Intron–Exon Organization of the Active Human Protein S Gene PSα and Its Pseudogene PSβ: Duplication and Silencing during Primate Evolution†,‡

Hans K. Ploos van Amstel,* Pieter H. Reitsma, C. Paul E. van der Logt, and Rogier M. Bertina

Haemostasis and Thrombosis Research Unit, University Hospital, 2300 RC Leiden, The Netherlands

Received January 30, 1990; Revised Manuscript Received May 2, 1990

ABSTRACT: The human protein S locus on chromosome 3 consists of two protein S genes, PSα and PSβ. Here we report the cloning and characterization of both genes. Fifteen exons of the PSα gene were identified that together code for protein S mRNA as derived from the reported protein S cDNAs. Analysis by primer extension of liver protein S mRNA, however, reveals the presence of two mRNA forms that differ in the length of their 5′-noncoding region. Both transcripts contain a 5′-noncoding region longer than found in the protein S cDNAs. The two products may arise from alternative splicing of an additional intron in this region or from the usage of two start sites for transcription. The intron–exon organization of the PSα gene fully supports the hypothesis that the protein S gene is the product of an evolutional assembling process in which gene modules coding for structural/functional protein units also found in other coagulation proteins have been put upstream of the ancestral gene of a steroid hormone binding protein. The PSβ gene is identified as a pseudogene. It contains a large variety of detrimental aberrations, viz., the absence of exon I, a splice site mutation, three stop codons, and a frame shift mutation. Overall, the two genes PSα and PSβ show between their exonic sequences 96.5% homology. Southern analysis of primate DNA showed that the duplication of the ancestral protein S gene has occurred after the branching of the orangutan from the African apes. A nonsense mutation that is present in the pseudogene of man also could be identified in one of the two protein S genes of both chimpanzee and gorilla. This implicates that silencing of one of the two protein S genes must have taken place before the divergence of the three African apes.

Protein S is a vitamin K dependent glycoprotein (DiScipio et al., 1977) that is synthesized in liver cells (Fair & Marlar, 1986), in endothelial cells (Stern et al., 1986), and in megakaryocytes (Ogura et al., 1987). The protein has a molecular mass of 70 kilodaltons and circulates in the blood as a single-chain molecule both free and bound to the C4b binding protein, a component of the complement system (Dahlbäck & Stenflo, 1981). Free protein S serves as a cofactor of activated protein C (Walker, 1980). It enhances the activity of this serine protease in the proteolytical degradation of the procoagulant cofactors Va (Walker, 1980) and VIIIa (Gardiner et al., 1984). The cofactor activity of protein S is inhibited when protein S is bound to the C4b binding protein (Dählback, 1986). The physiological significance of protein C and protein S is illustrated by the observation that a hereditary deficiency of protein C or protein S is a risk factor for the development of thrombotic disease (Giffin et al., 1981; Broekmans et al., 1983; Comp et al., 1984; Engesser et al., 1987).

The complete primary structure of bovine protein S has been established from protein and cDNA sequencing (Dahlbäck et al., 1986a) and that of human protein S from cDNA sequencing alone (Lundwall et al., 1986; Hoskins et al., 1987; Ploos van Amstel et al., 1987a). Human protein S is synthesized as a single-chain molecule of 676 amino acids. The first 87 amino acids are characteristic for all vitamin K dependent coagulation proteins (Furie & Furie, 1988). In analogy with these other proteins, it can be divided into a hydrophobic signal peptide of 25 amino acids (responsible for transport across the endoplasmic reticulum), a 16 amino acid long propeptide [the recognition site for the γ-carboxylase enzyme complex (Jorgensen et al., 1987)], and a vitamin K dependent region (Gla region) that contains the 11 glutamic acid residues that are γ-carboxylated (DiScipio & Davie, 1979; Dahlbäck, 1986). The Gla region is terminated by a short hydrophobic region containing four aromatic residues (so-called aromatic stack or H region). Before protein S is secreted from the cell, the pre- and propeptides are cleaved off. The vitamin K dependent domain is followed by a region that is sensitive to thrombin cleavage (Dahlbäck et al., 1986b). Among the vitamin K dependent proteins, this structure is unique for protein S. Thrombin cleavage destroys the cofactor activity of protein S (Walker, 1984).

The thrombin-sensitive region is connected to four so-called epidermal growth factor like (EGF-like) domains which are typical for a subset of coagulation proteins (Rees et al., 1988). The next region of protein S is, however, atypical for the vitamin K dependent coagulation proteins since it does not contain a catalytic domain. In the nonenzymatic cofactor protein S, the entire 380 amino acid long carboxy-terminal region shares significant homology with plasma steroid hormone binding proteins (Gershagen et al., 1987; Baker et al., 1987).

The gene structures of the vitamin K dependent proteins factor VII (O'Hara et al., 1987), factor IX (Yoshitake et al., 1985), factor X (Leytus et al., 1986), prothrombin (Friezner-Degen & Davie, 1987), and protein C (Foster et al., 1985) have been resolved. It has been shown that the domain structure of these proteins is reflected in the intron–exon organization of these genes. Each exon seems to code for a functional/structural unit of the protein.

† This work was supported by a grant from the Trombose Stichting Nederland (86.015).

‡ The nucleic acid sequence in this paper has been submitted to GenBank under Accession Number J02918.

* Correspondence should be addressed to this author at the Haemostasis and Thrombosis Research Unit, University Hospital, Building 1, C2-R, P.O. Box 9600, 2300 RC Leiden, The Netherlands.

0006-2960/90/0429-7853$02.50/0

The human genome has been shown to contain two protein S genes (Ploos van Amstel et al., 1987b) that both are located on chromosome 3 (Ploos van Amstel et al., 1987b) near the centromere (Watkins et al., 1988). From both genes, the exon corresponding to the 3′-untranslated region of the protein S mRNA has been cloned and sequenced (Ploos van Amstel et al., 1988). The two genes, PSα and PSβ, share a high degree ($\sim$97%) of homology in this region. Only the PSα gene seems transcriptionally active. In this paper, we report the cloning and intron–exon organization of the two protein S genes. It is shown that of the PSα gene 15 exons can be identified that code for the protein S mRNA as represented by the reported cDNAs. As has been found for other plasma proteins, the domain structure of human protein S is reflected in the intron–exon organization of the PSα gene. The PSβ gene is a genuine pseudogene: the gene lacks the 5′ exon containing the initiation methionine, it has a splice site mutation, several stop codons, and a frame shift mutation. In primate evolution, the duplication of the protein S gene seems to have occurred after the branching of the orangutan. Silencing of one of the two genes must have occurred before the divergence of the African apes.

Materials and Methods

Construction and Screening of Genomic Libraries in Phage λEMBL. Two human genomic libraries were constructed from DNA isolated from peripheral blood leukocytes. DNA was partially digested with either the enzyme *Sau*3A or the enzyme *Eco*RI and ligated in the arms of the phage λEMBL3 and -4, respectively (Promega Biotech, Madison), essentially as described (Maniatis et al., 1982). After in vitro packaging, the recombinant phages were plated on *Escherichia coli* NM539. Each library contained approximately 5×10^5 independent recombinant phages. Screening of the libraries was performed by plaque in situ hybridization with protein S cDNA fragments that together span the 5′-untranslated region, the complete coding region, and the 3′-untranslated region of the cDNA (Ploos van Amstel et al., 1987a). The DNA fragments were radiolabeled by random priming using [α-^{32}P]dCTP (New England Nuclear, Boston). Positive clones were plaque-purified, and DNA was isolated by the plate lysate method (Maniatis et al., 1982).

