BIOCHEMISTRY
including biophysical chemistry & molecular biology

Article

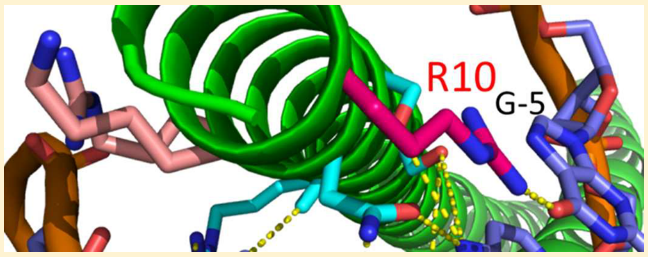
pubs.acs.org/biochemistry

Effect of Flanking Bases on the DNA Specificity of EmBP-1

Antonia T. De Jong*

Department of Chemistry, University of Toronto, Toronto, Ontario, Canada M5S 3H6

Ⓢ *Supporting Information*



ABSTRACT: EmBP-1 is a basic region leucine zipper (bZIP) protein found in many types of plants. In general, plant bZIP proteins bind selectively to DNA sequences containing ACGT core sequences with different immediate flanking nucleotides preferred by different proteins. I report that the distant flanking sequence also has a strong effect on the preference of EmBP-1 for internal bases and determine the residue governing this effect. EmBP-1 binds selectively to the 10 bp gcG-box palindrome **G**CC*ACGT*GG**C** 18-fold more tightly than the gcC-box **G**TG*ACGT*CA**C**, but when the outer flanking G/C residues were changed to A/T (i.e., **A**CC*ACGT*GG**T** and **A**TG*ACGT*CA**T**), an only 1.2-fold preference for G-box binding was observed. Analysis of a series of single-residue alanine mutants of EmBP-1 revealed that this effect is mediated by arginine 10. Mutation of this residue to alanine greatly reduces the affinity for the gcG-box while leaving the affinity for other sequences relatively unchanged. Partial retention of G-box specificity upon mutation of R10 to lysine indicates that the effect is reliant on the basic nature of the residue. Additional studies with other EmBP-1 protein mutants and with oligonucleotides containing the T/A and C/G flanking sequences demonstrate the complexity of the protein−DNA interaction and demonstrate that the mechanism of sequence selective DNA binding is highly dependent on the flanking sequence.

Basic region leucine zipper (bZIP) proteins make up a class of transcription factors found in many eukaryotes. They are structurally very simple, with each monomer in the functional dimer forming a single continuous α-helix and binding to DNA in a manner comparable to a pair of forceps.[1−6] At the C-terminus, residues of the leucine zipper form a coiled coil dimerization interface, while the N-terminal basic regions bind within the major groove of the DNA and are the primary determinants of sequence specificity.[7−10]

In plants, bZIP proteins mainly bind to promoters containing the ACGT core sequence.[11,12] In plant nomenclature, the base following the final thymine of this core region is used to name the DNA sequence.[11] Thus, C*ACGT***G** is designated as the G-box, while G*ACGT***C** is known as the C-box. EmBP-1 is one such plant bZIP, originally isolated from wheat but with analogues in many other plant species. It was isolated from wheat embryo cDNA libraries by phage selection with the Abscisic acid response element (ABRE) that contains two G-box sequences, and it may be involved in responses to the plant hormone abscisic acid.[13] As many similar ACGT core elements are found in the *cis*-acting DNA elements that regulate a wide range of genes in plants,[11] the question of how specific elements are recognized by different bZIP proteins is important.

In a study by Izawa et al.,[12] EmBP-1 exhibited the strongest G-box preference of the 10 plant bZIPs studied, with a 100-fold preference for the G-box sequence GCC*ACGT***G**GC over the C-box sequence GTG*ACGT***C**AC. In a subsequent study, Niu et al.[14] measured a 14-fold preference with the same set of sequences. Outer −5/5 flanking residue effects were demonstrated by Izawa et al.[12] and Foster et al.[11] with gcG-box and taG-box (see Figure 1 for naming and numbering) among the tightly bound sequences. These studies as well as binding site selections by Niu et al.[15] provide extensive knowledge of the high-affinity binding sequences for EmBP-1. However, aside from the gcG-box and gcC-box, only qualitative, single-point, binding assays were performed, and binding to the other sequences examined in this work was near or below the detection limit of the assay; therefore, detailed analysis of the binding preference was not possible. To further explore this interesting flanking sequence effect, systematic, quantitative studies were conducted. When the same 10 bp sequence used in previous studies was examined (here designated gcG-box and gcC-box), an 18-fold preference was observed, in agreement with past data. However, upon comparison of binding to the atG-box to atC-box, virtually no sequence preference was detected. While the gcG-box was bound with low nanomolar affinity by wild-type EmBP-1, the gcC-box, atG-box, and atC-box were all bound with similar low micromolar affinity despite the fact that gcC-box differs by only two residues from the optimal sequence while the atC-box differs at 6 of 10 bases.

While the flanking effects quantified here had been suggested by previous studies, these papers did not attempt to relate the sequence preference to specific protein residues. To explore the origins of this flanking sequence preference, I created various single-residue alanine mutants. Unfortunately, no crystal

