Characterization of the Mouse cDNA and Gene Coding for a Hepatocyte Growth Factor-like Protein: Expression during Development[†,‡]

Sandra J. Friezner Degen,* Lorie A. Stuart, Su Han, and C. Scott Jamison
Division of Basic Science Research, Children's Hospital Research Foundation and Developmental Biology Graduate Program, University of Cincinnati, Cincinnati, Ohio 45229

Received May 15, 1991; Revised Manuscript Received July 23, 1991

ABSTRACT: The cDNA and gene coding for mouse hepatocyte growth factor-like protein (HGF-like protein) were isolated and characterized. The size of the gene from the site of initiation of transcription to the polyadenylation site is 4613 bp in length and is composed of 18 exons separated by 17 intervening sequences. The exons range in size from 36 to 227 bp in length, while the intervening sequences range in size from 78 to 613 bp in length. The site of initiation of transcription was identified by primer extension analysis using total RNA isolated from mouse liver. On the basis of these results, the first exon is 146 bp in length and includes 94 bp of 5′-noncoding sequence. The sequence $^{5'}$TATGTG$^{3'}$ is present between 34 and 39 bp upstream of the transcription start site and could potentially be the TATA sequence found for many constitutively expressed eukaryotic genes to be the promoter for RNA polymerase II. The sequence $^{5'}$GCAAT$^{3'}$ at −96 to −92 may be the CCAAT sequence responsible for stimulation of transcription of some eukaryotic genes. The same sequences in the Genbank and NBRF databases were homologous to similar regions in the genes coding for both human and mouse HGF-like protein (Han et al., 1991). The acyl-peptide hydrolase gene is 410 bp downstream of the mouse HGF-like protein, but is transcribed from the complementary strand. The mouse cDNA for HGF-like protein codes for a putative protein with the same domain structure as its human homologue with four kringle domains followed by a serine protease-like domain. On the basis of the translated sequence of the cDNA, the mouse HGF-like protein would be 716 amino acids in length with a molecular weight of 80K. There are four potential N-linked carbohydrate attachment sites. The DNA and amino acid sequences of mouse HGF-like protein are compared to the human protein. Overall, the two proteins are about 80% identical with each other. In contrast to mRNA for human HGF-like protein, which is 2.4 and 3.0 kilobases in length in human liver, only the smaller species is seen in mouse and rat liver. The expression pattern of mRNA coding for HGF-like protein during development and in maternal rats was determined by Northern analysis. It is apparent that the majority of mRNA coding for HGF-like protein is expressed in liver. Messenger RNA is also expressed at a lower level in lung, adrenal, and placenta.

In the preceding paper (Han et al., 1991), we described the isolation and characterization of a human gene and cDNA coding for a protein with similar domain structure as hepatocyte growth factor (HGF)[1] with four kringle domains followed by a serine protease domain (Nakamura et al., 1989). HGF functions as a growth factor for a broad spectrum of tissue and cell types (Tashiro et al., 1990). Although we do not know the function of HGF-like protein, on the basis of the similar domain structure of this protein with HGF, we proposed to tentatively call it HGF-like protein.

The kringle domains in human HGF-like protein are 33–66% identical with kringles in other human kringle-containing proteins (Magnusson et al., 1975), while the serine protease-like domain is 30–45% identical with other serine proteases. The active-site amino acids have been changed from His to Gln, from Asp to Gln, and from Ser to Tyr in human HGF-like protein, so it is unlikely that this protein has proteolytic activity.

Database searches identified the presence of sequence for the DNF15S1 and DNF15S2 loci at the 3′ end of the gene. DNF15S1 and DNF15S2 are homologous loci found on human chromsomes 1 and 3, respectively (Welch et al., 1989). On the basis of sequence similarity and extensive restriction map information for the DNF15S2 locus (Welch et al., 1989), we inferred that the gene for HGF-like protein was present at this locus at 3p21.

Probes from the DNF15S2 locus have been used as restriction fragment length polymorphism markers for deletions in the short arm of human chromosome 3 that are associated with various carcinomas. This region is deleted in small cell lung carcinoma (SCLC; Whang-Peng et al., 1982; Naylor et al., 1987), other lung cancers (Kok et al., 1987; Brauch et al., 1987), renal cell carcinoma (Zbar et al., 1987; Kovacs et al., 1988), and von Hippel–Lindau syndrome (Seizinger et al., 1988), which suggests that one or more tumor suppressor genes are at this locus. When expressed, tumor suppressor genes play regulatory roles in cell proliferation, differentiation, and other cellular processes. Oncogenesis occurs when these genes are inactivated or lost due to chromosomal deletion. Genes that have presently been identified as tumor suppressors are involved in cell cycle control, signal transduction, angiogenesis, and development (Sager, 1989).

Since a tumor suppressor gene(s) may be located at or near the DNF15S2 locus on human chromosome 3, it is of interest

[†] This work was supported in part by the Pew Memorial Trust, by Research Grant HL38232 from the National Institutes of Health, by a grant-in-aid from the American Heart Association, and by Molecular and Cellular Biology Post-Doctoral Training Grant NIH HL07527 (C.S.J.). S.J.F.D. is a Pew Scholar in the Biomedical Sciences and an Established Investigator of the American Heart Association.

[‡] The nucleic acid sequences in this paper have been submitted to GenBank under Accession Numbers M74181 (cDNA) and M74180 (gene).

[1] Abbreviations: bp, base pair(s); EDTA, ethylenediaminetetraacetic acid; HGF, hepatocyte growth factor; kbp (kb in figures), kilobase pair(s); kDa, kilodalton(s); SDS, sodium dodecyl sulfate; Tris-HCl, tris(hydroxymethyl)aminomethane hydrochloride.

0006-2960/91/0430-9781$02.50/0

to learn more about the biology of HGF-like protein. Its similarity in structure to a known growth factor is interesting since cell proliferation is regulated by growth factor receptors. It is interesting to speculate that HGF-like protein functions as a competitive inhibitor for a growth factor receptor. When this protein is absent due to a chromosomal deletion, the growth factor is free to bind to its receptor, and uncontrolled growth may occur that results in neoplasia.

In this paper, we present the DNA sequences of the gene and cDNA coding for mouse HGF-like protein. Probes were isolated from the cDNA in order to determine the developmental expression pattern and the tissue distribution of mRNA coding for HGF-like protein in the rat. Maternal tissues were also analyzed in order to determine the effects of pre- and postparturitional stress on HGF-like protein.

Materials and Methods

General cloning procedures, restriction enzyme analysis, plasmid purification procedures, and phage DNA preparation have been described previously (Degen et al., 1983; Degen & Davie, 1987).

Probes. A 340 bp fragment from the cDNA coding for the human HGF-like protein (cDNA 33; Han et al., 1991) was isolated after digestion with *Eco*RI and *Kpn*I. This fragment coded for the amino-terminal portion of the protein including eight amino acids of the first kringle. The 740 bp insert from the cDNA coding for mouse HGF-like protein (pML5-2/740) was isolated after digestion with *Eco*RI and coded for the amino-terminal portion of the protein including all of the first kringle and most of the second kringle domain (Figure 1). A 1450 bp insert was isolated after digestion of the mouse cDNA coding for the HGF-like protein (pML5-2) with *Eco*RI. This probe coded for eight amino acids of the second kringle of the mouse protein, all of the third and fourth kringles, and the serine protease-like domain (Figure 1). A fragment containing exon 1 from the gene coding for mouse HGF-like protein was isolated after digestion of the subclone pmgL5-12Bam1.6 with *Bam*HI and *Eco*RI. The resulting 396 bp fragment contained sequence from −105 to +291 as shown in Figure 5. The 2000 bp insert from a full-length human prothrombin cDNA was isolated after digestion with *Eco*RI. All fragments were isolated after polyacrylamide gel electrophoresis followed by electroelution and purification over Elutip-D columns (Schleicher & Schuell). Fragments were radiolabeled with [^{32}P]αCTP (NEN Dupont) by using the random primer labeling procedure (Feinberg & Vogelstein, 1984).

