Structural Organization of Upstream Exons and Distribution of Transcription Start Sites in the Chicken *c-myb* Gene

SOONJUNG L. HAHN, MOUNOU HAHN, AND WILLIAM S. HAYWARD*

Graduate Program in Molecular Biology, Sloan-Kettering Institute for Cancer Research, and Cornell University Graduate School of Medical Sciences, 1275 York Avenue, New York, New York 10021

Received 1 August 1988/Accepted 3 November 1988

We mapped and sequenced three upstream exons of the chicken *c-myb* gene and the regions flanking the first coding exon. We found multiple potential binding sites for transcription factors in the 5'-noncoding region, a T-rich stretch of 78 base pairs (bp) (68% T) in the first intron, and four fairly long open reading frames in the antisense direction of the first coding exon and its flanking regions. Three major transcription start sites, contained within a single 11-bp region, were identified by S1 nuclease analysis and primer extension. A sequence comparison of the avian and murine *c-myb* genes revealed a highly conserved sequence of 124 bp in the 5'-noncoding region. Its location between the putative transcription factor binding sites and the major transcription start sites suggests that it may play an important regulatory role in *c-myb* expression.

The proto-oncogene *c-myb*, which encodes a 75-kilodalton nuclear protein, is the cellular counterpart of the *v-myb* genes carried by avian myeloblastosis virus (AMV) and E26 (5, 31). AMV produces a *gag-myb* fusion protein of 45 kilodaltons, causing myeloid leukemia in chickens; E26 produces a 135-kilodalton *gag-myb-ets* fusion protein, inducing myeloid and erythroid leukemia (6, 18, 25, 29). Both the *c-myb* and *v-myb* proteins are localized in the nucleus and bind to DNA (6, 19, 20). The level of *c-myb* expression is highest in immature hematopoietic cells and decreases with differentiation (9, 15, 31, 34).

Extensive truncations have occurred both at the 5' and 3' ends of *c-myb* during transduction by retroviruses (12, 13, 17, 18, 30, 32). The sequences homologous to *v-myb* are present in seven exons of the *c-myb* gene, which has additional exons both 5' and 3' of the *v-myb* homologous exons (17, 22).

In addition to the extensive truncation, there are 11 amino acid substitutions in the *v-myb* gene of AMV (17). However, a number of observations suggest that truncation of the protein is more important for the oncogenic activation of *c-myb* than are amino acid substitutions. First, the *v-myb* gene carried by E26, which is shorter than that of AMV, encodes a protein with only one amino acid substitution, which is not shared with the *v-myb* gene of AMV (28) (although the *gag* and *ets* sequences in the E26 fusion protein may contribute to its oncogenic activity). Second, Stober-Grässer and Lipsick (39) have shown that an AMV construct containing a *v-myb* gene with no amino acid substitutions still induces transformation of yolk sac cells. Furthermore, *c-myb* has been shown to be insertionally activated in myeloid leukemias induced by Moloney murine leukemia virus (14, 35). The proviruses in these tumors are integrated downstream of the first two coding exons of *c-myb*, resulting in extensive truncation of the protein.

Our laboratory has recently demonstrated insertional activation of the chicken *c-myb* gene in a new class of rapidly appearing B-cell lymphomas (16) that are induced by a recombinant avian leukosis virus, EU-8 (16, 36). The integrated proviruses in this system map farther upstream than they do in the murine leukemia virus-induced myeloid leu-

kemias, suggesting that the *c-myb* protein is truncated to a lesser extent. These observations prompted us to investigate the structural organization of the 5' exons of chicken *c-myb*.

The exon organization of the *c-myb* proto-oncogene has been studied most extensively in the murine system, in which three additional exons have been mapped upstream of the *v-myb* homologous exons (22). Although it could be assumed that the upstream exons of the chicken *c-myb* gene are organized in a similar pattern based on the highly conserved nature of proto-oncogenes, the possibility of divergence could not be ruled out, because two chicken *c-myb* cDNA sequences with different 5' ends have been described (12, 30). The possibility of alternative splicing in a tissue-specific fashion has been proposed to account for the difference (12). In this study, we mapped and sequenced three upstream exons of chicken *c-myb* and identified several transcription start sites. A comparison between the chicken and murine 5'-noncoding regions is also presented.

Genomic cloning. The clones used in this study, pSH1 and pMK1 (Fig. 1), were isolated from normal peripheral blood DNA and liver tumor DNA, respectively, of chickens infected with EU-8 virus (16, 36). pMK1 has been described previously (16), and pSH1 was isolated as follows. Genomic DNA was digested to completion with *Bgl*II and was size-fractionated on sucrose gradients. Fractions containing fragments of approximately 13 kilobase pairs (kbp) were pooled, and the fragments were digested with *Bam*HI. Since *Bam*HI does not cut the 13-kb *Bgl*II fragment derived from the *c-myb* locus, the double digestion scheme results in an enrichment for the fragment of interest. After a second size fractionation the 13-kb fragments were ligated to *Bam*HI-digested and dephosphorylated EMBL3, a λ replacement vector (*Bam*HI and *Bgl*II ends are compatible; Promega). Ligated phage DNAs were packaged in vitro with packaging extract (Gigapack; Stratagene). *Escherichia coli* MB406 (*recB21 recC22 sbcB15 supE hflA hflB hsdR*) was used for plating the phages (as recommended by Promega) because of its *RecB*⁻ *RecC*⁻ phenotype, in an effort to obtain an unbiased genomic library (42).

