Published online 3 September 2009 *Nucleic Acids Research, 2009, Vol. 37, No. 20* ***e135***
doi:10.1093/nar/gkp718

A generalized conformational energy function of DNA derived from molecular dynamics simulations

Satoshi Yamasaki[1], Tohru Terada[2,*], Kentaro Shimizu[1,2], Hidetoshi Kono[3,4] and Akinori Sarai[5]

[1]Intelligent Modeling Laboratory, The University of Tokyo, 2-11-16 Yayoi, Bunkyo-ku, Tokyo 113-8656, [2]Agricultural Bioinformatics Research Unit and Department of Biotechnology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, [3]Computational Biology Group, Neutron Biology Research Center, Quantum Beam Science Directorate, [4]Quantum Bioinformatics Team, Center for Computational Science and e-Systems, Japan Atomic Energy Agency, 8-1 Umemidai, Kizugawa, Kyoto 619-0215 and [5]Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

Received December 24, 2008; Accepted August 14, 2009

ABSTRACT

Proteins recognize DNA sequences by two different mechanisms. The first is direct readout, in which recognition is mediated by direct interactions between the protein and the DNA bases. The second is indirect readout, which is caused by the dependence of conformation and the deformability of the DNA structure on the sequence. Various energy functions have been proposed to evaluate the contribution of indirect readout to the free-energy changes in complex formations. We developed a new generalized energy function to estimate the dependence of the deformability of DNA on the sequence. This function was derived from molecular dynamics simulations previously conducted on B-DNA dodecamers, each of which had one possible tetramer sequence embedded at its center. By taking the logarithm of the probability distribution function (PDF) for the base-step parameters of the central base-pair step of the tetramer, its ability to distinguish the native sequence from random ones was superior to that with the previous method that approximated the energy function in harmonic form. From a comparison of the energy profiles calculated with these two methods, we found that the harmonic approximation caused significant errors in the conformational energies of the tetramers that adopted multiple stable conformations.

INTRODUCTION

Sequence-specific recognition of DNA by proteins plays a critical role in regulating gene expression. Accurate recognition is achieved by a combination of two different mechanisms (1,2). First, the affinity of a protein to a DNA sequence depends on the number of favorable interactions formed between the DNA and the protein. The interaction sites of DNA include the functional groups of the bases exposed on the surface of the minor and major grooves. Since the arrangement of the functional groups of the bases differs between DNA sequences, the protein can recognize a specific DNA sequence. This recognition mechanism is so-called direct readout. Second, the binding affinity also depends on the strengths of the interactions. Therefore, proteins may prefer a specific conformation of DNA, or DNA that can easily deform its conformation to strengthen the interactions. As a result, the proteins recognize the DNA bases that change the DNA conformation to a specific one and/or provide deformability without directly interacting with the bases. In contrast to direct readout, this recognition mechanism is called indirect readout. Biochemical studies have demonstrated that indirect readout is as important as direct readout for determining the specificity for some protein–DNA complexes (3–7). However, compared with direct readout, it is difficult to evaluate the contribution of indirect readout for a given protein–DNA complex.

Thermodynamically, protein–DNA binding can be virtually decomposed into two processes: free DNA changes its conformation to the one to be adopted in the complex, and the protein binds to the deformed

*To whom correspondence should be addressed. Tel: +81-48-462-1111 (Ext. 3179); Fax: +81-48- 462-1625; Email: tterada@riken.jp
Present addresses:
Satoshi Yamasaki, Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan.
Tohru Terada, Molecular Scale Team, Computational Science Research Program, RIKEN, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan.

DNA. The free-energy change for the former process corresponds to the contribution of indirect readout to the total free-energy change when the complex is formed. Statistics-based methods (8–15) and a molecular-mechanics-based method (16) have been used to evaluate the contribution of indirect readout. In the statistics-based studies, base-pair step parameters (shift, slide, rise, tilt, roll and twist) that represent the relative configuration between two successive base pairs are often used to describe the internal degrees of freedom of DNA with a reduced number of variables (8,17). Olson *et al.* (8) analyzed the dependence of the distributions of the step parameters on sequence using DNA structures bound to proteins. They demonstrated that the conformational energy of a given base-pair step could be estimated using a harmonic potential, whose force constant matrix and average geometry were calculated from the distribution of the step parameters of dinucleotides having the same sequence in the set of DNA structures. Gromiha *et al.* (11,12) extended this method to quantify the contribution of indirect readout to specificity. They used the *Z*-score defined as $\left(E(s,\boldsymbol{\Theta}) - \langle E(\boldsymbol{\Theta})\rangle\right)/\sigma(\boldsymbol{\Theta})$ as a measure of specificity, where $E(s, \boldsymbol{\Theta})$ is the conformational energy of DNA having sequence s and step parameters $\boldsymbol{\Theta}$, and $<E(\boldsymbol{\Theta})>$ and $\sigma(\boldsymbol{\Theta})$ correspond to the average and the standard deviation of the conformational energies, obtained by threading random sequences onto the DNA structure. A large negative *Z*-score indicates that indirect readout plays a large role in determining specificity for the protein–DNA complex. Although these studies used experimental structures to obtain the parameters of the harmonic potentials, Araúzo-Bravo *et al.* (13) and Fujii *et al.* (15) instead used the structures generated by molecular dynamics (MD) simulations. They carried out MD simulations for all of the 136 unique tetramer sequences, embedding them at the centers of DNA dodecamers and calculated the conformational energies as a function of the tetramer sequence and the step parameters of the central base-pair step. It should be noted that it is difficult to determine the harmonic-potential parameters for all possible tetramer sequences with experimental structures because of an insufficient number of data for some tetramer sequences. Fujii *et al.* found that the deformability of a base-pair step depends on its flanking base pairs. This dependency of deformability on flanking base pairs has also been reported in a series of molecular mechanical studies (18–20). Therefore, it is important to consider such a ‘long-range effect’ in calculating the conformational energy.

