Differences and Similarities in DNA-Binding Preferences of MyoD and E2A Protein Complexes Revealed by Binding Site Selection

T. Keith Blackwell and Harold Weintraub

A technique was developed for studying protein-DNA recognition that can be applied to any purified protein, partially purified protein, or cloned gene. From oligonucleotides in which particular positions are of random sequence, that subset to which a given protein binds is amplified by the polymerase chain reaction and sequenced as a pool. These selected and amplified binding site (SAAB) "imprints" provide a characteristic set of preferred sequences for protein binding. With this technique, it was shown that homo- and heterooligomers of the helix-loop-helix proteins MyoD and E2A recognize a common consensus sequence, CA- -TG, but otherwise bind to flanking and internal positions with different sequence preferences that suggest half-site recognition. These findings suggest that different combinations of dimeric proteins can have different binding sequence preferences.

Expression of MyoD can induce myogenesis and expression of muscle-specific genes in various cell types (*1*). Within MyoD (and other related myogenic gene products) is a conserved group of basic amino acids adjacent and amino-terminal to residues that are proposed to form a helix-loop-helix (HLH) structure (*2*). This basic-HLH (bHLH) domain is necessary and sufficient for myogenic conversion and it defines a large family of proteins, including many that regulate differentiation of specific cell lineages (*3*). Through the HLH domain, these proteins can form dimers with themselves and with related family members (*2–5*). Many bHLH proteins appear to directly regulate gene expression by binding to specific DNA sequences. For example, MyoD and other myogenic gene products bind in vitro to sites in the regulatory regions of muscle-specific genes (*6–9*). The binding of bHLH proteins to DNA seems to require oligomerization (*2, 4, 5*), which is thought to position and orient the 13–amino acid basic regions from each protomer so that they make specific contacts with DNA (*5*).

The binding sites that have been identified for bHLH proteins, including one from yeast (*10*), contain a consensus CA- -TG motif (*6*) that is present in the regulatory regions of many tissue-specific genes (*8, 9, 11–13*). Together with the product of the widely expressed bHLH gene *E2A* (*2, 14*), MyoD or the analogous bHLH inducers of the peripheral nervous system—the products of the *achaete-scute* genes (*15*)—bind in vitro to the same muscle-specific sequences, as well as to related immunoglobulin enhancer sequences (*4*). Thus, it seems paradoxical that different bHLH proteins apparently can bind similar DNA sequences and yet act in vivo on tissue-restricted sets of genes (*4, 5*). It is therefore crucial to determine whether there might be important differences in the sequence-specificity with which MyoD and other bHLH proteins recognize DNA. In addition, very little is known about how the basic regions of bHLH proteins contact DNA. This issue is of particular importance for MyoD because its basic region seems to mediate not only DNA binding, but also the subsequent transcriptional activation of muscle-specific genes (*5*).

A number of biochemical assays that involve protection of bound DNA or interference with the ability to bind DNA have provided important information as to how and where specific proteins bind. Mutational analyses have revealed much about the sequence specificity of binding, but a comprehensive study requires that a large number of mutants be generated and separately characterized. More recently, the range of information available from mutational analyses has been expanded by use of random-sequence mutagenesis (*16, 17*). For example, sequences within a protein-binding site (*16*) or protein (*17*) that function in a biological or biochemical assay are selected from a pool of random DNA sequence. Such studies can yield significant information about protein-binding sites in DNA but, again, require isolation and sequencing of a large number of mutants.

We have now developed a strategy for studying the DNA sequence specificity of protein-DNA binding that is based on random-sequence selection, but also allows simultaneous analysis of bound sequences as a population (Fig. 1A). From oligonucleotides in which specific binding site positions are random in sequence, those that are bound are isolated in an electrophoretic mobility shift assay (EMSA), amplified by the polymerase chain reaction (PCR), reiteratively re-bound and reamplified, and finally sequenced directly as a pool. The nucleotide sequence patterns of these "selected and amplified binding sites" (SAAB's) provide a characteristic "imprint" of protein binding. We have used SAAB imprints to compare and contrast the sequence-preferences of DNA binding by MyoD and E2A homo- and heterooligomers. Our results indicate that these different protein species preferentially bind the consensus CA- -TG motif, but select different sequence patterns at internal and surrounding positions. Their SAAB imprints suggest that in bHLH complexes individual protomers recognize half-sites on the DNA and, therefore, that the combinatorial interactions between these proteins can, in fact, define new binding preferences. Subtle nucleotide sequence differences among bHLH protein–binding sites may

The authors are in the Department of Genetics, Fred Hutchinson Cancer Research Center, and at the Howard Hughes Medical Institute, Seattle, WA 98104. Correspondence should be addressed to H. Weintraub.

thus be more important than previously realized. In addition, we have shown that proteins produced by reticulocyte lysates programmed with in vitro transcribed RNA can yield high-quality SAAB imprints, opening up the possibility that proteins mutagenized in vitro can be rapidly screened by this type of analysis.

Selection of the CA--TG consensus by MyoD homooligomers. We initially attempted to determine whether, in the SAAB assay, specification of one half of the palindromic consensus CA--TG can define the other half. The binding templates were based on a site in the muscle creatine kinase (MCK) enhancer (*11*) (Fig. 1B) that is required for its activation and to which MyoD and E2A protein complexes bind in vitro (*4–6*). Dimers of this site specifically direct cotransfected MyoD-dependent expression of a reporter gene in a variety of cell types (*18*). In the D1 and D2 templates, one-half of the consensus and core is specified and the other half is random (Fig. 1B). In the D2 template, position +4 is also random because methylation-interference studies suggest that this position is contacted by MyoD and E12-MyoD oligomers (*4, 6*) (E12 and E47 are differentially spliced E2A proteins).