Isolation of the cDNA Coding for Mouse HGF-like Protein. A C57BL/6 mouse liver cDNA library (Stratagene, La Jolla, CA) was screened for the cDNA coding for mouse HGF-like protein. The library was originally constructed with cDNAs greater than 1000 bp in length cloned into the *Eco*RI site of λgt10 after addition of *Eco*RI linkers. Approximately 10^6 phage were screened with a 340 bp probe isolated from the 5′ end of the cDNA coding for human HGF-like protein (see Probes) at reduced stringency to allow for cross-species hybridization (Degen et al., 1990). Ten positives were identified, and eight were plaque-purified. Most phage contained two *Eco*RI inserts of 1450 and 740 bp. These fragments from phage ML5-2 were individually subcloned into the *Eco*RI site of pBR322. In addition, a 1520 bp *Xho*I–*Kpn*I fragment from phage ML5-2 (Figure 1) that contained the internal *Eco*RI site was subcloned into Bluescript SK +/− (Stratagene).

Isolation of the Gene Coding for Mouse HGF-like Protein. A Balb/c mouse liver genomic DNA library (Clontech) was screened for the gene coding for mouse HGF-like protein. This library was constructed with partial Sau 3A genomic fragments ranging in size from 8 to 21 kbp in length cloned into the *Bam*HI site of EMBL-3 SP6/T7. Approximately 10^6 phage from the library were screened with a 1450 bp probe isolated from the cDNA coding for mouse HGF-like protein (see Probes). On the initial screen, 65 positives were identified; 9 were rescreened and plaque-purified. Phage DNA was purified, and restriction fragments were subcloned into pBR322.

Northern Analysis. Total RNA was isolated from human liver, rat tissues, and HepG2 cells following the procedure of Chomczynski and Sacchi (1987). Details of the Northern analysis procedures have been described previously (Jamison & Degen, 1991). Briefly, samples of total RNA (20 μg) were subjected to electrophoresis in a 1% agarose gel containing 2.4 M formaldehyde. RNA was transferred from the gel to a Biotrans membrane (ICN Biochemicals). Blots were hybridized with random-primer-labeled 1450 bp insert from the cDNA coding for mouse HGF-like protein and at a later date with a 2000 bp insert from the cDNA coding for human prothrombin (see Probes).

Determination of the Site of Initiation of Transcription. Primer extension of total RNA isolated from mouse liver (10 μg) was performed as described previously (Bancroft et al., 1990) using a 5′-end-labeled oligonucleotide that was complementary to nucleotides 769–798 in the second exon of the gene coding for mouse HGF-like protein (Figure 5). Hybridization of oligonucleotide to RNA was performed at either 45 °C or 60 °C. Products were resolved on 6 and 20% sequencing gels alongside a M13 sequencing ladder for determination of size. *Escherichia coli* tRNA was used as a control during hybridization and primer extension procedures.

DNA Sequence Analysis. DNA sequence was determined by a combination of the chemical modification procedures of Maxam and Gilbert (1980) and the quasi-end-labeling modification of the enzymatic dideoxy-chain termination procedure (Duncan, 1985). All sequences were analyzed on an IBM-AT computer using the Microgenie program (Queen & Korn, 1984; Beckman Instruments).

Developmental Studies in the Rat. Northern blots previously used to study the developmental expression of rat prothrombin in various pre- and postnatal tissues as well as maternal tissues taken at the same time points were used for studies on the developmental expression of rat HGF-like protein (Jamison & Degen, 1991). Day 17 timed pregnant female SD rats were obtained from Harlan Sprague Dawley (Indianapolis, IN). Surgical procedures, RNA isolation, and Northern analysis have been described previously (Jamison & Degen, 1991). A 1450 bp *Eco*RI insert from the cDNA coding for mouse HGF-like protein was used as a probe (see Probes).

Total RNA was isolated from brain, heart, aorta, lung, diaphragm, liver, spleen, stomach, small and large intestine, kidney, and adrenal tissues at the developmental stages indicated in Table II.

Total RNA was isolated from brain, heart, lung, diaphragm, liver, spleen, stomach, small intestine, large intestine, kidney, adrenal, ovary, uterus, placenta, and urinary bladder from a single maternal rat for each of the stages of pregnancy and after delivery indicated in Table III.

Results

Characterization of the cDNA Coding for Mouse HGF-like Protein. A partial restriction map and sequencing strategy for the longest cDNA (pML5-2) coding for mouse HGF-like protein are shown in Figure 1. This cDNA is 2188 bp in length and includes an open reading frame of 2104 bp followed

Table I: Comparison of Genes Coding for Mouse and Human HGF-like Protein

exon	size (bp) human	size (bp) mouse	sequence identity (%)	intron	size (bp) human	size (bp) mouse	type[a]	sequence identity (%)
1	52+[b]	146	78.8[c]	A	697	613	I	59.4
2	148	148	84.5	B	80	92	II	60.9
3	113	113	78.8	C	77	84	I	66.7
4	115	115	81.7	D	77	84	II	60.7
5	137	137	83.9	E	79	81	I	59.0
6	121	121	76.9	F	144	81	II	42.4
7	119	119	80.7	G	202	143	I	47.0
8	169	196	77.5[d]	H	120	159	II	55.0
9	131	131	76.3	I	97	78	I	60.8
10	103	103	84.5	J	127	122	II	54.3
11	137	137	76.6	K	89	88	I	69.2
12	36	36	88.9	L	88	79	I	58.4
13	121	109	79.8[e]	M	175	161	II	62.7
14	78	78	87.2	N	127	119	II	60.5
15	147	147	81.6	O	81	80	II	70.4
16	107	107	86.0	P	95	98	I	69.3
17	140	140	82.9	Q	119	141	O	62.4
18	242	227	76.0					

[a] Type I intervening sequences interrupt codons between the first and second base, type II between the second and third base, and type O between codons (Sharp, 1981). [b] The size of this exon was based on the distance from the codon for the initiator methionine to the 3′ end of exon 1; the length of the 5′-flanking region of the human gene has not been determined. [c] Sequence was compared from the codon for the initiator methionine to the 3′ end of exon 1 for both genes. [d] The mouse gene has 27 additional bases at the 5′ end; these were not included in the comparison. [e] The human gene has 12 extra bases that were not included in the comparison.

by a stop codon and a 3′-noncoding region of 65 bp (Figure 2). The 5′CATAAA3′ sequence present 20–25 bases upstream of the poly(A) tail is the apparent polyadenylation signal (Figure 2). This is also conserved in the mRNA coding for human HGF-like protein (Han et al., 1991). This cDNA was not full-length since the opening reading frame was present at the 5′ end of the sequence with no codon for the initiator methionine in-frame with the coding sequence. After determination of the sequence of the gene coding for mouse HGF-like protein (see below), it was determined that the cDNA lacked 44 bp of coding and 94 bp of 5′-noncoding sequence at its 5′ end. Several attempts were made to isolate a full-length cDNA without success. The sequence of the cDNA (including the sequence from exon 1) and its translated amino acid sequence are shown in Figure 2 compared to the sequence of the human cDNA. The mouse cDNA codes for four kringle domains followed by a serine protease-like domain.

Northern analysis of mouse liver total RNA indicated that there was one species of mRNA coding for mouse HGF-like protein with a size of approximately 2400 bases (Figure 3A). This is in agreement with the size of the mouse cDNA plus the additional 138 bp identified in the gene to be part of the mRNA. This is in contrast to human liver where at least two sizes of mRNA coding for HGF-like protein are present (2.4 and 3.0 kilobases; Han et al., 1991). The cDNA coding for mouse HGF-like protein hybridizes to similar size mRNA in rat and mouse liver and the same multiple banding pattern in human liver as seen with the human HGF-like cDNA probe. The mouse probe did not detect any hybridizing mRNA in HepG2 cells. A human cDNA probe also detected very little mRNA coding for HGF-like protein in HepG2 cells [Figure 4 in Han et al. (1991)]. The human prothrombin cDNA was hybridized to the same Northern blot to show that RNA was present in the HepG2 lane (Figure 3B).

