Characterization of the DNF15S2 Locus on Human Chromosome 3: Identification of a Gene Coding for Four Kringle Domains with Homology to Hepatocyte Growth Factor[†,‡]

Su Han, Lorie A. Stuart, and Sandra J. Friezner Degen*

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: A human genomic DNA library was screened by using conditions of reduced stringency with a bovine cDNA probe coding for the kringle domains in prothrombin in order to isolate the human prothrombin gene. Twelve positives were identified, three of which coded for prothrombin (Degen & Davie, 1987). Phage L5 was characterized in more detail because of its strong hybridization to the cDNA probe and its unique restriction map compared to the gene coding for human prothrombin. The gene in L5 was sequenced and found to code for a kringle-containing protein. A human liver cDNA library was screened by using a genomic probe from the gene in L5. cDNAs were isolated that contained sequence identical with regions in the gene in L5. Comparison of the cDNA with the gene indicated that the gene in L5 was composed of 18 exons separated by 17 intervening sequences and is 4690 bp in length. Exons ranged in size from 36 to 242 bp in length while intervening sequences ranged from 77 to 697 bp in length. The putative protein encoded by the gene in L5 contains four kringle domains followed by a serine protease-like domain. This domain structure is identical with that found in hepatocyte growth factor (HGF), although the two proteins are only about 50% identical. On the basis of the similarity of the protein encoded by L5 and HGF, we propose that the putative L5 protein be tentatively called HGF-like protein until a function is identified. The DNA sequence of the gene and cDNA and its translated amino acid sequence were compared against GenBank and NBRF databases. Sequences homologous to DNF15S1 and DNF15S2, human DNF15S2 lung mRNA, and rat acyl-peptide hydrolase were identified in exon 17 to the 3′ end of the characterized sequence for the gene. From our results, it is apparent that the gene coding for human HGF-like protein is located at the DNF15S2 locus on human chromosome 3 (3p21). The gene for acyl-peptide hydrolase is 444 bp downstream of the gene coding for HGF-like protein, but on the complementary strand. The DNF15S2 locus has been proposed to code for one or more tumor suppressor genes since this locus is deleted in DNA from small cell lung carcinoma, other lung cancers, renal cell carcinoma, and von Hippel-Lindau syndrome.

The amino-terminal noncatalytic region of the proteases involved in blood coagulation and fibrinolysis contains several types of domains that are responsible for the regulatory function of these proteins, a feature that distinguishes them from the digestive proteases trypsin and chymotrypsin (Davie et al., 1986; Furie & Furie, 1988). These regulatory regions are organized with various combinations of domains that include kringles, epidermal growth factor-like structures, finger structures, vitamin K dependent calcium-binding regions, and apple structures (McMullen et al., 1991). These domains appear to be essential for the biological specificity of the enzymatic portion of the molecule.

Kringle-containing proteins that have been identified were until recently confined to plasma proteins that functioned in either blood coagulation or fibrinolysis. These included factor XII, prothrombin, tissue plasminogen activator (t-PA),[1] urokinase (u-PA), and plasminogen with one, two, two, one, and five kringle domains, respectively. In the past several years, apolipoprotein(a) and hepatocyte growth factor (HGF) have also been identified to contain kringle domains. Apolipoprotein(a) which contains up to 38 kringles is implicated to be involved in atherosclerosis since there is a correlation with the occurrence of this disease and increased levels of this protein (McLean et al., 1987). HGF has four kringle domains and is a growth factor with broad target specificity (Nakamura et al., 1989).

Kringle domains were first identified in bovine prothrombin as an internal duplication of a triple-disulfide-bonded structure containing approximately 80 amino acids (Magnusson et al., 1975). Kringle structures are thought to function autonomously in plasminogen and t-PA (Trexler & Patthy, 1983; van Zonneveld et al., 1986). They also fold independently as shown by the three-dimensional structure of the first kringle in prothrombin that has been determined at 2.8-Å resolution (Tulinsky et al., 1988a). Functional differences between kringles probably are related to conformational differences

[†] This paper is dedicated in the memory of Dr. Staffan Magnusson. This work was supported in part by the Pew Memorial Trust, by Research Grant HL38232 from the National Institutes of Health, and by a grant-in-aid from the American Heart Association. 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 M74178 (cDNA) and M74179 (gene).

[1] Abbreviations: bp, base pair(s); EDTA, ethylenediaminetetraacetic acid; HGF, hepatocyte growth factor; kbp (kb in figures), kilobase pair(s); kDa, kilodalton(s); PAP, preactivation peptide; PCR, polymerase chain reaction; SDS, sodium dodecyl sulfate; Tris-HCl, tris(hydroxymethyl)aminomethane hydrochloride; t-PA, tissue plasminogen activator; u-PA, urokinase-type plasminogen activator.

resulting from amino acid substitutions in the variable regions of this domain. On the basis of homology between kringles, it has been hypothesized that all kringles have been derived from a single gene containing a kringle most similar to the fourth kringle of plasminogen (Castellino & Beals, 1987).

Kringles appear to be protein-binding domains (Patthy et al., 1984). They have been shown to be essential for the function of prothrombin (Esmon & Jackson, 1974), plasminogen (Wiman & Wallen, 1977; Wiman et al., 1979), and t-PA (van Zonneveld et al., 1986) since they regulate prothrombin–thrombin conversion by binding factor Va in prothrombin and they regulate plasmin-catalyzed fibrin degradation by mediating binding to fibrin in plasminogen and t-PA. The functions of all other kringles have not been identified, but since kringle structures are over 50% identical with each other, it is reasonable to assume that they are involved in binding interactions with other proteins essential for their regulation.

In this paper, we present the DNA sequence of a newly identified human cDNA and gene that were isolated on the basis of cross-hybridization with a probe coding for the kringle domains in bovine prothrombin. The putative protein encoded by this cDNA resembles HGF in that it contains four kringle domains and a serine protease-like domain but is only about 50% identical with HGF. Until the protein is isolated and a function identified, we propose to call it HGF-like protein solely based on similarity of domain structure. The gene is located at the DNF15S2 locus on human chromosome 3 (3p21) that has been proposed to code for one or more tumor suppressor genes since this locus is deleted in DNA from several types of carcinomas.

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. The bovine prothrombin cDNA probe was isolated after digestion of the cDNA with *Ava*I and *Bam*HI (MacGillivray & Davie, 1984). The fragment was 1200 bp in length and coded for amino acids 109–500 of bovine prothrombin which includes DNA coding for part of the first kringle, the entire second kringle, and most of the serine protease domain. A 1950 bp *Hin*dIII fragment was isolated from the genomic subclone pL5Hind1.85 (nucleotides 918–2868 in Figure 2) that contains eight exons of the human HGF-like gene coding for the amino-terminal portion of the protein including three kringle domains. A 680 bp fragment coding for part of the second kringle and all of the third was also isolated from pL5Hind1.85 by digestion with *Bam*HI and *Hin*dIII (nucleotides 2190–2868; Figure 2). A 340 bp fragment was isolated from a human cDNA coding for HGF-like protein (46; Figure 5) after digestion with *Eco*RI and *Nco*I. This fragment codes for part of kringles 1 and 2. A 1470 bp fragment was isolated from a human cDNA coding for HGF-like protein (33; Figure 5) after digestion with *Eco*RI and *Bam*HI. This fragment codes for part of the second kringle, kringles 3 and 4, and the entire serine protease-like domain. A 400 bp fragment containing the first exon of the gene coding for mouse HGF-like protein was isolated by digestion of the genomic subclone pMGL5-12Bgl3.3 with *Bam*HI and *Eco*RI (Degen et al., 1991). This probe contains 100 bp of 5′-flanking sequence, the first exon, and 145 bp of the first intervening sequence. All probes were labeled by nick-translation or by the random primer labeling procedure using [^{32}P]αCTP (Feinberg & Vogelstein, 1984).

Screening of Human Genomic DNA Libraries. An *Alu*I/*Hae*III fetal human liver genomic DNA library (kindly provided by Dr. Tom Maniatis of Harvard University; Lawn et al., 1978) was screened at reduced stringency using a cDNA probe coding for bovine prothrombin (see Probes). Reduced-stringency conditions included hybridization overnight at 60 °C in 2× Denhardt's solution [0.04% poly(vinylpyrrolidone), 0.04% Ficoll, and 0.04% bovine serum albumin] containing 6× SSC (1× SSC = 150 mM NaCl and 15 mM sodium citrate, pH 7.0), 0.5% SDS, 1 mM EDTA, and 1 × 10^6 cpm/mL probe followed by washing at 60 °C in 6× SSC with 0.5% SDS. Approximately 2 × 10^6 phage were screened; 12 positives were identified and plaque-purified, and phage DNA was prepared.

A second human genomic DNA library prepared from placental DNA (Clontech) was screened for the 5′ end of the gene coding for HGF-like protein with a mouse genomic clone containing the first exon of the gene for HGF-like protein (see Probes). Approximately 500 000 phage were screened under identical reduced-stringency conditions discussed above. Thirteen positives were identified; three were plaque purified, and phage DNA was prepared.

Screening of Human Liver cDNA Library. A λgt11 cDNA library prepared from human fetal liver mRNA (kindly provided by Dr. Vincent Kidd, University of Alabama, Birmingham) was screened for the human cDNA coding for HGF-like protein by using a probe isolated from the human gene coding for HGF-like protein (680 bp *Bam*HI/*Hin*dIII fragment; see Probes). Approximately 1 × 10^5 phage were screened at high stringency using standard techniques (Degen & Davie, 1987). Six positives were identified. The longest (46) was 1.9 kbp in length (Figure 5). A 5′-end fragment from this cDNA (340 bp *Eco*RI/*Nco*I fragment; see Probes) was used to rescreen the library to obtain clones with longer 5′ ends. Two clones were identified and characterized (33 and 19; Figure 5).

Southern Analysis. Genomic DNA isolated from human lymphocytes was kindly provided by Dr. Dan Wiginton. DNA (10 μg) was digested with various restriction enzymes (New England Biolabs or Bethesda Research Labs), fractionated on a 0.8% agarose gel, and transferred to Genescreen-plus (DuPont-NEN; Smith & Summers, 1980). The membrane was prehybridized in 50 mM Tris-HCl (pH 7.4) containing 1.0 M NaCl and 200 μg/mL *Escherichia coli* tRNA at 68 °C for 4 h. Hybridization was overnight a 68 °C in prehybridization solution containing 10% dextran sulfate (Pharmacia), 1% SDS, 200 μg/mL calf thymus DNA, and 5 × 10^6 cpm of a ^{32}P-labeled fragment isolated from the gene coding for human HGF-like protein (680 bp *Bam*HI/*Hin*dIII fragment; see Probes). The membrane was washed in 0.1× SSC/1% SDS at 68 °C and exposed to X-ray film (Kodak XAR-5) with an intensifying screen at −70 °C overnight.

Northern Analysis. Total RNA was isolated from confluent monolayers of HepG2 cells or human liver by the method of Glisen et al. (1974). RNA was denatured and electrophoresed on formaldehyde–agarose gels (Ausubel et al., 1989) followed by transfer to nitrocellulose. The filter was hybridized with a ^{32}P-labeled 1950 bp *Hin*dIII fragment from the gene coding for HGF-like protein (see Probes).