The library was screened as described by Maniatis et al. (23) by using a 2.5-kb *Oxa*NI (an isoschizomer of *Mst*II) fragment of pMK1 (Fig. 1) that was labeled with [α -³²P] dCTP by random priming (11). Four positive clones were

* Corresponding author.

FIG. 1. Structure of the *c-myc* locus in the chicken (CH) and the mouse (MU). *v-myc* homologous exons (■) and upstream exons (□) are shown. The exon map of the murine *c-myc* locus is from Watson et al. (41); the locations of exons and restriction sites of the chicken *c-myc* gene downstream of UE1 obtained from previously published reports (13, 16, 17); those farther upstream were determined in this study. The clones used in this study are shown below the corresponding regions. pMK1 contains a part of the 3' long terminal repeat sequence (sideways W in the pMK1 map) of the EU-8 provirus (16); pSH1 carries the *Bgl*II-*Hind*III fragment of λSH3. Restriction sites are abbreviated as follows: Ac, *Acc*I; B, *Bam*HI; Bg, *Bgl*II; E, *Eco*RI; H, *Hind*III; Hc, *Hinc*II; N, *Not*I; Nc, *Nco*I; O, *Oxa*NI; P, *Pst*I; Pv, *Pvu*II; PI, *Pvu*I; S, *Sac*I. The order of the 5'-most *Bam*HI and *Hind*III sites was not determined. *Pst*I sites are shown only in cases in which they are relevant to the sequence analyses. Fragments used for the sequencing analyses of upstream exons are marked with double-headed arrows. For fragment S1, only the region from the 5' *Sac*I site to the *Not*I site was sequenced.

isolated from 1.1×10^5 plaques of recombinant phages. One of these, called λSH3 (Fig. 1), was used for further characterization. The insert of λSH3 was digested with *Hind*III, and the upstream fragment of 7.5 kb was subcloned into pUC19 to make pSH1.

Mapping and sequencing of upstream exons. pMK1 contains a 3.8-kb fragment that comprises the junction between the 3' long terminal repeat of an integrated EU-8 provirus and sequences downstream of the integration site in the *c-myc* locus (Fig. 1) (16). pMK1 was previously shown to include the first and second *v-myc* homologous exons (16). To test whether upstream exons were also present in this clone, we carried out Southern blot analysis (38) of pMK1 using *c-myc* cDNA (30) as a probe. Two fragments of pMK1, the 0.6-kb *Eco*RI-*Pst*I and 0.5-kb *Pvu*II-*Acc*I fragments, hybridized strongly to the *c-myc* cDNA sequence upstream of the first *v-myc* homologous exon (data not shown). We subcloned these fragments into M13 mp18 and mp19 for sequence analysis (24).

Nucleotide sequences were determined for both strands by the dideoxy chain termination method by using Sequenase (U.S. Biochemical Corp.) in the presence of [α - 35 S]-dATP plus unlabeled dCTP, dTTP, and either dGTP or dITP (40). Sequences thus obtained were compared with those of chicken *c-myc* cDNAs (12, 30) and murine *c-myc* genomic DNA (22).

For purposes of consistency we designated the *c-myc* exons as described by Lavu and Reddy (22), with the *v-myc* homologous exons (v1 to v7) numbered from 5' to 3' and the upstream exons (UE1 to UE3) numbered in the reverse order, beginning upstream of v1 (Fig. 1).

The chicken upstream exon UE1 was localized to the 0.5-kb *Pvu*II-*Acc*I fragment of pMK1, with its splice donor site being a few bases upstream of the *Acc*I site (Fig. 2). The 0.6-kb *Eco*RI-*Pst*I fragment was found to contain the 3' portion of the viral long terminal repeat (downstream of the *Eco*RI site in the long terminal repeat), the 3' half of UE2, and some intron sequences (the provirus integration site is indicated in Fig. 2). In the region corresponding to UE2, there was a one-base difference between the two published cDNA sequences (12, 30), and our result agrees with that of Rosson and Reddy (30) (Fig. 2). The difference may be due to an allelic polymorphism; the base change occurred at the third position of a codon, and did not alter the amino acid sequence.

Analyses of sequences farther upstream in the *c-myc* gene were performed with a genomic *c-myc* clone (pSH1) that was isolated from normal chicken cells as described above. The restriction map of the chicken *c-myc* locus and the relative positions of our clones are shown in Fig. 1. The restriction map of pSH1 agreed with that of pMK1 in the region shared by the two clones, including the position of an *Oxa*NI site. pMK1 carries two *Oxa*NI sites; the upstream site is located within UE2 (Fig. 1). To localize UE2 in pSH1, we cut the plasmid with *Pst*I to generate multiple fragments; only one of these, with a length of 0.5 kb, was cleaved by *Oxa*NI, indicating the presence of the UE2 sequence. The 0.5-kb *Pst*I fragment was thus subcloned into M13 and sequenced, revealing the entire UE2 sequence (Fig. 2). The splice acceptor site for UE2 was located at the extreme 5' end of the *Pst*I fragment. (The *Pst*I recognition site is a perfect subset of the splice acceptor consensus sequence.)

