Proc. Natl. Acad. Sci. USA
Vol. 83, pp. 9636–9640, December 1986
Genetics

Human c-*myb* protooncogene: Nucleotide sequence of cDNA and organization of the genomic locus

(nuclear protein/DNA-binding domain/chromosome 6)

BARBARA MAJELLO, LAWRENCE C. KENYON, AND RICCARDO DALLA-FAVERA

Department of Pathology and Kaplan Cancer Center, New York University School of Medicine, New York, NY 10016

Communicated by H. Sherwood Lawrence, September 3, 1986

ABSTRACT **We have isolated cDNA clones of the human c-*myb* mRNA that contain ≈3.4 kilobases of the ≈3.8-kilobase mRNA sequence. Nucleotide sequence analysis shows that the c-*myb* mRNA contains an open reading frame of 1920 nucleotides, which could encode a 72-kDa protein. The cDNA nucleotide sequence and the predicted amino acid sequence of the c-*myb* protein are highly homologous to the corresponding chicken and mouse proteins. In particular, a region toward the NH_2 terminus of the protein containing a 3-fold tandem repeat of 51 residues is evolutionarily conserved and is the only region of homology with the *Drosophila* c-*myb* protein. This region may represent a functionally important structure, most likely the DNA-binding domain. cDNA clones have been used to isolate genomic clones and to define a preliminary intron/exon organization of the c-*myb* gene. Identification of 5′ and 3′ coding and noncoding exons indicates that the human c-*myb* locus spans a 40-kilobase region.**

The c-*myb* protooncogene is an evolutionarily conserved locus identified by its homology with the transforming gene v-*myb*, of avian myeloblastosis virus (AMV) and avian leukemia virus E26 (1–3). The products of the chicken c-*myb* gene and the AMV v-*myb* gene have been preliminarily identified as nuclear DNA-binding proteins of 75 kDa and 45 kDa, respectively (4–6). Among different protooncogenes, and in particular among the ones coding for nuclear proteins, c-*myb* is unique in that its expression is apparently tissue-specific. Although c-*myb* transcripts have occasionally been found in nonhematopoietic tumors (7, 8) in various animal species, c-*myb* mRNA has primarily been found in hematopoietic cells where abundant transcripts are detectable in both myeloid and lymphoid precursors (9–10). In these cells c-*myb* expression appears to be regulated during terminal differentiation since c-*myb* mRNA is found in immature myeloid and lymphoid cells but not in mature cells of the same lineage (10). Accordingly, *in vitro* induction of terminal differentiation is accompanied by an early disappearance of *myb* transcripts in several myeloid cell lines (10, 11), suggesting that the protein encoded by c-*myb* may be involved in the control of growth and/or differentiation of hematopoietic cells.

The tissue-specificity of c-*myb* expression is also reflected by the association between different alterations of *myb* sequences and the development of hematopoietic neoplasms in different species. The v-*myb* oncogene is responsible for the ability of AMV to cause myeloblastic leukemia in chickens and to transform avian myelomonocytic cells in culture (12, 13). The disruption of cellular *myb* sequences by viral insertional mutagenesis appears to be associated with the pathogenesis of mouse hematopoietic tumors (14–16). The mechanism leading to the oncogenic effect both of v-*myb* in AMV and of c-*myb* in murine leukemia appears to be analogous, involving the truncation of 5′ and/or 3′ coding sequences of the protooncogene (5, 16).

The association between c-*myb* activation and hematopoietic neoplasms in several animal species suggests that analogous mechanisms may be involved in the pathogenesis of human leukemias and lymphomas. Two observations provide preliminary support to this hypothesis. First, the human c-*myb* locus (*MYB*, in standard human gene nomenclature) has been mapped on chromosome 6 (17) in a region (6q21–23) (18) that is involved in chromosomal aberrations found in both myeloid and lymphoid neoplasms (19). Second, in some of these tumors, amplifications (20) and rearrangements (8) of the c-*myb* locus have been found. However, a definitive assessment of the frequency, type, and significance of these genetic abnormalities is severely hampered by the incomplete characterization of the c-*myb* gene. In particular, in human, as well as in various animal species, the characterization of the genomic locus is limited to v-*myb*-homologous sequences, accounting for only approximately one-third of the 3.8-kilobase (kb) c-*myb* mRNA.

In an effort to complete the structural and functional characterization of the c-*myb* gene and to define its role in normal and leukemic cells, we isolated and determined the nucleotide sequence of human c-*myb* cDNA clones containing the entire c-*myb* coding domain and 5′ and 3′ noncoding sequences. These clones allow the preliminary characterization of the human c-*myb* genomic locus, which appears to be significantly larger (≈40 kb) than previously reported (21, 22). Determining the nucleotide sequence of c-*myb* cDNA allowed us to predict the amino acid sequence and some structural and functional characteristics of the c-*myb* protein.

MATERIALS AND METHODS

Isolation of c-*myb* cDNA and Genomic Clones. A T-lymphoma (Fro 2.2) cDNA library (a gift from D. Littman; ref. 23) was screened by plaque-hybridization using a 2.6-kb *Eco*RI fragment derived from a genomic clone (a gift of D. Leprince and D. Stehelin; ref. 21). Five hundred thousand phage plaques were screened by the plaque-hybridization method (24). Sixteen positive clones were isolated and analyzed by restriction enzyme analysis. The longest c-*myb* cDNA (λCM8) and different portions of other phages were subcloned into the vector plasmid pGEM3 (Promega Biotec, Madison, WI). One million λ Charon 4A bacteriophage plaques of a genomic library constructed from human placental DNA (25) were screened by hybridization, using the λCM8 cDNA clone as probe.

