Predicting indirect readout effects in protein–DNA interactions

Yongli Zhang*, Zhiqun Xi†, Rashmi S. Hegde‡, Zippora Shakked§, and Donald M. Crothers*†¶

Departments of †Chemistry and *Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511; ‡Division of Developmental Biology, Children’s Hospital Research Foundation, 3333 Burnet Avenue, Cincinnati, OH 45229; and §Department of Structural Biology, Weizmann Institute of Science, Rehovot 76100, Israel

Contributed by Donald M. Crothers, April 2, 2004

Recognition of DNA by proteins relies on direct interactions with specific DNA-functional groups, along with indirect effects that reflect variable energetics in the response of DNA sequences to twisting and bending distortions induced by proteins. Predicting indirect readout requires knowledge of the variations in DNA curvature and flexibility in the affected region, which we have determined for a series of DNA-binding sites for the E2 regulatory protein by using the cyclization kinetics method. We examined 16 sites containing different noncontacted spacer sequences, which vary by more than three orders of magnitude in binding affinity. For 15 of these sites, the variation in affinity was predicted within a factor of 3, by using experimental curvature and flexibility values and a statistical mechanical theory. The sole exception was traced to differential magnesium ion binding.

Because many proteins deform DNA upon binding (1, 2), it is reasonable to expect that protein–DNA association might be facilitated by enhanced ease of DNA deformation and a match between intrinsic DNA shape in solution and the strained DNA conformation in the complex. The lack of a simple correspondence code between amino acids and DNA bases in hydrogen bonding, or direct readout, makes a general prediction of protein–DNA affinity impossible at this time. However, variations in contributions from indirect readout due to sequence-dependent shape and mechanical properties may be predictable, yielding a partial structural code for protein–DNA interaction (3). Many attempts to test this idea have been performed during the last two decades with different proteins, such as nucleosome (4–6), cAMP-binding protein (7), 434 repressor (8, 9), TATA box-binding protein (10), and E2 protein (11).

Among the impediments to realization of this objective is the difficulty of accurately determining the multiple DNA parameters involved, including magnitude and direction of curvature, helical twist, and bending and torsional flexibilities, for a variety of sequences corresponding to the region of indirect readout. The DNA cyclization method (1, 12, 13), in the high-throughput format we recently described (14), coupled with a statistical mechanical theory (15) for extracting the curvature and flexibility parameters from the data, provides a solution to this problem. Although the extent of variation of DNA flexibility with sequence remains controversial, recent results from the cyclization kinetics method show that a DNA sequence with high histone affinity (TATAAAGCGCC) has a nearly 2-fold smaller bending force constant and 35% less torsional rigidity than generic DNA (6). Also, an AT repeating sequence has 28% lower bending rigidity (14).

Experimental testing of the accuracy of prediction of indirect readout requires that the nucleotides involved not be also engaged in direct interactions. A system that meets this requirement is the noncontacted spacer region in the DNA-binding site for the E2 protein encoded by the human papillomavirus (HPV) type 16 genome (11). E2 proteins bind at multiple sites on the viral genome, regulate all papillomavirus genes, and are involved in viral DNA replication. The E2-binding sites of all viral strains incorporate a highly conserved dodecameric sequence of the general form ACCGNNNNGCGGT, where N4 (referred to as the “spacer”) is variable. Their ordered binding by E2 is thought to be crucial for viral activities. The DNA-binding domains (DBDs) of the E2 proteins from different papillomavirus strains are highly conserved. Comparison between crystal structures of the HPV-18 E2 DBD bound by high- and low-affinity DNA sequences shows that protein configurations, DNA trajectories, and direct protein–DNA hydrogen bonds are nearly the same, independent of the spacer sequences (16). The DNA molecules are bent by 45°, with an overall bending direction toward the minor groove of the spacer. There are no direct contacts between the protein and the spacer sequences. However, variations in spacer-dependent DNA-binding affinity for E2 proteins from different papillomavirus strains do not exclusively result from different DNA global structures and mechanical properties. It has been postulated that depending on which E2 proteins were used, the spacer sequence may be discriminated to variable extent by the distribution of positive charges on the DNA-interaction surface of the proteins due to long-range electrostatic interactions (11). The DBD of HPV-16 E2 was adopted for our study because the free protein crystal structure shows that this E2 protein has the minimal positive charges adjacent to the inferred spacer-sequence-binding site, which leads to a stronger dependence of the binding affinity on the spacer sequence compared with its homologues (17), thus fulfilling the requirement of maximal decoupling of indirect from direct readout.