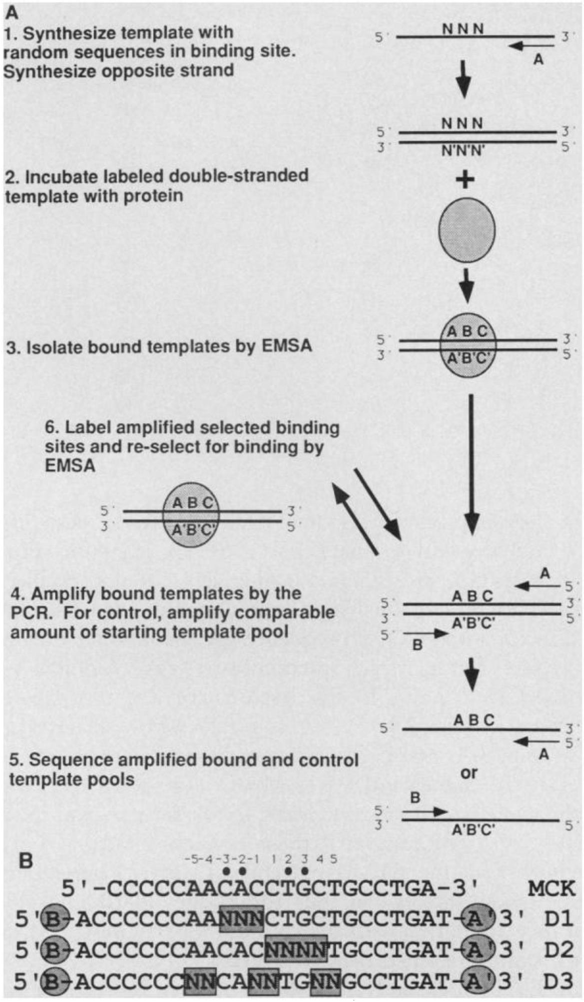

A
1. Synthesize template with random sequences in binding site. Synthesize opposite strand
5'
NNN
3'
A
5'
3'
NNN
N'N'N'
3'
5'
2. Incubate labeled double-stranded template with protein
+
3. Isolate bound templates by EMSA
ABC
A'B'C'
6. Label amplified selected binding sites and re-select for binding by EMSA
ABC
A'B'C'
4. Amplify bound templates by the PCR. For control, amplify comparable amount of starting template pool
A
ABC
A'B'C'
B
5. Sequence amplified bound and control template pools
ABC
A
or
B
A'B'C'
B
-5-4-3-2-1 1 2 3 4 5
5'-CCCCCAACACCTGCTGCCTGA-3' MCK
5'B-ACCCCCCAANNNNCTGCTGCCTGAT-A'3' D1
5'B-ACCCCCCAACACNNNNNTGCCTGAT-A'3' D2
5'B-ACCCCCCNNCANNTGNNGCCTGAT-A'3' D3

Fig. 1. Protocol for the SAAB imprint assay. (**A**) An oligonucleotide is synthesized with random sequences (designated as N) at positions of interest in the protein-binding site, which is flanked by sequences that correspond to primers A and B (*30*). A labeled double-stranded template generated from this oligonucleotide is incubated with the binding protein (or proteins). Bound templates are isolated by EMSA and then amplified by PCR with primers A and B. As a control for fidelity of template synthesis and amplification, the starting template population is similarly analyzed. The nucleotide sequences of the bound and starting template populations are determined with the use of either primer A or primer B. The background from nonspecific protein-DNA binding can be decreased and the stringency of selection for specific binding can be increased by subjecting the bound template pools to multiple additional rounds of selection, in which they are labeled and reselected for binding to the same protein preparation in an EMSA and then amplified as before. After the desired number of rounds of binding selection, the nucleotide sequence preferences of the bound population are determined as before. (**B**) Core sequences of the random sequence oligonucleotide templates D1, D2, and D3. The 22-bp core sequences of D1, D2 and D3 are based on the 3′ (right) MyoD–binding site of the MCK enhancer (*11*) (read 5′ to 3′ toward the MCK promoter). The sequences shown are flanked on the 5′ end by primer B and on the 3′ end by the complement of primer A (indicated as A′) (*30*). The consensus CA--TG motif is indicated by dots. Nucleotides in and around this motif are assigned positions so that −1 and +1 represent the center.

Bacterially produced glutathione-*S*-transferase–MyoD fusion protein (GluMyoD) binds to D1 and D2 to a proportionally much smaller extent than to the wild-type MCK site (Fig. 2A), which suggests that only a subset of D1 or D2 is bound with high affinity. The D1- and D2-GlyMyoD SAABs (Fig. 2B) show preferences for CA and TG at positions −3, −2 and +2, +3, respectively, and the D2-GluMyoD sequence reveals an additional T preference at position +4. These preferences are derived from only one round of EMSA selection and at most positions are not absolute, but they nevertheless confirm the importance of the CA--TG motif (*19*).

Binding of MyoD and E2A proteins with different sequence preferences within the context of the CA--TG consensus. We used the D3 template, in which the CA--TG consensus is specified but surrounding positions are random in sequence (Fig. 1B), to determine whether bHLH proteins have additional binding sequence preferences. Binding of the MCK, D2, and D3 templates by various complexes of MyoD and E12 and E47 (*2*), which were synthesized by in vitro translation, was investigated (Fig. 3, A and B). With the MCK template, multiple complexes that consisted of dimers and higher-order oligomers were formed in most cases (*5*). A much smaller relative fraction of the D2 and D3 templates were bound, and significant background signal was present in each lane.

However, the SAAB's derived from three successive rounds of selection and amplification (Fig. 3C) were bound by their respective complexes to a greater relative extent, and with less nonspecific background. Each SAAB was generated by successive selection for binding by a single complex (Fig. 3B) but was bound by the same set of multimeric complexes evident in Fig. 3B (that is, the initial binding pattern is serially regenerated), indicating that the members of each set of complexes can bind the same DNA sequences. Some of these SAAB's were also bound by specific but unidentified factors present in the reticulocyte lysate (Fig. 3C, lanes 3, 5, 9, and 13). Binding by a lysate factor is most striking in the D3-E12 SAAB (Fig. 3C, lanes 3 and 4); the proportion of binding is the same in the control lane, which suggests that this SAAB was derived primarily from the lysate factor.

These SAAB sequences were determined (Fig. 3, D and E) and compiled (Fig. 4). Under our experimental conditions, the sequence preferences derived from the first selection round (*20*)—when the proteins were in excess of their potential sites and all possible sequences were bound—were usually less specific than those derived from the third round (Fig. 3, D and E). After multiple selection rounds, when potential binding sites appear to be in excess and only a fraction of the labeled templates are bound (Fig. 3C), it is likely that only the binding sequences with the highest affinities are selected. Hence, this technique, as we have applied it, really provides the most preferred binding sites, and it would be unwise to conclude that a given sequence necessarily cannot bind with reasonable affinity if its sequence does not match the preferred one. Clearly, the technique can be used to detect sequences that bind with less affinity if the protein is maintained in excess at each round of selection; however, this is difficult to achieve with in vitro translated products.

