MOLECULAR AND CELLULAR BIOLOGY, May 1989, p. 2114–2123
0270-7306/89/052114-10$02.00/0
Copyright © 1989, American Society for Microbiology
Vol. 9, No. 5

Alternative Processing of mRNAs Encoding Mammalian Chromosomal High-Mobility-Group Proteins HMG-I and HMG-Y

KENNETH R. JOHNSON,[1] DONALD A. LEHN,[2] AND RAYMOND REEVES[1,2]*

Programs in Genetics and Cell Biology[1] and Biochemistry and Biophysics,[2] Washington State University, Pullman, Washington 99164

Received 19 December 1988/Accepted 13 February 1989

The high-mobility-group protein HMG-I is a well-characterized nonhistone chromosomal protein that is preferentially expressed in rapidly dividing cells, binds to A · T-rich regions of DNA in vitro, and has been localized to particular regions of mammalian metaphase chromosomes. We isolated eight cDNA clones encoding HMG-I and its isoform HMG-Y from a human Raji cell cDNA library and detected blocks of nucleotide sequence rearrangements in the 5′-untranslated regions of these clones. In addition to this leader sequence variation, five of the eight cDNA clones had either a 33- or 36-base-pair in-frame deletion in their open reading frame (ORF); we found that this shortened ORF encodes the HMG-Y protein isoform. We present evidence that the 5′-untranslated-region and ORF heterogeneity of the cDNA clones is the result of alternative processing of RNA transcripts from a single functional gene. Several additional but probably nonfunctional HMG-I or HMG-Y gene copies exist in the human genome; we isolated and partially sequenced one of these pseudogenes and found that it is a processed HMG-Y retropseudogene.

The high-mobility-group protein HMG-I is a well-characterized nonhistone chromosomal protein that is preferentially expressed in undifferentiated, rapidly proliferating tissues (8–10, 18, 30). HMG-I mRNA levels are also elevated in neoplastic tissues (11). Thus, HMG-I proteins probably play an important role in cell division and may be involved in the condensation of chromosomes during metaphase (18) or in maintaining the undifferentiated state of chromatin (9, 30).

HMG-I has been shown to specifically bind to A · T-rich regions of double-stranded DNA in vitro (5, 26). This DNA-binding characteristic has led to other postulated functions for the protein in nucleosome phasing (26) and in 3′-end-processing of transcripts (23). HMG-I may also be involved in transcriptional regulation of genes containing, or in proximity to, A+T-rich regions of DNA such as are found in the 3′-untranslated regions (3′-UTRs) of several lymphokine genes (21) and in the spacer regions of rRNA genes (32). In these instances, HMG-I (21) or an HMGI-like protein (32) might affect gene regulation by changing the chromatin conformation and hence the accessibility of A · T-rich regions of DNA to binding by other transcription factors. In vitro binding of HMG-I to the A · T-rich regions of the bovine interleukin-2 cDNA 3′-UTR changes its DNA conformation and thermal stability (16).

The A · T-specific DNA ligands Hoechst 33258 and distamycin A effectively compete in vitro with HMG-I for binding to A · T-rich mouse satellite DNA (unpublished data). We recently showed that HMG-I proteins localize to G(Q) bands, centromeres, and telomeres of mammalian metaphase chromosomes (unpublished data). These results strongly implicate the HMG-I and HMG-Y proteins as in vivo structural components involved in the condensation of A · T-rich regions of mammalian chromosomes.

Efficient methods have been developed in our laboratory for separating and purifying to homogeneity large quantities of HMG-I and other members of the high-mobility-group proteins (6–8). We have shown by amino acid composition and partial peptide sequence analyses (16) that the protein designated HMG-Y by Lund et al. (18) is an isoform of the protein they designated HMG-I. We now refer to the entire family of HMG-I-like proteins as HMGI and designate the two known isoforms as HMG-I and HMG-Y. In a previous paper, we reported the isolation and sequence of a full-length murine cDNA encoding an HMGI protein (11). The deduced amino acid sequence of this murine cDNA agreed with the reported peptide fragment sequence analysis of the human placental HMG-I protein (18), except that it lacked 11 internal amino acids present in the human sequence.

In this paper we report the isolation and nucleotide sequences of eight independent HMGI cDNA clones from a human Raji cell cDNA library. Three of the cDNA clones encode a HMGI protein whose amino acid sequence is virtually identical to that of HMG-I reported by Lund et al. (17). Four of the other cDNAs encode a HMGI protein with the same internal 11-amino-acid deletion that we previously found in the murine cDNA. We found by peptide fragment sequence analysis that high-pressure liquid chromatography-purified murine HMG-Y protein contains this same 11-amino-acid deletion but is otherwise identical to HMG-I. These complementary findings in both the protein sequences and cDNA sequences confirm the previously postulated isoformic relation of the HMG-I and HMG-Y proteins (16) and cDNAs (11).

In addition to observing this HMG-I and HMG-Y protein-coding sequence dimorphism, we detected sequence heterogeneity in the 5′-UTRs of the cDNA clones. We present evidence here that the observed cDNA heterogeneity is the result of alternative processing of pre-mRNA transcripts from a single functional gene. HMGI is the first structural component of mammalian chromosomes shown to be so encoded by alternatively processed mRNAs. We also report the isolation and partial sequencing of a HMGI pseudogene and show that it is a processed HMG-Y retropseudogene, one of several, probably nonfunctional, HMGI gene copies present in the human genome.

* Corresponding author.

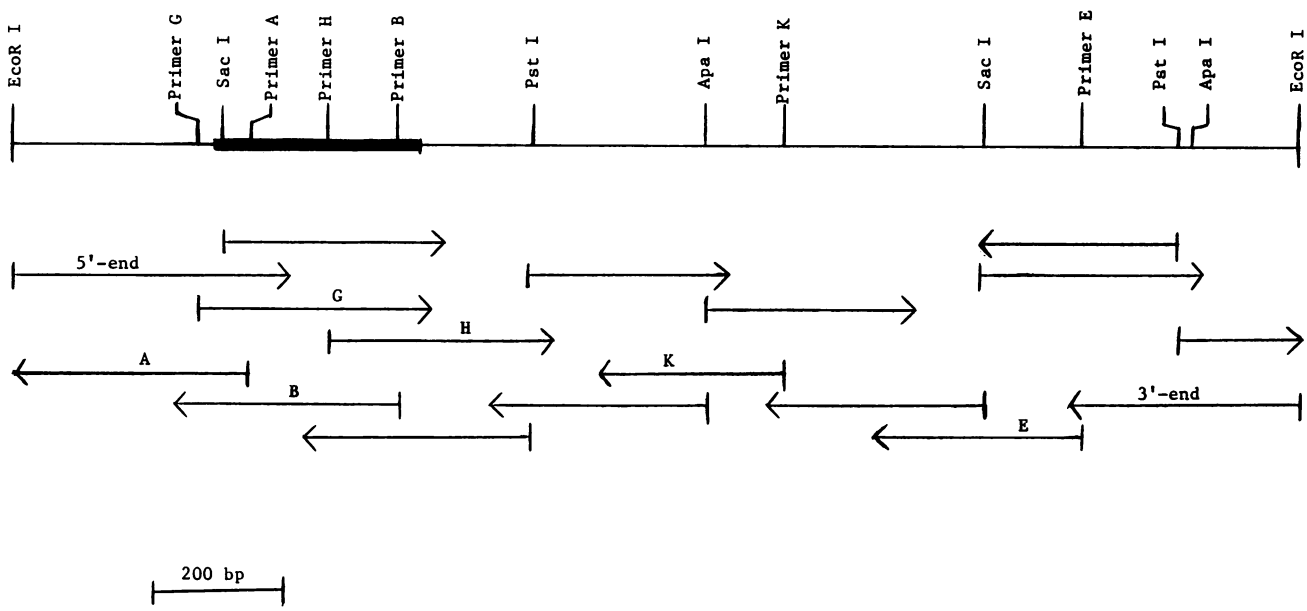

FIG. 1. Restriction enzyme map, hybridization locations of synthetic primers, and cloning strategy used to sequence the human HMGI cDNAs. The black box represents the HMGI protein-coding sequence of the cDNAs. Each of the eight cDNA clones were partially sequenced, as shown by the arrows labeled 5'-end, 3'-end, A, B, G, H, and E. The complete cDNA sequences of clones 6A and 8A were determined by using an additional primer (arrow K) and by sequencing subcloned restriction enzyme fragments (unlabeled arrows). The region 5' to primer G on the map varies among clones; it contains an *Apa*I site (not shown) in clone 6A and a *Sac*I site (not shown) in clones 1A, 2B, 3B, 7C, and 8A.

