© 1992 Oxford University Press Nucleic Acids Research, Vol. 20, No. 9 2257–2263

Determination of the DNA sequence recognized by the bHLH-zip domain of the N-Myc protein

Regina Alex, Osman Sözeri, Sandra Meyer and Renate Dildrop*
Institute for Genetics, University of Cologne, Weyertal 121, 5000 Cologne 41, Germany

Received February 10, 1992; Accepted March 31, 1992

ABSTRACT

The DNA-binding domain of the murine N-Myc protein, comprising the basic helix-loop-helix-zipper (bHLH-zip) region was expressed as a fusion protein in *E.coli.* The affinity purified glutathione-S-transferase-N-Myc fusion protein (GST-N-MYC) was used to select the N-Myc specific DNA-recognition motif from a pool of random-sequence oligonucleotides. After seven rounds of binding-site selection, specifically enriched oligonucleotides were cloned and sequenced. Of 31 individual oligonucleotides whose sequences were determined, 30 contained a common DNA-motif, defining the hexameric consensus sequence CACGTG. We confirm by mutational analysis that binding of the N-Myc derived bHLH-zip domain to this motif is sequence-specific.

INTRODUCTION

The N-*myc* gene is conserved as a distinct sequence in vertebrate species, and belongs to the *myc* gene family of oncogenes (1). Myc proteins are involved in the control of cell proliferation and differentiation, and aberrant expression of these proteins has been implicated in the genesis of a variety of neoplasms (2,3). The N-*myc* gene product is a phosphorylated protein which is located in the cell nucleus and has the ability to bind to DNA, although sequence-specific DNA-binding has so far not been demonstrated (4,5).

Recently, the proteins of the Myc family have been shown to possess structural similarities with a number of other DNA-binding proteins (6). The structural homology is limited to approximately 100 amino acids located at the C-terminus of the Myc proteins and contains a common DNA-binding and dimerization motiv, the basic-Helix-Loop-Helix motif (bHLH). Proteins containing this motif are of widespread occurrence among eucaryotic species (7,8), and for several of these proteins specific DNA-binding has been demonstrated. The integrity of the basic region is essential for sequence-specific DNA-binding (9,10), and the two amphipathic helices of the bHLH motif mediate the formation of homo- or heterodimeric protein complexes (9,11). In addition to the bHLH motif, the members of the Myc protein family contain a leucine-zipper (Leu-zip) dimerization motif (12), which is located directly adjacent to the bHLH motif.

Sequence-specific DNA-binding has been established for a number of bHLH proteins, and the commonly recognized DNA-motif defines a hexameric consensus sequence CANNTG, a sequence motif (E-box), originally identified in immunoglobulin enhancer elements (13).

Recently, the DNA-binding specificities for the c-Myc oncoprotein (14–17), for transcription factor TFE3 (18), and for the USF protein (19) have been determined. These proteins recognize the palindromic hexamer motif CACGTG. Since the N-Myc protein is strongly homologous to the c-Myc protein, and also shares considerable homology with the other bHLH proteins, we expected the protein to recognize a related E-box motif. As demonstrated by the binding-site selection described below, the N-Myc derived bHLH-zip domain recognizes the sequence CACGTG.

MATERIALS AND METHODS

Construction of GST-N-MYC fusionproteins

A *Sau*3A fragment extending from position 5961 to position 6285 in the murine N-*myc* gene (20) was ligated in frame into the *Bam*HI-cleaved glutathione-S-transferase expression vector pGEX-3X (21). The hybrid gene codes for the glutathione-S-transferase-N-Myc fusion protein GST-N-MYC. A derivative of this protein (GST-N-MYC-HD), lacking part of the basic region within the bHLH-zip domain, was constructed by deleting a *Hae* II fragment encoding amino acids 380 to 390 of the N-Myc protein. Expression constructs were transformed into *E.coli* K 12 strain BL21 DE3 (22), and purification of GST-wildtype and GST-N-MYC fusion proteins was performed as described previously (21). Purified proteins were dialysed against MTPBS (21) and stored in 50% Glycerin/50% MTPBS. Protein concentrations were determined by measuring the optical density (OD 280), using the equation 1OD 280 ≈0,5 mg/ml (21), and by estimating the concentration of coomassie-stained proteins which were electrophoretically resolved by SDS-PAGE.

DNA fragments and oligonucleotides

All DNA oligonucleotides were synthesized on an Applied Biosystems Synthesizer Model 394. The structure of the 63 base oligonucleotide mixture is as follows: 5′-GTTCTCGCATGCA-GGCTTGG $(N)_{23}$ GGAACTCAGGATCCGTGACC-3′. The

* To whom correspondence should be addressed

sequence of the 20 base oligonucleotides used as primers for PCR amplification is as follows: primer A: 5′-GTTCTCGCATGCA-GGCTTGG-3′; primer B: 5′-GGTCACGGATCCTGAGTT-CC-3′; the *Bam*HI and *Sp*HI restriction enzyme recognition sites, which allow cloning of individual fragments, are underlined. Doublestranded 63 bp oligonucleotides were generated by annealing the 63 base singlestranded DNA mixture to an equimolar amount of primer B, followed by a fill in reaction using the Klenow fragment of *E.coli* DNA polymerase. Labeling of oligonucleotide mixtures was performed by primer extension using TAQ-polymerase under standard PCR conditions: approximately 5 ng of purified template DNA was labeled in PCR reaction buffer (Amersham) supplemented with 20 μCi of ^{32}P dCTP (3000 Ci/mmol), 200 μM of each dATP, dGTP and dTTP, and 20 pmoles of primer A and B, each. The synthetic oligonucleotide o12(CG) was produced by annealing the following single stranded fragments: 5′-AGCTTCCCAACGCACGTGT-CGTCGTC-3′ and 5′-CTAGGACGACGACACGTGCGTTG-GGA-3′. Three derivatives of this oligonucleotide (o12AT, o12GC, o12TA) were generated by annealing corresponding single stranded fragments whose sequences in the underlined positions were exchanged to AT, GC, and TA, respectively.

