 μs were reported (Perez *et al.* 2007a, b). In addition, longer sequences had become accessible to MD study including a DNA minicircle (Lankas *et al.* 2006). Thus it became feasible to consider running MD on tetranucleotide sequences at the microsecond time frame, which is adopted for ABC III. We can envisage that MD on much longer time frames will be necessary in studies of biological phenomena, and so the issue of the performance of MD on DNA simulation capabilities upon access each new decade of time will likely be necessary. It is interesting to note the similarity between the results on base pair step parameters at 100 ns and 1 μs, but this may not hold for all sequences. The Cheatham lab has now pushed MD on DNA beyond the 10 μs timescale for several sequences, using an *AMBER* implemented on the Anton machine at the Pittsburgh Supercomputing Center. The results show reversible base pair opening on microsecond timescales and convergence of structural properties in 2–3 μs. A full account of these results will be forthcoming.

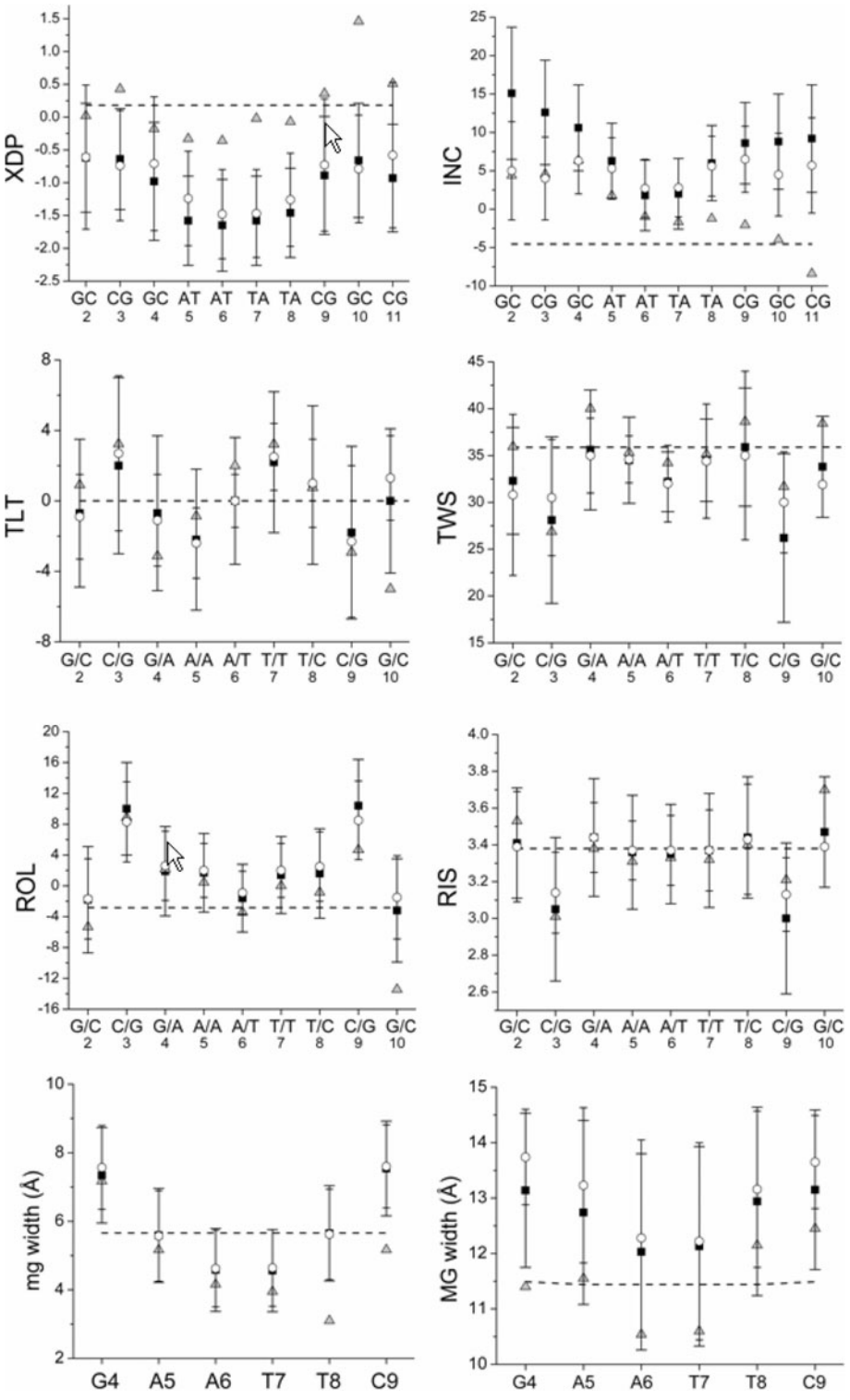

Figure 5. A comparison of base pair step parameters calculated from MD simulations on duplex d(CGCGAATTCGCG) based on parm99/parmbsc0 force field with corresponding experimentally observed crystal structure values from PDB file #1bna. Values for a canonical B-form structure are indicated by dotted lines. Triangles: PDB #1bna; squares: 100 ns MD trajectory; circles**:** 1 μsec MD trajectory (Perez *et al.* 2007).