FIG. 2. Sequences of the chicken *c-myc* upstream exons and regions flanking UE3. Splice donor (┐) and acceptor (└) sites, potential Sp1-binding sites (double underscore), CACCC boxes (single underscore), and ATG start codon (triple diamonds) are indicated. Transcription start sites mapped by S1 nuclease analysis (smaller arrowheads; the most prominent site, at position 910, is indicated by a larger arrowhead) or by primer extension experiments (○) are shown with an accuracy of ± 2 bp. Coding sequences are written in triplets, 5'-noncoding sequences are in units of 10 bases, and introns are shown without spaces. The 5' ends of two published cDNA sequences are aligned below the genomic sequences determined in this study: sequence 1 from Rosson and Reddy (30) and sequence 2 from Gerondakis and Bishop (12) (eight bases at the 5' end, probably a linker sequence, are not included for the latter sequence). In the UE2 sequence, nucleotide 70 (marked with an asterisk) differs in the two cDNA sequences. The provirus integration site in UE2 in the EU-8-induced B-cell lymphoma from which pMK1 was derived (16) is indicated with an arrow.

UE3 and Flanking Sequences

SacI
 GAGCTCAGGG CACGGCCGGC CCGCCGAGCC CCCTCCCCGC CGTAGGTCGG CCGCCCCGCA 60
 GGGCTGGGAG CGCCCCGACG GGGCGGCCCC GGGGGCCGCG GCGCTGCGCC CTCACCCGGC 120
 TCCGTCCCCG CCGGACCCCG CCGCTTGGCG CCGCCACCCG CCGGAGCGCC CCGGGCCCCG 180
 CGCGGGGAGG CGCTGCTCCG CCCCGTCCG CCCCGCCCG CCGGGGCTCA TTGATGATTC 240
 GGAGCCTTTG AGGAAAGGGC GAGCGCGGTG CGAGAGCCGC GCGGCCCCGG TTTTACAAAG 300
 TCGCGCGCGG AGCCTGGGAA CTCGGCGCGT TTCGCCATTT TTTACGTCC TGAATCATTT 360
 CTGTAGCAAA AAAAACCGTG TTCTCTGGTG CCGATCGGAG ACCCACCGCC GCTCCCCGGC 420
 ACATCCCCGC GGGGAACAT CGCCAAATC CCCTCGGCAC CGCGCCGCGG CCCCACCCG 480
 CAACACGCAC GCAAAGCGCG CTCCTGTCCG GGGCCGCTCT GCCCGCAGCA GGTGACCCG 540
 GGGAGCGAGC CCGAGGGGC GGGAGGCGCG GGGCGTCCG CCGAGGTCTC CGAGACGCGG 600
 CCGTTCTCT CGATGGGGGG GTCCGACTCC CCGCTCCCGA CCGCCGCTTC CTGACAGCCT 660
 TACCCCGCG GGGGGCTGAG GCGGGGCGAG AGGGGGGGGG TGCCGGGCGC CGTGCCCCGC 720
 TCTACCCCGC TCCGCCCGCC CCGCGCCCGC CCGCCGCGGC ACGGGGCCGC GGTGCGCGCG 780
 CTGCGCACTG CAGGGGCGCC AGATTTGGCG GGAGGGGGAG TGTCCAAAG. CCCTTTGTTT 840
 (1) TTCCGGCGGC CGTAGCGGGC CATCTGCACG CGGAGCTCGC GGCCATCCAG CACGGCCCCG
 (2) GATGGCATCT CTGTTACAG AGTTTACTCT TTAATCTCAA CCTGTTTCCT CCTCCTCCTC 900
 (1) TCCATGGCGT CCATCGCGTC CTCGGCGTCG CGTTTGTCTT GGAAACGGAC GAATGCGAAC
 (2) AA CCTGTTTCCT CCTCCTCCTC
 (1) CCGCGGGCTCC TCGGCACTTC GCTGCGCGGC GCGGGAGCGC CTGGGCGCAG CCGGGCAGCA 960
 (2) CCGCGGGCTCC TCGGCACTTC GCTGCG.GGC GCGGGAG.G. CTGGGC..AG CC.GGCAGCA
 (1) GGGCGCGGCG GCGGGCGCGG GCTCCCCCGT CCTCTTCCCC CAGCGGCCGC CCGGAGG 1017
 (2) GGGCGCGGCG GCGGGCG.GG GCTCCCCC..ACCGGCCGC CG.GAGG
 NotI
 (1) ATG GCC CGG AGA CCC CGG CAC AG GTAAACGGGGGGGGTCCGCGCTGTCACA 1070
 (2) ATG GCC CGG AGA CCC CGG CAC AG
 ♦♦♦
SacI
 CCGGGCGCGGGGCTGCCGGAGAGAGCTCTCCCTGAGCGGGCGGCGGCGCGCCACCCCGCTCG 1135
 TACCTGAGCCGGCCCCGGCCCCCGAGCGTCCCTCGCTCCGAGGGGACGACACCGGGCGCCCCGCGCT 1200
 GCCCGCGCTCCCGGGGACCGGGGCTCGCTGCCGCCCCCGGCGGGGCTGCCCGCGGGAGCTGT 1265
 CGGGGCGGCGGGCAGCGCGTCCCGTCCCATTCATTCATTGCGGGCGGGCGGTGACGGCAGG 1330
 TGGGCGGCGGGCGGGGCGGGGCTCTCTCTTTCTCTTATTCCTTTCTTTCTTTCTTTCTTTCTT 1395
 TTTTGTGTTTGTGTTTGTGTTTGTGTTTCCCCACTTTATTTCCGGCCCGC 1441
 NotI