When confronted with the symmetrical arrangement of random

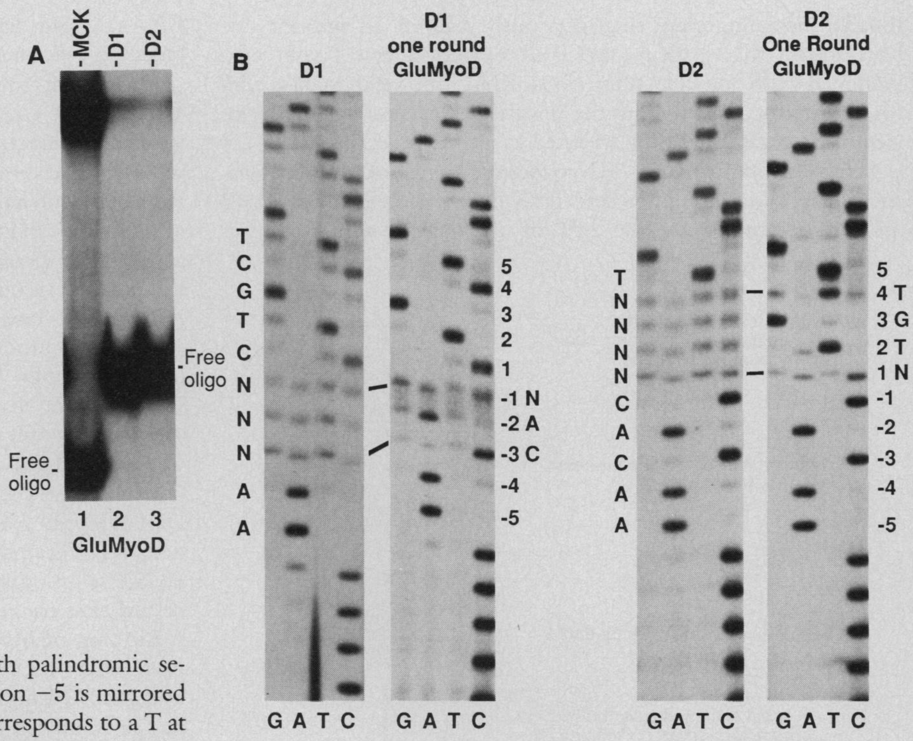


Fig. 2. SAAB imprints of GluMyoD binding to templates D1 and D2. (**A**) EMSA of GluMyoD fusion protein binding to the indicated templates (*31*). (**B**) The sequences of the starting templates D1 and D2 are compared with those selected in (A) for binding (*32*). D1 and D2 control sequences were also determined on amplified DNA. A sample (1 pg) of each control template was amplified for 35 cycles, and the product was purified (*32*) for sequencing. For D2 [and D3 (Fig. 3)], 1 pg of this amplified DNA was again amplified by 35 cycles before sequencing (*33*). Sequences were determined by the dideoxy method with a modified version of the Sequenase (U.S. Biochemical) protocol (*34*). Those shown were determined with primer B (*30*), which gives fewer compression artifacts than primer A and yields the sequence of the DNA strand shown in Fig. 1B. Lanes correspond to G, A, T, and C termination reactions are shown. Binding site positions and random sequences are designated as in Fig. 1. Experimentally derived sequence preferences are indicated at appropriate site positions. Lines connect corresponding positions in different samples.

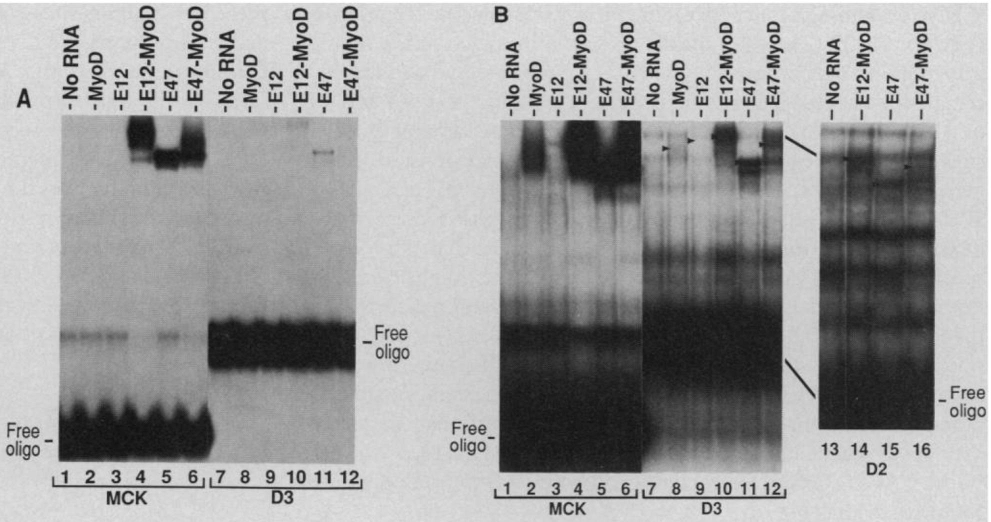
bases in D3, MyoD homooligomers bound with palindromic sequence preferences. A purine preference at position −5 is mirrored by a pyrimidine at +5, an A preference at −4 corresponds to a T at +4, and the −1 and +1 preferences are G and C, respectively (Fig. 3D and Fig. 4) (*21*). Identical sequence preferences were obtained with GluMyoD (*20*), indicating that they are not influenced by other proteins present in the reticulocyte lysate. The palindromic character of the preferred sequences is suggestive of symmetrical binding by MyoD homooligomers and indicates that defined sequences on the template do not influence the orientation of MyoD binding. The results are also consistent with the MyoD complex, itself, being symmetrical. However, conclusive proof that the MyoD complex is symmetrical would require determination that each individual preferred site is, itself, symmetrical. GluMyoD bound to a template with a sequence that corresponds to the D3-MyoD SAAB (CCCCCAACAGCTGTTGCCTGA) (Fig. 4) (*20*) at least as well as it bound to the MCK template (Fig. 1B), confirming that the SAAB protocol identified high-affinity binding sites.

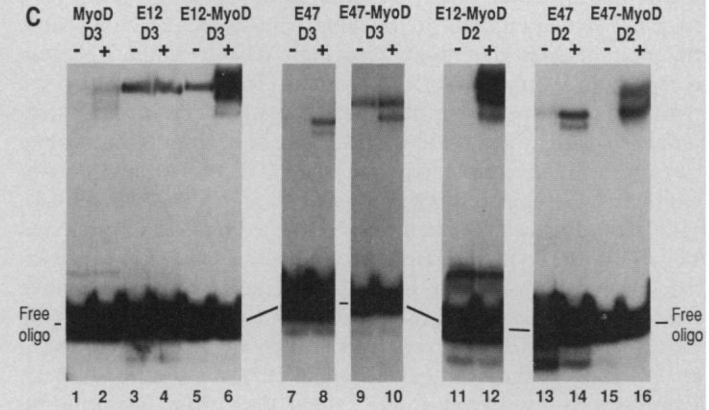
In contrast, although E47 homooligomers selected the palindromic CA−−TG motif from D2 (Figs. 3E and 4), they bound asymmetrically at certain other positions. At −1 and +1, and asymmetrical CC pair predominated (*22*), indicating binding in a preferred orientation relative to the rest of the template. Although the lack of a T at position −4 is complemented by a symmetrical lack of an A at +4 (Fig. 3, D and E, and Fig. 4), additional asymmetry in E47 binding is suggested by the near lack of an A nucleotide at +5, but not at −5 (Figs. 3D and 4). Selection against these A residues might actually derive from interference with binding by the 5-methyl group of the T on the other strand. Confirming the effect of these bases on binding, a template that is identical to MCK (Fig. 1B) except for a T at −4 and an A at +4 was not bound at all by E47 homooligomers (*20*). The unexpected asymmetry of binding sites selected by the E47 homooligomers, which we assume are symmetrical, suggests that sequences distal to positions +5 and −5 direct the orientation of asymmetrical binding on the template DNA. Presumably, as a result of interaction with these distal sequences, the protein "responds" (that is, changes conformation) by choosing an asymmetrical set of sequences at some positions, particularly in the center of the site. An alternative explanation is that the E47 complex, itself, can adopt an asymmetrical conformation.