Characterization of Protein S Genomic Clones. The positive clones were characterized by Southern analysis. *Eco*RI and *Pst*I digests of the recombinants were fractionated by agarose gel electrophoresis in TAE buffer (40 mM Tris–acetic acid/2 mM EDTA), and the DNA fragments were transferred onto a Gene Screen Plus membrane (New England Nuclear). Screening for the exons was performed with exon-specific oligonucleotides (Table I) which were at the 5′ end labeled with T4 polynucleotide kinase (Boehringer Mannheim, Mannheim, FRG) using [γ-^{32}P]dATP (Amersham International, Amersham, U.K.). Hybridization was performed at 42 °C in 6× SSC (900 mM sodium chloride/90 mM sodium citrate), 5× Denhardt's solution [0.1% Ficoll, 0.1% poly(vinylpyrrolidone), and 0.1% BSA], 0.5% SDS (sodium dodecyl sulfate), and 100 μg/mL salmon sperm DNA. The filters were washed 2 times for 10 min at 48 °C with 6× SSC/0.5% SDS.

The nucleotide sequences of the oligonucleotides were derived from the nucleotide sequence of the reported protein S cDNA (Ploos van Amstel et al., 1987a). Oligonucleotides were synthesized on a Cyclone DNA synthesizer (Millipore, Bedford).

DNA Sequencing. The *Pst*I or *Eco*RI fragments of the protein S genomic clones, that hybridized to the exon-specific oligonucleotides, were excised from low or ultra low gelling temperature agarose (Sigma Chemical Co., St. Louis) and subcloned in either phage M13 mp18 or plasmid pUC 18 vectors. The nucleotide sequences of the exons and the intron–exon splice junctions were determined by the dideoxy chain termination reaction (Sanger et al., 1977) using [α-^{35}S]dATP (Amersham International). The sequencing reactions were primed with the exon-specific oligonucleotides. Additional oligonucleotides, the sequences of which were based on the obtained nucleotide sequences, were synthesized and used to prime the reaction in the opposite directions.

Analysis of Primate DNA. High molecular weight DNA was isolated according to established procedures from the peripheral blood leukocytes of African green monkey, Rhesus monkey, orangutan, chimpanzee, man and from the spleen of a gorilla. After digestion with various restriction enzymes, the DNA fragments were separated on 0.8% agarose gels in TAE buffer and blotted onto Gene Screen Plus membranes. Prehybridization and hybridization were performed at 65 °C in 1 M sodium chloride, 1% SDS, 50 mM Tris-HCl, pH 7.5, 100 μg/mL salmon sperm DNA, and 10% dextran sulfate.

The filters were probed with various restriction fragments of the protein S cDNA. Labeling of the fragments was performed by random priming using [α-^{32}P]dCTP. The filters were washed at 65 °C 2 times for 30 min in 2× SSC/1% SDS.

Amplification of Primate DNA. The exon coding for amino acid residues 281–344 of mature protein S was amplified by using the polymerase chain reaction (Saiki et al., 1988) with the primers PS281 (CTCAAAGCTTGGATCTCTCTTTGTCCATTG) and PS344 (AGAACGGATCCCAGACTGCATCAAAGTGGG) located respectively upstream and downstream from this exon (exon X, Table II).

The amplification was performed in 32 cycles, each cycle consisting of a denaturation step at 94 °C for 1 min, an annealing step at 55 °C for 1 min, and a primer extension reaction at 65 °C for 2 min. The reaction mixture contained 500 ng of high molecular weight primate DNA (orangutan, gorilla, chimpanzee, and man), 400 ng of each primer, a 150 mM aliquot of each deoxynucleotide, 100 μg of BSA/mL, 67 mM Tris-HCl, pH 8.8, 6.7 mM $MgCl_2$, 10 mM β-mercaptoethanol, 6.7 μM EDTA, 16.6 mM $(NH_4)_2SO_4$, 10% dimethyl sulfoxide, and 4 units of Taq DNA polymerase (Cetus Corp., Boston) in a final volume of 100 μL. The amplified DNA was purified by ultra low gelling temperature agarose gel electrophoresis. The excised DNA fragments were subjected to restriction analysis and directly sequenced without further purification by the dideoxy chain termination reaction.

RNA Analysis. Total RNA was isolated from a human liver with the LiCl–urea method (Auffray & Rougeon, 1980). Northern analysis was performed on 25 μg of total RNA fractionated by electrophoresis on formaldehyde (6%)–agarose gels (1%) in 12 mM Tris-HCl, pH 7.5, 6 mM sodium acetate, and 0.3 mM EDTA. The RNA was transferred onto a Gene Screen Plus filter and handled according to the instructions of the manufacturer. Prehybridization and hybridization were performed at 60 °C in 1 M sodium chloride, 1% SDS, 10% dextran sulfate, and 100 μg of salmon sperm DNA/mL. The filter was probed with protein S cDNA radiolabeled by random priming. Washing was performed at 60 °C in 2× SSC/1% SDS twice for 30 min.

Primer Extension. The length of the 5′-noncoding region of the protein S mRNA was determined by primer extension (Calzone et al., 1987). Oligonucleotide PS5S (5′TCGGTCTGAGCCGTG) complementary to nucleotides 18–32 of the 5′UT region of the protein S cDNA (Ploos van

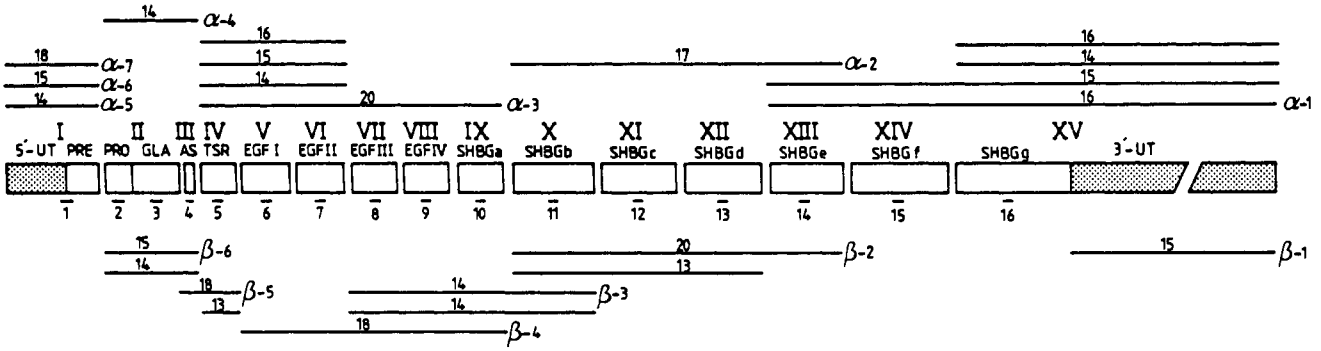

FIGURE 1: Schematic presentation of the exons of the protein S gene. The exons are numbered from I to XV and are drawn according to size. Stippled parts indicate untranslated regions of the mRNA. Oligonucleotides used to identify the exons are numbered 1–16 (see also Table I) and are indicated by dashes below each exon. The 13 genomic PSα clones (varying in length from 14 to 20 kbp) are drawn above the exons that they were found to contain. Clones α-1 to α-7 were used for the determination of the PSα gene organization. The 10 genomic PSβ clones (insert length varying from 13 to 20 kbp) are drawn below the exons that they contain. Clones β-1 to β-6 were used for the determination of the PSβ gene organization. Abbreviations: 5'-UT, 5'-untranslated region; pre, prepeptide; pro, propeptide; GLA, Gla domain; AS, aromatic stack; TSR, thrombin-sensitive region; EGF, epidermal growth factor; SHBG, sex hormone binding globulin; 3'-UT, 3'-untranslated region.

