Role of DNA Sequence in the Binding Specificity of Synthetic Basic-Helix-Loop-Helix Domains

Amy C. Beltran,[a] Philip E. Dawson,*[b] and Joel M. Gottesfeld*[a]

The basic helix-loop-helix (bHLH) domain defines a class of transcription factors that are essential for the regulation of many genes involved in cell differentiation and development. To determine the role of the DNA sequence in driving dimerization specificity of bHLH transcription factors, we analyzed the DNA sequence in and around a consensus hexanucleotide binding site (E-box). The bHLH domains of two transcription factors, E12 and TAL1, were chemically synthesized. The minimal DNA binding domain for both the E12 homodimer and the E12–TAL1 heterodimer was determined, thereby extending the E-box by two base pairs. Additional studies indicate that the presence of a thymine in the first flanking position 5′ to the E-box prevents DNA binding of both dimer complexes. The presence of a thymine or cytosine in a flanking position two bases 5′ to the E-box decreases the affinity for the E12 homodimer twofold but completely inactivates DNA binding for the E12–TAL1 heterodimer. Access to synthetic DNA and protein enabled the analysis of specific interactions between a conserved arginine residue in the basic helix of each bHLH domain and adenine in a flanking position two bases 5′ to the E-box. Our results indicate a key role of the DNA sequence in driving dimerization specificity among bHLH transcription factors.

Introduction

The basic helix-loop-helix (bHLH) proteins comprise a family of over 240 transcriptional regulators, which are involved in diverse biological functions such as cell-growth control and transformation, neurogenesis, sex determination, muscle differentiation, and other essential processes.[1–5] A classification scheme based upon structural, biochemical, and genetic data divides bHLH proteins into four distinct classes.[6] Studies have shown that class A bHLH proteins dimerize with many different bHLH proteins; this allows them to have a profound effect on the genesis and maintenance of many different cell lineages.[7–11] Class B bHLH proteins show a tissue-restricted pattern of expression. With a few exceptions, these proteins are incapable of forming homodimers and preferentially heterodimerize with members of class A.[7,12] Class C bHLH proteins are structurally unique in that they contain a leucine zipper on the C-terminal side of the bHLH motif. The proteins of the last class of bHLH proteins, class D, are also structurally unique, lacking the basic region and being unable to bind DNA. Members of this class are negative regulators of class A and class B bHLH proteins.[13–16] The functionality of bHLH proteins in transcriptional regulation is imparted by their ability to specifically dimerize with class A bHLH proteins.

TAL1, a class B bHLH, is a pivotal regulatory protein expressed in pluripotent hematopoietic stem cells.[3,17] The normal pattern of TAL1 expression is restricted to cells of early hematopoietic lineages, notably erythroid, mast, and early myeloid cells.[18–20] Other studies indicate that TAL1 is also required for proper B- and T-lineage development.[3,21] TAL1 was first identified through its involvement in a T-cell acute lymphoblastic leukemia (T-ALL).[4] Heterodimers of TAL1 and members of class A (such as E47, a member of the E family of bHLH protiens) have been found in normal hematopoietic cells that are undergoing erythroid differentiation as well as in leukemic cells of T-ALL patients.[9,12,22] This indicates that both normal and malignant properties of TAL1 are mediated by interaction with class A bHLH proteins such as E47. As with most class B bHLH proteins, TAL1 has been shown not to dimerize with itself or interact with ID, a class D HLH.[23] Thus, heterodimer formation with class A bHLH proteins is essential for the DNA-binding activity and functional properties of class B bHLH proteins.

Formation of a bHLH dimer is facilitated by stabilizing van der Waals interactions between conserved hydrophobic residues located in the HLH region of each monomer.[24] Upon dimerization, the HLH regions of both monomers form a parallel, left-handed, four-helix bundle with a stable hydrophobic core. Formation of this parallel four-helix bundle allows the basic helix to contact the major groove of DNA, with each monomer interacting with a DNA half-site.[24–27] Conserved amino acids within the basic helix of each monomer make base-specific and phosphate contacts over the DNA recognition sequence. The putative DNA-recognition sequence of bHLH proteins was first discovered from the kE2 site of the immunoglobulin kappa gene enhancer.[7,28] A consensus DNA site, the "E-box", is defined by the hexanucleotide sequence CANNTG, here the inner two nucleotides can vary but the outer bases, CA- and -TG, are absolutely conserved.

[a] *Dr. A. C. Beltran, Dr. J. M. Gottesfeld*
Department of Molecular Biology, The Scripps Research Institute
La Jolla, California 92037 (USA)
E-mail: joelg@scripps.edu

[b] *Dr. P. E. Dawson*
Departments of Cell Biology and Chemistry
The Scripps Research Institute
La Jolla, California 92037 (USA)
Fax: (+1) 858-784-8965
E-mail: dawson@scripps.edu

  DOI: 10.1002/cbic.200400184 *ChemBioChem* **2005**, *6*, 104 – 113

The bHLH motif provides an interesting opportunity for studying dimerization specificity among class A and B bHLH proteins. Binding studies conducted on bHLH proteins have primarily dealt with the recombinant full-length proteins.[9,14,29–32] However, the bHLH motif, of roughly 60 amino acids, has also been synthetically generated and shown to be capable of folding and binding DNA.[33–36] We have taken advantage of the small size of the bHLH domain and used solid-phase peptide-synthesis methods to generate the bHLH domains of both E12 and TAL1. The correlation between our results with a synthetic bHLH domain and previous studies with recombinant full-length bHLH proteins[9,14,29–32] allows for further investigation of this domain in the absence of the additional sequence elements in the full-length protein.

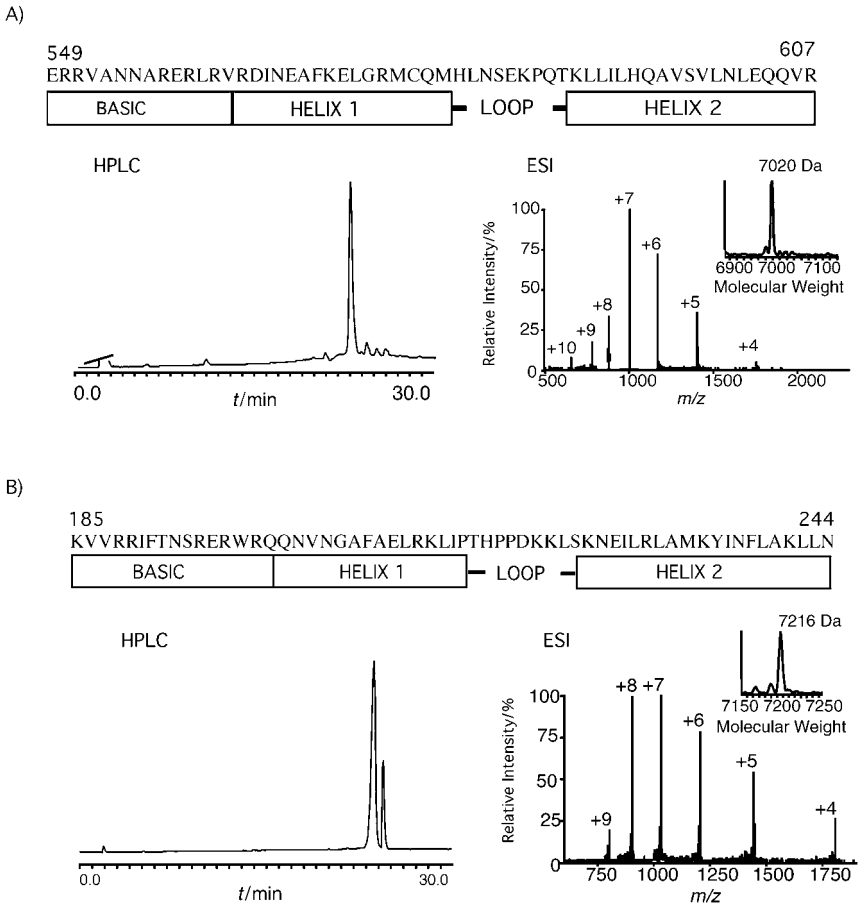


Figure 1. *A) Amino acid sequence of the bHLH domain of the human E12. Regions of secondary structure are indicated underneath. RP-HPLC trace obtained from pure material is shown to the left. The first peak that is shortened corresponds to acetic acid. Electrospray ionization (ESI) mass spectrum of the HPLC purified product with the observed protonation states labeled is also shown to the right. Reconstructed mass analysis of this spectrum is shown as inset. B) Amino acid sequence of the bHLH domain of the human TAL1. Regions of secondary structure are indicated underneath. RP-HPLC trace of pure material is shown to the left. The first peak that is shortened corresponds to acetic acid. ESI mass spectrum of the HPLC purified product with the observed protonation states labeled is shown to the right. Reconstructed mass analysis of this spectrum is shown as inset.*