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 dideoxy chain termination method (Duncan 1985). Sequences were analyzed by using the Microgenie DNA sequence analysis program from Beckman Instruments

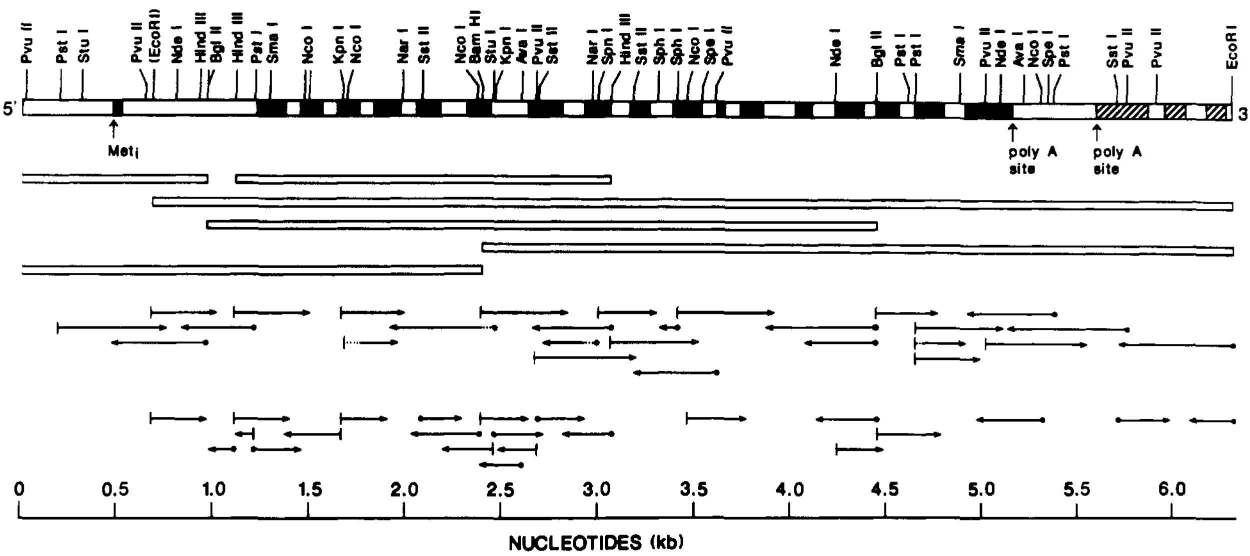

FIGURE 1: Restriction map and sequencing strategy of the gene coding for human HGF-like protein. A partial restriction map is shown above the bar. The *Eco*RI in parentheses is at the 5′ end of L5 and was from the linker added during construction of the library. Blackened parts of the bar represent exons in the gene coding for HGF-like protein, and hatched bars represent putative exons in the acyl-peptide hydrolase gene. Blank areas between the boxes represent intervening sequences except the area between the final black box and first hatched box which represents 3′-flanking sequence for both genes. The site for the codon of the putative initiator methionine for the gene coding for human HGF-like protein is indicated. The polyadenylation sites for both genes are indicated. The orientation of transcription for the gene for human HGF-like protein is indicated by the presence of the 5′ and 3′ at the ends of the bar. Transcription of the acyl-peptide hydrolase gene is in the opposite direction (on the complementary strand). The extent of subclones obtained from L5 and L5/3 phage is indicated below the restriction map by the open bars. Open-ended bars indicate that the fragment continues beyond the map shown. Subclones obtained from L5/3 are open-ended at their 5′ ends; all others were obtained from L5. The sequencing strategies for both M13 (top set of arrows) and Maxam and Gilbert (bottom set of arrows) sequencing procedures are shown by the extent of the arrows. The sequence determined on the coding strand is indicated by arrows ending with vertical bars and that determined on the complementary strand with closed circles at the ends of the arrows. Dotted parts of arrows represent regions not sequenced. The sequence was determined 2 times or more for 88% of the sequence, and 61% was determined on both strands. Only one overlap was not obtained but was present in an exon where the overlap had been obtained in the cDNA. The scale is shown in kilobase pairs.

on an IBM-AT computer (Queen & Korn, 1984).

RESULTS

Isolation of the Human L5 Gene. An *Alu*I/*Hae*III fetal human liver genomic DNA library was screened with a bovine prothrombin cDNA probe using conditions of reduced stringency in order to isolate the gene coding for human prothrombin. This experiment was performed before a cDNA coding for human prothrombin was available and before the human prothrombin gene had been isolated (Degen et al., 1983). Twelve positive phage were identified that hybridized to varying degrees with the probe. Characterization by restriction enzyme mapping and DNA sequence analysis identified three of the positive phage as coding for the human prothrombin gene. These have been discussed in detail previously (Degen et al., 1983; Degen & Davie, 1987).

Recombinant phage L5 was characterized in detail because of its strong hybridization to the bovine prothrombin cDNA probe and its unique restriction map compared to the gene coding for human prothrombin (Degen & Davie, 1987; Figure 1). Preliminary DNA sequence analysis of L5 indicated that it did not code for the prothrombin gene but did code for at least one kringle domain that was unique from any other previously characterized kringle.

Organization of the Gene Coding for HGF-like Protein. The L5 phage contained almost the entire gene coding for HGF-like protein, but did not contain the first exon (the 5′ end of the characterized part of L5 starts at nucleotide 501 in Figure 2) since 36 bp of cDNA sequence could not be found in the sequence upstream from the second exon (see Isolation of the cDNA for Human HGF-like Protein). Since the first exon had already been identified in the gene for the mouse homologue to the gene in L5 (Degen et al., 1991), a probe containing this exon was used to screen another human genomic DNA library (see Materials and Methods). Thirteen positive phage were identified; three were characterized and found to overlap with each other and phage L5. Fragments from phage L5/3 were subcloned, and the DNA sequence was obtained for the region that hybridized with the exon 1 probe. This sequence overlapped with the 5′-most sequence determined from phage L5 and contained an exon homologous to the first exon in the mouse gene. From the codon for the initiator methionine to the 3′ end of exon 1 in both the human and mouse genes, the sequences are 79% identical.

The entire sequence of the gene present in phage L5 and L5/3 is shown in Figure 2. The gene is 4690 bp in length (from the codon for the putative initiator methionine to the polyadenylation site) and contains 18 exons interrupted by 17 intervening sequences. The exons range in size from 36 to 242 bp in length while the intervening sequences range in size from 77 to 697 bp in length [see Table I in Degen et al. (1991)]. Excluding the first intervening sequence, the remaining 16 intervening sequences range from 77 to 202 bp in length. In addition to the sequence of the gene, 273 and 1137 bp of sequence were determined upstream of the initiator methionine codon and downstream of the polyadenylation site, respectively. All splice junctions of intervening sequences agree with the $^{5'}$GT-AG$^{3'}$ rule of Breathnach et al. (1978) and the consensus sequence of Mount (1982) except for the presence of a GC at the 5′ end of intervening sequence C (nucleotides 1364–1365; Figure 2) and an AAG at the 3′ end of intervening sequences G and O (nucleotides 2432–2434 and 4258–4260, respectively; Figure 2).

The sequence surrounding the proposed codon for the initiator methionine is $^{5'}$GAAGG$\underline{\text{ATG}}$G$^{3'}$ at positions 269–277 in Figure 2. 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. These two bases in the gene agree with the consensus. There is another ATG codon upstream in the same reading frame as the proposed initiator codon (nucleotides 232–234; Figure 2). The sequence surrounding this ATG ($^{5'}$CCCGGATGG$^{3'}$) is also identical in five of the eight bases of the consensus, but only one of the critical bases (at +4) is present. On the basis of the sequence of the mouse gene where this upstream ATG is not in the same reading frame as the coding sequence, we infer that the downstream ATG is the initiator methionine codon [see Degen et al., (1991)].

Primer extension experiments have been performed several times in order to determine the site of initiation of transcription. No conclusive results have been obtained at the present time, so we cannot rule out the presence of an upstream exon containing 5′-noncoding sequence. We have chosen to number the sequence determined for the gene starting with +1 as the first base in Figure 2 since negatively numbered sequences typically are part of upstream 5′-flanking regions that we have not definitively located.

Comparison of the sequence of the bovine prothrombin cDNA probe that hybridized with the HGF-like gene indicates several regions of homology. Nucleotides 1650–1734, 1657–1768, 2725–2783, 2727–2779, 3201–3271, and 3213–3324 in the HGF-like gene (Figure 2) are 69, 69, 77, 77, 72, and 67% identical with nucleotides 454–538, 776–887, 459–520, 776–831, 450–520, and 777–888 in the bovine prothrombin cDNA, respectively (MacGillivray & Davie, 1984). This high degree of homology agrees with the strong hybridization signal observed with L5 phage DNA.

Southern Analysis of Human Genomic DNA. Southern hybridization analysis of human genomic DNA digested with several restriction enzymes was performed to determine the approximate copy number of the gene and whether all fragments could be accounted for on the basis of sites deduced from the gene sequence (Figure 3). There were many more bands than expected for each enzyme digest, and therefore homologous sequences appear to be present in the genome. Although the intervening sequences in the gene do not appear to contain repetitive sequences (see Search of the DNA and Protein Database), we cannot rule out the possibility that low-copy repetitive sequences are present (see Discussion).

Northern Analysis of RNA Isolated from Human Liver and HepG2 Cells. In order to determine where mRNA coding for HGF-like protein might be expressed so that the appropriate cDNA library could be screened, Northern analysis was performed with total RNA isolated from human liver, HepG2 cells, and human placenta using a probe isolated from the human gene (1950 bp *Hin*dIII fragment from pL5Hind1.85; Materials and Methods). Messenger RNA for HGF-like protein was detected in human liver and to a much less extent in HepG2 RNA (Figure 4) but was undetectable in human placenta (data not shown). In human liver, mRNA species between 2.4 and 3.0 kbp were identified.

Isolation of the cDNA for Human HGF-like Protein. A λgt11 human liver cDNA library was screened for cDNAs coding for HGF-like protein, and several cDNAs were isolated and characterized. A partial restriction map for these cDNAs is shown in Figure 5. The longest cDNA (33) is 2200 bp in length excluding the poly(A) tail and is not full-length since its 5′ end starts 16 bp downstream from the putative initiator methionine codon in the first exon of the gene (starting at nucleotide 290 in Figure 2). The sequence of this cDNA is shown in Figure 6 (combined with the 16 additional upstream nucleotides from the gene). The cDNA has an open reading frame of 2118 bp followed by a stop codon and a 3′-noncoding region of 80 bp. The sequence $^{5'}$CATAAA$^{3'}$ is 30 bp upstream from the poly(A) tail (nucleotides 2187–2192; Figure 6); this resembles the typical $^{5'}$AATAAA$^{3'}$ polyadenylation sequence found in the majority of mRNAs. The $^{5'}$CATAAA$^{3'}$ sequence has approximately 20% of the polyadenylation activity [including cleavage and poly(A) addition] compared to the more common $^{5'}$AATAAA$^{3'}$ polyadenylation signal as determined by in vitro experiments (Wickens, 1990).

One cDNA (46) contained intervening sequences H and I (indicated in Figure 5; nucleotides 2604–2723 and 2855–2951, respectively, in Figure 2) and thus was partially spliced. Another cDNA (19) had two parts of the coding region deleted (regions A and B in Figure 5). Deleted region A was a complete exon (exon 13; nucleotides 3532–3652 in Figure 2) which could be spliced out of the RNA transcript if the 3′ end of the intervening sequence 5′ to exon 13 was not recognized as a splice junction. Deleted region B was from the 5′ end of exon 15 (nucleotides 4033–4081; Figure 2). The sequence at the 3′ end of this deleted region is similar to the consensus sequence for the 3′ end of intervening sequences (Mount, 1982). Therefore, it is possible that this was recognized as a splice site during RNA processing rather than the upstream site following nucleotide 4032 (Figure 2). If this cDNA represents a translated mRNA, it would code for the 4 kringle domains followed by only 22 amino acids since there are 2 in-frame stop codons at that point.

The polymerase chain reaction (PCR) was used to determine whether deleted regions A and B identified in only one cDNA (19) reflected a population of alternatively spliced mRNAs. Oligonucleotides spanning the combined regions (the 5′ primer was identical with nucleotides 1350–1379 while the 3′ primer was complementary to 1695–1724 in Figure 6) were used as primers while the template was single-stranded cDNA synthesized from total RNA isolated from human liver. The primary bands present on the sequencing gel used to separate the PCR reaction products were a very large band and another of 378 bp in length (data not shown). The larger fragment probably represents amplified product from contaminating genomic DNA which would include three intervening sequences (L, M, and N) as well as part of exons 12 and 15 and all of exons 13 and 14 (predicted size is 768 bp). The other fragment of 378 bp is equivalent in size to that predicted from all cDNAs except cDNA 19 with intact regions A and B. A faint fragment of 208 bp was detectable by this procedure that would represent a mRNA species with both A and B deleted. Comparison of the intensities of the bands indicates that the 378 bp fragment is at least 10 times more intense than the 208 bp fragment (data not shown).

