Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982) *Molecular Cloning. A Laboratory Manual*, p 486, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.
Marky, L. A., & Breslauer, K. J. (1987) *Biopolymers 26*, 1601-1620.
Mirau, P. A., & Kearns, D. K. (1984) *Biochemistry 23*, 5439-5446.
Morvan, F., Rayner, B., Imbach, J.-L., Lee, M., Hartley, J. A., Chang, D.-K., & Lown, J. W. (1987) *Nucleic Acids Res. 15*, 7027-7044.
Patel, D. (1980) in *Nucleic Acid Geometry and Dynamics* (Sarma, R. H., Ed.) Pergamon Press, New York.
Quaedflieg, P. J. L. M., Broeders, N. L. H. L., Koole, L. H., van Genderen, M. H. P., & Buck, H. M. (1990a) *J. Org. Chem. 55*, 122-127.
Quaedflieg, P. J. L. M., van der Heiden, A. P., Koole, L. H., van Genderen, M. H. P., Coenen, A. J. J. M., van der Wal, S., & Buck, H. M. (1990b) *Proc. K. Ned. Akad. Wet. 93*, 33-38.
Remin, M., & Shugar, D. (1972) *Biochem. Biophys. Res. Commun. 48*, 636-642.
Ribas-Prado, F., & Giessner-Prettre, C. (1981) *J. Mol. Struct.: THEOCHEM 76*, 81-92.
Roth, K., Kimber, B. J., & Feeney, J. (1980) *J. Magn. Reson. 41*, 302-309.
Sarma, M. H., Gupta, G., & Sarma, R. H. (1986) *FEBS Lett. 205*, 223-229.
Thuong, N. T., Asseline, U., Roig, V., Takasugi, M., & Hélène, C. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 5129-5133.
Tsuboi, M. (1967) in *Conformation of Biopolymers*, Vol. II, pp 689-702, Academic Press, New York.
van Genderen, M. H. P., Koole, L. H., Aagaard, O. M., van Lare, C. E. J., & Buck, H. M. (1987a) *Biopolymers 26*, 1447-1461.
van Genderen, M. H. P., Koole, L. H., & Buck, H. M. (1987b) *Proc. K. Ned. Akad. Wet., Ser. B 90*, 181-187.
van Genderen, M. H. P., Koole, L. H., & Buck, H. M. (1988) *Proc. K. Ned. Akad. Wet., Ser. B 91*, 171-178.
von Hippel, P. H., & McGhee, J. D. (1972) *Annu. Rev. Biochem. 41*, 231-300.
Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S., Jr., & Weiner, P. K. (1984) *J. Am. Chem. Soc. 106*, 765-784.
Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, p 208, Wiley and Sons, New York.

Organization of the Human Protein S Genes[†,‡]

Dyann K. Schmidel, Alicia V. Tatro, Lisa G. Phelps, Jennifer A. Tomczak, and George L. Long*
Department of Biochemistry, University of Vermont, Burlington, Vermont 05405
Received January 30, 1990; Revised Manuscript Received May 2, 1990

ABSTRACT: Human genomic clones that span the entire protein S expressed gene (PSα) and the 3′ two-thirds of the protein S pseudogene (PSβ) have been isolated and characterized. The PSα gene is greater than 80 kilobases in length and contains 14 introns and 15 exons, as well as 6 repetitive "Alu" sequences. Exons I and XV contain 112 and 1139 bp 5′ and 3′ noncoding segments in addition to the amino and carboxyl termini, respectively. Exons I–VIII encode protein segments that are homologous to the vitamin K dependent clotting proteins and are bounded by introns whose position and type are identical with other members of this protein family. Exons IX–XV encode protein segments homologous to sex hormone binding globulin (SHBG) and are bounded by introns of identical type and position as in the SHBG gene. Genomic clones for the PSβ gene cover a distance of greater than 55 kilobases and contain segments corresponding to amino acids 46–635 of the mature protein and the 1.1-kb 3′ noncoding region of the cDNA. The presence of multiple base changes in the coding portions of this gene, resulting in termination codons and frame shifts, suggests that it is a pseudogene. Comparison of DNA sequences for the two genes reveals 97% identity for coding and 3′ noncoding, and 95.4% for intronic regions, suggesting divergence of the two genes is a relatively recent event.

Human protein S is a 69 000-Da vitamin K dependent plasma glycoprotein (Di Scipio & Davie, 1979) that acts as a cofactor for activated protein C in the coagulation cascade to inactivate factors Va and VIIIa (Walker, 1981; Suzuki et al., 1983; Solymoss et al., 1988). Protein S is synthesized in hepatocytes (Fair & Marlar, 1986), endothelial cells (Fair et al., 1986; Stern et al., 1986), and the megaloblastic cell line MEG-01 (Ogura et al., 1987). It is found circulating in the blood in equimolar amounts free and bound to the complement protein C4b binding protein (C4BP), in a 1:1 ratio (Dahlback, 1983). Protein S in the bound form is not available as a cofactor for APC (Dahlback, 1986).

Hereditary protein S deficiency has been reported by several groups and is often associated with symptoms found in protein C deficient individuals [for a review, see Engesser et al. (1987)], including familial thrombophilia. Recently, a molecular alteration in the protein S gene has been reported in a family exhibiting protein S deficiency (Ploos van Amstel et al., 1989). We have also recently described different alterations in the expressed gene from four independent families (Schmidel et al., 1989).

The cDNA for human protein S has been cloned and fully characterized (Lundwall et al., 1986; Hoskins et al., 1987; Ploos van Amstel et al., 1987). The translated precursor

[†] This work was supported in part by National Institute of Heart, Lung and Blood Grants R01 HL 38899 and C06 39745.
[‡] The nucleic acid sequence in this paper has been submitted to GenBank under Accession Number J02917.
* To whom correspondence should be addressed.

0006-2960/90/0429-7845$02.50/0

protein consists of 676 amino acid residues from which a 41-AA leader peptide is cleaved to produce the single-chain mature protein. Ploos van Amstel and co-workers (1987, 1988) were the first to report that there are two copies of the gene for protein S, designated PSα and PSβ. Evidence from several laboratories suggests that both copies of the gene are on chromosome 3 (Ploos van Amstel et al., 1988; Long et al., 1988; Watkins et al., 1988).