In these studies, harmonic potentials were used to evaluate the conformational energies, based on the assumption that the probability distribution function (PDF) of the step parameters could be approximated with a Gaussian function. However, Fujii *et al.* (15) pointed out that the PDFs of some sequences were not Gaussian. Although such an assumption is inevitable when the number of available structures is limited, we can obtain a large number of structures from the MD simulations and can use a more general function to accurately describe the distribution of the step parameters. We therefore developed a new generalized energy function in the present study using the same MD data as Fujii *et al.* Here, the energy function was simply defined as the logarithm of the PDF. The purpose of developing the energy function was to clarify the mechanism responsible for protein–DNA recognition in terms of the energetics. To evaluate the accuracy of the present method, we examined the capability of the new energy function to distinguish the native sequence from random ones by using free DNA structures, and compared it with that of the previous method.

MATERIALS AND METHODS

Calculation of energy functions

We used a set of structural ensembles derived from MD simulations conducted by Fujii *et al.* for all 136 unique tetramer sequences embedded at the centers of B-DNA dodecamers (5′-CGCG–$n_1n_2n_3n_4$ –CGCG–3′; n_i is either A, T, G, or C) (13,15). Although the simulations were carried out for 10 ns, we used the data from the last 9 ns, as they did. Since the snapshot structures were recorded at every picosecond, we had 9000 structures for each sequence. The six base-pair step parameters (shift, slide, rise, tilt, roll and twist) were calculated for the n_2–n_3 step of each snapshot structure by using the X3DNA program (21). Note that although there were $4^4 = 256$ possible tetramer sequences, there were 136 unique tetramer sequences and the remaining 120 sequences could be described by taking sequence complementarity into consideration.

The PDF of step parameters $\boldsymbol{\Theta}$ in the six-dimensional space was calculated for each of the 256 possible tetramer sequences ($s = n_1n_2n_3n_4$). The six-dimensional space was divided into 13^6 cells with cell sizes of $[\max(\theta_i) - \min(\theta_i)]/13$, where $\max(\theta_i)$ and $\min(\theta_i)$ were the maximum and minimum values of the i-th component of $\boldsymbol{\Theta}$ in the whole structural ensemble. The PDF is given by

$$P(s,\boldsymbol{\Theta}) = \begin{cases} \frac{n(s,\boldsymbol{\Theta})}{N\Delta V J(\boldsymbol{\Theta})}, & n(s,\boldsymbol{\Theta}) \neq 0, \\ \frac{1}{N\Delta V}, & n(s,\boldsymbol{\Theta}) = 0, \end{cases} \qquad 1$$

where N is the number of snapshot structures (i.e. 9000), $n(s, \boldsymbol{\Theta})$ is the number of snapshot structures having sequence s and the step parameters that fall within the cell at $\boldsymbol{\Theta}$, $J(\boldsymbol{\Theta})$ is the Jacobian, and ΔV is the volume of the cell. When $n(\mathrm{s}, \boldsymbol{\Theta})$ is zero, $P(s, \boldsymbol{\Theta})$ is set to $1/N\Delta V$ to avoid taking the logarithm of zero. By using the PDF, the conformational energy, or strictly speaking, the free-energy difference between a state having step parameter $\boldsymbol{\Theta}$ and the equilibrium state is calculated as

$$E_{\mathrm{G}}(s,\boldsymbol{\Theta}) = -kT \ln P(s,\boldsymbol{\Theta}), \qquad 2$$

where k is the Boltzmann constant and T is the temperature. Here, we refer to this function as the generalized energy function, E_{G}, since this can take an arbitrary form. Following Fujii *et al.*, we used a reduced unit system with $kT = 1$. By using the relation between the step parameters and Euler angles (21), the Jacobian is calculated as

$$J(\boldsymbol{\Theta}) = \frac{\sin\sqrt{\tau^2+\rho^2}}{\sqrt{\tau^2+\rho^2}}, \tag{3}$$

where τ and ρ correspond to the tilt and roll angles. Note that $J(\mathbf{\Theta})$ is between 0 and 1 for the values that the tilt and roll angles normally assume. In the present study, however, the Jacobian had no effect on the Z-scores because the energies were compared between different sequences for a fixed structure. If the energies are compared between different structures, it is of course necessary to explicitly consider the Jacobian. The energy function used by Fujii *et al.* is given by

$$E_{\mathrm{HA}}(s,\boldsymbol{\Theta}) = \frac{kT}{2}\Delta\boldsymbol{\Theta}^{\mathrm{T}}\mathbf{F}_s\,\Delta\boldsymbol{\Theta} = \frac{kT}{2}\Delta\boldsymbol{\Theta}^{\mathrm{T}}\mathbf{M}_s^{-1}\Delta\boldsymbol{\Theta}, \tag{4}$$

where $\Delta\boldsymbol{\Theta} = (\Delta\theta_1, K, \ldots, \Delta\theta_6)$, $\Delta\theta_i = \theta_i - \langle\theta_i\rangle_s$ and $\mathbf{F}_s$ is the force constant matrix for sequence s and is the inverse of the variance–covariance matrix of step parameters $\mathbf{M}_s$ calculated from the last 9-ns trajectory of the 10-ns MD simulation undertaken on the DNA embedding sequence, s (15). Here, $<\ldots>_s$ denotes taking the average of a variable over the 9-ns trajectory of sequence s. This function is referred to as the harmonic-approximation (HA) energy function, E_{HA}.

Evaluation of accuracy

We evaluated the accuracy of E_{G} based on its capability to distinguish the native sequence from random ones, and compared it with that of E_{HA}. Its capability was quantified by the Z-score obtained by threading random sequences onto a DNA structure. A superior energy function should give a more negative Z-score. To eliminate the effect of the interactions with proteins on the DNA structures (i.e. direct readout), we only used free B-DNA structures for the evaluation. We obtained a list of crystal structures of free B-DNA structures from the Nucleic Acid Database (NDB) (22), and downloaded the coordinates of the first biological unit for each entry from the Protein Data Bank (PDB). We excluded structures with less than five base pairs and those containing nonstandard bases or non Watson-Crick base pairs from the dataset, and finally obtained 103 structures. After the base pairs had been removed at the 5′ and 3′ ends, 718 tetramers were obtained from these structures, allowing for overlaps. The step parameters of their central base-pair steps were calculated by using the X3DNA program (21).

