Szebenyi, D. M. E., & Moffat, K. (1986) *J. Biol. Chem. 261*, 8761–8777.
Szebenyi, D. M. E., Obendorf, S. F., & Moffat, K. (1981) *Nature (London) 294*, 327–332.
Tsai, M.-D., Drakenberg, T., Thulin, E., & Forsén, S. (1987) *Biochemistry 26*, 3635–3643.
Vogel, H. J., & Forsén, S. (1987) in *Biological Magnetic Resonance* (Berliner, L. J., & Reuben, J., Eds.) Plenum, New York (in press).
Vogel, H. J., Drakenberg, T. D., Forsén, S., O'Neil, J. D. J., & Hoffmann, T. (1985) *Biochemistry 24*, 3870–3876.
Wasserman, R. H., & Fullmer, C. S. (1982) in *Calcium and Cell Function* (Cheung, W. Y., Ed.) Vol. 2, pp 175–216, Academic, New York.
Wasserman, R. H., Fullmer, C. S., & Taylor, A. N. (1978) in *Vitamin D* (Lawson, E. E. M., Ed.) pp 133–166, Academic, New York.
Weber, G. (1975) *Adv. Protein Res. 29*, 1–83.

Nucleotide Sequence and Organization of the Human S-Protein Gene: Repeating Peptide Motifs in the "Pexin" Family and a Model for Their Evolution

Dieter Jenne*,‡,§ and Keith K. Stanley§

Institute of Medical Microbiology, Justus-Liebig-University in Giessen, 6300 Giessen, FRG, and European Molecular Biology Laboratory, 6900 Heidelberg, FRG

Received March 11, 1987; Revised Manuscript Received June 4, 1987

ABSTRACT: The S-protein/vitronectin gene was isolated from a human genomic DNA library, and its sequence of about 5.3 kilobases including the adjacent 5′ and 3′ flanking regions was established. Alignment of the genomic DNA nucleotide sequence and the cDNA sequence indicated that the gene consisted of eight exons and seven introns. The intron positions in the S-protein gene and their phase type were compared to those in the hemopexin gene which shares amino acid sequence homologies with transin and the S-protein. Three introns have been found at equivalent positions; two other introns are very close to these positions and are interpreted as cases of intron sliding. Introns 3–7 occur at a conserved glycine residue within repeating peptide segments, whereas introns 1 and 2 are at the boundaries of the Somatomedin B domain of S-protein. The analysis of the exon structure in relation to repeating peptide motifs within the S-protein strongly suggests that it contains only seven repeats, one less than the hemopexin molecule. A very similar repeat pattern like that in hemopexin is shown to be present also in two other related proteins, transin and interstitial collagenase. An evolutionary model for the generation of the repeat pattern in the S-protein and the other members of this novel "pexin" gene family is proposed, and the sequence modifications for some of the repeats during divergent evolution are discussed in relation to known unique functional properties of hemopexin and S-protein.

Several biological functions have been ascribed to a single 75-kilodalton (kDa)[1] glycoprotein of human plasma (Jenne & Stanley, 1985). In serum culture media, this protein, called vitronectin or serum spreading factor, is the principal agent mediating the adhesion and spreading of cells and facilitating their proliferation on surfaces in vitro (Barnes et al., 1980; Hayman et al., 1985). When complement is activated, the same protein, called complement S-protein, participates in the fluid phase assembly of the terminal complement proteins to form the soluble SC5b-9 complex; however, it is not found in the C5b-9 membrane complex that is assembled only on a lipid bilayer and generates transmembrane channels (Kolb & Müller-Eberhard, 1975; Bhakdi et al., 1976; Bhakdi & Tranum-Jensen, 1983). S-Protein incorporation into nascent C5b-7 complexes prevents their membrane attachment and the formation of cytolytically active complement complexes on lipid membranes after successive addition of C8 and C9. Therefore, S-protein/vitronectin appears to protect innocent bystander cells from membrane damage by nascent complement complexes (Podack et al., 1977, 1978).

S-Protein also forms stable ternary complexes with antithrombin III and thrombin during thrombin inactivation. It binds through a cryptic site in antithrombin III and exerts a net protective effect on thrombin toward its inactivation by the inhibitor (Jenne et al., 1985a; Ill & Ruoslahti, 1985). The molecule further possesses a heparin and glycosaminoglycan binding site (residues 348–380) (Suzuki et al., 1984) which is not exposed in the native plasma protein (Hayashi et al., 1985; Barnes et al., 1985). Binding to heparin-like molecules contributes to its procoagulatory role in the clotting process (Preissner et al., 1985).

The Arg-Gly-Asp (R-G-D) peptide sequence of S-protein/vitronectin (residues 45–47) interacts specifically with a cell surface receptor, the vitronectin receptor (Pytela et al., 1985). The molecule can therefore function as a cross-linker between cells and the extracellular matrix by virtue of its two different binding sites.

S-Protein has evidently specialized for a variety of binding functions which reside in different regions of the protein.

* Address correspondence to this author at the Institut de Biochimie, Université de Lausanne, CH-1066 Epalinges, Switzerland. Supported in part by a fellowship from the European Molecular Biology Organization.
‡ Justus-Liebig-University in Giessen.
§ European Molecular Biology Laboratory.

[1] Abbreviations: SDS, sodium dodecyl sulfate; HTF, *Hpa*II tiny fragments; CAT, chloramphenicol acetyltransferase; NaH_2PO_4, sodium dihydrogen orthophosphate; pEX, expression plasmid (Stanley & Luzio, 1984); kDa, kilodalton(s); bp, base pair(s); TK, thymidine kinase; kb, kilobase(s).

0006-2960/87/0426-6735$01.50/0

Analysis of the primary structure of the hemopexin and S-protein suggests that these molecules are composed of repeating structural motifs related to each other. A 10-repeat model was therefore inferred for the hemopexin (Altruda et al., 1985) and the S-protein (Stanley, 1986), although the homologies within the protein sequences are only evident for eight and six peptide segments, respectively. In this study, we report on the exon–intron structure of the human S-protein gene compared to that of the hemopexin gene in order to disclose the genetic basis for these amino acid repeats. Furthermore, we describe homologous repeating peptide segments in two other related proteins, transin and interstitial collagenase, indicating that these repeating peptide motifs have arisen by divergent evolution.

