MOLECULAR AND CELLULAR BIOLOGY, Sept. 1993, p. 5216–5224
0270-7306/93/095216-09$02.00/0

Vol. 13, No. 9

Binding of Myc Proteins to Canonical and Noncanonical DNA Sequences

T. KEITH BLACKWELL,[1] JING HUANG,[1] AVERIL MA,[2] LEO KRETZNER,[1] FREDERICK W. ALT,[2,3] ROBERT N. EISENMAN,[1] AND HAROLD WEINTRAUB[1,4]*

Division of Basic Sciences[1] and Howard Hughes Medical Institute,[4] Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, Washington 98104, and The Children's Hospital[2] and Howard Hughes Medical Institute,[3] Boston, Massachusetts 02115

Received 12 February 1993/Returned for modification 23 March 1993/Accepted 9 June 1993

Using an in vitro binding-site selection assay, we have demonstrated that c-Myc–Max complexes bind not only to canonical CACGTG or CATGTG motifs that are flanked by variable sequences but also to noncanonical sites that consist of an internal CG or TG dinucleotide in the context of particular variations in the CA--TG consensus. None of the selected sites contain an internal TA dinucleotide, suggesting that Myc proteins necessarily bind asymmetrically in the context of a CAT half-site. The noncanonical sites can all be bound by proteins of the Myc-Max family but not necessarily by the related CACGTG- and CATGTG-binding proteins USF and TFE3. Substitution of an arginine that is conserved in these proteins into MyoD (MyoD-R) changes its binding specificity so that it recognizes CACGTG instead of the MyoD cognate sequence (CAGCTG). However, like USF and TFE3, MyoD-R does not bind to all of the noncanonical c-Myc–Max sites. Although this R substitution changes the internal dinucleotide specificity of MyoD, it does not significantly alter its wild-type binding sequence preferences at positions outside of the CA--TG motif, suggesting that it does not dramatically change other important amino acid-DNA contacts; this observation has important implications for models of basic–helix-loop-helix protein-DNA binding.

Members of the Myc family of proteins (c-, N-, and L-Myc) have been implicated in oncogenesis, in progression through the cell cycle (39), and in induction of apoptosis (25, 45); however, the direct targets of their activity remain unknown. Myc proteins are members (21) of the basic helix-loop-helix (bHLH) protein family (42), in which the HLH domain is responsible for dimerization (42, 43) and the adjacent basic region mediates sequence-specific DNA binding (20, 56). They belong to a bHLH protein subgroup (bHLH-LZ proteins) in which a leucine zipper motif (37) that is located immediately C terminal to the HLH domain seems to participate in the dimerization process (7, 26, 31, 40). bHLH proteins bind DNA as dimers (20, 26, 42) to sequences that contain the palindromic consensus CA--TG (38), with each respective basic region recognizing half of the site (14). While Myc proteins can bind DNA in vitro as homodimers (1, 13, 35, 40), they dimerize and bind DNA more efficiently with Max, a widely expressed bHLH-LZ protein (15, 48) that may play a critical role in their functions (2, 16, 34, 41, 47, 58).

By analogy to other bHLH proteins (57), it would be predicted that Myc complexes might be involved in transcriptional regulation (18, 39), and recent evidence supports this idea (3, 33, 36). However, very little is known about what genes they might regulate (10, 22, 39). How their specificity of action is achieved is also unclear, since all Myc proteins can bind to the same CACGTG or CATGTG sequence (1, 11, 13, 34, 40, 46). Moreover, a number of related bHLH proteins, including the bHLH-LZ transcriptional regulatory proteins USF, TFE3, and TFEB, also bind to these sites (8, 17, 29), although some subtle differences among them have been identified (28, 30).

To explore these issues, we have used the strategy of sequential selection and amplification of binding sites (SAAB) (14), in which binding sites are isolated in vitro from a pool of random-sequence oligonucleotides by a process of reiterative selection and polymerase chain reaction (PCR) amplification (14, 24, 51, 53). We have selected a pool of c-Myc–Max binding sites, most of which contain the canonical CACGTG or CATGTG motif; however, to our surprise, others that bind strongly are noncanonical. All of the various types of noncanonical sites also share an internal CG or TG dinucleotide and can be bound specifically by Max and Myc family proteins but not necessarily by USF and TFE3. We have furthermore demonstrated that substitution of a single c-Myc basic region arginine (R) residue into the corresponding position in MyoD (20) changes its binding specificity so that it will now fail to recognize the MyoD preferred site (CAGCTG) and instead binds to canonical Myc family sites (CACGTG). However, like USF and TFE3, this MyoD mutant will bind to only a subset of the noncanonical sites. Thus, the ability to bind to an internal CG dinucleotide is not sufficient to allow binding to all of the noncanonical sequences, which appears to require specificity determinants that are shared by the Myc and Max proteins. An internal CG binding preference can be conferred upon MyoD without a dramatic alteration in its preferences at positions that flank the CA--TG consensus, suggesting that both the mutant and wild-type MyoD proteins make contact with DNA in an analogous fashion.

MATERIALS AND METHODS

Protein preparation. The IPmax preparation, which is a Max preparation that had been immunoprecipitated at low stringency from K562 cells, contains both the p21 and p22 forms of Max, thus representing binding by both (16) (not shown). Although it contains an excess of Max relative to c-Myc (16), our results are consistent with the finding that

* Corresponding author.

Max is phosphorylated by casein kinase II in mammalian cells (12, 16) and that this phosphorylation inhibits DNA binding by Max homodimers but not by c-Myc–Max heterodimers (12). The c-, N-, and L-Myc–glutathione *S*-transferase fusion proteins and protocols for their preparation are described elsewhere (15, 40). bHLH-LZ Max was produced by in vitro translation (in reticulocyte lysates; Promega) of RNA that encodes amino acids 13 to 96 of human Max preceded by the initiation methionine (32a) and was transcribed in vitro by standard procedures (14). MyoD and its various mutants, which had been generated by oligonucleotide-directed mutagenesis (Amersham), were similarly produced by in vitro translation and quantitated as described previously (20) so that equivalent amounts could be used in each DNA-binding reaction.

DNA-binding and SAAB assays. DNA-binding reactions, including antibody supershifts, were performed essentially as described previously (16) except for omission of single-stranded DNA. Only reactions containing MyoD proteins were performed under different conditions (14). IPmax reactions each contained 1.5 μl of protein (16), reactions performed with bacterially expressed proteins contained about 250 ng of protein, and those performed with proteins that were produced by in vitro translation contained 5 or 10 μl of lysate each. Each reaction contained 5×10^{-10} M double-stranded DNA probe that had been labeled with ^{32}P as described previously (14). Oligonucleotide probes were labeled by kinase reactions, and then the opposite strand was annealed. Sites that were selected in vitro from random-sequence templates (see below) were labeled by PCR incorporation. In binding competition assays, unlabeled DNA was added to binding reactions immediately prior to addition of the labeled probe. Electrophoretic mobility shift assays (EMSAs) were performed on 5% polyacrylamide gels as described previously (14).