By using the cyclization method, we have determined the variation in DNA curvature and flexibility for a series of 4-nt spacer sequences in the E2-binding site context. For completeness, we measured the DNA-binding affinity of the variable-spacer sequence set under the buffer conditions used for cyclization. Finally, we extended our recent theory of DNA cyclization (15) to replace circle formation constraints by constraints that place each end of the binding sequence at a specific position on the protein, enabling calculation of the relative binding affinity as a function of the structure and mechanical properties of the spacer sequence.

Materials and Methods

DNA Cyclization and Data Analysis. Detailed protocols were published elsewhere (14, 15). The same library of top template strands was used to make constructs by PCR and bottom template strands containing different test sequences were synthesized. The test sequences are three repeats of cgnmnggct, where mnnn represents the 4-bp spacer sequence.

To measure E2-induced DNA bending and flexibilities, the test sequence contains a single E2-binding site, i.e., AAGTGC-CACGNNNNGGCTGCAAGCTAG, where the binding site is shown in lowercase bold with the spacer either aacc or agct. Recombinant HPV-16 E2 DBD (18) was added to ~10 nM for AAAAC constructs and 20 nM for AGCT constructs, and incu-

Abbreviations: HPV, human papillomavirus; DBD, DNA-binding domain.

*To whom correspondence should be addressed. E-mail: donald.crothers@yale.edu.

© 2004 by The National Academy of Sciences of the USA
Chemically synthesized oligos were desalted, and then immediately loaded onto a 10% gel. DNA bands were detected by autoradiography and quantified. DNA corresponding to 2,000 cpm, with final concentration does not increase the cyclization rate. A total of 10 nM E2-DBD can only achieve partial occupancy of the major groove of ACCG, holding the DNA in an arc and leaving the spacer uncontacted by the E2 protein. (C) Calculated equilibrium DNA axes in the free states in solution and in the constrained states in E2–DNA complexes. The scaffold conformation is derived from measured global DNA curvature and twist-induced by HPV-16 E2 DBD binding. All of the DNA-binding sites are aligned by global translation and rotation to overlap their base pairs at one end and then deformed with minimum energy such that the base pairs at the other end coincide with that of the scaffold. The strained state is modeled with six constraints at this end, which is similar to the case in DNA cyclization, with its corresponding J factor calculated. The bending angles shown here are amplified by 1.5-fold for better clarity.

**Results and Discussion**

**The Model.** Fig. 1 shows the model system that was used to calculate the relative binding constants for the different binding sites. 

\[ K = \lambda J_{\text{loop}} K_{\text{half}} \]

where \( \lambda \) is a parameter that describes the cooperative binding between the two half-binding sites (ACCG), and \( J_{\text{loop}} \) is related to the J factor involved in DNA cyclization, but with an extended definition here. \( J \) was originally described by Jacobson and Stockmayer (21) as the effective concentration of one end of polymer chain at another to characterize the size-dependent circularization due to chain entropy effects. Here, we have extended its definition to non-zero end-to-end distance: Given one end of a DNA molecule at the origin, the \( J_{\text{loop}} \) factor is the effective concentration of another end at a specified location and orientation of the last base pair. It is proportional to the probability density of the end-constrained DNA configuration within unconstrained configuration space, and completely describes the accompanying DNA elastic energy and entropy changes upon its configurational transition from the free state in solution to the constrained state.

