The Gene for Human Protein Z Is Localized to Chromosome 13 at Band q34 and Is Coded by Eight Regular Exons and One Alternative Exon[†,‡]

Koichiro Fujimaki,[§] Tomio Yamazaki,[§] Masafumi Taniwaki,[||] and Akitada Ichinose*[,§]

Department of Molecular Patho-Biochemistry, Yamagata University School of Medicine, Iida-Nishi 2-2-2, Yamagata, 990-9585 Japan, and Department of Medicine III, Kyoto Prefectural Medical School, Shogo-in, Sakyo-ku, Kyoto, Japan

Received August 12, 1997; Revised Manuscript Received January 7, 1998

ABSTRACT: Human protein Z is a vitamin K-dependent plasma glycoprotein, deficiency of which leads to a mild bleeding tendency. Protein Z appears to assist hemostasis by binding thrombin and promoting its association with phospholipid vesicles. In this study, to characterize the gene for protein Z, its organization and structure were determined by a combination of PCR amplification of leukocyte DNA and isolation of phage clones from a human genomic library. The gene spanned about 14 kb and consisted of 9 exons including one alternative exon. It was of note that the gene organization was essentially identical to that of other vitamin K-dependent proteins, such as factors VII, IX, and X and protein C. The nucleotides in introns at exon/intron boundaries for eight regular exons were the consensus GT-AG sequences. In contrast, the sequence at an optional exon/intron junction was found to be GC rather than GT. The extra exon inserts a unique peptide consisting of 22 amino acids in the prepro-leader sequence. A similar situation was previously observed in factor VII, but not in other vitamin K-dependent plasma proteins. We also assigned the gene for protein Z to chromosome 13 by PCR amplification of genomic DNAs from human/hamster cell hybrids. Fluorescence in situ hybridization, employing a genomic clone coding for human protein Z, further localized the gene to band q34, where the genes of three other vitamin K-dependent proteins are clustered. These genes may have evolved via duplication of an ancestral gene at this locus.

Human protein Z (M_r = 62 000) is a single-chain glycoprotein that is synthesized in the liver and secreted into the blood (*1, 2*). Its concentration in plasma varies widely between individuals, with an average of 2.9 μg/mL. The protein's half-life is reported to be approximately 2.5 days (*2, 3*).

In this group's previous study, the mRNA for human protein Z was characterized as cDNAs and shown to be approximately 1.6 kb in length (*4*). It codes for a leader sequence containing the typical hydrophobic residues for secretion, and a propeptide region for recognition of the protein by the γ-carboxylase. The C-terminus of the propeptide also contains an Arg-Trp-Lys-Arg sequence for the processing enzyme (furin) which requires the −4 Arg and −1 Arg residues for cleavage. Protein Z contains 13 γ-carboxyglutamic acid (Gla)[1] residues (*4*−*6*) located in the N-terminal region of the protein, and these residues require vitamin K for their biosynthesis.

Mature human protein Z consists of 360 amino acids. Starting with the N-terminus, there are a Gla domain, two EGF domains, and a hinge region between the second EGF domain and a homologue of the catalytic B chain for serine proteases. Neither human nor bovine protein Z contains the active site His and Ser residues in the catalytic triad, whereas the active site Asp residue is conserved (*4*−*6*). The region around the typical activation cleavage site is also absent in both the human and bovine proteins. Thrombin treatment of human protein Z results in a decrease of its apparent molecular weight from 62 000 to 56 000 (*2*) and the loss of a Gla domain (*3*). The Trp-Arg-Arg-Tyr sequence starting at position 42 in human protein Z is the best candidate for the cleavage site by thrombin (*3*).

The structural features of protein Z suggest that it binds to phospholipid in the presence of calcium ions. In fact, it has been reported that protein Z binds thrombin and promotes its association with phospholipid vesicles in a calcium ion-dependent manner (*7, 8*). It is likely that protein Z plays an important role in vivo in hemostasis, since several cases of protein Z deficiency have been identified in association with bleeding symptoms (*9, 10*).

Since it has been documented that bleeding disorders accompany cases of protein Z deficiency, it is important to determine the structure and organization of the normal gene to compare it with abnormal genes. Knowledge regarding the protein Z gene could also provide some insight into its regulation as well as its evolution in relation to other closely related genes, such as those of factors VII, IX, and X and proteins C and S (*11*).

[†] This study was supported in part by research grants from the Ministry of Education, Science and Culture, Japan (09770808), Naito Foundation (Japan), and the Japan Research Foundation for Clinical Pharmacology.

[‡] The nucleotide sequence in this paper has been submitted to DDBJ, Accession Number 98041514353500142.

* Address correspondence to this author. Telephone: (81) 236-28-5275. Fax: (81) 236-28-5280. E-mail: aichinos@med.id.yamagata-u.ac.jp.

[§] Yamagata University School of Medicine.

[||] Kyoto Prefectural Medical School.

[1] Abbreviations: BGP, bone Gla protein; EGF, epidermal growth factor; GAS6, growth arrest-specific gene 6; Gla, γ-carboxyglutamic acid; FISH, fluorescence in situ hybridization; MGP, matrix Gla protein; PCR, polymerase chain reaction.

S0006-2960(97)02002-3 CCC: $15.00
Published on Web 04/28/1998

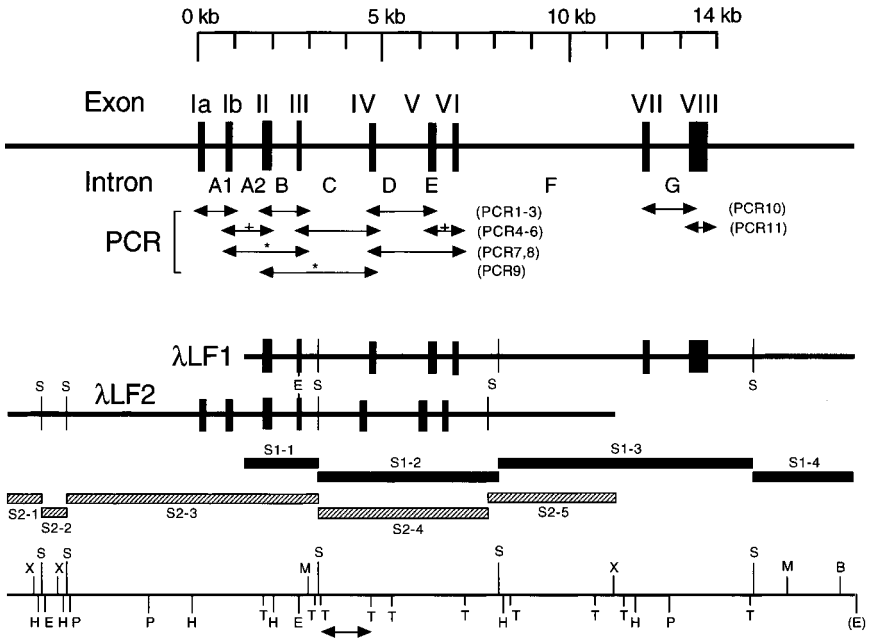

FIGURE 1: Organization and *Sac*I restriction map of the gene for human protein Z. (Top) The nine exons are shown with wide vertical bars and are numbered with roman numerals, while the eight introns are depicted by capitals. The 5′ and 3′ portions of the gene were amplified by PCR employing the specific primers shown in Figure 2. Two PCR products (PCR7 and 9) used for Southern blotting are indicated by arrows with asterisks, and those (PCR 4 and 6) used for chromosomal assignment are indicated by arrows with "+" symbols. (Middle) Two λ phage clones with DNA inserts coding for protein Z are also shown, and narrow vertical bars depict *Sac*I restriction sites. Closed and hatched bars represent subcloned plasmids for λLF1 and λLF2, respectively, containing individual *Sac*I fragments. (Bottom) Capitals in the restriction map stand for restriction sites: B, *Bam*HI; E, *Eco*RI; H, *Hind*III; M, *Sma*I; P, *Sph*I; S, *Sac*I; T, *Pst*I; X, *Xba*I. The polymorphic *Pst*I fragment is shown by an arrow.

