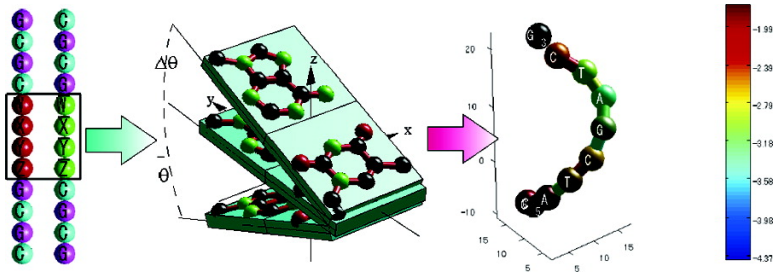
Article

Sequence-Dependent Conformational Energy of DNA Derived from Molecular Dynamics Simulations: Toward Understanding the Indirect Readout Mechanism in Protein–DNA Recognition

Marcos J. Arazo-Bravo, Satoshi Fujii, Hidetoshi Kono, Shandar Ahmad, and Akinori Sarai

J. Am. Chem. Soc., **2005**, 127 (46), 16074-16089• DOI: 10.1021/ja053241l • Publication Date (Web): 29 October 2005



More About This Article

Additional resources and features associated with this article are available within the HTML version:

- Supporting Information
- Links to the 5 articles that cite this article, as of the time of this article download
- Access to high resolution figures
- Links to articles and content related to this article
- Copyright permission to reproduce figures and/or text from this article

View the Full Text HTML

J|A|C|S
ARTICLES
Published on Web 10/29/2005

Sequence-Dependent Conformational Energy of DNA Derived from Molecular Dynamics Simulations: Toward Understanding the Indirect Readout Mechanism in Protein−DNA Recognition

Marcos J. Araúzo-Bravo,[†] Satoshi Fujii,[‡] Hidetoshi Kono,[§,||] Shandar Ahmad,[†,⊥] and Akinori Sarai*[,†]

Contribution from the Department of Biosciences and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka, 820-8502, Japan, Department of Chemistry and Biochemistry, Kyushu University, Fukuoka, Japan, Neutron Research Center and Center for Promotion of Computational Science and Engineering, Japan Atomic Energy Research Institute, 8-1, Umemidai, Soraku-gun, Kyoto, 619-0215, Japan, PRESTO, Japan Science and Technology Agency, and Advanced Technology Institute, Shiki, Saitama, Japan

Received May 18, 2005; E-mail: sarai@bse.kyutech.ac.jp

Abstract: Sequence dependence of DNA conformation plays a crucial role in its recognition by proteins and ligands. To clarify the relationship between sequence and conformation, it is necessary to quantify the conformational energy and specificity of DNA. Here, we make a systematic analysis of dodecamer DNA structures including all the 136 unique tetranucleotide sequences at the center by molecular dynamics simulations. Using a simplified conformational model with six parameters to describe the geometry of adjacent base pairs and harmonic potentials along these coordinates, we estimated the equilibrium conformational parameters and the harmonic potentials of mean force for the central base-pair steps from many trajectories of the simulations. This enabled us to estimate the conformational energy and the specificity for any given DNA sequence and structure. We tested our method by using sequence-structure threading to estimate the conformational energy and the *Z*-score as a measure of specificity for many B-DNA and A-DNA crystal structures. The average *Z*-scores were negative for both kinds of structures, indicating that the potential of mean force from the simulation is capable of predicting sequence specificity for the crystal structures and that it may be used to study the sequence specificity of both types of DNA. We also estimated the positional distribution of conformational energy and *Z*-score within DNA and showed that they are strongly position dependent. This analysis enabled us to identify particular conformations responsible for the specificity. The presented results will provide an insight into the mechanisms of DNA sequence recognition by proteins and ligands.

Introduction

Protein−DNA recognition plays an essential role in the regulation of gene expression. The idea that the DNA sequence-dependent conformation provides indirect readout[1,2] for protein−DNA recognition complementing the direct readout[3] provided by the pattern of noncovalent binding sites in the major and minor groove was proposed by Dickerson.[4] The regulatory proteins recognize specific DNA sequences mainly by way of direct readout through base−amino acid contact and indirect readout through DNA conformation and flexibility. For the direct readout, the intermolecular interaction free energy between base and amino acid determines the stability and specificity of the protein−DNA complex. For the indirect readout, the change in intramolecular conformational energy of DNA upon complex formation determines the sequence specificity. The indirect readout may result from the recognition of intrinsic conformation of DNA by proteins and/or deformability of DNA upon complex formation. In both ways, DNA conformation may work as a potential long-range signal for molecular recognitions.[5] Thus, it is important to evaluate DNA conformational energy and its sequence dependence.

Two kinds of approaches exist to tackle the sequence-dependent conformational energy problem: knowledge-based (or statistical, or empirical) and computational ones (ab initio, molecular dynamics, etc.).[6] The knowledge-based methods analyze known protein−DNA complex structures to derive

[†] Kyushu Institute of Technology.
[‡] Kyushu University.
[§] Japan Atomic Energy Research Institute.
[||] PRESTO.
[⊥] Advanced Technology Institute.

(1) Drew, H. R.; Travers, A. A. *Nucleic Acids Res.* **1985**, *13* (12), 4445−4467.
(2) Otwinowski, Z. R.; Schevitz, W.; Zhang, R.-G.; Lawson, C. L.; Joachimiak, A.; Marmorstein, R. Q.; Luisi, B. F.; Sigler, P. B. *Nature* **1988**, 335, 321−329.
(3) Kono, H.; Sarai, A. *Proteins* **1999**, *35*, 114−131.
(4) Dickerson, R. E. *Sci. Am.* **1983**, *249*, 94−111.
(5) Olson, W. K.; Gorin, A. A.; Lu, X. J.; Hock, L. M.; Zhurkin, V. B. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 11163−11168.
(6) Sarai, A.; Kono, H. *Annu. Rev. Biophys. Biomol. Struct.* **2005**, *34*, 379−398.

10.1021/ja053241l CCC: $30.25

statistical potentials for base−amino acid interactions or DNA conformational energy.[3,7−9] A DNA molecule is often approximated as an elastic object, with several degrees of freedom within a fixed geometry of bases. The local conformation of the DNA is identified at each location of a base pair (from complementary strands) in terms of known deformations such as base-pair step translations: Shift, Slide, Rise, and base-pair step rotations: Tilt, Roll, and Twist.[8,10] Each of these degrees of freedom is characterized by different degrees of flexibility.[11] When a real DNA structure is observed, its stability or energy can be estimated by the amount of deformations compared to a typical or an average structure. Thus, to transform conformational parameters data into energy data, three steps are required viz. (1) determination of an average value of that conformation from a representative nonredundant database, (2) calculation of the deviation of a conformational parameter of a target structure from an average conformation, and (3) derivation of a statistical potential, which can transform these deviations into energy values based on the elastic properties of the corresponding base-pair step for the particular conformational parameter. The total conformational energy of DNA can be approximated as a summation of all base-pair step energies. Olson et al.[5] used a harmonic function to calculate the conformational energy along each conformational parameter. They obtained the force constants from covariance matrices between all pairs of conformational parameters. Gromiha et al.[9] applied the method of Olson et al.[5] to quantify the specificity of the indirect readout. They calculated the conformational energy of DNA involved in 62 protein−DNA complex structures and used sequence-structure threading, in which the original sequence in the DNA structure was replaced by random sequences, to calculate a *Z*-score, i.e., the energy with respect to the mean, normalized by standard deviation as an estimate of the sequence specificity with respect to the conformational state. This technique, together with a similar analysis of direct readout, enabled them to compare the contribution of direct and indirect readouts in protein−DNA recognition. This method has also proven useful in the prediction of target sequences in genome recognized by transcription factors.

Although the statistical potentials derived from the structures of a set of protein−DNA complexes are useful for quantifying the specificity of protein−DNA recognition, they suffer from some inherent problems. For instance, since the amount of available structural data remains limited, the statistical confidence is poor for some structures. Moreover, the structural data may be biased toward certain classes of proteins and DNA sequences. The size problem has restricted the analysis of DNA deformation to dinucleotide steps. It has been established theoretically[12−15] and found experimentally[16−18] that the dinucleotide steps display sequence-dependent conformation.[12−14] Based on the observed standard deviation in Slide, Roll, and Twist, El Hassan and Calladine[19] classified three rigid steps, AA/TT, AT, and GA/TC, and subdivided the loose steps into bistable (all G|C steps: CG, GC, and GG/CC) and flexible (CA/TG and TA) types. Other authors[20] observed that TA, AT, and AA/TT are sequence-context-independent steps and so can be derived from dinucleotide models. But all the G|C steps are strongly context-dependent, and the remaining mixed steps show weakly context-dependent behavior.

For a better understanding of the sequence-dependent conformational energy of DNA, it would be interesting to examine the longer-range effect of the DNA sequence, i.e., to make analysis at the tetranucleotide level (three base-pair steps). The consideration of longer sequences allows the study of the DNA conformational changes with a higher level of cooperativity. For each central dinucleotide of each tetranucleotide, the longer-range analysis takes into account the effect of the neighboring bases on the modulation of the conformational energy associated with the central dinucleotide. For example, conformational properties of base-pair step AC can be resolved to deformations in WACZ tetranucleotides, where W and Z are any of the four possible nucleotides.[21] More recent works have shown that GGC and GCC sequences tend to confer bistability, low stability, and a predisposition to A-form DNA, whereas AA steps strongly prefer B-DNA and inhibit A-DNA structures.[22] The TA step stands out as the most flexible sequence element with respect to decreasing Twist and increasing Roll. This behavior is highly context-dependent, and some TA steps are very straight.[22] This type of high resolution analysis would require much more data of conformational parameters than those usually available in structure databases.

To overcome the limited-data problem in knowledge-based approach for studying the DNA sequence-dependent conformational energy of DNA, a systematic approach generating corresponding data via molecular dynamics (MD) simulations was first set forth by the Ascona B-DNA Consortium,[23] and they have presented some results for d(CpG) steps. The possible approaches to the problem by the MD simulations of this magnitude are not necessarily unique. We generated independently our own systematic set of MD data, by setting the 136 unique tetranucleotides in a different way from Beveridge et al.[23] (see Materials and Methods section). We carried out MD simulations to produce a canonical ensemble, where the trajectories are populated around equilibrium values of conformational parameters. We used the MD trajectory data to calculate the potentials of mean force (PMF) conformational parameters of DNA. A similar protocol as that in the knowledge-based method was followed to calculate the conformational energies of the base-pair steps of the whole DNA. We generalized the dinucleotide based approach (with 10 unique dinucleotide steps) to a longer range technique based on

(7) Selvaraj, S.; Kono, H.; Sarai, A. *J. Mol. Biol.* **2002**, *322* (5), 907−915.
(8) Olson, W. K. et al. *J. Mol. Biol.* **2001**, *313* (1), 229−237.
(9) Gromiha, M. M.; Siebers, J.; Selvaraj, S.; Kono, H.; Sarai, A. *J. Mol. Biol.* **2004**, *337*, 285−294.
(10) Dickerson, R. E.; Bansal, M.; Calladine, C. R.; Diekmann S.; Hunter, W. N.; Kennard, O.; Kitzing, E.; Lavery, R.; Nelson, H. C. M.; Olson, W. K.; Saenger, W. *Nucleic Acids Res.* **1989**, *17* (5), 1797−1803.
(11) Hagerman, P. J. *Annu. Rev. Biophys. Biophys. Chem.* **1988**, *17*, 265−286.
(12) Pedone, F.; Mazzei, F.; Santoni, D. *Biophys. Chem.* **2004**, *112*, 77−88.
(13) Olson, W. K.; Swigon, D.; Coleman, B. D. *J. Mol. Biol.* **2001**, *313* (1), 229−237.
(14) Coleman, B.; Olson, W. K.; Swigon, D. *J. Chem. Phys.* **2003**, *118* (15), 7127−7140.
(15) Zhang, Y.; Crothers, M. D. *Biopolymers* **2003**, *61* (1), 84−95.
(16) Scipioni, A.; Anselmi, C.; Zuccheri, G.; Samori, B.; De Santis, P. *Biophys. J.* **2002**, *83*, 2408−2418.
(17) Crothers, D. M.; Haran, T. E.; Nadeau, J. G. *J. Biol. Chem.* **1990**, *265*, 7093−7096.
(18) Hagerman, P. J. *Annu. Rev. Biochem.* **1990**, *59*, 755−781.
(19) El Hassan, M. A.; Calladine, C. R. *Philos. Trans. R. Soc. London, Ser. A* **1997**, *355*, 43−100.
(20) Packer, M. J.; Dauncey, M. P.; Hunter, C. A. *J. Mol. Biol.* **2000**, *295*, 85−103.
(21) Yanagi, K.; Privé, G. G.; Dickerson, R. E. *J. Mol. Biol.* **1991**, *217*, 201−214.
(22) Gardiner, E. J.; Hunter, C. A.; Packer, M. J.; Palmer, D. S.; Willett, P. *J. Mol. Biol.* **2003**, *332*, 1025−1035.
(23) Beveridge, D. L. et al. *Biophys. J.* **2004**, *87*, 3799−3813.

tetranucleotides (using 136 unique tetranucleotides). We tested whether the sequence-dependent DNA conformation can discriminate target sequences in experimentally observed DNA structures against random sequences by performing the sequence-structure threading to calculate *Z*-score. We used free DNA crystal structures as a template to evaluate our method. If the MD simulations produce a realistic ensemble of DNA conformations and PMFs, we would expect to obtain negative *Z*-scores. That would mean that the particular structure of free DNA can discriminate a particular sequence against random sequences.

