We have isolated an active gene encoding the mouse HMG1 protein among a multitude of cross-hybridizing sequences, which most likely are retrotransposed pseudo-genes. The hmgl gene contains five exons, of which the first is not translated, and the last contains a long 3'-untranslated sequence and three alternative polyadenylation sites. We found no evidence for a sequence encoding a membrane localization signal in the hmgl gene, despite the presence of HMG1 protein on the surface of several cell types. The hmgl promoter coincides with a CpG island, contains no TATA sequence, and drives the expression of reporter genes placed under its control. The hmgl gene may be a member of a family of closely related genes but appears to be the major or the only active gene coding for HMG1 protein.

High mobility group 1 protein (HMG1) is a very abundant and highly conserved carboxatin protein, which is present in all vertebrate nuclei. HMG1-like proteins exist in invertebrates, yeast, protozoa, and plants and are probably present in all eukaryotic cells (reviewed in Refs. 1 and 2). The ubiquity and sequence conservation of HMG1-like proteins suggest that they play fundamental functions, roles in DNA replication, chromatin assembly, and transcription have been proposed but so far have not been proved unequivocally.

In vitro mammalian HMG1 and its two DNA-binding domains bind with low affinity and no specificity to single-stranded, linear duplex and supercoiled DNA (3, 4). They also bind with high specificity and in a sequence-independent manner to DNA containing a sharp bend or kinks (5–7). More generally, HMG1 has the ability to introduce bends or kinks into linear DNA and therefore is functionally (but not structurally) similar to the prokaryotic proteins HU and IHF (reviewed in Ref. 8). The main role of HMG1 may be to facilitate the formation of specific nucleoprotein complexes (9) and perhaps to modulate the structure of chromatin (10).

Finally, HMG1 has been shown to be present on the surface of neurons and other cell types (11), where it is probably bound to the polysaccharide moiety of proteoglycans and may play roles in adhesion and tissue remodeling. Several cell types also display on their surface other proteins that differ from HMG1 only by a few amino acids (12). Thus, the hmgl gene may be a member of a gene family.

Despite this wealth of biochemical information, the analysis of the physiological role of HMG1 has been impaired by the lack of mutations in the gene coding for it. In fact, whereas HMG1 cDNAs have been cloned from a variety of sources (1), the genomic locus coding for HMG1 had not been identified so far. In this paper we describe the identification and the organization of the mouse hmgl gene.

**MATERIALS AND METHODS**

Oligonucleotides and Enzymes—All oligonucleotides were purchased from Genset. DNA modification and restriction enzymes were from Boehringer Mannheim, Promega, and New England Biolabs.

PCR—Intron 3 and intron 4 of gene hmgl were obtained by PCR on genomic DNA from mouse NIH3T3 cells, using the following oligonucleotides: INT3for (coding strand), 5'-ACCCAAGGCGCTCGC-3'; INT3rev (non-coding strand), 5'-CAAGAAAGGCGGCAACT-3'; INT4for (coding strand), 5'-GGGAAGGTATGAAAGG-3'; INT4rev (non-coding strand), 5'-GTTAGGAGGAAATTCC-3'. PCR mixes (50 µl) contained 50 pmol each of oligonucleotides INT3for and INT3rev for amplification of intron 3 and oligonucleotides INT4for and INT4rev for amplification of intron 4. 0.2 mM dNTPs, 5 µM of Taq polymerase 10 x buffer, 20 ng of mouse genomic DNA, and 1 unit of Taq polymerase (Promega). Thirty-five cycles of denaturation (80 s at 94°C), annealing (60 s at 55°C for INT3 and 60 s at 51°C for INT4), and extension (80 s at 72°C) were performed on a Perkin-Elmer instrument.

Isolation of Genomic hmgl Clones—10^6 phage plaques of the 129SV mouse genomic library in the λ-FIXII vector (5×10^6 primary recombinant phages, insert size of 9–22 kb, provided by Stratagene) were screened initially with the 3'-untranslated sequence of the mouse HMG1 cDNA, obtained by PCR and labeled by random priming. The same number of plaques was later screened with two probes obtained from PCR products INT3 and INT4. Fragments from the positive clones were subcloned into Bluescript KS(+) and sequenced with T7 DNA polymerase (Pharmacia Biotech Inc.).