In this study, the gene for human protein Z was characterized, and its gene structure was compared to those of other vitamin K-dependent proteins, such as factors VII, X, and IX and protein C. We also localized the gene to chromosome 13 at band q34, where three other members of this same gene family are clustered.

MATERIALS AND METHODS

Southern Blot Hybridization of Genomic DNAs. Genomic DNA samples were prepared from the leukocytes obtained from normal individuals by a standard technique (*12*). Ten micrograms of genomic DNA was digested with 20 units of either *Eco*RI or *Pst*I restriction enzyme and applied to a 0.4% agarose gel. The DNA fragments were transferred to a nylon membrane (Stratagene, La Jolla, CA), and the membrane was hybridized with a ^{32}P-labeled cDNA probe (*4*) or a ^{32}P-labeled amplified DNA fragment of 2.0 kb containing exons Ib, II, and III and introns A2 and B (PCR7, Figure 1, top) or 2.8 kb containing exons II, III, and IV and introns B and C (PCR9).

In Vitro Amplification of Genomic DNAs Employing Gene-Specific Primers. The gene for protein Z was first amplified by PCR of genomic DNA obtained from normal individuals, employing specific primers (Figure 2). These primers for putative exons were designed from the appropriate regions of the cDNA by hypothesizing that the gene organization of protein Z is identical to that of other vitamin K-dependent proteins (*11*). Genomic DNA, 0.1−1.0 μg was amplified in a 50-μL reaction mixture as described (*13*) employing 2.5−5.0 units of *Thermus aquaticus* DNA polymerase (*Taq* polymerase, Stratagene).

Nucleotide Sequence Analysis. The amplified DNA samples were subcloned into M13mp18 or M13mp19 (GIBCO-BRL, Gathersburg, MD). The DNA sequence of an insert was then obtained using the dideoxynucleotide method (*14*) with deoxyadenosine 5′-[α-[^{35}S]thio]triphosphate (Amersham, Arlington Heights, IL). The DNA sequence of the phage clones isolated by genomic screening was also obtained using the dideoxynucleotide method with an ABI 373A sequence analyzer (Perkin-Elmer). Oligonucleotides were synthesized as sequencing primers to obtain the DNA sequence of the second strand. The nucleotide sequence was determined two or more times, and more than 90% of the nucleotide sequence was confirmed on both strands.

Screening of Genomic Phage Library. A cDNA coding for human protein Z (*4*) was employed to isolate genomic clones containing the gene for the protein by screening a human lung fibroblast genomic library in λFIX (Stratagene). To select the correct genomic clones coding for protein Z, the isolated phage clones were first amplified by PCR employing the gene-specific primers, as described above. Several portions of the amplified DNA were then subjected to DNA sequencing analysis. Genomic DNA inserts were released by digestion of the phage DNA with *Sac*I endonuclease, subcloned into plasmid pUC19, and then subjected to sequencing analysis.

Rapid Amplification of 5′-cDNA Ends. To determine the transcription initiation site of the gene for protein Z, rapid amplification of 5′-cDNA ends (5′RACE) was performed using the 5′RACE System, version 2.0 (GIBCO-BRL) and total RNA from Huh7, HLF, HepG2, and Chang liver cells, following the manufacturer's instructions. The cDNA was first synthesized from 500 ng of total RNA using Superscript II reverse transcriptase (GIBCO-BRL) and an antisense primer specific for protein Z (GSP-1 in Figure 2) at 42 °C