The sequence of the CM1 probe has been described previously (13). All oligonucleotide probes that were generated by annealing of complementary strands contain 5′-GATC-3′ overhangs at both 5′ ends. The sequences of their double-stranded portions are as follows: MD1, 5′-CCCCCA ACAGCTGTTGCCTGA-3′; CM3, 5′-CCCCCAACACGTG TTGCCTGA-3′; M2, 5′-ATCGCAACACGCGGTTTTATG-3′; M10, 5′-AAGGTAACACGAGGTTGTAGA-3′; and M45, 5′-TCAGCCACGTTGCACAA-3′.

The various procedures and primers used in the SAAB assay have been described in detail previously (14) and were modified only in that selected DNA was eluted from polyacrylamide gels for 2 h to minimize elution of impurities that tended to generate PCR artifacts. Selections were performed by EMSA except for the first three rounds of the experiment shown in Fig. 6, which were done by immunoprecipitation with an antiserum against MyoD. Double-stranded D6 and D9 templates were generated in Klenow reactions and isolated from preparative polyacrylamide gels (14). The pool of amplified DNA that resulted from three rounds of selection for binding to IPmax (IPmaxD9^3) was digested with *Eco*RI and *Bam*HI, which cut in the primer sequences, for subcloning into Bluescript plasmids. Single-stranded DNA (strand A) produced from individual clones was analyzed by dideoxy sequencing by the Sequenase method (U.S. Biochemical). Double-stranded DNA was then generated from these single-stranded DNA samples by PCR, using primers A and B, and labeled by PCR incorporation (14) for DNA-binding assays. The B strands of the clones analyzed in Fig. 3 were sequenced with ^{32}P-labeled primer B as described previously (14).

The pooled sequences shown in Fig. 7B were obtained by the same SAAB protocol but determined from a site pool that had been expanded to a size of 124 bp by PCR amplification with primers C (5′-TCGTAAGCTGACCTAG CATGCTACGCAATGCTGTAGACGGATCCATTGCA-3′) and D (5′-CATCGACGCTCGTACACACTGTCCGTCTA GATGACTCCGAATTCCTACAG-3′), which overlap primers A and B in the underlined sequences. This PCR expansion was performed only on DNA that was used for sequencing and resulted in higher DNA yields and sequence data of generally better quality.

Methylation interference. Labeled DNA for methylation interference analysis was produced from individual selected-site clones by PCR amplification (see above) in which primer A or B had been labeled with ^{32}P in a kinase reaction. These templates were eluted from preparative polyacrylamide gels as described previously (14). Methylation and cleavage reactions were performed as described elsewhere (38). DNA-binding reactions were scaled up threefold, and the DNA and protein concentrations were doubled for these preparative EMSAs, in which less than half of the labeled material was present in the bound fraction. The bound and free DNA fractions were excised from dried-down EMSA gels and eluted as described previously (14), and cleavage products were analyzed on 8% denaturing polyacrylamide gels as described elsewhere (38).

RESULTS

DNA binding by c-Myc–Max complexes derived from mammalian cells. In IPmax, the predominant detected complex that bound to a CACGTG c-Myc binding site could be supershifted by antibodies against either Max or c-Myc, indicating that it was a heteromer containing both proteins (16). To perform a comprehensive analysis of DNA binding by c-Myc–Max complexes that are expressed in mammalian cells, we used the SAAB technique to isolate IPmax binding sequences from a pool of completely random DNA, employing an EMSA for selection. As a starting template for binding, we used the D9 oligonucleotide (Fig. 1A), in which 35 bp of random-sequence DNA are flanked by fixed sequences that permit PCR amplification and cloning. In the first selection round, the gel was run a short distance, and DNA was isolated by excision of the entire lane above the unbound template, making it possible to recover DNA that would be present in bound complexes that might not be readily detected (13). In the second and third selection rounds, DNA was isolated from a single apparent protein-DNA complex (not shown) (Fig. 1B). An antibody supershift analysis demonstrated that this complex contained both c-Myc and Max proteins (Fig. 1B) and thus that the isolated DNA represented a pool of c-Myc/Max binding sites.

Of 48 sites that were cloned and sequenced after the third selection round, 12 contained a CACGTG motif and 14 contained a CATGTG motif (Fig. 2A). The representation of CATGTG sites is probably artificially high because in five of these sites, half of this motif was contributed by fixed sequences that border the random sequences in D9 (Fig. 1A and 2A). In a separate SAAB analysis, IPmax binding sites were isolated from an oligonucleotide template in which positions within and flanking a fixed CA--TG motif were random (not shown) (13). In this latter experiment, both CACGTG and CATGTG sites were present after initial rounds, but with further selection, the CACGTG sites were present at an increasingly higher relative level and comprised most of the selected sites (not shown). These findings

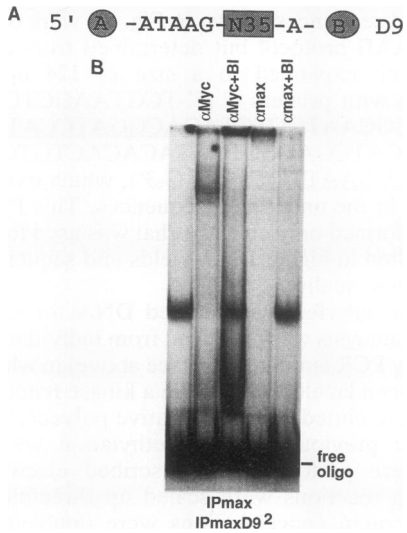

FIG. 1. SAAB analysis of IPmax DNA binding. (A) The D9 template consists of the sequences shown, with A and B′ indicating primer A and the complement of primer B, respectively (14). N35 indicates 35 bases of random sequence. (B) Binding of IPmax to selected sequences (IPmaxD9^2) was analyzed by antibody supershift EMSA. Antibodies that were added to particular binding reactions are indicated above the corresponding lanes. Where designated by BI, these antigen-antibody reactions were specifically blocked by addition of either the c-Myc peptide immunogen or the C-terminal fragment of Max that was used as the immunogen (16).

suggest that c-Myc–Max complexes can bind to CACGTG or CATGTG sites and that CACGTG sites constitute the greater proportion of their highest-affinity sites. None of the selected sites contained a symmetric CATATG motif (Fig. 2A), and no such sites were identified for IPmax in experiments in which a CAT motif had been fixed in the starting template (not shown), suggesting that c-Myc–Max complexes bind asymmetrically in the context of a CAT half-site. Presumably, c-Myc has a high affinity for one half-site in the hexamer and Max does for the other.

