Pure Appl. Chem., Vol. 76, Nos. 7–8, pp. 1579–1590, 2004.
© 2004 IUPAC

Minimalist proteins: Design of new molecular recognition scaffolds*

Jumi A. Shin[‡]

Department of Chemistry, University of Toronto, Mississauga, Ontario L5L 1C6, Canada

Abstract: We hypothesize that we can exploit what Nature has already evolved by manipulating the α-helix molecular recognition scaffold. Therefore, minimalist proteins capable of sequence-specific, high-affinity binding of DNA were generated to probe how proteins are used and can be used to recognize DNA. The already minimal basic region/leucine zipper motif (bZIP) of GCN4 was reduced to an even more simplified structure by substitution with alanine residues—hence, a generic, Ala-based, helical scaffold. The proteins generated, **wt bZIP**, **4A**, **11A**, and **18A**, contain 0, 4, 11, and 18 alanine mutations in their DNA-binding basic regions, respectively. All alanine mutants still retain α-helical structure and DNA-binding function, despite loss of virtually all Coulombic protein-DNA interactions. Mass spectrometry allowed characterization of proteins and post-translational modifications. Fluorescence anisotropy and DNase I footprinting were used to measure in situ binding of these mutant proteins to DNA duplexes containing target sites AP-1 (5′-TGACTCA-3′), ATF/CREB (5′-TGACGTCA-3′), or nonspecific DNA. The roles of van der Waals and Coulombic interactions toward binding specificity and affinity are being investigated. Thus, both DNA-binding specificity and affinity are maintained in all our bZIP derivatives. This Ala-rich scaffold may be useful in design and synthesis of small, α-helical proteins with desired DNA-recognition properties.

INTRODUCTION

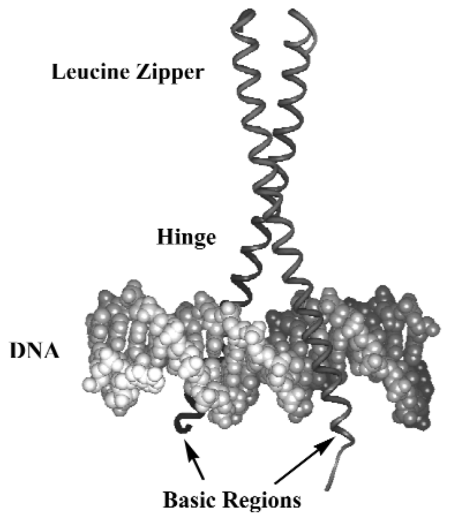
Our work aims to contribute to understanding the relationship between a protein's structure and its DNA-binding function—specifically, recognition of the DNA major groove by design of short, simplified α-helices based on the basic region/leucine zipper motif (bZIP). We exploit the protein α-helix, a structure used ubiquitously for sequence-specific DNA recognition, and one that chemists have successfully used in design and synthesis studies for many years (examples include refs. [1–6]).

Nature's use of the protein α-helix for specific DNA recognition is ubiquitous and maximally utilized by the bZIP, which comprises a pair of short, basic α-helices that recognize the DNA major groove with sequence-specificity and high affinity (Fig. 1). In order to begin to probe how Nature uses a pair of α-helices to bind DNA, we ask how DNA can be recognized by the bZIP structure: what is the relationship between a protein's structure and its DNA-binding function? Can we understand specificity and affinity in protein-DNA complexes by concentrating on simplified natural systems? Can we create a minimalist α-helical protein structure that retains desired DNA-binding function?

To begin investigation of the minimal protein determinants for sequence-specific, high-affinity recognition of the DNA major groove, proteins with α-helical structure and DNA-recognition capabil-

*Lecture presented at the symposium "Chemistry of nucleic acids", as part of the 39th IUPAC Congress and 86th Conference of the Canadian Society for Chemistry: Chemistry at the Interfaces, Ottawa, Canada, 10–15 August 2003. Other Congress presentations are published in this issue, pp. 1295–1603.

[‡]Tel.: (905) 828-5355; Fax: (905) 828-5425; E-mail, jshin@utm.utoronto.ca.



```
GCN4 basic regions
     226                        252
wt    DPAALKRARNTEAARRSRARKLQRMKQ
4A    ARAAAARARNTAAARRSRARKLQRMKQ
11A   ARAAAARARNTAAARRSRAAKAAAAAA
18A   AAAAAAAAANAAAAAAARAAKAAAAAA

 AP-1 DNA site      ATF/CREB DNA site
   3210               3210
5'-TGACTCA-3'      5'-TGAC•GTCA-3'
3'-ACTGAGT-5'      3'-ACTG•CAGT-5'

Oligonucleotides for Fluorescence Anisotropy
   AP-1 site, 20-mer
    5'-(FAM)TCCGGATGACTCATTTTTTG-3'

   ATF/CREB site, 21-mer
    5'-(FAM)TCCGGATGACGTCATTTTTTG-3'

   nonspecific duplex, 20-mer
    5'-(FAM)AACCGGTTACGTCAGACTGT-3'
```