When sequences from the characterized regions of the cDNAs shown in Figure 6 are compared, there are five sites where polymorphisms occur; only one results in an amino acid substitution. At nucleotide 635, a G is present in three of the cDNAs while a T is present in one. This results in an amino acid substitution of Cys to Phe at residue 212. This Cys is most probably involved in a disulfide bond with the Cys at position 251 since it would comprise one of the three disulfide bonds present in the second kringle. At position 1227, two cDNAs have A while one cDNA has a G. Nucleotide 1749 is a C in one cDNA and a T in two others. At 1905, a G is present in one cDNA while an A is present in another cDNA. Finally, at 1923 a T is present in one cDNA while a G is

```
CTGCAGAGGG GTTTCACCCC AACCCCAGGG CACCTCAAGT GTCCCCACCA AACCTTCCTA ACACCTGTCC ACTAAGCTGT ACTAGGCCCT TGCAACTGAC   100
CTATGGGACC CTGAGGCCTG GCCCCTCATG GCTCCTGTCA CCAGGTCTCA GGTCAGGGTC CAGCAGGGCC CTGAGCTGAC GTGTGGAGCC AGAGCCACCC   200

                                                                                MetGlyT rpLeuProLe uLeuLeuLeu
AATCCCGTAG ACAGGTTTCA CAACTTCCCG GATGGGGCTG TGGTGGGTCA CAGTGCAGCC TCCAGCCAGA AGGATGGGGT GGCTCCCACT CCTGCTGCTT   300

LeuThrGlnC ysLeuGlyVa lProG
CTGACTCAAT GCTTAGGGGT CCCTGGTAAG TGCCCCCAAC CCTGATCCCC ATCTGCCTTC AGGAGGGGGT TGGCCCCATT CTCCTATTCT AGGATGAGAA   400
AAAAGTCGGG AGCAGAGGCT CAGTGGGCAT GGGGCAGTGA CCTTGCCCTC TTGAGCACAG CTGGGAAGCC CTAGGAACAC ATAGACATTG CCCACTTAGG   500
CCTCTATTAG CACGTCTGCT CTAGCACTGA AGCAGTGTCA GGACCACACA GATGCACGCA CACAGCAGGC AGTGACCCCT CCTGAGCCTG ATCTACCCCT   600
CTAACCTAGC ATATGCCTTT GTGCAGGTGA GAGCCCAGAT TTGGAGTCTG AATGCCTAGC CAGGGCCCTT GGCTGGGTAA TGTGATGGCT CTGAGCCTTA   700
GCATTCTCAT TTGAGAGATG AGGTGGGGCA AGCTTCATCA CCCACTGCTC TCACAGAGCG TATGTGTTAG ATCTGAGCCC GGTGCCTGGG CCACTAAACA   800
GAGGCACCGG TGATAACTAC CAAGTCTGGG CCTGCTTCCC AGGGGAAATT TTTTTCACAA GTATCTGTGC AGGGGGCTAG ACTGGCCCTT GAAAGTGCAT   900
ACAGGGTCCA TCCCAGAAGC TTGTAGCTTT GATCCCCTGA ATGAACAAAG TGTGGACATG CCAATACACA TTACTGACAT GTATGCCCAC CTGACCTGCA  1000

                        lyGlnArg SerProLeuA snAspPheGl nValLeuArg GlyThrGluL euGlnHisLe uLeuHisAla ValValProG
CCCACTCATG CCTACTCTGC AGGGCAGCGC TCGCCATTGA ATGACTTCCA AGTGCTCCGG GGCACAGAGC TACAGCACCT GCTACATGCG GTGGTGCCCG  1100

lyProTrpGl nGluAspVal AlaAspAlaG luGluCysAl aGlyArgCys GlyProLeuM etAspCysAr
GGCCTTGGCA GGAGGATGTG GCAGATGCTG AAGAGTGTGC TGGTCGCTGT GGGCCCTTAA TGGACTGCCG GTGAGTGGCC ACTGGGCTAG ATAAGACTGG  1200

                                                       gAlaPheHis TyrAsnValS erSerHisGl yCysGlnLeu LeuProTrpT
GGGCAGGGAA GCCTGGGCTG TGGCGTTACC CTGTGCCTTC TTCTCTCCAG GGCCTTCCAC TACAACGTGA GCAGCCATGG TTGCCAACTG CTGCCATGGA  1300

hrGlnHisSe rProHisThr ArgLeuArgA rgSerGlyAr gCysAspLeu PheGlnLysL ysA
CTCAACACTC GCCCCACACG AGGCTGCGGC GTTCTGGGCG CTGTGACCTC TTCCAGAAGA AAGGCAAGTG GGGGTGGAGA GGGGCAGGGT GGGAGACAGG  1400

                                            spTyrValAr gThrCysIle MetAsnAsnG lyValGlyTy rArgGlyThr MetAlaThrT
GGACCTCAGC CCAAGTTGAT CTTCTGTCTC TTGCTCCCAG ACTACGTACG GACCTGCATC ATGAACAATG GGGTTGGGTA CCGGGGCACC ATGGCCACGA  1500

hrValGlyGl yLeuProCys GlnAlaTrpS erHisLysPh eProAsnAsp HisLy
CCGTGGGTGG CCTGCCCTGC CAGGCTTGGA GCCACAAGTT CCCGAATGAT CACAAGTGAG ACAAACACCT TCCCTCCGTC CCGGCCTGGG GCTTCCCCCA  1600

                                   sTyrThrP roThrLeuAr gAsnGlyLeu GluGluAsnP heCysArgAs nProAspGly AspProGlyG
GCACACACTA TAGTGATGCT CTGGGCCCTC AGGTACACGC CCACTCTCCG GAATGGCCTG GAAGAGAACT TCTGCCGTAA CCCTGATGGC GACCCCGGAG  1700

lyProTrpCy sTyrThrThr AspProAlaV alArgPheGl nSerCysGly IleLysSerC ysArgGluA
GTCCTTGGTG CTACACAACA GACCCTGCTG TGCGCTTCCA GAGCTGCGGC ATCAAATCCT GCCGGGAGGG TAAGCGGCGC CGGGTCAAGC TGGGAGAGTG  1800

                                                    la AlaCysValT rpCysAsnGl yGluGluTyr ArgGlyAlaV alAspArgTh
GAGACAAGCC CACGTCCATC CACGAACCCA CTGGCTCTTT GTCTCCAGCC GCGTGTGTCT GGTGCAATGG CGAGGAATAC CGCGGCGCGG TAGACCGCAC  1900

rGluSerGly ArgGluCysG lnArgTrpAs pLeuGlnHis ProHisGlnH isProPheGl uProGlyLy
GGAGTCAGGG CGCGAGTGCC AGCGCTGGGA TCTTCAGCAC CCGCACCAGC ACCCCTTCGA GCCGGGCAAG TACGCGTAGG CGGTATCGGC GTCCTGGGGG  2000
CCGGGCTAGG GAAGGTCCAG GACTCCAGGG GCAGGGCTCC GTGTAGGGCA ATTGGGCGGG GCCAGATAAG CCAGAGTCCC AGGGTCTTGT TCACGCCCCA  2100

              sPheLeu AspGlnGlyL euAspAspAs nTyrCysArg AsnProAspG lySerGluAr gProTrpCys TyrThrThrA spProGlnIl
TTACCGCCCC CAGGTTCCTC GACCAAGGTC TGGACGACAA CTATTGCCGG AATCCTGACG GCTCCGAGCG GCCATGGTGC TACACTACGG ATCCGCAGAT  2200

eGluArgGlu PheCysAspL euProArgCy sG
CGAGCGAGAG TTCTGTGACC TCCCCCGCTG CGGTAGGCGG CGGGGACCAG GCCTGGGAGG GTACCTGGGA ACCTTGGGGA GGGGCGTGGC TTGGCCGGGG  2300
AGGTAAGAGG GGCTGGGCGT GACCTGAGAG CATACCCCGT GGAGTACCGT ACACCTGGGA AAGGCGGGTT TGGTCCCAGC CCGAGAGGGA TCTCAGCTCT  2400

                                     lySerG luAlaGlnPr oArgGlnGlu AlaThrThrV alSerCysPh eArgGlyLys GlyGluGlyT
CGCTCGGGGC CCGACCTATC TCGGTCCATC TAAGGGTCCG AGGCACAGCC CCGCCAAGAG GCCACAACTG TCAGCTGCTT CCGCGGGAAG GGTGAGGGCT  2500

yrArgGlyTh rAlaAsnThr ThrThrAlaG lyValProCy sGlnArgTrp AspAlaGlnI leProHisGl nHisArgPhe ThrProGluL ysTyrAlaCy
ACCGGGGCAC AGCCAATACC ACCACTGCGG GCGTACCTTG CCAGCGTTGG GACGCGCAAA TCCCTCATCA GCACCGATTT ACGCCAGAAA AATACGCGTG  2600

sLy
CAAGTGAGGT GGGGGGGGGG GGCGGGCGTT GGGACGTGCT GCTGCGGGTG AGACGGGAGG AAGGTAGTCA CGGGCTCAAG GCTGGAGGCT GGCGGGCTAG  2700

                         sAspLeu ArgGluAsnP heCysArgAs nProAspGly SerGluAlaP roTrpCysPh eThrLeuArg ProGlyMetA
GGCTGAGTGG AGCGCCTGCT TAGAGACCTT CGGGAGAACT TCTGCCGGAA CCCCGACGGC TCAGAGGCGC CCTGGTGCTT CACACTGCGG CCCGGCATGC  2800

rgAlaAlaPh eCysTyrGln IleArgArgC ysThrAspAs pValArgPro GlnA
GCGCGGCCTT TTGCTACCAG ATCCGGCGTT GTACAGACGA CGTGCGGCCC CAGGGTGAGG CCCAAGCTTG GGGGCTACAG AGCCGGGCTG GAAGCTGGAA  2900

                                                        spCysTyrH isGlyAlaGl yGluGlnTyr ArgGlyThrV alSerLysTh
CCGGAGGCCG GGGCGAGGTC TCGGCCTGAT GGCTGCCCGC ACCCGCCACA GACTGCTACC ACGGCGCAGG GGAGCAGTAC CGCGGCACGG TCAGCAAGAC  3000

rArgLysGly ValGlnCysG lnArgTrpSe rAlaGluThr ProHisLysP roGl
CCGCAAGGGT GTCCAGTGCC AGCGCTGGTC CGCTGAGACG CCGCACAAGC CGCAGTGAGT CCCTGGTGCT CCCGGCCCCG CCAGGGCCCT AACCCTGGGG  3100

                                                                                         nPheThrPh eThrSerGlu
CGGCATGCTT TGGTGTCTGG GACCAGAGCC TGGAAATGGT TGAGACTACC CTGCCACGAT TTTGCTCCCG CTTCCGCCTA GGTTCACGTT TACCTCCGAA  3200

ProHisAlaG lnLeuGluGl uAsnPheCys ArgAsnProA spGlyAspSe rHisGlyPro TrpCysTyrT hrMetAspPr oArgThrPro PheAspTyrC
CCGCATGCAC AACTGGAGGA GAACTTCTGC CGGAACCCAG ATGGGGATAG CCATGGGCCC TGGTGCTACA CGATGGACCC AAGGACCCCA TTCGACTACT  3300

ysAlaLeuAr gArgCysA
GTGCCCTGCG ACGCTGCGGT GAGCACTAGT GACGCTTCCC CCATGACCCT GCCTCAGCCC CCACCCAAAG GCTGGCTCCC TTAACCCCAG TGAACTTTGT  3400

       laA spAspGlnPr oProSerIle LeuAspProP roA
CTTTCAGCTG ATGACCAGCC GCCATCAATC CTGGACCCCC CAGGTTAGGA GTTGGGCCAG TTATGGGTCA GGCCCTTTAG CCCACGACAT CCACACAGTC  3500