Nucleotide Sequence Analysis. DNA sequence analysis was performed on restriction fragments derived from the λCM8 phage insert and subcloned into the pGEM3 plasmid vector (Promega Biotec). The entire sequence analysis was per-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: AMV, avian myeloblastosis virus; kb, kilobase(s).

formed on both strands, and 90% of the entire cDNA sequence was independently determined by two different methods, Gemseq "transcript" and "double strand (ds)" sequencing systems, as recommended by Promega Biotec. Nucleotide and amino acid sequences were analyzed using an alignment program (26) modified by P. R. Smith and M. Plotnick and the Diagon program of Staden (27).

RESULTS

Isolation and Nucleotide Sequence Analysis of Human c-*myb* cDNA Clones. A cDNA library prepared using mRNA from the human T-cell lymphoma line Fro 2.2 (23), which expresses a 3.8-kb c-*myb* mRNA (data not shown), was screened by hybridization with a v-*myb*-homologous fragment of the human c-*myb* gene (clone p178, ref. 21). Subsequent rounds of screening, carried out using fragments of the newly isolated cDNA clones as probes, yielded a total of 16 clones. Restriction enzyme analysis of the isolated clones indicated that inserts of various lengths were overlapping on 3.4 kb of sequence and that one clone, λCM8, contained the entire sequence. This clone was chosen for nucleotide sequence analysis, and several portions of the analysis were repeated on overlapping clones.

