Harbor Laboratory, Cold Spring Harbor, NY.
Matrisian, L. M., Glaichenhaus, N., Gesnel, M.-C., & Breathnach, P. (1985) *EMBO J. 4*, 1435–1440.
Matrisian, L. M., LeRoy, P., Ruhlmann, C., Gesnel, M. C., & Breathnach, R. (1986) *Mol. Cell. Biol. 6*, 1679–1686.
Mignatti, P., Robbins, E., & Rifkin, D. B. (1986) *Cell (Cambridge, Mass.) 47*, 487–498.
Mount, S. M. (1982) *Nucleic Acids Res. 10*, 459–472.
Nagase, H., Brinckerhoff, C. E., Vater, C. A., & Harris, E. D., Jr. (1983) *Biochem. J. 214*, 281–288.
Perlman, D., & Halvorson, H. O. (1983) *J. Mol. Biol. 167*, 391–409.
Pless, D. D., & Lenarz, W. J. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 134–138.
Quinto, C., Quiroga, M., Swain, W. F., Nikovits, W. C., Jr., Standring, D. N., Pictet, R. L., Valenzuela, P., & Rutter, W. J. (1982) *Proc. Natl. Acad. Sci. US.A. 79*, 31–35.
Rogers, J. (1985) *Nature (London) 315*, 458–459.
Salo, T., Liotta, T., & Tryggvason, K. (1983) *J. Biol. Chem. 258*, 3058–3063.
Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*; 5463–5467.
Shaw, G., & Kamen, R. (1986) *Cell (Cambridge, Mass.) 46*, 659–667.
Takahashi, N., Takahashi, Y., & Putnam, F. W. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 73–77.
Ullrich, A., Shine, J., Chirgwin, J., Pictet, W. J., Tischer, E., Rutter, W. J., & Goodman, H. M. (1977) *Science (Washington, D.C.) 196*, 1313–1319.
Vater, C. A., Nagase, H., & Harris, E. D., Jr. (1983) *J. Biol. Chem. 258*, 9374–9382.
Welgus, H. G., Gant, G. A., Sacchettini, J. C., Roswit, W. T., & Jeffrey, J. J. (1985) *J. Biol. Chem. 260*, 13601–13606.
Whitham, S. E., Murphy, G., Angel, P., Rahmsdorf, H.-J., Smith, B. J., Lyons, A., Harris, T. J. R., Reynolds, J. J., Herrlich, P., & Docherty, A. J. P. (1986) *Biochem. J. 240*, 913–916.
Wooley, D. E., & Evanson, J. J. (1980) *Collagenase in Normal and Pathological Connective Tissues*, Wiley, New York.

Nucleotide Sequence of the Gene for Human Prothrombin†

Sandra J. Friezner Degen*,‡ and Earl W. Davie§

Children's Hospital Research Foundation and Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio 45229, and Department of Biochemistry, University of Washington, Seattle, Washington 98195

Received April 16, 1987; Revised Manuscript Received May 29, 1987

ABSTRACT: A human genomic DNA library was screened for the gene coding for human prothrombin with a cDNA coding for the human protein. Eighty-one positive λ phage were identified, and three were chosen for further characterization. These three phage hybridized with 5′ and/or 3′ probes prepared from the prothrombin cDNA. The complete DNA sequence of 21 kilobases of the human prothrombin gene was determined and included a 4.9-kilobase region that was previously sequenced. The gene for human prothrombin contains 14 exons separated by 13 intervening sequences. The exons range in size from 25 to 315 base pairs, while the introns range from 84 to 9447 base pairs. Ninety percent of the gene is composed of intervening sequence. All the intron splice junctions are consistent with sequences found in other eukaryotic genes, except for the presence of GC rather than GT on the 5′ end of intervening sequence L. Thirty copies of *Alu* repetitive DNA and two copies of partial *Kpn*I repeats were identified in clusters within several of the intervening sequences, and these repeats represent 40% of the DNA sequence of the gene. The size, distribution, and sequence homology of the introns within the gene were then compared to those of the genes for the other vitamin K dependent proteins and several other serine proteases.

The coagulation of blood in mammals is the result of many enzymatic reactions involving the sequential activation of a number of specific serine proteases by limited proteolysis. The end result of this series of reactions is the formation of an insoluble fibrin clot (Davie et al., 1979). Prothrombin participates in the final stage of this process when it is activated to thrombin by factor Xa in the presence of factor Va, calcium ions, and a phospholipid surface. Thrombin, in turn, cleaves fibrinopeptides A and B from the α and β chains of fibrinogen, respectively, to form the fibrin clot. Thrombin is also involved in the activation of factors V, VIII, and XIII and protein C, in the stimulation of platelets to undergo a change of shape, and in the regulation of proliferation of certain cell types, most notably endothelial cells (Davie et al., 1979; Esmon, 1983; Davey & Luscher, 1967; Chen & Buchanan, 1975).

Prothrombin is synthesized in the liver as a single-chain glycoprotein with a M_r of 72 000 (Barnhart, 1960; Mann & Elion, 1980). Vitamin K is required for the γ-carboxylation of prothrombin, as well as several other plasma proteins participating in coagulation. The carboxylation involves 10 amino-terminal glutamic acid residues that are converted to γ-carboxyglutamic acid (Gla;[1] Stenflo et al., 1974; Nelsestuen et al., 1974; Magnusson et al., 1975). These residues are required for prothrombin to bind calcium and thus aid in the binding of the protein to phospholipid surfaces. The biosyn-

†This work was supported in part by the Pew Memorial Trust, by Research Grants HL16919 and HL38232 from the National Institutes of Health, and by start-up funds from the Children's Hospital Research Foundation, Cincinnati, OH. S.J.F.D. is a Pew Scholar in the Biomedical Sciences.

*Author to whom correspondence should be addressed.
‡University of Cincinnati College of Medicine.
§University of Washington.

[1] Abbreviations: bp, base pairs; kb, kilobase pairs; Gla, γ-carboxyglutamic acid; t-PA, tissue plasminogen activator; u-PA, urokinase-type plasminogen activator; EDTA, ethylenediaminetetraacetic acid; SDS, sodium dodecyl sulfate.

0006-2960/87/0426-6165$01.50/0

thesis of prothrombin also includes carbohydrate addition and removal of the prepro leader sequence before secretion into the blood.

The entire primary structure of bovine and human prothrombin has been determined by protein sequence analysis (Magnusson et al., 1975; Butkowski et al., 1977; Walz et al., 1977; Thompson et al., 1977; Seegers, 1979) and sequence analysis of the cDNA (Degen et al., 1983; MacGillivray & Davie, 1984). Mature human prothrombin that circulates in the blood contains 579 amino acids, 8% carbohydrate (Degen et al., 1983; DiScipio et al., 1977), and 10 residues of γ-carboxyglutamic acid (Stenflo et al., 1974; Nelsestuen et al., 1974; Magnusson et al., 1975). The apparent carbohydrate attachment sites are Asn residues in position 78, 100, and 373. Two regions of internal homology called kringle structures are also present in prothrombin (Magnusson et al., 1975). Kringle structures are triple disulfide-bonded domains containing approximately 80 amino acids and are highly conserved. They have been implicated as domains involved in the interactions with a number of other proteins (Patthy et al., 1984). For example, the second kringle of prothrombin appears to bind to factor V, its cofactor (Esmon & Jackson, 1974), and the second kringle in tissue plasminogen activator participates in the binding of this protein to fibrin (van Zonneveld et al., 1986). The carboxy-terminal region of prothrombin contains the catalytic region, which is homologous to trypsin, as well as other serine proteases (Elion et al., 1977).

Prothrombin is synthesized with an amino-terminal extension of 43 amino acids (Degen et al., 1983; MacGillivray & Davie, 1984). This leader sequence ends with two arginine residues preceding the amino-terminal Ala present in the mature protein. Since Arg-Ala is not a typical signal peptidase cleavage site, it has been proposed that the leader sequence contains a prepeptide or signal peptide and a propeptide (Degen et al., 1983; MacGillivray & Davie, 1984). Propeptides have also been identified in other vitamin K dependent proteins (Kurachi & Davie, 1982; Fung et al., 1984; Dahlbäck et al., 1986; Hagen et al., 1986; Long et al., 1984; Pan & Price, 1985) and have been proposed to be involved in the γ-carboxylation process (Pan & Price, 1985; Suttie et al., 1987; Jorgensen et al., 1987). The signal sequence has the typical hydrophobic stretch of amino acids (Blobel et al., 1979) and is removed before secretion from the liver while the pro piece is removed by additional processing in the liver or blood (Bentley et al., 1986; Diuguid et al., 1986; Ware et al., 1986).

An almost full-length cDNA coding for human prothrombin has been previously isolated and found to code for a leader sequence of at least 36 amino acids, 579 amino acids coding for the mature molecule found in plasma, and a 97-bp 3′ noncoding region (Degen et al., 1983). The 10 glutamic acid residues that are converted to γ-carboxyglutamic acid are coded by GAG.

The genomic DNA coding for amino acid residues 144–448 in human prothrombin has been characterized by DNA sequence analysis (Degen et al., 1983). Coding regions were divided into five exons by six intervening sequences. Four copies of *Alu* repetitive DNA sequence were present in this portion of the gene. This paper presents the remaining DNA sequence (approximately 15 kb) for the gene for human prothrombin, which spans approximately 21 kilobases of DNA.

Materials and Methods

Preparation of Probes. All probes used for the isolation of the human prothrombin gene were prepared from a plasmid containing a 2005-bp cDNA inserted into the *Pst*I site of pBR322 (pHII-3; Degen et al., 1983). The cDNA codes for residues −36 to 579 and includes the 3′ noncoding region of human prothrombin. Probes 1 and 2 were isolated on a 3.5% polyacrylamide gel after digestion of pHII-3 with *Pst*I. Appropriate bands were excised, and the DNA was electroeluted, further purified, and concentrated on Elutip-d columns (Schleicher & Schuell, Keene, NH) as described by Degen et al. (1986). Probe 1 is 1570 bp in length and codes for residues −36 to 481, while the smaller 435-bp probe 2 codes for residues 482–579 and includes the entire 3′ noncoding region.

A 290-bp fragment (probe 3) coding for the 5′ end of human prothrombin (residues −36 to 56) was isolated after digestion of pHII-3 with *Pst*I and *Hin*dIII as discussed above. Probe 4 was isolated after digestion of pHII-3 with *Hin*dIII and *Bam*HI. The 1327-bp fragment codes for amio acid residues 56–499 of human prothrombin. DNA was labeled either by nick translation (Thomashow et al., 1980) or by a random-labeling procedure (Feinberg & Volgelstein, 1983, 1984).

Screening of Genomic Library. A recombinant phage library constructed by partial *Mbo*I digestion of human placental DNA and inserted into the *Bam*HI site of Charon 28 was kindly provided by Dr. Philip Leder (Harvard University). Approximately 2×10^6 phage were screened by the in situ hybridization technique of Benton and Davis (1977) as modified by Woo (1979) using combined nick-translated probes 1 and 2. Nitrocellulose filters were prehybridized 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) for at least 2 h at 68 °C. Hybridization was performed overnight at 68 °C in 2× Denhardt's solution containing 6× SSC, 1 mM EDTA, 0.5% SDS, and $\sim 1 \times 10^6$ cpm of probe per filter. Filters were washed 3 times at 68 °C in 1× SSC containing 0.5% SDS for 1 h each. After drying, the filters were exposed to X-ray film (Kodak XAR-5) overnight at −20 °C with an intensifying screen. Eighty-two positive phage were identified. Eighty-one of these remained positive on subsequent screenings. Thirty-four of the positive phage were plaque purified, and the DNA was isolated from 14 (Degen et al., 1983; Maniatis et al., 1982).

To determine the extent of hybridization of the 34 purified phage, each was plated out to a density of 10–100 phage per plate, and phage from each plate were transferred to two nitrocellulose filters and amplified overnight at 37 °C as described by Woo (1979). Filters were prepared for hybridization as discussed above. One set of filters was hybridized with a 5′ probe (probe 3), while the other set was hybridized with a probe coding for the 3′ end of human prothrombin (probe 2) by employing the same conditions used to screen the library.

Southern Hybridization Analysis. Restriction enzyme digests of positive phage DNA and subclones were electrophoresed on 0.8% agarose and transferred to nitrocellulose (Schleicher & Schuell, Keene, NH) by the Southern procedure (Southern, 1975) as modified by Smith and Summers (1980). Hybridization conditions were identical with those used for screening the genomic DNA library.

Subcloning and Restriction Enzyme Mapping. All restriction enzymes were obtained from New England Biolabs or Bethesda Research Laboratories and were used according to the manufacturer's instructions. Subcloning and transformation conditions were discussed previously (Degen et al., 1986). *Escherichia coli* containing the appropriate plasmid were identified by colony hybridization (Grunstein & Hogness, 1975) or by small-scale plasmid isolation (Birnboim & Doly,

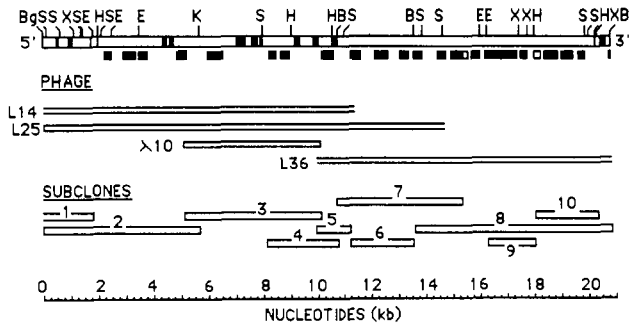

FIGURE 1: Recombinant phage, subclones, and an abbreviated restriction map for the gene coding for human prothrombin. On the top line is a representation of the gene coding for human prothrombin. The 14 coding regions or exons are indicated by solid bars and are separated by 13 open boxes representing intervening sequences. The 5′ to 3′ orientation of transcription is shown. Sites for several restriction enzymes are shown: Bg, *Bgl*II; E, *Eco*RI; S, *Sst*I; X, *Xba*I; H, *Hin*dIII; K, *Kpn*I; B, *Bam*HI. Below the map the 30 *Alu* repetitive DNA sequences and the two *Kpn* repeats are indicated by a series of solid and open boxes, respectively. Four overlapping recombinant phage containing human DNA coding for prothrombin are shown. In most cases the actual length of each insert has not been determined, and therefore, the bars have been left open ended. λ 10 was previously sequenced by Degen et al., (1983). Fragments isolated from each phage insert and cloned into pBR322 are shown under the subclone heading. The names of the subclones are 1-pL14Eco3, 2-pL25Bg15, 3-λ 10, 4-pL25Bam2.8, 5-pL25Bg11.3, 6-pL36BBg2a, 7-pL36Pst4.5, 8-pL36Bam8, 9-pL36EH2, and 10-pL36Hind2.

1979). Large-scale preparation of plasmids were performed by the method of Katz et al. (1973, 1977).