Organization of the Gene Coding for Mouse HGF-like Protein. A partial restriction map and sequencing strategy for the gene coding for mouse HGF-like protein are shown in Figure 4. The complete sequence of the gene was determined (Figure 5). The size of the gene from the site of initiation of transcription (see below) to the polyadenylation site is 4613 bp in length. In addition, 1191 and 947 bp of 5′- and 3′-flanking sequence were determined, respectively, for

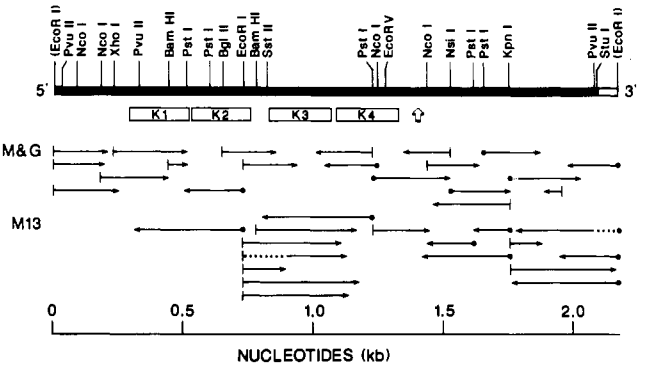


FIGURE 1: Partial restriction map and sequencing strategy for the cDNA coding for mouse HGF-like protein. Restriction sites are shown above the bar representing the cDNA. The *Eco*RI sites at each end are in parentheses since these are present only because of the linkers added during construction of the cDNA library. The darkened area represents the open reading frame potentially coding for protein, and the open bar represents the 3′-noncoding region. The 5′ and 3′ indicate the orientation of transcription. Domains that are coded for by the cDNA are shown below the bar. The four kringle domains are labeled K1, K2, K3, and K4; the potential activation site is indicated by an arrow. The DNA sequencing strategy is shown for both the chemical modification (M & G) and dideoxy sequencing procedures (M13). Sequences determined on the coding or complementary strand are indicated with vertical lines or circles at the end of the arrow, respectively. Dotted parts of arrows indicate regions not determined for that labeling. One hundred percent of the sequence was determined 2 times or more, and 76% was determined on both strands. All overlaps were obtained. The scale in kilobases is indicated.

a total of 6751 bp of contiguous sequence. Comparison of the DNA sequence of the gene with the cDNA indicated that the gene was composed of 18 exons separated by 17 intervening sequences. The exons range in size from 36 to 227 bp in length while the intervening sequences range in size from 78 to 613 bp in length (Table I). Exons 1 and 18 contain 5′- and 3′-noncoding sequence, respectively.

The first exon was identified by comparison with the 5′ end of the cDNA coding for human HGF-like protein (Han et al., 1991) since the mouse cDNA was shorter and would only include 8 bp in this exon (compared to 36 bp in the human cDNA). There is 67% homology between the 36 bp at the 5′ end of the human cDNA and the 3′ end of exon 1 (nucleotides 111–146; Figure 5). Upstream of this region is an

FIGURE 2: Comparison of cDNAs coding for human and mouse HGF-like protein and their translated sequences. The cDNA and translated amino acid sequence for human and mouse HGF-like protein are presented (since neither the human or the mouse cDNAs were full-length, additional sequence is included from exon 1 of both genes; Han et al., 1991). Nucleotides and amino acids in the cDNA and its translated sequence for mouse HGF-like protein that differ from the human sequences are shown below and above the human sequences, respectively. Numbering of the nucleotide sequence is in the right margin for the nucleotide at the end of each line. Every 50 amino acids are numbered starting with the putative initiator methionine at 1. The stop codon is indicated by three asterisks. The 5′ end of the cDNA for mouse HGF-like protein starts at nucleotide 45. Amino acids that are deleted when the two proteins are compared are represented by xxx. Deleted nucleotides are represented by an x. The four kringle domains are marked, and the putative activation site is indicated by the arrow. Amino acids at 531, 577, and 670 correspond to the active-site His, Asp, and Ser, respectively, in the active site of serine proteases (in boxes). The four potential N-glycosylation sites in the mouse protein are residues 72, 173, 305, and 624. Residues 72, 305, and 624 are potential glycosylation sites in the human protein. Amino acid residue 19 is a Pro in the cDNA for mouse HGF-like protein while Gln is coded for in the gene.

in-frame ATG codon coding for methionine (nucleotides 95–97; Figure 5). In addition, immediately 3′ to the 36 bp of homology between the human cDNA and the gene is the traditional GT splice site sequence found at the 5′ end of intervening sequences (Breathnach et al., 1978). Location of the splice site at this position makes it a type I site; a type I site is also present at the 3′ end of this intervening sequence which keeps the open reading frame intact.

Northern analysis of total RNA isolated from mouse liver using a probe from the mouse gene coding for HGF-like protein that includes exon 1 and its flanking sequences but no other exons (see Materials and Methods; Probes) identified one hybridizing band at 2400 bases, the same size as the mRNA observed when cDNA probes for mouse HGF-like

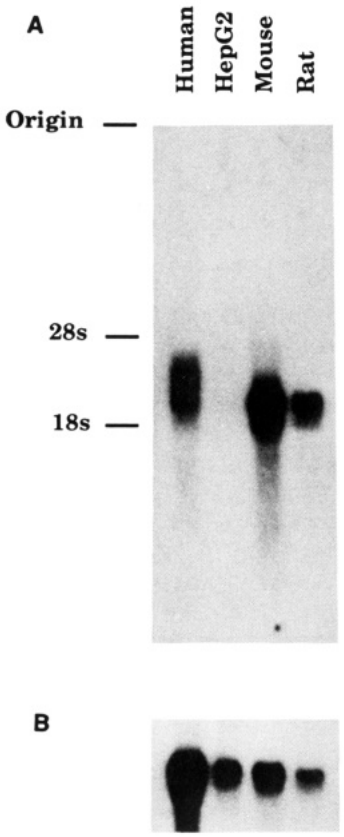


FIGURE 3: Northern analysis of total RNA isolated from mouse, rat, and human liver and HepG2 cells. Samples of total RNA (20 μg) isolated from human liver (Human), human hepatoblastoma cells (HepG2), mouse liver (Mouse), and rat liver (Rat) were subjected to electrophoresis, transferred to a Biotrans membrane, and hybridized with (A) ^{32}P-labeled 1450 bp cDNA probe coding for mouse HGF-like protein (see Materials and Methods) or (B) ^{32}P-labeled cDNA probe coding for human prothrombin. The migration of 28S and 18S ribosomal RNA is indicated.

protein (data not shown) were used. Thus, this sequence is part of the mRNA and therefore belongs to the exon sequence.

The sequence surrounding the proposed codon for the initiator methionine is $^{5'}$GGAGAATGG$^{3'}$ at positions 90–98 in Figure 5. Five of the eight bases agree with the consensus sequence compiled by Kozak (1986) of $^{5'}$CCACCATGG$^{3'}$. According to Kozak (1986), positions −3 and +4 (with the A of ATG being +1) have been found to be critical for the use of this ATG as the initiator methionine codon. These two bases in the gene coding for mouse HGF-like protein agree with the consensus. There is one other ATG upstream of the proposed initiator codon, but out-of-frame with the coding sequence (bases 52–54; Figure 5). The sequence surrounding this ATG is six out of eight bases identical with the consensus with the important bases for recognition conserved. There is a stop codon in-frame with this ATG 34 bp downstream (bases 88–90; Figure 5).

The sequence of splice junctions at the 5′ and 3′ ends of each intervening sequence agree with the $^{5'}$GT-AG$^{3'}$ rule of Breathnach et al. (1978) and the consensus sequences compiled by Mount (1982) except in two cases. The 5′ end of intervening sequence C has a GC at this site (nucleotides 1113–1114; Figure 5) rather than a GT, and an AAG is present at the 3′ end of intervening sequence O rather than C/T AG (nucleotides 3898–3900; Figure 5). These same sequences are also present in the human gene (Han et al., 1991). The human gene has one additional difference from the consensus with an AAG present at the 3′ end of intervening sequence G. The corresponding sequence in the mouse gene is a CAG (nucleotides 2075–2077; Figure 5).

