Published online 22 May 2012

Nucleic Acids Research, 2012, Vol. 40, No. 17 ***e129***
doi:10.1093/nar/gks462

A new method for evaluating the specificity of indirect readout in protein–DNA recognition

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

[1]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, [2]Agricultural Bioinformatics Research Unit, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657, [3]Molecular Modeling and Simulation Group, Quantum Beam Science Directorate, Japan Atomic Energy Agency, 8-1-7 Umemidai, Kizugawa, Kyoto 619-0215, [4]Department of Biotechnology, Graduate School of Agricultural and Life Sciences, The University of Tokyo, 1-1-1 Yayoi, Bunkyo-ku, Tokyo 113-8657 and [5]Department of Bioscience and Bioinformatics, Graduate School of Computer Science and Systems Engineering, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

Received December 6, 2011; Revised March 29, 2012; Accepted April 30, 2012

ABSTRACT

Proteins recognize a specific DNA sequence not only through direct contact (direct readout) with base pairs but also through sequence-dependent conformation and/or flexibility of DNA (indirect readout). However, it is difficult to assess the contribution of indirect readout to the sequence specificity. What is needed is a straightforward method for quantifying its contributions to specificity. Using Bayesian statistics, we derived the probability of a particular sequence for a given DNA structure from the trajectories of molecular dynamics (MD) simulations of DNAs containing all possible tetramer sequences. Then, we quantified the specificity of indirect readout based on the information entropy associated with the probability. We tested this method with known structures of protein–DNA complexes. This method enabled us to correctly predict those regions where experiments suggested the involvement of indirect readout. The results also indicated new regions where the indirect readout mechanism makes major contributions to the recognition. The present method can be used to estimate the contribution of indirect readout without approximations to the distributions in the conformational ensembles of DNA, and would serve as a powerful tool to study the mechanism of protein–DNA recognition.

INTRODUCTION

DNA-binding proteins play important roles in transcription, DNA replication, translation and ligation. Although some of these proteins bind to DNAs in a non-specific manner, most of them bind to specific regions of their target DNAs. Proteins recognize specific DNA sequences either through direct interactions between amino acids and base pairs, i.e. 'direct readout', or through sequence-dependent conformation and/or flexibility of DNA, i.e. 'indirect readout' (1). Hereafter, we will use the term 'indirect readout' to indicate only the readout through sequence-dependent conformation and/or flexibility of DNA, and the readout through water-mediated contacts is not included.

Experiments have indicated the existence of the indirect readout. In some protein–DNA complexes, changes in the binding affinity through the replacement of the DNA bases could not be fully explained by the changes in the direct protein–DNA interactions (1–5). However, it is rather difficult to measure the contribution of indirect readout to the binding affinity separately from other factors. Therefore, there have been few experimental results that could quantify the contribution of indirect readout.

In our previous studies (6–9), we derived the probability distribution functions (PDFs) of the base-pair step parameters (shift, slide, rise, tilt, roll and twist) from conformational ensembles obtained either from known structural data (6) or from molecular dynamics (MD) simulations (7–9). Since the PDFs were derived for all possible dimer

*To whom correspondence should be addressed. Tel: +81 3 3599 8676; Fax: +81 3 3599 8081; Email: s.yamasaki@aist.go.jp

or tetramer sequences, they represent sequence-dependent conformational propensities of the DNAs. In the case of knowledge-base approach using the known structural data (6), we could only consider the dimer sequences and had to use a harmonic approximation to calculate the potential of mean force for the sequence-dependent conformations from the PDF due to the limitation of available structural data. On the other hand, we also used the trajectories of MD simulations to calculate the PDF (7,8). In this case, we can consider tetramer sequences to examine longer-range effects. We first used the harmonic approximation for the derivation of the potential of mean force (7,8). However, we found that some PDFs exhibit non-Gaussian behavior. Therefore, we calculated the conformational energy of the central base-pair step within each tetramer sequence by taking the logarithm of the PDF value of its sequence at the step parameters (9). To test how well the PDFs can describe the experimentally observed sequence dependence of the step parameters, the conformational energies of various known B-DNA structures were compared with those calculated by threading non-native sequences into the structures. The results revealed that our method performed better at discriminating the native sequence from the others than the previous methods where the distribution of the step parameters was approximated with a Gaussian function (7–9). Although the PDFs can provide an intuitive measure of the conformational preference of DNA in terms of the energy profile, there are still some difficulties with this method. The energy goes to infinity when the probability goes to zero. To avoid such a divergence, the PDF was set to a small positive constant value when it was less than the value (9). To compare the sequence specificities among different conformations, we used the *Z*-scores of the conformational energies of the native sequences calculated from the distributions of the energies of all possible sequences (9). This introduces an unnecessary assumption that the distribution of the energy follows a normal distribution. It is therefore desirable to devise a more straightforward method of evaluating the sequence specificity of the DNA conformation and for identifying the potential regions that are recognized by proteins through the indirect readout mechanism.