DNA Sequence Analysis. Labeling, isolation, and purification of DNA fragments has been described previously (Degen et al., 1983, 1986). Labeled fragments were subjected to base modification and cleavage by the methods of Maxam and Gilbert (1980) and analyzed by electrophoresis on 40 cm long 6% and 20% polyacrylamide gels (Sanger & Coulson, 1978). DNA sequence was analyzed on an IBM-AT computer using the programs of Queen and Korn (1984).

RESULTS

Isolation of the Human Prothrombin Gene. A portion of the gene coding for human prothrombin (coding for residues 144–448) has been previously identified and characterized (λ phage 10, Figure 1; Degen et al., 1983). λ 10 was isolated from the *Alu*I/*Hae*III genomic DNA library of Lawn et al. (1978) and was the only phage detected containing portions of the gene for human prothrombin. Therefore, a second human genomic DNA library was screened for genomic phage containing DNA coding for the 5′ and 3′ ends of the gene for human prothrombin. This library was constructed by partial *Mbo*I digestion of human placental DNA cloned into the *Bam*HI site of Charon 28 (Rimm et al., 1980).

Two million recombinant phage were screened under stringent conditions with nick-translated probes 1 and 2 (see Materials and Methods) that code for residues −36 and 579 and the entire 3′ noncoding region of the cDNA for human prothrombin (Degen et al., 1983). Of the 82 positive phage identified on the first screening, 34 were plaque purified, and the DNA was prepared from 14.

Selective screening of the 34 phage with either probe 3 from the 5′ end of the cDNA for human prothrombin or probe 2 for the 3′ end indicated that 23 of the phage hybridized with both probes. Five phage hybridized with only the 5′ end probe, while three hybridized with only the 3′ end probe. Another three phage did not hybridize with either probe, indicating that they contained DNA coding for only the middle portion of the gene.

Analysis of restriction enzyme digests of DNA prepared from 14 of the phage indicated that they overlapped since many common restriction fragments were shared among the different phage. Hybridizing fragments indicated that the size of the gene for human prothrombin was about 24 kb of DNA. This result was similar to that from a genomic Southern hybridization analysis, which indicated that the gene was approximately 19 kb in length. Three of the phage (L14, L25, and L36) were chosen for further study (Figure 1). L14 hybridized with only the probe for the 5′ end, while L25 hybridized with probes from both ends and L36 hybridized with only the probe for the 3′ end. The actual length of human DNA contained within each phage was not determined, due to the fact that upon construction of the phage library the *Bam*HI sites of Charon 28 are usually lost (Rimm et al., 1980). Consequently, the human DNA insert could not be removed intact from the phage DNA.

Restriction endonuclease fragments from each phage were subcloned into pBR322 for further characterization (see Figure 1). All subclones overlapped with each other or with that of the previously characterized λ 10 (Degen et al., 1983).

Nucleotide Sequence of the Human Prothrombin Gene. The DNA sequence of the gene coding for human prothrombin was determined 5′ and 3′ to the previously sequenced portion of the gene (λ 10 in Figure 1) by the chemical modification and degradation procedures of Maxam and Gilbert (1980). A partial restriction map of the gene is shown in Figures 1 and 2. Eighty-nine percent of the sequence was determined 2 or more times, and 52% was sequenced on both strands (Figure 2). DNA sequence overlapping λ 10 was also determined. All hybridizing bands observed by Southern hybridization analysis of human genomic DNA with human prothrombin cDNA probes was accounted for in the sequence of the gene. This provides further proof that human and bovine prothrombin are coded for by a single-copy gene (Degen et al., 1983; Irwin et al., 1985).

The complete DNA sequence of 20 801 bp comprising the human prothrombin gene is shown in Figure 3.[2] This sequence extends 447 bp upstream from the initiator methionine and 144 bp downstream from the polyadenylation site. Comparison of the DNA sequence of the gene with the cDNA sequence (Degen et al., 1983) indicated that the gene consists of 14 exons interrupted by 13 intervening seqeunces (Figures 1–3). The exons ranged in size from 25 to 315 bp (Table I), while the intervening sequences ranged from 84 to 9447 bp in length. At the present time it is not known how far the gene extends upstream from the initiator methionine since a cDNA containing an extended 5′ noncoding region has not been isolated for comparison. The average exon size was 145 bp. This calculation excludes exon I since its actual size is not known. This is similar to the average size of 150 bp in the exons found in other eukaryotic genes (Naora & Deacon, 1982; Blake, 1983). Approximately 90% of the human prothrombin gene was intervening sequence. Unlike exons, no periodicity of intervening sequence length was observed (Naora & Deacon, 1982).

The distribution of purine and pyrimidine bases was fairly evenly divided between the four nucleotides in the gene coding for human prothrombin as calculated from the initiator methionine to the polyadenylation site. The composition included 23.0% A, 25.0% C, 25.6% G, and 26.4% T. The frequency

[2] This DNA sequence includes that of λ 10 and will be submitted to GenBank. Details on the operation of GenBank can be obtained from Dr. Walter Goad, Los Alamos National Laboratory, P.O. Box 1663, Los Alamos, NM 87545.

Table I: Location and Size of Exons and Intervening Sequences in the Gene for Human Prothrombin

exon	nucleotide positions	length (bp)	amino acids	intervening sequence	type[a]	nucleotide positions	length (bp)	no. of *Alu* repeats
I	+1–79	79+[b]	−43 to −17	A	1	80–465	386	
II	466–626	161	−17 to 37	B	0	627–1 285	659	
III	1 286–1 310	25	38–46	C	1	1 311–1 552	242	
IV	1 553–1 603	51	46–63	D	1	1 604–3 929	2326	4
V	3 930–4 035	106	63–98	E	2	4 036–4 131	96	
VI	4 132–4 268	137	98–144	F	1	4 269–6 606	2338	3
VII	6 607–6 921	315	144–249	G	1	6 922–7 245	324	
VIII	7 246–7 374	129	249–292	H	1	7 375–7 458	84	
IX	7 459–7 585	127	292–334	I	2	7 586–8 742	1157	2
X	8 743–8 910	168	334–390	J	2	8 911–9 407	497	
XI	9 408–9 581	174	390–448	K	2	9 582–10 123	542	1
XII	10 124–10 305	182	448–509	L	1	10 306–19 752	9447	20[c]
XIII	19 753–19 823	71	509–532	M	0	19 824–19 969	146	
XIV	19 970–20 210	241	533–poly(A) site					

[a] Intervening sequence placement as discussed by Sharp (1981) where a type 0 indicates placement between two codons, a type 1 interrupts a codon between the first and second base, and a type 2 occurs between the second and third base of the codon. [b] The length of the 5′ noncoding region of the mRNA for human prothrombin in unknown; therefore, the length of exon I is measured from the initiator methionine. [c] This intervening sequence also has two copies of partial *Kpn* repeats.

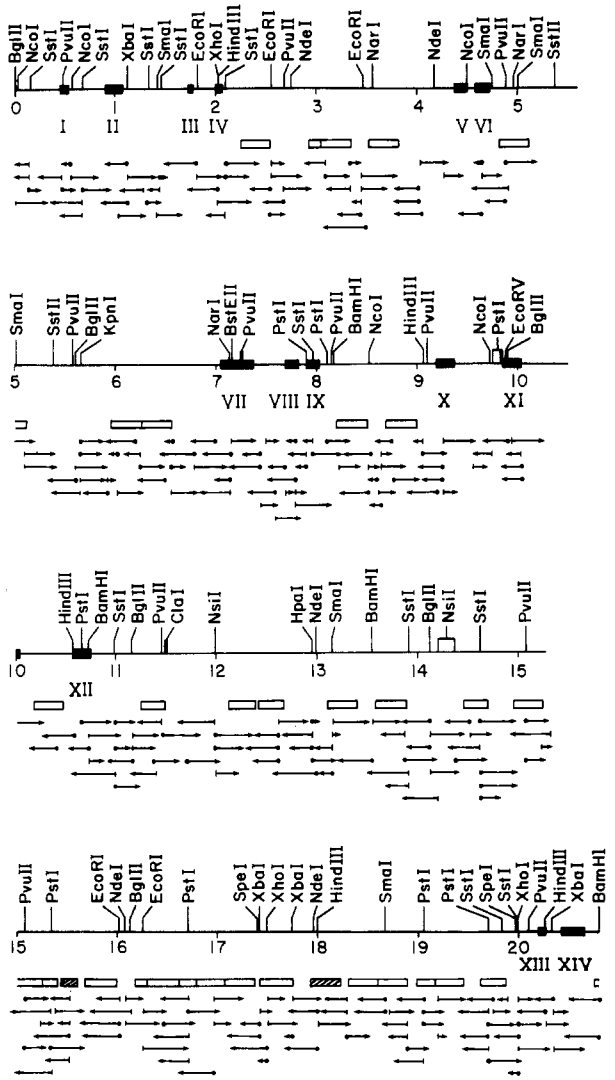

FIGURE 2: Detailed restriction map and sequencing strategy for the gene for human prothrombin. The locations of each of the 14 exons is shown with solid bars and are labeled with Roman numerals I–XIV. The 30 *Alu* and 2 *Kpn* repeats are shown with open and slashed bars, respectively. The length and direction of sequencing by the chemical degradation method are shown by the arrows. Vertical lines and closed circles at the base of each arrow represent labeling on the mRNA-like and transcribed strands, respectively. The nucleotides are listed in kilobases from 0 to 20. The sequencing strategy for the nucleotides shown between 5110 and 10050 bp was published previously by Degen et al. (1983).

Table II: Intron–Exon Splice Junction Sequences in the Gene for Human Prothrombin

intron	exon | 5′	splice junction sequences intron	3′ | exon
A	ATG	GTAAGG-------CCACCGCCTTTACAG	T
B	ACG	GTGAGC-------GCCCTTGTTTTTCAG	G
C	CAG	GTGAGC-------CTGGGTCTTTTCCAG	C
D	AAG	GTGAGC-------GTGGGGTCTCCGCAG	G
E	TGA	GTGAGT-------AATTTCCTCTTCCAG	A
F	GTG	GTAGGC-------CCCCTCACCCACCAG	G
G	GTG	GTGAGC-------CCTGGGTCCCAACAG	A
H	CAG	GTGAGG-------TGGCTTGCTCTGCAG	A
I	TTG	GTGTGT-------TGCTGCCCCTCCCAG	G
J	AAG	GTACAG-------TTGGGGTCTCTGCAG	G
K	CAG	GTGGGC-------CTTCCTTCCCCAAAG	C
L	CTG	GCAAGT-------CTGTTCTCTTTCAAG	G
M	AAG	GTAAGC-------ATCTTTCTTCTTCAG	A
consensus sequence[a]	C/A AG	GT(A/G)AGT-------(T)TTTTTTTTTT(C/T)NCAG	G

[a] From Mount (1982).

of dinucleotides within this region was typical of other eukaryotic genes (Nussinov, 1981) and was as follows: AA (5.9%), AC (4.8%), AG (7.5%), AT (4.9%), CA (7.3%), CC (7.5%), CG (2.0%), CT (8.2%), GA (5.8%), GC (6.4%), GG (7.9%), GT (5.4%), TA (4.0%), TC (6.2%), TG (8.2%), and TT (8.0%).

The sequence of splice junctions at the 5′ and 3′ ends of each intervening sequence agrees with the GT-AG rule of Breathnach et al. (1978) and with the consensus sequence of Mount (1982), except for one splice site (Table II). This sequence was GCAAGT and was located at the 5′ end of intervening sequence L. The GC is very unusual (Padgett et al., 1986; Sharp, 1987) but has been observed in the α A-crystallin gene (King & Piatigorsky, 1983), the adult chicken and duck α-globin genes (Dodgson & Engel, 1983), and the mouse, hamster, and human adenine phosphoribosyltransferase genes

(Broderick et al., 1987). This region was sequenced 4 times in the gene for prothrombin and on both strands to confirm these results. Studies performed in vitro by others indicate that genes containing mutations at the GT in the 5′ junction of the intervening sequence splice normally at that site but intermediates accumulate (Padgett et al., 1986).

From the translated DNA sequence it was determined that the prepro leader sequence for human prothrombin is 43 residues in length. Residue −43 is a methionine, and there are no other ATG codons upstream from that position in the proper reading frame. This agrees with the amino-terminal sequence predicted from the prothrombin cDNA of Jorgensen et al. (1986). Bovine prothrombin also has a leader sequence of 43 amino acids, and 36 of these residues are conserved in the two species (MacGillivray & Davie, 1984).

When the DNA sequence was determined that overlapped with the 5′ and 3′ ends of λ 10 (Degen et al., 1983), it was found that the first nine (GAATTCATG) and last five bases (CATGA) of the published sequence were not present in the prothrombin gene sequence established in the present experiments. This suggests that these nucleotides were part of the *Eco*RI linker used during construction of the genomic DNA library of Lawn et al. (1978).

Several differences were found when the gene was compared to the human prothrombin cDNA sequence (Degen et al., 1983; MacGillivray et al., 1986; Jorgensen et al., 1986). Two of these differences were reported previously by Degen et al. (1983) and do not result in an amino acid substitution. They are located at positions 7322 and 8908 in the gene (Figure 3). Two other silent substitutions at amino acids −42 and 13 (nucleotides 6 and 554) were also found. In the codon for residue −42 a G was present in the third position in the gene, while a C was identified in the cDNA by MacGillivray et al. (1986). For amino acid 13 a G was reported by MacGillivray et al. (1986), while the cDNA of Degen et al. (1983) and the gene described in the present studies have an A at the third position of the codon. Three differences that do result in an amino acid substitution, however, were identified including a His at residue −41 that was coded for in the gene and one of the cDNAs (Jorgensen et al., 1986). An Arg, however, was found in this position in the cDNA of MacGillivray et al. (1986). This results from an A to G substitution at position 8 in the gene. Previously it was reported by Degen and Davie (1986) that an Arg was present in this position, but on further investigation it was determined that the original assignment was incorrect and should have been a His. At residue −40 an A at the first position of the codon (nucleotide 10 in Figure 3) results in an Ile in the cDNA of MacGillivray et al. (1986), while a G was found in the gene and cDNA of Jorgensen et al. (1986) resulting in a Val at this position. Finally, differences were found in the codon for amino acid 121 (nucleotide 4200 in Figure 3). The ACC codon in the gene codes for Thr while the AAC codon in the cDNA codes for Asn (Degen et al., 1983). It is interesting to note that an Ile was observed in this position by amino acid sequence analysis (Walz et al., 1977; Seegers, 1979). Isoleucine could be coded for by an ATC codon. This may be the result of a polymorphism in the human prothrombin gene.