Results

Chemical synthesis of E12 and TAL1

The bHLH domains of E12 and TAL1 were manually synthesized by using in situ neutralization protocols for Boc solid-phase peptide synthesis (Figure 1a, E12 amino acids 549–607, and Figure 1b, TAL1 amino acids 184–244).[37] These synthetic bHLH domains are denoted E12 and TAL1, corresponding to the names of the full-length polypeptide chains. E12 and TAL1 were purified by RP-HPLC and folded. The purified peptides were analyzed by electrospray ionization-mass spectrometry (ESI-MS) and the observed masses matched the masses that were calculated by isotopic composition. In the case of E12, the observed mass was 7020.6 ± 1.0 Da with a calculated mass of 7018.2 Da. TAL1 yielded an observed mass of 7219.9 ± 1.0 Da with a calculated mass of 7217.6 Da.

Dissociation constants for the E12 homodimer and the E12–TAL1 heterodimer binding to DNA

Determination of the true equilibrium dissociation constants for the interaction of bHLH proteins with DNA requires analysis of the coupled equilibria of protein dimerization and DNA binding. Since bHLH dimerization occurs at micromolar protein concentrations and DNA binding is observed at low nanomolar concentrations, the contribution of protein dimerization to the observed dissociation constant will be negligible under the experimental conditions used herein.[10] A series of oligonucleotides was used for our binding studies (Figure 2). The apparent dissociation constants (K_d) for both the E12 homodimer and the E12–TAL1 heterodimer were measured for binding to the kE2 E-box found in the immunoglobulin kappa chain enhancer.[28] Quantitative electrophoretic mobility shift assays (EMSA) were used to investigate the interactions between E12 homodimers and E12–TAL1 heterodimers and DNA (Figure 3). The K_d obtained for the E12 homodimer, 11.5 ± 0.4 nM, reflects a slightly higher affinity for DNA than was previously determined in the context of the recombinant protein (52 nM) (see row A in Table 1).[32] Additionally, the K_d obtained for the E12–TAL1 heterodimer, 2.4 ± 0.3 nM, correlates well with previously obtained results of an E12–TAL1 heterodimer that consisted of recombinant proteins (3.6 nM to 4.2 nM; see row A in Table 2).[30]

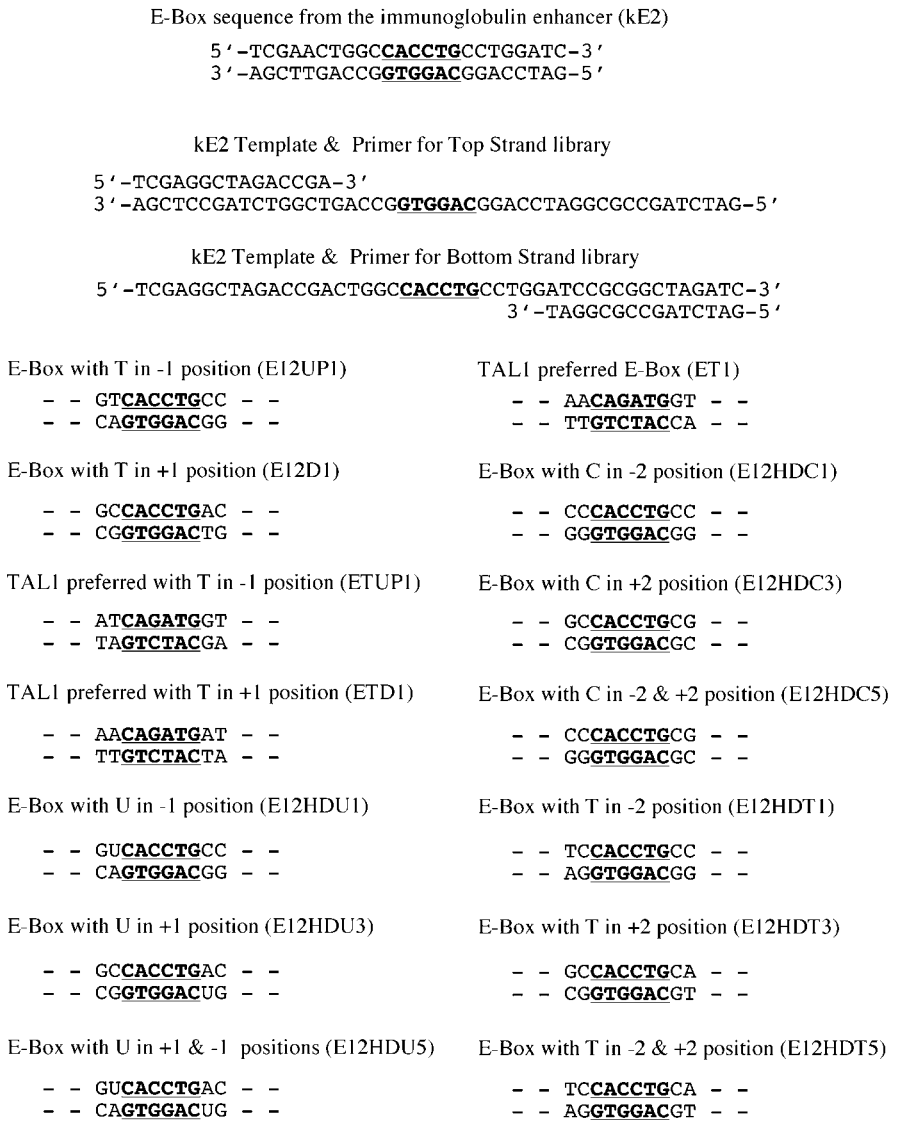


Figure 2. *Oligonucleotides used in this study. E-boxes are underlined and shown in boldface letters. Dashed lines indicate additional bases not shown that correspond to bases indicated in kE2 DNA sequence. The sequences are illustrated so as to allow for emphasis on important flanking sequences. Oligonucleotides are represented 5′ to 3′ on the top strand oligonucleotide and the complimentary oligonucleotide represented 3′ to 5′. Positions indicated by a "– number" indicate the number of bases 5′ from the E-box on the top strand. Positions shown by a "+ number" indicate the number of bases 5′ from the E-box on the bottom strand.*

Minimal binding domain of both the E12 homodimer and the E12–TAL1 heterodimer

While the E-box sequence is strictly defined, the minimal binding site required for bHLH-dimer binding has not been determined. Examination of three dimensional bHLH–DNA structures shows binding outside the E-box by each bHLH monomer.[24,27] This binding could both target certain flanking sequences and expand the DNA binding site. A number of assays have been used to determine the preferred binding site for bHLH dimers.[29,30,38,39] These assays target the composition of the E-box flanking sequences and show divergent length requirements. In some cases, the two outer bases seem to be important; however, it has also been suggested that as many as five bases outside of the E-box are necessary for high-affinity binding.[29] These studies, while inconclusive, suggest that the minimal binding sequence for a bHLH dimer extends beyond the six base pairs of the E-box.