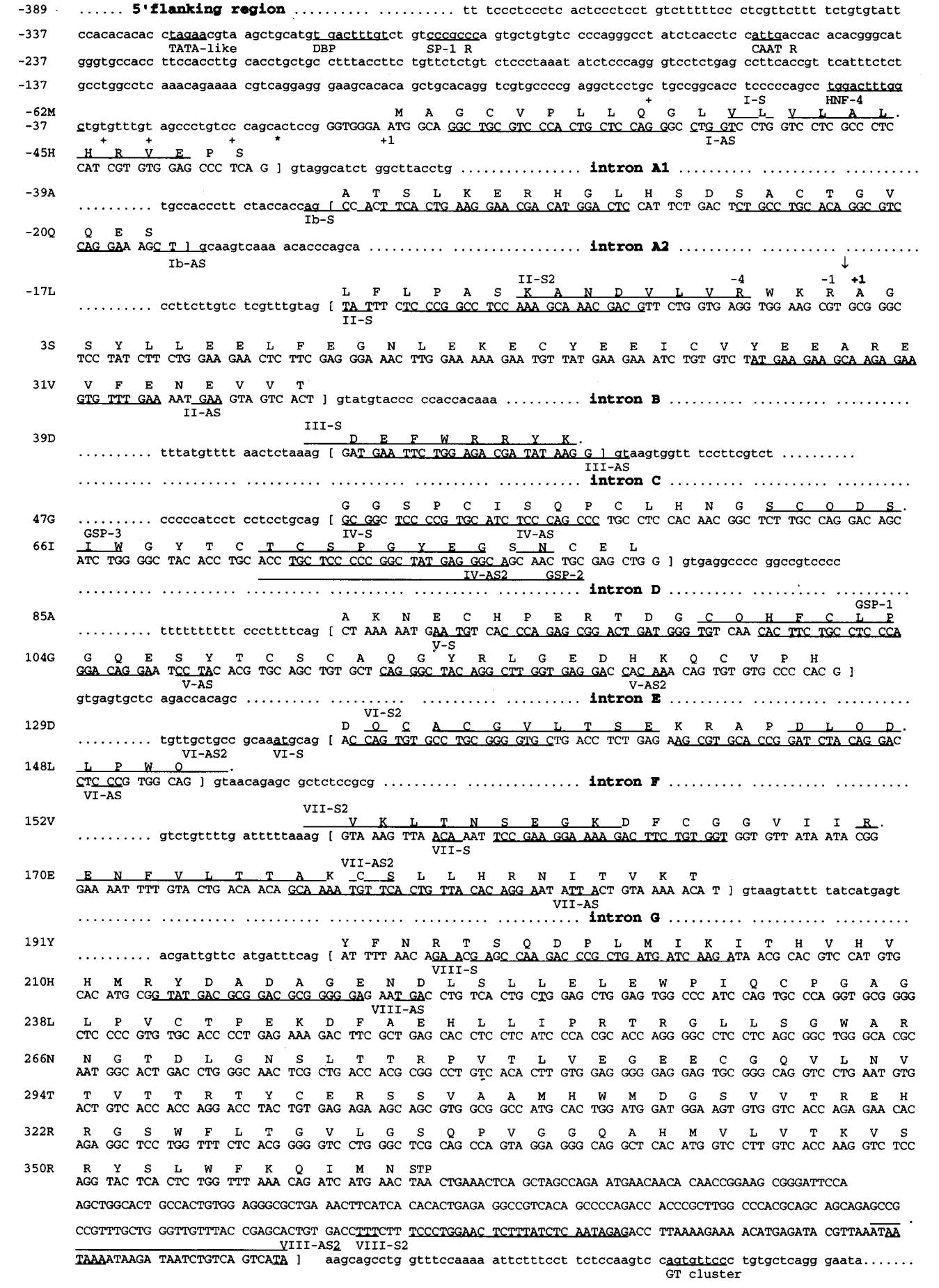
-389 ..... 5'flanking region .... tt tccctccctc actccctcct gtctttttcc ctcgttcttt tctgtgtatt
-337 ccacacacac ctagaacgta agctgcatgt gactttgtct gtcccgccca gtgctgtgtc cccagggcct atctcacctc cattgaccac acacgggcat
TATA-like DBP SP-1 R CAAT R
-237 gggtgccacc ttccaccttg cacctgctgc ctttaccttc tgttctctgt ctccctaaat atctcccagg gtcctctgag ccttcaccgt tcatttctct
-137 gcctggcctc aaacagaaaa cgtcaggagg gaagcacaca gctgcacagg tcgtgccccg aggctcctgc tgccggcacc tcccccagcc tggactttgg
I-S HNF-4
-62M M A G C V P L L Q G L V L V L A L.
-37 ctgtgtttgt agccctgtcc cagcactccg GGTGGGA ATG GCA GGC TGC GTC CCA CTG CTC CAG GGC CTG GTC CTG GTC CTC GCC CTC
I-AS
-45H H R V E P S
CAT CGT GTG GAG CCC TCA G] gtaggcatct ggcttacctg ..... intron A1 ....
-39A A T S L K E R H G L H S D S A C T G V
..... tgccaccctt ctaccaccag [CC ACT TCA CTG AAG GAA CGA CAT GGA CTC CAT TCT GAC TCT GCC TGC ACA GGC GTC
Ib-S
-20Q Q E S
CAG GAA AGC T] gcaagtcaaa acacccagca .... intron A2 ....
Ib-AS
II-S2 -4 -1 +1
-17L L F L P A S K A N D V L V R W K R A G
..... ccttcttgtc tcgtttgtag [TA TTT CTC CCG GCC TCC AAA GCA AAC GAC GTT CTG GTG AGG TGG AAG CGT GCG GGC
II-S
3S S Y L L E E L F E G N L E K E C Y E E I C V Y E E A R E
TCC TAT CTT CTG GAA GAA CTC TTC GAG GGA AAC TTG GAA AAA GAA TGT TAT GAA GAA ATC TGT GTC TAT GAA GAA GCA AGA GAA
31V V F E N E V V T
GTG TTT GAA AAT GAA GTA GTC ACT] gtatgtaccc ccaccacaaa ..... intron B ....
II-AS
III-S
39D D E F W R R Y K.
..... tttatgtttt aactctaaag [GAT GAA TTC TGG AGA CGA TAT AAG G] gtaagtggtt tccttcgtct .....
III-AS
.... intron C ....
G G S P C I S Q P C L H N G S C Q D S.
47G ..... cccccatcct cctcctgcag [GC GGC TCC CCG TGC ATC TCC CAG CCC TGC CTC CAC AAC GGC TCT TGC CAG GAC AGC
GSP-3 IV-S IV-AS
66I I W G Y T C T C S P G Y E G S N C E L
ATC TGG GGC TAC ACC TGC ACC TGC TCC CCC GGC TAT GAG GGC AGC AAC TGC GAG CTG G] gtgaggcccc ggccgtcccc
IV-AS2 GSP-2
.... intron D ....
GSP-1
85A A K N E C H P E R T D G C Q H F C L P
..... ttttttttt ccctttcag [CT AAA AAT GAA TGT CAC CCA GAG CGG ACT GAT GGG TGT CAA CAC TTC TGC CTC CCA
V-S
104G G Q E S Y T C S C A Q G Y R L G E D H K Q C V P H
GGA CAG GAA TCC TAC ACG TGC AGC TGT GCT CAG GGC TAC AGG CTT GGT GAG GAC CAC AAA CAG TGT GTG CCC CAC G]
V-AS V-AS2
gtgagtgctc agaccacagc .... intron E ....
VI-S2
129D D Q C A C G V L T S E K R A P D L Q D.
..... tgttgctgcc gcaaatgcag [AC CAG TGT GCC TGC GGG GTG CTG ACC TCT GAG AAG CGT GCA CCG GAT CTA CAG GAC
VI-AS2 VI-S
148L L P W Q.
CTC CCG TGG CAG] gtaacagagc gctctccgcg .... intron F ....
VI-AS
VII-S2
152V V K L T N S E G K D F C G G V I I R.
..... gtctgttttg atttttaaag [GTA AAG TTA ACA AAT TCC GAA GGA AAA GAC TTC TGT GGT GGT GTT ATA ATA CGG
VII-S
VII-AS2
170E E N F V L T T A K C S L L H R N I T V K T
GAA AAT TTT GTA CTG ACA ACA GCA AAA TGT TCA CTG TTA CAC AGG AAT ATT ACT GTA AAA ACA T] gtaagtattt tatcatgagt
VII-AS
.... intron G ....
191Y Y F N R T S Q D P L M I K I T H V H V
..... acgattgttc atgatttcag [AT TTT AAC AGA ACG AGC CAA GAC CCG CTG ATG ATC AAG ATA ACG CAC GTC CAT GTG
VIII-S
210H H M R Y D A D A G E N D L S L L E L E W P I Q C P G A G
CAC ATG CGG TAT GAC GCG GAC GCG GGG GAG AAT GAC CTG TCA CTG CTG GAG CTG GAG TGG CCC ATC CAG TGC CCA GGT GCG GGG
VIII-AS
238L L P V C T P E K D F A E H L L I P R T R G L L S G W A R
CTC CCC GTG TGC ACC CCT GAG AAA GAC TTC GCT GAG CAC CTC CTC ATC CCA CGC ACC AGG GGC CTC CTC AGC GGC TGG GCA CGC
266N N G T D L G N S L T T R P V T L V E G E E C G Q V L N V
AAT GGC ACT GAC CTG GGC AAC TCG CTG ACC ACG CGG CCT GTC ACA CTT GTG GAG GGG GAG GAG TGC GGG CAG GTC CTG AAT GTG
294T T V T T R T Y C E R S S V A A M H W M D G S V V T R E H
ACT GTC ACC ACC AGG ACC TAC TGT GAG AGA AGC AGC GTG GCG GCC ATG CAC TGG ATG GAT GGA AGT GTG GTC ACC AGA GAA CAC
322R R G S W F L T G V L G S Q P V G G Q A H M V L V T K V S
AGA GGC TCC TGG TTT CTC ACG GGG GTC CTG GGC TCG CAG CCA GTA GGA GGG CAG GCT CAC ATG GTC CTT GTC ACC AAG GTC TCC
350R R Y S L W F K Q I M N STP
AGG TAC TCA CTC TGG TTT AAA CAG ATC ATG AAC TAA CTGAAACTCA GCTAGCCAGA ATGAACAACA CAACCGGAAG CGGGATTCCA
AGCTGGCACT GCCACTGTGG AGGGCGCTGA AACTTCATCA CACACTGAGA GGCCGTCACA GCCCCAGACC ACCCGCTTGG CCCACGCAGC AGCAGAGCCG
CCGTTTGCTG GGTTGTTTAC CGAGCACTGT GACCTTTCTT TCCCTGGAAC TCTTTATCTC AATAGAGACC TTAAAAGAAA ACATGAGATA CGTTAAATAA
VIII-AS2 VIII-S2
TAAAATAAGA TAATCTGTCA GTCATA] aagcagcctg gtttccaaaa attctttcct tctccaagtc cagtgttccc tgtgctcagg gaata.....
GT cluster