Forty percent of the gene for human prothrombin is composed of repetitive DNA. The repetitive sequences belong to the *Alu* and *Kpn* family of repeats and are present in the intervening sequences of the gene. *Alu* repetitive DNA is ubiquitous throughout the human genome with over 500 000 copies (Rinehart et al., 1981; Schmid & Jelinek, 1982). These repeats of approximately 300 bp are dimeric and usually have a poly(A) sequence at the 3′ end. Some repeats have been found to be transcribed by RNA polymerase III, but as yet their function, if any, is unknown. Thirty-nine percent of the human prothrombin gene is composed of *Alu* repetitive DNA including 30 copies present within the gene. The characteristics of these repeats are shown in Table III. On the average they are 83% homologous to the consensus sequence of Deininger et al. (1981) and are present in both orientations. Direct-repeat sequences ranging from 5 to 17 bp in length were found at both ends of these repeats. There is evidence that the site of integration of *Alu* repetitive DNA is in A+T-rich regions (Daniels & Deininger, 1985). Comparison of the direct repeats flanking the *Alu* sequences in the prothrombin gene agree with this observation (Table III). Although it has been reported that *Alu* repeats are interdispersed throughout the genome at 3000–5000-bp intervals (Schmid & Jelinek, 1982), the repeats within the human prothrombin gene are highly clustered. These clusters include five sets of tandem repeats (*Alu* repeats 2 and 3, 6 and 7, 12 and 13, 26 and 27, and 28 and 29 in Table III), and these *Alu* repeats occur in a head-to-tail orientation as well as with direct repeats flanking the entire unit. This suggests that these repeats were probably inserted together into the genome, since direct repeats are the result of the integration process. Intervening sequence L, which is 9.5 kb in length, contains 20 *Alu* repeats, and 5 of these (*Alu* repeats 20–24 in Table III) occur in head-to-tail orientation with no additional DNA between them. These five tandem repeats are flanked by 6-bp direct repeats.

Two copies of partial *Kpn*I repeats also occur within the human prothrombin gene and are located in intervening sequence L (Table III). The *Kpn* family of repetitive DNA is notable for its length polymorphism. One partial *Kpn* repeat is 170 bp in length and is 68% homologous to sequences reported by DiGiovanni et al. (1983) and Potter and Jones (1983). Also, a 10-bp direct repeat is present flanking this sequence (see Table III). The other *Kpn*I repeat is 326 bp in length and is flanked by 5-bp direct repeats. It is 59% homologous with several other human and primate *Kpn*I sequences (Potter, 1984; DiGiovanni et al., 1983).

Discussion

It has been proposed that the genes coding for serine proteases are the result of duplication and divergence from a common ancestral gene (Katayama et al., 1979; Stone et al., 1985; Neurath, 1984). Further diversification has been proposed to occur by recombination of exons from different ancestral genes and may include exon shuffling (Gilbert, 1985; Rogers, 1985; Patthy, 1985). In this way, proteins having common domains occur even when other regions of the protein are different. The organization of the prothrombin gene, as well as other coagulation and fibrinolytic proteins, reflects these observations (Degen et al., 1985; Degen & Davie, 1986).