This paper describes the isolation and characterization of genomic DNA segments representing the two protein S genes. Clones for one of the genes (PSα) represent over 80 kb of DNA and contain segments that are in complete agreement with the entire cDNA sequence, and are considered to represent the expressed gene. Portions of the second gene (PSβ) spanning approximately 55 kb and containing elements corresponding to amino acids 46–635 of the mature protein as well as 1.1 kb of 3′ untranslated cDNA sequence have also been characterized. On the basis of multiple base changes in the second gene compared to the cDNA sequence (some resulting in termination codons and frame shifts), it is thought to be a pseudogene.

Materials and Methods

cDNA Screening of a λEMBL3 Library. A hemophiliac human liver library of partially *Sau*3A-digested genomic DNA in bacteriophage λEMBL3 was kindly supplied by Dr. R. T. A. MacGillivray, Department of Biochemistry, University of British Columbia, Canada, and is described elsewhere (Geddes et al., 1989). The library was screened by the procedure of Benton and Davis (1977). Nitroplus 2000 filters (MSI, Micronsep, Westboro, MA) were prehybridized in hybridization solution [6× SSC (0.9 M NaCl and 0.09 M sodium citrate, pH 7.0), 2× Denhardt's [100 mL: 2 g of ficoll, 2 g of poly(vinylpyrrolidone), and 2 g of bovine serum albumin], 1 mM EDTA, pH 8.0, and 0.5% SDS] prior to addition of probe. Various human protein S cDNA fragments (spanning nucleotides 116–3275; Hoskins et al., 1987) were radiolabeled by the random hexamer primer method of Fienberg and Vogelstein (1983) and purified on Bio-Rad (Richmond, CA) P-60 columns. The filters were hybridized overnight at 65 °C, rinsed, and exposed to X-ray film. Agar plugs corresponding to the positive signals on the film were pulled from the master plates and dispensed onto 1 mL of SM (0.1 M NaCl, 8 mM $MgSO_4 \cdot 7H_2O$, 50 mM Tris, pH 7.5, and 0.01% gelatin) containing one to three drops of chloroform.

Isolation of EMBL3 λ DNA. The plate lysis method of Maniatis et al. (1982, pp 65–66) was used to prepare phage stocks of positive clones. These were stored at 4 °C with one to three drops of chloroform. Rapid, small-scale isolation of bacteriophage λ DNA (Maniatis et al., pp 371–372, 1982) was used to prepare DNA for restriction digests and Southern blots. For mapping purposes, phage DNA stock was routinely digested with *Sal*I, *Eco*RI, *Bam*HI, *Hin*dIII, *Sst*I, and *Xba*I (Bethesda Research Labs, Gaithersburg, MD) and run on a horizontal submerged 0.7% agarose gel in 1X TAE (0.04 M Tris–acetate and 0.002 M EDTA, pH 8).

Southern Hybridization. Following ethidium bromide staining and photography, DNA in agarose gels was transferred to Nitroplus 2000 filters by the method of Smith and Summers (1980) and probed as described above.

Subcloning into pUC19. Electroeluted subfragments were ligated into appropriately cleaved plasmid pUC19 and used to transform competent DH5α cells (Bethesda Research Laboratories) according to the supplier. Cells were cultured overnight at 37 °C on 1.5% agar–LB plates containing ampicillin (50 μg/mL), 5-bromo-4-chloro-3-indolyl β-D-galactopyranoside (X-gal) (80 μg/mL), and isopropyl β-D-thiogalactopyranoside (IPTG) (100 mg/mL). Plasmid from positive colonies was mini-prepped by the method of Birnboim and Doly (1979) and analyzed by standard procedures (Maniatis et al., 1982).

DNA Sequencing. Mini-prepped DNA (~3 μg) suitable for sequencing was generated by RNase A digestion followed by phenol extraction and ethanol precipitation. Sequencing was done by the Sanger dideoxy method (Sanger et al., 1977) using the Sequenase 1.0 kit (United States Biochemical Corp., Cleveland, OH) with modifications (Zhang et al., 1988). Custom primers were synthesized on an Applied Biosystems Incorporated (Foster City, CA) Model 381 DNA synthesizer in the trityl-off mode using β-cyanoethyl phosphoramidite derivatives, and used without purification (0.5–5 pmol) in sequencing reactions. A sequencing strategy of "intron jumping" was employed whereby primers at the extreme ends of newly established adjacent exons were synthesized and used. Consequently, all of the exonic segments, including splice junctions, were sequenced on both strands. Exon III (Figure 1) because of its short length was sequenced in both directions using intronic primers. All of the 5′ flanking (hatched in Figure 1) and 3′ noncoding regions were also sequenced on both strands. Generally, intronic sequences were determined on only one strand.

Analysis of Alu Sequences. Highly repetitive Alu segments (Deininger et al., 1981) were identified by Southern hybridization with a radiolabeled 150 bp *Pst*I/*Bgl*II Alu fragment from intron E of the gene for human protein C (Foster et al., 1985). Sequence analysis of the Alu segments was performed with a custom oligonucleotide primer (5′CCCAGCTA-CTCGGGAGGCTGACCG3′) corresponding to consensus nucleotides −5 to +19 reported by Deininger et al. (1981).

Computer Analysis. All DNA and protein sequences were analyzed by using the Sequence Analysis Software Package of the University of Wisconsin Genetics Computer Group (GCG), version 5 (Devereux et al., 1984).

Results

Human Genomic Clones for Protein S. A human liver EMBL3 library was repetitively screened with cDNA probes to obtain overlapping clones of the two protein S genes. A total of 14 genomic clones were isolated and characterized for the gene PSα and 19 for PSβ. Our designation of expressed gene and pseudogene is based upon the observation of complete agreement with the cDNA sequence for expressed gene clones and multiple base substitutions, including termination codons and deletions, for pseudogene clones, as discussed further below. A restriction map of the protein S genes for six restriction endonucleases is shown in Figure 1. The restriction map of the genes is similar, with only 20% restriction fragment length polymorphisms for the overlapping regions of the PSα and PSβ genes. The sum of the regions represented by genomic clones corresponding to the PSα gene is greater than 80 kb in length and contains three gaps. The gaps are in introns A, C, and I. Those in introns A and C are of unknown size, but based upon the clones represented in Figure 1, introns A and C must be minimally ≥20 kb and ≥10 kb, respectively. The gap in intron I has been determined to be ~150 bp, as determined by restriction mapping. Clones for the PSβ gene span approximately 55 kb of DNA and do not contain any regions upstream of the 3′ end of intron C. There is one gap in the clones of the PSβ gene that begins 46 nucleotides before the *Sal*I site in exon 14 and extends downstream into intron N for approximately 3.5 kb, based upon restriction mapping presented by Edenbrandt and Stenflo (1990). The restriction

