Root Mean Square Deviation Probability Analysis of Molecular Dynamics Trajectories on DNA

Surjit B. Dixit, Sergei Y. Ponomarev, and David L. Beveridge*

Department of Chemistry and Molecular Biophysics Program, Hall-Atwater Laboratories, Wesleyan University, Middletown, Connecticut 06459

Received November 11, 2005

The comparison and detection of the commonalities and differences in multiple structural ensembles is an important step in the use of molecular simulations to gain insight into the conformation and dynamics of complex biomacromolecules. While the average structure is often employed as the representative of an ensemble of structures in such comparisons, dynamic molecular systems with multiple conformational substates call for a more accurate representation that captures the complete dynamical range of the ensemble. We present a probability analysis procedure based on the root-mean-square differences among the structural ensembles that efficiently and accurately performs the relevant comparison.

INTRODUCTION

Molecular dynamics (MD) computer simulation is now capable of providing computational models of DNA oligonucleotides in solution that compare reasonably well with corresponding experimentally determined structures from crystallography or NMR spectroscopy.[1] These studies are typically based on one or at most several individual MD trajectories. Recent developments in high performance computing now make it possible to generate many trajectories in a single study for more extensive comparisons required for validation of MD force fields and other methodological aspects of simulation.[2] Such comparisons to date have relied on reduced measures of difference, particularly the root-mean-square deviation (RMSD) of one structure from another computed following optimal structural alignment. When comparing different MD simulations, the ensemble average is typically taken as representative of the individual structural (snapshots) generated in the trajectory. However, in the case of highly flexible structures such as DNA, we have noted that the ensemble average structure may not be adequately representative of the ensemble. In this article, we demonstrate the nature of the problem and propose an alternative method, "RMSD probability analysis", *P*(RMSD). Results are provided that indicate this is more informative way of comparing MD simulations, particularly when the native dynamical structure involves substates. *P*(RMSD) analysis also obviates the problem of treating MD distributions with a statistical parameter derived from normal distributions when their behavior is not in fact normal.

The simulations used to test the *P*(RMSD) approach are on the DNA sequence $d(CGCGAATTCGCG)_2$ at room temperature. We have recently reported an MD of ∼100 ns of MD on this sequence including explicit solvent and using an AMBER (AM) force field.[3] A 3 ns MD trajectory was made available on the same sequence using the recent CHARMM (CH) force field for DNA.[4] Results on the comparison of AM and CH MD models of DNA with each other and with experiment are provided. In addition, we have performed several new MD simulations on this sequence using various Generalized Born (GB) methods for approximating solvation and use this as a basis for comparing MD on DNA based on an explicit solvent model versus a continuum dielectric model. The demonstrations here all involve MD on DNA, but the ideas are generally applicable to all simulations on biological macromolecules.

BACKGROUND

The most common current method of comparing any two structures involves the optimal alignment of either full structures or prescribed substructural elements using a least-squares fit procedure, followed by computation of the RMSD for some or all corresponding atoms.[5,6] In the analysis of MD, ensemble averaged structures are obtained by optimally superimposing a representative set of structural snapshots and calculating mean values from the arithmetic sum of each of the Cartesian coordinates for each atom divided by number of structures. The comparison of results from two different MD trajectories on the same molecule involves calculating the RMSD between the ensemble averaged structures of the two trajectories.

The hidden assumption in comparing ensemble averaged structures from MD is that the averaged structure is a sufficiently good representation of the ensemble. One problem is that the representations obtained from this process are not guaranteed to conform to the basic physical and chemical constraints of molecular structure. This problem is readily addressed numerically by simple application of a few cycles of energy minimization, letting the force field parametrized on proper physical models of molecules take care of the problem. However, there is still no guarantee that such an average structure is really representative of the ensemble, especially if the system being studied is highly flexible and exhibits non-Gaussian distributions and/or features multiple conformational substates. The problem is exemplified in Figure 1, which presents the RMSD of a 3 ns long DNA trajectory with reference to the average

* Corresponding author e-mail: dbeveridge@wesleyan.edu.

10.1021/ci0504925 CCC: $33.50
Published on Web 04/26/2006

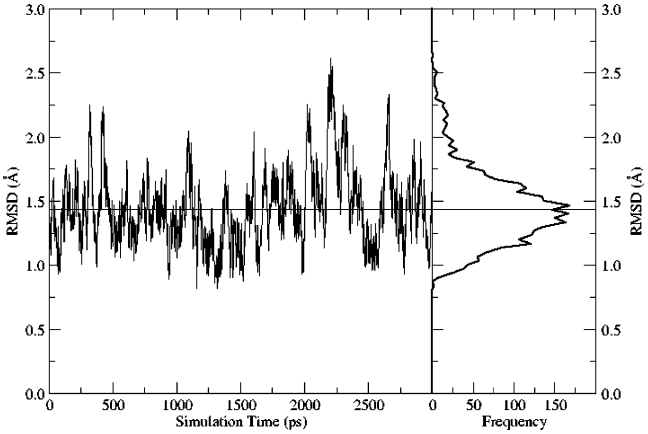

Figure 1. Plot to the left presents the RMSD of a 3 ns long MD trajectory of the d(CGCGAATTCGCG)$_2$ sequence with reference to the average structure of the same trajectory. The graph to the right presents the normalized probability distribution of the RMSD data. The straight line passing through 1.4 Å in the two graphs is the numerical average of the RMSD data.

structure derived from the trajectory. The plot on the right presents the frequency profile of this RMSD data. The data presented here shows that in comparison to the statistical concept of a mean, in fact none of the snapshot structures coincide with the average and the distribution profile of the RMSD values is clearly non-Gaussian.

A calculated RMSD is essentially a radial vector in structure space with a length r given by the absolute magnitude of the RMSD. The nature of a radial problem is that the larger the radius, the more volume of configurational space there is between a given r and $r+dr$. At large r, both similar and quite different structures can be captured by the same value of RMSD, and useful information about comparisons is compromised. Methods which rely on relatively small RMSD values provide a more reliable measure of difference. Another important problem with regard to the use of RMSD comes up when two or more structural substates are present as a result of the intrinsic flexibility of the molecule. One needs a method that can represent these dynamical aspects in a reliable manner without losing information.

Given this concern, the specific questions we address in this article are the following: (a) What is a proper approach to comparing two (or more) MD simulations based on the actual structural data rather than average structures? (b) How does this apply in the comparison of MD results based on two (or more) force fields or physical models? (c) How can this approach be useful in comparing the results of MD simulation with experiment? The analysis procedure described here has been used extensively in recent research efforts to identify the effect of sequence context on the dynamical structures of the 10 unique dinucleotide steps in DNA.[2] In particular, the project called for a rigorous comparison of structures from multiple trajectories involving the 136 unique tetranucleotide sequences to estimate the effect of the flanking base pairs on the structural behavior of the central dinucleotide of each of the tetranucleotides. Apart from providing insight on the occurrence of structural substates, the method also helped us to quantitatively classify the large data set on the basis of differences in structure and flexibility.

CALCULATIONS

MD Simulations. The AM MD trajectory on d(CGCGAATTCGCG)$_2$ DNA including explicit solvent used for analysis in this project is described in detail by Ponomarev et al.[7] The system for simulation comprised of the dodecamer DNA solvated with 3949 TIP3P water molecules[8] together with 22 Na^+ cations[9] in a rectangular cell with periodic boundary conditions. Long-range electrostatic interactions were treated by particle mesh Ewald.[10] The simulation was performed using the AMBER 7.0 suite of programs[11] and the parm94 force field.[3] The starting point for the simulation was the canonical B-form structure. The reported simulation was performed in an NPT ensemble at 300 K for a total of 60 ns. A full description of the dynamical structure of the DNA based on this force field has been provided by Cheatham and Kollman[12] and Young et al.[13]

The CH all atom explicit solvent trajectory on d(CGCGAATTCGCG)$_2$ DNA sequence was provided to us by Prof. Alex MacKerell, and the simulation is described in the literature by MacKerell and co-workers.[4,14,15] This CH trajectory is based on simulation protocol similar to the one used with AM. The system for simulation comprised of DNA and TIP3P water molecules[8] together with 22 Na^+ cations[16] in a rectangular cell with periodic boundary conditions. The starting point for the simulation was a canonical B-form structure of d(CGCGAATTCGCG)$_2$. The simulation was performed in an NPT ensemble at 300 K for a total of 3 ns with the long-range electrostatics being treated using the PME[17] in the CHARMM suite of programs[18] with CHARMM27 force field[4] for DNA. The detailed structural analysis of this MD is reported in the original articles.[4,14,15] Since the total length of the available AM MD (60 ns) and CH MD trajectories (3 ns) differ, we have based this analysis on 3 ns and use a segment of the AM MD from midway through the trajectory. Our previous analysis[7] of the convergence profile of the AM DNA simulation has shown that all the internal structural parameters of DNA converge rapidly within the time scale of about 500 ps leaving most of the trajectory well equilibrated for production analysis. MacKerell et al.[4,14,15] indicate that the CH trajectory we employ is also equilibrated.