Prothrombin is structurally homologous to several families of proteins including the other vitamin K dependent proteins and the pancreatic serine proteases. The vitamin K dependent coagulation factors (prothrombin, proteins C and S, and factors VII, IX, and X) all share a homologous Gla domain spanning approximately 50 amino acids that is located in the N-terminal region of the mature protein. This region contains all of the γ-carboxyglutamic acid residues. Two kringle domains follow this region in prothrombin and are present in other coagulation and fibrinolytic proteins including factor XII, plasminogen, tissue plasminogen activator (t-PA), and urokinase (u-PA). In the other vitamin K dependent proteins, the Gla region is followed by two to four growth factor like domains (Banyai et al., 1983). The catalytic or serine protease domain follows

```
TCTGATTAAG AACTTACGAT ATTCCATGGA CATTCCATTC CTAATCTCCT TTAGTCCTCA CAACAAAGTA TTATTCCCAT TGTATAGATG AGGAAACTGA  -348
GGCACACAGA GATGACAAGC AACCACCGCT ATATGTTAGG ATTCGAAGGA GCTCCAGGAA AGTCTCATAG CCCCACTGGC CAGAATGGGC TAAATCTCAG  -248
AGGGGGAGGG TGGGAGATGG GGGTGACAGT GACCTTTTTT GTGACTCCTC CTAGACCATC CATCCCTGCT CCCAGGAGGA CCTGTCCTCC CAGATGGTGG  -148
AGATGGACAG GAGGACTATC TACCCACCCG TCCCCACGGC CCTGACCCTC TGACCTCACC CTCTCCGCTG ATTTCTTCAT GTTAGTTCAA CATTACCCAG   -48

                                                  Met AlaHisValA rgGlyLeuGl nLeuProGly CysLeuAlaL euAlaAlaLe
AGGGGTCAGG ACAGACAATT CCTCAGTGAC CCAGGAGCTG ACACACTATG GCGCACGTCC GAGGCTTGCA GCTGCCTGGC TGCCTGGCCC TGGCTGCCCT    53

uCysSerLeu ValHisSerG lnHisV
GTGTAGCCTT GTGCACAGCC AGCATGGTAA GGGAGTGCTT GCAGGCTGGA ACAGGCTGGA GGACTGGGGT GTGGGCCCAT GGGCTGGGGT CTCCTGGCTG   153
GACAGAGCAC ACAGAGCTGG CCCCTAAGTA GGTCTCAGCC CCAGGCGGCC AGCTTAGGGA AGAAGTCAGG AGCTCAGGGC TGGAAAGAGA ATGGCTGCTT   253
CTCTCTTCCA ATATAGGGAG CAGGCTGGGG GCAAGGGGCA GTGTAGGAGG GGCACAGGGG GCCACATTTA GCAGCCTTCC AGGCCTTCCA CCAGCCCAGA   353
CTGCCTCTCT CAGAAGCCAG CAGGGGAGGG TGGGCTTGCT TCATGCCCCC AGATGGCCAA GACTGCCTGT TCCTGAGGTC GCTGTTCCAT GACCCCCCCA   453

             alPheLeu AlaProGlnG lnAlaArgSe rLeuLeuGln ArgValArgA rgAlaAsnTh rPheLeuGlu GluValArgL ysGlyAsnLe
CCGCCTTTAC AGTGTTCCTG GCTCCTCAGC AAGCACGGTC GCTGCTCCAG CGGGTCCGGC GAGCCAACAC CTTCTTGGAG GAGGTGCGCA AGGGCAACCT   553

uGluArgGlu CysValGluG luThrCysSe rTyrGluGlu AlaPheGluA laLeuGluSe rSerThrAla Thr
AGAGCGAGAG TGCGTGGAGG AGACGTGCAG CTACGAGGAG GCCTTCGAGG CTCTGGAGTC CTCCACGGCT ACGGTGAGCC TGGGCTGCTC GGACGGTGCC   653
GGGGCCTCAG ACCGGGCCCA ACTCTAGACA CTTCCACAGA GAAGCAAGCG AGGAACGCCA CAGCCCCTTC GCTGCTCACA GCCTCATTTC AACTCTGAGC   753
CCCTCCTCAC AGGGCTGGCA AGAGGAGCGG CCTCAGCCTT TCCTGGGGGT CTCTGTGCCT GGACTGTGTC CCTGTGCAGC TCCATGACAT GGGGAGGCCT   853
CCACAGTCTT CAGACATCCA CCTGCCTTGG AGCTCTGTGT CCACATGGCC TCCTCAGCGG CAGACTCCCA CACCACCCTT GAGGGGTGGG ACTCTGGGGA   953
GGCCACCACA AGCCCCCGGG CTCAAGACTC AGTGTTCCTG GAGCTCTGTG TCGCCTTTCC TGTCTGTAGG GACTCTGCCA GGGACCCACT GCCCCCTCTC  1053
CTCCCATCTC CCCCAGCCTC TTTCAGACTC GGTGTGTGTG TTGGAGGAAC TCCCCTATCC TCAAATATTC TTCTCCTTTT GGAAACAAAA GTAGGAAACT  1153
CTGCCACAAA CCTCCCCAGA GCCTGCCCCC TGCGTGACCA GGGTAAGGAA AGTGTGAGGA GGAGCATAAC ATTTACTAAA ACAACACAAA ACAGGAGCTG  1253

                                  AspValPh eTrpAlaLys TyrThrA
CCGTAGCCTC ACTCCCAGCC CTTGTTTTTC AGGATGTGTT CTGGGCCAAG TACACAGGTG AGCACCGGGA AGGATTTGCC CCAGGAAGGG AGGCCTGGGG  1353
ACCCCAGTGA GAGAATTCTA CCCAGAGAAT CTTCTGCTGC ACCTAGCCAT CCACCCATCC ACCCCTTCCC CACTCCTTCC TTGGTCCCTC CCATCTGTTC  1453

                                                                                                            l
ATCCATCTTT CTGTTTCTCA CCAACATCCC ATCCACCCTG ACTCCAGCTC ATCCTGGCCA TACCCCAATC CCAAAGGTAA ACACCTGGGT CTTTTCCAGC  1553

aCysGluThr AlaArgThrP roArgAspLy sLeuAlaAla CysLeuGluG
TTGTGAGACA GCGAGGACGC CTCGAGATAA GCTTGCTGCA TGTCTGGAAG GTGAGCAACT GACACGGGTT TGGGGAGCAG GACATGGAGG GGAGCTTGGG  1653
AGAAGAGCTC AGGGGTGGGT TTGGAGTGTG GCTGGTGGAG GCCGAGGCAG TCCCCAGCAT CTGACATTGC TCCCATTCCT GGGGTCAAGA TGTCTCTTTG  1753
TACCTGGCTC TGTGTCTGGC ATGCGAACGA ATGAATGAAT GAATGGACTA ATGAATTAAT GTTTTTTTTT TTGAGACAGA GTCTCGCTCT GTTGCCCAGG  1853
CTGGAGTGCA GTGGCACGAT CTTGGCTCAC TGTAAACTCC GCCTCCCGGA TTCAAGCAAT TCTCTGCCTC AACCTCCCAA GTAGCTGGGA TTACAGGTGC  1953
TCGCCACCAC GCCTAGCTAA TTTTTGTATT TTTAGTAGAG ACGGGGTTTC ACCATGTTGG CCAGGCTGGT CTTGAACTCC TGACCTCGTG ATCCACCCAC  2053
CTCGGCCTCA AAGTGCTGGG ATTATAGAAG TGAGCCACCG CGCCTGGCCA TGAATTCATG TTTAAGGCTT CATTCTCCTT TGCCTGACCC GAGTCTCTGC  2153
CCCCACCTAG TCAGAGCTTT GATGATGTCA CATTCCCCTT CTAGCTTTAG GTGTCACTGA ACCAAACAGG AACCCAAACC CCCAGCTGCT CTGACACCAA  2253
GGACTTCCCT AAGCATGCCA AGGTGTTTCT AGCACCTGGC CTTGCATATG TTGTCAATTT CCTCTGGAGC GACCATCACA TCTACTGAAC ACTTTCCTAT  2353
CCTTCAAGGA CTGCTTCAAA TGTCACCACT TTTGCTGAGA CTTCAGGGAG CACCCTCCCT CCTGCACTGT GTCTGAAGGC ACCTTTAGCA CGACAAAAAT  2453
GGAACTCTTT GTTTATTTAT AAGAGCAGGG TCTCCCTTTT TTGCCAGGCT GATCTTGAAC TCTTGGGCTC AGGCAATTCT CCCATCTCAG TCTCCCAAAG  2553
GAGTGGATTA TAAGTGTGAG CCACCATGCC TGGCTGCCAT ACTTTCATTT TTTTTTTTTT TTTTTGAGG TGGAGTCTCA CTCTGTCGCC TAGGCTGGAG  2653
TGCAGTGGCG CGATCTCGGC TCGCTGCAAC CTCCGCCTGG CGGTTCAAGT GATTCTCCTG CCTTAGCCTC CTGAGTAGCT GGGATTACAG GCACACACTA  2753
CCATGCCCAG CTAATTTTTT GTATTTTTTA GTAGAGACGG GGTTTCACCA TGTTGGCCAG GCTGGTCACA AACTCCTGAC CTCAGGTGAT CCACCAGCCT  2853
CAGCCTCCCA GAGTGCTGGG ATTACAGGTG TAATCCACTG CGCCCAGCCT CATTTGTTAA ATTACGTACT CAACAGACAT TTTACAAAGT TCCTGCTACG  2953
TGCCAGGCAC TATATCAGGT GCTGGGGATT TTAAGAGAAT CAAATACAGT CTCTGCCTTC AAGGAATTCA AAATCTCAAA AGAGAACAAA AATACAAAAT  3053
ATTAAAATGA TTGCGGCCGG GTGTGGTGGC TCAAGCCTGT AATCCTAGCA CTTTGGGAGC TGAGGTGGGC GCCCAGGCCA GGGGTTTGAG ACCATCTTGG  3153
CCAACATAGT GAAACCCCCA ACCTCTACTA AAAATACAAA AATTAGCTGG GGTGTGGTGG CACGCGCCTG TAATCCTAGC TACTAGGGAG GCTGATGGGG  3253
AGAATTTCTT GAATCTGGGA AGCGAAGGTT GCAGTGAGCT GAGATCATGC CACTGCACCT TCAGTCTCGG CATCAGATCA AGACTCATCT CAAATACATA  3353
ATAAATAATA ATTCAATAAG TGATTGCAAG AAAGTTCTGT TCAAGGCACC AAGAGACCAC AGGAAAATGA GTGTCTGGTT TGCCAGAAAA TGAGAGATGG  3453
CTTCCCAGGA GAGGCAGAGT TCTGCCTGGC CTAGTGGGAT GCATGGATGA ACAAACAAGT GGGCATTCCA GTCAGAAGAA ACAATCCGTG GAAAGACCCA  3553
GAGGCATGAG AAGCTGAGCT AGCAGGGACA GGTAGACCAG GGCCAGTTGA AAAGGACCTT CATCACTTTT TCATCCTGCT GGCCAAGAGA AGCCACAGAA  3653
TGGAAGCTCC ATGAGGGCAG GGCTGTGACT GTCCTATTGG TTGATGTGTA CTGAGCACCC GACAGTGCCT GTCATATGGT AGGCACTTAG CGAATATTTG  3753
GAGGCCACTG TTGAGTGAAT GGGAGAACTG CTGGTTGCAG AGGAAGAGGG GCTGGGTGAA TGCAGGTTCA GGATTGTGGA CCTGCATGAG CTGGGAGGTG  3853

                                                                                    lyAs nCysAlaGlu GlyLeuGlyT
GGGGATAGAC AACTTTGCAG GGAGAGAGGA AATAAGTCCC CAGGCTCCAA GGCTGACCGG GGTGGGGTCT CCGCAGGTAA CTGTGCTGAG GGTCTGGGTA  3953

hrAsnTyrAr gGlyHisVal AsnIleThrA rgSerGlyIl eGluCysGln LeuTrpArgS erArgTyrPr oHisLysPro Gl
CGAACTACCG AGGGCATGTG AACATCACCC GGTCAGGCAT TGAGTGCCAG CTATGGAGGA GTCGCTACCC ACATAAGCCT GAGTGAGTGA GGGGTCGGCC  4053

                                                                                        uI leAsnSerTh rThrHisPro
TTCCCACCAT GGGCTGAGAA CAGGGAGCAA GCGTACCTCA AGTTCAACAG CCTCCTGTTG GGCAATTTCC TCTTCCAGAA TCAACTCCAC TACCCATCCT  4153

GlyAlaAspL euGlnGluAs nPheCysArg AsnProAspS erSerThrTh rGlyProTrp CysTyrThrT hrAspProTh rValArgArg GlnGluCysS
GGGGCCGACC TACAGGAGAA TTTCTGCCGC AACCCCGACA GCAGCACCAC GGGACCCTGG TGCTACACTA CAGACCCCAC CGTGAGGAGG CAGGAATGCA  4253

erIleProVa lCysG
GCATCCCTGT CTGTGGTAGG CTGGGGGCAG TGGGGCGACC CATGACCAAG CCCGGGGGCT TCATGGGGCC TGGCAGCCTG GGATGGGAAC CAAGAATACT  4353
GGCTACCCAG GCACAGTGGC TCATGCCCGT AATCCCAGCA CTTTGGGAGG CTGAGGCAGG CAGATCACCT GAGGTCAGGG GTTTGAGACC AGCTGGGCCA  4453
ACATGGCAAA ACCCCGTCTC TACTAAAAAT ACAAAAATTG CCAGGCGTGG TGGTGGGCGC CTGTAATCCC AACTACTCTG GAGGCTGAGG CACGAGAATC  4553
GCTTGAACCC GGGAGGCGGA GTTTGCAGTG AGCTGAGATC CTGCCACTGT ACTTCAGCCT AGGCGACAAG AGCAAAACTC TGTCTCAAAG AAAAAAAAAA  4653
GATGCTGGCC ACCTTCAGAG CTGGCGTCAG TCATTCAGAT CATATCTGTG CCTATTGCTC AGTAAAGTCA GGGAATCAGG GGATCTGAGT GGGGGGATCT  4753
GCCAGCCTCC TCCTCCCCCT CCCCACTCTT GACTTCCTTA TGGTCTAGGC TGTGGCTCAT TCCAAACATG CCTCCTTTCT GATCAAGGCA CTCCTCCCTC  4853
CGGGAAGCCC TCCCTAGCCA TTTCAGTCCA CACACCCTGT TCTGAGTATC ACAGAGCAAG CCTTGTGCAG TTTGGCCCGC GGGATTCTGT CATTATTATT  4953
TCCTTGGTGT GTTAAGTAGC TATAGCCACC CCTTCCCTGA GGCAGACCAC AATAAGCATT TCTTTTTCCC ATGAGGGTTG GCAGGTGTGG CTGCACTCGC  5053
TAATGCGTCT GTAGGGTCAA CTGACGGAGG TTGGCCCTGG CTGGGTGGCT CTGATTCAAA TAATGGGTCC AGCTGAGTCT GGCTCCTCGT TGAGGGTTGG  5153
GCCTAGATCT GCTCCACGTG CGTTCATGCT GGGGCTGAGG CTGAAAGAAG GTACCTGGGA AAACTCTTCT TATGCTGATG ACAGACACAG AAAACAATGA  5253
ACAGAAAAGC GTCTTCTGTC CTGAAGGCCT GGCTCAGAAC AGGCACAGTC AGCCCTGCCC ACGTTCCATT GGCCAGAGCA AGTATATGTT CAAGGCCAGG  5353
GTCAAGAGGT AAACTACACC TCAGCCTGTA AAATCACAGA GCAAGGGATG TGGATGCAGG CAGGGGTAAA GAATTTGTGC CGATTACCAG TCCACAAACA  5453
TGCGTTAGTG TTTGTTCTCT AGGCAACCCT GTCGGGCCCA TTGCTCATTC CTGGGGTTGG TCTTTTTTTT TTTTCTTTCT AAGAAGGAGT CTCACTCCCT  5553
```

```
TGCCCAGGCT GTTGGAGTGC AGTGGCCCTA TCTCAGCTCA CTGCAACCTC CGCCTCCTGG GTTCAAGCGA TTCCCCTGCT TCAGCCTCCT GAGTAGCTAG  5653
GATTACAGGC GTGTGCCACC ACTCCTGGCT AATTTTTTTT TATGTTAGTA GAGACGGGGT TTCACCATGT TGGCCAGGCT GATCTCAAAC TCCTGACCTT  5753
GTGATCCTCC CGCCTCGGCC TCCCAAACTG CTGAGATTAC AGGGGTGAGG CACTGCGCCC AGCCATTTTT TTTTTTTTTT TTTTTGAGA TGGAGTCTCA  5853
CTCTCACCCA GGCTGGAGTG CAGTGGCATA ATCTTGGCTC ACTGCAACCT CCACCTCCTG GGTTCAGGCG ATTCTCTGCC TCAGCCTCTC ATATAGCTGG  5953
GATTACAGGC ACACGCCACC ACGCCTTGCT AATTTTGTAT TTTTAGTAGA GACGGGGTTT CTTCATGTTG GCCTTGCCTG ACTTGAACTC CTTGTTCCGG  6053
TGATCTGCCC AGCTCGGCCT CCCAAAGTTC TGGGATTACA GGTGTAAGCC ACTGCGCCTG GCCCCTGGTA TTGGTCTTAT AGCAAGTTTA TCCCAACAAA  6153
AACAGCTACT ATTTACTCCC CAACCCCCAT ACACACGCAC ACACATTGAT GATAAATAAG TTGCAGGCTT GCAGAAATTG GCCCATCCAG GTGAACAGCC  6253
TAGTGATCCG AGCAAGCGTC CTGCTGTGCA GCTATAAAAA CATGACTCCT CCAGCAGCTC CAGGCAGCCA CTACCAGTTG GTTACAGATG GCCTAGGAGG  6353
CCAAACCTGG TTACTATCTC TGGTTTATTA TGTGCCAGAC ACTTATGCTG TATATTTTGT TTAATCCTCT CAACAAACCT GCAAAAGTGG CATTAGTAAC  6453
CCCTTTAAAG GCAAACGGTC AGAAGCCCAG AGAGGTTAAG TAACCTGAGG TCACACAGGC AGAAAGCAGC AAGACCGGGG TTCACACCCC TGTCTGTTCC  6553

                                                    lyGlnAs pGlnValThr ValAlaMetT hrProArgSe rGluGlySer
GGTCCATGTG TGGTCTCACT CACTCTGCTG CCTCCTTGCC CCTCACCCAC CAGGCCAGGA TCAAGTCACT GTAGCGATGA CTCCACGCTC CGAAGGCTCC  6653

SerValAsnL euSerProPr oLeuGluGln CysValProA spArgGlyGl nGlnTyrGln GlyArgLeuA laValThrTh rHisGlyLeu ProCysLeuA
AGTGTGAATC TGTCACCTCC ATTGGAGCAG TGTGTCCCTG ATCGGGGGCA GCAGTACCAG GGGCGCCTGG CGGTGACCAC ACATGGGCTC CCCTGCCTGG  6753

laTrpAlaSe rAlaGlnAla LysAlaLeuS erLysHisGl nAspPheAsn SerAlaValG lnLeuValGl uAsnPheCys ArgAsnProA spGlyAspGl
CCTGGGCCAG CGCACAGGCC AAGGCCCTGA GCAAGCACCA GGACTTCAAC TCAGCTGTGC AGCTGGTGGA GAACTTCTGC CGCAACCCAG ACGGGGATGA  6853

uGluGlyVal TrpCysTyrV alAlaGlyLy sProGlyAsp PheGlyTyrC ysAspLeuAs nTyrCysG
GGAGGGCGTG TGGTGCTATG TGGCCGGGAA GCCTGGCGAC TTTGGGTACT GCGACCTCAA CTATTGTGGT GAGCTGCCTG GGTAGGGGGC CTGAGTTGCA  6953
GGGACAAATC CTAGTGGGAA TAACAACAGC CGCTTCTGCT TATCGAACGC TTACCTCATT GAGTGCGCTC ATTACAGCCT TACAGTAACC AGGTGGGGGG  7053
TAAGGTCCTG TGCCCATTTC ACAGATAAGT ACACTGAGGC CCCAGGAGGT TATTGCCTAG TAGCCCAACT GTGCATGCAC GCTTAACCTC TGCACCAAAT  7153

                                                                                                    luGluAla
GGCCTCCAAG GCCCGTAGGG GAACTGGGGG GATCTAGGGG ATGGGTGAGG AATGGCCCAG CCCAGTCCCG GCCGGTGCCT GGGTCCCAAC AGAGGAGGCC  7253

ValGluGluG luThrGlyAs pGlyLeuAsp GluAspSerA spArgAlaIl eGluGlyArg ThrAlaThrS erGluTyrGl nThrPhePhe AsnProArgT
GTGGAGGAGG AGACAGGAGA TGGGCTGGAT GAGGACTCAG ACAGGGCCAT CGAAGGGCGT ACCGCCACCA GTGAGTACCA GACTTTCTTC AATCCGAGGA  7353

hrPheGlySe rGlyGluAla A
CCTTTGGCTC GGGAGAGGCA GGTGAGGTAG TGGGCATCCG AGGGGATGCG GGGCTGCGGG GCTGGTGGCC AGGACTTGCC CCTCACTGCT TGGCTTGCTC  7453

     spCys GlyLeuArgP roLeuPheGl uLysLysSer LeuGluAspL ysThrGluAr gGluLeuLeu GluSerTyrI leAspGlyAr gIleValGlu
TGCAGACTGT GGGCTGCGAC CTCTGTTCGA GAAGAAGTCG CTGGAGGACA AAACCGAAAG AGAGCTCCTG GAATCCTACA TCGACGGGCG CATTGTGGAG  7553

GlySerAspA laGluIleGl yMetSerPro Tr
GGCTCGGATG CAGAGATCGG CATGTCACCT TGGTGTGTCC TGGAGCCCTG CGCTACCATT CACTCCTGGG GGCAGGTGTG CTGCTGGACC CCCACCCTCA  7653
GGCCCTGCCT GCAGGCCTGG GCTTTACAGA TGACAACAGC TGAGCATCCA GGATCCCACC AACTCCACAC AGCAGCCACA TGAGATGGGT TGTTTACTTC  7753
TTTTTTTTTT GTTTCTTAGA TGGAGTCTTG CTCTGTCACC TAGGCTGGAG TGCAGTGCTG CAATCTCGGC TCACTACCTC GATCTCAGCT CACTGCAACT  7853
TCTGCCTTCC GGGTTCAAAC GATTCTCTTG CCTCAGCCTC CTGAGTAGCT GAATTTACAG ACATGCGCCA CCACACCCGG CTAATTTTTG TATTTTAAGT  7953
AGAGACAGGG TTTCACCATG TTGGCCAGGC TGGTCTTGAA CTCCTGACCT CAAGTGATCC ACCTGCCTCA GCCTCCCAAA GTGCCGGGAT TACAGGCATG  8053
AGCCACCACA CCCGGCCCAT GGGTCCTTTA CTTCTAAGCA GATGGTAAAG CTGAGACTGA CGGAGCTGGT GGCTCACCTC CGCGCACAGC TAATGGGTTT  8153
GAATCCAGTT CTTCTGATTC CAGAGCTGTG CTACGCTATG TGAACTCTGG ACTGGAAGGA CCTAGTTAGG GGGTGCAAAA AGCAGGAGGC AGGTCAGGTG  8253
CAGTGGCTCA CCCCTGTAAT CCCAGCACTT TGGGAGGCCA AGACAGGAAG ATCACTTGAG GGCAGGAGTT CGAGGCCAGC TTGGGCAAAA TGGTAAAACC  8353
CCGTCTCTAC TAAAAATGCA AAAATTAGCC AGGTGTAGCA GCATGTCCCT GTAGTCCCAG CTACTAAGGA GGCTGAGGCG GGAGGATCGC CTGAGCCCAA  8453
GAGGCTGAGG CTTCAGTAAG CTGTGACTGT ACCATTGCAC TCCAGCCTGG GTGACAAGAG TGAGACCCTG TCTCAAAAAT AAATAAATAA ATAAATAAAA  8553
AGTGTGAGGC AGCCCCTCAG CATCACACGG AGGCTCCAGC CCCAAAGGCG GCCAGCCCAA GCTTGGATCT GGGCCCCGGA GGCAGCTCTG CCCAGCTGGG  8653

                                                                                                    p GlnValMetL
TTCTTAGACC TGGGATTGTT ACTTCTAGGG CTGGTGTAGA GGCAGCCCCC TCATCCTCAG CTCCTAATGC TTCCTGCTGC CCCTCCCAGG CAGGTGATGC  8753

euPheArgLy sSerProGln GluLeuLeuC ysGlyAlaSe rLeuIleSer AspArgTrpV alLeuThrAl aAlaHisCys LeuLeuTyrP roProTrpAs
TTTTCCGGAA GAGTCCCCAG GAGCTGCTGT GTGGGGCCAG CCTCATCAGT GACCGCTGGG TCCTCACCGC CGCCCACTGC CTCCTGTACC CGCCCTGGGA  8853

pLysAsnPhe ThrGluAsnA spLeuLeuVa lArgIleGly LysHisSerA rgThrAr
CAAGAACTTC ACCGAGAATG ACCTTCTGGT GCGCATTGGC AAGCACTCCC GCACAAGGTA CAGAACTGGT GGCCCGTGGG TGTCTGGCAG GGGTCTGAGT  8953
CCTCCAAAGC GATCATGAGG GGCCTTGGTG GCTCCGGGAC ACATAGGATG TTCTGTATAC CCCCCAGAAT ATAACATCCC AGCAGTCTCT GCTGGAAAGC  9053
CATTTGGTCA CGTCCTGACT GAGGCTTGGA GCGCGGGGAG AATCCGTCTG TCTCTGGTCC CTCCAACACT AGGATATAGC CCATGTGGGA GTCTCTGAAA  9153
ATAGAGTCTG TCTGGACTAG GGCGTGCAGC CTGTGCCCCT GTCCCCGTCC TCCAGGCTGT CTGACTCCAA AGCCCTGCAC GGCTTTAGGC CCAGGAAGAA  9253
ACACCCAGGG GGCTGCCATG GCAGGAACCA GCCCTATCCC CTCCCTGGTG GCCTGCAGGA CACACTGTCT CCCAGAACCC CAAGGGCAGG CAGTTTCCTG  9353

                                                    gTyrGl uArgAsnIle GluLysIleS erMetLeuGl uLysIleTyr
CTCCTTGCTG GGTGAACCTG CAGCTTCTCC ATTTCTTTCT TGGGGTCTCT GCAGGTACGA GCGAAACATT GAAAAGATAT CCATGTTGGA AAAGATCTAC  9453

IleHisProA rgTyrAsnTr pArgGluAsn LeuAspArgA spIleAlaLe uMetLysLeu LysLysProV alAlaPheSe rAspTyrIle HisProValC
ATCCACCCCA GGTACAACTG GCGGGAGAAC CTGGACCGGG ACATTGCCCT GATGAAGCTG AAGAAGCCTG TTGCCTTCAG TGACTACATT CACCCTGTGT  9553

ysLeuProAs pArgGluThr AlaAlaSe
GTCTGCCCGA CAGGGAGACG GCAGCCAGGT GGGCCACCAG ATGCTTGTTA GCTGAGGGGC AGAAGCCAAG TTCTGGGCCT GGCTCTGATA CCAAGTAGCC  9653
TTGCAAGAGC CCCTTTCCCT TTTCCAGGCC TCGGTTTCTT GGAGTGAACC CAAAAGTTCT TTTCAGTACT GGCGTTTTAT TTTTTATTTA TATTTATTTA  9753
TTTACTGACG GAGTTCCACT CTTGTCTCCC AGGCTGGAGT GTAGTTGTGC GATCTTGGCT CACTGCAACC CCACCTCCTG GGTTCAAGCG ACTCTCCTGC  9853
CTCAGTCTCC TGAGTAGCTG GGATTACAGG CTAATTTTTG TATTTTTAGT AGAGACTGGT GGGTTTCACC GTGTCGGCCA GGTTGGTCTC GAACCCCTGA  9953
CCTCAAGTGA TTCACCCGCC TCGGCCTCCC AAAGTGCCGA GACCACAGGC GTGAACGTCT GTGCCCAGCC AGCTCTGGCG TTTTAGATTC TGGTCTCTAA 10053

                                                                                    rLeuLeuGln AlaGlyTyrL ysGlyArgVa
GAAATGGCGT TGGGGCCAGG CGGCTCCTGT GGGGGTTGGC TCTCACTAGG CCCTTCTTCC TTCCCCAAAG CTTGCTCCAG GCTGGATACA AGGGGCGGGT 10153

lThrGlyTrp GlyAsnLeuL ysGluThrTr pThrAlaAsn ValGlyLysG lyGlnProSe rValLeuGln ValValAsnL euProIleVa lGluArgPro
GACAGGCTGG GGCAACCTGA AGGAGACGTG GACAGCCAAC GTTGGTAAGG GGCAGCCCAG TGTCCTGCAG GTGGTGAACC TGCCCATTGT GGAGCGGCCG 10253

ValCysLysA spSerThrAr gIleArgIle ThrAspAsnM etPheCysAl aG
GTCTGCAAGG ACTCCACCCG GATCCGCATC ACTGACAACA TGTTCTGTGC TGGCAAGTCT GTGCAGGGCG GGCTGAGGGA ACAGTGGGGC CCAAGCTGGG 10353
AGAACTGAGT TGTGCCTGGG TTCAAGCCAT GTGACTTTGA GCAAGTTGCC TAACCTCTTG GTGGCTCAGT TTCTTCCTCT GTAAAATGGA GGTAAAAGTC 10453
```