Received: October 16, 2012
Revised: December 9, 2012
Published: January 10, 2013

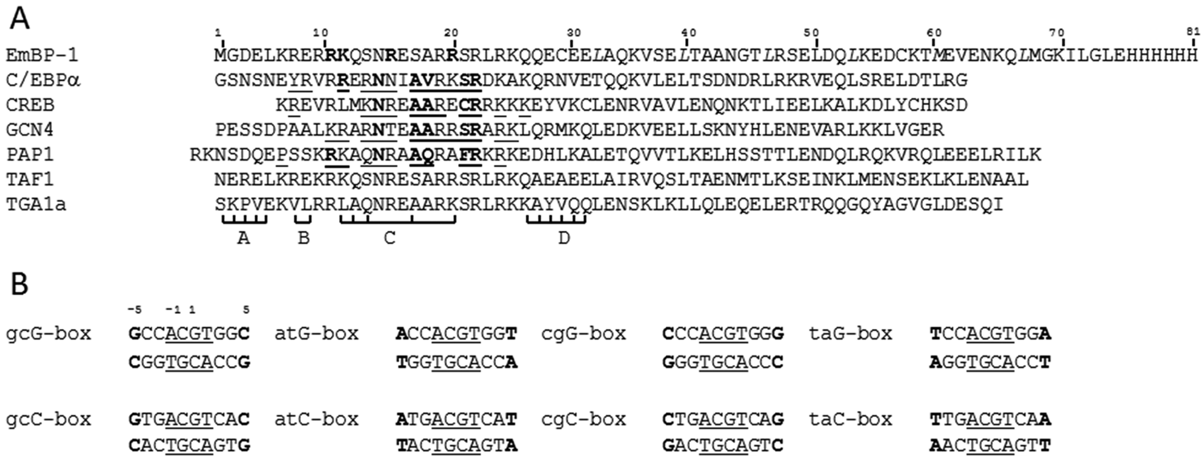

A
1
10
20
30
40
50
60
70
81
EmBP-1 MGDELKRERRKQSNRESARRSRLRKQQECEELAQKVSELTAANGTLRSELDQLKEDCKTMEVENKQLMGKILGLEHHHHHH
C/EBPα GSNSNEYRVRRERNNIAVRKSRDKAKQRNVETQQKVLELTSDNDRLRKRVEQLSRELDTLRG
CREB KREVRLMKNREAARECRRKKKEYVKCLENRVAVLENQNKTLIEELKALKDLYCHKSD
GCN4 PESSDPAALKRARNTEAARRSRARKLQRMKQLEDKVEELLSKNYHLENEVARLKKLVGER
PAP1 RKNSDQEPSSKRKAQNRAAQRAFRKRKEDHLKALETQVVTLKELHSSTTLENDQLRQKVRQLEEELRILK
TAF1 NERELKREKRKQSNRESARRSRLRKQAEAEELAIRVQSLTAENMTLKSEINKLMENSEKLKLENAAL
TGA1a SKPVEKVLRRLAQNREAARKSRLRKKAYVQQLENSKLKLLQLEQELERTRQQGQYAGVGLDESQI
A
B
C
D
B
−5 −1 1 5
gcG-box GCCACGTGGC CGGTGCACCG
atG-box ACCACGTGGT TGGTGCACCA
cgG-box CCCACGTGGG GGGTGCACCC
taG-box TCCACGTGGA AGGTGCACCT
gcC-box GTGACGTCAC CACTGCAGTG
atC-box ATGACGTCAT TACTGCAGTA
cgC-box CTGACGTCAG GACTGCAGTC
taC-box TTGACGTCAA AACTGCAGTT

Figure 1. Protein and DNA sequences. (A) Sequence alignment of protein EmBP-1 (GenBank entry CAA67298.1) and bZIP proteins with known structures. For the sake of clarity, throughout the text the numbering of residues for all proteins follows the numbering for the bZIP EmBP-1 construct. Residues in EmBP-1 that were mutated are shown in bold, and d residues of the leucine zipper heptad repeat are italicized. Mass spectrometry indicates that the N-terminal methionine is cleaved from the expressed protein. For C/EBPα[5], CREB[6], GCN4[3], and PAP1[1], residues that contact the DNA are underlined. Residues that make contacts with DNA bases are also shown in bold. Regions for segment swap experiments by Niu et al.[15] are indicated underneath the sequence of TGA1a. (B) DNA palindromes studied. The ACGT core of each sequence is underlined. Flanking bases mentioned in naming scheme are shown in bold.

structure exists for a bZIP protein bound to any G-box DNA sequence, but structures for proteins recognizing other closely related sequences are available and were consulted to determine residues likely to mediate this effect.[1−6] In PAP-1, Arg10 makes a hydrogen bond with O6 of G-5,[1] suggesting the same contact in EmBP-1 may be responsible for the observed influence of the outer flanking base. Upon mutation of this residue in EmBP-1 to alanine, it was found to significantly reduce the affinity for the gcG-box while having an only minor effect on binding to other sequences. Mutation to lysine partially maintained the wild-type binding profile, indicating that a basic residue at residue 10 is critical for flanking sequence specificity. Studies with other single-residue alanine mutants, K11A and R15A, with the DNA sequences described above demonstrated that the flanking bases also influence interactions at the core of the sequence. Additionally, binding to the remaining T/A- and C/G-flanked sequences demonstrated that the full 10 bp sequence contributes to the protein−DNA complex.

EXPERIMENTAL PROCEDURES

Creation of Genes for EmBP and Its Mutants. The gene for wild-type EmBP-1 (GenBank entry CAA67298.1, residues 69−140) in pET28a(+) (Novagen, Mississauga, ON) was generously provided by J. Shin (Department of Chemistry, University of Toronto). This gene contained *Nco*I and *Sac*I restriction sites flanking the basic region, allowing for excision and replacement of the basic region. The sequences for the K11A and R15A basic regions were created by Klenow assembly from pairs of oligos. These fragments were digested and ligated between the *Nco*I and *Sac*I sites of the vector described above. R10A, R10K, and R20A were created by the QuikChange polymerase chain reaction mutagenesis method (Stratagene) with pET28a(+)/EmBPwt as the template. Recombinant plasmids were transformed into *Escherichia coli* strain NEB Turbo (New England Biolabs) and isolated, and then the sequences were confirmed by DNA sequencing (The Centre for Applied Genomics, The Hospital for Sick Children, Toronto, ON).

Expression and Purification of EmBP and Its Mutants. Plasmids for WT EmBP (EmBPwt) and its mutants were transformed into *E. coli* BL21 Star(DE3) cells. Protein expression was induced by 1 mM IPTG at an OD_{600} of 0.6, and cells were harvested after having grown for 2 h at 37 °C. Proteins were first purified by TALON metal ion affinity chromatography (Clontech) and then further purified by reversed phase HPLC with a C8 column (Agilent). Details of the purification are given in the Supporting Information. Protein concentrations were determined using the Bradford assay with BSA as a standard. Concentrations were later corrected by quantitative amino acid analysis (Advanced Protein Technology Centre, The Hospital for Sick Children). WT and R10A were corrected by direct analysis, while a conversion factor of 10.5, generated by averaging the conversion factors for three individual samples, was used for the remainder of the samples.

Electrophoretic Mobility Shift Assay. Although reactants are technically not at equilibrium during EMSAs, previous studies have successfully used EMSAs for determination of dissociation constants for bZIP proteins.[16−18] Fluorescein-labeled DNA probes with gcG-box sequence 5′-TGCAGGA**G**CCACGTGG**C**CGAAGGTT, atG-box sequence 5′-TGCAGGA**A**CCACGTGG**T**GAAGGTT, gcC-box sequence 5′-TGCAGGA**G**TGACGTCA**C**GAAGGTT, and atC-box sequence 5′-TGCAGGA**A**TGACGTCA**T**GAAGGTT were used for EMSA experiments. Labeled and complementary, unlabeled oligonucleotides were commercially synthesized (Operon Biotechnologies, Huntsville, AL). The 6-carboxyfluorescein fluorophore (6-FAM) was incorporated at the 5′ end of the labeled strand, and the oligonucleotide was purified by HPLC. Strands were mixed with a 1:2 ratio of labeled to unlabeled oligonucleotide and annealed by being heated at 80 °C and then slowly cooled to room temperature. Annealed DNA was then used without further purification. Samples containing 2 nM DNA and protein were prepared in buffer (adapted from ref 19) containing 24 mM Tris (pH 7.9), 70 mM KCl, 0.7 mM EDTA, 15 mM $MgCl_2$, 2 mM DTT, 500 μg/mL sheared salmon DNA, 100 μg/mL BSA, and 24% glycerol and were

