doi:10.1016/j.jmb.2004.01.033 J. Mol. Biol. (2004) **337**, 285–294

JMB

Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Intermolecular and Intramolecular Readout Mechanisms in Protein–DNA Recognition

M. Michael Gromiha[1], Jörg G. Siebers[2], Samuel Selvaraj[3] Hidetoshi Kono[4] and Akinori Sarai[2]*

[1]*Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Aomi Frontier Building 17F, Aomi Koto-ku, Tokyo 135-0064 Japan*

[2]*Department of Biochemical Engineering and Science Kyushu Institute of Technology (KIT), 680-4 Kawazu, Iizuka 820-8502, Japan*

[3]*Department of Physics Bharathidasan University Tiruchirapalli 620 024 Tamilnadu, India*

[4]*Neutron Science Research Center and Center for Promotion of Computational Science and Engineering, Japan Atomic Energy Research Institute (JAERI), 8-1 Umemidai, Kizu-cho Souraku-gun, Kyoto 619-0215 Japan*

**Corresponding author*

Protein–DNA recognition plays an essential role in the regulation of gene expression. Regulatory proteins are known to recognize specific DNA sequences directly through atomic contacts (intermolecular readout) and/or indirectly through the conformational properties of the DNA (intramolecular readout). However, little is known about the respective contributions made by these so-called direct and indirect readout mechanisms. We addressed this question by making use of information extracted from a structural database containing many protein–DNA complexes. We quantified the specificity of intermolecular (direct) readout by statistical analysis of base–amino acid interactions within protein–DNA complexes. The specificity of the intramolecular (indirect) readout due to DNA was quantified by statistical analysis of the sequence-dependent DNA conformation. Systematic comparison of these specificities in a large number of protein–DNA complexes revealed that both intermolecular and intramolecular readouts contribute to the specificity of protein–DNA recognition, and that their relative contributions vary depending upon the protein–DNA complexes. We demonstrated that combination of the intermolecular and intramolecular energies derived from the statistical analyses lead to enhanced specificity, and that the combined energy could explain experimental data on binding affinity changes caused by base mutations. These results provided new insight into the relationship between specificity and structure in the process of protein–DNA recognition, which would lead to prediction of specific protein–DNA binding sites.

Keywords: protein–DNA recognition; direct and indirect readouts; specificity; statistical potential; structural data

Introduction

Protein–DNA interactions play a key role in many vital processes, including regulation of gene expression, DNA replication and repair, and packaging. The remarkable specificity with which proteins recognize target DNA sequences is of considerable theoretical and practical importance, and its basis has been demonstrated through structural analysis of large numbers of protein–DNA complexes.[1–5] Within these structures recognition involves, in part, direct contacts between amino acid residues and base-pairs (direct readout mechanism). That these contacts are both redundant and flexible suggests there is no simple code for the specificity of DNA–protein interactions.[6,7] In addition, the fact that mutation of bases not in direct contact with amino acid residues often affects the binding affinity[8] implies that water molecules bridging between amino acid residues and bases,[9] conformational changes in the DNA (e.g. bending),[10] and/or flexibility[11–13] also affects protein–DNA binding specificity (indirect readout mechanism). In terms of the energy contributed to the binding affinity, the direct readout and

Abbreviations used: IHF, integration host factor; PDB, Protein Data Bank; TBP, TATA-binding protein.

E-mail address of the corresponding author: sarai@bse.kyutech.ac.jp water-mediated contacts are intermolecular energies, whereas DNA deformation is associated with intramolecular energies. In order to avoid confusion, we hereinafter use the terms "intermolecular readout" and "intramolecular readout" to represent recognition *via* direct protein–DNA contact and *via* DNA deformations, respectively. Precisely how these intermolecular and intramolecular readout mechanisms contribute to the overall specificity is unknown; however, no good methods for quantifying those contributions have been available. The specificity of protein–DNA binding has been commonly predicted with a sequence-based method that uses sequence information from observed binding sites.[14] In addition, there have been several attempts to incorporate physical properties of DNA, such as the conformational properties and the stability, into the prediction scheme.[15–18] Still, it is difficult to separate the intermolecular and intramolecular contributions to the specificity using the sequence-based method.

We have developed a method for quantifying the specificity of intermolecular readout based on the statistical analysis of the structures of protein–DNA complexes.[19] We derived empirical potential functions for the specific interactions between bases and amino acids, and used these potentials to calculate the interaction energy, E_{PD}, for the protein–DNA complex. By threading different DNA sequences on the protein–DNA framework and calculating the total energy, we were able to quantify the difference in the fitness of various DNA sequences against the protein–DNA complex structure. This sequence-structure threading of random DNA sequences enabled us to calculate a Z-score defined by $(E_{\mathrm{PD}} - \langle E_{\mathrm{PD}}\rangle)/\sigma$, where $\langle E_{\mathrm{PD}}\rangle$ is the average interaction energy and σ is the standard deviation. This normalized quantity serves as a measure of the specificity of the protein–DNA interaction within a complex, and enabled us to examine the relationship between structure and specificity in cognate/noncognate,[19,20] symmetric/asymmetric,[20] and cooperative[19] binding. We have also used this method successfully to predict DNA target sites for regulatory proteins.[19]

The specificity of the intramolecular readout mechanism was quantified using the structural data from the protein–DNA complexes.[21,22] To evaluate intramolecular readout, we needed to evaluate the internal energy of the DNA within the complex to determine how the sequences fit into the DNA structure within the complex. For simplicity, we used only six conformational parameters (shift, slide, twist, rise, roll and tilt) to characterize the local geometry of each base-pair step. Then using a method developed for calculating the conformational energy from the structural data of protein–DNA complexes,[13] the internal DNA energy was approximated as the sum of harmonic functions along conformational coordinates. The corresponding force parameters and equilibrium geometries were estimated from the culled distributions using the aforementioned conformational variables in the protein–DNA complexes. We followed this protocol, adding a self-consistent scheme, to calculate the DNA energy, and then calculated a Z-score that represented the specificity of the intramolecular readout mechanism as we did for the intermolecular readout (see Methods); and because this is a normalized quantity, it could be directly compared with that of the intermolecular readout. We have compared these specificities for a large number of protein–DNA complexes. The results have revealed that both intermolecular and intramolecular readouts contribute to the specificity of protein–DNA recognition, and that their relative contributions vary depending upon the proteins within the complex. We show that combination of the intermolecular and intramolecular readout energies derived from the statistical analyses leads to enhanced specificity. We discuss the relationship between structure and specificity in protein–DNA recognition by considering several examples.