Amstel, 1987a) was labeled at the 5′ end with T4 polynucleotide kinase using [γ-^{32}P]dATP. Annealing of 5 ng of radiolabeled PS5S to 25 μg of liver total RNA was performed by heating for 10 min at 70 °C and cooling to room temperature.

The extension reaction was carried out for 60 min at 42 °C in a final volume of 35 μL containing 10 mM dithiothreitol, 1 mM each of the deoxynucleotides, 75 mM KCl, 4 mM sodium pyrophosphate, 35 units of RNasin (Promega, Biotech), 30 units of reverse transcriptase (Promega, Biotech), 50 mM Tris-HCl, pH 8.3, 10 mM $MgCl_2$, and 0.5 mM spermidine.

RESULTS

Isolation of Protein S Genomic Clones and Assignment of the Clones to the PSα and PSβ Genes. Twenty-three positive clones were identified after screening of the two genomic libraries. The positive clones were assigned to the various parts of the protein S gene by screening with exon-specific oligonucleotides (Table I) that were selected on the basis of a best guess of the intron–exon organization. The content of the clones with respect to the exons of the protein S gene is depicted in Figure 1. No attempt was made to assign the genomic clones to either the PSα or the PSβ gene by making detailed restriction maps. An approach was chosen in which the clones were assigned to the PSα or PSβ gene by nucleotide sequence analysis of the exons on sublconed restriction fragments. The PSα gene is defined as the protein S gene whose coding nucleotide sequence is identical with that of the reported liver protein S cDNA. The PSβ gene has been shown to contain 3% of divergence at the nucleotide level with the PSα gene for the 3′-untranslated region (Ploos van Amstel et al., 1988). We therefore assumed a similar degree of divergence in other parts of the PSβ gene with the PSα gene. In one case (exon III), an exon showing no divergence was assigned to the PSβ gene based on the divergence of the other exons that were contained within the same clone (clone λβ-5, λβ-6). The 23 positive clones varied in length from 13 kilobase pairs (kbp) to 20 kbp. As Figure 1 shows, 13 genomic clones could be assigned to the PSα gene, and 10 clones were assigned to the PSβ gene. The 13 PSα gene clones encompass the exons that correspond to the 5′-noncoding region, the complete coding region, and the 3′-untranslated region of the protein S cDNA. They span approximately 80 kbp of the human genome. The 10 PSβ gene clones contain the exons corresponding to exons II to XIII of the PSα gene and the 3′-untranslated region of exon XV.

Table I: Nucleotide Sequence and Position in the Protein S cDNA of the Oligonucleotides Used for the Analysis of the Protein S Genomic Clones[a]

oligonucleotide	nucleotide sequence	position
PS1	CCACCCAGGACCCTCATTTC	121–140(a)[b]
PS2	CTTCCTAACCAGGACTTGTG	221–240(a)
PS3	TCAAAGACCTCCCTG	326–341(a)
PS4	GATTATTTTTATCCAAAATA	358–377
PS5	CTTAGGTCAGGATAAGCATT	436–455(a)
PS6	GATGGAAAAGCTTCTTTTAC	529–548
PS7	AAGCATAACAAAACC	688–702(a)
PS8	TATCTGTAGCCTTGG	810–824(a)
PS9	GAAGAGTTGTGAGGTTGTTT	960–979
PS10	GCAGGGGTTGTTTTATATTA	1042–1060
PS11	CCTCCAGTTGTGATTTTGGA	1228–1247(a)
PS12	AGCAATCCATTTTCCGGC	1368–1385(a)
PS13	CAAGGAGCTTCTGGAATAAA	1495–1514
PS14	CTGGTAACAACACAGTGCCC	1706–1725
PS15	GAAGACCTTCAAAGACAACT	1915–1934
PS16	GGTGTACAGTTGGATCTGGA	2059–2078

[a] Position of the oligonucleotides is according to the protein S cDNA (Ploos van Amstel et al., 1987a). [b] "a" indicates that the nucleotide sequence is antisense.

The exon of the PSβ gene that corresponds to exon I of the PSα gene (coding for the mRNA 5′-noncoding region and the prepeptide) and those exons that correspond to exon XIV and to the protein-coding region of exon XV of the PSα gene were not present in the isolated genomic fragments. From previous hybridization experiments (Ploos van Amstel et al., 1987b), it was evident that exons XIV and XV were at least in part present in the PSβ gene. We therefore made an additional attempt to clone these exons by screening a third genomic library. However, although 5×10^5 recombinants were screened, none contained this region of the PSβ gene.

The possibility of the presence of an exon in the PSβ gene corresponding to exon I in the PSα gene was investigated by Southern analysis of the human genome (Figure 2). We used a *Hin*dIII–*Pst*I genomic DNA fragment containing the 5′-noncoding region and the prepeptide coding region of exon I of clone λα-5. Figure 2 shows that DNA digested with each of the enzymes *Eco*RI, *Hin*dIII, *Pvu*II, *Bgl*II, *Sac*I, and *Bam*HI gave one single hybridizing band. When we probed similar blots with protein S cDNA fragments, located downstream from the prepeptide coding region, invariably a number of enzymes gave hybridizing fragments that differed in length for the PSα and PSβ genes (Ploos van Amstel et al., 1987b). Therefore, the single hybridizing band in each of the lanes of

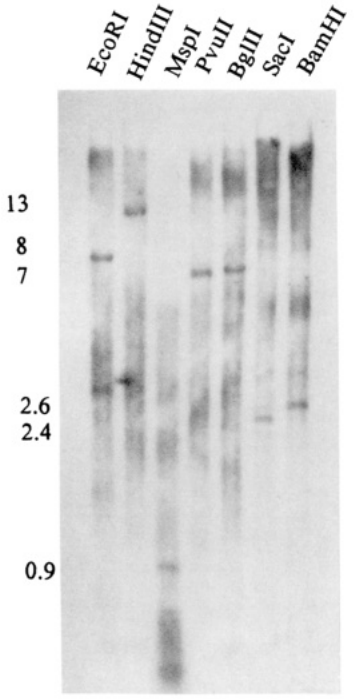

FIGURE 2: Southern blot analysis of the 5′ region of the human protein S genes. DNA from blood leukocytes was digested with the restriction endonucleases *Eco*RI, *Hin*dIII, *Msp*I, *Pvu*II, *Bgl*II, *Sac*I, and *Bam*HI and hybridized with a *Hin*dIII–*Pst*I fragment of clone λPSα-5 encompassing the exon coding for a part of the 5′-noncoding region and the prepeptide of protein S (see Figure 4). The calculated length of the fragments is shown in kilobase pairs.

