J|A|C|S
ARTICLES
Published on Web 00/00/0000

Free Energy Landscape of A-DNA to B-DNA Conversion in Aqueous Solution

Nilesh K. Banavali* and Benoît Roux

Contribution from the Department of Physiology, Biophysics, and Systems Biology, Weill Medical College of Cornell University, 1300 York Avenue, New York, New York 10021

Received January 24, 2005; E-mail: nilesh.banavali@cornell.edu

Abstract: The interconversion between the well-characterized A- and B-forms of DNA is a structural transition for which the intermediate states and the free energy difference between the two endpoints are not known precisely. In the present study, the difference between the Root Mean Square Distance (RMSD) from canonical A-form and B-form DNA is used as an order parameter to characterize this free energy difference using umbrella sampling molecular dynamics (MD) simulations with explicit solvent. The constraint imposed along this order parameter allows relatively unrestricted evolution of the intermediate structures away from both canonical A- and B-forms. The free energy difference between the A- and B-forms for the hexamer DNA sequence CTCGAG in aqueous solution is conservatively estimated to be at least 2.8 kcal/mol. A continuum of intermediate structures with no well-defined local minima links the two forms. The absence of any major barriers in the free energy surface is consistent with spontaneous conversion of the A-form DNA to B-form DNA in unconstrained simulations. The extensive sampling in the MD simulations (>0.1 μs) also allowed quantitative energetic characterization of local backbone conformational variables such as sugar pseudorotation angles and BI/BII state equilibria and their dependence on base identity. The absolute minimum in the calculated free energy profile corresponds closely to the crystal structure of the hexamer sequence, indicating that the present method has the potential to identify the most stable state for an arbitrary DNA sequence in water.

Introduction

The structural behavior of DNA in aqueous solution is directly linked to important biological processes such as DNA replication, DNA repair, and RNA transcription. The conformation of DNA primarily observed in aqueous solution is the B-form, but DNA is also frequently seen to deviate from this canonical conformation, especially when bound to proteins.[1] The A-form of DNA is observed in environments with low relative humidity such as solutions with high salt or high ethanol content,[2] and the free energy difference between the A- and B-forms is believed to be a linear function of the water activity.[3] Enzyme binding can induce local conformational changes to the A-form.[4−7] Such a conversion to the A-form structure, which exposes sugar−phosphate atoms in the backbone by a change in minor groove conformation, is primarily seen in enzymes such as polymerases and endonucleases that cut and seal phosphodiester bridges.[8] It has been suggested that the natural sequence-dependent tendency to form A-form structure is utilized by these enzymes in optimizing their association with DNA.[8] Possible A-form to B-form conformational transition pathways have been characterized using structural studies of stable intermediates[9,10] or with chemical manipulations such as bromination or methylation to trap intermediate states.[11,12] These studies indicate that a pathway of discrete intermediates can link the A- and B-forms of DNA. Transitions from the B- to the A-form induced by addition of ethanol are known to occur on the microsecond time scale and are suggested to have large activation barriers and a clear separation between the helical states.[13] It is possible to predict the propensity of localized DNA sequences to adopt the A-form using an empirical approach with a dimer or trimer code,[14,15] but the detailed atomic level interactions giving rise to these trends remain hidden in this empirical approach.

Previous theoretical attempts at characterizing the free energy difference between the A- and B-forms of DNA have used explicit solvent MD simulations to get representative ensembles

(1) Olson, W. K.; Zhurkin, V. B. *Curr. Opin. Struct. Biol.* **1992**, *10*, 286−297.
(2) Ivanov, V. I.; Krylov, D. Y. *Methods Enzymol.* **1992**, *211*, 111−127.
(3) Minchenkova, L. E.; Schyolkina, A. K.; Chernov, B. K.; Ivanov, V. I. *J. Biomol. Struct. Dyn.* **1986**, *4*, 463−476.
(4) Wlassoff, W. A.; Dymshits, G. M.; Lavrik, O. I. *FEBS Lett.* **1996**, *390*, 6−9.
(5) Bebenek, K.; Beard, W. A.; Darden, T. A.; Li, L. P.; Prasad, R.; Luxon, B. A.; Gorenstein, D. G.; Wilson, S. H.; Kunkel, T. A. *Nat. Struct. Biol.* **1997**, *4*, 194−197.
(6) Doublie, S.; Tabor, S.; Long, A. M.; Richardson, C. C.; Ellenberger, T. *Nature* **1998**, *391*, 251−258.
(7) Kiefer, J. R.; Mao, C.; Braman, J. C.; Beese, L. S. *Nature* **1998**, *391*, 304−307.
(8) Lu, X. J.; Shakked, Z.; Olson, W. K. *J. Mol. Biol.* **2000**, *300*, 819−840.
(9) Ng, H. L.; Kopka, M. L.; Dickerson, R. E. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 2035−2039.
(10) Ng, H. L.; Dickerson, R. E. *Nucleic Acids Res.* **2002**, *30*, 4061−4067.
(11) Vargason, J. M.; Henderson, K.; Ho, P. S. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 7265−7270.
(12) Dickerson, R. E.; Ng, H. L. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 6986−6988.
(13) Jose, D.; Porschke, D. *Nucleic Acids Res.* **2004**, *32*, 2251−2258.
(14) Basham, B.; Schroth, G. P.; Ho, P. S. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 6464−6468.
(15) Tolstorukov, M. Y.; Ivanov, V. I.; Malenkov, G. G.; Jernigan, R. L.; Zhurkin, V. B. *Biophys. J.* **2001**, *81*, 3409−3421.

of structures in the two endpoints. Further analysis to characterize energetics used free energy component analysis based on continuum models of solvation in the MM/PBSA approach.[16,17] These studies show that the A- and B-form equilibrium of DNA is related to a complex balance dominated by electrostatic effects and that the presence of salt electrostatically favors the A-form. These calculations provided a satisfying physical explanation for the relative abundance of the B-form in aqueous solution. Nevertheless, a caveat in these studies is that they do not address the details of the transition between the two forms of DNA and do not include effects of explicit solvent dynamics in the free energy estimates. Another study of spontaneous and reversible A- to B-form transitions used multiple simulations with water drops of various sizes surrounding the DNA.[18] Transitions from the B- to the A-form were observed in smaller water drops, while the opposite transitions occurred upon increase in the water drop size, mirroring the experimental trend of A-form prevalence in low water content environments. These calculations provided insight into A- to B-form transition intermediates but did not directly address the issue of relative free energy associated with these intermediates. Furthermore, whether the results obtained by this method can be extended to describe DNA properties in dilute aqueous solution is unclear due to finite-size effects.[19]

In the present study, umbrella sampling MD simulations are used to determine the free energy profile for the A- to B-form transition of the CTCGAG hexamer in aqueous solution. This is currently the most detailed approach for determining free energies of structural transitions and has been utilized before in studying base flipping in DNA.[20−22] The usual difficulties with this method are computational expense and slow convergence. The choice of a proper order parameter as reaction coordinate is a critical factor in the success of the approach. Unlike previous studies of DNA base flipping,[20−22] the change separating the A- and B-DNA forms is delocalized, and this should be reflected in the order parameter. The RMSD between two structures can be used to construct appropriate order parameters in such cases.[23] Here, we implement a 1D difference RMSD constraint to monitor the gradual conversion between canonical A- and B-DNA for efficient characterization of the transformation.

Methods

(a) Choice of Order Parameter. Constraints using RMSD reaction coordinates are routinely used to get a qualitative understanding of structural transitions using a targeted molecular dynamics (TMD) approach.[24−28] A straightforward application of a one-dimensional (1D) RMSD constraint[28] involves starting from the initial structure and gradually pulling toward the final structure by imposition of a harmonic constraint:

$$w_j = K_{\mathrm{rms}}(D(\boldsymbol{X},\boldsymbol{X}_{\mathrm{final}}) - D_{\mathrm{min}})^2 \quad (1)$$

where w_j is the constraint energy, K_{rms} is the force constant in kcal/mol/Å^2, $\boldsymbol{X}$ are the coordinates of a structure at a given dynamics step, $\boldsymbol{X}_{\mathrm{final}}$ are the coordinates of the final structure, $D(\boldsymbol{X},\boldsymbol{X}_{\mathrm{final}})$ is the RMSD of the intermediate structure from the final structure, and D_{min} is the RMSD value corresponding to the minimum value for the constraint. Since this constraint is only imposed with respect to the final structure, there is a possibility of bias or hystereses. A two-dimensional (2D) order parameter consisting of constraints on RMSD from both the initial and final structures would reduce such a bias through its symmetry, but at greater computational expense due to a larger number of umbrella sampling windows. To get good approximations for the intermediate states efficiently, it is desirable to retain both the simplicity of a 1D reaction coordinate and the connection to both endpoints without sacrificing conformational sampling ability. A 1D RMSD reaction coordinate satisfying these criteria constrains the difference between the RMSD values of each structure from the initial and final reference structures:

$$w_j = K_{\mathrm{rms}}(\Delta D_{\mathrm{rmsd}} - D_{\mathrm{min}})^2 \quad (2)$$

where ΔD_{rmsd} is the difference between the RMSD values of each intermediate structure from the final structure and initial structure, respectively, and is given by:

$$\Delta D_{\mathrm{rmsd}} = D(\boldsymbol{X},\boldsymbol{X}_{\mathrm{final}}) - D(\boldsymbol{X},\boldsymbol{X}_{\mathrm{initial}}) \quad (3)$$

and D_{min} specifies the minimum value for the harmonic potential (i.e., the value around which ΔD_{rmsd} is restrained). One advantage of a difference between the RMSD values from two endpoints is that the intermediate structures can, if necessary, sample conformational regions away from both endpoints. This motion away from the diagonal joining the two endpoints in RMSD conformational space is likely to allow rapid equilibration of intermediate states.