Individually cloned fragments, that were excised by restriction digestion as well as the control oligonucleotide O$_1$ (5′-CTAG-AATTGTTATCCGCTCACAATT-3′; 5′-CTAGAATTGTGA-GCGGATAACA-ATT-3′) (23) which exhibits 3′-recessive ends, were endlabeled using the Klenow subunit of *E.coli* DNA polymerase.

DNA binding analysis

The South Western analysis was performed as described previously (24), using *Hinc*II digested phage λ-DNA which was labeled by nick-translation to a specific activity of 100 000 cpm/ng.

Selection of oligonucleotides, specifically bound by the GST-N-MYC protein, was done by subsequent rounds of filterbinding. 2ng of the 63 bp oligonucleotide mixture were incubated with 200 ng of the GST-N-MYC protein (approximately 100ng full-length protein) in binding buffer (20mM HEPES (pH 7.6), 50mM KCl, 1mM DTT, 1mM EDTA, 8% Glycerol), which contained 100 ng poly (dI-dC) (Boehringer Mannheim) as nonspecific competitor DNA. Incubation was performed in a volume of 20 μl for 20 minutes at room temperature. After the incubation step, the reaction volume was increased to 100 μl by adding binding buffer, and then, the sample was filtered through cellulose nitrate filters (poresize 0.45 μm, Sartorius). Filters were washed with 1ml binding buffer, dried, cut into slices, and subsequently boiled in 100 μl water for 2 minutes to elute the bound DNA. A PCR amplification in standard buffer (Amersham) was performed with 50 μl of this solution, using 20 pmoles of each of primers A and B, and 200 μM of each of the four nucleotides (Boehringer Mannheim). The concentration of the PCR products was estimated from agarose gels, and approximately 2ng of the amplified DNA was used for the next round of binding-site selection without any further purification. After four rounds of filterbinding multimeric PCR products appeared, an effect, caused by high DNA concentrations during the first PCR cycles, which was subsequently circumvented by lowering the amount of recovered template oligonucleotides added to the PCR reaction. The PCR amplification after the fourth round of binding-site selection was therefore repeated with 10 μλ instead of 50 μl, as was the 5th PCR. In the PCR reaction after the 6th selection step 0.5 μl of the solution, containing the eluted template oligonucleotides, were amplified, and after the 7th step, amplification was performed with 0.1 μl of the DNA solution.

Binding reactions for retardation assays were performed with 0.2 ng of labeled oligonucleotides (100 000 cpm/ng), 200 ng GST-N-MYC or GST-N-MYC-HD protein, respectively, 100 ng poly (dI-dC) as nonspecific competitor DNA, in 20 μl binding buffer. The mixture was incubated for 20 min at room temperature and immediately loaded onto 5% polyacrylamide gels (acrylamide: bisacrylamide, 30:1). Electrophoresis was performed in 0,5×TBE (1×TBE: 89 mM tris-borate, 2mM EDTA; pH 8.3) for 90 minutes at 11 mA. Prior to loading, gels were prerun for 3 hours at 11 mA. The retarded DNA-protein complexes formed with oligonucleotides recovered after the 7th round of binding-site selection (fraction S7) were cut out of the dried gel, and the DNA eluted over night at 56°C in 333-buffer (300 mM NaCl, 30 mM Tris/Cl pH 8.0, and 3mM EDTA pH 8.0). After phenol extraction and ethanol precipitation, one fifth of the eluted DNA was reamplified by PCR (fraction SR7).

Cloning of PCR products

1) Blunt end ligation: PCR-amplified oligonucleotides were flush ended (25), using T$_4$-DNA polymerase, and ligated into Hinc II digested pUC 19 vector (sequences derived from such blunt ended inserts are marked with a 'b' in Figure 4). 2) Sticky end ligation: Amplified oligonucleotides were incubated with Proteinase K at 37°C for 1 hour (26), and then restricted with *Bam*HI and *Sp*HI. The digested fragments were ligated into a linearized pUC 19 derivative (pUC 19-I) with compatible ends, and transformed into DH5α cells. PUC 19-I (C. Kresse, personal communication) was constructed as follows: a 629 bp Dra I fragment of pBR322 was cloned into the *Hinc* II site of pUC 19. Within this plasmid, the sequence coding for the α-peptide of the β-galactosidase is out of frame, and therefore, DH5α cells carrying this plasmid do not have the α-complementation activity. Complete digestion of pUC19-I with the two restriction enzymes (*Bam*HI and *Sp*HI) which flank the pBR322 insert, produces compatible ends for sticky end ligation, and the cloning of enriched oligonucleotides into this vector DNA restores the reading frame of the α-peptide. This leads to the formation of blue colonies in the DH5α-background, when grown on X-Gal containing plates. Single blue colonies were picked and plasmid DNA prepared (Quiagen tip 100; Diagen). Sequencing reactions were performed according to the instructions of the manufactor (Sequenase-Kit, USB).