FIGURE 2: Nucleotide sequence of the 5′ and 3′ flanking regions, the exons, and the intron/exon boundaries of the gene coding for human protein Z. Nucleotides in the exons are shown in capitals and those in the introns and the 5′ and 3′ flanking regions are shown in lowercase. The DNA sequences upstream from the A in the codon for the initiator Met listed as +1 are shown in the left margin with negative numbers. The amino acids in the prepro-leader sequence are also shown with negative numbers in the left margin, while those in the mature protein are shown with positive numbers, starting with the N-terminal Ala residue as +1. The 5′ and 3′ ends of each exon are enclosed in brackets. The sequences used for the preparation of amplifying primers are underlined or overlined and labeled with their names at their 5′ ends. Putative regulatory elements in the 5′ flanking region and consensus sequences around the polyadenylation site are also underlined (R stands for "reverse"). The sequences used for the preparation of first strand cDNA and amplifying primers in 5′RACE are also underlined or overlined and labeled with their names at their 5′ ends. An apparent major transcription start site is indicated by an asterisk, and minor ones are indicated by a plus sign.

for 60 min. The cDNA was used in the TdT-tailing, and PCR was then performed with a gene-specific internal antisense primer (GSP-2 in Figure 2) and abridged anchor primer (5′-GGCCACGCGTCGACTAGTACGGGIIGGGII-GGGIIG-3′; I stands for inosine), followed by PCR using the nested gene-specific antisense primers (III-AS and GSP-3 in Figure 2) and the nested adapter primer, abridged universal amplification primer (5′-GGCCACGCGTCGACTAGTAC-3′). PCR products were analyzed by 2% agarose gel electrophoresis and by the dideoxy sequencing method after subcloning into pBluescript vectors (Invitrogen).

Chromosomal Assignment of the Gene for Human Protein Z by PCR Screening. Two DNA panels of human/hamster somatic cell hybrids (PCRable panels 1 and 2, lot I-4, BIOS Lab, New Haven, CT) were used to localize the gene coding for human protein Z. PCR was performed under stringent conditions (94 °C for 30 s, 64−66 °C for 60 s, 72 °C for 60 s) for 30 cycles, employing two pairs of PCR primers, Ib-S & II-AS; V−S & VI-AS2 (Figure 2), synthesized using the cDNA sequence for human protein Z (*4*). Amplified DNA was applied to a 0.8% agarose gel and electrophoresed. Buffer and water were also added as negative controls without genomic DNA. The identity of PCR products was confirmed by sequencing analysis.

Localization of the Gene for Human Protein Z by FISH. Genomic phage λLF1 (Figure 1, middle), obtained by screening a human fibroblast library with the cDNA for human protein Z, was used to localize the gene by FISH, as described previously (*15*). Briefly, the entire phage DNA was labeled by nick translation with biotin-16-dUTP (Boehringer, Mannheim, Germany). The biotin-labeled DNA was hybridized to normal metaphase chromosomes from phytohemagglutinin (PHA)-stimulated lymphocytes synchronized with bromodeoxyuridine and then incubated with fluorescein-conjugated avidin (Vector Laboratories, Burlingame, CA). Chromosomes were counterstained with propidium iodide and 4,6-diaminido-2-phenylindole and then photographed. Subsequently, the chromosomes were R-banded using the fluorochrome-photolysis technique.

RESULTS AND DISCUSSION

Genomic Southern Blotting. Two *Eco*RI fragments of 19 and 7.5 kb from human genomic DNA hybridized with a cDNA probe for protein Z by Southern blot analysis (Figure 3, left). When an amplified DNA containing exons Ib,[2] II, and III, and introns A2 and B (PCR7 in Figure 1, top) was employed as a probe, only the 7.5-kb band hybridized (data not shown). These data suggested that the 5′ part of the gene for protein Z was present in the 7.5-kb *Eco*RI fragment and that the remaining 3′ part resided in the 19-kb fragment. Since there is only one internal *Eco*RI site in the cDNA (*4*), it is very likely that the size of the gene for protein Z is less than 26.5 (= 19 + 7.5) kb.

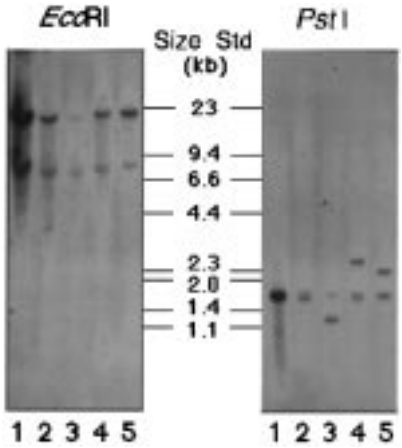

FIGURE 3: Southern blot hybridization of genomic DNAs with a radiolabeled cDNA for human protein Z (left) or an amplified DNA fragment coding for exons II−IV (right). Each lane contains genomic DNA from five normal individuals (1−5) digested with either *Eco*RI (left) or *Pst*I (right) endonuclease. Size markers (Size Std) were purchased from GIBCO-BRL.

Southern blotting analysis employing an amplified DNA fragment coding from exons II−IV and introns B and C (PCR9 in Figure 1, top) as a probe revealed that two *Pst*I fragments hybridized (Figure 3, right); the band of 1.7 kb was observed in common among all five individuals. When the PCR7 fragment was employed as a probe, only this 1.7-kb band hybridized (data not shown). Therefore, the common *Pst*I fragment of 1.7 kb contains exons II and III, and exon IV must be present on the remaining *Pst*I fragment (Figure 1, bottom). It was of note that the size of the second *Pst*I fragment differed (about 1.3−2.5 kb) from one to another among the 5 individuals, indicating the presence of extensive polymorphism around exon IV of the protein Z gene. This hypothesis is discussed later.

In Vitro Amplification of Genomic DNAs. To determine the gene organization of human protein Z, leukocyte DNA obtained from normal individuals was amplified by PCR employing specific primers for putative exons (Figure 2). Altogether, nine overlapping DNA fragments (PCR1−9) from the 5′ end of the gene and two fragments (PCR10 and 11) from the 3′ end were amplified (Figure 1, top). These fragments covered approximately 7 and 2 kb of genomic DNA from the 5′ and 3′ portions of the gene, respectively. DNA sequence analysis of these fragments revealed that the 5′ portion contained the genomic sequence coding from the signal peptide to the hinge region between the second EGF and the pseudo-serine protease domains (including exons

[2] This extra exon Ib encoded a 66-bp insert which was found in half of the amplified products by PCR when the HepG2 cDNA library was examined in the previous study.