Some base-pair steps had parameter values that were rarely observed during the MD simulations. Such abnormal structures can be caused by interactions with metal ions or with DNA in other biological units. In addition, structures with very small probabilities of appearance might not be sampled during the limited simulation time. Since these structures cause large errors in both methods, we excluded tetramers having such step parameters from the dataset as follows. We first divided each 9-ns trajectory into three 3-ns blocks. Then, we calculated $n_i(s,\mathbf{\Theta})$ for each block, where subscript i stood for the index of the block [i.e. $i = 1$, 2 and 3 and $n(s,\boldsymbol{\Theta}) = \sum_{i=1}^{3} n_i(s,\Theta)$]. We let $\mathbf{\Theta}'$ be the parameters of the central base-pair step of a given tetramer. If $n_{\mathrm{i}}(s, \mathbf{\Theta}')$ was larger than zero for all the blocks for more than 26 of the 256 possible tetramer sequences, the tetramer was retained in the dataset. Otherwise, it was removed. After this operation, 496 tetramers were finally obtained.

We calculated the Z-scores for all tetramers thus obtained using our energy function [Equation (2)] and that of Fujii *et al.* [Equation (4)]. The Z-score is defined as

$$Z(s_0,\boldsymbol{\Theta}_0) = \frac{E(s_0,\boldsymbol{\Theta}_0) - \bar{E}(\boldsymbol{\Theta}_0)}{\sigma(\boldsymbol{\Theta}_0)}, \tag{5}$$

where s_0 and $\mathbf{\Theta}_0$ correspond to the sequence of a given tetramer and the step parameters of its central base-pair step. The average and the standard deviation are calculated as

$$\bar{E}(\boldsymbol{\Theta}_0) = \frac{1}{256}\sum_{i=0}^{255} E(s_i,\boldsymbol{\Theta}_0), \tag{6}$$

$$\sigma(\boldsymbol{\Theta}_0) = \left(\frac{1}{256}\sum_{i=0}^{255}\left[E(s_i,\boldsymbol{\Theta}_0) - \bar{E}(\boldsymbol{\Theta}_0)\right]^2\right)^{1/2}, \tag{7}$$

where s_i with i being greater than zero stands for the non-native sequences in the 256 possible tetramer sequences.

RESULTS AND DISCUSSION

Comparison of performance

Figure 1 shows a scatter plot of Z-scores obtained by threading random sequences onto 496 tetramer structures and calculating their conformational energies with E_{G} and E_{HA}. Since 297 of the 496 tetramers (59.9%) are located in the upper triangle of this plot, E_{G} yielded more negative Z-scores for more structures than E_{HA} did. Under the null hypothesis, where the two energy functions have equal capabilities to distinguish the native sequence from non-natives, the probability of producing such a biased distribution by chance (p-value) was 6.2×10^{-6}. This suggests that E_{G} is significantly superior to E_{HA} in this capability.

We next tried to find why E_{G} yielded better results than E_{HA}. It should be noted here that the state used as a reference differed for the two methods. As previously described, E_{G} used the equilibrium state as a reference. The equilibrium state contained various structures and the probability for the existence of these structures followed a canonical distribution. In contrast, E_{HA} used a single average structure as the reference [Equation (4)]. Assume that step parameters $\mathbf{\Theta}$ follow normal distributions, then PDF is written as

$$P(s,\boldsymbol{\Theta}) = \frac{1}{(2\pi)^3|\mathbf{M}_s|^{1/2}}\exp\left(-\frac{1}{2}\Delta\boldsymbol{\Theta}^{\mathrm{T}}\mathbf{M}_s^{-1}\Delta\boldsymbol{\Theta}\right). \tag{8}$$

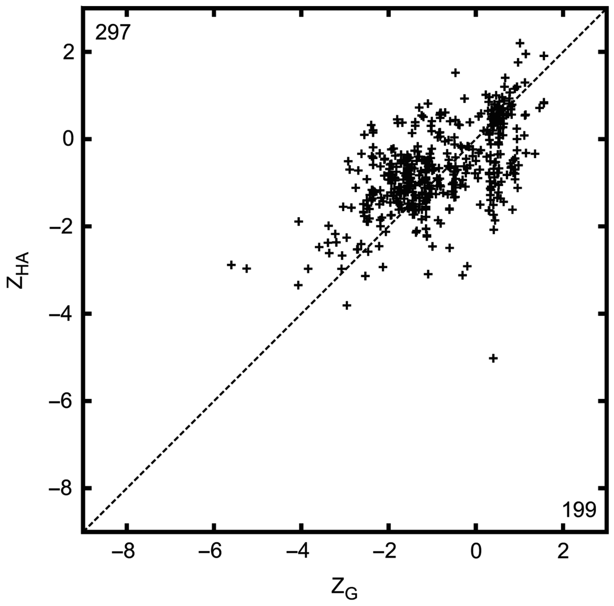

Figure 1. Scatter plot of *Z*-scores calculated for 496 tetramers in test dataset by using E_G (Z_G, *x*-axis) and E_{HA} (Z_{HA}, *y*-axis). The numbers of points within the upper and lower triangles of the plot are shown within the respective areas.