RESULTS

Expression of the N-Myc DNA-binding domain in bacteria

To express the DNA-binding domain of the N-Myc protein in bacteria we used the pGEX expression system (21). We fused a *Sau*3A restriction fragment derived from exon 3 of the murine N-*myc* gene (positions 5961–6285 in reference 20) to the 3′-end of the glutathione-S-transferase (GST) gene in the expression vector pGEX-3X (21). The resulting GST-N-MYC fusion protein contains a 109 amino acid region spanning residues 348 to 456 of the N-Myc protein. This region contains the complete bHLH-zip domain and an additional stretch of 30 amino acids N-terminal to the basic region (Figure 1). In a second construct we deleted codons 380 to 390 of the N-*myc* coding sequence, which leads to a mutant protein (GST-N-MYC-HD) that lacks part of the basic region (Figure 1). Deletions within the basic region of the bHLH-zip

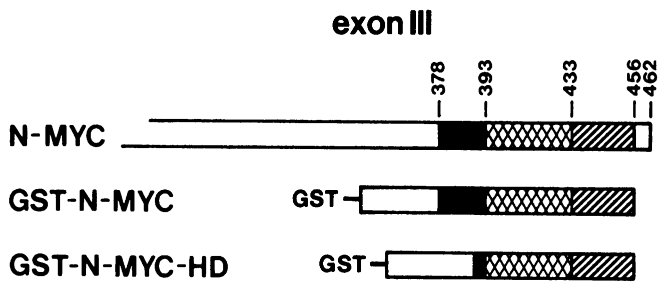

Figure 1. Structure of glutathione-S-transferase-N-Myc fusion proteins. On top, exon 3 of the murine N-*myc* gene, containing the bHLH-zip domain, is shown. Positions of the basic region (codons 378–392), the helix-loop-helix motif (codons 393–432), and the Leu-zipper (433–456) are indicated. Below, the GST-N-MYC fusion protein, used for the binding-site selection, is shown, and on the bottom, the structure of a mutant protein (GST-N-MYC-HD), lacking amino acids 380 to 390 within the basic region.

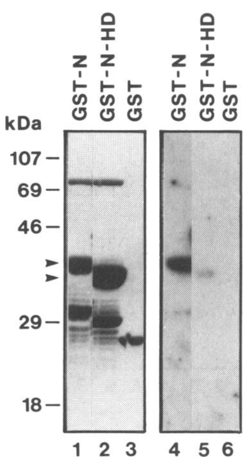

Figure 2. South-Western analysis of purified fusion proteins. Coomassie stained proteins resolved by SDS-PAGE are shown in lanes 1 to 3. Arrow heads indicate the positions of intact fusion proteins. Lanes 4 to 6 show the result of the South-Western analysis performed in parallel. The GST fusion protein containing the complete bHLH-zip domain of the N-Myc protein (GST-N) binds to phage λ DNA appoximately 50 times more efficiently than the mutant protein (GST-N-HD).

domain should reduce or abolish the ability of the protein to bind to DNA as already demonstrated for other bHLH proteins (9).

The GST-N-MYC fusion proteins were expressed in *E. coli* and purified by affinity-chromatography on glutathione-agarose beads. SDS-PAGE analysis (Figure 2, lanes 1 and 2) reveals that we copurified a bacterial protein with an apparent molecular weight of 70 KDa. This protein, which made up about 10% of the total protein preparation, did not copurify with the GST protein alone (lane 3), nor with other GST-N-MYC fusion proteins containing regions of the N-Myc protein other than the bHLH-zip domain (data not shown). Although the SDS-PAGE analysis reveals partial degradion of the GST-N-MYC fusion proteins, the apparent molecular weights (38 KD and 36 KD, respectively) of approximately half of the total protein present in the two preparations (lanes 1 and 2) corresponds to that of the intact fusion proteins. In order to test, whether the N-Myc derived bHLH-zip domain in conjunction with the GST protein has the ability to bind to DNA, a South-Western analysis was performed (Figure 2; lanes 4 to 6). For this purpose, the proteins were electrophoretically resolved by SDS-PAGE and blotted onto a nitrocellulose filter, which was then incubated with radiolabeled DNA. We used *Hinc* II digested phage λ-DNA, since we expected the phage DNA to be complex enough to contain suitable sequences for DNA-binding. Figure 2 shows that the GST-N-MYC-protein exhibits DNA-binding in this assay (lane 4), and that the DNA-binding of the mutant protein (GST-N-MYC-HD) is at least 50 fold weaker than that of the wildtype protein (lane 5). DNA-binding of the GST protein is not detectable in this assay (lane 6), nor do we see DNA-binding of the 70 KD protein copurified with the GST fusion proteins (lanes 4 and 5).