We did not consider the effect of water molecules explicitly in this study. The contribution of water is implicitly involved in our intermolecular readout, as some of the protein–DNA complexes used to derive the statistical potentials contain water molecules at the interface, but the statistics for water within the protein–DNA complex are weak. Here, the "intramolecular readout" is meant to represent only the DNA effect, not including protein conformation changes.

Results and Discussion

Z-scores for the intermolecular and intramolecular readouts of protein–DNA complexes

Table 1 shows the Z-scores for the intermolecular and intramolecular readouts of various protein–DNA complexes. We found that the complexes were roughly grouped into two clusters: one with a larger Z-score for intermolecular readout and the other with a larger Z-score for the intramolecular readout. The complexes listed in Table 1 were sorted according to $|Z(\text{intermolecular}) - Z(\text{intramolecular})|$, i.e. those with larger contributions from intermolecular readout are listed at the top. To derive the results in Table 1, we used a random sequence of 50,000 evenly distributed bases, i.e. with a ratio of 0.25 for each base, as a reference. Of course, within the actual genomic sequences, the base compositions are not uniform. We therefore examined the bias in random sequences with GC contents of 40% and 60%. Although there are some variations in Z-score values, the Z-score deviations fell within the standard error of the bootstrap test (figures in parenthesis in Table 1).

Comparison of the Z-scores with some of the

Table 1. The calculated *Z*-scores for intermolecular and intramolecular readout mechanisms

PDB code	Protein name	Motif	Sec. str.	Bending		Prokaryote/ eukaryote	*Z*-score	
				Angle	Type		Intermolecular	Intramolecular
1A74	Homing endonuclease I	Enzyme	na	64	kink2	E	**−1.6** (0.6)	**0.7** (0.6)
3CRO	434 Cro (OR1)	HTH	α	36	kink2	P	**−2.0** (0.9)	**0.3** (0.5)
1HCR	Hin recombinase	HTH	na	na	na	P	**−1.8** (0.7)	**0.4** (0.5)
1RV5	Endonuclease *Eco* RV	Enzyme	na	38	kink1	P	**−2.3** (0.5)	**−0.3** (0.7)
1FJL	Paired homeodomain	HTH	α	62	kink2	E	**−2.7** (0.8)	**−1.0** (0.6)
1BHM	Endonuclease *Bam* HI	Enzyme	na	4	Straight	P	**−2.9** (1.0)	**−1.3** (0.3)
1CDW	Human TBP core domain	β-Ribbon	β	104	kink2	E	**−2.2** (0.5)	**−0.6** (0.4)
1YRN	MAT-a1/α2	HTH	α	62	kink2	E	**−4.4** (0.6)	**−2.9** (0.5)
1A02	NFAT/Fos/Jun	na	Loop	15	kink1	E	**−3.4** (0.6)	**−1.8** (0.4)
1MEY	Consensus zinc finger protein	ZF	α	29	kink2	na	**−3.6** (0.6)	**−2.2** (0.5)
1PER	434 repressor (OR3)	HTH	α	41	kink2	P	**−2.5** (0.7)	**−1.1** (0.4)
1BER	Catabolite gene activator protein	HTH	α	94	kink2	P	**−2.0** (0.6)	**−0.8** (0.3)
1YSA	GCN4	LZ	α	48	kink2	E	**−3.0** (0.8)	**−2.1** (0.5)
1TF3	Transcription factor Iiia	ZF	α	12	kink1	E	**−3.2** (0.7)	**−2.3** (0.5)
1MNM	MAT-α2/MCM1	HTH	α	64	kink2	E	**−4.4** (0.7)	**−3.0** (0.5)
1SRS	Serum response factor	Coiled coil	Loop	na	Circular	E	**−3.0** (0.6)	**−2.4** (0.6)
1SVC	Transcription factor Nfkb	na	Loop	28	kink1	E	**−2.6** (0.8)	**−2.2** (0.4)
1GDT	Recombinase-resolvase	HTH	α	75	kink2	P	**−2.0** (0.6)	**−1.7** (0.5)
1BL0	Multiple antibiotic resistance	HTH	na	56	kink2	P	**−2.7** (0.5)	**−2.5** (0.7)
1D66	GAL4	ZF	α	na	na	E	**−1.8** (0.7)	**−1.7** (0.6)
1DP7	MHC class II transcription factor	Winged HTH	β	48	kink2	E	**−0.8** (1.0)	**−0.7** (0.3)
1ECR	Replication terminator protein	β-Ribbon	β	35	kink2	P	**−1.1** (0.7)	**−1.1** (0.5)
1GLU	Glucocorticoid receptor	ZF	α	25	kink2	E	**−1.1** (0.7)	**−1.1** (0.3)
1MHD	Smad Mh1 domain	na	na	9	kink1	E	**−1.9** (1.0)	**−1.9** (0.5)
1TSR	p53 tumor suppressor	LSH	Loop	22	kink1	E	**−1.1** (0.7)	**−1.2** (0.4)
1CJG	Lac repressor	HTH	α	38	kink1	P	**−1.1** (0.9)	**−1.4** (0.9)
1XBR	T Protein	β-Barrel	na	30	kink2	E	**−2.0** (0.7)	**−2.4** (0.5)
1PDN	Paired domain	HTH	α	18	kink1	E	**−2.0** (0.5)	**−2.5** (0.5)
1OCT	Oct-1 POU homeodomain	HTH	α	45	kink2	E	**−1.6** (0.7)	**−2.1** (0.6)
1B3T	Nuclear Protein Ebna1	α-Helix	na	49	kink2	E	**−1.4** (0.8)	**−2.1** (0.7)
1HDD	Engrailed homeodomain	HTH	α	27	kink2	E	**−1.1** (0.6)	**−1.8** (0.4)
1HRY	Human SRY	HMG box	α	na	Circular	E	**−0.2** (0.4)	**−0.9** (0.3)
1HCQ	Estrogen receptor	ZF	α	28	kink2	E	**−1.7** (0.6)	**−2.5** (0.5)
1UBD	Human YYI	ZF	α	26	kink2	E	**−1.3** (0.9)	**−2.1** (0.8)
2BOP	Bovine Papillomavirus-1 E2	α-Helix	na	52	kink2	E	**−0.9** (0.6)	**−1.7** (0.6)
1TC3	Transposase	HTH	na	66	kink2	E	**−1.7** (0.6)	**−2.5** (0.7)
1IHF	Integration Host Factor (IHF)	β-Ribbon	β	174	kink2	P	**−1.2** (0.4)	**−2.3** (0.6)
2DRP	Tramtrack protein	ZF	α	38	kink2	E	**−1.2** (0.8)	**−2.3** (0.5)
1REP	Replication Initiation Protein	HTH	α	27	kink2	P	**−2.0** (0.6)	**−3.2** (0.4)
1IF1	Interferon Regulatory Factor 1	HTH	α	na	na	E	**−0.4** (0.8)	**−1.7** (0.4)
1GAT	GATA-1	ZF	α	na	Circular	E	**−0.4** (0.7)	**−1.7** (0.5)
1CMA	Met repressor	β-Ribbon	α	41	kink2	P	**−0.2** (0.8)	**−1.6** (0.8)
1LMB	λ Repressor	HTH	α	36	kink2	P	**−2.9** (0.6)	**−4.3** (1.7)
1PUE	PU.1 ETS domain	HTH	α	39	kink2	E	**−1.1** (0.4)	**−2.7** (0.5)
1MSE	Myb	HTH	α	na	Circular	E	**−0.4** (0.7)	**−2.0** (0.4)
1HLO	Transcription factor Max	HLH	na	na	Circular	E	**0.1** (0.7)	**−1.6** (0.5)
1TRO	Trp repressor	HTH	α	31	kink2	P	**−1.3** (0.7)	**−3.1** (0.6)
1MDY	MyoD bHLH domain	HLH	α	32	kink2	E	**−0.7** (0.6)	**−2.5** (0.5)
1IGN	Rap1	HTH	α	31	kink2	E	**0.0** (0.6)	**−2.2** (0.5)
6CRO	Cro repressor	HTH	α	na	Circular	P	**0.0** (0.6)	**−2.3** (0.8)
1PAR	Arc repressor	β-Ribbon	β	33	kink2	P	**0.6** (0.7)	**−1.7** (0.6)

