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

al., 1986), and megakaryocytes (Ogura et al., 1987).

The precursors of the vitamin K dependent clotting factors VII, IX, and X and that of the anticoagulant protein C all have a similar structure and are composed of functional regions, a signal peptide followed by a propeptide, a γ-carboxyglutamic acid (Gla)[1] containing region, two domains that are homologous to the epidermal growth factor (EGF)[1] precursor, and finally a serine protease domain in the COOH-terminal half of the molecule [see Furie and Furie (1988) for a review]. The genes for these proteins also have a similar structure encompassing a mosaic of exons encoding different functional domains (O'Hara et al., 1987; Yoshitake et al., 1985; Leytus et al., 1986; Foster et al., 1985; Plutzky et al., 1986). Single exons encode the signal peptide, the EGF-like domains, and a small connection peptide between the Gla and the EGF-like domains. Furthermore, the propeptide and the Gla-containing region, which constitute a functional unit, are coded on a separate exon. This organization of the genes suggests that they have been assembled by exon shuffling (Gilbert, 1978; Patthy, 1987).

Protein S, like the other vitamin K dependent plasma proteins, has an NH_2-terminal signal peptide/propeptide–Gla region structural motif, but beyond this region, the domain structure of protein S differs from those of the vitamin K dependent serine proteases (Dahlbäck et al., 1986). It has a unique thrombin-sensitive domain that is followed by four consecutive EGF-like domains, instead of two such domains as in the vitamin K dependent serine proteases. The first EGF-like domain contains β-hydroxyaspartic acid and the following three β-hydroxyasparagine (Drakenberg et al., 1983; Stenflo et al., 1987). The COOH-terminal part of protein S has no resemblance to the serine proteases but is homologous to a steroid hormone binding protein in human plasma, called sex hormone binding globulin (SHBG) (Gershagen et al., 1987; Long et al., 1988), and to an androgen binding protein (ABP) of rat testis (Baker et al., 1987). Recently Ploos van Amstel et al. (1988) demonstrated that there are two protein S genes, called α and β, per haploid genome located on chromosome 3 (Ploos van Amstel et al., 1987b; Long et al., 1988; Watkins et al., 1988). Sequence analysis of the 3′ untranslated region of both genes demonstrated 97% homology and assigned the cDNA to the α-gene (Ploos van Amstel et al., 1988).

In order to relate functional units of protein S to the gene structure and the relationship of the 3′ part of the protein S genes to the SHBG gene, we describe the cloning and exon organization of 50 kb from the protein S α-gene that codes for the thrombin-sensitive region, the four EGF-like domains, and the COOH-terminal SHBG-like domain. In addition, we characterize 25 kb of the protein S β-gene, a pseudogene with a nucleotide sequence related to the SHBG-like domain of protein S.

Experimental Procedures

Preparation of Probes. Restriction endonuclease fragments of protein S cDNA were isolated from a clone, M117S, of 3344 bp (Lundwall et al., 1986). The first probe, M117-1, was an *Eco*RI fragment of 2196 bp (nucleotides 1–2196) encoding amino acids −15 to 635 and also containing parts of a 5′ intron. Probes for the 5′ end of the gene were a 140 bp *Mnl*I fragment of M117S (nucleotides 236–376) coding for the Gla region and also a *Hinc*II fragment of 341 bp from the bovine protein S cDNA, BS-2400 (Dahlbäck et al., 1986), corresponding to the signal peptide and the Gla region (nucleotides 1–341). A probe for the middle part of the cDNA was a *Dde*I fragment of 470 bp (nucleotides 1206–1676). Identification of 3′ end fragments was performed with the two *Eco*RI fragments M117S-2 (nucleotides 2190–2954) and M117S-3 (nucleotides 2955–3344). The probes were labeled with [^{32}P]dCTP by using the nick translation method or the random primer extension method (Feinberg & Vogelstein, 1983) according to kit specifications (Amersham, U.K.), to yield specific activities of approximately 10^9 cpm/μg of DNA.

Screening of Genomic DNA Libraries. Three human genomic DNA libraries constructed in EMBL3 (Frischauf et al., 1983), in the cosmid vector Lorist X (Little & Cross, 1985; Cross & Little, 1986), and in Charon 4A were investigated. The Charon 4A library was made from a partial digest of human genomic DNA with *Alu*I/*Hae*III. *Eco*RI linkers were added, and the material was cloned into Charon 4A. The libraries were screened by using standard techniques, positive clones that hybridized upon secondary and tertiary screening were purified and isolated according to the plate lysis or the liquid culture method, and the DNA was extracted as described (Maniatis et al., 1982).

Restriction Endonuclease Mapping of Genomic Clones. The DNA of isolated genomic clones was digested with one or more restriction endonucleases (see Figure 1), and the resulting fragments were separated by electrophoresis in agarose gels and blotted onto nylon membrane filters. The filters were hybridized with cDNA fragments or specific oligonucleotides. The oligodeoxyribonucleotides were synthesized in accordance with the cDNA sequence of human protein S (Lundwall et al., 1986) using an Applied Biosystems Model 380A DNA synthesizer (Foster City, CA). The filters for cDNA hybridization and the filters for oligonucleotide hybridization were prehybridized, hybridized, washed, and autoradiographed according to standard procedures (Maniatis et al., 1982).

Sequencing of Exons and Intron/Exon Junctions. DNA sequencing was performed directly on the isolated recombinant clones with the dideoxy chain termination reactions (Sanger et al., 1977) using reversed transcriptase according to Zagursky et al. (1985) with minor modifications. Sequences were analyzed with available computer programs (Staden, 1982; Orcutt et al., 1984).