Here, we propose a new method for this purpose. Instead of calculating the potential of mean force from the PDF, we convert the PDF into the probability of a sequence, given a step conformation, by using Bayes' theorem. When the probabilities of all sequences are equal for a given step conformation, this conformation is not sequence specific and the indirect readout does not make any contribution to the recognition. On the other hand, when the probability of a sequence is equal to 1 and those of the other sequences are 0 (i.e. the step conformation is only observed for a specific sequence), the conformation is completely sequence specific and the sequence can be recognized through indirect readout. Such a bias in the probabilities of the sequence can be quantified with the information entropy (10). The information entropy is superior to *Z*-score in that the information entropy does not require the assumption of normal distribution as *Z*-score does. Thus, the probability-/entropy-based approach is more straightforward than the energy-based approach.

Using the probability and information entropy, we predicted the potential regions in DNAs recognized through the indirect readout mechanism and assessed how well the native DNA sequence fits to the given step conformation. We compared the predicted results with experimental data. In addition, the information entropy has the advantage in that it can be decomposed into contributions from parts of the sequence. Taking this advantage, we evaluated the effect of the step conformation on the variety of the bases neighboring to the central dimer sequence by comparing the results from the dimer- and tetramer-based analyses.

MATERIALS AND METHODS

Probability of a sequence given step parameters

Using Bayes' theorem, we can derive the probability of finding sequence s, given step parameters $\mathbf{\Theta}$ (shift, slide, rise, tilt, roll and twist), $P(s|\mathbf{\Theta})$ (hereafter referred to as PST), from the PDF, the probability of the step parameter $\mathbf{\Theta}$, given the sequence s, $P(\mathbf{\Theta}|s)$, as,

$$P(s|\mathbf{\Theta}) = \frac{P(\mathbf{\Theta}|s)P(s)}{P(\mathbf{\Theta})} = \frac{P(\mathbf{\Theta}|s)P(s)}{\sum_s P(\mathbf{\Theta}|s)P(s)}, \quad (1)$$

where $P(s)$ is the prior probability of the sequence s. The native sequences are expected to have large PST values when they are applied to known protein–DNA complex or free DNA structures. The distribution of the PST also provides important information about the specificity, because it originates from the unevenness of the PST. Therefore, we shall introduce a measure of unevenness to quantify the specificity.

Information entropy for given step parameters

Using the PST, we can calculate the probability of each possible sequence for a given step of a DNA structure. When a sequence $s = s'$ gives $P(s'|\mathbf{\Theta}) = 1$ and the others give $P(s|\mathbf{\Theta}) = 0$ for a given step parameter $\mathbf{\Theta}$, this step conformation only accepts sequence s'. This indicates that the step conformation is highly sequence specific and the sequence is unambiguously 'readout' solely from the conformation. On the other hand, when the sequence and the step conformation are independent, i.e. $P(s|\mathbf{\Theta}) = P(s)$, or equivalently, $P(\mathbf{\Theta}|s) = P(\mathbf{\Theta})$, indirect readout at this site does not work at all. Therefore, the sequence specificity, i.e. unevenness of the PST, can be adequately described by the information entropy (10) relative to the independent case, which has been often used for measuring sequence conservation (11). Given a step parameter $\mathbf{\Theta} = \mathbf{\theta}_j$, the 'specificity score' I for a set of all possible sequences is defined by

$$I = \sum_s P\left(s\middle|\mathbf{\Theta} = \theta_j\right) \log_2 \frac{P\left(s\middle|\mathbf{\Theta} = \theta_j\right)}{P(s)}. \quad (2)$$

The score I is 0 when the sequence is independent of the step conformation and it increases as the specificity increases.