Recently, the idea that MD on DNA might be carried out efficiently using a model in which the solvent water is not treated explicitly but represented as a polarizable dielectric continuum (implicit solvent) has been advanced.[19] A number of formulations for molecular simulations using implicit solvent based on the Generalized Born (GB) model are currently available. This topic has been reviewed by Bashford and Case.[20] There have been few papers on the performance of GB MD on DNA, although a comparison with regard to proteins has been published recently.[21] One of the objectives of GB MD is to offer a computationally efficient model approximating to all-atom MD. Comparison of trajectories obtained from the GB and explicit solvent simulations is thus a point of validation, and we provide some leading results on this matter. We consider three different parameter sets implemented in AMBER version 8. The GB methods are all based on the Hawkins, Cramer, and Truhlar model.[22,23] IGB1 is based on parameters of Tsui and Case,[19] and IGB2 and IGB5 are based on different sets of empirical parameters developed by Onufriev, Bashford, and Case.[24]

***P*(RMSD) Calculations**. We investigate in this project an analysis procedure based *only on the actual snapshots* generated in the MD ensemble. This involves an RMSD comparison of every structure in the trajectory with every other structure in the same trajectory, i.e., performing $n*(n\text{-}1)/2$ RMSD comparisons where n is the number of snapshots in the trajectory. Similarly, the comparison of two different MD simulations would involve comparing every structure of the first trajectory to each snapshot in the second and so on. The elements of the analysis we pursue is presented in a matrix form and has been referred to as a 2D-RMSD.[25] The characteristics of a 2D-RMSD plot have been useful in the identification of substates.[26]

RMSD calculations as described up to this point are based on the Cartesian coordinates of two structures. However the basic idea may be readily extended to derived structural parameters. An RMSD comparison of DNA structures in terms of conformational angles can be calculated as

$$\mathrm{RMSD} = \sqrt{\frac{1}{n}\sum(\theta_{ij1} - \theta_{ij2})^2}$$

where the summation runs over all the positions (i) in the two strands (j) of the DNA, θ is any of the sugar phosphate backbone conformational angles of DNA, and the subscripts 1 and 2 denote the corresponding values of the angle in structures 1 and 2, respectively. The value of n in the denominator is the total number of variables included in the RMSD calculation i.e., the product of the number of positions (i), the number of strands (j), and the number of angular parameters considered. In the context of proteins, one could execute such an analysis in terms of the backbone and side-chain conformational angles. In such reduced descriptions, one could easily deconvolute the origin of the differences in the structure at the individual parameter level. Angular RMSD are useful in comparing the results of explicit solvent MD with MD using a GB solvent to pinpoint any differences in the dynamics.

The particular form of analysis we emphasize in this article involves the generation of a plot of the probability of observing a given RMSD between all pairs of snapshots in the MD simulation, which we call an RMSD probability analysis, *P*(RMSD). It is of interest to distinguish two cases at this point: (a) the "P_{intra}(RMSD)" in which the RMSD of all structures with all other structures in a single MD trajectory are analyzed and (b) the "P_{inter}(RMSD)" in which the structures from one MD ensemble are compared with those of another. Inspection of P_{intra}(RMSD) plots provides information on the extent of thermal motions or dynamical range of the MD model. The presence of multiple conformational substates in the trajectory would show up as multimodal distributions of *P*(RMSD). For two simulations in which the thermal fluctuations and flexibility are similar, the *P*(RMSD) distributions should be similar if both the trajectories have converged.

Determining an index of the similarity of two MD simulations involves application of statistical inference in the comparison of the respective *P*(RMSD) distributions, using statistical tests to determine the confidence level with which it may be inferred that the two sets of structures have been drawn from the same general population. The standard statistical test for the similarity of two distributions is the χ^2 test.[27] The calculated χ^2 values express the confidence level at which the null hypothesis that the two *P*(RMSD) distributions are equivalent is regularly true. The χ^2 test is ideally applicable to a categorical data set while the *P*(RMSD) distributions are not of that genre. An alternative, more rigorous information theoretic approach applicable in the case of such complex distributions is to calculate the "Kullback-Liebler (KL) Distance"[28] D_{KL}, which is a measure of the divergence between a "true" probability distribution, p, and a "target" probability distribution, q. For discrete probability distributions, $p = \{p_1, ..., p_n\}$ and $q = \{q_1, ..., q_n\}$, D_{KL} is defined as

$$D_{\mathrm{KL}}(p,q) = \sum_i p_i \log_2\left(\frac{p_i}{q_i}\right)$$

The value D_{KL} is always positive and equal to zero only if $p_i = q_i$. D_{KL} is not, in general, symmetric, and hence we employ the mean of $D_{\mathrm{KL}}(p,q)$ and $D_{\mathrm{KL}}(q,p)$. This equation based on expected log likelihood ratio between the two distributions is a metric of the relative entropies and can be viewed as the bits of information required to convert one distribution to another. Such an approach to compare the RMSD probability distributions provides a single index for examining the difference between two MD results and avoids the necessity of working with possibly problematic average structures. We have adopted the KL method in the comparisons of DNA structures in this project.[2]

RESULTS AND DISCUSSION

***P*(RMSD) Comparison of MD on DNA based on AMBER and CHARMM Force Fields**. 2D RMSD plots for the AM and CH simulations on d(CGCGAATTCGCG) duplex used in this study are shown in Figure 2. Each map represents a matrix of RMSD values between structures taken at constant time intervals from 3 ns trajectories. Note that the RMSD matrix is symmetrical in the upper and lower triangle. The bar on the right defines high and low RMSD values in gray scale. The self-comparisons of AM (bottom left quadrant) and CH (top right quadrant) have relatively low RMSD values in contrast to those comparing AM to CH trajectory structures (top left quadrant) indicating, as expected that the fluctuations of RMSD values between individual structures of the same trajectory are relatively low compared to those obtained from two different force fields.