The nucleotide sequence of the λCM8 insert (Fig. 1) contains a single long open reading frame that begins at the first base of the sequence and extends to an in-frame TGA termination codon at position 2034. An in-frame ATG initiation codon is found at position 114. The termination codon is followed by a 3′ untranslated region, which in clone λCM8 spans ≈1200 nucleotides. A poly(A) tail of ≈140 nucleotides is found at the 3′ end of the clone. Because the open reading frame extends upstream from the ATG codon found at position 114, we cannot formally exclude the possibility that this codon encodes an internal methionine residue. However, several observations argue against this possibility and indicate that the ATG at position 114 represents the translation start site of the human c-*myb* protein: (*i*) this ATG is surrounded by sequences that match the consensus sequence (CCRCCATGG) for initiation of translation in eukaryotes (28); (*ii*) the reading frame opened by this ATG would code for a protein of 640 amino acids with an approximate molecular mass of 72 kDa, in good agreement with that reported for chicken c-*myb* protein (5); and (*iii*) the sequence

```
GGCGGCAGCGCCCTGCCCACGCCCGCGGGAGGCACGCAGCCAGGCGGCGGGCAGCGGGAGGCGGCAGCCCCGGTGCTCCCCCGCGGCTCTCGGCGGAGCCCCGC   100
CGCCCGCCGCGCCATGGCCCGAAGACCCCGGCACAGCATATATAGCAGTGACGAGGATGATGAGGACTTTGAGATGTGTGACCATGACTATGATGGGCTG   200
          MetAlaArgArgProArgHisSerIleTyrSerSerAspGluAspAspGluAspPheGluMetCysAspHisAspTyrAspGlyLeu

CTTCCCAAGTCTGGAAAGCGTCACTTGGGGAAAACAAGGTGGACCCGGGAAGAGGATGAAAAACTGAAGAAGCTGGTGGAACAGAATGGAACAGATGACT   300
LeuProLysSerGlyLysArgHisLeuGlyLysThrArgTrpThrArgGluGluAspGluLysLeuLysLysLeuValGluGlnAsnGlyThrAspAspT

GGAAAGTTATTGCCAATTATCTCCCGAATCGAACAGATGTGCAGTGCCAGCACCGATGGCAGAAAGTACTAAACCCTGAGCTCATCAAGGGTCCTTGGAC   400
rpLysValIleAlaAsnTyrLeuProAsnArgThrAspValGlnCysGlnHisArgTrpGlnLysValLeuAsnProGluLeuIleLysGlyProTrpTh

CAAAGAAGAAGATCAGAGAGTGATAGAGCTTGTACAGAAATACGGTCCGAAACGTTGGTCTGTTATTGCCAAGCACTTAAAGGGGAGAATTGGAAAACAA   500
rLysGluGluAspGlnArgValIleGluLeuValGlnLysTyrGlyProLysArgTrpSerValIleAlaLysHisLeuLysGlyArgIleGlyLysGln

TGTAGGGAGAGGTGGCATAACCACTTGAATCCAGAAGTTAAGAAAACCTCCTGGACAGAAGAGGAAGACAGAATTATTTACCAGGCACACAAGAGACTGG   600
CysArgGluArgTrpHisAsnHisLeuAsnProGluValLysLysThrSerTrpThrGluGluGluAspArgIleIleTyrGlnAlaHisLysArgLeuG

GGAACAGATGGGCAGAAATCGCAAAGCTACTGCCTGGACGAACTGATAATGCTATCAAGAACCACTGGAATTCTACAATGCGTCGGAAGGTCGAACAGGA   700
lyAsnArgTrpAlaGluIleAlaLysLeuLeuProGlyArgThrAspAsnAlaIleLysAsnHisTrpAsnSerThrMetArgArgLysValGluGlnGl

AGGTTATCTGCAGGAGTCTTCAAAAGCCAGCCAGCCAGCAGTGGCCACAAGCTTCCAGAAGAACAGTCATTTGATGGGTTTTGCTCAGGCTCCGCCTACA   800
uGlyTyrLeuGlnGluSerSerLysAlaSerGlnProAlaValAlaThrSerPheGlnLysAsnSerHisLeuMetGlyPheAlaGlnAlaProProThr

GCTCAACTCCCTGCCACTGGCCAGCCCACTGTTAACAACGACTATTCCTATTACCACATTTCTGAAGCACAAAATGTCTCCAGTCATGTTCCATACCCTG   900
AlaGlnLeuProAlaThrGlyGlnProThrValAsnAsnAspTyrSerTyrTyrHisIleSerGluAlaGlnAsnValSerSerHisValProTyrProV

TAGCGTTACATGTAAATATAGTCAATGTCCCTCAGCCAGCTGCCGCAGCCATTCAGAGACACTATAATGATGAAGACCCTGAGAAGGAAAAGCGAATAAA  1000
alAlaLeuHisValAsnIleValAsnValProGlnProAlaAlaAlaAlaIleGlnArgHisTyrAsnAspGluAspProGluLysGluLysArgIleLy

GGAATTAGAATTGCTCCTAATGTCAACCGAGAATGAGCTAAAAGGACAGCAGGTGCTACCAACACAGAACCACACATGCAGCTACCCCGGGTGGCACAGC  1100
sGluLeuGluLeuLeuLeuMetSerThrGluAsnGluLeuLysGlyGlnGlnValLeuProThrGlnAsnHisThrCysSerTyrProGlyTrpHisSer

ACCACCATTGCCGACCACACCAGACCTCATGGAGACAGTGCACCTGTTTCCTGTTTGGGAGAACACCACTCCACTCCATCTCTGCCAGCGGATCCTGGCT  1200
ThrThrIleAlaAspHisThrArgProHisGlyAspSerAlaProValSerCysLeuGlyGluHisHisSerThrProSerLeuProAlaAspProGlyS

CCCTACCTGAAGAAAGCGCCTCGCCAGCAAGGTGCATGATCGTCCACCAGGGCACCATTCTGGATAATGTTAAGAACCTCTTAGAATTTGCAGAAACACT  1300
erLeuProGluGluSerAlaSerProAlaArgCysMetIleValHisGlnGlyThrIleLeuAspAsnValLysAsnLeuLeuGluPheAlaGluThrLe

CCAATTTATAGATTCTTTCTTAAACACTTCCAGTAACCATGAAAACTCAGACTTGGAAATGCCTTCTTTAACTTCCACCCCCCTCATTGGTCACAAATTG  1400
uGlnPheIleAspSerPheLeuAsnThrSerSerAsnHisGluAsnSerAspLeuGluMetProSerLeuThrSerThrProLeuIleGlyHisLysLeu

ACTGTTACAACACCATTTCATAGAGACCAGACTGTGAAAACTCAAAAGGAAAATACTGTTTTTAGAACCCCAGCTATCAAAAGGTCAATCTTACAAAGCT  1500
ThrValThrThrProPheHisArgAspGlnThrValLysThrGlnLysGluAsnThrValPheArgThrProAlaIleLysArgSerIleLeuGluSerS

CTCCAAGAACTCCTACACCATTCAAACATGCACTTGCAGCTCAAGAAATTAAATACGGTCCCCTGAAGATGCTACCTCAGACACCCTCTCATCTAGTAGA  1600
erProArgThrProThrProPheLysHisAlaLeuAlaAlaGlnGluIleLysTyrGlyProLeuLysMetLeuProGlnThrProSerHisLeuValGl

AGATCTGCAGGATGTGATCAAACAGGAATCTGATGAATCTGGATTTGTTGCTGAGTTTCAAGAAAATGGACCACCCTTACTGAAGAAAATCAAACAAGAG  1700
uAspLeuGlnAspValIleLysGlnGluSerAspGluSerGlyPheValAlaGluPheGlnGluAsnGlyProProLeuLeuLysLysIleLysGlnGlu

GTGGAATCTCCAACTGATAAATCAGGAAACTTCTTCTGCTCACACCACTGGGAAGGGGACAGTCTGAATACCCAACTGTTCACGCAGACCTCGCCTGTGC  1800
ValGluSerProThrAspLysSerGlyAsnPhePheCysSerHisHisTrpGluGlyAspSerLeuAsnThrGlnLeuPheThrGlnThrSerProValA

GAGATGCACCGAATATTCTTACAAGCTCCGTTTTAATGGCACCAGCATCAGAAGATGAAGACAATGTTCTCAAAGCATTTACAGTACCTAAAAACAGGTC  1900
rgAspAlaProAsnIleLeuThrSerSerValLeuMetAlaProAlaSerGluAspGluAspAsnValLeuLysAlaPheThrValProLysAsnArgSe

CCTGGCGAGCCCCTTGCAGCCTTGTAGCAGTACCTGGGAACCTGCATCCTGTGGAAAGATGGAGGAGCAGATGACATCTTCCAGTCAAGCTCGTAAATAC  2000
rLeuAlaSerProLeuGlnProCysSerSerThrTrpGluProAlaSerCysGlyLysMetGluGluGlnMetThrSerSerSerGlnAlaArgLysTyr

GTGAATGCATTCTCAGCCCGGACGCTGGTCATGTGAGACATTTCCAGAAAAGCATTATGGTTTTCAGAACAGTTCAAGTTGACTTGGGATATATCATTCC  2100
ValAsnAlaPheSerAlaArgThrLeuValMet***
TCAACATGAAACTTTTCATGAATGGGAGAAGAACCTATTTTTGTTGTGGTACAACAGTTGAGAGCACCACCAAGTGCATTTAGTTGAATGAAGTCTTCTT  2200
GGATTTCACCCAACTAAAAGGATTTTTAAAAATAAATAACAGTCTTACCTAAATTATTAGGTAATGAATTGTAGCCAGTTGTTAATATCTTAATGCAGAT  2300
TTTTTTAAAAAAAAACATAAAATGATTTATCTGGTATTTTAAAGGATCCAACAGATCAGTATTTTTTCCTGTGATGGGTTTTTTGAAATTTGACACATTA  2400
AAAGGTACTCCAGTATTTCACTTTTCTCGATCACTAAACATATGCATATATTTTTAAAAATCAGTAAAAGCATTACTCTAAGTGTAGACTTAATACCATG  2500
TGACATTTAATCCAGATTGTAAATGCTCATTTATGGTTAATGACATTGAAGGTACATTTATTGTACCAAACCATTTTATGAGTTTTCTGTTAGCTTGCTT  2600
TAAAAATTATTACTGTAAGAAATAGTTTTATAAAAAATTATATTTTTATTCAGTAATTTAATTTTGTAAATGCCAAATGAAAAACGTTTTTTGCTGCTAT  2700
GGTCTTAGCCTGTAGACATGCTGCTAGTATCAGAGGGGCAGTAGAGCTTGGACAGAAAGAAAAGAAACTTGGTGTTAGGTAATTGACTATGCACTAGTAT  2800
TTCAGACTTTTTAATTTTATATATATATACATTTTTTTTCCTTCTGCAATACATTTGAAAACTTGTTTGGGAGACTCTGCATTTTTTATTGTGGTTTTTT  2900
TGTTATTGTTGGTTTATACAAGCATGCGTTGCACTTCTTTTTTTGGGAGATGTGTGTTGTTCATGTTCTATGTTTTGTTTTGTGTGTAGCCTGACTGTTTT  3000
ATAATTTGGGAGTTCTCGATTTGATCCGCATCCCCTGTGGTTTCTAAGTGTATGGTCTCAGAACTGTTGCATGGATCCTGTGTTTGCAACTGGGGAGACA  3100
GAAACTGTGGTTGATAGCCAGTCACTGCCTTAAGAACATTTGATGCAAGATGGCCAGCAGTCAACTTTTTGAGATATGACGGTGTACTTACTGCCTTGTAG  3200
CAAAATAAAGATGTGCCCTTATTTTAAAAAAAAAAAAAA  3239
```