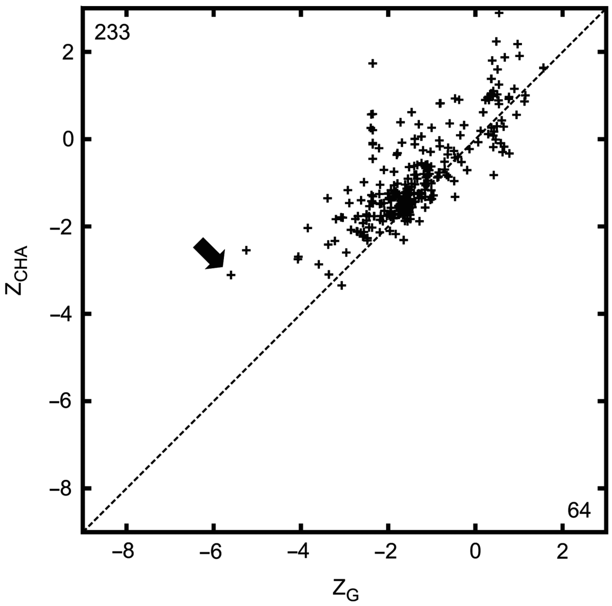

Figure 2. Scatter plot of *Z*-scores calculated for tetramers that gave better results with E_G than with E_{HA}, by using E_G (Z_G, *x*-axis) and E_{CHA} (Z_{CHA}, *y*-axis). The numbers of points within the upper and lower triangles of the plot are shown within the respective areas. The point of tetramer GTTA from 1ZF0 is indicated by the arrow.

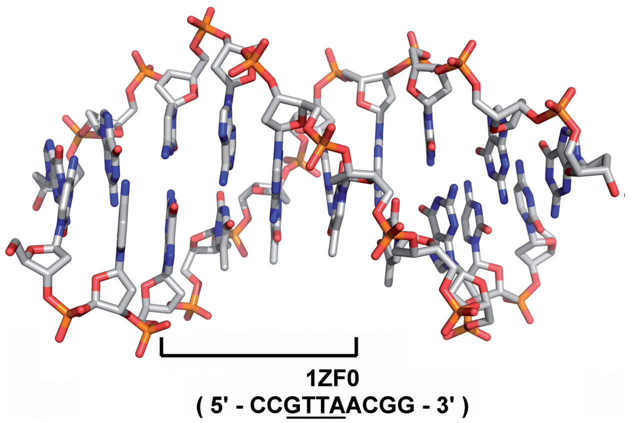

Figure 3. Crystal structure of 1ZF0. Carbon, nitrogen, oxygen and phosphorus atoms are colored gray, blue, red and orange, respectively. The nucleotide sequence of one chain is shown below. Tetramer GTTA is indicated by the bracket in the structural image and is underlined in the nucleotide sequence.

The energy function corresponding to this PDF is

$$E_{\mathrm{CHA}}(s,\boldsymbol{\Theta}) = kT\ln\left[(2\pi)^3|\mathbf{M}_s|^{1/2}\right] + \frac{kT}{2}\Delta\boldsymbol{\Theta}^{\mathrm{T}}\mathbf{M}_s^{-1}\Delta\boldsymbol{\Theta}. \qquad 9$$

Note that the reference for this energy function is the equilibrium state. Comparing Equations (4) and (9), we can see that the first term in Equation (9) represents the free-energy difference between the average structure and the equilibrium state that has been ignored in Equation (4). We refer to this energy function as the corrected harmonic-approximation (CHA) energy function, E_{CHA}. Since the first term varied between 6.3 and 9.4 in the sequences, it is possible that the difference in the reference states caused the difference in capability. Of course, there is another possibility that the use of the harmonic approximation is the primary factor for the lower capability of E_{HA}.

To distinguish these two possibilities, we calculated the *Z*-scores with E_{CHA} for the 297 tetramers, for which our method derived better results. We found that E_G still provided better results for 233 of the 297 tetramers (Figure 2). The *p*-value for this distribution was 5.0×10^{-24}, indicating that E_G was still superior to E_{CHA}. Consequently, using raw PDF without harmonic approximation was a major factor in the improvements we achieved with the present method. Since E_{CHA} yielded slightly better scores than E_{HA} did for 320 of 496 tetramers, we will compare E_G with E_{CHA} after this.

Comparison of energy profiles

The previous section discussed the importance of using raw PDF for the conformational energy function. Here, we selected a tetramer from the test dataset and compared the energy profiles between the two methods to further clarify why the present method improved the *Z*-scores. We chose the structure of tetramer GTTA extracted from a structure whose PDB ID was 1ZF0 (Figure 3), because this tetramer significantly improved the *Z*-scores with E_G by more than 2 (Figure 2). Figure 4 shows the energy surfaces of native (GTTA) and non-native (GTCG) tetramer sequences for each energy function sectioned at the position of the tetramers' step parameters along the principal axes of the variance–covariance matrix calculated from the simulation of DNA in which the native sequence was embedded. With E_G, the