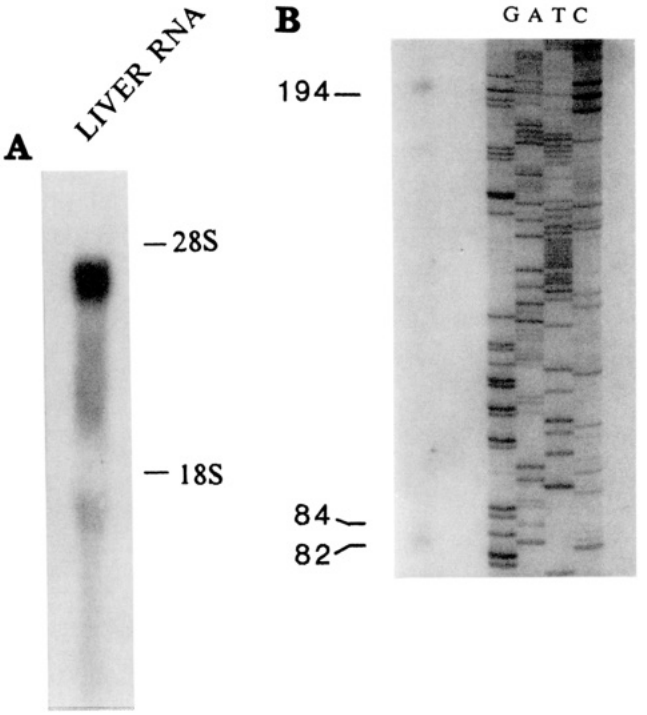

FIGURE 3: Northern analysis and primer extension of liver protein S mRNA. (A) 25 μg of total RNA from a human liver was probed with the coding region of protein S cDNA. The positions of the 28S and 18S RNA are indicated. (B) Primer extension reaction was run along with a sequencing reaction as standard: 25 μg of total RNA was used as template in the extension reaction primed with PS5S. The length of the extension products is indicated.

Figure 2 may be considered as a serious indication that exon I is missing in the PSβ gene.

Organization of the Protein S Genes PSα and PSβ. Nucleotide sequences of the exons and the intron–exon splice junctions of both genes were determined by using exon-specific primers (see Table I). For the PSα gene, the sequences of intron/exon junctions are in agreement with the GT/AG rule and splice site consensus sequences as first formulated by Breathnach and Chambon (1981). The longest reported liver protein S cDNA (Ploos van Amstel et al., 1987a) is encoded by 15 exons (I–XV). The domains of protein S that can be functionally/structurally identified are encoded by separate exons (Figure 1, Table II). The 5′-noncoding region and the propeptide are encoded by a single exon (exon I) as are the propeptide/vitamin K dependent region (exon II), the hydrophobic region (exon III), the thrombin-sensitive region (exon IV), and the four epidermal growth factor domains (exons V–VIII). The region homologous with the steroid hormone binding proteins is encoded by seven exons (exons IX–XV). The phase of the splice junctions of this part of the protein S gene varies from 0 to 2 whereas in the first part of the gene the junctions are predominantly phase 1 (Table II).

The positions of the introns are similar in the PSβ gene. At the splice junctions of the PSβ gene, two significant alterations were detected (Table II). First, the conserved AG dinucleotide at the splice acceptor site (Breathnach & Chambon, 1981) has been changed in AC at the acceptor site of the PSβ exon which corresponds to exon II of the PSα gene. This G → C transversion destroys the consensus sequence for splicing and will prevent correct splicing at this position. The second alteration concerns a deletion of two nucleotides in the codon for Met^{344} in exon X of the PSβ gene, (ATG → ..G), which means that the phase of the splice junction at the donor site of intron J (phase I, Table II) is incompatible with that of the junction at the acceptor site of intron J (phase 0). The consensus for splicing is not necessarily disrupted, but a shift in the reading frame of the PSβ gene is introduced.

Analysis of the 5′ Region of the PSα Gene. The 5′-noncoding region that is present in the reported liver protein S cDNAs (Hoskins et al., 1987; Ploos van Amstel et al., 1987a) and the region coding for the prepeptide are encoded by one single exon (exon I). Analysis of liver RNA, however, suggests that the mRNA for protein S is larger than that represented by the liver protein S cDNAs. Figure 3A depicts the Northern analysis of liver RNA showing a protein S mRNA with a length of approximately 3.5–4 kilobases [also reported by Hoskins et al. (1987)]. To assess the length of the mRNA more precisely, we analyzed the 5′ region of liver protein S mRNA by primer extension (Figure 3B). As primer, we used oligonucleotide PS5S (Figure 4) complementary to nucleotides 18–32 of the reported protein S cDNA (Ploos van Amstel et al., 1987a). Two major extension products are identified with lengths of 194 and 82 nucleotides, respectively. In addition, an extension product of 84 nucleotides can be detected which has a much lower intensity. The two nucleotide length difference between the latter two transcripts might be caused by methylation of the first residues of the mRNA which interferes with the primer extension by reversed transcription (Calzone et al., 1987). The observation of two extension products reveals that two major forms of protein S mRNA exist: (1) a mRNA with a 5′-noncoding region of approximately 286 nucleotides and (2) a mRNA with a 5′-noncoding region of 174 nucleotides. The two putative start positions of the mRNA synthesis are indicated in Figure 4 by asterisks. They both do not fall in a consensus sequence [$(Py)_nA$] that is found for the start site for mRNA synthesis (Breathnach & Chambon, 1981).

Comparison of the Nucleotide Sequence of the Exons of the PSα Gene and the PSβ Gene. A comparison was made between the nucleotide sequence of the exons of the PSα and

Table II: Summary of the PSα and PSβ Exons and Splice Junctions[a]