The abbreviations used are: HTH, helix-turn-helix; HLH, helix-loop-helix; ZF, zinc finger; LZ, leucine zipper; LSH, loop-sheet-helix; na, not available. Sec. str, recognizing secondary structure; kink1, single kink; kink2, double kink; E, eukaryote; P, prokaryote. (The angle and type of bending were taken from http://www.imb-jena.de/Piet/html/). The values in the parentheses following *Z*-scores represent bootstrap standard errors (see Methods).

structural and functional features of these complexes showed enzymes to be ranked at the top in the list, indicating a major contribution of intermolecular readout to their DNA binding. In general, however, the structural motif of proteins does not seem to have any preference for intermolecular or intramolecular readout; nor do the secondary structural elements involved in the recognition. In addition, the relation between structural deformation (e.g. bending) and specificity is rather complex. Minor-groove binding proteins, which often severely distort DNA geometry, have been thought to use an intramolecular readout mechanism, but the present results indicate a subtle interplay between intermolecular and intramolecular readouts. Below we discuss these issues further in the context of several examples.

Major role of intermolecular readout in Zn fingers

DNA recognition by zinc finger proteins has been studied extensively[23–26] and provides us a good model system for the validation of the calculation. We have already shown[19] that the specificity of DNA sequence recognition by Zn fingers can be explained by intermolecular readout quite well. Indeed, the calculated *Z*-scores show that the intermolecular readout makes the larger contribution to specificity (Table 1). We have now further analyzed the designed Zn finger (PDB code, 1MEY).[27] The complex structure contains three fingers, each recognizing a three base-pair DNA subsite with a different sequence. In order to dissect the contribution of the intermolecular readout mechanism, we fixed the DNA sequence (GAGGCAGAA) in the protein–DNA complex structure and swapped the seven amino acid residues (positions from −1 through 6) responsible for sequence recognition among the three fingers. Thus, there are $3 \times 3 \times 3 = 27$ possible combinations of the finger configuration. We then calculated the *Z*-score for each structure. The calculated *Z*-scores range from −3.6 to −1.9, with the original finger combination in the crystal structure giving the lowest *Z*-score (−3.6). Since this result is based on protein mutations without changes in DNA sequence, the selection reflects only the intermolecular readout. The finding that the lowest *Z*-score corresponds to the original finger combination indicates an important role for the intermolecular readout mechanism in DNA recognition by Zn fingers. It is also notable that calculated recognition sequences of the proteins with different order of the three zinc fingers are almost perfectly matched (average 8.4 out of nine bases) to the target sequences deduced by experimental data.[25,26]

In addition, the same finger-swapping calculations were carried out for zif268. The *Z*-score ranges from −3.3 to −2.3. The co-crystallized DNA sequence (GCGTGGGCG) yielded a *Z*-score of −3.1, which was the fifth lowest among the 27 combinations. The lowest *Z*-score against the co-crystallized sequence was obtained for the modeled protein having finger 1, finger 2, and finger 1 sequences in the finger 1, finger 2 and finger 3 positions, respectively. This result is reasonable, since the fingers 1 and 3 have the same key residues to recognize GCG sequence.[23,24] The other low *Z*-score finger combinations may be explained by residues other than the key residues, as well as the effect of intramolecular readout, which can affect the binding affinity.

Cognate and non-cognate forms of *Eco*RV

Being a restriction enzyme, *Eco*RV's recognition of its target sequence is very stringent. Comparison of the structures of this enzyme shows that there are significant differences in the conformations of the free form and the forms in complex with cognate and non-cognate DNAs,[28] and that there is also significant deformation of the DNA.[28,29] We found that the *Z*-scores for intermolecular and intramolecular readouts were, respectively, −2.3 and −0.3 for 1RV5 and −1.1 and −0.1 for 4RVE, both of which complex with cognate DNA. These values are indicative of the major role played by intermolecular readout in the recognition. On the other hand, the *Z*-scores for intermolecular and intramolecular readouts of a non-cognate complex (2RVE) were 1.0 and 0.6, respectively, indicating intramolecular readout to contribute more substantially in this case. We have shown that the conformational change in *Eco*RV from the non-cognate to the cognate form contributes to the specificity of the intermolecular readout.[20] This suggests that conformational changes in this protein and the DNA may be necessary to bring amino acid residues and bases into intermolecular contact for recognition of the target sequence.