(b) Choice of DNA Sequence. The hexamer sequence chosen for this study is CTCGAG. In the crystal structure determination study of this sequence, it was suggested to be a model for the A- to B-form transition.[29] This hexamer sequence was also used for conducting tests on the CHARMM DNA force field during the parametrization process,[30] allowing direct comparison of the present study to previous results. An additional advantage is its relatively small size, which reduces computational cost and allows extensive conformational sampling. Thus, the chosen sequence fulfills the multiple criteria of ability to interconvert between the A- and B-forms of DNA, comparability to previous studies, and computational tractability.

(c) Simulation Methodology. All calculations were performed using the c30a1 academic version of the program CHARMM,[31] modified to incorporate the relevant constraint. The CHARMM27 nucleic acid force field,[30,32] the CHARMM-modified TIP3P water model,[33] and sodium parameters from Beglov and Roux[34] were used. The canonical A- and B-forms of this sequence were generated using the 3DNA program package.[35] Initial system setup involved taking the two canonical forms of DNA and overlaying them with a pre-equilibrated 43 × 34 × 38 Å^3

(16) Srinivasan, J.; Cheatham, T. E., III; Cieplak, P.; Kollman, P. A.; Case, D. A. *J. Am. Chem. Soc.* **1998**, *120*, 9401−9409.
(17) McConnell, K. J.; Beveridge, D. L. *J. Mol. Biol.* **2000**, *304*, 803−820.
(18) Mazur, A. K. *J. Am. Chem. Soc.* **2003**, *125*, 7849−7859.
(19) Mazur, A. K. *J. Am. Chem. Soc.* **2002**, *124*, 14707−14715.
(20) Banavali, N. K.; MacKerell, A. D., Jr. *J. Mol. Biol.* **2002**, *319*, 141−160.
(21) Huang, N.; Banavali, N. K.; MacKerell, A. D., Jr. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 68−73.
(22) Varnai, P.; Lavery, R. *J. Am. Chem. Soc.* **2002**, *124*, 7272−7273.
(23) Elber, R. A. *J. Chem. Phys.* **1990**, *93*, 4312−4321.
(24) Schlitter, J.; Engels, M.; Kruger, P.; Jacoby, E.; Wollmer, A. *Mol. Simul.* **1993**, *10*, 291−309.
(25) Schlitter, J.; Engels, M.; Kruger, P. *J. Mol. Graphics* **1990**, *12*, 84−89.
(26) Ma, J. P.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 11905−11910.
(27) Young, M.; Gonfloni, S.; Superti-Furga, G.; Roux, B.; Kuriyan, J. *Cell* **2001**, *105*, 115−126.
(28) Yang, L. J.; Beard, W. A.; Wilson, S. H.; Roux, B.; Broyde, S.; Schlick, T. *J. Mol. Biol.* **2002**, *321*, 459−478.
(29) Wahl, M. C.; Rao, S. T.; Sundaralingam, M. *Biophys. J.* **1996**, *70*, 2857−2866.
(30) MacKerell, A. D., Jr.; Banavali, N. K. *J. Comput. Chem.* **2000**, *21*, 105−120.
(31) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187−217.
(32) Foloppe, N.; MacKerell, A. D., Jr. *J. Comput. Chem.* **2000**, *21*, 86−104.
(33) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.
(34) Beglov, D.; Roux, B. *J. Chem. Phys.* **1994**, *100*, 9050−9063.
(35) Lu, X. J.; Olson, W. K. *Nucleic Acids Res.* **2003**, *31*, 5108−5121.

solvent box consisting of water molecules and enough randomly distributed sodium ions to neutralize the system. The solvent box extended at least 7 Å beyond the DNA in all directions. All solvent molecules having a non-hydrogen atom within 2.6 Å of the DNA non-hydrogen atoms were deleted. Periodic boundary conditions were used in all subsequent calculations with the images generated using the CRYSTAL module in CHARMM.[36] All calculations were performed using SHAKE[37] to constrain covalent bonds containing hydrogen enabling use of an integration time step of 0.002 ps. Long-range electrostatic interactions were treated using the particle mesh Ewald (PME) approach[38] with a B-spline order of 4 and a fast Fourier transform grid of 1 point/Å and a real-space Gaussian-width κ of 0.3 $Å^{-1}$. Real space and Lennard-Jones (LJ) interaction cutoffs of 10 Å were used with nonbond interaction lists maintained and heuristically updated out to 15 Å. Weak (0.5 kcal/mol) center-of-mass translational and rotational constraints applied using the MMFP module of CHARMM[34] were used to prevent spurious translation or rotation of DNA. The final neutral system consisted of 5563 atoms, including 1725 waters and 10 sodium ions. After minimization for 500 steps using steepest descent with mass-weighted harmonic constraints of 5.0 kcal/mol on the non-hydrogen DNA atoms, a 20 ps constant volume, isothermal (NVT) ensemble MD simulation keeping the same harmonic constraints was carried out to equilibrate the solvent around the DNA. Since spontaneous transitions are known to occur from the A-form to the B-form in unconstrained simulations, these structures were not equilibrated any further and were used as the reference structures for the subsequent umbrella sampling MD simulations. The initial structures for the A to B transition were generated starting from the canonical A-form using a simple 1D RMSD coordinate shown in eq 1 by pulling all non-hydrogen atoms in the structure gradually toward the reference canonical B-form. Each intermediate was generated at 0.1 Å RMSD intervals with 10 ps of NVT sampling before pulling it to the next intermediate value having an RMSD 0.1 Å smaller than the one before. The final structure was constrained to a RMSD value of 0 Å from the canonical B-form reference structure. Thus, 31 intermediate configurations were generated spanning the range of 3 Å RMSD separating the canonical A-form and B-form structures (considering only non-hydrogen atoms). These initial configurations were then used as starting points for umbrella sampling MD simulations with the 1D difference RMSD constraint (ΔD_{rmsd}) shown in eq 2, which was implemented in CHARMM. The initial separation of 0.1 Å intervals in the constraint used for pulling to generate initial coordinates in each window corresponds to a separation of 0.2 Å for the same structures in ΔD_{rmsd}. The force constant for this harmonic constraint was gradually reduced from 500 to 20 kcal/mol/$Å^2$ over a period of 0.5 ns in the NVT MD simulation. Each window was then allowed to evolve with the final ΔD_{rmsd} constraint of 20 kcal/mol/$Å^2$ using an NVT MD simulation for a further 3.5 ns, out of which the last 3 ns were used for further analysis. The total simulation time for all windows combined together added up to about 0.125 μs. The 1D weighted histogram analysis method (WHAM) algorithm[39,40] was used to get the potential of mean force (PMF) along ΔD_{rmsd} from the time series of this variable saved at every step of the 3-ns production dynamics. For comparison, unconstrained MD simulations of the A-form, B-form, and experimental structures were also carried out in the constant pressure and temperature thermodynamic (NPT) ensemble[41] for a total sampling time of 3.5 ns. All analysis of torsion angle related parameters was done using the FREEHELIX98 program[42] modified to read CHARMM trajectories, and analysis of DNA base step parameters was done using the 3DNA program.[35] Molecular pictures were produced using DINO (http://www.dino3d.org), and graphs were made using either gnuplot or OPENDX software.

(d) PMF along Additional Degrees of Freedom. The MD simulations biased with the chosen ΔD_{rmsd} order parameter contain information about many individual degrees of freedom that contribute to the overall transition. If the convergence of the ΔD_{rmsd} free energy profile indicates proper sampling of these individual degrees of freedom, then it is possible to estimate the free energy profile along these additional degrees of freedom using the present umbrella sampling MD simulations. For example, ΔD_{rmsd} can be decomposed into the two components that are used to calculate it, namely the RMSD from canonical A-form and the RMSD from the canonical B-form. The procedure to obtain such extensions of the free energy surface begins by considering the biased probability distribution, $\langle \rho_{\text{bias}}(\xi_1,\xi_2,\xi_3)\rangle$ for the three variables: $\xi_1 = \Delta D_{\text{rmsd}}(t)$, ξ_2 = RMSD from canonical A-DNA, and ξ_3 = RMSD from canonical B-DNA stored at each time point t in all windows. The unbiased probability distribution $\langle \rho(\xi_1,\xi_2,\xi_3)\rangle$ can then be obtained from $\langle \rho_{\text{bias}}(\xi_1,\xi_2,\xi_3)\rangle$ using the WHAM equation

$$\langle \rho(\xi_1,\xi_2,\xi_3)\rangle = \frac{\sum_i^N n_i \langle \rho_{\text{bias}}(\xi_1,\xi_2,\xi_3)\rangle}{\sum_j^N n_j \exp[F_j - w_j(\xi_1)/k_B T]} \quad (4)$$

where N is the number of windows, n_i and n_j are the number of time points (the different subscripts i and j indicating that the number of bins for the histogram needs not be the same as the number of simulation windows), F_j is the free energy constants for each window, and $w_j(\xi_1)$ is the biasing harmonic potential imposed along ΔD_{rmsd} indicated by eq 2. The values of F_j can be obtained from the equation

$$\exp(F_j/k_B T) = \int \langle \rho(\xi_1,\xi_2,\xi_3)\rangle \exp[-w_j(\xi_1)/k_B T]\, d\xi_1\, d\xi_2\, d\xi_3 \quad (5)$$