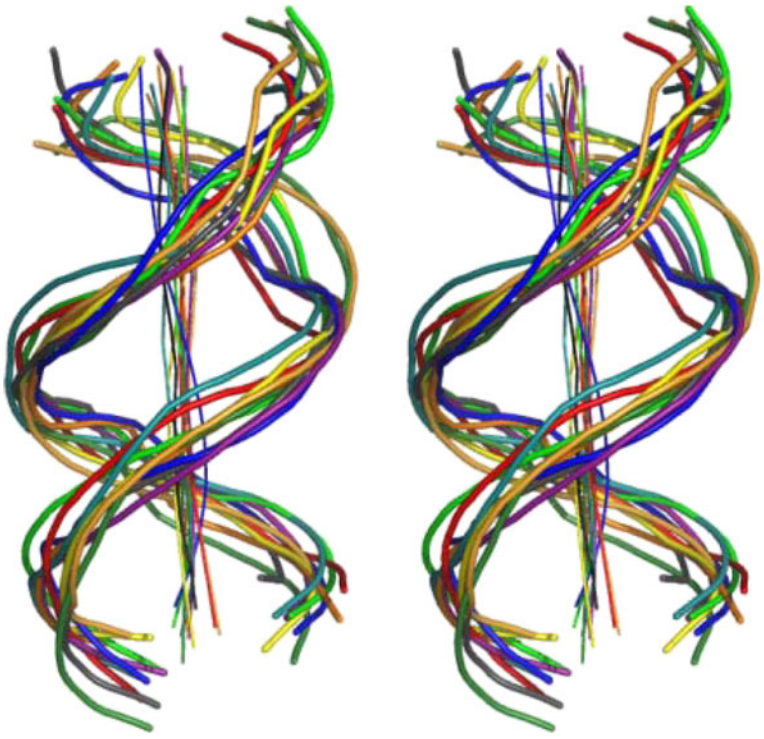

Figure 6. Superposition of MD structures from a 100 ns trajectory on d(CGCGAATTCGCG) including solvent.

5. ABC-enabled and related projects

5.1 *New programs for analysis of MD on DNA*

The analysis of the results of MD simulations on DNA involves the calculation of all the conformational, helicoidal and morphological (major and minor groove widths, axis bending and persistence length) properties of each snapshot in each the trajectory. The ABC initiative coincided with the development of a new and improved version of the analysis program *Curves* by Lavery and coworkers called *Curves+*(Lavery *et al.* 2009). In addition, a new program called *Canal* was created that reads *Curves+*analysis of an MD trajectory and organizes further analysis including statistics due to thermal dispersions. The programs are described in detail in a corresponding publication (Lavery *et al.* 2009) and are available as a Web server at *http://gbio-pbil.ibcp.fr/cgi/Curves_plus/* (Blanchet *et al.* 2011).

5.2 *Informatics*

MD carried out for ABC sequences involve running a number of simulations simultaneously. *MANYJOBS,* a python-based tool for managing this task on widely distributed computing resources, has been created by Bishop and coworkers and has been implemented on a number of supercomputing resources. The code is available able for distribution and can be obtained at *http://dna.engr.latech.edu/.*

For ease of sharing results, a suite of programs has been developed to compress MD trajectories using principal component analysis (PCA) and can be downloaded at *http://mmb.pcb.ub.es/software/pcasuite/pcasuite.html.* This technique offers a trade-off of compression ratio at the expense of losing some precision in the trajectory. PCA involves a linear transformation into modes that can be ordered with respect to the amount of motional variance. The idea here is to back transform with only the PC modes that have a significant effect on the overall dynamics. This makes a significant difference in the size of the data set. The ABC II results compressed by this method are available at *http://holmes.cancres.nottingham.ac.uk/~charlie/ABC2.*

Analysis of the large ABC datasets requires efficient methods of data management and distribution. A Web-based SQL database and a server for interactive analysis and structure prediction was developed for ABC I (Dixit and Beveridge 2006). Results from the structural analysis of all the trajectories using *Curves* and also the *3DNA* program (Lu and Olson 2003) were stored in the database, indexed with respect to nucleotide position, the time step in the

simulation, the sequence composition such as dinucleotide step, tetranucleotide step and accessible analysis via a structured query language. Assuming the individual bases to be planar and rigid, the helicoidal parameters can be employed to predict the average structure of a DNA sequence of any length. Some examples of some structure predictions are shown in figure 7. These account for some key features of sequence-dependent DNA bending noted in section 2. While this original server is no longer active, an alternative server for DNA structure prediction based on ABC II results as well as other dinucleotide models is available at *www.wesleyan.edu/bendna* (Liu and Beveridge 2001).

5.3 *Statistical significance of sequence effects*

The derived parameters calculated from MD on DNA come out as distribution functions that describe the thermal fluctuations of structures that comprise the MD ensemble. Since DNA has notable flexibility, the fluctuations may be relatively large, and thus in studies comparing MD results with experiment or investigating sequence effects within and MD simulations, it is important to consider not only the differences but also the statistical significance of the differences as well. A particular problem case arises when two means are quite different in magnitude, but the difference is not statistically significant. A number of situations like this have been identified in an MD on d(CGCGAATTCGCG) (Lee *et al.* unpublished). The mean groove widths in a DNA sequence vary by ~± 1 Å, and sequence effects here are particularly vulnerable to misinterpretation. The output from a *Canal* analysis of a MD trajectory on DNA provides all the information necessary for means testing as long the distributions are normal or nearly so. Non-normal, bimodal or polymodal distributions, of course, require special consideration.