By introducing base mutations and base analogs into the central TA base step of a target sequence (GATATC) located where the DNA exhibits a sharp (50°) bend into the major groove, Martin *et al*.[30] were able to dissect the structural and energetic origins of site-specific DNA cleavage by *Eco*RV in terms of intermolecular and intramolecular readouts. Their analysis showed that intermolecular readout provides 5 kcal/mol toward catalytic specificity, whereas intramolecular readout contributes 6–10 kcal/mol. We have examined all possible base substitutions at those two positions and estimated the energetic contributions of the intermolecular and intramolecular readouts. We found that the energies of both were the lowest for the target sequence. Upon substitution of TA with CG, the *Z*-score values of the intermolecular and intramolecular readouts increased by 2.0 and 1.8, respectively, i.e. specificity was entirely lost. Although the catalytic and binding specificities may not be compared directly, the calculated result agrees with experiment in that intramolecular readout makes a substantial contribution to the specificity for the recognition of the central TA sequence.

Proteins causing severe DNA bending

Many DNA-binding proteins deform the structure of the DNA, e.g. they may cause the DNA to bend. Integration host factor (IHF) is one such protein; indeed the structure of the IHF–DNA complex[31] (PDB code, 1IHF) shows the DNA to be severely bent (overall bending angle ~160°), and it has been suggested that the conformation of the DNA is crucial to the protein–DNA interactions in this case.[10,32] Consistent with that idea, the calculated Z-score for the intramolecular readout was -2.3, twice that of the intermolecular readout ($Z = -1.2$).

The TATA-binding protein (TBP) binds to the minor groove, severely bending the DNA.[33] In the absence of specific contact with base-pairs in the major groove, TBP is believed to recognize a specific conformation or property of the DNA. However, calculation of Z-scores for the crystal structure of the TBP–DNA complex (1CDW) gives an intramolecular readout Z-score of only -0.6, as compared to -2.2 for the intermolecular readout (Table 1), indicating the latter to make an important contribution to the recognition. Amino acid–base contacts in the minor groove are believed to play only a minor role in the sequence recognition. This result may seem surprising at first, but the specificity is the result of interplay between different forces. When TBP binds to DNA, the extensive contacts between TBP and DNA favor the interaction energy, but deformation of the DNA sacrifices the DNA energy. Thus, the extreme bending of the DNA caused by TBP actually makes the Z-score for the intramolecular readout rather moderate. On the other hand, the binding of TBP produces a wide open, underwound, shallow minor groove,[33] completely rearranging the positions of H-bond-forming hetero-atoms of base-pairs in the minor groove. This geometric rearrangement may enable TBP to use intermolecular readout from the minor groove. Serum response factor (1SRS) presents a similar case, and together these two proteins highlight the potential importance of a deformed minor groove in sequence recognition by the intermolecular readout mechanism.

Major role of intramolecular readout in ETS proteins and *trp* repressor

Evidence for intramolecular readout has sometimes been derived experimentally. For instance, Szymczyna & Arrowsmith[8] measured the binding affinity between ETS family proteins and various DNAs containing mutations in their flanking regions outside the core GGA trinucleotide sequence. They concluded that the recognition mechanism used by ETS proteins is partially governed by an intramolecular readout. Our calculated Z-scores for intermolecular and intramolecular readouts were -1.1 and -2.7, respectively, for the PU.1 ETS complex (1PUE) and -0.5 and -3.1 for SAP-1 ETS (1HBX), confirming the major role played by intramolecular readout in recognition.

The experiment by Szymczyna & Arrowsmith[8] provides an opportunity to validate our calculations for the contribution of intramolecular readout. The second C of the target sequence (ACCGGAAGT) of the SAP-1 ETS domain has no contact with any amino acid residues, while the third C has a contact with Arg69. As expected from the presence of Arg69, substitution of C by G at the third position substantially lowered the calculated energy due to intermolecular readout (by 1.0 Z-score unit), though C to G mutation at the second position slightly lowered the energy (by 0.3 Z-score unit). These results are not consistent with the experimental observation that these are the two most deleterious mutations,[8] indicating that intermolecular readout cannot explain the experimental result. On the other hand, the same substitutions increased the energy due to intramolecular readout (by 0.6 and 1.3 Z-score units, respectively), which is consistent with the experimental affinity changes. A similar result was obtained for mutations at the seventh and eighth positions. Experimentally observed affinity changes caused by these mutations were much smaller than those at the other end. The calculated energy changes due to intermolecular readout were opposite to the experimental affinity changes. On the other hand, A to T mutation at the seventh position caused a small but positive energy change, and G to C mutation at the eighth position gave slightly lower energy (by 0.1 Z-score unit) due to intramolecular readout.

In the case of PU.1 ETS, conservation in the flanking sequence is not very obvious. Experimental results from affinity measurements with individual mutations, multiplex binding and SELEX data, and consensus promoter sequences do not agree, indicating that individual bases do not independently contribute to complex stability.[8] The free energy changes due to the mutations were much smaller than in SAP-1, with the lowest binding free energy for ACGGGAAGT. The effect of the mutations on the calculated energy due to intermolecular readout was insignificantly small, and the energy for ACGGGAAGT was slightly lower than that for ACCGGAAGT. The calculated energy due to intramolecular readout was the lowest for ACCGGAAGT, although the variation of energy was smaller than in SAP-1. The specificity of PU.1 thus appears to be more complex than SAP-1.