Gene	Exon (#)	Length (bp)	Splice acceptor		Splice donor	Amino acid interrupted	Intron phase class
					199		
PSα	I	199			ACTgtgagt	Phe (-16)	I
PSβ				NOT CLONED			
				-5 -2 2 26 35	37		
			200	Val Arg Asn Glu Pro	Thr 357		
PSα	II	158	tttccttcagTTT	GTT CGT AAT GAA CCG	ACGgtaagc	Thr (37)/Asp (38)	0
PSβ		158	atttttttcacTTT	TTT TGT AGT GCA CCT	ATGgtaagc	Met (37)/Asp (38)	0
			* ** *	Phe Cys Ser Ala Pro	Met		
			358		V 382		
PSα	III	25	ttctttctagGAT		TAGgtaagt	Val (46)	I
PSβ		25	ttctttctagGAT		TAGgtaagt	Val (46)	I
					V		
			46	49 60 61			
			383 al	Arg Arg Gln	469		
PSα	IV	84	ttattttcagTTT	CGC CGT CAG	ATGgtaagc	Ala (75)	I
PSβ		84	ttattttcagTGT	CTC CAT TAG	ATGgtaagc	Ala (75)	I
			al	Leu His Stop			
				79 87 92 95 103			
			470	Gln Glu Ser Asp Thr	592		
PSα	V	123	tttccttcagCCA	CAG GAA AGC GAT ACT	TTGgtacgt	Asp (116)	I
PSβ		123	tttccttcagCCA	CAA GCA AGT GAA ATT	TTGgtatgt	Asp (116)	I
				Gln Ala Ser Glu Ile	*		
				120 147 157			
			593	Cys Asn Asp	724		
PSα	VI	132	cctgttttagACA	TGC AAT GAT	AAGgtaaga	Asp (160)	I
PSβ		132	cctgttttagACA	CGC AGT GAC	AAGgtaaga	Asp (160)	I
				Arg Ser Asp			
				188			
			725	Pro	850		
PSα	VII	126	ttatttatagATG	CCC	AAGgtagga	Asp (202)	I
PSβ		126	ttatttatagATG	CCT	AAGgtagaa	Asp (202)	I
				Pro	*		
				206 209 225			
			851	Cys Asn Tyr	972		
PSα	VIII	122	tttacctcagATA	TGC AAC TAT	GAGgtaaac	Glu (242)/Val (243)	0
PSβ		122	tttacctcagATA	CGC AAT TAC	GAGgtaaac	Glu (242)/Val (243)	0
				Arg Asn Tyr			
				250 255 262 266 275			
			973	Leu Lys Ala Ala Arg	1088		
PSα	IX	116	tttattccagGTT	TTG AAG GCG GCA CGT	CAGgtgagg	Arg (281)	II
PSβ		116	tttattccagGTT	CTG CAG ACA TCA CAT	CAGgtgagg	Arg (281)	II
				Leu Gln Thr Ser His			
				287 289 296 299 303 304 307 335	344		
			1089	Asp Arg Val Tyr Ile Asp Ala Asp	Met 1278		
PSα	X	190	cattgtttagATT	GAT CGG GTG TAC ATC GAT GCG GAT	ATGgtacgt	Met (344)/Val (345)	0
PSβ		188	cattgtttagATT	CAT TGG ATG TAA GTC AAT GCA GAC	Ggtacgt	?/Val (345)	I/0
				His Trp Met Stop Val Asn Ala Asp			
				355 375 391 393			
			1279	Ser Pro Arg Val	1446		
PSα	XI	168	ttaattgtagGTG	AGC CCG CGG GTG	CCGgtaatg	Pro (400)/Ile (401)	0
PSβ		168	ttaattgtagGTG	AAC CCT CAG GCG	CCGgtaatt	Pro (400)/Ile (401)	0
				Asn Pro Gln Ala	*		
				404 406 410 411 433 448 453 454			
			1447	Arg Asp Arg Ser His Ser His Ile	1615		
PSα	XII	169	aatttgttagATT	CGT GAT CGA AGC CAT TCT CAC ATA	ATAgtaagt	Asn (457)	I
PSβ		169	aatttgttagATT	TGT GAC TGA GGC CAC TTT CGC GTA	ATAgtaagt	Asn (457)	I
				Cys Asp Stop Gly His Phe Arg Val			
				467 474 477 506			
			1616	Val Arg Thr Ser	1767		
PSα	XIII	152	ttttaaatagATA	GTA CGT ACG TCA	CAGgtaact	Gln (507)/Asp (508)	0
PSβ		152	ttttaaatagATA	ATA CAT ATG TCG	CAGgtaact	Gln (507)/Asp (508)	0
				Ile His Met Ser			
			1768		1993		
Psα	XIV	226	tgctcttcagGAT		CAGgtatct	Asp (583)	I
PSβ				NOT CLONED			
			1994				
PSα	XV	1297	ccttttacagATG				
PSβ				NOT CLONED			

[a] Exons are numbered from I–XV, and their length is indicated. Intron sequences are shown in lower case letters and exon sequences in capitals. Differences in intronic sequences between the PSα and PSβ genes are marked by asterisks. Since the codons of the PSα gene fully agree with the cDNA (Ploos van Amstel et al., 1987a), only the codons that are different between the two genes are listed. Numbering of the nucleotide position at the splice junctions follows the cDNA sequence reported by Ploos van Amstel et al. (1987a) as does the numbering of the amino acid residues.

```
                                                    *
AACGTCACACTGTGGAGGAAAAGCAAGCAACTAGGGAGCTGGTGAAGAAGGATGTCTCAGCAGTGTTTACTAGGCCTCCA   80
                          HindIII
ACACTAGAGCCCATCCCCCAGCTCCGAAAAGCTTCCTGGAAATGTCCTTGTTATCACTTCCCCTCTCGGGCTGGGCGCTG  160

* *
GGAGCGGGCGGTCTCCTCCGCCCCCGGCTGTTCCGCCGAGGCTCGCTGGGTCGCTGGCGCCGCCGCGCAGCACGGCTCAG  240

ACCGAGGCGCACAGGCTCGCAGCTCCGCGGCGCCTAGCGCTCCGGTCCCCGCCGCGACGCGCCACCGTCCCTGCCGGCGC  320
PS5S
                  M  R  V  L  G  G  R  C  G  A  L  L  A  C  L  L  L  V  L  P  V
CTCCGCGCGCTTCGAAATGAGGGTCCTGGGTGGGCGCTGCGGGGCGCTGCTGGCGTGTCTCCTCCTAGTGCTTCCCGTCT  400

S  E  A  N
CAGAGGCAAACTGTGAGTAATCAATAGCGTCTCTTCTCCCTTCCCCAGCATTGTCGACTGAACTGCGTCCCTGGTTGGTA  480
                    PstI
GGATTTTCTTCTCTAGAGCTGCAG                                                            504
```

FIGURE 4: Nucleotide sequence of exon I and the region flanking the exon for the 5′-noncoding region and the prepeptide of protein S. The asterisks indicate the hypothetical positions of the termini of the extension products with primer PS5S. The first nucleotide of the 5′-noncoding region of the protein S cDNA (Ploos van Amstel et al., 1987a) is indicated by the arrow. The two potential alternatively spliced acceptor sites of the first intron are underlined. Note that the two sites are located 113 base pairs from each other.

PSβ genes (Table II). The PSβ gene shows at four positions a stop codon. The exon of the PSβ gene corresponding to exon IV of the PSα gene contains a TAG stop codon instead of CAG coding for Glu^{61} of mature protein S. The codon for Tyr^{299} (TAC) in exon X of the PSα gene has been changed in a TAA stop codon in the corresponding exon of the PSβ gene. The codon for Arg^{410} (CGA) in exon XII of the PSα gene has been changed in the PSβ gene in a TGA stop codon. As already mentioned, a two-nucleotide deletion has occurred in the codon corresponding to the codon for Met^{344} in exon X of the PSα gene; this deletion will introduce a shift in the reading frame. Furthermore, the complete region corresponding to exon I of the PSα gene has probably been deleted from the PSβ gene. Each of the aforementioned mutations will prevent the correct expression of the protein Sβ gene. Apart from these nonsense mutations, the PSβ gene shows 33 missense mutations and 15 nonreplacement mutations (Table II) in the protein-coding exons. By use of a correction for potential superimposed substitutions (Ueda et al., 1986), the degree of divergence between the PSα and PSβ genes at the replacement sites is calculated to be 3.4%. The degree of divergence between the two genes at the silent sites of the protein-coding region is 3.9%. The degrees of divergence for the two classes of sites are of the same magnitude as that found for the 3′-untranslated region of both genes (2.9%) (Ploos van Amstel et al., 1988).

The nucleotide sequences of both genes are in agreement with the sequences in the accompanying papers by Edenbrandt et al. (1990) and Schmidel et al. (1990) with the exception of the codon for Leu-250 of the PSβ gene (CTG instead of TTG found by Edenbrandt et al.).