TCTATCCCAT AAGGTTATGG GAGGGTTAAA TGAAGTAGTA TATATTAATG TACTTGGCAT AGTATCAGTC ACCAGTGAGC TCAGATAGCA GCAAGAGGCT 10553
GCGGGTAGGG AAATGCCATT CATTCAGTCA CTCAGCAAAT ATTTATTGAG CGCCTATCAC GTTCCAGGCA GCGTTCTAGG GTATACAGCA GGGACCCAGA 10653
CGGACAATGT CTGTGCCCTC AGAGAGCTTC CTTCCTAGGA GGGCACATCC ATAAACAGAT CTAAAACAGC AATCCCTGAC CAGTGCTGTG AAGAAAAATG 10753
AAGCACAGGG AGAGAGAACG GCTGATGAAG TGGGCTTCTA AATAGGGTGG CCAGACAAGG TGGGCAGATC ACTTGAGGTC AGGAGTTCAA GACCAGCCTG 10853
GCCAACATGG TGAAACCCCG TCTCTACTAA AAATACAAAA ATTAGCTGGT CATGGTGACG CATGCCTGTA GTCGCAGCTA CTCAGGAGGC TGAGGCAGGA 10953
GAATTGCTTG AGCCAGGGAG GCGGAGGTTG CAGTGAGCTG AGATCGGGCA TCATTGCACT CCAGCTGGGC AACACAGCAA GACTCCATTG ATCGATCGAT 11053
CAATCAATCA ATCAGGTGGC CAGAGAAGGT TGGAGAAGGC CTCCCTGAGA AGGTGATGTC TGGGCAGGGA CTGGAAGAGG GGAAGGAAGG AGTGAGCAGG 11153
CATATCTAGG GGAGGAGCAC CGCAGGCTGG GGGCATGGCA GGCACTAAGG CCCTGAGGTG GGAGCACTCT TGGCTTGTCT GGGGAGCAGT AGGGAGGCCT 11253
GGGGGGCTGA GGAGGGGCAG CAGTGGGTGA GGGGAGAGAG GGGGGCAGGC AGAGGACAGC CACTTCCTTT AGGGCCTGGA AGGACTTTAT TGAGTGAGAT 11353
GGGAAGTTAT TGAGGGGCTT GAGGCAGGTT AAGAAATGAT GTGACTGACT TTAAAAGTAA AAAATAAAAA AATTTAGTGT AATTTCAGAC TCACAGAAAA 11453
GTTGTAAAAA TAATACAAAG ATTTCCTGTA TACTGTCATC CAGATTGTCC TCCATTCTGT GGATGTGTGG GAATTTTTAT ATATATATAT GCATAGTTTG 11553
AGAGCAAATC ATGAATATGG TTTCTTTTTA CCCATAAATA CTTGAGTATT TCCAAAAAAA AAAAAAATAC CCAAGGATGT TCTCTTATGC AACCACAATA 11653
CAAATATTAA AACCCGGAAA TTTTTTTTTG ACATAGCTTC GCGTCACCCA GGCTTGAGTG CAGTGGCACA ATCTCGGCTC ACTGCAACCT CCTGCTCCCA 11753
GGTTCAAGTG ATTCTCCTGC CTCAGCCTCC TGAGTAGCTG GAATCACAGG CATGTACTAC CATGCCTAGC TAATTTTTGT ATTTGTAGTA GAGACAGGGT 11853
TTCACCATGT TGGCCAGGTC GGTCTTGAAC TCCGACCTCA GGTGATTCAC CTGCCTCGGC CTCCCAAAGT GCTGGGATTA CAGGCGTGAA CCACTGTACT 11953
CGGCCAAAAC CAGGAAATTT TTTTTTTTTT TTGAGATGGA ATCTTGCTCT GTTGCCCAGG CTAGAGTGCA GTGGCATGGT CTCGGCTTAC TTGGAATTAC 12053
AGGTGCCTGC CACCACGCCC GGCTAACTTT TTGTATTTTT AGTATTTTTA GTAGTGATGG GGTTTCACCA TGTTGGCCAG GCTGGTCTTG AACTCCTGAC 12153
CTCGGGTAAT CCACCCACCT CGGCTTCCCA AAGTGCTGGG ATTACAGGCG TGAGCACCAG CACCTGGCCC AAAACCAGGA AATTAATGAT GATACAATAT 12253
TATTGTCTAA TCTATAGACC TTATTCAAAT TTTTGTTAGT CTTGCTAATG TCTTTTATAG GGAAAAAAAA AAAAAAAAGC GTGTTTCTCA CCCAGGATTC 12353
AATGAAGGAT CTTTCTTTGT CTTCTATGAC CTTGACATGT CTGATGAGTG CAGTCTGGTT ATTTTGTACA CTGGCCCTGA ATCCGGGTTT GTCTAAGGTT 12453
TCCTCACGGT CAGGTTCGGG CTCAGTGGTG CCATGTCCTT CTTGGTGCAT CCTGTTAACT GGCACATGAG AACAATTTGT CTCATATGTG GTGAGTCTAA 12553
CTCTGACCTC TTGAGGAAGG CAATGTCTGC CAAGTTTCTT GCTGTAACTT CTGTTTTTCC CTTTGTAATT AATAAGAATC TGGTAAAGAG ACACTTTGAT 12653
GTTTTTTTTT TTTTTTTTTT TTGTGATGGA GTCTCCCTCT ATCACCCGGG CTGGAGTGTG TGGTGCGATC TCGGCTCACT GCAACCTCCA TCCCCCAGGT 12753
TCAAGTGATT CTCCTGCCTC AGCCTCCCAA GTAGCAGGGA TTACAGGCAT GTGCCACCAC ACCCAGCTAA TTTTTGTATT TTTAGTAGAG ATGGGGTTTC 12853
ACCATGTTGG CCAGGATGGT CTCGAACTCC TGACCTTGTG ATCCGTCTGC CTCAGCCTCC CAAAGTGCTG GGATTACAGG TGTGAGCCAA TACGCCTGGC 12953
CTACTTTGAT ATTTTGTATT CTGTTTGCAT CAAAACCTTC TCCCAACTAG GGTGACTACC AAATGGCACT TATCTAATTC TGTCATTCCT TCTACATTTG 13053
TTAGTTACTT TATTGCTTTC CTTCCTTTCA TTCTATCAGT GTGGACTTAA GGATCCTTAC TTTATTCTAA GGGTTCACCT TTTTTTTCTT TTTTTTGAG 13153
ATGGAGTTTC GCCCATGTTG CCCAGGCTGG ATGGAGTGCA ATGGCGTGAT CTCGGCTCAC TGCAACGTCC TCCTCCCAGG TTCAAGCAAT TCTCCTGCCT 13253
CAGCCTCCTG AGTAGCTGGG ATTACAGGCA TGTGCCACCA CGCCTGGCTA ATTTTTTGTA TTTTTAGTAG AGACAGGGTT TCACCATGTT GGCCAGGCTG 13353
GTCTCGAACT CCTGACCTCA GGTGATCCGC CCGCCTCAGC CTTCCAAGGT TCTGGGATTA TAAGCGTGAG CTCTACCGTG CCAGGCCATA CTTTGTTACT 13453
ACTGTTATTT TTTCTGATGC TCAGATGATC CCAAGTTTGG CCTGTGGAAG TCCCTTCAAG CTGGCTTCTG TGACTTGGGG AGATGTTCTG TCATTCTTTG 13553
AGTACTTTCT TTCTTTCTGG CACAGCAAAA TGATTCAGGT TAATCCTACT TTCCTTACTG TAGTGTTGGA ACCAGCCATT TCTCCAGGGA ACCCTTGTAG 13653
TCAAGAGTGG AATTTAGAAC TGAGATCTGG GTGCTGGCGT GTGCACATTG CTAGTGGGAT GTCATTACTT CTAGGCTCTC TTAGTGGACA GAACCAGAAA 13753
AAAATTATAT GATGCATATA CCAATATCTC TATCATCTAT ATAAAAAACC ATGAGTTCCT ACTGAAACCT CCAATTCCAT TCTAACACCA CAGGATTAAT 13853
TTTAGCTTTT CCTTTTCCAT ATTTGTAACT CTCTCTGTTG ACAGTGAGAA ACCTGACCCT CATTATCTGT AATGCATTTG CCTATTTGAA CAATACTAGA 13953
ATATAGTTTC AAAATCCTCC ATCCATAACA CTATTAAAAC CAATCCTATG GCTGGGCTCA GCCCACTGCA ACCTCTGCCT CCTGGACTCA AGCCAGCCTC 14053
CCACTTTAGC CTCCCGAGTA GCCAGGGCTA CAGGCACACA CCACCATGCC CAGCTAATTT TTGTATTTTT TGTAGAGACT GGGTCTCACT GTGTTGCCCA 14153
GACAGGTCTT GAACTCTGAG CTCAAGTGAT CCATCCAACT CAGCCTCCCA AAGTGCTAGG ATTACAGGTG TGAGTCACCA TGCCTGGCCT CTCCTAGTAA 14253
ATTTTTAGAA GTGGTGTTGT TAGGTCAAAA GGCAAACATG TATGTCATTT TTTAGAGATT TTTAAATTTC TTTCCATAAG GGTTGTACCA GTTTGCATTT 14353
CCATCACAGT GTATGAGAAT GCCTGTTTCC CCACAACCTT GCCAAAAGAA TGTCACAGTT TAAATTTTAC CAATCTGAGA GGTGAGAAAT AGTACCTGAA 14453
ATTGTTTAAC GGACATCTTC AAATTGAAAT TGAGGTTGAC AACGAATCAT AGTTAGGACC TTTTTTTTTT TTTTTTTGA GTGGGTCTCC TCGTCACCAA 14553
GCTGAGTGCA TGGCACGATT TGCTCACTGC AACTTCCGCC TTCTGGGTTC AAGCGATTCT CCTGCTTCAG CCTCCCAAGC AGCTGGGACT CCAGGCGCGA 14653
GTCACCATGC CCGCTAATTT TTGTATTTTT AGTAGAGACA GGGTTTTACC AGATTGGCCA GGCTGGTCTC GAACTCCTTA CCTTGTGATC CTCCCGCCTC 14753
GGCCTCCCAA AGTGCTGAGA TTACAGGCAT GAGCCACCAC GCCTGGCCTA AGGACCATTT TTATATAATT TTTTTTTTGA GACAGAGTCT TGCTTTGTCA 14853
CCCAGGCTGG AGTGCAATGG TGCAATCTTG GCTCACTGCA GCCTCCACTT CCCTGGTTCA AGTGATTCTC CTGCCTCAGC CTCCCGAGTA GCTGGTTCCA 14953
CAGGTGCGTG CCTGGCTAGT ATTTGTATTA TATAATTTTT TTGTGAATTG TCTCTTCATG GTTTTTGCC CATTTTTGG TCCCTTTCTT ATCAATTTTT 15053
GTGAGTTCTT CGTATTTATA TTAGGCCTTT ATTTGTGATA TACATTGCAA ATGTTTTCTC CTAGTTTGTC AGTTTTTTTA ACCTCATGTA TAATTTTTCT 15153
GGCCATGCAG TTTAAAAAAT TACTAGGTAG TCAAATTTAT CAATCATTAT TTATAAATCT GGTTTGAACA GAGATAAACT TTCCTGGCCA AGTGTGGTGT 15253
TTACACCTGT AATCCCAGCA CTCTGAGAGG CTGAGGTGGG GATCACCTGA GGTCAGAAGT TCAAGACCAG CCTGGCCAAC ATGGTGAAAC CCTGTCTCTA 15353
CTAAAAATAC AAAAATTAGC TGGGCGTGGT GGCTGATGCC TGTAGTCCCA GCTACTCAGG AGACTGAGGC TGGAGAATTG CTTGAACCTG GGAGGCGGAG 15453
GTTGCAGTGA GCAGAGATCG TGCCGCTGCA CTCCAGCCTG GGTGACAGAG CAAGACTCTG TCTCAAAAAC AAAACGACAA AAAACAACAA CAGAAAAGCC 15553
TTTCCTGATA GCTAGGTCAT TGAGGAATTC ACTCATGTTT TCTTCTAGTA CCTGATTTCA TTTTTCTGCA CTTAGATTCC TGACTCATAT GGAGTTTATT 15653
TTTGTATCTG ATGTGAGGCA TAGATCTAAT TTATTATTTT CCAAATGGCT AACTAGCTGT CTCTAAACCC TTTATTAAAA ATTATTGGCC AAGTGCGGTA 15753
GCCACACCTG TAATCCCAGC AGTTTGGAAG GCTGAGGCAG GATTGCTTGA GGCCAGGAAT TCAAAACCAG CCCAGACAAC ATAGCAAGAC CCTGTCTCTA 15853
CAAGAAAATA TTGGTCAGGT GTGGTGGCTC ACGCCTATAA TCCCAGCACT TTGGGAGGCT GAGGCAGGTG GATCATGAGG TCAGGAGATA GAGACCATCC 15953
TGGCCAACAT GGTGAAACCC TCGTCTCTAC TAAAATACAA AAAATTAGCT GGGTGTGGTG GCGCATGCCT GTAGTACCAG CTACTCAGGA GGCTGAGGCA 16053
GGGGAATCAT TTGAACCCAG GAGGTGGAGG TTGCAGTGAG CTGAGATCAC GCCATTGAAC TCCAGCCTGG CGACAGAGCA AGACTCCATC TCAAAAAAAA 16153
AAGGAAAAAG AAAATATTTT AAAAATTAGC TGGGCATGGT GGCATGTGCC TTGTAGTCTC AGCTACTTGA GAGGCTGAGT TAGGAGGATT GCTTGAGCCT 16253
AGGAGTTCAA TACTGCAGTG AGCTATGACC GCACCATTGC ACTCCAGCCT GGGCAACAGA GTGAGACCCT GTTTCTATTA AAAAAAAAAA ATCGGCTGGG 16353
CGCGGTGGCT CACGCCTGTA ATCCTAGCAC TTTGGGAGGC CGAGGCGAGC GGATCACCTG AGGTCAGGAG TTCAAGACCA GCCTGACCAA CATGGAAAAA 16453
CCCTGTCTCT GCTAAAAATA CAAAATTAGC CAGACATGGA GGCACATGTC TGTAATCCCA GCTACTCGGG AGGCTGAGGC AGGAGAATCG CTTGAACCTG 16553
GGAGACGGAG GTTGCAGTGA GCTGAGATCC CTCCATTGCA CTCCAGCCTG GGCAACAAGA GTAAAAACTC CGTTTCGCCA GGTGCGGTGA CTCACACCTG 16653
TAATCCCAGC ACTTTGGGAG GCCGAGGTGG GTGAATCACA AGGTCAGGAG TTTGAGACAA GCCTGGCCGA CATGGTGAAA CCCCATCTCT ACTAAAATAC 16753
AAAAAATTAG CCTGGCATGG TGGTGTGCGC CTGTAATCCC AGCTACTTGG GAGGCTGAGG CAGGGGAATC ACTTGAACCT GGGAGGAGGA GGTTGCAGTG 16853
AGCCGAGATG GTGCCACTGC ACTCCAGCCT GGCAACAGAG CGAGACTCTA TCTCAAAATC AATCAATCAA TCAATCAATC TTTGAACTAG TGATTTGAGA 16953
TTTCACCTTT ATCACATTCT AGATTGTATC TTATTTTCAT TTATTTATTT GAAATATAGA CAAGTCTCCC TGTGCTGCCC AGGCTGATTT CAAACTCCTG 17053
GCTGGGCTCG AGCAAGTCTC CCGCCTTGGC CTCCCAAACT GCTGGGATTA CAGACGTGAG CCACCATACC TGACCCAGGT TTTATTTTTT AGTTTTATTT 17153
TTTCCTGCAT CCAGCTAATT TGATTTGATT TGTAGAGACG GGGTCTTGCT ATGTTACCTA GGCTGGTCTC GAACTCTTGG GCTCAAGTGA TCCTCCTACC 17253
TTGGCCTCCC AAAGTGTTGG GATTACCAGC ATGAGCCACG GTGCCCAGCC CCACGTTCTA GATTTCTATG GATAGAGTAT GCTTAAGGAT GAGTATGTTT 17353
CTGGATGTTC GACTCGGCTT TCCTGGTCTG TTGTCTGTCT GTGTACAGCG TCACATTGTT TTAATGATAG AGGCTTTAGC GTACATAGCT GGGAAGGCTA 17453
ATGTTCTCTT TTAGTTTTTC TTTCCAGTGG TTTCCTGGCA ATTCTTGCAT GTTTGTTTTT CCATATGAAC TTTAGTGTCA ACATGCCTAG GTCTATAAAA 17553
AAGCTTGGTG GTAATTTTAT TGGGATTATG ACACTTCAAC AAATTAACTG GGAGAATGAA CATATTTTTG ATGTTGAGTC ATTTTATCCA AGGATAAGAA 17653
ACGTTTTCCT ATTTGCTCAA GTCTATTATT GTATCTTTCT TGACTGCTGC AATGTATTCT CTTATAATTT TTTCTATTGG TATCTTATTT TATGTATTTG 17753
TAATATCTTA TTTTTCTTGA GTAAATTAGT TAATGGCTTG CCGGTTTTCT CAAAACAAAT ATCTAGGGAT TTGATTTATG AAATTATTAG GCCTATTATT 17853
TTTCTTTTTT TTGAGATGGA GTCTCACTCT GTCGCCCAGG CTGGAGTGCA GTGGCGTGAT CTCAGCTCAC TGCAACCTCC ACCTCCTGGG TTCAAGTGAT 17953
TCTCCTGCCT CAGCCTCCCC AGTAGCTGGG GTTACAGGTG CACGCCACCA TGCCCGGCTA ATTTTTTTAT ATTTTTAGTA GAGACGGGGT TTCACCATGT 18053
TAGCCAGGCT GGTCTCGAAC TCCTGACCTC ATGATCCGAC TGCCTCAGCC TCCCAAAGTG CTGGGATTAC AGGTGTGAGC CACCGTGCCT GGCCTTTTTT 18153
TTTTTTTTTT GAGACAGAGT CTTGCTCTGT CACCCAGGCT GGAGTGCAGT GGTGCGATCT CGGCTCACTG AAAGCTCCAC CTCCCGGGTT CACGCCATCC 18253
TCCTGCCTCA GCCTCCCGAG TAGCTGGGAC TACAGGTGTA CACTGCCACG CCCAGCTAAT TTTTTGTATT TAGTAGAGAC AGGGTTTCAC CGCGTTCGCC 18353
AGGATGGTCT CGATCTCCTG ACCTTGTGAT CCGCCTGCCT CAGCCTCCCA AAGTGCTGGT ATTACGGGCG TGAGCCACTG CGCCCGGCCA GGCCTATTAT 18453
TTTTCTATTG TGGTTCATTA ATTTCTGCTT TTTTCTCTTA AAAAGTTTGC TTACGTTTTT GTCTGGTTTA CTTTGCTGTT CTCTTGCTAG CTTTTTTTTT 18553
TTTCAGATAG GGTCTTGCTC TGTTGCCCAG GCTGGAGTGC AGTGGCACAG TCATAGCTCA CTGCAGCCTT GAACTCCTGG GCTCAAGCAA TCCTCTTCTT 18653
GCTTCAGCCT CCCACGTAGC TAGGATCAGA GGTACATGCC ACCATGTTCG GCTAATTTTT TTTTTTCGAG ACAGAGTCTT GTTCTGTCGC TCAGGCGGTA 18753
GTGCAGTGGT GCAATCCCGG CTCACTGCAA CCTCCACCTC CACCTCCCAG GTTCAAGCAA TTCTACCTCA GTCTCCTGAG TAGCTGGGAT TATAGGCGCA 18853

```
CACCAACATG TCTGGCTAAT TTTTGTATTT TTAGTAGAGA CAGGGTTTCA CCACGTTGGC TAGGCTGGTC TTAAACTCCT GACTTCATGA TCCGCCCGCC 18953
TTGGCCTCCC AAAGTGCTGA GATTACAGGT GTGAGCCACA GCACCTAGTG AAAGTGTGGT TTTTTTGTGT AGGTTTTACT GTTGTTAGTG TTGTTCTGTA 19053
TTGTTTGTAG AGGATACGTG GGGAGATTTG GATAAAAGCA ACTATCATTA TTATCCTCAT CAGACTTGTA GGTCTAACTT TTTAATTTTT TAATTTTTAA 19153
TTTAAATTTT TTTCTTGGTC TTTTATCATT AATTAATTTT TTCGAGACAG GGTCTCACTC TGTTGCCCAG GCTGGAGTGT GGTGACATGA TCACGGCTCA 19253
CTGCAGCCTT AACCTCCCAG GTGCAAGTGA TCCTCCTCTC TTAGCCTCCC GAGTAGCTGG GACTCCAGGC ATGTGCCACC ATGCCCAGCT AATTTTTTGT 19353
AGAGAGAGGG TTTTGCCATA TTGCCCAGGC TGGTCTTGAA CTGCTGAGCT CAAGTGATCC ACCCGGCTTG GGCATGAGCC ACCTCCCCTG GTCTGGTCCA 19453
ACTTTTTAAA AGCATTATTC TGCCTGTTGG GTGGAGAATA GACTGTAGGT GGGCAAAGAA TGAAGGAAAC TAGTGGGTTC AGGAGCTCGA GCTAGAAGTG 19553
GTGAGAAGGG TTTGGATTTG GGGTCTATGC TGAAGGTAGA GCCGACAAGA TTTGCTAGGA TTGGATGTGT AGGGTGAGGA AGTGGGGACA GCAAGAATGA 19653

                                                                                                            1
CTGGAGGGGT AAGTGGACTC TCACCAGCTG TGTCTCGTGA AGGGGCGTGG CTGGGCTATG AGCTATGCTC CTGAGCACAG ACGGCTGTTC TCTTTCAAGG 19753

yTyrLysPro AspGluGlyL ysArgGlyAs pAlaCysGlu GlyAspSerG lyGlyProPh eValMetLys
TTACAAGCCT GATGAAGGGA AACGAGGGGA TGCCTGTGAA GGTGACAGTG GGGGACCCTT TGTCATGAAG GTAAGCTTCT CTAAAGCCCA GGGCCTGGTG 19853
AACACATCTT CTGGGGGTGG GGAGAAACTC TAGTATCTAG AAACAGTTGC CTGGCAGAGG AATACTGATG TGACCTTGAA CTTGACTCTA TTGGAAACCT 19953

                SerP roPheAsnAs nArgTrpTyr GlnMetGlyI leValSerTr pGlyGluGly CysAspArgA spGlyLysTy rGlyPheTyr
CATCTTTCTT CTTCAGAGCC CCTTTAACAA CCGCTGGTAT CAAATGGGCA TCGTCTCATG GGGTGAAGGC TGTGACCGGG ATGGGAAATA TGGCTTCTAC 20053

ThrHisValP heArgLeuLy sLysTrpIle GlnLysValI leAspGlnPh eGlyGlu***
ACACATGTGT TCCGCCTGAA GAAGTGGATA CAGAAGGTCA TTGATCAGTT TGGAGAGTAG GGGGCCACTC ATATTCTGGG CTCCTGGAAC CAATCCCGTG 20153
AAAGAATTAT TTTTGTGTTT CTAAAACTAT GGTTCCCAAT AAAAGTGACT CTCAGCGAGC CTCAATGCTC CCAGTGCTAT TCATGGGCAG CTCTCTGGGC 20253
TCAGGAAGAG CCAGTAATAC TACTGGATAA AGAAGACTTA AGAATCCACC ACCTGGTGCA CGCTGGTAGT CCGAGCACTC GGGAGGCTGA GGTGGGAGGA 20353
T                                                                                                           20354
```