Southern Blot Analysis of Genomic DNA. Human genomic DNA was purified from peripheral blood leukocytes of healthy individuals by proteinase K digestion and phenol extraction followed by restriction enzyme digestion, agarose electrophoresis, Southern blotting, hybridization with radiolabeled protein S cDNA, and autoradiography (Bell et al., 1981; Maniatis et al., 1982).

Results

Isolation and Characterization of Human Protein S Genomic Clones. The human protein S cDNA fragment M117S-1, corresponding to the coding region except for the signal peptide, was used as the first probe to screen the three genomic libraries. Among approximately 10^6 recombinants of the EMBL3 library, 5 positive clones were found, and among 10^6 recombinants of the Charon 4A library, 21 positive clones were found. Furthermore, 200 000 colonies of the cosmid library were screened, and 1 positive clone was found. In addition, several clones unrelated to the protein S genes were isolated. The reason for this was suggested by a computer search that showed the postulated 5′ intron of the protein S cDNA (M117) to contain a copy of type O dispersed repetitive DNA (Sun et al., 1984). Restriction enzyme mapping of the

[1] Abbreviations: EGF, epidermal growth factor; Gla, γ-carboxyglutamic acid.

genomic protein S clones showed that they represented two different genes. The isolated clones contained coding sequences representing the thrombin-sensitive region, the four EGF-like domains, the SHBG homology region of protein S, and also the 3′ untranslated end. The libraries were rescreened several times with restriction fragments from the 5′ end of both human and bovine cDNA as well as with synthetic oligonucleotides, but no clone encoding material 5′ to the thrombin-sensitive region was found. Several clones were found to overlap and cover approximately 50 kb of the protein S α-gene as well as overlapping clones covering approximately 25 kb of the protein S β-gene (Figure 1). These clones were chosen for further DNA sequencing of exons and exon/intron junctions and for mapping the distances between exons.

Sequencing of Exons and Exon/Intron Junctions. The nucleotide sequences of the exons and the exon/intron junctions were established by direct sequencing of the λ clones, since in several occasions genetic material was deleted when subcloned into the plasmid vector pUC 18. A set of oligonucleotides (17–22 bases) were synthesized according to the cDNA sequence at 50–100 bp intervals. These were then used both to identify clones that contained the appropriate exon and as primers for sequencing the genomic DNA. The exact location of the exon/intron boundaries was determined by aligning the genomic sequences with the corresponding cDNA sequence. A new set of oligonucleotides was subsequently synthesized in order to get complete sequences of delineated exons and exon/intron boundaries.

The part of the protein S α-gene that has been characterized (Figure 1A) is divided into 12 exons by introns (Table I). The DNA sequences of the splice donor and acceptor sites all follow the GT-AG rule proposed by Breathnach and Chambon (1981) and Mount (1982). The exons vary in size from 87 to 1295 bp. The nucleotide sequences of the exons are identical with the cDNA sequence of Hoskins et al. (1987) and Ploos van Amstel et al. (1987a,b). Compared to the cDNA sequence of Lundwall et al. (1986), there are three differences. Codons for amino acid 180 are CCA not CTA, for 222 TAC not CAC, and for 304 GAT not TAT. The nucleotide sequences of the exons and the splice junctions (Table I) were found to be identical with the sequences reported by Ploos van Amstel et al. (1990) and Schmidel et al. (1990). The distances between exons were estimated by double restriction endonuclease digestion and ranged between 0.1 and >10 kb. The 50 kb of cloned protein S α-gene corresponds to 3 kb of the protein S transcript, and thus approximately 95% of the gene consists of intervening sequences.

The Domain Structure of Protein S Correlates with the Structure of the Gene. We did not find genomic clones containing exons coding for the signal peptide, the Gla region, or the connecting peptide. However, the unique thrombin-sensitive region of protein S and also the four EGF-like domains were found to be coded by separate exons. The introns between these exons are all type I; i.e., they interrupt between the first and the second nucleotides of the codon. The SHBG homology region of the protein S is coded by seven exons and encompasses approximately 30 kb of the protein S gene. The last exon of 1295 bp also includes the 3′ untranslated region. The introns between these exons are of all three possible types, i.e., 0, I, or II.

The Genes for Protein S and SHBG Are Phylogenetically Related. Table II shows a comparison of the size of the exons and the phase of the exon/intron boundaries of the protein S gene with those of the genes coding for the homologous proteins, i.e., the human sex hormone binding globulin