Materials and Methods

Isolation of Genomic S-Protein Clones. A human genomic library of blood leukocyte DNA, constructed in the cosmid vector pcos 2EMBL (Poustka et al., 1984), was kindly provided by A. Frischauf. Colony screening and plasmid isolation were done according to standard procedures (Maniatis et al., 1982).

RNA Blot Analysis. Poly(A+) RNA from human liver was prepared by ethanol precipitation from guanidine hydrochloride solutions (Deeley et al., 1977) and affinity chromatography on oligo(dT)–cellulose. The RNA was denatured with glyoxal and size-separated by electrophoresis in a 1.2% agarose gel (Carmichael & McMaster, 1980). The RNA was transferred onto nylon membranes in high salt solutions according to the manufacturer's instructions (NEN Research Products) and UV-cross-linked (Church & Gilbert, 1984).

DNA Blot Analysis. Total human DNA was extracted from peripheral blood lymphocytes following standard procedures (Maniatis et al., 1982). Restriction fragments were fractionated on 0.8% agarose gels and immobilized on nylon membranes according to Church and Gilbert (1984).

Hybridization Analysis. A complete cDNA clone for S-protein, S203 (Jenne & Stanley, 1985), was used as a probe in RNA, DNA, and in situ colony hybridizations. High specific activity of the probe was achieved by using a random oligo-primed labeling procedure (Feinberg & Vogelstein, 1983). Hybridization and washing were conducted under stringent conditions.

Primer Extension Analysis. A 132 bp *Nco*I, *Sac*I fragment of the cDNA was gel-purified, end-labeled with *Escherichia coli* DNA polymerase (Klenow fragment), and used as a primer for the reverse transcriptase reaction according to published procedures (Maniatis et al., 1982).

DNA Sequence Analysis. Genomic fragments were subcloned in the M13 derivative mp18 in opposite orientations. Overlapping M13 subclones were generated by a DNase deletion method (Labeit et al., 1987) and sequenced by the dideoxy chain termination procedure (Sanger et al., 1977).

Homology Comparison between Aligned Amino Acids. Protein sequences were aligned by using a matrix comparison procedure based on six physical residue characteristics as well as on the Dayhof relatedness odds matrix (Barker et al., 1978; Altruda et al., 1985; Argos, 1987). Longer stretches of aligned homologous sequences were found with this procedure by using multiple window lengths in the range of 7–35 amino acids, and the homologies were subsequently quantified by calculating the mean correlation coefficient over the aligned amino acids only.

CAT Assays. A *Bam*HI/*Bal*I genomic fragment of the 5′ region of the gene (Figure 2) was inserted into a special pTK-CAT vector (Miksicek et al., 1986) upstream of the thymidine kinase (TK) promoter in both orientations, and the chloramphenicol acetyltransferase (CAT) activity was measured after transient infection of Hep G2 and HeLa cells. Control plasmids were the parental plasmid pTK-CAT which gives a low level of CAT enzyme activity and a pTK-CAT derivative containing the intact SV 40 enhancer upstream of the TK promoter for high expression levels of the CAT enzyme.

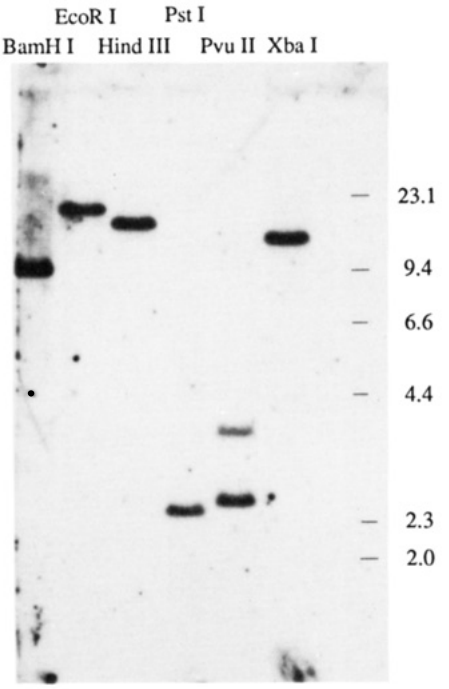

FIGURE 1: Southern blot hybridization of total human DNA isolated from peripheral blood lymphocytes using a S-protein cDNA probe. Five micrograms of DNA digested with the restriction enzymes *Bam*HI, *Eco*RI, *Hin*dIII, *Pst*I, *Pvu*II, and *Xba*I, respectively, was separated on a 0.8% agarose gel, treated with 0.25 M HCl for 20 min, and transferred to a nylon membrane in 0.5 M NaOH/1.5 M NaCl. After UV-cross-linking, the nylon filter was prehybridized and hybridized in 0.5 M NaH_2PO_4/7% SDS (pH 7.2). The final washes were done in 20 mM NaH_2PO_4/1% SDS (pH 7.2). λ DNA, digested with *Hin*dIII and end-labeled, was used as the size marker (shown in kilobases).

Results

Isolation of the Human S-Protein Gene. Total human genomic DNA was analyzed with several different restriction enzymes in a Southern blot experiment using a full-length cDNA probe. Digestions with four enzymes, *Bam*HI, *Eco*RI, *Hin*dIII, and *Xba*I, revealed in each case one unique fragment to which the probe hybridized under stringent conditions, showing that a single copy of the gene resides in the human haploid genome. The size of this gene cannot exceed the 9.4 kb *Bam*HI fragment (Figure 1).