FIG. 1. Nucleotide sequence of the human c-*myb* cDNA clone λLMC8. The translation of this sequence from the putative initiation codon at position 114 to the termination codon (***) is shown below the nucleotide sequence. The poly(A)-addition signal at position 3204 is underlined.

reported in Fig. 1 is highly homologous to the one reported for mouse c-*myb* cDNA clones (29, 30), and the murine sequences contain initiation, termination, and poly(A)-addition signals located at positions analogous to the ones identified for the human sequence.

Analysis of evolutionary conservation of *myb* cDNA nucleotide sequences between human and mouse reveals a high degree of overall homology (80%). In particular, the coding domain displays 87% homology (see below for an analysis of predicted amino acid products), and the 5′ and 3′ untranslated regions, 82% and 72% homology, respectively. Within the long 3′ untranslated region, segments of relatively high homology (e.g., nucleotides 2610–2810, 84% homology) are interrupted by areas of relative divergence (e.g., nucleotides 2990–3190, 55% homology), suggesting that the highly conserved regions may contain functionally important domains for transcriptional and/or posttranscriptional regulation of c-*myb* expression.

Analysis and Evolutionary Conservation of the Predicted Human c-*myb* Protein. The amino acid sequence predicted by the coding portion of the c-*myb* cDNA is shown in Fig. 2 *Left* (standard one-letter amino acid symbols are given). Computer-assisted analysis of this sequence revealed a series of potentially important features: (*i*) three tandemly organized direct repeats of 51–52 amino acids (positions 38–89, 90–141, and 142–192) in the NH_2-terminal portion of the protein; (*ii*) a generally hydrophobic profile (not shown); and (*iii*) an even distribution of acidic and basic amino acids, with the exception of a predominantly basic region toward the NH_2-terminal region where the three repeats are located (22% arginine plus lysine between residues 38 and 192; 8% in the remainder of the protein). These features are found to be generally conserved when the human sequence is aligned (Fig. 2) with the corresponding sequences of mouse (29) and chicken (31) c-*myb* protein. The degree of evolutionary conservation is high overall, but it varies for different parts of the protein. Residues 1–200 appear to be particularly conserved (99.5% homology between mouse and human). A second region of high conservation is present at positions 261–410 (96% homology between mouse and human). Regions of relatively higher divergence are residues 201–260 and the COOH-terminal region of the molecule. When the human, mouse, and chicken sequences are compared to the v-*myb*-homologous portion of the *Drosophila* c-*myb* protein (32) (Fig. 2 *Right*), the NH_2-terminal portion containing the three repeats is found to be the only one conserved during evolution. These structural and topographical homologies indicate that this region of the c-*myb* protein is a critical functional domain.