UE2

PstI
 CTGCAGC ATA TAC AGC AGC GAT GAC GAT GAA GAA GAT GTT GAG ATG TAC 49
 ↓GAC CAC GAT TAC GAC GGC CTG CTT CCT AAG GCT GGG AAA CGT CAC CTA 97
 GGG AAA ACC AGG TGG ACC CGT GAA GAG GTAAATGCGGGTTTGGT 141

UE1

ATTGTGATATTAATTATTTAATTGTTTATTTATTTTAAATGATTCAG GAT GAG AAA CTG 60
 AAG AAA CTT GTG GAA CAG AAT GGC ACA GAA GAC TGG AAA GTC ATT GCC 108
 AGT TTC CTT CCT GTAGGTAGAC 130
 AccI

The approximate position of UE3 was determined by using the *NotI* site as a marker. Both of the available chicken *c-myb* cDNAs (12, 30) contain a *NotI* site just upstream of the ATG start codon, which should be present in UE3. Our restriction mapping of pSH1 indicated the presence of two *NotI* sites about 430 bases apart (Fig. 1). It was surprising to find two *NotI* sites in such proximity, because *NotI* has an eight-base recognition site and includes two CpG pairs, which are known to occur in eucaryotic DNA at only about one-fifth the expected frequency (4). We subcloned two *SacI* fragments of pSH1, S1 and S2, each of which contained one *NotI* site, and sequenced them in both directions from the common *SacI* site. The ATG start codon was localized in the S2 fragment, which carried the upstream *NotI* site. The nucleotide sequence of the entire S2 was determined from seven overlapping M13 subclones. The UE3 sequence included the first 23 bases of the *c-myb*-coding region of the cDNA sequences; a consensus splice donor site was located at the 3' end of this sequence (Fig. 2). The location of this splice donor site and the boundaries of chicken exons UE2 and UE1 are the same as they are in the murine *c-myb* gene (22).

Comparison of genomic and cDNA sequences in the UE3 region. The two published chicken *c-myb* cDNA sequences are shown below our UE3 sequence in Fig. 2 for comparison. The published cDNA sequences of Gerondakis and Bishop (12) and Rosson and Reddy (30) are in good agreement throughout most of their length, but they diverge at the 5' ends, beginning about 100 bases upstream of the ATG in UE3 (Fig. 2, position 899). Our sequence of genomic DNA was in complete agreement with the cDNA sequence of Gerondakis and Bishop (12) within the region of divergence. There is no consensus splice acceptor sequence (26) at the point of divergence; thus, it is unlikely that the divergence reflects alternative splicing. The closest potential splice acceptor upstream of the divergence point was more than 100 base pairs (bp) away, at position 793/794. Downstream of the divergent region there are several small deletions (the longest of which is 13 bp) in the sequence of Gerondakis and Bishop (12) relative to that of Rosson and Reddy (30). In this region our genomic sequence matched the cDNA sequence of Rosson and Reddy (30). The deletions are in a noncoding region and may simply represent polymorphisms.

In an effort to determine the origin of the divergent upstream cDNA sequences, we used the first 124 bp of the Rosson and Reddy (30) cDNA clone (from the 5' end to the *SacII* site at position 904, which lies a few bases downstream of the divergence point; Fig. 2) as a hybridization probe for Southern and Northern blot analyses. The probe detected one or more fragments in each digest of genomic DNA. However, these sequences were not present on pSH1, which extended more than 3 kb upstream of the ATG start codon, nor were they present on restriction fragments of genomic DNA that extended more than 8 kb upstream of the ATG (data not shown). Furthermore, it did not hybridize to the 4-kb *c-myb* mRNA from chicken thymus (data not shown). We do not have an explanation for these sequences in the cDNA sequence of Rosson and Reddy (30).

In the 5'-noncoding region of our UE3 sequence there were stop codons in all reading frames: at nucleotides 249, 351, and 840 (frame 1); 677 and 872 (frame 2); and 43, 232, 235, 364, and 652 (frame 3). The distribution of stop codons is consistent with the suggestion of Gerondakis and Bishop (12) that the ATG codon at position 1018 (Fig. 2) is the start codon for the *c-myb* protein.