An alignment of the selected CACGTG and CATGTG sites suggested some sequence preferences at positions flanking these hexamers (Fig. 2B). In both types of sites, these preferences were asymmetric, perhaps again deriving from a c-Myc preference on one half of the site and a Max preference on the other. However, with the exception of the A preference at −5 in both types of sites (Fig. 2B), none of them corresponded to the 5′-GAC-3′ flanking sequence in each half-site of a proposed 12-bp complete c-Myc binding site (30). We also did not isolate any individual sites that had a GAC flanking sequence, although we isolated other flanking motifs (most frequently TAG) multiple times (Fig. 2A; not shown).

Twenty-two of the selected sequences did not include either a CACGTG or a CATGTG motif (Fig. 2A). None of them (Fig. 2A and 3C) corresponded to a proposed TCAT TCA site for cellular c-Myc–Max complexes (5, 44), and under our conditions, neither IPmax nor bacterially expressed c-Myc–Max heterodimers bound to an oligonucleotide that contains this sequence (not shown). To distinguish sequences that might contain a noncanonical c-Myc–Max binding site from those that might have been isolated because of nonspecific binding, several were analyzed by EMSA for binding to IPmax (Fig. 2C). Under these conditions, the overall affinity of binding by the selected pool (IPmaxD9^3) was severalfold greater than that of D9 (Fig. 2C). Significantly, many of the individual selected sequences that did not contain a CA--TG hexamer bound to IPmax with affinities that are comparable to those of the CACGTG and CATGTG sites (Fig. 2C).

We used a guanine methylation interference assay to map the site of c-Myc–Max binding within several of the high-affinity noncanonical sequences and to obtain an indication of some of the crucial DNA contacts (Fig. 3A and B; summarized in Fig. 3C). Most of the alternative sites contained a noncanonical hexamer motif in which a CACGTG or CATGTG sequence had been substituted at position +2 with C or A (Fig. 3C). The single exception, M45, appeared to be based on a CA--TG heptamer sequence (Fig. 3C). Each of these motifs appeared in more than one selected sequence: CACGCG in M2 and M27; CACGAG in M10 and M42; CATGCG in M5, M11, M18, and M36; and CACGTTG in M13 and M45 (Fig. 2A).

A striking feature of these interference patterns was the relative importance of interactions with G residues in the central 2 bp of each site (Fig. 3). While the degree of interference at other G residues within these motifs varied, in every site examined, methylation of the internal G residues disrupted binding (Fig. 3), a finding that is consistent with presence of a CG or TG dinucleotide in all of these types of IPmax sites and in all of the known bHLH-LZ protein binding sites (see above). Surprisingly, methylation of the G residue at +3 in M18 and in M36 did not appear to dramatically disrupt binding (Fig. 3), even though this G residue defines part of the hexamer and is conserved in every site isolated (Fig. 2A). In contrast, methylation of the G residue at +4 in the heptameric sequence M45 does interfere, suggesting that this sequence may actually be read as a canonical CA--TG heptamer (Fig. 3). Although binding interference was most striking within the consensus motifs, it was also apparent in flanking sequences in most of these sites. The conservation of flanking sequences between M2 and M10, and between M18 and M36, similarly suggests that these sequences may be important, as do the differences between the methylation interference patterns of M18 and M36.

Binding of bHLH-LZ proteins to selected c-Myc–Max sites. To more accurately assess the relative affinities with which IPmax binds to the noncanonical sites, three of them were assayed for binding in competition with CM1 (13, 15), a high-affinity CACGTG binding site for c-Myc and c-Myc–Max complexes (Fig. 4). The M2 and M10 sequences competed about as well at a 100-fold molar excess as did unlabeled CM1 at a 25-fold excess, with M45 binding only slightly less strongly (Fig. 4, lanes 1 to 4 and 8 to 16). In contrast, a CAGCTG MyoD binding site (MD1) (14) competed only slightly better than did nonspecific DNA (Fig. 4, lanes 5 to 7, and data not shown). The observation that these noncanonical sites bound to IPmax at affinities that were high but slightly lower than those of optimal CACGTG sites (Fig. 2C and 4) is consistent with their relative levels of representation among the selected sequences (Fig. 2A).

Because all bHLH-LZ proteins examined so far bind to CACGTG (and CATGTG) sites, we have examined how well the noncanonical sites are bound by a panel of different bHLH-LZ proteins (for brevity, these results are presented in summary form in Fig. 3C). Homodimers of a Max protein fragment that contains only the bHLH-LZ region bind with apparently equal affinity to each of the sites indicated in Fig. 3C (data not shown). Similarly, the relative binding affinities

FIG. 2. Selected IPmax binding sites. (A) Grouping of individual sequences that were cloned from the IPmaxD9^3 population according to whether they contain a CACGTG or CATGTG motif. Strand A, the continuation of primer A, is shown, with six fixed bases (Fig. 1A) underlined on each side of the selected sequences. In some sequences, these were deleted because of an internal restriction site. M46a and M46b, two inserts derived from the same clone; N, ambiguous bases; #, site that was analyzed by EMSA for IPmax binding. Consensus binding motifs and noncanonical c-Myc–Max binding motifs that were identified in this study (see text) are shown by outlined letters. (B) Tabulation of sequences flanking the canonical c-Myc–Max sites. In the upper panel, these have been oriented to maximize apparent preferences; in the lower panel, those CA CATG (CATGTG) sites in which half was contributed by fixed sequences (see panel A) have been omitted. The CA--TG consensus is boxed. A line over a letter indicates an apparent preference for lack of that base at that position. (C) EMSA in which the indicated sequences were PCR labeled to approximately equal specific activities and analyzed for binding to IPmax. Sequences with canonical binding motifs are indicated.

of these sites for c-, N-, and L-Myc–glutathione *S*-transferase fusion proteins, each of which contains only the C-terminal bHLH-LZ region, are comparable (Fig. 3C and data not shown). Max also efficiently forms DNA-binding heterodimers with the bHLH-LZ proteins Mad and Mxi-1, which are closely related to each other (6, 59), and Mad-Max heterodimers bind to each of these sites with similar affinities (5a). In contrast, while USF binds well to M2, M10, and M45, in addition to the tested CACGTG and CATGTG sites, it binds only very weakly to M18 and not at all to M36 (Fig. 3C and data not shown). TFE3 is more restrictive in its pattern of binding and does not bind at all to M18, M36, or M45 (Fig. 3C and data not shown). Although the basic regions of these proteins differ to comparable extents (Fig. 5), our results suggest that Max and its dimerization partners exhibit relatively similar binding properties with respect to these sites, while USF and TFE3 are different from them and from each other.