Table 1. Dissociation Constants Determined by EMSAs

	K_D (nM)[a]				C/G ratio[b]	
protein	gcG-box (GCCACGTGGC)	gcC-box (GTGACGTCAC)	atG-box (ACCACGTGGT)	atC-box (ATGACGTCAT)	GC	AT
WT	65 ± 12	1100 ± 200	940 ± 50	1100 ± 100	17	1.2
R10A	1570 ± 90	1800 ± 370	2500 ± 320	1600 ± 180	1.1	0.65
R10K	97 ± 4	1100 ± 200	510 ± 35	650 ± 45	11	1.3
K11A	76 ± 9	9280 ± 60	4800 ± 1200	10100 ± 120	120	2.0
R15A	111 ± 7	860 ± 40	6900 ± 200	1170 ± 90	7.7	0.17
R20A	48 ± 5	4200 ± 1200			87	

[a] K_D values for WT, R10A, and R10K are the average of fits to three independent titrations ± the standard deviation. K_D values for the remaining proteins are the average of fits to two independent titrations ± the standard deviation. [b] The C/G ratio is the ratio of the dissociation constant for the C-box divided by the value for the G-box with the indicated flanking sequence.

incubated at room temperature for 15 min. Nondenaturing, 8% polyacrylamide, 1:37.5 cross-linked gels were prerun in 1× TAE at 300 V and 4 °C for 1 h, and then samples were loaded and run at 300 V and 4 °C for 105 min. Gels were then imaged using a Molecular Imager PharosFX instrument (Bio-Rad Laboratories) using the FITC application (excitation at 488 nm, emission unfiltered). Band percent intensities were determined for each lane independently using Image Lab version 4.0 (Bio-Rad Laboratories). Dissociation constants were determined by plotting fraction bound versus monomeric concentration and fitting the data to the following equation (eq 1) in Origin 8 Pro:

$$\Theta = \Theta_{\min} + (\Theta_{\max} - \Theta_{\min})[[\mathrm{M}]^2/(K_D{}^2 + [\mathrm{M}]^2)] \qquad (1)$$

where Θ is the fraction bound, [M] is the monomeric protein concentration, and K_D corresponds to the apparent dissociation constant of binding of protein to DNA. It should be noted that this apparent dissociation constant likely includes the dimerization of the protein as dimerization constants for isolated bZIP proteins are typically in the micromolar range.[20−22] Assays were performed in triplicate for the wild type, R10A, and R10K and in duplicate for K11A, R15A, and R20A. The K_D values are the average of separate fits (R values of >0.95) ± the standard deviation.

Competition EMSA. Fluorescein-labeled double-stranded gcC-box DNA was prepared as described above. Oligonucleotides for the preparation of unlabeled competitor DNA were commercially synthesized (IDT, Coralville, IA). The 24 bp sequences contained the 10 bp sequences taG-box, taC-box, cgG-box, and cgC-box (Figure 1) with the same flanking sequences as the labeled probes. Complementary strands were mixed at a 1:1 ratio and annealed by being heated at 80 °C and then cooled at a rate of 0.1 °C/min to 10 °C. Annealed DNA was then used without further purification. Samples containing 2 nM FAM-gcC-box and 0−150 μM competitor DNA were prepared in the same buffer that was used in the direct EMSA. After approximately 30 min, 6.5 μM protein was added to the reaction mixture followed by a second 30 min incubation at room temperature. Gel running and imaging were performed as described above. Data for taC-box, cgC-box, and cgG-box were fit to an equation based on the work of Metallo et al.[23] (eq 2):

$$\begin{aligned}\Theta = \{&K_C{}^2K_L + 4K_LK_C[\mathrm{C}][\mathrm{M}] + 2K_C{}^2[\mathrm{M}]^2 \\ &+ [K_C{}^2K_L{}^2(K_C{}^2 + 8K_C[\mathrm{M}][\mathrm{C}])]^{1/2}\} \\ &/[2(4K_L{}^2[\mathrm{C}]^2 + K_LK_C{}^2 + 4K_LK_C[\mathrm{M}][\mathrm{C}] \\ &+ K_C{}^2[\mathrm{M}]^2)]\end{aligned} \qquad (2)$$

where Θ is the fraction bound, [M] is the monomeric concentration of the protein in molar, [C] is the concentration of competitor DNA, K_L is the previously determined dissociation constant for labeled gcC-box DNA, and K_C is the dissociation constant for the competitor DNA. Because of differences in the definitions of dissociation constants between the reference and this paper, K_D values reported here are the square root of those used in the equation. The full derivation is included in the Supporting Information. For cgG-box, cgC-box, and taC-box, values presented are the average of fits for three independent titrations ± the standard deviation, and all fits have R values of >0.96. However, the derivation of this equation is based on the assumption that competitor DNA is in excess of protein concentration and thus the total DNA concentration is far in excess of the concentration of bound DNA. In the case of taG-box binding, the affinity for taG-box is apparently tighter than that for gcC-box, and this assumption is no longer valid with the protein concentrations necessary for initial full binding of the labeled probe. Thus, for these data sets, DynaFit,[24] which does not utilize specific equations but performs nonlinear least-squares regression based on chemical equilibria input into the program, was used to fit data with a custom DynaFit script (see the Supporting Information for the full script). Data sets were fit separately, and values presented are the average of fits for three independent titrations ± the standard deviation. As different fitting methods were used, K_D ratios for T/A-flanked sequences may not be accurate and are presented only as a guide in this case.

Circular Dichroism. Samples contained 10 μM protein monomer, 15.08 mM Na_2HPO_4, 4.92 mM KH_2PO_4 (pH 7.4), 50 mM NaCl, and 10 μM DNA where appropriate. CD data were collected on an Olis spectrometer using a cylindrical quartz cell with a 1 mm path length at 20 °C. Spectra were acquired between 190 and 260 nm in 1 nm increments with a sampling time of 1 s. Each spectrum is the average of three scans, and the buffer control, containing DNA where appropriate, was subtracted from each protein spectrum. Curves were then smoothed in Origin 8 Pro. Mean residue ellipticities are presented. The protein helix content was calculated from Θ_{222} using the method described by Chen et al.[25] assuming only helical structure was present. Data for WT protein with taG-box, gcG-box, and atG-box were collected from 190 to 300 nm to probe differences in DNA structure around 280 nm. Raw data are presented in Figure S1 of the Supporting Information.

■ RESULTS

Sequence Specificity Is Dependent on Flanking Sequence. EmBP-1 is a bZIP protein that has been shown to bind specifically to the G-box.[11,12,14] When trying to reproduce G-box selectivity data from Izawa et al.[12] and Niu et al.[14] with atG-box and atC-box probes (Figure 1), I could not detect significant selectivity by an EMSA despite varying the solution conditions and incubation times (data not shown). Upon examination of the DNA sequences used in the previous papers, it was noted that both studies utilized the 10 bp sequences designated here as the gcG-box and gcC-box with probes differing outside of this core region. Thus, new probes were employed with the same 10 bp core region to test the effect of outer flanking sequence on selectivity. EMSA titrations revealed that the G-box selectivity was dependent on the outer flanking sequence with an 18-fold preference for G-box with G/C flanking sequences but an only 1.2-fold G-box preference with A/T flanking sequences (Table 1 and Figures 2 and 3A).

