MOLECULAR AND CELLULAR BIOLOGY, Dec. 1993, p. 7612–7624
0270-7306/93/127612-13$02.00/0

Vol. 13, No. 12

CTCF, a Conserved Nuclear Factor Required for Optimal Transcriptional Activity of the Chicken c-*myc* Gene, Is an 11-Zn-Finger Protein Differentially Expressed in Multiple Forms

ELENA M. KLENOVA,[1,2] ROBERT H. NICOLAS,[2] HUGH F. PATERSON,[2] ALEXANDER F. CARNE,[2] CAROL M. HEATH,[2] GRAHAM H. GOODWIN,[2] PAUL E. NEIMAN,[3] AND VICTOR V. LOBANENKOV[3]*

Institute of Carcinogenesis, Russian Cancer Research Center, 115478 Moscow, Russia[1]; Chester Beatty Laboratories, Institute of Cancer Research, 237 Fulham Road, London SW3 6JB, United Kingdom[2]; and Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, Washington 98104[3]

Received 12 July 1993/Returned for modification 11 August 1993/Accepted 2 September 1993

A novel sequence-specific DNA-binding protein, CTCF, which interacts with the chicken c-*myc* gene promoter, has been identified and partially characterized (V. V. Lobanenkov, R. H. Nicolas, V. V. Adler, H. Paterson, E. M. Klenova, A. V. Polotskaja, and G. H. Goodwin, Oncogene 5:1743–1753, 1990). In order to test directly whether binding of CTCF to one specific DNA region of the c-*myc* promoter is important for chicken c-*myc* transcription, we have determined which nucleotides within this GC-rich region are responsible for recognition of overlapping sites by CTCF and Sp1-like proteins. Using missing-contact analysis of all four nucleotides in both DNA strands and homogeneous CTCF protein purified by sequence-specific chromatography, we have identified three sets of nucleotides which contact either CTCF or two Sp1-like proteins binding within the same DNA region. Specific mutations of 3 of 15 purines required for CTCF binding were designed to eliminate binding of CTCF without altering the binding of other proteins. Electrophoretic mobility shift assay of nuclear extracts showed that the mutant DNA sequence did not bind CTCF but did bind two Sp1-like proteins. When introduced into a 3.3-kbp-long 5′-flanking noncoding c-*myc* sequence fused to a reporter CAT gene, the same mutation of the CTCF binding site resulted in 10- and 3-fold reductions, respectively, of transcription in two different (erythroid and myeloid) stably transfected chicken cell lines. Isolation and analysis of the CTCF cDNA encoding an 82-kDa form of CTCF protein shows that DNA-binding domain of CTCF is composed of 11 Zn fingers: 10 are of C2H2 class, and 1 is of C2HC class. CTCF was found to be abundant and conserved in cells of vertebrate species. We detected six major nuclear forms of CTCF protein differentially expressed in different chicken cell lines and tissues. We conclude that isoforms of 11-Zn-finger factor CTCF which are present in chicken hematopoietic HD3 and BM2 cells can act as a positive regulator of the chicken c-*myc* gene transcription. Possible functions of other CTCF forms are discussed.

The c-*myc* proto-oncogene encodes a nuclear phosphoprotein with leucine zipper and helix-loop-helix structural motifs which appears to be important in the molecular biology of normal and abnormal cellular proliferation. Recent reviews of *myc* function and regulation (45, 63) summarized the literature supporting this conclusion. For example, *myc* is implicated in the control of both differentiation and replication (14), and recent reports link *myc* to apoptotic cell death (2, 19, 49). Myc and its dimerization partner Max form stable heterodimers through their helix-loop-helix and leucine zipper domains and bind specifically to a core E-box CACGTG DNA sequence (9). Max homodimers may serve as transcriptional repressors, whereas *myc-max* heterodimers can activate transcription (38). Therefore, at least some of the biological functions of *myc* may be mediated by transcriptional regulation of putative target genes.

Despite this progress in defining the mechanism of *myc* action on downstream events, less progress has been made in defining the proteins regulating the expression of c-*myc* itself. Both transcriptional and posttranscriptional mechanisms appear to play a role in regulation of c-*myc* gene expression (14, 45, 63). Maintenance of the level of the c-*myc* mRNA is achieved by regulation of both transcriptional initiation and elongation. Both initiation and elongation of the c-*myc* mRNA depend upon promoter elements which interact specifically with particular nuclear factors (62, 63). Data from several laboratories have provided a general map of mouse and human c-*myc* transcription elements, and a number of nuclear factors binding to them have been identified. In some cases, novel cDNAs encoding such factors have been isolated and sequenced, including ZF87 (also called MAZ), a proline-rich six-Zn-finger protein binding to ME1a1 and ME1a2 elements within the P2 promoter of the murine c-*myc* gene (11, 53); a 37-kDa protein, MBP-1, which appears to be a negative regulator of the human c-*myc* promoter (54); and nuclease-sensitive element protein 1, which binds to a region necessary for efficient P2 initiation (36). In addition, an Rb binding protein, E2F, which recognizes an E1A transactivation site in the human c-*myc* promoter (68) has also been cloned recently (24).

The chicken c-*myc* 5′-flanking region is at least 10-fold enriched in CpG pairs compared with those in total chicken DNA. Therefore, it is a member of the family of CpG-rich islands thought to be important for regulation of housekeeping genes (8). Overall high GC content (~80%) of the

* Corresponding author. Electronic mail address: vlobanen@fred.fhcrc.org.

5′-flanking region predicts that most of the potential regulatory DNA elements will be GC rich. Our previous analysis of DNA-protein interactions within the 5′-flanking region of the chicken c-*myc* gene revealed multiple GC-rich sequences which specifically interact with nuclear proteins (44). Proteins binding to one specific region within a hypersensitive site approximately 200 bp upstream of the start of transcription have been analyzed in detail (41, 43). Three nuclear factors were found to bind to several overlapping sequences within 180 to 230 bp upstream of the start of transcription. Two proteins resemble the transcription factor Sp1, and the other is a protein which binds to a GC-rich sequence containing three regularly spaced repeats of the core sequence CCCTC. This CCCTC-binding factor was termed CTCF (41, 43).

In this report we have characterized nucleotides within the −180 to −230 region which specifically contact CTCF, created mutants of these contact bases which no longer bind CTCF but continue to bind other factors recognizing overlapping sequences, and used these mutants to test the effects of the absence of CTCF binding on c-*myc* transcription. We have also isolated and sequenced a CTCF cDNA and analyzed the forms of CTCF RNA and protein present in chicken cells and their conservation and intracellular localization in other vertebrates. Our results show that CTCF can function as a positive regulator of c-*myc* transcription in chicken hematopoietic cells and suggest other possible roles for different forms of CTCF protein.

MATERIALS AND METHODS

Isolation of CTCF cDNA clones. The sequence-specific chromatography procedures were essentially as previously described (29) and were carried out as published earlier (43). Purified CTCF produced one polypeptide band of about 130 kDa on sodium dodecyl sulfate (SDS)-polyacrylamide gel electrophoresis (PAGE) and was used for the missing-contact-point analysis. To determine partial amino acid sequence of CTCF, purified protein was cleaved with cyanogen bromide, and the peptides were separated by SDS-PAGE and blotted onto polyvinyl difluoride membrane as described in detail elsewhere (50). The stained bands were sequenced, and three peptide sequences were obtained. Degenerate oligonucleotides were synthesized and used to probe Northern (RNA) blots containing gel-fractionated total RNA isolated from several chicken cell lines. One of these degenerate probes gave a band of sufficient length to encode 130-kDa protein, and this was used to screen a cDNA library prepared from the chicken myeloid cell line BM2 (48) and was kindly made available to us by U. Kruse and A. E. Sippel. One of four positive clones (p900) when sequenced contained a protein open reading frame (ORF) containing the three CTCF peptide sequences. This partial cDNA was used to isolate two other overlapping cDNAs (p910 and p911) which when sequenced and combined gave a total cDNA length of about 3.7 kbp. By screening a chicken genomic DNA library (a kind gift of A. Begue and V. Laudet) with the most 5′-distal fragment of the combined CTCF cDNA, a genomic clone was isolated and its shorter fragments were subcloned and sequenced. One such fragment overlapped the p910 cDNA, and its sequence has been joined onto the 5′ end of the cDNA sequence shown in Fig. 4. Cloned DNAs were sequenced as double-stranded plasmids by the dideoxynucleotide chain termination method, using Sequenase (United States Biochemical). Sequence was obtained from both strands.

Northern blot analysis. Total cellular RNA was isolated from different chicken cell lines and from tissues by guanidinium thiocyanate extraction (13). Taking precautions to avoid loss of resolution as previously described (69), 10 μg of RNA was separated on 1.5% agarose gels containing 6% formaldehyde, blotted to nylon membranes (Hybond N; Amersham), and hybridized to DNA probes labelled as described by the supplier of the random priming kit (Amersham). The final blot wash was in 0.1 × SSPE (1× SSPE is 0.18 M NaCl, 10 mM $NaPO_4$, and 1 mM EDTA [pH 7.7]) at 60°C.

Generation of antibodies, Western immunoblot, and indirect immunofluorescence microscopy. Rabbit antisera against EGEAVEAIVEES and HIAPNQADGGEV synthetic peptides with C-terminal cysteines (amino acids 2 to 13 and 39 to 50 of the CTCF sequence [see Fig. 4]) were raised as previously described (39), and corresponding polyclonal antibodies (operationally termed Ab1 and Ab2, respectively) were affinity purified by using Sulfo-SMCC (Pierce Chemical Co., Rockford, Ill.) according to the protocol described by the manufacturer. Specificity of Ab1 and Ab2 for CTCF was verified by blocking experiments with peptides used for immunizations. Preliminary characterization also showed that Ab1 recognizes CTCF proteins in cells of different species, whereas Ab2 is specific for chicken CTCF. Both Ab1 and Ab2 had an identical reactivity towards CTCF proteins in chicken cells and tissues. For Western immunoblot analysis, small-scale crude nuclear preparations from chicken tissues or from cell cultures were dissolved in lysis buffer containing 62.5 mM Tris-HCl (pH 6.8), 1 mM $MgCl_2$, and 2.3% SDS supplemented with the proteinase inhibitor cocktail composed of 86 μg of phenylmethylsulfonyl fluoride, 2 μg of aprotinin per ml, 40 μg of bestatin per ml, 1 μg of leupeptin per ml, 1.5 μg of pepstatin per ml, 78.5 μg of benzamidine per ml, and 1 μg of DNase I per μl (at this concentration, DNase I was found to degrade DNA in SDS-containing buffers) to reduce viscosity of samples before mixing with equal volume of 2× reducing SDS gel-loading buffer. Gel-fractionated proteins were transferred to Immobilon-P membranes (Millipore, Bedford, Mass.) by semidry blotting and probed with either Ab1 or Ab2 at a 1:100 dilution. CTCF protein bands were visualized by the enhanced chemiluminescence (ECL) procedure by using an ECL detection system (Amersham International plc) according to manufacturer's instructions. Absence of protein degradation was confirmed by reprobing immunoblots with other antibodies (anti-tubulin or pan-anti-*myc*). For indirect immunofluorescence staining, the protocol of Harlow and Lane (23) was adapted as described elsewhere (1).