                                  spGlnValG lnPheGluLy sCysGlyLys ArgValAspA rgLeuAspGl nArgArgSer LysLeuArgV
TGGGTTTCAT CCAGCCCACC CCATCCTACA GACCAGGTGC AGTTTGAGAA GTGTGGCAAG AGGGTGGATC GGCTGGATCA GCGGCGTTCC AAGCTGCGCG  3600

alValGlyGl yHisProGly AsnSerProT rpThrValSe rLeuArgAsn Ar
TGGTTGGGGG CCATCCGGGC AACTCACCCT GGACAGTCAG CTTGCGGAAT CGGTGAGGCA CAACTGCCTG TCTCCCACAG AGAGGAGCTG AGGTTGTGTC  3700
CTCTGTGGTT ATCCACTGGG GCTGGGAATC TATCCGTGCC CCTTGAGAGG TCCTAGCCAA GAAGATGGCA GGTCTTACGA ATCTGTCCCA GGAGTCTGTT  3800

                             gGl nGlyGlnHis PheCysGlyG lySerLeuVa lLysGluGln TrpIleLeuT hrAlaArgGl nCysPheSer
ACCTGTCCTA ATTCCCCACT CCTCTAGGCA GGGCCAGCAT TTCTGCGGGG GGTCTCTAGT GAAGGAGCAG TGGATACTGA CTGCCCGGCA GTGCTTCTCC  3900
```

```
SerCy
TCCTGGTGAG CCTCCCTTGT GTTTGGGGAC CCAGTCTCAT CCCACCTTCC CCCTTCCCCA GGCAAGCTAA CAAGTGAGCC TTGGGGCAAT GGACTGAGAG  4000

                                   sHisMetP roLeuThrGl yTyrGluVal TrpLeuGlyT hrLeuPheGl nAsnProGln HisGlyGluP
TCACAAATGA CCTAGCAGAG CTTCTCTCCC AGCCATATGC CTCTCACGGG CTATGAGGTA TGGTTGGGCA CCCTGTTCCA GAACCCACAG CATGGAGAGC  4100

roSerLeuGl nArgValPro ValAlaLysM etValCysGl yProSerGly SerGlnLeuV alLeuLeuLy sLeuGluAr
CAAGCCTACA GCGGGTCCCA GTAGCCAAGA TGGTGTGTGG GCCCTCAGGC TCCCAGCTTG TCCTGCTCAA GCTGGAGAGG TATGTGGACA ACCTGGGAGG  4200

                                                                  gSerValThr LeuAsnGlnA rgValAlaLe uIleCysLeu
GTGTGAGGTG GGGCTGGGCC TTGTGGCCTC AGACCCTGAG TGCCCCCATT CTTGCTAAAG ATCTGTGACC CTGAACCAGC GTGTGGCCCT GATCTGCCTG  4300

ProProGluT rpTyrValVa lProProGly ThrLysCysG luIleAlaGl yTrpGlyGlu ThrLysG
CCCCCTGAAT GGTATGTGGT GCCTCCAGGG ACCAAGTGTG AGATTGCAGG CTGGGGTGAG ACCAAAGGTA AGAGCACAGT GCACAGGACT GCTGGTGGCC  4400

                                                                    lyThrGly AsnAspThrV alLeuAsnVa lAlaLeuLeu
AGGAGGCCAG CCCTGGATCT TCCTGCAGGA CCCTCTCCCT CTCCCCATTC CCCTCACTGC AGGTACGGGT AATGACACAG TCCTAAATGT GGCCTTGCTG  4500
                                                             |||| |||||||||| |||||||||| |||||||||| ||||||||||
                                                             CTGC AGGTACGGGT AATGACACAG TCCTAAATGT GGCCTTGCTG

AsnValIleS erAsnGlnGl uCysAsnIle LysHisArgG lyArgValAr gGluSerGlu MetCysThrG luGlyLeuLe uAlaProVal GlyAlaCysG
AATGTCATCT CCAACCAGGA GTGTAACATC AAGCACCGAG GACGTGTGCG GGAGAGTGAG ATGTGCACTG AGGGACTGTT GGCCCCTGTG GGGGCCTGTG  4600
|| ||||||| |||||||||| |||||||||| |||||||||| ||| |||||| |||||| ||| |||||||||| |||||||||| |||||||||| ||   |||||
AACGTCATCT CCAACCAGGA GTGTAACATC AAGCACCGAG GACATGTGCG GGAGAGCGAG ATGTGCACTG AGGGACTGTT GGCCCCTGTG GGxxxCTGTG

lu
AGGTTGGTGG CAGGGCCTGG GCAGCCCTGG AAGTATGGGG GGCTAGAAAT GAACTATTTT ATCATGAAGC AGGCTAGTCA TTGCTGTGGC CCGGGGCCTC  4700
|||||||| | |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| | |||||| | | ||| ||||
AGGTTGGTAG CAGGGCCTGG GCAGCCCTGG AAGTATGGGG GGCTAGAAAT GAACTATTTT ATCATGAAGC AGGCTAGTCA TGGCTGTGCC CGGGGCCCTC

                       GlyAspTyr GlyGlyProL euAlaCysPh eThrHisAsn CysTrpValL euGluGlyIl eIleIlePro AsnArgValC
ATCAGTTCTC CTACCTGCCA GGGTGACTAC GGGGGCCCAC TTGCCTGCTT TACCCACAAC TGCTGGGTCC TGGAAGGAAT TATAATCCCC AACCGAGTAT  4800
|||||||||| |||||||||| | |||||||| |||||||||| |||||||||| |||||||||| |||||||||| || ||||||| || ||||||| ||| ||||||
ATCAGTTCTC CTACCTGCCA GAGTGACTAC GGGGGCCCAC TTGCCTGCTT TACCCACAAC TGCTGGGTCC TGAAAGGAAT TAGAATCCCC AACTGAGTAT

ysAlaArgSe rArgTrpPro AlaValPheT hrArgValSe rValPheVal AspTrpIleH isLysValMe tArgLeuGly ***
GCGCAAGGTC CCGCTGGCCA GCTGTCTTCA CGCGTGTCTC TGTGTTTGTG GACTGGATTC ACAAGGTCAT GAGACTGGGT TAGGCCCAGC CTTGATGCCA  4900
| |||||||| ||||||||| || ||||||| ||| |||||| |||||| ||| |||||||||| |||||||||| |||||||||| |||||||||| ||||| ||||
GTGCAAGGTC GCGCTGGCCA GCCGTCTTCA CGCTTGTCTC TGTGTTxGTG GACTGGATTC ACAAGGTCAT GAGACTGGGT TAGGCCCAGC CTTGACGCCA

TATGCCTTGG GGAGGACAAA ACTTCTTGTC AGACATAAAG CCATGTTTCC TCTTTATGCC TGTACAGATG CTTCTTAGCC TTTGCTTCCA GGAAATGTGT  5000
||||| |||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||| ||||| ||||||||||
TATGCTTTGG GGAGGACAAA ACTTCTTGTC AGACATAAAG CCATGTTTCC TCTTTATGCC TGTACAGATG CTTCTTAGCC TTTGATTCCA GGAAATGTGT

CAGTGACTCC TTGCTAGGGC TCGGGTGGCT TGAGCCCAGC ACACCCTGGG CTAGGTGATC TGTCCAGCCT AGGGGCTTCC CCAACCAAGG CAATGTCCCT  5100
|||||||||| |||||||||| |  ||||||| |
CAGTGACTCC TTGCTAGGGC TGCGGTGGCT T

GGGACTACTT TTGCCCATGG GTGCCGTGGA AAGACAGGGC CTCACACTAG TCCTCCAGAC ATACTCTTGG GAAGGGTGGT ACAGAGTAGT TGCTAATGGA  5200
AGGGGCTGCA GCAGGGAAGC TAGGCTGGTA CAGAGTCCTG GTTGCCAGGA CAGGCAGAGG CTAAGCCTCT CACTGTTCCC TCCCTTCTCA CACTGGAGGC  5300
AGATGAAGCC CTTGTGGCTG CCACACCCAG AACCTAGGGT CTCTGCACCC CAGAGTGGGA GGTGGGGTTG GGGATGGTTT GGTACAAAGT ACCAGCAGGA  5400

ACCAGGCTCT GTGTCCTAAT TTATTATGAC TACATAGCCC ACATTCCTCT GCCCATGCAT CCGTGGAGTC CAGAGCCCAG AAAGCCTCCT GCTGCCCTGC  5500
       ||| |||||||||| |||||||||| |||||||||| |||||||||| | ||| |||| |||||||||| |||||||||| |||||||||| ||  ||||||
       TCT GTGTCCTAAT TTATTATGAC TACATAGCCC ACATTCCTCT GxCCACGCAT CCGTGGAGTC CAGAGCCCAG AAAGCCTCCT GCxxCCCTGC

CAGACCGTTG AGCTCCTCAA GAGGAAGTGT GGCACAGGCT GATCAGCTCA TGCAGAATGG CAGGGCTTCA GCTGCCCAAG TGTGTGCGTA GCCAGAGCAC  5600
|||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| ||||||||||   |||||||
CAGACCGTTG AGCTCCTCAA GAGGAAGTGT GGCACAGGCT GATCAGCTCA TGCAGAATGG CAGGGCTTCA GCTGCCCAAG TGTGTGCGTA CGCAGAGCAC

AGCATTCATG AAGCTGTCTG ACTCCACCTC CACCTCTGAT AATGCGTGGG TGCTTTTGGG ATAGAGCAGG AGCCTGTAGG GATTAGTCAG CAACATTTAA  5700
|||||||||| ||||||     |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||
AGCATTCATG AAGCTGxxxx ACTCCACCTC CACCTCTGAT AATGCGTGGG TGCTTTTGGG ATAGAGCAGG AGC

GGTTGGAGGG TCCTCCTGTG CTCACCTGCC CACCAGCTGC CAGGGCCTTC ATGCTGCACT CACCGAACAG GCACATTCCG GGTCTTGAGG GCACGGTAAT  5800
                                                                     ||||||| ||||||||| | ||||||||   ||| || |
                                                                     CGAACAG xCACATTCCG GxTCTTGAGG xxACGxTAxT

ACTCCATGCC CTGCTTGAAG GGCACACGCC GGTCCTCCTG GCCCAACATC AGTAACAGTG GTGTCTTCAC CTGGGTGTTT GGGGAAGAGT GGGGAGCTGT  5900
||| ||||||  |||| || | | |||||  | | ||||||||  ||||||||| |||||||||| ||||||||||
ACTxCATGCC xTGCTxGAxG GxCACACCGC GxTCCTCCTG xCCCAACATC AGTAACAGTG GTGTCTTCAC

GTTGAGCTGG GCCCTGGATT CTGGATGGAT GGGCAGCACA CAGGGCAAGC AGGGGGCTGC ATACCTGAGG GATGTATCTG ATGGGCGATT TGTCCAGCAT  6000
                                                                      ||||||  | ||||||| |||||||||| ||||||||||
                                                                      CTGAGG xAxGTATCTG ATGGGCGATT TGTCCAGCAT

CTCAGCCCAC ACGCTGAGGT CTGGCAGGCA GTCACTGCTG AAAGGAAAGC CAGCCTCCAC CACGCACCTG CAAGACACCG AGCTGTTGCA GCCCCAGGAA  6100
|||||||||| |||||||||| |||||||||| |||||||||| || |||| || |||||||||| |||||||
CTCAGCCCAC ACGCTGAGGT CTGGCAGGCA GTCACTGCTG AAxGGAAxGC CAGCCTCCAC CACGCAC
```