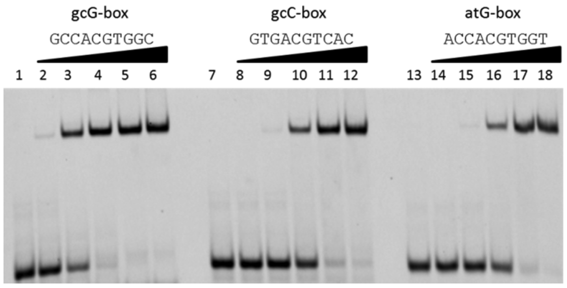


Figure 2. Sample EMSA gel for EmBPwt. All samples contained 2 nM FAM-DNA of the indicated sequence. Lanes 1, 7, and 13 contained no protein. Lanes 2−6, 8−12, and 14−18 contained 10−10000 nM EmBPwt.

This selectivity is due to the high affinity of binding to the gcG-box (65 ± 12 nM), while low affinities of 1100 ± 200, 940 ± 50, and 1100 ± 100 nM are measured for the gcC-box, atG-box, and atC-box, respectively.

The gcG-Box Sequence Preference Is Dependent on the Basic Residue at Position 10. Several crystal structures were examined to predict which residues might be responsible for the observed flanking sequence effect. In the PAP1 structure, a direct contact is made between Arg82 and O6 of G-5 of the DNA palindrome.[1] This residue corresponds to Arg10 of the EmBP-1 bZIP construct, and thus, the R10A mutant was created to test the involvement of this residue in sequence specificity. Upon mutation, sequence specificity was lost with dissociation constants measured for all sequences falling within a 1.6-fold range. The affinity for the gcG-box was greatly reduced from 65 ± 12 to 1570 ± 90 nM, while the affinities for the other three sequences were only slightly diminished (Table 1 and Figure 3B).

To test whether the interaction between R10 and DNA is an electrostatic contact, the amino acid was changed to a lysine. In contrast to the R10A mutation, which eliminated sequence preference among the sequences that were studied, mutation of the arginine to a lysine maintained sequence specificity (Table 1 and Figure 3C). The level of binding to the gcG-box was slightly reduced to 97 ± 4 nM, while gcC-box binding was within error of that for WT, resulting in a reduction from an 18-fold G-box preference to an 11-fold preference. With the A/T flanking sequence, R10K binding affinities were slightly tighter than those for WT but the C/G ratio was maintained (1.2 for WT and 1.3 for R10K). This result supports the hypothesis that the effect of Arg10 on binding affinities is related to the basic nature of its side chain.

Competition Experiments Demonstrate R10 Interaction Is Critical for only gcG-Box Binding. To further examine the effect of flanking bases, after data had been collected for the G/C- and A/T-flanked sequences by direct EMSA experiments, binding to the four remaining palindromic sequences, cgG-box, cgC-box, taG-box, and taC-box (Figure 1), was assayed by a competition EMSA for both EmBPwt and R10A (Table 2 and Figure 4). Experiments with the C/G-flanked sequences produced results similar to those with A/T-flanked sequences with an only minor preference detected for the cgG-box and binding affinity ratios unchanged from that of WT by the R10A mutation (Table 2). The T/A-flanked sequences, on the other hand, exhibited behavior different from that of all other studied sequences. Like the G/C-flanked sequences, a preference is detected for G-box binding with the WT protein. However, unlike the G/C-flanked sequences, this preference in not eliminated by the R10A mutation.

Circular Dichroism Spectroscopy Suggests Equal Folding for High- and Low-Affinity Sequences. Circular dichroism (CD) spectra were measured for both EmBPwt and R10A alone in solution and in the presence of DNA (Figure 5). The concentrations used for CD are much higher than the dissociation constants, and thus, all proteins should be fully bound to DNA. As with other bZIP proteins, both EmBPwt and R10A are largely unfolded in solution in the absence of DNA. R10A is more folded (37% helical structure) than WT (20% helical structure), which may be due to the higher helical

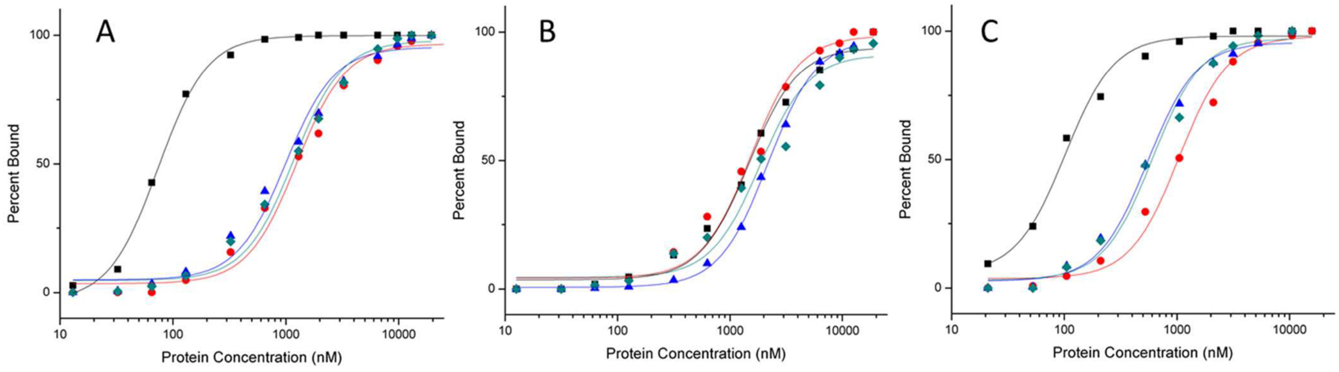


Figure 3. EMSA binding isotherms. Sample binding isotherms for (A) WT EmBP bZIP, (B) EmBP R10A, and (C) EmBP R10K with gcG-box (black squares), gcC-box (red circles), atG-box (blue triangles), and atC-box (green diamonds). Each data set is the result of a single titration. Curves are fits to eq 1 generated in Origin 8 Pro.

Table 2. Competition Dissociation Constants

	K_d (nM)[a]				C/G ratio[b]	
protein	cgG-box (**C**CCACGTGG**G**)	cgC-box (**C**TGACGTCA**G**)	taG-box (**T**CCACGTGG**A**)	taC-box (**T**TGACGTCA**A**)	CG	TA
WT	1500 ± 70	2200 ± 200	120 ± 30[c]	5500 ± 800	1.5	46[d]
R10A	1720 ± 60	2800 ± 200	370 ± 50[c]	8000 ± 500	1.6	22[d]

[a] Values are the average of fits to three independent titrations ± the standard deviation. [b] The C/G ratio is the ratio of the dissociation constant for the C-box divided by the value for the G-box with the indicated flanking sequence. [c] Values were determined using Dynafit. The remaining values were determined by fitting to eq 2 in Origin 8 Pro. [d] Because of the discrepancy in methods for determining dissociation constants, these ratios may be inaccurate and are presented only as a guide.