Since the essential movements occurring in free B-DNA should be similar to those necessary to deform DNA in protein−DNA complexes,[24] this initial analysis of the specificity of the conformational state and its context-dependence can be used to study the recognition process of protein−DNA binding complexes.

Materials and Methods

2.1. Data Selection. For the *Z*-scores evaluation, the lists of experimental free B-DNA and A-DNA crystal structures were taken (March 2005) from the Nucleic Acid Database (NDB) http://ndbserver.rutgers.edu/.[25] For every file in the lists we took the first biological unit structure from the Biological Units repository of the Protein Data Bank (PDB) (http://www.rcsb.org/pdb/).[26] The dataset constructed by this procedure contains many redundant A-DNA octanucleotides, which are recognized to be highly distorted and potentially bias the dataset. Therefore, we have constructed an A-DNA dataset, consisting of 52 structures, that excludes all octanucleotides, and this more limited dataset was used for all subsequent analyses. The 3DNA software[27] was used to obtain the conformational parameters of each central dinucleotide and prefilter them, deleting the steps with nonstandard nucleotides and the sequences with less than four base steps. As a result, 147 free B-DNA and 51 free A-DNA crystal structures remained.

For comparison, we also calculated the force field matrices for the NDB free B-DNA dataset of Olson et al.[5] These matrices were obtained from the covariance matrices, average values, and dispersion of base-pair step parameters for dimer steps in B-DNA (http://rutchem.rutgers.edu/∼olson/pdna.html) using eq 4. To overcome a possible bias in the calculation of the *Z*-scores of the free B-DNA crystal structures from free B-DNA Olson force field matrices, we excluded the structures used to produce the Olson force fields from the free B-DNA crystal structures.

2.2. Dinucleotide and Tetranucleotide Conformational Analysis. Generally, for a range R of analysis ($R = 2$ for dinucleotides, $R = 4$ for tetranucleotides) $2^{2\cdot R}$ force field matrices are necessary. However, considering the symmetries, the number #R-mer of different R-mers (dinucleotides or tetranucleotides) is equal to the number of different elements (one-half plus the diagonal) of a symmetric matrix of dimension $2^R \times 2^R$,

$$\#R\text{-mers} = 2^{2\cdot R-1} + 2^{R-1} \qquad (1)$$

Special care should be taken in the case of Shift and Tilt conformational coordinates when dealing with symmetries, since the conformational coordinates are calculated using one of the DNA strands.[27] The Shift and Tilt coordinates of the other DNA strand are inverted for the symmetric steps. In the force field calculation, when a symmetric step appears, the signs of Shift and Tilt coordinates are changed. In the energy calculation, we have to take into account that the force field matrices were calculated in the direction of one of the DNA strands. Then, for the case of a symmetric step, the calculation of the perturbation in the Shift and Tilt coordinates should follow the same convention used in the force fields, and the corresponding signs in the measured values should be changed. With the symmetric reduction, for the theoretical 16 possible dinucleotides, applying eq 1 with $R = 2$ results in 10 unique dinucleotide classes. For the theoretical 256 possible tetranucleotides, the application of eq 1 with $R = 4$ provides 136 unique tetranucleotide classes. The 256 → 136 tetranucleotide mapping is shown in the Supporting Information.

To reproduce the dinucleotide force field matrices from the MD data, the dinucleotide *XY* MD data are calculated as the union of all the tetranucleotides W_iXYZ_j that have the dinucleotide *XY* in their center, $\{W, X, Y, Z\} \in \mathcal{N} = \{A,C,G,T\}$

$$XY = \bigcup_{i}^{\mathcal{N}} \bigcup_{j}^{\mathcal{N}} W_i XY Z_j \qquad (2)$$

2.3. Molecular Dynamics Simulation. We have generated dodecamer B-DNA sequences 5′-CGCG-*WXYZ*-CGCG-3′, where $\{W, X, Y, Z\} \in \mathcal{N} = \{A,C,G,T\}$. Each sequence has one of the 136 unique tetranucleotide at its center, and the terminals are always the CGCG tetranucleotide that gives higher stability to the ensemble. This is in contrast with the analysis by Beveridge et al.,[23] where each oligomer was composed of 15 base pairs long built by repeating tetranucleotide sequences and capping the ends with a single G or C to avoid fraying (5′-G-*WXYZ*-*WXYZ*-*WXYZ*-*WXYZ*-*W*-C-3′).

Initial DNA structures were built based on the Arnott B-DNA model[28] with the nucgen module in the AMBER packages 6 and 7.[29,30] Using the Leap module of the package, the initial DNA structures were solvated with the TIP3P water molecules,[31] so that the DNA molecule could be covered with at least a 9 Å water layer in each direction in a truncated octahedral unit cell of size 60 × 60 × 60 Å^3. For the neutralization of the system, 22 K^+ ions were added at favorable positions, and then 17 K^+ and 17 Cl^- ions were added so that the salt concentration of the system would be 0.15 M.

First, we took a 1000-step minimization for water molecules and ions with fixed DNA structure, followed by a further 2500-step minimization for the entire system to remove the large strains in the system. The cutoff used for the van der Waals interactions was 9.0 Å. The particle mesh Ewald method (PME)[32] was used to calculate the full electrostatic energy of a unit cell. After the minimization, the entire system was linearly heated from 0 to 300 K with a weak harmonic restraint to the initial coordinates on DNA (10 kcal/mol) during 20 ps of MD simulation under NVT conditions. We further carried out 100 ps of molecular simulation, keeping the weak DNA restraint for the equilibration of the system under NPT conditions at 300 K. Molecular dynamics simulation for each of 136 unique sequences was then carried out to sample the DNA conformations for 2 ns with NPT conditions with a time constant of 0.2 ps for the pressure control. The temperature was controlled to be 300 K by Berendsen's algorithm[33] with a coupling time of 1 fs, which was set to be the same as the time step of MD simulation. We have used smaller time constants for the pressure and temperature controls for simulations because simulations under such conditions produce an ensemble closer to the canonical ensemble in

(24) Pérez, A.; Noy, A.; Lankaš, F.; Luque, F. J.; Orozco, M. *Nucleic Acids Res.* **2004**, *32* (20), 6144−6151.
(25) Berman, H. M.; Olson, W. K.; Beveridge, D. L.; Westbrook, J.; Gelbin, A.; Demeny, T.; Hsieh, S. H.; Srinivasan, A. R.; Schneider, B. *Biophys. J.* **1992**, *63*, 751−759.
(26) Deshpande, N.; Addess, K. J.; Bluhm, F. W.; Merino-Ott, J. C.; Townsend-Merino, W.; Zhang, Q.; Knezevich, C.; Lie, L.; Chen, L.; Feng, Z.; Kramer-Green, R.; Flippen-Anderson, J. L.; Westbrook, J.; Berman, H. M.; Bourne, P. E. *Nucleic Acids Res.* **2005**, *33* (1), D233−237.
(27) Lu, X. J.; Olson, W. K. *Nucleic Acids Res.* **2003**, *31*(17), 5108−5121.
(28) Arnott, A.; Hukins, D. W. *J. Mol. Biol.* **1973**, *81*(2), 93−105, 1973.
(29) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. R.; Cheatham, T. E., III; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. *Comput. Phys. Commun.* **1995**, *91*, 1−41.
(30) Cheatham, T. E., III; Young, M. A. *Biopolymers* **2001**, *56*, 232−256.
(31) Jorgensen, W. L. *J. Am. Chem. Soc.* **1981**, *103*, 335−340.
(32) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577−8593.
(33) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A. *J. Chem. Phys.* **1984**, *81*, 3684−3690.

the configurational space of DNA though the fluctuation of kinetic energy is suppressed.[34,35] The SHAKE algorithm[36] was used on bonds involving hydrogen. The force field parameters used for MD was from Wang et al. (parm99).[37] To obtain the ensemble, we used the last 1 ns trajectories, where the conformation was sampled at every picosecond (1000 conformations). The average conformational coordinates, the force field matrices, and energy calculations for crystal structures of DNA were programmed in Matlab 7, with the procedures described below.

2.4. Derivation of Dinucleotide and Tetranucleotide Potentials of Mean Forces. The calculation of the DNA conformational energy needs to generate the force field matrices $\mathbf{F}^p$ for each type p of #R-mer steps (where for dinucleotide steps $R = 2$, #R-mer $= 10$, and p is of type XY, and for dinucleotide steps $R = 4$, #R-mer $= 136$, and p is of type $WXYZ$). This is accomplished by approximating the energy of each central dinucleotide six-dimensional conformational fluctuation $\Delta\Theta^p$ of a step of type p with a harmonic function $E = 1/2\Delta\Theta^{pT}\mathbf{F}^p\Delta\Theta^p$ as in Olson et al.[5] Due to an anomalous behavior of the terminal base-pair steps originated by boundary effects, we ignored the 5′ and 3′ terminals from our conformational energy and Z-score calculations of dinucleotide steps. From the six-dimensional matrix that contains the six conformational coordinates of each step of type p, and for each nonterminal step $s = 2,..., L_D - 1$, where L_D is the number of base pairs of the sequence, the following steps are undertaken.

1. For each of the different types of tetranucleotide steps p and for each of the six types of conformational coordinates θ_i of the corresponding central dinucleotide, the initial averages $\bar{\theta}_i^p(k)$ and the standard deviations $\sigma_i^p(k)$ are calculated (where $k = 0$ stands for the first iteration of the statistics calculation). Finally, the conformational coordinates are classified into the #R-mer different types of tetranucleotide classes.

2. Each class of tetranucleotide step is made symmetrical to reduce the numerical problems associated with the inversion of the covariance matrices (that is necessary for the calculation of the force field matrices). An iterative procedure is implemented so that for each of the six conformational coordinates θ_i of each base-pair step $s = 2,..., L_D - 1$ of base-pair type $p(s) = 1,...,$ #R-mer in the target DNA sequence, the average value $\bar{\theta}_i^{p(s)}$ and the standard deviation value $\sigma_i^{p(s)}$ of the associated tetranucleotide are looked for, the base-pair step instantaneous fluctuation from its equilibrium as $\Delta\theta_i^{p(s)} = \theta_i^{p(s)} - \bar{\theta}_i^{p(s)}$ is calculated, and a culling method based on filtering the step s with high fluctuation is applied: if $|\Delta\theta_i^{p(s)}| \leq 3\sigma_i^{p(s)}$, the base-pair step s is rejected. With the remaining nonrejected base-pair steps, the statistics averages $\bar{\theta}_i^p(k+1)$ and the standard deviations $\sigma_i^p(k+1)$ are recalculated, and the rejection and statistics calculation processes $k = k + 1$ are repeated until no more rejection events happen.

3. Once no more base-pair steps are rejected, the covariance matrices $cov^p = \langle\theta_i^p\ \theta_j^p\rangle$, $i, j = 1,..., 6$, for each type of base-pair steps p are calculated.