As mentioned in 'Results and Discussion' section, the conformational ensembles from the MD simulations of each tetramer sequence were not enough for reliable estimation of PST. Thus, we have reduced the original tetramer sequence space into dimer AGTC sequence space and dimer and tetramer RY [purines (R) and pyrimidines (Y)] sequence spaces, in order to increase the number of conformations in each of ensembles per sequence. A tetramer sequence can be decomposed into two parts (e.g. the central dimer sequence and the flanking sequence) and the corresponding information entropies and specificity scores can be calculated. In general, when the sequence is decomposed into two parts ($s = s_1 s_2$), the specificity score of the whole sequence, I_{12}, is expressed as:

$$\begin{aligned} I_{12} &= \sum_{s_1}\sum_{s_2} P(s_1|\mathbf{\Theta}=\theta_j) P(s_2|s_1,\mathbf{\Theta}=\theta_j) \log_2 \frac{P(s_1|\mathbf{\Theta}=\theta_j) P(s_2|s_1,\mathbf{\Theta}=\theta_j)}{P(s_1)P(s_2|s_1)} \\ &= \sum_{s_1} P(s_1|\mathbf{\Theta}=\theta_j) \log_2 \frac{P(s_1|\mathbf{\Theta}=\theta_j)}{P(s_1)} \\ &\quad + \sum_{s_1} P(s_1|\mathbf{\Theta}=\theta_j) \sum_{s_2} P(s_2|s_1,\mathbf{\Theta}=\theta_j) \log_2 \frac{P(s_2|s_1,\mathbf{\Theta}=\theta_j)}{P(s_2|s_1)} \\ &= I_1 + \Delta I, \end{aligned} \tag{3}$$

where I_1 is the marginal relative entropy of the sub-sequence, s_1 and ΔI is the conditional relative entropy of the residual sequence, s_2, given s_1. ΔI is a measure of the restriction on the residual sequence imposed by the step conformation and the sub-sequence. Note that even a 1-nt 'sequence' can be decomposed into two parts: one representing the size of the base (R or Y) and another representing the number of hydrogen bonds in Watson–Crick-type base pairing (two or three). For example, tetramer sequence AGTC is expressed as R2R3Y2Y3, which can be decomposed into s_1 = RRYY and s_2 = 2323. This decomposition allows us to calculate the entropy in the RY sequence space; the PSTs in the RY space are obtained by summing the original PSTs over s_2.

Test dataset

First, we downloaded the coordinates of 52 protein–DNA complex structures listed and discussed in Gromiha's article (6) from Protein Data Bank (12). Then, we excluded the structures containing non-canonical bases (1DP7), nicks or bulges in the nucleotide strands (1BER, 1SVC, 1GDT, 1IHF and 6CRO). Consequently, we obtained 46 protein–DNA complex structures. After overhangs of unpaired nucleotides were removed, 619 tetramers were extracted from the structures, allowing overlap. The step parameters were calculated for the central base-pair steps of the tetramers by using the 3DNA program (13). The denominators of Equation (1), $P(\mathbf{\Theta})$, were very small ($<4 \times 10^{-6}$) for some tetramers; we excluded these tetramers and obtained a dataset composed of 588 tetramers.

In comparison, the numbers of the direct and water-mediated hydrogen bonds between the protein and DNA bases were counted for each DNA base pair by using the LIGPLOT program (14). For the thymine bases, the hydrophobic contacts at the methyl group were also counted when the protein carbon atoms were within 3.9 Å from the carbon atom of the methyl group. The number of direct contacts is the sum of the direct hydrogen bonds and hydrophobic contacts.

Conformational ensemble of DNA

In this work, we used, as an example, the conformational ensembles of DNAs generated by 10-ns MD simulations for B-DNA dodecamers containing the 136 kinds of unique tetramer sequences at its center (5′–CGCG–$n_1n_2n_3n_4$ –CGCG–3′; n_i is A, T, G, or C) [see (6,9) for details]. The ensembles contain 9000 structures derived from each simulation. The 1-ns MD simulation took ~5 h on two Intel X5670 (2.93 GHz) chips.

Inference of PSTs and specificity scores

Performing an MD simulation corresponds to collecting samples from the population in the conformational space. Since the PDFs, $P(\mathbf{\Theta}|s)$, in Equation (1) are the probabilities of the step parameter $\mathbf{\Theta}$ in the populations that are not *a priori* known, we inferred them from the sets of the samples using Bayesian statistics. We consider the probability distribution of $x = P(\mathbf{\Theta}|s)$. When the step parameter $\mathbf{\Theta}$ was observed $n(\mathbf{\Theta}|s)$ times during the MD simulation producing $N(s)$ samples, the probability distribution of x can be expressed as,

$$P(x|n(\mathbf{\Theta}|s)) = P(n(\mathbf{\Theta}|s)|x)P(x)/P(n(\mathbf{\Theta}|s)). \tag{4}$$