**Plasmid Construction**—To obtain the pHMG1-neo plasmid, the 4.5-kb NsiI-NsiI fragment (containing a 2-kb region upstream of the transcription start site, exon 1, intron 1, and a part of exon 2) and a 2-kb XhoI-XhoI fragment (containing the aph gene from T5 transposon) were cloned into the pBlueScript KS(+) vector. This plasmid expresses a chimeric protein in which the first 2 residues of the bacterial aph gene product are substituted by 15 amino acids from HMG1 protein, followed by 11 amino acids coded by polylinker sequences. The pBarnHI-neo and the pEcoRI-neo constructs were obtained from pHMG1-neo by internal deletion of a 1.5-kb BarnHI-BarnHI fragment and a 3-kb EcoRI-EcoRI fragment, respectively.

**Cell Culture and Transfection**—The mouse NIH3T3 cell line was grown in high glucose Dulbecco's modified Eagle's medium, supplemented by 10% newborn calf serum and antibiotics. Cells were transfected by calcium phosphate co-precipitation in 6-cm dishes as indicated in the legend to Fig. 5.

ARA Extraction, Primer Extension, and RNA Protection—Total RNA was prepared as described by Chomczynski and Sacchi (13). For primer extension, the [32P]-labeled oligonucleotide MnII (5'-GGCATCGCTG-GCTTCTTAG-3') was hybridized to 20 µg of total RNA extracted from NIH3T3 cells in 10 µl of hybridization buffer containing 10 µM Tris-Cl, pH 7.5, 2 mM EDTA, and 60 µM NaCl. After 3 h of hybridization at 42°C, 40 µl of reverse transcriptase reaction mixture containing 10 µM Tris-Cl, pH 8.4, 5 mM MgCl2, 10 µM dithiothreitol, 1 µM dNTPs, 20 units of RNAse inhibitor, and 12.5 units of avian myeloblastosis virus reverse transcriptase (Boehringer Mannheim) were added to the samples, and incubation was continued for 30 min at 42°C. Reactions were stopped.
The hmgl Gene

RESULTS AND DISCUSSION

The Mouse Genome Contains a Large Number of Sequences Similar to the HMG1 cDNA—The presence of several sequences with homology to the HMG1 cDNA has been previously reported in human and pig genomes (14, 15). Southern blot analysis using as probes different regions of the rat HMG1 cDNA (HMG1boxA and the 3'-untranslated region) confirmed the presence of many HMG1-related sequences in mouse, rat, and hamster (Fig. 1 and data not shown). We reasoned that the presumptive hmgl gene contained introns in the same position as the presumptive hmg1 gene contained introns in the same position as the presumptive hmg2 gene.

We designed two couples of oligonucleotides with sequences corresponding to the HMG1 cDNA sequence flanking the presumptive intron 3 and intron 4 and ending at their 3' terminus with one or two bases corresponding to the canonical splicing junction 3'-GT and 3'-CT. PCR reactions on genomic mouse DNA yielded two fragments of about 350 and 850 base pairs, which we named INT3 and INT4, respectively. Control PCRs performed on the cDNA clone gave no amplified band. Direct sequencing of the 350-base pair fragment showed that it had features characteristic of introns (AT-rich sequence, lack of any ORF, and presence of 3'-splicing consensus site). Finally, a PCR reaction performed with the primers 5' to the presumptive intron 3 and 3' to the presumptive intron 4 gave an amplified band of 1.4 kb, indicating that INT3 and INT4 are colinear on the mouse genome.

Evidence that INT3 and INT4 fragments were represented by a single-copy sequence in the mouse genome was provided by Southern blot analysis of mouse genomic DNA, shown in Fig. 1. We then isolated from the phage mouse genomic library four positive clones (HMG1-ϕ1 to HMG1-ϕ4), which hybridized to both INT3 and INT4 probes and had overlapping restriction maps (Fig. 3). HindIII and NotI fragments from phages HMG1-ϕ1 and HMG1-ϕ4 were subcloned in plasmid pBlueScript KS(+).

Fig. 2 shows the sequence we obtained, which includes entirely the 5'-untranslated region and the coding region of the mouse HMG1 cDNA (16). The cDNA sequence contains four base substitutions, one of which causes the conservative replacement of a glutamic acid for an aspartic acid residue in the acidic tail of HMG1 protein. These sequence divergences most probably represent genetic polymorphisms between inbred mouse strains; the cDNA was derived from P19 cells (corresponding to the C3H strain), while our gene was isolated from a bank made with the DNA of SV129 mice.