The essence of *P*(RMSD) analysis is to cast the 2D RMSD results into probability distributions. *P*(RMSD) plots from each quadrant of the 2D RMSD plot are shown in Figure 3. The mean RMSD value for the AM structures compared to other structures in the same trajectory, denoted AM/AM, is 2.0 ± 0.5 Å. The corresponding value for the "intra" comparison of the CH DNA model, denoted CH/CH, is 1.8 ± 0.4 Å. The minimum and maximum RMSD values for AM/AM distribution are 0.7 and 4.3, while those for CH/CH are 0.7 and 3.8, respectively. Thus both the mean and width of the *P*(RMSD) distribution in the case of AM trajectory is slightly larger than that for CH, indicating that the dynamical range of structures in the AM MD of DNA is larger than in the CH model. The distribution of RMSD values for AM/CH cross comparison ranges from ~2.5 to 5.0 Å. This index indicates the extent to which the MD model obtained from the AM simulation is significantly different

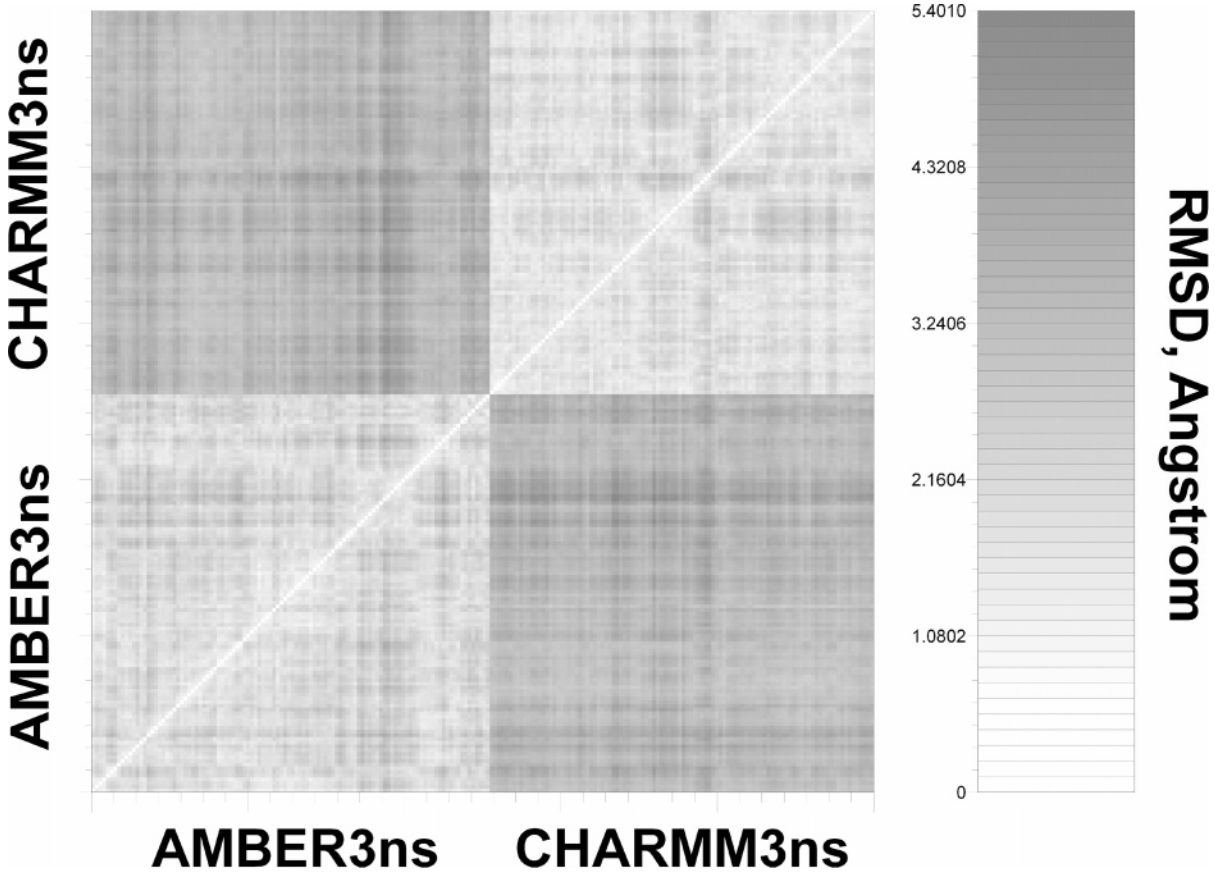

Figure 2. 2D RMSD analysis of MD on DNA based on 3 ns trajectories using the AMBER and CHARMM force fields.

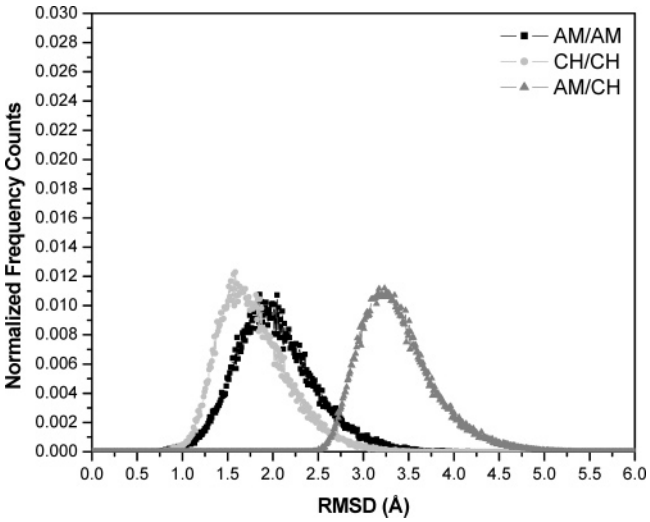

Figure 3. *P*(RMSD) analysis of MD trajectory on d(CGCGAATTCGCG) duplex DNA: (a) AM/AM self-comparison (black squares), (b) CH/CH self-comparison (grey circles), and (c) comparison of results of AM with the CH force fields, AM/CH (grey triangles).

from that of CH. The value of the net RMSD per se does not indicate exactly how the two models differ, but further decompositions can reveal this. The broadening of the *P*(RMSD) distribution and the presence of an incipient tail is indicative of the presence of substates.

Figure 4 shows the *P*(RMSD) distribution curves based on two AM MD trajectories of $d(CGCGAATTCGCG)_2$, one beginning with an A-form conformation in 85% v/v ethanol/water mixture (denoted "eth") and one from the B-form in water (denoted "wat"). The P_{intra}(RMSD) of the A-form structure in ethanol peaks around 1.5 Å RMSD, while the P_{intra}(RMSD) for the B-form in water peaks about 2 Å. The P_{inter}(RMSD) between the A and B forms peaks about 5 Å. Notice a small population in the P_{intra} of the A_{eth}/A_{eth} distribution near RMSD value of ~5 Å indicative of the presence of a small population of B-type structures in the A-form simulation. While the RMSD between the predominant state (A-form) and substate (B-form) is sufficiently separated in this case to clearly observe the subpopulations, this is often not so clear and manifests as a broadening of the distribution. A better resolution of *P*(RMSD) for substates is found when the structures are described in the internal coordinates rather than the Cartesian coordinates (see below). Supporting the idea that the width and height of the *P*(RMSD) distribution curve can be employed as a measure of flexibility in the structural ensemble, we observe *P*(RMSD) curve of the A-form structures to be more sharply peaked and narrower than the B-form distribution, in accord with the experimental observations that the A-form structures are more compact and thus stiffer than B-form structures.[29]

We next address the issue of which of the two MD models for $d(CGCGAATTCGCG)_2$ is closer to experimentally determined structures for this sequence from X-ray crystallography and NMR spectroscopy. There are 6 different high-resolution crystal structures of intact d(CGCGAATTCGCG) duplex in the nucleic acid databases (PDB ID: 1BNA, 1FQ2, 2DAU, 355D, 2BNA, 428D).[30−34] The *P*(RMSD) between 6 XRAY structures and all the AM MD structures is shown in Figure 5a. Corresponding comparisons for CH are shown