MATERIALS AND METHODS

Isolation of HMGI cDNA clones. A human lymphocyte (Raji cell) cDNA library in bacteriophage lambda *gt*10 (HL1002a; Clontech, Palo Alto, Calif.) was chosen for screening because murine and human lymphoblastic cells were shown to express high levels of HMGI mRNAs (11). Approximately 120,000 plaques were screened for HMGI DNA sequences by in situ hybridization on nitrocellulose membranes with the ^{32}P-labeled, 450-base-pair (bp) *Sac*I-to-*Pst*I fragment of murine HMGI cDNA which contains the open reading frame (ORF) (11). Radiolabeling of probe DNA and hybridization conditions were as previously described (11). Eight positively hybridizing plaques that contained different-sized cDNA inserts greater than 1,500 bp were isolated and amplified. Each of the insert cDNAs from these eight plaques was subcloned into the *Eco*RI site of a plasmid vector, Bluescript (Stratagene, La Jolla, Calif.) and the bacteriophage vector M13.mp18.

Isolation of HMGI pseudogene. About 300,000 plaques from a human leukocyte genomic DNA library in phage EMBL-3 (HL1006d; Clontech) were screened for HMGI DNA sequences by hybridization with the 2,000-bp insert from human cDNA clone 6A by using previously described procedures (11). Several positively hybridizing plaques were isolated; one was analyzed in detail. A 6,000-bp *Eco*RI-to-*Sal*I DNA fragment containing an entire human HMGI pseudogene was isolated from this plaque and subcloned into the Bluescript plasmid vector. We were unable to isolate a functional HMGI gene from this screen of the library.

DNA sequencing. Single-stranded DNAs in both orientations from each of the eight HMGI cDNAs subcloned into the *Eco*RI site of M13.mp18 were sequenced (Fig. 1, arrows 5'- end and 3'-end) by the dideoxynucleotide termination method (24) with the M13 universal primer, [^{35}S]ATP, and Sequenase (U.S. Biochemical Corp., Cleveland, Ohio). Previously described and synthesized 39-mer oligonucleotides (11) corresponding to the antisense strand of HMGI cDNA nucleotides 42 to 81 (primer A) and nucleotides 265 to 304 (primer B) were used to obtain further sequence information for all eight clones (Fig. 1, arrows A and B). On the basis of these results, three additional oligonucleotides were synthesized corresponding to the sense strand of HMGI cDNA nucleotides −35 to −16 (primer G) and nucleotides 171 to 191 (primer H) and to the antisense strand of nucleotides 1341 to 1358 (primer E); these synthetic oligonucleotides were also used as primers to further sequence all eight cDNAs (Fig. 1, arrows G, H, and E). Two of the cDNA plasmid clones (pBS6A and pBS8A) were restricted with *Apa*I, *Pst*I, and *Sac*I; cDNA fragments from these digests were individually subcloned into M13.mp18 or M13.mp19 or both. Single-stranded DNAs from these fragments were sequenced by using the M13 universal primer (Fig. 1, unlabeled arrows). Single-stranded cDNAs from clones 6A and 8A were also sequenced by using an additional synthetic oligonucleotide primer corresponding to the antisense strand of HMGI cDNA nucleotides 874 to 890 (Fig. 1, Primer K).

Double-stranded DNA from the 6,000-bp HMGI pseudogene insert subcloned into the Bluescript plasmid vector was sequenced in both directions by using the T_3 and T_7 vector primers. The 5' end of the pseudogene was further sequenced by using the synthetic HMGI cDNA primers A, B, G, and H (described above).

Peptide fragment sequencing of HMG-Y. About 200 μg of HMG-Y purified by high-pressure liquid chromatography (8) was cleaved with endoproteinase Glu-C (20 μg) in 100 mM ammonium bicarbonate (pH 7.8). The resulting fragments were separated on a Vydac C_{18} column as previously described (16). Purified peptides were then subjected to gas-phase sequence analysis (model 470A, Applied Biosystems).

Isolation of nucleic acids and blot hybridizations. Isolation of genomic DNA and total cellular RNA and procedures used for Southern and Northern (RNA) blot hybridizations

```
5'-UTR of cDNA clone 11D:  45 bp

                                                          1
  -45  GCATCCCAGCCATCACTCTTCCACCTGCTCCTTAGAGAAGGGAAGATG                3


5'-UTR of cDNA clone 10A:  135 bp

-135  GCTTTTTAAGCTCCCCTGAGCCGGTGCTGCGCTCCTCTAATTGGGACTCCGAGCCGGGGC   -76
 -75  TATTTCTGGGCTGGCGCGGCTCCAAGAAGGCCATCCCAGCCATCACTCTTCCACCTGCTC   -16
 -15  CTTAGAGAAGGGAAGATG                                             3


5'-UTR of cDNA clones 1A and 8A:  242 bp

-242  GCTTTTTAAGCTCCCCTGAGCCGGTGCTGCGCTCCTCTAATTGGGACTCCGAGCCGGGGC  -183
-182  TATTTCTGGGCTGGCGCGGCTCCAAGAAGATCCGCATTTGCTACCAGCGGCGGCCGCGCG  -123
-122  GAGCCAGGCCGGTCCTCAGCGCCCAGCACGGCTCCCGGCAACCCGGAGCGCGCACCGCAG   -63
 -62  CCGGCGGCCGAGCTCGCGCATCCCAGCCATCACTCTTCCACCTGCTCCTTAGAGAAGGGA    -3
  -2  AGATG                                                          3


5'-UTR of cDNA clones 2B, 3B, and 7C:  207 bp

                                        splice acceptor site
-207  GAGCACGCGGCGGCGGCGGTCTCTGAGCGCCTCTGCTCTCTCTCCCGGTTTCAGATCCGC  -148
-147  ATTTGCTACCAGCGGCGGCCGCGCGGAGCCAGGCCGGTCCTCAGCGCCCAGCACGGCTCC   -88
 -87  CGGCAACCCGGAGCGCGCACCGCAGCCGGCGGCCGAGCTCGCGCATCCCAGCCATCACTC   -28
 -27  TTCCACCTGCTCCTTAGAGAAGGGAAGATG                                 3


5'-UTR of cDNA clone 6A:  312 bp

-312       GAGTTGTGCACCCGGCGGAGCCGGAGGAGCCCGGCGCACCTCGCCGGGCGGCCGCT  -257
-256  CCCGCTGGAGCCGGAGCCCGAGCCCGAGCCCGAGCCCGGGCCCGGGTGAGGGGCGGGGAG  -197
-196  AGACACGGGCTGCGGCGCGGCAGGAGAGTGGGGGGTCGGGCGCCCCCCCGCAGCTCAGGAC  -137

      MetGlyValProProTyrCysProGlyTrpSerArgThrProAspLeuLysGlnSerSer
-136  ATGGGAGTCCCACCGTATTGTCCAGGCTGGTCTCGAACTCCTGACCTCAAGCAGTCCTCC   -77

      CysPheGlyIleProLysCysTrpAspTyrSerIleProAlaIleThrLeuProProAla
 -76  TGCTTTGGCATCCCAAAGTGCTGGGATTACAGCATCCCAGCCATCACTCTTCCACCTGCT   -17

      ProEnd
 -16  CCTTAGAGAAGGGAAGATG                                           3
```