In vitro transcription and translation. To obtain the pCITE/CTCF1 construct for efficient in vitro translation, the CTCF Zn finger domain was fused downstream of the cap-independent translation enhancer (CITE) of the encephalomyocarditis virus. The *Hin*dIII-*Xba*I fragment of the p900 plasmid encoding all 11 Zn fingers (amino acids 211 to 585 [see Fig. 4]) was ligated in frame to the *Nco*I-*Xba*I ends of the pCITE-1 vector (Novagen, Madison, Wis.) by using *Nco*I-*Hin*dIII adapter. The pCITE/CTCF1 plasmid was linearized at the unique *Xba*I site downstream of the CTCF, transcribed with phage T7 RNA polymerase (Boehringer Mannheim Co., Indianapolis, Ind.), and translated in a rabbit reticulocyte lysate system (Promega Co., Madison, Wis.) as described in the manufacturers' manuals. For electrophoretic mobility shift assay (EMSA) and methylation interference, 1 to 5 μl of the translation product was used.

Transient expression in COS-7 cells and dexamethasone-

inducible expression in stable NIH 3T3 clones. The *Not*I-*Xba*I fragment containing the full-length ORF of the CTCF form shown in Fig. 4 was subcloned in the pcDNA I Neo vector (Invitrogen Co., San Diego, Calif.) for high-level transient expression in COS-7 cells as previously described (59). For immunoblot analysis, nuclear and cytoplasmic fractions of transfected cells were prepared essentially as previously described (16). A second expression vector, pLK/SXneo, was made to obtain stable clones of NIH 3T3 cells conditionally expressing CTCF cDNA. In this construct, the *Sma*I-*Xba*I fragment containing the same ORF was subcloned into the pLK-neo vector (25) downstream of the new variant of a dexamethasone-inducible mouse mammary tumor virus long terminal repeat. The pLK-SXneo plasmid was transfected into NIH 3T3 and several G418-resistant clones, including the NIH 3T3(C14) clone used in this work, and examined for inducible expression of the CTCF cDNA by Western immunoblot analysis and immunofluorescence staining with Ab1 and Ab2.

Missing-contact-point analysis of CTCF-DNA interaction. To determine all DNA bases involved in DNA recognition by purified CTCF, a modified missing-contact analysis procedure (12) was carried out by using either *Acc*T-*Eco*RI or *Hin*dIII-*Eco*RI DNA fragments of the FpV oligonucleotide subcloned into the pUC12 labelled either (i) on its top (CT-rich) strand at the *Acc*I site by reverse transcriptase and [α-^{32}P]dCTP or (ii) on its bottom (G-rich) stand at the *Hin*dIII site by T4 polynucleotide kinase. Gel-purified ^{32}P-labelled DNA fragments were modified by the C+T or G+A reactions (46) as previously described (42), and a preparative-scale EMSA gel was used to separate base-modified [^{32}P]DNA probes bound and unbound to purified CTCF protein. Following identification of CTCF-bound bands, DNA from retarded and free bands was isolated and cleaved by piperidine, and equal radioactivity from each sample was resolved on a sequencing gel.

Agarose gel EMSA, DNase I footprinting, DNA probes, and competitors. The agarose gel EMSA technique has been described in detail (43). EMSA with polyacrylamide gels was carried out essentially as previously described (52). DNase I footprinting analysis of purified CTCF binding to the c-*myc* promoter was also carried out as described earlier (44). The chicken c-*myc* genomic DNA subclone pC*c-myc*-SA19 (44) was used to obtain an Nsi mutation (see Fig. 3A) within the CTCF-binding sequence by the polymerase chain reaction-mediated site-directed mutagenesis procedure (65) resulting in the CTCF site-mutated plasmid pC(Nsi)SA19. The DNA sequence of the mutant plasmid was verified by both restriction analysis and dideoxy DNA sequencing. These two plasmids were used to produce wild-type and Nsi mutant ^{32}P-labelled DNA probes for agarose gel EMSA experiments by isolating 152-bp *Hin*dIII-*Apa*I DNA fragments end labelled with [α-^{32}P-nucleoside triphosphates (NTPs) and Klenow polymerase (59). The same cold DNA fragments were also used as the wild-type DNA and the Nsi mutant DNA competitors.

Stable transfection of the *myc*-CAT constructs. A test reporter plasmid containing wild-type 5′-flanking noncoding sequence of the chicken c-*myc* gene fused to the coding sequence of the bacterial *cat* gene, pPst2CAT (see Fig. 3A), was constructed by ligating the 3.33-kbp *Bam*HI-*Hin*dIII DNA fragment of the pC*c-myc*Pst2 plasmid (44) into *Bam*HI-*Hin*dIII ends of the pKK232-8 vector (Pharmacia). The *Ecl*XI-*Apa*I fragment of the pPst2CAT was substituted for the *Ecl*XI-*Apa*I fragment of pC(Nsi)SA19, resulting in the *myc*-CAT construct pPst2NsiCAT (see Fig. 3A) containing CTCF site-mutated sequence identical to that of Nsi mutant DNA probe used for EMSA. Each of the two reporter plasmids was cotransfected together with the pCMV/β-gal plasmid and with the pRSV/Neo plasmid into chicken erythroleukemia HD3 cells (5) and myeloid leukemia BM2 cells (48, 66) by the lipofection method (20). The molar ratio of the chloramphenicol acetyltransferase (CAT)-expressing plasmid to the β-galactosidase (β-Gal) expressing plasmid and the Neo-expressing plasmid was about 10:1:1, respectively. A promoterless pKK232-8 plasmid and the pTK/CAT plasmid were also included in these experiments as test constructs. Cells containing stably integrated transfected plasmids were selected in culture medium supplemented with geneticin (G418; GIBCO/BRL) at 1 mg/ml and established as separate Neo-resistant polyclonal cultures by passaging for about 5 weeks in the presence of G418. CAT activity, normalized to the internal control β-Gal activity in cell extracts prepared from equal amount of these cells, was assayed as previously described (59) and quantitated by a direct ^{14}C-image analysis of the thin-layer chromatography plates.

Nucleotide sequence accession number. The DNA sequence shown in Fig. 4 has been submitted to the GenBank/EMBL data base under accession number Z22605.

RESULTS

Determination of nucleotides contacting CTCF. The DNA region protected by CTCF from cleavage generated by orthophenanthroline-copper (OP-Cu^{2+}) is unusually long (about 50 bp). It extends from −180 to −230 position relative to the start of transcription (41). Within this region, three other proteins were also shown to bind: two Sp1-like factors and a poly(dG)-binding protein (43). In order to design a mutation that would specifically knock out binding of CTCF without altering binding of other proteins, we had to distinguish nucleotides contacting CTCF from nucleotides contacting these other proteins. Although we have found previously that most of those nucleotides in the noncoding (G-rich) strand which give strong methylation interference with CTCF binding overlie one of the two Sp1 binding sequences, we hoped that a more detailed contact point probing of all four bases in both DNA strands might reveal more crucial CTCF-DNA contacts outside these two Sp1-binding sequences.

To facilitate contact point analysis of the CTCF-DNA interaction, we have purified homogeneous 130-kDa CTCF protein by sequence-specific chromatography as described earlier (43). Purified CTCF produced only the characteristic doublet of shifted bands during EMSA on both agarose and acrylamide gels and gave exactly the same DNase I footprint as it did in nuclear extracts (43, 44; data not shown). Therefore, we used purified CTCF protein in our contact point analysis to avoid contamination of CTCF-DNA complexes by other complexes containing either an Sp1-like or a poly(dG)-binding protein(s). This approach also improved recovery of partially depurinated or depyrimidated DNA specifically bound by CTCF (33).

The technique of missing-contact probing (12) has been modified to analyze all four bases contacting CTCF (43). DNA bases which on their removal or modification reduced binding by the affinity-purified CTCF protein resulted in bands of decreased intensity in lanes displaying protein-bound DNA (Fig. 1, lanes B) compared with the free DNA lane (Fig. 1, lanes F).

There were no adenines or guanines in the coding strand

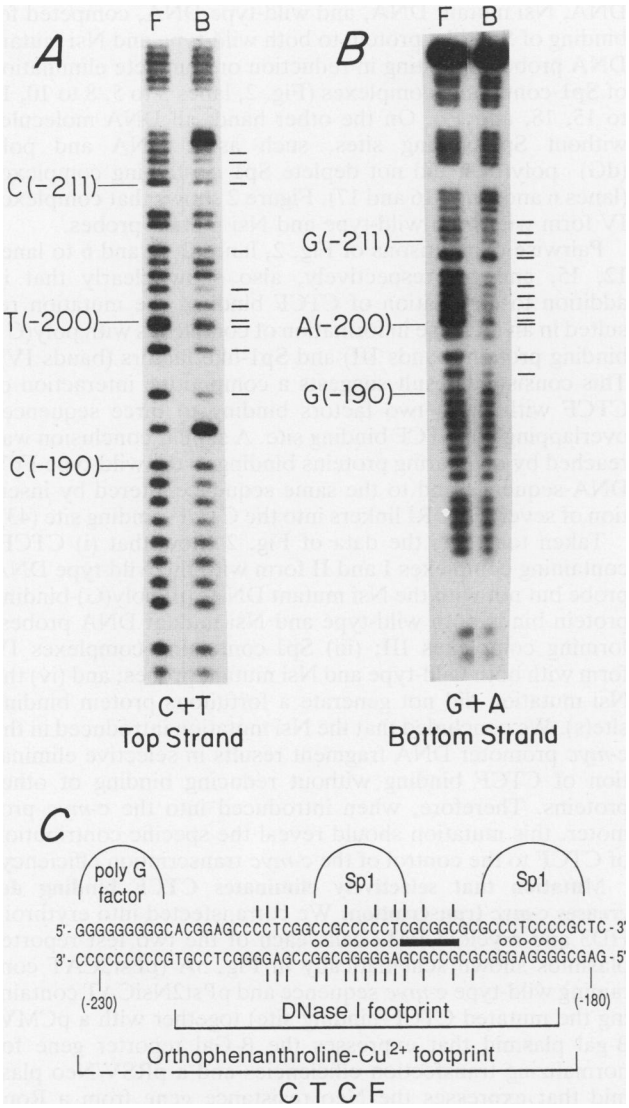

FIG. 1. Missing-contact-point analysis of purified CTCF binding to the footprint V region of the *c-myc* promoter. The DNA probe was 3′ end labelled at the top (coding) strand and modified by the C+T-reaction (A) or 5′-end labelled at the bottom strand and modified by the G+A-reaction (B). Lanes F, modified bases of the free DNA probe; lanes B, modified bases of DNA molecules bound by CTCF. Small bars at the right of panels (A and B) and above and below the nucleotide sequence (C) indicate DNA bases involved in recognition by purified CTCF. (C) To display the relative disposition of bases contacting different proteins binding to the same DNA region, data shown in panels A and B are combined with data of missing-point analysis obtained for two Sp1-like proteins (shown by small circles), the CTCF-binding DNA region determined by DNase I or *O*-phenanthroline-copper footprinting is also shown (41, 43). Six nucleotides substituted to create a new *Nsi*I site (see explanations in the text) are shown by the black bar.

and no cytidines or thymidines in the noncoding strand involved in recognition of the DNA sequence by CTCF. Therefore, Fig. 1 presents data on the missing-contacts analysis only for cytidines and thymidines in the top DNA strand and for guanines and adenines in the bottom strand of the CTCF-binding DNA region. The results display 10 pyrimidines in the top strand and 15 purines in the bottom strand involved in specific CTCF-DNA interaction (Fig. 1C). Although most of the important CTCF-contacting nucleotides required for tight binding correspond to two of the three direct CCCTC repeats implicated earlier in DNA sequence recognition by CTCF (43), additional nucleotides, especially CG pairs of the CGCGGCGCG sequence with four CpG dinucleotides between the second and the third CCCTC repeat (i.e., between two Sp1-binding sites [Fig. 1]), also appear to be necessary.