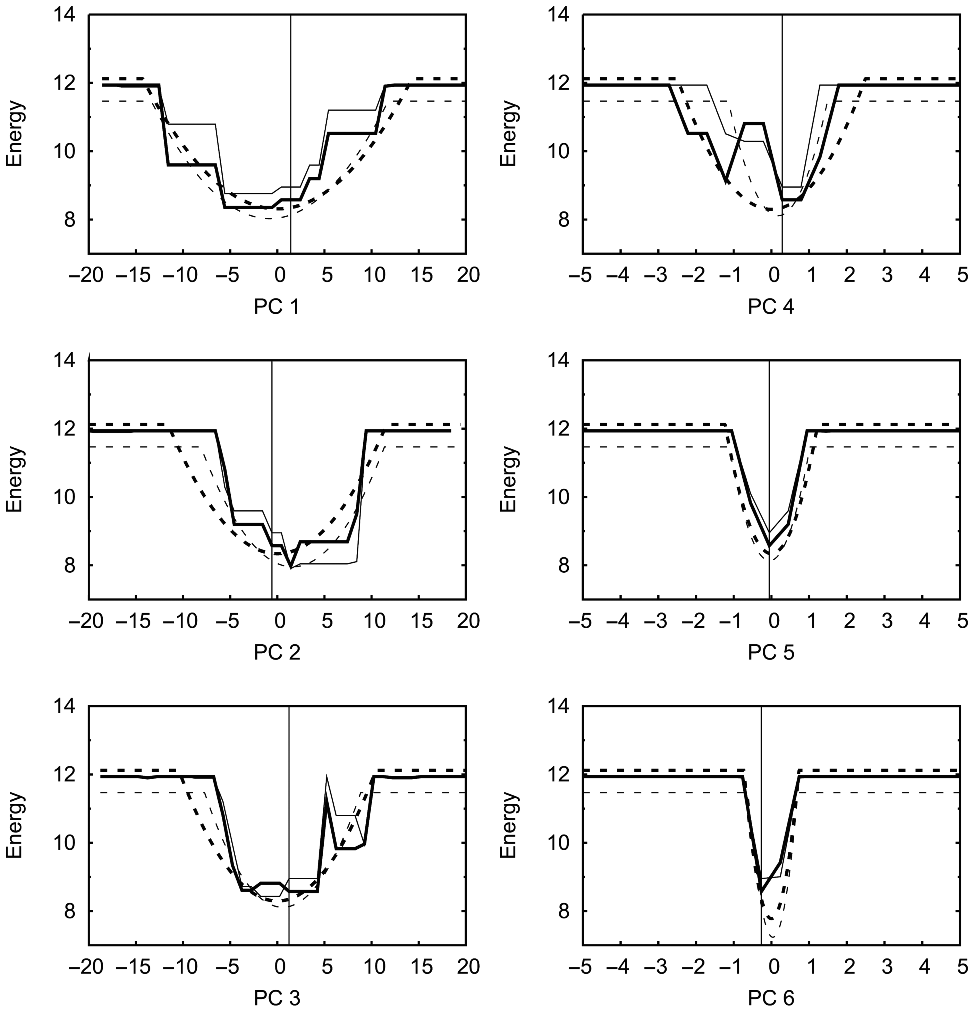

Figure 4. Profiles of E_G for native (GTTA, thick solid line) and non-native (GTCG, thin solid line) sequences and of E_{CHA} for GTTA (thick dashed line) and GTCG (thin dashed line). The profiles are produced by sectioning energy surfaces at the position of the step parameters of tetramer GTTA from 1ZF0 along the principal axes of the variance–covariance matrix calculated from the simulation of DNA in which the same sequence (GTTA) was embedded. PC *n* ($n = 1, 2, \ldots, 6$) stands for the profile along the *n*-th principal axis. The vertical line indicates the position of the step parameters of the tetramer.

conformational energy of the native sequence was lower than that of a non-native sequence, while E_{CHA} yielded an erroneous result where the latter was lower than the former. As can be seen from the profile of E_G along the fourth principal axis (Figure 4), the native sequence has at least two stable conformations. In such cases, E_{CHA} is calculated by approximating the PDF with a broad Gaussian function, even though it substantially deviates from the actual profile of the PDF. This example clearly demonstrates that the harmonic approximation used in E_{CHA} causes significant error in the conformational energy profile, where the tetramer adopts multiple stable conformations.

Since the fourth principal axis of the above example was almost on the plane of shift and slide, we calculated free-energy maps as a function of shift and slide by integrating the six-dimensional PDF with respect to other variables. Of 136 unique tetramer sequences, seven sequences (GTTA, ATAA, ATAG, CGGG, TGGT, TTAA and TTAG) had two obvious peaks in their free-energy maps. Figure 5 shows the free-energy maps for GTTA, ATAA and ATAG selected from the seven sequences. The scatter plots of the shift-slide parameter values of tetramers having these sequences in their crystal structures have been overlaid on the free-energy maps of the respective sequences for comparison. The structures of the central dimers of crystal structures that have shift-slide values close to the peaks of the free-energy maps are also shown. Although only a few experimental structures were available for some tetramer sequences, we found that these tetramers tended to have broad distributions in their shift-slide parameters in the crystal structures, which is consistent with the two-peak distributions in the ensembles generated by the MD simulations.

Effect of flanking base pairs on central base-pair step

Fujii *et al.* (15) have shown that the flexibility of the central base-pair step of a tetramer is affected by the first and the fourth base pairs in the tetramer from their