4. The conformational energies are estimated approximating the energy of the instantaneous fluctuations of every type p of base-pair step, by a harmonic function as in the case of Olson et al.[5]

$$E^p = E_0^{\ p} + \frac{1}{2}\sum_{i=1}^{6}\sum_{j=1}^{6} f_{ij}^{\ p}\Delta\theta_i^{\ p}\Delta\theta_j^{\ p} \qquad (3)$$

where E_0^p is the minimum energy, and the f_{ij}^p, elastic constants impeding the deformations of the given step of type p. Setting arbitrarily the minimum value energies E_0^p equal to 0, assuming that the conformational ensemble is under thermal equilibrium, each element f_{ij}^p of the force field matrix $\mathbf{F}^p$ of each tetranucleotide step p is calculated through matrix inversion of the covariance matrix of the tetranucleotide step, as in the case of Olson et al.[5] for dinucleotide steps,

$$f_{ij}^{\ p} = \frac{1}{k_B T}\langle\Delta\theta_i^p\Delta\theta_j^p\rangle^{-1} = \frac{1}{k_B T}(\langle\theta_i^p\ \theta_j^p\rangle - \langle\Delta\theta_i^p\rangle\langle\Delta\theta_j^p\rangle)^{-1} \qquad (4)$$

where k_B is the Boltzmann constant and T is the temperature in Kelvin. Both are taken equal to one, since during the Z-scores calculation their values in the numerator and denominator of the Z-scores function cancel themselves.

2.5. Derivation of *Z*-Scores of the Conformational Energies. The Z-score method was applied to normalize the energies given by eq 3 for each whole target DNA sequence and for each step of the sequence with the objective of calculating the specificity of the sequence and its steps with respect to the conformational state. The following steps were involved during the Z-score calculations:

1. For each of the six conformational coordinates $\theta_i^{p(s)}$ of each base-pair step $s = 2,..., L_D - 1$ of the target DNA sequence, look for the statistical parameters (the average value $\bar{\theta}_i^{p(s)}$ and the standard deviation value $\sigma_i^{p(s)}$) of the corresponding tetranucleotide step p obtained during the force field calculation using the index p. If the base-pair step is symmetric, change the Shift and Tilt signs of the average value. Calculate the base-pair step instantaneous fluctuation from its equilibrium as $\Delta\theta_i^{p(s)} = \theta_i^{p(s)} - \bar{\theta}_i^{p(s)}$ and saturate the fluctuation:

$$\Delta\theta_i^{p(s)} = \begin{cases} \Delta\theta_i^{p(s)} & \text{if } |\Delta\theta_i^{p(s)}| \leq 3\sigma_i^{p(s)} \\ \text{sign}(\Delta\theta_i^{p(s)})\cdot 3\sigma_i^{p(s)} & \text{otherwise} \end{cases} \qquad (5)$$

2. Calculate the energy $E^{p(s)}$ of each tetranucleotide step $p(s)$ using the eq 3 and the total energy of the sequence as $E = \sum_{s=2}^{L_D-1} E^{p(s)}$.

3. For the sequence-structure threading, generate $rnd = 1,..., N$ random sequences (with discrete uniform distribution of the four bases) of the same length as the original one, build the random sequence, and calculate the six conformational coordinates $\theta_i^{p(s)}$ of each base-pair step. These coordinates are taken from the conformational coordinates of the target sequence step in the same position s.

4. For each tetranucleotide step of each random sequence, calculate the fluctuations of the six conformational coordinates of the central dinucleotide step and saturate them using eq 5.

5. For each random sequence rnd, calculate the energy of every step $E_{rnd}^{p(s)}$, the total energy E_{rnd} (taking into account that, in the case of the symmetric base-pair steps, the averages of the Shift and Tilt coordinates are taken with opposite signs), the statistics of the random step ($E_{rnd}^{p(s)}$ and $\sigma_{rnd}^{p(s)}$), the statistics of the whole random sequence ($\bar{E}_{rnd}$ and σ_{rnd}), the Z-scores of the target step energies $Z^{p(s)} = (E^{p(s)} - E_{rnd}^{p(s)})/\sigma_{rnd}^{p(s)}$, and the Z-scores of the total energy of the target sequence $Z = (E - \bar{E}_{rnd})/\sigma_{rnd}$.

The Z-scores of the step energies $Z^{p(s)}$ are used to analyze the distribution of the energy in the conformational state of the DNA strands, using the "worm" graphs in subsection 3.4.

A time evolution analysis of the Z-scores method was performed with an increasing number of random sequences to analyze the number of random sequences necessary to produce a stable estimate of the Z-scores, and their Z-scores running from $N = 10,..., 5000$ were calculated both for MD dinucleotide (MD2) and MD tetranucleotide (MD4) based force fields. We concluded that, for both cases, $N = 1000$ random sequences are enough to stabilize the Z-scores values, and this was the number of random sequences that we used in all the analysis.

Results

3.1. MD Data Analysis and Resulting Potentials of Mean Force. We are interested in the extent to which the MD simulations can produce ensembles that can predict the sequence-

(34) Morishita, T. *J. Chem. Phys.* **2000**, *113*(8), 2976−2982.
(35) Nose, S. *Prog. Theor. Phys. Suppl.* **1991**, *103*, 1−46.
(36) Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J. Comput. Phys.* **1977**, *23*, 372−336.
(37) Wang, J. M.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049−1074.

dependence of DNA conformation. In this respect, it is important to assess the range in which the PMFs derived from such ensembles can take effect. Therefore, we have examined the difference between dinucleotide and tetranucleotide step models by analyzing the correlation coefficient between their PMFs.

To analyze the difference between the PMFs calculated from dinucleotide and tetranucleotide steps, the correlations of the more salient properties that influence the behavior of the PMFs, the average coordinates $\bar{\theta}_i^p$ corresponding to the average of the data (involved in the conformational perturbation calculation) and the force field parameters f_{ij}^p were calculated. The procedure to calculate both MD2 and MD4 force fields is shown in the Materials and Method section. The data for generating MD2 force fields were obtained from the MD4 data using eq 2.

We estimated the discrepancy between the average values of the conformational coordinates calculated by the MD simulation for dinucleotide steps $\bar{\theta}_i^{XY}$ and tetranucleotide steps $\bar{\theta}_i^{WXYZ}$, where i is the index of each of the six conformational coordinates. For this purpose, we calculated the Pearson's correlation coefficients, $R(\bar{\Theta}^{XY},\bar{\Theta}^{WXYZ(XY)})$, where $XY = 1,...,10$ is the set of the unique dinucleotide steps and $WXYZ(XY) \subset WXYZ$ is the subset of the $WXYZ = 1,...,136$ unique tetranucleotide steps that have the dinucleotide XY at their center. For each XY of the 10 unique dinucleotide steps we built the six-dimensional vector with the six average conformational coordinates $\bar{\Theta}^{XY}$ and their corresponding $\bar{\Theta}^{WXYZ(XY)}$ vectors. Then we calculated the correlation coefficients $R(\bar{\Theta}^{XY},\bar{\Theta}^{WXYZ(XY)})$ and sorted them in increasing order. Their values are represented in Figure 1a. If the correlation coefficient is close to 1, it means that the conformational parameters calculated from the tetranucleotide model are almost equivalent to those from the dinucleotide model. Thus, except for a few cases, the interactions from neighboring bases do not seem to be very significant. We can see that the locations of the gravity centers of the CA and CG base-pair steps are more sensitive to their neighbors, since the tetranucleotides with these centers have more variability in the correlation coefficients $R(\bar{\Theta}^{CA},\bar{\Theta}^{WCAZ(CA)})$ and $R(\bar{\Theta}^{CG},\bar{\Theta}^{WCGZ(CG)})$. At least in the case of the CG base-pair step, this variability is due to the bistable character of the step. The lowest correlation occurs for the tetranucleotide A**CG**A.

We also examined the discrepancy between force field parameters obtained by considering short-range interactions f_{ij}^{XY} and long-range interactions $f_{ij}^{WXYZ(XY)}$, where i and j are the indices of each of the six conformational coordinates. In this case, we calculated the correlation coefficients, $R(f^{XY},f^{WXYZ(XY)})$. For each XY of the 10 unique dinucleotide steps we built the 21-dimensional vector with the 21 nonsymmetric force field parameters of the 6×6 matrix $\mathbf{F}^{XY}$ and their corresponding $f^{WXYZ(XY)}$ vectors. Then we calculated the correlation coefficients $R(f^{XY},f^{WXYZ(XY)})$ and sorted them in increasing order. Their values are represented in Figure 1b. Clearly, the interactions from neighboring bases have a more obvious effect on the force field parameters, since the correlation coefficients are more variable depending on the sequence. This clearly shows that the adjacent base pairs affect more the fluctuations of the conformations rather than the average conformations. For example, the force field parameters of the TA base pair are more sensitive to a neighboring influence, since the tetranucleotides with this center have more variability in the correlation coefficients of their force field parameters. The lowest correlation happens for the tetranucleotide A**TA**T.

To analyze the reason of the differences between dinucleotide and tetranucleotide force fields, the pattern of the distribution of the data was studied using scatterplots. Figures 2 and 3 show the scatterplots of each one of the 10 unique possible central dinucleotides (in the left) and the scatterplot of the corresponding tetranucleotide (in the right), which in Figure 1b had shown the lowest correlation coefficient with its 6×6 force field parameters f_{ij}^p. All the scatterplots correspond to the no-data-rejection state $k = 0$ (the final force field matrices $\mathbf{F}^p$, $p = 1,...,\#R$-mer were obtained typically after four or five rejecting cycles, depending on the base-pair step type). The bidimensional scatterplots of the coordinates pairs with more salient features were chosen from all 15 possible pairs of combinations of the six conformational coordinates θ_i, shown in Figures 2 and 3. The histograms and the equipotential ellipses were also calculated in the scatterplots. The ellipses are projections of the six-dimensional equipotential surfaces on the respective base-pair plane obtained from the 2×2 covariance matrices; these contours correspond to energies of 4.5 k_BT ("$3\Delta\theta$ ellipses").[27] The contours are chosen at 4.5 k_BT energies because the acceptance range is 3 in units of standard deviations to select (cull) the data. Assuming linear independent coordinates which are decoupled and using the harmonic potential hypothesis, $E_{max} = {}^1/_2\, f_{ii}\Delta\theta_{imax}\Delta\theta_{imax} = f_{ii}\cdot 0.5\cdot 3\cdot 3 = f_{ii}\cdot 4.5\cdot k_BT$.

Three main reasons for the discrepancy in the dinucleotide and tetranucleotide force field matrices can be envisaged from the scatterplots. First, some dinucleotide scatterplots present bimodal (GA, GG, CG) and trimodal (TA) distributions due to the superposition of tetranucleotide modes with a different gravity center. This means that the modes of some dinucleotide data are split by their tetranucleotide conformation, an observation in conformity with earlier reported results of analysis using X-ray crystal data,[20] as well as by MD simulations.[23] A second reason is that in the tetranucleotide scatterplots there are cases with bimodal behavior in the bidimensional projections of the conformational coordinates (e.g., Twist coordinate in C**CG**G). This nonharmonical behavior is apparently canceled in the dinucleotide set by an averaging effect that compensates the different trends of its tetranucleotide components. For example, in the C**CG**G case the summation eq 2 of the 10 tetranucleotides corresponding to the central dinucleotide CG makes the data distribution more Gaussian. A third reason is that the average of the dinucleotide conformational coordinates is not always close to the gravity center of all the tetranucleotides. For example, the bimodal gravity center of the Twist of C**CG**G is at 30.5°, for CG Twist it is at 27.0°, for the Tilt of C**CG**G it is at 2.5 Å, and for CG it is at 0.0 Å.

The bistable behavior of the steps involving G|C nucleotides has already been reported based both on computational models and on MD simulations. Packer et al.[38] proposed the electrostatic interactions as the reason for this behavior. Our MD simulation results agree with the results of Packer et al.[38] In the CG case, Packer et al.[38] suggested that the bistable behavior is due to the necessity of introducing a positive or a negative Shift to neutralize the overlap between one of the negative G charges and one of the positive C charges. These Shift changes are also

(38) Packer, M. J.; Dauncey, M. P.; Hunter, C. A. *J. Mol. Biol.* **2000**, *295*, 71−83.