Having identified all of the nucleotides in the CTCF-binding site which contact CTCF, and taking into account contact nucleotides for two Sp1 proteins defined previously, we mapped the contacts for these proteins as shown schematically in Fig. 1C and designed specific mutations to eliminate binding of CTCF without altering the binding of other proteins to the same DNA segment. Six nucleotides, shown in Fig. 1C by a black bar within the CTCF-binding site, were substituted to create a new *Nsi*I restriction site. This Nsi mutant DNA sequence has three CTCF-contacting CG base pairs substituted for AT pairs or for the inverse GC pair (as shown in Fig. 3A) but does not alter any of Sp1-contacting nucleotides (Fig. 1C). Therefore, it was expected to bind the same proteins as the wild type, except for CTCF.

Nsi mutation selectively eliminates binding of CTCF. In order to compare proteins binding to the wild-type CTCF-binding DNA region and to the Nsi mutant DNA, two 152-bp *Hin*dIII-*Apa*I DNA fragments [derived from the pC*c-myc*-SA19 and the pC(Nsi)SA19 plasmids, respectively] were end labelled and used as DNA probes for modified EMSA in agarose gels with nuclear extracts from HD3 cells. As shown earlier (41), agarose gel EMSA allows resolution of multiple DNA-protein complexes formed with the wild-type DNA sequence, including high-molecular-weight complexes composed of several proteins which do not penetrate polyacrylamide gels. Characteristic gel-retarded bands formed with the wild-type DNA fragment were designated according to their characteristic electrophoretic mobilities as previously described (43): complexes I and II contain CTCF, complex III contains poly(dG)-binding protein, and complex IV contains Sp1-like proteins.

To compare binding of these factors to wild-type and Nsi mutant DNA fragments and to rule out the possibility that mutation of the CTCF-binding site may fortuitously result in generation of a new protein binding sequence(s), we redefined all DNA-protein complexes detectable by agarose gel EMSA by using both wild-type and Nsi mutant ^{32}P-labelled DNA probes. Figure 2 shows that agarose gel-retarded bands I and II, III, and IV are formed in precise agreement with the expected binding and competition properties characteristic for CTCF, Sp1-like factors, and poly(dG)-binding protein (as summarized below) and that no new protein binds to the mutant DNA probe.

(i) CTCF-containing complexes I and II. We found previously (41, 43) that tight CTCF binding requires additional DNA sequence flanking its recognition site: it does not bind to the 44-bp-long oligonucleotide which contains all nucleotides contacting CTCF (oligonucleotide FpV monomer) but does bind to the same sequence when it is made longer by extension with some flanking DNA (e.g., by polylinker DNA in the subcloned FpV DNA fragment) or when it is ligated to itself to form FpV-oligonucleotide multimers. CTCF binding is also resistant to competition with Sp1-binding oligonucleotide, λ DNA and double-stranded poly(dG) · poly(dC). These binding properties allowed us to distinguish CTCF-containing complexes I and II from closely migrating com-

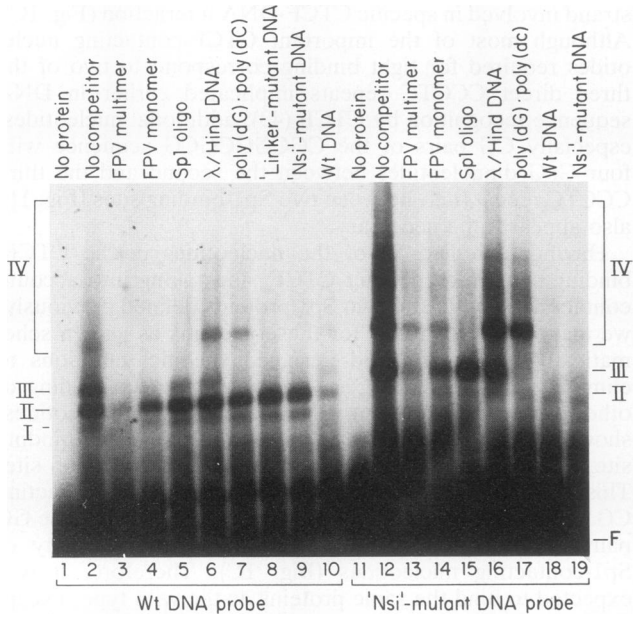

FIG. 2. Comparative agarose gel EMSA analysis of proteins binding to the wild-type and to the Nsi mutant DNA sequence of the c-*myc* promoter. Lane 1, DNA probes alone (no protein added) which migrate as band F. Binding reaction mixtures contained 1 μl of HD3 nuclear extract. Competitor DNAs are indicated at the top of each lane and were added in 200-fold molar amounts (for Sp1-binding oligonucleotide, double-stranded poly(dG) · poly(dC) homopolymer, and FpV monomer oligonucleotide) or in a 100-fold weight excess (for DNA fragments and multimerized FpV oligonucleotide) over the ^{32}P-labelled DNA probes. Specific DNA-protein complexes are designated I to IV as described earlier (43) and explained in the text.

plex III. Comparison of bands I and II in lanes 2 to 7 and in lanes 13 to 17 shows that CTCF-containing complexes I and II form with the wild-type DNA probe but do not form with the Nsi mutant DNA. In accord with this conclusion, Nsi mutant DNA used as a cold competitor in EMSA experiment efficiently competes for binding of poly(G)-binding protein and Sp1-like protein to both wild-type DNA probe and Nsi mutant probe (Fig. 2, compare lanes 9 and 19) but does not compete for binding of CTCF to the wild-type DNA probe, leaving bands I and II intact (Fig. 2, lane 9). In addition, linker mutant DNA [which has *Eco*RI linkers inserted into the CTCF-binding site and, therefore, does not bind CTCF but can bind Sp1 and poly(G)-binding protein (43)] does not compete for CTCF but depletes binding of other proteins (Fig. 2, lane 8).

(ii) Poly(G)-binding protein-containing complex III. Poly(dG)-binding protein does not efficiently compete for Sp1-binding oligonucleotides (Fig. 2, lanes 5 and 15) or for λ DNA (lanes 6 and 16), but it could be inhibited by the poly(dG) · poly(dC) homopolymer (lanes 7 and 17). It also competes with the wild-type DNA, linker mutant DNA, or Nsi mutant DNA (lanes 8, 9, 10, 18, and 19) because all of these DNA fragments contain the same sequence of nine consecutive deoxyguanosine residues. Figure 2 shows that complex III forms with both wild-type and Nsi mutant probes.

(iii) Complex IV containing Sp1-like proteins. All DNA molecules containing Sp1-binding sites, such as FpV mono- or multimers, Sp1-binding oligonucleotide, linker mutant DNA, Nsi mutant DNA, and wild-type DNA, competed for binding of Sp1-like protein to both wild-type and Nsi mutant DNA probes, resulting in reduction or complete elimination of Sp1-containing complexes (Fig. 2, lanes 3 to 5, 8 to 10, 13 to 15, 18, and 19). On the other hand, all DNA molecules without Sp1-binding sites, such as λ DNA and poly (dG) · poly(dC), did not deplete Sp1 containing complexes (lanes 6 and 7 and 16 and 17). Figure 2 shows that complexes IV form with both wild-type and Nsi mutant probes.

Pairwise comparisons of Fig. 2, lanes 2, 5, and 6 to lanes 12, 15, and 16, respectively, also show clearly that in addition to elimination of CTCF binding, the mutation resulted in an increase in formation of complexes with poly(G)-binding protein (bands III) and Sp1-like factors (bands IV). This consistent result suggests a competitive interaction of CTCF with these two factors binding to three sequences overlapping the CTCF binding site. A similar conclusion was reached by comparing proteins binding to the wild-type FpV DNA sequence and to the same sequence altered by insertion of several *Eco*RI linkers into the CTCF-binding site (43).

Taken together, the data of Fig. 2 show that (i) CTCF-containing complexes I and II form with the wild-type DNA probe but not with the Nsi mutant DNA; (ii) poly(G)-binding protein binds both wild-type and Nsi mutant DNA probes, forming complexes III; (iii) Sp1-containing complexes IV form with both wild-type and Nsi mutant probes; and (iv) the Nsi mutation did not generate a fortuitous protein binding site(s). We concluded that the Nsi mutation introduced in the c-*myc* promoter DNA fragment results in selective elimination of CTCF binding without reducing binding of other proteins. Therefore, when introduced into the c-*myc* promoter, this mutation should reveal the specific contribution of CTCF to the control of the c-*myc* transcription efficiency.

Mutation that selectively eliminates CTCF binding decreases c-*myc* transcription. We cotransfected into erythroid HD3 and myeloid BM2 cells each of the two test reporter plasmids shown schematically in Fig. 3A (pPst2CAT containing wild-type c-*myc* sequence and pPst2NsiCAT containing the mutated CTCF-binding site) together with a pCMV/β-gal plasmid that expresses the β-Gal reporter gene for normalizing transfection efficiencies and a pRSV/Neo plasmid that expresses the Neo resistance gene from a Rous sarcoma virus long terminal repeat. The thymidine kinase gene promoter fused to the CAT gene (pTK/CAT plasmid) was also used as a test plasmid in these experiments in order to monitor relative strength of the wild-type c-*myc* promoter in cells of different lineage. For each test plasmid, all G418-resistant clones were pooled, established as cell cultures growing in the presence of G418, and assayed for CAT activity by using β-Gal expression as an internal control. Figure 3B demonstrates that the Nsi mutation of CTCF binding site resulted in a 3- and 10-fold reduction of transcription directed by the c-*myc* promoter in two different stably transfected cell lines. These results show that in hematopoietic cells, the CTCF-binding sequence can act as a positive element of the chicken c-*myc* gene promoter and also suggest that the extent of stimulation of the c-*myc* transcription by CTCF could be cell type specific.