There was only one difference found when the sequences of the cDNA and gene coding for mouse HGF-like protein were compared. There is an A at position 763 in exon 2 of the gene (Figure 5) while there is a C at this position in the cDNA (nucleotide 56 in Figure 2). The amino acid coded for in the gene is a Gln while a Pro is coded for in the cDNA at residue 19. This difference is probably in the signal peptide of the synthesized protein (see Discussion). A Gln is coded for at this same position in the cDNA for human HGF-like protein (Figure 2).

Identification of the Site of Initiation of Transcription. Primer extension analysis of mouse liver RNA using an oligonucleotide complementary to the sequence in exon 2 of the gene coding for mouse HGF-like protein revealed a band of 185 bp in length (Figure 6). This corresponds to a start site

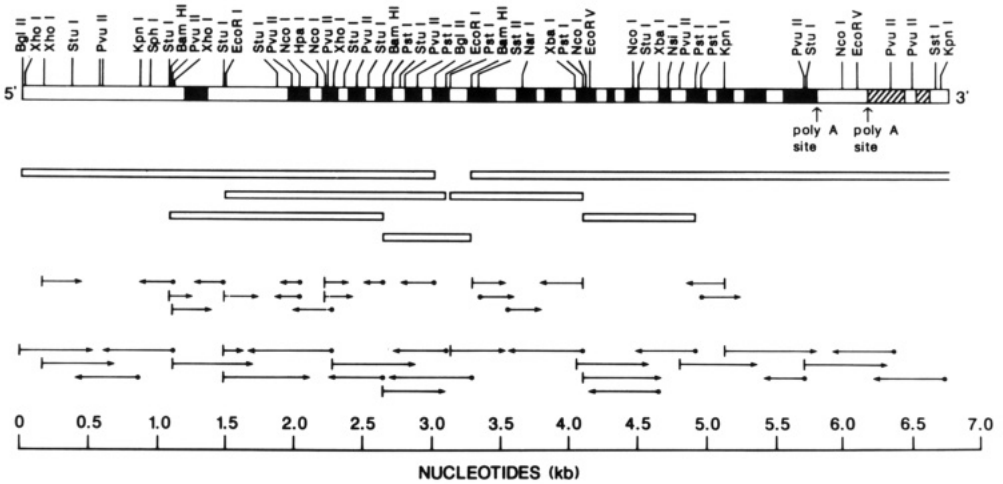
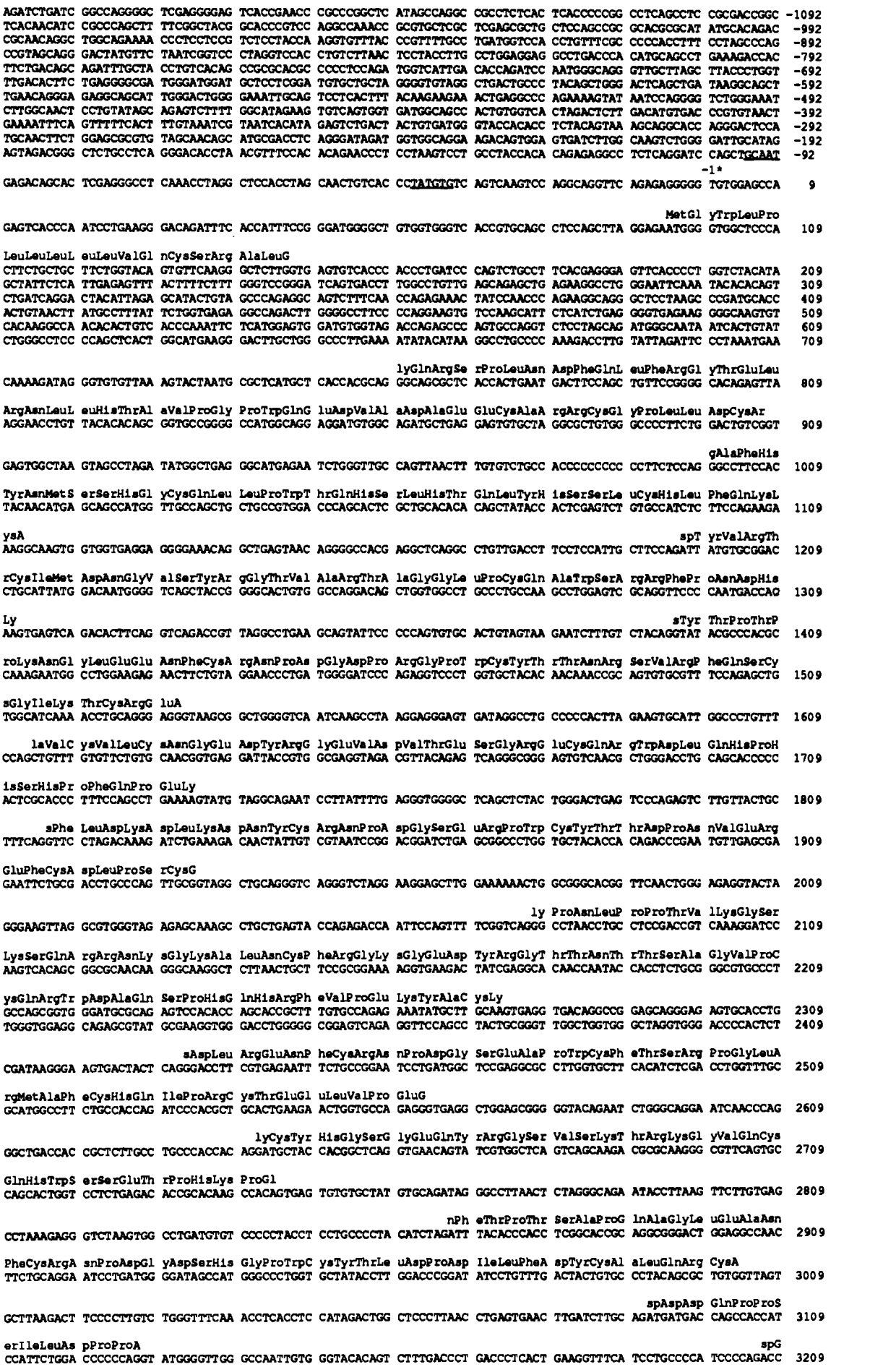


FIGURE 4: Restriction map and sequencing strategy for the gene coding for mouse HGF-like protein. A partial restriction enzyme map is indicated above the bar. Darkened areas on the bar represent exons coding for mouse HGF-like protein, and the spaces in between represent the intervening sequences. Hatched boxes represent exons coding for the acyl-peptide hydrolase gene. Polyadenylation sites for both genes are indicated. The orientation of transcription of the gene for HGF-like protein is indicated by the 5′ and 3′ at each end. The extent of gene fragments subcloned into pBR322 is shown by the open bars. One subclone (indicated by the open-ended bar at its 3′ end) extends beyond the restriction map. The sequencing strategies for both chemical modification (top set of arrows) and dideoxy procedures (bottom set of arrows) are indicated. Sequences determined on the coding and complementary strands are indicated by a vertical bar or circle at the labeling or cloning site, respectively. Dotted parts of arrows indicate regions where the sequence was not determined for that experiment. The sequence was determined 2 times or more for 87% of the sequence; 60% was determined on both strands. The scale in kilobases is indicated.