FIG. 2. Nucleotide sequences of the variable 5'-UTRs of the human HMGI cDNA clones. Nucleotide numbering for all clones begins with the ATG start codon of the HMGI ORF (double underlined). The 45 single-underlined nucleotides are common to the 5'-UTRs of all eight cDNA clones. For clones 2B, 3B, and 7C, the nucleotide sequence (−172 to −154) conforming to an RNA 3'-splice acceptor consensus sequence is double underlined. For clone 6A, the deduced amino acid sequence is shown above an additional ORF (nucleotides −136 to −14) that is 5' to, but out of frame with, the ORF encoding HMGI (nucleotides 1 to 321).

were done as previously described (11). Poly(A)$^+$ RNA was selected from total RNA on oligo(dT) cellulose columns.

Primer extension. Primer extension methods were modified from those of Nissen and Friesen (19). About 50 ng of a synthetic oligonucleotide 39-mer corresponding to the antisense strand of HMGI cDNA nucleotides 42 to 81 (primer A, described above) was end labeled with ^{32}P. About 5×10^5 cpm of this labeled primer was incubated with either 20 μg of total RNA or 5 μg of poly(A)$^+$ RNA in 50 μl of hybridization solution (60% formamide, 0.1% sodium dodecyl sulfate, 20 mM Tris hydrochloride [pH 7.4], 400 mM NaCl, 1 mM EDTA) for 15 min at 65°C and then allowed to hybridize for 20 h at 22°C. The RNA and annealed primer were precipitated with ethanol, dried, and suspended in 50 μl of an aqueous solution containing 0.5 mM dNTP, 50 μg of actinomycin D per ml, 1 mM vanadyl ribonucleotide complex, 50 μg of bovine serum albumin per ml, 625 U of Moloney murine leukemia virus reverse transcriptase (Bethesda Research Laboratories, Inc., Gaithersburg, Md.), and 10 μl of 5× reverse-transcriptase buffer (Bethesda Research Laboratories). The extension reaction was stopped after 1 h at 37°C by the addition of 5 μl of 500 mM EDTA. The solution was extracted with phenol-chloroform, and nucleic acids in the aqueous phase were precipitated with ethanol, dried, suspended in sequencing buffer, denatured, and run on a 6% sequencing gel. The same primer A with HMGI cDNA from clone 8A as template was used to generate a sequencing ladder by the dideoxynucleotide termination method (24) with Sequenase and [^{35}S]ATP.

RESULTS

A restriction enzyme map illustrating the overall size and structure of the human HMGI cDNAs is shown in Fig. 1. The cDNA clones varied in length from 1,670 to 2,050 bp and contained characteristic G+C-rich (70%) 5'-UTRs that varied in size from 45 to 312 bp (Fig. 2); a protein coding sequence of 297, 300, or 333 bp; and identical 3'-UTRs of 1,340 bp excluding poly(A) tracts ranging in length from 0 to 84 bp (Fig. 3). Because of the 5'-UTR heterogeneity, nucleotide numbering for all clones begins with the ATG start codon of the HMGI ORF. Of the eight human HMGI cDNA clones isolated, the sequence of clone 2B is most similar to the previously reported murine HMGI cDNA sequence (11). The nucleotide sequence similarity between this human cDNA and that of the murine cDNA was 74% for the 5'-UTRs, 91% for the protein-coding sequences, and 64% for the 3'-UTRs.

5'-UTR variation among cDNA clones. The nucleotide sequences of the 5'-UTRs varied among clones (Fig. 2).

```
                                                    MetSerGluSerSer     5
GCATCCCAGCCATCACTCTTCCACCTGCTCCTTAGAGAAGGGAAGATGAGTGAGTCGAGC
-45                                                 1                  15

SerLysSerSerGlnProLeuAlaSerLysGlnGluLysAspGlyThrGluLysArgGly    25
TCGAAGTCCAGCCAGCCCTTGGCCTCCAAGCAGGAAAAGGACGGCACTGAGAAGCGGGGC    75

ArgGlyArgProArgLysGlnProProValSerProGlyThrAlaLeuValGlySerGln    45
CGGGGCAGGCCGCGCAAGCAGCCTCCGGTGAGTCCCGGGACAGCGCTGGTAGGGAGTCAG   135

LysGluProSerGluValProThrProLysArgProArgGlyArgProLysGlySerLys    65
AAGGAGCCCAGCGAAGTGCCAACACCTAAGAGACCTCGGGGCCGACCAAAGGGAAGCAAA   195

AsnLysGlyAlaAlaLysThrArgLysThrThrThrThrProGlyArgLysProArgGly    85
AACAAGGGTGCTGCCAAGACCCGGAAAACCACCACAACTCCAGGAAGGAAACCAAGGGGC   255

ArgProLysLysLeuGluLysGluGluGluGluGlyIleSerGlnGluSerSerGluGlu   105
AGACCCAAAAAACTGGAGAAGGAGGAAGAGGAGGGCATCTCGCAGGAGTCCTCGGAGGAG   315

GluGlnEnd                                                       107
GAGCAGTGACCCATGCGTGCCGCCTGCTCCTCACTGGAGGAGCAGCTTCCTTCTGGGACT   375

GGACAGCTTTGCTCCGCTCCCACCGCCCCCGCCCCTTCCCCAGGCCCACCATCACCACCG   435
CCTCTGGCCGCCACCCCCATCTTCCACCTGTGCCCTCACCACCACACTACACAGCACACC   495
AGCCGCTGCAGGGGCTCCCATGGGCCTGAGTGGGGAGCAGTTTTCCCCTGGCCTCAGTTC   555
ACAGCTCCCCCCGCCCACCCACGCATACACACATGCCCTCCTGGACAAGGCTAACATCCC   615
ACTTAGCCGCACCCTGCACCTGCTGCGTCCCCACTCCCTTGGTGGTGGGGACATTGCTCT   675
CTGGGCTTTTGGTTTGGGGGCGCCCTCTCTGCTCCTTCACTGTTCCCTCTGGCTTCCCAT   735
AGTGGGGCCTGGGAGGGTTCCCCTGGCCTTAAAAGGGGCCCAAGCCATCTCATCCTGGCA   795
CGCCCTACTCCACTGCCCTGGCACAGCAGGTGTGGCCAATGGAGGGGGGTGCTGGCCCCC   855
AGGATTCCCCCAGCCAAACTGTCTTTGTCACCACGTGGGGCTCACTTTTCATCCTTCCCC   915
AACTTCCCTAGTCCCCGTACTAGGTTGGACAGCCCCCTTCGGCTACAGGAAGGCAGGAGG   975
GGTGAGTCCCCTACTCCCTCTTCACTGTGGCCACAGCCCCCTTGCCCTCCGCCTGGGATC  1035
TGAGTACATATTGTGGTGATGGAGATGCAGTCACTTATTGTCCAGGTGAGGCCCAAGAGC  1095
CCTGTGGCCGCACCTGAGGTGGGCTGGGGCTGCTCCCCTAACCCTACTTTCGTTCCGCCA  1155
CTCAGCCATTTCCCCCTCCTCAGATGGGGCACCAATAACAAGGAGCTCACCCTGCCCGCT  1215
CCCAACCCCCCTCCTGCTCCTCCCTGCCCCCCAAGGTTCTGGGTTCCATTTTTCCTCTGT  1275
TCACAAACTACCTCTGGACAGTTGTGTTGTTTTTTGTTCAATGTTCCATTCTTCGACATC  1335
CGTCATTGCTGCTGCTACCAGCGCCAAATGTTCATCCTCATTGCCTCCTGTTCTGCCCAC  1395
GATCCCCTCCCCCAAGATACTCTTTGTGGGAAGAGGGGCTGGGGCATGGCAGGCTGGGTG  1455
ACCGACTACCCCAGTCCCAGGGAAGGTGGCCCTGCCCCTAGGATGCTGCAGCAGAGTGAG  1515
CAAGGGGGCCCGAATCGACCATAAAGGGTGTAGGGGCCACCTCCTCCCCCTGTTCTGTTG  1575
GGGAGGGGTAGCCATGATTTGTCCCAGCCTGGGGCTCCCTCTCTGGTTTCCTATTTGCAG  1635
TTACTTGAATAAAAAAAATATCCTTTTCTGGAAAAAAAAAAAAAAAAAAAA.....
                                1666
```