5.4 *Base pair sequence effects on DNA solvation*

The ABC I dataset of was used by Dixit *et al.* (2012) for detailed analyses of the sequence-dependent hydration and ion atmosphere of DNA for all the 136 unique tetranucleotide steps. Proximity analysis (Mehrotra and Beveridge 1980; Mezei and Beveridge 1986; Makarov and Pettitt 2002) was employed to obtain sequence-dependent differences. Significant sequence effects on solvation and ion localization are indicated by these simulations. A representative MD-calculated hydration density for DNA is shown in figure 8. Energetic analysis of solute–solvent interactions based on proximity analysis of solvent reveals that the GC or CG base pairs interact more strongly with water molecules in the minor groove of DNA that the A-T or T-A base pairs, while the interactions of the A-T or T-A pairs in the major groove are stronger than the G-C or C-G pairs. The MD results for DNA–cation distribution functions show a structured (condensed) region of mobile ions within a radius of 10 Å from the DNA surface. This is dominated by interactions with the DNA anionic phosphates, and is thus essentially independent of sequence. However, The G-C and C-G pairs tend to associate with cations in the major groove of the

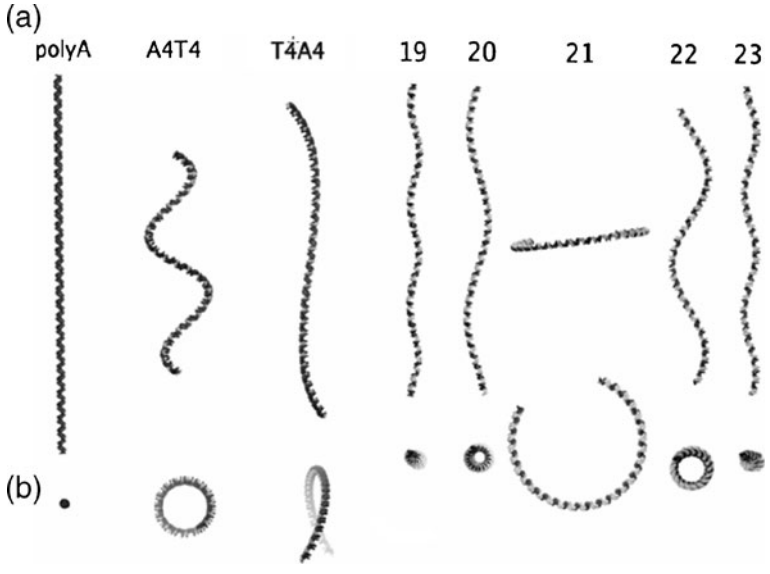

Figure 7. DNA structures for several oligonucleotide sequences predicted from ABC parameters using the server *www.wesleyan.edu/bendna.* This view is (**a**) perpendicular to the helix axis in the top set of structures, and (**b**) the same structures looking down the helix axis.

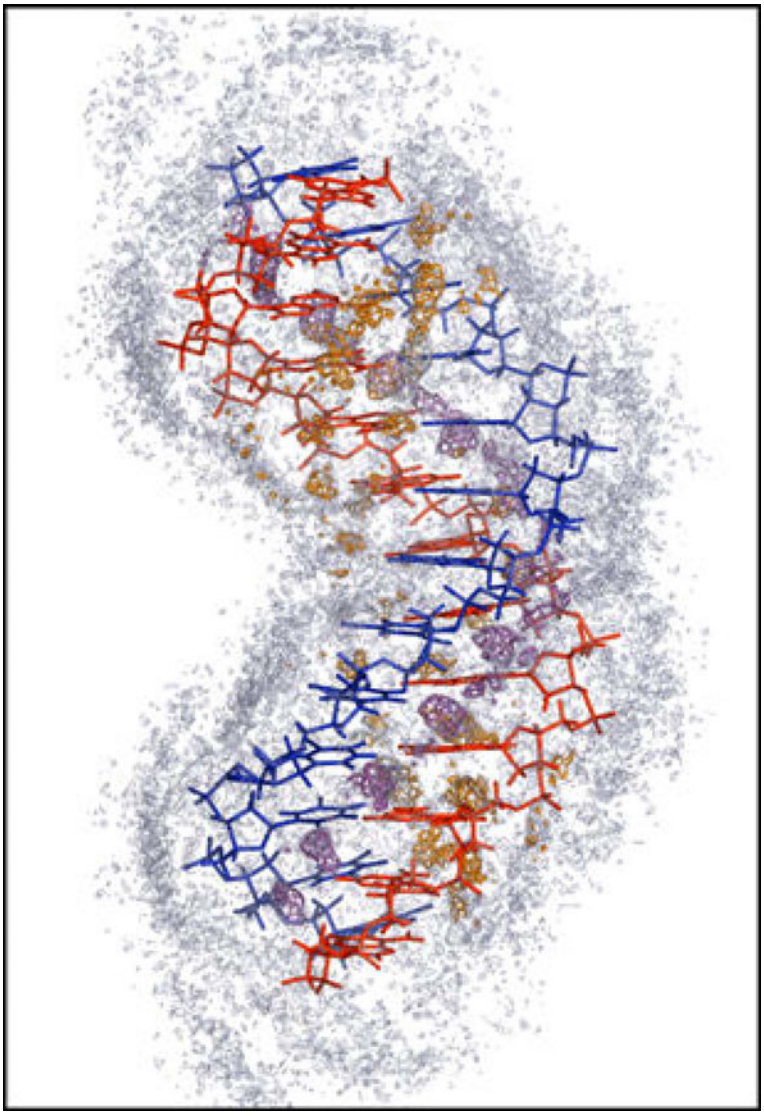

Figure 8. An example of a MD calculated hydration density (poly-A) DNA (Dixit *et al.* 2012).

DNA structure to a greater extent than the A-T and T-A pairs. Cation association is more frequent in the minor groove of A-T than the G-C pairs. However, as noted above, the fraction of counterions in the grooves is relatively low. The contribution this might make to the thermodynamics of counterion release on ligand binding to the grooves remains to be established. A comparison of solvent-accessible surface areas of the nucleotide units with results derived from analysis of crystallographic structures was found to be quite good. Time-resolved Stokes-shift experiments measure the dynamics of DNA and solvent on subnanosecond time scales, and MD was used as an aid in interpreting the results (Sen *et al.* 2009). The simulations were found to account for the magnitude and unusual power-law dynamics of the Stokes shift. Water is found to have the largest contribution to power-law dynamics, with counterions having a smaller but non-negligible contribution. The contribution to the signal of the DNA itself is only minor.