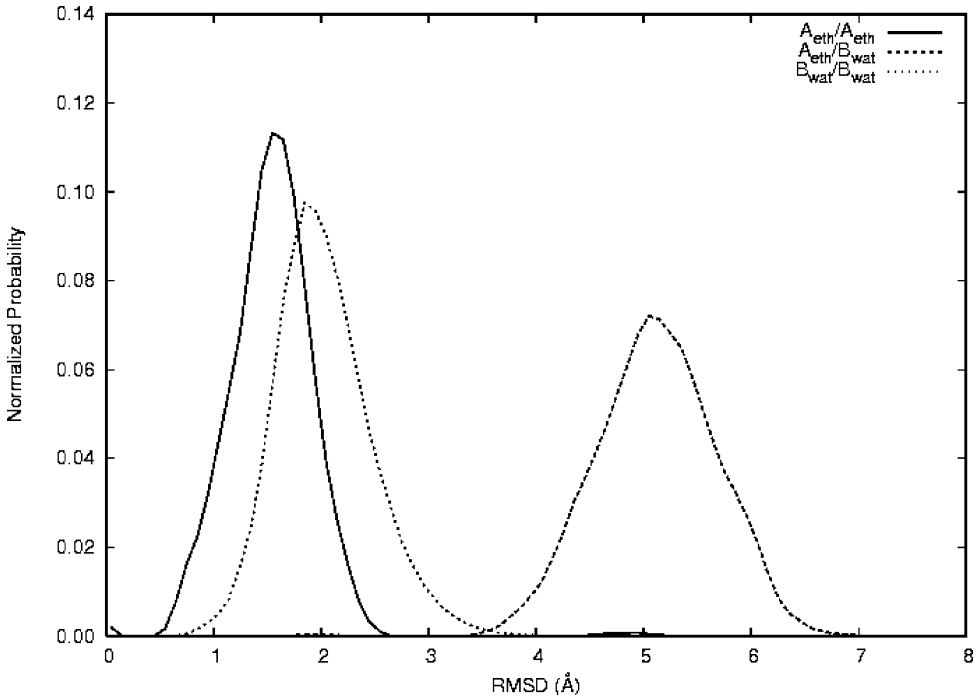

Figure 4. Normalized *P*(RMSD) curves based on the comparison of the A and B forms of DNA adopted by the d(CGCGAATTCGCG)$_2$ sequence in ethanol and water. The solid curve presents the P_{intra}(RMSD) of A form in ethanol, the dotted line is P_{intra}(RMSD) of the B-form in water, and the P_{inter}(RMSD) between the DNA structures in ethanol and water is shown with the dashed line. Notice the small population in the A_{eth}/A_{eth} curve near RMSD of 5 Å, in the region of the A_{eth}/B_{wat} distribution indicating the presence of a small subset of structures similar to the B-form.

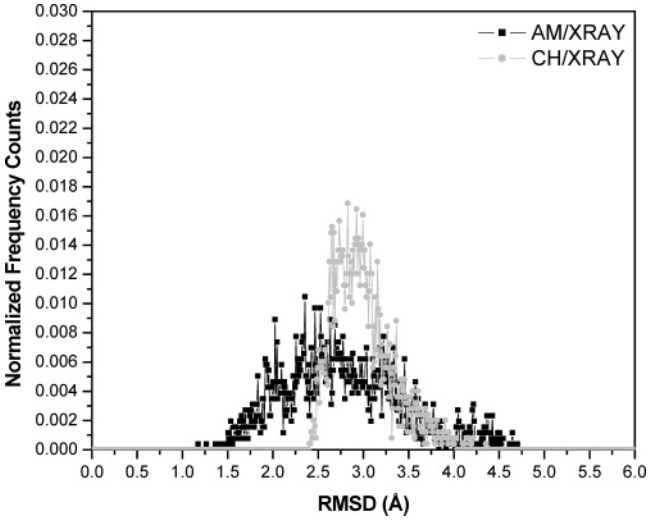

Figure 5. *P*(RMSD) analysis of AMBER and CHARMM MD trajectories on the d(CGCGAATTCGCG) duplex DNA compared to 6 different high-resolution crystal structures of the same sequence: (a) AM/XRAY comparison (black squares) and (b) CH/XRAY comparison (grey circles).

in Figure 5b. The results of P_{inter} for the AM/XRAY comparison lies in the range of 1.5 and 4.7 Å, while the dispersion in the CH/XRAY comparison ranges between 2.4 and 4.2 Å, with a more pronounced skew toward higher end values. The AM/XRAY shows a slightly broader distribution than CH/XRAY, and presents certain conformations of DNA that are individually closer to the XRAY structures. The mean difference is 2.8 ± 0.7 Å for AM/XRAY and 3.0 ± 0.3 Å for CH/XRAY, and the uncertainties indicate that the difference is not sufficient to conclude one force field is better than the other for this sequence. The values of skew and kurtosis for AM/XRAY distribution are 1.298 and 0.656, while those for CH/XRAY are 2.410 and 4.719, respectively. The results do indicate that the AM MD model of DNA is more flexible, i.e., exhibits a larger dynamic range of motion than the CH model. We note that preliminary analysis indicates that the flexibility of the DNA model obtained even from an AM trajectory of a 30 base pair long DNA is not enough to explain the high FRET efficiency observed in the experiments of Weiss and co-workers[35] (Dixit, S. B.; Ponomarev, S. Y.; Beveridge, D. L. Eaton, W., unpublished results).

NMR structure determinations on d(CGCGAATTCGCG)$_2$ have been reported from several different laboratories in recent years,[36−38] the most recent being that of Bax and co-workers[38] using residual dipolar coupling (RDC) data. The NMR structure determination is reported as an ensemble of structures which fit the experimental data within a certain tolerance and are refined with energy minimization. The number of structures reported in such a study is arbitrary, and the RMSD among the structures in the NMR ensemble is usually very small, < 1 Å. The NMR ensemble defined in this manner does not represent a Boltzmann distribution but reflects an admixture of information on molecular flexibility and experimental uncertainty in the NMR structures. A *P*(RMSD) comparison of results from the AM and CH MD models with the 12 structures provided as the NMR ensemble (PDB ID: 1DUF, 1GIP, 1NAJ) is shown in Figure 6. As in the comparison with crystal structure data, the dispersion of RMSD from AM is much wider than that of CH. The results of AM/NMR indicate a range of RMSD values between 1.2 and 4.2 Å, while the corresponding range of the CH/NMR *P*(RMSD) comparison is 2.2−3.5 Å. Thus, as in the comparison with XRAY structures, the flexibility of the AM structures permits DNA conformations that are about 1 Å closer to the NMR data than the corresponding

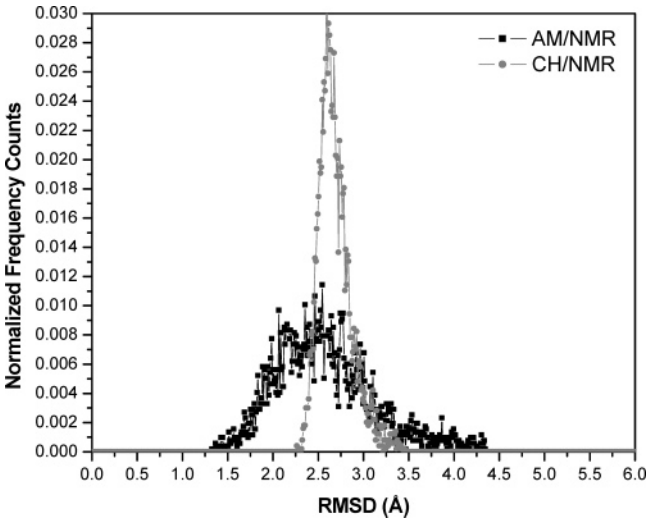

Figure 6. *P*(RMSD) analysis of AMBER and CHARMM MD trajectories on the d(CGCGAATTCGCG) duplex DNA compared to 12 different RDC NMR structures of the same sequence: (a) AM/NMR comparison (black squares) and (b) CH/NMR comparison (grey circles).

Table 1. Calculated Kullback-Leibler Distance between Different RMSD Probability Distributions

	AM/ AM	CH/ CH	AM/ CH	AM/ NMR	CH/ NMR	AM/ XRAY	XH/ XRAY
AM/AM	0.0	0.18	0.83	0.53		0.64	
CH/CH	0.18	0.0	0.83		1.23		1.06

closest CH structures. The distribution of the AM/NMR also presents a more normal tendency with the skew and kurtosis values of 1.6 and 1.4, respectively, in comparison to the CH/NMR distribution which presents very high skew and kurtosis values of 3.6 and 13.0, respectively. The mean of the AM/NMR and CH/NMR *P*(RMSD) distributions are 2.6 ± 0.5 Å and 2.7 ± 0.2 Å, respectively, but given the non-Gaussian nature of these distributions, especially the CH/NMR, the overall distributions are clearly not equivalent. This point would be obscured if only RMSD values were compared.