The \( J_{\text{loop}} \) factor can be accurately calculated from a statistical mechanical approach we recently developed for constrained DNA molecules when their curvature, helical repeat, and flexibilities are known (15). In this method, the mechanical equilibrium configuration of the end-constrained DNA molecule is first calculated and its \( J_{\text{loop}} \) factor is then evaluated by integrating the small configurational fluctuations around it. Although not formulated with a \( J \) factor, Eq. 1 was derived by Crothers and Metzger (22) to investigate the thermodynamic linkage between monomeric antibody-binding strengths and the overall association constants of multivalent antibodies, and was recently revisited for study of protein–DNA interactions for proteins with two domains connected by flexible linkers (23). Our approach introduces flexibility by means of the more accurate worm-like chain model. It is worth emphasizing that enhanced DNA flexibility does not always guarantee higher affinity, because flexible chains have a greater entropy loss due to the constraints imposed by complex formation. The theory used here takes into account both bending energy and chain entropy effects.

Our strategy is to detect DNA global structures and mechanical properties in solution for a variety of E2-binding sites with the high-throughput DNA cyclization approach, which measures \( J \) factors for cyclized DNA molecules (13). The test sequences contain three repeats of the spacer sequence (for heightened sensitivity), but because the binding site spans 12 bp, maintaining a 10-bp spacing requires overlap of the consensus sequences. Because the repeat used was CGNNNCGGT, the curvature and other properties determined do not correspond directly to the entire binding site. However, there is good correspondence between the overall bend angles we measured and the available crystallographic data on the free DNA targets (3, 24). The essential point in our measurement is to determine the variation of curvature and mechanical properties with spacer sequence, which is rigorously performed by our constructs.

With sequence-dependent properties in hand, the next step is to calculate the \( J_{\text{loop}} \) factors for the constrained state in the complex by using the derived DNA curvature and flexibility parameters, and finally compare their relative values, \( J_{\text{rel}} \), with the experimental relative binding constants \( K_{\text{rel}} \). The binding constant for the half-site, \( K_{\text{half}} \), is estimated to be \( < 10^{-5} \text{M}^{-1} \), but...
Global DNA Structure in the Complex. Because the crystal structure of the HPV16 E2–DNA complex is not available, we measured the protein-induced global curvature and twist by DNA cyclization (19) for single copies of high- and low-affinity full-binding sites, represented by AAAC and AGCT, respectively. Results from the total length assay are shown in Fig. 2. Matches in the measured J factors for the two binding sites indicate that their global DNA structures and flexibilities are nearly identical and independent of affinity (Table 1), which is consistent with DNA bending of high- and low-affinity sites measured by circular permutation assays using full-length HPV-16 E2 protein (26) and the crystal structures of HPV-18 E2–DBD-DNA complexes (16), but in contrast with cAMP-binding protein–DNA complexes in which the magnitude of DNA curvature correlates with affinity (7). DNA twist barely changes upon E2 binding. Protein binding dramatically reduces the apparent local DNA bending affinity (7). DNA twist barely changes upon E2 binding. Protein binding dramatically reduces the apparent local DNA bending affinity (7).

Derived DNA Properties and Binding Affinities. Table 2 exhibits the experimental global structures and flexibilities for 16 binding sites in their free state, along with their relative binding constants, determined by the electrophoretic mobility shift assay. Here again, as for the bound state, DNA curvature is modeled by a single roll angle that approximates the global bending of the double helix. Hence, such values may be either larger or smaller than individual roll angles observed in the crystal structures, where DNA curvature is rather smooth and partitioned along several base pairs. The solution-derived roll angles of the E2–DNA targets, in a coordinate frame at the center of the spacer (except for AAAA; see Table 2), vary from −10° to AAAA to +2.5° for GGCC; the negative sign indicates bending into the minor groove, as in the protein-bound state. The bending flexibilities, expressed as rms fluctuations in roll and tilt angles, vary from 4.45° for GGCC to 6.40° for ATAT. Note that the bending force constant varies with the inverse square of the rms fluctuation angle, so the force constant for GGCC is ∼2-fold larger than for ATAT. The twisting flexibilities have a similar range, but as discussed below, twist flexibility has only a minor effect on the binding constant when DNA twist is changed little upon binding.