5.5 *Using ABC for parameterization of coarse-grained models*

The set of ABC MD trajectories contains a full complement of details about the time evolution of the structure and energetics for each sequence. The MD model in ABC simulations is an all-atom representation of the DNA and includes explicit consideration of solvent water, counterions and co-ions. In many instances, such a detailed model is not necessary for the interpretation of

a particular experiment or explanation of a phenomenon. Thus, quite a range of models for DNA have been devised that are based on a reduced representation or coarse-grained model of the system (DeMille *et al.* 2011). As in MD, these models typically involve functions with adjustable parameters chosen with respect to experimental data or calculations on small prototypes, and the problem then is how to choose the parameters. Thus, the ABC database can serve as a basis for parameterizing coarse-grained models. We review below several instances from the recent literature in which the ABC data has been used in this manner. Note also that a force field based on crystal structure results on DNA and protein DNA complexes has been constructed and used for normal mode calculations on DNA (Olson *et al.* 1998).

A general approach to parameterization of sequence-dependent rigid base and rigid base pair models of DNA from MD has been described (Gonzalez and Maddocks 2001; Lankas *et al.* 2009). This method treats the internal energy as a quadratic function of internal coordinates, with sequence-dependent shape, stiffness and mass parameters. What is unique about this method is incorporation of the kinetic energy as a quadratic function of the linear and angular velocities, which permits construction of a coarse-grained model for use in statistical mechanics on the full phase space of the system. An implementation of the method was parameterized on the basis of the *parm94* force field and a demonstration case on the sequence $G(TA)_7C$ was provided. A parameterization of this method based on ABC is in progress.

5.6 *Genome annotation*

A coarse-grained trinucleotide model was used for the characterization of the physicochemical properties of DNA codons from ABC results (Singhal *et al.* 2008). In this study, the hydrogen bonding involved in DNA base pairing and the base pair step stacking energies were parsed from the energetics of the ABC phase I sequences. These two datasets referred to the base pair level plus a third parameter assigned based on the conjugate rule previously proposed to account for the wobble hypothesis with respect to degeneracies in the genetic code (Crick 1966; Young *et al.* 1997a, b). The third parameter values were found to correlate well with *ab initio* MD-calculated solvation energies and flexibility of codon sequences. Assignment of these three parameters enables the calculation of the magnitude and orientation of a 'j-vector' for each codon and a cumulative 'J-vector' for a DNA sequence of any length in each of the six genomic reading frames. Analysis of 372 genomes comprising 350,000 genes using this method (figure 9) shows that the orientations of the gene and non-gene vectors were found to be well differentiated. Thus, a method based on parameters derived from the ABC database results in a clear distinction between gene and non-gene sequences comparable to knowledge-based methods for gene finding (Singhal *et al.* 2008).

5.7 *Partition functions and free energies from ABC parameters*

Energy parameters were obtained from ABC I to construct statistical mechanics partition functions and calculate thermodynamic properties of DNA sequences (Khandelwal *et al.* unpublished). The methodology derives from and parallels closely the *COREX* method used extensively by Hilser *et al.* (2006) to study the statistical thermodynamics of proteins. In this approach, called *GAREX*, individual base pairs are presumed to assume either of two states: a 'closed' Watson Crick paired form or an open state. The microstates of a DNA sequence are formed in terms of blocks of base pairs that are themselves fully closed or open. This gives rise to a finite set of 'bubble-like' microstates, which are fully enumerable in the calculation of a partition function. From the *GAREX* partition function, thermodynamic properties can be calculated for DNA sequences of any composition and length. The method has been found to describe the thermodynamic stability of oligonucleotides with a correlation coefficient of .95 over a sample of 93 cases for which melting data is available.

The *GAREX* method was designed to be extensible, and has now been used the hypothesis that genes, introns and exons differ in their thermodynamic stability. Nucleotide stability constants are calculated for genomic sequences by a moving window average over 75 base pair sequences. Some preliminary results from a more comprehensive study are shown in figure 10, which shows that (a) promoter

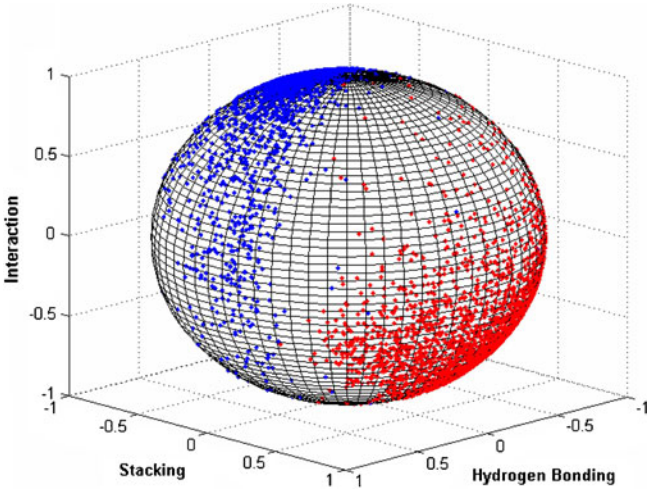

Figure 9. Results of prokaryotic gene finding based on MD on DNA (Singhal *et al.* 2008): Cumulative physicochemical codon vectors projected onto a unit sphere for 4250 genes (blue) and equal number of frame-shifted non-genes (red) in *E. Coli*.

sequences in stretches of the *E. coli* K-12 genome exhibit lower stabilities than their corresponding genes, and (b) the introns in granule-bound starch sythetase genes exhibit lower stabilities than exons. This idea that genes are the most thermodynamically stable of the sequence elements has significant implications with respect to evolutionary genomics.