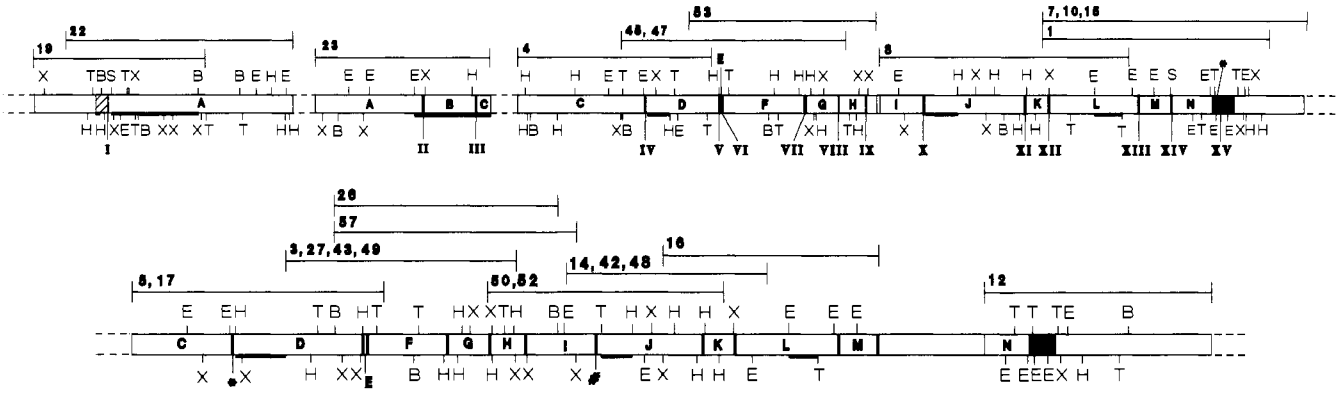


FIGURE 1: Organization of the human protein S genes. Broad open horizontal bar maps represent the PSα (expressed, top) and PSβ (pseudo, bottom) genes. Horizontal lines (with numbers) above each bar map show individual genomic inserts in λEMBL3, with vertical end lines representing vector *Sal*I cloning sites. Dashed lines denote gaps of unknown size. Intron (bold letters) and exon (roman numerals) positions are noted. Positions of the termination codon in the expressed gene and the first termination codon in the pseudogene are shown with asterisks. A two-nucleotide deletion in the PSβ gene, resulting in a coding frame shift, is also noted (#). Regions of the genes containing repetitive "Alu" sequences are represented by a heavy horizontal line under the broad open bars. Flanking 5′ sequence is shown with hatched lines. Abbreviations for enzymes are as follows: *Eco*RI, *E*; *Hin*dIII, H; *Sal*I, S; *Bam*HI, B; *Xba*I, X; *Sst*I, T.

Table I: Comparison of Intron Position and Type for Human Protein S with Homologous Proteins[a]

intron[b]	A	B	C	D	E	F	G	H	I	J	K	L	M	N
protein S	−16[c]	37/38	46	75	116	160[d]	202[d]	242/243	281	344/345	400/401	457	507/508	583
factor VII	−17	37/38	46	84	131			167/168						
factor IX	−17	38/39	47	85	128			195/196						
factor X	−17	37/38	46	84	128			209/210						
protein C	−19	37/38	46	92	137			184/185						
prothrombin	−17	37/38	46											
human SHBG								8/9	39	102/103	156/157	210	255/256	325
intron type[e]	I	O	I	I	I	I	I	O	II	O	O	I	O	I

[a] Homologous proteins include human plasma coagulation proteins prothrombin (Degan & Davie, 1987), factor VII (O'Hara et al., 1987), factor IX (Anson et al., 1984; Yoshitake et al., 1985), factor X (Leytus et al., 1986), protein C (Foster et al., 1985; Plutsky et al., 1986), and human sex hormone binding globulin (SHBG) (Gershagen et al., 1989). [b] Intron letters are for protein S, this paper. [c] Numbers refer to amino acid positions reported in the cited sources. A blank space denotes that no region exists that is homologous to that in protein S. [d] Exons following introns F and G are two additional domains in protein S that are homologous to epidermal growth factor, in addition to those following introns D and E. [e] Based upon the convention of Sharp (1981).

maps are in excellent agreement with those independently generated by Edenbrandt and Stenflo (1990).

Organization of the Protein S Gene. Figure 2 presents partial DNA sequences and resulting translation products for both protein S genes, including segments in the PSα gene corresponding to the entire reported cDNA sequence (Hoskins et al., 1987). A total of 8752 and 5414 nucleotides are presented for the PSα and PSβ genes, respectively. The sequence for the PSα gene is in complete agreement with that of the cDNA, whereas that of the PSβ gene contains several point mutations, including the generation of a termination codon at amino acid residue 61 of the mature protein. Comparison of the two DNA sequences with each other reveals that they are 96.8% identical for coding regions, 96.9% for the 3′ noncoding cDNA region, and 95.4% for intervening sequence. (An insertional or deletion event regardless of the number of nucleotides was scored as one substitution). With the exception of one nucleotide difference (G → T at the fifth position from the 5′ end of PSα gene intron K; Edenbrandt & Stenflo, 1990), the sequences are in complete agreement with those independently determined by two other laboratories (Edenbrandt & Stenflo, 1990; Ploos van Amstel et al., 1990).