The likelihood function, $P(n(\mathbf{\Theta}|s)|x)$, was given by the binomial distribution,

$$P(n(\mathbf{\Theta}|s)|x) = \frac{N(s)!}{n(\mathbf{\Theta}|s)![N(s)-n(\mathbf{\Theta}|s)]!} x^{n(\mathbf{\Theta}|s)}(1-x)^{N(s)-n(\mathbf{\Theta}|s)}. \tag{5}$$

Assuming uniform distribution for $P(x)$ and substituting Equation (5) into Equation (4), we obtain,

$$P(x|n(\mathbf{\Theta}|s)) = \frac{x^{n(\mathbf{\Theta}|s)}(1-x)^{N(s)-n(\mathbf{\Theta}|s)}}{B(n(\mathbf{\Theta}|s)+1, N(s)-n(\mathbf{\Theta}|s)+1)}, \tag{6}$$

where $B(\alpha,\beta)$ is a normalization factor, which can be calculated with the beta function. We drew 1000 samples from the posterior distribution for each sequence and calculated the averages and the standard deviations of the PSTs and the specificity scores using the samples.

RESULTS AND DISCUSSION

Probabilities of native sequences and specificity scores from known protein–DNA complex structures

In order to apply the present method, we first considered the 588 tetramers derived from 46 known protein–DNA complex structures, and calculated the PST, information entropy and specificity scores using Equations (1)–(3), based on the PDF derived from the 10-ns MD simulations. In the calculations, we assumed that $P(s)$ is the same for all sequences. We first checked the statistical

errors of the PST and information entropy according to the procedure described in 'Materials and Methods' section. The ensembles of 9000 conformations from each of the 136 simulations were not large enough to estimate the PST and information entropy with statistical significance. Thus, we reduced the sequence space from tetramer, or increased the size of ensembles. We considered dimer sequences within the 46 protein–DNA complex structures, where larger ensembles per sequence can be used to calculate the above quantities. We also considered a reduced space composed of purine (R) and pyrimidine (Y). In this RY space, we can examine the tetramer effect on the specificity. Hereafter, we denote PSTs in dimeric AGTC, dimeric RY and tetrameric RY sequence spaces by $P(\{\mathrm{AGTC}\}_2|\mathbf{\Theta})$, $P(\{\mathrm{RY}\}_2|\mathbf{\Theta})$ and $P(\{\mathrm{RY}\}_4|\mathbf{\Theta})$ or by $\mathrm{PST}(\{\mathrm{AGTC}\}_2)$, $\mathrm{PST}(\{\mathrm{RY}\}_2)$, $\mathrm{PST}(\{\mathrm{RY}\}_4)$. Similarly, the specificity scores calculated in dimeric AGTC, dimeric RY and tetrameric RY sequence spaces are denoted by $I(\{\mathrm{AGTC}\}_2)$, $I(\{\mathrm{RY}\}_2)$, $I(\{\mathrm{RY}\}_4)$. The maximum values of the scores are 4, 2 and 4 bits, respectively.

For the sake of comparison with the contribution of the direct readout to the recognition specificity in a rather qualitative manner, we calculated the numbers of the direct and water-mediated contacts with protein atoms for each base pair from the protein–DNA complex structures. All the data were tabulated in Supplementary Tables S1 and S2.

Figure 1 shows the scatter plots of $\mathrm{PST}_{\mathrm{native}}$ [$P(s|\mathbf{\Theta})$ with the native sequence for s] against I for all the dimers in the AGTC space and for all the tetramers in the RY space from the 46 protein–DNA complex structures. Many points fell into the bottom-left region with small I and $\mathrm{PST}_{\mathrm{native}}$ values; their steps may be outside the recognition regions, or otherwise the indirect readout plays less important role than the direct readout at those positions. A significant number of points are located in the top-right region with large I and $\mathrm{PST}_{\mathrm{native}}$ values and there is a correlation between these values; i.e. their native sequences are adapted to their step conformations. On the other hand, a significant number of points are clustered in the bottom-right region with large I and small $\mathrm{PST}_{\mathrm{native}}$ values; this is apparently contradictory, because the native sequences are unfavorable for these step conformations even though these conformations are highly specific to certain sequences. However, a number of possible reasons for this contradiction can be raised. One possibility is the effect of the interactions with the proteins. In many of these cases, direct or water-mediated contacts with the proteins are observed at the bases of their two central nucleotides, indicating that the sequences are mainly recognized through the direct readout mechanism. The second possibility is the effect of the interaction with the additives of the crystallization solution or the effect of the crystal packing. However, for some points, there are no obvious indications that contacts at the bases alter their step conformation. Therefore, there is another possibility that the replacement with the sequence with a larger PST increases the affinity to the target protein. Of course, it is also possible that the contradiction is caused by the reduction of sequence space or the sampling errors