FIGURE 3: Nucleotide sequence of the gene for human prothrombin. Nucleotide 1 is the first nucleotide of the codon for the initiator methionine. The length of the 5′ noncoding region of the gene has not been established. DNA sequence upstream from the initiator methionine codon is numbered in a negative fashion. 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 3′ noncoding region is underlined. Intervening sequences that interrupt a codon are shown by the presence of a partial amino acid triplet. The previously sequenced region of the gene is from nucleotide 4662 to nucleotide 9604 (λ 10; Degen et al., 1983).

Table III: Characteristics of *Alu* and *Kpn* Repetitive Sequences in the Gene for Human Prothrombin

	nucleotide position	strand[a]	% homology[b]	size of direct repeat	direct repeat	mismatch[c]
Alu						
1	2253–2548	−	88.7	14	ATGAATT$^{A}_{C}$ATGTTT	1
2	2903–3034	−[f]	73.9	10	CTC$^{-}_{A}$TTTGTT	1
3	3049–3348	−	89.7			
4	3516–3810	+	79.6	7	TGATTGC	0
5	4806–5105	+	84.3	9	CTGGC$^{T}_{C}$ACC	1
6[d]	5962–6263	−	86.7	8	TTGGTCTT	0
7[d]	6264–6562	−	85.0			
8[d]	8198–8516	−	88.0	17	ATGGGT$^{TG}_{CC}$TTTACTTCT	2
9[d]	8693–8999	+	79.3	16	AAAAAG$^{CAG}_{TGT}$GAGGCAG	3
10	10179–10469	−	77.7	11	CTGGCGTTTTA	0
11[e]	11248–11487	+	87.9	6	AGAGAA	0
12	12123–12404	−	84.5	14	AAAACC$^{C}_{A}$GGAAATT	1
13	12418–12668	−	72.0			
14	13104–13400	−	87.3	13	ACTTTGAT$^{G}_{A}$TTTT	1
15	13579–13885	−	88.0	9	TACTTT$^{A}_{G}$TT	1
16[e]	14457–14688	−	82.0	5	TCCTA	0
17	14961–15248	−	80.7	12	AGGACC$^{T}_{A}$TTTTT	1
18	15260–15412	−[f]	83.0	15	TTATATAATTTTTTT	0
19	15687–15983	+	87.3	8	CTTTCCTG	0
20	16188–16307	+[f]	75.0			
21	16314–16611	+	87.7			
22	16621–16794	+[f]	77.5	6	AACTAG	0
23	16795–17076	+	87.5			
24	17077–17369	+	87.0			
25	17431–17749	−	64.6	7	AGATTTC	0
26	18294–18593	−	90.7	14	AGGCCTATTATTTT	0
27	18594–18888	−	85.2			
28	18985–19155	−[f]	77.4	6	TGGTTT	0
29	19156–19448	−	85.0			
30[e]	19625–19871	−	77.3	5	CTTGG	0
Kpn						
1	14993–15163		68.0[g]	10	TTA$^{T}_{A}$A$^{T}_{A}$AATT	2
2	17910–18236		59.2[h]	5	TAATG	0

[a] With respect to direction of transcription from the RNA polymerase III promoter; (+) represents transcription from 5′ to 3′ on the mRNA-like strand, and (−) represents transcription from the complementary strand. [b] Compared to the consensus sequence of Deininger et al. (1981), insertions are not considered in this calculation. [c] Number of mismatches between the direct repeats. [d] *Alu* repeats 6–9 were published previously by Degen et al. (1983) and were numbered 1–4. [e] Alu repeats that are incomplete. [f] *Alu* repeats represent only half of a complete *Alu* sequence. [g] Compared to sequences of DiGiovanni et al. (1983) and Potter and Jones (1983).

the kringle structures in prothrombin. This domain is present in all of the above-mentioned coagulation and fibrinolytic proteins, except for protein S.