The calculated Kullback-Leibler distance between the different *P*(RMSD) distributions discussed above are given in Table 1. The KL distance between the AM/AM and CH/CH *P*(RMSD) distributions is 0.18 which is the average of D_{KL}(CH/CH, AM/AM) = 0.20 and D_{KL}(AM/AM, CH/CH) = 0.16. The KL distance indicates that the AM/NMR and AM/XRAY distributions are closer to the AM/AM in comparison to CH/NMR and CH/XRAY distributions with respect to the CH/CH data. This supports the earlier observation that the set of XRAY and NMR structures are slightly closer to the AM ensemble of structures than those from the CH ensemble in this example. Other selective studies of the relative performance of various force fields for MD simulations of nucleic acids are the quantum mechanical calculations of Hobza et al.,[39] MD simulations by Reddy et al.,[40] de Souza and Ornstein,[41] and the reviews of MacKerell[42] and Cheatham and Young.[43]

Comparison of All-Atom Explicit Solvent and Generalized Born Solvent Models MD in AMBER. The Generalized Born (GB) method is a procedure for replacing fully explicit solvent in MD with an approximation based on continuum electrostatics. Such an implicit description of the solvent environment can save considerable time in the production of a trajectory since the computation of nonbonded interactions involving solvent molecules usually constitutes the largest fraction of the calculations being performed at every step of the MD simulation. The question we address here using *P*(RMSD) analysis is how well GB approximates the explicit solvent results, and whether the approximation is accurate enough to safely use GB MD in subsequent applications. We test three GB variants readily available with the AMBER8 software package,[44] namely, IGB1, IGB2, and IGB5.

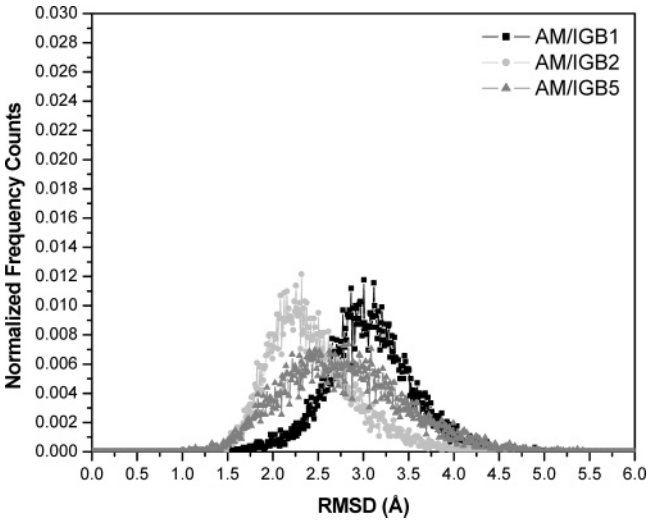

Figure 7. *P*(RMSD) analysis comparing the MD trajectory on d(CGCGAATTCGCG)$_2$ duplex DNA in AMBER (AM) simulation with explicit solvent to 3 different generalized Born based implicit solvent models: (a) comparison of AM to 1GB1 model (black squares), (b) comparison of AM to 1GB2 model (light gray circles) and (c) comparison of AM to 1GB5 model structures (dark gray triangles).

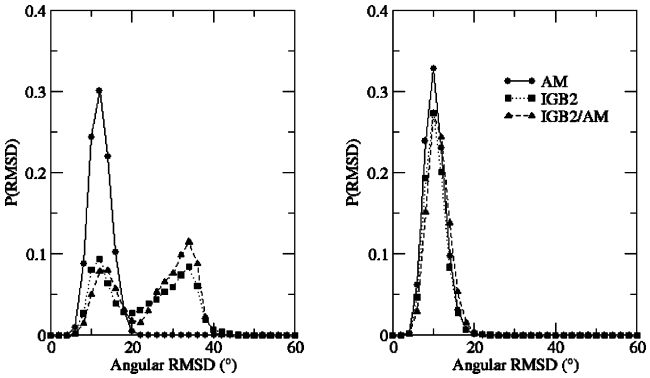

Figure 8. *P*(RMSD) analysis of the explicit solvent AMBER (AM) simulation and the IGB2 generalized Born model in terms of buckle, opening and propeller twist base pairing parameters of the DNA structure. (left) Data for all the 12 base pairs in the dodecamer DNA sequence and (right) data for the central 8 base pairs of the duplex DNA are shown. Symbols used: $P_{AM/AM}$(Intra AM): circles; $P_{IGB2/IGB2}$(Intra IGB2): squares; $P_{IGB2/AM}$(Inter IGB2/AM): triangles.

The *P*(RMSD) distributions comparing AM trajectory with three Generalized Born trajectories (IGB1, IGB2, and IGB5) are shown in Figure 7. The distribution of RMSD values for IGB1/AM structures ranges from ~1.5 to 4.7 Å, having the highest (out of three GB models considered here) mean value of 3.1 Å with a standard deviation of 0.5 Å. The distribution for IGB2/AM spreads out from 1.5 to 4 Å, and the mean RMSD value in this case is 2.4 ± 0.5 Å, the

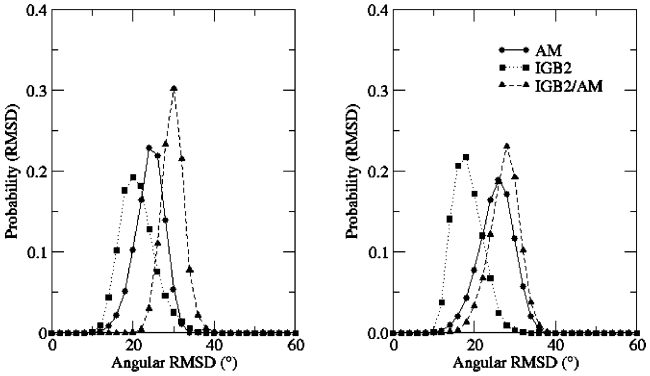

Figure 9. *P*(RMSD) analysis of the explicit solvent AMBER (AM) simulation and the IGB2 generalized Born model in terms of the backbone conformational angles of the DNA structure. (Left) data for all the backbone positions in the dodecamer DNA sequence and (right) data for the central 8 nucleotide positions of the duplex DNA. Symbols used: $P_{AM/AM}$(Intra AM): circles; $P_{IGB2/IGB2}$(Intra IGB2): squares; $P_{IGB2/AM}$(Inter IGB2/AM): triangles.

smallest among the three GB models considered. The widest distribution is seen for IGB5/AM: it stretches from 1.0 to 5.0 Å with a mean RMSD of 2.8 ± 0.7 Å. The P_{intra} KL distance values of the $P_{IGB1/IGB1}$, $P_{IGB2/IGB2}$, and $P_{IGB5/IGB5}$ compared to $P_{AM/AM}$ are 0.29, 0.11, and 0.06, respectively, suggesting that the ensemble of structures obtained in the IGB5 simulation is the closest of the three to the AM explicit solvent simulation, closely followed by IGB2. The P_{inter} KL distance of $P_{IGB1/AM}$, $P_{IGB2/AM}$, and $P_{IGB5/AM}$ curves directly compare the structures in the two different trajectories and turn out to be 1.57, 0.39, and 0.67, respectively. This indicates that although the flexibility and fluctuations in the IGB5 simulation are representative of the AM simulation, the ensemble of structures in the IGB2 trajectory is overall closer to the AM trajectory.