Isolation and characterization of cDNA encoding the 82-kDa form of CTCF protein. We screened a cDNA library derived from poly(A)-containing RNAs of the chicken myeloid BM2 cell line with a degenerate oligonucleotide corresponding to one of three CTCF peptide sequences, MEGEAVEAIVEE. This was selected as a probe because it detected on Northern blots an RNA that might be long enough to encode 130-kDa CTCF protein. By screening about 7×10^5 phage from a

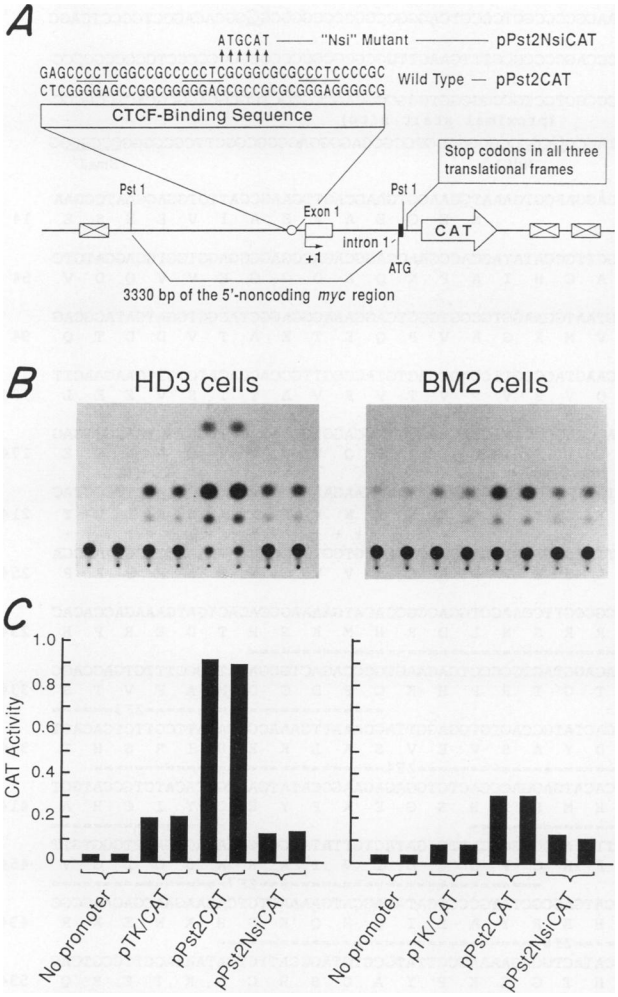

FIG. 3. (A) Schematic description of control (pPst2CAT) and mutant (pPst2NsiCAT) *myc*-CAT fusion plasmid constructs used in stable transfection experiments. Crossed boxes show transcriptional terminators. Although in these test indicator constructs the first ATG of the c-*myc* coding sequence is just upstream of the CAT coding sequence, translational read-through from this ATG was prohibited by three stop codons shown in the figure. (B) Autoradiograms of CAT assay analysis of transcription of test plasmids in stably transfected erythroid HD3 and myeloid BM2 cells. (C) Direct quantitation of data shown in panel B by ^{14}C-image analysis. Note that black columns in are aligned in a row with the thin-layer chromatography TLC dots shown above in panel B.

library representing about 1.4×10^6 recombinants (57), a positive clone, p900, was isolated. Translated sequence of the p900 clone gave an ORF containing all three CTCF peptide sequences obtained from purified 130-kDa protein. Two other overlapping cDNAs, p910 and p911, were isolated by screening the same library with the p900 DNA probe. The combined sequence of these three cDNAs gave a total cDNA length of 3,779 bp. Figure 4 shows that the 3′ end of the cDNA has a poly(A) tract and a consensus polyadenylation signal. To map the 5′ end of the gene, a genomic clone was isolated, and its fragments were subcloned and sequenced. Preliminary Southern blot analysis of chicken genomic DNA indicates that CTCF is a single-copy gene locus (data not shown). One genomic fragment overlapped the p910 cDNA, and its sequence has been joined onto the 5′ end of the cDNA sequence in Fig. 4. Primer extensions were carried out to define the putative start of transcription by using several primers encoded at the 5′ end of the longest cDNA clone. Using poly(A)$^+$ RNA from BM2 or HD3 cells, two unequally employed start sites were mapped: a minor distal site at designated base +1 in Fig. 4 and a major proximal site at base 332 (data not shown). Genomic sequence aligned with the ORF as shown in Fig. 4 appears to represent authentic 5′-flanking region of the gene, including the 5′ end(s) of CTCF mRNAs because (i) transcription start sites were mapped by extending different primers from inside of the cDNA sequence, (ii) none of the several independently obtained cDNAs and none of the products of the 5′-RACE procedure had longer 5′ ends, and (iii) when used to probe Northern blots, genomic DNA clones representing more than 12 kbp of sequence upstream of the distal start site did not detect CTCF RNA (data not shown). However, the formal possibility of an additional 5′ exon more than 12 kbp upstream is not ruled out. Although there was no obvious TATA box upstream of either of the start sites, multiple Sp1-binding GC-rich consensus sequences (about 40, therefore not shown in Fig. 4) and a pyrimidine-rich initiator element (Inr) with a perfect match to the YY1(UCRBP)-binding site (21, 61), the common features of many TATA-less promoters of housekeeping genes (reviewed in reference 71), are present. In addition, a typical E-box sequence, located between distal and proximal start sites (Fig. 4), suggests that cooperative interaction of an initiator-binding factor TFII-I and a helix-loop-helix activator (56a) could be involved in regulation of the CTCF promoter.

Northern blot analysis of RNA isolated from several chicken cell lines and from different adult chicken tissues by using the p900 plus p911 cDNA probes (Fig. 5, and data not shown) revealed four major CTCF mRNA species. These included closely migrating species of about 3.7 and 4.0 kb and a second doublet of about 6.5 and 7.0 kb. These RNA doublets were often difficult to resolve on agarose gels. Figure 5 shows that doublets of 4- and 7-kb CTCF RNAs are detected in erythroid leukemia HD3 cells (5), in the DT40 B-cell leukemia line (31), in acutely v-*myc*-transformed chicken embryo fibroblasts (CEF/mc29), in a v-*myc*-transformed macrophage-like HD11 cell line (6), and in myeloid leukemia BM2 cells (48). The ratio between smaller and larger bands of the doublets varies in different cell type and was regulated by phorbol myristate acetate and other stimuli (Fig. 5, compare lanes 1 and 2) (34).

The ORF consisting of 728 amino acids is shown in Fig. 4. The first ATG codon (at position 460) downstream of the transcription start site is flanked by a sequence favorable for efficient translation (37) and is preceded by two stop codons in the same reading frame. No intron-exon junctions were discernible in the sequences between this methionine and these upstream stop codons. All three peptide sequences obtained from the purified 130-kDa CTCF protein were found at amino acid positions 1 to 19, 229 to 248, and 424 to 438. Examination of the sequence reveals 10 C2H2-type and 1 C2HC-type Zn finger motifs (for a review, see references 4, 18, and 35). On the basis of the structure of the 11-Zn-finger domain, CTCF could not be classified as *GLI*-Krüppel-like because not all Zn fingers conform to the C2H2 type and not all of them are consecutively connected by a 6-amino-acid conserved, reiterated H/C link, a structural feature defining the *GLI*-Krüppel-like family (58). Neither the nucleotide nor the predicted amino acid sequences outside of the Zn finger domain showed any significant homology to genes or pro-