There are three possible splice acceptor sites upstream of

FIG. 3. Mapping of the transcription start sites of the chicken *c-myb* gene by S1 nuclease analysis. A uniformly labeled, single-stranded, antisense probe, extending from nucleotides 501 to 1008 in the 5'-noncoding region, was hybridized with 100 μ g of total RNA from chicken thymus (lane 1) and from yeast cells (lane 2). After digestion with S1 nuclease, protected fragments were fractionated on a 6% polyacrylamide-8 M urea gel, along with *MspI*-digested pBR322 DNA as a marker (lane M; sizes are indicated in base pairs). The largest band in each group of protected fragments is marked with an arrowhead. When an end-labeled, double-stranded probe was used, we detected only the three major sites (map positions 910, 916, and 920; Fig. 2) because of the low specific activity of the probe; the map positions were shifted by about two bases downstream relative to those obtained with the single-stranded probe.

the ATG codon, located at positions 45/46, 793/794, and 1003/1004. The first site is not present in the corresponding position of the murine sequence (1, 41). Nevertheless, it is an exceptionally good consensus splice acceptor sequence, which makes it difficult to rule out the possibility that additional upstream exons are present in the chicken *c-myb* gene.

Mapping of transcription start sites. S1 nuclease analysis and primer extension experiments were performed in order to map the transcription start sites of chicken *c-myb*. A uniformly labeled single-stranded probe (10^5 cpm), encompassing the sequences between the *Bss*HII (position 501) and *NotI* (position 1008) sites, was hybridized at 53°C with 100 μ g of total thymus RNA or, as a control, with the same amount of yeast RNA; then it was digested with 1,000 U of S1 nuclease per ml at 37°C for 30 min under the conditions described by Maniatis et al. (23). The results are shown in Fig. 3. Five transcription start sites were identified, each of which appeared as more than one band, probably because of nibbling at the ends of the duplex. We assume that the largest fragment in each group (Fig. 3, arrowheads) represents the correct start site and indicated these positions in Fig. 2.

The three major start sites, accounting for more than 95%

```

M616 . [CGGCCGCG] ccCtccc [CAGCC] cgggCt [CCCCGC] ccGcccgcggtcCcGcCgCgtGcG
C640 a [CGGCCGCG] ttCctga [CAGCC] ..ttCa [CCCCGC] ggGgggctgaggCgGgCgGgGgGgGg

M673 cGcaGgagCagctccgc [GGCGC] gagaggCgaTgggggCgGCGgCagggcggtgCagCtccggc
C696 gGggGtgcCg..... [GGCGC] cgtgccCgcTctaccCgCtCgcccgcgccCgagccCgagccCg

                                  ↓ DNase I HS
M739 gaCcGCGgaGgCggcgGGCag [GGCGC] g [TCGCACTGCAGGGGCGCCAGATTTCGCGGAGGGGG
C755 cgCgGCacgGGgCcgCGGCtc [GGCGC] c [TCGCACTGCAGGGGCGCCAGATTTCGCGGAGGGGG

                                     ****
M803 AGTGTCCAAA) .cCt [CTTTGTTTGATGGCA] .. [TCTGTTTACAGATT) .ACa [CTTTAA)
C819 AGTGTCCAAA) agCc [CTTTGTTTGATGGCA] tc [TCTGTTTACAGATT) tACt [CTTTAA)

*
M856 g [TCAACCTGTTTCTCTCTCTCT) tCtCtCtCTTCTCtCtCtCtCtCtCtCtCtCtCtCtCtCtCt
C876 c [TCAACCTGTTTCTCTCTCTCTCT) cCgCgGCTCtCTGgGCaacttcgC.....

M922 ccccttttctccggagaaacttgcccccggggtgCggagcgcg [TGCGC] aGcCgGgagGgGcGgc
C923 ..... [TGCGC] gGcGcGgagcGcCtgg

M988 GgcAAGgCggag [GGCAGC] gGgag [GCGGC] aaCcGGtGCGG.. [TCCCC] gGg [CTCTT)
C945 GgcAAGcCc... [GGCAGC] aGgGc [GCGGC] ggCgGGcGCGGgc [TCCCC] cGtc [CTCTT)

M1041 ggCggAGcc [CGGCC] cgCctcgcc [ATGGCCCGAGACCCCG] a [CACAGGTAA) tGGGgGg
C997 ccCccAG.. [CGGCC] gcCgCgagg [ATGGCCCGAGACCCCG] g [CACAGGTAA) cGGGcGg

                                  ← coding region →

M1099 ctGagagGgCggcctcgggcta [GGGCCGCG] cGGggatCc.....
C1053 ggGgtccGcgctgtcacacg.. [GGGCCGCG] gGcCtgcCggagagagctc

```

FIG. 4. Alignment of sequences of chicken and murine UE3 sequences and 5'-noncoding regions. Only the sequences surrounding the 124-bp homologous region (underlined) and the coding region of UE3 are shown. The murine sequence (prefix M) is from Bender and Kuehl (1). Numbering of the chicken sequence (prefix C) is as in Fig. 2. Identical nucleotides are in capital letters; regions of identity for five bases or more are bracketed. A DNase I-hypersensitive (HS) site (downward-pointing arrow) (41) and the TATA box (asterisks) (1) in murine *c-myb* are indicated.

of the total signal, were located within a single 11-bp region (positions 910, 916, and 920). Two minor sites, mapping farther upstream at positions 873 and 900, represented approximately 1 and 3%, respectively, of the total signal. The site at position 873 is just upstream of the 5' end of the cDNA of Gerondakis and Bishop (12); the next site (position 900) corresponds to the point of divergence between the two cDNAs.