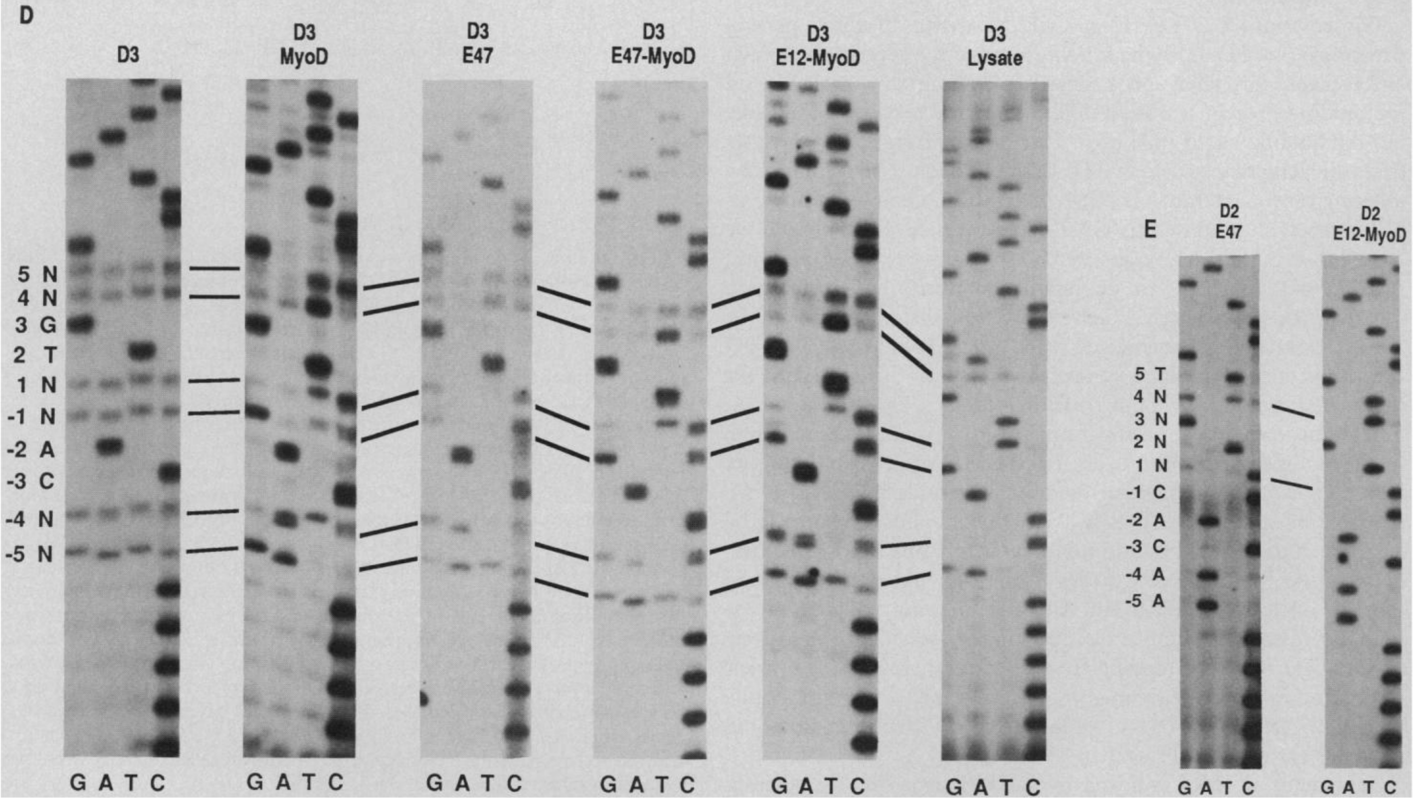
These experiments did not permit an evaluation of binding by E12 homooligomers. The sequence preferences (*20*) of templates present in the appropriate region of a D3–"No RNA" lane (Fig. 3B, lane 7) are identical to the D3-E12 SAAB sequence (Fig. 3D, labeled as lysate), except for a higher level of nonspecific background. This finding further indicates that the D3-E12 SAAB's were derived primarily from sequence-specific binding by a reticulocyte lysate factor, which predominates over binding by the translated E12. Binding by this lysate factor was also apparent to some extent in the E47-MyoD and E12-MyoD heterooligomer SAAB's (Fig. 3, D and E), as indicated by a GT pair at −1 and +1 (Fig. 4) (*23*). Significantly, E12-MyoD heterooligomers did not bind the lysate SAAB, and the lysate factor did not bind the MCK template (*20*), indicating that these respective factors contribute independently to the E47-MyoD and E12-MyoD heterooligomer SAAB's. Lysate factors can thus complicate SAAB imprints, but only if they comigrate with the complexes of interest; and, in any case, the unprogrammed lysate provides a control.

Both E47-MyoD and E12-MyoD heterooligomers select the CA−−TG motif from D2 (Fig. 3E and Fig. 4) but otherwise bind asymmetrically. At positions −1 and +1, both select an asymmetrical CC pair (Fig. 3, D and E, and Fig. 4). Most strikingly, their preferences at −4 and −5 are almost identical to those of the E47 homooligomer, and at +4 and +5 are similarly close to those of MyoD homooligomers (Fig. 3, D and E). This similarity to the corresponding MyoD preference at +5 is especially apparent in the D3–E12-MyoD SAAB's (*23*). The sequence preference patterns of MyoD and E47 homooligomers at positions −4 and +4 and −5 and +5 thus give a diagnostic "imprint" of binding by those particular proteins, and indicate that in heterooligomeric complexes the respective DNA-binding regions of the MyoD and E2A proteins bind to distinct half-sites (Fig. 4). This suggests that the basic regions from each monomer do not interdigitate in some way to present a common binding surface; instead, each seems to have its own target. The asymmetry of these heterooligomer imprints indicates further that these complexes are binding in a defined orientation relative to the rest of the template, and thus that base pairs more distal to the core than positions −5 and +5 are having a significant effect on binding.

Fig. 3. SAAB imprints of binding by HLH proteins synthesized by translation in vitro. (**A**) EMSA of binding of the indicated protein complexes (*35*) to the MCK and D3 templates. "No RNA" indicates an assay of binding by the products of an in vitro translation reaction to which no RNA had been added. (**B**) A longer exposure of the gel shown in (A), and a similar exposure of a gel showing the binding of the indicated complexes to D2. Black arrowheads designate where templates were isolated from the gel (*32*). The location of the excised area was determined on the basis of either the appearance of a binding activity above background in the corresponding "No RNA" lane or the mobility of the desired complex bound to the MCK template (or both). (**C**) EMSA of template pools that have undergone three rounds of binding and selection by EMSA. Template pools isolated in (B) were amplified and labeled by PCR (*36*), and then subjected to EMSA with the protein complex with which they had been selected initially. In each selection round, the protein-DNA complex indicated in (B) was isolated. The SAAB's that were thus assayed in (C) had undergone two such rounds of EMSA selection and amplification subsequent to their initial isolation in (B). The protein complexes used in EMSA selection are indicated above the lanes. "No RNA" control reactions for each template pool are designated as −, and reactions containing the indicated protein complex as +. (**D**) Nucleotide sequences of the D3 SAAB's that were assayed in (C). The D3 control template was amplified and all sequencing was performed as described in Fig. 2B. The sequence derived by selection with E12 is designated "lysate" because it was derived primarily from binding by a reticulocyte lysate factor. Binding site positions are indicated as in Figure 1B. (**E**) Nucleotide sequences of the indicated D2 SAAB's [also described in (C)]. The D2–E47-MyoD SAAB imprint is not shown because it is identical to that of E12-MyoD.