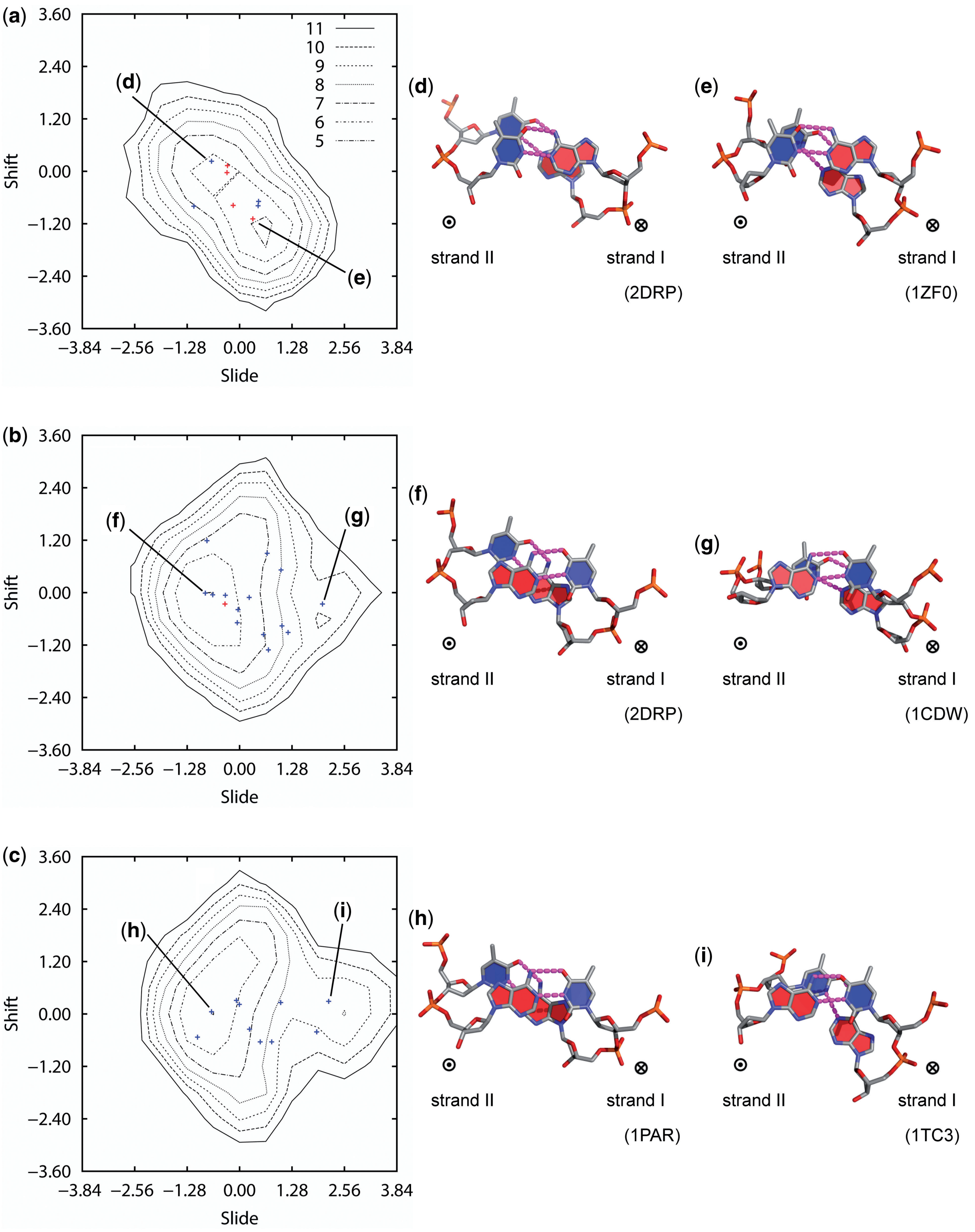

Figure 5. Free-energy maps calculated for tetramer sequences GTTA (**a**), ATAA (**b**) and ATAG (**c**) as a function of shift and slide parameters. Same contour levels are used in the three maps. Red and blue plus marks indicate that shift-slide parameter values of the tetramers having the respective sequences in free- and complex-DNA crystal structures, respectively. Structures of the central dimers of the tetramers having shift-slide parameter values indicated with (**d–i**) are shown with stick model. PDB IDs of the structures are shown in parentheses. In the structural images, Watson–Crick base pair hydrogen bonds are indicated with dashed lines. Carbon, nitrogen, oxygen and phosphorus atoms are colored gray, blue, red and orange, respectively. Circled-times and circled-dot marks indicate the direction from 5′- to 3′-end of DNA strands. Circled-dot points out of the plane of the paper, while circled-times points into the plane of the paper.

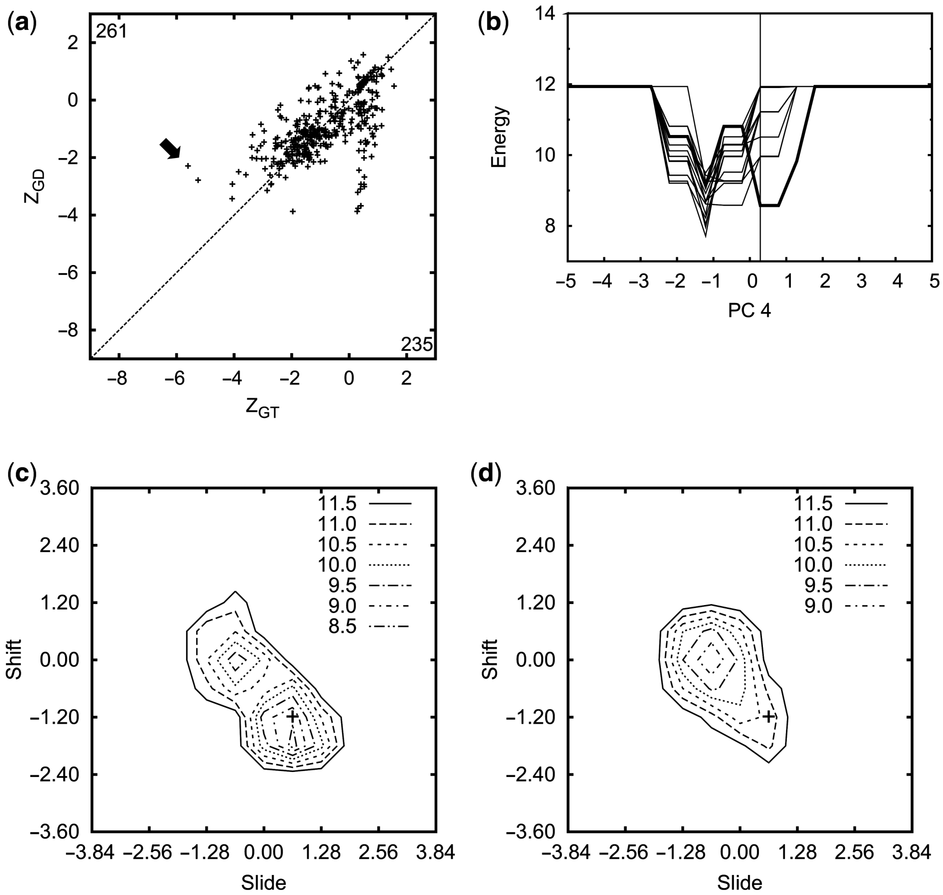

Figure 6. (**a**) Scatter plot of *Z*-scores calculated for each tetramer by using E_{GT} (Z_{GT}, *x*-axis) and E_{GD} (Z_{GD}, *y*-axis). Note that E_{GT} and Z_{GT} correspond to aliases of E_G and Z_G. The numbers of points within the upper and lower triangles of the plot are shown within the respective areas. The point of tetramer GTTA from 1ZF0 is indicated by the arrow. (**b**) The profiles of E_{GT} for the native sequence (thick line) and the ones that have different bases at the first and fourth positions (thin lines). The profiles are produced by sectioning energy surfaces at the position of the step parameters of tetramer GTTA from 1ZF0 along the fourth principal axis of the variance–covariance matrix calculated from the simulation of DNA in which the same sequence (GTTA) was embedded. The vertical line indicates the position of the step parameters of the tetramer. The contour plots of E_{GT} (**c**) and E_{GD} (**d**) as a function of shift and slide parameters, fixing the values of other dimensions at those of the step parameters of the tetramer. The plus mark indicates the position of the step parameters of the tetramer.

analysis of the conformational entropies, which they calculated as the determinants of variance-covariance matrices. In the present study, we examined what effect these base pairs had in terms of the energy profile. We calculated the PDF, which only depends on the central dimer sequences (n_2n_3), by averaging the original PDFs [Equation (1)] over the first (n_1) and the fourth (n_4) nucleotides and converted them into an energy function by using Equation (2). After this, this function will be referred to as E_{GD}, whereas the original energy function, E_G, has been designated as E_{GT} to distinguish it from the former. Figure 6a shows a scatter plot of *Z*-scores obtained by using E_{GD} and E_{GT}. We found that E_{GT} provided better results for 261 of the 496 tetramers (52.6%). This suggests that the flanking base pairs indeed have an effect on the conformation of the central base-pair step.