trp Repressor provides another example of the major role played by intramolecular readout, as indicated by the Z-score in Table 1. In fact, the crystal structure of the *trp*–DNA complex involves only one direct contact between the guanidino group of Arg69 at the N terminus of helix D and the G at position 9/−9 of the 18 bp 2-fold symmetric DNA operator sequence, GTACTAGTT AACTAGTAC.[34] The protein–DNA complex structure exhibits a significant bend between positions −4 and −5, as well as a smaller bend at the central TA step.[34] The limited direct contacts between *trp* repressor and the DNA, the presence of interfacial water molecules, and the conformation of the DNA apparent in the crystal structure led Sigler and co-workers to propose an indirect readout mechanism for repressor–DNA recognition. *In vitro* binding experiments showed that mutations at 5/−5 and 6/−6 reduced the affinity significantly.[35,36] We have calculated the effect of mutations at these positions. G to C mutation at position −9 increased the energy for intermolecular readout by 0.3 *Z*-score unit, whereas mutations at other positions increased energy only by less than 0.1 *Z*-score unit. On the other hand, we found that the wild-type operator sequence gave the lowest energy due to intramolecular readout and most mutations increased energy significantly (by up to 1.1 *Z*-score units). Furthermore, replacement of the central TTAA by AATT, which would affect the bending of the DNA, destabilized the complex, as reflected by a 2 kcal/mol change in binding free energy.[36] Our calculation for this mutation resulted in an increase in intramolecular readout energy by 1.2 *Z*-score units. Interfacial water molecules have been proposed to play an important role in the recognition between *trp* repressor and DNA. However, the present result indicates that the conformational properties of DNA can also make a significant contribution to the specificity of *trp*-operator recognition.

Proteins of similar structures with distinct recognition modes

Families of proteins with similar structures often show distinct differences in their modes of sequence recognition. For instance, the structures of DNA–protein complexes involving the estrogen (1HCQ) or glucocorticoid (1GLU) receptor are similar. Nevertheless, the contributions of the intermolecular and intramolecular readouts differ in the two complexes. The estrogen receptor binds to DNA with higher specificity (Z(intermolecular) = −1.7 and Z(intramolecular) = −2.5) than the glucocorticoid receptor (Z(intermolecular) = −1.1 and Z(intramolecular) = −1.1). Moreover, in the estrogen receptor the intramolecular readout makes a stronger contribution to the binding than the intermolecular readout. We also compared the modes of recognition of λ repressor (1LMB), λ cro (6CRO), 434 repressor (1PER) and 434 cro (3CRO). The *Z*-scores for the intermolecular readout were −2.9 (1LMB), 0.0 (6CRO), −2.5 (1PER) and −2.0 (3CRO), while those for the intramolecular readout were −4.3 (1LMB), −2.3 (6CRO), −1.1 (1PER) and 0.3 (3CRO). This means that although they are similar in structure, repressors apparently bind with more specificity than Cro, and λ complexes and 434 complexes use intermolecular and intramolecular modes differently.

Combination of intermolecular and intramolecular readouts

So far, we have quantified the specificities of intermolecular and intramolecular readout mechanisms separately. We are also able to combine the two energies to calculate the total energy. However, because the derivations of these empirical energies are based on different statistics, we cannot simply make a summation. Instead, we must introduce a weighting factor: $E_{\mathrm{tot}} = cE_{\mathrm{PD}} + (1 - c)E_{\mathrm{DNA}}$, where E_{DNA} is the energy of the intramolecular readout and c is a weighting coefficient ranging between 0 and 1. This coefficient is determined by maximizing the total *Z*-score, i.e. the *Z*-score is calculated from random sequences, and a value of c is sought that gives the highest total *Z*-score. As an example, we considered 1YRN, a complex of DNA with MAT-a1 and α2, two proteins involved in determining mating type in yeast.[37] Figure 1 shows the total *Z*-score obtained from the combination of intermolecular and intramolecular readouts as a function of the weight factor c. Interestingly, the total *Z*-score (−5.3 at $c = 0.32$) was higher than either the *Z*-score for the intermolecular (−4.4) or intramolecular (−2.9) readout. One interpretation of this result is that the energies of the intermolecular and intramolecular readouts each contain independent information that in combination enhances the specificity of the recognition. If both energies are totally dependent or correlated, the

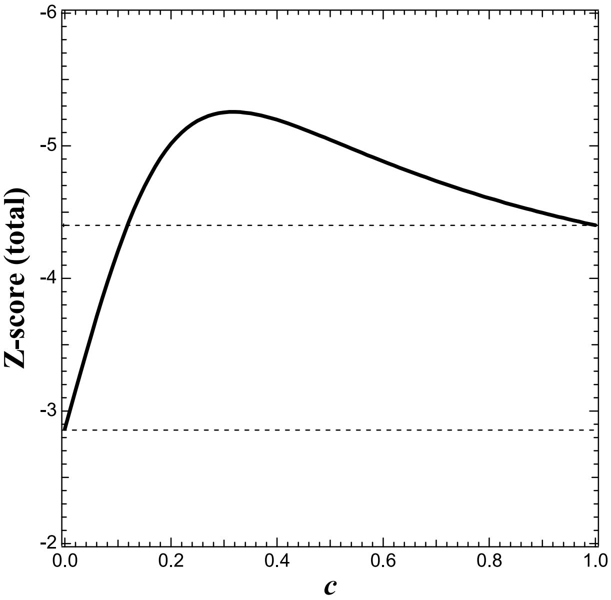

Figure 1. Enhanced specificity caused by the combination of intermolecular and intramolecular readouts for the MAT-a1/α2/DNA complex (1YRN). The *Z*-score was calculated for the combined total energy, $E_{\mathrm{tot}} = cE_{\mathrm{PD}} + (1 - c)E_{\mathrm{DNA}}$, and plotted as a function of the weighting coefficient c.

total *Z*-score would not increase. We examined this effect for all the systems listed in Table 1 and found that the total *Z*-score did indeed increase for those having two negative *Z*-scores.