FIGURE 2: Sequence of the gene coding for human HGF-like protein. The DNA sequence of the gene coding for human HGF-like protein is shown along with its 3′-flanking sequence. Since the site of initiation of transcription has not been determined, the first nucleotide is +1. The numbers in the right margin correspond to the last nucleotide on each line. The amino acid sequences of the exons are 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 in this gene or 3′-flanking region are shown below the sequence of the gene. Identical nucleotides are indicated by a vertical line. Deletions are indicated by an x. The sequence from nucleotides 4457–4924 is similar to DNF15S1 (Welch et al., 1989), nucleotides 4849–5031 are similar to DNF15S2 (Welch et al., 1989), and nucleotides 5408–5673, 5764–5870, and 5965–6067 are similar to the complementary strand of DNF15S2 lung mRNA (nucleotides 1565–2014; Naylor et al., 1989). Additional nucleotides present in these homologous sequences are not indicated. There were five additional bases in DNF15S1, three in DNF15S2, and two in DNF15S2 lung mRNA.

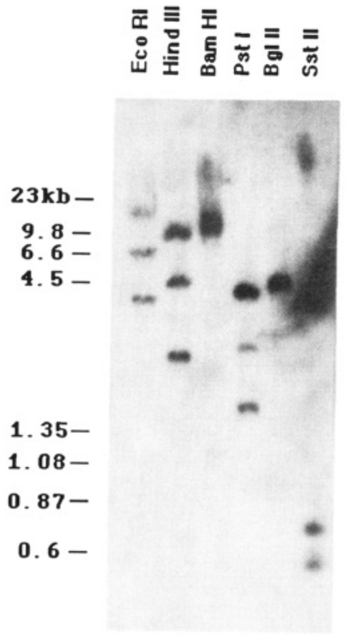

FIGURE 3: Southern analysis of genomic DNA isolated from human lymphocytes. Human lymphocyte genomic DNA (10 μg) was digested with the indicated restriction enzymes, electrophoresed, transferred to a Genescreen-plus membrane, and hybridized with a ^{32}P-labeled human genomic probe from the gene coding for HGF-like protein (680 bp *Bam*HI/*Hin*dIII probe; Materials and Methods). The migration of size markers is indicated.

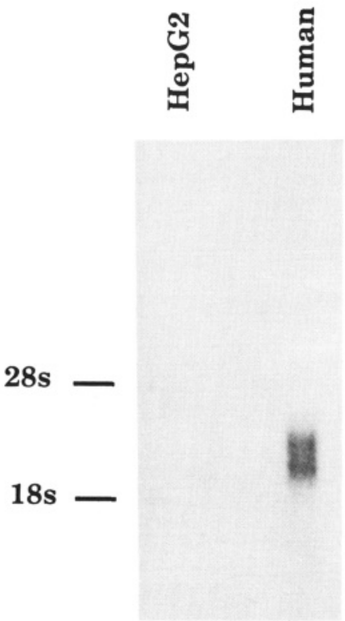

FIGURE 4: Northern analysis of total RNA isolated from human liver and HepG2 cells. Total RNA (20 μg) isolated from HepG2 cells (HepG2) and human liver (Human) was subjected to electrophoresis, transferred to a Biotrans membrane, and hybridized with ^{32}P-labeled 1950 bp *Hin*dIII fragment from the gene coding for human HGF-like protein (Materials and Methods). The migration of 28S and 18S ribosomal RNA is indicated.

present in another. When the sequences of the exons in the gene are compared to the cDNA sequences, the following nucleotides are present at the polymorphic sites found in the cDNAs: at 635, G; 1227, A; 1749, T; 1905, G; 1923, G (using the cDNA numbering scheme in Figure 6; nucleotides 1918, 3200, 4282, 4533, and 4551, respectively, in Figure 2). There was one additional difference between the cDNA and the gene where an A is present at nucleotide 38 in the cDNA (Figure 6; nucleotide 311 in the gene, Figure 2) and a G is present in the gene. This results in an amino acid substitution at residue 13 in the signal sequence of a Tyr in the cDNA to a Cys in the gene.

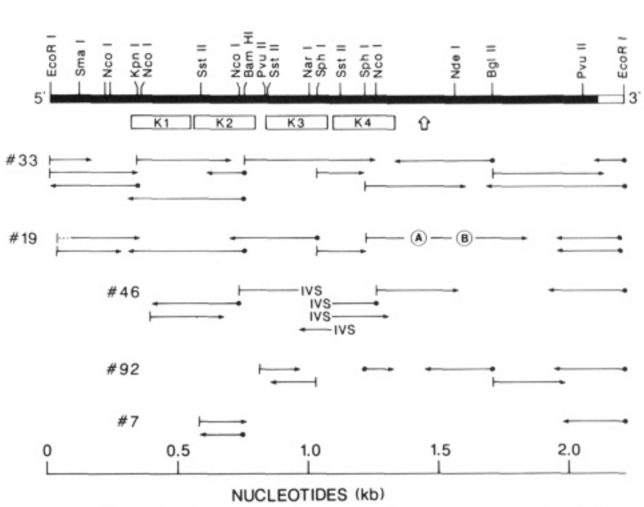

FIGURE 5: Restriction map and sequencing strategy of the cDNA coding for human HGF-like protein. A partial restriction map is indicated above the bar. The coding sequence is indicated by the blackened bar while the 3′-noncoding region is represented by the open bar. The orientation of transcription is indicated by the placement of 5′ and 3′ at the ends of the bar. Regions coding for kringle domains are schematically represented by the bars labeled K1, K2, K3, and K4. The arrow following these domains represents the putative activation site. The sequencing strategy for five cDNAs (33, 19, 46, 92, and 7) is shown by the extent of each arrow. The sequence determined on the coding strand is represented by arrows with vertical lines, and those determined on the complementary strand have closed circles at the end of arrows. Dotted parts of arrows represent regions not sequenced. A and B indicate two regions missing in cDNA 19 (see Results). IVS represents placement of intervening sequences in cDNA 46 (see Results). All overlaps were obtained. One hundred percent of the sequence was determined 2 times or more, and 98% of the sequence was determined on both strands. The scale is shown in kilobase pairs.

The translated amino acid sequences of the gene and cDNA predict a protein of 80 325 molecular weight containing 711 amino acids. The amino acid composition is Ala-35, Arg-60, Asn-27, Asp-37, Cys-44, Gln-41, Glu-40, Gly-64, His-24, Ile-15, Leu-53, Lys-22, Met-10, Phe-25, Pro-56, Ser-32, Thr-43, Trp-20, Tyr-19, and Val-44. There are three potential N-linked carbohydrate addition sites at asparagines in the sequence Asn-X-Thr/Ser at positions 72, 296, and 615 (Figure 6). 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. There is a stretch of five leucines that is a typical hydrophobic stretch found in signal sequences which are responsible for spanning the membrane during secretion. Comparison of the amino-terminal sequence to a consensus sequence compiled for known signal peptidase cleavage sites (von Heijne, 1983; Watson, 1984) predicts that the cleavage site could be between residues Gly-31 and Thr-32 (Figure 6). Amino acids with small uncharged side chains are typically found at the cleavage site while hydrophobic or small neutral residues are found at positions −3 and −4 (cleavage sites are between −1 and +1). If the cleavage site is after Gly-31, then Val and Leu are at positions −4 and −3, respectively. There is no apparent propeptide following the signal sequence like that found in vitamin K dependent proteins since there are no Arg-X-Arg/Lys-Arg sequences near the amino-terminal end of the putative protein.

On Northern analysis of the human liver RNA, there are two predominate mRNA species of 2.4 and 3.0 kilobases for human HGF-like protein (Figure 4). All cDNAs were 2.2 kbp

```
                                    10                                      20                                      30                                      40
Met Gly Trp Leu Pro Leu Leu Leu Leu Leu Thr Gln Tyr Leu Gly Val Pro Gly Gln Arg Ser Pro Leu Asn Asp Phe Gln Val Leu Arg Gly Thr Glu Leu Gln His Leu Leu His Ala
ATG GGG TGG CTC CCA CTC CTG CTG CTT CTG ACT CAA TAC TTA GGG GTC CCT GGG CAG CGC TCG CCA TTG AAT GAC TTC CAA GTG CTC CGG GGC ACA GAG CTA CAG CAC CTG CTA CAT GCG   120

                                    50                                      60                                      70      *                               80
Val Val Pro Gly Pro Trp Gln Glu Asp Val Ala Asp Ala Glu Glu Cys Ala Gly Arg Cys Gly Pro Leu Met Asp Cys Arg Ala Phe His Tyr Asn Val Ser Ser His Gly Cys Gln Leu
GTG GTG CCC GGG CCT TGG CAG GAG GAT GTG GCA GAT GCT GAA GAG TGT GCT GGT CGC TGT GGG CCC TTA ATG GAC TGC CGG GCC TTC CAC TAC AAC GTG AGC AGC CAT GGT TGC CAA CTG   240

                                    90                                      100                                     110                                     120
Leu Pro Trp Thr Gln His Ser Pro His Thr Arg Leu Arg Arg Ser Gly Arg Cys Asp Leu Phe Gln Lys Lys Asp Tyr Val Arg Thr Cys Ile Met Asn Asn Gly Val Gly Tyr Arg Gly
CTG CCA TGG ACT CAA CAC TCG CCC CAC ACG AGG CTG CGG CGT TCT GGG CGC TGT GAC CTC TTC CAG AAG AAA GAC TAC GTA CGG ACC|TGC ATC ATG AAC AAT GGG GTT GGG TAC CGG GGC   360
                                                                                                                   └K1

                                    130                                     140                                     150                                     160
Thr Met Ala Thr Thr Val Gly Gly Leu Pro Cys Gln Ala Trp Ser His Lys Phe Pro Asn Asp His Lys Tyr Thr Pro Thr Leu Arg Asn Gly Leu Glu Glu Asn Phe Cys Arg Asn Pro
ACC ATG GCC ACG ACC GTG GGT GGC CTG CCC TGC CAG GCT TGG AGC CAC AAG TTC CCG AAT GAT CAC AAG TAC ACG CCC ACT CTC CGG AAT GGC CTG GAA GAG AAC TTC TGC CGT AAC CCT   480

                                    170                                     180                                     190                                     200
Asp Gly Asp Pro Gly Gly Pro Trp Cys Tyr Thr Thr Asp Pro Ala Val Arg Phe Gln Ser Cys Gly Ile Lys Ser Cys Arg Glu Ala Ala Cys Val Trp Cys Asn Gly Glu Glu Tyr Arg
GAT GGC GAC CCC GGA GGT CCT TGG TGC TAC ACA ACA GAC CCT GCT GTG CGC TTC CAG AGC TGC GGC ATC AAA TCC TGC|CGG GAG GCC GCG|TGT GTC TGG TGC AAT GGC GAG GAA TAC CGC   600
                                                                                                    K1┘              └K2

                                    210                                     220                                     230                                     240
Gly Ala Val Asp Arg Thr Glu Ser Gly Arg Glu Cys Gln Arg Trp Asp Leu Gln His Pro His Gln His Pro Phe Glu Pro Gly Lys Phe Leu Asp Gln Gly Leu Asp Asp Asn Tyr Cys
GGC GCG GTA GAC CGC ACG GAG TCA GGG CGC GAG TGC CAG CGC TGG GAT CTT CAG CAC CCG CAC CAG CAC CCC TTC GAG CCG GGC AAG TTC CTC GAC CAA GGT CTG GAC GAC AAC TAT TGC   720

                                    250                                     260                                     270                                     280
Arg Asn Pro Asp Gly Ser Glu Arg Pro Trp Cys Tyr Thr Thr Asp Pro Gln Ile Glu Arg Glu Phe Cys Asp Leu Pro Arg Cys Gly Ser Glu Ala Gln Pro Arg Gln Glu Ala Thr Thr
CGG AAT CCT GAC GGC TCC GAG CGG CCA TGG TGC TAC ACT ACG GAT CCG CAG ATC GAG CGA GAG TTC TGT GAC CTC CCC CGC TGC|GGG TCC GAG GCA CAG CCC CGC CAA GAG GCC ACA ACT   840
                                                                                                            K2┘

                                    290                     *               300                                     310                                     320
Val Ser Cys Phe Arg Gly Lys Gly Glu Gly Tyr Arg Gly Thr Ala Asn Thr Thr Thr Ala Gly Val Pro Cys Gln Arg Trp Asp Ala Gln Ile Pro His Gln His Arg Phe Thr Pro Glu
GTC AGC|TGC TTC CGC GGG AAG GGT GAG GGC TAC CGG GGC ACA GCC AAT ACC ACC ACT GCG GGC GTA CCT TGC CAG CGT TGG GAC GCG CAA ATC CCT CAT CAG CAC CGA TTT ACG CCA GAA   960
       └K3

                                    330                                     340                                     350                                     360
Lys Tyr Ala Cys Lys Asp Leu Arg Glu Asn Phe Cys Arg Asn Pro Asp Gly Ser Glu Ala Pro Trp Cys Phe Thr Leu Arg Pro Gly Met Arg Ala Ala Phe Cys Tyr Gln Ile Arg Arg
AAA TAC GCG TGC AAA GAC CTT CGG GAG AAC TTC TGC CGG AAC CCC GAC GGC TCA GAG GCG CCC TGG TGC TTC ACA CTG CGG CCC GGC ATG CGC GCG GCC TTT TGC TAC CAG ATC CGG CGT  1080

                                    370                                     380                                     390                                     400
Cys Thr Asp Asp Val Arg Pro Gln Asp Cys Tyr His Gly Ala Gly Glu Gln Tyr Arg Gly Thr Val Ser Lys Thr Arg Lys Gly Val Gln Cys Gln Arg Trp Ser Ala Glu Thr Pro His
TGT|ACA GAC GAC GTG CGG CCC CAG GAC|TGC TAC CAC GGC GCA GGG GAG CAG TAC CGC GGC ACG GTC AGC AAG ACC CGC AAG GGT GTC CAG TGC CAG CGC TGG TCC GCT GAG ACG CCG CAC  1200
K3┘                                └K4
                                    410                                     420                                     430                                     440
Lys Pro Gln Phe Thr Phe Thr Ser Glu Pro His Ala Gln Leu Glu Glu Asn Phe Cys Arg Asn Pro Asp Gly Asp Ser His Gly Pro Trp Cys Tyr Thr Met Asp Pro Arg Thr Pro Phe
AAG CCG CAG TTC ACG TTT ACC TCC GAA CCG CAT GCA CAA CTG GAG GAG AAC TTC TGC CGG AAC CCA GAT GGG GAT AGC CAT GGG CCC TGG TGC TAC ACG ATG GAC CCA AGG ACC CCA TTC  1320

                                    450                                     460                                     470                                     480
Asp Tyr Cys Ala Leu Arg Arg Cys Ala Asp Asp Gln Pro Pro Ser Ile Leu Asp Pro Pro Asp Gln Val Gln Phe Glu Lys Cys Gly Lys Arg Val Asp Arg Leu Asp Gln Arg Arg Ser
GAC TAC TGT GCC CTG CGA CGC TGC|GCT GAT GAC CAG CCG CCA TCA ATC CTG GAC CCC CCA GAC CAG GTG CAG TTT GAG AAG TGT GGC AAG AGG GTG GAT CGG CTG GAT CAG CGG CGT TCC  1440
                            K4┘
            ↓                       490                                     500                                     510                                     520
Lys Leu Arg Val Val Gly Gly His Pro Gly Asn Ser Pro Trp Thr Val Ser Leu Arg Asn Arg Gln Gly Gln His Phe Cys Gly Gly Ser Leu Val Lys Glu Gln Trp Ile Leu Thr Ala
AAG CTG CGC GTG GTT GGG GGC CAT CCG GGC AAC TCA CCC TGG ACA GTC AGC TTG CGG AAT CGG CAG GGC CAG CAT TTC TGC GGG GGG TCT CTA GTG AAG GAG CAG TGG ATA CTG ACT GCC  1560

                                    530                                     540                                     550                                     560
Arg Gln Cys Phe Ser Ser Cys His Met Pro Leu Thr Gly Tyr Glu Val Trp Leu Gly Thr Leu Phe Gln Asn Pro Gln His Gly Glu Pro Ser Leu Gln Arg Val Pro Val Ala Lys Met
CGG CAG TGC TTC TCC TCC TGC CAT ATG CCT CTC ACG GGC TAT GAG GTA TGG TTG GGC ACC CTG TTC CAG AAC CCA CAG CAT GGA GAG CCA AGC CTA CAG CGG GTC CCA GTA GCC AAG ATG  1680

                                    570                                     580                                     590                                     600
Val Cys Gly Pro Ser Gly Ser Gln Leu Val Leu Leu Lys Leu Glu Arg Ser Val Thr Leu Asn Gln Arg Val Ala Leu Ile Cys Leu Pro Pro Glu Trp Tyr Val Val Pro Pro Gly Thr
GTG TGT GGG CCC TCA GGC TCC CAG CTT GTC CTG CTC AAG CTG GAG AGA TCT GTG ACC CTG AAC CAG CGC GTG GCC CTG ATC TGC CTG CCC CCT GAA TGG TAT GTG GTG CCT CCA GGG ACC  1800

                                    610                 *                   620                                     630                                     640
Lys Cys Glu Ile Ala Gly Trp Gly Glu Thr Lys Gly Thr Gly Asn Asp Thr Val Leu Asn Val Ala Leu Leu Asn Val Ile Ser Asn Gln Glu Cys Asn Ile Lys His Arg Gly Arg Val
AAG TGT GAG ATT GCA GGC TGG GGT GAG ACC AAA GGT ACG GGT AAT GAC ACA GTC CTA AAT GTG GCC TTG CTG AAT GTC ATC TCC AAC CAG GAG TGT AAC ATC AAG CAC CGA GGA CGT GTG  1920

                                    650                                     660                                     670                                     680
Arg Glu Ser Glu Met Cys Thr Glu Gly Leu Leu Ala Pro Val Gly Ala Cys Glu Gly Asp Tyr Gly Gly Pro Leu Ala Cys Phe Thr His Asn Cys Trp Val Leu Glu Gly Ile Ile Ile
CGT GAG AGT GAG ATG TGC ACT GAG GGA CTG TTG GCC CCT GTG GGG GCC TGT GAG GGT GAC TAC GGG GGC CCA CTT GCC TGC TTT ACC CAC AAC TGC TGG GTC CTG GAA GGA ATT ATA ATC  2040

                                    690                                     700                                     710
Pro Asn Arg Val Cys Ala Arg Ser Arg Trp Pro Ala Val Phe Thr Arg Val Ser Val Phe Val Asp Trp Ile His Lys Val Met Arg Leu Gly ***
CCC AAC CGA GTA TGC GCA AGG TCC CGC TGG CCA GCT GTC TTC ACG CGT GTC TCT GTG TTT GTG GAC TGG ATT CAC AAG GTC ATG AGA CTG GGT TAG GCCCAGCCTTGATGCCATATGCCTTGGGGAG  2167

GACAAAACTTCTTGTCAGACATAAAGCCATGTTTCCTCTTTATGCCTGTAAAAAAAAAAAAAAAAAAA                                                                                           2236
```