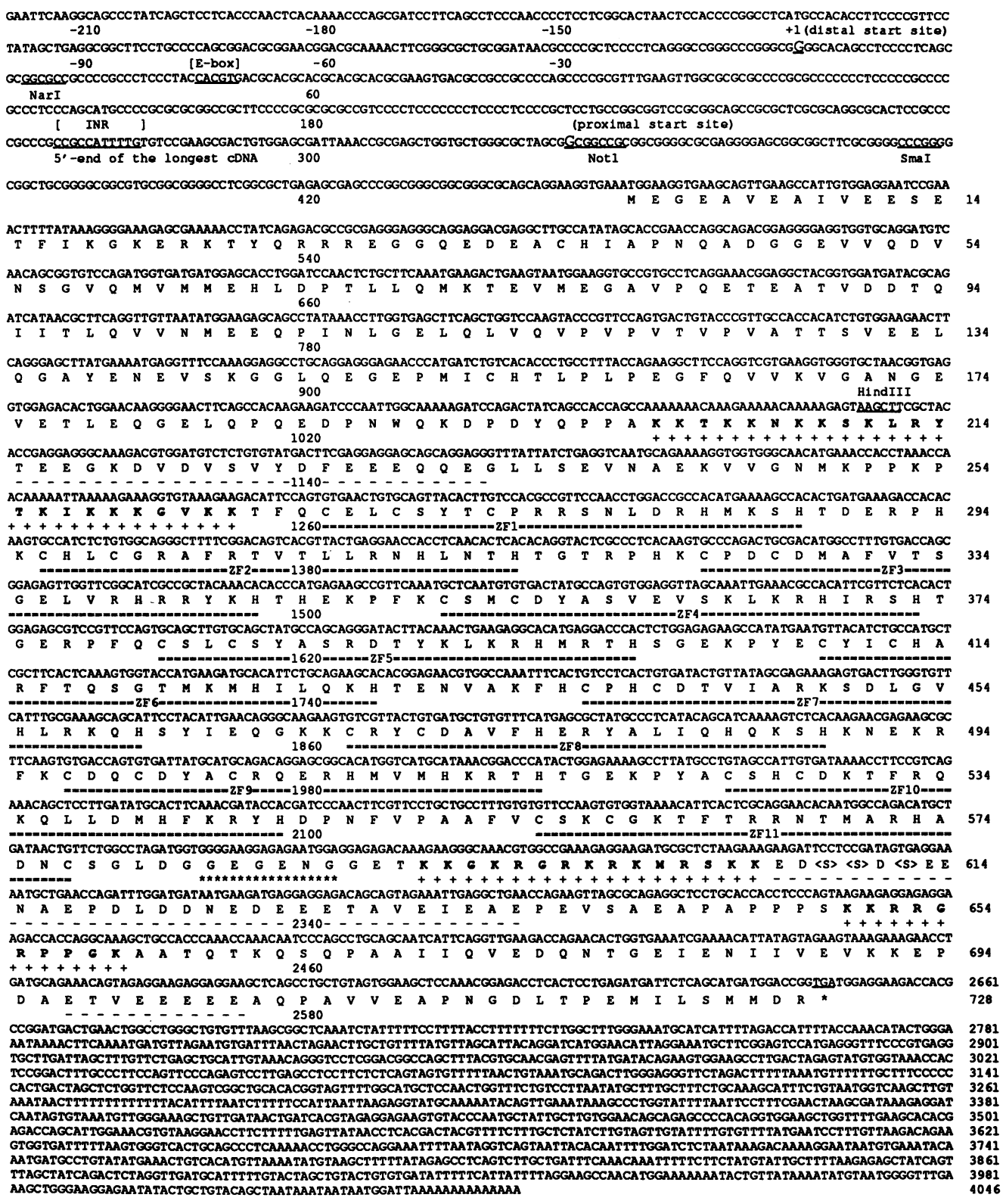

FIG. 4. Combined nucleotide and predicted amino acid sequences of the 82-kDa form of CTCF cDNA joined at the 5′ end to the genomic DNA sequence. Nucleotide sequence is numbered under each line (designated +1 nucleotide is at the distal transcription start site); beginning from nucleotide 2661, numbering goes at the right side. E-box sequence and initiator (Inr) elements are underlined. Amino acid sequence is numbered at the right side of each line. Zn finger motifs are underlined by a double broken line and are numbered from ZF1 to ZF11. Highly acidic and basic domains are underlined by minus and plus signs, respectively. Putative nucleotide-binding site is indicated by asterisks. Ser residues within a potential NLS-associated phosphorylation sites are indicated in brackets. Amino acid sequences of all three sequenced peptides derived from purified 130-kDa CTCF were found throughout the ORF at positions 1 to 19, 229 to 248, and 424 to 438. Amino acid sequences of two synthetic peptides used to generate polyclonal antibodies Ab1 and Ab2 are at positions 2 to 13 and 39 to 50, respectively.

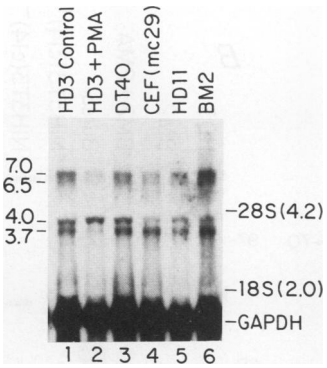

FIG. 5. Northern blot analysis of CTCF expression. Total RNA isolated from chicken cells listed at the top of the blot was resolved in a 1.5% agarose gel. Hybridization was carried out simultaneously with the CTCF cDNA probe and with the chicken glyceraldehyde-3-phosphate dehydrogenase (GAPDH) cDNA probe. Two probes were used together in this experiment, because we determined earlier that there was no CTCF RNA comigrating with the GAPDH RNA. Positions (and size in kilobases) of CTCF RNAs and rRNAs are shown at the left and at the right, respectively.

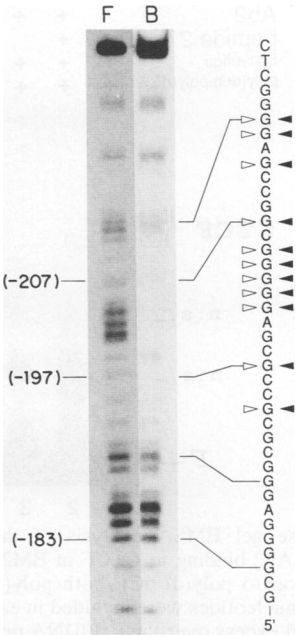

FIG. 6. Methylation interference assay of in vitro-translated 11-Zn-finger domain binding to the footprint V DNA sequence. The DNA probe was 5′ end labelled at the noncoding strand and methylated by dimethyl sulfate. Lane F, modified guanines of the free DNA probe; lane B, methylated G's of DNA molecules bound by the 11-Zn-finger polypeptide synthesized in vitro. Open arrowheads at the left side and closed arrowheads at the right side of the DNA sequence indicate G residues recognized by in vitro-translated Zn fingers and by endogenous CTCF (Fig. 4) (43), respectively. Difficult reading of the sequencing gel bands is in part due to a very uneven modification efficiency of G residues around −200 position of the CTCF-binding sequence. Numbering of nucleotides is as in Fig. 1.

teins in available data bases. The 11-Zn-finger domain is flanked on each side by two positively charged regions with a K/R-rich amino acid sequence (indicated in Fig. 4). A glycine-rich motif, GXGXXG, followed by a conserved lysine, which has been identified as a common nucleotide binding fold in many ATP- and GTP-binding proteins (60, 70), is also present in CTCF immediately C terminal to the Zn finger domain. The presence of this putative nucleotide-binding domain in CTCF could be of functional importance because we have found that ATP and other ribonucleotide triphosphates stimulate binding of CTCF to the FpV DNA sequence and have exploited this property to facilitate purification of the CTCF protein (41, 43). This domain is followed by the most positively charged site in the CTCF polypeptide with a K/R-rich amino acid sequence characteristic of a nuclear localization signal (NLS) (reviewed in reference 17). The region of CTCF adjacent to its eleventh C2HC-type Zn finger thus harbors consensus sequences for a putative ATP-binding domain, NLS, and also includes potential target sites for phosphorylation by casein kinase II and cyclic AMP-dependent protein kinase (Fig. 4, potential target Ser residues of these sites are shown in brackets) (51). The overall arrangement (including spacing) of the eleventh C2HC-type C-terminal CTCF Zn finger situated immediately next to the NLS followed by potential phosphorylation sites (Fig. 4) is remarkably similar to the arrangement of the C-terminal third C2HC-type Zn finger of SWI5 which is also next to the NLS and Cdc2 phosphorylation sites (47). Nuclear translocation of several transcription factors including SWI5 is regulated by phosphorylation of the NLS-adjacent sites (for a review, see reference (27). The similarity of arrangement of these domains between SWI5 regulatory region and CTCF raises a possibility that nuclear import of CTCF is regulated by phosphorylation. The significance of the presence of a putative ATP-binding domain within the same region is not clear. There are also three highly acidic domains (underlined by minus signs in Fig. 4) composed of residues favorable for α-helix formation. A helical wheel representation of these three regions reveals a common pattern of negative residues arranged in a single stripe on one side of a putative helix (data not shown), a feature of some strong acidic transcriptional activators (22).

Since guanines involved in recognition of the DNA-binding domain synthesized in vitro (Fig. 6) are exactly the same as those previously shown to be involved in binding of endogenous 130-kDa CTCF by both methylation interference assay (see Fig. 4 in reference 43) and by missing-contact analysis (Fig. 1), we concluded that the in vitro-translated 11-Zn-finger domain binds to the c-*myc* promoter DNA sequence in a manner indistinguishable from that of endogenous 130-kDa CTCF protein. As expected for a smaller protein, the in vitro-synthesized 11-Zn-finger domain binding to the FpV DNA fragment produces DNA-protein complexes which migrate during EMSA faster than do complexes with endogenous CTCF. Nevertheless, these complexes demonstrate competition properties listed above as diagnostic for CTCF, including the requirement for additional DNA sequence-flanking contact nucleotides (data not shown).

In order to obtain additional reassurance that cloned sequence is directly related to 130-kDa CTCF, we tested whether the Ab2 polyclonal sera against N-terminal peptide (amino acids 39 to 50) derived from the cDNA sequence shown in Fig. 4 would specifically cross-react with the endogenous CTCF bound to DNA in nuclear extracts. Data from Fig. 7 show that the Ab2 anti-peptide antibody does ablate specifically the formation of the characteristic complexes I and II containing 130-kDa endogenous CTCF.

Thus, the cDNA described has the expected properties of

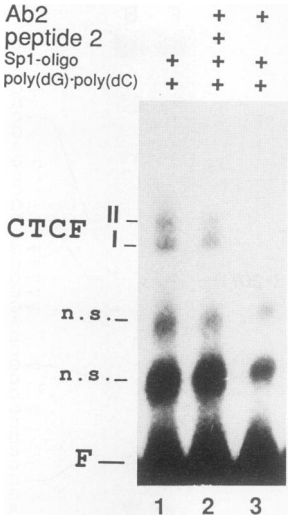

FIG. 7. Agarose gel EMSA analysis of the polyclonal anti-peptide antibody Ab2 binding to CTCF in BM2 cell crude nuclear extract. In addition to poly(dI-dC), both poly(dG) · poly(dC) and Sp1-binding oligonucleotides were included in each binding mixture in about a 100-fold excess over the [^{32}P]DNA probe. Antibody Ab2 blocked by preincubation with the peptide 2 (lane 2) or Ab2 alone (lane 3) was added to the binding reaction mixture together with 2 μl of the nuclear extract 30 min prior to mixing with the DNA-probe and loaded onto a gel after an additional 15 min of incubation. Control binding reaction (lane 1) was assembled in the same order but with the antibody excluded. Free DNA probe (F), nonspecific complexes (n.s.), and CTCF-DNA complexes (I and II) are indicated.

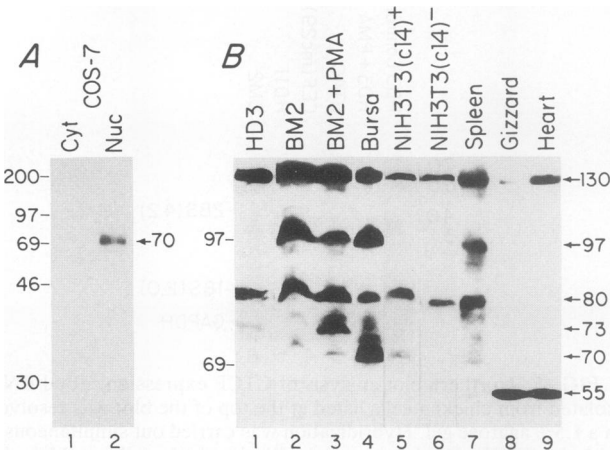

FIG. 8. Western immunoblot analyses of recombinant CTCF expressed in COS-7 cells (A) or in stably transfected NIH 3T3(C14) cells before (B, lane 6) and after (B, lane 5) dexamethasone induction; (B) endogenous CTCF in nuclear preparations from different chicken cells and tissues. Nuc, nuclear fraction; Cyt, cytoplasmic fraction. Either Ab2-antibodies specific for chicken CTCF proteins (A) or pan-specific Ab1 antibodies (B) were used. Arrows at the right of panel indicate CTCF proteins and their apparent molecular masses; positions of the rainbow markers are shown at the left of both panels.