Table 1: Nucleotide Sequence and Type of the Splice Junctions and Size of Exons[a]

Region	Size(bp)	Exon	Boundary sequence	Intron	Boundary sequence	Exon	type[b]
5'NC+Signal	(88[d])	Ia	..GAGCCCTCAG GTAGGC...	A1	...CACCAG CCACTTCACT..	Ib	I
Extra	66	Ib	..CAGGAAAGCT GCAAGT...	A2	...CTGTAG TATTTCTCCC..	II	I
Pro+Gla	164	II	..AGTAGTCACT GTATGT...	B	...CTAAAG GATGAATTCT..	III	0
ThrSR	25	III	..CGATATAAGG GTAAGT...	C	...CTGCAG GCGGCTCCCC..	IV	I
1st EGF	114	IV	..TGCGAGCTGG GTGAGG...	D	...TTTCAG CTAAAAATGA..	V	I
2nd EGF	132	V	..GTGCCCCACG GTGAGT...	E	...ATGCAG ACCAGTGTGC..	VI	I
Hinge	68	VI	..CCCGTGGCAG GTAACA...	F	...TTAAAG GTAAAGTTAA..	VII	0
AAHC Loop	118	VII	..GTAAAAACAT GTAAGT...	G	...TTTCAG ATTTTAACAG..	VIII	I
C-Term+3'NC	788	VIII					
		CONSENSUS[c]	C A AAG GTGAGT-		TT T -CCNCAG GT		

[a] NC, noncoding; Pro, propeptide; Gla, γ-carboxyglutamic acid; ThrSR, thrombin-sensitive region; EGF, epidermal growth factor; Term, terminus. The exact size of exon Ia (shown in parentheses) is not known. An unusual nucleotide C of an extra intron A2 at the obligatory GT sequence is underlined. [b] Sharp (1981). [c] Senapathy et al. (1990). [d] 5′-end of the longest cDNA (Ichinose, unpublished data).

I−VI), and the 3′ portion contained the genomic sequence coding for the pseudo-serine protease domain (exons VII and VIII). These results strongly suggested that the gene for protein Z was coded by nine exons and interrupted by eight introns.

A region corresponding to putative intron F was never obtained despite repeated attempts at PCR under varying conditions. Since the maximum size of the entire gene was estimated to be 26.5 kb and the sum of the 5′ and 3′ portions was 9 (= 7 + 2) kb, the size of intron F could be up to 17.5 kb, which appeared to be beyond the limit of the PCR technique available at the time these experiments were performed. Accordingly, we carried out conventional screening of the genomic phage library in the following experiments.

Isolation of Genomic Clones for Human Protein Z. Two million recombinant phage were screened under stringent conditions with labeled cDNA probes. Two positive clones were identified and plaque purified (termed λLF1 and λLF2 in Figure 1, middle). Restriction mapping, Southern blotting, and nucleotide sequencing analysis of amplified DNAs from λLF1 and λLF2 by PCR revealed that these clones contained portions coding for protein Z from the propeptide (exon II) to the 3′ noncoding region (exon VIII), and the portions encoding from the signal peptide (exon Ia) to the hinge region (exon VI) that is present between the second EGF and the pseudo-serine protease domain (Figure 1, middle and Table 1), respectively. These results indicated that the two isolated clones covered the entire gene for protein Z.

Further characterization of the two isolated phage clones and their subcloned plasmids by restriction mapping, Southern blotting, PCR, and nucleotide sequencing analysis revealed that the size of intron F was 5.8 kb. Accordingly, it was concluded that the gene for human protein Z spanned 14 kb.

It was of note that the two genomic clones each contained a *Sac*I fragment which differs in size from that of the other clone by 0.4 kb (S1-2 and S2-4 in Figure 1, bottom). When the S1-2 and S2-4 fragments were digested with *Pst*I, unique bands of 1.7 and 1.3 kb were generated, respectively (data not shown). These *Pst*I fragments corresponded exactly to the 1.3-kb band and one band of the 1.7-kb doublet in Figure 3. These results were confirmed by a PCR−RFLP analysis of the two genomic clones employing amplified intron C (PCR5 in Figure 1, top) and *Pst*I. Accordingly, there is a 0.4 kb insertion in intron C of the λLF1 clone. Since Southern blotting analysis also demonstrated *Pst*I bands of about 2.5 and 2.2 kb (Figure 3, right), the size of this insertion may vary among individuals. Alternatively, one or more adjacent *Pst*I sites could be absent.

Nucleotide Sequence of the Human Protein Z Gene. The nucleotide sequence for each exon and its boundaries determined from the isolated clones was consistent with that obtained from the amplified DNA. The DNA sequences of the gene for human protein Z obtained by these experiments are summarized in Figure 2. The sequence of the nine exons was in complete agreement with that of the cDNA for protein Z (*4*). Exon Ia and a part of exon II encoded the preproleader sequence (Figure 4) that is removed during posttranslational processing by signal peptidase and furin to produce the mature protein Z. The remaining part of exon II encoded the Gla domain, and exons III, IV, V, and VI coded for the thrombin-sensitive region (ThrSR), the first EGF, the second EGF, and the hinge region, respectively. The pseudo-serine protease domain was encoded by exons VII and VIII (Figure 2 and Table 1).

Nucleotide sequence analysis also revealed that all splice donor and acceptor sequences are consistent with the GT-AG rule (*16*) and the consensus sequence (*17*), except for one at the junction between the extra exon Ib and intron A2 (GC rather than GT; Table 1). However, this unusual dinucleotide has also been observed in a few other genes including that for human prothrombin at the 5′ end of intron L (*18*). The difference in the nucleotide sequence between the unusual C and typical T leads to a difference in the consensus values for splicing (*17*): 0.578 vs 0.752. Thus, the GC dinucleotide is likely less favorable than GT for recognition by the splicing machinery. This hypothesis is consistent with the fact that genes containing the GC sequence at the 5′ end of introns splice normally, but the efficiency of splicing decreases. This may result in a much smaller amount of the mRNA containing an extra exon and a greater amount of the mRNA lacking it (*19*).