Table 1. Global structural parameters of E2 protein–DNA complexes

<table>
<thead>
<tr>
<th>Spacer sequence</th>
<th>Bend angle,°</th>
<th>Bending flexibility,°</th>
<th>Twist angle,°</th>
<th>Bending flexibility,°</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAAC</td>
<td>−41.6 ± 0.8</td>
<td>3.41 ± 0.06</td>
<td>34.39 ± 0.12</td>
<td>3.7 ± 0.3</td>
</tr>
<tr>
<td>AGCT</td>
<td>−42.3 ± 1.2</td>
<td>3.38 ± 0.08</td>
<td>34.36 ± 0.15</td>
<td>3.7 ± 0.5</td>
</tr>
</tbody>
</table>

*The bend is by roll at the central dinucleotide step.

The persistence length \( P \) (when expressed in radians) are related to the persistence length \( P \) by \( P = 1/\alpha \) with the helical rise \( l = 3.4 \) Å, or 1/\( \alpha \), as unit length (1, 15). Twist fluctuations \( \alpha \) are related to the torsional force constant \( C \) by \( C = 1/\alpha l \).
sites may be related to the nonsymmetrical conformation of this target, where the bending direction is one step away from the center, unlike all of the other binding sites. Such an arrangement may not be optimal for a symmetric protein dimer, and the DNA duplex would have to writhe to bind E2, leading to free-energy loss relative to the symmetrically bent DNA targets. In contrast to AT-rich sequences, GC-rich sequences curve slightly into the major groove, and are more rigid. Similar bending characteristics were observed in crystal structures of free E2 DNA targets with spacers ACGT (3), GTAC (3) AATT (24), and AAAA (H. Rozenberg, M. Kitayner, and Z.S., unpublished work).

Comparison Between Experiment and Theory. The apparently complex dependence of binding affinity on DNA properties is accurately predicted by the theory. Table 2 shows $J_{rel}$ and $K_{rel}$ values, which give the $J_{loop}$ and $K$ values in terms of their ratio to the value for the reference ACGT sequence. Because $K$ is proportional to $J_{loop}$, the ratio $J_{rel}/K_{rel}$ should be unity if the prediction is accurate. Even though the binding affinities span three orders of magnitude, the calculated $J_{rel}/K_{rel}$ ratio is indeed close to unity; the average (excluding the outlying TATA sequence) is $1.2 \pm 0.7$. As is evident from Table 2, the one striking exception is the TATA sequence, whose measured affinity is 44-fold lower than predicted. We found that this disparity is much reduced at $Mg^{2+}$ concentrations $<10$ mM, and nearly eliminated in $Mg^{2+}$-free buffers. Preliminary experiments show that the TATA sequence releases approximately twice as many $Mg^{2+}$ ions as the average of the other sequences upon E2 binding (Y.Z., Z.X., and D.M.C., unpublished work). Also, the anomaly of this sequence may be linked to its propensity to adopt an A-DNA-like conformation (i.e., bending into the major groove) as a result of the two T-A hinges (32).

The contrasting sensitivity of the affinity to the multiple parameters is further confirmed by the calculated dependence of the $J$ factor on these parameters, shown in Fig. 3. The $J$ factor dramatically changes with DNA curvature and bending flexibility. In contrast, the $J$ factor changes are $<2$-fold when DNA twist and torsional flexibility vary in their accessible regions. In a case such as the E2–DNA complex, in which there is virtually no twist change upon binding, there is nothing to be gained by an increase in the torsional flexibility, but there is entropy to be lost because the increased twisting fluctuations are quenched in the complex. For example, an increase of the torsional flexibility of ATAT from $4.0^\circ$ to $5.1^\circ$ decreases its calculated relative $J$ factor from 372 to 292. On the other hand, in a case such as 434 repressor, in which the protein twists the DNA upon binding, the $J$ factor and binding constant should show high sensitivity to DNA torsional flexibility (9).