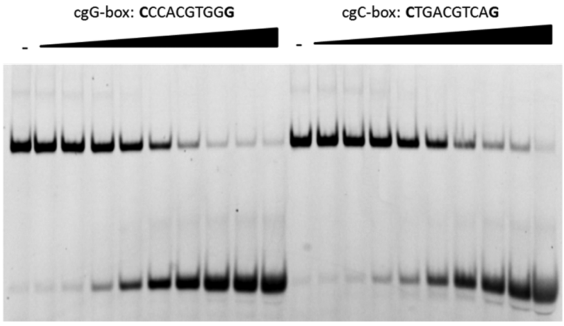


Figure 4. Competition EMSA. Representative competition EMSA gel. Reaction mixtures contain 6.5 μM EmBPwt, 2 nM FAM-labeled gcC-box, and 0−150 μM unlabeled competitor DNA containing the cgG-box or cgC-box as indicated.

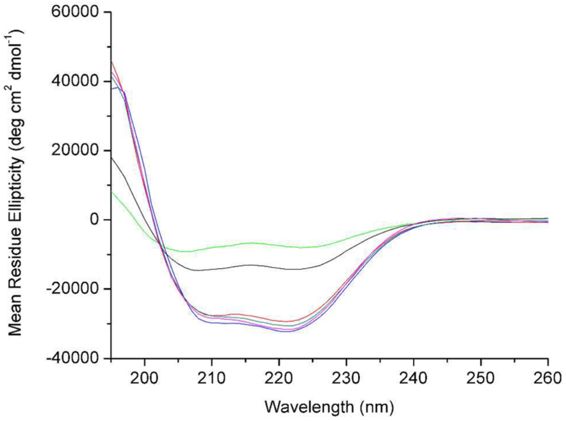


Figure 5. Circular dichroism spectra of EmBPwt and R10A. Spectra for EmBPwt without DNA (green), with gcG-box (blue), and with atG-box (teal) and R10A without DNA (black), with gcG-box (red), and with atG-box (pink) are shown. Samples contained 10 μM protein monomer, 15.08 mM Na_2HPO_4, 4.92 mM KH_2PO_4 (pH 7.4), 50 mM NaCl, and 10 μM DNA where indicated. Each spectrum is the average of three scans, and curves were smoothed using Origin 8 Pro. The buffer control, containing DNA where appropriate, was subtracted from each protein spectrum. Mean residue ellipticities are presented.

propensity of alanine versus arginine.[26] Despite the differences in helicity between the unbound proteins, WT and R10A exhibit nearly identical helicity upon the addition of DNA. Under these conditions, the degree of helix induction is not dependent on the sequence of the DNA, with WT being 83% helical when incubated with the high-affinity gcG-box DNA and 80% helical upon addition of low-affinity atG-box DNA. Similarly, R10A displayed 76 and 82% helical structure with gcG-box and atG-box DNA, respectively, although both sequences are bound with low affinity by the mutant protein.

One possible explanation for the high-affinity binding between EmBPwt and gcG-box is that it is the only protein−DNA pair that allows proper folding of the basic region, and that steric clashes between the protein and DNA interfere with the folding of the basic regions in the other protein−DNA complexes. This does not seem to be the case because all protein−DNA mixtures studied produce similar spectra with approximately 80% helical character despite affinities ranging from 65 nM to 2.5 μM. Overall, sequence discrimination does not appear to take place at the gross structural level but is likely due to placement of side chains. This observation is consistent with studies of proteins from the closely related bHLH and bHLHZ families. For bHLHZ proteins USF and Max and bHLH protein Pho4, it has been proposed that nonspecific association of the protein−DNA complex is followed by a slow conformational rearrangement involving protein side chain adjustments when bound to specific DNA.[27−30] However, nonspecific binding is likely only transient, as observed for binding of Pho4 and GCN4 to nonspecific DNA,[30,31] and stable DNA binding would be achieved only upon formation of favorable side chain interactions with the appropriate target DNA.

The CD signal in the 240−300 nm range was also examined for taG-box, gcG-box, and atG-box with EmBPwt (Figure S1 of the Supporting Information). In this range, protein contributions should be minimal and differences in DNA structure may be visible. For all of the DNA probes, similar changes were observed upon addition of protein. While the raw ellipticities of isolated DNA were variable, the ratio of signal increase and the observed red shift of the maxima upon protein binding were similar. All sequences showed maxima at approximately 272 nm in the absence of DNA. Upon addition of protein, the maxima were slightly shifted to approximately 274 nm with a small (1.1-fold) increase in ellipticity at the maximum.

K11A, R15A, and R20A Mutations Increase Sequence Specificity. On the basis of sequence alignment, structures, and past mutational data, as detailed below for each residue, the selectivity of three other mutants, K11A, R15A, and R20A, was examined. Unlike the R10A mutation, which reduces specificity, these mutations increase specificity among the sequences studied (Table 1). For all three mutations, the gcG-box affinity remains within a 2-fold range while affinities for the other sequences are greatly weakened.

Lysine 11 was chosen for mutation on the basis of the alignment with C-box binding protein TGA1a. This residue is in the "C region" (Figure 1) found by Niu et al.[14] to be critical for sequence specificity in segment swap experiments with C-box binding protein TGA1a. The K11A mutation reduces the binding affinity for all three low-affinity sequences, with the affinity for the atG-box reduced to 4800 ± 1200 nM, and gcC-box and atC-box bound extremely poorly with affinities of 9280 ± 60 and 10100 ± 120 nM, respectively. The gcG-box has an affinity 120 times stronger than that of the gcC-box and 64 times stronger than that of the atG-box.

Arginine 15 was targeted because although it was not mutated by Niu et al.,[14] it is also present in C-box binders CREB and TGA1a and was found by Montclare et al.[32] to be important for CREB half-site spacing specificity. The impact of the R15A mutation of EmBP-1 is more focused than that of the K11A mutation. This mutation significantly affects only the atG-box affinity, reducing it from 940 ± 50 nM with WT to 6900 ± 200 nM with the mutant. This results in a reversal of specificity with a 5.8-fold preference for C-box binding with A/T flanking sequence, while a 7.7-fold preference for G-box binding is maintained with G/C flanking sequence.

Arginine 20 interacts with the 5′-phosphate of A-5 in the GCN4 crystal structure[3] and thus was considered along with R10 as a possible source of flanking sequence preference. However, upon mutation of Arg20 to alanine, specificity was enhanced rather than diminished. The gcG-box affinity is slightly strengthened, while the gcC-box affinity is reduced approximately 3-fold, resulting in a 78-fold preference for binding to gcG-box over gcC-box. Single-point assays with concentrations near the gcC-box dissociation constant (i.e., ~50% bound protein) suggest that the binding constants for atG-box and atC-box are similar to that of gcG-box (data not shown). As this result is in line with the trend observed for WT protein, further quantitative studies were not pursued.