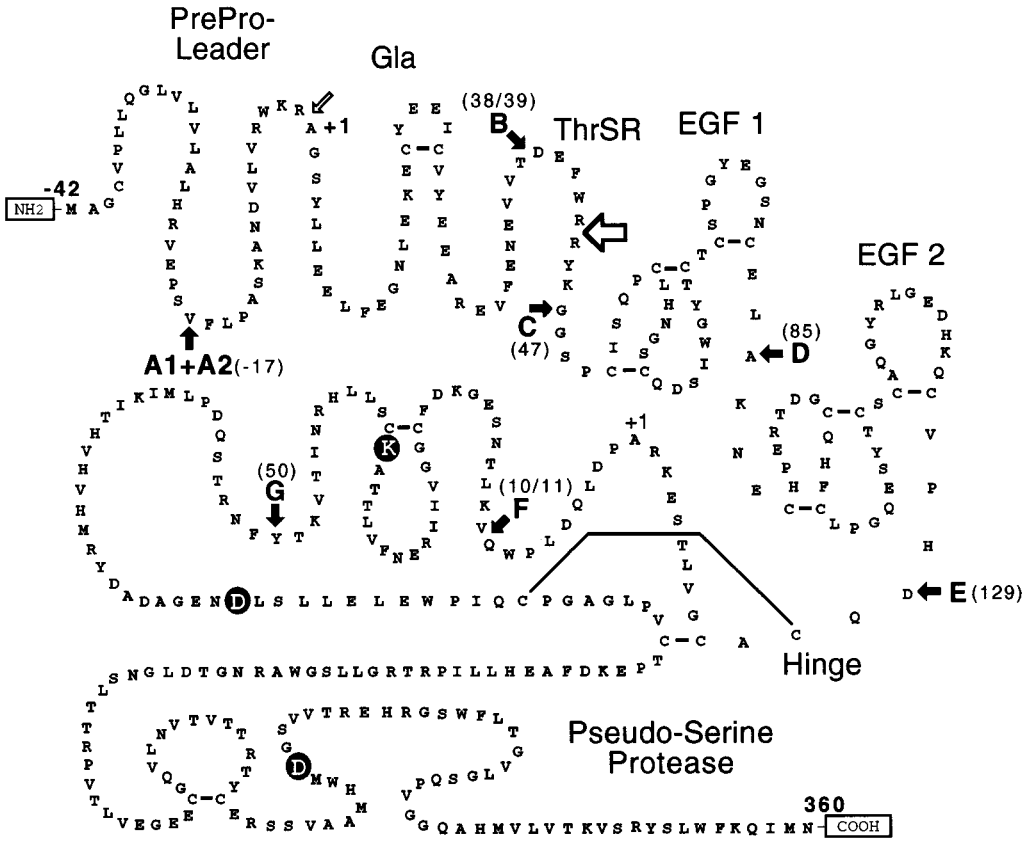

FIGURE 4: Location of introns in the structure of human protein Z. The positions of the introns (A−F) are indicated by solid arrows at or between specific amino acids with numbers in parentheses. Numbering starts with the N-terminus Ala residue and resumes at Ala142. Small and wide open arrows depict the sites of cleavage by furin and thrombin, respectively. Residues at positions for the active sites His, Asp, and Ser for serine-proteases are highlighted, although two of these are replaced by other amino acids. Gla, γ-carboxyglutamic acid; ThrSR, thrombin-sensitive region; EGF, epidermal growth factor.

It is worth noting that extra exon Ib (66 bp) inserts 22 amino acids between the signal peptide important for secretion and the propeptide essential for recognition by γ-carboxylase. The insertion of 66 nucleotides leads to a change of Val to Leu at position −17 (Figures 2 and 5). This position is also where 66 nucleotides coding for 22 amino acids are inserted in a cDNA coding for human factor VII obtained from the same HepG2 library (*20*) as the cDNA for protein Z (*4*). The 66-bp insert in the factor VII cDNA is also coded by an extra exon in intron A of its gene (*21*). Thus, the nucleotide sequence coding for the 22 amino acids in both protein Z and factor VII has resulted from an alternative splicing of the precursor mRNA. No alternatively spliced product was detected in the HepG2 library by extensive PCR for the human prothrombin gene (Ichinose, unpublished data).

A similar insertion (termed the α-helical region) had been reported in the genes for bone Gla protein (BGP) and matrix Gla protein (MGP) (*22, 23*). Although the inserted region of protein Z shares no homology with other proteins, that of factor VII shows some amino acid sequence identity with the α-helical region of BGP, but not with that of MGP (Figure 5, bottom). To date, it is unknown how this region functions in any of these proteins. It is interesting that recombinant factor VII containing the insertion is reported to be as active as the wild type (*19*) and that mature MGP retains this region since its N-terminus is Tyr20 (*24*; Figure 5, bottom). Thus, the inserted peptide seems not to inhibit γ-carboxylation of these proteins.

As described above, all discrete regions (or domains) in protein Z were coded by separate exons (Figures 2 and 4), and introns were inserted at positions precisely identical to those in the gene coding for human factor VII (*21*). Furthermore, the organization of the genes for human protein Z, in terms of the locations of introns and the types of splice junctions (*25*), was essentially identical to that for the genes coding for other vitamin K-dependent proteins, such as factors IX and X, protein C, 5 introns of protein S, three introns of prothrombin, etc. (*11*; Table 2). In particular, the exons in the 5′ portion of the genes in the vitamin K-protein family encode highly homologous protein regions, e.g., the signal peptide, the propeptide, the Gla domain, and the EGF domains. Thus, homologies between the N-terminal half of protein Z and the other family members including factors IX, X, and VII, prothrombin, and proteins C and S are strikingly high, 45%, 47%, 39%, 23%, 38%, and 26%, respectively (*4*). These findings support the hypothesis that these genes derived from a common ancestor through duplication.

In contrast to the 5′ portion, the placement of introns with respect to the serine protease domain in the 3′ portion of the genes is somewhat diverse. For instance, the genes for proteins Z and C and factors VII, IX, and X have only two exons for the serine protease domain, while this domain is encoded by four exons in the prothrombin gene (*18*). The latter situation has also been observed in a number of other genes which contain the serine protease domain, such as the genes for plasminogen and apolipoprotein(a) (*26, 27*).

Table 2: Comparison of Intron Location, Splice Junction Type, and Size of the Genes for Vitamin K-Dependent Plasma Proteins[a]

intron	gene	location (amino acid)	junction type	size (kb)	intron	gene	location (amino acid)	junction type	size (kb)
A	protein Z	−17	I	1.3[b]	D	protein Z	85	I	1.5
	factor VII	−17	I	2.6[b]		factor VII	84	I	1.9
	factor IX	−17	I	6.2		factor IX	85	I	7.2
	factor X	−17	I	5.0		factor X	84	I	1.8
	protein C	−19	I	1.3		protein C	92	I	0.1
	prothrombin	−17	I	0.4		protein S	75	I	4.4
	protein S	−16	I	?					
					E	protein Z	129	I	0.4
B	protein Z	38/39	0	0.8		factor VII	131	I	1.0
	factor VII	37/38	0	1.9		factor IX	128	I	2.6
	factor IX	38/39	0	0.2		factor X	128	I	2.9
	factor X	37/38	0	7.4		protein C	137	I	2.7
	protein C	37/38	0	1.5		protein S	116	I	0.1
	prothrombin	37/38	0	0.7					
	protein S	37/38	0	?	F	protein Z	151/152 (10/11)[c]	0	5.8
						factor VII	167/168 (15/16)	0	0.6
C	protein Z	47	I	1.7		factor IX	195/196 (15/16)	0	9.5
	factor VII	46	I	0.1		factor X	209/210 (15/16)	0	3.4
	factor IX	47	I	3.7		protein C	184/185 (15/16)	0	0.9
	factor X	46	I	1.0					
	protein C	46	I	0.1	G	protein Z	191 (50)[c]	I	1.0
	prothrombin	46	I	0.2		factor VII	209 (57)	I	0.8
	protein S	46	I	>10		factor IX	234 (54)	I	0.7
						factor X	249 (55)	I	1.7
						protein C	224 (55)	I	1.1

[a] Only common parts among all members of the gene family are included. [b] Exon Ib and its 5′- and 3′-introns are treated as a single intron A. [c] Numbering restarts at Ala142 as +1 in protein Z, and at the N-termini of the B chain in other proteins.

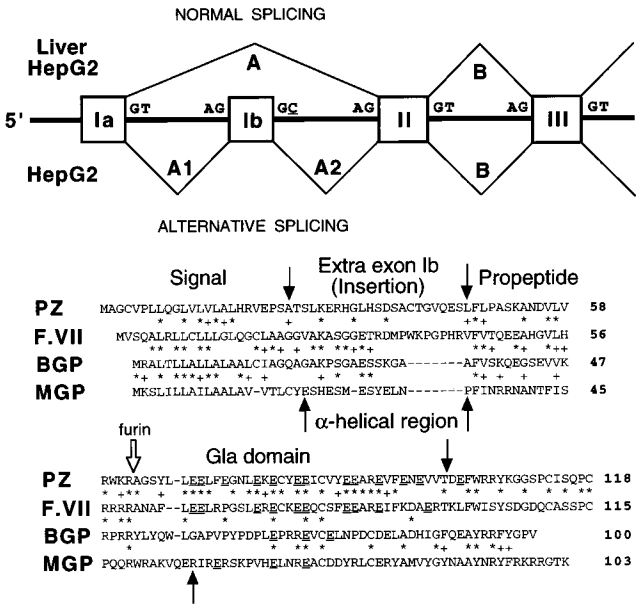

FIGURE 5: Alternative splicing of intron Ib in the genes for human protein Z and factor VII. (Top) Exon Ib and introns A1 and A2 may be spliced out together, resulting in exon Ib skipping. (Bottom) A short stretch of amino acids is inserted in the prepro-leader sequence of human protein Z and factor VII. (*) Symbols: identical, (+) similar amino acid, (-) gap. Numbering is relative to the initiator Met residue and is shown in the right margin. Gla residues are underlined (13 in protein Z, 10 in factor VII, 3 in BGP, and 4 in MGP). Arrows above the amino acid sequences indicate the location of introns in the genes for protein Z and factor VII, while those below the sequences demonstrate the positions of introns in the genes for BGP and MGP.