Fig. 1 (Left) GCN4 bZIP in complex with the AP-1 DNA site, 5′-TGACTCA [13]. DNA is the horizontal double helix at the bottom of the figure, and the bZIP is the vertical α-helical dimer. The leucine zipper dimerizes into the coiled-coil structure shown at the top of the figure; the helical zipper then smoothly forks to either side of the DNA major groove. **(Right top)** Sequences of the bZIP basic regions. Sequences of the full expressed proteins comprise 30–35 residues from the expression vector, the GCN4 basic region and derivatives, C/EBP leucine zipper, plus a linker for chemical derivatization. Alanine substitutions are underlined, and highly conserved bZIP residues are in bold. **(Right middle)** Sequences of the AP-1, ATF/CREB, and nonspecific DNA sites. Numbering begins at the central CG base pair. Filled circle denotes division between abutting half sites in ATF/CREB. **(Right bottom)** Sequences of the oligonucleotide duplexes used in fluorescence anisotropy titrations. "FAM" is fluorescein phosphoramidite, and the AP-1 and ATF/CREB sites are underlined. For titrations with uracil containing duplexes, the sequences of u-AP-1 and u-NS are the same as for the FAM-labeled 20-mer duplexes, except that all thymines are replaced with uracil.

ities were generated from a core scaffold based on the GCN4 bZIP (Fig. 1) [7,8]. We focus on the bZIP domain of GCN4, a homodimeric transcriptional regulatory protein that governs histidine biosynthesis in yeast under conditions of amino acid starvation [9]. The α-helical bZIP motif was chosen for examination, as it is the smallest, simplest protein structure that recognizes specific DNA sites with high binding affinity [10,11]. The full-length GCN4 monomer is 281 amino acids, and the bZIP structure comprises a dimer of ~60-residue monomers. The compact bZIP is solely responsible for the protein dimerization and DNA-recognition activities of GCN4. Because the α-helix is well characterized and so commonly used by Nature, examination of how α-helices interact with specific DNA sites promises to be a manageable and informative strategy to explore protein-DNA recognition.

Crystal structures of the bZIP domain of GCN4 bound to two different DNA sites [12–14] and the Jun-Fos heterodimer bZIP-DNA crystal [15] show that a continuous α-helix of ~60 amino acids provides the basic region interface for binding to specific DNA sites, as well as the amphipathic leucine zipper coiled-coil dimerization structure (Fig. 1). The hinge region orients the basic region for binding of DNA. In GCN4, the hinge retains the flexibility to bind both the pseudopalindromic 7 base-pair (bp) AP-1 site, 5′-TGACTCA-3′, which is the in vivo target site of native GCN4 in yeast, and the palindromic 8 bp ATF/CREB site, 5′-TGAC•GTCA-3′, which is recognized by the cAMP-response element binding factor family [16], with similar affinities; ATF/CREB contains two abutting half sites with a filled circle marking the axis of C_2 symmetry, whereas AP-1 comprises two overlapping half sites (Fig. 1). Remarkably, these crystal structures also demonstrate astonishing conservation of protein backbone structure between the two yeast GCN4 and avian Jun-Fos structures. The protein scaffold is comprised of extremely regular and predictable α-helices that lie similarly in the major groove—the protein backbones in all three structures are virtually superimposable.

Thus, the simplicity and tractability of the bZIP make it an ideal molecular recognition scaffold for protein design and quantitative analysis of binding specificity and affinity. Although the wealth of information about the bZIP provides a sound basis for exploration of novel proteins based on a well-characterized motif, these data have demonstrated that a detailed understanding of protein interactions is a supremely complex issue that will be a challenge to "solve". No simple code exists for protein-DNA recognition, and this fact has made design of sequence-specific DNA-binding proteins a major challenge.

Despite these difficulties, several groups have developed productive strategies for design of proteins based on native motifs with desired, in some cases predictable, DNA-binding capabilities. Alanna Schepartz's group has designed miniature proteins based on the small α-helical avian pancreatic polypeptide, aPP [17]. They dissect the residues required for sequence-specific DNA binding and graft them onto the well-folding, but nonfunctional, aPP α-helix. A library of proteins is thus generated by solid-phase peptide synthesis or phage display [18], and functional selection follows. Using their protein-grafting strategy, the Schepartz group has generated an aPP variant that mimics GCN4's high DNA-binding affinity and specificity [17], another variant that comprises the DNA contact residues from the engrailed homeodomain also capable of high DNA-binding affinity and specificity [19], and miniature proteins that bind to human proteins Bcl-2 and Bcl-X_L [20]. This strategy has successfully yielded small helical proteins capable of native function, whether targeting specific DNA sites or serving as protein-binding ligands.

Other successful strategies have utilized the zinc-finger motif for recognition of specific DNA sites. Like the bZIP, each 30-amino acid Zn finger comprises an α-helix for sequence-specific recognition of the DNA major groove, but it also includes a β-sheet secondary structure as well as a necessary Zn^{+2} chelated by two Cys and two His residues critical for proper folding of the finger structure [21]. Because each individual Zn finger recognizes a 3-bp DNA subsite, Zn finger design lends itself well to modular assembly of proteins capable of targeting desired sites. Carl Pabo's group has shown that three-finger proteins that target a 10-bp sequence from the *erbB2* gene and six-finger proteins that target an 18-bp recognition sequence within the promotor of checkpoint kinase 2 gene (*CHK2*) can be selected from randomized libraries [22,23]. Extensive in vitro and in vivo characterization of these engineered Zn finger proteins' activities showed that protein-DNA recognition was very specific and high affinity (low nanomolar binding affinities) even in mammalian systems: these results underscore their strong potential as clinical therapeutics [24]. Carlos Barbas and his group have also been exploring three-finger and six-finger binding of DNA targets. Recently, they examined a modular strategy for design of Zn finger combinations whose DNA-binding abilities may possibly be routinely predicted. Extensive characterization of the six-finger proteins showed that their ability to predict in detail DNA-binding specificity and affinity is limited, and that possibly, affinity and specificity may be opposing forces, one coming at the expense of the other [25].