Comparison of calculated energies with experimental activity data

In the previous sections, we showed that there is good agreement between the calculated energies and experimental binding affinity data for *Eco* RV, ETS and *trp* repressor. To test further our ability to predict target specificity based on the calculated energies, we considered MAT-α2 and MCM1, which are also involved in determining mating type in yeast.[37] Analysis of the three-dimensional structure of their heterodimeric complex with DNA (1MNM)[38] showed the DNA to be bent to a significant degree, and the calculated *Z*-scores showed that the specificity is contributed by both the intermolecular and intramolecular readouts (Table 1). Zhong & Vershan[39] also investigated the binding specificity of the complex by systematically substituting all bases at each position of the target sequence (CATGTAATT). We predicted the effect of single base mutations within 1MNM on the affinity change based on the combined energies of the intermolecular and intramolecular readouts. Those data are presented in Figure 2(a); for comparison the experimentally derived repression activity values[39] are also presented. In this result, the lowest calculated energies agreed with the strongest experimental repression activities in eight of nine cases. Thus, the present method can predict quite well the correct activity among others. When we inspect the results more carefully, we are able to dissect individual contributions made by the intermolecular and intramolecular readouts. We found that energy changes due to intermolecular readout agree with the experimental data better than those of intramolecular readout. This result is reasonable, as there are many intermolecular contacts between amino acid residues and base-pairs, which is reflected in the higher *Z*-score for intermolecular readout than for intramolecular readout (−4.4 and −3.0, respectively). However, within the mutated region (CGTGTAAAT) of the MAT-α2/MCM1–DNAcomplex, the base-pair at position 2 is not in contact with any amino acid residues. At this position, the energy of intramolecular readout is the lowest for A, which is in agreement with the experimental data, whereas the energy of intermolecular readout is the lowest for G. Thus, at this position, intramolecular readout can explain the experimental observation better than intermolecular readout. Although the predicted patterns of energy changes are similar to the experimental data, the matching of detailed values is rather modest. Since transcriptional repression involves many steps, we may not be able to compare the binding energy to experimental repression data directly.

We also compared the calculations with the binding affinity data for λ repressor. We calculated the energy changes for all the single mutations of the consensus sequence of λ operator and compared them with the binding free energy changes measured using a filter-binding assay.[40] As shown in Figure 2(b), agreement with the experimental data was similar to that obtained with MAT-α2/MCM, i.e. the lowest energies were predicted well, but the agreement of the mutant energies was rather modest.

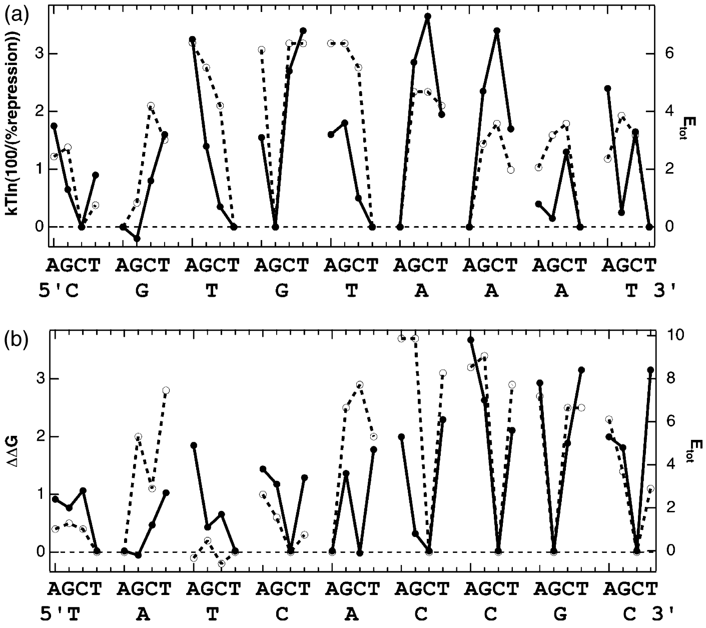

Figure 2. (a) Comparison of the calculated total energies (continuous line with filled circles) with experimental repression activity data (dotted line with open circles: scaled as kT ln(100/(% repression))) for base mutations within the MAT-α2/MCM1/DNA complex. The sequence used in the experiment[39] was CATGTAATT; the sequence in crystal structure (1MNM) was CGTGTAAAT, which is shown here. The total energy, E_{tot}, was calculated using the equation, $cE_{PD} + (1 - c)E_{DNA}$, with $c = 0.21$; $kT = 0.6$ kcal/mol and the energy for the experimental wild-type sequence was set to zero. (b) Comparison of calculated total energies (continuous line with filled circles) with experimental binding free energy data (dotted line with open circles) for base mutations in the λ repressor–DNA complex. Here, $c = 0.15$. The energy changes for all single mutations within the consensus sequence of OR1 operator were compared with the binding free energy changes measured using a filter-binding assay.[40] The energy for the consensus sequence was set to zero.

In general, the statistical potentials were derived from protein–DNA complexes that occur naturally and rarely contain unfavorable interactions between amino acid residues and bases. The statistical potential thus predicts correct targets (wild-type) against incorrect ones (mutants) quite well. On the other hand, the statistical potential may not be as good at predicting activity values among mutants.

Conclusions

In summary, we have quantified the specificities of the intermolecular and intramolecular readout mechanisms and compared them in various protein–DNA complexes. This has enabled us to show that, generally, both intermolecular and intramolecular readouts contribute to the specificity of protein–DNA recognition; that their relative contributions vary depending upon the proteins within the complex; that combination of the intermolecular and intramolecular readout energies leads to enhanced specificity; and that target sites for DNA binding proteins can be predicted based on analysis of the structure–specificity relationship.

The present method still has a number of limitations, however. For instance, the amount of available structural data remains limited. Consequently, statistical confidence is not very high for some structures. Some protein–DNA complexes have 0 or positive *Z*-scores for the intermolecular or intramolecular readout, which, respectively, suggest that the intermolecular or intramolecular readout mechanism provides no specificity in these structures. At present, we do not know whether this is true or due to an artifact. It is possible that the interactions involved in such structures are not well reflected in the structural dataset used for deriving the statistical potentials, or that unfavorable interactions are involved in those structures.

The role of water in determining specificity remains to be explored. The contribution made by water is partly reflected in our intermolecular readout, as some of the protein–DNA complexes used to derive statistical potentials contain water molecules at the interface. The statistical potentials thus include some effect of the mediating water molecules. However, the number of complexes containing such water molecules is still small, and the statistics for water within the protein–DNA complex are weak. Once additional high-resolution structures become available, especially within structures determined by neutron scattering, where water molecules are visible, we would like to develop the statistical potential for water molecules separately. We calculated the energy of DNA conformation based on simple harmonic functions. This approximation may break down if the distortion of DNA conformation is very severe. We also neglected the effect of protein conformation changes, but it may play an important role in protein–DNA recognition.

Despite these caveats, the present results provide new insight into the respective roles of intermolecular and intramolecular readout mechanisms in protein–DNA recognition. We anticipate that new structural data made available by the ongoing structural genomics projects will further enhance our understanding of the structure–specificity relationships that are the key determinants of protein–DNA recognition.