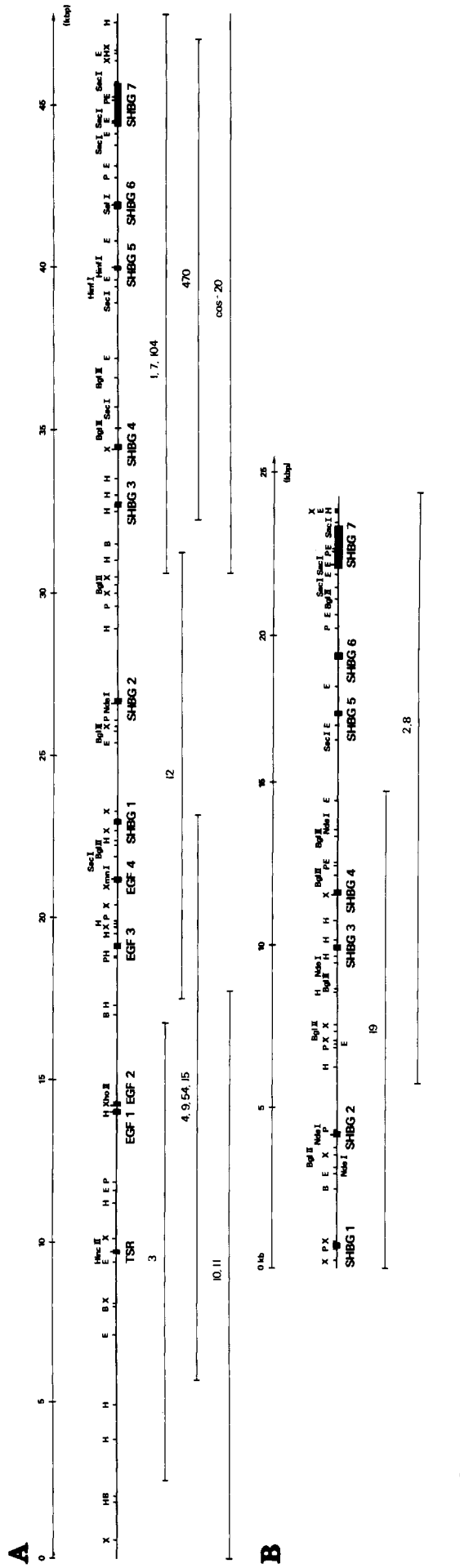


FIGURE 1: Organization of the human protein S genes and schematic representation of the λ and cosmid clones used. All cleavage sites found for restriction enzymes *Eco*RI (E), *Xba*I (X), *Pst*I (P), and *Hin*dIII (H) are indicated. The other enzymes were used only when necessary for precise exon localization. Exons are indicated with filled boxes and correspond to the thrombin-sensitive region (TSR), the EGF-like region (EGF), and the SHBG-like region (SHBG). λ clones are identified by their number in the figure. More than one number denotes identical clones. (A) Protein S α-gene. (B) Protein S β-gene.

Table I: Exon–Intron Splice Junctions of the Human Protein S α-Gene and Protein S β-Gene[a]

Exon		Exon size (bp)	5′splice donor		Intron size (kb)	3′splice acceptor		Intron type	Exon
							46 al		
					> 10	ttttcag	TT	I	TSR
				75			75		
			Asn	A			Ia		
TSR	α	87	AAT	G gtaagc	4.4	cctccag	CC	I	EGF1
				116			116		
			Phe	a			sp		
EGF1	α	123	TTT	G gtacgt	0.1	tttcag	AC	I	EGF2
				160			160		
			Lys	A			sp		
EGF2	α	132	AAA	G gtaaga	4.8	ttatag	AT	I	EGF3
				202			202		
			Glu	A			sp		
EGF3	α	126	GAA	G gtagga	1.8	cctcag	AT	I	EGF4
				242			243		
			Cys	Glu			Val		
	α	122	TGT	GAG gtaaac	2.3	tttcag	GTT	0	
EGF4									SHBG1
	β					tttcag	GTT	0	
				281			281		
			Ser	Ar			g		
	α	116	AGC	AG gtgagg	3.3	gtttag	A	II	
SHBG1									SHBG2
	β	116	AGC	AG gtgagg	3.4	gtttag	A	II	
				344			345		
			Asn	Met			Val		
	α	190	AAT	ATG gtacgt	5	ttgtag	GTG	0	
SHBG2									SHBG3
	β	188	AAT	..G gtacgt	6.8	ttgtag	GTG	I/0	
				400			401		
			Lys	Pro			Ile		
	α	168	AAA	CCG gtaatg	1.5	tgttag	ATT	0	
SHBG3									SHBG4
	β	168	AAA	CCG gtaatt	1.5	tgttag	ATT	0	
				457			457		
			Tyr	A			sn		
	α	169	TAT	A gtaagt	5.3	aaatag	AT	I	
SHBG4									SHBG5
	β	169	TAT	A gtaagt	5.1	aaatag	AT	I	
				507			508		
			Ser	Gln			Asp		
	α	152	TCA	CAG gtaact	1.8	cttcag	GAT	0	
SHBG5									SHBG6
	β	152	TCA	CAG gtaact	1.8	not sequenced		0	
				583			583		
			Pro	A			sp		
	α	226	CCA	G gtatct	2.4	ttacag	AT	I	
SHBG6									SHBG7
	β		not sequenced		2.5	not sequenced			
				635					
			Ser	Stop					
	α	1295	TCT	TAA	- 3′-untranslated - TTTTAAAAtgcatg				
SHBG7									
	β		not sequenced						

[a] Exons correspond to the thrombin-sensitive region (TSR), the epidermal growth factor like domains (EGF1–EGF4), and the sex hormone binding globulin like regions (SHBG1–SHBG7). Type of intron refers to the position of the intron in the codon triplet. Type 0 introns are located between codons, type I introns after the first nucleotide, and type II introns after the second nucleotide of the codon. Exon sequences are in capital letters. The intron sizes were estimated from the gene map (Figure 1). (..) denotes the 2 bp deletion in the protein S β-gene.

(SHBG) (Gershagen et al., 1989; Hammond et al., 1989) and the rat androgen binding protein (ABP) (Joseph et al., 1988). It is noteworthy that the size of the corresponding exons in the three genes is almost identical. After alignment of the amino acid sequences of the two proteins using a computer program (Orcutt et al., 1984), the introns are found to be in almost identical positions within the amino acid sequences of protein S and SHBG (Figure 2), and furthermore, the corresponding exon–intron boundary phases are identical. A diagonal plot comparison of the amino acid sequences of protein S versus SHBG reveals similarity between the two proteins, but also three stretches of amino acids where the