Since there are two unknowns, F_j and $\langle \rho(\xi_1,\xi_2,\xi_3)\rangle$, to get optimal values of $\langle \rho(\xi_1,\xi_2,\xi_3)\rangle$, eq 4 and eq 5 have to be solved by self-iteration using appropriate convergence criteria (change of less than 0.0001 in successive F_j values is used as a convergence criteria here). Finally, the unbiased density $\langle \rho(\xi_2,\xi_3)\rangle$ for the two coordinates ξ_2 and ξ_3 can be obtained from:

$$\langle \rho(\xi_2,\xi_3)\rangle = \int \langle \rho(\xi_1,\xi_2,\xi_3)\rangle\, d\xi_1 \quad (6)$$

and the free energy surface represented by $W(\xi_2,\xi_3)$ can obtained from:

$$W(\xi_2,\xi_3) = -k_B T \ln(\langle \rho(\xi_2,\xi_3)\rangle) + C \quad (7)$$

where C is an arbitrary constant. In interpreting $W(\xi_2,\xi_3)$, it is important to remember that the sampling along these two dimensions is not enforced directly; therefore, the entire range possible may not have been fully explored. The assumption is that whatever range was visited is relevant to the given transition and that it was sufficiently sampled to give an accurate free energy surface. The 1D free energy profiles along single unconstrained degrees of freedom, such as pseudorotation angles or backbone torsion angles, are also determined in a similar fashion from coordinates saved at every 500 steps of umbrella sampling MD simulations.

Results and Discussion

(a) The Order Parameter. The ΔD_{rmsd} order parameter introduced here can discriminate appropriately between the canonical A- and B-form structures. The advantage of ΔD_{rmsd}

(36) Field, M. J.; Karplus, M. *CRYSTAL: Program for Crystal Calculations in CHARMM*; Harvard University: Cambridge, MA, 1992.
(37) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 327−341.
(38) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(39) Kumar, S.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A.; Rosenberg, J. M. *J. Comput. Chem.* **1992**, *13*, 1011−1021.
(40) Souaille, M.; Roux, B. *Comput. Phys. Commun.* **2001**, *135*, 40−57.
(41) Feller, S. E.; Zhang, Y. H.; Pastor, R. W.; Brooks, B. R. *J. Chem. Phys.* **1995**, *103*, 4613−4621.
(42) Dickerson, R. E. *Nucleic Acids Res.* **1998**, *26*, 1906−1926.

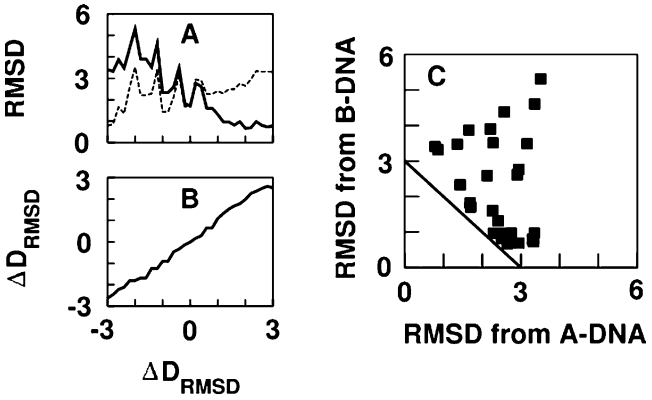

Figure 1. (A) Structures averaged over the last 10 ps of sampling in each window shown in their deviation from canonical A- and B-form structures along the ΔD_{rmsd} windows. It can be seen that some structures deviate as far as 5.5 Å away in RMSD from the B-form and up to 3.5 Å in RMSD from the A-form. (B) ΔD_{rmsd} value for the same structures plotted against their respective ΔD_{rmsd} windows in the free energy profile. This is the parameter that is constrained and therefore lies mostly on the diagonal connecting the two endpoints. (C) Deviation of the same structures (shown as ■) from the diagonal line connecting canonical A- and B-form structures in 2D RMSD conformational space. All values are in Å. The 1D reaction coordinate constraining the difference in RMSD from the canonical A- and B-forms allows structures in all windows of the free energy profile to simultaneously evolve away from both endpoint states.

is the possibility of faster identification of relevant intermediates by allowing the structures to evolve away from both endpoints simultaneously. Panel A of Figure 1 shows the RMSD of structures averaged over the last 10 ps of each window in comparison to each endpoint individually as they progress along ΔD_{rmsd}. Wherever a deviation occurs away from the diagonal in this plot, the deviation is reflected in simultaneous changes in RMSD from both endpoints in the same direction. A clear example of this phenomenon is the presence of peaks in the lines denoting RMSD from both A-form and B-form in the window corresponding to a ΔD_{rmsd} value of −2.0 Å. The only constrained parameter in the umbrella sampling calculations is the difference in RMSD from the endpoints as shown in panel B of Figure 1. ΔD_{rmsd} extends from −3 Å (0 Å from A-DNA minus 3 Å from B-DNA) to 3 Å (3 Å from A-DNA minus 0 Å from B-DNA) as the structures morph from the A-form to the B-form. Since this parameter is constrained, the average value in the last 10 ps of sampling is close to the value it is constrained to and most of the structures lie on the diagonal, as expected. Panel C of Figure 1 shows the location of the same structures in 2D RMSD conformational space where one dimension is RMSD from canonical A-DNA and the other dimension is RMSD from canonical B-DNA. Most of the intermediates (represented as points) lie away from the diagonal connecting the two canonical endpoints. Since this diagonal is the shortest separation distance between the endpoints, a straightforward linear interpolation of all atomic coordinates without any energetic considerations would lie exactly along it. Therefore, structures along the diagonal are less likely to be the most relevant intermediates, and the observed evolution away from the diagonal is reasonable.

(b) Unconstrained Behavior and the Free Energy Profile of A to B Transition. Previous studies[30,43,44] have shown that A-form DNA sequences placed in aqueous solution make spontaneous transitions to the B-form, while B-form DNA sequences in the same environment remain stable. These observations are generally consistent with the perception that the B-form of DNA is prevalent in aqueous solution.[45]To ascertain the behavior of the reference structures used in this study, unconstrained NPT ensemble MD simulations were carried out up to 3.5 ns for both endpoint states as well as the crystal structure configuration. Figure 2 shows the unconstrained evolution of the structures by comparing their RMSD from the canonical A-form structure (red line), the canonical B-form structure (green line), and the crystal structure (blue line). The A-form structure shows spontaneous transition toward the B-form after about 500 ps of sampling, while the B-form structures remain closer to the B-form over the entire sampling range. The unconstrained simulations starting from the crystal structure also remain close to this structure until about 1.8 ns, at which point a base flipping event occurs for the 5′-terminal cytosine of one strand, leading to a gradual shift of RMSD from all reference structures. This deviation due to base flipping is not surprising given the limited stability of short DNA sequences especially at terminal base pairs not stabilized by base stacking on both sides. If this base pair is excluded from the RMSD calculation (results not shown), the rest of the structure remains close to the crystal structure configuration. If the canonical A-DNA NPT MD simulation is continued beyond 3.5 ns (also not shown), the same excursion from the fully double helical structure occurs, indicating that it is not simply a chance occurrence. If this dissociating terminal base pair is ignored, all three unconstrained simulations converge to the same general configurational space in which the crystal structure is located. The 2D plot shown in Figure 2 compares the RMSD in all three simulations with respect to both canonical A-DNA and canonical B-DNA. Even in this more informative 2D RMSD coordinate space, all three simulations converge toward the region of the crystal structure, indicating that the MD simulation structures are not only making the transitions in the correct direction but also are converging toward the right structure as long as the DNA remains completely double helical. Even the terminal base flipping deviation is manifested in the 2D conformational space as a motion away from the diagonal joining canonical A-form and B-form DNA structures. This is precisely the kind of conformational change that remains unrestricted in the present umbrella sampling MD simulations. The present unconstrained MD simulations show similar behavior as those for the same sequence in previous studies.[30] The RMSD time series in Figure 2 also show a substantial amount of fluctuation (about 0.5−1 Å) about their average RMSD, testifying to the dynamic nature of DNA at room temperature.

The primary goal of the present study is, however, not to simulate unconstrained evolution of the DNA structures, but to characterize the previously unknown free energy profile of the A to B transition in aqueous solution. The convergence problems generally true for umbrella sampling calculations could be exacerbated in the present case since it involves projection of a large number of degrees of freedom onto one dimension specified by ΔD_{rmsd}. The ability of the unconstrained MD simulations to easily find the region neighboring the crystal structure starting from two very different starting structures

(43) Langley, D. R. *J. Biomol. Struct. Dyn.* **1998**, *16*, 487−509.
(44) Cheatham, T. E., III; Kollman, P. A. *J. Mol. Biol.* **1996**, *259*, 434−444.
(45) Saenger, W. *Principles of Nucleic Acid Structure*; Springer-Verlag: New York, 1984.