FIGURE 6: DNA sequence and translated amino acid sequence of the cDNA coding for human HGF-like protein. The nucleotide sequence is shown below the translated amino acid sequence. The first 16 bp of sequence were determined in the first exon of the gene coding for human HGF-like protein. The longest cDNA (33) starts at nucleotide 17. Amino acids are numbered every 10 residues starting with the putative initiator methionine as 1. The last nucleotide of each line is numbered. The three potential N-linked glycosylation sites are indicated by an asterisk. Each kringle domain is marked, and the putative activation site is indicated by an arrow. Amino acids (in boxes) at 522, 568, and 661 correspond to the active-site amino acids that have been changed from His to Gln, from Asp to Gln, and from Ser to Tyr, respectively, in the active site of serine proteases. Polymorphic sites are at nucleotide positions 38, 635, 1227, 1749, 1905, and 1923 where nucleotides G, T, G, T, A, and G are present in other cDNAs or the gene at these sites, respectively.

or smaller. The 3′-noncoding regions of all cDNAs were the same length, so it appears unlikely that the size difference is due to differential use of polyadenylation sites. In mouse liver, only one transcript size of 2.4 kilobases is identified using human or mouse cDNA probes for HGF-like protein (Degen et al., 1991). A likely explanation is that both mRNA species are transcribed from the same gene using different promoters but alternative splicing of exons occurs. The larger transcript may contain additional 5′-noncoding sequence in an as yet unidentified exon(s) or may contain additional coding sequence. Isolation of longer length cDNAs will help address this question. It is also possible that the larger mRNA species represents an mRNA transcribed from a closely related gene. Our inability to accurately determine a transcription start site for this gene may be due to the variable size of the mRNA.

Search of the DNA and Protein Database. The DNA sequence of the gene and cDNA and its translated amino acid sequence were compared against the GenBank (release 64) and NBRF databases (release 25) to search for sequences in common with other genes and proteins and to locate repetitive sequences. Homology was found with kringles in all kringle-containing proteins (33–66% identical) and numerous serine proteases (30–45% identical). From this analysis, we determined that the gene in phage L5 and L5/3 codes for a putative protein with four kringle domains followed by a serine protease-like domain. This domain structure is identical with that found in HGF (Nakamura et al., 1989). The kringles are between 37 and 66% identical with the kringles in HGF while the serine protease domain is 45% identical. Therefore, L5 protein is a HGF-like protein. Comparison of the four kringle domains in HGF-like protein to other human kringles is shown in Table I. The amino acids found in the active site of serine proteases have been changed from His to Gln, Asp to Gln, and Ser to Tyr at positions 522, 568, and 661, re-

Table I: Comparison of the Amino Acid Sequence of the Four Kringle Domains in Human HGF-like Protein to Other Human Kringle Domains

		homology to kringles in human HGF-like protein (%)			
protein	kringle	1	2	3	4
prothrombin	1	47	44	43	49
prothrombin	2	41	34	35	41
plasminogen	1	53	44	43	53
plasminogen	2	43	55	46	41
plasminogen	3	47	50	51	46
plasminogen	4	51	44	47	53
plasminogen	5	55	48	48	50
urokinase		35	36	33	37
t-PA[a]	1	36	40	33	35
t-PA	2	36	38	35	41
factor XII		35	34	40	37
HGF[b]	1	46	44	50	43
HGF	2	48	60	49	46
HGF	3	48	51	66	45
HGF	4	38	37	40	49
HGF-like	1		43	46	48
HGF-like	2			51	46
HGF-like	3				54

[a] Tissue plasminogen activator. [b] Hepatocyte growth factor.

spectively. Therefore, we anticipate that this protein has no proteolytic activity. Between the kringle domain region and the serine protease-like domain is amino acid sequence that is typically found at the activation sites of other coagulation and fibrinolytic proteins with serine protease activity. Residue 483 is an Arg (Figure 6) followed by the sequence Val-Val-Gly-Gly that is typically found at the amino-terminal end of serine proteases. On the basis of this sequence, it is possible that HGF-like protein is proteolytically cleaved to yield a two-chain molecule held together by disulfide bonds or cleaved into two separate polypeptides.

Other significant findings from the database search were the presence of sequences homologous to human 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) in exon 17 to the 3′ end of the sequence presented in Figure 2. DNF15S2 and DNF15S1 are homologous loci found on human chromosomes 3 and 1, respectively. DNF15S2 lung mRNA is transcribed from the DNF15S2 locus. The region from 4457 to 4924 (Figure 2) is 94.3% identical with DNF15S1. This homology overlaps with the DNF15S2 sequence at 4849 (Figure 2). DNF15S2 is homologous to nucleotides 4849–5031 (Figure 2) and is 97% identical with the gene sequence. Sequence coding for DNF15S2 lung mRNA is homologous to three regions in the gene at nucleotides 5408–5673 (94.8%), 5764–5870 (86.0%), and 5965–6067 (96.1%; Figure 2) on the complementary strand. These same three regions are 76.9, 91.6, and 79.6% identical with the 3′ end of the rat acyl-peptide hydrolase mRNA.

The L5 gene contains no other repetitive sequences such as those in the Alu family. There is only one intervening sequence large enough to contain a copy of this repeat.

Discussion

Growth factors are important for normal developmental processes, as well as healing of wounds. Their abnormal expression has been implicated in neoplasia and other proliferative disorders. The kringle-containing protein HGF was originally identified as a potent growth factor involved in liver regeneration after liver injury or partial hepatectomy. It is now known that HGF functions as a growth factor for a broad spectrum of tissues and cell types. Messenger RNA for HGF has been identified in lung, kidney, spleen, thymus, brain, and liver in rats (Tashiro et al., 1990). Although HGF is not a tyrosine kinase, it does induce autophosphorylation of a 145-kDa protein in target cells that has been identified as the c-met tyrosine kinase (Rubin et al., 1991; Bottaro et al., 1991). This protein binds HGF as its cell-surface receptor and mediates the mitogenic signal of HGF. The transcript for c-met has been identified in many tissues and correlates with the broad cell specificity of HGF (Bottaro et al., 1991).

HGF is a heterodimer of 82 kDa composed of an α- and a β-subunit of 69K and 34K molecular weight, respectively. Both subunits are encoded in the same 6-kilobase mRNA and result from proteolytic processing of the synthesized single polypeptide chain. The cDNAs for human and rat HGF have been cloned and characterized by several groups (Miyazawa et al., 1989; Nakamura et al., 1989; Okajima et al., 1990; Seki et al., 1990; Tashiro et al., 1990; Rubin et al., 1991).

HGF has no obvious homology with other known growth factors but is 38% homologous to plasminogen. It contains four kringle domains followed by a serine protease-like domain where the active-site His and Ser have been changed to Gln and Tyr, respectively. HGF has no detectable protease activity. At present, the function of the kringle domains in HGF is unknown.