FIG. 3. Human HMGI cDNA sequence common to all eight clones but excluding the variable regions 5′ to nucleotide −45 of Fig. 2. The deduced amino acid sequence of HMGI is given above its ORF (nucleotides 1 to 321). The single-underlined 33 nucleotides (103 to 135) in the HMGI ORF are present in clones 3B, 6A, and 7C but deleted in clones 1A, 2B, and 8A; clone 11D has 36 nucleotides (103 to 138) deleted. The nine nucleotides in boldface (100 to 108) conform to the consensus sequence of an RNA 5′-splice donor site. The consensus sequence signal for poly(A) addition is double underlined.

However, this variation appeared to be mostly due to different arrangements of identical blocks of nucleotides. Nucleotides −45 to −1 were identical in all eight clones. Nucleotides −135 to −47 of clone 10A were identical to nucleotides −242 to −154 of clones 1A and 8A. Nucleotides −153 to −46 of clones 1A and 8A were identical to nucleotides −153 to −46 of clones 2B, 3B, and 7C. Nucleotides −178 to −154 of clones 2B, 3B, and 7C conformed to the consensus sequence for a RNA 3′ splice acceptor site (20). This canonical RNA splice site coincided exactly with the site of cDNA sequence divergence among clones 2B, 3B, and 7C and clones 1A and 8A (Fig. 2). Nucleotides −312 to −46 of clone 6A formed a unique sequence that contained an additional short ORF (nucleotides −136 to −14) that was 5′ to, but out of phase with, the HMGI ORF (nucleotides 1 to 321).

We detected a broad band of HMGI mRNAs by Northern blot analysis of RNA extracts from human K562 and HUT78 cells (Fig. 4A). The estimated size range of these mRNAs (1,700 to 2,300 nucleotides), although not precise, agreed with the size range of the cloned HMGI cDNAs. HMGI mRNA expression was almost 10-fold higher in K562 cells than in HUT78 cells, although both cell types expressed HMGI mRNAs of similar sizes.

By primer extension analysis we detected three distinct size classes of 5′-UTRs in HMGI mRNAs from K562 cells (Fig. 4B). These size classes corresponded to three of the 5′-UTR sizes observed in the cloned cDNAs. The most abundant type of 5′-UTR in HMGI mRNAs from K562 cells corresponded in size to the 5′-UTR of cDNA clones 2B, 3B, and 7C. The next most abundant type corresponded to the 5′-UTR of cDNA clones 1A and 8A. A faint band was detected corresponding to HMGI mRNAs with the same 5′-UTR size as cDNA clone 10A. However, we detected no significant band representing the larger 5′-UTR size of cDNA clone 6A.

Amino acid sequences of isoforms HMG-I and HMG-Y. The protein-coding sequences of the human HMGI cDNAs varied among clones. Three of the clones (2B, 6A, and 7C) encoded a 107-amino-acid protein (Fig. 3) whose sequence is virtually identical to the reported amino acid sequence of the human placental HMG-I protein determined by peptide fragment analysis (17). Four of the other cDNA clones (1A, 2B, 8A, and 10A) encoded a similar but slightly smaller 96-amino-acid protein identical in sequence to HMG-I except for an internal 11-amino-acid deletion (Fig. 3, residues 35 to 45, corresponding to the in-frame deletion of nucleotides 103 to 135). Clone 11D encoded a 95-amino-acid protein

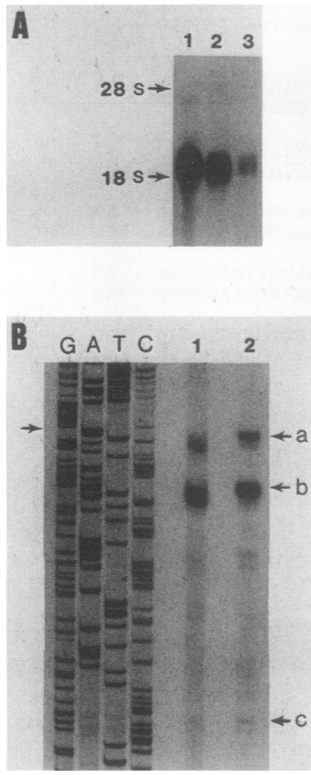

FIG. 4. Analysis of HMGI mRNAs from human cell extracts. (A) Northern blot of total RNA extracted from K562 cells (20 μg; lane 1) and poly(A)$^+$ RNA extracted from K562 cells (3 μg; lane 2) and HUT78 cells (5 μg; lane 3). The blot was hybridized with a ^{32}P-labeled, 490-bp, *Sac*I-to-*Pst*I fragment (nucleotides 15 to 505 in Fig. 3) of human HMGI cDNA. Migration positions of 28S and 18S rRNA chains are shown on the left. (B) Primer extension analysis of total RNA (20 μg; lane 1) and poly(A)$^+$ RNA (5 μg; lane 2) extracted from K562 cells. A synthetic 39-mer that is antisense to nucleotides 43 to 81 in Fig. 3 was end labeled with ^{32}P and used to prime DNA extension on the cellular RNA templates. This same primer was used to produce a sequencing size ladder of HMGI cDNA from clone 8A (lanes G, A, T, and C). The arrow on the left indicates the primer-to-5′-end size of this cDNA (325 nucleotides, including the primer). The arrows on the right point to bands corresponding to mRNAs that have the same primer-to-5′-end lengths as HMGI cDNA clones 8A and 1A (arrow a, 325 nucleotides); clones 2B, 3B, and 7C (arrow b, 290 nucleotides); and clone 10A (arrow c, 215 nucleotides).