In order to determine the minimal DNA sequence for bHLH-dimer binding, E12 and TAL1 were allowed to homodimerize or heterodimerize and then bind to a library of double-stranded DNA fragments that differed successively in length by one base pair.[40] This DNA library included oligonucleotides that contained parts or all of the kE2 E-box and flanking sequences. Only DNA fragments containing the necessary elements for bHLH-dimer binding are selected and bound in this assay. The selection experiment was performed at limiting protein concentrations—where less than 50% of the DNA was bound—in order to select for DNA sequences containing high-affinity binding sites. Libraries were generated from either the top or bottom strand DNA templates; this enabled the determination of flanking bases important for DNA binding both 5′ and 3′ to the E-box, respectively (Figure 2). Selected DNA fragments from each library were then separated and identified under denaturing sequencing gel conditions (Figure 4). Individual bands correspond to a unique DNA fragment and are represented by the last incorporated base. Quantitation of individual bands allowed for graphical analysis of the binding data (Figure 5). Values for the natural log of $[DNA_{Bound}]/[DNA_{Total}]$ were plotted as a function of the last incorporated base. Values that are significantly negative correspond to DNA fragments that do not contain a full E-box and necessary flanking sequences. These are represented on the right side of each graph as the length of the DNA fragments decreases from left to right. A characteristic dip in the beginning of the graphs corresponds to longer DNA fragments that are under-represented in the unbound lanes due to an increase in binding of these longer DNA fragments. Analysis of the data for the E12 homodimer shows that binding of this dimer is influenced by two base pairs flanking the 3′ side of the E-box and only one base pair flanking the 5′ side of the E-

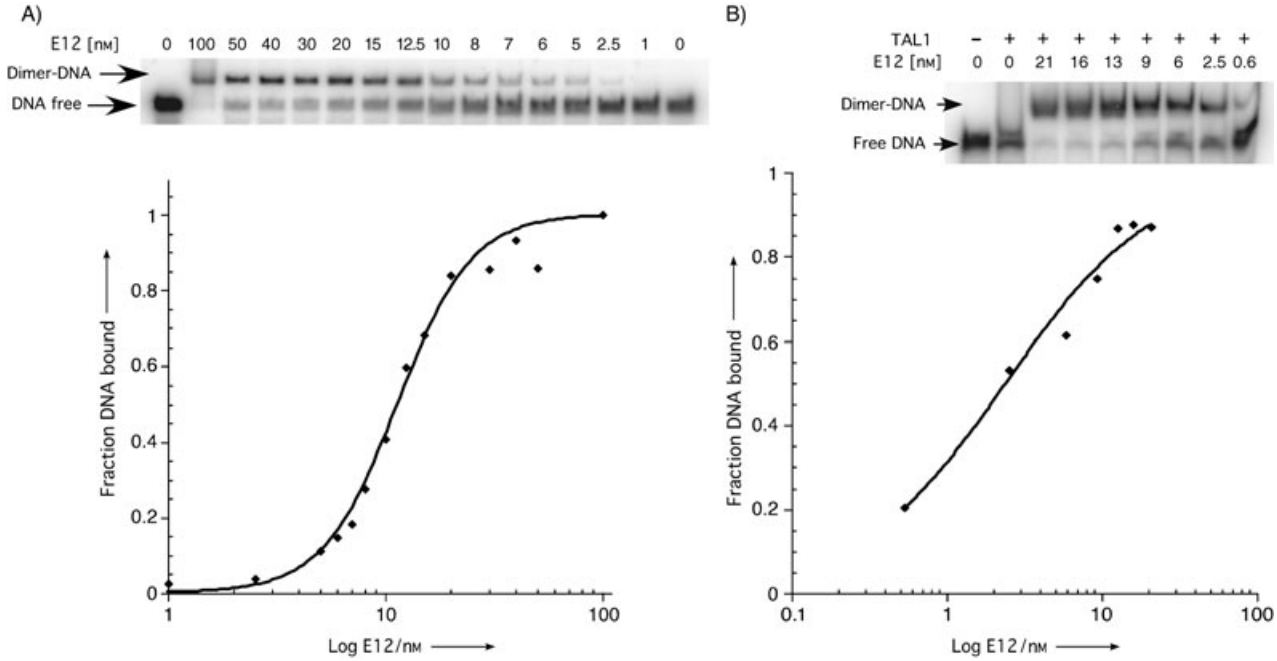


Figure 3. *A) EMSA of E12 homodimer binding to kE2 DNA oligonucleotide. The K_d was determined to be 11.5 nM. B) EMSA of E12–TAL1 heterodimer binding to kE2 DNA oligonucleotide in the presence of 75 nM TAL1. The K_d was determined to be 2.4 nM.*

Table 1. *Dissociation constants of the E12 homodimer bound to DNA oligonucleotides of varying sequences. Only the E-box and flanking sequences are shown for the top strand (5′–3′). E-boxes are indicated by bold typeface and underlined. Columns with N/A are shown for DNA oligonucleotides unable to bind E12 homodimer. The apparent K_ds and standard deviation for at least three determinations are given.*

	Oligo	E-box and flanking	E12–E12 [nM]
A	kE2 "wild-type"	-GC**CACCTG**CC-	11.5 ± 0.4
B	ET1 TAL1 preferred	-AA**CAGATG**GT-	10.7 ± 0.3
C	E12HDU1 uracil in −1 position	-GU**CACCTG**CC-	12.7 ± 0.4
D	E12HDU3 uracil in +1 position	-GC**CACCTG**AC-	13.2 ± 0.7
E	E12HDU5 uracil in both	-GU**CACCTG**AC-	12.0 ± 0.2
F	E12HDC1 cytosine in −2 pos.	-CC**CACCTG**CC-	26.4 ± 1.5
G	E12HDC3 cytosine in +2 pos.	-GC**CACCTG**CG-	23.2 ± 1.4
H	E12HDC5 cytosine in both	-CC**CACCTG**CG-	28.4 ± 0.5
I	E12HDT1 thymine in −2 pos.	-TC**CACCTG**CC-	25.1 ± 2.0
J	E12HDT3 thymine in +2 pos.	-CC**CACCTG**CA-	20.7 ± 1.3
K	E12HDT5 thymine in both	-TC**CACCTG**CA-	21.6 ± 1.2
L	E12UP1 thymine in −1 pos.	-GT**CACCTG**CC-	N/A
M	E12D1 thymine in +1 pos.	-GC**CACCTG**AC-	N/A
N	ETUP1 thymine in −1 pos.	-AT**CAGATG**GT-	N/A
O	ETD1 thymine in +1 pos.	-AA**CAGATG**AT-	N/A

Table 2. *Dissociation constants (K_d) of the E12–TAL heterodimers bound to DNA oligonucleotides of varying sequences. Only the E-box and flanking sequences are shown for the top strand (5′–3′).*

	Oligo	E-box and flanking	E12–TAL1 [nM]
A	kE2 "wild-type"	-GC**CACCTG**CC-	2.4 ± 0.3
B	ET1 TAL1 preferred	-AA**CAGATG**GT-	1.0 ± 0.2
C	E12HDU1 uracil in −1 position	-GU**CACCTG**CC-	1.5 ± 0.2
D	E12HDU3 uracil in +1 position	-GC**CACCTG**AC-	1.7 ± 0.2
E	E12HDU5 uracil in both	-GU**CACCTG**AC-	2.2 ± 0.2
F	E12D1 thymine in +1 pos.	-GC**CACCTG**AC-	N/A
G	ETUP1 thymine in −1 pos.	-AT**CAGATG**GT-	N/A
H	ETD1 thymine in +1 pos.	-AA**CAGATG**AT-	N/A

box. Analysis of the data for the E12–TAL1 heterodimer shows that the binding of this dimer is also influenced by two base pairs flanking the 3′ side of the E-box and only one base pair flanking the 5′ side of the E-box. The minimal binding domains, 5′-GCCACCTGC-3′, determined for both homodimer– and heterodimer–DNA complexes, have an asymmetric sequence with 5′ overhangs on either side of the E-box. The requirement of additional bases outside the E-box for both the E12 homodimer and the E12–TAL1 heterodimer is consistent with earlier studies suggesting that protein–DNA contacts occur outside the E-box.[24,27]