Since tetramer GTTA from 1ZF0 again demonstrated a large difference [Figure 6a], we compared the profiles of the two energy functions. Figure 6b shows the energy surfaces of E_{GT} for various tetramer sequences sectioned at the position of the tetramer's central base-pair step parameters along the fourth principal axis used in Figure 4. The energy profile of the native sequence (GTTA) was minimum near the position of the tetramer's step parameters, whereas the profiles of sequences with different bases at the first and the fourth positions had minima at different positions. Comparing the energy profiles between E_{GT} and E_{GD} [Figure 6c and d], we can see that the minimum of E_{GT} near the position of the tetramer's step parameters disappeared in E_{GD} due to averaging and the conformational energy of the tetramer obtained with E_{GD} became higher than that obtained with E_{GT}. The conformational energy of the central base-pair step of a tetramer is thus dependent on the first and the fourth base pairs. The averaging caused a loss of information on sequence dependence and degraded the ability to distinguish a native sequence from random ones.

Factors limiting accuracy of present method

Since numerous structural data are required to calculate PDFs, it was necessary to use molecular simulations in the present method. Therefore, its accuracy was limited by the accuracy of the simulations. There are generally two major factors that limit the accuracy of MD simulations. The first is the error in the potential energy function, which causes biases in the PDFs. It has already been pointed out that the distribution of the slide parameter in the ensembles from MD simulations was biased toward more negative values than those calculated using crystal structures (15). In addition, it has been revealed that the backbone structure of B-DNA is often severely distorted due to the imbalance of force-field parameters, when simulation is executed for prolonged periods (23). It is possible for these biases to cause larger conformational energy in the native sequence than those in the non-native sequence and to cause positive Z-scores. However, we obtained negative Z-scores for most of the tetramers and could not further improve the positive Z-scores by using PDFs calculated without the slide parameter. In addition, no distortion in the backbone structure was observed during the 10-ns simulations. Therefore, we believe that the error in the potential energy function did not cause significant errors in our conformational energy function.

Another factor is the statistical error caused by the sampling problem. Although the native DNA structure should be in the global free-energy minimum, its local structure, such as the tetramer structure extracted from the whole DNA structure, is not necessary in the energy minimum. In a canonical distribution, the probability of appearance drastically decreases as the energy increases. Therefore, high-energy conformations that often exist in the crystal structures are rarely sampled during MD simulations of limited duration and the accuracy of the energy function is severely degraded in the high-energy region. The positive Z-scores and the worse Z-scores with the present method than those with the previous were mainly caused by this problem. The tetramer CGCG from 287D is a typical example that demonstrates how the sampling problem affects the accuracy of the energy function. This tetramer yielded Z-scores of 0.40 and –5.02 with E_{G} and E_{HA}, and is located in the lowest-rightmost region in Figure 1. The energy profile of the native sequence around the tetramer's step parameters was very rugged with very high-energy values. Similar energy profiles were obtained for non-native sequences. This indicates that structures having the tetramer's step parameters were not sufficiently sampled during the MD simulations. As a result, the PDFs were not sufficiently converged in this region and the energy values had large errors. In contrast, a very small Z-score for this tetramer was obtained by using E_{HA}. However, this does not mean that the previous method is superior to the present approach. The previous methodology is based on the assumption that the distribution of the step parameters can be approximated with a Gaussian distribution. The actual distribution, however, significantly deviated from the Gaussian distribution as we determined from a Kolmogorov-Smirnov test. Therefore, this indicates that it is necessary to obtain more samples especially for high-energy structures. Generalized ensemble methods, such as the multicanonical MD method, should be useful to efficiently sample structures from wider conformational space (24) and to improve the accuracy of our method. The force-field parameters were recently revised to solve the problem of distortion in the DNA backbone structure (25). The use of such force-field parameters is also important to minimize the bias in PDF caused by error in the potential energy function.

CONCLUSIONS

We developed a new generalized energy function to estimate how dependent the deformability of DNA was on sequence. A function was defined for all possible tetramer sequences as a logarithm of the PDF of the central base-pair step parameters of a tetramer in an ensemble derived from an MD simulation on a B-DNA dodecamer that had the tetramer sequence embedded at its center. The accuracy of the energy function was evaluated using Z-scores that measured the ability to distinguish the native sequence from random ones, and it was compared with that of the previous method that approximated the energy function with a harmonic potential. Sequence-structure threading was performed on 496 tetramer structures extracted from the experimental free B-DNA structures to calculate the Z-scores. Our method yielded better Z-scores for 297 tetramers than the previous approach. By comparing the energy profiles of the two methods, we found that harmonic approximation caused serious error in conformational energy where the tetramer adopted multiple stable conformations. The efficiency of the energy function was also compared with that of the energy function obtained by averaging the PDFs over the first and the fourth nucleotides. The original energy function provided better results for more than half the test dataset.

With these results, we concluded that the energy function should simply be defined as the logarithm of the PDF and should take into account the long-range effect on the deformability of a base pair step caused by its flanking base pairs.