Although facile prediction and design of proteins capable of targeting any desired DNA sequence is complicated and may never be achieved, research has shown we can manipulate the natural protein scaffold for our own design and generate systems with native structure and function. Even the Zn finger structure can be further simplified: Barbara Imperiali's group has shown that a stable, well-folded, metal-independent 23-residue "Zn finger" can be generated [26]. Such simplified scaffolds, like the Zn finger and bZIP motifs, are worthy of exploration and exploitation in chemical design of novel proteins with desired capabilities.

RESULTS AND DISCUSSION

We hypothesize that the elegantly minimal bZIP structure can be reduced to an even more simplified structure by substitution with alanines to afford a preorganized, helical scaffold. These Ala-based proteins serve to begin investigation of the minimal protein determinants for sequence-specific, high-affinity recognition of the DNA major groove. Of the naturally occurring amino acids, Ala possesses the highest propensity for forming and stabilizing α-helical protein structures [27,28]. Significantly, the bZIP basic region is disordered until binding to DNA: NMR and circular dichroism (CD) demonstrate that while the leucine zipper is intrinsically stable and helical, the basic region remains only loosely helical until binding to a specific DNA sequence [29–32]. Nature may employ this folding transition to enhance control of gene transcription. Thus, the bZIP basic region requires site-specific DNA binding to achieve stability and helicity, and this energetic requirement may be circumvented by design of preorganized Ala-based scaffolds.

Four bZIP mutants containing increasing numbers of alanines were constructed to explore the relationship between a preorganized protein structure and DNA-binding function (Fig. 1). The leucine zipper contains C/EBP, residues 312–338, and the basic region comprises Ala-based derivatives of GCN4, residues 226–252 [7]. The **wt bZIP** (wild-type) is the "native" variant comprising the GCN4 basic region fused to the C/EBP leucine zipper at the same junction used by Agre et al.; their fusion was demonstrated to mimic the DNA-binding function of native GCN4 bZIP [33].

The GCN4 bZIP-DNA crystal structures show that only four amino acids in each bZIP basic region monomer make direct contacts to bases in the DNA major groove: Asn^{235}, Ala^{238}, Ala^{239}, and Arg^{243} [12–14]. These four amino acids are also highly conserved among bZIP proteins [34]. Our basic region mutant with the highest Ala content, **18A**, retains only these four amino acids from native GCN4, plus Lys^{246} due to concerns about solubility of hydrophobic proteins (Fig. 1); the refined crystal structure of the GCN4 bZIP with the ATF/CREB site shows that Lys^{246}, which lies in the hinge region, is involved in a water-mediated hydrogen-bonding network in the major groove and may improve protein solubility [14]. For our proteins, however, solubility was not noticeably enhanced by Lys^{246}, as all the proteins, even **wt bZIP** with the native GCN4 basic region, suffered from some solubility problems; the C/EBP leucine zipper likely contributes to the expressed proteins' hydrophobicity.

Note that only 3 of 27 amino acids in the **18A** basic region are non-alanine. Only base-specific interactions are conserved with **18A**, and Coulombic protein-DNA interactions have been virtually abolished. **4A** and **11A** contain 4 and 11 Ala substitutions, respectively: in these proteins, both specific interactions with DNA bases and nonspecific electrostatic interactions with the DNA phosphodiester backbone are maintained [12–14]. **11A** is also mutated in the hinge region, which is important for spacing the basic region monomers properly on the DNA site. In in vivo experiments on the bZIP protein C/EBP, Sera and Schultz found that mutations of amino acids in the hinge region can affect DNA-binding function [35]. Amino acid 227 is arginine in both **4A** and **11A**; this is a cloning artifact, and this residue has no interactions with DNA [12–14].

Temperature-leap tactic maintains protein solubility

These bacterially expressed mutants are unusual proteins for expression in that they are short (~100 amino acids) and hydrophobic (Ala-mutated basic regions, leucine-zipper dimerization domains). Hydrophobicity was a significant issue throughout the expression and purification stages. We overcame major problems with inclusion body formation and protein aggregation by careful manipulation of conditions for cell growth and induction of protein expression, use of high concentrations of denaturant in all steps of protein isolation, purification, and storage (at least 4 M denaturant), and temperature modulation techniques on protein stocks and working solutions.

We cannot conduct experiments, including footprinting and mass spectrometry, in 4 M denaturant and high salt, however. In order to maintain protein solubility at denaturant concentrations <4 M, a temperature-leap tactic (T-leap) that aids in maintaining protein solubility when concentration of denaturant drops to <400 mM was utilized [36]. In studies by Xie and Wetlaufer on the renaturation of bovine carbonic anhydrase, when guanidine hydrochloride concentration was reduced below 1 M, protein aggregation occurred; however, if refolding was allowed to occur at 4 °C for 2 h, aggregation was significantly suppressed (37 % enzyme activity). Moreover, if the enzyme was then rapidly warmed to 36 °C, the activity increased to 95 %. Their explanation was that there are two sequential, slow-folding intermediates, the first being prone to aggregation, the second leading to native enzyme. At 4 °C, the aggregation-prone intermediate is depleted after 120 min, and a rise in temperature allows the second intermediate to convert rapidly to the native, active form [36]. Although our bZIP mutants are much smaller than carbonic anhydrase and not similar in structure or function, we decided to try the T-leap tactic.