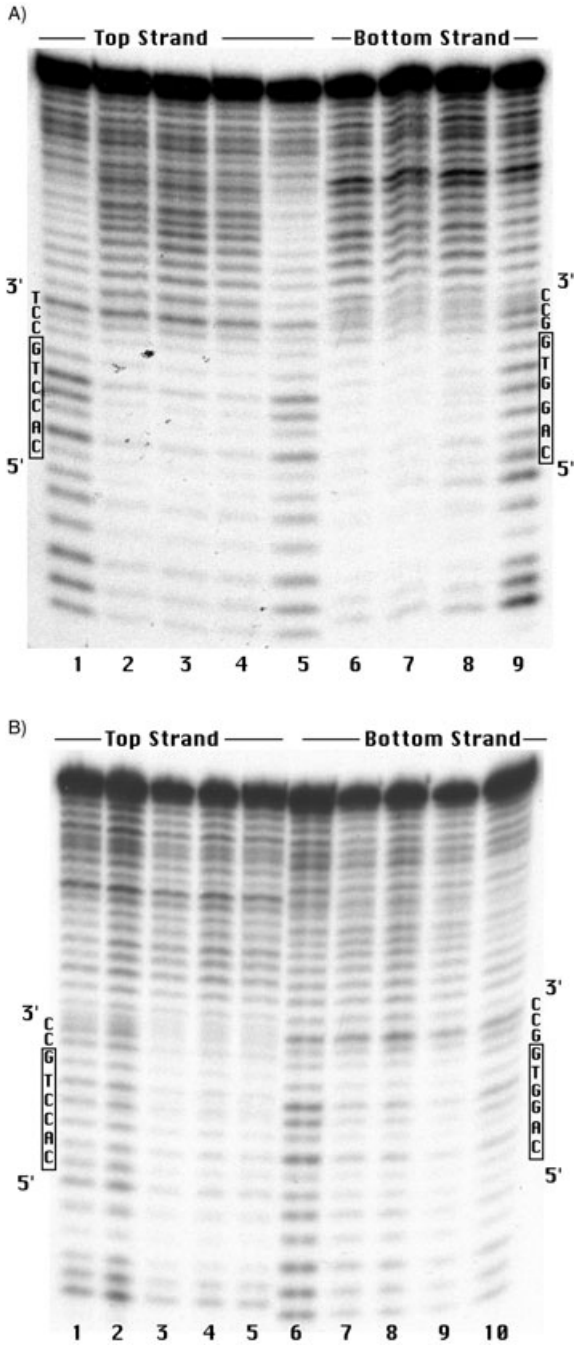


Figure 4. *A) Sequencing ladder of bound and unbound DNA for E12 homodimer titrations with two DNA libraries. Lanes 1–5 correspond to DNA fragments from a top-strand library while lanes 6–9 correspond to DNA fragments from a bottom-strand library (Figure 2). The E-box is indicated with a box. Lanes 1, 5, and 9 correspond to DNA fragments in the unbound fractions. Lanes 2, 3, 4 and 6, 7, 8 correspond to DNA fragments that were able to bind E12 homodimers at varying E12 monomer concentrations; up to 44% (lane 2) or as low as 24% of DNA found to bind (lane 8). B) Sequencing ladder of bound and unbound DNA for E12–TAL1 heterodimer titrations with two DNA libraries. Lanes 1–5 correspond to DNA fragments from a top-strand library and lanes 6–10 correspond to DNA fragments from a bottom-strand library (Table 1). The E-box is indicated with a box. Lanes 5 and 9 correspond to DNA fragments in the unbound fractions while lanes 1 and 10 correspond to total input DNA. Lanes 2, 3, 4 and 6, 7, 8 correspond to DNA fragments that were able to bind E12 homodimers at varying E12 monomer concentrations; up to 28% (lane 6) or as low as 16% DNA was found to bind (lane 8).*

Role of E-box-flanking sequences in dimerization and DNA binding

Quantitative EMSAs were used to measure binding constants for either the E12 homodimer or the E12–TAL1 heterodimer with E-box oligonucleotides of differing flanking sequences. An E-box "half site" showing preferential binding to a TAL1 monomer has previously been determined to be 5′-AACAGATGGT-3′, with the E-box underlined and the preferred half site for TAL1 being 5′-ATG-3′.[30] The E12–TAL1 heterodimer binds to an E-box sequence without the preferred half site, with an observed dissociation constant of 2.4 ± 0.3 nM, while the same complex yields a dissociation constant of 1.0 ± 0.2 nM for an E-box sequence with the preferred half site (row B in Tables 1 and 2). The E12 homodimer showed comparable affinity for both DNA sequences, with K_ds of 10.7 ± 0.3 nM for the preferred sequence and 11.5 ± 0.4 nM for the kE2 binding sequence. Therefore, the preferred TAL1 sequence does not show discrimination between E12 homodimer binding and E12–TAL1 heterodimer binding.

The role of flanking base pairs in binding affinity for both the E12 homodimer and the E12–TAL1 heterodimer was examined next. The convention 5′−1 denotes the base position one base from the 5′ end of the E-box, and 5′−2 denotes the base two bases 5′ to the E-box. The presence of a thymine base at 5′−1, on the top or bottom strand of either the kE2 E-box or the TAL1-preferred E-box sequence, inhibited DNA binding with protein concentrations of up to of 150 nM (Figure 6A). Interpretation of results with protein concentrations above 150 nM was not possible due to multiple binding complexes, determined by smearing and supershifiting of the DNA. Substitution of uracil for thymine in either the 5′−1 position of the top or bottom strand restores DNA binding of both dimer complexes; this indicates that the thymine methyl group interferes with DNA binding for both the E12 homodimer and E12–TAL1 heterodimer. Binding affinities measured for the E12 dimer complex to uracil containing oligonucleotides mirror binding affinities previously obtained with the kE2 oligonucleotide (see rows C, D, and E in Table 1). Additionally, binding affinities measured for the E12–TAL1 dimer complex to uracil-containing oligonucleotides were also consistent with binding affinities obtained for the kE2 oligonucleotide (see rows C, D, and E in Table 2).

Arginine contacts from a bHLH monomer to bases flanking the E-box have been observed in both the E47 and MyoD crystal structures.[24,27] A water-mediated hydrogen bond between the δ amino group of this arginine and the N7 of adenine at 5′−2 has been suggested to facilitate DNA binding. The role of N7 in the purine ring, found in either adenine or guanine at a position two bases removed from the E-box, was then assayed. The flanking sequences were probed by determining the binding affinities of the E12 homodimer for oligonucleotides with either a cytosine (see rows F, G, and H in Table 1) or thymine (see rows I, J, and K in Table 1) at 5′−2. The observed dissociation constants revealed a twofold decrease in DNA-binding affinity for the E12 homodimer with this substitution.

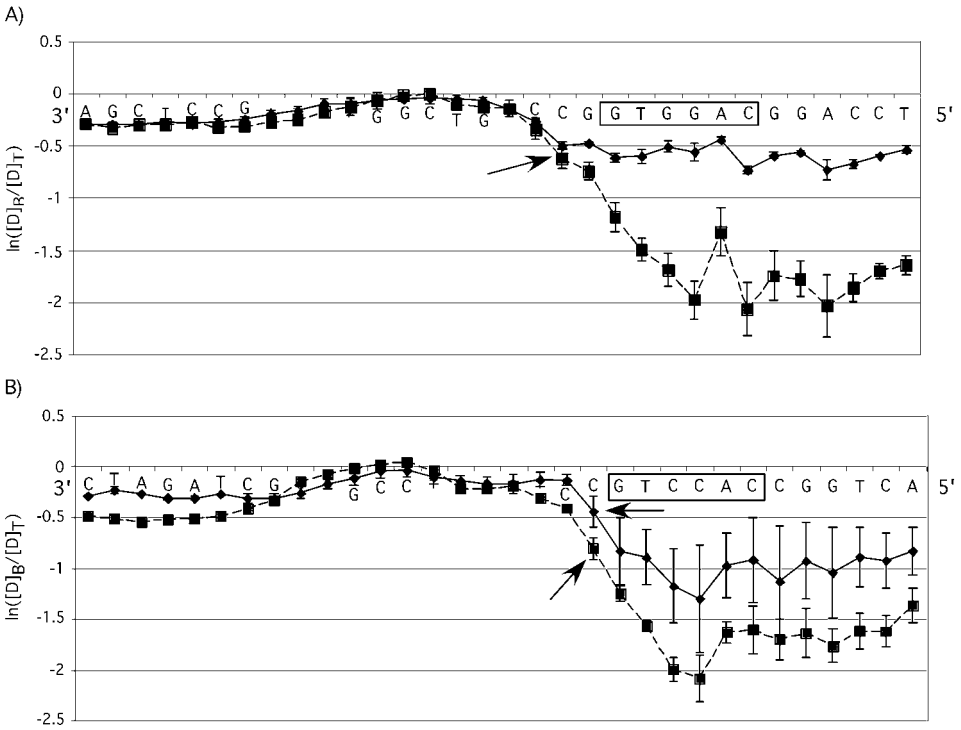


Figure 5. *Graphical analysis of bands from Figure 3 for both the E12 homodimer and the E12–TAL1 heterodimer. The E-box is underlined. In both cases the full-length DNA oligonucleotide is the first data point and the DNA fragments become shorter as the graph progresses left to right. ■ corresponds to data points for the E12 homodimer, while ◆ corresponds to data points for the E12–TAL1 heterodimer. Arrows indicate base endpoints obtained from each of the curves. A) This graph shows binding reactions carried out with a bottom-strand library. The data indicate that two base pairs flanking the 3′ end of the E-box on the bottom strand contribute to DNA binding of both the E12 homodimer and the E12–TAL1 heterodimer. A) This graph shows binding reactions performed with a top-strand library. The data indicate that only one base pair flanking the 3′ end of the E-box on the top strand are necessary for DNA binding of both the E12 homodimer and the E12–TAL1 heterodimer.*