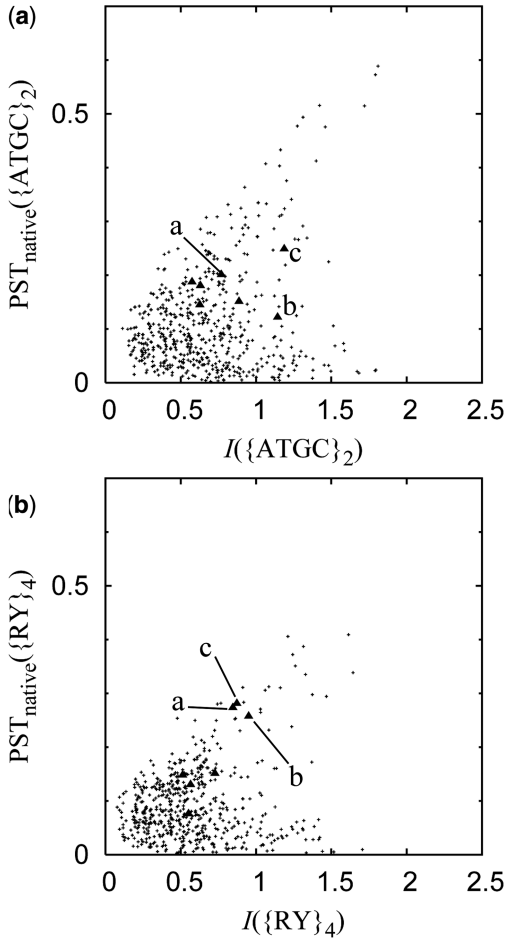

Figure 1. Scatter plot of $\mathrm{PST}_{\mathrm{native}}$ [$P(s|\mathbf{\Theta})$ with the native sequence for s] against I calculated in dimeric ATGC space (**a**) and in tetrameric RY space (**b**). Triangles correspond to the tetramer sequence steps for which experimental data obtained by base mutations suggested the involvement of the indirect readout. The labeled points ('a' to 'c') are the examples discussed in the manuscript.

of the MD simulations. It is therefore necessary to improve the accuracy of the PDFs by extending the simulation time and/or by using more efficient sampling algorithms before making a reliable prediction for the tetramers with small $\mathrm{PST}_{\mathrm{native}}$ values.

Although the contact with the protein may deform the structure of the DNA, the deformed structure is not always ill-suited to the native sequence. In fact, for 416 out of 588 points direct or water-mediated contacts with the proteins were observed at their two central base pairs. This is reasonable because the direct and indirect readouts come from different origins: the former comes from intermolecular interactions with the protein whereas the latter comes from intramolecular interactions within the DNA structure. Therefore, the same nucleotide can contribute to both the direct- and the indirect-readout mechanisms. This is probably the reason why it has been difficult to quantify the contributions of the direct and indirect readouts separately in experiments. With the present method, we can highlight any nucleotide that contributes

to the indirect readout even if it is involved in the direct interactions with the protein.

Comparison with experimental observations

To date, mutation results are available for seven locations in four protein–DNA complexes, which were suggested to be important for the indirect readout. As shown in Figure 1 (labeled 'a' to 'c'), the calculated I and PST_{native} values for these data fell in the middle region in the plot. This suggests that the indirect readout may play some role in the sequence recognition by these proteins. Below, we make some detailed comparisons of our data with the experimental observations for a to c showing higher PST_{native} values.

The Steps a, b and c in Figure 1 are derived from the crystal structure of homing endonuclease I-PpoI in complex with DNA (PDB ID: 1A74) (15). Figure 2 plots the PST_{native}, I, and number of contacts against the DNA sequence. The Step 'a' is located at the center of the DNA and corresponds to the cleavage site of the endonuclease. The DNA bends here, with a large roll parameter of 11.2° at the central base-pair step. The native sequence of this region is TTAA. The PST_{native} is the largest among the PSTs of all possible dimer sequences in the AGTC and RY spaces (TA and YR are the native sequences) and of all possible tetramer sequences in the RY space (YYRR is the native sequence). The specificity scores are also larger compared to those in the adjacent steps, indicating that the conformation is specific to the native sequence. An experiment has consistently shown that the native sequence is strongly preferred at this position (95% preference to T for the first position of the tetramer sequence, 100% to T for the second, 95% to A for the third and 100% to A for the fourth) (16). Since direct contacts between the protein and the DNA are mainly formed outside this tetramer (15,17), the native sequence contributes to the binding through the indirect readout mechanism, by enabling the DNA to bend and fit the shape of the protein surface.