The Generalized Born model IGB2 results are closest to the explicit solvent in terms of the *P*(RMSD) distribution, and we consider this MD for a closer comparison with the explicit solvent trajectory based on the backbone and sugar torsional angles and the helicoidal parameters defining the base pairing geometries. A comparison of *P*(RMSD) of the three angular base pair parameters,[45,46] buckle, propeller twist, and opening, for the DNA trajectories from the explicit (AM) and implicit (IGB2) solvent simulations are provided in Figure 8. These three parameters characterize base pair fraying which is an important issue of concern in the GB simulation of DNA. While the plot to the left in Figure 8 presents the data for all 12 base pairs in the dodecamer DNA, the plot to the right presents the data for the central eight base pair positions in the DNA. Note that while secondary peaks indicating the presence of substates are observed in these angular *P*(RMSD) plots, there was no hint of such structural differences in Figure 7 which is based on Cartesian coordinates. These differences in the *P*(RMSD) curves in Figures 8 indicate that the terminal two base pairs on each strand of the DNA in the generalized Born model exhibit a high degree of structural changes, contributing to the large RMSD values in the secondary peak observed in Figure 8. The fraying of the terminal base pairs in the explicit solvent

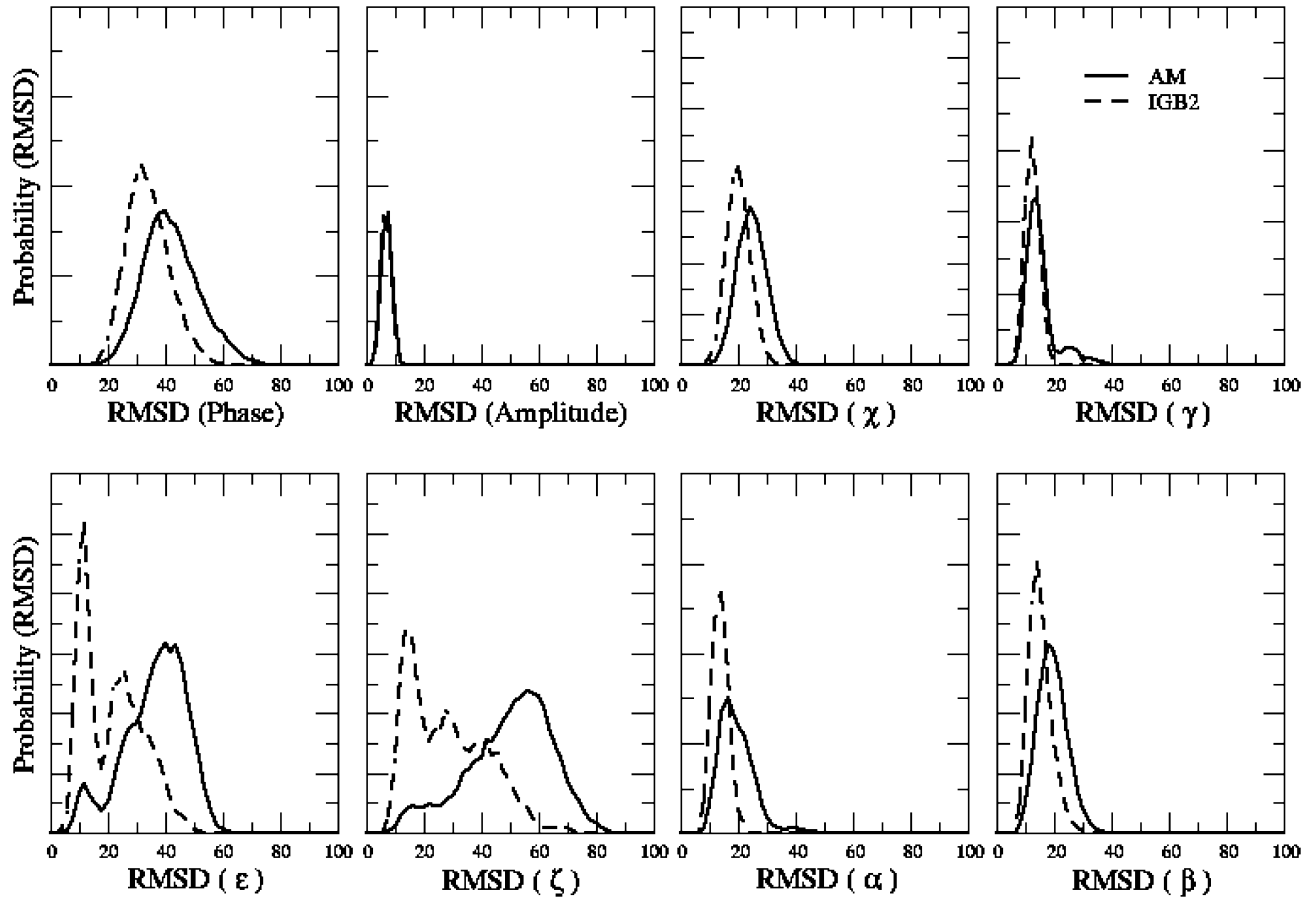

Figure 10. *P*(RMSD) analysis of individual conformational angles in the backbone of the d(CGCGAATTCGCG)$_2$ sequence plotted for the ensemble of structures in the explicit solvent simulation ($P_{AM/AM}$(Intra AM): solid line) and the implicit solvent IGB2 generalized Born model ($P_{IGB2/IGB2}$(Intra IGB2): dashed line). The RMSD values along the ordinate are in degrees.

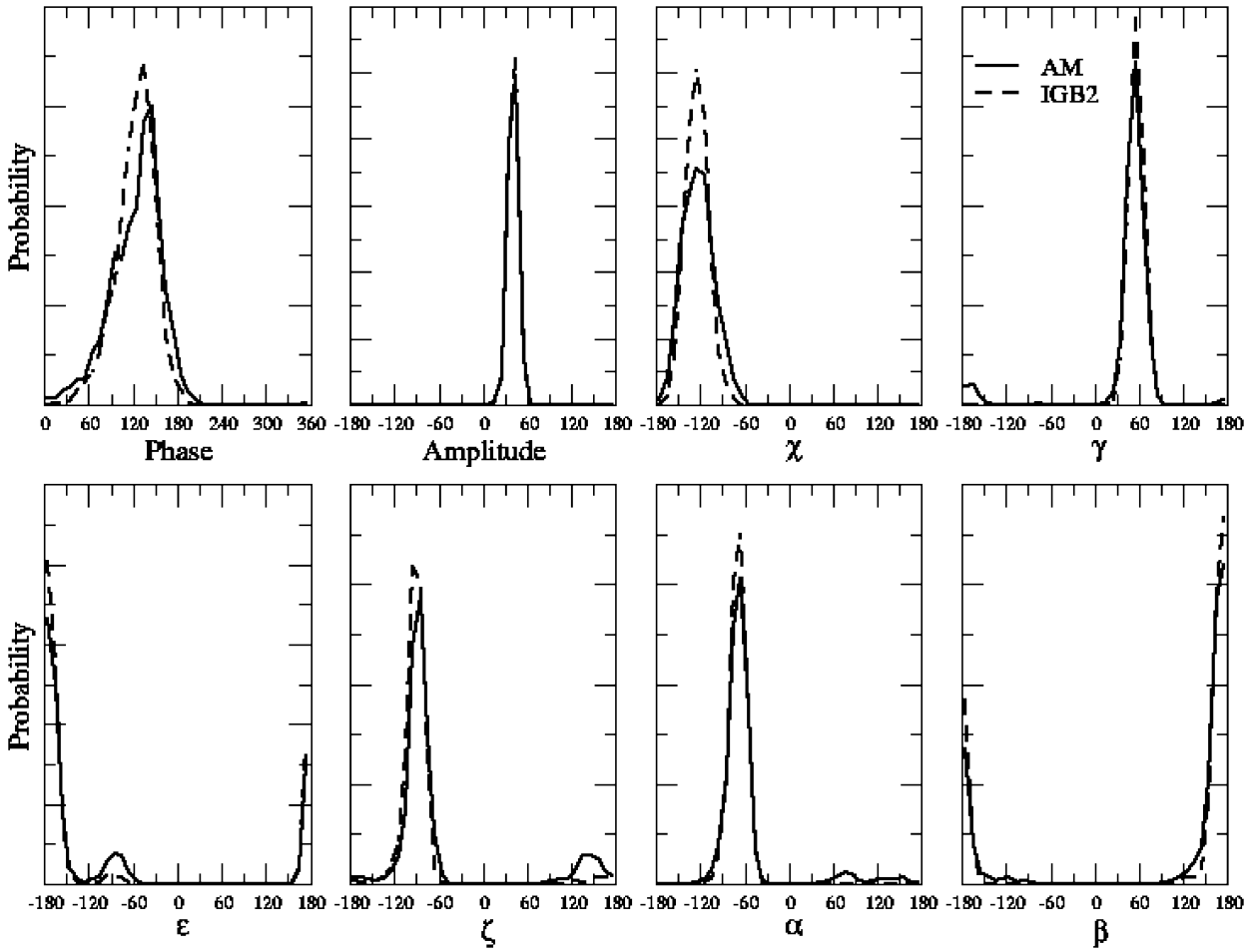

Figure 11. Normalized probability distribution of backbone conformational angles in the explicit solvent AMBER (AM) simulation of the DNA sequence (solid line) compared to the values in simulation with the IGB2 generalized Born model (dashed line). The angular parameters along the ordinate are in degrees.

simulation is much less. Figure 8(right) indicates that the fluctuations in the base pair parameters for the central 10 base pairs are similar in the implicit and explicit solvent simulations. This illustrates how *P*(RMSD) analysis of internal coordinate data can distinguish small differences in distribution.