About 250 000 colonies of a human genomic library from blood lymphocyte DNA were screened with a full-length cDNA probe for the S-protein. Eight independent cosmid clones were obtained after colony purification. Southern blot analysis of these clones showed that the entire gene was contained in three *Kpn*I fragments of 2.5, 1.4, and 4.6 kb. No differences in hybridizing fragments were seen among all eight clones, thus providing additional evidence for a single-copy gene. Since the restriction enzyme *Kpn*I recognizes two sites in the cDNA sequence of the S-protein, three genomic *Kpn*I fragments represent the minimal number to be expected. The 9.4 kb *Bam*HI fragment was subcloned into and amplified in the pUC 19 vector. This fragment was further mapped with single and double restriction enzyme digests using *Cla*I and

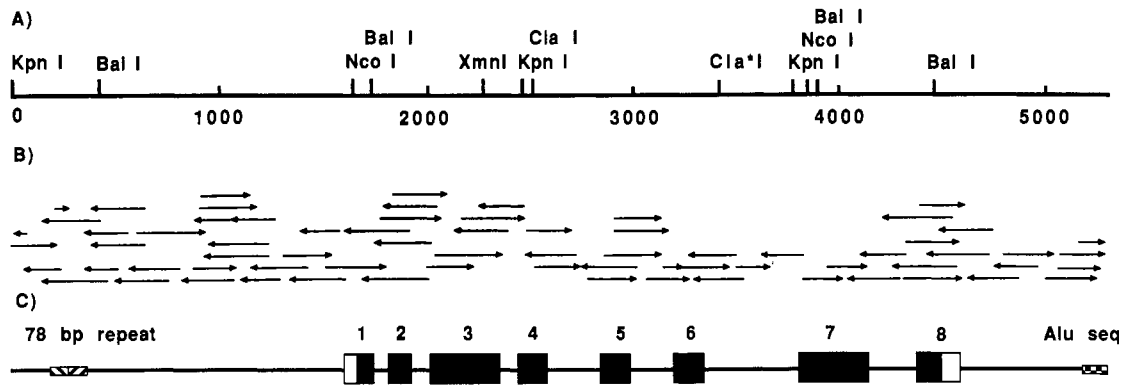

FIGURE 2: (A) Restriction sites. Note that cleavage of the second *Cla**I recognition sequence is inhibited by *E. coli dam* methylation. (B) Sequence assembly for the S-protein gene. The two *Kpn*I fragments and the 3.3 kb *Kpn*I/*Bam*HI fragment were cloned in M13 in both orientations, and sets of ordered deletion subclones were sequenced by dideoxy chain termination as shown by arrows. (C) Structure of the human S-protein gene shown schematically. Solid boxes present coding exons and open boxes untranslated regions. The position of the 78 bp long direct repeat in the 5′ region and the location of an Alu sequence are indicated as hatched boxes. The Alu sequence copy corresponds to the right monomer of the human Alu consensus sequence.

*Xmn*I in combination with *Kpn*I. The former two enzymes cut once in the genomic *Bam*HI fragment as well as in the cDNA sequence and allowed the arrangement of three *Kpn*I fragments (Figure 2).

Nucleotide Sequence of the S-Protein Gene. Nucleotide sequences were determined for the 1.4 and 2.5 kb *Kpn*I fragments and partially for the 3.3 kb *Kpn*I/*Bam*HI genomic fragment on both strands in the M13 phage vector mp19. Overlapping M13 subclones were generated by DNase deletion (Labeit et al., 1987) in both orientations (Figure 2). The transcribed nucleotide sequence for the S-protein covers about 3 kb of genomic DNA. Flanking regions of 1.6 kb at the 5′ terminus and 0.7 kb at the 3′ terminus were sequenced in addition. Alignment of the genomic sequence with that of the cDNA revealed the position and size of exons and introns. The gene consists of eight exons interrupted by introns in the size range of 75 bp (intron 2) to 459 bp (intron 6). All exon/intron junctions conform to the GT-AT rule (Breathnach & Chambon, 1981; Mount, 1982). Six introns are positioned in phase 1, splitting the codon between the first and second nucleotide. Intron 4 occurs between codons in phase 0. Phase 2 introns are not present in the gene.

Inspection of the sequence upstream of the 5′ end of the cDNA revealed neither a TATA nor a CCAAT box at an appropriate position; however, a possible CAATCT (position 1126) and TAAATAAA (position 1158) box region (underlined in Figure 3) as well as good consensus donor and acceptor splice sites is found within a region 512 bp upstream of the cDNA sequence start. The resulting further 5′ exon (underlined in Figure 3) would contribute only 10 untranslated base pairs to the transcribed mRNA. However, primer extension experiments (Figure 4) and Northern blots with genomic fragments 3.7 kb upstream of the cDNA start could not provide evidence for the existence of such a 5′ exon. The length of the primer-extended products (Figure 4, see arrow) rather suggests that the transcription initiation site is located 2 bp upstream of position 1636 where the cDNA sequence starts. This putative mRNA start site (caggcATCAGAG) would be compatible with the base frequencies around transcription start points derived from other genes (Breathnach & Chambon, 1981).

A novel 78 bp perfect direct repeat (see Figure 2, underlined in Figure 3) is located at position 220. It has a high CG content and falls in a region of six clustered *Hpa*II sites, possibly indicating an HTF island at the 5′ boundary of the gene (Bird, 1986). Because of the resemblance of this region to viral long terminal repeats, we inserted a 414 bp fragment including the 78 bp repeat into the plasmid pTK-CAT immediately upstream of the HSV TK promoter. No increased levels of CAT expression were measured, however, in transfected HepG2 cells. This nucleotide sequence repeat therefore does not appear to be functionally related to the simian virus 40 enhancer.

At the 3′ end of the gene, there is only one polyadenylation site 64 bases downstream from the translation stop codon; 35 bases further downstream, a sequence of GACTG may correspond to a signal sequence directing the cleavage of the primary transcript at the 3′ end before polyadenylation (Berget, 1984). In the 3′ flanking region of the gene starting at position 5144, a partial copy of a human Alu sequence—the right half (underlined in Figure 3)—is interspersed (Schmid & Jelinek, 1982). It is not included in the RNA transcription unit for the S-protein. The DNA sequence for the coding region of the gene and the cDNA sequence for the S-protein differ only in the triplet coding for Ala-347 (GCC) instead of Thr-347 (ACC). The genomic sequence at this position thus agrees with the vitronectin cDNA sequence and the protein sequence data (Suzuki et al., 1984, 1985).