Oligonucleotides flanked with cytosine or thymine at 5′−2 were also assayed with E12 in the presence of TAL1 under conditions in which E12–TAL1 heterodimers were previously observed. However, even in the presence of TAL1, only E12 homodimers were observed with dissociation constants similar to that of the E12-only titrations (Figure 7, and Table 3). The absence of E12–TAL1 heterodimers indicates a decreased ability of the TAL1 monomer to bind an E-box with either cytosine or thymine at 5′−2. Thus, the presence of either a guanine or adenine two bases 5′ to the E-box is important for high-affinity binding of the E12 homodimer and the E12–TAL1 heterodimer.

Role of arginine, R557, in DNA binding assessed by using amino acid analogues

Since DNA contacts outside the E-box are mediated though an arginine residue, R557, located within the basic helix,[24] the role of this amino acid residue in binding affinity for both the homo- and heterodimer was examined. E12 was synthesized by using solid-phase peptide synthesis and resin-splitting methods, which allowed the generation of both wild-type and E12 with unnatural amino acid substitutions in parallel. In this synthesis, R557, was substituted with either ornithine or citruline (Table 4). E12 protein titrations

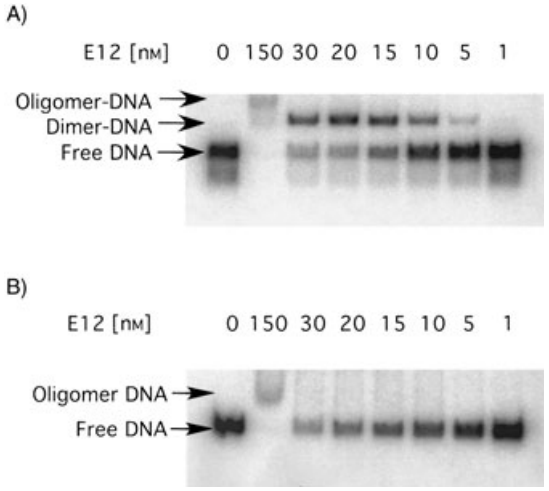


Figure 6. *A) EMSA of the E12 homodimer bound to the TAL1 preferred DNA oligonucleotide. Titration of the protein to a concentration of 150 nM induces binding greater than one dimer–DNA complex. The smearing of the DNA in this lane indicates more than one bound species, characteristic of nonspecific DNA binding. B) EMSA of E12 homodimer titration identical to the one observed in (A). In this case the DNA oligonucleotide, E12UP1, contained a thymine base upstream of the E-box in the first flanking base position. No binding was observed until 150 nM where a smeared lane indicates the presence of unspecific DNA–protein binding.*

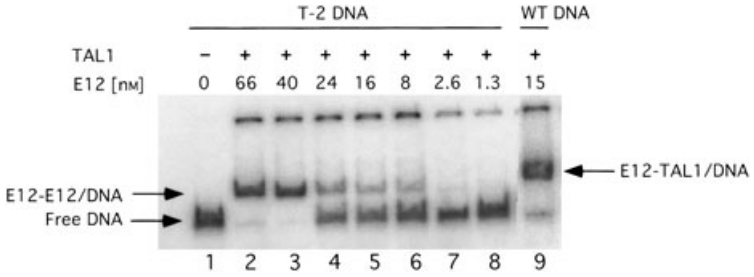
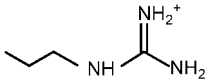
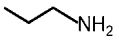
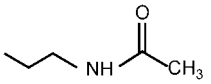
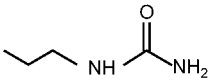


Figure 7. *These EMSAs illustrate the difference in migration of the two dimer–DNA complexes. Lanes 1 to 8 are an EMSA in which only E12 homodimers are observed in the presence of excess TAL1 (150 nM). Lane 1 is DNA only. Lanes 2–8 are E12 monomer titrations to an oligonucleotide, E12HDT1, that contains a thymine base two flanking bases 5′ to the E-box (Table 1). The presence of this thymine base inhibits E12–TAL1 heterodimer formation and only E12 homodimers are observed. Lane 9 is from an EMSA titration to kE2, wild type, DNA oligonucleotide. TAL1 concentration is in excess, 150 nM, and E12–TAL1 heterodimers are observed. E12–TAL1 heterodimers migrate slower in the acrylamide gel allowing for separation of the two dimer–DNA complexes.*

Table 3. *Dissociation constants (K_d) of the E12 homodimers bound to DNA oligonucleotides of varying sequences in the presence of excess TAL1. Only the E-box and flanking sequences are shown for the top strand (5′–3′) oligonucleotide.*

	Oligo	E-box and flanking	E12–E12 [nM]
A	E12HDC1 cytosine in −2 pos.	-CC**CACCTG**CC-	18.2 ± 1.4
B	E12HDC3 cytosine in +2 pos.	-GC**CACCTG**CG-	21.2 ± 1.5
C	E12HDC5 cytosine in both	-CC**CACCTG**CG-	20.1 ± 1.0
D	E12HDT1 thymine in −2 pos.	-TC**CACCTG**CC-	27.7 ± 2.5
E	E12HDT3 thymine in +2 pos.	-CC**CACCTG**CA-	25.6 ± 2.5
F	E12HDT5 thymine in both	-TC**CACCTG**CA-	20.5 ± 1.4

Table 4. *Unnatural amino acid substitutions for R557 in E12. The side-chain structures of arginine, ornithine, citruline, and acelylated ornithine are shown. For all DNA-binding experiments, the kE2 DNA oligonucleotide was used.*

	Amino acid	Side chain	DNA binding
A	arginine		3.0 nM
B	ornithine		>150 nM
C	acetylated ornithine		>150 nM
D	citruline		>150 nM

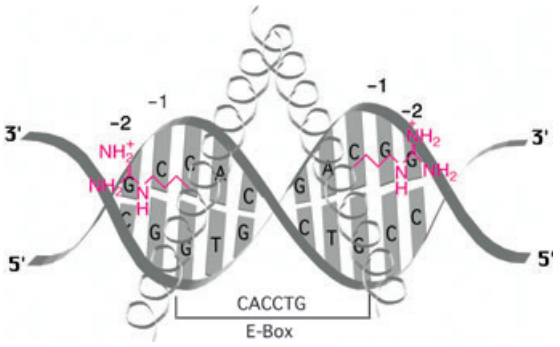


Figure 8. *Cartoon illustration of the basic helix and helix one of two bHLH monomers bound to DNA sequence kE2. The arginine residue conserved in all bHLH proteins and shown to bind outside the E-box is shown in magenta. This arginine stretches across the 5′−1 position to make base specific and phosphate backbone contacts with the 5′−2 base.*

with DNA oligonucleotides yielded no protein binding by either the E12–ornithine or the E12–citruline proteins when using protein concentrations of up to 1 μM. This suggests at least a 100-fold loss in binding affinity by these substitutions. Acetylating the ornithine side chain did not restore binding. These results suggest an important role of the charge and hydrogen-bonding capability of the arginine guanidino group in mediating DNA binding (Figure 8).