MyoD half-sites are present in muscle-specific regulatory regions. Of 22 CA– –TG motifs present in muscle-specific regulatory regions that were examined (*24*), nucleotides that meet the criteria for a MyoD half-site are present at −5 and −4 or +4 and +5 in 11 (50 percent), including MCK. In contrast, among 31 such sites present in regulatory regions of surveyed non-muscle cellular genes (*24*), these half-sites are present in only four (13 percent), or at about the statistically predicted frequency of one in eight. These observations are more striking if it is considered that the muscle-specific sites that do not fit the MyoD consensus might be recognized by the products of other myogenic determination genes—such myogenin, Myf5, or MRF4-Myf6-herculin (*8, 25*)—that could potentially have slightly different binding specificities; moreover, some of these sites might not be used by muscle-specific factors. Of potential additional significance, a GCTG motif that frequently overlaps or flanks muscle-specific CA– –TG elements (*6, 24*) is part of the CAGCTG consensus identified by the MyoD homooligomer SAAB imprints (Fig. 4).

Applications of the SAAB imprint assay. By combining the power of random-sequence selection with pooled sequencing, the SAAB imprint assay makes possible simultaneous screening of a large number of binding site mutants. This technique allows identification of sites that bind with high relative affinity, because competition is inherent in the protocol. It can also identify site positions that are neutral, or specific bases that can interfere with binding, such as a T at −4 in the E47 half-site (Fig. 4). When several bases are selected at more than one position in a SAAB imprint (for example, C or G at positions −1 and +1 in the E47 consensus), some may actually be coupled with each other in the actual preferred sites. This can be investigated by cloning and sequencing individual selected templates or by judicious choice of an alternative random oligonucleotide template. In the above example, the D2-E47 SAAB clearly shows that the C at +1 is coupled to the C at −1 as, by inference, are G and G (*22*). In addition, SAAB imprints can complement and extend data from footprinting, binding-interference, and other biochemical assays of protein-DNA interaction.

We envision that this assay will be useful in three general situations: (i) when both the binding site and the protein are known and available; (ii) when only a consensus binding site is available and the binding protein is not; and (iii) when the protein is available, but the binding site is unknown. Our initial efforts apply primarily to the first situation, where we establish this as a useful technique by focusing on the interactions of MyoD and E2A homo- and hetero-oligomers with a consensus CA– –TG sequence. In doing so, we have shown that in vitro–translated proteins can be used and, hence, that mutant proteins can be rapidly analyzed. For the second category, we have detected activities in reticulocyte lysates and in nuclear extracts from several cell types (*20*) that bind to CA– –TG sites, thus demonstrating a general usefulness of this technique for identifying factors that bind to variations of a known consensus. Finally, in most cases the third category will require an approach that is modified from the one described here. When the binding site is not known, the number of random nucleotide positions in the template must be large. If, as is likely, this number turns out to be larger than the actual binding site, then a "phasing" problem arises because the binding site can begin at many positions along the random nucleotide stretch and direct sequencing will not give a unique sequence. Cloning and sequencing of individual reiteratively selected and amplified binding species would therefore be in order. On the other hand, beginning with a randomized oligonucleotide (NNNNNCANNTGNNNNN), we have recently identified a binding sequence for the bHLH protein c-Myc (*26*).

Specificity of DNA binding by GHLH proteins. By isolating preferred binding sites for homo- and heterooligomers of MyoD and E2A proteins, we have established that CA– –TG is, in fact, the preferred consensus; that MyoD appears to bind symmetrically; that E2A-MyoD and, surprisingly, E47 bind asymmetrically; and that flanking and internal sequences can influence binding. However, although we have identified differences and similarities in the preferred binding sites for these proteins, we have not evaluated the extent to which their respective ranges of binding specificity actually overlap. To do so would require more extensive binding studies with specific templates, for which SAAB imprinting can be useful in providing direction. On the other hand, our findings have shown that different versions of the CA– –TG motif can have very different

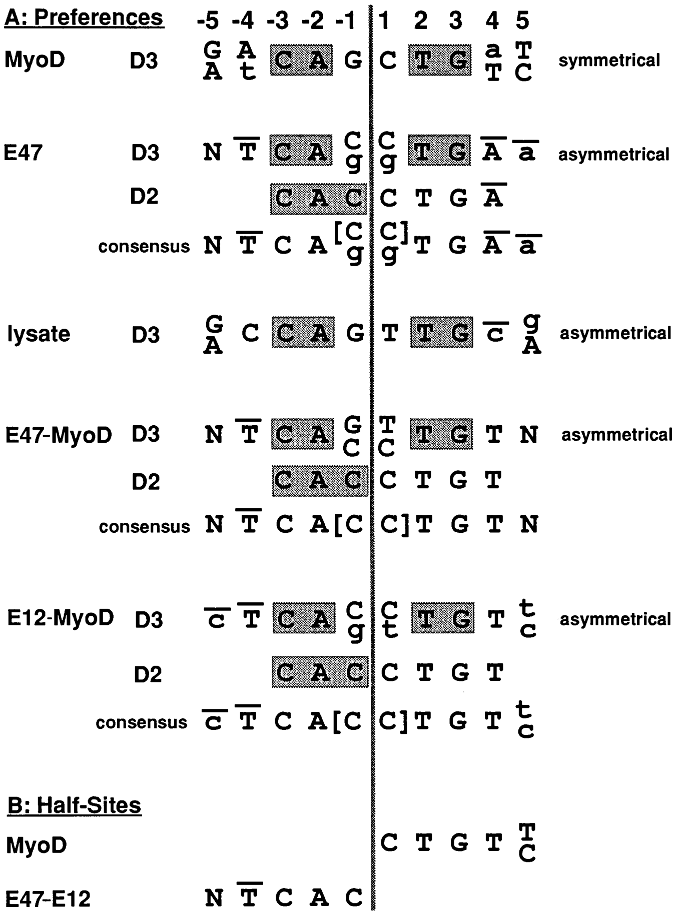

A: Preferences
-5 -4 -3 -2 -1 | 1 2 3 4 5
MyoD D3 $\overset{G}{A}$ $\overset{A}{t}$ C A G | C T G $\overset{a}{T}$ $\overset{T}{C}$ symmetrical
E47 D3 N $\overline{T}$ C A $\overset{C}{g}$ | $\overset{C}{g}$ T G $\overline{A}$ $\overline{a}$ asymmetrical
D2 C A C | C T G $\overline{A}$
consensus N $\overline{T}$ C A [$\overset{C}{g}$ | $\overset{C}{g}$] T G $\overline{A}$ $\overline{a}$
lysate D3 $\overset{G}{A}$ C C A G | T T G $\overline{c}$ $\overset{g}{A}$ asymmetrical
E47-MyoD D3 N $\overline{T}$ C A $\overset{G}{C}$ | $\overset{T}{C}$ T G T N asymmetrical
D2 C A C | C T G T
consensus N $\overline{T}$ C A [C | C] T G T N
E12-MyoD D3 $\overline{c}$ $\overline{T}$ C A $\overset{C}{g}$ | $\overset{C}{t}$ T G T $\overset{t}{c}$ asymmetrical
D2 C A C | C T G T
consensus $\overline{c}$ $\overline{T}$ C A [C | C] T G T $\overset{t}{c}$
B: Half-Sites
MyoD | C T G T $\overset{T}{C}$
E47-E12 N $\overline{T}$ C A C |