Circular dichroism and DNase I footprinting: Structure and function

We utilized the T-leap in our footprinting, mass spectrometry, and fluorescence anisotropy experiments. DNase I footprinting studies showed that all of our bZIP proteins bind specifically to AP-1 and ATF/CREB. Our **wt bZIP** strongly footprints at 5 μM monomer concentration, **4A** at 10 μM, **11A** at 20 μM, and **18A** at 100 μM (Fig. 2). We note that high concentrations of protein give clear footprints, not merely coating DNA nonspecifically. The mutants consistently produced a DNA-binding pattern mimicking that of **wt bZIP** on a ~650 base-pair restriction fragment. Despite elimination of numerous Coulombic interactions, these proteins still bind specific sequences [7,8]. These results are especially surprising for the heavily mutagenized **18A**, in which 24 of 27 amino acids in the basic region are alanine: only those amino acids making *specific* contacts to DNA bases—Asn^{235}, Ala^{238}, Ala^{239}, and Arg^{243}—are maintained (nonbonding Arg^{246} was retained to aid solubility); all nonspecific electrostatic contacts have been eliminated. **18A** demonstrates that extraordinarily few amino acids in the bZIP are necessary to confer high-affinity, sequence-specific DNA binding.

We took advantage of the fact that the native GCN4 basic region is intrinsically disordered, and therefore, CD can be utilized to monitor changes in bZIP helical structure upon DNA binding. Our bZIP mutants should become increasingly helical as Ala content increases, and therefore, structural stability, which we equate with α-helicity, also increases. Mean residue ellipticity values at Θ_{222} for these mutants may be compared to Θ_{222} for an ideal α-helix, calculated to be –37 500 deg·cm^2·dmol^{-1} [37]. The **wt bZIP** and **4A** have intrinsic helical character of 27 and 38 %, respectively, whereas **11A** and **18A** possess substantially more helicity of 59 and 71 %, respectively [8]. **18A** is maximally helical in the bZIP domain (the expressed proteins contain an extra ~35 residues from the expression vector). Therefore, increasing Ala content in the bZIP basic region generates proteins of higher α-helical stability with potentially more favorable energetics for binding to DNA.

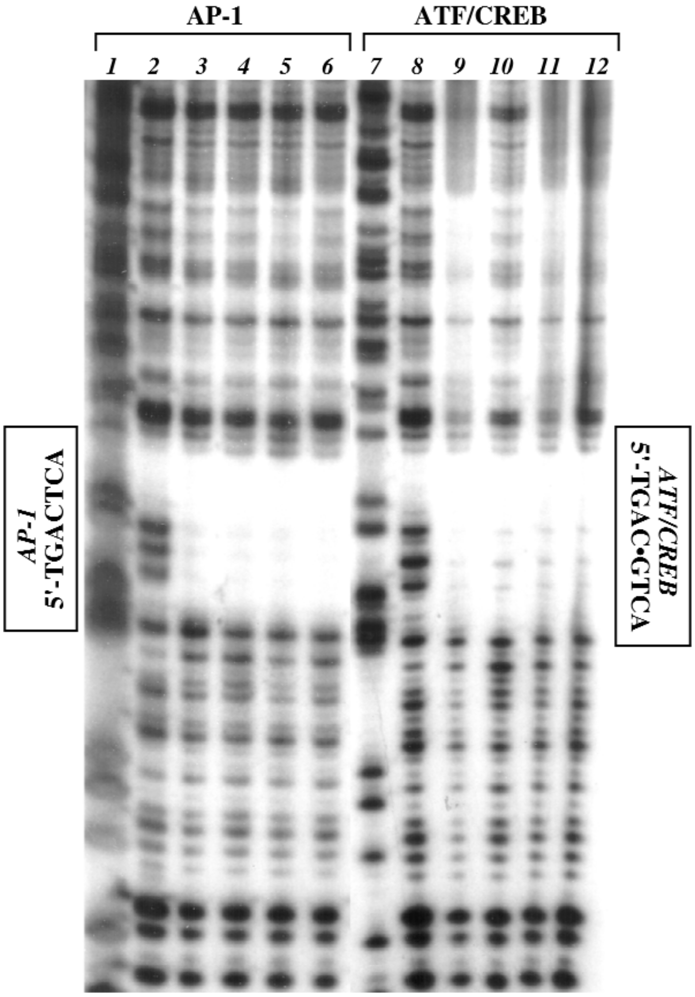


Fig. 2 Autoradiogram of a high-resolution denaturing polyacrylamide gel of DNase I footprinting reactions on **wt bZIP**, **4A**, **11A**, and **18A** proteins bound to the AP-1 and ATF/CREB DNA sites. Footprint sites are indicated by boxes on either side of autoradiogram. Data presented for 3′ end-labeled DNA. Lanes 1–6, footprinting at AP-1 site; lanes 7–12, footprinting at ATF/CREB site. Lanes 1 and 7, chemical sequencing G reaction [57]; lanes 2 and 8, DNase I cleavage control. Lanes 3–6 and 9–12, DNase I cleavage reactions in the presence of various concentrations of protein. Lanes 3 and 9, 5 μM wt bZIP. Lanes 4 and 10, 10 μM **4A**. Lanes 5 and 11, 20 μM **11A**. Lanes 6 and 12, 100 μM **18A**. The bars drawn on the left and right sides of the autoradiogram indicate the AP-1 and ATF/CREB sites, respectively.