The results obtained in the primer extension experiments (data not shown) were in good agreement with the S1 nuclease mapping data for the three major start sites (see positions indicated in Fig. 2). However, the minor sites at positions 873 and 900 were not detected by this technique. Two other possible start sites (positions 933 and 938) were detected only in the primer extension experiments and may represent premature terminations (strong stops) rather than actual transcription start sites. Primer extensions were performed essentially as described by Calzone et al. (7), using 4 μ g of chicken thymus poly(A)⁺ RNA for each assay. In order to eliminate artifacts caused by fortuitous binding of the primer to other RNAs, primer extensions were performed with two different 24-base oligonucleotides (one hybridizing to UE3 and the other hybridizing to UE2); only sites identified with both of the primers were considered valid.

Comparison between chicken and murine 5'-noncoding regions. In the 5'-noncoding region of the chicken UE3 sequence, which is highly GC-rich, we found a multitude of potential binding sites for two transcription factors, Sp1 (10) and the CACCC box-binding factor (8, 33), which were distributed over a broad region extending from positions 81

to 741. The 5'-noncoding region of chicken *c-myb* contained eight Sp1-binding sites (CCGCCC or its inverted complement) with two pairs of overlapping sites (Fig. 2), whereas the same region of the murine gene has three Sp1-binding sites, two of which are overlapping (1, 41). We found seven CACCC boxes in this region of the chicken gene and five in the published sequence of the murine gene. CACCC boxes have been identified previously in the upstream regions of the β -globin gene and β -like globin genes (8) and the rat tryptophan oxygenase gene (33). The CACCC box-binding factor is thought to be a ubiquitous transcription factor like Sp1 (33).

Overall, there was little homology between the 5'-noncoding regions of the chicken and murine *c-myb* genes; the sequences diverged from the first base upstream of the ATG start codon (Fig. 4). Farther upstream, however, we found a strikingly high homology in a very confined region. The homologous sequence was 124 bp long in the chicken 5'-noncoding region, with 93% identity to the murine counterpart; it was located 11 bp upstream of the major transcription start sites and 35 bp downstream of the region covered by potential factor-binding sites (between positions 776 and 899; Fig. 2 and 4). According to the data of Watson et al. (41), there is a DNase I-hypersensitive site 23 bp upstream of the main cluster of transcription start sites in the murine gene. This hypersensitive site maps near the 5' end of the homologous region described above. It is noteworthy that there is a TATA box, which is not associated with a CAAT box (1), within the region of homology in the murine gene. The corresponding sequence in the chicken gene has one base substitution (A to C at nucleotide 876), resulting in the

absence of a consensus TATA box. There is also no CAAT box in chicken *c-myb*, either in the homologous region or elsewhere in the 5' region.

The first 401 bp of the intron sequence between the UE3 splice donor site of chicken *c-myb* and the downstream *NotI* site are also included in Fig. 2. Near the *NotI* site there is a T-rich stretch of 78 bp (68% T). In the human and murine *c-myc* genes, T-rich regions near the exon 1-intron 1 boundary and in intron 1 have been implicated in gene regulation at the level of transcription elongation (3, 27). A similar elongation control point has recently been identified in the first intron of murine *c-myb* (2). It remains to be determined whether the T-rich sequence in the chicken *c-myb* gene plays a similar role.

In the antisense direction of the sequence covering chicken UE3 and its flanking regions, we found four fairly long open reading frames, starting with ATG codons located at positions 1303, 1299, 1295, and 848 and potentially coding for polypeptides of 193, 77, 147, and 206 amino acids, respectively. The fourth and longest open reading frame was particularly interesting in that its start codon satisfied the rule of Kozak (21) by having a purine at position -3. Starting 199 bp upstream of this ATG (in the antisense orientation) is a region that contains six potential Sp1-binding sites and one CACCC box. For murine *c-myb*, only the sequence of the first 50 bp of the corresponding intron has been published (1); in this sequence we detected one potential Sp1-binding site; however, antisense RNA transcripts from the UE3 region have been detected in murine cells (2).

While this report was in preparation, Soret et al. (37) reported similar mapping and sequencing data for the three upstream exons of chicken *c-myb*; our results are in good agreement with theirs. However, based on preliminary data they suggested the presence of sequences homologous to the 5' end of the cDNA sequence of Rosson and Reddy (30) within several kilobases upstream of the ATG start codon in UE3, whereas we found no evidence for these sequences within 8 kb upstream of the ATG.