The Steps b and c are located near the Step a. Wittmayer *et al*. reported that the cleavage by the enzyme, i.e. the affinity to the enzyme, was completely abolished by the substitution of C for A at Step b. The efficiency of the cleavage was reduced to 70–90% of the wild-type when G was substituted for T at Step c (3). These substitutions reduced the PSTs to nearly zero in both cases. Since the base pairs at these positions do not directly interact with the protein (Figure 2), our results clarify that the observed change in affinity is due to the reduction of the fitness of the sequence to the structure that is favorable to bind to the protein, indicating that these positions are also involved in the indirect readout mechanism.

Lu *et al*. suggested that B to A deformations of DNA in this complex is important for the recognition (18). The large PST and the specificity scores of Steps a, b and c are correspondent to the large deformation at the TCT/AGA trinucleotides, which they pointed out the structural junctions of A/B deformation. Thus, the deformability of

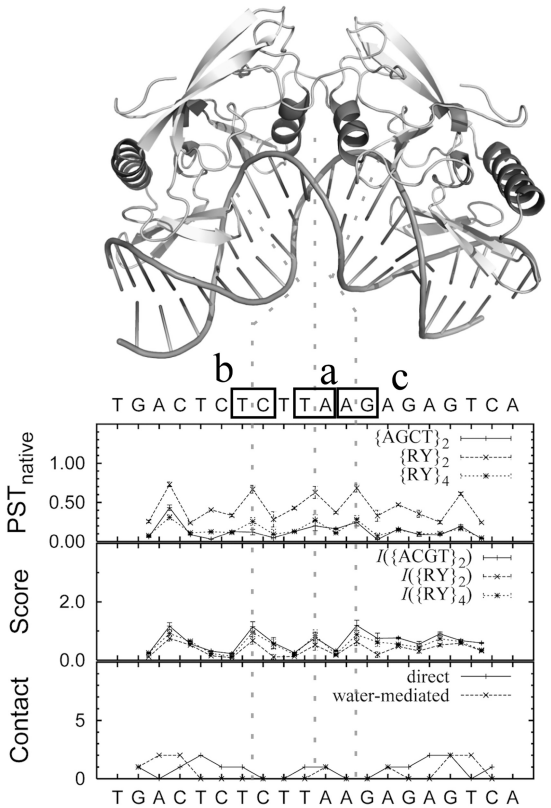

Figure 2. Schematic representation of the crystal structure of homing endonuclease I-PpoI in complex with DNA (PDB ID: 1A74) (top). Plots of PST_{native} [$PST_{native}(\{ATGC\}_4)$: solid line, $PST_{native}(\{RY\}_2)$: dashed line, $PST_{native}(\{RY\}_4)$: dotted line], specificity scores [$I(\{ATGC\}_2)$: solid line, $I(\{RY\}_2)$: dashed line, $I(\{RY\}_4)$: dotted line], number of contacts (direct: solid line with plus marks, water-mediated: dashed line with x marks) versus DNA sequence (bottom). Positions of Steps a, b and c are indicated with boxes in the sequence, and positions of their central base-pair steps are indicated with gray dashed lines in the structure image and in the plots.

DNA at these positions may play important role in the indirect readout.

Comparison of specificity scores between different sequence spaces

When the sequence can be decomposed into two parts, we can calculate the marginal entropy of the sub-sequence and the conditional entropy of the residual sequence, given the sub-sequence. Based on these values, we determined to what extent the step conformation restricts the variety of the sequence. Figure 3a and b show scatter plots of differences between $I(\{AGTC\}_2)$ and $I(\{RY\}_2)$ (ΔI_1) and between $I(\{RY\}_4)$ and $I(\{RY\}_2)$ (ΔI_2) against $I(\{RY\}_2)$, respectively. In the plot of ΔI_1, of 96 points with $I(\{AGTC\}_2) > 1$, 78 showed larger $I(\{RY\}_2)$ than ΔI_1, being located below the diagonal line representing $\Delta I_1 = I(\{RY\}_2)$. In these cases, transversions (exchanges between purines and pyrimidines) more severely affect the indirect readout mechanism than transitions (exchanges between A and G and between T and C) do. Similarly, in the plot of ΔI_2, of 62 points with $I(\{RY\}_4) > 1$, 61 showed larger $I(\{RY\}_2)$ than ΔI_2. This indicates that the