Selection of DNA sequences specifically bound by the N-Myc bHLH-zip domain

To determine the 'DNA sequences which are specifically recognized by the bHLH-zip domain of the N-Myc protein, we chose the strategy of binding-site selection from random-sequence oligonucleotides (27). For this purpose an oligonucleotide mixture, 63 bases in length, was generated, containing a central region of 23 nucleotides of totally random sequence, flanked on either side by 20 nucleotides of fixed sequence. Two oligonucleotides (20 bases in length) corresponding to the flanking sequences were synthesized, to serve as 3′ and 5′ primers for PCR amplification. The GST-N-MYC protein was incubated with the double stranded oligonucleotide mixture (see Materials and Methods), and oligonucleotides specifically bound by the protein were subsequently separated from the bulk of unbound DNA by filter binding (28). The reaction mixture, containing free protein, free DNA and protein-DNA complexes, was applied to a nitrocellulose-filtering device. While protein-DNA complexes and free proteins are held back by the nitrocellulose filter, the unbound DNA is not retained and can be washed off. The DNA recovered from the filter was amplified by PCR, and subjected to the next round of binding-site selection. Seven successive rounds of binding-site selection were performed, to ensure enrichment of specifically bound oligonucleotides. The PCR products of the sixth and seventh step (S6 and S7) were radiolabeled and tested in a gel mobility shift assay for their binding to the GST-N-MYC fusion protein. Incubation of the GST-N-MYC protein with labeled S6 or S7 oligonucleotides both resulted in DNA-protein complexes (data not shown). The GST-N-MYC protein-DNA complex of the S7 fraction was isolated from the gel, and the recovered DNA reamplified by PCR (fraction SR7). Figure 3 (lanes 1 and 2) shows the gel mobility shift assay performed with radiolabeled SR7 DNA and the unselected starting oligonucleotide population (S0). In contrast to oligonucleotides of the S0 fraction, which are not visibly retarded upon incubation with the GST-N-MYC protein (lane 1), oligonucleotides of the SR7 fraction give rise to DNA-protein complexes, indicating successful enrichment of specifically bound oligonucleotides.

Sequence analysis

The DNA of the SR7 fraction was digested with *Bam*HI and *Sph*I,the two enzymes whose recognition sites lie within the fixed 5′ and 3′ sequence of the oligonucleotide mixture. After ligation into a pUC 19 derivative (see Materials and Methods), recombinant plasmids were transformed into DH5α cells. Plasmid DNA from single bacterial colonies was isolated and the DNA sequences of individual inserts determined.

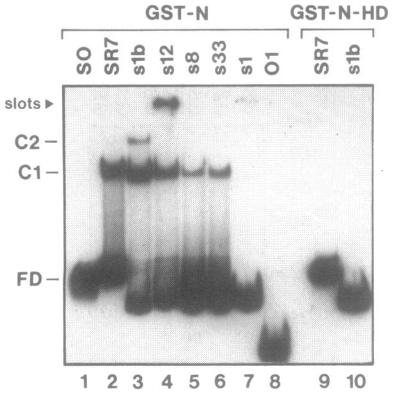

Figure 3. Gel retardation analysis with oligonucleotides enriched by the binding-site selection. Retardation assays performed with the GST-N-Myc fusion protein (GST-N) are shown in lanes 1 to 8, and retardation assays performed with the mutant protein (GST-N-HD) are shown in lanes 9 and 10. The type of radiolabeled DNA, present in the individual binding reactions, is indicated on top of the autoradiogram, the positions of retarded DNAs (C1 and C2), and the positions of free DNA (FD), is indicated on the left. An arrow head marks the origin of the gel, were part of some of the samples was retained in the slots. Lane 1, unselected starting oligonucleotide mixture (S0); lanes 2 and 9, selected oligonucleotide mixture after 7 rounds of binding-site selection (SR7); lanes 3 to 6 and 10, individual oligonucleotides derived from fraction SR7 (s1b: two CACGTG motifs; s12: one CACGTG motif; s8: one CACGCG and one CACATG motif; s33: one CTCGTG motif); lane 7, oligonucleotide s1, which was derived from fraction SR7, but which lacks the hexamer-motif; lane 8, control oligonucleotide O1, also lacking the hexamer-motif (see Materials and Methods).

A

```
s1b     ttccGCCCACGTGACTAATCTCACGTGccaa

s42     ttccCGCGTGTAACCGCTGCACGTGGAccaa
s5      ttccCTCGTGGTGCCTACACGTGTCCAccaa
s24     ttccCACGTGCAGCAAACTCGTGTGGCccaa
s39     ttccGCCGCGTGAAACCCTGTCACGTGccaa
s9      ttccATCACGTGTTACGGGCATGTGTTccaa
s40     ttccTACGCGTGCGTAACACGTGCCTCccaa
s19     ttccCAAAGCACGTGGCACCGCGTGTAccaa

s8      ttccCACGCGTGGCTTACACATGCCTCccaa
s43     ttccCACGACGTGAAAGTCACGCGCACccaa
```