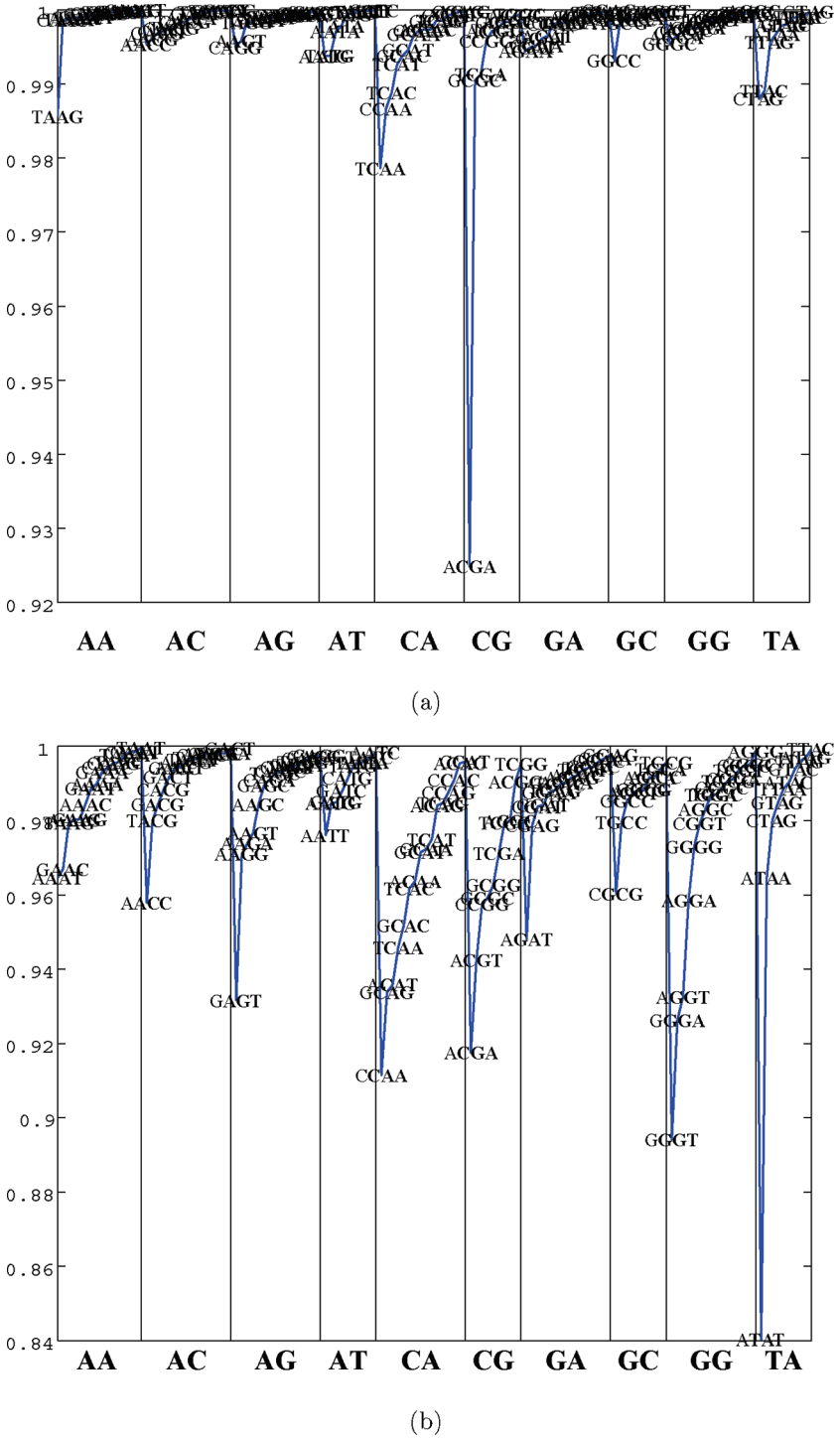


Figure 1. Pearson's correlation coefficients of each tetranucleotide *WXYZ*(*XY*) and each corresponding central dinucleotide *XY* (a) $R(\bar{\Theta}^{XY}, \bar{\Theta}^{WXYZ(XY)})$ of the six average conformational coordinates $\bar{\theta}_i^p$, (b) $R(f^{XY}, f^{WXYZ(XY)})$ of the 21 nonsymmetric *XY* force field coefficients f_{ij}^p. The correlation coefficients are sorted in increasing order for each central dinucleotide step.

performed by our MD simulated data, as reflected in the histograms and scatterplots of the CG MD evolution of Figure 3. In this figure the Shift oscillates between a negative local minimum around −0.4 Å and a positive one around 0.6 Å. The transitions between both states were performed passing through 0 Å Shift states, with simultaneous oscillations of the Twist between 30.5° and 27.0°. Packer et al.[38] also suggested that the reason for the bistable behavior of the GC steps is again rearrangement of the electrostatic charges. But in the GC case it is a movement to a negative Slide which reduces the repulsion and brings the G negative charges close to the C positive charges. The increase or decrease in the Shift coordinates reduces the interaction between the two charges. As in the case of the conformational maps calculated in Packer et al.,[38] a clear preference for high Shift due to the lower interacting electrostatic energy does not emerge in our MD simulations. Nevertheless, negative Slide appears around −0.6 Å, and a more diffused positive Slide, around 0.5 Å (see case GC in Figure 3). Finally, for the GG case Packer et al.[38] proposed that the interaction between the van der Waals and the electrostatical forces can be

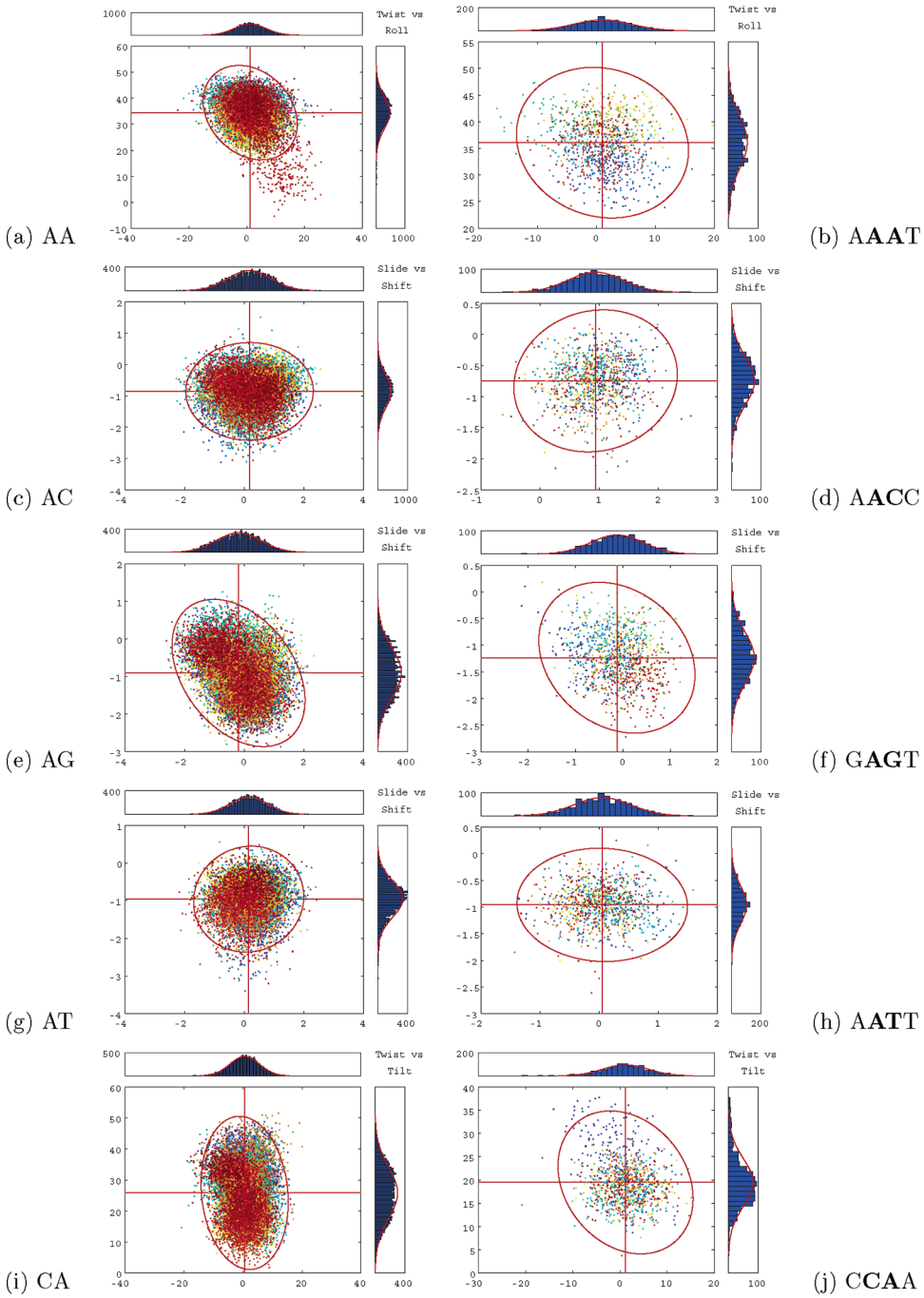


Figure 2. Scatterplots of MD2 data AA, AC, AG, AT, and CA (from top to bottom on the left) and MD4 data with lower correlations A**AA**T, A**AC**C, G**AG**T, A**AT**T, and C**CA**A (from top to bottom on the right). The ellipse is the projection of the six-dimensional equipotential on the bidimensional plane.

reduced through the movement of the step between positive and negative Slides simultaneously with negative and positive Shift movements, thus reducing the repulsion between the G−G interactions. In Packer et al.[38] results, one of the local minima is at Slide = 2.8 Å, Shift = −1.7 Å, and the other at Slide = −2.2 Å, Shift = 0.4 Å. In our MD simulations (see case GG in Figure 3) one minimum appears at Slide = 0.1 Å, Shift = −0.9 Å, and the other at Slide = −1.8 Å, Shift = 0.5 Å.

The deviations from the normal distribution can be due either to an underlying nonharmonic potential surface with respect to the MD motion or as a result of a superposition of thermally accessible substates. Either could be influenced by context effects. We will describe the details of differences between dinucleotide and tetranucleotide conformational properties and non-Gaussian behavior in a following article. Here, we calculate the conformational energy of DNA and *Z*-scores by approximating the distributions by a Gaussian one. We take into account that the culling method for the filtering of the data during the force field calculation, described in Material and Methods, and the saturation method using eq 5 can alleviate in part the nonharmonic behavior of the data and the target conformational states.

3.2. Conformational Energy and *Z*-Scores by Sequence-Structure Threading. To test the ability of the PMF, obtained from MD simulations, to predict realistic sequence-dependent conformation, we estimated the conformational energies and