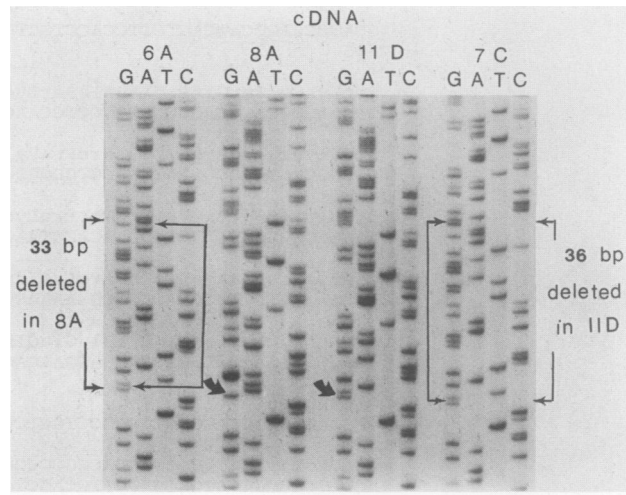

FIG. 5. Variation of the HMGI protein-coding sequence among cDNA clones. cDNAs from clones 6A, 8A, 11D, and 7C were sequenced by priming with a synthetic oligonucleotide corresponding to the sense strand of nucleotides −35 to −16 in Fig. 3. The large arrows indicate the locations of a 33-bp deletion in clone 8A and a 36-bp deletion in clone 11D. The 33 bp deleted from clone 8A (nucleotides 103 to 135 in Fig. 3) are demarcated by arrows in the sequencing lanes for clone 6A; the 36 bp deleted from clone 11D (nucleotides 103 to 138) are demarcated by arrows in the sequencing lanes for clone 7C. Clones 6A and 7C encode the HMG-I isoform and clones 8A and 11D encode the "deleted" isoform HMG-Y.

also identical in sequence to HMG-I except for a 12-amino-acid deletion (Fig. 3, residues 35 to 46, corresponding to the in-frame deletion of nucleotides 103 to 138).

Nucleotides 100 to 108 (CCGGTGAGT) of cDNA clones 3B, 6A, and 7C conform to the consensus sequence for an RNA 5′ splice donor site (20). The point of postulated RNA cleavage within this site was between nucleotides 102 and 103; this point corresponded exactly to the location of the observed deletions in cDNA clones 1A, 2B, 8A, 10A, and 11D. An autoradiogram of the different sequencing gel patterns observed among clones 6A, 8A, 11D, and 7C illustrates the two types of in-frame deletions detected and their locations in the protein-coding sequences of these clones (Fig. 5). Note that the deletions in clones 8A and 11D occur precisely at the postulated internal RNA cleavage site—between nucleotides CCG and GTGAGT—of clones 6A and 7C.

From partial peptide sequence data, the protein designated HMG-Y by Lund et al. (18) was previously postulated to be an isoform of HMG-I (16), but the entire protein had not been completely sequenced. Particularly missing from the results of these previous HMG-Y analyses was the peptide sequence of the region containing the 11-amino-acid deletion that was predicted from the cDNA sequences (Fig. 3). We therefore sequenced additional peptide fragments of HMG-Y purified from murine ascites cells by high-pressure liquid chromatography. The amino acid sequence of one of these peptides (Fig. 6A, bottom rows, single-underlined residues 23 to 52) confirmed that the HMG-Y isoform does indeed contain the predicted 11-amino-acid deletion. We conclude, therefore, that human cDNA clones 3B, 6A, and 7C encode HMG-I (Fig. 6A, top rows) and that human cDNA clones 1A, 2B, 8A, and 10A (Fig. 6A, middle rows) and the previously reported murine cDNA clone (Fig. 6A, bottom rows) encode HMG-Y. A HMGI protein with a 12-amino-acid deletion corresponding to the deduced product of human cDNA clone 11D may also exist but would be difficult to separate from HMG-Y.

The two identical palindromic amino acid sequences (PRGRP) reported by Lund et al. (17) to be potential DNA-binding motifs of HMG-I (Fig. 6A, top rows, residues 57 to 61 and 83 to 87) were also present in human and murine HMG-Y (Fig. 6A, bottom and middle rows, residues 46 to 50 and 72 to 76). A third similar motif (RGRP) was also present in both HMG-I and HMG-Y (Fig. 6A, all rows, residues 26 to 29). The internal 11 amino acids deleted in HMG-Y but present in HMG-I represented the only hydrophobic region in an otherwise highly hydrophilic protein (Fig. 6B); this deletion occurred between the first two of the three postulated DNA-binding motifs.

Human genome contains several HMGI-coding genes; most are probably nonfunctional pseudogenes. Southern blot analysis of restricted human placental DNA hybridized with two

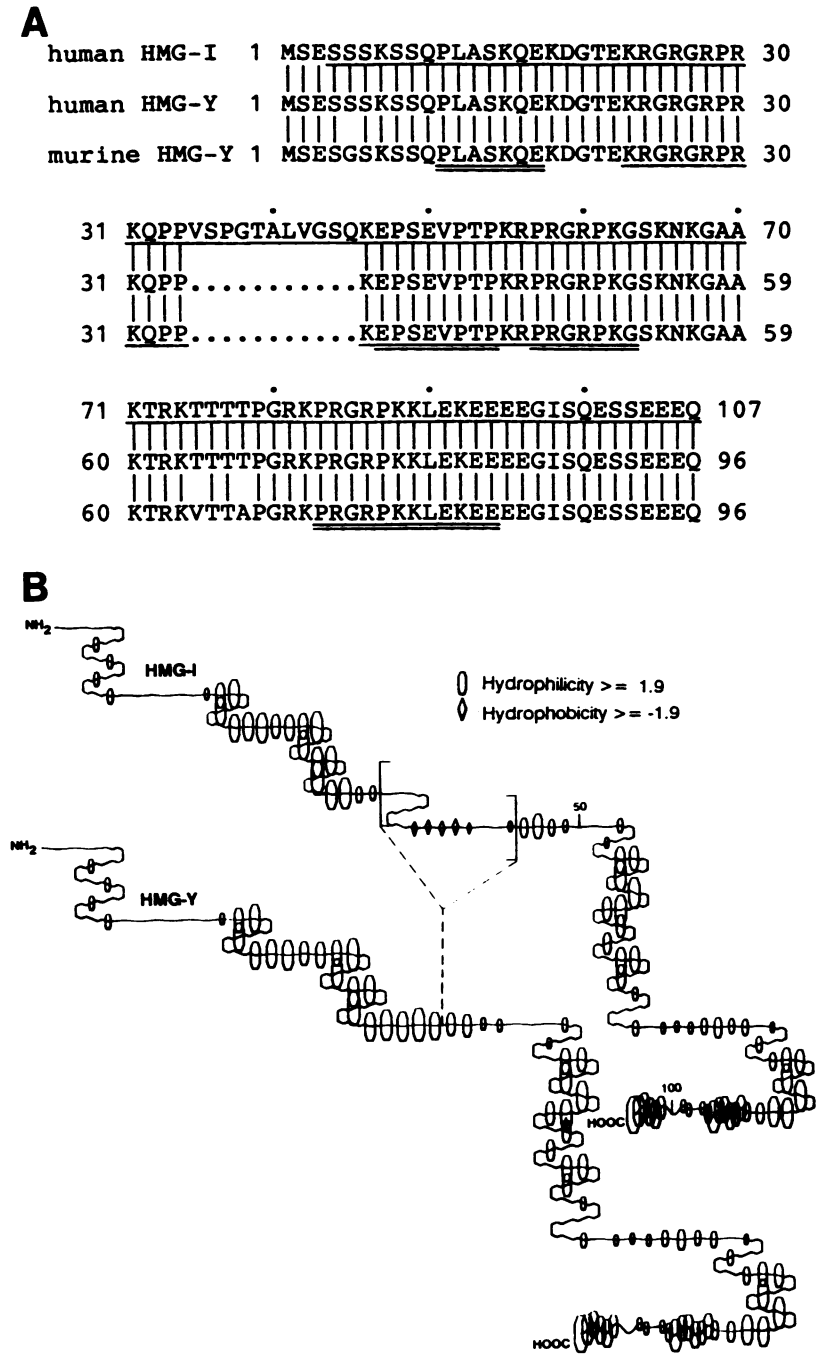

FIG. 6. (A) Comparison of the amino acid sequences of human and murine HMG-I and HMG-Y. (Top rows) Amino acid sequence of human HMG-I, as deduced from the cDNA sequences reported in this paper (residues 1 to 107) and from the peptide fragment analysis reported by Lund et al. (17) (underlined residues 4 to 107). (Middle rows) Human HMG-Y amino acid sequence, as deduced from the cDNA sequences reported in this paper. (Bottom rows) Murine HMG-Y amino acid sequence, as deduced from the cDNA sequence reported by Johnson et al. (11) (residues 1 to 96) and from the peptide fragment analysis reported in this paper (single-underlined residues 23 to 52) and by Lehn et al. (16) (double-underlined residues 11 to 17, 36 to 43, 46 to 52, and 72 to 83). (B) Predicted secondary structure (2) of the human HMG-I and HMG-Y proteins generated by the PLOTSTRUCTURE program of the University of Wisconsin Genetics Computer Group (4). The bracketed region of HMG-I represents the internal 11 amino acids that are deleted at the point indicated in HMG-Y.