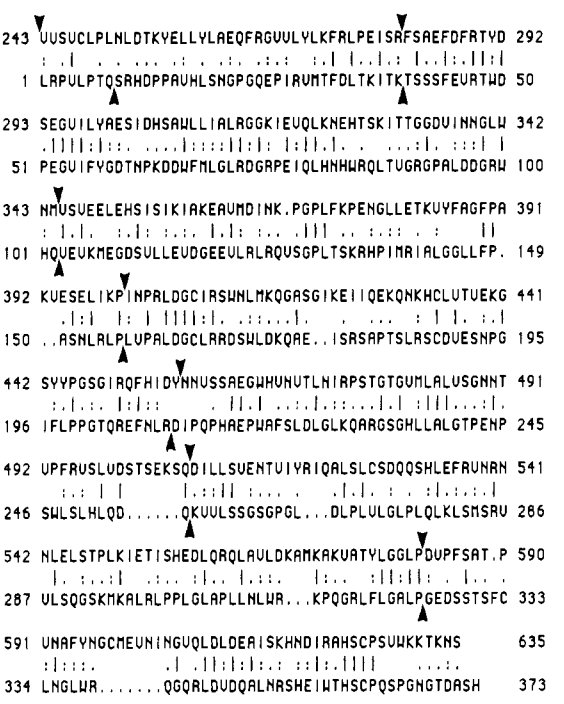


FIGURE 2: Sequences of human SHBG (bottom) and the SHBG homology region of human protein S (top) were aligned at the nucleotide level by the computer program GAP (Orcutt et al., 1984) using a gap weight of 5.00 and a length weight of 0.100. The corresponding amino acid sequences are shown with the intron positions indicated with arrowheads.

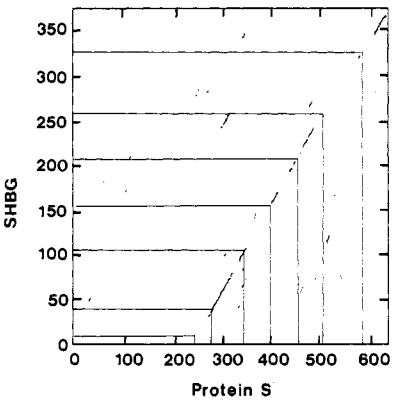


FIGURE 3: Amino acid sequences of human protein S and SHBG were compared by using the DIAGON program (Staden, 1982). The span length was set to 15 and the proportional score to 175. The intron positions are indicated by horizontal and vertical lines.

similarity is less pronounced (Figure 3). When the introns are positioned in the plot, it is obvious that the parts of the sequences that show little similarity are largely confined to three separate exons, i.e., exons 1, 3, and 6 of the SHBG homology region of protein S, corresponding to exons 2, 4, and 7 of SHBG. It is noteworthy that the part coded by exons 4 and 7 in SHBG has been suggested to take part in the steroid hormone binding (Picado-Leonard & Miller, 1988; Petra et al., 1988).

The Protein S β-Gene Is a Pseudogene. Three genomic clones were found to hybridize with the human protein S cDNA and cover almost 25 kb of the protein S β-gene. These clones included all exons of the SHBG homology region. The restriction map of the protein S β-gene is clearly different from that of the protein S α-gene (Figure 1B). Furthermore, when oligonucleotides, synthesized according to the cDNA sequence, were hybridized to the Southern-blotted clones, several oligonucleotides did not hybridize or hybridized only weakly.

Table II: Comparison of Exon Size (bp) and Type of Introns between the SHBG Homology Regions of the Two Protein S Genes and the Human Sex Hormone Binding Globulin (SHBG) Gene (Gershagen et al., (1989) and the Rat Androgen Binding Protein (ABP) Gene (Joseph et al., 1988)[a]

	protein S α-gene		protein S β-gene		SHBG		rat ABP	
exon	exon size	intron type	exon size	intron type	exon size	intron type	exon size	intron type
1	(EGF4)	0	(EGF4)	0	111	0	114	0
2	116	2	116	2	92	2	92	2
3	190	0	188	I/0	190	0	190	0
4	168	0	168	0	162	0	162	0
5	169	1	169	1	160	1	160	1
6	152	0	152	0	137	0	137	0
7	226	1	*b*	*b*	208	1	208	1
8	158/1295*		*b*	*b*	146		146	

[a] To make comparisons between the four genes easier, no frame shift has been introduced in the protein S β-gene after the third exon. The asterisk denotes the exon size including the untranslated 3′ end. [b] Not sequenced.

Those oligonucleotides that did hybridize were used to sequence five out of seven exons. However, the last two exons did not hybridize to any of our oligonucleotides and also gave a weaker hybridization signal to the cDNA probes as compared to the corresponding fragments from the protein S α-gene. When the cDNA was hybridized to the Southern-blotted genomic clones of the pseudogene and the washing stringency was lowered to 1 × SSC, 55 °C, a strong signal was also obtained from the last two exons. This suggests that the last two exons have a considerable mismatch compared to the cDNA sequence and much more than the other exons of the pseudogene (≈15%; Andersson & Young, 1988). The nucleotide sequence of the 5 exons of the protein S β-gene differs from that of the protein S α-gene in 29 positions, and 20 of these base substitutions result in amino acid differences (Table III). Most importantly, the nucleotide changes also give rise to an in-phase stop codon (from TAC to TAA) in the second exon. The 3′ end of this exon also has a deletion of two nucleotides compared to the protein S α-gene. This alteration changes the splice junction from phase 0 to phase I and introduces a frame shift in the following exons resulting in nine stop codons in exon 3, three in exon 4, and three in exon 5. The multiple stop codons and the deletion of two bases as compared to the protein S α-gene suggest that the protein S β-gene is incapable of producing a functioning protein product and is a pseudogene. However, the pseudogene is organized similarly to the active α-gene, and all exon/intron boundaries of the pseudogene that have been analyzed obey the GT-AG rule (Table I). The nucleotide sequence corresponding to the amino acid in position 250 was, in our clone, identical with the cDNA sequence, in contrast to what was found by Ploos van Amstel et al. (1990) and Schmidel et al. (1990).