Moment of Duplication. The high degree of homology (~97%) between the two protein S genes indicates that the duplication of the ancestral protein S gene has occurred recently during the evolution of the primates (Britten, 1986; Ueda et al., 1986). The genomes of the African green monkey, rhesus monkey, orangutan, gorilla, and chimpanzee were therefore investigated by Southern analysis for the presence of one or two protein S genes. Genomic DNA of these primates was digested with several restriction enzymes and sequentially probed with adjoining nonoverlapping fragments of the 3′-UT region of the human protein S cDNA (Fragments a and b in Figure 5). Figure 5 shows the hybridization pattern for the enzyme *Hin*dIII. Both the gorilla and the chimpanzee show two *Hin*dIII fragments that hybridized with the two probes. The African green monkey, rhesus monkey, and orangutan all show one hybridizing fragment. In analogy with the identification of the two genes for protein S in man (Ploos van Amstel et al., 1987b), we conclude from these results that

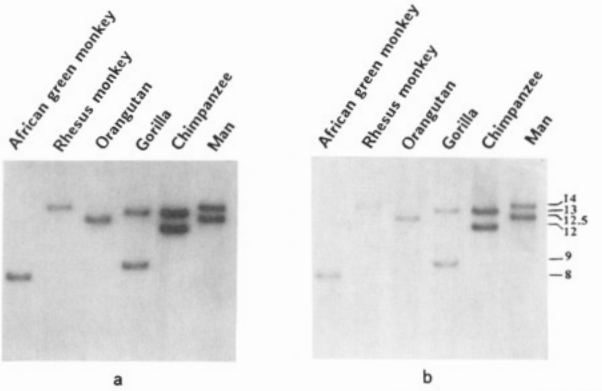

FIGURE 5: Southern analysis of the genome of the primates. *Hin*dIII digests of DNA from African green monkey, rhesus monkey, orangutan, gorilla, chimpanzee, and man are probed with the proximal (a) and distal (b) part of the 3′-untranslated region of the protein S cDNA (probes DS 600 P and DS 400.2, respectively; Ploos van Amstel et al., 1987b). The calculated length of the hybridizing bands is indicated in kilobase pairs.

both the gorilla and chimpanzee have, like man, two protein S genes per haploid genome. African green monkey, rhesus monkey, and orangutan all have only one protein S gene. The duplication of the ancestral protein S gene seems therefore to have occurred after the branching of the orangutan from the African apes, which is estimated to be 13–17 million years ago both by fossil records (Andrews & Cronin, 1982) and by DNA sequencing (Koop et al., 1986).

The extent of divergence between the PSα and PSβ genes of man is of the same magnitude at the replacement sites as at the silent sites. This indicates that no selective pressure has been exerted to prevent the occurrence of mutations at the replacement sites of the PSβ gene, in order to conserve the amino acid sequence of the protein S molecule. We therefore hypothesized that at or shortly after the moment of duplication of the protein S gene one of the two genes has been silenced (this is in man, by definition, PSβ). To investigate whether the gorilla and chimpanzee also contain one active protein S gene and one protein S pseudogene, we amplified exon X of these primates. Exon X, of the PSβ gene of man, contains a TAA stop codon at the position of the codon for Tyr^{299} in the PSα gene. By direct sequence analysis of the amplified DNAs of orangutan, chimpanzee, gorilla, and man (Figure 6), the stop codon at this position is shown to be present in the genome of the chimpanzee, gorilla, and man, but absent in that of the orangutan. The cytidine to adenine transversion in the pseudogenes of the three primates could be confirmed by restriction analysis of the amplified fragment (data not shown) since the transversion destroys the recognition sequence

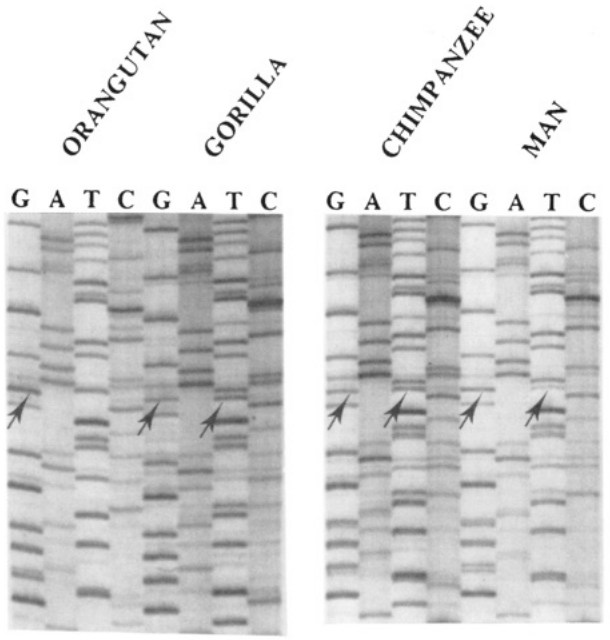

FIGURE 6: Nucleotide sequence of a part of the PCR-amplified exon X of the primates orangutan, gorilla, chimpanzee, and man. The sequence reactions were primed with PS11. The arrows point to the position corresponding to the third nucleotide of the codon for Tyr^{299} that in man is transversed from C/G → A/T in the PSβ gene, thereby changing the TAC for Tyr^{299} to a TAA stop codon.

for the restriction enzyme *Rsa*I (GTAC → GTAA).

These data further confirm the presence of the two protein S genes in the haploid genome of gorilla and chimpanzee and of one gene in the orangutan. The stop codon prohibits expression of the protein S genes in gorilla and chimpanzee, and therefore these apes, like man, have only one active protein S gene.

Discussion

Organization of the Protein S Gene. Fifteen exons have been identified that code for the protein S mRNA as derived from the reported protein S cDNAs (Lundwall et al., 1986; Hoskins et al., 1987; Ploos van Amstel et al., 1987a).

The part of the protein S gene which codes for the 5′-noncoding region, the propeptide, and the vitamin K dependent domain of protein S is similarly organized as the corresponding regions of the vitamin K dependent serine protease precursors factors VII, IX, and X, protein C, and prothrombin. The introns of the genes of these six vitamin K dependent proteins are of the same phase and located at identical positions in such a way that the exon organization corresponds to the domain structure of these proteins. The prepeptide that is involved in the transport of the protein across the endoplasmic reticulum is encoded by a separate exon (exon I). The propeptide that contains the recognition site for the enzyme carboxylase and the Gla domain that contains the glutamic acid residues, that form the substrate for the carboxylase, are organized in a single exon (exon II). The short stretch of aromatic residues located next to the Gla domain is encoded by a separate exon (exon III) in the genes for all six vitamin K dependent proteins. Protein S has a region sensitive to thrombin cleavage following the aromatic stack. Among the coagulation proteins, this region is unique for protein S and encoded by a separate exon (exon IV) flanked by phase 1 introns. The four EGF domains of protein S are each encoded by a single exon (exons V–VIII). Phase and position of the introns that separate these exons are identical with those occurring in the genes for factors VII, IX, and X and protein C that each contain two EGF domains. Prothrombin lacks these EGF structures but instead contains two so-called Kringle domains at the corresponding position. A phase 0 intron separates the exon coding for the fourth EGF domain of protein S from the exons coding for the region homologous with steroid hormone binding proteins (Baker et al., 1987; Gershagen et al., 1987). The vitamin K dependent serine proteases show at this position the exons for the activation peptide and the serine protease domain. The SHBG region of protein S is encoded by seven exons (exons IV–XV).

The homology extends from amino acid 274 of mature protein S to the carboxy-terminal amino acid (Gershagen et al., 1987). Recently, the organization of the rat androgen binding protein gene (Joseph et al., 1988) and of the human sex hormone binding globulin gene (Gershagen et al., 1989) has been elucidated. Both genes show a similar intron–exon organization as their homologous counterpart of the human protein S gene. Position and phases of introns are similar for exons II–VIII of both the ABP gene and the SHBG gene and for exons IX–XV of the protein S gene. We have no insight in the functions of the various parts of the SHBG region of protein S nor of those of the sex hormone binding globulin itself. It therefore remains to be resolved whether the intron–exon structure of this region corresponds to functional domains of this part of the protein. It is interesting to note that the small loops formed by internal disulfide bridges (Cys^{408} to Cys^{434} and Cys^{597} to Cys^{625}, respectively) are each encoded by a single exon. Furthermore, the three potential sites for N-linked glycosylation (Asn^{458}, Asn^{468}, and Asn^{489}) are encoded by a single exon.