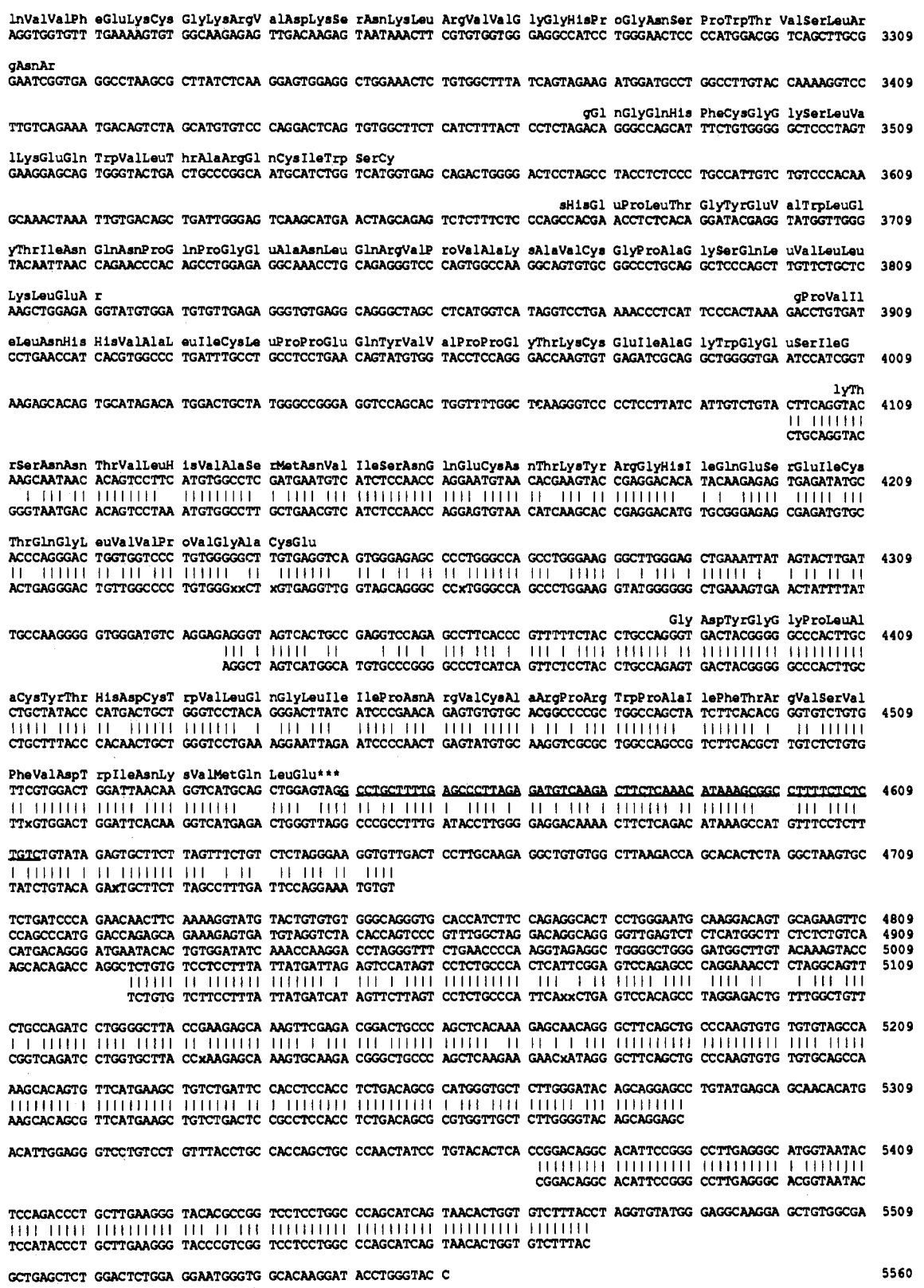


FIGURE 5: Sequence of the gene coding for mouse HGF-like protein. The nucleotide sequence of the gene coding for mouse HGF-like protein is shown along with 5′- and 3′-flanking regions. An asterisk indicates the putative site of initiation of transcription (+1) and therefore the beginning of exon 1. Inferred promoter sequences ("TATA" and "CCAAT") are underlined. DNA sequences upstream of the start site are numbered in a negative manner. The number in the right margin corresponds to the last nucleotide on each line. The amino acid sequence of the exons is indicated above the DNA sequence; splice junctions occur immediately 5′ and 3′ to these sequences. The stop codon is indicated by three asterisks. The 3′-noncoding region is underlined. Intervening sequences that interrupt a codon are shown by the presence of a partial amino acid triplet. Sequences in the database found to be homologous to regions of this gene are shown below the sequence of the gene. Identical nucleotides are indicated by a vertical line. Sequence from 4100–4309 is homologous to nucleotides 1–207 of DNF15S1 (Welch et al., 1989); nucleotides 4335–4654 are homologous to 216–431 and 1–153 of DNF15S1 and DNF15S2, respectively (Welch et al., 1989); nucleotides 5024–5288 and 5371–5477 are homologous to nucleotides 2099–2360 and 1992–2098, respectively, on the complementary strand of rat acyl-peptide hydrolase (Kobayashi et al., 1989). Insertions are not indicated; there is one insertion in the region of DNF15S1 homology, six in the DNF15S2 region, and one in the first region of homology with the rat acyl-peptide hydrolase gene.

Table II: Summary of Northern Analyses for HGF-like Protein: Fetal and Newborn Tissues

day of gestation[b]	fetal/newborn tissues[a]										
	BR	H	A	LG	D	L	SP	ST	I	K	AD
18	ND[c]	–	ND	ND	ND	+	ND	–	–	–	ND
20	–	–	ND	+	–	+	–	–	–	–	–
22	–	–	ND	+	–	+	–	–	–	–	–
B	ND	–	–	ND	ND	+	–	–	–	–	ND
B+5	–	–	ND	+	+	+	–	–	–	–	–
B+13	–	–	ND	–	ND	+	–	–	–	–	–
B+30	–	–	ND	–	–	+	–	–	–[d]	–	ND

[a] Abbreviations: BR, brain; H, heart; A, aorta; LG, lungs; D, diaphragm; L, liver; SP, spleen; ST, stomach; I, intestine; K, kidney; AD, adrenal. The presence (+) or absence (–) of a band for HGF-like protein mRNA on Northern analysis is indicated. [b] Abbreviations: 18, 18th day of gestation; 20, 20th day of gestation; 22, 22nd day of gestation; B, less than 12 h after birth; B+5, 5 days after birth; B+13, 13 days after birth; B+30, 30 days after birth. [c] Not determined. [d] Includes data for RNA isolated from small intestine and large intestine. The large and small intestines were only separated at 30 days after birth. At that time point, there was no detectable HGF-like protein mRNA expression in either small or large intestine.

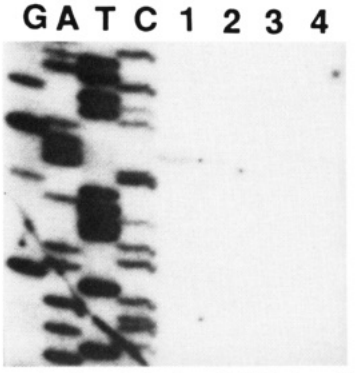


FIGURE 6: Determination of the site of initiation of transcription in the gene coding for mouse HGF-like protein. Primer extension of total RNA isolated from mouse liver (lanes 1 and 2) or *E. coli* tRNA (lanes 3 and 4) was performed by using an oligonucleotide complementary to nucleotides 769–798 in the second exon of the gene coding for mouse HGF-like protein. Lanes 1 and 3 represent hybridizations at 45 °C, and lanes 2 and 4 were hybridized at 60 °C. An M13 sequencing ladder was used for size determination (G, A, T, and C).

at +1 in Figure 5. The sequence at this site does not agree with the consensus sequence derived for transcription initiation sites (a CA followed by several pyrimidines; Bucher & Trifonov, 1986). On the basis of these results, the first exon is 146 bp in length with a 94 bp 5′-noncoding region. The sequence $^{5'}$TATGTG$^{3'}$ is present between 34 and 39 bases upstream of this putative transcription start site and could potentially be the TATA sequence found for many constitutively expressed eukaryotic genes to be the promoter for RNA polymerase II. The sequence $^{5'}$GCAAT$^{3'}$ at positions –96 to –92 (Figure 5) upstream from the putative start site could potentially by the CCAAT sequence responsible for stimulation of transcription of some eukaryotic genes (Bucher & Trifonov, 1986).