different regions of human HMGI cDNA suggested that there are more than 2 but probably fewer than 10 HMGI-coding genes in the haploid human genome (Fig. 7). We screened a human genomic EMBL3 library and isolated a number of clones containing HMGI pseudogenes (Fig. 8). Partial sequence analysis of one of these pseudogenes revealed characteristics of a processed HMG-Y retropseudogene (31): it contained no introns, it contained a poly(A) tract at its 3′ terminus, and it was flanked by 13-bp direct repeats (Fig. 8). The pseudogene showed considerable divergence

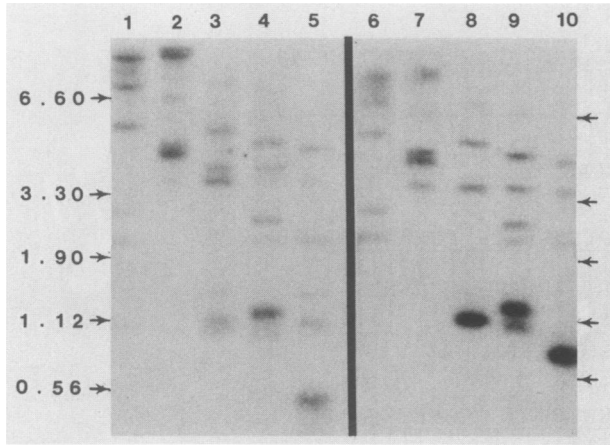

FIG. 7. Southern blot analysis of DNA extracted from human placenta. DNA (15 μg per lane) was restricted with *Eco*RI (lanes 1 and 6), *Hin*dIII (lanes 2 and 7), *Pst*I (lanes 3 and 8), *Sac*I (lanes 4 and 9), or *Pst*I plus *Sac*I (lanes 5 and 10). Lanes 1 to 5 were hybridized with the ^{32}P-labeled, 490-bp, *Sac*I-to-*Pst*I fragment (nucleotides 15 to 505 in Fig. 3), and lanes 6 to 10 were hybridized with the 695-bp, *Pst*I-to-*Sac*I fragment (nucleotides 506 to 1201 in Fig. 3) of human HMGI cDNA. Molecular weight markers (in kilobases) appear on the left.

from the cDNA sequence to which it was most similar (clone 10A). Of 674 nucleotides compared with this cDNA, the pseudogene had 50 single-nucleotide substitutions, one 36-bp deletion, and one 6-bp duplication, giving an overall similarity of 92%.

DISCUSSION

The HMGI cDNA sequence variation that we observed among clones was most likely the result of alternative pre-mRNA processing from a single functional gene. Coffino (3) cautioned that some cDNA sequence heterogeneity may be artifactual, resulting from the infidelity of reverse transcriptase, particularly on the 5′ leader sequences of template mRNAs. However, the results of our primer extension and peptide sequence analyses show that the observed HMGI cDNA variation corresponds to observed HMGI mRNA and protein variation and is, therefore, not a cDNA-cloning artifact. Our primer extension analysis detected HMGI mRNAs with three distinct leader sequence lengths corresponding in size to those of the cloned cDNAs. Our peptide fragment sequence analysis of the HMG-I and HMG-Y protein isoforms revealed the same 11-amino-acid addition-deletion dimorphism as was deduced from our cDNA sequence analysis.

The presence of internal RNA splice acceptor and donor sites in the variable cDNA sequences is additional evidence that alternative pre-mRNA processing generated the cDNA heterogeneity. The predicted locations of RNA cleavages within the canonical splice sites occurred precisely at the observed points of cDNA sequence divergence among clones. Regions of nucleotide sequence identities and differences among the HMGI cDNA clones are shown in Fig. 9. Most of the HMGI cDNA variation is due to the different arrangements of identical blocks of nucleotides and is not the result of evolutionary sequence divergence. Evolutionary divergence between separate genes would result in an accumulation of nucleotide differences that would be distributed randomly throughout the gene sequence (12). All eight cloned cDNAs have identical nucleotide sequences except for their variable 5′ leader sequences and the presence or absence of a 33- or 36-bp deletion in their ORFs.

Several HMGI pseudogenes were isolated, and one of these was analyzed in detail. In contrast to the virtual identity in nucleotide sequences among HMGI cDNA clones, the DNA sequence of this pseudogene showed substantial divergence from the cDNA sequence to which it was most similar (Fig. 8). This degree of sequence divergence agrees closely with that found among HMG-17 cDNAs and their corresponding pseudogenes (27) and represents an evolutionary divergence time of roughly 10 million years, assuming no selective constraints on pseudogene mutations (12). Also like HMG-17 pseudogenes (27), the HMGI pseudogene contains all the characteristics of a processed retropseudogene—the absence of introns, the presence of a poly(A) tract, and flanking direct-repeat sequences (Fig. 8). HMGI retropseudogenes are probably not functional: they would have to have been fortuitously inserted near some type of transcriptional promoter element to be expressed (31). Nucleotide sequences under functional constraints are expected to be more conserved than are nonfunctional sequences. Since the degree of nucleotide sequence divergence between the human HMGI cDNA and HMGI pseudogene is the same in the noncoding regions as in the ORFs (Fig. 8), the ORF of the pseudogene probably does not encode a functional protein. Indeed, the deduced amino acid sequence similarity between the murine and human HMGI cDNAs was 97% (Fig. 6A), whereas the deduced amino acid similarity between the human HMGI cDNA and human pseudogene ORFs was only 86%. Loss of the poly(A) addition signal is additional evidence that the human HMGI pseudogene is not functional (Fig. 8). We conclude from these results and from our Southern blot analysis (Fig. 7) that the haploid human genome contains several HMGI gene copies. Most likely only one of these copies is a functional HMGI gene; the rest are probably nonfunctional retropseudogenes.

Many examples have been reported of alternatively processed mRNAs (1, 15) that are similar to what we observed for HMGI (Fig. 9), including those that utilize alternative 5′-terminal exons and that contain internal donor and acceptor splice sites. However, the only other example of a nuclear DNA-binding protein whose diversity results from alternative RNA processing is the CTF-NF-I family of human CCAAT-box-binding proteins (25). HMGI is the only structural component of mammalian chromosomes for which such a mechanism for generating multiple protein isoforms has been reported.