Organization of the c-*myb* Gene: Exons Define a 40-kb Genomic Locus. In order to isolate genomic clones containing the human c-*myb* gene, a recombinant library made from normal human placental DNA was screened using the 3.4-kb λCM8 insert as a probe. Five clones overlapping over 60-kb of genomic DNA were isolated (G0, G2, G3, G5, and G6; Fig.

```
hu-c-myb MARRPRHSIY SSDEDDEDFE MCDHDYDGLL PKSGKRHLGK TRWTREEDEK LKKLVEQNGT 60
mu-c-myb ---------- --------I- ---------- ---------- ---------- ----------
ch-c-myb ---------- ---D-E--V- -Y-------- --A------- ---------- ----------

hu-c-myb DDWKVIANYL PNRTDVQCQH RWQKVLNPEL IKGPWTKEED QRVIELVQKY GPKRWSVIAK 120
mu-c-myb ---------- ---------- ---------- ---------- ---------- ----------
ch-c-myb E-------SF- ---------- ---------- ---------- ---------- ----------

hu-c-myb HLKGRIGKQC RERWHNHLNP EVKKTSWTEE EDRIIYQAHK RLGNRWAEIA KLLPGRTDNA 180
mu-c-myb ---------- ---------- ---------- ---------- ---------- ----------
ch-c-myb ---------- ---------- ---------- ---------- ---------- ----------

hu-c-myb IKNHWNSTMR RKVEQEGYLQ ESSKASQPAV ATSFQKNSHL MGFAQAPPTA QLPATGQPTV 240
mu-c-myb ---------- ---------- -P-----TP- -------N-- ---GH-S-PS --SPS--SS-
ch-c-myb ---------- ---------- -----GL-SA T-G---S--- -A--HN--AG P--GA--APL

hu-c-myb NNDYSYYHIS EAQNVSSHVP YPVALHVNIV NVPQPAAAAI QRHYNDEDPE KEKRIKELEL 300
mu-c-myb -SE-P----A ----I----- ------A--- ---------- ---------- ----------
ch-c-myb GS--P----A -P---PGQI- ------V--V ---------- ----N----- ----------

hu-c-myb LLMSTENELK GQQVLPTQNH TCSYPGWHST TIADHTRPHG DSAPVSCLGE* HHSTPSLPAD 360
mu-c-myb ---------- ---A------ ---------- S-V-Q----- ----------* --A-------
ch-c-myb ---------- ---A------ -AN------- -V--N--TS- -N--------H --C---P-V-

hu-c-myb PGSLPEESAS PARCMIVHQG TILDNVKNLL EFAETLQFID SFLNTSSNHE NSDLEMPSLT 420
mu-c-myb ---------- ---------- ---------- ---------- ---------- S-G-DA-T-P
ch-c-myb H-C------- ---------S N--------- -------L-- ---------- -LN-DN-A--

hu-c-myb STPLIGHKLT VTTPFHRDQT VKTQKENTVF RTPAIKRSIL ESSPRTPTPF KHALAAQEIK 480
mu-c-myb ---------- ***-C*---- -------SI- ---------- ---------- ----------
ch-c-myb ---VC---MS ------R--- F-------H-- ---------- ---------- -N--------

hu-c-myb YGPLKMLPQT PSHLVEDLQD VIKQESDESG FVAEFQENGP PLLKKIKQEV ESPTDKSGNF 540
mu-c-myb ---------- ---A------ ---Q------ I-R----S-- --------A- ----E-----
ch-c-myb ---------- -T-------- -----SE--A I--GLH-S-- ---------- ------A---

hu-c-myb FCSHHWEGDS LNTQLFTQTS PVRDAPNILT SSVLMAPASE DEDNVLKAFT VPKNRSLASP 600
mu-c-myb ---N--AEN- -S----S-A- --A------- -----T-V-- ---------- -----P-VG-
ch-c-myb ---N----EN -------HA- TME-V--L-- --I-KM-V-- E-GSFH--A --R--P----

hu-c-myb LQPCSSTWEP ASCGKMEEQM TSSSQARKYV NAFSARTLVM                        660
mu-c-myb -----GA--- -----T-D-- -A-GP----- ----------
ch-c-myb --HLNNA-GS ---E-T-D-- ALTD-----M A--PT-----
```