Mutation of MyoD to a Myc-like DNA-binding specificity. All bHLH and bHLH-LZ proteins that bind to CACGTG sites have in their basic regions a hydrophobic residue at position 8 and an arginine (R) at position 13 (15) (Fig. 5). This R-13 residue is required for c-Myc binding to CACGTG sites (31), its substitution into the bHLH protein AP4 allows binding to such a site (19), and its substitution into MyoD allows heterodimers of this mutant with the bHLH protein E2A to bind to a CACGTG site (54). The role of the conserved hydrophobic residue at 8 is unknown, but with a single exception (AP4 [32]), bHLH proteins that lack R-13 have an R instead at position 8 (9). We have investigated to what extent these substitutions allow MyoD homodimers to bind to the various Myc family binding sites described above. To more clearly delineate the consequences of these substitutions, we have also analyzed the extent to which they affect the DNA-binding preferences of MyoD at various site positions.

MyoD bound preferentially to the symmetrical sequence AACAGCTGTT (MD1; Fig. 6B, lane 3) (14) but bound only very weakly to an otherwise identical site in which the center dinucleotide is CG (CM3; Fig. 6B, lane 7). Conversely, c-Myc homodimers could bind specifically to CM3 but not to

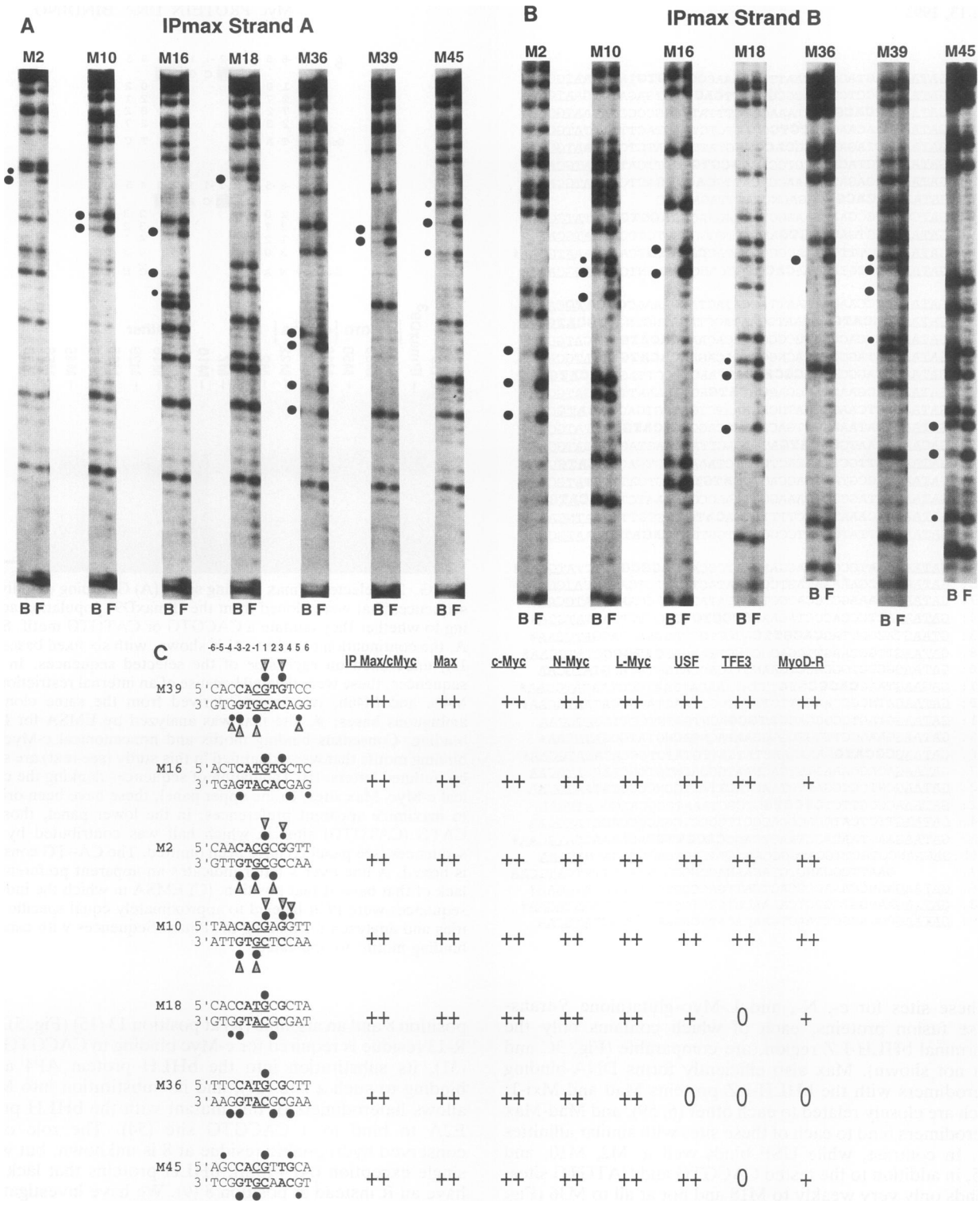

FIG. 3. Binding of various bHLH proteins to representative c-Myc–Max sites. (A) Identification of binding sites within the indicated DNA sequences by methylation interference. DNA in which strand A was end labeled was partially methylated (one hit per molecule) and then incubated with IPmax in a DNA-binding reaction. The bound (B) and free (F) fractions were isolated by EMSA, cleaved at methylated G residues with piperidine, and electrophoresed on a denaturing polyacrylamide gel as shown. Site positions at which at least 90% of the methylated DNA was in the free fraction are indicated by large dots, and those with a lesser degree of interference (as indicated by multiple gels and exposures) indicated by small dots. (B) Experiment identical to that in panel A but performed with DNA that was end labeled on strand B. (C) Summary of the results in panels A and B (dots are as defined above) along with results of methylation interference studies performed with MyoD-R (not shown), which are indicated in an analogous fashion with large and small triangles. Sequences that correspond to the CA--TG consensus and CG or TG internal dinucleotides (which are also underlined) are in boldface type. The M10, M16, and M36 sites are indicated in orientation opposite that in Fig. 2A. The levels at which the designated proteins bound to these sequences (as measured by EMSA; not shown) are summarized as follows: ++, binding within a fivefold range of the level of binding to M39; +, barely detectable but specific binding; 0, no binding detected.

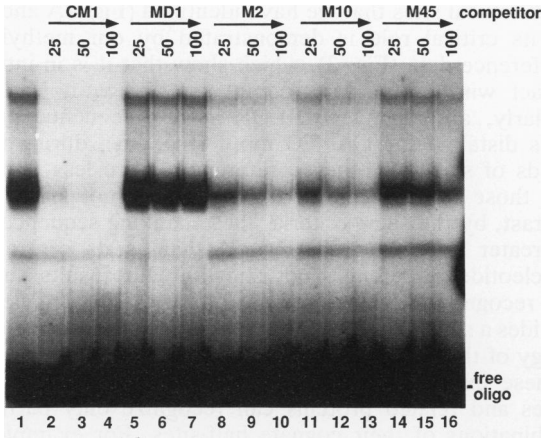

FIG. 4. Competition of canonical and noncanonical sites for binding to IPmax. In this EMSA, binding of IPmax to labeled CM1 was competed for with the indicated unlabeled DNA sites. The amount of competitor added is indicated above each lane as a multiple of the concentration of labeled probe. The most slowly migrating species evident in this autoradiogram contains Max protein (15) and appears at variably low levels among different experiments (not shown).