The domain structure and placement of intervening sequences with respect to the amino acid sequence of human prothrombin is shown in Figure 4. Intervening sequences A, B, and C in the prothrombin gene are located in almost the same position in the amino acid sequence as the introns in the genes for factor VII, factor IX, factor X, and protein C (Davie et al., 1983; Anson et al., 1984; Foster et al., 1985; Yoshitake et al., 1985; Leytus et al., 1986; Plutsky et al., 1986; O'Hara et al., 1987). Intervening sequence A, which interrupts the

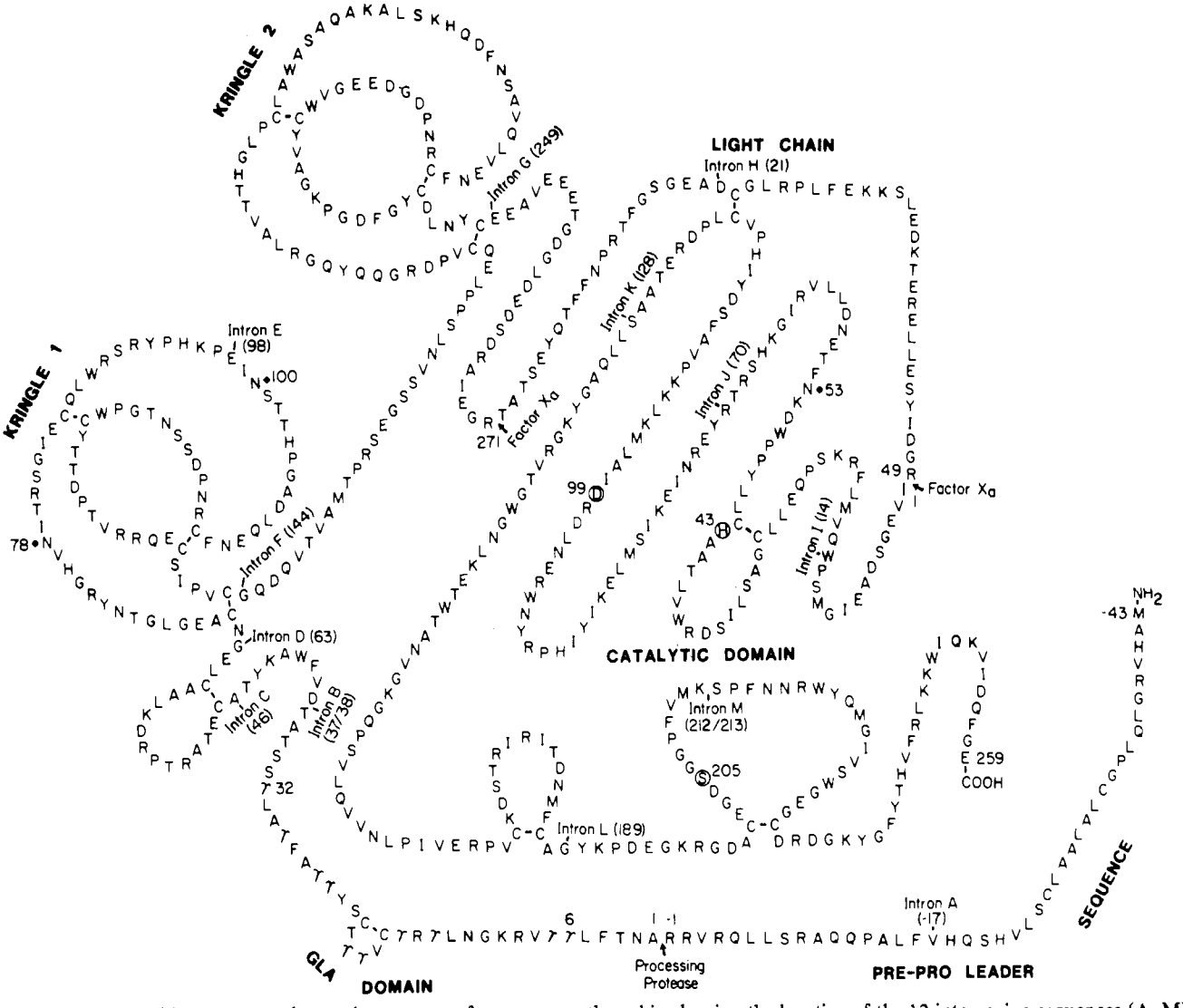

FIGURE 4: Amino acid sequence and tentative structure for prepro prothrombin showing the location of the 13 intervening sequences (A–M). The prepro leader sequence is removed during the biosynthesis of the protein by signal peptidase and a protease that hydrolyzes an R-A peptide bond between −1 and 1. The Gla domain and two kringle structures are located in the amino-terminal region of the protein that is removed after proteolytic activation by factor Xa. Thrombin is composed of a light chain (49 amino acids) linked to a catalytic domain (259 amino acids). Thrombin undergoes further self-autolysis removing 13 amino acids from the N-terminal of the light chain. The amino acids that participate in catalysis are circled (H-43, D-99, and S-205). The three carbohydrate attachment sites are represented by solid diamonds. The positions of the intervening sequences are indicated. The amino acid residues are numbered as follows starting with the amino-terminal end: prepro leader sequence, −43 to −1; prothrombin fragment 1, 1–271; light chain of thrombin, 1–49; catalytic chain of thrombin, 1–259. The Asn at position 121 is coded by an AAC codon in the cDNA (Degen et al., 1983), while an ACC codon at this position in the gene codes for a Thr.

leader sequence at residue −17 in prothrombin, factor VII, factor IX, and factor X and at position −19 in protein C, is near the proposed signal peptidase recognition site in all five proteins (Bentley et al., 1986). All of these intervening sequences are classified as type I (Sharp, 1981) since they interrupt the codon between the first and second base. Intervening sequence B is located immediately following the Gla domain and occurs between residues 37 and 38 in prothrombin, factor VII, factor X, and protein C and between residues 38 and 39 in factor IX. These introns are type O since the intervening sequence occurs between codons. Intervening sequence C occurs at residue 47 in factor IX and at residue 46 in the other four proteins, and all are type I intervening sequences. The minor differences seen for factor IX are due to the presence of an additional amino acid in the amino-terminal end of the protein.

Intervening sequences A and B separate the Gla domain into an exon that includes 17 amino acids of the leader sequence. This region is involved in the γ-carboxylation process (Jorgensen et al., 19877; Suttie et al., 1987; Foster et al., submitted for publication). The Gla domains is the region most conserved between the vitamin K dependent proteins and the placement of intervening sequences in this region indicates that this conservation extends to the gene organization.

Although there is similar placement of intervening sequences A, B, and C in the genes for the vitamin K dependent proteins, there is no similarity in the sizes of the respective intervening sequences and little detectable sequence homology. Intervening sequence A is 386 bp in length in prothrombin compared to 2574, 6206, and 1263 bp in the genes for factor VII (O'Hara et al., 1987), factor IX (Yositake et al., 1985), and protein C (Foster et al., 1985), respectively. The exact size of this intervening sequence in factor X has not been determined. Intervening sequence B in prothrombin is 659 bp in length compared to 1919, 188, 1462, and 7400 bp for factor VII, factor IX, protein C, and factor X (Leytus et al., 1986), respectively. Intervening sequence C is 242 bp in length in prothrombin and 68, 3689, 92, and 950 bp in factor VII, factor

IX, protein C, and factor X, respectively. The size difference and lack of sequence homology of these intervening sequences indicates that there is no apparent evolutionary constraint on these regions of the gene, while there is major conservation of the sequences of the exons.

The placement of intervening sequences with respect to the two kringle regions is shown in Figure 4. Intervening sequences D, F, and G flank these domains. The second kringle of prothrombin is encoded by one exon, while the first kringle is encoded by two exons due to the presence of intervening sequence E within the DNA coding for this kringle. Both kringles of t-PA, the one kringle present in u-PA, and at least one kringle in plasminogen are also encoded by two exons (Degen et al., 1986; Ny et al., 1984; Riccio et al., 1985; Malinowski et al., 1984). The placement of the internal intervening sequnce with respect to the amino acid sequence of the first kringle in prothrombin is different from those found in t-PA, u-PA, and plasminogen. This is interesting since the DNA coding for the kringles of t-PA and u-PA have identical placement of intervening sequences. From these data and the homology of their DNA and amino acid sequences, it is generally accepted that t-PA and u-PA are more closely related to each other than to prothrombin (Patthy, 1985).

The placement of intervening sequences with respect to the catalytic domain in prothrombin is also shown in Figure 4. Each of the amino acids that make up the active site (His-43, Asp-99, and Ser-205) is in individual exons. This has also been observed in the genes coding for t-PA (Degen et al., 1986; Ny et al., 1984), u-PA (Riccio et al., 1985), kallikrein (Mason et al., 1983), factor B (Campbell & Porter, 1983), trypsin (Craik et al., 1984), chymotrypsin (Bell et al., 1984), and elastase (Swift et al., 1984). In contrast, the genes for factor VII, factor IX, factor X, and protein C have only the active site His region in an individual exon while the active site Asp and Ser regions are encoded in the same exon. Differences in the number of intervening sequences may be attributed to insertions or deletions occurring throughout evolution, and slight variations in their placement may be due to the sliding of intron–exon junctions (Stone et al., 1985; Craik et al., 1983).

The sizes of genes coding for serine proteases vary tremendously with the size of the gene coding for human prothrombin with the latter falling in the middle size range. The prothrombin gene is ~21 kb in length, while t-PA is 32.7 kb (Degen et al., 1986), factor IX is 33.5 kb (Yoshitake et al., 1985), factor X is ~22 bp (Leytus et al., 1986), complement component C2 is ~18 kb (Campbell & Bentley, 1985), factor VII is 12.8 kb (O'Hara et al., 1987), elastase is ~12 kb (Swift et al., 1984), protein C is ~11 kb (Foster et al., 1985), complement factor B is 6 kb (Campbell & Porter, 1983), trypsin is ~4.5 kb (Craik et al., 1984), and chymotrypsin is ~4.7 kb (Bell et al., 1984). Accordingly, it is clear that there is no correlation of gene size to the size of the encoded serine protease.

It is not known whether an intervening sequence is present within the 5′ noncoding region of the human prothrombin gene. There is no DNA sequence homology of the 5′ noncoding regions of factor IX and protein C with the DNA upstream from the initiator methionine in the prothrombin gene. The 5′ noncoding region in the gene coding for human factor IX is 29 bp (Yoshitake et al., 1985) with no intervening sequence interrupting this region, while the gene coding for human protein C contains a 5′ noncoding region of 75 bp that is interrupted by a 1426-bp intervening sequence (Foster et al., 1985; Plutzky et al., 1986). The longest human prothrombin cDNA sequence available (MacGillivray et al., 1986) includes 27 bp of 5′ noncoding sequence. This sequence is present immediately 5′ to the initiator methionine in the gene except for the first four bases in the cDNA, which are CCCT. The sequence in the gene in this region is CCTC. Also, this sequence in the gene does not conform with the consensus sequence for a 3′ end of an intervening sequence (Mount, 1982). These conflicting results can only be resolved when a longer cDNA is isolated. Preliminary results of primer extension experiments indicate that the 5′ noncoding region of the mRNA for human prothrombin is 166 bases in length. This would place the site of initiation of transcription near nucleotide −166 (Figure 3). Unfortunately, there is no sequence 30 bp upstream of this site that conforms with a "TATA" box sequence for RNA polymerase II recognition. By including a 166-base 5′ noncoding region, the mRNA for human prothrombin would be approximately 2130 bases in length, excluding the poly(A) tail. The average length of a poly(A) tail is 200 nucleotides (Perry, 1976), and thus the prothrombin mRNA would be approximately 2300 nucleotides. This size is consistent with the size of mRNA coding for human and bovine prothrombin as determined by Northern analysis (Degen et al., 1983; MacGillivray & Davie, 1984). The actual placement of the site of transcription initiation will have to await further experiments.

The size of the human prothrombin gene is 20 210 bp while the bovine gene is approximately 15 kb in length (Irwin et al., 1985). A comparison of the cDNA sequences of human and bovine prothrombin indicates that they are 88% homologous in the protein coding region and 70% in the 3′ noncoding region. The human and bovine genes appear to be organized in a similar manner in regard to the number and placement of intervening sequences. Intervening sequences D, F, and L in the human gene, however, are significantly larger than their bovine counterparts. These three intervening sequences also contain the highest number of *Alu* repetitive DNA sequences within the gene (Table I). The high content of repetitive DNA within the human gene could account for the size difference between the human and bovine genes, although the repetitive DNA content of the bovine gene has not been reported.

The human prothrombin gene contains 30 *Alu* repeats and two *Kpn* repeats that constitute 40% of the gene. This high content of repetitive DNA is very unusual. Of the characterized genes coding for coagulation and fibrinolytic proteins, only t-PA has a large number of *Alu* repeats (27 copies) that make up 24% of the gene sequence (Degen et al., 1986). Factor IX has four *Alu* repeats, while protein C and u-PA have two each. Factor VII, however, has no *Alu* repeats (O'Hara et al., 1987). Three other characterized genes have been found to contain clusters of *Alu* repetitive DNA (Slagel et al., 1986). These include the β-tubulin gene with 10 repeats in a ~5-kb intervening sequence (Lee et al., 1984), the thymidine kinase gene with 13 copies in a 10-kb region (Flemington et al., 1987), and 21 *Alu* repeats in 32 kb of the human adenosine deaminase gene (Wiginton et al., 1986). As yet there is not obvious explanation for the high repetitive DNA content of these genes, including prothrombin.