Southern Blot Analysis of Genomic DNA. Figure 4 shows a Southern blot of human genomic DNA hybridized with the cDNA probe M117-1. All restriction fragments found in our maps of the two protein S genes are also found in the blot. We also hybridized the blot with the 3′ third of the probe (an *Xba*I fragment of nucleotides 1504–2196) including the SHBG-like exons 4–7. We then found six *Eco*RI bands (12, 6.0, 3.4, 2.2, 1.3, and 0.3 kb), two *Xba*I bands (13 and 11 kb), two *Hin*dIII bands (15 and 13 kb), and four *Pst*I bands (12, 8, 6, and 2.4 kb). All bands except the 3.4 kb *Eco*RI fragment are found in the maps. This band was consistently found in genomic blots from several individuals.

Discussion

We have isolated λ and cosmid clones from three human genomic libraries covering approximately 50 kb of the expressed protein S α-gene and approximately 25 kb of the protein S β-gene which appears to be a pseudogene. The part of the protein S α-gene that has been analyzed has a complex organization and consists of 12 exons separated by intervening sequences. DNA sequence analysis of the exons showed three differences compared to the cDNA sequence previously reported by Lundwall et al. (1986), but was identical with the cDNA sequences reported by Hoskins et al. (1987) and Ploos van Amstel et al. (1987a). The introns comprise approximately 95% of the gene and vary in size from 0.1 kb to >10 kb. The exon/intron boundaries were all found to follow the consensus sequences proposed by Breathnach and Chambon (1981) and Mount (1982). The sequences in Tables I and III have been

Table III: Nucleotide Sequence in Five out of Seven Exons in the SHBG-like Region of the Protein S α-Gene and the Protein S β-Gene for comparison[a]

SHBG-like 1.

```
α  ag GTTGTTTCAGTGTGCCTTCCCTTGAACCTTGACACAAAGTATGAATTACTTTACTTGGCGGAG
β  -- ----------------------------------------C-------------------A-A---

α  CAGTTTGCAGGGGTTGTTTTATATTTAAAATTTCGTTTGCCAGAAATCAGCAG gt
β  ------T---------------------------A------------------ --
```

SHBG-like 2.

```
α  ag ATTTTCAGCAGAATTTGATTTCCGGACATATGATTCAGAAGGCGTGATACTGTACGCAGAATC
β  -- ----------------C-----T-------------------A---------A--------

α  TATCGATCACTCAGCGTGGCTCCTGATTGCACTTCGTGGTGGAAAGATTGAAGTTCAGCTTAAGAA
β  -G--A----------A-----------------------------------------------

α  TGAACATACATCCAAAATCACAACTGGAGGTGATGTTATTAATAATGGTCTATGGAATATG gt
β  ---------------------------------C-------------------------- --
```

SHBG-like 3.

```
α  ag GTGTCTGTGGAAGAATTAGAACATAGTATTAGCATTAAAATAGCTAAAGAAGCTGTGATGGAT
β  -- ------------------------------A-------------------------------

α  ATAAATAAACCTGGACCCCTTTTTAAGCCGGAAAATGGATTGCTGGAAACCAAAGTATACTTTGCA
β  ------------------------------T-------------------------------

α  GGATTCCCTCGGAAAGTGGAAAGTGAACTCATTAAACCG gt
β  ----------A---------------------------- --
```

SHBG-like 4.

```
α  ag ATTAACCCTCGTCTAGATGGATGTATACGAAGCTGGAATTTGATGAAGCAAGGAGCTTCTGGA
β  -- ---------T-------C---------T--G-------------------------------

α  ATAAAGGAAATTATTCAAGAAAAACAAAATAAGCATTGCCTGGTTACTGTGGAGAAGGGCTCCTAC
β  ----------------------------------C-----------------------------

α  TATCCTGGTTCTGGAATTGCTCAATTTCACATAGATTATA gt
β  ----------T----------------G-G--------- --
```

SHBG-like 5.

```
α  ag ATAATGTATCCAGTGCTGAGGGTTGGCATGTAAATGTGACCTTGAATATTCGTCCATCCACGG
β  -- ------------------------------A-------------------A--------T--

α  GCACTGGTGTTATGCTTGCCTTGGTTTCTGGTAACAACACAGTGCCCTTTGCTGTGTCCTTGGTGG
β  ------------------------------------------------------------------

α  ACTCCACCTCTGGAAAAATCACAG gt
β  --------------------G--- --
```

[a] (–) denotes identity between the two genes, STOP codons are underlined, and (▼) denotes deletion.

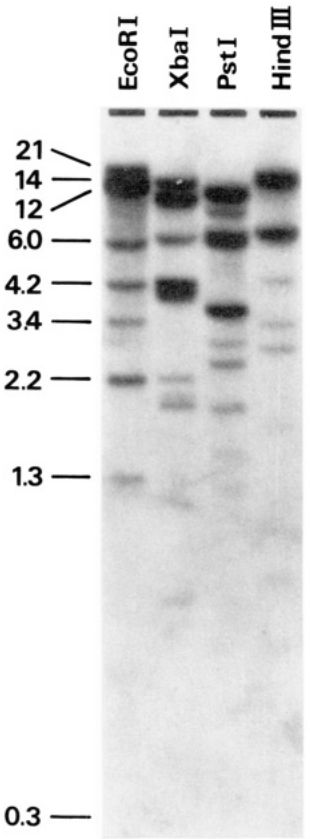


FIGURE 4: Southern blot of human genomic DNA digested with *Eco*RI, *Xba*I, *Pst*I, and *Hin*dIII, hybridized with human protein S cDNA (probe M117-1). The sizes of the *Eco*RI fragments are indicated.

compared to the sequences in the preceding papers (Ploos van Amstel et al., 1990; Schmidel et al., 1990) and were found to be identical except in one instance (see Results).