Database Search. The DNA sequence of the gene coding for mouse HGF-like protein was compared against the Genbank (release 64) and NBRF databases (release 25) to search for sequences in common with other genes and to locate repetitive sequences. The same sequences were found to be homologous as when the human gene was compared to these databases (Han et al., 1991). Sequences homologous to DNF15S1 and DNF15S2 (Welch et al., 1989), human DNF15S2 lung mRNA (Naylor et al., 1989), and rat acyl-peptide hydrolase mRNA (Kobayashi et al., 1989) were identified in exon 17 to the 3′ end of the sequence presented in Figure 5. DNF15S1 and DNF15S2 are loci on human chromosomes 1 and 3, respectively, while human DNF15S2 lung mRNA is transcribed from the DNF15S2 locus. The region from 4100 to 4309 (Figure 5) in the mouse gene is 78% identical with part of the sequence for DNF15S1. After a gap of 26 bp, the DNF15S1 homology continues from nucleotides 4335–4551 and is 80% identical with DNF15S1. The DNF15S1 homology overlaps with the DNF15S2 repetitive sequence at 4514 (Figure 5). DNF15S2 is homologous to nucleotides 4514–4654 (Figure 5) and is 70% homologous to the mouse gene. The 3′ end of the sequence coding for rat acyl-peptide hydrolase mRNA is homologous to two regions in the gene for mouse HGF-like protein at nucleotides 5024–5288 (85.7%) and 5371–5477 (96.3%; Figure 5) on the complementary strand. These appear to be two exons present at the 3′ end of the gene coding for acyl-peptide hydrolase. There are typical splice junction sequences present for the one complete intervening sequence (5289–5370; Figure 5) and the 3′-splice site of the partial intervening sequences at 5478–5560 (Figure 5). The sequence of human DNF15S2 lung mRNA is homologous to the rat acyl-peptide hydrolase mRNA and is most probably the human homologue of acyl-peptide hydrolase (Han et al., 1991). Human DNF15S2 lung mRNA is 77% and 78% identical with nucleotides 5024–5288 and 5371–5477, respectively (Figure 5). The polyadenylation site in the gene coding for mouse HGF-like protein is 410 bp upstream of the polyadenylation site for the acyl-peptide hydrolase gene. These genes are transcribed on opposite strands.

Developmental Expression. In order to characterize the developmental expression pattern of mRNA coding for HGF-like protein as well as its tissue distribution, total RNA was isolated from various rat tissues and subjected to Northern analysis. In liver, mRNA coding for HGF-like protein was expressed at low levels during gestation until just before birth when levels dramatically increased (Figure 7). As judged by the intensity of autoradiographic bands, levels of HGF-like protein mRNA in liver continued to increase after birth to an apparent maximum at 13 days after birth (Figure 7). At 30 days after birth, HGF-like protein mRNA appeared to be solely expressed in liver and not in any of the other tissues analyzed (Figure 8). Northern analysis was performed for tissues from other stages of development, and these results are summarized in Table II. It is of interest to note that mRNA coding for HGF-like protein is expressed in liver at all of the developmental stages analyzed and in lung at three stages. Messenger RNA coding for HGF-like protein was not expressed in brain, heart, aorta, spleen, stomach, intestine, kindey, or adrenal gland at any stage of development analyzed.

In order to determine the effects of pre- and postparturitional stress on HGF-like protein expression, total RNA was isolated from pregnant and postpartum female rats and subjected to Northern analysis. At day 20 of gestation, mRNA coding for HGF-like protein was expressed mostly in liver and also in lung, diaphragm, adrenal, and placenta (Figure 9). Northern analysis was performed for tissues at other pre- and

Table III: Summary of Northern Analyses for HGF-like Protein: Maternal Tissues

day of gestation[b]	maternal tissues[a]														
	BR	H	LG	D	L	SP	ST	SI	LI	K	AD	O	U	P	UB
P18	ND[c]	ND	ND	ND	+	ND	ND	ND	ND	ND	ND	ND	ND	ND	ND
P20	–	–	+	+	+	–	–	–	ND	–	+	–	–	+	ND
P22	–	–	+	–	+	–	+	–	+	–	+	–	–	+	ND
D	ND	–	ND	–	+	–	–	–	ND	–	ND	ND	–	NP[d]	–
D+5	–	–	+	–	+	–	–	–	–	–	+	–	–	NP	ND
D+13	–	–	+	–	+	–	–	–	–	–	–	–	–	NP	ND

[a] Abbreviations: BR, brain; H, heart; LG, lungs; D, diaphragm; L, liver; SP, spleen; ST, stomach; SI, small intestine; LI, large intestine; K, kidney; AD, adrenal; O, ovary; U, uterus; P, placenta; UB, urinary bladder. Presence (+) or absence (–) of a band for HGF-like protein mRNA on Northern analysis is indicated. [b] Abbreviations: P18, 18th day of pregnancy; P20, 20th day of pregnancy; P22, 22nd day of pregnancy; D, less than 12 h after delivery; D+5, 5 days after delivery; and D+13, 13 days after delivery. [c] Not determined. [d] Tissue not present.

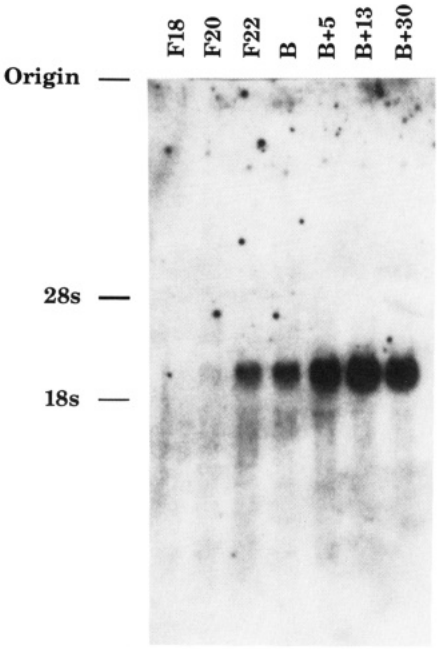


FIGURE 7: Northern analysis of total RNA isolated from rat liver at various stages of fetal and postnatal development. Samples (20 μg) of total RNA isolated from rat liver at fetal days 18, 20, and 22 (F18, F20, and F22), less than 12 h after birth (B), and 5, 13, and 30 days after birth (B+5, B+13, and B+30) were subjected to electrophoresis, transferred to a Biotrans membrane, and allowed to hybridize with a ^{32}P-labeled 1450 bp cDNA probe coding for mouse HGF-like protein (see Probes under Materials and Methods). Migration of 28S and 18S ribosomal RNA bands is indicated.

postpartum stages (Table III). Messenger RNA coding for HGF-like protein was expressed mostly in liver, but also in lung, adrenal, and placenta at several pre- and postpartum stages. HGF-like protein mRNA was not expressed in brain, heart, spleen, kidney, ovary, uterus, or urinary bladder at any pre- or postparturitional stage analyzed (Table III).