MALDI-TOF mass spectrometry confirms identity and protein modifications

We experienced great difficulty in obtaining mass spectrometry data for our proteins due to the presence of high concentrations of salt and denaturant necessary to keep protein from aggregating. By incorporating the T-leap tactic to aid in protein-matrix cocrystallization and by using high matrix-to-protein ratios of 50 000 or 100 000 to 1, we were able to detect strong mass spectrometric signals for all four mutants by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS), even in the presence of salts and denaturant. No spectra were generated in the absence of T-leap. The percent errors between the observed and calculated masses for all four proteins was <0.05 % [38].

Enzymatic digestion mapping combined with MALDI-TOF MS characterization of protein fragments allowed us to resolve mass discrepancies between the expected and observed molecular mass measurements [39]. Once high-quality, calibrated mass spectra were generated on the full intact bZIP proteins, it became evident that these proteins were modified [38]. Proteins can undergo a wide variety of co- and post-translational modifications during growth, isolation, and purification. We therefore performed enzymatic digestion on each protein followed by MALDI-TOF MS analysis. By using an Endo-Lys carboxypeptidase on-plate digestion reaction reported by Roepstorff and coworkers [40], we were able to vary systematically the amounts of protein, enzyme, and reaction times to optimize the intensities of the protein fragments. The on-plate digestion procedure involved spotting sequentially the matrix solution, protein solution, and enzyme solution onto a 100-well gold MALDI-TOF MS sample plate and allowing each spot to dry completely at room temperature before adding the next spot. After the enzyme solution was spotted, the MS plate was placed on a crystallizing dish in a 37 °C water bath for 1–2 h for the enzymatic cleavage reaction [38,39].

The enzymatic digestion reactions combined with MALDI proved that all four proteins were missing their initiating methionines. Protein translation is initiated by *N*-formyl methionine, which can be removed by methionine aminopeptidase, especially if the adjacent amino acid possesses a small side chain, as in our proteins, which contain glycine as the penultimate amino-terminal residue [41–44]. Also evident was adduct formation: we suspected carbamylation at the amino terminus and perhaps lysine side chains, for the proteins were in high contact with urea. However, acetylation by $N\alpha$-acetyltransferase at the amino terminus is also a possibility [41]: MALDI is unable to distinguish between the addition of an acetyl group (+42.037 Da) or a carbamyl group (+43.025 Da). New urea solutions were purified by deionization on mixed-bed ion-exchange resin. Fresh proteins were expressed and purified with deionized urea and digested on plate. Spectra were retaken: adduct formation at the *N*-terminus or lysines was not present. Therefore, the adduct formation observed previously was due to carbamylation, not acetylation [39].

We also observed up to four or five sequential additions of approximately 16 amu. This was attributed to formation of 2-oxohistidine. Because our proteins possess a six-histidine tag that aids in purification, reactive species, such as metal-generated radicals, can oxidize the histidine side chains. Indeed, Uchida and Kawakishi found that 2-oxoHis was the main product in oxidative modification of proteins by free radicals: for example, inactivation of Cu,Zn-superoxide dismutase occurs readily upon radical oxidation of His^{118} to 2-oxoHis in the active site [45]. Because the His_6 tag on our proteins is utilized to chelate a metal ion affinity column during purification, generation of 2-oxoHis under these conditions is likely. Except during metal ion affinity chromatography, all protein solutions contain EDTA to chelate unwanted metal ions.

Changes in mass were attributed to posttranslational modifications by proteolytic cleavage of the initiating methionine residue, carbamylation at the amino terminus, oxidation of histidine side chains, and in some cases, oxidative addition of β-mercaptoethanol at the cysteine side chain (Table 1) [39]. Although they changed the expected molecular masses of our proteins, these modifications did not affect their α-helical structure or DNA-binding function, as these changes were not in the bZIP regions of our proteins [8]. Gaining MALDI MS for these proteins was extremely challenging, yet highly informative. Such changes in mass can only be detected by a high-resolution technique such as MALDI, which in conjunction with enzymatic digestion mapping, becomes a powerful methodology for characterization of protein structure.

Table 1 Mass spectrometry results[a].

bZIP	Expected mass[b] (amu)	–Met[c] (amu)	+*N*-Carbamylation[d] (amu)	+BME[e] (amu)	Calculated mass (amu)	Observed mass (amu)	% Error
wt	11073.16	–132.08	+43.03	+76.11	11060.22	11062.85	0.024
4A	11731.85	–132.08	+43.03	NA	11642.80	11637.20	0.048
11A	11288.27	–132.08	NA	NA	11156.19	11158.61	0.022
18A	10816.76	–132.08	+43.03	+76.11	10803.82	10806.39	0.024

[a] Data from ref. [46].
[b] For full intact protein with no posttranslational modifications.
[c] Loss of initiating methionine at amino terminus.
[d] Carbamylation at amino terminus.
[e] BME is β-mercaptoethanol.

Fluorescence anisotropy: Thermodynamics of protein-DNA complexation

We used fluorescence anisotropy spectroscopy to measure dissociation constants of our bZIP derivatives binding oligonucleotide duplexes labeled with fluorescein [46]. Fluorescence anisotropy measures the tumbling motion of molecules containing a fluorophore; the anisotropic tumbling of our short DNA duplexes increases measurably upon protein binding, as the 5′-fluorescein-labeled duplex is ~13 kD and bZIP monomer is ~11–12 kD. True thermodynamic/equilibrium binding vs. stoichiometric binding can be achieved due to the sensitivity of fluorescence for detection of fluorescein [47]. Therefore, protein concentration can be maintained in excess throughout the titration (at least 25-fold excess protein over DNA in our work).