■ DISCUSSION

EmBP-1 is a bZip protein that has exhibited high-affinity binding to the gcG-box in past studies.[11,12] In these past studies, although the effect of outer flanking bases on affinity was explored qualitatively, the effect on specificity was not determined. The quantitative approach taken here, combined with the analysis of mutant proteins, allows for a more complete view of the sequence recognition process of EmBP-1. The highest affinity was observed for the gcG-box, which exhibits an 18-fold preference relative to the gcC-box. Furthermore, the R10 residue is critical for this binding preference with specificity lost upon mutation to alanine. This effect is restricted to G/C-flanked sequences. A/T- and C/G-flanked sequences produce little preference for binding to the G-box over the C-box, and this selectivity is unchanged upon mutation of R10. T/A-flanked sequences do exhibit G-box preference, but unlike G/C-flanked sequences, this preference is unaffected by R10. In contrast to R10A, protein mutants K11A, R15A, and R20A acted to increase sequence specificity by reducing affinity for weakly bound sequences. Possible factors contributing to sequence specificity are discussed below.

R10 Is the Only Residue Found To Significantly Affect Binding to the gcG-Box. Analysis of mutational data for the gcG-box, which is the most tightly bound sequence, reveals the fewest confirmed interactions (Table 1). The K11, R15, and R20 mutations all have a negligible effect on the binding affinity. Of the residues mutated, only R10A, which has no noticeable effect on binding to any other sequence studied, has a notable impact on gcG-box binding. This result appears to be linked to the basic nature of the arginine side chain because lysine is an adequate substitute at this position. On the basis of the crystal structure of PAP1,[1] it is proposed that this residue forms a hydrogen bond with O6 of G-5 (Figure 6).

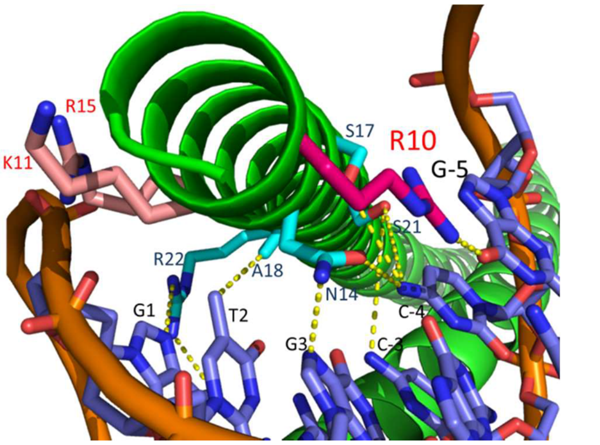


Figure 6. Three-dimensional model of interaction of EmBP with gcG-box. The model is based on protein chain E and DNA strands A and B from the crystal structure of PAP1 (PDB ID: 1GD2). DNA mutations at positions −3 and 3 and positions −4 and 4 (T and A to C and G in both cases) and A17S, Q18A, and F21S amino acid mutations were performed using Discovery Studio 3.1. Rotamers were chosen on the basis of maximal interaction with DNA bases. This figure is intended solely for the purpose of visualization, and no structural optimization has been performed.

The PAP1 crystal structure is the only structure of a bZIP protein with an interaction between the amino acid corresponding to residue 10 and a DNA base, and it reveals symmetrical deformation of the DNA at the center of the palindrome toward the protein. While the mean axial rise is consistent with B-DNA, the helical twist suggests A-type character. Both CREB and C/EBP produce very little deformation from standard B-DNA structure.[5,6] However, when rotamers of R10 were searched within the CREB and C/EBP structures, distances of >6 Å from the −5 flanking base were found for all rotamers. Thus, DNA deformation may be necessary for R10 to contact G-5. A conformational change in the DNA upon binding by EmBP-1 is consistent with the CD data obtained in the 240−300 nm range. A slight red shift and an increase in the magnitude of the signal in the 275 nm range, where DNA three-dimensional structure dominates the signal, are observed upon addition of protein. Similar small changes in this region are observed for binding of GCN4 to both AP-1 and C-box[33] as well as binding of Jun/Fos to AP-1 and C-box sequences,[34,35] with some variation in the response observed for C-box sequences depending on flanking nucleotides.[33,35] In EmBP-1, this change does not seem to be related to the flanking DNA sequence as the changes detected for the low-affinity atG-box sequence are similar to those of the tightly bound gcG-box and taG-box. Similarly, the C-box/CRE sequence has an intrinsic bend that is removed upon binding of CREB,[36] and in the case of CREB, the bend of the DNA is not related to sequence specificity.[23] However, different DNA sequences are known to have differences in propensity for deformations from standard B-form DNA, so while differences in the DNA structure based on flanking sequence are not apparent from the CD spectra, the inherent flexibility of the DNA may be significantly affected by the flanking sequence, in turn affecting the energetics of binding. This is supported by a recent modeling study for binding of GCN4 to varying DNA sequences where a significant interactive term between DNA bases, proposed to have a physical origin in $\pi-\pi$ stacking, was required for agreement between experimental data and computer models.[37]

Phosphate Contacts Have a Significant Effect on Specificity. From the analysis of mutant proteins, it appears that residues that are likely to make phosphate contacts have a significant effect on affinities and the importance of these contacts varies with the binding sequence. From the crystal structures of PAP1 and GCN4, K11 is capable of interaction with the 5′-phosphate of either base 2[3] or base 3.[1] This interaction appears to be important for interaction of EmBP-1 with both C-boxes, with a significant reduction in affinity upon mutation to alanine (see Table 1). K11 is also important for interaction with atG-box, but it may be involved in a base contact instead of a phosphate contact in this case. In the C/EBP structure, R11 hydrogen bonds with N7 of A3. No such interaction is possible between K11 of EmBP-1 and C3, but the direct base contact may be possible with N7 of G3. This hydrogen bond might be less energetically favorable than an electrostatic interaction with a phosphate group,[38] which may explain the lower energetic cost of the mutation with the atG-box. R20 has a significant effect on the affinity for gcC-box and appeared to have a similar effect on the affinity for the atC-box and atG-box, although full titrations were not performed with these sequences. R20 of GCN4 interacts with the 5′-phosphate of base −5,[3] as does K20 of C/EBP.[5] Thus, R20 of EmBP-1 is assumed to interact with the 5′-phosphate of G-5 (gcC-box) or A-5 (atC-box). The R15A mutation had a significant effect on only the affinity for the atG-box. For R15, an interaction with the 5′-phosphate of base 2 or 3 is anticipated. N15 of C/EBP,[5] T15 of GCN4,[3] and R15 of PAP1[1] all interact with this phosphate in their respective crystal structures. In CREB, R15 interacts with the same phosphate as well as making a second contact with the 5′-phosphate of base 3.[6] Although both K11 and R15 are predicted to contact the DNA far from the −5/5 bases, the outer sequence has a significant effect on the interactions with core bases. For both residues in EmBP-1, while a dramatic reduction in the level of binding to the atG-box is observed upon mutation to alanine, the level of binding to the gcG-box is reduced <2-fold. Overall, residues that are anticipated to make phosphate contacts have a significant effect on binding to the lower-affinity binding sequences, the gcC-box, atG-box, and atC-box.