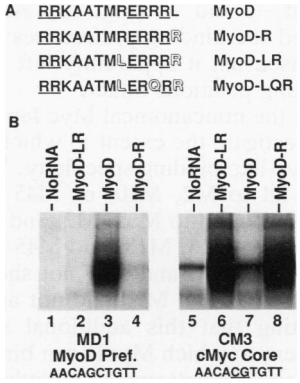

FIG. 6. Changing MyoD to a Myc-like DNA-binding specificity. (A) Basic regions are shown as in Fig. 5 but with the conserved amino acids underlined. Amino acids that have been substituted with the corresponding c-Myc residues (Fig. 5) are indicated by open letters. (B) The indicated mutants and unprogrammed reticulocyte lysate (NoRNA) were analyzed by EMSA for binding to equal amounts of the MD1 and CM3 probes, which differ only in the central 2 bp (underlined in the partial CM3 sequence shown). Only the bound species are shown. MD1 corresponds to the MyoD binding preferences identified previously (14).

the MD1 site (not shown). Substitution of L-8 and R-13 into MyoD (MyoD-LR; Fig. 6A) resulted in high-affinity binding to the CM3 site but not to MD1 (Fig. 6B, lanes 2 and 6). A MyoD protein in which only R-13 was introduced (MyoD-R; Fig. 6A) also did not bind to MD1 and bound well to CM3, although not quite as well as did MyoD-LR (Fig. 6B, lanes 4 and 8). In contrast, MyoD-R bound at higher affinity to CATGTG sites than did either MyoD-LR or MyoD (not shown). The R-13 mutation thus appeared to allow MyoD to switch its specificity for the center 2 bp in the site from GC to CG (or TG), with the additional substitution of L-8 stabilizing interaction with the symmetric CACGTG site.

To test whether this change in specificity involves a difference in overall binding conformation, we used the SAAB technique to assay how such mutations affect the preferred binding sequences at positions within and flanking the CA--TG motif. We replaced the RERRRL sequence in MyoD with the c-Myc sequences LERQRR (MyoD-LQR; Fig. 6A) and compared the binding preferences that MyoD and MyoD-LQR selected at positions within and flanking a fixed CA--TG consensus (Fig. 7B). We assayed these preferences by sequencing the pool of selected sites, a procedure that yields a characteristic image, or imprint, of binding by a given protein (14). After six selection rounds, the two respective proteins bound to their selected site pools with comparable affinities (Fig. 7A). The MyoD binding-site preferences were identical to those previously determined between positions −5 and +5 (14), most notably the A at −4, the central GC dinucleotide, and the T at +4, with additional preferences apparent at +6 and +7 (Fig. 7B). MyoD-LQR instead preferred a central CG dinucleotide but was identical to MyoD in its preferences at −4 and +4 and very similar in

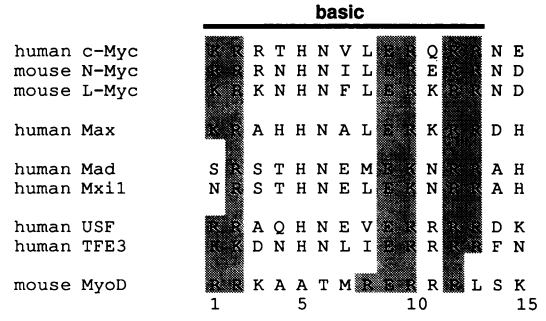

FIG. 5. bHLH protein basic regions. Sequences of the indicated proteins were adapted from references 6, 15, and 59. Shaded regions designate conserved basic and acidic amino acids.

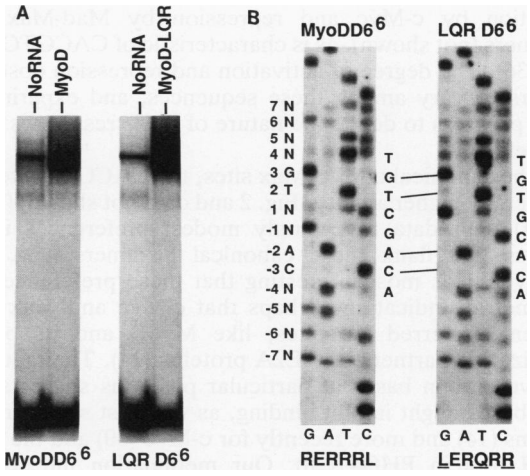

FIG. 7. SAAB analysis of DNA binding by MyoD and MyoD-LQR. (A) The indicated proteins were assayed by EMSA for binding to the respective products of six rounds of binding-site selection from the random sequence template D6 (13). (B) The indicated selected-site pools were analyzed by pooled sequencing using primer B. Random-sequence positions in the original D6 template are indicated at the left and selected preferences are indicated at the right of each sequence. The CA--TG consensus positions are connected by lines. The basic region residues that were substituted in the LQR mutant are indicated below the sequence.

its preferences at −5 and +5 (Fig. 7B). Although the LQR mutation changed the binding preferences of MyoD within the CA--TG consensus, it apparently left the most critical contacts at flanking positions intact.

Availability of the noncanonical Myc family sites enabled us to further investigate the extent to which such mutations give MyoD a Myc-like binding specificity. Wild-type MyoD did not bind at all to M2, M10, or M45 (not shown). In contrast, MyoD-R bound to M39, M2, and M10 with significant affinities and to M16, M18, and M45 at low levels but not at all to M36 (Fig. 3C and data not shown). MyoD-LR bound less well to M2 and M10 and not at all to M45 (not shown), suggesting that this additional substitution also weakens the extent to which MyoD can bind to asymmetric noncanonical sites. The patterns of methylation interference with MyoD-R binding to M39, M2, and M10 were essentially the same as those of c-Myc–Max binding (not shown; summarized in Fig. 3C), indicating that it is able to make very similar contacts with these sites. Thus, the R-13 substitution can confer upon MyoD a Myc-like recognition of CG or TG center dinucleotides, but, like USF and TFE3, MyoD-R is more restricted than are the Myc-Max family proteins in the contexts in which it can do so.

DISCUSSION

DNA binding by c-Myc–Max complexes. Using the SAAB technique (14), we have isolated sequences that are bound specifically by c-Myc–Max complexes that were isolated by low-stringency immunoprecipitation of Max from a mammalian cell line (16). The predominant DNA-binding activity that we detect in this preparation consists of heteromeric c-Myc–Max complexes (Fig. 1). Our most surprising result is that in addition to binding to the canonical CACGTG or CATGTG sites that were identified in previous studies (see above), these complexes bind to certain noncanonical DNA sequences (Fig. 2A and 3C). Preliminary results indicate that when linked to a basal promoter element, the noncanonical sequences that we have tested can mediate transcriptional activation by c-Myc and repression by Mad-Max heterodimers (not shown), as is characteristic of CACGTG sites (3, 6, 36). The degree of activation and repression observed appears to vary among these sequences, and experiments are in progress to define the nature of these responses more precisely.