We conducted fluorescence anisotropy measurements to generate binding isotherms of all four bZIP proteins bound to the AP-1 20-mer and ATF/CREB 21-mer duplexes (Fig. 1). Table 2 lists the dissociation constants. The dissociation constants of the protein-DNA complexes were measured under equilibrium conditions. K_d values reveal strong, sequence-specific binding in the low nanomolar range; binding to a nonspecific sequence was at least three orders of magnitude weaker than specific binding [46]. All four proteins displayed similar binding affinities to both the AP-1 and ATF/CREB sites; likewise, the native GCN4 bZIP does not discriminate between AP-1 and ATF/CREB [31]. The **wt bZIP** binds to the AP-1 and ATF/CREB sites with K_d values of 9.1 nM and 14 nM, respectively. These numbers correlate very well with measurements made by other groups for GCN4 and other bZIP domains [12,31,48–53].

Table 2 Dissociation constants for GCN4 bZIP derivatives bound to the AP-1, ATF/CREB, nonspecific DNA, u-AP-1, or u-NS sites[a].

	K_d (10^{-9} M)				
bZIP	AP-1[b]	ATF/CREB[b]	Nonspecific[b,c]	u-AP-1[d]	u-NS[c,d]
wt	9.1 ± 1.2	14 ± 1.5	>1 μM	48.9 ± 15.7	>1 μM
4A	78 ± 5.7	64 ± 5.0	>1 μM	18.5 ± 2.14	not determined
11A	4.8 ± 0.65	5.8 ± 0.53	>10 μM	12.8 ± 2.08	not determined
18A	15 ± 1.3	7.8 ± 1.3	>10 μM	16.5 ± 1.34	>1 μM

[a] Each titration experiment contained 250 pM specific DNA duplex or 1 nM nonspecific DNA duplex in buffer (4.3 mM Na_2HPO_4, 1.4 mM KH_2PO_4, pH 7.4, 150 mM NaCl, 2.7 mM KCl, 1 mM EDTA, 800 mM urea, 20 % glycerol, 0.4 mg/ml acetylated BSA, 1 mM DTT, and 100 μM in base-pair calf thymus DNA) [46].
[b] Data from ref. [46].
[c] Saturation protein binding was not achieved in any of the nonspecific duplex DNA titrations.
[d] Data from ref. [55].

4A is the weakest binder to AP-1 and ATF/CREB; thus, the few alanine replacements in **4A** are enough to weaken its DNA-binding ability. However, more Ala substitutions, as in **11A** and **18A**, regain high-affinity DNA-binding function. These binding results closely parallel those of Takemoto and Fisher, who made Ala replacements in bHLH proteins Myc and TFEB [54]. Likewise, the K_d values of their bHLH mutant-DNA complexes varied within one order of magnitude, as do our values for bZIP mutant-DNA complexes, despite extensive mutations. Interestingly, they also found that just a few Ala mutations were detrimental to protein-DNA binding affinity, similar to **4A**, and therefore, any helical stability conferred by such few Ala replacements was not enough to compensate for losses in protein-DNA interactions [54]. Yet with increasing Ala substitutions, they regained binding affinities, similar to **11A** and **18A**. In the bHLH mutants, increasing Ala mutations in the 18-residue DNA-binding basic region conferred more α-helical stability to the disordered native basic region, which, like our bZIP mutants, becomes helical and stable upon sequence-specific DNA binding; this significant increase in structural stability can now compensate for losses in protein-DNA interactions. The results of Takemoto and Fisher and our work demonstrate a general trend that increasing α-helicity in basic regions by alanine mutagenesis can generate preorganized basic regions capable of high-affinity, sequence-specific DNA binding [46].

Given that the nonpolar **18A** binds strongly and specifically to AP-1 and ATF/CREB, thereby mimicking native GCN4 behavior, we examined the importance of van der Waals interactions between our mutants and DNA duplexes in which thymines were replaced with uracils, thereby replacing the C5 methyl group on thymine with hydrogen [55]. Substitution of thymines with uracils removes specific nonpolar interactions between the thymine C5 methyl and protein groups, and importantly, uracil substitutions do not affect DNA structure. Therefore, meaningful, quantitative comparisons of the binding energetics of these mutants bound to native DNA or uracil-substituted DNA can be made. Both the u-AP-1 and t-AP-1 duplexes are of the same sequence, except for T → U substitutions at all thymines; the duplexes are both 20-mers labeled with fluorescein (Fig. 1).

The structures of the GCN4 bZIP in complex with the AP-1 [13] or ATF/CREB [12,14] sites show that the methyl side chains of Ala^{238} and Ala^{239} are within van der Waals contact distance of the C5 methyl groups of thymines T3 and T1′ (Fig. 1), respectively. The Ser^{242} side chain also makes a more distant van der Waals contact with the C5 methyl on T3. Our mutants can maintain these nonpolar interactions, for Ala^{238} and Ala^{239} are retained in all our proteins; Ser^{242} is replaced with Ala in the **18A** mutant, but the Ala methyl should similarly maintain the native van der Waals interaction. Possibly, our mutants may replace native Coulombic and hydrogen-bonded contacts with new van der Waals interactions between Ala methyl side chains and nonpolar groups on DNA.