The structure of the protein S gene demonstrates that the proposed domain structure of protein S correlates with the structure of the gene. It has previously been shown that the positions of introns in the genes of the other vitamin K dependent coagulation factors and protein C divide the coding region that corresponds to the NH_2-terminal half of the molecule into modular units, i.e., the signal peptide, the propeptide, and the Gla region, which form a functional unit, the so-called connecting peptide and the EGF-like domains. In the protein S gene, we found that the four EGF-like domains are coded on separate exons as in the genes coding for the other vitamin K dependent factors (Furie & Furie, 1988) which each has two such domains. The thrombin-sensitive region, which is unique to protein S, is also coded on a separate exon. Patthy (1985) has suggested that the proteins involved in coagulation and fibrinolysis, which share types of presumably independently folded domains, have been assembled by exon shuffling. A prerequisite for exon shuffling is that the introns are of the same type, i.e., interrupt the codon at the same position to avoid a frame shift when exons are inserted or deleted. All of the exon/intron junctions that separate the regions coding for the thrombin-sensitive region and the four EGF-like domains were found to be of type I, i.e., the same type as found between exons in the 5′ half of the genes of the other proteins which contain domains of this type and are involved in blood coagulation and fibrinolysis. Our results thus support the hypothesis that the protein S genes have been assembled by exon shuffling.

In contrast to the vitamin K dependent clotting factors, the carboxy-terminal region of protein S is unrelated to the serine proteases. The amino acid sequence of this part of protein S has been found to be homologous to the sex hormone binding globulin (SHBG) found in plasma (Gershagen et al., 1987) and the rat androgen binding protein (Baker et al., 1987). The SHBG homology region of protein S, SHBG, and rat ABP have a similar size, i.e., 393, 373, and 373 amino acids, respectively. Furthermore, the four half-cysteine residues in SHBG are in positions that are homologous to the four half-cysteine residues in the carboxy-terminal region of protein S that are linked by intrachain disulfide bonds. The other two half-cysteine residues in this part of protein S have no counterpart in SHBG or ABP.

The SHBG-like region of protein S is coded by seven exons. It is noteworthy that the size of the exons in this part of protein S and in SHBG and ABP is almost identical. Furthermore, all three types of exon/intron splice junctions are found, and the phases are identical in the corresponding position in all three proteins. This would suggest that the SHBG homology region of protein S, SHBG, and rat ABP have evolved from a common ancestral gene. This parrt of the protein S gene has presumably been linked to the 5′ part of the protein S gene as one block by a shuffling mechanism that occurred prior to the duplication of the ancestral protein S gene [see also Ploos van Amstel et al. (1990) and Schmidel et al. (1990)].

We also found three clones which covered the SHBG-like region of the protein S β-gene, which is a pseudogene with in-phase stop codons, and a two base pair deletion in the second exon of the SHBG-like region which gives rise to a frame shift and several stop codons in the following exons. The nucleotide sequence difference was 3.6% when compared to the protein S α-gene. The two 3′ exons of the pseudogene hybridized only weakly to the cDNA, and oligonucleotides that were synthesized on the basis of the cDNA sequence did not hybridize. On the basis of hybridization, the difference between these two exons and the cDNA was estimated to about 15%. Ploos van Amstel et al. (1990) have only cloned the untranslated part of the 3′ exon and did not have clones covering the exon immediately upstream. However, Schmidel et al. (1990) have sequenced this part of the protein S β-gene and only found few nucleotide substitutions. The reason for this discrepancy is not clear.

The overall amino acid sequence similarity between SHBG and the homologous region in protein S is about 30% (Gershagen et al., 1987). A graphic representation of the similarity can be obtained when the primary structures are compared in a diagonal plot as in Figure 3. Segments of pronounced and weak sequence similarity are visualized as the presence or the absence of a diagonal line in the plot. If the intron positions are marked in such a plot of SHBG and its homologous region in protein S, it becomes apparent that the amino acids encoded by exons 1, 3, and 6 in protein S (amino acids 243–281, 345–400, and 508–582) and by exons 2, 4, and 7 (amino acids 9–39, 103–156, and 256–324) of SHBG are much less similar than those encoded by the other related exons. In this context, it is noteworthy that exons 4 and 7 in SHBG have been implicated in steroid hormone binding (Petra et al., 1988; Picado-Leonard & Miller, 1988). The weak similarity between SHBG and protein S in this region suggests that SHBG has acquired or protein S lost steroid binding capacity due to extensive mutations in this region. However, to the best of our knowledge, no systematic evaluation of the binding of steroid hormones or similar hydrophobic compounds to protein S has been performed.

Another clue to the function of the SHBG-like region of protein S might come from studies of a second translation

product of the SHBG gene called SHBGgrp (Gershagen et al., 1989). The transcript for this still unknown protein is derived from the SHBG gene by an alternative splicing mechanism which includes the deletion of the seventh exon and translation of the eight exon in another reading frame. The postulated SHBGgrp protein also has an amino-terminal structure that differs from that of SHBG. Since this molecule is supposed to lack sex hormone binding property, it resembles protein S more than SHBG in that respect.

ACKNOWLEDGMENTS

The human genomic DNA libraries were kindly provided by Dr. B. Forget, Yale University, New Haven, CT (Charon 4A), Drs. G. Andersson and L. Rask, The Biochemical Center, Uppsala, Sweden (EMBL3), and Dr. P. F. R. Little, Chester Beatty Laboratories, London, U.K. (Lorist X). The expert technical assistance of Mrs. Monica Jönsson, Mrs. Ann-Marie Thämlitz, and Ms. Ingrid Dahlqvist is gratefully acknowledged. We thank Drs. R. M. Bertina, H. K. Ploos van Amstel, P. H. Reitsma, and G. L. Long for fruitful discussion of their and our data on the protein S gene at a meeting in Leiden in December 1989.