In addition to the three upstream exons, we sequenced a region of about 1.4 kb flanking the coding region of UE3 and mapped transcription start sites. Three major start sites were identified by both S1 nuclease analysis and primer extension. Although we might have missed start sites with very weak signals or those that were located farther upstream, our data suggest that the transcription start sites in the chicken *c-myb* gene are more clustered and smaller in number than those of the murine gene (1, 41). Approximately 60% of chicken *c-myb* transcripts begin at a single site, and more than 95% initiate within a single 11-bp region. Beginning 11 bp upstream of the major start sites there is a highly conserved sequence of 124 bp that is shared by chicken and murine *c-myb*. It is likely that this region of homology plays an important role in the regulation of *c-myb* expression.

We thank M. R. Kanter for providing the plasmid pMK1 and for helpful discussions; E. Lacy, N. Neff, and F. Alonso-Caplen for valuable comments; and D. Rosson, E. P. Reddy, S. Gerondakis, and J. M. Bishop for cDNA clones. S.L.H. is especially grateful to Martha Howe.

This research was supported by Public Health Service grants CA43250 and CA16599 from the National Institutes of Health.

LITERATURE CITED

1. Bender, T. P., and W. M. Kuehl. 1986. Murine *myb* protooncogene mRNA: cDNA sequence and evidence for 5' heterogeneity. *Proc. Natl. Acad. Sci. USA* 83:3204-3208.
2. Bender, T. P., C. B. Thompson, and W. M. Kuehl. 1987.

- Differential expression of *c-myb* mRNA in murine B lymphomas by a block to transcription elongation. *Science* 237:1473-1476.
3. Bentley, D. L., and M. Groudine. 1988. Sequence requirements for premature termination of transcription in the human *c-myc* gene. *Cell* 53:245-256.
4. Bird, A. P. 1986. CpG-rich islands and the function of DNA methylation. *Nature (London)* 321:209-213.
5. Bister, K., M. Nunn, C. Moscovici, B. Perbal, M. A. Baluda, and P. H. Duesberg. 1982. Acute leukemia viruses E26 and avian myeloblastosis virus have related transformation-specific RNA sequences but different genetic structures, gene products, and oncogenic properties. *Proc. Natl. Acad. Sci. USA* 79:3677-3681.
6. Boyle, W. J., M. A. Lampert, J. S. Lipsick, and M. A. Baluda. 1984. Avian myeloblastosis virus and E26 virus oncogene products are nuclear proteins. *Proc. Natl. Acad. Sci. USA* 81:4265-4269.
7. Calzone, F. J., R. J. Britten, and E. H. Davidson. 1987. Mapping of gene transcripts by nuclease protection assays and cDNA primer extension. *Methods Enzymol.* 152:611-632.
8. Dierks, P., A. van Ooyen, M. D. Cochran, C. Dobkin, J. Reiser, and C. Weissmann. 1983. Three regions upstream from the cap site are required for efficient and accurate transcription of the rabbit β -globin gene in mouse 3T6 cells. *Cell* 32:695-706.
9. Duprey, S. P., and D. Boettiger. 1985. Developmental regulation of *c-myb* in normal myeloid progenitor cells. *Proc. Natl. Acad. Sci. USA* 82:6937-6941.
10. Dynan, W. S., and R. Tjian. 1983. The promoter-specific transcription factor Sp1 binds to upstream sequences in the SV40 early promoter. *Cell* 35:79-87.
11. Feinberg, A. P., and B. Vogelstein. 1983. A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. *Anal. Biochem.* 132:6-13.
12. Gerondakis, S., and J. M. Bishop. 1986. Structure of the protein encoded by the chicken proto-oncogene *c-myb*. *Mol. Cell. Biol.* 6:3677-3684.
13. Gonda, T. J., and J. M. Bishop. 1983. Structure and transcription of the cellular homolog (*c-myb*) of the avian myeloblastosis virus transforming gene (*v-myb*). *J. Virol.* 46:212-220.
14. Gonda, T. J., S. Cory, P. Sobieszczuk, D. Holtzman, and J. M. Adams. 1987. Generation of altered transcripts by retroviral insertion within the *c-myb* gene in two murine monocytic leukemias. *J. Virol.* 61:2754-2763.
15. Gonda, T. J., D. K. Sheiness, and J. M. Bishop. 1982. Transcripts from the cellular homologs of retroviral oncogenes: distribution among chicken tissues. *Mol. Cell. Biol.* 2:617-624.
16. Kanter, M. R., R. E. Smith, and W. S. Hayward. 1988. Rapid induction of B-cell lymphomas: insertional activation of *c-myb* by avian leukosis virus. *J. Virol.* 62:1423-1432.
17. Klempnauer, K.-H., T. J. Gonda, and J. M. Bishop. 1982. Nucleotide sequence of the retroviral leukemia gene *v-myb* and its cellular progenitor *c-myb*: the architecture of a transduced oncogene. *Cell* 31:453-463.
18. Klempnauer, K.-H., G. Ramsay, J. M. Bishop, M. G. Moscovici, C. Moscovici, J. P. McGrath, and A. D. Levinson. 1983. The product of the retroviral transforming gene *v-myb* is a truncated version of the protein encoded by the cellular oncogene *c-myb*. *Cell* 33:345-355.
19. Klempnauer, K.-H., and A. E. Sippel. 1986. Subnuclear localization of proteins encoded by the oncogene *v-myb* and its cellular homolog *c-myb*. *Mol. Cell. Biol.* 6:62-69.
20. Klempnauer, K.-H., G. Symonds, G. I. Evan, and J. M. Bishop. 1984. Subcellular localization of proteins encoded by oncogenes of avian myeloblastosis virus and avian leukemia virus E26 and by the chicken *c-myb* gene. *Cell* 37:537-547.
21. Kozak, M. 1984. Compilation and analysis of sequences upstream from the translational start site in eukaryotic mRNAs. *Nucleic Acids Res.* 12:857-872.
22. Lavu, S., and E. P. Reddy. 1986. Structural organization and nucleotide sequence of mouse *c-myb* oncogene: activation in ABPL tumors is due to viral integration in an intron which results in the deletion of the 5' coding sequences. *Nucleic Acids Res.* 14:5309-5320.