Fluorescence anisotropy titrations showed that K_d values to the uracil-containing AP-1 duplex (u-AP-1) were strong for all proteins, in the low-to-mid nanomolar range, very similar to binding to the original thymine-containing duplexes (Table 2) [55]. The **wt bZIP** binding to u-AP-1 is five-fold weaker than binding to t-AP-1; **4A** binds four-fold stronger to u-AP-1 than to t-AP-1, corresponding to ~0.6 kcal/mol, again a very modest change. **11A** and **18A** bind to u-AP-1 and t-AP-1 with essentially the same, low-nanomolar dissociation constants. Changes in binding affinity upon thymine-to-uracil substitution are not dramatic, and this may underscore the multivalent nature of protein-DNA interactions. Proteins can compensate for mutations that affect enthalpic and entropic contributions toward DNA-binding in order to maintain overall free energies similar to that of the native protein-DNA complex [56]. Therefore, compensatory changes in our proteins may alleviate the losses of van der Waals interactions with thymine C5 methyl groups or Coulombic interactions removed upon Ala substitutions. At the same time, however, this compensation can complicate our efforts to create proteins with widely variable binding affinities for DNA.

CONCLUSION

Our data demonstrate the important result that the bZIP can be greatly simplified with Ala substitution—such that only those few amino acids responsible for *direct, specific* interactions with the DNA major groove are maintained—and retain native structure and function. Traditional views about protein design hold that because a protein's structure is maintained by cooperative interactions of individual amino acids, mutations will often be disruptive of the entire structure, particularly for small proteins. Though cognizant of this fact, our work showed that the α-helix is a stable, reproducible scaffold-tolerating substitutions more predictably than other protein structures.

The bZIP-DNA crystal structures show that only four amino acids per monomer make base-specific contacts with the major groove [12–15]: these amino acids can be considered responsible for bZIP function, i.e., sequence-specific recognition of the DNA major groove. Of the 20 proteinogenic amino acids, alanine is sufficient to maintain the minimalist bZIP backbone α-helix. Those amino acids necessary for maintenance of the α-helix can be considered responsible for bZIP structure. Therefore, Ala-rich α-helices with judiciously placed residues conferring DNA-binding function may comprise the minimal protein determinants for high-affinity, sequence-specific recognition of the DNA major groove. These results suggest that short, predictable peptides can make design, synthesis, and characterization of biomolecular assemblies a more tractable problem.

ACKNOWLEDGMENTS

I especially wish to acknowledge my coworkers who contributed to this work: Ajay Lajmi, Greg Bird, Mike Lovrencic, and Ken Kise. Generous funding has been provided by NSF (CAREER MCB-9733410) and NIH (1RO1GM63220).

REFERENCES

1. P. E. Dawson and S. B. H. Kent. *J. Am. Chem. Soc.* **115**, 7263–7266 (1993).
2. A. Grove, M. Mutter, J. E. Rivier, M. Montal. *J. Am. Chem. Soc.* **115**, 5919–5924 (1993).
3. S. Marqusee and R. L. Baldwin. *Proc. Natl. Acad. Sci. USA* **84**, 8898–8902 (1987).
4. T. Sasaki and E. Kaiser. *J. Am. Chem. Soc.* **111**, 380–381 (1989).
5. R. P. Wharton and M. Ptashne. *Nature* **316**, 601–605 (1985).
6. M. R. Ghadiri, C. Soares, C. Choi. *J. Am. Chem. Soc.* **114**, 4000–4002 (1992).
7. A. R. Lajmi, T. R. Wallace, J. A. Shin. *Prot. Exp. Purif.* **18,** 394–403 (2000).
8. A. R. Lajmi, M. E. Lovrencic, T. R. Wallace, R. R. Thomlinson, J. A. Shin. *J. Am. Chem. Soc.* **122**, 5638–5639 (2000).
9. D. E. Hill, I. A. Hope, J. P. Macke, K. Struhl. *Science* **234**, 451–457 (1986).
10. K. Struhl. *Trends Biochem. Sci.* **14**, 137–140 (1989).
11. W. H. Landschulz, P. F. Johnson, S. L. McKnight. *Science* **240**, 1759–1764 (1988).
12. P. König and T. J. Richmond. *J. Mol. Biol.* **233**, 139–154 (1993).
13. T. E. Ellenberger, C. J. Brandl, K. Struhl, S. C. Harrison. *Cell* **71**, 1223–1237 (1992).
14. W. Keller, P. König, T. J. Richmond. *J. Mol. Biol.* **254,** 657–667 (1995).
15. J. N. M. Glover and S. C. Harrison. *Nature* **373**, 257–261 (1995).
16. P. K. Brindle and M. R. Montminy. *Curr. Opin. Gen. Dev.* **2**, 199–204 (1992).
17. N. J. Zondlo and A. Schepartz. *J. Am. Chem. Soc.* **121**, 6938–6939 (1999).
18. J. W. Chin, R. M. Grotzfeld, M. A. Fabian, A. Schepartz. *Bioorg. Med. Chem. Lett.* **11**, 1501–1505 (2001).
19. J. K. Montclare and A. Schepartz. *J. Am. Chem. Soc.* **125**, 3416–3417 (2003).
20. J. W. Chin and A. Schepartz. *Angew. Chem., Int. Ed.* **40**, 3806–3809 (2001).
21. C. O. Pabo and R. T. Sauer. *Annu. Rev. Biochem.* **61**, 1053–1095 (1992).