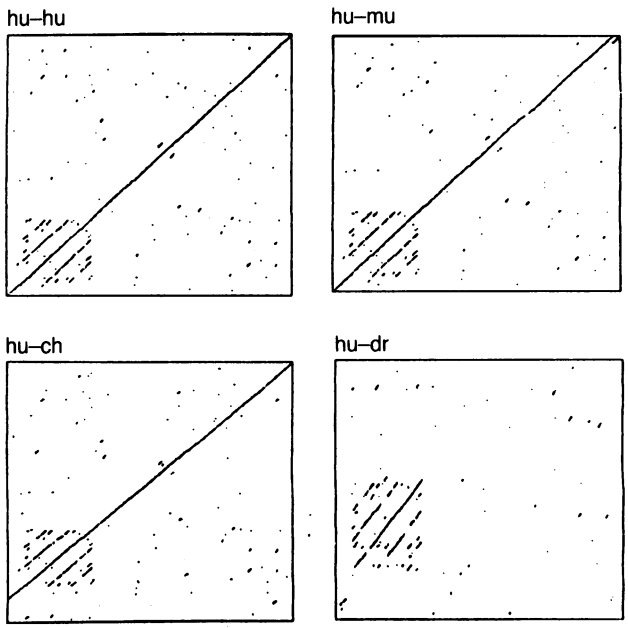

FIG. 2. (*Left*) Amino acid sequence of the predicted human (hu) c-*myb* protein and alignment with its murine (mu) (29) and chicken (ch) (31) counterparts. Residues that are different among species are indicated. Asterisks indicate the absence of a given residue. Dashes (−) indicate identity. Horizontal arrows underline the three tandem repeats (see text). (*Right*) Dot matrix analysis (27) for internal repeats within the human c-*myb* protein (hu–hu) and comparative analysis between human and mouse (hu–mu), chicken (hu–ch), and *Drosophila* (hu–dr). Segments of 11 amino acids were compared sequentially with each 11-residue-long segment of the protein. A dot was placed on the matrix at the appropriate position when the total mutation data matrix score for the comparison was >20.

3). By use of different 5′ and 3′ fragments of the λCM8 insert as probes in Southern blot (33) experiments (data not shown), hybridizing fragments were identified and positioned within the genomic clones (Fig. 3). In order to confirm that the putative initiation and poly(A)-addition sites identified in the cDNA clone were also present within the genomic clones, the regions of the genomic clones hybridizing to the 5′- and 3′-terminal fragments of the cDNA clone (Fig. 3) were sequenced (data not shown). Within these regions, the positions of the initiation (ATG) and poly(A)-addition (AATAAA) sites is marked on Fig. 3. This analysis defined the approximate boundaries of the c-*myb* transcriptional domain and provided a preliminary characterization of the intron/exon arrangement of this gene (Fig. 3). Based on the analysis of cDNA-hybridizing fragments and restriction sites present both in genomic and in cDNA clones, a minimum of 15 exons have been identified. However, further subdivision in additional exons and introns, which may escape restriction enzyme analysis, can be anticipated. Since the 5′ terminus of the RNA is not represented in the λCM8 cDNA insert, it is presently impossible to define the limits of the first 5′ exon beyond a tentative mapping within the 5′ 2.3-kb *Eco*RI fragment (Fig. 3). This position is analogous to the one identified for the transcription initiation site of the murine c-*myb* gene (21). These data indicate that the human c-*myb* gene spans ≈40 kb of the human genome and therefore extends significantly beyond the limits of the previously identified v-*myb*-related domains (21, 22) to include additional 5′ and 3′ coding and noncoding exons.

DISCUSSION

The isolation of human c-*myb* cDNA clones has allowed us to determine the nucleotide sequence of the c-*myb* mRNA, to predict the amino acid sequence of the c-*myb* protein, and to preliminarily define the boundaries and the organization of the human c-*myb* locus.

Our longest cDNA clone is 3.4 kb long, approximately 0.4 kb shorter than the 3.8-kb full-length c-*myb* mRNA detected by blot hybridization of electrophoretically fractionated RNA (data not shown). This 0.4 kb may be accounted for by additional nucleotides in the poly(A) tail of mRNA and/or by additional 5′ sequence. Within the cloned sequence, the cDNA clone contains a long open reading frame flanked by 5′ and 3′ noncoding regions. We have already discussed several arguments that indicate that translation is initiated by the methionine codon at nucleotide 114, but this cannot be proved until the sequence further upstream from this ATG codon is known. Several observations suggest that the 5′ region of this gene is structurally and functionally complex and requires additional detailed analysis. Preliminary analysis of the 5′ region of the mouse c-*myb* gene suggests that transcription may be initiated at multiple sites within a segment of the murine c-*myb* locus (21) that is homologous by sequence and topography to the one that we tentatively identified as the first c-*myb* exon (see Fig. 3). Analogous studies on chicken c-*myb* cDNA clones have reported a sequence that extends further upstream from the open reading frame reported in Fig. 3 and significantly diverges from both murine and human sequences (31). Finally, our and others' (36) preliminary studies indicate that additional c-*myb* mRNA species may be present, most likely originating through a differential-splicing mechanism. These observations could account for the size heterogeneity reported for c-*myb* mRNA and for the occasional observation of two mRNA species in some cell types (22). While the present report provides a preliminary characterization of the human c-*myb* RNA coding domain in lymphoid cells, additional studies are required to fully elucidate the different patterns of expression of the c-*myb* locus in different cells.