Differences in the number of introns may be attributable to insertions and deletions during evolution. Homologies between the C-terminus of protein Z and the other family members are lower than those of the N-terminus, 25%, 24%, 22%, 16%, and 26% for factors IX, X, and VII, prothrombin, and protein C, respectively (*4*). Since the sizes of 3′ exons are somewhat diverse when compared with those of 5′ exons, mature proteins of this gene family differ slightly in their overall sizes. For example, exon VI coding for the hinge region of protein Z is smaller than those of factors VII, IX, and X and protein C, which results in its having a shorter mature polypeptide chain than that found in other proteins. Overall homology of protein Z is highest to factors IX, X, and VII among the members of this gene family.

The striking degree of similarity in length and sequence between the exons of the genes encoding vitamin K-dependent proteins is in contrast to the lack of resemblance not only between the nucleotide sequences but in the sizes of the introns of these genes as well (Table 2). Accordingly, there is no correlation between the gene size and the size of the synthesized peptide chain. At present, it remains unknown why most of the intron sequences have not been maintained through evolutionary selection.

5′ and 3′ Flanking Regions and Transcription Initiation Site. The 5′ flanking region of the gene for protein Z contained a cluster of putative regulatory elements, including a TATAA-like sequence, a GC box (SP-1 site), and a reverse CAAT box around a region 300 bp upstream from the initiator Met codon (Figure 2, underlined). A sequence homologous to the HNF-4 element (TGAACT/CTTGCC) was present 47 bp upstream from the Met codon; this element has also been found in the genes for factors X, IX, and VII (*28*). At present, it is not known whether either of these sequences functions as a promoter/enhancer element.

Rapid amplification of the 5′-cDNA ends was performed to identify the transcription start site of the gene for protein Z. A single major band and several minor bands were observed in both HepG2 and Huh7 cells (data not shown); no PCR product was visible in HLF and Chang liver cells.

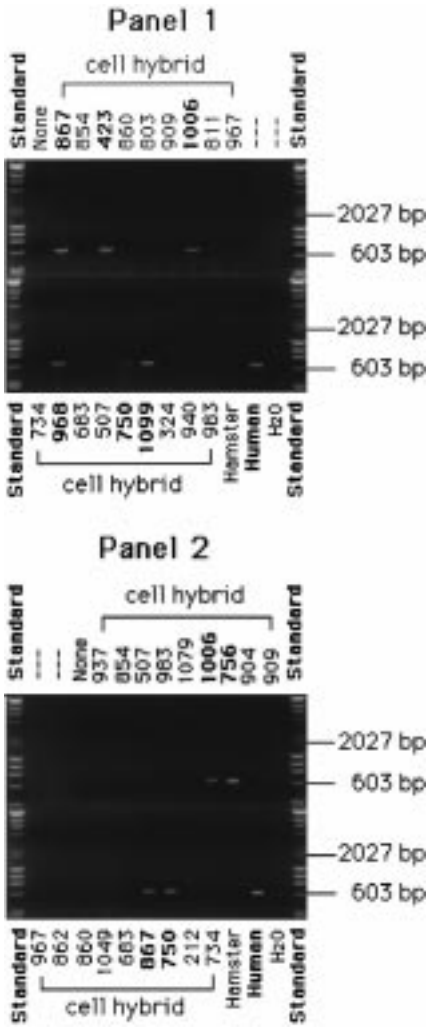

FIGURE 6: Chromosomal assignment of the gene for human protein Z. The PCR amplified products of genomic DNA samples from human, hamster, and human−hamster hybrid cell lines were electrophoresed in a 0.8% agarose gel. The lanes labeled None and H_2O are the negative controls without added DNA (buffer or water). Panels 1 and 2 contain different combinations of genomic DNA samples obtained from various cell lines, as indicated by the numbers. Cell lines 423, 750, 756, 867, 968, 1006, and 1099 contain chromosome 13 (shown in boldface).

Sequencing of the resulting transcripts revealed an apparent major transcription initiation site starting at an A nucleotide located 13 nucleotides upstream from the translation initiation codon (position −13, Figure 2), in 11 subclones. In addition, minor transcription initiation sites started at the C of position −18 and the C of position −68 in HepG2, and at the T of position −28, the T of position −34, and the C of position −68 in Huh7, in two or three subclones for each. It is unusual that the major transcription start site is located at only 13 nucleotides upstream of the initiator codon. Since the 5′RACE method tends to underestimate amounts of larger fragments but not those of smaller ones, the apparent minor start sites upstream from this major site may be utilized much more. The C at −68 bp is a good candidate since it is adjacent to the putative HNF-4 site. These results will likely be confirmed in a future study on the gene regulation of protein Z.

The 3′ end of the gene contained a polyadenylation signal, AATAAA, and a T(A) polyadenylation site. The consensus sequence of YGTGTTYY (GT cluster), which is required for efficient formation of the 3′ terminus of mRNA, was present 64 bp downstream from the AATAAA sequence. No other sequences, such as CAYTG, that commonly surround transcription termination sites were observed.

FIGURE 7: Fluorescence in situ hybridization (FISH) for the gene for human protein Z. Partial metaphase spreads were hybridized with the genomic clone λLF1 coding for human protein Z (top). The same metaphase was counterstained with propidium iodide and 4,6-diaminido-2-phenylindole (bottom). Specific fluorescent signals and the locus of 13q34 are indicated by arrows.

Localization of the Gene for Human Protein Z to 13q34. A PCR product of the expected size (0.6 kb) was generated using a pair of primers (V-S anf VI-AS, Figure 2) when human DNA and those of hybrid cell lines containing chromosome 13 (867, 423, 1006, 968, 750, and 1099) were used as a template, while no bands were observed with hamster DNA or those of other cell lines (Figure 6, panel 1). The same band was obtained only from the human DNA and those of the 1006, 756, 867, and 750 cell lines, all of which contained chromosome 13 (Figure 6, panel 2). A hybrid cell line (423) did not contain chromosome 13, since no band was detected when two other lots (107, 145) were examined later (data not shown). Nucleotide sequence analysis justified the finding that this DNA fragment contained exons Ib and II of the protein Z gene. These results were confirmed by similar experiments employing another set of primers (Ib-S and II-AS, Figure 2), which produced a 1-kb fragment for exons V and VI (data not shown). Thus, the gene for protein Z was localized to chromosome 13 with 0% discordance (0 out of 25 cell lines). The genes for factors VII and X were also localized to chromosome 13 by PCR employing their gene-specific primers (data not shown).