22. J. A. Hurt, S. A. Thibodeau, A. S. Hirsch, C. O. Pabo, J. K. Joung. *Proc. Natl. Acad. Sci. USA* **100**, 12271–12276 (2003).
23. S. Tan, D. Guschin, A. Davalos, Y.-L. Lee, A. W. Snowden, Y. Jouvenot, H. S. Zhang, K. Howes, A. R. McNamara, A. Lai, C. Ullman, L. Reynolds, M. Moore, M. Isalan, L.-P. Berg, B. Campos, H. Qi, S. K. Spratt, C. C. Case, C. O. Pabo, J. Campisi, P. D. Gregory. *Proc. Natl. Acad. Sci. USA* **100**, 11997–12002 (2003).
24. A. C. Jamieson, J. C. Miller, C. O. Pabo. *Nat. Rev. Drug Discov.* **2**, 361–368 (2003).
25. D. J. Segal, R. R. Beerli, P. Blancafort, B. Dreier, K. Effertz, A. Huber, B. Koksch, C. B. Lund, L. Magnenat, D. Valente, C. F. Barbas III. *Biochemistry* **42**, 2137–2148 (2003).
26. M. D. Struthers, R. P. Cheng, B. Imperiali. *Science* **271**, 342–345 (1996).
27. K. T. O'Neil and W. F. DeGrado. *Science* **250**, 646–651 (1990).
28. I. Luque, O. L. Mayorga, E. Freire. *Biochemistry* **35**, 13681–13688 (1996).
29. K. T. O'Neil, J. D. Shuman, C. Ampe, W. F. DeGrado. *Biochemistry* **30**, 9030–9034 (1991).
30. V. Saudek, H. S. Pasley, T. Gibson, H. Gausepohl, R. Frank, A. Pastore. *Biochemistry* **30**, 1310–1317 (1991).
31. M. A. Weiss, T. E. Ellenberger, C. R. Wobbe, J. P. Lee, S. C. Harrison, K. Struhl. *Nature* **347**, 575–578 (1990).
32. J. A. Shin. *Bioorg. Med. Chem. Lett.* **7**, 2367–2372 (1997).
33. P. Agre, P. F. Johnson, S. L. McKnight. *Science* **246**, 922–926 (1989).
34. P. F. Johnson. *Mol. Cell. Biol.* **13**, 6919–6930 (1993).
35. T. Sera and P. G. Schultz. *Proc. Natl. Acad. Sci. USA* **93**, 2920–2925 (1996).
36. Y. Xie and D. B. Wetlaufer. *Prot. Sci.* **5**, 517–523 (1996).
37. J. M. Scholtz, H. Qian, E. J. York, J. M. Stewart, R. L. Baldwin. *Biopolymers* **31**, 1463–1470 (1991).
38. G. H. Bird, A. R. Lajmi, J. A. Shin. *Anal. Chem.* **74**, 219–225 (2002).
39. G. H. Bird and J. A. Shin. *Biochim. Biophys. Acta* **1597**, 252–259 (2002).
40. M. Kussmann, E. Nordhoff, H. Rahbek-Nielsen, S. Haebel, M. Rossel-Larsen, L. Jakobsen, J. Gobom, K. Mirgorodskaya, A. Kroll-Kristensen, L. Palm, P. Roepstorff. *J. Mass Spectrom.* **32**, 593–601 (1997).
41. S. M. Arfin and R. A. Bradshaw. *Biochemistry* **27**, 7979–7984 (1988).
42. P. Wingfield, P. Graber, G. Turcatti, N. R. Movva, M. Pelletier, S. Craig, K. Rose, C. G. Miller. *Eur. J. Biochem.* **180**, 23–32 (1989).
43. C. Miller, K. L. Strauch, A. M. Kukral, J. L. Miller, P. T. Wingfield, G. J. Mazzei, R. C. Werlen, P. Graber, N. R. Movva. *Proc. Natl. Acad. Sci. USA* **84**, 2718–2722 (1987).
44. S. Huang, R. C. Elliott, P.-S. Liu, R. K. Koduri, J. L. Weickmann, H.-H. Lee, L. C. Blair, P. Ghosh-Dastidar, R. A. Bradshaw, K. M. Bryan, B. Einarson, R. L. Kendall, K. H. Kolacz, K. Saito. *Biochemistry* **26**, 8242–8246 (1987).
45. K. Uchida and S. Kawakishi. *J. Biol. Chem.* **269**, 2405–2410 (1994).
46. G. H. Bird, A. R. Lajmi, J. A. Shin. *Biopolymers* **65**, 10–20 (2002).
47. V. LeTilly and C. A. Royer. *Biochemistry* **32**, 7753–7758 (1993).
48. C. Park, J. L. Campbell, W. A. Goddard III. *J. Am. Chem. Soc.* **117**, 6287–6291 (1995).
49. T. Morii, J. Yamane, Y. Aizawa, K. Makino, Y. Sugiura. *J. Am. Chem. Soc.* **118**, 10011–10017 (1996).
50. S. J. Metallo and A. Schepartz. *Chem. Biol.* **1**, 143–151 (1994).
51. G. J. Foulds and F. A. Etzkorn. *Nucleic Acids Res.* **26**, 4304–4305 (1998).
52. J. J. Hollenbeck and M. G. Oakley. *Biochemistry* **39**, 6380–6389 (2000).
53. J. W. Sellers, A. C. Vincent, K. Struhl. *Mol. Cell. Biol.* **10**, 5077–5086 (1990).
54. C. Takemoto and D. E. Fisher. *Gene Express.* **4**, 311–317 (1995).
55. K. J. Kise, Jr. and J. A. Shin. *Bioorg. Med. Chem.* **9**, 2485–2491 (2001).

56. L. Jen-Jacobson, L. E. Engler, L. A. Jacobson. *Structure* **8**, 1015–1023 (2000).
57. A. Maxam and W. Gilbert. *Methods Enzymol.* **65**, 499–560 (1980).