Discussion

Our studies demonstrate synthetic access to members of each of the major classes of bHLH DNA binding domains and shed new light on the mechanisms of dimerization specificity amongst bHLH proteins. Our synthetic DNA-binding domains exhibit affinities for their respective DNA targets that are comparable to those obtained with full-length recombinant proteins. The minimal binding domain determined in this study showed a DNA sequence that extended beyond the E-box by two base pairs and the importance of these flanking sequences in dimerization specificity. The molecular basis for homodimerization and heterodimerization among the bHLH proteins must originate within the amino acid sequences of these proteins. An asymmetric binding pattern is observed in the E47 homodimer structure when bound to the E-box, 5′-AA CACCTGGC-3′, with the E-box underlined. An arginine within the basic helix of one E47 monomer makes a contact outside of the E-box to the adenine at position 5′−2 to the E-box (Figure 9). Additionally, selection studies on bHLH–DNA complexes reveal a discrimination against thymine at a 5′−1 position. In order for an arginine in the basic helix of an E47 monomer to make contacts with the adenine two bases outside the E-box, the side chain of arginine must stretch across the first flanking base. Binding assays of both the E12 homodimer and the E12–TAL1 heterodimer revealed an inability of either complex to bind DNA when thymine was located 5′−1 to the E-box. The substitution of uracil in this position completely restores high-affinity binding; this indicates that the thymine methyl group inhibits the correct orientation of the arginine side chain.

The arginine responsible for making contacts outside the E-box is strictly conserved in all bHLH proteins but is absent in bHLHZip. Comparison of structures between Max, a bHLHZip, and E47 (another class A bHLH), reveal a different DNA-binding pattern.[41] While E47 makes contacts outside the E-box, Max does not. Binding affinities are 1.3 nM for the Max homodimer, 270 pM for the Max–Myc heterodimer, and 160 nM for the Myc homodimer to a DNA sequence 5′-GCCACGTGAC-3′.[42] The presence of an adenine immediately 3′ to the E-box positions a thymine on the bottom strand 5′ to the E-box, a position earlier determined to interfere with bHLH–DNA binding.

A water-mediated contact between the conserved arginine in the basic helix of bHLH proteins (R557 in E12) and the N7 of adenine 5′−2 was observed in the E47 and MyoD structures.[24,27] Substitution of guanine for this adenine did not decrease binding affinities of either the E12 homodimer or the E12–TAL1 heterodimer complex. Substitution of cytosine or thymine for adenine decreased E12 homodimer binding twofold. No E12–TAL1 heterodimer binding was observed in titrations of E12 with excess TAL1 to oligonucleotides containing

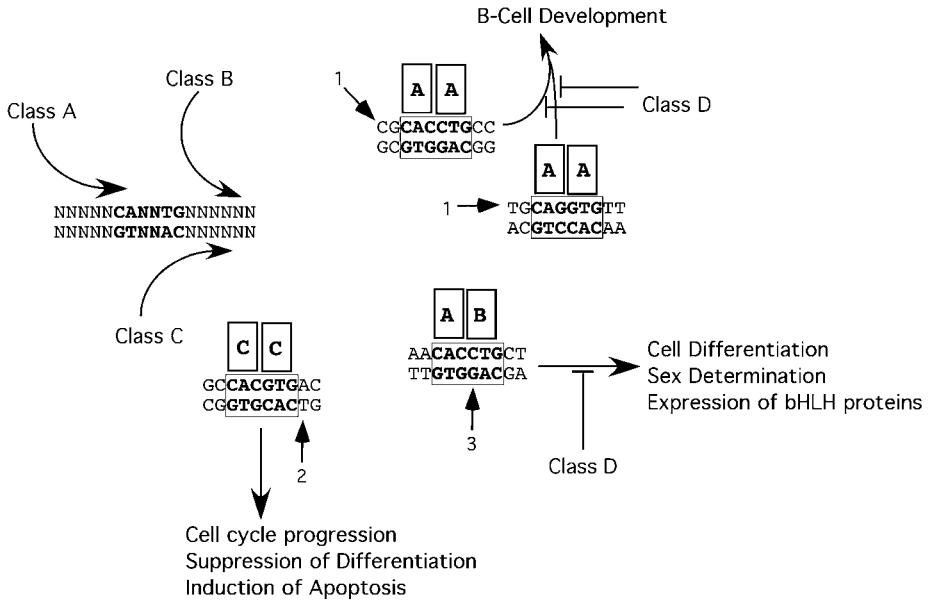


Figure 9. *Schematic of transcriptional regulation imparted through the DNA sequence. In this network, class A, B, and C bHLH proteins are binding to the E-box and flanking sequences shown to be targeted by these specific dimers in vivo. DNA bases within and flanking the E-box that are responsible for specifying dimer binding are shown by arrows. Class A bHLH homodimers preferentially bind over class A–class B heterodimers to E-box flanking sequences containing a cytosine or thymine, arrow 1. Class C dimers preferentially bind over both class A homodimers and class A–class B heterodimers to E-box sequence flanking sequence containing a thymine, arrow 2. Class A homodimers and class A–class B heterodimers preferentially bind over class C dimers to the E-box sequence with varying sequence of the inner two nucleotides, while class C dimers are restricted by the inner base sequences, arrow 3. Both class A homodimers and class A–class B heterodimers are inhibited by the class D proteins. Small changes in binding affinities can have profound consequences for the ultimate read-out of transcriptional responses.*

cytosine or thymine in place of adenine. Instead, only E12 homodimer binding is observed. In the crystal structure of E47, only one monomer is observed to make this base-specific contact through arginine to adenine while the other monomer makes only phosphate contacts with the DNA outside the E-box.[24] Substitution of cytosine or thymine on only one side of the E-box decreased DNA binding. This was comparable to substituting cytosine or thymine in both positions. The structure of E47 is a static representation of the structure, and it is likely that in solution there is a contribution from both monomers that make contacts to the purine ring of either adenine or guanine.

The guanidino side chain group of the conserved arginine residue was shown to be important for DNA binding.[24] Mutagenesis of this arginine residue (R557 in E12) interfered with DNA binding. This further suggests the importance of a specific protein–DNA contact outside the E-box sequence. Examination of the bHLHZip protein sequences showed no conserved arginine residue in a position comparable to the bHLH proteins. The high affinity of bHLHZip complexes for DNA and the absence of the conserved arginine responsible for DNA contacts outside the E-box reveal a different binding pattern for bHLHZip proteins compared to bHLH proteins.

E12 homodimers are able to bind an E-box with cytosine or thymine at 5′−2, albeit with a twofold decrease in affinity. An extra helical turn located at the end of helix one in E12 extends the dimer interface between two E12 monomers as well as providing a salt bridge. E47 has also been shown to be able to bind DNA sequences with a thymine at position 5′−2.[28] This study compared qualitative binding of the E47 homodimer to E-boxes found in both the IgH gene enhancer, μ1–μ5, and in the Ig kappa chain enhancer, kE1-kE3. The binding of E47 to the kE2 sequence, 5′-GCCACCTGCC-3′, was not competed by the addition of μE1, 5′-GTCAAGTGGC-3′, or μE3, 5′-GTCATGTGG-3′; both sequences contain a thymine in the first flanking position. However, the binding of E47 was competed by the addition of μE5, 5′-TGCAGGTGTT-3′, which contains a thymine in the second position flanking the E-box.

The ability of bHLH dimers to discriminate against DNA binding based on the presence of base pairs flanking the E-box affords the transcriptional process another level of regulation (Figure 9). Additionally, the extension of the E-box into a ten-base-pair sequence decreases the number of available sites. The regulation of dimer formation on the DNA by both the sequence of the inner two nucleotides and the flanking sequences allows an E-box to specifically target a dimer pair, thereby activating specific genes. The spatial and temporal expression of both the tissue-specific bHLH and inhibitory HLH proteins, in addition to the higher affinity of a heterodimer complex to DNA, ensures that expression of class B bHLH proteins guarantees the formation of heterodimer complexes. However, the introduction of an E-box sequence capable of preferentially binding a class A homodimer complex in the presence of class B bHLH proteins, targets that E-box for class A homodimers exclusively. Additionally, E-boxes can discriminate between binding of bHLH dimers and bHLHZip proteins.