6. Summary and conclusions

The ABC consortium has completed two phases of MD on DNA. The latest being a set of 50–100 ns MD simulations on all tetranucleotide steps based on the *AMBER ff99* force field with the *parmbsc0* modification. The general characteristics of the simulations are described in the context of the literature on experimental and computational modeling studies of sequence effects on DNA structure. The ABC project and database has inspired or enabled related research on analysis of MD trajectories, informatics, compression of trajectories for efficacy of distribution, DNA solvation by water and ions, parameterization of coarse-grained models as well as gene finding and genome annotation. A future phase of ABC is in progress, and involves producing MD trajectories of at least 1 μs for sequences containing all tetranucleotide steps.

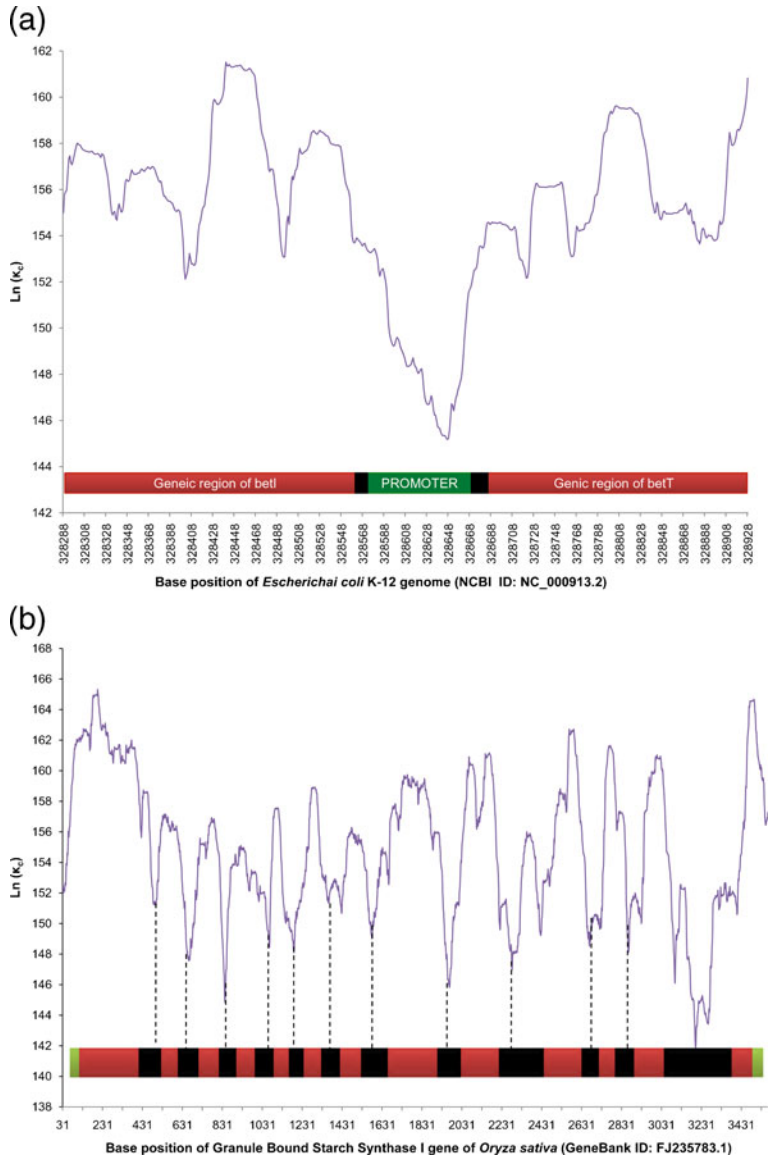

Figure 10. Genome Stability profiles based on *GAREX* parameterized with MD results (Khandelwal *et al.* unpublished): (**a**) Nucleotide stability profile for a stretch of 641 bases of *Escherichia coli* K-12 genome, encompassing betT promoter and betT gene regions along with the preceding betI region. (**b**) Nucleotide stability profile for *Granule Bound Starch Synthase I* gene of *Oryza sativa* comprising 13 exons and 12 introns.

Acknowledgements

We described in this review original research performed by the many participants in the ABC Consortium whose names and affiliations are listed in the articles published to date (Beveridge *et al.* 2004; Dixit *et al.* 2005; Lavery *et al.* 2010); special thanks to those who provided us with critical feedback on earlier drafts of this article. The authors are grateful to Professors B Jayaram, Adittya Mittal, the students and staff of SCFBio, and the organizers the Conference on Nucleic Acids in Disease and Disorder at IIT Delhi for their kind hospitality during our visit. The preparation of this article was supported in part by NIH grant NIGMS 37909 (DLB, PI), NIH GM079383, and from the NSF TeraGrid MCA01S027 (TEC III, PI). Access to high-performance computing was provided to us via the NSF Extreme Science and Engineering Discovery Environment (XSEDE) CHE10040 (DLB) and MCA01S027 (TEC III). DLB acknowledges support from Dr Joshua Boger and the WE Coffman Family. We thank Ms Elizabeth Wheatley (Wesleyan University), who contributed MD trajectories to be published elsewhere and provided help with the preparation of the figures.