B

```
s6      ttccCACGTGTACTAACCCCCATCCGATccaa
s13     ttccCACGTGGTGCGAAGGCCTTCTACccaa
s17     ttccCACGTGCACACATCGATAAGTGCccaa
s45     ttccCCCACGTGATACAACATCAGCCCccaa
s25     ttccCCACACGTGTCCCGGGACGGTGGccaa
s32     ttccGACCACGTGTAACCCCTGCAACCccaa
s43b    ttccCGCCACGTGGTACTGAACCGAACccaa
s38     ttccTGCCCACGTGTATCTCCGCGTACccaa
s12     ttccCAACGCACGTGTCGTCGTCCGTCccaa
s36     ttccCACCAAAACACGTGGGAGGCCAAccaa
s7      ttccGGAGTCGTAGCACGTGCCTCTCCccaa
s22     ttccGCAAGCCCCCCACGTGTTGTCCccaa
s14     ttccCGCAATGCTTGTCACGTGTCGCCccaa
s23     ttccCCCTCCAGCATCGCACGTGCCGTccaa
s46     ttccCGCCTGAGTAAAGTACACGTGACccaa
s31     ttccACCCAGAGTATTAGTACACGTGGccaa

s47     ttccACGCGACGTCAAATTCCGGTCACccaa
s4      ttccCACATGGCGCGCCATCTGCTGTAccaa
s51     ttccACGCAAACCTCGTGCGTATGGTCccaa
s33     ttccAACCGGACTATCCTCGTGTTCTGccaa
```

C

```
s1      ttccGAACGATATATCCTGCATCCGCCccaa
```

Figure 4. Sequences of individual oligonucleotides, selected by the GST-N-MYC protein from the pool of randomized oligonucleotides. Nucleotides shown in uppercase letters display the central region of the oligonucleotide mixture, originally constituting the randomized sequence area, and lowercase letters indicate the constant sequences flanking the randomized area. Hexamer motifs, which are common to the oligonucleotides, display the consensus sequence CACGTG (underlined positions). (A) Nucleotide sequences of selected oligonucleotides, which contain two hexamer motifs. (B) Nucleotide sequences containing one hexamer motif. (C) Nucleotide sequence of oligonucleotide s1, lacking the consensus motif.

The DNA sequences from 31 individual oligonucleotides derived from the SR7 fraction were determined (Figure 4). All but one of these sequences exhibit a common hexameric motif that is in at least five of the six positions identical to the sequence CACGTG (underlined positions in Figure 4). The only sequence lacking the consensus CACGTG motif is sequence s1 (Figure 4C). In one third of the sequences (10 out of 30) the hexamer motif is present twice (Figure 4A), although in most cases (7 out of 10) one intact palindromic CACGTG motif is accompanied by a hexameric motif which deviates from the consensus sequence at one position. In addition, two out of the ten sequences (s8 and s43) each displays two incomplete motifs. Sequence s1b is the only sequence containing two complete CACGTG motifs. Figure 4B shows the sequences of oligonucleotides containing a single hexamer motif. In contrast to the sequences containing two hexamers, the majority of these sequences (16 out of 20) exhibit the intact palindromic motif CACGTG.

The comparison of all hexamer motifs and their flanking regions is shown in Figure 5. Counting the base composition for every position clearly reveals the hexameric consensus sequence CACGTG. Within this palindromic hexamer only certain nucleotide exchanges are found which differ from the consensus sequence. In position 1 of the CAC-halfsite thymidine (T) occurs once, which corresponds to the adenine (A) exchanges in the GTG-halfsite. In position 2 guanine (G) and thymidine (T) exchanges are found, corresponding to the cytosine (C) exchanges present in the GTG-halfsite. In position 3 we only found one guanine (G) substitution within the 40 hexamers compared. Comparison of the positions which flank the hexameric motif (positions 4 to 7) reveals no significant sequence preferences, although at position 4 of the CAC-halfsite cytosine (C) occurs in one half of the sequences examined (Figure 5). From this analysis we conclude that under the conditions used for the binding-site selection, the minimal sequence motif which is specifically recognized by the GST-N-MYC protein exhibits the palindromic sequence CACGTG.

Binding specificities

In order to prove that the CACGTG motif is specifically recognized by the bHLH-zip domain of the N-MYC protein, retardation assays were performed with a subset of the oligonucleotides derived from the SR7 fraction (Figure 3). For this purpose the following oligonucleotides were chosen: s1b (two CACGTG motifs), s12 (one CACGTG motif), s8 (one CACGCG plus one CACATG motif), s33 (one CACGAG motif) and s1 (no hexameric motif). DNA fragments were excised from the pUC vector by endonuclease digestion, radiolabeled, and binding

Pos.	6	5	4	**3**	**2**	**1**	**1′**	**2′**	**3′**	4′	5′	6′
A	7	10	9	**0**	**32**	**0**	**2**	**0**	**0**	6	11	11
C	11	15	20	**39**	**0**	**39**	**0**	**3**	**0**	11	16	11
G	6	6	6	**1**	**4**	**0**	**38**	**0**	**40**	8	5	11
T	16	9	5	**0**	**4**	**1**	**0**	**37**	**0**	15	8	7
	N	N	c	**C**	**A**	**C**	**G**	**T**	**G**	N	N	N

Figure 5. Sequence comparison of selected oligonucleotides reveals a common DNA motif. The hexameric consensus sequence CACGTG was deduced from the comparison of 40 CACGTG-related sequence motifs present in the selected oligonucleotides shown in Figure 4. The base composition at each position of the hexamer motif, and the 3 flanking positions are given. Strongly conserved positions are printed in bold letters, and a position conserved less well, immideately 5′ to the CACGTG motif, is printed in lower case letter.