Correspondence between Exons of the S-Protein Gene and Repeating Peptide Motifs in S-Protein and Hemopexin. Analysis of the self-homology of S-protein using physical residue parameters which reflect protein folding has suggested that the protein may have evolved from a short peptide segment repeated 10 times in the present day structure of the protein. The internal homologies, however, are discernible only for six segments of the S-protein (Stanley, 1986). The exon/intron structure of both the hemopexin and S-protein genes supports the hypothesis of a repeat pattern in that five introns of the S-protein gene and six introns of the hemopexin gene fall at corresponding positions within repeated peptide motifs (Figure 6).

The original window of the repeat was described with a conserved glycine residue in a hydrophobic environment at the amino-terminal end and an Asp within a strongly conserved hydrophobic pentapeptide sequence at the carboxy-terminal end. In view of the exon/intron pattern of the hemopexin and S-protein genes which clearly shows that the repeated motif starts with the FDAFT in exon 3, we have now redefined the window of the repeated peptide motifs (Figure 6): therefore, in order to generate, e.g., four complete amino acid repeats in half of the molecule, as in the case of hemopexin, part of a fifth exon is required to encode the second portion of con-

FIGURE 3: Complete nucleotide sequence of the human S-protein gene including 5′ and 3′ flanking regions. Underlined sequences are from 5′ to 3′: the 78 bp repeat, a putative CAAT and TATA box, a hypothetical 5′ mini-exon, the cDNA sequence start, and the Alu sequence homology. The amino acid sequences encoded by the exons are shown beneath the DNA sequences in a three-letter code.

served hydrophobic residues following the intron split glycine. The location of the interrupted amino acid within a highly conserved segment and not at the boundary of the repeated peptide motif is peculiar and contrary to the view that exons evolved to contain discrete structural units.

A comparison of the homologous exons of the hemopexin and S-protein genes is shown in Figure 5. The first exon of the S-protein encompasses the complete leader peptide and

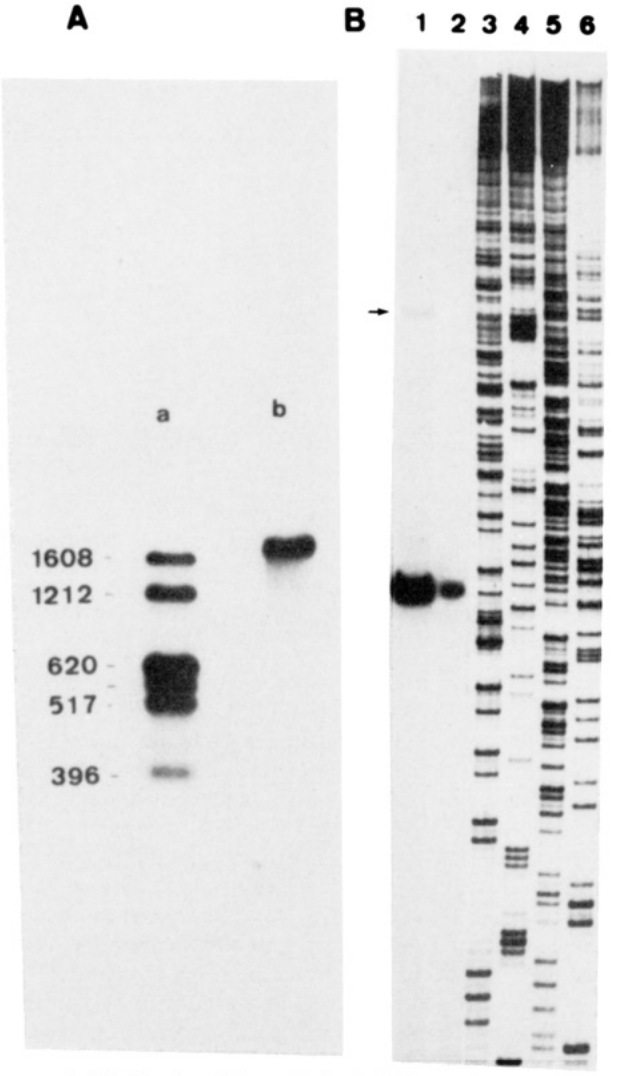

FIGURE 4: (A) Northern blot analysis of the S-protein mRNA. One microgram of human liver poly(A+) RNA was denatured with glyoxal, size-separated on a 1.2% agarose gel, and transferred to a nylon membrane. The filter was hybridized under the same conditions as in the Southern blot experiment. (a) Molecular weight markers (pEX, *Hin*fI fragments, sizes given in base pairs); (b) human poly(A+) RNA. (B) 5′ end of the S-protein mRNA. Human liver poly(A+) RNA (3 μg) was extended by using a cloned cDNA fragment as a primer. The extension products were separated on a 10% polyacrylamide/6 M urea sequencing gel. Lane 1, primer extension products [(→) 189 bp]; lane 2, the end-labeled primer (125 bp); lanes 3–6, the four sequencing reactions for a known C9 M13 clone.

two amino acids of the mature protein. The second exon (residues 3–42) encodes the Somatomedin B peptide except for the two carboxy-terminal residues. This is evidence for an independent functional protein domain in addition to the fact that all eight cysteine residues in this region are interlinked by disulfide bonds. Exon 3 (residues 44–157) carries the codons for the cell receptor recognition determinant (R-G-D) at its amino-terminal boundary only one residue away from an interrupting intron. At the carboxy-terminal end of this exon is the first repeat unit homologous with exon 3 and exon 7 of the hemopexin gene (see S3, H3, and H7 in Figure 5). Exons 4, 5, and 6 clearly represent homologous peptide repeats, and in each case, the introns are found in predicted strand–turn–strand configurations nearby or within a glycine codon for both genes (see S4–S6 and H4–H6; Altruda et al., 1985). The phase type of the introns which split these conserved peptide motifs is different for the intron/exon boundaries of exons 3–6 but corresponds precisely between the genes. Exon