Of the canonical c-Myc–Max sites, the CACGTG sites are of generally higher affinity (Fig. 2 and data not shown) (1, 11, 13, 34). Our data reveal only modest preferences in sequences that flank these canonical hexamers (Fig. 2B). However, it is most interesting that these preferences are asymmetric, indicating perhaps that c-Myc and Max have different preferred half-sites, like MyoD and its bHLH dimerization partners, the E2A proteins (14). The failure to observe certain bases at particular positions suggests that these bases might inhibit binding, as was first seen for E2A proteins (14) and more recently for c-Myc (30) and the yeast bHLH protein PHO4 (28). Our methylation interference results (Fig. 3) and those of other investigators (26), further suggest that these flanking site positions are contacted by bHLH-LZ proteins. These findings are consistent with the idea that although flanking sequences can be important for binding by c-Myc–Max complexes, a large number of different sequence combinations will give high-affinity binding.

A striking observation from this analysis is the relative importance of the internal CG or TG dinucleotide. This sequence is present in every one of the canonical and noncanonical sites that we have identified (Fig. 2A and 3C), and its critical role is demonstrated by our methylation interference data (Fig. 3), which show that it is in intimate contact with c-Myc–Max complexes and with MyoD-R. Similarly, although MyoD-LQR selects sequence preferences distal to the CA--TG motif (Fig. 7B), during initial rounds of selection, these preferences were less dramatic than those for the central CG sequence (not shown). In contrast, by the same criteria, these flanking sequences are of greater importance for MyoD than is its central GC dinucleotide preference (not shown). Our results suggest that recognition of the internal CG (or TG) dinucleotide provides a substantial and critical contribution to the binding energy of these protein-DNA complexes.

These experiments also suggest that c-Myc–Max complexes and related proteins can recognize only particular combinations of their cognate half-sites. For example, although they can utilize a CAT half-site that is paired with GTG, binding to a symmetric CATATG sequence has not been observed (1, 40, 49) (data not shown). Thus, as is the case with E2A protein homodimers (14), in these dimeric complexes, bHLH protein basic regions sometimes do not recognize their respective half-sites in a completely independent manner. In the noncanonical half-sites, only certain substitutions within the CA--TG motif seem to be allowed. For example, in the noncanonical hexamers (Fig. 2A and 3C), only A or C is substituted for T at position +2, and only one half-site is thus substituted in each site, indicating that the corresponding promoter (either c-Myc or Max, but not both simultaneously) can tolerate some degree of variation. It is not known whether other bHLH proteins will generally allow such substitutions in the CA--TG motif, but the Enhancer-of-split bHLH protein (which also contains the R-13 residue) appears to bind to a CACGAG site (52). A variant form of the noncanonical sequences is represented by the apparent heptamers (M13 and M45; Fig. 2A and 3C), a type of site that has also been isolated previously for the E2A proteins (50). The methylation interference that we have observed at position +4 in M45 (Fig. 3) is consistent with the idea that they are true heptameric CA--TG sites, with a difference in spacing of the two basic regions allowing them to make analogous contacts with each half of the consensus.

Significantly, the different Myc family protein complexes that we have tested all bind specifically in vitro to the various types of canonical and noncanonical c-Myc–Max sites (Fig. 3C), although some differences exist among them (for example, c-Myc homodimers are less likely to bind to CATGTG sites (13) (data not shown). In contrast, while USF, TFE3, and MyoD-R bind to the canonical sites, they are more restricted and do not bind to all of the noncanonical sites (Fig. 3C). A substitution at position +2 in the CA--TG motif has the most dramatic effect on binding by bHLH-LZ proteins other than the Myc-Max family, with USF, TFE3, and MyoD-R not binding to the selected CATGCG sites (M18 and M36; Fig. 3C). At position +3 in these two sites, the degree to which methylation interfered with c-Myc–Max binding was less than in other sites, suggesting that a difference in base-specific contacts may have been required for binding to them (Fig. 3C). Our results indicate that presence of R-13 (and resulting preference of an internal CG or TG dinucleotide) is not sufficient to mediate recognition of all of the noncanonical sites. Max and its dimerization partners thus appear to share a determinant of binding specificity that allows them to bind to M18, M36, and M45; this determinant is lacking in the other R-13 bHLH proteins

that we tested and could be identified by mutagenesis experiments.

General aspects of bHLH protein-DNA interactions. bHLH protein basic regions appear to adopt an α-helical conformation when binding to DNA (4, 27), suggesting a structure analogous to that of LZ protein basic regions, which cross the major groove as α helices that make both base-specific and backbone-phosphate contacts (23). When the basic region residues of bHLH proteins are thus arrayed, they display a "face" that would allow conserved amino acids access to the major groove (26, 27), as would be consistent with methylation interference and mutational analyses (Fig. 3) (4, 26, 27). Two general types of helical models have been proposed for how bHLH proteins might bind to DNA. In one type, the R-13 residue within each basic region promoter would directly contact one of the CG (or TA) base pairs in the center of the site, and contacts with the remainder of the site would be maintained in an analogous fashion by all bHLH proteins (19, 31, 55). In the other type of model, R-13 does not contact the central bases (27, 28) but instead mediates this specificity indirectly, by changing the basic region conformation relative to the DNA so that other amino acids bind to them directly (28). This latter model thus appears to require that some of the contacts by which the R-13 bHLH proteins recognize bases at other site positions are different from those made by the rest of the bHLH family.

Our experiments have shown that homodimers of the MyoD-R mutant do not bind at high affinity to a cognate MyoD site (CAGCTG) but instead recognize an otherwise identical site in which the only the central base pairs have been changed (to CG; Fig. 6B). This finding confirms and extends the conclusion of previous studies that the presence of this residue allows binding to CACGTG sites (19, 31, 54). Significantly, we have furthermore shown that although MyoD-LQR preferentially binds to CACGTG sites, its preferred binding sequences at positions ±4 and ±5 remain almost identical to those of wild-type MyoD (Fig. 7B). The most likely explanation for this latter result is that the LQR substitution has changed sequence recognition specifically at the central dinucleotide (from GC to CG) and that the basic regions of the R-13-substituted and wild-type MyoD molecules otherwise prefer to make analogous base-specific contacts over at least a 5-bp half-site. These findings suggest that the R-13 substitution does not cause the MyoD basic region to shift or ratchet along the major groove and instead are consistent with the idea that each R-13 residue instead contacts one of the central base pairs directly. The methylation interference patterns for binding of MyoD-R to the M2, M10, and M39 sites are virtually identical to those of c-Myc–Max (not shown), suggesting that these respective protein-DNA complexes also involve very similar base-specific interactions. Together, our findings suggest that all bHLH proteins make generally analogous contacts with the CA--TG consensus and flanking regions, but in the absence of sufficient structural information, such speculations and models should be interpreted with caution.