The nucleotide sequence of the cDNA clones allowed us to predict the amino acid sequence of the putative human c-*myb* protein. Given the high degree of evolutionary conservation of this protein, several features regarding the properties and structural features of the murine c-*myb* protein (22) can be confirmed for its human counterpart. These include (*i*) a general hydrophobic profile; (*ii*) an even distribution of acidic and basic amino acids, with the exception of a basic region

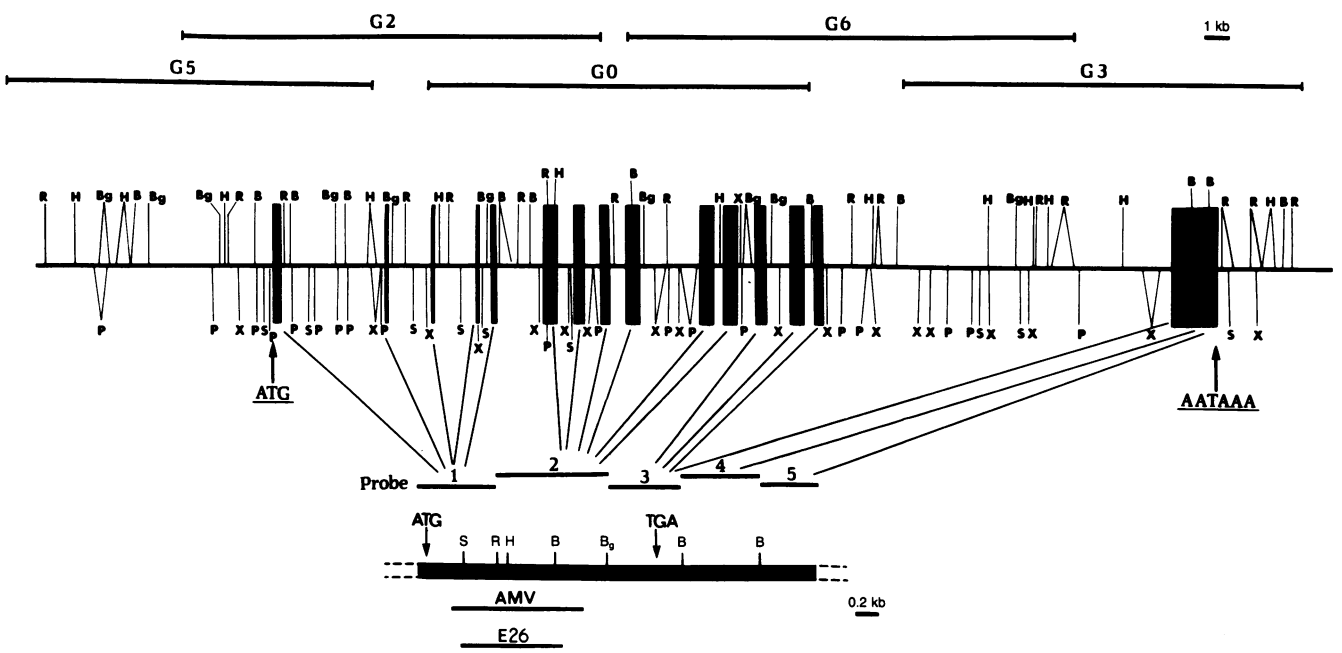

FIG. 3. Genomic organization of the human c-*myb* locus. Restriction maps of clone inserts spanning the human genomic locus (clones G0, G2, G3, G5, and G6, at top) and cDNA clone (λCM8, at bottom) are shown. cDNA probes 1–5 used to identify *myb* coding sequences in the genomic clones are also illustrated. Black boxes in the genomic map indicate the minimum number of exons identified by hybridization and by the presence of restriction sites found in the cDNA clone. The length of each exon is arbitrarily determined in order to approximately accommodate the length of the corresponding probe. Below the map of the cDNA clone, the regions of homology with AMV and murine leukemia virus E26 are indicated. Restriction enzyme sites: R, *Eco*RI; B, *Bam*HI; Bg, *Bgl* II; H, *Hin*dIII; P, *Pst* I; S, *Sst* I; X, *Xba* I.

toward the NH_2 terminus, containing the three tandem repeats; (*iii*) a secondary structure involving a high proportion of α-helix; and (*iv*) a tertiary structure tending toward an overall globular configuration. Although presumably located in the nucleus, as has been demonstrated for its avian counterpart (6), the human c-*myb* protein does not contain domains significantly homologous to nuclear proteins encoded by other protooncogenes—namely, c-*myb*, N-*myc*, or c-*fos*. It also does not appear to contain the consensus sequence for nuclear localization that has been found in polyoma and simian virus 40 large tumor (T) antigens (34). One feature of the c-*myb* protein that may be critical for function is the evolutionarily conserved, tandemly repeated sequences at the NH_2 terminus of the protein. The addition of the human sequences to the comparative analysis of various species and the observation that these sequences are the ones selectively conserved from *Drosophila* to man lend further support to previous suggestions (30) that these sequences may represent the DNA-binding domain of the protein. While this work was in progress, experimental studies have documented this hypothesis for the chicken c-*myb* protein (35).

Finally, we have preliminarily identified the boundaries and the organization of the human genomic c-*myb* locus. Previous characterizations were limited to the identification of the v-*myb*-related domain (14 kb; ref. 21) and of additional sequences identified by their hybridization to *myb* mRNA (30 kb, ref. 22). Our data indicate the presence of additional 5′ and 3′ coding and noncoding exons, suggesting that the c-*myb* gene spans >40 kb of the human genome.