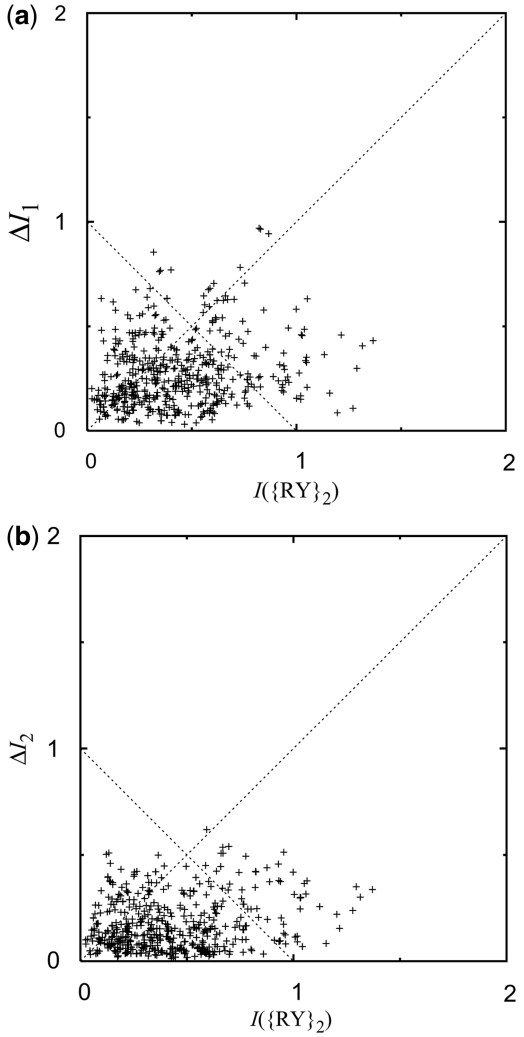

Figure 3. Scatter plots of $\Delta I_1 = I(\{ATGC\}_4) - I(\{RY\}_2)$ (**a**) and of $\Delta I_2 = I(\{RY\}_4) - I(\{RY\}_2)$ (**b**) against $I(\{RY\}_2)$. Diagonal dotted lines represent $\Delta I_1 = I(\{RY\}_2)$ and $\Delta I_2 = I(\{RY\}_2)$. Dotted lines downward to the right represent $I(\{ATGC\}_4) = 1$ and $I(\{RY\}_4) = 1$.

variety of the sequence of the central base-pair step is more strongly limited by the step conformation than that of the flanking sequence.

Limitations and prospects

The present method enables us to assess how the sequence adapts to a particular DNA conformation and quantify the specificity of indirect readout in terms of information entropy. We converted the probability of step conformations (PDF) in the sequence space directly into the probability of a sequence, given a step conformation (PST), by using Bayes' theorem. The PDF can be derived from conformational ensembles either obtained from known structural data or generated by MD simulations. Thus, the real application of the present method relies on the quality of conformational ensembles available. Unfortunately, it is the availability of conformational ensembles that is actually limiting the application. We attempted to use the trajectories of 10-ns MD simulations for DNA dodecamers containing 136 kinds of unique tetramer sequences (each simulation produced an ensemble of 9000 conformations). However, this amount of ensembles in 6D conformational space did not satisfy the statistical test for the probability (posterior distribution) of a sequence given a step conformation. Thus, we have reduced the sequence space into the dimer space or RY space so that we have a larger ensemble per sequence. This reduction in sequence space may result in the loss of some information, and the calculated probability of sequence and information entropy may provide only a low-resolution picture of the whole problem.

Another issue in the MD simulation is whether the simulation time is long enough to produce an equilibrium ensemble of conformations or not. We assume that the ensemble derived from the 10-ns MD simulations is a random sample extracted from a true population. If the structure is trapped in particular conformations or stay in some local minima surrounded by energy barriers during the 10-ns MD simulation, the trajectory will produce unreliable count $n(\mathbf{\Theta}|s)$ assigned to the conformation. In order to check how well an ensemble from a 10-ns MD simulation reflects the population, we carried out a 100-ns simulation for a DNA containing AGCC. Then, we evaluated the probability (*P*-value) that the count derived from the 10-ns simulation was equal to or less than $n(\mathbf{\Theta}|\text{AGCC})$, by using the probability $P(\mathbf{\Theta}|\text{AGCC})$ derived from the 100-ns simulation. As a result, the *P*-values were >5% for 94% of step conformations in the protein–DNA complexes, indicating that the difference between the ensembles from 10- and 100-ns simulations is not statistically significant for most of step conformations. We could present such comparison for one sequence, as the simulation is very time-consuming. However, in order to apply the present method to general problems, we need to carry out longer simulations for all the tetramer sequences. Such larger ensembles would enable us to obtain more reliable probability and information entropy, so that we can examine longer-range effect and assess the role of indirect readout in protein–DNA recognition.