We proceed to a comparison of the generalized Born simulation with the explicit solvent trajectory in terms of the backbone conformational angles of the DNA. The *P*(RMSD) of the sugar−phosphate backbone conformational angles for the DNA trajectories from the explicit and implicit solvent simulations are shown in Figure 9. Figure 9(left) presents the data for all the 24 backbone positions (12 nucleotides * 2 strands), and Figure 9(right) presents the data for the eight central base pair positions in the DNA. While the data here again shows the differences in the behavior of the central eight base pair positions as against the complete dodecamer sequence, this figure points to another important distinction not captured in Figure 8. Although the base pair parameters showed a good match for the explicit and implicit solvent models in Figure 8(right), Figure 9(right) shows that there are differences in the two trajectories in the eight central nucleotide positions with regard to the backbone conformational angles. The DNA structures in the GB trajectory show smaller RMSD within the ensemble in comparison to the explicit solvent trajectory implying that the fluctuations in this trajectory are less, resulting in a slightly more rigid dynamics of the central eight base pairs. This is likely to be a consequence of a lesser degree of solvation forces in the GB model.

The origin of differences can be analyzed in detail at the level of the individual conformational angles as shown in Figure 10. Notable differences in terms of the RMSD are observed in the case of ϵ and ζ and to a smaller degree in the case of α, β, γ, phase, and χ. In comparison to the normalized distribution of backbone conformational angles shown in Figure 11, one can immediately notice the ability of the *P*(RMSD) plots (Figure 10) to unambiguously highlight the differences in the two trajectories. The dissimilarities in the *P*(RMSD) of ϵ and ζ is the result of differences in proportion of B_I and B_{II} conformational substates of DNA structure in the two simulations. The B_I to B_{II} conformational transitions are described when the value of ϵ/ζ dihedral angles in the DNA backbone changes from values around t/g- in the B_I state to those about g-/t in the case of B_{II}. Such transitions between the B_I and B_{II} states are often observed in crystallographic structures,[47,48] and it is believed that such conformational changes could be an important component of the protein-DNA recognition process. Extensive MD simulations with explicit solvent have shown that transitions between the B_I and B_{II} states occur reversibly, and the B_I and B_{II} populations constitute approximately 92% and 6% of the backbone states, respectively,[2] while the analysis of crystal structure data indicates

that the B_I and B_{II} population ratios might be as high as 79% and 18%, respectively.[47] Given this high accessibility of B_{II} conformations and their potential significance in the protein-DNA recognition process,[47] a force field capable of thoroughly sampling these less probable conformational states may be considered more efficient. Figures 10 and 11 show that the GB model samples the values of ϵ and ζ corresponding to the B_{II} state to a much smaller extent than the simulation with explicit solvent. A different kind of crankshaft motion involves conformational changes in α and γ dihedrals which are observed in the explicit solvent simulation but not in the simulation with the GB model. These backbone conformational substates are also observed in crystal structures of DNA bound to proteins.

SUMMARY AND CONCLUSIONS

The MD average structure typically utilized for the comparison of MD results with experiment and of one MD model with another is not necessarily a reliable index of difference. In response, an analysis based on the probability distributions of RMSDs between all snapshots in the MD trajectory is proposed. The procedure was tested on comparisons of MD on DNA based on the AM and CH force fields with experiment and the results of MD with an implicit generalized Born solvent model against all-atom explicit solvent model simulation. The average RMSD of individual snapshots from AM and CH MD simulations are within thermal uncertainty, but the AM model exhibits a larger range of dynamical motion. The AM trajectory overlaps better than CH trajectory with the experimentally observed structures in both XRAY and NMR, but overall the difference is not sufficient to make firm conclusions. Among the generalized Born models assessed here, the best agreement between the implicit solvent and the explicit solvent AM trajectory was for the IGB2 model proposed by Onufriev et al.[24] The difference in results from all-atom MD compared to GB MD is significant. While the terminal base pairs in the GB model exhibit large unnatural fraying, the base pairs in the center of DNA present much less dynamic range of motion. The GB trajectory also does not present much sampling of the less populous but possibly significant conformational substates of DNA structure.

ACKNOWLEDGMENT

We thank Drs. Richard Lavery, Kelly Thayer for useful discussions, and Dr. Bethany Kormos for comments on the manuscript. Funding for this work came from the National Institutes of Health Grant GM37909 to D.L.B. We thank the anonymous reviewers for valuable comments on the manuscript.

REFERENCES AND NOTES

(1) Arthanari, H.; McConnell, K. J.; Beger, R.; Young, M. A.; Beveridge, D. L.; Bolton, P. H. 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* **2003**, *68* (1), 3−15.

(2) Dixit, S. B.; Beveridge, D. L.; Case, D. A.; Cheatham, T. E., III; Giudice, E.; Lankas, F.; Lavery, R.; Maddocks, J. H.; Osman, R.; Sklenar, H.; Thayer, K. M.; Varnai, P. 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.* **2005**, *89* (6), 3721−40.

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

(4) MacKerell, A. D., Jr.; Banavali, N. All-atom empirical force field for nucleic acids: II. Application to molecular dynamics simulations of DNA and RNA in solution. *J. Comput. Chem.* **2000**, *21* (2), 105−120.

(5) Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.* **1976**, *32A*, 922−923.

(6) Kabsch, W. A discussion of the solution for the best rotation to related two sets of vectors. *Acta Crystallogr.* **1978**, *34A*, 827−828.

(7) Ponomarev, S. Y.; Thayer, K. M.; Beveridge, D. L. Ion motions in molecular dynamics simulations on DNA. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101* (41), 14771−5.

(8) Jorgensen, W. L. Transferable Intermolecular Potential Functions. Application to Liquid Methanol Including Internal Rotation. *J. Am. Chem. Soc.* **1981**, *103*, 341−345.

(9) Aqvist, J. Ion−Water Interaction Potentials Derived from Free Energy Perturbation Simulations. *J. Phys. Chem.* **1990**, *94*, 8021−8024.

(10) York, D. M.; Darden, T. A.; Pedersen, L. G. The effect of long-range electrostatic interactions in simulations of macromolecular crystals: A comparison of the Ewald and truncated list methods. *J. Chem. Phys.* **1993**, *99* (10), 8345−8348.

(11) Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham, T. E., III; Wang, J.; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R.; Cheng, A. L.; Vincent, J. J.; Crowley, M. F.; Tsui, V.; Gohlke, H.; Radmer, R. J.; Duan, Y.; Pitera, J.; Massova, I.; Seibel, G.; Singh, U. C.; Weiner, P.; Kollman, P. A. *AMBER 7*; University of California: San Fransisco, 2002.

(12) Cheatham, T. E., III; Kollman, P. A. Molecular dynamics simulation of nucleic acids. *Annu. Rev. Phys. Chem.* **2000**, *51*, 435−71.

(13) Young, M. A.; Ravishanker, G.; Beveridge, D. L. A 5-Nanosecond Molecular Dynamics Trajectory for B-DNA: Analysis of Structure, Motions and Solvation. *Biophys. J.* **1997**, *73* (5), 2313−2336.

(14) Foloppe, N.; MacKerell, A. D., Jr. All-atom empirical force field for nucleic acids: I. Parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comput. Chem.* **2000**, *21* (2), 86−104.