When the sequence of the gene for prothrombin was compared to sequences compiled in the Genetic Sequence Data Bank (GenBank, Sepember 1986 compilation), the only significant homologies observed were with other serine proteases and repetitive sequences. A large number of short (~20 nucleotides) homologies were identified, but their significance is unknown. Three regions were found that did not appear to be homologous with *Alu* or *Kpn* repeats but were repeated in several other genes. For instance, nucleotides 14295–14397

(Figure 3) were 72% homologous to a clone of a rat repetitive sequence (Whitney & Furano, 1984). This repeat was also found in a cytochrome *c* pseudogene (77% homologous; Scarpulla, 1985) and rat insulin I gene (76% homologous; Cordell et al., 1979). Nucleotides 10378–10524 (Figure 3) were homologous to regions in the genes for human β-tubulin [76% to nucleotides 3714–3787 in Lee et al. (1983)], the γ-chain of human fibrinogen [66.7% to nucleotides 4834–4981 in Rixon et al. (1985)], human adenosine deaminase [68% to nucleotides 8278–8386 in Wiginton et al. (1986)], and bovine acetylcholine receptor β-subunit [73% to nucleotides 2168–2229 in Tanabe et al. (1984)]. In all of these genes the repeated sequence is present in an intervening sequence. Nucleotides 10571–10632 (Figure 3) are also 75% homologous to regions in several β-tubulin pseudogene clones (Lee et al., 1983) and 70% homologous to the human immune interferon gene [nucleotides 5859–5921 in Gray and Goeddel (1982)]. Whether these sequences are repeated elsewhere in the genome remains to be determined.

ACKNOWLEDGMENTS

We thank Dr. Philip Leder for kindly providing us with his human genomic library. We also thank Lorie Schaefer for excellent technical assistance, Jan Hagedorn for preparation of the manuscript, and Dr. Roger Ganschow for assistance with computer data processing.

REFERENCES

Anson, D. S., Choo, K. H., Rees, D. J. G., Giannelli, F., Gould, K., Huddleston, J. A., & Brownlee, G. G. (1984) *EMBO J. 3*, 1053–1060.

Bányai, L., Váradi, A., & Patthy, L. (1983) *FEBS Lett. 163*, 37–41.

Barnhart, M. I. (1960) *Am. J. Physiol. 199*, 360–366.

Bell, G. I., Quinto, C., Quiroga, M., Valenzuela, P., Craik, C. S., & Rutter, W. J. (1984) *J. Biol. Chem. 259*, 14265–14270.

Bentley, A. K., Rees, D. J. G., Rizza, C., & Brownlee, G. G. (1986) *Cell (Cambridge, Mass.) 45*, 343–348.

Benton, W. D., & Davis, R. W. (1977) *Science (Washington, D.C.) 196*, 180–182.

Birnboim, H. C., & Doly, J. (1979) *Nucleic Acids Res. 7*, 1513–1523.

Blake, C. (1983) *Nature (London) 306*, 535–537.

Blobel, G., Walter, P., Chang, C. N., Goldman, B. M., Erikson, A. H., & Lingappa, R. (1979) *Soc. Exp. Biol. Symp. 33*, 9–36.

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

Broderick, T. P., Schaff, D. A., Bertino, A. M., Dush, M. K., Tischfield, J. A., & Stambrook, P. J. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 3349–3353.

Butkowski, R. J., Elion, J., Downing, M. R., & Mann, K. G. (1977) *J. Biol. Chem. 252*, 4942–4957.

Campbell, R. D., & Porter, R. R. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 4464–4468.

Campbell, R. D., & Bentley, D. R. (1985) *Immunol. Rev. 87*, 19–37.

Chen, L. B., & Buchanan, J. M. (1975) *Proc. Natl. Acad. Sci. U.S.A. 72*, 131–135.

Cordell, B., Bell, G., Tischer, E., DeNoto, F. M., Ullrich, A., Pictet, R., Rutter, W. J., & Goodman, H. M. (1979) *Cell (Cambridge, Mass.) 18*, 533–543.

Craik, C. S., Rutter, W. J., & Fletterick, R. (1983) *Science (Washington, D.C.) 220*, 1125–1129.

Craik, C. S., Choo, Q.-L., Swift, G. H., Quinto, C., MacDonald, R. J., & Rutter, W. J. (1984) *J. Biol. Chem. 259*, 14255–14264.

Dahlbäck, B., Lundwall, A., & Stenflo, J. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 4199–4203.

Daniels, G. R., & Deininger, P. L. (1985) *Nucleic Acids Res. 13*, 8939–8954.

Davey, M. G., & Luscher, E. F. (1967) *Nature (London) 216*, 857–858.

Davie, E. W., Fujikawa, K., Kurachi, K., & Kisiel, W. (1979) *Adv. Enzymol. Relat. Areas Mol. Biol. 48*, 277–318.

Davie, E. W., Degen, S. J. F., Yoshitake, S., & Kurachi, K. (1983) *Dev. Biochem. 25*, 45–52.

Degen, S. J. F., & Davie, E. W. (1986) *Ann. N.Y. Acad. Sci. 485*, 66–72.

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

Degen, S. J. F., Rajput, B., Reich, E., & Davie, E. W. (1985) *Protides Biol. Fluids 33*, 47–50.

Degen, S. J. F., Rajput, B., & Reich, E. (1986) *J. Biol. Chem. 261*, 6972–6985.

Deininger, P. L., Jolly, D. J., Rubin, C. M., Friedmann, T., & Schmid, C. W. (1981) *J. Mol. Biol. 151*, 17–33.

DiGiovanni, L., Haynes, S. R., Misra, R., & Jelinek, W. R. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 6533–6537.

DiScipio, R. G., Hermodson, M. A., Yates, S. G., & Davie, E. W. (1977) *Biochemistry 16*, 698–706.

Diuguid, D. L., Rabiet, M. J., Furie, B. C., Liebman, H. A., & Furie, B. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 5803–5807.

Dodgson, J. B., & Engel, J. D. (1983) *J. Biol. Chem. 258*, 4623–4629.

Elion, J., Downing, M. R., Butkowski, R. J., & Mann, K. G. (1977) in *Chemistry and Biology of Thrombin* (Lundblad, R., Fenton, J. W., II, & Mann, K. G., Eds.) pp 97–111, Ann Arbor Science, Ann Arbor, MI.

Esmon, C. T. (1983) *Blood 62*, 1155–1158.

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

Feinberg, A. P., & Vogelstein, B. (1983) *Anal. Biochem. 132*, 6–13.

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

Flemington, E., Bradshaw, H., Traina-Dorge, V., Slagel, V., & Deininger, P. (1987) *Gene 52*, 267–277.

Foster, D. C., Yoshitake, S., & Davie, E. W. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 4673–4677.

Fung, M. R., Campbell, R. M., & MacGillivray, R. T. A. (1984) *Nucleic Acids Res. 12*, 4481–4492.

Gilbert, W. (1985) *Science (Washington, D.C.) 228*, 823–824.

Gray, P. W., & Golddel, D. V. (1982) *Nature (London) 298*, 859–863.

Grunstein, M., & Hogness, D. S. (1975) *Proc. Natl. Acad. Sci. U.S.A. 72*, 3961–3965.

Hagen, F. S., Gray, C. L., O'Hara, P., Grant, F. J., Saari, G. C., Woodbury, R. G., Hart, C. E., Insley, M., Kisiel, W., Kurachi, K., & Davie, E. W. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 2412–2416.

Irwin, D. M., Ahern, K. G., Pearson, G. D., & MacGillivray, R. T. A. (1985) *Biochemistry 24*, 6854–6861.

Jorgensen, M. J., Cantor, A. B., Furie, B. C., & Furie, B. (1986) *Circulation 74 (Suppl. II)*, Abstr. 1637.

Jorgensen, M. J., Cantor, A. B., Furie, B. C., Brown, C. L., Shoemaker, C. B., & Furie, B. (1987) *Cell (Cambridge, Mass.) 48*, 185–191.

Katayama, K., Ericsson, L. H., Enfield, D. L., Walsh, K. A., Neurath, H., Davie, E. W., & Titani, K. (1979) *Proc. Natl. Acad. Sci. U.S.A. 76*, 4990–4994.
Katz, L., Kingsbury, D. T., & Helinski, D. R. (1973) *J. Bacteriol. 114*, 577–591.
Katz, L., Williams, P. H., Sato, S., Leavitt, R. W., & Helinski, D. R. (1977) *Biochemistry 16*, 1677–1683.
King, C. R., & Piatigorsky, J. (1983) *Cell (Cambridge, Mass.) 32*, 707–712.
Kurachi, K., & Davie, E. W. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 6461–6464.
Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G., & Maniatis, T. (1978) *Cell (Cambridge, Mass.) 15*, 1157–1174.
Lee, M. G.-S., Lewis, S. A., Wilde, C. D., & Cowan, N. J. (1983) *Cell (Cambridge, Mass.) 33*, 477–487.
Lee, M. G.-S., Loomis, C., & Cowan, N. J. (1984) *Nucleic Acids Res. 12*, 5823–5836.
Leytus, S. P., Foster, D. C., Kurachi, K., & Davie, E. W. (1986) *Biochemistry 25*, 5098–5102.
Long, G. L., Belagaje, R. M., & MacGillivray, R. T. A. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 5653–5656.
MacGillivray, R. T. A., & Davie, E. W. (1984) *Biochemistry 23*, 1626–1634.
MacGillivray, R. T. A., Irwin, D. M., Guinto, E. R., & Stone, J. C. (1986) *Ann. N.Y. Acad. Sci. 485*, 73–79.
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.
Malinowski, D. P., Sadler, J. E., & Davie, E. W. (1984) *Biochemistry 23*, 4243–4250.
Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*, pp 371–372, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
Mann, K. G., & Elion, J. (1980) in *CRC Handbook Series in Clinical Laboratory Science* (Schmidt, R. M., Ed.) Section I, Vol. 3, pp 15–31, CRC Press, Boca Raton, FL.
Mason, A. J., Evans, B. A., Cox, D. R., Shine, J., & Richards, R. I. (1983) *Nature (London) 303*, 300–307.
Maxam, A. M., & Gilbert, W. (1980) *Methods Enzymol. 65*, 499–560.
Mount, S. M. (1982) *Nucleic Acids Res. 10*, 459–472.
Naora, H., & Deacon, N. J. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 6196–6200.
Nelsestuen, G. L., Zytkovicz, T. H., & Howard, J. B. (1974) *J. Biol. Chem. 249*, 6347–6350.
Neurath, H. (1984) *Science (Washington, D.C.) 224*, 350–357.
Nussinov, R. (1981) *J. Mol. Biol. 149*, 125–131.
Ny, T., Elgh, F., & Lund, B. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 5355–5359.
O'Hara, P. J., Grant, F. J., Haldemann, B. A., Gray, C. L., Insley, M. Y., Hagen, F. S., & Murray, M. J. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 5158–5162.
Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S., & Sharp, P. A. (1986) *Annu. Rev. Biochem. 55*, 1119–1150.
Pan, L. C., & Price, P. A. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 6109–6113.
Patthy, L. (1985) *Cell (Cambridge, Mass.) 41*, 657–663.
Patthy, L., Trexler, M., Vâli, Z., Bânyai, L., & Vâradi, A. (1984) *FEBS Lett. 171*, 131–136.
Perry, R. P. (1976) *Annu. Rev. Biochem. 45*, 605–629.
Plutzky, J., Hoskins, J. A., Long, G. L., & Crabtree, G. R. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 546–550.
Potter, S. S. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 1012–1016.
Potter, S. S., & Jones, R. S. (1983) *Nucleic Acids Res. 11*, 3137–3153.
Queen, C., & Korn, L. J. (1984) *Nucleic Acids Res. 12*, 581–599.
Riccio, A., Grimaldi, G., Verde, P., Sebastio, G., Boast, S., & Blasi, F. (1985) *Nucleic Acids Res. 13*, 2759–2771.
Rimm, D. L., Hogness, D., Kucera, J., & Blattner, F. R. (1980) *Gene 12*, 301–309.
Rinehart, F. P., Ritch, T. G., Deininger, P. L., & Schmid, C. W. (1981) *Biochemistry 20*, 3003–3010.
Rixon, M. W., Chung, D. W., & Davie, E. W. (1985) *Biochemistry 24*, 2077–2086.
Rogers, J. (1985) *Nature (London) 315*, 458–459.
Sanger, F., & Coulson, A. R. (1978) *FEBS Lett. 87*, 107–110.
Scarpulla, R. C. (1985) *Nucleic Acids Res. 13*, 763–775.
Schmid, C. W., & Jelinek, W. R. (1982) *Science (Washington, D.C.) 216*, 1065–1070.
Seegers, W. H. (1979) *Prog. Chem. Fibrinolysis Thrombolysis 4*, 241–254.
Sharp, P. A. (1981) *Cell (Cambridge, Mass.) 23*, 643–646.
Sharp, P. A. (1987) *Science (Washington, D.C.) 235*, 766–771.
Slagel, V., Flemington, E., Traina-Dorge, V., Bradshaw, H., & Deininger, P. (1987) *Mol. Biol. Evol. 4*, 19–29.
Smith, G. E., & Summers, M. D. (1980) *Anal. Biochem. 109*, 123–129.
Southern, E. M. (1975) *J. Mol. Biol. 98*, 503–517.
Stenflo, J., Fernlund, P., Egan, W., & Roepstorff, P. (1974) *Proc. Natl. Acad. Sci. U.S.A. 71*, 2730–2733.
Stone, E. M., Rothblum, K. N., & Schwartz, R. J. (1985) *Nature (London) 313*, 498–500.
Suttie, J. W., Hoskins, J. A., Engelke, J., Hopfgartner, A., Ehrlich, H., Bang, N. U., Belagaje, R. M., Schoner, B., & Long, G. L. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 634–637.
Swift, G. H., Craik, C. S., Stary, S. J., Quinto, C., Lahaie, R. G., Rutter, W. J., & MacDonald, R. J. (1984) *J. Biol. Chem. 259*, 14271–14278.
Tanabe, T., Noda, M., Furntani, Y., Takai, T., Takahashi, H., Tanaka, K., Hirose, T., Inayama, S., & Numa, S. (1984) *Eur. J. Biochem. 144*, 11–17.
Thomashow, M. F., Nutter, R., Montoya, A. L., Gordon, M. P., & Nester, E. W. (1980) *Cell (Cambridge, Mass.) 19*, 729–739.
Thompson, A. R., Enfield, D. L., Ericsson, L. H., Legaz, M. E., & Fenton, J. W., II (1977) *Arch. Biochem. Biophys. 178*, 356–367.
Van Zonneveld, A.-J., Veerman, H., & Pannekoek, H. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 4670–4674.
Walz, D. A., Hewett-Emmett, D., & Seegers, W. H. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 1969–1972.
Ware, J., Liebman, H. A., Kasper, C., Graham, J., Furie, B. C., Furie, B., & Stafford, D. (1986) *Blood 68 (Suppl.)*, 343a.
Whitney, F. R., & Furano, A. V. (1984) *J. Biol. Chem. 259*, 10481–10492.
Wiginton, D. A., Kaplan, D. J., States, J. C., Akeson, A. L., Perme, C. M., Bilyk, I. J., Vaughn, A. J., Lattier, D. L., & Hutton, J. J. (1986) *Biochemistry 25*, 8234–8244.
Woo, S. L. C. (1979) *Methods Enzymol. 68*, 389–395.
Yoshitake, S., Schach, B. G., Foster, D. C., Davie, E. W., & Kurachi, K. (1985) *Biochemistry 24*, 3736–3750.