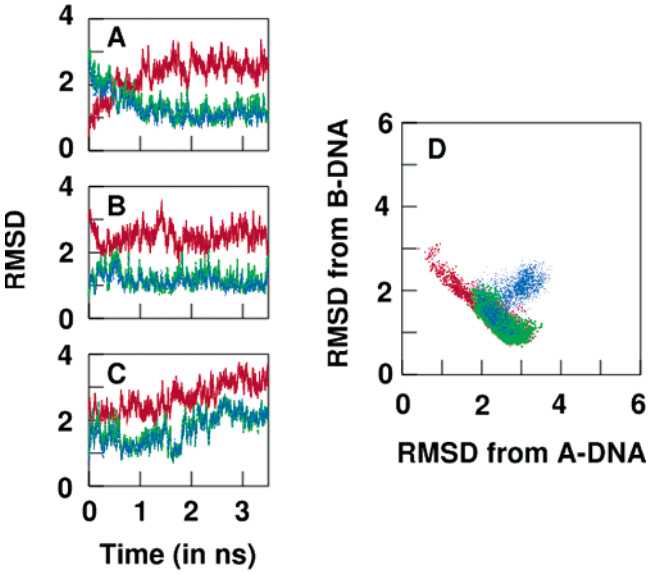

Figure 2. Evolution of (A) canonical A-form, (B) canonical B-form, and (C) crystal structure in unconstrained NPT MD simulations. The canonical A-form makes a transition close to the crystal structure within 500 ps, and the canonical B-form and the crystal structure are both also stabilized close to the crystal B-form. In (A−C), RMSD from canonical A-form is shown by a red line, the RMSD from the canonical B-form is shown by a green line, and RMSD from the crystal structure is shown by a dotted blue line, all values in Angstroms. (D) Distribution of structures from each simulation in the 2D RMSD conformational space. In (D) the A-form simulation is shown in red, the B-form simulation is shown in green, and the crystal structure simulation is shown in blue. All RMSD values are in angstroms. The deviation from both A-form and B-form structures seen for the crystal structure simulation after 1.5 ns (panel C) is due to a terminal base pair fraying event (see text).

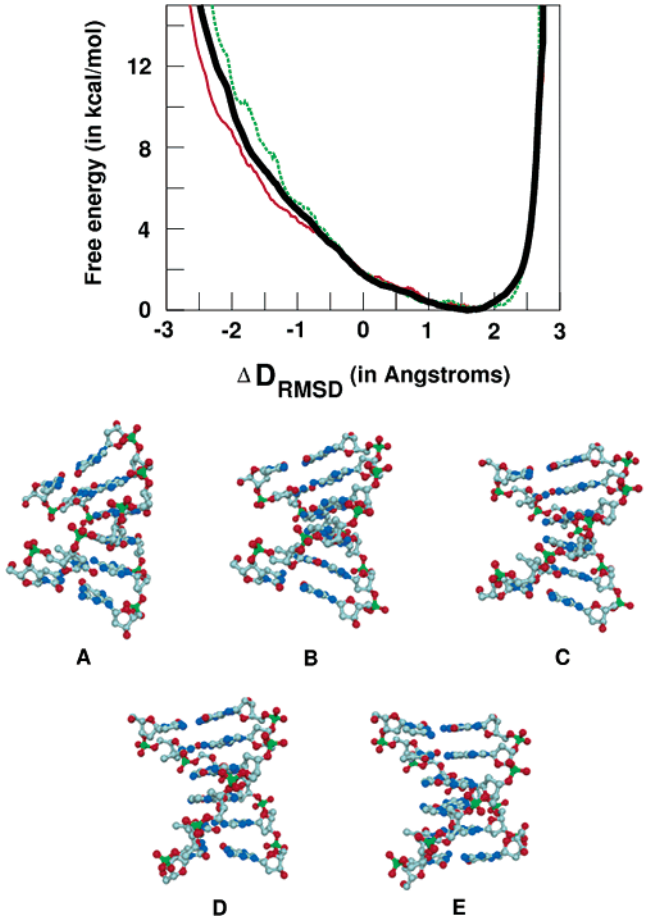

Figure 3. (A) Free energy profile for conversion of A-form to B-form for hexamer sequence CTCGAG along ΔD_{rmsd} for the final 3 ns of sampling calculated using the WHAM algorithm.[39,40] The total sampling range is divided into 10 equal segments with a length of 0.3 ns each; the bold red line represents the profile corresponding to the first 0.3 ns, and the dotted green line represents the profile corresponding to the last 0.3 ns. The bold black line represents the profile for the entire sampling range. Added sampling changes the free energy profile mainly in the region of the A-form structure near −3 Å, making it more unfavorable. The energy well near the most stable B-form structure is quite wide and shallow, indicating a large amount of flexibility in the structure. Selected structures involved in the free energy profile are also shown with reaction coordinate values of, from left to right: A = −3 Å, B = −1 Å, C = 0 Å, D = 1 Å, and E = 3 Å, respectively. Each structure is averaged over the last 10 ps of the 3-ns production sampling with one structure per picosecond included in the averaging. The gradual change visible most prominently in change in relative size of minor and major grooves mirrors the gradual change in the free energy profile.

suggests that this may not be a major problem in the present system. Indeed, the free energy profile calculated for the A to B transition shown in Figure 3 shows remarkably good convergence, especially in the lower energy region. It shows two important characteristics: the presence of a wide energy well representing the global B-form structure, and the absence of a high energy barrier separating the A- and B-forms. An exact magnitude for the free energy difference between A-form and B-form structures cannot be stated unambiguously because there is no defined energy minimum well region that can be used to group a set of structures as exclusively A-form structures. If a simple categorization of all structures is made based on the midpoint of the ΔD_{rmsd} range, then all structures on the left of the midpoint are of higher energy than the lowest free energy state (corresponding to the 1.8 Å window) by 1.8 kcal/mol or higher. If the PMF is integrated over many windows as illustrated in eqs 6 and 7, with A-form DNA extending from ΔD_{rmsd} values of −2.7 to 0.0 Å and B-form DNA extending from ΔD_{rmsd} values of 0.0 to 2.7 Å, then the resulting two-state free energy difference between the A- and B-forms would be about 2.8 kcal/mol. Since the change in the free energy profile upon further sampling seems to be in the direction of destabilizing the A-form structures, the actual free energy difference could be slightly higher. Structures corresponding to both canonical endpoint forms of DNA appear to be high energy, but since the endpoints are unique structures, the Jacobian[46] involved in transformation from Cartesian atomic coordinates to RMSD coordinates is expected to play a role in this apparent destabilization. The volume in the multidimensional configurational space becomes increasingly small as the RMSD goes to zero, which always translates into an apparent rise in free energy for RMSD values smaller than about 1 Å. It should be noted, however, that sequence,[47] base composition,[48] and solvation all influence DNA structure. The canonical forms generated from fiber diffraction data[35] do not adequately take into account these effects and may be high energy forms irrespective of the influence of the Jacobian.

The two base step parameters of slide and zp[49] permit a tighter characterization, than the reaction coordinate alone, of the conformational transition between the A- and B-forms of DNA.[8]

(46) Boresch, S.; Karplus, M. *J. Chem. Phys.* **1996**, *105*, 5145−5154.

(47) Lankas, F.; Sponer, J.; Langowski, J.; Cheatham, T. E., III. *Biophys. J.* **2003**, *85*, 2872−2883.
(48) Foloppe, N.; MacKerell, A. D., Jr. *Biophys. J.* **1999**, *76*, 3206−3218.
(49) ElHassan, M. A.; Calladine, C. R. *J. Mol. Biol.* **1998**, *282*, 331−343.

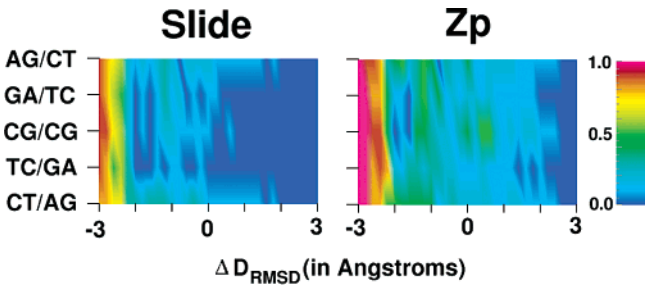

Figure 4. Color maps of change in fraction of A-form conformations sampled for the parameters slide and zp at each base step along ΔD_{rmsd} windows. Color bar shown on right side of graph shows the color gradient over the range 0.0 (0% A-form) to 1.0 (100% A-form) in fraction of structures in A-form sampled over 3 ns. The cutoff criteria for inclusion of values as A-form are: slide less than −0.8 Å and zp greater than 1.5 Å. The base step positions shown on the *Y*-axis are increasing in 5′- to 3′-direction from bottom to top.