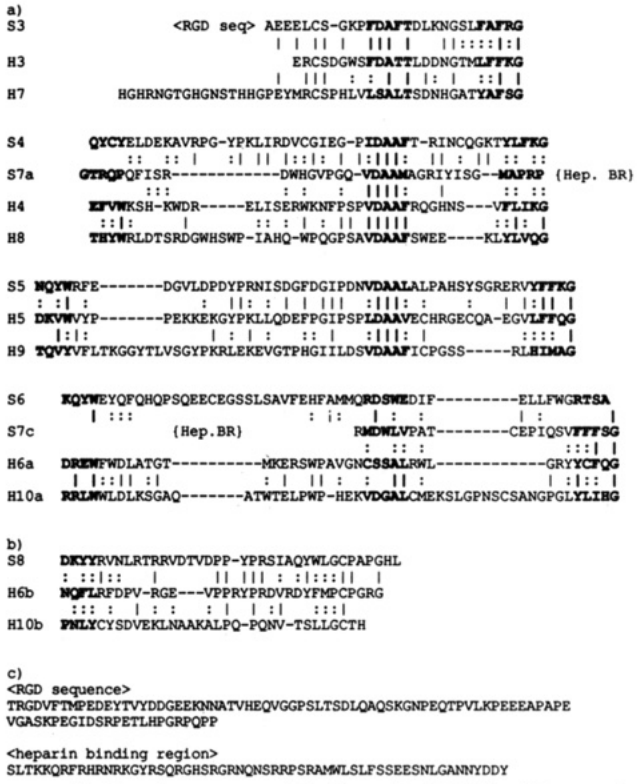

FIGURE 5: Homologous exons of the S-protein (S3–S8) and hemopexin (H3–H10) genes. Exon-encoded protein segments were aligned in order to show the close correspondence between genetic units and repeating peptide units. (A) Exons coding for "pexin" repeats; note that exon 7 of the S-protein is composed of S7a, <Hep. BR> (the heparin binding region of the S-protein), and S7c and codes for three repeats and that exons 6 and 10 of hemopexin (H6, H10) are divided into two portions (H6a, H6b, H10a, H10b). (b) Proline-rich regions of S8 and at the carboxy-terminal end of exons H6 (H6b) and H10 (H10b). (c) Sequence of <Hep. BR> (heparin binding region) and of <RGD seq> in detail. Solid lines indicate identities, and dashed lines indicate conserved amino acids according to the scheme: (RK); (DENQ); (ST); (PG); (MCYWFLAIVH). Bold characters have been used to emphasize the typical features of peptide repeats.

7 of S-protein appears to span two typical repeat units separated by the unique heparin binding domain. The first one shows closest homology with the repeat of exon 4 (data not shown); the carboxy-terminal one corresponds to amino acid sequences encoded by exons 6 (H6a) and 10 (H10a) of hemopexin. Exon 8 has no counterpart in the amino-terminal half of the S-protein; nevertheless, the peptide segment resembles a proline-rich sequence found in exons 6 and 10 of the hemopexin molecule (Figure 5).

Repeating Peptide Motifs in Other Members of the Pexin Gene Family. Searching a protein sequence data bank (National Biomedical Research Foundation, Bethesda, MD) with the S-protein sequence identified two other proteins, rat transin (Matrisian et al., 1985) and human skin fibroblast collagenase (Goldberg et al., 1986), which share short homologous regions with human S-protein in addition to hemopexin. Whereas an overall homology of 25% for the second half of transin to hemopexin has already been reported (Matrisian et al., 1986), we now identified a very similar pattern of repeating units in transin and in collagenase like those in S-protein and hemopexin. Dot plot matrix comparisons with the S-protein indicated four strongly homologous repeating motifs in the carboxy-terminal half and at least one well-conserved repeating motif in the amino-terminal half of the transin and collagenase molecule (TR3, CR3; see Figure 6). Examination of the transin and collagenase sequences by eye suggests the existence of a further three repeat sequences in the amino-terminal half

Repeat	Residues No.		Correl.Coeff. Compar. to HR2
S-protein			
SR1	142-165	FDAFTDLK (X)$_4$ FAFRG QYCYELD	0.42
SR2	187-212	IDAAFT-R (X)$_7$ YLFKG NQYWRFE	0.65
SR3	235-264	VDAALALP (X)$_{10}$YFFKG KQYWEYQ	0.60
SR4	291-316	RDSWEDIF (X)$_6$ RTSAG TRQPQFI	-0.12
SR6	327-351	VDAAMAGR (X)$_5$ MAPRP SLTKKQR	0.37
SR7	377-401	PSRAMWLS (X)$_5$ ESNLG ANNYDDY	-0.17
SR8	403-429	MDWLVPAT (X)$_7$ FFFSG DKYYRVN	0.57
Haemopexin			
HR1	33- 56	FDATTLDD (X)$_4$ LFFKG EFVWKSH	0.45
HR2	74- 97	VDAAFRQG (X)$_4$ FLIKG DKVWVYP	1.00
HR3	120-148	LDAAVECH (X)$_9$ LFFQG DREWFWD	0.61
HR4	165-187	CSSALRWL (X)$_3$ YCFQG NQFLRFD	0.51
HR5	240-263	LSALTSDN (X)$_4$ YAFSG THYWRLD	0.51
HR6	285-307	VDAAFSWE (X)$_3$ YLVQG TQVYVFL	0.63
HR7	337-361	VDAAFICP (X)$_5$ HIMAG RRLWWLD	0.57
HR8	380-414	VDGALCME (X)$_{15}$YLIHG PNLYCYS	0.59
Collagenase			
CR1	12- 59	VDLVQKYL (X)$_{29}$QEFFG LKVTGK-	0.22
CR2	60- 93	PDAETLKV (X)$_{14}$VLTEG NPRWEQT	0.26
CR3	111-154	VDHAIEKA (X)$_{24}$SFVRG DHRDNSP	0.58
CR4	174-197	GDAHFDED (X)$_4$ NNFRE YNLHRVA	0.24
CR5	265-287	FDAITTIR (X)$_3$ MFFKD RFYMRTN	0.37
CR6	309-333	LEAAYEFA (X)$_5$ RFFKG NKYWAVQ	0.54
CR7	358-382	IDAALSEE (X)$_5$ YFFVA NKYWRYD	0.50
CR8	407-429	VDAVFMKD (X)$_3$ YFFHG TRQYKFD	0.45
Transin			
TR1	29- 76	MEVLQKYL (X)$_{29}$QKFLG LKMTGK-	0.10
TR2	77-110	LDSNTMEL (X)$_{14}$STFPG SPKWRKN	0.23
TR3	128-171	VDSAIERA (X)$_{24}$SFAVE EHGDFIP	0.38
TR4	191-214	GDAHFDDD (X)$_4$ DDVTG TNLFLVA	0.37
TR5	294-316	FDAVSTLR (X)$_3$ LFFKD RHFWRKS	0.46
TR6	338-352	MDAAYEVT (X)$_5$ FILKG NQIWAIR	0.61
TR7	386-410	IDAAISLK (X)$_5$ YFFVE DKFWRFD	0.43
TR8	435-457	VDAVFEAF (X)$_3$ YFFSG SSQLEFD	0.41