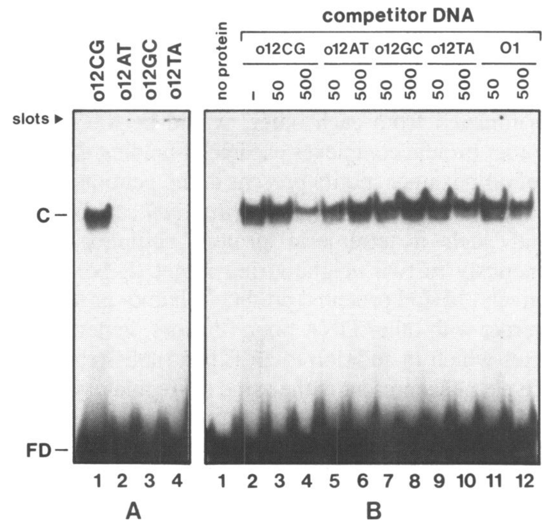

Figure 6. The CACGTG motif is specifically bound by the GST-N-MYC protein. (A) Gel retardation analysis of oligonucleotides bearing the CACGTG motif (o12CG, lane 1), and of 3 derivatives (o12AT, o12 GC, o12TA, lanes 2–4), which contain the indicated dinucleotide exchanges at the inner CG-positions of the hexamer. Only oligonucleotide o12CG, which contains the intact CACGTG motif, is retarded by the GST-N-MYC protein. This is also demonstrated by the competition experiments shown in (B). To reaction mixtures, containing the GST-N-MYC protein and a constant amount of radiolabeled oligonucleotide o12CG (0,2 ng), increasing amounts (50× and 500×molar excess) of unlabeled competitor DNA was added. Only oligonucleotide o12CG is able to compete for binding to the GST-N-MYC protein (lanes 2–4), all other competitor DNAs, which do not contain an intact CACGTG motif, fail to compete for binding.

to the GST-N-MYC protein was tested by the gel shift analyses shown in Figure 3 (lanes 3 to 8). All fragments containing a hexamer motif were retarded by the GST-N-MYC protein, although with different efficiencies (lanes 3 to 6). Those fragments which contain intact palindromic sequences, especially fragment s1b (two CACGTG motifs) exhibit stronger binding than those with hexamer motifs, where only one of the halfsites of the palindromic sequence is kept intact. The retardation assay performed with fragment s1b reveals a second protein complex (C2), probably due to the simultaneous occupation of both CACGTG motifs present in this fragment. In contrast, fragment s1 (lane7), which does not contain the hexamer motif, is not retarded by the GST-N-MYC protein, nor is the oligonucletide O1 (lane 8) which only contains one halfsite of the motif (see Materials and Methods).

We also performed retardation assays with derivatives of the s12 oligonucleotide (one CACGTG motif) containing base substitutions in the central two positions (CG) of the hexamer sequence. Four oligonucleotides (o12CG, o12AT, o12GC, o12TA) were synthesized: oligonucleotide o12CG contains the unmutated CACGTG motif, while in the other three oligonucleotides the central CG-dinucleotide of the hexamer motif was replaced by AT, GC or, TA, respectively. Retardation assays performed with these four oligonucleotides show that only oligonucleotide o12CG, which contains the unmutated CACGTG motif, is retarded by the GST-N-MYC protein (Figure 6A). Competition assays (Figure 6B) also demonstrate that these nucleotide exchanges abolish DNA recognition by the protein. None of the oligonucleotides containing mutated hexameric motifs were able to compete for binding of the GST-N-MYC protein to the oligonucleotide o12CG (Figure 6B).

The ability of the GST-N-MYC protein to recognize its DNA target is dependent on the integrity of the basic region within the bHLH-zip domain of the protein. We constructed a derivative of the GST-N-MYC fusion protein by deleting codons 380 to 390 of the N-*myc* coding sequence (Figure 1). The mutant protein (GST-N-MYC-HD) which lacks part of the basic region was tested for its ability to bind to oligonucleotides containing the CACGTG motif (Figure 3). In contrast to the wildtype protein (GST-N-MYC) which binds to oligonucleotides of the SR7 fraction and to the oligonucleotide s1b (lanes 2 and 3), the mutant protein (GST-N-MYC-HD) does not retard these fragments (lanes 9 and 10).

DISCUSSION

In order to determine the DNA sequence which is specifically recognized by the N-Myc protein, we performed a binding-site selection from random-sequence oligonucleotides using a bacterially expressed glutathione-S-transferase-N-Myc fusion protein (GST-N-MYC). The DNA sequences of 31 selected oligonucleotides were determined. In all but one of the sequences analysed we identified a common motif displaying the hexameric consensus sequence CACGTG (Figure 4). Several lines of evidence support our finding that binding of the N-Myc bHLH-zip domain to this sequence is specific: (1) Gel retardation assays performed with individual oligonucleotides demonstrate, that only DNA fragments containing the hexamer motif are retarted by the GST-N-MYC protein, while DNA fragments devoid of the CACGTG sequence are not retarted (Figure 3). (2) Mutations introduced into the CACGTG motif by exchanging the central dinucleotide (CG) by either GC, AT or TA, abolish recognition by the GST-N-MYC protein, as demonstrated by retardation and competition analyses (Figure 6). (3) We generated a mutant GST-N-MYC protein which carries a deletion within the basic region, the integrity of which is essential for DNA binding of other bHLH proteins. In retardation experiments this mutant protein (GST-N-MYC-HD) is unable to recognize DNA-fragments carrying the CACGTG motif (Figure 3).