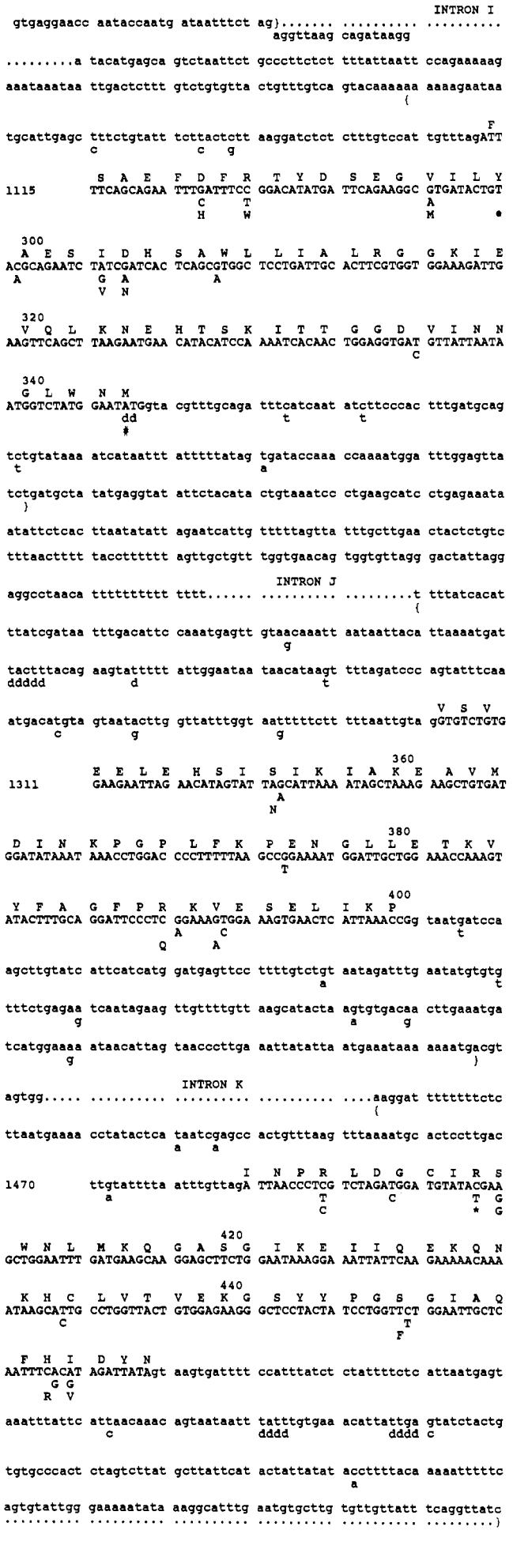
The human PSα gene consists of 15 exons separated by 14 introns. Exons I and XV include 112 and 1139 bp of 5′ and 3′ untranslated sequence, respectively. The exons code for segments ranging from 9 to 76 amino acids in length. All intron–exon junctions were found to correspond to consensus sequences (Mount, 1982) and obey the 5′GT...AG3′ rule (Breathnach et al., 1978). Sequences resembling the proposed lariot branch site generally located 20–50 nucleotides upstream of the 3′ splice site of the intron (Sharp, 1987) are also present. The intron–exon junctions of protein S with other human vitamin K dependent plasma proteins and sex hormone binding globulin (Gershagen et al., 1989) are shown in Table I. Table I reveals that intron position and type are identical for protein S and homologous proteins in the regions of homology when the sequences are maximally aligned.

By Southern hybridization and subsequent DNA sequencing, six repetitive Alu sequences have been identified in the protein S gene. The approximate location of each Alu sequence is noted in Figure 1. Three additional Alu sequences have been identified in the PSβ gene, the positions of which are the same as in the PSα gene (see Figure 1). Overall similarity of the Alu sequences (data not shown) is about 82% identity with each other and the consensus sequence of Deininger et al. (1981). Consequently, the expected average base substitution level is about 18%. Subclones, each containing an Alu region shown in Figure 1, yielded clean, unique sequences, suggesting that in each region only one Alu segment exists.