Summary and Discussion. In general, the relationships between DNA sequence, global DNA structure and mechanical properties, and DNA-protein affinity studied here, indicate that AT-rich sequences tend to bind protein better than GC-rich sequences when bending toward the minor groove is required, as seen for nucleosomes (4–6), cAMP-binding protein (7), and 434 repressor (8). The high-affinity sequences all bend toward the minor groove and simultaneously have high bendability compared with the low-affinity sites, which are straight, bent into the major groove, or rigid. Hence, DNA curvature and bending flexibility are the principle components determining, and equally contributing to, the affinity changes. Both the magnitude and direction of curvature are important. In contrast to DNA bending and bendability, DNA twist and torsional flexibility only slightly modulate the affinities.

Table 2. Comparisons between the predictions of the relative binding constants and experimental measurements

<table>
<thead>
<tr>
<th>Spacer</th>
<th>Roll, °</th>
<th>Twist, °</th>
<th>Bending flex, °</th>
<th>Twisting flex, °</th>
<th>$J_{rel}$</th>
<th>$K_{rel}$</th>
<th>$J_{rel}/K_{rel}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAAT</td>
<td>−6.5 (0.4)</td>
<td>34.26 (0.16)</td>
<td>5.73 (0.06)</td>
<td>4.8 (0.4)</td>
<td>281 (41)</td>
<td>340 (60)</td>
<td>0.8</td>
</tr>
<tr>
<td>ATAT</td>
<td>−4.6 (0.6)</td>
<td>34.28 (0.18)</td>
<td>6.40 (0.04)</td>
<td>5.1 (0.4)</td>
<td>292 (42)</td>
<td>320 (34)</td>
<td>0.9</td>
</tr>
<tr>
<td>AATT</td>
<td>−7.5 (0.5)</td>
<td>34.32 (0.14)</td>
<td>5.41 (0.09)</td>
<td>4.8 (0.3)</td>
<td>262 (49)</td>
<td>300 (52)</td>
<td>0.9</td>
</tr>
<tr>
<td>AATG</td>
<td>−6.2 (0.6)</td>
<td>34.47 (0.17)</td>
<td>5.61 (0.05)</td>
<td>4.9 (0.4)</td>
<td>223 (43)</td>
<td>296 (65)</td>
<td>0.8</td>
</tr>
<tr>
<td>AAAC</td>
<td>−8.4 (0.6)</td>
<td>34.52 (0.16)</td>
<td>4.94 (0.04)</td>
<td>4.4 (0.4)</td>
<td>198 (43)</td>
<td>260 (22)</td>
<td>0.8</td>
</tr>
<tr>
<td>AAGT</td>
<td>−5.6 (0.6)</td>
<td>34.61 (0.18)</td>
<td>5.45 (0.06)</td>
<td>4.8 (0.4)</td>
<td>156 (29)</td>
<td>214 (47)</td>
<td>0.7</td>
</tr>
<tr>
<td>AAAA</td>
<td>−10.5 (0.5)</td>
<td>34.85 (0.15)</td>
<td>4.86 (0.05)</td>
<td>4.5 (0.5)</td>
<td>121 (23)</td>
<td>160 (25)</td>
<td>0.8</td>
</tr>
<tr>
<td>TTAA</td>
<td>−0.2 (0.3)</td>
<td>34.51 (0.13)</td>
<td>5.31 (0.06)</td>
<td>5.2 (0.4)</td>
<td>19 (4)</td>
<td>9.4 (5.3)</td>
<td>2.0</td>
</tr>
<tr>
<td>TATA</td>
<td>3.0 (0.3)</td>
<td>34.61 (0.13)</td>
<td>5.02 (0.06)</td>
<td>5.4 (0.5)</td>
<td>132 (20)</td>
<td>3.0 (1.1)</td>
<td>44†</td>
</tr>
<tr>
<td>GATC</td>
<td>0.9 (0.5)</td>
<td>34.21 (0.14)</td>
<td>4.71 (0.04)</td>
<td>4.6 (0.5)</td>
<td>2.7 (0.7)</td>
<td>1.5 (0.1)</td>
<td>1.8</td>
</tr>
<tr>
<td>ACGT†</td>
<td>1.7 (0.4)</td>
<td>34.08 (0.12)</td>
<td>4.53 (0.05)</td>
<td>4.1 (0.3)</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GGCC</td>
<td>2.5 (0.5)</td>
<td>34.04 (0.19)</td>
<td>4.45 (0.03)</td>
<td>4.2 (0.5)</td>
<td>0.45 (0.12)</td>
<td>0.71 (0.04)</td>
<td>0.6</td>
</tr>
<tr>
<td>GACC</td>
<td>1.9 (0.5)</td>
<td>34.42 (0.11)</td>
<td>4.52 (0.05)</td>
<td>3.9 (0.3)</td>
<td>0.92 (0.29)</td>
<td>0.7 (0.2)</td>
<td>1.3</td>
</tr>
<tr>
<td>AGCT</td>
<td>2.2 (0.7)</td>
<td>34.40 (0.21)</td>
<td>4.52 (0.07)</td>
<td>4.3 (0.5)</td>
<td>0.72 (0.33)</td>
<td>0.6 (0.3)</td>
<td>1.2</td>
</tr>
<tr>
<td>GTAC</td>
<td>2.4 (0.5)</td>
<td>34.33 (0.15)</td>
<td>4.60 (0.04)</td>
<td>4.7 (0.4)</td>
<td>0.85 (0.26)</td>
<td>0.5 (0.2)</td>
<td>1.7</td>
</tr>
<tr>
<td>CGGG</td>
<td>2.3 (0.6)</td>
<td>34.29 (0.17)</td>
<td>4.51 (0.04)</td>
<td>5.1 (0.5)</td>
<td>0.55 (0.18)</td>
<td>0.18 (0.03)</td>
<td>3.1</td>
</tr>
<tr>
<td>Average</td>
<td>−2.4 (4.5)</td>
<td>34.39 (0.21)</td>
<td>5.10 (0.62)</td>
<td>4.7 (0.4)</td>
<td>106 (115)</td>
<td>119 (143)</td>
<td>1.27 (0.7)</td>
</tr>
</tbody>
</table>