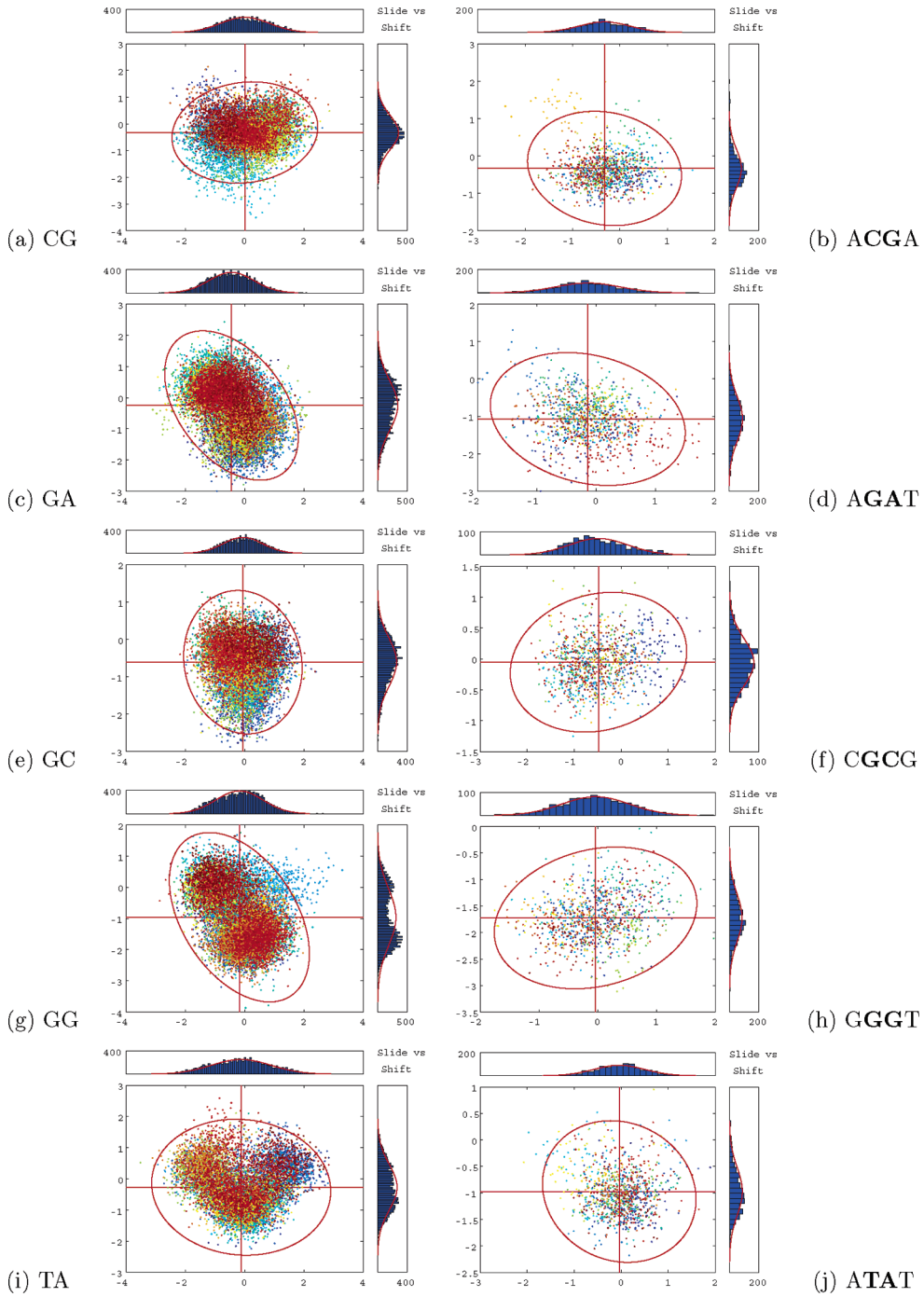


Figure 3. Scatterplots of MD2 data CG, GA, GC, GG, and TA (from top to bottom on the left) and MD4 data with lower correlations A**CG**A, A**GA**T, C**GC**G, G**GG**T and A**TA**T (from top to bottom on the right). The ellipse is the projection of the six-dimensional equipotential on the bidimensional plane.

Z-scores for experimental free DNA crystal structures by sequence-structure threading. This technique can estimate how specific a DNA conformation is with respect to its sequence. As an example, Figure 4 shows the distribution of the *Z*-scores of the random sequences generated with the MD4 force fields for the free B-DNA crystal structure of the PDB file 424d. Clearly, the distribution of the random sequences is approximately Gaussian; the arrow over the black bin marks the position of the *Z*-scores (−3.16) of the conformational energy associated to this structure. Since the *Z*-scores are quite low, we deduce that the conformational state of this structure is very specific of its native sequence 5′-ACCGACGTCGGT-3′.

We considered free DNA crystal structures since these are the standard structures closest to those in MD simulations. Although in theory the MD simulations have to produce structures more similar to the DNA in solution obtained with NMR techniques, it is difficult to judge whether the NMR NOE data are good enough to construct reliable sequence-dependent base-pair conformations. Therefore, here we present only results based on X-ray crystal structures.

As we will illustrate with the "worm" graphs, the *Z*-scores of each step of a sequence based on MD4 seem to have more variability than the *Z*-scores based on MD2. To test this hypothesis, we calculate, for each sequence, the standard

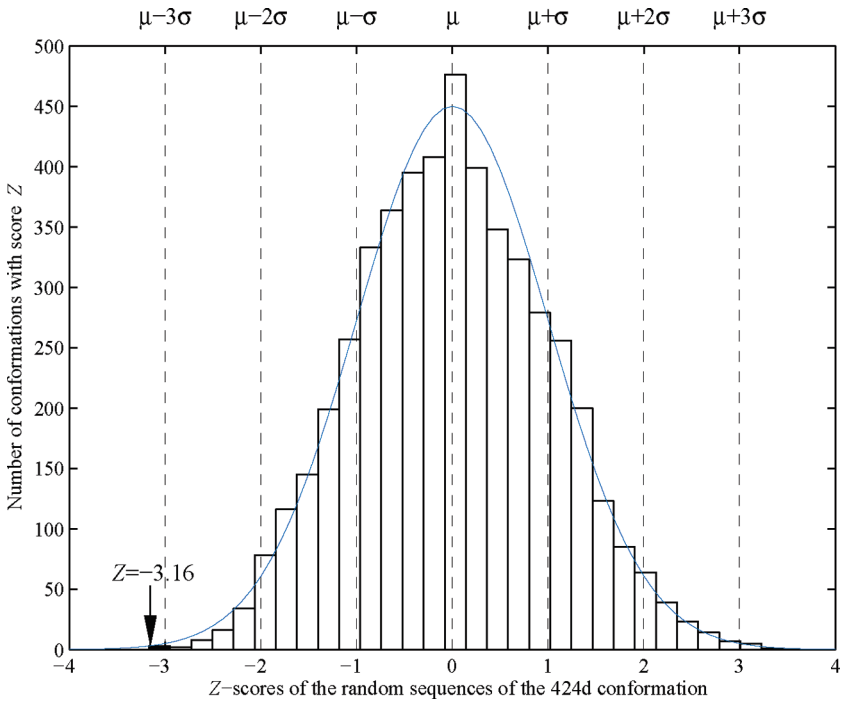


Figure 4. MD4 *Z*-scores distribution for the DNA conformation of the 424d PDB structure and its fit to a Gaussian function. The *Z*-score distribution corresponds to a threading process run over 5000 random sequences. The arrow over the black bin shows the position of the *Z*-scores of the conformation of the DNA native sequence in the 424d structure.

deviation of the step *Z*-scores of the whole sequence, $\sigma_{Z^s_{MD2}}$ and $\sigma_{Z^s_{MD4}}$, using MD2 and MD4 force field matrices, respectively. Then, we calculate the difference $\sigma_{Z^s_{MD2}} - \sigma_{Z^s_{MD4}}$, order the values in increasing order of $Z_{MD2} - Z_{MD4}$, and approximate them with a linear regression; see Supporting Information. From this comparison, we see that the $\sigma_{Z^s_{MD4}}$ values show a trend higher than that of the $\sigma_{Z^s_{MD2}}$ values. This means that the MD4 *Z*-scores of each sequence have more variability, more contrast, than the MD2 *Z*-scores. Since the linear regression has a positive slope, this higher contrast of the MD4 force fields is stronger in the DNA sequences in which the MD4 force fields produce higher *Z*-scores than the MD2 force fields. The contrast of the MD4 force field is attenuated as the Z_{MD4} becomes smaller than the Z_{MD2}.

3.3. Comparison of *Z*-Scores Based on Knowledge-Based and MD Methods. To compare our *Z*-scores based on MD force fields with knowledge-based force fields, we performed several *Z*-score calculations for free B-DNA and A-DNA crystal structures. Figure 5 shows the results for free B-DNA crystal structures. For the knowledge-based Olson force fields from free B-DNA crystal structures, to avoid possible biases in the *Z*-scores calculation we deleted the data used to produce the Olson force fields from the dataset. The Olson *Z*-scores versus the MD2 based *Z*-scores are shown in Figure 5a. We observed that the Olson based *Z*-scores are approximately in the range from −2.5 to 0, with some exceptional cases of high positive *Z*-scores in the PDB files: 1d80, 1dc0, 425d. The *Z*-score distribution is more dispersed in the case of the MD2 based *Z*-scores (from −3.5 to 1.5). The comparison between the MD2 and MD4 based *Z*-scores is shown in Figure 5b. In this case, the comparison is done over the whole free B-DNA crystal dataset. Higher correlation than that in the previous ones is observed, and the MD4 based *Z*-scores tend to show higher values than the MD2 ones.

Figure 6 shows the results for free A-DNA crystal structures. The comparisons between the *Z*-scores produced by the Olson force fields and by the MD2, and between the *Z*-scores by the MD2 force fields and by the MD4, are shown in Figures 6a and b, respectively. Compared with the results of the Olson force fields for free B-DNA (Figure 5a), the Olson force field *Z*-scores reveal a higher dispersion in the free A-DNA case than in the free B-DNA case. Putting Olson versus MD2 (Figure 6 a), opposite to the free B-DNA case (Figure 5a), the Olson *Z*-scores are more dispersed for free A-DNA than the MD2 ones. These observations suggest that the Olson force fields (obtained from B-DNA crystal structures) are more specific for B-DNA, but the MD produces feasible *Z*-scores for both A-DNA and B-DNA structures. During the MD simulations, transitions occur in the trajectories between the B-DNA and the A-DNA conformations, as it had also been observed in other B-DNA MD simulations.[23] As a result of these transitions, MD generates data suitable to estimate the conformational energy not only of B-DNA structures but also of A-DNA. The same trend in the relation between MD4 versus MD2 observed for free A-DNA (Figure 6b) is presented in the free B-DNA case (Figure 5b). Generally, MD4 produces higher *Z*-scores than MD2 in both cases.

The dataset of free B-DNA crystal structures contains the same sequence with different structures. They are interesting examples to evaluate the extent of changes in *Z*-score for the structures with the same DNA sequences. We thus examined the distribution of *Z*-scores within the same sequences; e.g., whether multiple structures of the same sequence exhibit a distribution of *Z*-scores that all fall at the same point, or whether they are distributed. For this purpose, we grouped the free B-DNA crystal structures with the same DNA sequence and obtained 86 sets, out of which 23 sets contain more than one structural member (see Table 6 in the Supporting Information). We calculated the standard deviation of the Z_{MD2}-scores among

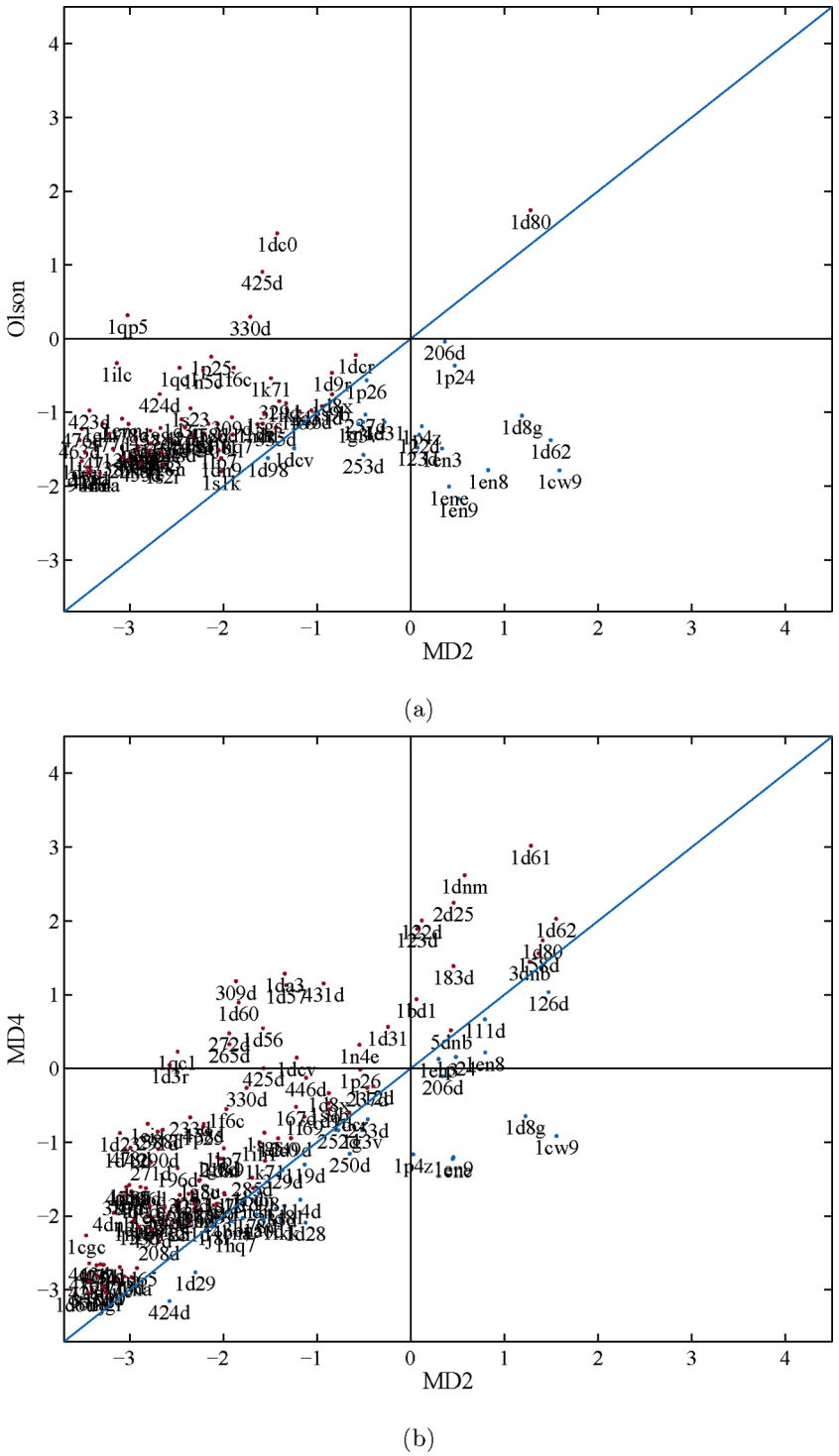


Figure 5. *Z*-scores of free B-DNA crystal structures. (a) *Z*-scores produced with Olson force fields versus MD2 for the free B-DNA set without the files used to obtain the Olson force fields. (b) *Z*-scores with MD4 versus MD2 for the whole free B-DNA set.

the members of the same set. Except for 4 sets, the standard deviations are smaller than 1, compared with the value of 1.34 for the entire B-DNA structure set. The maximum standard deviation is 1.72, and 3 of the remaining 4 sets contain DNA structures with anomalous bases. We can thus conclude that the multiple structures with the same sequence show a lower dispersion of the distribution of the *Z*-scores. This result is reasonable, supporting the reliability of *Z*-score calculations based on MD. The table with *Z*-scores of the whole B-DNA crystal database calculated with knowledge-based and MD methods grouped in sets with the same sequence is presented in the Supporting Information.