FIGURE 6: Peptide alignment of the conserved regions from the repeats in human hemopexin (HR 1–8), S-protein (SR1–8), collagenase (CR1–8), and rat transin (TR1–8). Amino acids interrupted by an intron between the first and second nucleotides of their coding triplet (phase 1) are marked by (▽); introns occurring between triplets (phase 0) are indicated by an arrow in the case of S-protein and hemopexin. The mean correlation coefficients over six physical residue characteristics between the aligned amino acids of each repeat and of repeat 2 of hemopexin (HR2) have been calculated and are listed on the right. Amino acids, indicated by $(X)_n$ or aligned to gaps (–), are not included in the calculations.

which, however, have diverged considerably. The most conserved regions of 31 repeating sequence segments from 4 proteins have been aligned and compared to each other by calculating the mean correlation coefficients over 6 residue physical characteristics among all possible pairs (Figure 6). The repeat unit HR2 of the hemopexin gives the highest values for the pairwise comparisons (shown in Figure 6) and therefore represents a consensus sequence for the repeating motif. As a rule, HR2 shows the highest correlation values with repeat R6 and with repeat R3 of all proteins except for those in regions which carry unique functional properties (SR6–SR7, CR1–CR4, and TR1–TR4; Matrisian et al., 1986).

Discussion

Several polypeptides of human S-protein have been observed in blood: 75K, 65K, 57K, and 5K (Somatomedin B) molecular weight species (Hayman et al., 1982, 1983; Jenne et al., 1985b; Fryklund & Sievertsson, 1978). We were interested in the mechanism by which this heterogeneity is generated especially as the cell attachment factor, fibronectin, has been shown to undergo differential splicing (Hynes, 1986). In this paper, we have shown that a single-copy gene codes for the human S-protein and gives rise to a homogeneous mRNA transcript of about 1700 bp in liver tissue and in HepG2 cells (Figure 4). Alternative splicing sites or polyadenylation sites are not found in the gene. In addition, the metabolically labeled protein from HepG2 liver cells isolated with monoclonal antibodies comigrates with the 75-kDa band of S-protein from human plasma in a Western blot (data not shown). Therefore, the most likely origin of multiple S-protein species is proteolytic processing in the circulation or in peripheral tissues after secretion.

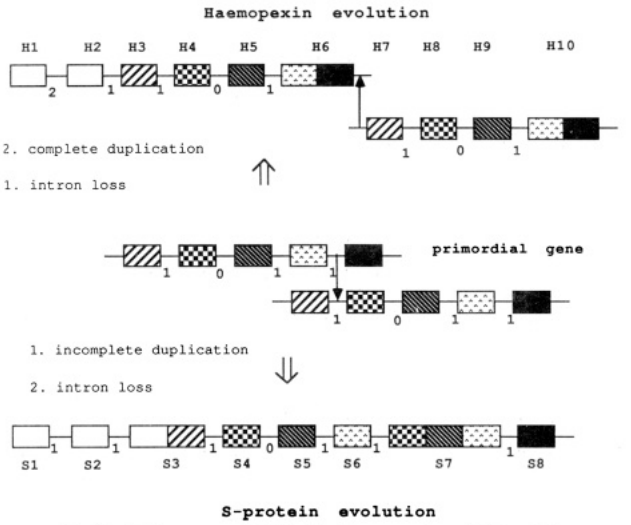

FIGURE 7: Evolutionary model for the hemopexin and S-protein genes based on the phase of introns and correlation coefficients for peptide segments encoded by corresponding exons. Exons are drawn as boxes and introns as connecting lines. Numbers below exons refer to the phase type of introns. See Discussion for more details.

Although the transcription start site for the gene has not unambiguously been determined, the size of the mRNA on Northern blots and the primer extension data are consistent with a short 5′ untranslated region. Two possibilities still exist: either transcription is initiated without the aid of a TATA box at position 1636 which fits a consensus sequence around transcription start sites or an additional small 5′ exon is present in the further upstream region. In the latter case, this missing exon might be located at positions 1985–1994, 28 bp downstream from a possible TATA box sequence.

S-Protein has been shown to bear an amino acid sequence homology with hemopexin which indicates the presence of a repeating amino acid stretch of about 38 residues. Since the window of the repeat was determined from a comparison based on physical residue parameters important in protein folding, it is likely that these repeats represent structural motifs of the protein. In agreement with this, the start point of each repeat occurs in a region of low conservation of sequence. We have been able to test this hypothesis by comparing the intron/exon structure of the gene with the amino acid repeat. While confirming the presence of a repeating peptide motif even at the genetic level, the window of the repeat in each exon is not the same as that predicted by computer. Indeed, the intron-interrupted amino acid is the most highly conserved residue in the repeating sequence. A similar situation exists in the hemopexin gene (Altruda et al., 1986).

Comparison of the S-protein gene structure with that of hemopexin can be used to formulate a model for their evolution (see Figure 7). In hemopexin, the phases of the introns and sequence comparisons suggest that the two halves of the gene originated by a gene duplication event. In each half of the molecule, there is evidence for four homologous segments in tandem followed by a proline-rich sequence fused to the last exon. A process of four elongative duplications of an ancestral

exon leading to the common primordial gene for the pexin protein family is consistent with the structural correlation coefficients between each repeating unit of exons 3–6 in hemopexin. The second half of the hemopexin molecule appears to be generated by a complete contiguous gene duplication after an intron loss between the last two exons. In the case of the S-protein, an incomplete duplication compatible with the phase type of the introns involved could account for the fact that the intron phases of exons 3–6 are identical in the S-protein and hemopexin gene and that the proline-rich linker sequence is only found once at the carboxy-terminal end of the S-protein molecule. According to our model, two introns have been deleted after the internal duplication. The proposed events are furthermore consistent with the phase type of intron 6 (phase 1) corresponding with introns 3 and 7 in the S-protein. It therefore appears that duplication to form a molecule with two halves occurred independently and by a different mechanism in hemopexin and S-protein (Figure 7).