Figure 4 monitors their change at each base step in response to change in the ΔD_{rmsd}. Any structure with a slide value less than −0.8 Å and zp greater than 1.5 Å is classified as A-form.[8] The slide parameter seems to be less discriminatory than zp between A- and B-forms with this cutoff. As expected, at the earlier windows corresponding to A-form DNA, both slide and zp show greater proportion of A-form values (fractional values close to 1.0 indicating almost 100% A-form). The occurrence of the A-form decreases in differing ways for each base step. The G/A steps show much greater tendency to shift to B-form values than the A/G or C/G base steps. The small size of the hexamer makes it difficult to make any general observations about the A-form propensity of these base steps, but the direction of the sequence is obviously important since G/A and A/G base steps have different propensity for the A-form. Both slide and zp parameters show complete conversion to B-form values in the final few windows of the reaction coordinate (2.6 to 3.0 Å). The zp values can be used to classify windows along ΔD_{rmsd} as either belonging to the A-form or the B-form to get a simple two-state free energy difference. If only windows with average A-form zp fractional value for all base steps equal to or greater than 0.5 are considered to be A-form, then the span of the A-form is from windows −3.0 to −2.2 Å along ΔD_{rmsd}. A Boltzmann-weighted sum of the free energies according to this more stringent classification yields a higher two-state free energy difference between A-form and B-form DNA of about 13.5 kcal/mol.

Structures obtained by averaging snapshots saved every 1 ps over the last 10 ps of sampling in selected windows of the umbrella sampling simulations shown in Figure 3A−E indicate that the structural changes involved in the transition are gradual. The wide energy well for the global minimum B-form structure is consistent with the observations of both the unconstrained MD simulations (Figure 2) and NMR relaxation measurements[50] that DNA can undergo significant fluctuations in aqueous solution. Although a particular closely related set of structures may correspond to the global energy minimum state, the free energy profile indicates that these structures could evolve transiently to other conformations without incurring prohibitively large energetic penalties.

The free energy profile shows good convergence in regions around the B-form global energy minimum in a relatively short sampling time per window (0.8 ns). When the sampling is extended to 10 times this length, the profile in this region remains virtually unchanged. In contrast, higher energy regions corresponding to A-form structures do not converge as readily. These regions of the PMF show a tendency to increase in free energy with progressively longer simulation times. The present profile, therefore, may actually be a lower boundary for true relative free energies of these structures. The free energy profile determined here is completely consistent with the observations made in unconstrained simulations of the reference states.

(d) Variation in Internal Degrees of Freedom. In the transition from the A- and B-forms, each nucleotide position must adjust its own conformation to a new local minimum. The backbone conformational parameters that describe the local dynamics of individual nucleotide strands are sugar pucker and the backbone torsions χ, α, β, γ, ϵ, and ζ. The free energy profiles along these parameters can be determined accurately if they are adequately sampled. By grouping the free energy profiles according to base type, it is possible to probe the effect of the identity of base and surrounding sequence on the conformational energetics of individual base positions.

1. The Sugar Pucker. The most representative local parameter for A-form versus B-form structure is the sugar pucker conformation of each nucleotide.[45] The C3′-endo sugar conformation corresponds to A-DNA, and the gradual conversion to B-DNA should result in increased sampling of the C2′-endo conformation. The transition from the C3′-endo to C2′-endo conformation in the five-membered deoxyribose sugar ring is due to combined rotations around the five bonds in the ring that can be characterized by the corresponding five torsion angles. Due to the highly correlated nature of these five torsions, a good approximation is to condense them into just two degrees of freedom called the pseudorotation angle and the pseudorotation amplitude.[51] In the pseudorotation angle coordinate (P), the C3′-endo conformation corresponds to a value around 10° ("north" conformation), while the C2′-endo conformation corresponds to a value around 180° ("south" conformation). NMR studies[52,53] and previous MD calculations[20,44] have illustrated that, at the level of a localized nucleotide in DNA, a conformational equilibrium exists between the C2′-endo and C3′-endo conformations.

Figure 5A shows the calculated 1D free energy profiles along the unconstrained dimension of the pseudorotation angle (P) of each individual base position sugar. The free energy profiles are cut off at 5 kcal/mol because the sampling of higher energy regions is insufficient. The overall nature of all profiles confirms the presence of two minima corresponding to the C3′-endo (P around 10−25°) and C2′-endo conformations (P around 130−170°). Interestingly, the relative energy, location, and height of the transition barrier for the minima can all vary depending on the base type and the surrounding sequence. The exact positions of the minima and the putative O4′-endo transition state and their corresponding calculated free energies are shown in Table 1. Quantum mechanical studies characterizing the potential energy profiles for the pseudorotation transition in vacuum show that, in general, the C2′-endo conformation is intrinsically more stable and the most likely pathway for the

(50) Kojima, C.; Ulyanov, N. B.; Kainosho, M.; James, T. L. *Biochemistry* **2001**, *40*, 7239−7246.
(51) Altona, C.; Sundaralingam, M. *J. Am. Chem. Soc.* **1972**, *94*, 8205−8206.
(52) Meints, G. A.; Karlsson, T.; Drobny, G. P. *J. Am. Chem. Soc.* **2001**, *123*, 10030−10038.
(53) LiWang, A. C.; McCready, D. E.; Drobny, G. P.; Reid, B. R.; Kennedy, M. A. *J. Biomol. NMR* **2003**, *26*, 249−257.

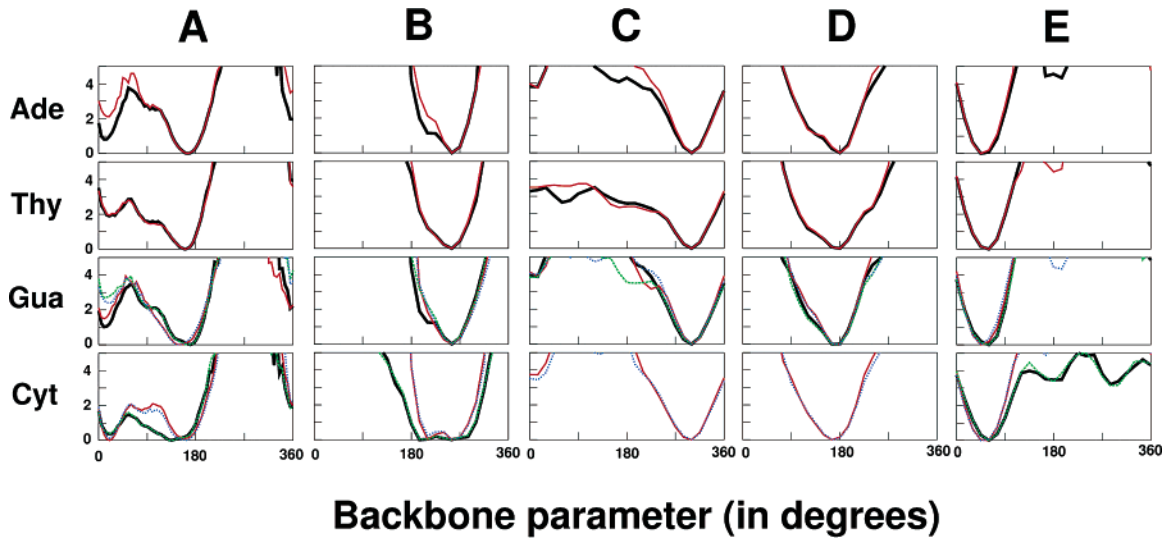

Figure 5. Calculated 1D free energy profiles along the (A) pseudorotation angle, (B) χ torsion, (C) α torsion, (D) β torsion, and (E) γ torsion determined from the umbrella sampling MD simulations and classified according to the base type. The coloring of the lines representing each base depends on its number in the 12 base positions (strand 1 CTCGAG numbered 1−6 and strand 2 CTCGAG numbered 7−12): first, black line; second, red line; third, dotted green line; fourth, dotted blue line. For example, the first guanine (position 4) is represented by a black line, the second (position 6) by a red line, the third (position 10) by a dotted green line, and the fourth (position 12) by a dotted blue line. The *Y*-axis shows the free energy in kilocalories per mole.

Table 1. Positions of Pseudorotation Angle Minima and Putative Transition States and Their Corresponding Free Energies[a]

base number	C3′-endo minimum	O4′-endo barrier	C2′-endo minimum	base type
1	0.4 (20)	1.5 (55)	0.0 (135)	cytosine
2	1.9 (20)	2.9 (60)	0.0 (160)	thymine
3	0.0 (20)	2.1 (60)	0.1 (160)	cytosine
4	1.0 (10)	3.5 (60)	0.0 (170)	guanine
5	0.8 (10)	3.8 (55)	0.0 (165)	adenine
6	1.5 (10)	3.9 (50)	0.0 (150)	guanine
7	0.4 (20)	1.6 (55)	0.0 (130)	cytosine
8	2.0 (20)	2.8 (55)	0.0 (160)	thymine
9	0.1 (20)	2.1 (60)	0.0 (165)	cytosine
10	2.7 (10)	4.0 (60)	0.0 (165)	guanine
11	2.1 (15)	4.6 (60)	0.0 (165)	adenine
12	2.4 (25)	3.9 (50)	0.0 (150)	guanine

[a] Free energies are in kilocalories per mole, and corresponding pseudorotation angles are in degrees (shown in parentheses).

transition to the C3′-endo conformation is through the O4′-endo transition state.[54] In the present results, interconversion between the two minima also occurs through the O4′-endo conformation, but the presence of this transition state conformation is consistently at a *P* value of around 50−60 instead of the higher *P* values near 90 anticipated by previous quantum mechanical studies on model compounds[54] or MD studies on free adenosine nucleoside.[55] There are also consistent differences in the free energy profile between the different base types irrespective of surrounding sequence. Adenine, thymine, and guanine positions show larger barriers to interconversion between the two minima and always favor the C2′-endo conformation. On the other hand, the cytosine in position 3 favors the C3′-endo conformation. Even in other positions, the relative free energy difference between the C3′-endo and C2′-endo conformations is not greater than k_BT for cytosine. Clearly, cytosine is the only base that shows almost equal preference for the C2′-endo and the C3′-endo conformations. The tendency for AT-rich DNA to adopt the B-form and for GC-rich DNA to adopt the A-form (or even the Z-form) seems to thus originate at the local nucleotide level due to the characteristics of deoxyribose sugars with attached cytosine bases. This predilection for the C3′-endo conformation in cytosine, suggested before based on quantum chemical calculations,[48] is confirmed here.