Discussion

In this paper, we report the cloning and characterization of the expressing gene (PSα) for human protein S, including all elements corresponding to the cDNA. We have also isolated and characterized genomic clones containing multiple nucleotide substitutions, including termination codons and frame shifts. Translation of the coding portions of this second copy (PSβ) of the gene would result in termination at amino acid position 61 of the mature protein, resulting in at the very

```
aggaaggctg agacaggaga atcacttgaa cccggagtgg aggttgcagt gagccgaaat
tgcgccactg ttctccagcc tgggcgacag agtgagactc tgtctcaaaa aaaaaaaaaa
tttttttaaa ttaataaata ggtagatgta tagatagata ttcactttca actgaagtct
ttatcggagc aagatttttt ttaaggtaga ttattctaat attccctctc tttcttcatg
ttctttacat tatttttatg cctgtatggc atacaagacc gaaaaaacat gtggatgatc
aaaatgaccc catttgcttt tactatcacc atagttcttc ctaaagtcct cattgacttc
caggttttgg ttaatatgtc ttcaggacaa ctcagtgtct cactgtttct gcttctgaac
ctagggatcc tgtcctcttg aaccctggaa gttgtcttga ccagtcagag aactgcgttc
cccacccctt cccctttgga aacgtcacac tgtggaggaa aagcaagcaa ctagggagct
ggtgaagaag gatgtctcag cagtgtttac taggcctcca acactagagc ccatccccca
gctccgaaaa gcttcctgga aatgtccttg ttatcacttc ccctctcggg ctgggcgctg
ggagcgggcg gtctcctccg cccccggctg ttccgccgag gctcgctggg tcgctggcgc
38         cgccGCGCAG CACGGCTCAG ACCGAGGCGC ACAGGCTCGC AGCTCCGCGG
CGCCTAGCGC TCCGGTCCCC GCCGCGACGC GCCACCGTCC CTGCCGGCGC CTCCGCGCGC

    -40
    M  R   V  L  G   G  R  C   G  A  L   L  A  C   L  L  L   V
TTCGAAATGA GGGTCCTGGG TGGGCGCTGC GGGGCGCTGC TGGCGTGTCT CCTCCTAGTG

     -20
L  P  V  S   E  A  N   L
CTTCCCGTCT CAGAGGCAAA CTgtgagtaa tcaatagcgt ctcttctccc ttccccagca
ttgtcgactg aactgcgtcc ctggttggta ggattttctt ctctagagct gcagctccta
        INTRON A
gaaa...... .......... ttagttcgat catactgatt ttaaatgtca tacaattcat
atgcagaaaa tgattttaac tcttattgtt taataaaaca atatatttta catggaaaaa
                                                  L  S   K  Q  Q
223        tgattaattc atataaactg attgtttcct tcagTTTTGT CAAAGCAACA

                                 +1
A  S  Q   V  L  V   R  K  R   R  A  N   S  L  L   E  E  T   K  Q
GGCTTCACAA GTCCTGGTTA GGAAGCGTCG TGCAAATTCT TTACTTGAAG AAACCAAACA

                                 20
G  N  L   E  R  E   C  I  E   E  L  C   N  K  E   E  A  R   E  V
GGGTAATCTT GAAAGAGAAT GCATCGAAGA ACTGTGCAAT AAAGAAGAAG CCAGGGAGGT

F  E  N   D  P  E   T
CTTTGAAAAT GACCCGGAAA CGgtaagcat ttatggaaac tatcaagttc acacatctag
                                      INTRON B
acatacaact acagactgaa ca........ .......... .......... ...ttggtat
atttacctta tataaaaatt ataatgtgaa aatgatggtt atatgtaact tatttgcatc
tttaatataa aaacacatta taaaattaag ttttaactct ataatgaaat ttaggtttgc
                                              40
                                              D  Y  F   Y  P  K   Y
381        taagatatgt tttctttttc ttctttctag GATTATTTTT ATCCAAAATA

L  V
CTTAGgtaag ttcaaaacat ctcaattata taatcttaga aatggaaggg aacttagata
tgttcctgtc taactctcca cctatctatt ccatccaatg taatttcatt atccttggga
gtgaggtttc attttatatg caattttgct gtctttatat ttaacacatg ttgagaaatt
                                            INTRON C
tggaacggtt ctttaattac catgtgacat tttaaatttt ........ag atttttaatt
tttcagaaag gtagatggaa acatttagtg ttgcattttg aagactgaat cttgtgaatc
tacaggagca taaatgtcct acctcttggg acagttccta ccatgaattc agatcaagta
                                    (gg   ig
                                                          C  L
406        tgtgtgtcta ctctaagaag attatgtttg tttttatttt cagTTTGTCT
            a           gc          c                     G

                                   60
R  S  F   Q  T  G   L  F  T   A   A  R  Q   S  T  N  A   Y  P  D
TCGCTCTTTT CAAACTGGGT TATTCACTGC TGCACGTCAG TCAACTAATG CTTATCCTGA
 T                                 A T
 L                                 H  *

L  R  S   C  V  N  A
CCTAAGAAGC TGTGTCAATG gtaagcactt ctaccatcaa ttgaaaaaac aaaacaaaaa
                                            c           t     )
ctctgtaggt aaagtacacc catggtaata taaactaacg tttaaaaatt gagaaatgat
cctttgatgg gttgtatgcc cagcaaggaa gtattttcaa actgcatttc taatacttgt
ttatagttgt aaatgctcct gtattactca aaatgaattt tt........ ..........
                 INTRON D
.......... .......... .......... .....(tttt tcattggttc taggcttcag
                               tttt tttttt                a
gatttttatt atagtacaca caattttatt tttccatgac atgagataaa aaaaataaat
            c  g                     a          t g

                                   80
                                 I  P  D   Q  C  S   P  L  P  C
493        agatgtctat ttccttcagc CATTCCAGAC CAGTGTAGTC CTCTGCCATG
                                             A

                                               100
N  E  D   G  Y  M   S  C  K   D  G  K   A   S  F  T   C   T  C  K
CAATGAAGAT GGATATATGA GCTGCAAAGA TGGAAAAGCT TCTTTTACTT GCACTTGTAA
    C                T           A                       T
  A                     E                                I

P  G  W   Q  G  E   K  C  E   F  D
ACCAGGTTGG CAAGGAGAAA AGTGTGAATT TGgtacgtat aataaccccc gccccccagc
                                   t          g          t
                                   INTRON E
tcatcaggat tggtctcctg aaaagttctc tgcaggttat attactttaa aaataattta
                                   t
                                      120
                                 I  N  E   C  K  D   P  S  N  I
616          tttttt cctgttttag ACATAAATGA ATGCAAAGAT CCCTCAAATA
                                           C
                                           R

                                                140
N  G  G   C  S  Q   I  C  D   N  T  P  G   S  Y  H   C  S  C  K
TAAATGGAGG TTGCAGTCAA ATTTGTGATA ATACACCTGG AAGTTACCAC TGTTCCTGTA

                                               160
N  G  F   V  M  L   S  N  K   K  D  C  K    D
AAAATGGTTT TGTTATGCTT TCAAATAAGA AAGATTGTAA AGgtaagagc aggatggtag
  G                                C          g
 S
                                                            INTRON F
aattaaaaca catttactat gtgagaataa )......... .......... ..........
                       g          tcccagttag agaaatttt
.......... .....gtgag cctaacatat gatgatagat ttaatgtttt tggtccaaag
                                             ( tt t     t   g
gccaatctgt tatctcatta catatattaa acaagatcca ggaaacacaa atcaagggtt
                             c g     t          a          a
ctttggtagt gtctgctggt atgggaaaaa tgttttttaa tatagtgatt ttatttatag
   c                                 d
            V  D  E   C  S  L   K  P  S  I   C  G  T   A  V  C
748        ATGTGGATGA ATGCTCTTTG AAGCCAAGCA TTTGTGGCAC AGCTGTGTGC

           180
K  N  I  P   G  D  F   E  C  E   C  P  E  G   Y  R  Y   N  L  K
AAGAACATCC CAGGAGATTT TGAATGTGAA TGCCCCGAAG GCTACAGATA TAATCTCAAA
                                     T

       200
S  K  S  C   E  D
TCAAAGTCTT GTGAAGgtag gatgatggtg gtatcattac tgatc).... ..........
                  a          g               ggtgga gaggttgggt
                                                       INTRON G
.......... .......... .......... .......... .......... ..........
gtgaacctga aatggtattt aaaagcattg gttt...... .......... ..........
.......... .......... .......... .......aat gtagttgatg tcatagtatt
                                             (  ag  t tc
cttccctaag gttcgattat cattgattat atcatactac aatcataata ttcctctgcc
t c          c t                                            g
tataagattg aacatttagg ggatattaaa gtttgtgtgc gtgtgttttt tttacctcag
                       a                 t        g gi[gt3,g,cg3,tg]
            I  D  E   C  S  E   N  M  C  A   Q  L  C   V  N  Y
874        ATATAGATGA ATGCTCTGAG AACATGTGTG CTCAGCTTTG TGTCAATTAC
                       C           T
                       R

   220
P  G  G  Y   T  C  Y   C  D  G   K  K  G  F   K  L  A   Q  D  Q
CCTGGAGGTT ACACTTGCTA TTGTGATGGG AAGAAAGGAT TCAAACTTGC CCAAGATCAG
                       C

  240
K  S  C  E
AAGAGTTGTG AGgtaaacat tttacaatgc ttaacttctc acctgttttc taaaatgaga
                           a                      d
gatcctagat acttattttc acatagctaa gtcaggaaaa tacagacgtt ctgcaataat
 t                            g  ittagg          t        a
tgctgtagct ttaagcaggt tatgagtcat gattatttac taattatggt catgtaagtc
   d                              t                             a
ataaacattg gggataatta gtggctgggt taaaaaagcg ggta...... ..........
                         a              )
                 INTRON H
.......... .......... .......... .......... ....tctaga tagcacattc
ttgctcagaa aacttttgct taataaatga atgagtgaat gaatgaggtg ctgtttattg
gtacattaat tcttacacct atttctgact tcaaataaaa ataattttca gacataattc
atgggagata atatacctga ctgttaatta aagaaatata tattggttct tgaagaaaga
gtttgtgttt aggacgaaaa tttgcaaagg aaggtattaa agacaaagat cgaaagccat
gcacattgaa cgaggcttta aaacacatgt attcttggag gttatactga tagatagact
atacaaggaa atggaagata tgtatttatt attaccaaaa atattcttaa ttattttagt
                                                    (
gtattacaga tgatacatta gtaaccaaac aaaaatgcat gacctcacac aaacattaag
                                  g  a
                                                           V  V
caataacctg tgcattttga ttttcttgtt gtttatttgg tttcttttat tccagGTTGT
     a                      a

                                                          260
            S  V  C   L  P  L  N   L  D  T   K  Y  E  L   L  Y  L
1001       TTCAGTGTGC CTTCCCTTGA ACCTTGACACA AAGTATGAAT TACTTTACTT
                                              C
                                              Q

                                                          280
A  E  Q   F  A  G  V   V  L  Y   L  K  F   R  L  P  E   I  S  R
GGCGGAGCAG TTTGCAGGGG TTGTTTTATA TTTAAAATTT CGTTTGCCAG AAATCAGCAG
A A         T                                A
T           S                                H
```