All three phases of intron are present in the region coding for the SHBG part of protein S (0, II, 0, 0, I, 0, and I, respectively). It has been hypothesized that nonrandom use of intron phases is a sign of gene assembly by exon shuffling (Patthy, 1987). This process, through which new genes are thought to be assembled, needs modules of the same phase. The occurrence of all three intron phases in the SHBG coding region of the protein S gene seems therefore to suggest that this region has not been assembled by exon shuffling. It seems more likely that the SHBG region as a whole (exons IX–XV) can be considered as one module. The protein S gene can therefore be considered as the product of an evolutional assembling process in which well-known modules (viz., signal peptide, Gla domain, EGF domains) have been put upstream of the ancestral gene of a steroid hormone binding protein.

The 5′ Region of the PSα Gene. Analysis of the 5′ region flanking exon I of the PSα gene revealed no regulatory elements for gene transcription (Mitchell & Tjian, 1989). Therefore, we further investigated the 5′ end of liver protein S mRNA by primer extension. Two major forms of protein S mRNA were identified respectively with a 175- and a 285-nucleotide-long 5′-noncoding region. These two mRNA forms may arise either from the existence of two start sites for mRNA synthesis or from the alternative splicing of an intron in the 5′-noncoding region of the PSα gene. If we simply align the primer extension products with the genomic sequence immediately upstream of the initiator methionine, the resulting two start sites do not adhere to a pyrimidine start site consensus $(Py)_nA$ (Breathnach & Chambon, 1981) nor are these sites flanked by known transcriptional control elements (Mitchell & Tjian, 1989). It is therefore suggestive that an additional intron exists in the 5′-noncoding region. Support for this hypothesis can be derived from the presence of two potential acceptor sites for splicing in this region. Interestingly, these putative acceptor sites are separated by 113 nucleotides which

corresponds to the difference in length of the 2 primer extension products. Therefore, it may well be that protein S mRNA synthesis is started from a single promotor connected to an unidentified 5′ exon and that the intervening sequence is differentially spliced at the indicated acceptor sites. Alternatively, the two transcription products may well be the result of the use of two start sites for mRNA synthesis in combination with an intron in the 5′-noncoding region. It is unfortunate that none of the cDNAs available in our laboratory have a sufficiently long 5′-noncoding region to support one of these hypotheses.

The Protein S Pseudogene PSβ and Its Origin. The human protein S locus on chromosome 3 consists of two protein S genes, PSα and PSβ, that were shown to have a high degree of homology (97%) in the 3′-untranslated region (Ploos van Amstel et al., 1988). The cloning of the two genes reveals that since the duplication of the ancestral protein S gene, the PSβ gene has accumulated several mutations that are incompatible with gene expression both at the transcriptional level and at the translational level. In summary, the PSβ gene lacks the exon I coding for the 5′-noncoding region and the prepeptide; it contains an acceptor splice site mutation in the second exon; furthermore, the gene contains three stop codons and one frame shift mutation, due to a two-nucleotide deletion. Overall, the amino acid coding region of the PSβ gene shows a degree of divergence from that of the PSα gene of approximately 3.5% both at replacement and at nonreplacement sites; this percentage is very similar to what was found for the 3′-untranslated region of the two genes (Ploos van Amstel et al., 1988). For the processed pseudogene $C_{\epsilon 3}$ of the human IgC_ϵ gene family, a divergence between man versus chimpanzee of 1.8% was found and between man and gorilla of 2.6% (Ueda et al., 1986). We therefore hypothesized that duplication of the ancestral protein S gene may have occurred recently during primate evolution. Southern analysis of the genomes of orangutan, gorilla, and chimpanzee revealed that duplication has occurred after the branching of the orangutan from the three African apes. Furthermore, a nonsense mutation found in the human PSβ gene (TAC for Tyr^{299} in PSα and TAA in PSβ) was also identified in one of the protein S genes of both chimpanzee and gorilla. This suggests that one of the protein S genes was already silenced before the divergence of the African apes. From the analysis of fossils, the three-way split of the African apes (6–8 million years ago) (Koop et al., 1986) has been estimated to have occurred 6–8 million years after the branching of the orangutan (Andrews & Cronin, 1982). Within this period, the ancestral protein S gene has been duplicated, and one of the two genes must have been mutated to a pseudogene.

Acknowledgments

We are indebted to R. Belterman and R. Bontrop for providing tissue samples and blood samples of the primates. We thank Mary Mentink for her excellent help in preparing the manuscript. During the preparation of this paper, the relevant sequence data were compared with those of C. M. Edenbrandt and J. Stenflo and with those of D. K. Schmidel, A. V. Tatro, and G. L. Long. We thank them for their helpful comments.

References

Andrews, P., & Cronin, J. (1982) *Nature 297*, 541–546.
Auffray, C., & Rougeon, F. (1980) *Eur. J. Biochem. 107*, 303–314.
Baker, M., French, F. S., & Joseph, D. R. (1987) *Biochem. J. 243*, 293–296.
Breathnach, R., & Chambon, P. (1981) *Annu. Rev. Biochem. 50*, 349–383.
Britten, R. J. (1986) *Science 231*, 1393–1398.
Broekmans, A. W., Veltkamp J. J., & Bertina, R. M. (1983) *N. Engl. J. Med. 309*, 340–344.
Calzone, F. J., Britten, R. J., & Davidson, E. H. (1987) *Methods Enzymol. 152*, 611–632.
Comp, P. C., Nixon, R. R., Cooper, M. R., & Esmon, C. T. (1984) *J. Clin. Invest. 74*, 2082–2088.
Dahlbäck, B. (1986) *J. Biol. Chem. 261*, 12022–12027.
Dahlbäck, B., & Stenflo, J. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 2512–2516.
Dahlbäck, B., Lundwall, A., & Stenflo, J. (1986a) *Proc. Natl. Acad. Sci. U.S.A. 83*, 4199–4203.
Dahlbäck, B., Lundwall, A., & Stenflo, J. (1986b) *J. Biol. Chem. 261*, 5111–5115.
DiScipio, R. G., & Davie, E. W. (1979) *Biochemistry 18*, 899–904.
DiScipio, R. G., Hermodson, M. A., Yates, S. G., & Davie, E. W. (1977) *Biochemistry 16*, 698–706.
Edenbrandt, C.-M., Lundwall, Å., Wydro, R., & Stenflo, J. (1990) *Biochemistry* (third of three papers in this issue).
Engesser, L., Broekmans, A. W., Briët, E., Brommer, E. J. P., & Bertina, R. M. (1987) *Ann. Intern. Med. 106*, 677–682.
Fair, D. S., & Marlar, R. A. (1986) *Blood 67*, 64–70.
Foster, D. J., Yoshitake, S., & Davie, E. W. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 4673–4677.
Friezner-Degen, S. J., & Davie, E. W. (1987) *Biochemistry 26*, 6165–6177.
Furie, B., & Furie, B. C. (1988) *Cell 53*, 505–518.
Gardiner, J. E., Mc Gamm, M. A., Berridge, C. W., Fulcher, C. A., Zimmerman, T. S., & Griffin, J. H. (1984) *Circulation 70*, 205.
Gershagen, S., Fernlund, P., & Lundwall, A. (1987) *FEBS Lett. 220*, 129–135.
Gershagen, S., Lundwall, A., & Fernlund, P. (1989) *Nucleic Acids Res. 17*, 9245–9258.
Griffin, J. H., Evatt, B., Zimmerman, T. S., Kleiss, A. J., & Wideman, C. (1981) *J. Clin. Invest. 68*, 1370–1373.
Hoskins, J., Norman, D. K., Beckmann, R. J., & Long, G. L. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 349–353.
Jorgensen, M. J., Cantor, A. B., Furie, B. C., Brown, C. L., Shoemaker, C. B., & Furie, B. (1987) *Cell 48*, 185–191.
Joseph, D. R., Hall, S. H., Conti, M., & French, F. S. (1988) *Mol. Endocrinol. 2*, 3–13.
Koop, B. F., Goodman, M., Xu, P., Chan, K., & Slightom, J. L. (1986) *Nature 319*, 234–237.
Leytus, S. P., Foster, D. C., Kurachi, K., & Davie, E. W. (1986) *Biochemistry 25*, 5098–5102.
Lundwall, A., Dackowski, W., Cohen, E., Shaffer, M., Mahr, A., Dahlbäck, B., Stenflo, J., & Wydro, R. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 6716–6720.
Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982) *Molecular cloning: A laboratory manual*, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
Mitchell, P. J., & Tjian, R. (1989) *Science 245*, 371–378.
Ogura, M., Tanabe, N., Nishioka, J., Suzuki, K., & Saito, H. (1987) *Blood 70*, 301–306.
O'Hara, P. J., Grant, F. J., Haldeman, 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.
Patthy, L. (1987) *FEBS Lett. 214*, 1–7.