Fig. 4. Binding site preferences and consensus sequence motifs derived from SAAB imprints of MyoD and E2A protein complexes. (**A**) Binding site preferences derived from D2 and D3 template pools, and consensus preferences based on information from both. These preferences were determined from the gels shown in Figure 3 (D and E) and from a range of exposures of multiple different gels (*20*). When necessary, two additional approaches (*20*) clarified the data. To control for potential artifacts in the sequencing reactions, we sequenced the templates on the opposite strand with primer A. In addition, gel artifacts from compressions could be eliminated by sequencing with dITP, but on these small templates such sequences were of poorer quality because of a higher level of nonspecific termination by the enzyme. Data that are not shown were helpful in clarifying the MyoD (*21*) and E47 (*22*) sequence preferences. A portion of the E47-MyoD and E12-MyoD preferences were apparently contributed by the binding activity present in the reticulocyte lysate (*23*) and is omitted from the consensus. Preferences that are absolute or nearly so are indicated with uppercase letters and incomplete preferences are indicated with lowercase letters. A line drawn over an uppercase letter indicates a base that is not present in that position, and a line over a lowercase letter similarly indicates a decrease in use of that base. If one of two nucleotides dominates at a given position, the less-preferentially utilized base is listed with a lowercase letter. Pairs of nucleotides that are coupled are bracketed. (**B**) Half-site motifs derived from the sequence preferences listed in (A).

binding properties and be recognized by different protein species. For example, the D3-lysate SAAB is not bound by E12-MyoD (*20*). Similarly, c-Myc homooligomers can bind to certain CA– –TG sites but not to either MCK or the D3-lysate SAAB (*26*).

Recent studies have revealed two paradoxes related to DNA binding and transcriptional activation by bHLH proteins. First, although *achaete-scute* products are important for neuronal determination (*15*), when complexed with E2A proteins, they bind to a muscle or immunoglobulin consensus sequence with affinities that are comparable to those of the corresponding E2A-MyoD complexes (*4, 5, 27*). Hence, why are muscle-specific or immunoglobulin genes not activated in nerve cells? Moreover, as indicated above, other cell types seem to contain protein species capable of recognizing the CA– –TG consensus (*10, 13*), indicating that the problem of specificity is even more complex. Consideration of the second paradox might offer some resolution. When the basic region of MyoD is replaced by that of an *achaete-scute* product, the resulting chimeric protein can bind muscle-specific enhancers but fails to activate muscle-specific genes in vivo (*5*). These results suggested that some combined property of the DNA-binding site and the bound basic region results in activation of muscle-specific transcription, and led to the prediction that DNA sequences exist that could support E2A-MyoD binding but not support activation of a MyoD-responsive gene (*5*). Analogous sites that support binding but not activation have been identified for other transcriptional regulatory proteins (*5*). The SAAB imprint technique has provided a spectrum of binding sites that can be used to test this prediction.

Our results have also provided some general insights into the structural aspects of DNA binding by bHLH proteins. The palindromic character of the CA– –TG motif indicates a degree of structural symmetry of protein-DNA contact, but the otherwise asymmetrical preferences of E12-MyoD and E47-MyoD heterooligomers suggest that in these complexes the respective basic regions each bind to a half-site on the DNA. The respective selected half-sites for MyoD and E2A proteins are nearly the same in the various homo- and heterooligomeric complexes, even though heterooligomeric complexes generally bind with significantly higher affinities (*4, 5*) (Fig. 3C). Although other explanations are possible, these differences in binding could reflect the relative efficiencies of dimerization. Our findings thus predict, for example, that MyoD and E12 proteins dimerize less efficiently with themselves than with each other. An unexpected result from our analysis is the preferred binding of putatively symmetrical E47 homooligomers to asymmetrical sites. Whether binding to these sites induces or stabilizes an asymmetrical conformation of E47 awaits direct verification.

The bHLH proteins that have been shown to bind a CA– –TG consensus site contain a conserved Glu-Arg-X-Arg (or Glu-Lys-X-Arg) sequence (where X represents any amino acid) in the basic region at exactly the same position vis-à-vis the HLH dimerization domain (*3, 10, 28*). In both homo- and heterooligomers, the Glu-Arg-X-Arg sequence within each monomer might, in fact, directly contact the conserved CA and TG residues, respectively. By analogy, only two amino acids are conserved between the DNA-recognition helices of lambda phage cI repressor and cro; these residues make contact with the only two nucleotides that are absolutely conserved among all of the half-sites within the six partially palindromic binding sites for these protein dimers (*29*). The other bases in these recognition sites vary, so that they are recognized by cI and cro with different relative affinities that have functional consequences (*29*). Perhaps analogous principles operate in recognition of the CA– –TG consensus by bHLH proteins.

Note added in proof: Techniques that are similar to the SAAB imprint assay have been reported since submission of this manuscript (*37*).