INTRON I
INTRON J
INTRON K
INTRON L
INTRON M
INTRON N

```
GGTTGGATAG GTGGGGAAAT CTGAAGTGGG TACATTTTTT AAATTTTGCT GTGTGGGTCA
          AT                        AC       G

CACAAGGTCT ACATTACAAA AGACAGAATT CAGGGATGGA AAGGAGAATG AACAAATGTG
    d                              iC

GGAGTTCATA GTTTTCCTTG AATCCAACTT TTAATTACCA GAGTAAGTTG CCAAAATGTG

ATTGTTGAAG TACAAAAGGA ACTATGAAAA CCAGAACAAA TTTTAACAAA AGGACAACCA
                                  G

CAGAGGGATA TAGTGAATAT CGTATCATTG TAATCAAAGA AGTAAGGAGG TAAGATTGCC
                                  A       A

ACGTGCCTGC TGGTACTGTG ATGCATTTCA AGTGGCAGTT TTATCACGTT TGAATCTACC

ATTCATAGCC AGATGTGTAT CAGATGTTTC ACTGACAGTT TTTAACAATA AATTCTTTTC
                                           G     G   A

3268        ACTGTATTTT ATATCACTTA TAATAAATCG GTGTATAATT TTAAAAtgca
                                        T  A

tgtgaatatc tttattatat caactgtttg aataaaac
                    g              }
```

FIGURE 2: Nucleotide sequence of the protein S genes. Nucleotides are presented in the 5′ → 3′ direction for the coding strand. Those corresponding to the cDNA are shown with capital letters. Corrections in the cDNA sequence from that originally reported (Hoskins et al., 1987) are underlined. Nucleotide numbers on the left side of lines correspond to the first cDNA nucleotide in the line and are taken from Hoskins et al. (1987). The translated amino acid sequence and numbering for the protein precursor are shown in one-letter code above the expressed gene sequence. Immediately below the PSα gene sequence are presented the corresponding PSβ gene sequence and the amino acid translation. Small braces represent the 5′- and 3′-most extreme points of comparison. The nucleotides at the position of the brace and blank spaces within the brace are identical with the shown sequence. Nucleotide substitutions, deletions (d), and insertions (i) in the pseudogene are shown below the expressed gene sequence. Inserted nucleotides are shown to the right of the insertion symbol (i) and are inserted in the upper sequence following the nucleotide directly above the i. The sequence of the insertion in brackets at the 3′ end of intron G is gtgtgtgcgcgcgtg. Unsequenced gaps are shown with dotted lines. Large brackets enclosing the 3′ end of exon 14 and the 5′ end of intron N denote a region of the PSβ gene not cloned. Predicted termination codons are shown with asterisks. The location of two-nucleotide deletion in the PSβ gene is noted (#) and would result in multiple subsequent termination codons because of the frame shift. The first 120 nucleotides shown are the 3′ end of an "Alu" sequence.