Discussion

The mouse cDNA for HGF-like protein codes for a putative protein with the same domain structure as its human homologue with four kringle domains followed by a serine protease-like domain. Translated sequence from the gene and cDNA coding for mouse HGF-like protein indicate that a protein of 716 amino acids with a molecular weight of 80 593 would be synthesized. The amino acid composition of this protein is Ala-32, Arg-50, Asn-35, Asp-36, Cys-45, Gln-39, Glu-39, Gly-58, His-26, Ile-18, Leu-57, Lys-28, Met-6, Phe-22, Pro-60, Ser-41, Thr-38, Trp-19, Tyr-20, and Val-47. There are four potential N-linked carbohydrate attachment sites at asparagines in the sequence Asn-X-Thr/Ser at positions 72, 173, 305, and 624 (Figure 2). Three of these sites are also potential sites in the human protein at residues 72, 305, and

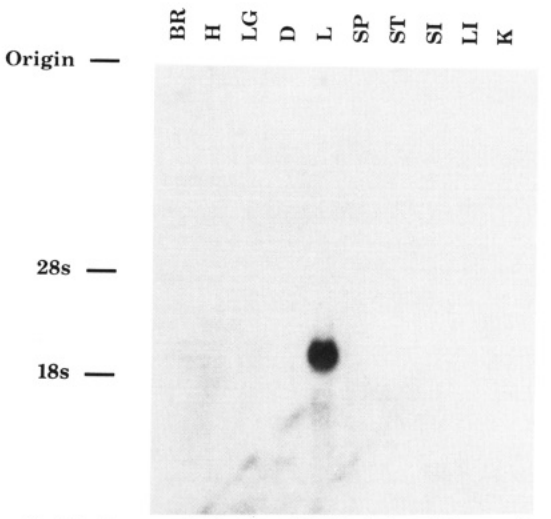


FIGURE 8: Northern analysis of total RNA isolated from various rat tissues during postnatal development. Samples of total RNA (20 μg) isolated from brain (BR), heart (H), lung (LG), diaphragm (D), liver (L), spleen (SP), stomach (ST), small intestine (SI), large intestine (LI), and kidney (K) tissues from day 30 after birth were subjected to electrophoresis, transferred to a Biotrans membrane, and allowed to hybridize to a 1450 bp mouse cDNA probe coding for HGF-like protein (see Probes under Materials and Methods). The migration of 28S and 18S ribosomal RNA bands is indicated.

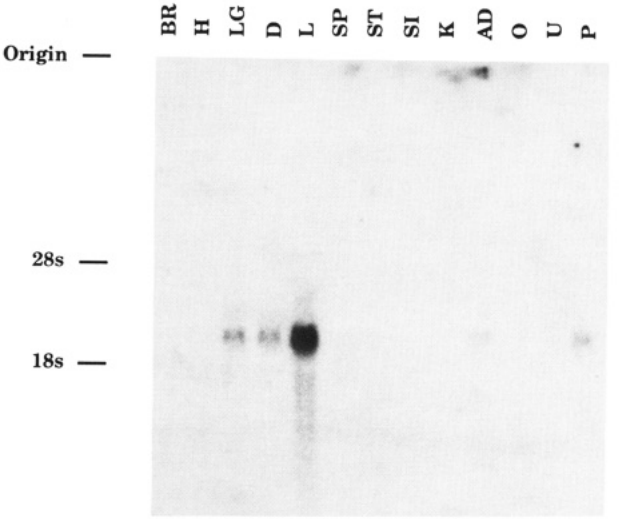


FIGURE 9: Northern analysis of total RNA from maternal rat tissues during pregnancy. Samples of total RNA (20 μg) isolated from brain (BR), heart (H), lung (LG), diaphragm (D), liver (L), spleen (SP), stomach (ST), small intestine (SI), kidney (K), adrenal (AD), ovary (O), uterus (U), and placenta (P) from day 20 of pregnancy were subjected to electrophoresis, transferred to a Biotrans membrane, and allowed to hybridize to a ^{32}P-labeled 1450 bp cDNA probe coding for mouse HGF-like protein (see Probes under Materials and Methods). The migration of 28S and 18S ribosomal RNA bands is indicated.

624 (positions 72, 296, and 615 in the human protein; Han et al., 1991). The sequence at the amino-terminal end of the putative protein is hydrophobic and therefore may be part of a signal sequence required for secretion of the protein from the cell. On the basis of homology with the translated sequence of the cDNA coding for human HGF-like protein, the signal peptidase cleavage site is possibly between amino acid residues Gly-31 and Thr-32 (Figure 2). Within the first 40 amino acids, this proposed site and surrounding sequences are the only ones conserved in the human HGF-like protein sequence that fits with the consensus sequence for signal peptidase recognition sequences (von Heijne, 1983; Watson, 1984; Han et al., 1991).

The translated amino acid sequence of the mouse cDNA is five residues longer than the human protein. The difference in length is attributed to additional residues in two regions that are between domains. In mouse HGF-like protein, there are 23 residues separating the second and third kringle domains while the human protein has 14 residues (Figure 2; residues 269–291). There are 31 and 35 amino acids between the fourth kringle and the putative activation site in the mouse and human protein, respectively (residues 458–492; Figure 2).

The kringle and serine protease domains coded for in the cDNA for mouse HGF-like protein are over 80% identical with their respective domains in the human protein at both the protein and the nucleic acid sequence level. Sequences at the putative activation site and at the amino terminal of the serine protease domain are the same in the translated sequence for both human and mouse HGF-like proteins. The same differences in active-site residues in the serine protease-like domain are found in the mouse and human proteins; His has been changed to Gln, Asp to Gln, and Ser to Tyr (residues 531, 577, and 670, respectively; Figure 2).

In contrast to mRNA for human HGF-like protein, which is 2.4 and 3.0 kilobases in length in human liver, only the smaller species is seen in mouse and rat liver. Primer extension analysis of mouse liver RNA resulted in only one band which placed the site of initiation of transcription for the mouse gene 94 bp upstream of the putative initiator methionine codon. This is consistent with the size of the mRNA. The same experiment performed on human liver always resulted in inconclusive results which might be indicative of the multiple size classes of the mRNA. It appears likely that alternative splicing of the RNA transcript occurs at the 5′ end of the human mRNA for HGF-like protein (Han et al., 1991) or that a closely related gene is transcribed in human liver.

Comparison of the genes coding for human and mouse HGF-like protein shows a high degree of identity in exons as well as in intervening sequences (Table I). All splice junctions are in identical positions with respect to coding sequence in both genes and are always the same type (Table I). Exons are between 76 and 89% identical, with the 3′-noncoding region in exon 18 being the least similar part of the two coding regions. Intervening sequences are unusually small and comparable in size between the two genes. These sequences are between 42 and 70% identical with each other when the two genes are compared (Table I).

The same sequences in the Genbank and NBRF databases were homologous to similar regions in the genes coding for both human and mouse HGF-like protein. We have been able to infer that the gene for human HGF-like protein is on human chromosome 3 at the DNF15S2 locus (3p21) based on the similarity of its sequence with that determined at the DNF15S2 locus and published restriction maps for this region (Han et al., 1991). As with the human gene, the acyl-peptide hydrolase gene is approximately 450 bp downstream of the gene for mouse HGF-like protein but is transcribed in the opposite direction (since it is encoded on the complementary strand).

The expression pattern of mRNA coding for HGF-like protein during development and in maternal rats was determined by Northern analysis. It is apparent that the majority of mRNA coding for HGF-like protein was expressed in liver. Messenger RNA coding for HGF-like protein was also expressed at a lower level in lung, adrenal, and placenta. In certain tissues, diaphragm, stomach, and small intestine, mRNA coding for HGF-like protein was only expressed at a single developmental timepoint. Messenger RNA coding for HGF-like protein was not expressed at any time point in brain, heart, aorta, spleen, kidney, or urinary bladder.

Several other proteins including prothrombin, plasminogen, apolipoprotein(a), and HGF contain kringle domains structurally similar to those found in HGF-like protein. The mRNA for each of these proteins is primarily expressed in liver, with lower amounts in other tissues. Prothrombin mRNA has been shown to be expressed in rats in several tissues including liver, diaphragm, stomach, intestine, kidney, adrenal, uterus, and placenta (Jamison & Degen, 1991). Plasminogen mRNA has been demonstrated to be expressed in rat tissues including liver, diaphragm, spleen, and stomach (unpublished results) and in rhesus monkey tissues including liver, testes, and kidney (Tomlinson et al., 1989). Apolipoprotein(a) mRNA has been demonstrated to be expressed in liver, testes, and brain (Tomlinson et al., 1989). HGF mRNA has been demonstrated to be expressed in rat tissues including liver, spleen, kidney, heart, lung, and brain (Okajima et al., 1990; Tashiro et al., 1990). It is apparent that the primary site of synthesis of these kringle-containing proteins is the liver and that these genes are expressed at lower levels in other tissues.

ACKNOWLEDGMENTS

We thank Mary Jo Danton for reviewing the manuscript and Sue McDowell for her excellent technical assistance and for preparation of the figures containing DNA sequence.