Methods

We considered a set of 62 non-redundant protein–DNA complexes; the interaction energies and *Z*-scores for intermolecular readout were calculated as described.[19,20] Briefly, the distant-dependent statistical potentials for the specific base-amino acid interactions were derived from the spatial distributions of C^{α} atoms of amino acid residues around a base. The potential function for each pairs of base and amino acid in a particular protein–DNA complex was summed to derive a total potential energy. By threading a set of random DNA sequences onto the template structure, we calculated the *Z*-score of the specific sequences against the random sequences, $(X - m)/\sigma$, where X is the energy of a particular sequence, m is the mean energy of 50,000 random DNA sequences, and σ is the standard deviation. The *Z*-score represents the specificity of the complex, with larger negative values corresponding to higher specificity.

To estimate the sequence-dependent DNA conformational energy, we mostly followed the approach described by Olson *et al.*[13] and added a self-consistent component (see below). The conformation energies were approximated using a harmonic function, $E_{\mathrm{DNA}} = 1/2\sum\sum f_{ij}\Delta\theta_i\Delta\theta_j$, in which θ_i represents the base-step parameters, and f_{ij} are the elastic force constants impeding deformation of the given base step $\Delta\theta_i = \theta_i - \theta_i^0$, in which θ_i^0 is the average base-step parameter. The base-step parameters used were shift, slide, rise, tilt, roll, and twist. The definitions of these parameters are given as in the literature.[41] Note that we only gave the parameters for the ten mutually distinct base steps, while the remaining parameters were derived from symmetry relations.[41] The unknown parameters f_{ij} and θ_i^0 were determined by statistical analysis of the same 62 non-redundant protein–DNA complexes. Setting up a co-variance matrix from observed distributions of θ_i thus refers to an effective inverse harmonic force-constant matrix. Inversion of this matrix transformed it to a force-constant matrix in the original coordinate basis. Exclusion of data anomalies plays an important role in this procedure, and we followed Olson's three standard deviation exclusion procedure: all parameters of a base step for which one parameter exceeded three standard deviations were removed from the data set. This procedure requires a re-calculation of averages and standard deviations before setting up the final co-variance matrix. Since there was no way of knowing about the behaviour of the remaining dataset we repeated the data culling procedure until no more data were assigned to the cut-off value. Typically three iterations were necessary to make the procedure self-consistent. Only then was the final force field calculated. The total intramolecular energy of a given complex structure was calculated as the sum of all the base steps. We assigned the energy corresponding to the threshold value when any parameter exceeded three standard deviations.

We carried out jack-knife and bootstrap tests to assess the statistical confidence in the *Z*-score calculations. To remove the effect of self-contributions, we always removed the self from the original dataset of complex structures when calculating its *Z*-score. We then examined the *Z*-score further by removing one additional randomly selected structure from the dataset and repeated this procedure. We found that the *Z*-scores were stable against these treatments, indicating that our dataset of protein–DNA complexes provides adequate information for a generally valid structure-based potential. To calculate the bootstrap standard errors, we prepared a set of 61 randomly selected complex structures in which the self was removed but duplications of the same structure were allowed. We created 200 such replications and calculated the standard errors. These standard errors are shown in the parentheses following *Z*-scores in Table 1.

Acknowledgements

This work was supported in part by a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports and Technology. Most of the work was done in the RIKEN Institute. We thank RIKEN Institute for support.

References

1. Mandel-Gutfreund, Y. & Margalit, H. (1998). Quantitative parameters for amino acid–base interaction: implications for prediction of protein–DNA binding sites. *Nucl. Acids Res.* **26**, 2306–2312.
2. Mandel-Gutfreund, Y., Margalit, H., Jernigan, R. L. & Zhurkin, V. B. (1998). A role for CH···O interactions in protein–DNA recognition. *J. Mol. Biol.* **277**, 1129–1140.
3. Nadassy, K., Wodak, S. J. & Janin, J. (1999). Structural features of protein–nucleic acid recognition sites. *Biochemistry*, **38**, 1999–2017.
4. Jones, S., van Heyningen, P., Berman, H. M. & Thornton, J. M. (1999). Protein–DNA interactions: a structural analysis. *J. Mol. Biol.* **287**, 877–896.
5. Luscombe, N. M. & Thornton, J. M. (2002). Protein–DNA interactions: amino acid conservation and the effects of mutations on binding specificity. *J. Mol. Biol.* **320**, 991–1009.
6. Matthews, B. W. (1988). Protein–DNA interaction. No code for recognition. *Nature*, **335**, 294–295.
7. Pabo, C. O. & Nekludova, L. (2000). Geometric analysis and comparison of protein–DNA interfaces: why is there no simple code for recognition? *J. Mol. Biol.* **301**, 597–624.
8. Szymczyna, B. R. & Arrowsmith, C. H. (2000). DNA binding specificity studies of four ETS proteins support an intramolecular readout mechanism of protein–DNA recognition. *J. Biol. Chem.* **275**, 28363–28370.
9. Schwabe, J. W. (1997). The role of water in protein–DNA interactions. *Curr. Opin. Struct. Biol.* **7**, 126–134.
10. Harrington, R. E. & Winicov, I. (1994). New concepts in protein–DNA recognition: sequence-directed DNA bending and flexibility. *Prog. Nucl. Acid Res. Mol. Biol.* **47**, 195–270.
11. Hogan, M. E. & Austin, R. H. (1987). Importance of DNA stiffness in protein–DNA binding specificity. *Nature*, **329**, 263–266.
12. Sarai, A., Mazur, J., Nussinov, R. & Jernigan, R. L. (1989). Sequence dependence of DNA conformational flexibility. *Biochemistry*, **28**, 7842–7849.
13. Olson, W. K., Gorin, A. A., Lu, X. J., Hock, L. M. & Zhurkin, V. B. (1998). DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proc. Natl Acad. Sci. USA*, **95**, 11163–11168.
14. Frech, K., Quandt, K. & Werner, T. (1997). Finding protein-binding sites in DNA sequences: the next generation. *Trends Biochem. Sci.* **22**, 103–104.
15. Karas, H., Knuppel, R., Schulz, W., Sklenar, H. & Wingender, E. (1996). Combining structural analysis of DNA with search routines for the detection of transcription regulatory elements. *Comput. Appl. Biosci.* **12**, 441–446.
16. Mishmar, D., Rahat, A., Scherer, S. W., Nyakatura, G., Hinzmann, B., Kohwi, Y. *et al.* (1998). Molecular characterization of a common fragile site (FRA7H) on human chromosome 7 by the cloning of a simian virus 40 integration site. *Proc. Natl Acad. Sci. USA.* **95**, 8141–8146.
17. Ponomarenko, J. V., Ponomarenko, M. P., Frolov, A. S., Vorobyev, D. G., Overton, G. C. & Kolchanov, N. A. (1999). Conformational and physicochemical DNA features specific for transcription factor binding sites. *Bioinformatics*, **15**, 654–668.
18. Liu, R., Blackwell, T. W. & States, D. J. (2001). Conformational model for binding site recognition by the *E. coli* MetJ transcription factor. *Bioinformatics*, **17**, 622–633.
19. Kono, H. & Sarai, A. (1999). Structure-based prediction of DNA target sites by regulatory proteins. *Proteins: Struct. Funct. Genet.* **35**, 114–131.
20. Selvaraj, S., Kono, H. & Sarai, A. (2002). Specificity of protein–DNA recognition revealed by structure-based potentials: symmetric/asymmetric and cognate/non-cognate binding. *J. Mol. Biol.* **322**, 907–915.
21. Sarai, A., Selvaraj, S., Gromiha, M. M., Siebers, J. G., Prabakaran, P. & Kono, H. (2001). Target prediction of transcriptor factors: refinement of structure-based method. *Genome Inform.* **12**, 384–385.
22. Steffen, N. R., Murphy, S. D., Tolleri, L., Hatfield, G. W. & Lathrop, R. H. (2002). DNA sequence and structure: direct and indirect recognition in protein–DNA binding. *Bioinformatics*, **18**, S22–S30.
23. Choo, Y. & Klug, A. (1994). Toward a code for the interactions of zinc fingers with DNA: selection of randomized fingers displayed on phage. *Proc. Natl Acad. Sci. USA*, **91**, 11163–11167.
24. Choo, Y. & Klug, A. (1994). Selction of DNA binding sites for zinc fingers using rationally randomized DNA reveals coded interactions. *Proc. Natl Acad. Sci. USA*, **91**, 11168–11172.
25. Desjarlais, J. R. & Berg, J. M. (1993). Use of a zinc-finger consensus sequence framework and specificity rules to design specific DNA binding proteins. *Proc. Natl Acad. Sci. USA*, **90**, 2256–2260.
26. Desjarlais, J. R. & Berg, J. M. (1994). Length-encoded multiplex binding site determination: application to