REFERENCES

Anderson, M. L. M., & Young, B. D. (1988) in *Nucleic Acid Hybridization* (Hames, B. D., & Higgins, S. J., Eds.) pp 73–111, IRL Press, Oxford.

Baker, M. E., French, F. S., & Joseph, D. R. (1987) *Biochem. J. 243*, 293–296.

Bell, G. I., Karam, J. H., & Rutter, W. J. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 5759–5763.

Breathnach, R., & Chambon, P. (1981) *Annu. Rev. Biochem. 50*, 349–383.

Broekmans, A. W., Bertina, R. M., Reinalda-Poot, J., Engesser, L., Muller, H. P., Leeuw, J. A., Michiels, J. J., Brommer, E. J. P., & Briet, E. (1985) *Thromb. Haemostasis 53*, 273–277.

Comp, P. C., & Esmon, C. T. (1984) *N. Engl. J. Med. 311*, 1525–1528.

Comp, P. C., Nixon, R. R., Cooper, M. R., & Esmon, C. T. (1984) *J. Clin. Invest. 74*, 2082–2088.

Cross, S. H., & Little P. F. R. (1986) *Gene 49*, 9–22.

Dahlbäck, B. (1984) *Semin. Thromb. Hemostasis 10*, 139–148.

Dahlbäck, B., & Stenflo, J. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 2512–2516.

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

Drakenberg, T., Fernlund, P., Roepstorff, P., & Stenflo, J. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 1802–1806.

Engesser, L., Broekmans, A. W., Briet, E., Brommer, E. J. P., & Bertina, R. M. (1987) *Ann. Intern. Med. 106*, 677–682.

Esmon, C. T. (1987) *Science 235*, 1348–1352.

Esmon, C. T. (1989) *J. Biol. Chem. 264*, 4743–4746.

Fair, D. S., & Marlar, R. A. (1986) *Blood 67*, 64–70.

Fair, D. S., Marlar, R. A., & Levin, E. G. (1986) *Blood 67*, 1168–1171.

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

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

Frischauf, A.-M., Lehrach, H., Poustka, A., & Murray, N. (1983) *J. Mol. Biol. 170*, 827–842.

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

Gershagen, S., Fernlund, P., & Lundwall, Å. (1987) *FEBS Lett. 220*, 129–135.

Gershagen, S., Lundwall, Å., & Fernlund, P. (1989) *Nucleic Acids Res. 17*, 9245–9258.

Gilbert, W. (1978) *Nature 271*, 501.

Hammond, G. L., Underhill, D. A., Rykse, R. M., & Smith, C. L. (1989) *Mol. Endocrinol. 3*, 1869–1876.

Heeb, M. J., & Griffin, J. H. (1988) in *Protein C and Related Proteins* (Bertina, R. M., Ed.) pp 55–70, Churchill Livingstone, New York.

Hoskins, J., Norman, D. K., Beckmann, R. J., & Long, G. L. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 349–353.

Joseph, D. R., Hall, S. H., Conti, M., & French, F. S. (1988) *Mol. Endocrinol. 2*, 3–13.

Leytus, S. P., Foster, D. C., Kurachi, K., & Davie, E. W. (1986) *Biochemistry 25*, 5098–5102.

Little, P. F. R., & Cross, S. H. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 3159–3163.

Long, G. L. (1987) in *Current Advances in Vitamin K Research* (Suttie, J. W., Ed.) pp 153–163, Elsevier, New York.

Long, G. L., Marshall, A., Gardner, J. C., & Naylor, S. L. (1988) *Somatic Cell Mol. Genet. 14*, 93–98.

Lundwall, Å., 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.

Mount, S. M. (1982) *Nucleic Acids Res. 10*, 459–472.

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.

Orcutt, B. C., Dayhoff, M. O., George, D. G., & Barker, W. C. (1984) *Protein Identification Resource*, National Biomedical Research Foundation, Georgetown University Medical Center, Washington, DC.

Patthy, L. (1985) *Cell 41*, 657–663.

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

Petra, P. H., Que, B. G., Namkung, P. C., Ross, J. B. A., Chabonneau, H., Walsh, K. A., Griffin, P. R., Shabanowitz, J., & Hunt, D. F. (1988) *Ann. N.Y. Acad. Sci. 538*, 10–24.

Picado-Leonard, J., & Miller, W. L. (1988) *Mol. Endocrinol. 2*, 1145–1150.

Ploos van Amstel, H. K., van der Zanden, L., Reitsma, P. H., & Bertina, R. M. (1987a) *FEBS Lett. 222*, 186–190.

Ploos van Amstel, J. 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.

Ploos van Amstel, H. K., Reitsma, P. H., van der Logt, C. P. E., & Bertina, R. M. (1990) *Biochemistry* (second of three papers in this issue).

Plutzky, J., Hoskins, J. A., Long, G. L., & Crabtree, G. R. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 546–550.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 5463–5467.

Schwarz, H. P., Fisher, M., Hopmeyer, P., Batard, M. A., & Griffin, J. H. (1984) *Blood 64*, 1297–1300.

Schmidel, D. K., Tatro, A. V., Phelps, L. G., Tomczak, J. A., & Long, G. L. (1990) *Biochemistry* (first of three papers in this issue).