On average, the pyrimidines have lower barriers to interconversion between pseudorotation minima than the purines; a rough ordering of bases according to barrier heights would be adenine > guanine > thymine > cytosine. The two adenines, despite having the same environment, show different free energy profiles especially with respect to the relative free energy difference between the C3′-endo and C2′-endo minima, which might be due to slightly different sampling in each position. Base stacking also seems to play a role in the pseudorotation angle free energy profile since bases at the ends of the oligonucleotide show consistent shifts as compared to the same bases stacked between two base pairs inside the double helix. In the present sequence, the cytosines in positions 1 and 7 and the guanines in positions 6 and 12 located at the ends of the oligonucleotide show a shift toward lower values of P for the C2′-endo minimum. In addition, the reduced sampling of the C3′-endo conformation observed for these terminal base positions could be a direct consequence of greater exposure to water, consistent with the link between high water activity and the reduction in A-form conformations.[3] Earlier predictions on energetics of the pseudorotational equilibrium were based on varied approaches such as energy minimizations of small model compounds,[54,56] statistical approaches based on crystallographic data,[57] NMR on solid hydrated DNA,[52] or MD studies on RNA in vacuum,[58] and single nucleosides in solution.[55] The present study provides quantitative pseudorotation free energy profiles for each nucleotide in a DNA double helix in an explicit solvent environment.

2. Nucleotide Torsions on the 5′-Side of the Sugar Moiety (χ, α, β, and γ). Of all nucleotide torsions figuratively located

(54) Foloppe, N.; MacKerell, A. D., Jr. *J. Phys. Chem.* **1998**, *102*, 6669−6678.
(55) Arora, K.; Schlick, T. *Chem. Phys. Lett.* **2003**, *378*, 1−8.
(56) Brameld, K. A.; Goddard, W. A. *J. Am. Chem. Soc.* **1999**, *121*, 985−993.
(57) Olson, W. K.; Sussman, J. L. *J. Am. Chem. Soc.* **1982**, *104*, 270−278.
(58) Harvey, S. C.; Prabhakaran, M. *J. Am. Chem. Soc.* **1986**, *108*, 6128−6136.

on the 5′-side of the sugar, the χ torsion (rotation of the bases around the glycosyl bond linking them to the sugar moiety) usually shows the strongest correlation to the sugar pucker conformation.[35] Figure 5B shows the calculated free energy profiles along the χ torsion for all base positions truncated at 5 kcal/mol. The overall behavior of all the bases is that they remain in the anti configuration with the global free energy minimum centered at 255°. Exceptional behavior is shown only by cytosines for whom there is a well-defined second minimum at lower values of χ; in fact, the cytosine at position 1 has its global free energy minimum at 195°. This greater sampling range in cytosine is consistent with its greater ability to sample the C3′-endo conformation since the lower range of χ around 180−210 is mostly sampled in A-DNA conformations.[59] It is clear that the syn conformation, associated with Z-DNA,[45] is not populated within the confines of the A-form to B-form transition. The free energy well around the anti conformation is, however, quite wide for the cytosine positions at the termini, and interconversion between the minima at around 210°(A-form) and the minima around 250°(B-form) should occur very easily for them at room temperature. In agreement with greater preference for the C2′-endo conformation, it is the thymine base positions that show the least tendency to adopt the χ value around 210.

The free energy profiles along the α torsion for all base positions truncated at 5 kcal/mol are shown in Figure 5C. The global free energy minima for all the bases are located at the gauche− conformation (300°); however, the thymine base positions show a relatively low energetic cost (<4 kcal/mol) for sampling the entire conformational range possible. In MD simulations of cytosine base flipping for DNA in water, the α torsion is seen to be the local parameter most correlated to the base flipping process.[20] The relative flexibility of the α torsion for the thymine base may play a role in facilitating local structural transitions in DNA involving AT base pairs. On the other hand, it is also possible that this flexibility is independent of the base identity and solely due to the location of the base position near the 5′-terminal end of the oligonucleotide.

Figure 5D shows the β torsion free energy profiles truncated at 5 kcal/mol for all base positions. All global free energy minima are in the range 165−180° with no alternate minima having free energies lesser than 5 kcal/mol. The only discernible pattern seems to be that the pyrimidines have slightly broader energy wells as compared to their corresponding base-paired purines, suggesting slightly greater flexibility of the β torsion at these positions. Drastic local transitions in the β torsion do not seem to be necessary for the overall A-form to B-form transition.

Figure 5E shows the calculated free energy profiles along the γ torsion for all base positions truncated at 5 kcal/mol. The γ torsion in isolated DNA mostly exists in the conformation around 60°; only in Z-DNA is a substantial conformational shift to trans conformation around 180° seen.[59] The present calculations indicate that the global free energy minimum remains between 45 and 60°, but for all bases, there is at least one base position where the trans conformation is less than 5 kcal/mol from the global energy minimum. For cytosine bases at the 5′-end position (positions 1 and 7), however, the entire sampling range of the γ torsion is below 5 kcal/mol, indicating that absence of a 5′-base pair leads to a greater flexibility. The agreement between the surfaces for both 5′-terminal cytosines in opposite strands indicates that the calculated surfaces are converged. For these cytosines, there are three minima located around 60°, 180°, and 285°. In contrast to the β torsion, the A-form to B-form transition may involve marginal variation of the γ torsion. The comparison to cytosine base flipping for DNA in water is inevitable given that these alternate minima are also known to be involved in that process.[20] It is noteworthy that it is one of these 5′-terminal cytosines that also undergoes spontaneous base flipping in the unconstrained MD simulations started from the canonical A-form and crystal structure configurations. The γ torsions in flipped DNA structures complexed to proteins also show a tendency to sample values around 180°,[60] which, taken together with the unconstrained MD simulation results, strongly suggest a connection between γ torsion variation and base flipping. A further description of correlated energetics of the $\alpha-\gamma$ torsion pair is provided in the Supporting Information.

3. The ϵ and ζ Torsions and the BI/BII Equilibrium. The B-form of DNA is known to accommodate two different correlated conformations of the $\epsilon-\zeta$ dihedral pair; these two conformational states are called the BI and BII forms of DNA.[61] The lower energy BI form is defined as ϵ in the range 130−210° and ζ in the range 235−295°; the higher energy BII form is defined as ϵ in the range 210−300° and ζ from 150 to 210°.[61] Since the $\epsilon-\zeta$ torsion pair spans the backbone between two base positions, its behavior also depends on the identity of the 3′ base. Figure 6 shows the calculated 2D free energy profiles along the ϵ and ζ torsions for all base steps truncated at 5 kcal/mol. The BI/BII equilibrium is observable for all base steps in the present sequence. In all cases, the BI state is the global energy minimum, and the BII state is the second prominent local minimum. The largest thermally accessible sampling of this 2D torsional space occurs at the two 5′-terminal C/T base steps (positions 1/2 and 7/8). For these base steps, unusual conformational regions lower than 5 kcal/mol in relative free energy include regions around ϵ = gauche+ and ζ = gauche+, ϵ = gauche+ and ζ = gauche−, and almost the entire range of ζ values for ϵ between 180 and 275°. Some ϵ = gauche+ and ζ = gauche+ conformations for interior G/A and A/G base steps also show free energies less than 5 kcal/mol. Table 2 shows the simple two-state energy difference obtained by calculating the Boltzmann-weighted sum over each state using the precise definitions of each state mentioned above and the putative transition state obtained as a single point free energy in the 2D space mapped. The minimum energy path in 2D space varies along with the magnitude of the putative barrier between BI and BII conformations. C/T, C/G, and G/A base steps show lower energy differences (<1.8 kcal/mol) and barrier heights (<2.6 kcal/mol). T/C and A/G base steps show relatively higher energy differences (>2.4 kcal/mol) and barrier heights (2.7 and 3.1 kcal/mol, respectively). Since all estimates are in duplicate and the difference between the two individual energy estimates is never greater than 0.2 kcal/mol, the convergence

(59) Foloppe, N.; MacKerell, A. D., Jr. *J. Phys. Chem. B* **1999**, *103*, 10955−10964.
(60) Wang, P.; Brank, A. S.; Banavali, N. K.; Nicklaus, M. C.; Marquez, V. E.; Christman, J. K.; MacKerell, A. D., Jr. *J. Am. Chem. Soc.* **2000**, *122*, 12422−12434.
(61) Schneider, B.; Neidle, S.; Berman, H. M. *Biopolymers* **1997**, *42*, 113−124.