Structure of the Mouse Hmgl Gene—The hmgl gene contains five exons, as indicated by the comparison of the cDNA and genomic sequences (Fig. 3). An untranslated first exon falls in a region of very high C and G content with the features of a CpG island. Exon 2, which is located about 2.5 kb 3' to exon 1, contains the translation start site. The DNA binding domain A is encoded by exon 2 and exon 3, and the DNA binding domain B is encoded by exons 3, 4, and 5. The relative positions of the introns within the segments coding for the two HMG boxes are different, which is somewhat unexpected if one supposes that the vertebrate hmgl gene arose by internal duplication of an ancestral gene containing a single HMG box, similar to the modern gene for HMG1-like proteins in lower eukaryotes, plants, and insects (18-22). The terminal acidic tail is encoded
The hmg1 Gene

by exon 5, which is the longest and contains a long untranslated region and multiple polyadenylation sites.

In order to map the transcription start site of the hmg1 gene, we performed a primer extension on total RNA extracted from NIH3T3 cells (Fig. 4), using an oligonucleotide that maps immediately downstream of the translation start site in exon 2. One major extended product and two minor ones were obtained. Two of them are not far upstream of the 5' terminus of the longest known cDNA for mouse HMGl (16); the third, shorter extended product may correspond to a weak start site or to a strong pause site for reverse transcriptase.

We found no evidence of alternatively spliced variants of the HMGl mRNAs, and we found no sequence in the genomic locus that could code for membrane-targeting signals. Thus, the presence of the HMGl protein outside of the cell membrane in several cell types probably does not depend on classical protein secretion routes.

The Promoter of the hmg1 Gene—To qualify as the authentic hmg1 gene, the genomic fragment must contain a region in cis capable of driving its transcription. We constructed a plasmid, pHMG1-neo, in which a fragment encompassing 2 kb upstream of the transcription start sites, exon 1, the entire intron 1, and part of exon 2, was fused in frame to the prokaryotic aminoglycoside 3'-phosphotransferase gene, aph. The expression of this construct results in the production of a chimeric protein capable of inactivating the G418 antibiotic. The plasmid was intro-
The hmg1 Gene

A, exon-intron organization of the hmg1 gene; exons are indicated by boxes (open for translated regions; filled for untranslated regions). H, HindIII sites; N, NotI sites. B, sequence alignment of HMG box A and HMG box B. The locations of the introns are indicated by arrows. The position of the intron in the single HMG box coding sequence in the TCF-1 gene (28) is different yet again.

Fig. 3. Structure of the hmg1 gene. A, exon-intron organization of the hmg1 gene; exons are indicated by boxes (open for translated regions; filled for untranslated regions). H, HindIII sites; N, NotI sites. B, sequence alignment of HMG box A and HMG box B. The locations of the introns are indicated by arrows. The position of the intron in the single HMG box coding sequence in the TCF-1 gene (28) is different yet again.

Fig. 4. Primer extension of the HMG1 mRNA. A, 20 pg of total RNA from NIH3T3 cells was hybridized with oligonucleotide MnlI, mapping to exon 2. B, major predicted transcription start sites are indicated by closed dots; an open dot indicates a possible weaker start site. An asterisk indicates the start of the longest mouse HMG1 cDNA clone. CCAAT sites are underlined.

duced in NIH3T3 cells, and G418-resistant clones were selected (Fig. 5). Plasmids pRSVneo and pSV2neo, in which the aph gene is under the control of the Rous sarcoma virus LTR or the SV40 promoter, respectively, were used as positive controls. In several independent experiments, transfection with the pHMG1-neo plasmid gave rise to a large number of resistant clones, almost comparable with the number of clones arising after transfection with the positive controls. Transfection with the bacterial aph gene with no eukaryotic sequences in cis, or with promoterless derivatives of the pHMG1-neo plasmid, gave rise to no resistant clones or to very few. Northern blot analysis of total RNA extracted from three clones stably transfected with pHMG1-neo and three clones stably transfected with pRSVneo confirmed that transcripts of the expected length and containing the aph sequence were indeed present (Fig. 5B and results not shown).

In order to estimate the strength of the hmg1 promoter, we compared by RNase protection the relative abundance of the transcripts produced under its control with the abundance of those produced under the control of the RSV long terminal repeat (Fig. 6). Total RNAs were isolated from NIH3T3 cells 48 h after transfection. Transcripts from the endogenous hmg1 genes give rise to a 45-nt band (not shown in Fig. 6), the RSV-neo mRNA gives rise to a 177-nt band, and the HMG1-neo chimeric mRNA gives rise to a 284-nt band. The intensities of the signals from a cotransfection (lane 1) point to a high level of expression of the chimeric mRNA and indirectly to a high activity of the hmg1 promoter.