CTCF as judged by the presence of the peptide sequences, by DNA-binding characteristics, and by specific binding of the anti-peptide antibodies to CTCF in the crude nuclear extracts. However, the ORF predicts a protein of 82 kDa. CTCF protein purified from chicken erythrocytes had an apparent molecular mass of 130 kDa. When the cDNA depicted in Fig. 4 was transiently expressed in COS-7 cells (Fig. 8A) or conditionally expressed in stably transfected NIH 3T3 cells (Fig. 8B, lanes 5 and 6), it produced on SDS gels protein migrating as a single band of 70 kDa. In Fig. 8A, the Ab2 polyclonal antibody specific for chicken CTCF proteins was used. In Fig. 8B, the Ab1 polyclonal antibody specific for CTCF proteins from different species was used. In NIH 3T3 (C14) cells, it detects, in addition to the inducible chicken 70-kDa recombinant protein (which is specifically induced at relatively low level in cells treated by dexamethasone (Fig. 8B, lane 5) two endogenous mouse CTCF forms of 130 and 80 kDa (lanes 5 and 6, respectively). If antibody Ab2 (which does not recognize mouse proteins) was used to immunostatin NIH 3T3 nuclear proteins, only the 70-kDa chicken recombinant CTCF protein was detected in dexamethasone-induced cells (Fig. 9; data not shown). Thus, the cDNA depicted in Fig. 4 encodes a CTCF protein with predicted 82-kDa and apparent 70-kDa molecular masses. Therefore, the difference in molecular weight between CTCF purified from erythrocytes and the product of the cDNA shown in Fig. 4 is unlikely to be due to anomalous migration in SDS gels. It was therefore reasonable to suggest that multiple mature CTCF mRNA species detected on Northern blots (Fig. 5) could give rise to several cDNAs, including one encoding a protein with electrophoretic mobility of 70 kDa.

Multiple differentially expressed forms of CTCF protein. Data from Western immunoblot analysis (Fig. 8B) demonstrated that the CTCF gene encodes several proteins of different size which all share an epitope recognized by anti-CTCF polyclonal antibodies. In highly proliferating HD3 cells, two major forms were detected, namely, two most abundant 130- and 80-kDa forms and one minor 73-kDa form (Fig. 8B, lane 1). The same 130- and 80-kDa forms were the major CTCF proteins detected in mouse NIH 3T3 (C14) cells (Fig. 8B, lanes 5 and 6). In dividing BM2 cells, three major forms (130, 97, and 80 kDa) and two minor forms (73 and 70 kDa) are present (Fig. 8B, lane 2). In BM2 cells induced by phorbol myristate acetate to terminally differentiate into macrophages (66), the 97-kDa form is down-regulated and the 73-kDa form is highly up-regulated (Fig. 8B, lane 3). Since no 7-kb RNA and two additional CTCF RNA isoforms (1.5 and 2.2 kb) were detected in phorbol myristate acetate-induced BM2-derived macrophages (42), we believe that alteration in relative proportion of different CTCF protein forms induced upon terminal differentiation of BM2 cells is due to an alteration in CTCF mRNA processing rather than to posttranslational processing. In addition to 80-, 97-, and 130-kDa forms, a 70-kDa protein accounts for almost one quarter of all CTCF proteins detected in lymphocytes of mature bursal follicles (Fig. 8B, lane 4). This 70-kDa form comigrates with the product of the chicken CTCF cDNA conditionally expressed in stably transfected NIH 3T3 (C14) cells (Fig. 8B, lane 5) or transiently expressed in COS-7 cells (Fig. 8A). Therefore, the cDNA clone shown in Fig. 4 is likely to represent a minor mature mRNA encoding in BM2 cells a minor form of CTCF protein with an apparent molecular mass of 70 kDa.

None of the 97-, 80-, 73-, and 70-kDa forms are expressed in muscle tissue of gizzard and heart which, nevertheless, produced the ubiquitous 130-kDa form and a tissue-specific 55-kDa form of CTCF (Fig. 8, lanes 8 and 9, respectively). While not formally ruled out, we believe that the smaller

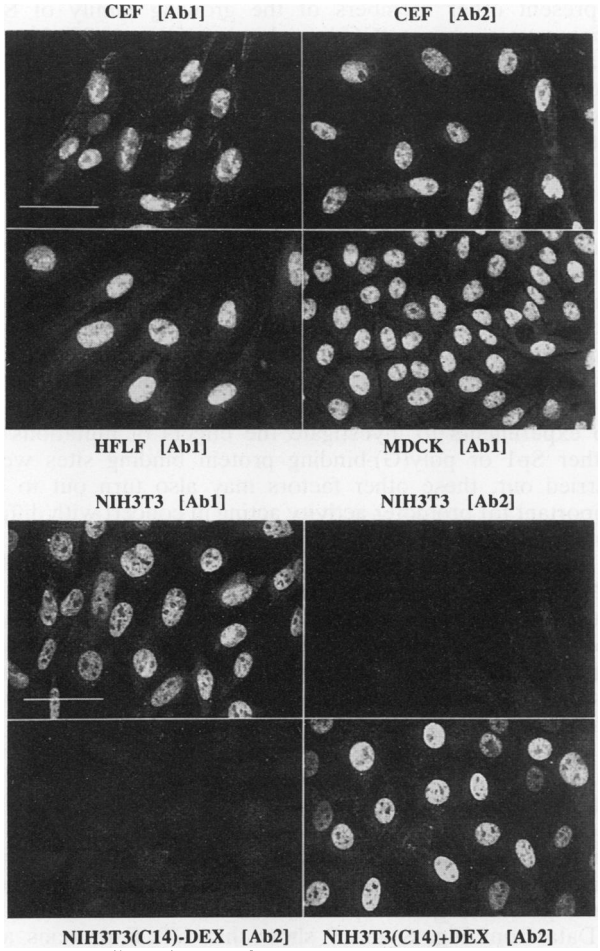

FIG. 9. Indirect immunofluorescence staining of CTCF in cells of different species. Type of cells and of primary antibody used are indicated above or below panels. CEF, chicken primary fibroblasts; HFLF, human fetal lung fibroblasts; MDCK, canine kidney cell line; NIH 3T3, mouse embryo cell line; NIH 3T3(C14), clone 14 of NIH 3T3 cells stably transfected by pLK/SXneo. Bars, 50 μm. For further explanation, see text.

forms of CTCF protein are unlikely to be due to in vitro proteolysis because (i) no degradation was detected by reprobing the same blots with other antibodies; (ii) both presence and relative proportion of individual forms in different cells or tissue samples were reproducible, and (iii) preincubation of samples at 37°C in the absence of proteinase inhibitors resulted in band smearing rather than in generation of additional discrete bands (data not shown).

CTCF is abundant and conserved in vertebrate species. Northern blots containing total RNA from human and mouse hematopoietic cell lines probed at moderate stringency with several representative chicken CTCF cDNA probes also revealed expression of the CTCF homolog(s) in these species (15). In addition, competition EMSA analyses of the CTCF DNA-binding activity in nuclear extracts prepared from cells of a number of species, from frog to human, have also shown conserved nature of the CTCF DNA-binding activity (data not shown). Finally, Fig. 9 shows that CTCF proteins are detected in nuclei of chicken embryo fibroblasts by both Ab1 and Ab2 polyclonal anti-CTCF antibodies. Figure 9 also demonstrates that, although Ab2 detects only chicken CTCF proteins [either in chicken embryo fibroblasts or in the mouse NIH 3T3(C14) cells conditionally expressing the chicken CTCF cDNA (also see data from Fig. 8)], Ab1 detects CTCF proteins in mouse NIH 3T3 cells, in canine kidney MDCK cells, and in human fetal lung fibroblasts. Thus, CTCF is a conserved nuclear factor highly expressed in a variety of vertebrate cells.

DISCUSSION

Our objectives in the experiments reported here were to determine whether there is a functional role for CTCF binding to one particular site within the chicken c-*myc* promoter and to provide an initial characterization of the CTCF protein by cloning and sequencing a cDNA for CTCF. By using affinity-purified CTCF for missing-contact analysis (Fig. 1), 25 nucleotides in both DNA strands were found to be required for DNA sequence recognition by CTCF. The experiments described in Fig. 1 also showed that, in addition to three CCCTC repeats found within the region protected by CTCF from DNase I or OP-Cu^{2+} cleavage, a sequence of reiterated cytosines and guanines between the second and third CCCTC repeat is required for efficient recognition by CTCF of the entire binding region. This result may explain why several DNA sequences composed of simple repetitions of 10- to 12-bp spaced CCCTC motifs were previously found to be inefficient binding sites for CTCF (67). Results of agarose gel EMSA (Fig. 2) showed that when three GC pairs between two Sp1 binding sites were altered by site-specific mutagenesis, the mutated sequence no longer binds CTCF but still bound Sp1-like factors and poly(G)-protein.

Assuming that the difference between proteins binding in vitro to the wild type and to the mutated DNA sequences is reproduced in cells stably transfected with the corresponding c-*myc*–CAT reporter constructs, mutations which selectively knock out CTCF binding in vitro should reveal the specific role of CTCF in *trans*-activation of the c-*myc* transcription in vivo. Data shown in Fig. 3 demonstrated that the mutation of the CTCF binding site results in 3- and 10-fold reduction of transcription directed by the 3.3 kbp of the 5′-noncoding c-*myc* sequence in stably transfected chicken myeloid and erythroid cell lines. This analysis did not distinguish between the contribution of CTCF to transcriptional regulation on the level of initiation versus on the level of pausing and/or attenuation. The different effect of the same CTCF binding site mutation on transcription in two cell lines representing two different hematopoietic lineages may reflect variations in relative proportion of different cell-type-specific forms of CTCF (see below) and also suggests that stimulation of the c-*myc* transcription by CTCF may depend on other, cell-type-specific nuclear factors.

To gain insight into how CTCF regulates the c-*myc* promoter, we have attempted to clone and sequence a full-length cDNA clone encoding it. From purified 130-kDa CTCF protein three peptides were isolated, sequenced, and used to isolate overlapping consecutive cDNA sequences which, when combined, revealed ORF encoding 728 amino acids corresponding to a polypeptide chain of 82 kDa with 11 Zn finger motifs (Fig. 4). When expressed in cells, it produced a nuclear protein with apparent molecular mass of 70 kDa as determined by SDS-gel electrophoresis (Fig. 8). Three lines of evidence prove that it has characteristics of CTCF: (i) it harbors amino acid sequences of all three peptides isolated from purified 130-kDa protein; (ii) its 11-Zn-finger domain translated in vitro specifically binds to the CTCF-binding DNA sequence recognizing exactly the

same nucleotides (Fig. 6); and (iii) antibodies raised against synthetic peptides derived from the cloned sequence recognize the endogenous 130-kDa nuclear protein (Fig. 8 and 9) and specifically interfere with the binding of CTCF to its DNA target in nuclear extracts (Fig. 7). Therefore, we concluded that cDNA shown in Fig. 4 is likely to represent one of several possible mature CTCF mRNA isoforms. Indeed, Northern (RNA) blot analysis (Fig. 5) demonstrated four major CTCF RNA species of 3.7, 4.0, 6.5, and 7.0 kb, indicating that CTCF gene may encode multiple proteins by generating a variety of mRNA isoforms. Consistent with this idea, Western (protein) blot analysis revealed at least six major differentially expressed forms of CTCF proteins with apparent molecular masses of 130, 97, 80, 73, 70, and 55 kDa (Fig. 8). It is tempting to speculate that alternative mRNA processing is involved in generating multiple isoforms of CTCF mRNAs and, respectively, of multiple forms of proteins, because among Zn finger proteins alternative mRNA splicing appears to be a common way to generate different isoforms of transcription regulators (7, 10, 26, 55). Alternative explanations for multiple forms of CTCF proteins would be a highly related multigene family, multiple widely separated transcriptional start sites, or proteolysis. Although we have not obtained any data which support these alternative explanations, cloning and sequencing of cDNAs corresponding to the different CTCF proteins together with complete sequencing of the genomic CTCF locus will be required in order to fully understand the molecular mechanism by which different cells produce different forms of CTCF.