A particularly interesting feature emerging from a comparison of the two genes is the disposition of functional sequences relative to the underlying repeated exon structure. In S-protein, the two regions with defined functional significance are the R-G-D sequence and the heparin binding site. While the latter could have evolved from an ancestral repeat unit, it appears that the Somatomedin B domain and the R-G-D-containing region at the amino-terminal boundary of repeat 1 have been inserted and fused into the 5′ end of exon 3. In the same relative exon position to the R-G-D peptide of S-protein, a histidine-rich peptide segment on exon 7 of the hemopexin gene is found. It is therefore a likely candidate for the interaction with heme iron. Exon 3 of the S-protein and exon 7 of the hemopexin, which correspond to each other (Figure 7), therefore appear to have acquired new specific functional regions during evolution.

In the S-protein, the structural modification of another repeat, the heparin binding region, may be crucial for its interaction with the negatively charged class A cysteine-rich domain present in the terminal complement proteins C8, C9, and probably C7. Since this region is implicated in the assembly of terminal complement proteins on lipid membranes leading to target cell lysis (Tschopp & Mollnes, 1986; Tschopp et al., 1986), S-protein binding to this region could account for its inhibitory function. However, the very positively charged heparin binding region (14 positively charged amino acids in a stretch of 30 residues) is not accessible in native S-protein. Therefore, it is likely that this region interacts with a negatively charged region in the native molecule. A cluster of seven negatively charged amino acids (residues 53–64) immediately following the R-G-D sequence in exon 3 may fulfill this function. Such a negatively charged region may bind to the heparin binding site of antithrombin III, thus enabling S-protein to form stable complexes with thrombin-bound antithrombin III.

We have also shown that two proteins, transin and collagenase, contain the same repeating unit as is found in hemopexin and S-protein. S-Protein, hemopexin, transin, and collagenase can therefore be regarded as members of a novel protein and gene family characterized by a unique repeating amino acid motif and the unusual position of introns within these conserved peptide motifs. Considerable modifications and sequence changes discernible for some of these repeats point to specific functional properties of these molecules.

ACKNOWLEDGMENTS

D.J. thanks Professors Wellensiek and Bhakdi for their continuing interest and support in these studies and Hans-Jürgen Thiesen for performing the CAT assays.

REFERENCES

Altruda, F., Poli, V., Restagno, G., Argos, P., Cortese, R., & Silengo, L. (1985) *Nucleic Acids Res. 13*, 3841–3859.

Altruda, F., Poli, V., Restagno, G., & Silengo, L. (1987) *Nucleic Acids Res.* (in press).

Argos, P. (1987) *J. Mol. Biol. 193*, 385–396.

Barker, W. C., Ketcham, L. K., & Dayhoff, M. O. (1978) in *Atlas of Protein Sequence and Structure* (Dayhoff, M. O., Ed.) Vol. 5, Suppl. 3, pp 359–362, National Biomedical Research Foundation, Washington, DC.

Barnes, D., Wolfe, R., Serrero, G., McClure, D., & Sato, G. (1980) *J. Supramol. Struct. 14*, 47–63.

Barnes, D. W., Reing, J. E., & Amos, B. (1985) *J. Biol. Chem. 260*, 9117–9122.

Berget, M. S. (1984) *Nature (London) 309*, 179–182.

Bhakdi, S., & Tranum-Jensen, J. (1983) *Biochim. Biophys. Acta 737*, 343–372.

Bhakdi, S., Ey, P., & Bhakdi-Lehnen, B. (1976) *Biochim. Biophys. Acta 419*, 448–457.

Bird, A. P. (1986) *Nature (London) 321*, 209–213.

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

Carmichael, G. G., & McMaster, G. K. (1980) *Methods Enzymol. 65*, 380–391.

Church, G. M., & Gilbert, W. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 1991–1995.

Deeley, R. G., Gordon, J. I., Burns, A. T. H., Mullinix, K. P., Binastein, M., & Goldberger, R. F. (1977) *J. Biol. Chem. 252*, 8310–8319.

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

Fryklund, L., & Sievertsson, H. (1978) *FEBS Lett. 87*, 55–60.

Goldberg, G. I., Wilhelm, S. M., Kronberger, A., Bauer, E. A., Grant, G. A., & Eisen, A. Z. (1985) *J. Biol. Chem. 261*, 6600–6605.

Hayashi, M., Akama, T., Kono, I., & Kashiwagi, H. (1985) *J. Biochem. (Tokyo) 98*, 1135–1138.

Hayman, E. G., Engvall, E., A'Hearn, E., Barnes, D., Pierschbacher, M., & Ruoslahti, E. (1982) *J. Cell Biol. 95*, 20–23.

Hayman, E. G., Pierschbacher, M. D., Öhgren, Y., & Ruoslahti, E. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 4003–4007.

Hayman, E. G., Pierschbacher, M. D., Suzuki, S., & Ruoslahti, E. (1985) *Exp. Cell Res. 160*, 245–258.

Hynes, R. (1986) *Annu. Rev. Cell Biol. 1*, 67–90.

Ill, C. R., & Ruoslahti, E. (1985) *J. Biol. Chem. 260*, 15610–15615.

Jenne, D., & Stanley, K. K. (1985) *EMBO J. 4*, 3153–3157.

Jenne, D., Hugo, F., & Bhakdi, S. (1985a) *Thromb. Res. 38*, 401–412.

Jenne, D., Hugo, F., & Bhakdi, S. (1985b) *Biosci. Rep. 5*, 343–352.

Kolb, W. P., & Müller-Eberhard, H. J. (1975) *J. Exp. Med. 141*, 724–735.

Labeit, S., Goody, R. S., & Lehrach, H. (1987) *Methods Enzymol.* (in press).

Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Matrisian, L. M., Glaichenhaus, N., Gesnel, M.-C., & Breathnach, R. (1985) *EMBO J. 4*, 1435–1440.

Matrisian, L. M., Leroy, P., Ruhlmann, C., Gesnel, M.-C., & Breathnach, R. (1986) *Mol. Cell. Biol. 6*, 1679–1686.

Miksicek, R., Heber, A., Schmid, W., Danesch, U., Posseckert, G., Beato, M., & Schütz, G. (1986) *Cell (Cambridge, Mass.)* *46*, 283–290.
Mount, S. M. (1982) *Nucleic Acids Res. 10*, 459–472.
Podack, E. R., Kolb, W. P., & Müller-Eberhard, H. J. (1977) *J. Immunol. 119*, 2024–2029.
Podack, E. R., Kolb, W. P., & Müller-Eberhard, H. J. (1978) *J. Immunol. 120*, 1841–1848.
Poustka, A., Rackwitz, H.-R., Frischauf, A.-M., Hohn, B., & Lehrach, H. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 4129–4133.
Preissner, K. T., Wassmuth, R., & Müller-Berghaus, G. (1985) *Biochem. J. 231*, 349–355.
Pytela, R., Pierschbacher, M. D., Suzuki, S., & Ruoslahti, E. (1985) *Proc. Natl. Acad. Sci. U.S.A. 60*, 245–258.
Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 5463–5467.
Schmid, C. W., & Jelinek, W. R. (1982) *Science (Washington, D.C.) 216*, 1065–1070.
Stanley, K. K. (1986) *FEBS Lett. 199*, 249–253.
Stanley, K. K., & Luzio, J. P. (1984) *EMBO J. 3*, 1429–1434.
Suzuki, S., Pierschbacher, M. D., Hayman, E. G., Nguyen, K., Öhgren, Y., & Ruoslahti, E. (1984) *J. Biol. Chem. 259*, 15307–15314.
Suzuki, S., Oldberg, A., Hayman, E. G., Pierschbacher, M. D., & Ruoslahti, E. (1985) *EMBO J. 4*, 2519–2524.
Tschopp, J., & Mollnes, T.-E. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 4223–4227.
Tschopp, J., Masson, D., & Stanley, K. K. (1986) *Nature (London) 322*, 831–834.

A Complete cDNA Sequence for the Major Epidermal Growth Factor Binding Protein in the Male Mouse Submandibular Gland†

Michael Blaber,‡ Paul J. Isackson,‡,§ and Ralph A. Bradshaw*,‡
Departments of Biological Chemistry and of Anatomy and Neurobiology, California College of Medicine, University of California, Irvine, California 92717
Received March 30, 1987

ABSTRACT: The complete cDNA sequence of the major epidermal growth factor binding protein (EGF-BP), isolated from the mouse submandibular gland, has been determined. Oligonucleotide probes complementary to unique, nonconserved, regions of homogeneous preparations of EGF-BP were used to identify the correct cDNA clone from a male mouse submandibular gland cDNA library. The nucleotide sequence codes for a glandular kallikrein that is the main arginine esterase complexed with epidermal growth factor. The mRNA coding for this EGF-BP is estimated at 0.24% of the total mRNA of the adult male mouse submandibular gland, thus representing an abundant member of the kallikrein family in this tissue. In addition, the cDNA sequence defines a putative transcription start site. The reported cDNA sequence is clearly different from, and not an allelic form of, a previously reported cDNA sequence for EGF-BP. The present work reconciles conflicting information in the literature regarding the identity of EGF-BP.

The mouse submandibular gland is a rich source of both epidermal growth factor (EGF)[1] and nerve growth factor (NGF) (Cohen, 1960, 1962). Both are isolated from the mouse submandibular gland as high molecular weight complexes containing the arginine-specific esterases EGF binding protein (EGF-BP) and γ-NGF, respectively (Taylor et al., 1970; Varon et al., 1968). These esterases are serine proteases of the glandular kallikrein family, 25 closely related members of which are localized on mouse chromosome 7 (Mason et al., 1983; Evans et al., 1987). Glandular kallikreins have been defined as being able to cleave kininogen to release kinins, potent vasodilators (Orstavik, 1980). However, this specific kallikrein activity may vary considerably among the various members of this family. Since both EGF and NGF are translated as high molecular weight precursors that are processed to yield the mature active growth factor (Scott et al., 1983a,b; Ullrich et al., 1983; Gray et al., 1983), these specific esterases that are bound to the mature growth factors in the mouse submandibular gland have been postulated to be involved (Angeletti & Bradshaw, 1971; Frey et al., 1979; Berger & Shooter, 1977). Thus, EGF-BP and γ-NGF may be involved in the regulation of the levels of the mature growth factors.

The partial amino acid sequence of a preparation of EGF-BP has been reported by Anundi et al. (1982). Two forms of EGF-BP, type "A" and type "B", were reported by this group to be present in the male mouse submandibular gland. This preparation of EGF-BP was characterized as being similar to the EGF-BP described by Taylor et al. (1970) by the criteria of amino acid composition and cross-reactivity to a polyclonal anti-EGF-BP antiserum, although neither form was demonstrated to complex with EGF in vitro. Partial and full-length cDNA clones corresponding to type B EGF-BP have been reported (Ronne et al., 1983; Lundgren et al., 1984).

A comparison of the partial amino acid sequences reported by Anundi et al. (1982) with a partial sequence determined in our laboratory (Silverman, 1977) revealed significant discrepancies that were subsequently confirmed (Isackson et al.,

† This work was supported by USPHS Research Grant NS19964 and by American Cancer Society Grant BC273.
* Address correspondence to this author.
‡ Department of Biological Chemistry.
§ Department of Anatomy and Neurobiology.

[1] Abbreviations: EGF, epidermal growth factor; EGF-BP, epidermal growth factor binding protein; NGF, nerve growth factor; γ-NGF, γ-subunit of mouse 7S NGF; β-NGF, β-subunit of mouse NGF; bp, base pair(s); Tris, tris(hydroxymethyl)aminomethane; EDTA, ethylenediaminetetraacetic acid; SDS, sodium dodecyl sulfate; Da, dalton(s).

0006-2960/87/0426-6742$01.50/0