The region immediately upstream of the proposed transcription start sites does not contain any sequence conforming to the consensus TATAA; this promoter is probably TATA-less, as is often the case for promoters of housekeeping genes (23). However, it does contain several CAAT boxes, which might promote transcription by binding any one of several factor types (24). Since its promoter can direct the expression of reporter genes placed under its control, we have no doubt that the gene we have isolated is active and encodes the HMG1 protein. The hmg1 gene is expressed at high but not identical levels in all tissues of the mouse embryo at day 10.5 post coitum; likewise, we and others found in all tissues that were investigated the typical set of three HMG1 mRNAs due to differential usage of three alternative polyadenylation signals (14, 25). Being quite active and compact, the hmg1 promoter should be useful to direct the ubiquitous expression of transgenes.

How Many hmg1 Genes Are Present?—Although the data reported in the preceding paragraphs indicate that we have isolated a bona fide hmg1 gene, additional hmg1-related genes may exist, especially if they are intronless like the genes coding for SRY and SOX HMG box proteins (26). We did not find direct evidence for additional active genes; the fragments amplified by PCR were all colinear and belonged to the same hmg1 gene

2 M. Gulisano and M. E. Bianchi, unpublished results.
number of G418-resistant clones

|       | exp I | exp II |
|-------|-------|--------|
| HMG1-neo| > 200 | > 200  |
| ΔBamHI-neo| 12    |        |
| ΔEcoRI-neo| 0     |        |
| RSV-neo | 65    |        |
| SV2-neo | > 200 |        |

Fig. 5. The hmg1 promoter is active in NIH3T3 cells. A, NIH3T3 cells were transfected in duplicate with the indicated plasmids. In experiment I 3.5 × 10⁶ cells were transfected with 20 μg of the expression vectors, and in experiment II 7 × 10⁶ cells were transfected with 5 μg of the expression vectors. Stably transfected clones were selected by adding 900 μg/ml G418 24 h after transfection and were counted 11 days later. Plasmid pHMG1-neo contains the aph gene driven by the NotI-NotI promoter fragment; plasmids pΔBamHI-neo and pΔEcoRI-neo are derived from pHMG1-neo by deletion of a 1.5-kb BamHI fragment and a 3-kb EcoRI fragment, respectively. pRSV-neo and pSV2-neo contain the RSV and SV40 promoters.

Later, they should have accumulated base changes and additional nucleotide sequences in a topology that minimizes the number of substitutions from an ancestral sequence. It is clear that the HMG1 cDNA sequence could not correspond to the sequence from which the other ones were derived independently (an example of the trees generated is shown in Fig. 7B).

By default, we conclude that some or all of the HMG1-related sequences may derive from a gene distinct from hmg1 and which we code name hmg-Z (Fig. 7C). hmg-Z cannot code for the HMG1-related cell surface proteins; it is nonetheless much more closely related to the hmg1 than to the hmg2 gene, which diverged from a common ancestor at least 200 million years ago (27). Northern analysis of total RNA from adult or embryonic mouse tissues gives no evidence of additional mRNA species beyond those deriving from the hmg1 gene, and the selection of HMG1-related clones from cDNA libraries consistently leads to the identification of sequences deriving from the hmg1 gene. This implies that the expression of the hmg-Z gene and/or of the genes coding for cell surface proteins very closely related to HMG1 may be restricted to limited districts or times or may be totally absent nowadays.
Fig. 7. Comparison of HMG1-related sequences. A, alignment of the nine HMG1-R sequences to HMGl cDNA. Only the conceptual translation of the codons with non-silent substitutions with respect to the HMGl cDNA is given here; most of the HMGl-R sequences contain frameshifts too. The nucleotide sequences are deposited at the EMBL data library with accession numbers X80459-X80467, and their similarity to the cDNA ranges between 90 and 99.5%. Z represents a stop codon; a dash indicates that the sequence corresponding to the indicated position was not determined. A group of five sequences contains the same pattern of mutations. B, the maximum parsimony algorithm TREECOM was applied to the HMGl cDNA and HMGl-R sequences. The predicted evolutionary tree shows that HMGl cDNA does not correspond to the sequence from which the HMGl-R sequences were derived, with the exception of HMGl-R-227. C, a model for the evolution of HMG1-R pseudogenes. hmgl and hmgl2 genes probably were derived from a common ancestral gene, hmgl-X, as indicated by the high sequence similarity and the identical exon-intron organization. We suppose that another gene, hmgl-Z, itself derived from hmgl or directly from hmgl-X might have originated some of the HMGl-R sequences.