It is conceivable that some of the forms of CTCF may be either repressors or activators. Even though most of the individual CTCF forms share identical DNA-binding specificity (34) and, therefore, could indeed behave as competitive transcription regulators possessing distinct *trans*-acting activity, the functional significance of multiple forms of CTCF protein binding to the same DNA target will remain unclear until we define the *trans*-activating and/or *trans*-repressing domain(s) of each individual differentially expressed CTCF isoform. In preliminary experiments we have found that (i) in cotransfection experiments carried out using recipient cells which express 130- and 80-kDa isoforms of the endogenous CTCF, expression of the recombinant 82-kDa isoform specified by the cDNA clone shown in Fig. 4 results in selective *trans* repression of an indicator gene fused to the c-*myc* promoter and, (ii) at least one domain of the 82-kDa CTCF isoform (amino acids from 115 to 210 at the N-terminal side to the 11-Zn-finger domain [Fig. 4]) behaves as a strong transcriptional repressor when fused to the GAL4 DNA-binding domain (34). These observations suggest that the major longer (130- and 80-kDa) CTCF isoforms which predominate in both HD3 and BM2 cells (Fig. 8) may lack the repressor domain and/or may possess an activating domain(s) absent in the cloned CTCF isoform, because in these cells the CTCF-binding DNA sequence of the c-*myc* promoter acts as a positive transcriptional element (Fig. 3).

Binding of the CTCF protein(s) to the FpV DNA sequence activates transcription directed by the c-*myc* promoter in stably transfected HD3 and BM2 cells more efficiently than does binding of two Sp1-like factors and poly(G)-binding protein to the same DNA region without CTCF (Fig. 3). This fact suggests a negative role for the two Sp1-like proteins binding to the promoter in the presence of 130- and 80-kDa or 130-, 97-, and 80-kDa CTCF protein isoforms detected in HD3 and BM2 cells, respectively. In our studies, these Sp1-like activities were not formally shown to be the positive transcription factor Sp1 as originally defined (30). They may represent other members of the growing family of Sp1 consensus sequence (GC box) binding factors (28, 32, 64) that may possess transcription repressor properties. It is also worth noting that a suggested negative role for Sp1-like factors is not unique to the chicken c-*myc* promoter. The interleukin-2 receptor α chain gene promoter contains a single Sp1 binding site that overlaps a CArG box binding serum response factor, SRF. In this promoter, Sp1 was also shown to serve a role of repressor by mutually exclusive binding mechanism similar to that described here (56). Transcription factors Sp1 and NF-E1 (also called YY-1, δ, and UCRBP) both belong to the Zn finger family and bind, exclusively of each other, to the initiator element in the cytochrome *c* oxidase subunit Vβ (COXVβ) gene promoter. Again, binding of NF-E1 but not Sp1 was found to be crucial for COXVβ promoter function (3). On the other hand, since no experiments to investigate the effects of mutations in either Sp1 or poly(G)-binding protein binding sites were carried out, these other factors may also turn out to be important for promoter activity acting in concert with different CTCF proteins.

Assuming that different forms of CTCF may have similar DNA-binding but distinct *trans*-activating characteristics, an intriguing mechanism of c-*myc* transcriptional regulation via competitive binding of several distinct Sp1-like factors and different forms of CTCF to the same promoter element could also be envisioned. Such a complicated but flexible mechanism would match appropriately the complexities of c-*myc* gene regulation during development and cell differentiation. Direct functional comparison of *trans*-activating domains of different forms of CTCF protein and Sp1-like proteins will be necessary to test these ideas. At present, the best evidence for functional role of the multiple CTCF proteins in situ remains the result of mutational analysis of their cognate binding DNA sequence presented here.

Data from Fig. 8 and 9 show that CTCF proteins are present in nuclei of cells from different species. Comparative quantitative EMSA titration of DNA-binding activity of Sp1 protein and CTCF in the same nuclear extract from dividing HD3 cells has shown that under moderate salt conditions, the two activities have practically identical binding constants and very similar concentrations (33). Since it is known that Sp1 is a highly abundant ubiquitous protein (40), this result suggests that some CTCF isoforms may be as abundant as general transcription factor Sp1. Since our data showed that chicken CTCF proteins are involved in regulation of the chicken c-*myc* promoter, mammalian CTCF proteins are likely to regulate mammalian c-*myc* promoters. Although the promoters of human and mouse c-*myc* genes contain sequences which match reasonably well to the essential contact nucleotides for CTCF binding, CTCF-binding sites in cloned DNA from these genes have not been mapped directly. It is also possible that c-*myc* promoters may not be unique targets for CTCF proteins. In addition to detailed study of the expression mechanisms of the CTCF gene itself, isolation of CTCF cDNA clones reported here will also allow us to identify other target genes regulated by CTCF.

ACKNOWLEDGMENTS

We thank G. Loring, C. Deminie, S.-J. Thomas, R. To, T. LaPorte, S. Fagerlie, M. Zaboikin, T. Zaitsevskaja, G. Filippova, S. Mirkin, L. Kretzner and B. McStay for assistance and support. Our thanks also to S. Collins, E. M. Blackwood, R. Eisenman, and M. Groudine for critical reading of the manuscript and to C. Marshall for interest and encouragement. We are indebted to U. Kruse and A.

Sippel for generously providing BM2 cDNA library and to A. Begue and V. Laudet for the chicken genomic library.

This work was funded by the NIH R03 TW00057 Fogarty International Research Collaboration Award to P. Neiman and V. Lobanenkov, by grant NIH/NCI RO1 CA20068 to P. Neiman, by the Cancer Research Compaign grants to G. Goodwin, by Yamagiwa-Yoshida Memorial International Cancer Study Grant to V. Lobanenkov, and by EMBO Fellowships for East Europeans to E. Klenova.

REFERENCES

1. **Adamson, P., H. F. Paterson, and A. Hall.** 1992. Intracellular localization of the P21*rho* proteins. J. Cell Biol. **119:**617–627.
2. **Askew, D. S., R. A. Ashmun, B. C. Simmonds, and J. L. Cleveland.** 1991. Constitutive *c-myc* expression in an IL-3-dependent cell line suppresses cell cycle arrest and accelerates apoptosis. Oncogene **6:**1915–1922.
3. **Basu, A., K. Park, M. L. Atchison, R. S. Carter, and N. G. Avadhani.** 1993. Identification of a transcriptional initiator element in the cytochrome c oxidase subunit Vβ promoter which binds to transcription factors NF-E1 (YY-1, δ) and Sp1. J. Biol. Chem. **268:**4188–4196.
4. **Berg, J. M.** 1990. Zinc fingers and other metal binding domains. J. Biol. Chem. **265:**6513–6516.
5. **Beug, H., G. Doederlin, C. Freudenstein, and T. Graf.** 1982. Erythroblast cell lines transformed by a *ts* mutant of AEV: a model system to study erythroid differentiation *in vitro*. J. Cell. Physiol. 1(Suppl.):195–207.
6. **Beug, H., A. von Kirchbach, G. Doederlein, J.-F. Conscience, and T. Graf.** 1979. Chicken hematopoietic cells transformed by seven strains of defective avian leukemia viruses display three distinct phenotypes of differentiation. Cell **18:**579–588.
7. **Bickmore, W. A., K. Oghene, M. H. Little, A. Seawright, V. van Heyningen, and N. D. Hastie.** 1992. Modulation of DNA binding specificity by alternative splicing of the Wilms tumor *wt1* gene transcript. Science **257:**235–237.
8. **Bird, A. P.** 1986. CpG-rich islands and the function of DNA methylation. Nature (London) **321:**209–213.
9. **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.
10. **Bordereaux, D., S. Fichelson, P. Tambourin, and S. Gisselbrecht.** 1990. Alternative-splicing of the *Evi-1* zinc finger gene generates mRNAs which differ by the number of zinc finger motifs. Oncogene **5:**925–927.
11. **Bossone, S. A., C. Asselin, A. J. Patel, and K. B. Marcu.** 1992. MAZ, a zinc finger protein, binds to *c-myc* and C2 gene sequences regulating transcriptional initiation and termination. Proc. Natl. Acad. Sci. USA **89:**7452–7456.
12. **Brunelle, A., and R. F. Schleif.** 1987. Missing contact probing of DNA-protein interactions. Proc. Natl. Acad. Sci. USA **84:**6673–6676.
13. **Chomczynski, P., and N. Sacchi.** 1987. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal. Biochem. **162:**156–159.
14. **Cole, M. D.** 1986. The *myc* oncogene: its role in transformation and differentiation. Annu. Rev. Genet. **20:**361–384.
15. **Collins, S., L. Mueller, and V. Lobanenkov.** Unpublished data.
16. **Dignam, J. D., R. M. Lebovitz, and R. G. Roeder.** 1983. Accurate transcription initiation of RNA polymerase II in a soluble extract from isolated mammalian nucleic. Nucleic Acids Res. **11:**1475–1489.
17. **Dingwall, C., and R. A. Laskey.** 1991. Nuclear targeting sequences—a consensus? Trends Biochem. Sci. **16:**478–481.
18. **El-Baradi, T., and P. Tomas.** 1991. Zinc finger proteins: what we know and what we would like to known. Mech. Dev. **35:**155–169.
19. **Evan, G. L., 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:**119–125.
20. **Felgner, P. L., T. R. Gadek, M. Holm, R. Roman, H. N. Chan, M. Wenz, J. P. Northrop, G. M. Ringold, and M. Danielsen.** 1987. Lipofection: a highly efficient, lipid-mediated DNA-transfection procedure. Proc. Natl. Acad. Sci. USA **84:**7413–7417.
21. **Flanagan, J. R., K. G. Becker, D. L. Ennist, S. L. Gleason, P. H. Driggers, B.-Z. Levi, E. Appella, and K. Ozato.** 1992. Cloning of a negative transcription factor that binds to the upstream conserved region of Moloney murine leukemia virus. Mol. Cell. Biol. **12:**38–44.
22. **Giniger, E., and M. Ptashne.** 1987. Transcription in yeast activated by a putative amphipathic α helix linked to a DNA binding unit. Nature (London) **330:**670–672.
23. **Harlow, E., and D. Lane.** 1988. Antibodies: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
24. **Helin, K., J. A. Lees, M. Vidal, N. Dyson, E. Harlow, and A. Fattaev.** 1992. A cDNA encoding a pRB-binding protein with properties of the transcription factor E2F. Cell **70:**337–350.
25. **Hirt, R. P., O. Poulain-Godefroy, J. Billotte, J.-P. Kraehenbuhl, and N. Fasel.** 1992. Highly inducible synthesis of heterologous proteins in epithelial cells carrying a glucocorticoid-responsive vector. Gene **111:**199–206.
26. **Hsu, T., J. A. Gogos, S. A. Kirsh, and F. C. Kafatos.** 1992. Multiple zinc finger forms resulting from developmentally regulated alternative splicing of a transcription factor gene. Science **257:**1946–1950.
27. **Hunter, T., and M. Karin.** 1992. The regulation of transcription by phosphorylation. Cell **70:**375–387.
28. **Imataka, H., K. Sogawa, K. Yasumoto, Y. Kikuchi, K. Sasano, A. Kobayashi, M. Hayami, and Y. Fujii-Kuriyama.** 1992. Two regulatory proteins that bind to the basic transcription element (BTE), a GC box sequence in the promoter region of the rat P-450 1A1 gene. EMBO J. **11:**3663–3673.
29. **Kadonaga, J. T.** 1991. Purification of sequence-specific binding proteins by DNA affinity chromatography. Methods Enzymol. **208:**10–23.
30. **Kadonaga, J. T., A. J. Courey, J. Ladika, and R. Tjian.** 1988. Distinct regions of Sp1 modulate DNA binding and transcriptional activation. Science **242:**1566–1570.
31. **Kim, S., E. H. Humphries, L. Tjoelker, L. Carlson, and C. B. Thompson.** 1990. Ongoing diversification of the rearranged immunoglobulin light-chain gene in a bursal lymphoma cell line. Mol. Cell. Biol. **10:**3224–3231.
32. **Kingsley, C., and A. Winoto.** 1992. Cloning of GT box-binding proteins: a novel Sp1 multigene family regulating T-cell receptor gene expression. Mol. Cell. Biol. **12:**4251–4261.
33. **Klenova, E. M., and V. V. Lobanenkov.** Unpublished data.
34. **Klenova, E. M., G. Loring, R. H. Nicolas, G. H. Goodwin, P. E. Neiman, and V. V. Lobanenkov.** Unpublished data.
35. **Klug, A., and D. Rhodes.** 1987. "Zinc fingers": a novel protein motif for nucleic acid recognition. Trends Biochem. Sci. **12:**464–467.
36. **Kolluri, R., and A. J. Kinniburgh.** 1991. Full length cDNA sequence encoding a nuclease-sensitive element DNA binding protein. Nucleic Acids Res. **17:**4771.
37. **Kozak, M.** 1987. An analysis of 5′-noncoding sequences from 699 vertebrate RNAs. Nucleic Acids Res. **15:**8125–8148.
38. **Kretzner, L., E. M. Blackwood, and R. N. Eisenman.** 1992. MYC and MAX proteins possess distinct transcriptional activity. Nature (London) **359:**426–429.
39. **Lachmann, P. J., L. Strangeways, A. Vyakarnam, and G. I. Evan.** 1986. Raising antibodies by coupling peptides to PPD and immunizing BCG-sensitized animals. CIBA Found. Symp. **119:**25–40.
40. **Letovsky, J., and W. S. Dynan.** 1989. Measurement of the binding of transcription factor Sp1 to a single GC box recognition sequence. Nucleic Acids Res. **17:**2639–2653.
41. **Lobanenkov, V. V., V. V. Adler, E. M. Klenova, R. H. Nicolas, and G. H. Goodwin.** 1989. CCCTC-binding factor (CTCF): a novel sequence-specific DNA-binding protein which interacts with the 5′-flanking sequence of the chicken *c-myc* gene, p. 45–68. *In* T. S. Papas (ed.), Gene regulation and AIDS: transcriptional activation, retroviruses and pathogenesis. Portfolio Publishing Corp., Woodlands, Tex.
42. **Lobanenkov, V. V., and P. E. Neiman.** Unpublished data.
43. **Lobanenkov, V. V., R. H. Nicolas, V. V. Adler, H. Paterson,**