ACKNOWLEDGMENTS

This work was supported by grants from the NIH to H.W., R.N.E. (CA 20525), and F.W.A. and by the Howard Hughes Medical Institute. T.K.B. was a Burroughs Wellcome Fund Fellow of the Life Sciences Research Foundation while most of this work was performed and is now supported by the Cardiovascular Pathology training grant in the Department of Pathology at the University of Washington. L.K. is supported by a fellowship from the ACS.

We thank Beth Blackwood for a very generous supply of IPmax, Robert Roeder for a USF expression vector, Tom Kadesch for a TFE3 expression vector, Jonathan Mitchell for technical assistance, and Mark Groudine for helpful discussions and for critically reading the manuscript.

REFERENCES

1. **Alex, R., O. Sozeri, S. Meyer, and R. Dildrop.** 1992. Determination of the DNA sequence recognized by the bHLH-zip domain of the N-Myc protein. Nucleic Acids Res. **20**:2257–2263.
2. **Amati, B., M. W. Brooks, N. Levy, T. D. Littlewood, G. Evan, and H. Land.** 1993. Oncogenic activity of the c-Myc protein requires dimerization with Max. Cell **72**:233–245.
3. **Amati, B., S. Dalton, M. W. Brooks, T. D. Littlewood, G. I. Evan, and H. Land.** 1992. Transcriptional activation by the c-Myc oncoprotein in yeast requires interaction with Max. Nature (London) **359**:423–426.
4. **Anthony-Cahill, S. J., P. A. Benfield, R. Fairman, Z. R. Wasserman, S. L. Brenner, W. F. Stafford, C. Altenbach, W. L. Hubbell, and W. F. DeGrado.** 1992. Molecular characterization of helix-loop-helix peptides. Science **255**:979–983.
5. **Ariga, H., Y. Imamura, and S. M. M. Iguchi-Ariga.** 1989. DNA replication origin and transcriptional enhancer in c-myc gene share the c-myc protein binding sequences. EMBO J. **8**:4273–4279.

5a. **Ayer, D.** Unpublished data.

6. **Ayer, D., L. Kretzner, and R. Eisenman.** 1993. Mad: a heterodimeric partner for Max that antagonizes Myc transcriptional activity. Cell **72**:211–222.
7. **Beckmann, H., and T. Kadesch.** 1991. The leucine zipper of TFE3 dictates helix-loop-helix dimerization specificity. Genes Dev. **5**:1057–1066.
8. **Beckmann, H., L. K. Su, and T. Kadesch.** 1990. TFE3: a helix-loop-helix protein that activates transcription through the immunoglobulin enhancer μE3 motif. Genes Dev. **4**:167–179.
9. **Benezra, R., R. L. Davis, D. Lockshon, D. L. Turner, and H. Weintraub.** 1990. The protein Id: a negative regulator of helix-loop-helix DNA binding proteins. Cell **61**:49–59.
10. **Benvenisty, N., A. Leder, A. Kuo, and P. Leder.** 1992. An embryonically expressed gene is a target for c-Myc regulation via the c-Myc binding sequence. Genes Dev. **6**:2513–2523.
11. **Berberich, S., D. N. Hyde, P. Espenshade, and M. Cole.** 1992. max encodes a sequence-specific DNA-binding protein and is not regulated by serum growth factors. Oncogene **7**:775–779.
12. **Berberich, S. J., and M. D. Cole.** 1992. Casein kinase II inhibits the DNA-binding activity of Max homodimers but not Myc/Max heterodimers. Genes Dev. **6**:166–176.
13. **Blackwell, T. K., L. Kretzner, E. M. Blackwood, R. N. Eisenman, and H. Weintraub.** 1990. Sequence-specific DNA binding by the c-Myc protein. Science **250**:1149–1151.
14. **Blackwell, T. K., and H. Weintraub.** 1990. Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. Science **250**:1104–1110.
15. **Blackwood, E. M., and R. N. Eisenman.** 1991. Max: a helix-loop-helix zipper protein that forms a sequence-specific DNA-binding complex with Myc. Science **251**:1211–1217.
16. **Blackwood, E. M., B. Luscher, and R. N. Eisenman.** 1992. Myc and Max associate in vivo. Genes Dev. **6**:71–80.
17. **Carr, C. S., and P. A. Sharp.** 1990. A helix-loop-helix protein related to the immunoglobulin E box-binding proteins. Mol. Cell. Biol. **10**:4384–4388.
18. **Collum, R. G., and F. W. Alt.** 1990. Are myc proteins transcription factors? Cancer Cells **2**:69–75.
19. **Dang, C. V., C. Dolde, M. L. Gillison, and G. J. Kato.** 1992. Discrimination between related DNA sites by a single amino acid residue of Myc-related basic-helix-loop-helix proteins. Proc. Natl. Acad. Sci. USA **89**:599–602.
20. **Davis, R. L., P. F. Cheng, A. B. Lassar, and H. Weintraub.** 1990. The MyoD DNA binding domain contains a recognition code for muscle-specific gene activation. Cell **60**:733–746.
21. **Davis, R. L., H. Weintraub, and A. B. Lassar.** 1987. Expression