Temporal and spatial regulation of alternatively processed mRNAs has been reported for the c-*abl* gene in mice (22) and the *antp* gene in *Drosophila melanogaster* (28), among others. The multiple alternatively processed mRNAs encoding HMGI may likewise be under temporal and spatial regulation. There may be a functional difference between the two protein isoforms, HMG-I and HMG-Y, that are the result of alternative RNA processing. The deletion of the 11 amino acids from HMG-I that results in the HMG-Y isoform brings closer two highly conserved amino acid sequences that may be DNA-binding motifs (17). This deletion could therefore change the DNA-binding affinity of HMG-Y relative to that of HMG-I. The relative abundance of these two protein isoforms may be somehow regulated by their variable mRNA leader sequences; however, comparisons of the HMGI cDNAs (Fig. 9) suggest that the splicing combina-

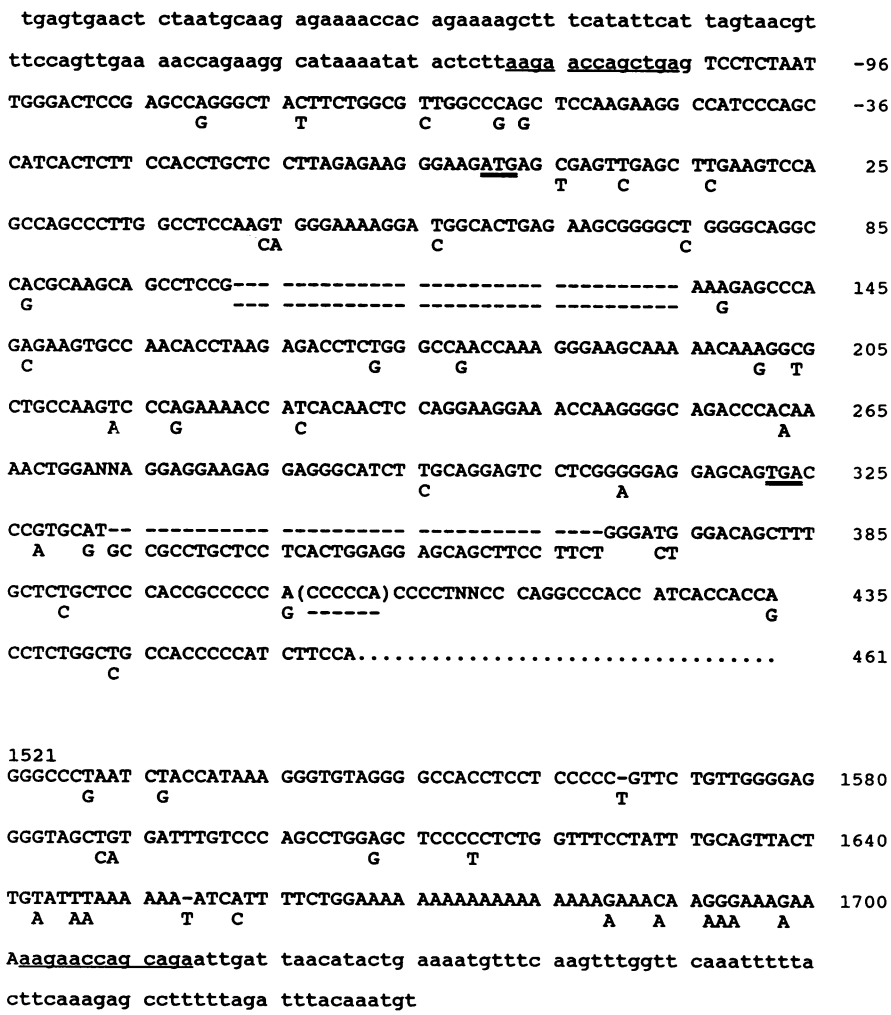

FIG. 8. Partial DNA sequence of a human HMGI retropseudogene. Nucleotides of a HMG-Y cDNA sequence (clone 10A) that differ from the pseudogene DNA sequence are listed below it. Nucleotide numbers are consistent with the cDNA numbers in Fig. 2 and 3; the internal region of the pseudogene corresponding to cDNA nucleotides 462 to 1520 was not sequenced. Lowercase letters designate flanking genomic sequences; 13-bp direct repeats are underlined. Translation start and stop codons are double underlined. Deleted nucleotides are designated by hyphens; both the HMG-Y cDNA and the pseudogene have a 33-bp deletion in their ORFs (nucleotides 103 to 135); the pseudogene has an additional 36-bp deletion (nucleotides 334 to 369). A duplication in the pseudogene of nucleotides 401 to 406 is enclosed in parentheses. The poly(A) addition signal (AATAAA) has been lost (nucleotides 1643 to 1648).

tions in the leader sequences are independent of the splicing pattern in the protein-coding sequence.

Synthesis of two proteins from nonoverlapping reading frames theoretically can occur either by "leaky" ribosome scanning or by internal initiation (13). The short ORF in the leader sequence of clone 6A encodes a 41-amino-acid peptide that precedes and is out of frame with the larger HMGI ORF (Fig. 2). It is possible that this small peptide is synthesized and serves some regulatory function. Alternatively, since a presumed RNA splice site occurs within the ORF of the leader sequence, it is also possible that some mRNAs are processed that contain this same leader sequence but with its ORF in phase with the HMGI ORF, thus encoding a larger, 148-amino-acid HMGI-like protein. Alternative initiations at two in-phase AUGs have been shown to produce the two forms of the Ia antigen-associated invariant chain in humans (29). In either case, the type of HMGI mRNA leader sequence exemplified by clone 6A is atypical of the other, much more abundant forms of HMGI mRNA leader sequences found in K562 cells by primer extension (Fig. 4B) and may represent an incompletely processed mRNA (14).

In conclusion, we have shown by cDNA sequence analysis that HMGI isoforms are encoded by at least six distinct mRNAs and that the HMG-I protein isoform contains 11 additional amino acids that are absent from the HMG-Y isoform. We have presented evidence that these multiple mRNAs are the result of alternative processing from a single functional gene. In addition, we have shown that the haploid human genome also contains several nonfunctional pseudogenes, one of which we partially sequenced and found to be a processed retropseudogene.

These findings establish the HMGI family of nonhistone chromatin proteins as among the most thoroughly characterized of all mammalian nuclear structural proteins. In addition, this is the first report of cDNAs encoding proteins that are specifically localized in the G(Q)-band regions of human mitotic chromosomes. The role(s) played by the HMGI protein family in chromosome structure is unknown, but all current evidence suggests that these proteins are involved in

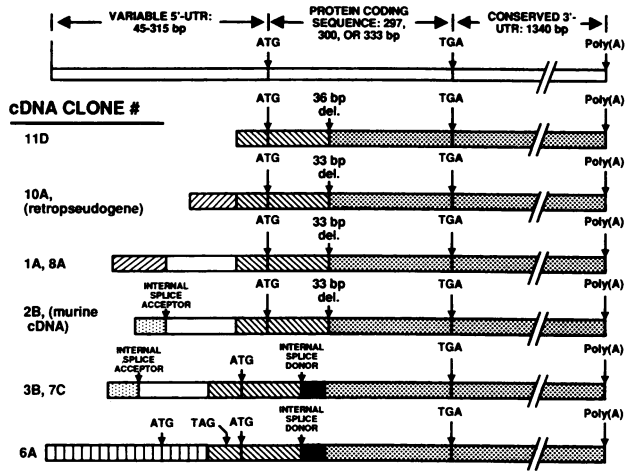

FIG. 9. Multiple HMGI mRNAs alternatively processed from a single gene. Each cDNA clone is represented by a bar; identical patterns on these bars represent identical cDNA sequences. Translation start (ATG) and stop (TGA or TAG) codons and the consensus signal for poly(A) addition are indicated by arrows, as are cDNA sequences conforming to the consensus sequences for RNA splice acceptor and donor sites. The black regions (33 or 36 bp) of the bars representing clones 3B, 6A, and 7C are deleted at the positions indicated on the bars representing clone 11D (36 bp) and clones 10A, 1A, 8A, and 2B (33 bp).

chromosome compaction, cell replication, or both. Alternative splicing of precursor HMGI mRNAs may well play an important role in regulating the functional expression of HMGI protein isoforms in mammalian chromosomes. These possibilities are being actively investigated.