Acknowledgments—We thank Dr. Pierre Ferrier for hospitality (to C.), Dr. Claudio Bandi for help in the evolutionary analysis, and all the members of our laboratory for helpful suggestions and critical reading of the manuscript.

REFERENCES
1. Bustin, M., Lehn, J.-K., and Landsman, D. (1990) Biochim. Biophys. Acta 1049, 231–243.
2. Bianchi, M. E., Beltame, M., and Fabriola, L. (1992) in Nucleic Acids and Molecular Biology (Eckstein, F., and Lilley, D. M. J., eds) Vol. 6, pp. 112–128, Springer-Verlag, Berlin.
3. Sheflin, L. G., and Spaulding, S. W. (1989) Biochemistry 28, 5658–5664.
4. Stros, M., Stokrovs, J., and Thomas, J. O. (1994) Nucleic Acids Res. 22, 1044–1051.
5. Bianchi, M. E., Beltame, M., and Ponzessa, G. (1989) Science 243, 1056–1059.
6. Bianchi, M. E., Fabriola, L., Ferrari, S., and Lilley, D. M. J. (1991) EMBO J. 11, 1055–1063.
7. Fial, P. M., and Lippard, S. J. (1992) Science 256, 254–256.
8. Bianchi, M. E. (1994) Mol. Microbiol. 13, in press.
9. Ottate, S. A., Prendergast, P., Wagner, J. P., Nissen, M., Reeves, R., Pettijohn, D. E., and Edwards, P. F. (1994) Mol. Cell. Biol. 14, 3376–3391.
10. Ner, S. S., and Travers, A. A. (1994) EMBO J. 13, 1817–1822.
11. Merenmisies, J., Pihlaskari, R., Laitinene, J., Wartiovaara, J., and Rauvala, H. (1991) J. Biol. Chem. 266, 16722–16729.
12. Parkkinnen, J., Raulo, E., Merenmisies, J., Nelo, H., Rajander, E. O., Baumann, M., and Rauvala, H. (1993) J. Biol. Chem. 268, 19726–19738.
13. Chomczynski, P., and Sacchi, N. (1987) Anal. Biochem. 162, 156–159.
14. Wen, L., Huang, J.-K., Johnson, B. H., and Reec, G. R. (1988) Nucleic Acids Res. 17, 1197–1214.
15. Truda, K., Kikuchi, M., Mori, K., Waga, S., and Yoshida, M. (1988) Biochemistry 27, 6159–6163.
16. Yotev, W. V., and St-Arnaud, R. (1992) Nucleic Acids Res. 20, 3516.
17. Shirakawa, H., and Yoshida, M. (1992) J. Biol. Chem. 267, 6641–6646.
18. Kolodrubets, D., and Burgum, A. (1989) J. Biol. Chem. 265, 3234–3239.
19. Roth, S. Y., Schulman, I. G., Cook, R. G., and Allis, C. D. (1987) Nucleic Acids Res. 15, 8112.
20. Graesser, R., and Feix, G. (1991) Nucleic Acids Res. 19, 2573–2577.
21. Wagner, C. R., Hanama, K., and Elgin, S. C. R. (1992) Mol. Cell. Biol. 12, 1015–1023.
22. Ner, S. S., Churchill, M. A., Searles, M. A., and Travers, A. A. (1993) Nucleic Acids Res. 21, 4369–4371.
23. Maniatis, T., Goodbourn, S., and Fischer, J. A. (1987) Science 236, 1237–1245.
24. Locke, J. (1993) in Gene Transcription: A Practical Approach (Hames, B. D., and Higgins, S. J., eds) pp. 320–345, Oxford University Press, Oxford.
25. Lee, K.-L., Pentecost, B. T., Amano, J. A., Tobey, R. A., Gurlay, L. R., and Dixon, G. H. (1987) Nucleic Acids Res. 15, 5051–5058.
26. Gubbi, J., Vivas, N., Economou, A., Jackson, D., Goodfellow, P., and Lovell-Badge, R. (1992) Proc. Natl. Acad. Sci. U. S. A. 89, 7953–7957.
27. Laudet, V., Schilden, H., and Clevers, H. (1993) Nucleic Acids Res. 21, 2498–2501.
28. van de Wetering, M., Oosterwegel, M., Holstege, F., Dooye, D., Suijk, R., van Kessel, A., and Clevers, H. (1992) J. Biol. Chem. 267, 8530–8536.