References

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

Arnott S, Campbell-Smith PJ and Chandrasekaran R 1976 Atomic coordinates and molecular conformations for DNA-DNA, RNA-RNA, and DNA-RNAHelices; in *CRC Handbook of biochemistry and molecular biology* (ed) G Fasman (Cleveland: CRC Press) pp 411–422

Arthanari H, McConnell KJ, Beger R, Young MA, Beveridge DL and Bolton PH 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

Berendsen HJ C, Grigera JR and Straatsma TP 1987 The missing term in effective pair potentials. *J. Phys. Chem.* **97** 6269–6271

Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin A, Demeny T, Hsieh, SH, Srinivasan AR and Schneider B 1992 The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.* **63** 751–759

Beveridge DL and McConnell KJ 2000 Nucleic acids: theory and computer simulation, Y2K. *Curr. Opin. Struct. Biol.* **10** 182–196

Beveridge DL, Barreiro G, Byun KS, Case DA, Cheatham TE III, Dixit SB, Giudice E, Lankas F, *et al.* 2004 Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d(CpG) steps. *Biophys. J.* **87** 3799–3813

Blanchet C, Pasi M, Zakrzewska K and Lavery R 2011 CURVES+ web server for analyzing and visualizing the helical, backbone and groove parameters of nucleic acid structures. *Nucleic Acids Res.* **39** W68–73

Calladine CR 1982 Mechanics of sequence-dependent stacking of bases in B-DNA. *J. Mol. Biol.* **161** 343–352

Case DA, Cheatham TE III, Darden T, Gohlke H, Luo R, Merz Jr KM, Onufriev A, Simmerling C, *et al.* 2005 The Amber biomolecular simulation programs. *J. Comput. Chem.* **26** 1668–1688

Case DA, Darden TA, Cheatham TE III, Simmerling CL, Wang J, Duke RE, Luo R, Merz KM, *et al.* 2006 AMBER 9 (San Francisco: University of California)

Cheatham TE III, Brooks BR and Kollman PA 2001 Molecular modeling of nucleic acid structure: electrostatics and solvation. *Curr. Protoc. Nucleic Acid Chem.* Chapter 7 Unit 7.9

Cheatham TE III and Kollman PA 1996 Observation of the A-DNA to B-DNA transition during unrestrained molecular dynamics in aqueous solution. *J. Mol. Biol.* **259** 434–444

Cheatham TE III and Kollman PA 1997 Molecular dynamics simulations highlight the structural differences among DNA:DNA, RNA:RNA and DNA:RNA Hybrid Duplexes. *J. Am. Chem. Soc.* **119** 4805–4825

Cheatham TE III and Kollman PA 1997 Insight into the stabilization of A-DNA by specific ion association: spontaneous B-DNA to A-DNA transitions observed in molecular dynamics simulations of d[ACCCGCGGGT]2 in the presence of hexaamminecobalt(III). *Structure* **5** 1297–1311

Cheatham TE III and Kollman PA 1998 Molecular dynamics simulation of nucleic acids in solution: how sensitive are the results to small perturbations in the force field and environment? *Struct. Motion, Interact. Express. Biol. Macromol., Proc. Conversation Discip. Biomol. Stereodyn. 10th* **1** 99–116

Cheatham TE III and Kollman PA 2000 Molecular dynamics simulation of nucleic acids. *Annu. Rev. Phys. Chem.* **51** 435–471

Cheatham TE III and Young MA 2001 Molecular dynamics simulation of nucleic acids: Successes, limitations, and promise. *Biopolymers* **56** 232–256

Cheatham TE III, Miller JL, Fox T, Darden TA and Kollman PA 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 TE III, Cieplak P and Kollman PA 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

Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, *et al.* 1995 A second generation force field for the simulation of proteins, nucleid acids and organic molecules. *J. Am. Chem. Soc.* **117** 5179–5197

Crick F 1966 Codon–anticodon pairing: the wobble hypothesis. *J. Mol. Biol.* **19**: 548–555

Dang LX 1995 Mechanism and thermodynamics of ion selectivity in aqueous solutions of 18-Crown-6 ether: A molecular dynamics study. *J. Am. Chem. Soc.***117** 6954–6960

DeMille RC, Cheatham TE III and Molinero V 2011 A coarse-grained model of DNA with explicit solvation by water and ions. *J. Phys. Chem. B* **115** 132–142

Dickerson RE 1983 The DNA helix and how it is read. *Sci. Am.* **249** 94–111

Dixit SB and Beveridge DL 2006 Structural bioinformatics of DNA: a web-based tool for the analysis of molecular dynamics results and structure prediction. *Bioinformatics* **22** 1007–1009

Dixit SB, Beveridge DL, Case DA, Cheatham 3rd TE, Giudice E, Lankas F, Lavery R, Maddocks JH, *et al.* 2005 Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides II: Sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophys. J.* **89** 3721–3740

Dixit SB, Mezei M and Beveridge DL 2012 Studies of base pair sequence effects on DNA solvation based on all-atom molecular dynamics simulations. *J. Biosci.* **37** 399–421

El Hassan MA and Calladine CR 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 MA and Calladine CR 1997 Conformational characteristics of DNA: Empirical classifications and a hypothesis for the conformational behaviour of dinucleotide steps. *Philos. Transac. R. Soc. London Ser. A – Math. Phys. Eng. Sci.* **355** 43–100

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

Farwer J, Packer MJ and Hunter CA 2006 Prediction of atomic structure from sequence for double helical DNA oligomers. *Biopolymers* **81** 51–61

Franklin RE and Gosling RG 1953 The structure of sodium thymonucleate fibers I. The influence of water content. *Acta Cryst.* **6** 673–677

Fujii S, Kono H, Takenaka S, Go N and Sarai A 2007 Sequence dependent DNA deformability studied using molecular dynamics simulations. *Nucleic Acids Res.* **35** 6063–6074

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

Gonzalez O and Maddocks JH 2001 Extracting parameters for base-pair level models of DNA from molecular dynamics simulations. *Theor. Chem. Acc.* **106** 76–82