3.4. Position Dependence of *Z*-Scores within a DNA Structure. The conformational energies and *Z*-scores can be calculated not only for the whole DNA but also for each strand step. This enables us to obtain more detailed information of the way the conformational changes affect the energy distribution within DNA. The *Z*-scores of the conformational energy

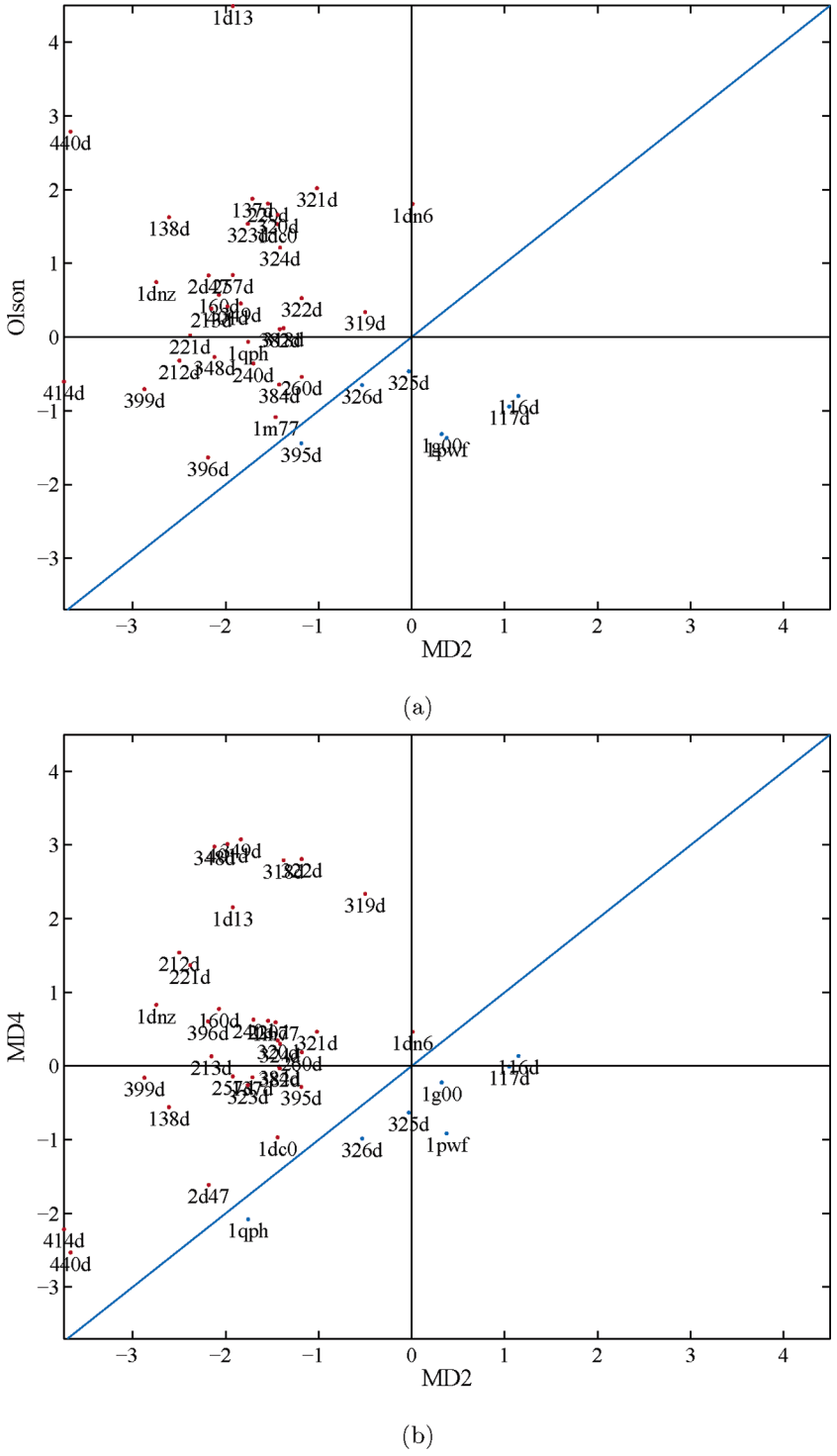


Figure 6. *Z*-scores of free A-DNA crystal structures. (a) *Z*-scores produced with Olson force fields versus MD2. (b) *Z*-scores with MD2 versus MD4 for the whole free A-DNA set.

of each dinucleotide or tetranucleotide step tell us how specific the dependence of a conformational state is on the sequence of bases that build it. The more negative the *Z*-score value is, the more specific the dependence of the conformational state is on the sequence. The variability of the *Z*-scores of a set of tetranucleotide steps with the same central dinucleotide step quantifies the context-dependence of the specificity of the central dinucleotide step. In other words, low standard deviations of the *Z*-scores mean low context-dependence of the specificity, since the different neighbors of a central dinucleotide step do not modify its specificity. This analysis of specificity variance is complementary to the analysis of sequence-dependent DNA structure done[19,20,22,38,39] with the final objective of a better understanding of the sequence-dependent conformational recognition code (indirect-readout) for protein−DNA recognition. To analyze the context-dependence of the central dinucleotide steps with respect to the neighbors, the *Z*-scores of all the tetranucleotide and dinucleotide steps of the whole free B-DNA

(39) Lankaš, F.; Sponer, J.; Langowski, J.; Cheatham, T. E., III. *Biophys. J.* **2003**, *85*, 2872−2883.

crystal structures were calculated using MD2 and MD4 force field matrices, respectively. The histograms of the *Z*-score of all the tetranucleotide steps with the same central dinucleotide step are shown in the Supporting Information.

A more detailed analysis of the statistics of the tetranucleotide steps is presented in Table 1. The last four rows of the table show the statistics of the *Z*-scores produced with the MD2 force field matrices for all the dinucleotide steps. The lower-right side of the table summarizes the results for the whole database, 930 dinucleotide steps, with an average *Z*-score of −0.74 and a standard deviation of 0.85. The rest of the table shows the statistics for the MD4 force field matrices for all the tetranucleotide steps. The tetranucleotide steps are classified according to their central dinucleotide steps. The last four rows of the MD4 show the statistics of the *Z*-scores of the aggregation of all the tetranucleotide steps with the same central dinucleotide step. The last column shows the aggregation of all the tetranucleotide steps with the same neighbors for all the possible different central dinucleotide steps. The lower-right side of the MD4 shows the summary of the results for the whole database, 930 tetranucleotide steps, with an average *Z*-score of −0.49 and a standard deviation of 0.88. We use this standard deviation value as a reference to classify the tetranucleotide steps according to the degree of their context-dependence. In this sense, from the sixth row (starting from the bottom) of Table 1, we see that the least context-dependent specificities are W**AA**Z ($\sigma_{MD4} = 0.64$), W**AG**Z ($\sigma_{MD4} = 0.71$), and W**AC**Z ($\sigma_{MD4} = 0.77$). The most context-dependent are W**CG**Z ($\sigma_{MD4} = 0.91$), W**GA**Z ($\sigma_{MD4} = 0.94$), and W**GG**Z ($\sigma_{MD4} = 1.02$). These results are in broad agreement with the results of Packer et al.[20] based on an ab initio treatment of the base-pair stacking interactions in conjunction with an empirical model for the backbone and the environment.

Compared to the average *Z*-score $\mu_{MD4} = -0.49$, the most specific tetranucleotide is C**CA**C ($\mu_{MD4} = -1.75$). On the other hand, the least specific tetranucleotide is T**AT**T ($\mu_{MD4} = 1.88$). From the average values μ_{MD4} of the *Z*-scores of the aggregation of tetranucleotides with the same central dinucleotide step, the most specific dinucleotide steps seem to be W**GC**Z ($\mu_{MD4} = -0.70$), W**CG**Z ($\mu_{MD4} = -0.74$), and W**GA**Z ($\mu_{MD4} = -0.78$). The least specific dinucleotide steps are W**AT**Z ($\mu_{MD4} = -0.12$), W**TA**Z ($\mu_{MD4} = -0.09$), and W**GG**Z ($\mu_{MD4} = -0.08$). Comparing these last results with the classification of dinucleotides of El Hassan and Calladine,[19] we see no correlation between them. Among the most specific steps, W**GC**Z and W**CG**Z are bistable, and W**GA**Z is rigid. Among the least specific steps, on the other hand W**AT**Z is rigid, whereas W**TA**Z is flexible, and W**GG**Z is bistable.

The statistics of the aggregation of tetranucleotides with the same central dinucleotide and of the corresponding central dinucleotides obtained with MD2 force fields (last four rows of Table 1) reveal that the results are correlated, except for **-CG-**, **-TA-**, and **-AG-** steps. The MD2 force fields produce lower *Z*-scores than the MD4, as we saw from the analysis of the *Z*-scores of the whole sequences (Figure 6b and Figure 5b).

To get insight into the contribution of specificity associated with sequence and conformation within a given DNA structure in a more visual manner, we developed the so-called "worm" graphs showing the distribution of specificity mapped on the 3D structure of a DNA strand in a color codification. In these graphs, the C_1 atoms of the nucleotides are used to represent the DNA backbone. These are drawn as spheres, with a color that is the average color of the surrounding base-pair steps. For the sake of simplicity, only the first DNA strand of each PDB file is drawn. The base steps are represented with cylinders, with a color codifying the *Z*-scores of base steps: high *Z*-scores (low specificity) in red and low *Z*-scores (high specificity) in blue. The "worm" graphs of some of the structures with significant differences between the MD2 and MD4 based *Z*-scores are shown in Figure 7. The structure 1p4z is an example with Z_{MD2} higher than Z_{MD4}. The *Z*-score distribution is quite uniform in the MD2 case, but the MD4 step *Z*-scores show more variability, with higher *Z*-scores in the central G**TA**C tetranucleotide than the corresponding **TA** dinucleotide, and lower *Z*-scores in the initial C**AG**T and proximal terminal A**CT**G steps. Because of these two steps with low *Z*-scores, the global result is more negative (more specific) for the MD4 based results. In contrast to 1p4z, the 1d23 structure is an example with Z_{MD4} higher than Z_{MD2}. The *Z*-score distribution is again quite uniform in the MD2 case, with higher variability in the MD4 step *Z*-scores, with higher *Z*-scores in the central A**TC**G tetranucleotide and in the initial C**GA**T and terminal A**TC**G steps. Structures 1d56 and 1d49 are other examples with Z_{MD4} higher than Z_{MD2}, both of them with very similar sequences. The structure 1d56 contains a TATA sequence, in the central T**AT**A step. Both force fields produce high *Z*-scores for this step (Z_{MD4} higher than Z_{MD2}). The main difference between the two force fields results is in the terminals. In both terminals C**GA**T and A**TC**G the MD4 *Z*-scores are higher than the MD2 *Z*-scores. The central tetranucleotide of 1d49 is T**TA**A instead of T**AT**A. Opposite to the 1d56 case, the force fields produce low *Z*-scores in the central T**TA**A step. In 1d49 both force field results differ in the terminals, as in 1d56. In both terminals C**GA**T and A**TC**G, the MD4 *Z*-scores are higher than the MD2 *Z*-scores. The interesting point in the 1d56−1d49 examples is that, even with similar sequences, both MD2 and MD4 force field matrices detect the dissimilarity, since they produce different global *Z*-scores. And the step *Z*-scores detect the steps which cause the difference in the global *Z*-scores.