It is necessary to obtain more conformational samples especially those that have high energies to achieve higher levels of accuracy. Generalized ensemble methods should be useful for this purpose. We are now planning to apply the present method to systems of protein–DNA complexes to evaluate the contribution of indirect readout to binding free-energy changes. Since deviations in the protein-bound DNA structure from the canonical B-DNA structure are significantly larger than those of free DNA structures, it will be quite important to obtain PDFs that cover wide ranges of step parameters with accurate values in such applications.

FUNDING

Scientific Research on Priority Areas 'Systems Genomics' (20016006 to K.S. and 20016022 to A.S.) and 'DECODE' (20052021 to A.S.) made available by the Ministry of Education, Culture, Sports, Science and Technology of Japan, and in part by the fund of the Institute for

Bioinformatics Research and Development (to K.S.). Funding for open access charge: The University of Tokyo.

Conflict of interest statement. None declared.

REFERENCES

1. Dickerson,R.E. (1983) The DNA helix and how it is read. *Sci. Am.*, **249**, 94–111.
2. Sarai,A. and Kono,H. (2005) Protein-DNA recognition patterns and predictions. *Annu. Rev. Biophys. Biomol. Struct.*, **34**, 379–398.
3. Hegde,R.S. (2002) The Papillomavirus E2 proteins: structure, function, and biology. *Annu. Rev. Biophys. Biomol. Struct.*, **31**, 343–360.
4. Horton,N.C., Dorner,L.F. and Perona,J.J. (2002) Sequence selectivity and degeneracy of a restriction endonuclease mediated by DNA intercalation. *Nature Struct. Biol.*, **9**, 42–47.
5. Koudelka,G.B. (1998) Recognition of DNA structure by 434 repressor. *Nucleic Acids Res.*, **26**, 669–675.
6. Lamoureux,J.S., Maynes,J.T. and Mark Glover,J.N. (2004) Recognition of 5′-YpG-3′ sequences by coupled stacking/hydrogen bonding interactions with amino acid residues. *J. Mol. Biol.*, **335**, 399–408.
7. Lawson,C.L., Swigon,D., Murakami,K.S., Darst,S.A., Berman,H.M. and Ebright,R.H. (2004) Catabolite activator protein: DNA binding and transcription activation. *Curr. Opin. Struct. Biol.*, **14**, 10–20.
8. Olson,W.K., Gorin,A.A., Lu,X.J., Hock,L.M. and Zhurkin,V.B. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl Acad. Sci. USA*, **95**, 11163–11168.
9. Steffen,N.R., Murphy,S.D., Tolleri,L., Hatfield,G.W. and Lathrop,R.H. (2002) DNA sequence and structure: direct and indirect recognition in protein-DNA binding. *Bioinformatics*, **18(Suppl. 1)**, S22–S30.
10. Lankaš,F., Šponer,J., Langowski,J. and Cheatham,T.E. III. (2003) DNA basepair step deformability inferred from molecular dynamics simulations. *Biophys. J.*, **85**, 2872–2883.
11. Gromiha,M.M., Siebers,J.G., Selvaraj,S., Kono,H. and Sarai,A. (2004) Intermolecular and intramolecular readout mechanisms in protein-DNA recognition. *J. Mol. Biol.*, **337**, 285–294.
12. Gromiha,M.M., Slebcrs,J.G., Selvaraj,S., Kono,H. and Sarai,A. (2005) Role of inter and intramolecular interactions in protein-DNA recognition. *Gene*, **364**, 108–113.
13. Araúzo-Bravo,M.J., Fujii,S., Kono,H., Ahmad,S. and Sarai,A. (2005) Sequence-dependent conformational energy of DNA derived from molecular dynamics simulations: toward understanding the indirect readout mechanism in protein-DNA recognition. *J. Am. Chem. Soc.*, **127**, 16074–16089.
14. Becker,N.B., Wolff,L. and Everaers,R. (2006) Indirect readout: detection of optimized subsequences and calculation of relative binding affinities using different DNA elastic potentials. *Nucleic Acids Res.*, **34**, 5638–5649.
15. Fujii,S., Kono,H., Takenaka,S., Go,N. and Sarai,A. (2007) Sequence-dependent DNA deformability studied using molecular dynamics simulations. *Nucleic Acids Res.*, **35**, 6063–6074.
16. Paillard,G. and Lavery,R. (2004) Analyzing protein-DNA recognition mechanisms. *Structure*, **12**, 113–122.
17. Olson,W.K., Bansal,M., Burley,S.K., Dickerson,R.E., Gerstein,M., Harvey,S.C., Heinemann,U., Lu,X.J., Neidle,S., Shakked,Z. *et al.* (2001) A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.*, **313**, 229–237.
18. Gardiner,E.J., Hunter,C.A., Packer,M.J., Palmer,D.S. and Willett,P. (2003) Sequence-dependent DNA structure: a database of octamer structural parameters. *J. Mol. Biol.*, **332**, 1025–1035.
19. Packer,M.J., Dauncey,M.P. and Hunter,C.A. (2000) Sequence-dependent DNA structure: dinucleotide conformational maps. *J. Mol. Biol.*, **295**, 71–83.
20. Packer,M.J., Dauncey,M.P. and Hunter,C.A. (2000) Sequence-dependent DNA structure: tetranucleotide conformational maps. *J. Mol. Biol.*, **295**, 85–103.
21. Lu,X.J. and Olson,W.K. (2003) 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.*, **31**, 5108–5121.
22. Berman,H.M., Olson,W.K., Beveridge,D.L., Westbrook,J., Gelbin,A., Demeny,T., Hsieh,S.H., Srinivasan,A.R. and Schneider,B. (1992) The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, **63**, 751–759.
23. Varnai,P. and Zakrzewska,K. (2004) DNA and its counterions: a molecular dynamics study. *Nucleic Acids Res.*, **32**, 4269–4280.
24. Mitsutake,A., Sugita,Y. and Okamoto,Y. (2001) Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers (Pept. Sci.)*, **60**, 96–123.
25. Perez,A., Marchan,I., Svozil,D., Sponer,J., Cheatham,T.E. III, Laughton,C.A. and Orozco,M. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophys. J.*, **92**, 3817–3829.