ACKNOWLEDGMENTS

We thank John Bollinger for helping to synthesize the oligonucleotide primers, Gerhard Munske for helping to sequence the HMG-Y peptides, Mark Nissen for advice on primer extension procedures, and Jane Disney for help with the photography.

This work was supported in part by National Science Foundation grant DCB-8602622 and Department of Agriculture grant 85-CRCR-1-1730 (both to R.R.).

LITERATURE CITED

1. **Breitbart, R. E., A. Andreadis, and B. Nadal-Ginard.** 1987. Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes. Annu. Rev. Biochem. **56**:467–495.
2. **Chou, P. Y., and G. D. Fasman.** 1978. Prediction of the secondary structure of proteins from their amino acid sequence. Adv. Enzymol. **47**:45–147.
3. **Coffino, P.** 1988. Probable cloning artefacts previously interpreted as unusual leader sequences of rodent ornithine decarboxylase mRNAs—a cautionary tale. Gene **69**:365–368.
4. **Devereux, J., P. Haeberli, and O. Smithies.** 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12**:387–395.
5. **Elton, T. S., M. S. Nissen, and R. Reeves.** 1987. Specific A-T DNA sequence binding of RP-HPLC purified HMG-I. Biochem. Biophys. Res. Commun. **143**:260–265.
6. **Elton, T. S., and R. Reeves.** 1985. Microheterogeneity of the mammalian high mobility group (HMG) proteins 1 and 2 investigated by reverse-phase high performance liquid chromatography. Anal. Biochem. **144**:403–416.
7. **Elton, T. S., and R. Reeves.** 1985. Microheterogeneity of the mammalian high-mobility group proteins 14 and 17 investigated by reverse-phase high-performance liquid chromatography. Anal. Biochem. **146**:448–460.
8. **Elton, T. S., and R. Reeves.** 1986. Purification and post-synthetic modifications of Friend erythroleukemic cell high mobility group protein HMG-I. Anal. Biochem. **157**:53–62.
9. **Giancotti, V., B. Pani, P. D'Andrea, M. T. Berlingieri, P. P. Di Fiore, A. Fusco, G. Veccio, R. Philip, C. Crane-Robinson, R. H. Nicolas, C. A. Wright, and G. H. Goodwin.** 1987. Elevated levels of a specific class of nuclear phosphoproteins in cells transformed with v-ras and v-mos oncogenes and by co-transfection with c-myc and polyoma middle T genes. EMBO J. **6**:1981–1987.
10. **Goodwin, G. H., P. N. Cockerill, S. Kellam, and C. A. Wright.** 1985. Fractionation by high-performance liquid chromatography of the low-molecular-mass high-mobility-group (HMG) chromosomal proteins present in proliferating rat cells and an investigation of the HMG proteins present in virus transformed cells. Eur. J. Biochem. **149**:47–51.
11. **Johnson, K. R., D. A. Lehn, T. S. Elton, P. J. Barr, and R. Reeves.** 1988. The chromosomal high mobility group protein HMG-I(Y): complete murine cDNA sequence, genomic structure, and tissue expression. J. Biol. Chem. **263**:18338–18342.
12. **Kimura, M.** 1983. The neutral theory of molecular evolution, p. 208–233. *In* M. Nei and R. K. Kohn (ed.), Evolution of genes and proteins. Sinauer Associates, Inc., Sunderland, Mass.
13. **Kozak, M.** 1986. Bifunctional messenger RNAs in eukaryotes. Cell **47**:481–483.
14. **Kozak, M.** 1987. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res. **15**:8125–8132.
15. **Leff, S. E., and M. G. Rosenfeld.** 1986. Complex transcriptional units: diversity in gene expression by alternative RNA processing. Annu. Rev. Biochem. **55**:1091–1117.
16. **Lehn, D. A., T. S. Elton, K. R. Johnson, and R. Reeves.** 1988. A conformational study of the sequence specific binding of HMG-I(Y) with the bovine interleukin-2 cDNA. Biochem. Int. **16**: 963–971.
17. **Lund, T., K. H. Dahl, E. Mork, J. Holtlund, and S. G. Laland.** 1987. The human chromosomal protein HMG I contains two identical palindrome amino acid sequences. Biochem. Biophys. Res. Commun. **146**:725–730.
18. **Lund, T., J. Holtlund, M. Fredriksen, and S. G. Laland.** 1983. On the presence of two new high mobility group-like proteins in HeLa S3 cells. FEBS Lett. **152**:163–167.
19. **Nissen, M. S., and P. D. Friesen.** 1989. Molecular analysis of the transcriptional regulatory region of an early baculovirus gene. J. Virol. **63**:493–503.
20. **Padgett, R. A., P. J. Grabowski, M. M. Konarska, S. Seiler, and P. A. Sharp.** 1986. Splicing of messenger RNA precursors. Annu. Rev. Biochem. **55**:1119–1150.
21. **Reeves, R., T. S. Elton, M. S. Nissen, D. Lehn, and K. R. Johnson.** 1987. Posttranscriptional gene regulation and specific binding of the nonhistone protein HMG-I by the 3' untranslated region of bovine interleukin 2 cDNA. Proc. Natl. Acad. Sci. USA **84**:6531–6535.
22. **Renshaw, M. W., M. A. Capozza, and J. Y. J. Wang.** 1988. Differential expression of type-specific c-*abl* mRNAs in mouse tissues and cell lines. Mol. Cell. Biol. **8**:4547–4551.
23. **Russnak, R. H., P. M. Candido, and C. R. Astell.** 1988. Interaction of the mouse chromosomal protein HMG-I with the 3' ends of genes in vitro. J. Biol. Chem. **263**:6392–6399.
24. **Sanger, F., G. Nicklen, and A. R. Coulsen.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
25. **Santoro, C., N. Mermod, P. C. Andrews, and R. Tijan.** 1988. A family of human CCAAT-box-binding proteins active in transcription and DNA replication: cloning and expression of multiple cDNAs. Nature (London) **334**:218–334.
26. **Solomon, M., F. Strauss, and A. Varshavsky.** 1986. A mammalian high mobility group protein recognizes any stretch of six A-T base pairs in duplex DNA. Proc. Natl. Acad. Sci. USA **83**:1276–1280.
27. **Srikantha, T., D. Landsman, and M. Bustin.** 1987. Retropseudogenes for human chromosomal protein HMG-17. J. Mol. Biol. **197**:405–413.

28. **Stroeher, V. L., J. C. Gaiser, and R. L. Garber.** 1988. Alternative RNA splicing that is spatially regulated: generation of transcripts from the antennapedia gene of *Drosophila melanogaster* with different protein-coding regions. Mol. Cell. Biol. **8**:4143–4154.
29. **Strubin, M., E. O. Long, and B. Mach.** 1986. Two forms of the Ia antigen-associated invariant chain result from alternative initiations at two in-phase AUGs. Cell **47**:619–625.
30. **Vartiainen, E., J. Palvimo, A. Mahonen, A. Linnala-Kankkunen, and P. H. Maenpaa.** 1988. Selective decrease in low-M_r HMG proteins HMG I and HMG Y during differentiation of mouse teratocarcinoma cells. FEBS Lett. **228**:45–48.
31. **Weiner, A. M., P. L. Deininger, and A. Efstratiadis.** 1986. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu. Rev. Biochem. **55**:631–661.
32. **Yang-Yen, H.-F., and L. I. Rothblum.** 1988. Purification and characterization of a high-mobility-group-like DNA-binding protein that stimulates rRNA synthesis in vitro. Mol. Cell. Biol. **8**:3406–3414.