The gene present in phage L5 and its cDNA code for a protein with a similar domain structure as HGF with four kringles followed by a serine protease-like domain (Figure 7). The four kringles are most similar to kringles in plasminogen and HGF (Table I). Interestingly, the fourth kringle in L5 has the highest degree of similarity with kringles 1 and 4 of plasminogen, both of which are lysine- and fibrin-binding domains (Vali & Patthy, 1984). Tulinsky et al. (1988b) have modeled lysine/fibrin-binding kringles after the three-dimensional structure of kringle 1 of prothrombin and have deduced that Asp-55, Asp-57, and Arg-71 (numbering is standardized to plasminogen kringle 5) are crucial residues in binding lysine and fibrin. In the fourth kringle in L5, only the two aspartic acids are present (residues 423 and 425; Figure 6) while a Pro is present in place of Arg-71 (residue 439; Figure 6). Therefore, it is unlikely that the fourth kringle of HGF-like protein binds lysine or fibrin.

The serine protease-like domain of HGF-like protein is 45% identical with the same domain in HGF. The amino acids traditionally at the active site of serine proteases have been changed from His to Gln, Asp to Gln, and Ser to Tyr (residues 522, 568, and 661, respectively, in Figure 6). In HGF, the same differences are present for His and Ser, but the active-site Asp has not been changed.

Amino acid residues 56–103 in human HGF-like protein are homologous to the preactivation peptide (PAP) in plasminogen and HGF (Figure 6). The PAP region in plasminogen is between the amino-terminal end of the mature protein (Glu-plasminogen) and the plasmin activation site between Lys-77 and Lys-78 (Wallen, 1978). Both lysines are conserved in HGF-like protein (residues 103 and 104 in Figure 6). Cleavage at this site would remove a peptide of 103 amino acids from the protein (including the putative signal peptide) if it is not disulfide-bonded to the remainder of the protein (there is one additional cysteine in this region). The four cysteines that form the two disulfide bonds in the PAP region of plasminogen are also conserved in HGF-like protein (Cys^{56}–Cys^{78} and Cys^{60}–Cys^{66} in Figure 6). The PAP region in HGF-like protein is only 27% homologous to the same

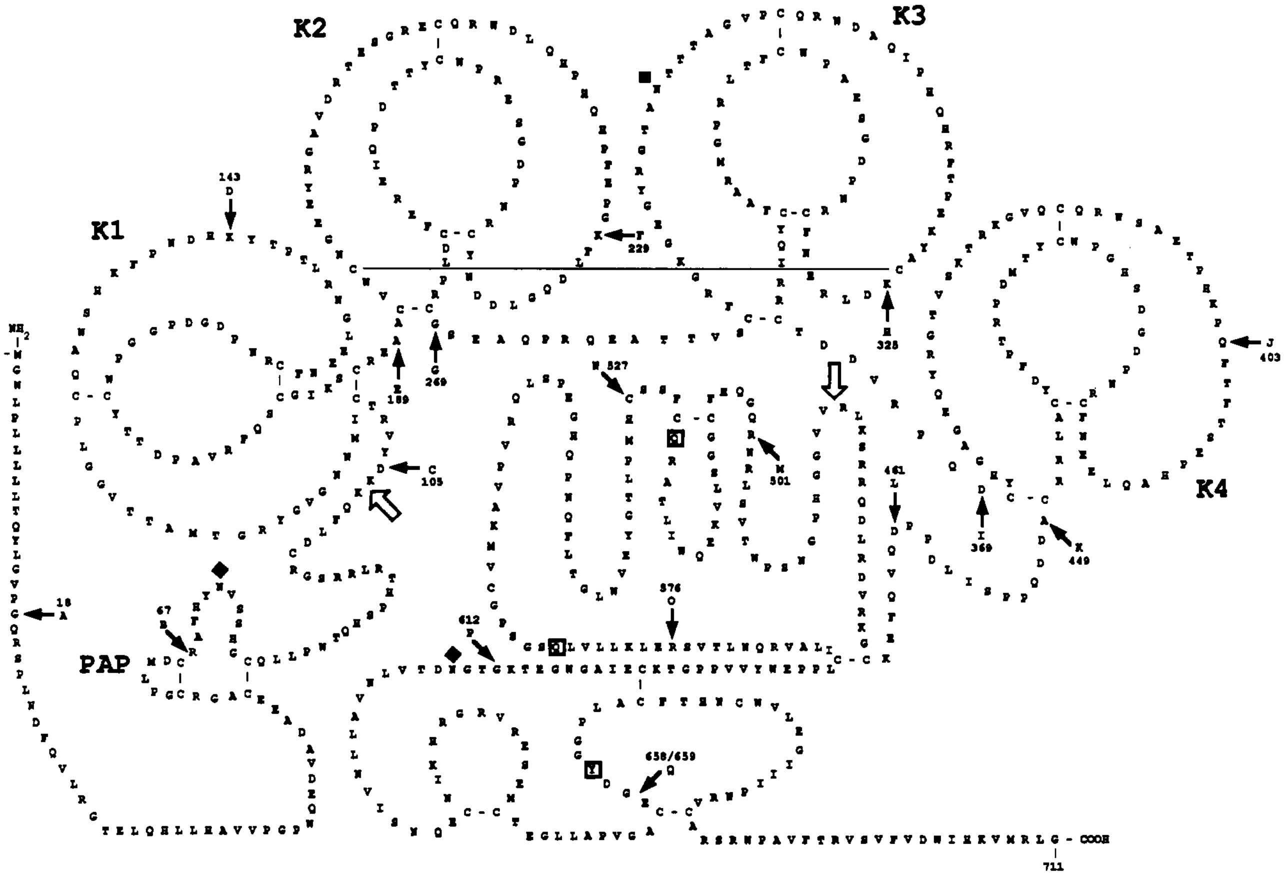

FIGURE 7: Amino acid sequence of human HGF-like protein and the location of intervening sequences with respect to the coding sequence in the gene coding for HGF-like protein. The amino acid sequence of human HGF-like protein is shown starting with residue 1 at the amino-terminal end and ending with residue 711 at the carboxy-terminal end. Placement of all disulfide bonds was determined solely on the basis of homology of this protein sequence to plasminogen, where placement of disulfides has been determined. The four kringle domains are indicated by K1, K2, K3, and K4. The region homologous to the preactivation peptide of plasminogen is indicated by PAP. The three potential N-linked glycosylation sites are indicated by diamonds. Two potential protease cleavage sites are indicated by open arrows. The sequence following the second open arrow is homologous to other serine proteases. The active-site amino acids His, Asp, and Ser have been changed to Gln, Gln, and Tyr, respectively, and are indicated in boxes. The positions of intervening sequences (A–Q) are represented by solid arrows at or between the indicated amino acid residues.

region in human plasminogen.

On the basis of homology with plasminogen and HGF, the gene for HGF-like protein codes for a protein with a putative signal peptide followed by a preactivation peptide, four kringle domains, and a serine protease-like domain (Figure 7).

Of the 44 cysteines in HGF-like protein, 40 are in comparable positions in plasminogen (Sottrup-Jensen et al., 1978). On the basis of this homology, putative disulfide bonds within domains of HGF-like protein would be between the following cysteine residues: in the PAP region, 56 and 78, 60 and 66; in the first kringle, 110 and 186, 131 and 169, 157 and 181; in the second kringle, 191 and 268, 212 and 251, 240 and 263; in the third kringle, 283 and 361, 304 and 343, 332 and 355; in the fourth kringle, 370 and 448, 391 and 431, 419 and 443; in the serine protease domain, 507 and 523, 602 and 667, 632 and 646, 657 and 685 (Figure 7). Cysteines involved in disulfide bonds between domains are residues 194 and 324 between the second and third kringles and residues 468 and 588 between the activation peptide following the fourth kringle domain and the serine protease domain (Figure 7). Cysteines at positions 98, 527, 562, and 672 are not conserved in plasminogen, and therefore no tentative disulfide pattern has been proposed.

The placement of intervening sequences in the gene coding for HGF-like protein with respect to the coding sequence and domain structure of the protein is very similar to the placement of intervening sequences in the gene coding for human plasminogen (Petersen et al., 1990). Intervening sequence B (nucleotides 1171–1250; Figure 2) is placed in the same position with respect to the PAP region as intervening sequence B in plasminogen. There are intervening sequences flanking all four kringle domains, which has been found to be the case in all the genes characterized to date coding for other kringle-containing proteins. The placement of the internal intervening sequence with respect to the amino acid sequence of each kringle is the same for the first and fourth kringles of HGF-like protein (intervening sequences D and J at nucleotides 1556–1632 and 3055–3181, respectively, in Figure 2). This placement is also identical with the placement of intervening sequence D in the first kringle of human plasminogen (Petersen et al., 1990). Intervening sequence F in the second kringle of HGF-like protein (nucleotides 1970–2113; Figure 2) is in an identical position as intervening sequence F in the second kringle of human plasminogen. Finally, intervening sequence H that interrupts the third kringle of HGF-like protein (nucleotides 2604–2723; Figure 2) is placed in an identical position as intervening sequences H and J that interrupt the third and fourth kringles in plasminogen, respectively. Intervening sequences M, O, P, and Q at nucleotides 3653–3827, 4180–4260, 4368–4462, and 4603–4721, respectively (Figure 2), in the serine protease domain of HGF-like protein are placed in identical positions with regard to the amino acid sequence as intervening sequences N, P, Q, and R in the gene for plasminogen, respectively. The only difference is that intervening sequence N (nucleotides 3906–4032; Figure 2) in the gene coding for HGF-like protein is shifted three base pairs downstream of intervening sequence O in the gene for plasminogen when the two genes are compared. The only significant difference between the structures of the genes coding for HGF-like protein and plasminogen is that the region between the last kringle domain and the activation site is encoded by one exon (exon 14) in plasminogen and two in HGF-like protein (exons 12 and 13; Figure 2).

All of these intervening sequences flanking kringles in the gene coding for HGF-like protein are of the type I class (Sharp, 1981). All four kringles in HGF-like protein are encoded by two exons each and are interrupted by type II intervening sequences. This arrangement of type I and type II intervening sequences has been found for all other kringle domains except in the case of the second kringle of prothrombin where there is no internal intervening sequence (Degen et al., 1983). Patthy (1987) originally observed that introns flanking all kringle domains were always in the type I phase class. Phase I intervening sequences appear to be dominant at domain junctions and are consistent with the importance of exon shuffling in the evolution of genes coding for multi-domain-containing proteins since the reading frame remains intact.

Recently, a small amount of amino acid sequence was reported for mouse scatter factor, a cytokine secreted from certain fibroblasts that enhances movement and causes the dissociation and scattering of epithelial cells (Gheradi & Stoker, 1990). These authors speculated that scatter factor and HGF might be the same or highly related proteins. The putative human HGF-like protein has only 9 identical residues when compared with the 22 reported. The sequence reported for mouse scatter factor is sequence typically found at the amino terminal of serine proteases and, therefore, is indicative that this protein is a serine protease. Its biological function is consistent with serine protease activity.

Chromosomal abnormalities have been found in a number of neoplastic diseases which in some cases have been found to be associated with the activation of a protooncogene or the loss of a gene that suppresses tumor growth. Tumor suppressor genes are genes expressed in normal cells that play regulatory roles in cell proliferation, differentiation, and other cellular events. Loss or inactivation of these genes is oncogenic. Tumor suppressor genes that have been identified are involved in cell cycle control, signal transduction, angiogenesis, and development (Sager, 1989).

Deletion of the short arm of human chromosome 3 has been implicated 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 reside on chromosome 3p which manifest their transformed phenotype upon their inactivation. The chromosomal locus DNF15S2 (formerly identified as D1S1) is a restriction fragment length polymorphism probe that most consistently is associated with loss of heterozygosity in SCLC, being detected in virtually 100% of SCLC.

DNF15S2 was first identified as a locus on human chromosome 1 by in situ hybridization using the anonymous DNA clone λH3 (also called λCh4A-H3; Harper & Saunders, 1981). It is now known that the probe H3H2 isolated from λH3 hybridizes with the locus DNF15S2 on chromosome 3 at 3p21 and two loci on chromosome 1 (DNF15S1A and B) at 1p36.