REFERENCES

Bancroft, J. D., Schaefer, L. A., & Degen, S. J. F. (1990) *Gene 95*, 253–260.

Brauch, H., Johnson, B., Hovis, J., Yano, T., Gazdar, A., Pettengill, O. S., Graziano, S., Sorenson, G. D., Poiesz, B. J., Minna, J., Linehan, M., & Zbar, B. (1987) *N. Engl. J. Med. 317*, 1109–1113.

Breathnach, R., Benoist, C., O'Hare, K., Gannon, F., & Chambon, P. (1978) *Proc. Natl. Acad. Sci. U.S.A. 75*, 4853–4857.

Bucher, P., & Trifonov, E. N. (1986) *Nucleic Acids Res. 14*, 10009–10026.

Chomczynski, P., & Sacchi, N. (1987) *Anal. Biochem. 162*, 156–159.

Degen, S. J. F., & Davie, E. W. (1987) *Biochemistry 26*, 6165–6177.

Degen, S. J. F., MacGillivray, R. T. A., & Davie, E. W. (1983) *Biochemistry 22*, 2087–2097.

Degen, S. J. F., Bell, S. M., Schaefer, L. A., & Elliott, R. W. (1990) *Genomics 8*, 49–61.

Duncan, C. H. (1985) *NEN Product News 4*, 6–7.

Feinberg, A. P., & Vogelstein, B. (1984) *Anal. Biochem. 137*, 266–267.

Han, S., Stuart, L. A., & Degen, S. J. F. (1991) *Biochemistry* (preceding paper in this issue).

Jamison, C. S., & Degen, S. J. F. (1991) *Biochim. Biophys. Acta 1088*, 208-216.
Kobayashi, K., Lin, L.-W., Yeadon, J. E., Klickstein, L. B., & Smith, J. A. (1989) *J. Biol. Chem. 264*, 8892-8899.
Kok, K., Osinga, J., Carritt, B., Davis, M. G., van der Hout, A. H., van der Veen, A. Y., Landsvater, R. M., de Leij, L. F. M. H., Berendsen, H. H., Postmus, P. E., Poppema, S., & Buys, C. H. C. M. (1987) *Nature 330*, 578-581.
Kovacs, G., Erlandsson, R., Boldog, F., Ingvarsson, S., Muller-Brechlin, R., Klein, G., & Sumegi, J. (1988) *Proc. Natl. Acad. Sci. U.S.A. 85*, 1571-1575.
Kozak, M. (1986) *Cell 44*, 283-292.
Magnusson, S., Petersen, T. E., Sottrup-Jensen, L., & Claeys, H. (1975) in *Proteases and Biological Control* (Reich, E., Rifkin, D. B., & Shaw, E., Eds.) pp 123-149, Cold Spring Harbor Laboratories, Cold Spring Harbor, NY.
Maxam, A. M., & Gilbert, W. (1980) *Methods Enzymol. 65*, 499-560.
Mount, S. M. (1982) *Nucleic Acids Res. 10*, 459-472.
Nakamura, T., Nishizawa, T., Hagiya, M., Seki, T., Shimonishi, M., Sugimura, A., Tashiro, K., & Shimizu, S. (1989) *Nature 342*, 440-443.
Naylor, S. L., Johnson, B. E., Minna, J. D., & Sakaguchi, A. Y. (1987) *Nature 329*, 451-454.
Naylor, S. L., Marshall, A., Hensel, C., Martinez, P. F., Holley, B., & Sakaguchi, A. Y. (1989) *Genomics 4*, 355-361.
Okajima, A., Miyazawa, K., & Kitamura, N. (1990) *Eur. J. Biochem. 193*, 375-381.
Queen, C., & Korn, L. J. (1984) *Nucleic Acids Res. 12*, 581-599.
Sager, R. (1989) *Science 246*, 1406-1412.
Seizinger, B. R., Rouleau, G. A., Ozelius, L. J., Lane, A. H., Farmer, G. E., Lamiell, J. M., Haines, J., Yuen, J. W. M., Collins, D., Majoor-Krakauer, D., et al. (1988) *Nature 332*, 268-269.
Sharp, P. A. (1981) *Cell 23*, 643-646.
Tashiro, K., Hagiya, M., Nishizawa, T., Seki, T., Shimonishi, M., Shimizu, S., & Nakamura, T. (1990) *Proc. Natl. Acad. Sci. U.S.A. 87*, 3200-3204.
Tomlinson, J. E., McLean, J. W., & Lawn, R. M. (1989) *J. Biol. Chem. 264*, 5957-5965.
von Heijne, G. (1983) *Eur. J. Biochem. 133*, 17-21.
Watson, M. E. E. (1984) *Nucleic Acids Res. 12*, 5145-5164.
Welch, H. M., Darby, J. K., Pilz, A. J., Ko, C. M., & Carritt, B. (1989) *Genomics 5*, 423-430.
Whang-Peng, J., Bunn, P. A., Jr., Kao-Shan, C. S., Lee, E. C., Carney, D. N., Gazdar, A. F., & Minna, J. D. (1982) *Cancer Genet Cytogenet. 6*, 119-134.
Zbar, B., Brauch, H., Talmadge, C., & Linehan, M. (1987) *Nature 327*, 721-724.

Sequence Specificity in Triple-Helix Formation: Experimental and Theoretical Studies of the Effect of Mismatches on Triplex Stability

Jean-Louis Mergny,‡ Jian-Sheng Sun, Michel Rougée, Thérèse Montenay-Garestier, Francisca Barcelo,§ Jacques Chomilier, and Claude Hélène*
Laboratoire de Biophysique, INSERM U201, CNRS UA481, Muséum National d'Histoire Naturelle, 43 rue Cuvier, 75005 Paris, France

Received April 25, 1991; Revised Manuscript Received July 16, 1991

ABSTRACT: The specificity of a homopyrimidine oligonucleotide binding to a homopurine–homopyrimidine sequence on double-stranded DNA was investigated by both molecular modeling and thermal dissociation experiments. The presence of a single mismatched triplet at the center of the triplex was shown to destabilize the triple helix, leading to a lower melting temperature and a less favorable energy of interaction. A terminal mismatch was less destabilizing than a central mismatch. The extent of destabilization was shown to be dependent on the nature of the mismatch. Both single base-pair substitution and deletion in the duplex DNA target were investigated. When a homopurine stretch was interrupted by one thymine, guanine was the least destabilizing base on the third strand. However, G in the third strand did not discriminate between a C·G and an A·T base pair. If the stretch of purines was interrupted by a cytosine, the presence of pyrimidines (C or T) in the third strand yielded a less destabilizing effect than purines. This study shows that oligonucleotides forming triple helices can discriminate between duplex DNA sequences that differ by one base pair. It provides a basis for the choice of antigene oligonucleotide sequences targeted to selected sequences on duplex DNA.

Sequence-specific recognition of nucleic acids is essential for the regulation of cellular functions including replication, transcription, and translation. In most cases, regulation of gene expression in living organisms is achieved by specific nucleic acid binding proteins. In a limited number of cases it has been demonstrated that nucleic acids could also play a regulatory role [see Hélène and Toulmé (1990) for a review]. Homopyrimidine oligodeoxynucleotides can be targeted to a homopurine–homopyrimidine tract of double-helical DNA (Le Doan et al., 1987; Moser & Dervan, 1987; Lyamichev et al., 1988). They form a local *triple helix*, a structure that has been first discovered for homopolynucleotides (Felsenfeld et al., 1957; Stevens & Felsenfeld, 1964). Homopyrimidine oligodeoxynucleotides could control transcription, using what we call the "antigene" strategy (Hélène & Toulmé, 1990). Intermolecular triplex formation occurs upon binding of the third oligopyrimidine strand to the major groove of double-

* Author to whom correspondence should be addressed.
‡ J.-L.M. was supported by a financial grant from the Institut de Formation Supérieure Biomédicale and Rhône-Poulenc.
§ Permanent address: Universitat de les Illes Balears, Dpto. de Biologia i Ciències de la Salut, Palma de Mallorca, Spain.

0006-2960/91/0430-9791$02.50/0