*The roll is for the middle dinucleotide step of the spacer sequence for all the binding sites, except for AAAA, for which the roll is at the step one base pair to the right. Values in parentheses signify SD of the measurement.

†The outlying value of 44 for $J_{rel}/K_{rel}$ for TATA differs by $>10$ SD from the mean and has been excluded from calculation of the average.

‡This sequence is used as a reference for the relative $J$ factor and binding constant. Its absolute dissociation constant is 12 nM.

![Fig. 3. Variation of the $J$ factor with each parameter for global DNA structure and mechanical properties. Unless indicated as variable, the default parameters used for calculations shown here are 0° kink, 34.45° twist, 4.678° bending flexibility, and 4.388° twisting flexibility, respectively. The kink is a roll at the middle of the binding site. Note the difference in scale for the $J$ factor between the two images.](image-url)
The spacer sequences of the natural DNA-binding sites of HPV-16 (AATT, TTTT, AAAT, and AAAC) are AT-rich, which showing considerable intrinsic bending into the minor groove (Table 2). The AATT sequence exhibits the highest affinity in this series (26) in agreement with the current study, most likely due to its symmetric configuration and enhanced bendability (Table 2). The dramatically lower affinity (by three orders of magnitude) shown by the wild-type TTTT target (ACCGTTTTGGGT) is mainly a consequence of substitution of a conserved cytosine by a guanine base at the 3’ end of the dodecameric target (26). The biological consequence (transcriptional activation/repression or viral DNA replication) of E2 binding to each of the four binding sites on the genome is different. Thus, both high- and low-affinity binding sites are critical to the life cycle of the pathogenic virus.

We show that differential binding affinity in the HPV-16 E2 system is determined by indirect readout and that these effects can be accurately predicted by theory. Such studies provide insights into structure–function relationships in gene regulation.

Y.Z. thanks Jing Wang for technical assistance. This work was supported by National Institutes of Health Grants GM 21966 (to D.M.C.) and CA66964 and GM63582 (to R.S.H.); the National Foundation for Cancer Research in support of the National Foundation for Cancer Research–Yale Center for Protein and Nucleic Acid Chemistry; and the Israel Science Foundation (to Z.S.).