Hamelberg D, Williams LD and Wilson WD 2001 Influence of the dynamic positions of cations on the structure of the DNA minor groove: sequence-dependent effects *J. Am. Chem. Soc.* **123** 7745–7755

Harvey SC, Tan RK-Z and Cheatham TE III 1998 The flying ice cube: Velocity rescaling in molecular dynamics leads to violation of energy equipartition. *J. Comput. Chem.* **19** 726–740

Hilser VJ, Garcia-Moreno EB, *et al.* 2006 A statistical thermodynamic model of the protein ensemble. *Chem. Rev.* **106** 1545–1558

Joung IS and Cheatham TE III 2009 Molecular dynamics simulations of the dynamic and energetic properties of alkali and halide ions using water-model-specific ion parameters. *J. Phys. Chem. B* **113** 13279–13290

Knee KM, Dixit SB, Aitken CE, Ponomarev S, Beveridge DL and Mukerji I 2008 Spectroscopic and molecular dynamics evidence for a sequential mechanism for the A-to-B transition in DNA. *Biophys. J.* **95** 257–272

Kollman PA, Caldwell JW, Ross WS, Pearlman DA, Case DA, DeBolt S, Cheatham TE III, Ferguson D and Seibel G 2002 AMBER: A program for simulation of biological and organic molecules; in *Encyclopedia of computational chemistry* (John Wiley & Sons Ltd)

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

Lankas F, Sponer J, Langowski J and Cheatham TE III 2003 DNA base pair step deformability inferred from molecular dynamics simulations. *Biophys. J.* **85** 2872–2883

Lankas F, Sponer J, Langowski J and Cheatham TE III 2004 DNA deformability at the base pair level. *J. Am. Chem. Soc.* **126** 4124–4125

Lankas F, Lavery R and Maddocks JH 2006 Kinking occurs during molecular dynamics simulations of small DNA minicircles. *Structure* **14** 1527–1534

Lankas F, Gonzalez O, Heffler LM, Stoll G, Moakher M and Maddocks JH 2009 On the parameterization of rigid base and basepair models of DNA from molecular dynamics simulations. *Phys. Chem. Chem. Phys.* **11** 10565–10588

Lankas F, Spackova N, Moakher M, Enkhbayar P and Sponer J 2010 A measure of bending in nucleic acid structures applied to A-tract DNA. *Nucleic Acids Res.* **38** 3414–3422

Laughton CA and Harris SA 2011 The atomistic simulation of DNA. *Wiley Interdisc. Rev. Comput. Mol. Sc.* **1** 590–600

Lavery R, Moakher M, Maddocks JH, Petkeviciute D and Zakrzewska K 2009 Conformational analysis of nucleic acids revisited: Curves+. *Nucleic Acids Res.* **37** 5917–5929

Lavery R, Zakrzewska K, Beveridge DL, Bishop TC, Case TA, Cheatham IIIT, Dixit S, Jayaram B, *et al.* 2010 A systematic molecular dynamics study of nearest neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res.* **38** 299–313.

Lefebvre A, Mauffret O, Hartmann B, Lescot E and Fermandjian S 1995 Structural behavior of the CpG step in two related oligonucleotides reflects its malleability in solution. *Biochemistry* **34** 12019–12028

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

Liu Y and Beveridge DL 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. Strut. Dyn.* **18** 505–526

Lu XJ and Olson WK 2003 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **31** 5108–5121

Mackerell Jr AD and Nilsson L 2008 Molecular dynamics simulations of nucleic acid-protein complexes. *Curr. Opin. Struct. Biol.* **18** 194–199

Madhumalar A and Bansal M 2003 Structural insights into the effect of hydration and ions on A-tract DNA: a molecular dynamics study. *Biophys. J.* **85** 1805–1816

Makarov V and Pettitt BM 2002 Solvation and hydration of proteins and nucleic acids: A theoretical view of simulation and experiment. *Acc. Chem. Res.* **35** 376–384

Manning GS 1978 The molecular theory of polyelectrolyte solutions with applications to the electrostatic properties of polynucleotides. *Quart. Rev. Biophys.* **11** 179–246

Matsumoto A and Olson WK 2002 Sequence-dependent motions of DNA: a normal mode analysis at the base-pair level. *Biophys. J.* **83** 22–41

McCammon JA, Gelin BR and Karplus M 1977 Dynamics of folded proteins. *Nature* **267** 585–590

McConnell KJ and Beveridge DL 2000 DNA structure: what's in charge? *J. Mol. Biol.* **304** 803–820

McConnell KJ and Beveridge DL 2001 Molecular dynamics simulations of B-DNA: sequence effects on A-tract-induced bending and flexibility. *J. Mol. Biol.* **314** 23–40

McConnell KM, Nirmala R, Young MA, Ravishanker G and Beveridge DL 1994 A nanosecond molecular dynamics trajectory for a BDNA double helix: Evidence for substates. *J. Am. Chem. Soc.* **116** 4461–4462

Mehrotra PK and Beveridge DL 1980 Structural analysis of molecular solutions based on quasi-component distribution functions. application to $[H_2CO]_{aq}$ at 25°C. *J. Am. Chem. Soc.* **102** 4287–4294

Mezei M and Beveridge DL 1986 Structural Chemistry of biomolecular hydration: The proximity criterion. *Methods Enzymol.* **127** 21–47

Miller JL, Cheatham TE III and Kollman PA 1999 Simulation of nucleic acid structure; in *Oxford handbook of nucleic acid structure* (ed) S Neidle (Oxford, New York: Oxford University Press) pp 95–115