The hexameric CACGTG motif which we have identified belongs to the family of E-box motifs, which exhibit the consensus sequence CANNTG. E-box motifs are present in a number of regulatory gene elements such as the immunoglobulin heavy chain (μE) enhancer, the adenovirus major late (AdML) promotor and the muscle creatine kinase (MCK) enhancer (13,29). All known target sequences of bHLH proteins contain the CANNTG motif (8), and the CACGTG motif has recently been shown to be recognized by several bHLH proteins. These proteins are transcription factor E3 (TFE3) which activates transcription through the immunoglobulin enhancer μE3 motif and also binds to the AdML motif (18), protein TFEB which is closely related to TFE3 (30), the upstream stimulatory factor (USF), a cellular factor required for efficient transcription of the AdML promoter *in vitro* (19), the yeast centromere-binding protein CBF1 (31), the c-Myc oncoprotein (14−16) which *in vitro* binds the perfectly palindromic sequence GACCACGTGGTC with the highest efficiency (17), and the MAX/Myn proteins which are capable of forming heterodimeric complexes with Myc proteins (32,33). Since the N-Myc protein is strongly homologous to the c-Myc protein and also shares considerable homology within the basic region when compared to the other bHLH proteins, we expected the N-Myc protein to recognize a related or, as demonstrated here, the identical DNA-motif.

Within the collection of selected oligonucleotides which we have analyzed (Figure 4), the majority of sequences (24 out of 30) contain the palindromic CACGTG motif. The symmetry of the recognized DNA sequence suggests, that each halfsite of the sequence is contacted by one subunit of an oligomeric (dimeric or tetrameric) protein complex. Potent dimerization domains (the HLH and the Leu-zip motifs) are present in the GST-N-MYC protein, and dimerization has been shown to be essential for stable 'DNA-binding of other bHLH-proteins (9,11). In addition to the palindromic CACGTG motif, incomplete hexamer motifs which deviate from the CACGTG consensus at one position are also present in the sequence collection (Figure 4). Within the 24 sequences displaying a CACGTG motif, seven contain in addition an incomplete hexamer motif, and six oligonucleotides contain incomplete hexamer motifs only. We have shown, that single nucleotide exchanges still allow binding by the GST-N-MYC protein as examplified by oligonucleotides s8 and s33 (Figure 3). From the type of nucleotide differences we observe in these hexamer motifs (Figure 5) we assume that there is a restricted set of exchanges which still allow recognition by the DNA-binding domain of the N-Myc protein, and that at least one halfsite present in those hexamers must exhibit the sequence CAC. The latter point is supported by the fact that we did not find hexamer motifs, bearing more than one exchange from the consensus sequence, and that mutated CACGTG motifs where both halfsites were affected, were no longer recognized by the GST-N-MYC protein (Figure 6).

Sequence comparison of the oligonucleotides enriched by our selection reveals a short, hexameric binding site. At positions adjacent to the hexamer, the only marked preference is the occurence of a cytosine (C) at position 4 of the CAC-halfsite (Figure 5). We find a cytosine (C) at this position (or guanine (G) adjacent to the GTG-halfsite) in more than half (24 out of 40) of the hexamer motifs (Figure 4). The preference for cytosine (C) at this position is also reflected by the fact that within the 30 oligonucleotide sequences depicted in Figure 4, nine hexamer motifs are positioned directly adjacent to the fixed border sequence at the 5′-site, which already provides cytosine residues (for example oligonucleotides s42, s5, s24, in Figure 4A). Otherwise comparison of oligonucleotides enriched by our selection does not reveal additional sequence preferences, and we thus conclude that under the *in vitro* assay conditions we used, the positions flanking the hexamer seem not to be important for DNA binding. These data correlate with the finding that the bacterially expressed bHLH domain of the c-Myc protein shows no strong preferences for bases flanking the hexameric core sequence in an SAAB (selected and amplified binding-site) imprinting assay, an *in vitro* binding-site selection, similar to the one used in the present study (14).

Under in vivo conditions, we assume that a hexameric DNA motif as such would not be sufficient to direct protein-binding to only a few sites within the genome, the true target sites relevant for transcriptional regulation. Assuming a statistical base composition, a hexameric sequence will appear about 1×10^6 times in the mammalian genome, and thus, discrimination must occur. Discrimination could be achieved by an extended target sequence, in which sequences adjacent to the core E-box motif might enhance specific recognition in vivo. It has been reported that the efficiency of protein-binding to a given core E-box motif is strongly influenzed by the neighbouring sequences (15,17), suggesting that the core E-box motif is neccessary, but not sufficient for specific recognition. An additional mechanism to enhance the selectivity by which potential DNA target sites can be discriminated from each other, would be the formation of higher order protein complexes capable of binding to only a small subset of all hexamer motifs present in the genome. It has been shown, that c-Myc protein produced in *E.coli* can form tetramers (24) and such a tetrameric protein complex could bind simultaneously to two neighbouring target (E-box) sites (34). Alternatively, bHLH proteins (present as homo- or heterodimers) may interact with other DNA bound factors, generating protein complexes which in addition to an E-box motif recognize other sequence elements present in the same cis-regulatory region (35).