It is also interesting to compare these results with the statistics for tetranucleotide steps given in Table 1. T**TA**A has an almost zero average *Z*-score $\mu_{MD4} = -0.04$ and the highest dispersion $\sigma_{MD4} = 1.36$ over all the tetranucleotide steps. T**AT**A has a positive average *Z*-score $\mu_{MD4} = 1.20$ and a low dispersion $\sigma_{MD4} = 0.24$. Both tetranucleotides are in the group of neighbors T**XY**A. This group disperses significantly ($\sigma_{MD4} = 1.04$) their central dinucleotide step. The high μ_{MD4} value for T**AT**A means that this tetranucleotide is unspecific of the conformational states in which it appears. Then, the conformational states in which it takes place can be occupied by other tetranucleotides. And since its dispersion is low, this behavior is maintained in the different conformational steps in which it appears. This confers the ubiquitous capability on the TATA tetranucleotide. These features of the TATA sequence are complementary to its properties to be very flexible, to have few Watson−Crick hydrogen bonds and weak stacking interactions.[20] This makes it an obvious candidate for unwrapping the double helix at an origin of replication.[40] With respect to the different behaviors

(40) Watson, J. D.; Backer, T. A.; Bell, S. P.; Gann, A.; Levine, M.; Losick, R. *Molecular Biology of the Gene*, 5th edition; Benjamin-Cummings: New York, 2003.

Table 1. Statistics of the Step *Z*-Scores Obtained with MD4 Force Fields for All the Tetranucleotide Steps (W**X**YZ) of the Free B-DNA Crystal Dataset[a]

	WCGZ	WCAZ	WTAZ	WAGZ	WGGZ	WAAZ	WGAZ	WATZ	WACZ	WGCZ	
AXY **A**	A**C**G**A**	A**C**A**A**	A**T**A**A**	A**A**G**A**	A**G**G**A**	A**A**A**A**	A**G**A**A**	A**AT**A	A**AC**A	A**GC**A	**A**XY**A**
μ_{MD4}	0.95	−0.74		0.05		−0.84	−0.83		−0.28		−0.60
σ_{MD4}	0.00	0.00		0.15		0.26	0.00		0.18		0.52
$\#_{MD4}$	1	1		2		13	1		2		20
AXY **C**	A**C**G**C**	A**C**A**C**	A**T**A**C**	A**A**G**C**	A**G**G**C**	A**A**A**C**	A**G**A**C**	A**AT**C	A**AC**C	A**GC**C	**A**XY**C**
μ_{MD4}	−0.39	−0.78	−1.36	0.61	−0.62	−0.83		−0.53			−0.72
σ_{MD4}	0.97	0.32	0.00	0.51	0.24	0.29		0.00			0.47
$\#_{MD4}$	3	24	1	3	2	28		1			62
AXY **G**	A**C**G**G**	A**C**A**G**	A**T**A**G**	A**A**G**G**	A**G**G**G**	A**A**A**G**	A**G**A**G**	A**AT**G	A**AC**G	A**GC**G	**A**XY**G**
μ_{MD4}	−0.57	−0.06		−0.43	−0.64	−0.61	0.42	0.71	−0.40	−0.69	−0.26
σ_{MD4}	0.99	0.73		0.65	0.59	0.58	0.70	1.39	0.60	0.27	0.83
$\#_{MD4}$	2	3		4	6	16	10	6	8	7	62
AXY **T**	A**C**G**T**	A**C**A**T**	A**T**A**T**	A**A**G**T**	A**G**G**T**	A**A**A**T**	A**G**A**T**	A**ATT**	A**AC**T	A**GC**T	**A**XY**T**
μ_{MD4}	−1.18		−0.25			−0.82	−0.76	−0.54		−0.43	−0.59
σ_{MD4}	0.22		0.38			0.30	0.59	0.36		1.11	0.43
$\#_{MD4}$	6		10			11	3	53		2	85
CXY **A**	C**C**G**A**	C**C**A**A**	C**T**A**A**	C**A**G**A**	C**G**G**A**	C**A**A**A**	C**G**A**A**	C**AT**A	C**AC**A	C**GC**A	**C**XY**A**
μ_{MD4}	−0.96	0.60			0.37	−0.87	−1.01	−0.24	−0.65	−1.00	−0.70
σ_{MD4}	0.65	0.36			0.96	0.68	0.89	0.00	0.17	0.69	0.92
$\#_{MD4}$	5	11			3	11	37	1	4	11	83
CXY **C**	C**C**G**C**	C**C**A**C**	C**T**A**C**	C**A**G**C**	C**G**G**C**	C**A**A**C**	C**G**A**C**	C**AT**C	C**AC**C	C**GC**C	**C**XY**C**
μ_{MD4}	0.64	−1.75	−1.10	−0.91	−0.20	0.80	−1.35		0.36	0.05	0.02
σ_{MD4}	1.01	0.00	0.00	0.37	0.98	0.13	0.77		0.82	0.67	0.96
$\#_{MD4}$	7	1	1	3	13	3	2		12	11	53
CXY **G**	C**C**G**G**	C**C**A**G**	C**T**A**G**	C**A**G**G**	C**G**G**G**	C**A**A**G**	C**G**A**G**	C**AT**G	C**AC**G	C**GC**G	**C**XY**G**
μ_{MD4}	−0.29	0.60	0.51	0.89	−0.44	1.49	−1.44	0.32	−1.14	−0.89	−0.49
σ_{MD4}	1.13	0.20	0.13	0.05	0.22	0.27	0.48	1.12	0.15	0.88	1.05
$\#_{MD4}$	6	15	3	2	3	4	4	10	4	85	136
CXY **T**	C**C**G**T**	C**C**A**T**	C**T**A**T**	C**A**G**T**	C**G**G**T**	C**A**A**T**	C**G**A**T**	C**ATT**	C**AC**T	C**GC**T	**C**XY**T**
μ_{MD4}	−1.39	−0.07		−0.84	0.05	−0.63	0.54	1.68	−0.86	−0.84	−0.17
σ_{MD4}	0.00	0.32		0.24	0.97	0.55	1.09	0.50	0.16	0.03	1.01
$\#_{MD4}$	1	3		4	6	10	10	4	6	5	49
GXY **A**	G**C**G**A**	G**C**A**A**	G**T**A**A**	G**A**G**A**	G**G**G**A**	G**A**A**A**	G**G**A**A**	G**AT**A	G**AC**A	G**GC**A	**G**XY**A**
μ_{MD4}	−1.02	−0.89		−0.88	−0.26	−0.64	−0.98	−0.43	−0.91		−0.88
σ_{MD4}	0.53	0.54		0.18	1.14	0.55	0.00	0.25	0.00		0.56
$\#_{MD4}$	31	16		3	3	5	1	3	1		63
GXY **C**	G**C**G**C**	G**C**A**C**	G**T**A**C**	G**A**G**C**	G**G**G**C**	G**A**A**C**	G**G**A**C**	G**AT**C	G**AC**C	G**GC**C	**G**XY**C**
μ_{MD4}	−0.55		0.70	−0.75	−0.55			−0.62		0.01	−0.27
σ_{MD4}	0.44		0.27	0.22	0.16			0.40		0.73	0.66
$\#_{MD4}$	14		6	2	2			5		5	34
GXY **G**	G**C**G**G**	G**C**A**G**	G**T**A**G**	G**A**G**G**	G**G**G**G**	G**A**A**G**	G**G**A**G**	G**AT**G	G**AC**G	G**GC**G	**G**XY**G**
μ_{MD4}	1.38	−0.56	−0.39	−0.80	−0.28	0.02	−1.37		0.43	−0.07	0.02
σ_{MD4}	0.34	0.00	0.00	0.30	0.68	1.09	0.23		0.67	0.58	0.87
$\#_{MD4}$	6	1	1	7	10	5	2		13	13	58
GXY **T**	G**C**G**T**	G**C**A**T**	G**T**A**T**	G**A**G**T**	G**G**G**T**	G**A**A**T**	G**G**A**T**	G**ATT**	G**AC**T	G**GC**T	**G**XY**T**
μ_{MD4}	−1.19	−0.87	−1.14	0.17	−0.72	−0.72		−0.04			−0.73
σ_{MD4}	0.00	0.45	0.00	0.34	0.00	0.42		0.79			0.50
$\#_{MD4}$	2	22	1	3	1	28		2			59
TXY **A**	T**C**G**A**	T**C**A**A**	T**T**A**A**	T**A**G**A**	T**G**G**A**	T**A**A**A**	T**G**A**A**	T**AT**A	T**AC**A	T**GC**A	**T**XY**A**
μ_{MD4}	−1.11		−0.04	−0.16	0.39	−1.28	−0.87	1.20	−0.38		−0.10
σ_{MD4}	0.62		1.36	0.05	0.02	0.04	0.50	0.24	0.16		1.04
$\#_{MD4}$	3		6	4	2	2	4	6	2		29
TXY **C**	T**C**G**C**	T**C**A**C**	T**T**A**C**	T**A**G**C**	T**G**G**C**	T**A**A**C**	T**G**A**C**	T**AT**C	T**AC**C	T**GC**C	**T**XY**C**
μ_{MD4}	−1.28	−0.61		0.24	1.68	−0.64	−0.98	−0.69	−0.59		−0.68
σ_{MD4}	0.44	0.69		1.20	1.61	0.10	0.00	0.50	0.00		1.10
$\#_{MD4}$	27	13		4	5	2	1	3	1		56
TXY **G**	T**C**G**G**	T**C**A**G**	T**T**A**G**	T**A**G**G**	T**G**G**G**	T**A**A**G**	T**G**A**G**	T**AT**G	T**AC**G	T**GC**G	**T**XY**G**
μ_{MD4}	−1.61	0.67	−1.29		−0.59	−0.85	−1.09	0.54	−0.63	−0.82	−0.67
σ_{MD4}	0.23	1.11	0.00		0.00	0.42	0.53	0.92	0.20	0.20	0.93
$\#_{MD4}$	3	13	1		1	6	34	2	3	10	73
TXY **T**	T**C**G**T**	T**C**A**T**	T**T**A**T**	T**A**G**T**	T**G**G**T**	T**A**A**T**	T**G**A**T**	T**ATT**	T**AC**T	T**GC**T	**T**XY**T**
μ_{MD4}	0.47		−0.64			−1.33		1.88	−0.25		−0.35
σ_{MD4}	0.00		0.00			0.34		0.00	0.21		1.13
$\#_{MD4}$	1		1			3		1	2		8
W**XY**Z	W**CG**Z	W**CA**Z	W**TA**Z	W**AG**Z	W**GG**Z	W**AA**Z	W**GA**Z	W**AT**Z	W**AC**Z	W**GC**Z	W**XY**Z
μ_{MD4}	−0.74	−0.32	−0.09	−0.32	−0.08	−0.66	−0.78	−0.12	−0.19	−0.70	−0.49
σ_{MD4}	0.91	0.86	0.85	0.71	1.02	0.64	0.94	0.90	0.77	0.83	0.88
$\#_{MD4}$	118	123	31	41	57	147	109	97	58	149	930
-**XY**-	-**CG**-	-**CA**-	-**TA**-	-**AG**-	-**GG**-	-**AA**-	-**GA**-	-**AT**-	-**AC**-	-**GC**-	-**XY**-
μ_{MD2}	−1.34	−0.07	−0.80	0.31	−0.49	−0.67	−1.16	−0.54	−0.39	−0.90	−0.74
σ_{MD2}	0.40	0.99	0.68	0.49	0.88	0.78	0.57	1.00	0.92	0.64	0.85
$\#_{MD2}$	140	95	31	43	39	175	111	94	35	167	930

[a] The last column shows the statistics for the same central dinucleotide step within the one neighbor context. The last four rows show the statistics of the *Z*-scores obtained with MD2 force fields for all of the dinucleotide steps (-**XY**-). μ, σ, and # stand for average, standard deviation, and number of steps found in the crystal structures, respectively.