Experimental Section

Synthesis of E12 and TAL1: The bHLH domain of E12 (amino acids 549–607)[28] (Figure 1a) was manually synthesized by using stepwise solid phase peptide synthesis (SPPS). In situ neutralization BOC chemistry protocols were followed according to Schnolzer et al. and *p*-methylbenzhydrylamine (MBHA) resin was used leaving an amide at the C terminus.[37] Synthesis was carried out for a 0.2 mmol scale by using 1.1 mmol of each Boc-amino acid to couple. Each amino acid was converted into active esters by using 2-(1-H-benzotriazol-1-yl)-1,1,3,3-tetramethyluronium (HBTU, 1.0 mmol), in DMF (0.5 M) and an excess (0.5 mL) of diisopropylethylamine. Coupling times averaged 15 min for each activated

amino acid with a longer coupling time (45 min) for the first amino acid. Due to the large amount of glutamine residues in both TAL1 and E12, dichloromethane (DCM) washes were preformed before and after treatment with trifluoroacetic acid (TFA).[37] Coupling times were increased (30 min) for residues directly proceeding glutamines to allow for resin swelling as well as amino acid coupling. Only one amino acid, Cys 576, seemed to continually present coupling problems. This reside is at the end of helix 1 and is incorporated after a glutamine half way through the synthesis. Proteins containing this cysteine deletion were cleanly separated from product on the reversed-phase high-performance liquid chromatography (RP-HPLC).

The bHLH domain of TAL1 (amino acids 185–244)[43] (Figure 1b) was also prepared on MBHA resin. The synthesis of TAL1 mirrored E12 in conditions and scale. TAL1 does not contain any cysteine residues and had no problematic couplings.

Analogues of the bHLH domain of E12 were synthesized by using a resin-splitting method. In this method, synthesis of E12 on MBHA resin was carried out as usual until the residue C-terminal to arginine, R557, was coupled. After coupling of E558 the resin was split into three separate reaction vessels. Each aliquot of resin was then deprotected with TFA followed by a DMF wash. In one vessel, arginine was coupled for wild type E12, and in another vessel ornithine was coupled, and in the last vessel citruline was coupled. Each of these couplings was allowed to proceed for 30 min. After coupling, the synthesis was carried out in each vessel as for wild-type E12. These three proteins were then cleaved and purified as outlined below. The typical yield from each synthesis was roughly 10%.

Cleavage and purification: The N terminus was acetylated on E12 and TAL1 prior to cleavage from the resin. Cleavage from the resin was carried out in anhydrous HF for 1.5 h at 4 °C in the presence of *p*-cresol (4% *v/v*), which acted as a scavenger. Following cleavage, peptides were precipitated with ice-cold ether and dissolved in acetic acid (50% *v/v*). Synthetic product polypeptides were purified by semipreparative RP-HPLC by using a linear gradient of 30–60% buffer B (90% acetonitrile, 0.1% TFA in H_20) over 60 min at a flow rate of 10 mL min^{-1}. Fractions were analyzed by electrospray ionization mass spectrometry (PE Sciex API-III). Peptide masses were calculated from the experimental *m/z* from all of the observed protonation states of the peptide. Fractions with the correct mass were pooled and lyophilized.

Protein folding: Pure lyophilized material of E12 and TAL1 was dissolved in 50% acetic acid. A Pharmacia PD-10 column was then used to separately exchange both proteins into storage buffer (20 mM Hepes, pH 6.8, 25 mM ammonium sulfate, 100 mM KCl, 1 mM EDTA). Protein fractions were collected and in all cases the proteins were eluted from the size exclusion columns before either guanidine (6 M) or acetic acid (50% *v/v*). Fractions containing protein were determined by using Bio-Rad protein stain. These fractions were then centrifuged to pellet-aggregated protein and the supernatants pooled. 100–150 nmol was then removed for amino acid analysis. DTT (10 mM) and glycerol (10% *v/v*) were added to each sample. These samples were then dispensed into aliquots and stored at −80 °C. All protein concentrations were determined by amino acid analysis carried out by AAA Laboratories (Seattle WA).

DNA labeling: The oligonucleotides used for gel shift assays were obtained from Genosys Biotechnologies, Inc. The "wild type" oligonucleotide sequence used was based on the immunoglobulin kappa chain enhancer with the E-box underlined.[28] The sequence of the top strand is: 5′-TCGAACTGGCCACCTGCCTGGATC-3′ (Figure 1). Complimentary single-stranded oligonucleotides were labeled with [χ-^{32}P] ATP in kinase buffer (70 mM Tris-HCl pH 7.6, 10 mM $MgCl_2$, 5 mM DTT) with 10 units of polynucleotide kinase at 37 °C for 1.5 h. Oligonucleotides were then precipitated in 2.5 M ammonium acetate, 20 μg glycogen, and 4 volumes of ethanol. Labeled oligonucleotides were then resuspended in annealing buffer (100 mM Tris-HCl pH 7.6, 10 mM $MgCl_2$). Complimentary labeled oligonucleotides were annealed as follows: 3 min at 95 °C, 10 min at 70 °C, 10 min at 37 °C, and 10 min at ambient temperature. Labeled double-stranded oligonucleotides were separated from single-stranded oligos on a 15% nondenaturing polyacrylamide gel. Double-stranded oligonucleotides were then eluted from the polyacrylamide gel, overnight at 37 °C. Purified labeled double-stranded oligonucleotides were resuspended in TE and stored at −20 °C.

Electrophoretic mobility shift assays (EMSA): Radiolabeled double-stranded oligonucleotides were equilibrated in the binding buffer (20 mM Hepes, pH 8, 100 mM KCl, 1 mM EDTA, 5% glycerol, and 0.01% NP40). DNA concentrations used ranged from 30–60 pM. For protein titrations, serial dilutions of protein stock solution were equilibrated with the DNA in assay buffer and bovine serum albumin (BSA, 2.5 μg μL^{-1}) for 30 min at ambient temperature in 20 μL aliquots. Samples were then loaded onto 10% nondenaturing polyacrylamide Bio-Rad minigel and run for 25 min at 125 V cm^{-1} in Tris-borate buffer (88 mM), pH 8.3. E12 was titrated as above.

E12–TAL1 titrations were carried out in two different ways. TAL1, in dilution buffer (20 mM Hepes, pH 6.8, 25 mM ammonium sulfate, 100 mM KCl, 1 mM EDTA, 0.01% NP40), was equilibrated with the DNA for 30 min. TAL1 concentration was held constant. E12, also diluted into the same buffer, was titrated into the TAL1–DNA complex and also allowed to equilibrate for 30 min. Alternatively, E12 was diluted and titrated with the DNA and allowed to equilibrate for 30 min, TAL1 was then added to each sample in a constant concentration and equilibrated for 30 min. No difference was observed when either TAL1 or E12 was added to the DNA first; therefore, in the presented data E12 was equilibrated with DNA and then TAL1 was added.

Radioactivity in a given band was quantitated with a Molecular Dynamics Phosphoimager System by using ImageQuant software. Quantitated data were plotted as fraction DNA bound versus total titrated monomer E12 protein concentration.

Apparent dissociation constants were determined with the following equation by using KaleidaGraph software:

$$f = \frac{[P]^n}{[P]^n + K_d{}^n}$$

where n is the cooperativity coefficient, K_d is the apparent dissociation constant, [P] is the monomer concentration of E12, and f is the fraction of DNA bound. All dissociation constants reported herein are the average of several determinations, and standard deviations are given.

Primer extension assay: DNA templates were designed to contain the kE2 E-box and flanking sequences (Figure 2). Complementary 15 base primers were designed to bind to these oligonucleotides at the 3′ ends. These DNA fragments and primers were obtained from Genosys Biotechnologies, Inc. The primers were purified by RP-HPLC and the DNA templates desalted. Primers were labeled with [χ −^{32}P] ATP. Labeled primers were then resuspended in TE buffer.

Labeled primers (2 pmol μL^{-1}) were annealed to the templates (5 $\mu g\,\mu L^{-1}$) in Sequenase buffer (40 mM Tris-HCl, pH 7.5, 20 mM $MgCl_2$, 5 mM NaCl) for 5 min at 95 °C, 1 min at ambient temperature and then transferred to ice. Annealed primer was extended 5′ to 3′ in the above buffer, which contained each dNTP (33 μM), each ddNTP (3 μM), DTT (10 mM), and 2.0 units of Sequenase (USB Sequenase v. 2.0). This extension was allowed to react for 10 min at 37 °C. The DNAs were then precipitated with ethanol and DNA fragments were resuspended in TE buffer and stored at −20 °C.