Alanine Point Mutants Do Not Overlap with Segment Swapping Studies. The data obtained from alanine point mutations paint a different picture than data from segment swapping experiments. In the study by Niu et al.,[14] the segment primarily responsible for sequence selectivity is the C region. This set of EmBP-1 mutations (K11L, Q12A, S13Q, S17A, and R20K) was found to reverse selectivity from a 14.5-fold preference for the gcG-box over the gcC-box to a 2.4-fold preference for binding to the gcC-box. However, when individual residues from this set, K11 and R20, were mutated in this study to alanine, gcC-box affinity was significantly reduced, resulting in a stronger preference for the gcG-box.

In the case of R20, the different mutant residue may play a role in this apparent discrepancy. The alanine mutant results in a nearly 4-fold reduction in affinity for the gcC-box. Presumably, the arginine to alanine mutation abolishes any side chain interactions at this position. However, in the segment swap, the conservative R20K mutation is created. Thus, it is likely this mutation has little effect on binding affinity, and the importance of this residue is masked in this set of mutations. In the case of the K11 mutation, the discrepancy is less easily explained. The K11A mutation reduces the affinity for the gcC-box approximately 8.5-fold to nearly 10 μM, while gcG-box affinity is maintained within error of the wild-type affinity. The K11L mutation in the segment swap[14] also replaces the lysine with a hydrophobic residue that should also abolish any electrostatic interactions or H-bonds at this residue. However, the other mutations seem to be able to compensate for this probable large reduction in binding energy. Thus, while the results of this study show that the DNA sequence, including flanking bases, can have a profound effect on the importance of individual residues for binding affinity, the surrounding protein sequence can also alter the apparent interaction pattern.

Other Indirect Factors May Contribute to Protein−DNA Recognition. In the analysis described above, only overall free energy changes are considered and changes are analyzed at the level of direct contacts. However, more complex factors beyond the scope of this paper have been noted for other bZIP proteins. First, enthalpy−entropy compensation has been seen for other bZIP proteins such as Jun/Fos[39,40] and thus may affect the analysis of protein−DNA contacts. This phenomenon may mask the enthalpic contribution of some protein−DNA contacts because of an increase in entropy upon loss of the interaction. Thus, maintaining a similar dissociation constant after mutation does not necessarily indicate a contact was absent in the wild-type protein complex.

Additionally, Seldeen et al.[40] have noted an effect of mutations on the cooperativity of binding. While mutations of only one of the monomers in the heterodimeric Jun/Fos pair were largely ineffective, with entropy−enthalpy compensation governing the retention of binding affinity, combining single-point mutations on each of the monomers resulted in large reductions in binding affinity for certain combinations of mutations. This result indicated that allosteric effects beyond simple protein−DNA contacts play a role in DNA binding. For homodimeric EmBP-1, mutation of a single residue results in mutation in both monomers; thus, similar allosteric effects may affect some of the dissociation constants observed for the various alanine mutants.

■ CONCLUSIONS

Key findings of this study are the clarification of the previously observed flanking residues on DNA binding and determination of the previously unstudied basis of this effect in the protein sequence. Specifically, here it is revealed that R10 is responsible for the G/C flanking base preference and that this effect is specific for the G/C sequence. While a T/A flanking also results in a preference for G-box binding, this effect is not mediated by R10, showing that multiple binding modes are encoded within a single basic region.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Detailed protein purification protocols, derivation of eq 2 for competition EMSAs, Dynafit script for taG-box competition EMSAs, and CD spectra of DNA (Figure S1). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Department of Chemistry, University of Toronto, 80 St. George St., Toronto, ON, Canada M5S 3H6. Telephone: (416) 946-8021. E-mail: antonia.dejong@mail.utoronto.ca.

Funding

Funding was provided by the Ontario Graduate Scholarship funding program and the University of Toronto.

Notes

The author declares no competing financial interest.

ACKNOWLEDGMENTS

I thank Jack Iwanczyk and Jumi Shin for kindly providing the pET28a(+)/EmBPwt plasmid and Deborah Zamble and Andrew Woolley for the use of equipment and helpful discussion and editing.

ABBREVIATIONS

6-FAM, 6-carboxyfluorescein; ABRE, abscisic acid response element; bHLH, basic region helix−loop−helix; bHLHZ, basic region helix−loop−helix leucine zipper; bZIP, basic region leucine zipper; BSA, bovine serum albumin; CD, circular dichroism; EMSA, electrophoretic mobility shift assay; HPLC, high-performance liquid chromatography; IPTG, isopropyl β-D-1-thiogalactopyranoside; TAE, Tris-actetate-EDTA buffer; WT, wild type.

REFERENCES

(1) Fujii, Y., Shimizu, T., Toda, T., Yanagida, M., and Hakoshima, T. (2000) Structural basis for the diversity of DNA recognition by bZIP transcription factors. *Nat. Struct. Biol. 7*, 889−893.

(2) Glover, J. N. M., and Harrison, S. C. (1995) Crystal-Structure of the Heterodimeric bZIP Transcription Factor C-FOS-C-JUN Bound to DNA. *Nature 373*, 257−261.

(3) Keller, W., Konig, P., and Richmond, T. J. (1995) Crystal-Structure of a bZIP/DNA Complex at 2.2 Angstrom: Determinants of DNA Specific Recognition. *J. Mol. Biol. 254*, 657−667.

(4) Konig, P., and Richmond, T. J. (1993) The X-ray Structure of the GCN4-bZIP Bound to ATF CREB Site DNA Shows the Complex Depends on DNA Flexibility. *J. Mol. Biol. 233*, 139−154.

(5) Miller, M., Shuman, J. D., Sebastian, T., Dauter, Z., and Johnson, P. F. (2003) Structural basis for DNA recognition by the basic region leucine zipper transcription factor CCAAT/enhancer-binding protein alpha. *J. Biol. Chem. 278*, 15178−15184.

(6) Schumacher, M. A., Goodman, R. H., and Brennan, R. G. (2000) The structure of a CREB bZIP·somatostatin CRE complex reveals the basis for selective dimerization and divalent cation-enhanced DNA binding. *J. Biol. Chem. 275*, 35242−35247.

(7) Johnson, P. F. (1993) Identification of C/EBP Basic Region Residues Involved in DNA-Sequence Recognition and Half-Site Spacing Preference. *Mol. Cell. Biol. 13*, 6919−6930.

(8) Ransone, L. J., Wamsley, P., Morley, K. L., and Verma, I. M. (1990) Domain Swapping Reveals the Modular Nature of FOS, JUN, and CREB Proteins. *Mol. Cell. Biol. 10*, 4565−4573.

(9) Sellers, J. W., and Struhl, K. (1989) Changing FOS Oncoprotein to a JUN-Independant DNA-Binding Protein with GCN4 Dimerization Specificity by Swapping Leucine Zippers. *Nature 341*, 74−76.

(10) Agre, P., Johnson, P. F., and McKnight, S. L. (1989) Cognate DNA-Binding Specificity Retained After Leucine Zipper Exchange Between GCN4 and C/EBP. *Science 246*, 922−926.

(11) Foster, R., Izawa, T., and Chua, N. H. (1994) Plant bZIP Proteins Gather at ACGT Elements. *FASEB J. 8*, 192−200.