ACKNOWLEDGEMENTS

We are grateful to B.Müller-Hill and S.Oehler for their constant support, and to B.Jack, K.Rajewsky and B.von Wilken-Bergmann for critically reading the manuscript. Oligonucleotides were kindly provided by B.Müller-Hill and S.Oehler. Photographies were kindly prepared by U.Ringeisen. This work was supported by the Bundesministerium für Forschung und Technologie and by a grant of the Konrad-Adenauer-Stiftung to R.A.

REFERENCES

1. Alt, F.W., DePinho, R., Zimmerman, K., Legouy, E., Hatton, K., Ferrier, P.,Tesfaye, A., Yancopoulos, G. and Nisen, P. (1986) *Cold Spring Harbor Symp. Quant. Biol.*, **51**, 931−941.
2. Cory, S. (1986) *Adv. Cancer Res.*, **47**, 189−234.
3. Lüscher, B. and Eisenman, R.N. (1990) *Genes Dev.*, **4**, 2025−2035.
4. Slamon, D.J., Boone, T.C., Seeger, R.C., Keith, D.E., Chazin, V., Lee, H.C. and Souza, L.M. (1986) *Science*, **232**, 768−772.
5. Ramsay, G., Stanton, L., Schwab, M. and Bishop, J.M. (1986) *Mol. Cell. Biol.*, **6**, 4450−4457.
6. Murre, C., Schonleber McCaw, P. and Baltimore, D. (1989) *Cell*, **56**, 777−783.
7. Jones, N. (1990) *Cell*, **61**, 9−11.
8. Olson, E.N. (1990) *Genes Dev.*, **4**, 1454−1461.
9. Davis, R.L., Cheng, P.-F., Lassar, A.B. and Weintraub, H. (1990) *Cell*, **60**, 733−746.
10. Weintraub, H., Dwarki, V.J., Verma, I., Davis, R., Hollenberg, S., Snider, L., Lassar, A. and Tapscott, S.J. (1991) *Genes Dev.*, **5**, 1377−1386.
11. Voronova, A. and Baltimore, D. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 4722−4726.
12. Landschulz, W.H., Johnson, P.F. and McKnight, S.L. (1988) *Science*, **240**, 1759−1764.
13. Calame, K. and Eaton, S. (1988) *Adv. Immunol.*, **43**, 235−270.

14. Blackwell, T.K., Kretzner, L., Blackwood, E.M., Eisenman, R.N. and Weintraub, H. (1990) *Science*, **250**, 1149–1151.
15. Prendergast, G.C. and Ziff, E.B. (1991) *Science*, **251**, 186–189.
16. Kerkhoff, E., Bister, K. and Klempnauer, K.H. (1991) *Proc. Natl.Acad. Sci. USA*, **88**, 4323–4327.
17. Halazonetis, T.D. and Kandil, A.N. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 6162–6166.
18. Beckmann, H., Su, L.-K. and Kadesch, T. (1990) *Genes Dev.*, **4**, 167–179.
19. Gregor, P.D., Sawadogo, M. and Roeder, R.G. (1990) *Genes Dev.*, **4**, 1730–1740.
20. DePinho, R.A., Legouy, E., Feldman, L.B., Kohl, N.E., Yancopoulos, G.D. and Alt, F.W. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 1827–1831.
21. Smith, D.B. and Johnson, K.S. (1988) *Gene*, **67**, 31–40.
22. Studier, F.W. and Moffatt, B.A. (1986) *J. Mol. Biol.*, **189**, 113–130.
23. Gilbert, W. and Maxam, A. (1973) *Proc. Natl. Acad. Sci. USA*, **70**, 3581–3584.
24. Dang, C., McGuire, M., Buckmire, M. and Lee, W.M.F. (1989) *Nature*, **337**, 664–666.
25. Hemsley, A., Arnheim, N., Toney, M.D., Cortopassi, G. and Galas, D.J. (1989) *Nucleic Acids. Res.*, **17**, 6545–6551.
26. Crowe, J.S., Cooper, H.J., Smith, M.A., Sims, M.J., Parker, D. and Gewert, D. (1991) *Nucleic Acids Res.*, **19**, 184.
27. Oliphant, A.R., Brandl, C.J. and Struhl, K. (1989) *Mol. Cell. Biol.*, **9**, 2944–2949
28. Riggs, A.D., Newby, R.F., Bourgeois, S. and Cohn, M. (1968) *J. Mol. Biol.*, **34**, 365–368.
29. Buskin, J.N. and Hauschka, S.D. (1989) *Mol. Cell. Biol.*, **9**, 2627–2640.
30. Carr, C.S. and Sharp, P.A. (1990) *Mol. Cell. Biol.*, **10**, 4384–4388.
31. Cai, M. and Davis, R.W. (1990) *Cell*, **61**, 437–446.
32. Blackwood, E.M. and Eisenman, R.N. (1991) *Science*, **251**, 1211–1217.
33. Prendergast, G.C., Lawe, D. and Ziff, E.B. (1991) *Cell*, **65**, 395–407.
34. Peterson, C.A. (1991) *The New Biologist*, **3**, 442–445.
35. Lamb, P. and McKnight, S.L. (1991) *TIBS*, **16**, 417–422.