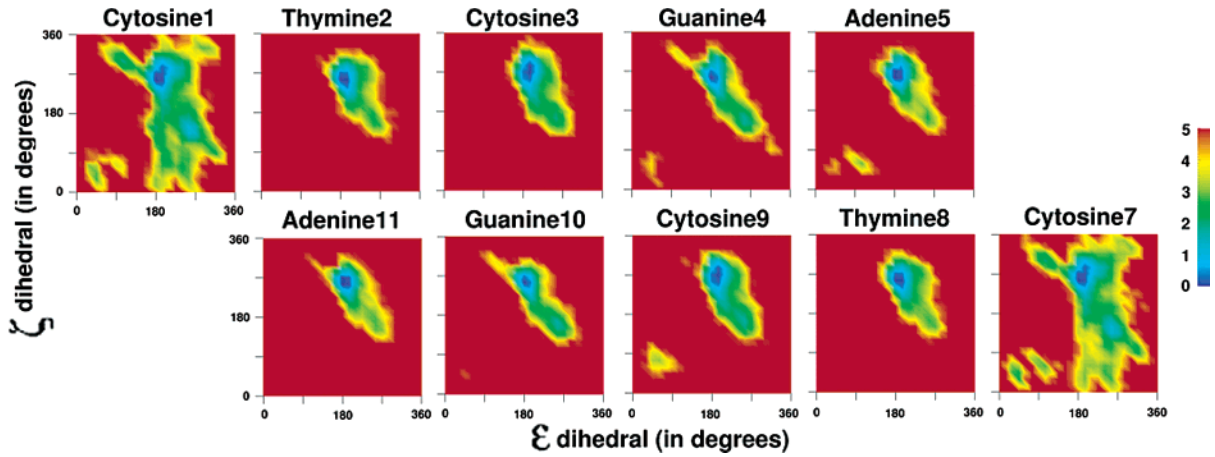

Figure 6. Calculated 2D free energy profile along the ϵ and ζ torsions determined from the umbrella sampling MD simulations for each base position. The arrangement of the bases is such that each column corresponds to a base pair in the oligonucleotide, and the missing positions correspond to the bases at the 3′-termini. The color-coded energy values are shown in kilocalories per mole, and the red-colored area has calculated free energy values beyond the limit of 5 kcal/mol.

Table 2. Two-State Free Energy Difference and Transition State for the BI/BII Equilibrium at Each Base Step[a]

base number	two-state energy difference	global minimum	transition state	base step
1	1.8	(180,255)	2.0 (180,150)	C/T
2	2.4	(195,255)	2.7 (225,195)	T/C
3	1.8	(195,270)	2.3 (225,195)	C/G
4	1.7	(180,255)	2.5 (225,195)	G/A
5	2.7	(180,255)	3.2 (225,195)	A/G
7	1.8	(180,255)	2.1 (210,180)	C/T
8	2.5	(195,255)	2.9 (225,195)	T/C
9	1.6	(195,270)	2.1 (225,195)	C/G
10	1.6	(180,255)	2.6 (225,195)	G/A
11	2.6	(180,255)	3.1 (225,195)	A/G

[a] Free energies calculated at 15° intervals are in kilocalories per mole, and torsions (shown in parentheses in the format (ϵ,ζ)) are in degrees.

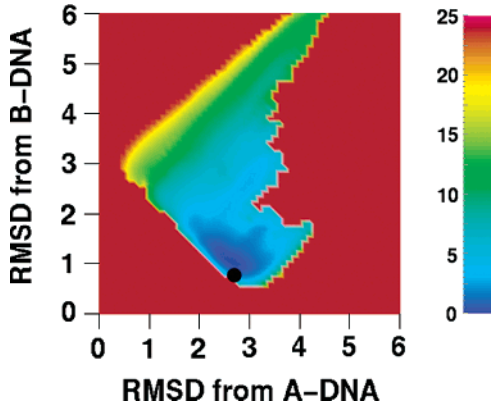

Figure 7. 2D free energy profile along the two dimensions of RMSD from canonical A-DNA and RMSD from canonical B-DNA. The crystal structure located in the broad energy well corresponding to the calculated global free energy minimum on this 2D map is indicated by a black filled circle. It should be noted that no information about this crystal structure was fed into the calculation, the only initial structures used were fiber-diffraction data derived canonical A- and B-form structures. The remarkable correspondence of the free energy minimum obtained by the present method using the CHARMM force field and the high-resolution crystal structure provides very strong validation of the accuracy of the force field. The RMSD values are shown in angstroms and the color-coded free energy is in kilocalories per mole.

is deemed satisfactory. Moreover, previous estimates for the same equilibrium in C/G base steps using unconstrained simulations of a different DNA sequence and force field[62] show a two-state free energy difference for the C/G base step of around 1.7 kcal/mol and a barrier height of about 3 kcal/mol. The BI/BII equilibrium in DNA is considered to be one of the indicators of inherent flexibility for the DNA backbone in water.[61] The present results indicate that this flexibility is prevalent for all base steps in a DNA oligonucleotide with some sequence dependence in the 2D energetic landscape specified by ϵ and ζ.

(d) The Global Free Energy Minimum and the Crystal Structure. The 2D free energy profile along the RMSD relative to canonical A- and B-forms, shown in Figure 7, is consistent with the 1D free energy profile shown in Figure 3. It shows in greater detail how the windows constrained to the A-form try to move away from this high energy region by simultaneously deviating from both canonical A-DNA and B-DNA structures. This spread of sampling toward the top right corner of the 2D profile does not occur in the low energy (blue) regions corresponding to the global energy minimum of the free energy profile. This global free energy region thus behaves like an attractive funnel that prevents structures from dispersing away to a large extent. A valid question to ask upon identification of this global minimum is: how close are the structures obtained in the 3-ns sampling at this window to the crystal structure?[29] Remarkably, the crystal structure (indicated as a black filled circle) lies in the center of this free energy well region. The exact coordinates of the minimized crystal structure are (2.7 Å, 0.8 Å) in the present 2D RMSD conformational space, while the coordinates of the absolute minimum of the calculated profile are (2.7 Å, 1.0 Å), showing the promixity of the two configurations.

The global energy minimum in the 1D free energy profile shown in Figure 3 is located at the window corresponding to a ΔD_{rmsd} value of 1.8 Å. Figure 8 shows the distribution of the RMSD values of structures sampled in this window from the crystal structure. The distribution (black line) is centered around the RMSD value of 1 Å, indicating that the global energy minimum window structures are all very close to the crystal structure. The overlay of the crystal structure with a structure averaged over the last 10 ps of sampling in this window shown

(62) Rauch, C.; Trieb, M.; Wellenzohn, B.; Loferer, M.; Voegele, A.; Wibowo, F. R.; Liedl, K. R. *J. Am. Chem. Soc.* **2003,** *125,* 14990−14991.

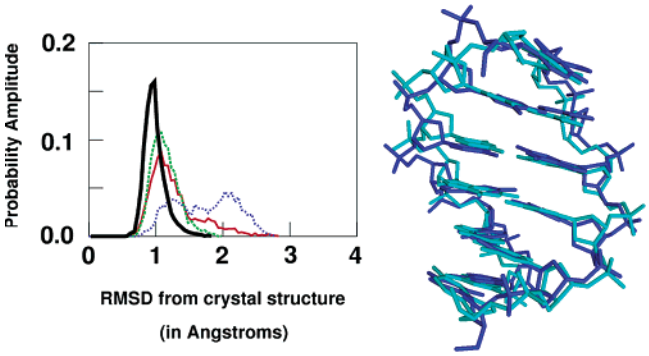

Figure 8. Probability distributions of RMSD from the CTCGAG crystal structure[29] for conformations sampled in four different simulations: (1) The global energy minimum window of the PMF (reaction coordinate value of 1.8 Å, black line), (2) unconstrained simulations started from canonical A-form (red line), (3) unconstrained simulations started from canonical B-form (dotted green line), and (4) unconstrained simulations started from the crystal structure (dotted blue line). The global energy minimum state identified in the free energy profile is clearly sampling conformational regions closest to the crystal structure; all the other unconstrained simulations seem to converge to the same global B-form while being able to sample a wider range of nearby structures. The overlay of the global energy minimum structure averaged over the last 10 ps (blue) onto the crystal structure (cyan) visually indicates that the crystal structure can be correctly anticipated based on the free energy profile.