Ploos van Amstel, H. K., Van der Zanden, A. L., Reitsma, P. H., & Bertina, R. M. (1987a) *FEBS Lett. 222*, 186–190.
Ploos van Amstel, H. K., Van der Zanden, A. L., Bakker, E., Reitsma, P. H., & Bertina, R. M. (1987b) *Thromb. Haemostasis 58*, 982–987.
Ploos van Amstel, H. K., Reitsma, P. H., & Bertina, R. M. (1988) *Biochem. Biophys. Res. Commun. 157*, 1033–1038.
Rees, D. J. G., Jones, I. M., Handford, P. A., Walter, S. J., Esnouf, M. P., Smith, K. J., & Brownlee, G. G. (1988) *EMBO J. 7*, 2053–2061.
Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., & Erlich, H. A. (1988) *Science 239*, 487–491.
Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 5463–5467.
Schmidel, D. K., Tatro, A. V., Phelps, L. G., Tomczak, J. A., & Long, G. L. (1990) *Biochemistry* (first of three papers in this issue).
Stern, D., Brett, J., Harris, K., & Nawroth, P. (1986) *J. Cell Biol. 102*, 1971–1978.
Sugo, T., Dahlbäck, B., Holmgren, A., & Stenflo, J. (1986) *J. Biol. Chem. 261*, 5116–5120.
Ueda, S., Watanabe, Y., Hayashida, H., Miyata, T., Matsuda, F., & Honjo, T. (1986) *Cold Spring Harbor Symp. Quant. Biol. 51*, 429–432.
Walker, F. J. (1980) *J. Biol. Chem. 255*, 5521–5524.
Walker, F. J. (1984) *J. Biol. Chem. 259*, 10335–10339.
Watkins, P. C., Eddy, R., Fukushima, Y., Byers, M. G., Cohen, E. H., Dackowski, W. R., Wydro, R. M., & Shows, T. B. (1988) *Blood 71*, 238–241.
Yoshitake, S., Schach, B. G., Foster, D. C., Davie, E. W., & Kurachi, K. (1985) *Biochemistry 24*, 3736–3750.

Molecular Analysis of the Gene for Vitamin K Dependent Protein S and Its Pseudogene. Cloning and Partial Gene Organization†,‡

Carl-Magnus Edenbrandt,§,‖ Åke Lundwall,§ Robert Wydro,⊥ and Johan Stenflo*,§

Department of Clinical Chemistry, University of Lund, Malmö General Hospital, S-214 01 Malmö, Sweden, and Genzyme Corporation, One Mountain Road, Framingham, Massachusetts 01701

Received January 30, 1990; Revised Manuscript Received May 2, 1990

ABSTRACT: Protein S is a vitamin K dependent plasma protein and a cofactor to activated protein C, a serine protease that regulates blood coagulation. The haploid genome contains two protein S genes (α and β) with the protein S α-gene corresponding to the cloned cDNA. We have now isolated and mapped overlapping genomic clones that cover an area of 50 kilobases of the protein S α-gene which code for the 3′ part of the gene, i.e., the thrombin-sensitive region, the four domains that are homologous to the epidermal growth factor (EGF) precursor, the COOH-terminal part of protein S that is homologous to a plasma sex hormone binding globulin (SHBG), and, finally, the 3′ untranslated region. The thrombin-sensitive region and the EGF-like domains are each coded on a separate exon. The sizes of the exons coding for the COOH-terminal half of protein S and the location of the introns are nearly identical with those in the homologous SHBG gene. Furthermore, the phase class of the splice junctions is the same in these two genes. We have also isolated and mapped genomic clones that cover 25 kilobases of the protein S β-gene, which was found to contain stop codons and a 2 bp deletion which introduces a frame shift, suggesting that it is a pseudogene. The structure of the two protein S genes and a comparison with the vitamin K dependent clotting factors support a model for their origin by exon shuffling and recruitment of the 3′ part of the gene from an ancestor shared with the sex hormone binding globulin.

Protein C is a precursor of a serine protease that after activation destroys coagulation factors Va and VIIIa by limited proteolysis (Esmon, 1987, 1989; Stenflo, 1988). Activated protein C requires a cofactor, protein S, for biological activity (Walker 1984; Heeb & Griffin, 1988). The role of protein S as a regulator of blood coagulation in vivo is illustrated by the association of familial protein S deficiency with an increased risk for thromboembolic disease in early adulthood (Comp et al., 1984; Comp & Esmon, 1984; Schwarz et al., 1984; Broekmans et al., 1985; Engesser et al., 1987). The concentration of protein S in blood plasma is approximately 25 mg/L, about half of which is in complex with the complement regulatory protein C4b binding protein (Dahlbäck & Stenflo, 1981; Dahlbäck, 1984). Both amino acid and cDNA sequences for bovine protein S (Dahlbäck et al., 1986) and the cDNA sequence for human protein S (Lundwall et al., 1986; Hoskins et al., 1987; Ploos van Amstel et al., 1987a,b) have been determined. The human protein S molecule consists of 635 amino acids and has an apparent molecular weight of approximately 70 000. It has three potential N-glycosylation sites. Protein S is synthesized by hepatocytes (Fair & Marlar, 1986), vascular endothelial cells (Fair et al., 1986; Stern et

† This work was supported by grants from the Swedish Medical Research Council (Project K-90-13P-08135-04B, B89-13X-08660-01A, and B89-13X-04487-15A), the Swedish Society of Medicine, the Crafoord Foundation, and Lund University.

‡ The nucleic acid sequence in this paper has been submitted to GenBank under Accession Number J02919.

* To whom correspondence should be addressed.
§ University of Lund, Malmö General Hospital.
‖ Present address: Department of Medicine, University of Lund, Lund, Sweden.
⊥ Genzyme Corporation.

0006-2960/90/0429-7861$02.50/0