Genomic fragments have been identified that hybridize with various probes isolated from λH3 containing the DNF15S2 locus (Harper & Saunders, 1981). Probe H3H2 (a 2.0 kbp HindIII fragment) hybridizes with genomic *Hin*dIII fragments of 8.0, 3.8, and 2.0 kbp; the 2.0 kbp fragment is present on chromosome 3 while the others represent the two loci on chromosome 1 (Carritt et al., 1986). Our results from Southern analysis of human genomic DNA agree with what others have found using the probe H3H2 (Figure 3). We saw the same complex hybridization pattern indicating the presence of multiple copies in the genome. The probe H3H2 from the DNF15S2 locus on human chromosome 3 is part of the human HGF-like gene (nucleotides 918–2868; Figure 2). Therefore, the gene coding for HGF-like protein is on human chromosome 3. This region contains eight exons, six of which code for kringle domains. Since H3H2 also hybridizes to two loci on human chromosome 1, it is likely that additional genes or pseudogenes coding for kringle-containing proteins are present at these loci.

Genomic clones have been isolated for all three DNF15 loci on human chromosomes 1 and 3. DNA sequence has been obtained for DNF15S2 and DNF15S1 where the loci diverge (approximately 2 kbp 3′ to the probe H3H2; Welch et al., 1989). These sequences are present at the 3′ end of the sequence determined for the gene coding for human HGF-like protein (Figure 2). The sequence that is overlapped by these two published sequences represents the junction where the two sequences diverge. DNF15S2 is 97% identical with nucleotides 4849–5031 in the gene coding for HGF-like protein while the sequence for DNF15S1 is 94% identical with nucleotides 4457–4924 in Figure 2.

Published restriction maps that include the DNF15S2 locus (Boldog et al., 1989; Welch et al., 1989) agree for the most part with our experimentally determined sites and those found based on the sequence of this region. The DNF15S1 loci on human chromosome 1 have quite different maps from that shown in Figure 1 (Welch et al., 1989). A high-frequency *Hin*dIII polymorphism on chromosome 3 has been detected with the probe H3H2. Fragments of 2.2 and/or 2.0 kbp are detected at the DNF15S2 locus in Caucasians, Chinese, Malays, and Indians of Dravidian origin (Goode et al., 1986; Saha et al., 1990). On the basis of the sequence in Figure 2, these two fragments can be accounted for if the polymorphic *Hin*dIII site is at nucleotide 917.

Naylor et al. (1989) isolated a cDNA from a human lung cDNA library using another probe from the DNF15S2 locus. At that time, there was no homology between the sequence of this cDNA called DNF15S2 lung cDNA and sequences in the DNA database. The corresponding 3.3-kilobase mRNA for the cDNA does not appear to be critically involved in SCLC since most SCLC cell lines express the normal size mRNA (Naylor et al., 1989). Naylor et al. (1989) comment on the correlation of loss of acyl-peptide hydrolase activity (encoded by the ACY1 gene located at 3p21; also called aminoacylase-1) and DNF15S2 lung mRNA in a subset of SCLC tumors that might indicate that deletion of the 3p21 locus could encompass other markers, but do not comment on the possibility that DNF15S2 lung mRNA codes for acyl-peptide hydrolase.

There are three regions on the complementary strand in the 3′-flanking region of the gene coding for human HGF-like protein that are homologous to human DNF15S2 lung cDNA and rat acyl-peptide hydrolase (Naylor et al., 1989; Kobayashi et al., 1989). The first region is 444 bp downstream of the polyadenylation site in the gene for HGF-like protein (nucleotide 4963; Figure 2). Nucleotides 5408–5673, 5764–5870, and 5965–6067 (Figure 2) are 94.8, 86.0, and 96.1% identical with contiguous regions in the cDNA for DNF15S2 lung mRNA [nucleotides 1565–2014 in Naylor et al. (1989)] and 76.9, 91.6, and 79.6% identical with the 3′ end of the rat acyl-peptide hydrolase mRNA [nucleotides 1889–2360 in Kobayashi et al. (1989)]. These three regions appear to be exons since they are flanked by sequences typically found at the 5′ and 3′ ends of intervening sequences. The sequence from 5408 to 5673 is the 3′-most exon for the gene coding for DNF15S2 lung mRNA even though there is approximately 1000 bp of sequence in the cDNA 3′ to this exon sequence not found anywhere in the gene sequence for HGF-like protein. There is an $^{5'}$AATAAA$^{3'}$ sequence (5420–5425; Figure 2) 13–19 bp upstream of the 3′ end of this exon. We propose that this is a likely polyadenylation site. The sequence at the 3′ end of DNF15S2 lung mRNA [nucleotides 2015–3025 in Naylor et al. (1989)] was searched against the database, and no homologous sequences were found. This additional sequence could conceivably be in an exon upstream of the gene coding for HGF-like protein but on the complementary strand.

The enzyme acyl-peptide hydrolase catalyzes hydrolysis of the amino-terminal peptide bond of an acetylated peptide to generate an N-acetylated amino acid and a peptide with a free amino terminus. The cDNA sequence of rat acyl-peptide hydrolase (nucleotides 13–2357; Kobayashi et al., 1989) is 86% identical with nucleotides −290 to 2014 in the human DNF15S2 lung mRNA (Naylor et al., 1989). Although it has not previously been suggested, it appears that these two cDNAs code for the same protein from two different species. The complete rat cDNA sequence compares with only part of the human cDNA since the human cDNA has approximately 1000 additional bases at its 3′ end. There is no poly(A) tail or polyadenylation signal at the 3′ end of DNF15S2 lung mRNA that would indicate that it is the true 3′ end of the mRNA. The two sequences are comparable at their 5′ ends which in the rat sequence starts near the initiator methionine and in the human sequence is reported as 5′-nontranslated sequence. On the basis of the rat cDNA, it appears that the 290 bp of 5′-noncoding sequence in the human cDNA is actually part of the open reading frame. Not enough sequence was obtained to code for the initiator methionine.

From these results, it is apparent that the gene for HGF-like protein is located at the DNF15S2 locus on human chromosome 3. The gene coding for acyl-peptide hydrolase/DNF15S2 lung mRNA is also at this locus downstream of the gene coding for HGF-like protein but would be transcribed in the opposite orientation.

Since we have not identified a function for the HGF-like protein, we do not know if this is a likely candidate for the tumor suppressor gene(s) that appear(s) to be present at the DNF15S2 locus on human chromosome 3. Since it has a similar domain structure as a known growth factor with broad tissue specificity, it is interesting to speculate that the putative HGF-like protein might be a competitive inhibitor for a growth factor receptor. When this protein is defective or absent due to a chromosomal deletion, the growth factor would be free to bind to its receptor, and uncontrolled growth may occur that results in neoplasia.

Acknowledgments

We thank Sue McDowell for excellent technical assistance and John Bancroft and C. Scott Jamison for isolation of RNA. We are grateful to Russell F. Doolittle for performing some of the database searches and to Earl W. Davie for his guidance and support, especially during the beginning stages of this project. We also thank C. Scott Jamison and Mary Jo Danton for reviewing the manuscript.

References

Ausubel, F. M., Brent, R., Kingston, R. E., Moore, D. D., Seidman, J. G., Smith, J. A., & Struhl, K. (1989) *Current Protocols in Molecular Biology*, Wiley, New York.

Boldog, F., Erlandsson, R., Klein, G., & Sumegi, J. (1989) *Cancer Genet. Cytogenet. 42*, 295–306.

Bottaro, D. P., Rubin, J. S., Faletto, D. L., Chan, A. M.-L., Kmiecik, T. E., Vande Woude, G. F., & Aaronson, S. A. (1991) *Science 251*, 802–804.

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.

Carritt, B., Welch, H. M., & Parry-Jones, N. J. (1986) *Am. J. Hum. Genet. 38*, 428–436.

Castellino, F. J., & Beals, J. M. (1987) *J. Mol. Evol. 26*, 358–369.

Davie, E. W., Ichinose, A., & Leytus, S. P. (1986) *Cold Spring Harbor Symp. Quant. Biol. 51*, 509–514.

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., Stuart, L. A., Han, S., & Jamison, C. S. (1991) *Biochemistry* (following paper in this issue).

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

Esmon, C. T., & Jackson, C. M. (1974) *J. Biol. Chem. 249*, 7791–7797.

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

Furie, B., & Furie, B. C. (1988) *Cell 53*, 505–518.

Gherardi, E., & Stoker, M. (1990) *Nature 346*, 228.

Glisin, V., Crkvenjakov, R., & Byus, C. (1974) *Biochemistry 13*, 2633–2637.

Goode, M. E., vanTuinen, P., Ledbetter, D. H., & Daiger, S. P. (1986) *Am. J. Hum. Genet. 38*, 437–446.

Harper, M. E., & Saunders, G. F. (1981) *Chromosoma 83*, 431–439.

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. B., 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.

Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G., & Maniatis, T. (1978) *Cell 15*, 1157–1174.

MacGillivray, R. T. A., & Davie, E. W. (1984) *Biochemistry 23*, 1626–1634.

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.

McLean, J. W., Tomlinson, J. E., Kuang, W.-J., Eaton, D. L., Chen, E. Y., Fless, G. M., Scanu, A. M., & Lawn, R.

M. (1987) *Nature 330*, 132–137.

McMullen, B. A., Fujikawa, K., & Davie, E. W. (1991) *Biochemistry 30*, 2056–2060.

Miyazawa, K., Tsubouchi, H., Naka, D., Takahashi, K., Okigaki, M., Arakaki, N., Nakayama, H., Hirono, S., Sakiyama, O., Takahashi, K., Gohda, E., Daikuhara, Y., & Kitamura, N. (1989) *Biochem. Biophys. Res. Commun. 163*, 967–973.

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.

Patthy, L. (1987) *FEBS Lett. 214*, 1–7.

Patthy, L., Trexler, M., Vali, Z., Banyai, L., & Varadi, A. (1984) *FEBS Lett. 171*, 131–136.

Petersen, T. E., Martzen, M. R., Ichinose, A., & Davie, E. W. (1990) *J. Biol. Chem. 265*, 6104–6111.

Queen, C., & Korn, L. J. (1984) *Nucleic Acids Res. 12*, 581–599.

Rubin, J. S., Chan, A. M.-L., Bottaro, D. P., Burgess, W. H., Taylor, W. G., Cech, A. C., Hirschfield, D. W., Wong, J., Miki, T., Finch, P. W., & Aaronson, S. A. (1991) *Proc. Natl. Acad. Sci. U.S.A. 88*, 415–419.

Sager, R. (1989) *Science 246*, 1406–1412.

Saha, N., Tay, J. S. H., & Carritt, B. (1990) *Hum. Hered. 40*, 250–252.

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.

Seki, T., Ihara, I., Sugimura, A., Shimonishi, M., Nishizawa, T., Asami, O., Hagiya, M., Nakamura, T., & Shimizu, S. (1990) *Biochem. Biophys. Res. Commun. 172*, 321–327.

Sharp, P. A. (1981) *Cell 23*, 643–646.

Smith, G. E., & Summers, M. D. (1980) *Anal. Biochem. 109*, 123–129.

Sottrup-Jensen, L., Claeys, H., Zajdel, M., Petersen, T. E., & Magnusson, S. (1978) *Prog. Chem. Fibrinolysis Thrombolysis 3*, 191–209.

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.

Trexler, M., & Patthy, L. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 2457–2461.

Tulinsky, A., Park, C. H., & Skrzypczak-Jankun, E. (1988a) *J. Mol. Biol. 202*, 885–901.

Tulinsky, A., Park, C. H., Mao, B., & Llinas, M. (1988b) *Proteins: Struct., Funct., Genet. 3*, 85–96.

Vali, Z., & Patthy, L. (1984) *J. Biol. Chem. 259*, 13690–13694.

van Zonneveld, A.-J., Veerman, H., & Pannekoek, H. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 4670–4674.

von Heijne, G. (1983) *Eur. J. Biochem. 133*, 17–21.

Wallen, P. (1978) *Prog. Chem. Fibrinolysis Thrombolysis 3*, 167–181.

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.

Wickens, M. (1990) *Trends Biol. Sci. 15*, 277–281.

Wiman, B., & Wallen, P. (1977) *Thromb. Res. 10*, 213–222.

Wiman, B., Lijnen, H. R., & Collen, D. (1979) *Biochim. Biophys. Acta 579*, 142–154.

Zbar, B., Brauch, H., Talmadge, C., & Linehan, M. (1987) *Nature 327*, 721–724.