Noy A, Pérez A, Lankas F, Luque FJ and Orozco M 2004 Relative flexibility of DNA and RNA: A molecular dynamics study. *J. Mol. Biol.* **343** 627–638

Olson WK and Zhurkin VB 2000 Modeling DNA deformations. *Curr. Opin. Struct. Biol.* **10** 286–297

Olson WK, Gorin AA, Xizng-Jung L, Hock LM and Zhurkin VB 1998 DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA* **95** 11163–11168

Olson WK, Colasanti A, Li Y, Ge W, Zheng G and Zhurkin V 2006 *DNA simulation benchmarks as revealed by X-ray structures. in computational studies of RNA and DNA* (Netherlands: Springer) pp 235–257

Orozco M, Noy A and Pérez A 2008 Recent advances in the study of nucleic acid flexibility by molecular dynamics. *Curr. Opin. Struct. Biol.* **18** 185–193

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

Pastor N and Weinstein H 2001 Protein-DNA interactions in the initiation of transcription: The role of flexibility and dynamics of the TATA recognition sequence and the TATA box binding protein; in *Theoretical and computational chemistry* (ed) AE Leif (Amsterdam, Elsevier) pp 377–407

Pearlman DA, Case DA, Cadwell JW, Ross WS, Cheatham TE III, Debolt S, Ferguson D, George S, *et al.* 1995 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. *Comp. Phys. Com.* **91** 1–41

Perez A, Luque FJ and Orozco M 2007 Dynamics of B-DNA on the microsecond time scale. *J. Am. Chem. Soc.* **129** 14739–14745

Perez A, Marchan I, Svozil D, Sponer J, Cheatham TE III, Laughton CA and Orozco M 2007 Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys. J.* **92** 3817–3829

Perez A, Luque FJ and Orozco M 2012 Frontiers in molecular dynamics simulations of DNA. *Acc. Chem. Res.* **45** 196–205

Ponomarev SY, Thayer KM and Beveridge DL 2004 Ion motions in molecular dynamics simulations on DNA. *Proc. Natl. Acad. Sci. USA* **101** 14771–14775

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

Rueda M, Cubero E, Laughton CA and Orozco M 2004 Exploring the counterion atmosphere around DNA: what can be learned from molecular dynamics simulations? *Biophys. J.* **87** 800–811

Saenger W 1984 Principles of nucleic acid structure (New York: Springer Verlag)

Sen S, Andreatta D, Ponomarev SY, Beveridge DL and Berg MA 2009 Dynamics of water and ions near DNA: comparison of simulation to time-resolved stokes-shift experiments. *J. Am. Chem. Soc.* **131** 1724–1735

Sherer EC, Harris SA, Soliva R, Orozco M and Laughton CA 1999 Molecular Dynamics Studies of DNAA-Tract Structure and Flexibility. *J. Am. Chem. Soc.* **121** 5981–5991

Singhal P, Jayaram B, Dixit SB and Beveridge DL 2008 Prokaryotic gene finding based on physicochemical characteristics of codons calculated from molecular dynamics simulations. *Biophys. J.* **94** 4173–4183

Sprous D, Young MA and Beveridge DL 1998 Molecular dynamics studies of the conformational preferences of a DNA double helix in water and in an ethanol/water mixture: Theoretical considerations of the A/B transition. *J. Phys. Chem.* **102** 4658–4667

Sprous D, Young MA and Beveridge DL 1999 Molecular dynamics studies of axis bending in $d(G_5\text{-}(GA_4T_4C)_2\text{-}C_5)$ and $d(G_5\text{-}(GT_4A_4C)_2\text{-}C_5)$: Effects of sequence polarity on DNA curvature. *J. Mol. Biol.* **285** 1623–1632

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

Svozil D, Sponer JE, Marchan I, Perez A, Cheatham TE III, Forti F, Luque FJ, Orozco M and Sponer J 2008 Geometrical and electronic structure variability of the sugar-phosphate backbone in nucleic acids. *J. Phys. Chem. B* **112** 8188–8197

Tidor B, Irikura KK, Brooks BR and Karplus M 1983 Dynamics of DNA oligomers. *J. Biomol. Struct. Dyn.* **1** 231–252

Ulyanov NB and Zhurkin VB 1984 Seqence-dependent anisotropic flexibility of B-DNA: A conformational study. *J. Biomol. Struct. Dyn.* **2** 361–385

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

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

Wang J, Cieplak P and Kollman PA 2000 How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput Chem.* **21** 1049–1074

Watson JD and Crick FH C 1953 A structure for deoxyribonucleic acid. *Nature* **171** 737–738

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

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

Young, MA and Beveridge DL 1998 Molecular dynamics simulations of an oligonucleotide duplex with adenine tracts phased by a full helix turn. *J. Mol. Biol.* **281** 675–687

Young MA, Ravishanker G, Beveridge DL and Berman HM 1995 Analysis of local helix bending in crystal structures of DNA oligonucleotides and DNA-protein complexes. *Biophys. J.* **68** 2454–2468

Young MA, Jayaram B and Beveridge DL 1997 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 MA, Ravishanker G and Beveridge DL 1997 A 5-nanosecond molecular dynamics trajectory for B-DNA: Analysis of structure, motions and solvation. *Biophys. J.* **73** 2313–2336

Zhurkin VB 1985 Sequence-dependent bending of DNA and phasing of nucleosomes. *J. Biomol. Struct. Dyn.* **2** 785–804

Zhurkin VB, Tolstorukov MY, Xu F, Colasanti AV and Olson WK 2005 Sequence dependent variability of B-DNA: An update on bending and curvature; in *DNA conformation and transcription* (ed) T Ohyama Springer US pp 18–34 *http://www.eurekah.com*