(12) Izawa, T., Foster, R., and Chua, N. H. (1993) Plant bZIP Protein-DNA Binding-Specificity. *J. Mol. Biol. 230*, 1131−1144.

(13) Guiltinan, M. J., Marcotte, W. R., and Quatrano, R. S. (1990) A Plant Leucine Zipper Protein That Recognizes an Abscisic Acid Response Element. *Science 250*, 267−271.

(14) Niu, X. P., Renshaw-Gegg, L., Miller, L., and Guiltinan, M. J. (1999) Bipartite determinants of DNA-binding specificity of plant basic leucine zipper proteins. *Plant Mol. Biol. 41*, 1−13.

(15) Niu, X. P., and Guiltinan, M. J. (1994) DNA-Binding Specificity of the Wheat bZIP Protein EmBP-1. *Nucleic Acids Res. 22*, 4969−4978.

(16) Hollenbeck, J. J., and Oakley, M. G. (2000) GCN4 binds with high affinity to DNA sequences containing a single consensus half-site. *Biochemistry 39*, 6380−6389.

(17) Metallo, S. J., and Schepartz, A. (1994) Distribution of labor among bZIP segments in the control of DNA affinity and specificity. *Chem. Biol. 1*, 143−151.

(18) Chan, I. S., Fedorova, A. V., and Shin, J. A. (2007) The GCN4 bZIP targets noncognate gene regulatory sequences: Quantitative investigation of binding at full and half sites. *Biochemistry 46*, 1663−1671.

(19) Guiltinan, M. J., and Miller, L. (1994) Molecular Characterization of the DNA-Binding and Dimerization Domains of the bZIP Transcription Factor, EmBP-1. *Plant Mol. Biol. 26*, 1041−1053.

(20) Dragan, A. I., Frank, L., Liu, Y. Y., Makeyeva, E. N., Crane-Robinson, C., and Privalov, P. L. (2004) Thermodynamic signature of GCN4-bZIP binding to DNA indicates the role of water in discriminating between the AP-1 and ATF/CREB sites. *J. Mol. Biol. 343*, 865−878.

(21) Worrall, J. A. R., and Mason, J. M. (2011) Thermodynamic analysis of Jun-Fos coiled coil peptide antagonists. Inferences for optimization of enthalpic binding forces. *FEBS J. 278*, 663−672.

(22) Carrillo, R. J., Dragan, A. I., and Privalov, P. L. (2010) Stability and DNA-Binding Ability of the bZIP Dimers Formed by the ATF-2 and c-Jun Transcription Factors. *J. Mol. Biol. 396*, 431−440.

(23) Metallo, S. J., Paolella, D. N., and Schepartz, A. (1997) The role of a basic amino acid cluster in target site selection and non-specific binding of bZIP peptides to DNA. *Nucleic Acids Res. 25*, 2967−2972.

(24) Kuzmic, P. (1996) Program DYNAFIT for the analysis of enzyme kinetic data: Application to HIV proteinase. *Anal. Biochem. 237*, 260−273.

(25) Chen, Y. H., Yang, J. T., and Chau, K. H. (1974) Determination of the helix and β form of proteins in aqueous solution by circular dichroism. *Biochemistry 13*, 3350−3359.

(26) O'neil, K. T., and Degrado, W. F. (1990) A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids. *Science 250*, 646−651.

(27) Sha, M., Ferre-D'Amare, A. R., Burley, S. K., and Goss, D. J. (1995) Anti-cooperative biphasic equilibrium binding of transcription factor upstream stimulatory factor to its cognate DNA monitored by protein fluorescence changes. *J. Biol. Chem. 270*, 19325−19329.

(28) Sauve, S., Naud, J.-F., and Lavigne, P. (2007) The mechanism of discrimination between cognate and non-specific DNA by dimeric b/HLH/LZ transcription factors. *J. Mol. Biol. 365*, 1163−1175.

(29) Cohen, S. L., Ferredamare, A. R., Burley, S. K., and Chait, B. T. (1995) Probing the solution structure of the DNA-binding protein Max by a combination of proteolysis and mass spectrometry. *Protein Sci. 4*, 1088−1099.

(30) Cave, J. W., Werner, K., and Wemmer, D. E. (2000) Backbone dynamics of sequence specific recognition and binding by the yeast Pho4 bHLH domain probed by NMR. *Protein Sci. 9*, 2354−2365.

(31) Okahata, Y., Niikura, K., Sugiura, Y., Sawada, M., and Morii, T. (1998) Kinetic studies of sequence-specific binding of GCN4-bZIP peptides to DNA strands immobilized on a 27-MHz quartz-crystal microbalance. *Biochemistry 37*, 5666−5672.

(32) Montclare, J. K., Sloan, L. S., and Schepartz, A. (2001) Electrostatic control of half-site spacing preferences by the cyclic AMP response element-binding protein CREB. *Nucleic Acids Res. 29*, 3311−3319.

(33) Votavova, H., Hodanova, K., Arnold, L., and Sponar, J. (1997) Interaction of a bZip oligopeptide model with oligodeoxyribonucleotides modelling DNA binding sites. The effect of flanking sequences. *J. Biomol. Struct. Dyn. 15*, 587−596.

(34) Weiss, M. A., Ellenberger, T., Wobbe, C. R., Lee, J. P., Harrison, S. C., and Struhl, K. (1990) Folding transition in the DNA-binding domain of GCN4 on specific binding to DNA. *Nature 347*, 575−578.

(35) John, M., Leppik, R., Busch, S. J., GrangerSchnarr, M., and Schnarr, M. (1996) DNA binding of Jun and Fos bZip domains: Homodimers and heterodimers induce a DNA conformational change in solution. *Nucleic Acids Res. 24*, 4487−4494.

(36) Paolella, D. N., Palmer, C. R., and Schepartz, A. (1994) DNA targets for certain bZIP proteins distinguished by an intrinsic bend. *Science 264*, 1130−1133.

(37) Wang, X., Zhang, A., Ren, W., Chen, C., and Dong, J. (2012) Genome-wide Inference of Transcription Factor−DNA Binding Specificity in Cell Regeneration Using a Combination Strategy. *Chem. Biol. Drug Des. 80*, 734−744.

(38) Privalov, P. L., Dragan, A. I., and Crane-Robinson, C. (2011) Interpreting protein/DNA interactions: Distinguishing specific from non-specific and electrostatic from non-electrostatic components. *Nucleic Acids Res. 39*, 2483−2491.

(39) Seldeen, K. L., McDonald, C. B., Deegan, B. J., Bhat, V., and Farooq, A. (2009) DNA Plasticity Is a Key Determinant of the Energetics of Binding of Jun-Fos Heterodimeric Transcription Factor to Genetic Variants of TGACGTCA Motif. *Biochemistry 48*, 12213−12222.

(40) Seldeen, K. L., Deegan, B. J., Bhat, V., Mikles, D. C., McDonald, C. B., and Farooq, A. (2011) Energetic coupling along an allosteric communication channel drives the binding of Jun-Fos heterodimeric transcription factor to DNA. *FEBS J. 278*, 2090−2104.