CONCLUSIONS

We have developed a method to evaluate the contribution of indirect readout to the protein–DNA recognition. We used Bayes' theorem to derive the probability of having a particular sequence for a given DNA structure directly from an ensemble of structures with various sequences, which can be obtained from known structures of DNA or the trajectories of MD simulations of DNA. We also quantified the sequence specificity for a given DNA structure based on the information entropy. The method enabled us to identify the potential regions in a protein–DNA complex that have high specificity of indirect readout, and to evaluate how well the actual sequences fit to the structure. The advantage of the present method is that it does not need to use approximations and assumptions to the distribution of the conformational ensemble, and that the longer-range effect on the specificity can be

examined. The application of the present method to available experimental data is successful to some extent in explaining the mutation effect on the binding affinity. The present method can also predict new regions which are involved in the recognition through the indirect readout mechanism, even if experimental analysis is difficult, and would serve as a powerful tool to study the mechanism of protein–DNA recognition.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–2.

FUNDING

Grants-in-Aid for Scientific Research on Priority Areas 'Systems Genomics' [20016006 to K.S. and 20016022 to A.S.] and 'DECODE' [20052021 to A.S.]; Grants-in-Aid for Scientific Research (B) [20300103 to H.K.]; Grants-in-Aid for Challenging Exploratory Research [21651087 to H.K.]; Ministry of Education, Culture, Sports, Science and Technology of Japan and Grant-in-Aid for Young Scientists (B) [22700317 to S.Y.]. Funding for open access charge: Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST).

Conflict of interest statement. None declared.

REFERENCES

1. Sarai,A. and Kono,H. (2005) Protein-DNA recognition patterns and predictions. *Annu. Rev. Biophys. Biomol. Struct.*, **34**, 379–398.
2. Kispert,A. and Herrmann,B.G. (1993) The *Brachyury* gene encodes a novel DNA binding protein. *EMBO J.*, **12**, 3211–3220.
3. Wittmayer,P.K., McKenzie,J.L. and Raines,R.T. (1998) Degenerate DNA recognition by I-*Ppo*I endonuclease. *Gene*, **206**, 11–21.
4. Grillo,A.O., Brown,M.P. and Royer,C.A. (1999) Probing the physical basis for *trp* repressor-operator recognition. *J. Mol. Biol.*, **287**, 539–554.
5. Szymczyna,B.R. and Arrowsmith,C.H. (2000) DNA binding specificity studies of four ETS proteins support an indirect read-out mechanism of protein-DNA recognition. *J. Biol. Chem.*, **275**, 28363–28370.
6. 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.
7. 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.
8. 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.
9. Yamasaki,S., Terada,T., Shimizu,K., Kono,H. and Sarai,A. (2009) A generalized conformational energy function of DNA derived from molecular dynamics simulations. *Nucleic Acids Res.*, **37**, e135.
10. Shannon,C.E. (1948) A mathematical theory of communication. *Bell Sys. Tech. J.*, **27**, 379–423, 623–656.
11. Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
12. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
13. 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.
14. Wallace,A.C., Laskowski,R.A. and Thornton,J.M. (1995) LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Eng.*, **8**, 127–134.
15. Flick,K.E., Jurica,M.S., Monnat,R.J. and Stoddard,B.L. (1998) DNA binding and cleavage by the nuclear intron-encoded homing endonuclease I-*Ppo*I. *Nature*, **394**, 96–101.
16. Argast,G.M., Stephens,K.M., Emond,M.J. and Monnat,R.J. (1998) I-PpoI and I-CreI homing site sequence degeneracy determined by random mutagenesis and sequential in vitro enrichment. *J. Mol. Biol.*, **280**, 345–353.
17. Galburt,E.A., Chadsey,M.S., Jurica,M.S., Chevalier,B.S., Erho,D., Tang,W.L., Monnat,R.J. and Stoddard,B.L. (2000) Conformational changes and cleavage by the homing endonuclease I-*Ppo*I: a critical role for a leucine residue in the active site. *J. Mol. Biol.*, **300**, 877–887.
18. Lu,X.J., Shakked,Z. and Olson,W.K. (2000) A-form conformational motifs in ligand-bound DNA structures. *J. Mol. Biol.*, **300**, 819–840.