REFERENCES AND NOTES

1. R. L. Davis, H. Weintraub, A. B. Lassar, *Cell* **51**, 987 (1987); H. Weintraub *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 5434 (1989).
2. C. Murre, P. S. McCaw, D. Baltimore, *Cell* **56**, 777 (1989).
3. R. Benezra, R. L. Davis, D. Lockshon, D. L. Turner, H. Weintraub, *ibid.* **61**, 49 (1990).
4. C. Murre *et al.*, *ibid.* **58**, 537 (1989).
5. R. L. Davis, P.-F. Cheng, A. B. Lassar, H. Weintraub, *ibid.* **60**, 733 (1990).
6. A. B. Lassar *et al.*, *ibid.* **58**, 823 (1989).
7. T. J. Brennan and E. N. Olson, *Genes Dev.* **4**, 582 (1990).
8. T. Braun, E. Bober, B. Winter, N. Rosenthal, H. H. Arnold, *EMBO J.* **9**, 821 (1990).
9. T. J. Baldwin and S. J. Burden, *Nature* **341**, 716 (1989); J. Piette, J.-L. Bessereau, M. Huchet, J.-P. Changeux, *ibid.* **345**, 353 (1990); S. Tapscott and H. Weintraub, unpublished data; V. Sartorelli, K. A. Webster, L. Kedes, *Genes Dev.* **4**, 1811 (1990).
10. M. Cai and R. W. Davis, *Cell* **61**, 437 (1990).
11. E. A. Sternberg *et al.*, *Mol. Cell. Biol.* **8**, 2896 (1988); J. N. Buskin and S. D. Hauschka, *ibid.* **9**, 2627, (1990).
12. X.-M. Wang, H.-J. Tsay, J. Schmidt, *EMBO J.* **9**, 783 (1990); N. Rosenthal, *Curr. Op. Cell Biol.***1**, 1094 (1990).
13. For representative sites see the following: G. M. Church, A. Ephrussi, W. Gilbert, S. Tonegawa, *Nature* **313**, 798 (1985); L. G. Moss, J. B. Moss, W. J. Rutter, *Mol. Cell. Biol.* **10**, 2620 (1988); K. B. Meyer and M. S. Neuberger, *EMBO J.* **8**, 1959 (1989); C. D. Reddy and E. P. Reddy, *Proc. Natl. Acad. Sci. U.S.A.* **86**, 7326, (1989); M. A. Magnuson and K. D. Shelton, *J. Biol. Chem.* **264**, 15936 (1989); A. Meister, S. L. Weinrich, C. Nelson, W. J. Rutter, *ibid.*, p. 20744; S. Pettersson, G. P. Cook, M. Brüggemann, G. T. Williams, M. S. Neuberger, *Nature* **344**, 165 (1990); H. Kleinert, S. Bredow, B.-J. Benecke, *EMBO J.* **9**, 771 (1990); P. W. Finn *et al.*, *ibid.*, p. 1543; J. M. Redondo, S. Hata, C. Brocklehurst, M. S. Krangel, *Science* **247**, 1225 (1990); J. Whelan, S. R. Cordle, E. Henderson, P. A. Weil, R. Stein, *Mol. Cell. Biol.* **10**, 1564 (1990).
14. M. P. Kamps, C. Murre, X. Sun, D. Baltimore, *Cell* **60**, 547 (1990).
15. M. C. Alonso and C. V. Cabrera, *EMBO J.* **7**, 2585 (1988); R. Villares and C. V. Cabrera, *Cell* **50**, 415 (1987); F. Gonzalez, S. Romani, P. Cubas, J. Modolell, S. Campuzano, *EMBO J.* **8**, 3553 (1989).
16. M. S. Z. Horwitz and L. A. Loeb, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 7405 (1986); A. R. Oliphant and K. Struhl, *Methods Enzymol.* **155**, 568 (1987); R. M. Gronostajski, *Nucleic Acids Res.* **15**, 5545 (1987); M. S. Z. Horwitz and L. A. Loeb, *J. Biol. Chem.* **263**, 14724, (1988); A. R. Oliphant and K. Struhl, *Nucleic Acids Res.* **16**, 7673 (1988); A. R. Oliphant, C. J. Brandl, K. Struhl, *Mol. Cell. Biol.* **3**, 1564 (1989); V. L. Singer, C. R. Wobbe, K. Struhl, *Genes Dev.* **4**, 636 (1990).
17. C. A. Kaiser, D. Preuss, P. Grisafi, D. Botstein, *Science* **235**, 312 (1987); J. Ma and M. Ptashne, *Cell* **51**, 113 (1987); D. K. Dube and L. A. Loeb, *Biochemistry* **28**, 5703 (1989).
18. H. Weintraub, R. Davis, D. Lockshon, A. Lassar, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 5623 (1990).
19. Preference for a CA– –TG motif might also account for the weak preference for C and A at positions 1 and 2, respectively, in the bound D2 template; these bases might thus be associated with the T and G at positions 5 and 6, respectively (Fig. 2B).
20. T. K. Blackwell and H. Weintraub, unpublished data.
21. In the D3-MyoD primer B sequence (Fig. 3D) the A preference at the −4 position appears as a doublet because of a compression, and seems to be weaker than the T preference at the +4 position. However, when the D3-MyoD SAAB was sequenced on the opposite strand with primer A (Fig. 1B) (*20*), the +4 T preference (on this gel an A) appeared as a doublet identical in relative intensity to the −4 A preference on the primer B gel (Fig. 3D). On the primer A gel the −4 A preference appeared as a strongly preferred T similar to that apparent at +4 on the primer B gel (Fig. 3D). Similarly, the +4 position is also characterized by a weak A preference that was only apparent on the primer A gel [by analogy to the weak T preference apparent at −4 on the primer B gel (Fig. 3D)]. These findings indicate that the preferences at −4 and +4 are symmetrical (Fig. 4). This symmetry was confirmed by sequencing with deoxyinosine triphosphate (dITP) [which eliminated the doublet A (*20*)]. The weak T preference apparent at the +1 position (Fig. 3D) was not confirmed by sequencing with primer A or with dITP and is omitted from Figure 4.
22. The E47 consensus at −1 and +1 (Fig. 4) was determined on the basis of two considerations. First, in the D3-E47 SAAB's derived by three binding-selection rounds, analysis of multiple sequences indicated that at both positions C nucleotides are at least two- to threefold more prominent, and that either C or G nucleotides are present to similar extents (Fig. 3D) (*20*). However, in D3 templates isolated by one round of selection, the levels of G and C nucleotides were about equivalent in both positions (*20*). These findings suggest that templates with a central CC pair are selected during multiple rounds, but they do not rule out the presence of templates with a central GC or CG. However, a C is overwhelmingly favored at the +1 position in the D2-E47 SAAB's, in which the −1 position is specified as a C (Fig. 3E and Fig. 4), indicating a greater preference for an asymmetrical CC than a symmetrical CG. The preference is thus listed as a coupled CC (Fig. 4). A coupled GG preference is also inferred by these data.
23. On D3, at −1 and +1 both E47-MyoD and E12-MyoD heterooligomer binding preferences are C or G and C or T, respectively (Fig. 3D and Fig. 4). However, on D2 the C specified at −1 results in a C being absolutely preferred at +1 (Fig. 3E and Fig. 4) (*20*), again suggesting CC coupling. In addition, in both D3–E47-MyoD and D3–E12-MyoD SAAB's, C nucleotides are preferred to approximately the same extent in both positions, and in the D3–E12-MyoD SAAB's they are greatly preferred over the corresponding G and T residues (Fig. 3D). G and T

nucleotides are preferred at -1 and $+1$, respectively, by the binding activity present in the reticulocyte lysate (Fig. 3D and Fig. 4). Significantly, in the D3–E47-MyoD and D3–E12-MyoD SAAB's the relative levels of G and T preference approximately parallel those of binding by the lysate activity (Fig. 3, C and D), which nearly comigrates with these species in the EMSA. Furthermore, MyoD-E12 heterooligomers do not bind the D3-lysate SAAB (see text). These observations suggest that the apparent G and T preferences actually derive from the lysate factor; they are consequently omitted from the consensus sequences indicated in Figure 4. The contribution of the lysate factor to the D3–E12-MyoD SAAB's is less, probably because the mobility of the species overlap less and because binding by E12-MyoD heterooligomers is stronger than that of E47-MyoD heterooligomers (Fig. 3C).