FISH analysis was next carried out using the genomic clone λLF1 coding for the protein Z gene (Figure 1, middle), and the results are shown in Figure 7. In 26 metaphases examined after FISH, chromosome 13 showed specific signals on both chromatids in a total of 25 cells (Figure 7, top). Thus, the protein Z locus was mapped to 13q34. No other bands hybridized with the protein Z probe. These results were consistent with those obtained by PCR, indicating that the locus of protein Z is 13q34, where three other vitamin K-dependent protein genes, factors VII, X, and GAS6, have been mapped (*29*−*31*; Table 3). These genes may have evolved via duplication of an ancestral gene at this locus. In contrast, the loci of the genes for the vitamin

Table 3: Comparison of the Genes for Vitamin K-Dependent Plasma Proteins[a]

gene	size (kb)	exons	introns	chromosome	authors
protein Z	14	8[b]	7[b]	13q34	present study
factor VII	13	8[b]	7[b]	13q34	Scambler and Williamson (1985)
factor X	27	8	7	13q34	Royle et al. (1986)
factor IX	34	8	7	Xq27	Boyd et al. (1984)
prothrombin	21	14	13	11p11−q12	Royle et al. (1987)
protein C	11	8[c]	7	2q14−21	Kato et al. (1988)
protein S	>80	15	14	3p11.1−q11.2	Watkins et al. (1988)
GAS6	?	?	?	13q34	Saccone et al.(1995)
BGP	1.2	4	3	1q	Johnson et al. (1991)
MGP	3.9	4	3	12p	Cancela et al. (1990)

[a] The genes for GAS6 (growth arrest-specific gene-6), BGP (bone Gla protein or osteocalcin), and MGP (matrix Gla protein) are also included. [b] An extra exon (Ib) and its 5′- and 3′-introns (A1 and A2) are treated as a single intron (A). [c] An extra exon for the 5′ noncoding region that is present in the protein C gene is not included.

K-dependent proteins other than these four are found randomly dispersed across chromosomes (Table 3). The gene for protein Z may exist adjacent to those for factors VII and X, which are separated by only 2.8 kb (*32*).

The results obtained here will directly contribute to future studies, including characterizing the genetic defects of protein Z deficiency and investigating regulation of the protein Z gene, plasma concentrations of which vary widely among individuals. The RFLP of *Pst*I around exon IV of the protein Z gene will be useful also for determination of haplotypes.

ACKNOWLEDGMENT

The first and second authors contributed equally to the completion of this study. The authors are indebted to Mr. J. Harris for synthesis of the oligonucleotides, Mr. E. Espling for his excellent assistance, Drs. T. Hashiguchi, T. Izumi, and T. Saito for their helpful discussion, and Ms. L. Boba for her help in preparation of the manuscript. This research was presented in part at the XIIIth ISTH meetings in Amsterdam in June, 1991.

REFERENCES

1. Prowse, C. V., and Esnouf, M. P. (1977) *Biochem. Soc. Trans. 5*, 255−256.
2. Broze, G. J., and Miletich, J. P. (1984) *J. Clin. Invest. 73*, 933−938.
3. Miletich, J. P., and Broze, G. J. (1987) *Blood 69*, 1580−1586.
4. Ichinose, A., Takeya, H., Espling, E., Iwanaga, S., Kisiel, W., and Davie, E. W. (1990) *Biochem. Biophys. Res. Commun. 172*, 1139−1144.
5. Hojrup, P., Jensen, M. S., and Petersen, T. E. (1985) *FEBS Lett. 184*, 333−338.
6. Sejima, H., Hayashi, T., Deyashiki, Y., Nishioka, J., and Suzuki, K. (1990) *Biochm. Biophys. Res. Commun. 171*, 661−666.
7. Hogg, P. J., and Stenflo, J. (1991) *J. Biol. Chem. 266*, 10953−10958.
8. Hogg, P. J., and Stenflo, J. (1991) *Biochem. Biophys. Res. Commun. 178*, 801−807.
9. Kemkes-Matthes, B., and Matthes, K. J. (1995) *Thromb. Res. 79*, 49−55.
10. Gross, J., Pindur, G., Morsdorf, S., Wagner, B., Seyfert, U. T., and Wenzel, E. (1997) *Thromb. Haemostasis*, Suppl., 640.
11. Ichinose, A., and Davie, E. W. (1994) in *Hemostasis and Thrombosis: Basic Principles and Clinical Practice* (Colman, R. W., Hirsh, J., Marder, V. J., and Salzman, E. W., Eds.) 3rd ed., pp 19−54, J. B. Lippincott Co., Philadelphia.
12. Bell, G. I., Karam, J. H., and Rutter, W. J. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 5759−5763.
13. Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., and Erlich, H. A (1988) *Science 239*, 487−491.
14. Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 5463−5467.
15. Taniwaki, M., Matsuda, F., Jauch, A., Nishida, K., Takashima, T., Tagawa, S., Sugiyama, H., Misawa, S., Abe, T., and Kashima K. (1994) *Blood 83*, 2962−2969,.
16. Mount, S. M. (1982) *Nucleic Acids Res. 10*, 459−472.
17. Senapathy, P., Shapiro, M. B., and Harris, N. L. (1990) *Methods Enzymol. 183*, 252−278.
18. Degen, S. J., and Davie, E. W. (1987) *Biochemistry 26*, 6165−6177.
19. Berkner, K., Busby, S., Davie, E., Hart, C., Insley, M., Kisiel, W., Kumar, A., Murray, M., O'Hara, P., Woodbury, R., and Hagen, F. (1986) *Cold Spring Harbor Symp. Quant. Biol. 51*, 531−541.
20. 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., and Davie, E. W. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 2412−2416.
21. O'Hara, P. J., Grant, F. J., Haldeman, B. A., Gray, C. L., Insley, M. Y., Hagen, F. S., and Murray, M. J. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 5158−5162.
22. Celeste, A. J., Rosen, V., Buecker, J. L., Kriz, R., Wang, E. A., and Wozney, J. M. (1986) *EMBO J. 5*, 1885−1890.
23. Cancela, L., Hsieh, C. L., Francke, U., and Price, P. A. (1990) *J. Biol. Chem. 265*, 15040−15048.
24. Pan, L. C., and Price, P. A. (1986) *Proc. Natl. Acad. Sci. U.S.A. 82*, 6109−6113.
25. Sharp, P. A (1981) *Cell 23*, 643−646.
26. Petersen, T. E., Martzen, M. R., Ichinose, A., and Davie, E. W. (1990) *J. Biol. Chem. 265*, 6104−6111.
27. Ichinose, A. (1995) *Biochem. Biophys. Res. Commun. 209*, 365−371.
28. Greenberg, D., Miao, C. H., Ho, W.-T., Chung, D. W., and Davie, E. W. (1995) *Proc. Natl. Acad. Sci. U.S.A. 92*, 12347−12351.
29. de Grouchy, J., Dautzenberg, M. D., Turleau, C., Beguin, S., and Chavin-Colin, F. (1984) *Hum. Genet. 66*, 230−233.
30. Royle, N. J., Fung, M. R., MacGillivray, R. T., and Hamerton, J. L. (1986) *Cytogenet. Cell Genet. 41*, 185−188.
31. Saccone, S., Marcandalli, P., Gostissa, M., Schneider, C., and Della Valle, G. (1995) *Genomics 30*, 129−131.
32. Miao, C. H., Leytus, S. P., Chung, D. W., and Davie, E. W. (1992) *J. Biol. Chem. 267*, 7395−7401.

BI972002A