of a single transfected cDNA converts fibroblasts to myoblasts. Cell **51**:987–1000.
22. **Eilers, M., S. Schirm, and J. M. Bishop.** 1991. The MYC protein activates transcription of the alpha-prothymosin gene. EMBO J. **10**:133–141.
23. **Ellenberger, T. E., C. J. Brandl, K. Struhl, and S. C. Harrison.** 1992. The GCN4 basic region-leucine zipper binds DNA as a dimer of uninterrupted α-helices: crystal structure of the protein-DNA complex. Cell **71**:1223–1237.
24. **Ellington, A. D., and J. W. Szostak.** 1990. In vitro selection of RNA molecules that bind specific ligands. Nature (London) **346**:818–822.
25. **Evan, G. I., A. H. Wyllie, C. S. Gilbert, T. D. Littlewood, H. Land, M. Brooks, C. M. Waters, L. Z. Penn, and D. C. Hancock.** 1992. Induction of apoptosis in fibroblasts by c-myc protein. Cell **69**:119–128.
26. **Fisher, D. E., C. S. Carr, L. A. Parent, and P. A. Sharp.** 1991. TFEB has DNA-binding and oligomerization properties of a unique helix-loop-helix/leucine-zipper family. Genes Dev. **5**:2342–2352.
27. **Fisher, D. E., L. A. Parent, and P. A. Sharp.** 1993. High affinity DNA-binding Myc analogs: recognition by an α helix. Cell **72**:467–476.
28. **Fisher, F., and C. R. Goding.** 1992. Single amino acid substitutions alter helix-loop-helix protein specificity for bases flanking the core CANNTG motif. EMBO J. **11**:4103–4109.
29. **Gregor, P. D., M. Sawadogo, and R. G. Roeder.** 1990. The adenovirus major late transcription factor USF is a member of the helix-loop-helix group of regulatory proteins and binds to DNA as a dimer. Genes Dev. **4**:1730–1740.
30. **Halazonetis, T. D., and A. N. Kandil.** 1991. Determination of the c-Myc DNA-binding site. Proc. Natl. Acad. Sci. USA **88**:6162–6166.
31. **Halazonetis, T. D., and A. N. Kandil.** 1992. Predicted structural similarities of the DNA binding domains of c-Myc and endonuclease Eco RI. Science **255**:464–466.
32. **Hu, Y. F., B. Luscher, A. Admon, N. Mermod, and R. Tjian.** 1990. Transcription factor AP-4 contains multiple dimerization domains that regulate dimer specificity. Genes Dev. **4**:1741–1752.
32a. **Huang, J.** Unpublished data.
33. **Kato, G. J., J. Barrett, G. M. Villa, and C. V. Dang.** 1990. An amino-terminal c-myc domain required for neoplastic transformation activates transcription. Mol. Cell. Biol. **10**:5914–5920.
34. **Kato, G. J., W. M. Lee, L. L. Chen, and C. V. Dang.** 1992. Max: functional domains and interaction with c-Myc. Genes Dev. **6**:81–92.
35. **Kerkhoff, E., K. Bister, and K. Klempnauer.** 1991. Sequence-specific DNA binding by Myc proteins. Proc. Natl. Acad. Sci. USA **88**:4323–4327.
36. **Kretzner, L., E. M. Blackwood, and R. N. Eisenman.** 1992. Myc and Max proteins possess distinct transcriptional activities. Nature (London) **359**:426–429.
37. **Landschulz, W. H., P. F. Johnson, and S. L. McKnight.** 1988. The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. Science **240**:1759–1764.
38. **Lassar, A. B., J. N. Buskin, D. Lockshon, R. L. Davis, S. Apone, S. D. Hauschka, and H. Weintraub.** 1989. MyoD is a sequence-specific DNA binding protein requiring a region of myc homology to bind to the muscle creatine kinase enhancer. Cell **58**:823–831.
39. **Luscher, B., and R. N. Eisenman.** 1990. New light on Myc and Myb. Part I. Myc. Genes Dev. **4**:2025–2035.
40. **Ma, A., T. Moroy, R. Collum, H. Weintraub, F. W. Alt, and T. K. Blackwell.** 1993. DNA-binding by N- and L-Myc proteins. Oncogene **8**:1093–1098.
41. **Mukherjee, B., S. D. Morgenbesser, and R. A. DePinho.** 1992. Myc family oncoproteins function through a common pathway to transform normal cells in culture: cross-interference by Max and trans-acting dominant mutants. Genes Dev. **6**:1480–1492.
42. **Murre, C., P. S. McCaw, and D. Baltimore.** 1989. A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins. Cell **56**:777–783.
43. **Murre, C., P. S. McCaw, H. Vaessin, M. Caudy, L. Y. Jan, Y. N. Jan, C. V. Cabrera, J. N. Buskin, S. D. Hauschka, and A. B. Lassar.** 1989. Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence. Cell **58**:537–544.
44. **Negishi, Y., A. S. Iguchi, and H. Ariga.** 1992. Protein complexes bearing myc-like antigenicity recognize two distinct DNA sequences. Oncogene **7**:543–548.
45. **Neiman, P. E., S. J. Thomas, and G. Loring.** 1991. Induction of apoptosis during normal and neoplastic B-cell development in the bursa of Fabricius. Proc. Natl. Acad. Sci. USA **88**:5857–5861.
46. **Papoulas, O., N. Williams, and R. E. Kingston.** 1992. DNA binding activities of c-Myc purified from eukaryotic cells. J. Biol. Chem. **267**:10470–10480.
47. **Prendergast, G. C., R. Hopewell, B. J. Gorham, and E. B. Ziff.** 1992. Biphasic effect of Max on Myc cotransformation activity and dependence on amino- and carboxy-terminal Max functions. Genes Dev. **6**:2429–2439.
48. **Prendergast, G. C., D. Lawe, and E. B. Ziff.** 1991. Association of Myn, the murine homolog of max, with c-Myc stimulates methylation-sensitive DNA binding and ras cotransformation. Cell **65**:395–407.
49. **Prendergast, G. C., and E. B. Ziff.** 1991. Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region. Science **251**:186–189.
50. **Sun, X. H., and D. Baltimore.** 1991. An inhibitory domain of E12 transcription factor prevents DNA binding in E12 homodimers but not in E12 heterodimers. Cell **64**:459–470.
51. **Thiesen, H. J., and C. Bach.** 1990. Target detection assay (TDA): a versatile procedure to determine DNA binding sites as demonstrated on SP1 protein. Nucleic Acids Res. **18**:3203–3209.
52. **Tietze, K., N. Oellers, and E. Knust.** 1992. Enhancer of splitD, a dominant mutation of Drosophila, and its use in the study of functional domains of a helix-loop-helix protein. Proc. Natl. Acad. Sci. USA **89**:6152–6156.
53. **Tuerk, C., and L. Gold.** 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science **249**:505–510.
54. **Van Antwerp, M. E., D. G. Chen, C. Chang, and E. V. Prochownik.** 1992. A point mutation in the MyoD basic domain imparts c-Myc-like properties. Proc. Natl. Acad. Sci. USA **89**:9010–9014.
55. **Vinson, C. R., and K. C. Garcia.** 1992. Molecular model for DNA recognition by the family of basic-helix-loop-helix-zipper proteins. New Biol. **4**:396–403.
56. **Voronova, A., and D. Baltimore.** 1990. Mutations that disrupt DNA binding and dimer formation in the E47 helix-loop-helix protein map to distinct domains. Proc. Natl. Acad. Sci. USA **87**:4722–4726.
57. **Weintraub, H., R. Davis, S. Tapscott, M. Thayer, M. Krause, R. Benezra, T. K. Blackwell, D. Turner, R. Rupp, S. Hollenberg, Y. Zhuang, and A. Lassar.** 1991. The myoD gene family: nodal point during specification of the muscle cell lineage. Science **251**:761–766.
58. **Wenzel, A., C. Cziepluch, U. Hamann, J. Schurmann, and M. Schwab.** 1991. The N-Myc oncoprotein is associated in vivo with the phosphoprotein Max(p20/22) in human neuroblastoma cells. EMBO J. **10**:3703–3712.
59. **Zervos, A. S., J. Gyuris, and R. Brent.** 1993. Mxi-1, a protein that specifically interacts with Max to bind Myc-Max recognition sites. Cell **72**:223–232.