(15) Banavali, N. K.; MacKerell, A. D., Jr. Reexamination of the intrinsic, dynamic and hydration properties of phosphoramidate DNA. *Nucleic Acids Res.* **2001**, *29* (15), 3219−30.

(16) Beglov, D.; Roux, B. Approximations for Incorporating Solvent Effects in Computer-Simulations of Biomolecules. *Biophys. J.* **1994**, *66* (2), A390-A390.

(17) Darden, T. A.; York, D. M.; Pedersen, L. G. Particle mesh Ewald: an *N* log(*N*) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.

(18) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187−195.

(19) Tsui, V.; Case, D. A. Molecular Dynamics Simulations of Nucleic Acids with a Generalized Born Solvation Model. *J. Am. Chem. Soc.* **2000**, *122* (11), 2489−2498.

(20) Bashford, D.; Case, D. A. Generalized born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.* **2000**, *51*, 129−52.

(21) Fan, H.; Mark, A. E.; Zhu, J.; Honig, B. Comparative study of generalized Born models: protein dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (19), 6760−4.

(22) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Pairwise Solute Descreening of Solute Charges from a Dielectric Medium. *Chem. Phys. Lett.* **1995**, *246*, 122−129.

(23) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium. *J. Phys. Chem.* **1996**, *100*, 19824−19839.

(24) Onufriev, A.; Bashford, D.; Case, D. A. Modification of the Generalized Born Model Suitable for Macromolecules. *J. Phys. Chem. B* **2000**, *104* (15), 3712−3720.

(25) McConnell, K. M.; Nirmala, R.; Young, M. A.; Ravishanker, G.; Beveridge, D. L. A Nanosecond Molecular Dynamics Trajectory for a B DNA Double Helix: Evidence for Substates. *J. Am. Chem. Soc.* **1994**, *116*, 4461−4462.

(26) Flatters, D.; Young, M. A.; Beveridge, D. L.; Lavery, R. Conformational Properties of the TATA Box Binding Sequence of DNA. *J. Biomol. Struct. Dyn.* **1997**, *14* (6), 1−9.

(27) NIST/SEMATECH, *e-Handbook of Statistical Methods*. http://www.itl.nist.gov/div898/handbook/prc/section4/prc45.htm (Accessed Feb, 2006).

(28) Kullback, S.; Leibler, R. A. On Information and sufficiency. *Ann. Math. Statistics* **1951**, *22* (1), 79−86.

(29) Wahl, M. C.; Sundaralingam, M. A. DNA duplexes in the crystal. In *Oxford handbook of nucleic acid* structure; Neidle, S., Ed.; Oxford Univeristy Press: Oxford, GB, 1999; pp 117−144.
(30) Drew, H. R.; Wing, R. M.; Takano, T.; Broka, C.; Tanaka, S.; Itakura, K.; Dickerson, R. E., Structure of a B-DNA dodecamer: conformation and dynamics. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78* (4), 2179−83.
(31) Sines, C. C.; McFail-Isom, L.; Howerton, S. B.; VanDerveer, D.; Williams, L. D., Cations mediate B-DNA conformational heterogeneity. *J. Am. Chem. Soc.* **2000**, *122*, 11048−11056.
(32) Denisov, A. Y.; Zamaratski, E. V.; Maltseva, T. V.; Sandstrom, A.; Bekiroglu, S.; Altmann, K. H.; Egli, M.; Chattopadhyaya, J., The solution conformation of a carbocyclic analog of the Dickerson-Drew dodecamer: Comparison with its own X-ray structure and that of the NMR structure of the native counterpart. *J. Biomol. Struct. Dyn.* **1998**, *16* (3), 547−568.
(33) Shui, X.; Sines, C. C.; McFail-Isom, L.; VanDerveer, D.; Williams, L. D. Structure of the Potassium Form of CGCGAATTCGCG: DNA Deformation by Electrostatic Collapse around Inorganic Cations. *Biochemistry* **1998**, *37*, 16877−16887.
(34) Drew, H. R.; Samson, S.; Dickerson, R. E. Structure of a B-DNA Dodecamer at 16 K. *Proc. Natl. Acad. Sci. U.S.A.* **1982**, *79*, 4040−4044.
(35) Deniz, A. A.; Dahan, M.; Grunwell, J. R.; Ha, T.; Faulhaber, A. E.; Chemla, D. S.; Weiss, S.; Schultz, P. G. Single-pair fluorescence resonance energy transfer on freely diffusing molecules: observation of Forster distance dependence and subpopulations. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96* (7), 3670−5.
(36) Tjandra, N.; Tate, S.-i.; Ono, A.; Kainosho, M.; Bax, A. The NMR Structure of a DNA Dodecamer in an Aqueous Dilute Liquid Crystalline Phase. *J. Am. Chem. Soc.* **2000**, *122* (26), 6190−6200.
(37) Kuszewski, J.; Schwieters, C.; Clore, G. M. Improving the accuracy of NMR structures of DNA by means of a database potential of mean force describing base-base positional interactions. *J. Am. Chem. Soc.* **2001**, *123* (17), 3903−18.
(38) Wu, Z.; Delaglio, F.; Tjandra, N.; Zhurkin, V. B.; Bax, A. Overall structure and sugar dynamics of a DNA dodecamer from homo- and heteronuclear dipolar couplings and 31P chemical shift anisotropy. *J. Biomol. NMR* **2003**, *26* (4), 297−315.
(39) Hobza, P.; Kabelač, M.; Šponer, J.; Mejzlik, P.; Vondrašek, J. Performance of empirical potentials (AMBER, CFF95, CVFF, CHARMM, OPLS, POLTEV), semiempirical quantum chemical methods (AM1, MNDO/M, PM3), and *ab initio* Hartree−Fock method for interaction of DNA bases: Comparison with nonempirical beyond Hartree−Fock results. *J. Comput. Chem.* **1997**, *18* (9), 1136−1150.
(40) Reddy, S. Y.; Leclerc, F.; Karplus, M. DNA polymorphism: a comparison of force fields for nucleic acids. *Biophys. J.* **2003**, *84* (3), 1421−49.
(41) de Souza, N. O.; Ornstein, R. L. Effect of a Warmup Protocol and Sampling Time on Convergence of Molecular Dynamics Simulations of a DNA Dodecamer Using AMBER 4.1 and Particle-Mesh Ewald Method. *J. Biomol. Struct. Dyn.* **1997**, *14* (5), 607−611.
(42) Mackerell, A. D., Jr. Empirical force fields for biological macromolecules: overview and issues. *J. Comput. Chem.* **2004**, *25* (13), 1584−604.
(43) Cheatham, T. E., III; Young, M. A. Molecular dynamics simulation of nucleic acids: Successes, limitations, and promise. *Biopolymers* **2001**, *56* (4), 232−56.
(44) Case, D. A.; Darden, Cheatham, T. A., III; T. E. C.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Mertz, K. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A. *AMBER 8*; University of California: San Francisco, 2004.
(45) Dickerson, R. E.; Bansal, M.; Calladine, C. R.; Diekmann, S.; Hunter, W. N.; O.Kennard; von Kitzing, E.; Lavery, R.; Nelson, H. C. M.; Olson, W. K.; Saenger, W.; Shakked, Z.; Sklenar, H.; Soumpasis, D. M.; Tung, C. S.; Wang, A. H. J.; Zhurkin, V. B. Definitions and Nomenclature of Nucleic Acid Structural Parameters. *EMBO J.* **1989**, *8*, 1−4.
(46) Lavery, R.; Sklenar, H. Defining the Structure of Irregular Nucleic Acids: Conventions and Principles. *J. Biomol. Struct. Dyn.* **1989**, *6*, 655−667.
(47) Djuranovic, D.; Hartmann, B. Conformational characteristics and correlations in crystal structures of nucleic acid oligonucleotides: evidence for sub-states. *J. Biomol. Struct. Dyn.* **2003**, *20* (6), 771−88.
(48) Madhumalar, A.; Bansal, M. Sequence preference for BI/BII conformations in DNA: MD and crystal structure data analysis. *J. Biomol. Struct. Dyn.* **2005**, *23* (1), 13−27.

CI0504925