E. M. Klenova, A. V. Polotskaja, and G. H. Goodwin. 1990. A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5′-flanking sequence of the chicken *c-myc* gene. Oncogene **5:**1743–1753.

44. **Lobanenkov, V. V., R. H. Nicolas, M. A. Plumb, C. A. Wright, and G. H. Goodwin.** 1986. Sequence-specific DNA-binding proteins which interact with (G+C)-rich sequences flanking the chicken *c-myc* gene. Eur. J. Biochem. **159:**181–188.
45. **Marcu, K. B., S. A. Bossone, and A. J. Pate.** 1992. Myc function and regulation. Annu. Rev. Biochem. **61:**809–860.
46. **Maxam, A., and W. Gilbert.** 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. **65:** 499–560.
47. **Moll, T., G. Tebb, U. Surana, H. Robitish, and K. Nasmyth.** 1991. The role of phosphorylation and the CDC28 protein kinase in cell cycle-regulated nuclear import of the *S. cerevisiae* transcription factor SWI5. Cell **66:**743–758.
48. **Moscovici, C., N. Zeller, and M. G. Moscovici.** 1982. Continuous lines of AMV-transformed non-producer cells: growth and oncogenic potential in the chick embryo, p. 435–449. *In* R. F. Revoltella, G. M. Pontieri, C. Basiliico, G. Rovera, R. C. Gallo, and J. H. Subak-Sharpe (ed.), Expression of differentiated function in cancer cells. Raven Press, New York.
49. **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.
50. **Nicolas, R. H., and G. H. Goodwin.** 1993. Purification and cloning of transcription factors, p. 81–104. *In* D. S. Latchman (ed.), Transcription factors: a practical approach. IRL Press, Oxford.
51. **Pearson, R. B., and B. E. Kemp.** 1991. Protein kinase phosphorylation site sequences and consensus specificity motifs: tabulations. Methods Enzymol. **200:**62–81.
52. **Plumb, M. A., V. V. Lobanenkov, R. H. Nicolas, C. A. Wright, S. Zavou, and G. H. Goodwin.** 1986. Characterization of chicken erythroid nuclear proteins which bind to the nuclease hypersensitive regions upstream of the β^A- and β^H-globin genes. Nucleic Acids Res. **14:**7675–7693.
53. **Pyrc, J. J., K. H. Moberg, and D. J. Hall.** 1992. Isolation of a novel cDNA encoding a zinc-finger protein that binds to two sites within the *c-myc* promoter. Biochemistry **31:**4102–4110.
54. **Ray, R., and D. M. Miller.** 1991. Cloning and characterization of a human *c-myc* promoter-binding protein. Mol. Cell. Biol. **11:**2154–2161.
55. **Read, D., and J. L. Manley.** 1992. Alternatively spliced transcripts of the *Drosophila ramtrack* gene encode zinc finger proteins with distinct DNA binding specificities. EMBO J. **11:**1035–1044.
56. **Roman, D. G., M. B. Toledano, and W. J. Leonard.** 1990. Sp1 represses IL-2 receptor α chain gene expression. New Biol. **2:**642–647.

56a. **Roy, A. L., M. Meisterernst, P. Pognonec, and R. G. Roeder.** 1991. Cooperative interaction of an initiator-binding transcription initiation factor and the helix-loop-helix actinator USF. Nature (London) **354:**245–248.

57. **Rupp, R. A., U. Kruse, G. Multhap, U. Gobel, K. Beyreuther, and A. E. Sippel.** 1990. Chicken NFI/TGGCA proteins are encoded by at least three independent genes: NFI-A, NFI-B, and NFI-C with homologues in mammalian genome. Nucleic Acids Res. **18:**2607–2616.
58. **Ruppert, J. M., K. W. Kinzler, A. J. Wong, S. H. Bigner, F.-T. Kao, M. L. Law, H. N. Seuanez, S. J. O'Brien, and B. Vogelstein.** 1988. The *GLI*-Kruppel family of human genes. Mol. Cell. Biol. **8:**3104–3113.
59. **Sambrook, J., E. F. Fritsch, and T. Maniatis.** 1989. Molecular cloning: a laboratory manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
60. **Saraste, M., P. R. Sibbald, and A. Wittinghofer.** 1990. The P-loop—a common motif in ATP- and GTP-binding proteins. Trends Biochem. Sci. **15:**430–434.
61. **Seto, E., Y. Shi, and T. Shenk.** 1991. YY1 is an initiator sequence-binding protein that directs and activates transcription *in vitro*. Nature (London) **354:**241–245.
62. **Spencer, C. A., and M. Groudine.** 1990. Transcription elongation and eukaryotic gene regulation. Oncogene **5:**777–785.
63. **Spencer, C. A., and M. Groudine.** 1991. Control of *c-myc* regulation in normal and neoplastic cells. Adv. Cancer Res. **56:**1–48.
64. **St.-Arnaud, R., and J. M. Moir.** 1993. *Wnt-1*-inducing factor 1: a novel G/C box-binding transcription factor regulating the expression of *Wnt-1* during neuroectodermal differentiation. Mol. Cell. Biol. **13:**1590–1598.
65. **Stappert, J., J. Wizsching, and R. Kemler.** 1992. A PCR method for introducing mutations into cloned DNA by joining an internal primer to a tagged flanking primer. Nucleic Acids Res. **20:**624.
66. **Symonds, G., K.-H. Klempnauer, G. I. Evan, and J. M. Bishop.** 1984. Induced differentiation of avian myeloblastosis virus-transformed myeloblasts: phenotypic alteration without altered expression of the viral oncogene. Mol. Cell. Biol. **4:**2587–2593.
67. **Tevosian, S. G., E. S. Gromova, V. V. Adler, and V. V. Lobanenkov.** 1991. Regulatory protein factor CTCF interacts with a segment of the chicken *c-myc* promoter capable of changing to a non-canonical DNA conformation. Mol. Biol. (Moscow) **25:**1013–1023.
68. **Thalmeier, K., H. Synovzik, R. Mertz, E.-L. Winnaker, and M. Lipp.** 1989. Nuclear factor E2F mediates basic transcription and trans-activation by E1A of the human MYC promoter. Genes Dev. **3:**527–536.
69. **Tsang, S. S., X. Yin, C. Guzzo-Arkuran, V. S. Jones, and A. J. Davison.** 1993. Loss of resolution in gel electrophoresis of RNA: a problem associated with the presence of formaldehyde gradients. BioTechniques **14:**380–381.
70. **Walker, J. E., M. Saraste, M. J. Runswick, and N. J. Gay.** 1982. Distantly related sequences in the α- and β-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. EMBO. J. **1:**945–951.
71. **Weis, L., and D. Reinberg.** 1992. Transcription by RNA polymerase II: initiator-directed formation of transcription-competent complexes. FASEB J. **6:**3300–3309.