zinc finger proteins. *Proc. Natl Acad. Sci. USA*, **91**, 11099–11103.

27. Kim, C. A. & Berg, J. M. (1996). A 2.2 Å resolution crystal structure of a designed zinc finger protein bound to DNA. *Nature Struct. Biol.* **3**, 940–945.
28. Winkler, F. K., Banner, D. W., Oefner, C., Tsernoglou, D., Brown, R. S., Heathman, S. P. *et al.* (1993). The crystal structure of EcoRV endonuclease and of its complexes with cognate and non-cognate DNA fragments. *EMBO J.* **12**, 1781–1795.
29. Horton, N. C. & Perona, J. J. (1998). Role of protein-induced bending in the specificity of DNA recognition: crystal structure of *Eco*RV endonuclease complexed with d(AAAGAT) + d(ATCTT). *J. Mol. Biol.* **277**, 779–787.
30. Martin, A. M., Sam, M. D., Reich, N. O. & Perona, J. J. (1999). Structural and energetic origins of indirect readout in site-specific DNA cleavage by a restriction endonuclease. *Nature Struct. Biol.* **6**, 269–277.
31. Rice, P. A., Yang, S., Mizuuchi, K. & Nash, H. A. (1996). Crystal structure of an IHF–DNA complex: a protein-induced DNA U-turn. *Cell*, **87**, 1295–1306.
32. Gromiha, M. M., Munteanu, M. G., Simon, I. & Pongor, S. (1997). The role of DNA bending in Cro protein–DNA interactions. *Biophys. Chem.* **69**, 153–160.
33. Nikolov, D. B., Chen, H., Halay, E. D., Hoffman, A., Roeder, R. G. & Burley, S. K. (1996). Crystal structure of a human TATA box-binding protein/TATA element complex. *Proc. Natl Acad. Sci. USA*, **93**, 4862–4867.
34. Otwinowski, Z., Schevitz, R. W., Zhang, R. G., Lawson, C. L., Joachimiak, A., Marmorstein, R. Q. *et al.* (1988). Crystal structure of trp repressor/operator complex at atomic resolution. *Nature*, **335**, 321–329.
35. Joachimiak, A., Haran, T. E. & Sigler, P. B. (1994). Mutagenesis supports water mediated recognition in the trp repressor–operator system. *EMBO J.* **13**, 367–372.
36. Grillo, A. O., Brown, M. P. & Royer, C. A. (1999). Probing the physical basis for trp repressor–operator recognition. *J. Mol. Biol.* **287**, 539–554.
37. Wilson, K. L. & Herskowitz, I. (1984). Negative regulation of STE6 gene expression by the alpha 2 product of *Saccharomyces cerevisiae*. *Mol. Cell Biol.* **4**, 2420–2427.
38. Tan, S. & Richmond, T. J. (1998). Crystal structure of the yeast MATalpha2/MCM1/DNA ternary complex. *Nature*, **391**, 660–666.
39. Zhong, H. & Vershon, A. K. (1997). The yeast homeodomain protein MATalpha2 shows extended DNA binding specificity in complex with Mcm1. *J. Biol. Chem.* **272**, 8402–8409.
40. Sarai, A. & Takeda, Y. (1989). Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically. *Proc. Natl Acad. Sci. USA*, **86**, 6513–6517.
41. Olson, W. K., Bansal, M., Burley, S. K., Dickerson, R. E., Gerstein, M., Harvey, S. C. *et al.* (2001). A standard reference frame for the description of nucleic acid base-pair geometry. *J. Mol. Biol.* **313**, 229–237.

Edited by Sir A. Klug

(Received 12 June 2003; received in revised form 16 January 2004; accepted 21 January 2004)