24. Examples of sites in muscle-specific genes: (*8, 9, 11, 12*). Examples of sites in other genes: (*13*). A compilation of these data is available on request.
25. W. E. Wright, D. A. Sassoon, V. K. Lin, *Cell* **56**, 607 (1989); D. G. Edmondson and E. N. Olson, *Genes Dev.* **3**, 628 (1989); S. J. Rhodes and S. F. Konieczny, *ibid.*, p. 2050; T. Braun, G. Buschhausen-Denker, E. Bober, E. Tannich, H. H. Arnold, *EMBO J.* **8**, 701 (1989); J. H. Miner and B. Wold, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 1089 (1990).
26. T. K. Blackwell, L. Kretzner, E. M. Blackwood, R. N. Eisenman, H. Weintraub, *Science* **250**, 1149 (1990).
27. R. L. Davis, A. B. Lassor, H. Weintraub, unpublished data.
28. H. Beckmann, L.-K. Su, T. Kadesch, *Genes Dev.* **4**, 167 (1990); Y. F. B. Lüscher, A. Admon, N. Mermod, R. Tijan, *ibid.*, p. 1741; P. Gregor, M. Sawadogo, R. G. Roeder, *ibid.*, p. 1730.
29. M. Ptashne, *A Genetic Switch* (Blackwell, Palo Alto, CA, 1987), pp. 33–48.
30. Nucleotide sequences of primer A and primer B are 5′-TCCGAATTCCTACAG-3′ and 5′-AGACGGATCCATTGCA-3′, respectively. Both contain restriction enzyme sites that allow cloning of individual templates if desired.
31. Double-stranded D1, D2, and D3 templates were generated by annealing each oligonucleotide to a tenfold molar excess of primer A, synthesizing the complementary strand with the Klenow fragment of *Escherichia coli* DNA polymerase, and purifying the template on a 12 percent polyacrylamide gel. These templates were then end-labeled by kinase reactions as described (*5*). The MCK template consisted of a 25–base pair (bp) oligonucleotide (*6*) that was annealed to its complement after labeling in a kinase reaction (*5*). Binding reactions were performed as described (*6*) but with 100 ng of poly(dI-dC). Each reaction contained approximately 200 ng of GluMyoD fusion protein (greater than 95 percent pure) prepared as described (*6*) and either 0.15 ng of the MCK template or 0.30 ng of a random-sequence template (about 6×10^4 cpm each), to give a protein concentration of about 1.5×10^{-7} M and a protein monomer to DNA ratio of about 360. The EMSA was performed on a 6 percent polyacrylamide gel as described (*5*).
32. To isolate templates, we excised a slice approximately 0.3 cm wide from the dried gel, including the 3MM (Whatman) paper backing. Gel slices were incubated at 37°C for 3 hours in 0.5 ml of 0.5 M ammonium acetate, 10 mM $MgCl_2$, 1 mM EDTA, and 0.1 percent SDS. Approximately 50 percent of the radioactivity present in the gel slice was recovered by this procedure. After addition of 5 μg of tRNA carrier, the eluate was extracted twice with phenol and twice with chloroform:isoamyl alcohol (24:1) and then precipitated with ethanol. These samples were then adjusted to 0.3 M sodium acetate and precipitated with ethanol again. Approximately one-fifth of the suspended sample was amplified for 35 cycles of PCR in a 100-μl reaction with primers A and B (*30*). PCR was performed under standard conditions (*33*) after optimization of the magnesium concentration. Care was taken to avoid cross-contamination of samples, and in all experiments a control reaction that did not contain a template did not yield a product. Under these conditions, a test reaction that contained 1 pg of starting template yielded approximately 100 ng of product. Reactions performed on material excised from EMSA gels gave comparable yields. DNA could be recovered when as few as 50 cpm of EMSA-isolated template were added to the reaction. The products of these reactions were purified on 14 percent polyacrylamide gels and then eluted and purified as above.
33. R. K. Saiki, in *PCR Technology*, H. A. Erlich, Ed. (Stockton Press, New York, 1989), pp. 7–16.
34. For the sequencing protocol, we used a labeled primer (A or B) (*30*) and the termination step of the Sequenase (U.S. Biochemical) procedure. Primers were labeled in a kinase reaction to a specific activity of 1×10^6 to 2×10^6 cpm per nanogram, and unincorporated label was removed by a Sephadex G50 spin column. Labeled primer (10 ng) was mixed with about 5 ng of a purified amplified template pool (*32*) in a 12-μl reaction volume that contained 1 μl of Sequenase Manganese buffer and 2 μl of 5× Sequenase buffer. This mixture was incubated at 95°C for 5 minutes and then allowed to cool at room temperature for 1 minute, during which time it was centrifuged for 1 second in an Eppendorf microfuge. The mixture was placed on ice, and 1 μl of 0.1 M dithiothreitol and 2 μl of diluted Sequenase 2.0 enzyme [(1:8) in ice-cold Tris-EDTA (TE) (pH 7.4)] were added. A portion of this mix (3.5 μl) was added to 2.5 μl of each of the four Sequenase deoxyguanosine triphosphate (dGTP) termination mixes and incubated at 45°C for 4 minutes. These reactions were terminated by adding 4 μl of Sequenase stop solution. Manganese buffer was omitted from reactions performed with dITP termination mixes. Reaction mixtures were subjected to electrophoresis on a 14 percent polyacrylamide denaturing sequencing gel containing 8 M urea in tris-borate-EDTA buffer (TBE). In each case, 1.5 μl of reaction mixture were loaded per well, with the exception of the C reaction in sequences generated with primer B. The nonrandom bases appearing in this C lane were generally fainter than those in the corresponding G, A, and T lanes (*20*), presumably because of terminations occurring in the six C residues 5′ of the site (Fig. 1B). This difference in intensity was compensated for by loading 2.5 μl of the C reaction in the sequencing gels shown in Figure 3. No such differences in intensity were observed among the lanes in sequences generated with primer A (*20*). Before fixing the gel in 10 percent acetic acid and 10 percent methanol, the unreacted primer (which was present in vast excess of incorporated product) was cut away to prevent its diffusion.
35. A mouse MyoD cDNA, a human E12 cDNA (E12 R) (*2*), and a human E47 cDNA (E47P) (*2*) were transcribed in vitro as described (*3*). Translations in vitro were performed and quantitated as described (*3, 5*). A portion (2.5 μl) of a 50-μl reticulocyte lysate (Promega) in vitro translation reaction was added to each binding reaction; those involving heterooligomeric complexes thus contained a total of 5 μl. Each reaction contained the in vitro–synthesized protein species at a concentration of approximately 6.9×10^{-11} M, and either 0.15 ng of MCK or 0.30 ng of labeled D2 or D3 templates (*31*), giving a protein to DNA molar ratio of about 0.18. To form heterooligomeric complexes, we mixed separate translation reactions before the DNA-binding reaction. Proteins were incubated at 37°C for 20 minutes and then added to a binding reaction mixture so that the final mix contained: 20 mM Hepes (pH 7.6), 50 mM KCl, 1 mM dithiothreitol, 1 mM EDTA, 8 percent glycerol, 0.1 mg of poly(dI-dC), and 2 mg of a 50-bp single-stranded oligonucleotide. The last two components were added as nonspecific competitors. Binding reactions were performed at room temperature for 20 minutes, and immediately after, samples subjected to EMSA as described (*31*).
36. For subsequent rounds of EMSA, amplified eluted templates were labeled by incorporation with PCR. Approximately 5 ng of the purified amplified template (*32*) were labeled for one cycle in a 20-μl reaction volume that contained 30 μCi of $[^{32}P]$dTTP (deoxythymidine triphosphate) (Dupont Biotechnology Systems), 50 mM each of dATP (deoxyadenosine triphosphate), dGTP, and dCTP (deoxycytidine triphosphate), and 100 ng each of primers A and B (*30*) in the standard PCR reaction buffer (*33*). The large excess of primers was added to ensure that synthesis occurred on all templates in the reaction and, thus, to prevent potential heteroduplex formation by annealing of denatured templates to each other. Unincorporated label was removed with a 1-ml Sephadex G50 spin column. In all cases, the reaction products were of full length, indicating complete synthesis. The binding reaction and EMSA were performed as described (*35*), but with approximately 0.1 ng of the PCR-labeled template pool (giving a protein to DNA molar ratio of about 0.54).
37. C. Turek and L. Gold, *Science* **249**, 511 (1990); R. Green, A. D. Ellington, J. W. Szostak, *Nature* **347**, 406 (1990).

38. We thank M. Groudine, S. Hahn, and members of the Weintraub laboratory for critically reading the manuscript, and R. Benezra and A. Lassar for reagents. Supported in part by a Burroughs Wellcome Fund Fellowship of the Life Sciences Research Foundation (T.K.B.).

25 June 1990; accepted 28 September 1990