EMSAs were used to separate the DNA–protein complexes and unbound DNA from the primer-extension generated DNA library. EMSA conditions for the binding were the same as outlined before with the exception that the reactions were run by using total DNA concentration of approximately 0.014 pmol per 20 μL reaction. E12 homodimers or E12–TAL1 heterodimers were allowed to bind the double-stranded DNA libraries and then separated on a 10% nondenaturing polyacrylamide gel run for 1.5 h at 223 $V\,cm^{-1}$. To determine the percentage DNA bound, radioactivity in a given band was quantitated with Molecular Dynamics PhosphorImager by using ImageQuant software. Bands corresponding to complexed and uncomplexed double-stranded DNAs were excised and soaked overnight at 37 °C in SDS stop buffer (25 mM Tris-HCl, 250 mM NaCl, 12.5 mM EDTA, and 0.05% SDS). The DNA fractions were then purified by phenol extraction (twice), ethanol precipitation, and resuspended in TE.

Loading buffer (95% formamide, 20 mM EDTA, 0.05% bromophenol blue, and 0.05% xylene cyanol) was added to the DNA samples from above. These were then incubated for 3 min at 70 °C and loaded onto a prerun (1500 $V\,cm^{-1}$) 12% denaturing polyacrylamide sequencing gel and resolved for 2.5 h (Figure 4). Radioactivity in each band was quantified as before. These data were then plotted as the natural log of the ratio between bound and total DNA to the nucleotide position for chain termination (Figure 5).

$$\ln\left(\frac{DNA_{bound}}{DNA_{total}}\right) = 0$$

indicates that the base pair is not required for binding. The more negative the number the greater the requirement for that base pair.

Acknowledgements

This work was supported by a grant from the National Institutes of Health (GM 45730).

Keywords: basic helix-loop-helix protein · DNA recognition · DNA–protein interaction · solid-phase synthesis · transcription factor

[1] W. R. Atchley, W. M. Fitch, *Proc. Natl. Acad. Sci. USA* **1997**, *10*, 5172–5176.
[2] F. Guillemot, L. C. Lo, J. E. Johnson, A. Auerbach, D. J. Anderson, A. L. Joyner, *Cell* **1993**, *75*, 463–476.
[3] C. Porcher, W. Swat, K. Rockwell, Y. Fujiwara, F. W. Alt, S. H. Orkin, *Cell* **1996**, *86*, 47–57.
[4] L. Robb, C. G. Begley, *Int. J. Biochem. Cell Biol.* **1996**, *28*, 609–618.
[5] H. Weintraub, *Cell* **1993**, *75*, 1241–1244.
[6] C. Murre, G. Bain, M. A. van Dijk, I. Engel, B. A. Furnari, M. E. Massari, J. R. Matthews, M. W. Quong, R. R. Rivera, M. H. Stuiver, *Biochim. Biophys. Acta* **1994**, *1218*, 129–135.
[7] C. Murre, P. S. McCaw, H. Vaessin, M. Caudy, L. Y. Jan, Y. N. Jan, C. V. Cabrera, J. N. Buskin, S. D. Hauschka, A. B. Lassar, H. Weintraub, D. Baltimore, *Cell* **1989**, *58*, 537–544.
[8] A. F. Voronova, F. Lee, *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 5952–5956.
[9] H. L. Hsu, J. T. Cheng, Q. Chen, R. Baer, *Mol. Cell. Biol.* **1991**, *11*, 3037–3042.
[10] H. Wendt, R. M. Thomas, T. Ellenberger, *J. Biol. Chem.* **1998**, *273*, 5735–5743.
[11] X.-H. Sun, D. Baltimore, *Cell* **1991**, *64*, 459–470.
[12] H.-L. Hsu, I. Wadman, R. Baer, *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 3181–3185.
[13] R. Benezra, R. L. Davis, D. Lockshon, D. L. Turner, H. Weintraub, *Cell* **1990**, *61*, 49–59.
[14] K. Langlands, X. Yin, G. Anand, E. V. Prochownik, *J. Biol. Chem.* **1997**, *272*, 19785–19793.
[15] Y. Jen, H. Weintraub, R. Benezra, *Genes Dev.* **1992**, *6*, 1466–1479.
[16] X. H. Sun, N. G. Copeland, N. A. Jenkins, D. Baltimore, *Mol. Cell. Biol.* **1991**, *11*, 5603–5611.
[17] L. Robb, N. J. Elwood, A. G. Elefanty, F. Kontgen, R. Li, L. D. Barnett, C. G. Begley, *EMBO J.* **1996**, *15*, 4123–4129.
[18] C. G. Begley, P. D. Aplan, S. M. Denning, B. F. Haynes, T. A. Waldmann, I. R. Kirsch, *Proc. Natl. Acad. Sci. USA* **1989**, *86*, 10128–10132.
[19] A. R. Green, E. Salvaris, C. G. Begley, *Oncogene* **1991**, *6*, 475–479.
[20] P. D. Aplan, K. Nakahara, S. H. Orkin, I. R. Kirsch, *EMBO J.* **1992**, *11*, 4073–4081.
[21] P. J. Quesenberry, N. N. Iscove, C. Cooper, G. Brady, P. E. Newburger, G. S. Stein, J. S. Stein, G. P. Reddy, S. Pearson-White, *J. Cell. Biochem.* **1996**, *61*, 478–488.
[22] G. Condorelli, L. Vitelli, M. Valtieri, I. Marta, E. Montesoro, V. Lulli, R. Baer, C. Peschle, *Blood* **1995**, *86*, 164–175.
[23] Y. Xia, L. Y. Hwang, M. H. Cobb, R. Baer, *Oncogene* **1994**, *9*, 1437–1446.
[24] T. Ellenberger, D. Fass, M. Arnaud, S. C. Harrison, *Genes Dev.* **1994**, *8*, 970–980.
[25] A. R. Ferré-D'Amaré, P. Pognonec, R. G. Roeder, S. K. Burley, *EMBO J.* **1994**, *13*, 180–189.
[26] A. R. Ferré-D'Amaré, G. C. Prendergast, E. B. Ziff, S. K. Burley, *Nature* **1993**, *363*, 38–45.
[27] P. C. M. Ma, M. A. Rould, H. Weintraub, C. O. Pabo, *Cell* **1994**, *77*, 451–459.
[28] C. Murre, P. S. McCaw, D. Baltimore, *Cell* **1989**, *56*, 777–783.
[29] T. K. Blackwell, H. Weintraub, *Science* **1990**, *250*, 1104–1110.
[30] K. A. Gould, E. H. Bresnick, *Gene Expression* **1998**, *7*, 87–101.
[31] S. J. Maleki, C. A. Royer, B. K. Hurlburt, *Biochemistry* **1997**, *36*, 6762–6767.
[32] M. Sieber, R. K. Allemann, *Biol. Chem.* **1998**, *379*, 731–735.
[33] P. Bishop, I. Ghosh, C. Jones, J. Chmielewski, *J. Am. Chem. Soc.* **1995**, *117*, 8283–8284.
[34] P. Bishop, C. Jones, I. Ghosh, J. Chmielewski, *Int. J. Pept. Protein Res.* **1995**, *46*, 149–154.
[35] R. L. Winston, D. P. Millar, J. M. Gottesfeld, S. B. H. Kent, *Biochemistry* **1999**, *38*, 5138–5146.
[36] R. L. Winston, J. M. Gottesfeld, *Chem. Biol.* **2000**, *7*, 245–251.
[37] M. Schnölzer, P. Alewood, A. Jones, D. Alewood, S. B. H. Kent, *Int. J. Pept. Protein Res.* **1992**, *40*, 180–193.
[38] H.-L. Hsu, L. Huang, J. T. Tsan, W. Funk, W. E. Wright, J.-S. Hu, R. E. Kingston, R. Baer, *Mol. Cell. Biol.* **1994**, *14*, 1256–1265.
[39] H.-L. Hsu, I. Wadman, J. T. Tsan, R. Baer, *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 5947–5951.
[40] H. N. Liu-Johnson, M. R. Gartenberg, D. M. Crothers, *Cell* **1986**, *47*, 995–1005.
[41] S. K. Nair, S. K. Burley, *Cell* **2003**, *112*, 193–205.
[42] W. Fieber, M. L. Schneider, T. Matt, B. Krautler, R. Konrat, K. Bister, *J. Mol. Biol.* **2001**, *307*, 1395–1410.
[43] P. D. Aplan, C. G. Begley, V. Bertness, M. Nussmeier, A. Ezquerra, J. Coligan, I. R. Kirsch, *Mol. Cell. Biol.* **1990**, *10*, 6426–6435.

Received: June 3, 2004

Early View Article
Published online on December 9, 2004