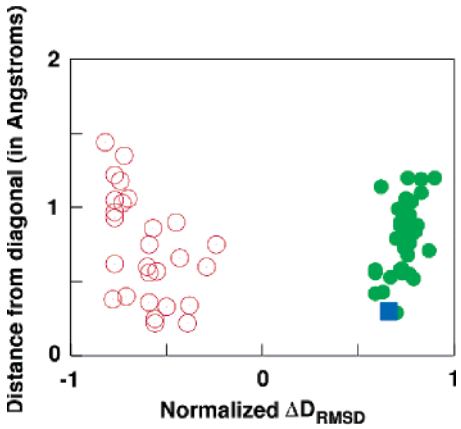

Figure 9. Distance of selected DNA crystal structures from the diagonal joining canonical A-DNA and canonical B-DNA in 2D RMSD space. This distance is calculated as the length of the perpendicular drawn from the point representing the 2D coordinates of each crystal structure to the diagonal corresponding to that structure. DNA crystal structures classified as A-form or B-form in the Nucleic Acid Database and not containing any mismatches, bulges, flipped base pairs, or modifications were selected. The *X*-axis is the 1D ΔD_{rmsd} normalized by dividing by the RMSD difference between canonical A-DNA and canonical B-DNA for each structure. Coloring is as follows. A-form: red, B-form: green, CTCGAG: blue. All structures lie within 1.5 Å of the diagonal, suggesting that excessive sampling may not be required for accessing and identifying these structures using the present umbrella sampling MD simulation method.

in Figure 8 illustrates this agreement. In comparison, the structures obtained during unconstrained simulations starting from the canonical A-form, B-form, and the experimental structure are all further in RMSD from the crystal structure. All unconstrained simulations show significant sampling near the crystal structure configuration. The unconstrained simulation started from the experimental structure also shows the presence of a second peak in the histogram and the wider spread of the distribution. This anomalous peak was identified to be due to a flipping 5′-terminal cytosine base in one strand and not due to a general structural change. In fact, when this fraying base pair is excluded from the RMSD analysis, the remaining overall structure maintains its proximity to the experimental form (not shown). The fraying of this terminal base pair illustrates the existence of additional accessible regions of the free energy landscape of the hexamer during unconstrained dynamics that are not sampled in the present study. A precise identification of the global energy minimum state based solely on unconstrained simulations would be extremely difficult due to the inability to sample higher energy regions closer to canonical A- and B-forms. In contrast, the umbrella sampling MD simulations provide a reasonably converged free energy profile within 0.8 ns of sampling per window or a total of less than 27 ns of sampling. This amount of sampling for the solvated hexamer sequence (about 5600 atoms) can be achieved in 7 days per window. If all 31 windows are run simultaneously on 31 separate 1 GHz Pentium processors, all the required sampling can be completed in 7 days. Thus, the present strategy, when combined with the CHARMM DNA force field, seems to display the potential to predict small DNA structures accurately and efficiently.

(e) Distance from 2D RMSD Diagonal. For the present sequence, the crystal structure is located very close to the diagonal line connecting the canonical forms in 2D RMSD space. This may help ensure adequate sampling of the regions surrounding the global energy minimum in the present calculations. It is not clear, however, if sufficient sampling would be automatically assured for arbitrary DNA sequences simply by generating intermediates spanning the range between canonical A- and B-forms. In Figure 9, an estimate is made for the distance deviation from the 2D RMSD diagonal for selected DNA crystal structures classified as either A-form or B-form in the Nucleic Acid Database[63] (the selection criteria for the structures were that they should not have any mismatches, bulges, overhangs, modifications, or flipped bases). The distance from the diagonal is calculated as the length of the perpendicular drawn from the 2D coordinates (RMSD from A-DNA, RMSD from B-DNA) of each structure to its respective diagonal. The diagonal varies for each structure since the RMSD between canonical A- and B-forms depends on the length and sequence of the structure. This distance is a measure of how much deviation from initial intermediates would be required to sample the conformational region corresponding to the crystal structures using the present method assuming that the initial intermediate configurations will lie at or close to the diagonal. It is clear from Figure 9 that none of the crystal structures deviate by more than 1.5 Å in RMSD from the diagonal. Figure 7 shows that the least deviation seen from the diagonal for the present umbrella sampling MD simulations is 1.5 Å. Thus, it seems likely that at least one of the intermediate umbrella sampling configurations between canonical A- and B-form structures would be close to the global energy minimum, which indicates that the present stategy may be applicable to most DNA sequences.

Conclusions

The ΔD_{rmsd} reaction coordinate used in this study demonstrates the applicability of a delocalized RMSD-based constraint in getting a quantitative free energy profile through umbrella sampling MD simulations. The calculated free energy profiles for conversion from A- to B-DNA (Figures 3 and 7) show that

(63) Berman, H. M.; Westbrook, J.; Feng, Z.; Iype, L.; Schneider, B.; Zardecki, C. *Methods Biochem. Anal.* **2003**, *44*, 199−216.

no significant barriers separate the canonical A- and B-form states from the lowest energy B-form minimum state. Previous MD simulation studies have characterized the BI/BII equilibrium[62] and the correlation between α and γ torsions[64] for specific base steps. In the present study, a more comprehensive characterization of free energy profiles of sugar pseudorotation transitions and backbone torsional variation for all nucleotide steps was obtained with enough accuracy to enable extraction of some of their sequence-dependent characteristics. This characterization yielded one common conclusion: the A- to B-form transition involves conformational changes in local variables that can be best described as changes in relative populations of local minima. The differences between the present local conformational free energy profiles and the corresponding potential energy surfaces obtained using quantum mechanical or molecular mechanical calculations for isolated small model compounds[32,48,54,59] confirm that factors such as sequence context and solvation can cause complex variations from results obtained in vacuum. This modulation of DNA oligonucleotides by its environment and sequence is also consistent with their variable sequence-dependent binding affinity and deformation capacity upon interaction with proteins.[8]

The energy well surrounding the global energy minimum is quite broad, illustrating that there are many closely related conformations accessible through thermal fluctuations. The two-state free energy difference estimated using a midpoint of the ΔD_{rmsd} reaction coordinate as a divider between A- and B-forms yields a two-state free energy difference of 2.8 kcal/mol. But if the more stringent zp base step parameter is used to discriminate the two forms, however, the same value jumps to 13.5 kcal/mol, indicating the near impossibility of an overall A-form conformation in aqueous solution. Unconstrained MD simulations started from the canonical A-form, canonical B-form, and the crystal structure are consistent with the calculated free energy profiles in that they all converge to the same region of 2D RMSD conformational space (Figure 2), except for a terminal base pair fraying event.

An important observation is that the absolute minimum of the calculated free energy surface corresponds very well with the minimized crystal structure. Thus, umbrella sampling with the ΔD_{rmsd} constraint using the CHARMM all27 DNA force field allows correct identification of the global free energy minimum from the extremely large number of sampled structures. The CHARMM nucleic acid force field used in the present study has been parametrized to correctly represent local conformational equilibria with respect to both the quantum mechanical potential energy surfaces and crystal structure torsion distributions in DNA oligonucleotides.[32] It is satisfying that the characteristics of the A- and B-form equilibrium for DNA in aqueous solution are accurately represented in the present study. Since the only external structural information used to obtain the present free energy profile is the canonical A-form and B-form structure derived from fiber diffraction data,[35] in principle, it is possible to replicate the present study for arbitrary DNA sequences without any a priori subjective biasing or human intervention. The present strategy could then constitute a method to directly convert DNA sequence information to accurate isolated DNA 3D structures through the identification of their global free energy minima.

While specific interactions that contribute to stabilization of various structures are included in the free energy determination, further studies using approaches such as mean force decomposition[65] are needed to comprehensively resolve the nature of these underlying mechanisms. The present framework could be also applied to DNA or RNA bound to other macromolecules to estimate conformational flexibility of local degrees of freedom as well as the energetic cost of the overall deformation. Recent studies have shown that the Z-form of DNA may have a biologically relevant role.[66] Determination of the free energy characteristics of the B- to Z-form transition that involves much larger structural perturbations such as base pair rotations would be an interesting future challenge for the present approach.

Acknowledgment. We would like to thank Simon Bernèche, Guillaume Lamoureux, Toby Allen, Hyung-June Woo, Yuqing Deng, Sergei Noskov, José Faraldo-Gómez, and Deniz Sezer for stimulating discussions. A project to calculate the free energy difference between A-DNA and B-DNA based on distance constraints initiated by Alexander MacKerell Jr. in N.K.B.'s graduate work had remained unfulfilled. Therefore, we are indebted to Alexander MacKerell Jr. for the initial concept for this study. This work was supported financially by Grant CA93577-01 from the National Institutes of Health and the Keck Postdoctoral Fellowship for N.K.B. Computational support from the Pittsburgh Supercomputing Center (PSC) obtained through the National Resource Allocation Committee (NRAC) was used for the calculations.

Supporting Information Available: List of PDB identifiers for structures included in Figure 9, figure similar to Figure 6 showing the free energy profile along the 2D $\alpha-\gamma$ torsional space, and a figure similar to Figure 4 showing fraction of A-form structure along the ΔD_{rmsd} coordinate as judged by sampling of a specific region of 2D $\delta-\chi$ torsional space. This material is available free of charge via the Internet at http://pubs.acs.org.

JA050482K

(64) Varnai, P.; Djuranovic, D.; Lavery, R.; Hartmann, B. *Nucleic Acids Res.* **2002**, *30*, 5398−5406.
(65) Allen, T.; Andersen, O. S.; Roux, B. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 117−122.
(66) Rich, A.; Zhang, S. G. *Nat. Rev. Genet.* **2003**, *4*, 566−572.