Staden, R. (1982) *Nucleic Acids Res. 10*, 2951–2961.
Stenflo, J. (1988) in *Protein C and Related Proteins* (Bertina, R. M., Ed.) pp 21–54, Churchill Livingstone, New York.
Stenflo, J., Lundwall, Å., & Dahlbäck, B. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 368–372.
Stern, D., Brett, J., Harris, K., & Nawroth, P. (1986) *J. Cell. Biol. 102*, 1971–1978.
Sun, L., Poulson, K. E., Schmid, C. W., Kadyk, L., & Leinwand, L. (1984) *Nucleic Acids Res. 12*, 2669–2690.
Walker, F. J. (1984) *Semin. Thromb. Hemostasis 10*, 131–138.
Watkins, P. C., Eddy, R., Fukushima, Y., Byers, 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.
Zagursky, R. J., Baumeister, K., Lomax, N., & Berman, M. L. (1985) *Gene. Anal. Techn. 2*, 89–94.

Oxidative Cleavage of DNA Mediated by Hybrid Metalloporphyrin–Ellipticine Molecules and Functionalized Metalloporphyrin Precursors†

Li Ding, Guita Etemad-Moghadam, and Bernard Meunier*
Laboratoire de Chimie de Coordination du CNRS, 205 route de Narbonne, 31077 Toulouse Cedex, France
Received February 14, 1990; Revised Manuscript Received May 11, 1990

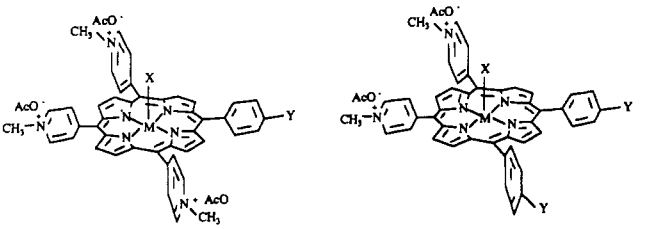
ABSTRACT: The nuclease activity of functionalized metalloporphyrins **1–8** and hybrid metalloporphyrin–ellipticine molecules **10–16** in the presence of potassium monopersulfate ($KHSO_5$) or magnesium monoperoxyphthalate (MMPP), water-soluble oxygen atom donors at physiological pH, toward double-stranded ϕX174 DNA is reported. The DNA cleavage efficiency as a function of the nature of functionalized metalloporphyrins, the length of the linkage between the two parts of the hybrid molecule, viz., metalloporphyrin and 9-methoxyellipticine, the nature of the central metal atom (Mn, Fe, or Zn), the ionic strength, and the nature of the oxygen donor has been studied. Single-strand breaks (SSBs) are observed on double-stranded DNA with a short incubation time of 2 min in the presence of manganese derivatives of both metalloporphyrins and hybrid molecules. Owing to their cytotoxic and nuclease activity, these new water-soluble hybrid molecules may be considered as efficient bleomycin models based on cationic metalloporphyrins.

The therapeutic activity of bleomycin, an efficient antitumoral antibiotic agent, is generally attributed to its DNA binding properties (Chien et al., 1977; Henichart et al., 1985) and to its ability to cleave DNA (Takeshita et al., 1978; d'Andrea & Haseltine, 1978). However, the DNA cleavage is observed in the presence of three cofactors: iron or copper salts, molecular oxygen, and an electron source (Sausville et al., 1978; Burger et al., 1981; Ehrenfeld et al., 1987). Such cleavage mainly occurs via single-strand breaks due to the abstraction of a hydrogen atom at the $C_{4'}$ position of the deoxyribose ring by a high-valent metal–oxo species strongly chelated by bleomycin (Povirk, 1983; Hecht, 1986; Stubbe & Kozarich, 1987; Pratviel et al., 1989a).

This mechanism involving several redox-active metals makes the modeling of bleomycin an interesting goal. All the modeling studies of bleomycin have been based on its structural duality: one part of the molecule is responsible for the DNA interaction (bithiazole, intercalating agent), whereas a second part is a strong chelator for metal ions [peptidic chain, EDTA[1] (Moser & Dervan, 1987; Youngquist & Dervan, 1987; Dervan, 1986), or a metalloporphyrin (Lown & Joshua, 1982; Lown et al., 1986; Hashimoto et al., 1983, 1986)].

Because our group is involved in oxidation reactions catalyzed by metalloporphyrins (Meunier, 1986), DNA cleavage by high-valent metal–oxo species (Fouquet et al., 1987; Bernadou et al., 1989; Pratviel et al., 1989a,b), and the mechanism of action of cytotoxic ellipticine derivatives (Meunier et al., 1988), we decided to synthesize hybrid metalloporphyrin–ellipticine molecules (Tadj & Meunier, 1988) on the basis of the association of a chelating agent, a metalloporphyrin, to an intercalating agent, an ellipticine. However, because of the presence of one hydrophobic metalloporphyrin moiety, these

Chart I: Structures of Cationic Functionalized Metalloporphyrins **1–8**



1: Y = NO_2; M = Mn
2: Y = NH_2; M = Mn
3: Y = NMe_3; M = Mn, Fe, Zn
4: Y = OH; M = Mn
5: Y = $NHCO(CH_2)_3N^+Me_3$; M = Mn
6: Y = $O(CH_2)_3N^+Me_3$; M = Mn
7: Y = OH; M = Mn
8: Y = $O(CH_2)_3N^+Me_3$; M = Mn

† This research was supported by the CNRS (DVAR) and by Pierre Fabre Médicaments (Castres). A fellowship from l'Association pour la Recherche sur le Cancer (ARC, Villejuif) to L. D. is gratefully acknowledged.

[1] Abbreviations: EDTA, ethylenediaminetetraacetic acid; MMPP, magnesium monoperoxyphthalate; 5-MF, 5-methylene-2-furanone.

0006-2960/90/0429-7868$02.50/0