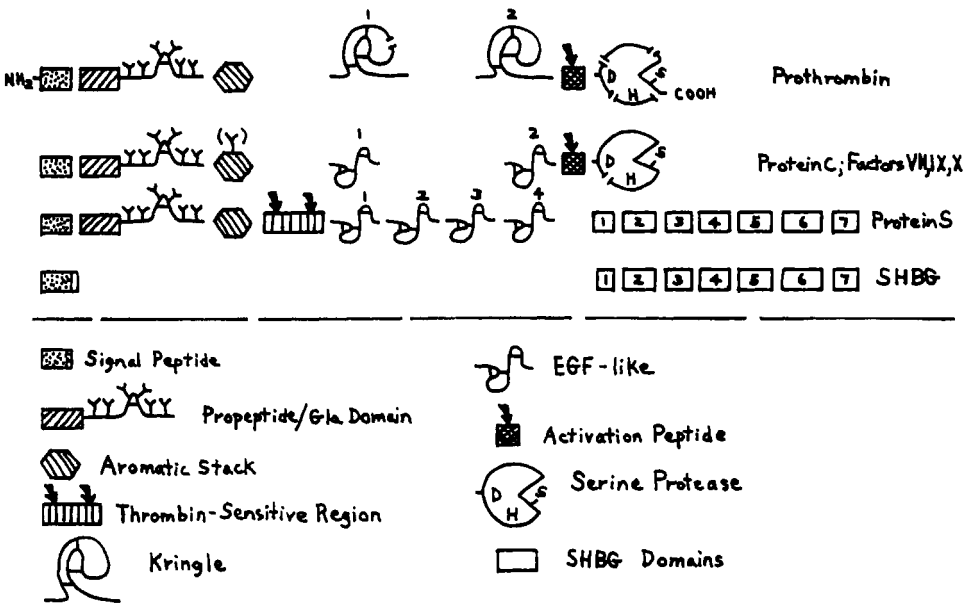


FIGURE 3: Exonic domains of protein S and homologous proteins. Shown schematically are the exonic domains in protein S and homologous proteins. Exon sizes are not shown to scale except for the sex hormone binding globulin (SHBG) exons in SHBG and protein S. Note that the propeptide and gla regions are located in one exon. Kringle 1 of prothrombin is composed of two exons. The serine protease domain in prothrombin is made up of four exons, in contrast to the case for protein C and factors VII, IX, and X where only two exons are involved. Details of the individual domains have been presented elsewhere as noted in the text.

most a truncated polypeptide containing only the leader peptide, the Gla domain, and a portion of the hydroxy amino acid rich thrombin-sensitive region of protein S. Consequently, we consider the second copy to be a pseudogene in the sense that it would not be expected to result in any product resembling plasma protein S.

Comparison of the nucleotide sequences for the two genes indicates that the duplication of the two genes may be a relatively recent event. Overall, the two genes show ~4% substitution, as compared to 18% for the six unique "Alu" sequences (data not presented) with one another and the consensus sequence, suggesting that the gene duplication occurred much later than the dispersal of Alu sequences into the human genome. Another indicator of the length of time since gene duplication is based upon the level of codon (nucleotide) substitution between the two genes. The average nucleotide substitution rate derived from comparison of mammalian protein amino acid substitutions is $\sim 1 \times 10^{-9}$ nucleotide substitutions per year per nucleotide position (King & Jukes, 1969; Fitch & Langley, 1976). At this rate, with a nucleotide substitution level in the coding portions of protein S of 3.2%, the evolutionary divergence time is approximately 30 million years. Because of great variation in the mutation rate of individual proteins and the archeological time clock upon which the average rate is based, the time of divergence of the two protein S genes can only be considered as a rough estimate. It is, however, in generally good agreement with data presented by Ploos van Amstel et al. (1990).

If one assumes that for an expressed gene the nucleotide sequence will be more conserved in regions corresponding to the cDNA than in intervening regions, one can judge as to whether two genes (one being the pseudogene in the case of protein S) have both been expressed and subject to evolutionary selective pressure during the time after their duplication. The common cDNA portions of the protein S genes have a 3.2% nucleotide substitution level, whereas the intronic regions have a 4.6% level of substitution. These data suggest that shortly after gene duplication ($\sim 10 \times 10^6$ years, based upon a divergence time of 30×10^6 years) the PSβ gene became inactivated and no longer expressed functional protein S.

The homology of the vitamin K dependent plasma coagulation factors and the similarity of the genes have been re-

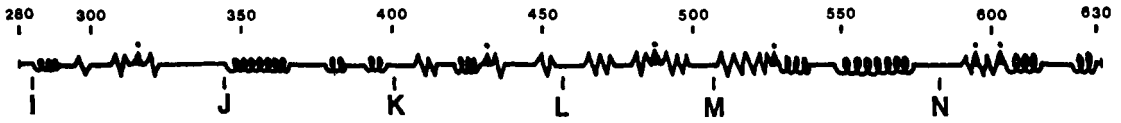


FIGURE 4: Secondary structure prediction for the carboxyl-terminal half of protein S. Amino acid numbering of the mature protein is shown above the predicted structure. Secondary structure prediction is by the method of Chou and Fasman (1978). Predicted α-helix is shown by a coil, β-strand by the sawtooth, β-turn by closed the dotted triangle, and irregular structure by the horizontal line. The position of introns (capital letters) is shown below the predicted structure.

viewed extensively elsewhere (Rogers, 1985; Patthy, 1985; Long, 1986). Protein S is a member of this family, and this report establishes the identity of location and type of the first seven introns (Table I) with the other vitamin K dependent proteins. The carboxy-terminal half of protein S has also been shown by amino acid sequence comparison (Baker et al., 1987; Gershagen et al., 1987; Long, 1988; Petra et al., 1988) to be homologous with plasma androgen binding protein also referred to as sex hormone binding globin (SHBG). As shown in Table I, the remaining seven introns are of identical position and type as those found in the gene for human SHBG (Gershagen et al., 1989). These results suggest that the two genes represent a gene family analogous to, but distinct from, the vitamin K dependent plasma protein genes and that the protein S gene is a mosaic combination of the two gene families. A related point is the location of Alu repeat sequences in introns J and L of the protein S genes, in contrast to the human SHBG gene which has an Alu sequence in the intron corresponding to intron M (Gershagen et al., 1989). This observation suggests that gene duplication resulting in the SHBG gene and the protein S gene occurred either prior to dispersal of Alu sequences into the human genome or subsequent to intronic recombinational events in the two genes.

The exonic domains in the amino-terminal half of protein S, as well as for other vitamin K dependent plasma proteins, are shown schematically in Figure 3. The structural and functional domains of the carboxy-terminal half of protein S and the homologous androgen binding proteins, also shown in Figure 3, are less well characterized. A prediction of secondary structure (Figure 4) by the method of Chou and Fasman (1978) does predict that all of the introns would fall within irregular or turn regions, consistent with the expected location of structural domain junctions. The complete agreement of location and type of introns for protein S and sex hormone binding globulin along with the location of introns relative to predicted secondary structure suggests that their location may reflect actual junctions, thereby providing a useful way to view these biomolecules for the interpretation of other experimental data. In this regard, the predicted β-strand for protein S residues 509–526 is of interest as this is the region corresponding to the proposed steroid binding site in androgen binding protein (Walsh et al., 1986; Petra et al., 1988). In the case of androgen binding protein, this region is also predicted to be a β-strand and contains eight alternating leucyl residues, providing a uniform hydrophobic "landing strip" for the steroid molecule. In protein S, this region does not contain the alternating leucine motif but instead contains several polar residues in the corresponding alternating positions at the carboxyl-terminal end, possibly explaining the apparent lack of steroid binding to protein S (Long, Que, and Petra, unpublished results; Stenflo and Fernlund, unpublished results).