23. Maniatis, T., E. F. Fritsch, and J. Sambrook. 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
24. Messing, J. 1983. New M13 vectors for cloning. *Methods Enzymol.* **101**:21-79.
25. Moscovici, C., J. Samarut, L. Gazzolo, and M. G. Moscovici. 1981. Myeloid and erythroid neoplastic responses to avian defective leukemia viruses in chickens and in quail. *Virology* **113**:765-768.
26. Mount, S. M. 1982. A catalogue of splice junction sequences. *Nucleic Acids Res.* **10**:459-472.
27. Nepveu, A., and K. B. Marcu. 1986. Intragenic pausing and anti-sense transcription within the murine *c-myc* locus. *EMBO J.* **5**:2859-2865.
28. Nunn, M., H. Weiher, P. Bullock, and P. Duesberg. 1984. Avian erythroblastosis virus E26: nucleotide sequence of the tripartite *onc* gene and of the LTR, and analysis of the cellular prototype of the viral *ets* sequence. *Virology* **139**:330-339.
29. Radke, K., H. Beug, S. Kornfeld, and T. Graf. 1982. Transformation of both erythroid and myeloid cells by E26, an avian leukemia virus that contains the *myb* gene. *Cell* **31**:643-653.
30. Rosson, D., and E. P. Reddy. 1986. Nucleotide sequence of chicken *c-myb* complementary DNA and implications for *myb* oncogene activation. *Nature (London)* **319**:604-606.
31. Roussel, M., S. Saule, C. Lagrou, C. Rommens, H. Beug, T. Graf, and D. Stehelin. 1979. Three new types of viral oncogenes of cellular origin specific for hematopoietic cell transformation. *Nature (London)* **281**:452-455.
32. Rushlow, K. E., J. A. Lautenberger, T. S. Papas, M. A. Baluda, B. Perbal, J. G. Chirikjian, and E. P. Reddy. 1982. Nucleotide sequence of the transforming gene of avian myeloblastosis virus. *Science* **216**:1421-1423.
33. Schüle, R., M. Muller, H. Otsuka-Murakami, and R. Renkawitz. 1988. Cooperativity of the glucocorticoid receptor and the CACCC-box binding factor. *Nature (London)* **332**:87-90.
34. Sheiness, D., and M. Gardinier. 1984. Expression of a proto-oncogene (proto-*myb*) in hemopoietic tissues of mice. *Mol. Cell. Biol.* **4**:1206-1212.
35. Shen-Ong, G. L. C., H. C. Morse III, M. Potter, and J. F. Mushinski. 1986. Two modes of *c-myb* activation in virus-induced mouse myeloid tumors. *Mol. Cell. Biol.* **6**:380-392.
36. Simon, M. C., W. S. Neckameyer, W. S. Hayward, and R. E. Smith. 1987. Genetic determinants of neoplastic diseases induced by a subgroup F avian leukosis virus. *J. Virol.* **61**:1203-1212.
37. Soret, J., M. Vellard, C. Martinerie, and B. Perbal. 1988. Organization of 5'-proximal *c-myb* exons in chicken DNA: implications for *c-myb* tissue-specific transcription. *FEBS Lett.* **232**:227-234.
38. Southern, E. M. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. Biol.* **98**:503-517.
39. Stober-Grässer, U., and J. S. Lipsick. 1988. Specific amino acid substitutions are not required for transformation by *v-myb* of avian myeloblastosis virus. *J. Virol.* **62**:1093-1096.
40. Tabor, S., and C. C. Richardson. 1987. DNA sequence analysis with a modified bacteriophage T7 DNA polymerase. *Proc. Natl. Acad. Sci. USA* **84**:4647-4771.
41. Watson, R. J., P. J. Dyson, and J. McMahon. 1987. Multiple *c-myb* transcript cap sites are variously utilized in cells of mouse haemopoietic origin. *EMBO J.* **6**:1643-1651.
42. Wyman, A. R., K. F. Wertman, D. Barker, C. Helms, and W. H. Petri. 1986. Factors which equalize the representation of genome segments in recombinant libraries. *Gene* **49**:263-271.