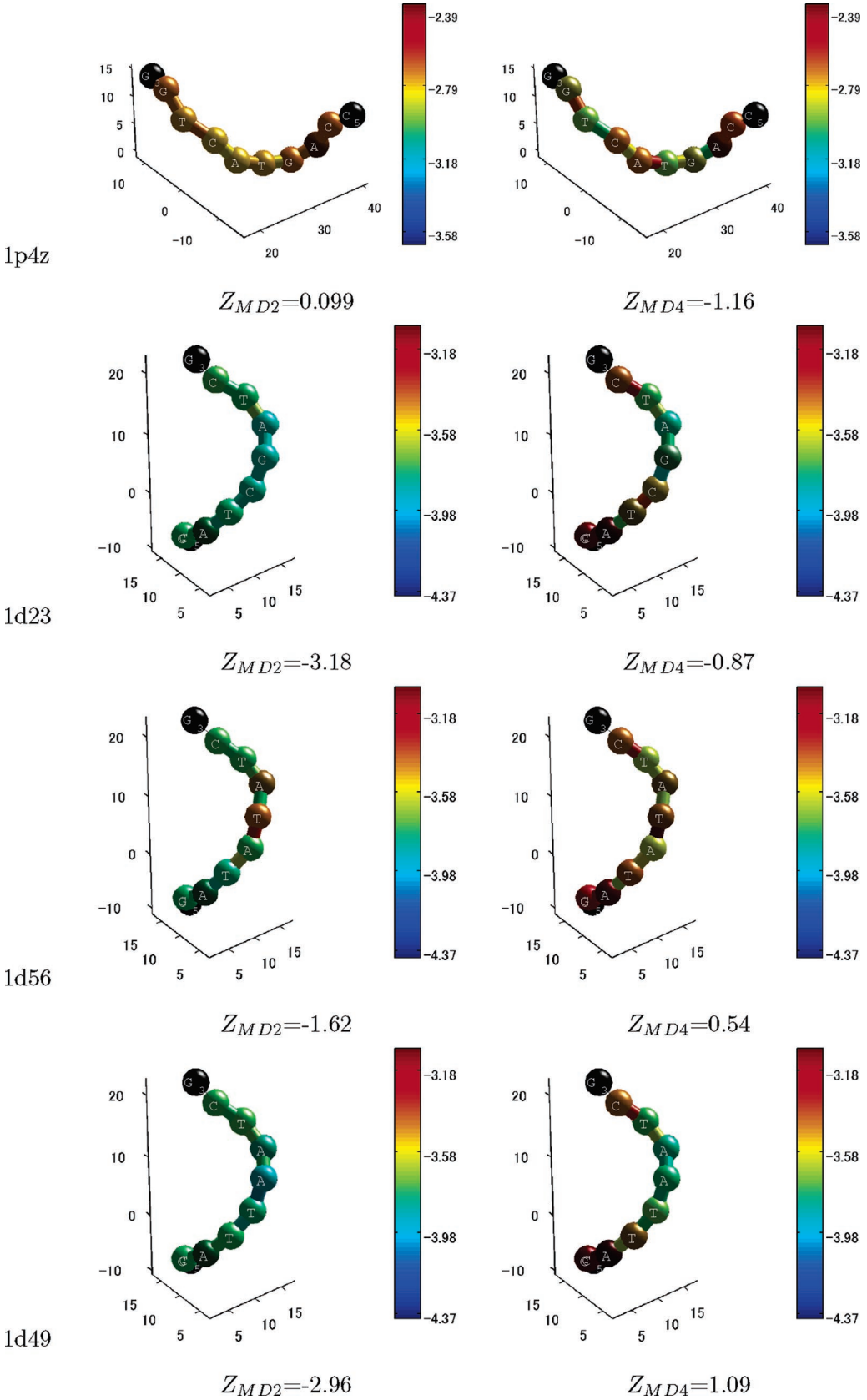


Figure 7. "Worm" representation of the free DNA sequences, using dinucleotide (left) and tetranucleotide (right) based force fields. The color codifies the *Z*-score value of the energy in each dinucleotide step (the blue color indicates high specificity, and the red, low). The tube passes across the backbone formed by the C_1 carbons of each nucleotide. The terminal bases are represented with black spheres and the terminal steps, for which the *Z*-scores are not calculated, with thin black lines. The subindices 5 and 3 stand for terminals 5′ and 3′ of the DNA sequences.

between MD2 and MD4 based *Z*-scores of the terminal A**TC**G, this tetranucleotide corresponds (is symmetric) to the tetranucleotide C**GA**T, which has an average *Z*-score in the whole database (Table 1) of $\mu_{MD4} = 0.54$ significantly higher than the

average *Z*-scores obtained with MD2 force fields of its corresponding central dinucleotide **-GA-** $\mu_{MD2} = -1.16$.

In general, the *Z*-scores of the whole sequence produced by MD2 force fields are lower (more specific) than the *Z*-score produced by MD4. However the "worm" graphs show that the MD4 *Z*-score distributions offer more contrast between the different steps than the MD2 *Z*-score distributions. It indicates that a possible reason for the global *Z*-scores based on MD2 force field matrices are smaller than those based on MD4 is a boundary effect at the tetranucleotide chain terminal steps. In several cases we observed higher *Z*-scores of the terminal tetranucleotide than for the corresponding central dinucleotide. For *Z*-scores based on MD2, we eliminated the terminal dinucleotides of the DNA sequence to alleviate the possible boundary effects. But, in the case of tetranucleotides analysis, to have the same number of steps as in the dinucleotide case, the terminal steps are preserved. Thus, if the terminal of a DNA sequence is *WXYZ*-3′, in the dinucleotide oriented analysis, the step *YZ*-3′ is eliminated, and the last dinucleotide step remains *XY*. But in the tetranucleotide oriented analysis, the preservation of the step *YZ*-3′ lets us compare the *Z*-scores of the terminal tetranucleotide step *WXYZ*-3′ with those of its equivalent central dinucleotide step *XY*. One possible method to overcome the boundary steps in the tetranucleotide analysis is to delete the last two dinucleotide steps of each terminal. But since the available free DNA crystal structures are of short length, this method produces a significant reduction in the available data. Therefore, we decided to eliminate only the last terminal dinucleotide steps.

Discussion

We have developed a PMF model for calculating the conformational energy and specificity of DNA. Here, we have taken a systematic approach, considering the 136 unique possible tetranucleotide sequences at the center of dodecamer DNA, to produce a large number of MD trajectories. Using a simplified conformational model with six parameters to describe the geometry of adjacent base pairs and harmonic potentials along these coordinates, we estimated the PMFs from those trajectories. The PMFs derived from MD trajectories made it possible to estimate the conformational energy and the specificity for any given DNA sequence and structure. We have used a sequence-structure threading method to estimate the *Z*-score as a measure of specificity for many B-DNA and A-DNA crystal structures. The average *Z*-scores were negative for both types of structures, indicating that the PMFs are capable of predicting sequence specificity for both types of DNA structures. We have also shown that the distribution of conformational energy and *Z*-score within DNA are strongly position dependent, indicating that this kind of analysis enables us to identify particular conformations responsible for the specificity.

A similar systematic approach to explore all the tetranucleotides by MD simulations was first set forth by the Ascona B-DNA Consortium of Beveridge et al.[23] We have compared the average conformational parameters derived from our MD trajectories with their results for the tetranucleotides X**CG**W, which were the only ones presented. We calculated the root-mean-square difference (RMSD) between the corresponding average conformational parameters over those sequences. Both of the results are similar, as the RMSDs are within the average standard deviations of the corresponding conformational parameters. Recently, Hays et al.[41] have reported a systematic approach to explore conformational space of DNA experimentally by crystallizing all possible trinucleotide sequence permutations within a defined inverted repeat sequence motif under nearly identical solution conditions. This would be the experimental complement to the present computational strategy, and such a dataset would provide a better benchmark to test our method.

The present approach has an advantage over the knowledge-based approach, which uses known structures of DNA to derive statistical potentials for DNA conformational parameters, in that we can avoid possible biases that appear when the potentials are calculated from a limited number of real structures. However, there are a number of limitations to the present approach: It remains to be checked rigorously whether the conformation samplings for each of 136 tetranucleotides are sufficient or not. As far as we have carried out 10-ns runs for several of the teteranuclotides and compared their ensembles with those obtained by 2-ns runs, for most of the cases, especially for sequences having a purine-pyrmidine dimer step at the center, the χ^2 tests did not reject, with 5% significance, the null hypothesis that two ensembles derived from 10-ns and 2-ns runs have the same distribution (data not shown). The PMFs were coarse-grained by the assumption of rigid-body base pair and a harmonic approximation. We observed that conformational parameters calculated from the MD trajectories obeyed the Gaussian distribution for most of the sequences, but some sequences such as ACGA and TCAA exhibited non-Gaussian behavior, e.g., bimodal or distorted Gaussian distributions. For such sequences, we need more careful examination of the MD trajectories with longer runs, although we expect that the results based on the coarse-grained potentials would not be affected so drastically. In future works, we will make more extensive analysis of MD trajectories and conformational parameters and consider more complex energetic models for the nonharmonicity due to the non-Gaussian distribution, based on the knowledge obtained from such analysis.

DNA is one of the most important biomolecules, carrying genetic information. There has been increasing evidence to suggest that the structures of DNA play an active role in the regulation of genetic information by proteins through sequence-dependent conformational properties of DNA (so-called indirect readout mechanism).[42,43] The present studies of DNA by molecular dynamics simulations are intended not only to characterize the structure and dynamics of DNA but also to extract information about sequence-dependent PMFs from those results and apply them to predict sequence-dependent conformational energy and specificity in protein−DNA recognition and ligand−DNA interactions. If the ensembles of DNA conformation obtained by the MD simulations are wide enough to cover the observed conformations of DNA in protein−DNA or ligand−DNA complex structures, the present results can be applied to evaluate the conformational energy and specificity of protein−DNA or ligand−DNA recognition. The present results have shown that the PMFs derived from MD simulations

(41) Hays, F. A.; Teegarden, A.; Jones, Z. J. R.; Harms, M.; Raup, D.; Watson, J.; Cavaliere, E.; Shing Ho, P. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (20), 7157−7162.
(42) Dickerson, R. E.; Chiu, T. K. *Biopolymers* **1997**, *44*, 361−403.
(43) Vandervliet, P. C.; Verrijzer, C. P. *Bioessays*, **1993**, *15*, 25−32.

can account for the sequence specificity of DNA structures solved by X-ray crystallography, indicating that the force field parameters used for the MD simulations are good enough to describe sequence-dependent DNA conformations. Although we need to examine other sets of parameters used in the MD simulation more carefully, the significant sequence specificity of DNA structure obtained by the current analysis indicates a predictive power of the introduced method. We would like to apply the obtained results based on free DNA structure to protein−DNA complexes, calculating the *Z*-scores for the complex structures and comparing them with the results obtained by the knowledge-based approach. We will also apply the present method to ligand−DNA interactions, to decipher the role of sequence-dependent DNA conformations in ligand recognition.

Acknowledgment. M.J.A.-B. would like to acknowledge the Japanese Society for the Promotion of Science (JSPS) for supporting him for this research. This work is supported in part by Grants-in-Aid for Scientific Research 16014219 and 16041235 (A. Sarai) and 16014226 (H. Kono) from Ministry of Education, Culture, Sports, Science and Technology in Japan. We thank Prof. N. Go for encouraging this work and providing useful comments. Part of the MD calculations were carried out using ITBL computer facilities at JAERI.

Supporting Information Available: An example of the meaning of the perturbation of the conformational coordinates.The 256→136 tetranucleotide mapping. The average and standard deviation of the conformational coordinates and the 10 dinucleotide and 136 tetranucleotide force field matrices. The differences between the *Z*-scores energies produced from the MD2 force field matrices and those produced from the MD4 force field matrices for the free B-DNA crystal structure dataset. Comparison of *Z*-scores from knowledge-based and MD methods. Histograms of the *Z*-score of all the tetranucleotide steps with the same central dinucleotide step. This material is available free of charge via the Internet at http://pubs.acs.org.

JA053241L