Approximately 0.7 kb of genomic sequence upstream from the cDNA 5′ end has been sequenced in an effort to identify possible transcriptional start sites and promoter elements. Analysis of the sequence (Figure 2) does not reveal a region resembling the "TATA" box described by Breathnach and Chambon (1981) and found for many constitutively expressed eukaryotic genes. There are, however, several clustered GC-rich segments approximately 30–80 nucleotides upstream from the start of the cDNA sequence. These segments resemble the transcription factor SP1 binding site consensus sequence described by Gidoni et al. (1984) and subsequently found in several eukaryotic genes as an alternative regulatory motif to the TATA box [for a review, see Wasylyk (1988)]. An alternative interpretation regarding the 5′ upstream sequence, based upon the data of Ploos van Amstel et al. (1990), is that it is part of a 5′ noncoding intron, as has also been reported for protein C (Plutsky et al., 1986).

ACKNOWLEDGMENTS

Sharing of manuscripts prior to submission for publication on the protein S genes from the laboratories of Drs. R. M. Bertina and J. Stenflo is gratefully acknowledged. The secretarial assistance of Lisa McNaney is also acknowledged.

Registry No. DNA (human glycoprotein S gene coding region), 107371-17-5; glycoprotein S (human precursor protein moiety reduced), 107371-37-9; glycoprotein S (human protein moiety reduced), 107371-38-0.

REFERENCES

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

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

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

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

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

Breathnach, R., Benoist, C., O'Hare, K., Gannon, F., & Chambon, P. (1978) *Proc. Natl. Acad. Sci. U.S.A. 75*, 48–53.

Chou, P. Y., & Fasman, G. D. (1978) *Adv. Enzymol. Relat. Areas Mol. Biol. 47*, 45–148.

Dahlback, B. (1983) *Biochem. J. 209*, 847–856.

Dahlback, B. (1986) *J. Biol. Chem. 261*, 12022–12027.

Degan, S. J. F., & Davie, E. W. (1987) *Biochemistry 26*, 6165–6177.

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

Devereux, J., Haeberli, P., & Smithies, O. (1984) *Nucleic Acids Res. 12*, 387–395.

DiScipio, R. G., & Davie, E. W. (1979) *Biochemistry 18*, 899–904.

Edenbrandt, C.-M., Lundwall, A., Wydro, R., & Stenflo, J. (1990) *Biochemistry* (third of three papers in this issue).

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

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

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

Fienberg, A. P., & Vogelstein, B. (1983) *Anal. Biochem. 132*, 6–13.
Fitch, W. M., & Langley, C. H. (1976) *Fed. Proc. 35*, 2092–2097.
Foster, D. C., Yoshitake, S., & Davie, E. W. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 4673–4677.
Geddes, V. A., Le Bonniec, B. F., Louie, G. V., Brayer, G. D., Thompson, A. R., & MacGillivray, R. T. A. (1989) *J. Biol. Chem. 264*, 4689–4697.
Gershagen, S., Fernlund, P., & Lundwell, A. (1987) *FEBS Lett. 220*, 129–135.
Gershagen, S., Lundwall, A., & Fernlund, P. (1989) *Nucleic Acids Res. 17*, 9245–9258.
Gidoni, D., Dynan, W. S., & Tjian, R. (1984) *Nature 312*, 409–413.
Hoskins, J., Norman, D. K., Beckmann, R. J., & Long, G. L. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 349–353.
King, J. L., & Jukes, T. H. (1969) *Science 164*, 788–798.
Leytus, S. P., Foster, D. C., Kurachi, K., & Davie, E. W. (1986) *Biochemistry 25*, 5098–5102.
Long, G. L. (1986) *Cold Spring Harbor Symp. Quant. Biol. 51*, 525–529.
Long, G. L. (1988) 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, A., Dackowski, W., Cohen, E., Shaffer, M., Mahr, A., Dahlback, B., Stenflo, J., & Wydro, W. (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. W. (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.
Patthy, L. (1985) *Cell 41*, 657–663.
Petra, P. H., Que, B. G., Namkung, P. C., Ross, J. B. A., Charbonneau, H., Walsh, K. A., Griffin, P. R., Shabanowitz, J., & Hunt, D. A. (1988) *Ann. N.Y. Acad. Sci. 536*, 10–24.
Ploos van Amstel, J. K., Van der Zanden, A. L., Bakker, E., Reitsma, P. H., & Bertina, R. M. (1987) *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., Huisman, M. V., Reitsma, P. H., Cote, J. W., & Bertina, R. M. (1989) *Blood 73*, 479–483.
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).
Plutsky, J., Hoskins, J. A., Long, G. L., & Crabtree, G. R. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 546–560.
Rogers, J. (1985) *Nature 315*, 458–459.
Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 5463–5467.
Schmidel, D. K., Tatro, A. V., Broxson, E. H., Comp, P. C., Marlar, R. A., & Long, G. L. (1989) *Blood 74 (Suppl. 1)*, 174a.
Sharp, P. A. (1981) *Cell 23*, 643–646.
Sharp, P. A. (1987) *Science 235*, 766–771.
Smith, G. E., & Summers, M. D. (1980) *Anal. Biochem. 109*, 123–129.
Solymoss, S., Tucker, M. M., & Tracy, P. B. (1988) *J. Biol. Chem. 263*, 14884–14890.
Stern, D. M., Brett, J., Harris, K. W., & Nawroth, P. (1986) *J. Cell Biol. 102*, 1971–1978.
Suzuki, K., Nishioka, J., & Hashimoto, S. (1983) *J. Biochem. 94*, 699–705.
Walker, F. J. (1981) *J. Biol. Chem. 256*, 11128–11131.
Walsh, K. A., Titani, K., Takio, K., Kumar, S., Hayes, R., & Petra, P. H. (1986) *Biochemistry 25*, 7584–7590.
Wasylyk, B. (1988) *CRC Crit. Rev. Biochem. 23*, 77–120.
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.
Zhang, H., Scholl, R., Browse, J., & Somerville, C. (1988) *Nucleic Acids Res. 16*, 1220.