We are grateful to Drs. D. Leprince and D. Stehelin for the gift of the human c-*myb* probe which was instrumental in generating the cDNA clones. We are indebted to William Ference for excellent technical assistance, to Francis G. Kern and Luigi Lania for critical review of the manuscript, and to Diane Nazario for careful editing of the manuscript. This work was supported by National Institutes of Health Grant CA16239 and by a Presidential Award from the Leukemia Society of America. B.M. is supported by a Fellowship from Associazione Italiana per la Ricerca sul Cancro. L.C.K. is supported by the Medical Scientist Training Program. R.D.-F. is a Leukemia Society of America Scholar.

1. Roussel, M., Saule, S., Lagrou, C., Rommens, C., Beug, H., Graf, T. & Stehelin, D. (1979) *Nature (London)* **281,** 452–455.
2. Souza, L. M., Strommer, J. N., Hillyard, R. L., Komaromy, M. C. & Baluda, M. A. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 5177–5181.
3. Bergmann, D. G., Souza, L. M. & Baluda, M. A. (1981) *J. Virol.* **40,** 450–455.
4. Boyle, W. J., Lipsick, S. J., Reddy, E. P. & Baluda, M. A. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 2834–2838.
5. Klempnauer, K. H., Ramsay, G., Bishop, J. M., Moscovici, M. G., Moscovici, C., McGrath, J. P. & Levinson, A. D. (1983) *Cell* **33,** 345–355.
6. Klempnauer, K. H., Symonds, G., Evan, G. I. & Bishop, M. J. (1984) *Cell* **37,** 537–547.
7. Alitalo, K., Winqvist, R., Lin, C. C., de la Chapelle, A., Schwab, M. & Bishop, M. J. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 4534–4538.
8. Yokota, J., Yokota, Y. T., Battifora, H., Le Fevre, C. & Cline, M. J. (1986) *Science* **231,** 261–265.
9. Gonda, T. J. Sheiness, D. K. & Bishop, J. M. (1982) *Mol. Cell. Biol.* **2,** 617–624.
10. Westin, E. H., Gallo, R. C., Arya, S. K., Eva, A., Souza, L. M., Baluda, M. A., Aaronson, S. A. & Wong-Staal, F. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 2194–2198.
11. Craig, R. W. & Bloch, A. (1984) *Cancer Res.* **44,** 442–446.
12. Moscovici, C. (1975) *Curr. Top. Microbiol. Immunol.* **71,** 79–101.
13. Duesberg, P. H., Bister, K. & Moscovici, C. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 5120–5124.
14. Mushinsky, J. F., Potter, M., Bauer, S. R. & Reddy, E. P. (1983) *Science* **220,** 795–798.
15. Shen-Ong, G. L., Potter, M., Mushinsky, J. F., Lavu, S. & Reddy, E. P. (1984) *Science* **226,** 1077–1080.
16. Shen-Ong, G. L., Morse, H. C., III, Potter, M. V. & Mushinsky, J. F. (1986) *Mol. Cell. Biol.* **6,** 380–392.
17. Dalla-Favera, R., Franchini, G., Martinotti, S., Wong-Staal, F., Gallo, R. C. & Croce, C. M. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 4714–4717.
18. Harper, M. E., Franchini, G., Loye, J., Simon, M. I., Gallo, R. C. & Wong-Staal, F. (1983) *Nature (London)* **304,** 169–171.
19. Mitelman, F. (1983) *Catalogue of Chromosome Aberration in Cancer*, ed. Klinger, H. P. (Karger, Basel).
20. Pelicci, P. G., Lanfrancone, L., Brathwaite, M. D., Wolman, S. R. & Dalla-Favera, R. (1984) *Science* **224,** 1117–1121.
21. Leprince, D., Saule, S., de Taisne, C., Gegonne, A., Begue, A., Rocjo, M. & Stehelin, D. (1983) *EMBO J.* **2,** 1073–1078.
22. Franchini, G., Wong-Staal, F., Baluda, M. A., Lengel, C. & Tronick, S. R. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 7385–7389.
23. Maddon, P. J., Littman, D. R., Godfrey, M., Maddon, D. E., Chess, L. & Axel, R. (1985) *Cell* **42,** 93–104.
24. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
25. Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K. & Efstratiadis, A. (1978) *Cell* **15,** 687–701.
26. Pustell, J. & Kafatos, F. C. (1982) *Nucleic Acids Res.* **10,** 51–59.
27. Staden, R. (1982) *Nucleic Acids Res.* **10,** 2951–2961.
28. Kozak, C. (1984) *Nucleic Acids Res.* **12,** 857–872.
29. Bender, T. P. & Kuehl, W. M. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 3204–3208.
30. Gonda, T. J., Gough, N. M., Dunn, A. R. & de Blaquiere, J. (1985) *EMBO J.* **4,** 2003–2008.
31. Rosson, D. & Reddy, E. P. (1986) *Nature (London)* **319,** 604–606.
32. Katzen, A. L., Kornberg, T. B. & Bishop, J. M. (1985) *Cell* **41,** 449–456.
33. Southern, E. M. (1975) *J. Mol. Biol.* **98,** 503–517.
34. Richardson, W. D., Roberts, B. L. & Smith, A. E. (1986) *Cell* **44,** 77–85.
35. Klempnauer, K.-H., Bonifer, C. & Sippel, A. E. (1986) *EMBO J.* **5,** 1903–1911.
36. Slamon, D. J., Boone, T. C., Murdock, D. C., Keith, D. E., Press, M. F., Larson, R. A. & Souza, L. M. (1986) *Science* **233,** 347–351.