The structure and evolution of a 461 amino acid human protein C precursor and its messenger RNA, based upon the DNA sequence of cloned human liver cDNAs

Robert J. Beckmann, Robert J. Schmidt, Robert F. Santerre, Jorge Plutzky¹, Gerald R. Crabtree¹ and George L. Long

Division of Molecular and Cellular Biology, Lilly Research Laboratories, Indianapolis, IN 46285, and ¹Department of Pathology, Stanford Medical School, Palo Alto, CA 94305, USA

Received 29 May 1985; Accepted 19 June 1985

ABSTRACT Human liver cDNA coding for protein C has been synthesized, cloned and sequenced. The abundance of protein C message is approximately 0.02% of total mRNA. Three overlapping clones contain 1,798 nucleotides of contiguous sequence, which approximates the size of the protein's mRNA, based upon Northern hybridization. The cDNA sequence consists of 73 5'-noncoding bases, coding sequence for a 461 amino acid nascent polypeptide precursor, a TAA termination codon, 296 3'-noncoding bases, and a 38 base polyadenylation segment. The nascent protein consists of a 33 amino acid "signal", a 9 amino acid propeptide, a 155 amino acid "light" chain, a Lys-Arg connecting dipeptide, and a 262 amino acid "heavy" chain. Human protein C and Factor IX and X precursors possess about one third identical amino acids (59% in the γ -carboxyglutamate domain), including two forty-six amino acid segments homologous to epidermal growth factor. Human protein C also has similar homology with prothrombin in the "leader", γ -carboxyglutamate and serine protease domains, but lacks the two "kringle" domains found in prothrombin.

INTRODUCTION

Protein C is a blood plasma protein which plays a key role in hemostasis. Two brief reviews have recently covered the topic of protein C (1,2). Esmon and coworkers have shown that circulating, inactive protein C zymogen precursor is converted at the endothelial cell surface to an active serine protease by limited thrombin cleavage; and that the activation process is dramatically potentiated by the protein cofactor, thrombomodulin (3,4). Activated protein C in the presence of phospholipid and Ca^{2+} (i.e. cell surface) acts by limited proteolysis to inactivate both factor Va (5-7) and VIIIc (8), thereby down regulating in a feedback mode both the intrinsic and the extrinsic pathways of coagulation. Evidence exists suggesting that protein C may also be involved in enhanced fibrinolysis (9,10). Clinical evidence for the antithrombotic role of protein C comes from observations of reduced plasma levels of protein C in human patients with congenital thrombotic disease (11,12) and other causes of intravascular coagulation (13).

Human protein C has been purified and partially characterized by Kisiel (14). In plasma, it circulates predominately as a two-chain ($M_r = 21,000$ and $41,000$) inactive precursor linked by disulfide bridging. Recently, it has been reported that a significant portion (5-15%) of protein C circulates as a single chain ($M_r = 62,000$) molecule (15). Human protein C undergoes extensive post-translational modification. It contains 23% carbohydrate (14) and undergoes vitamin K-dependent γ -carboxylation of amino terminal glutamyl residues in the light chain (16). Based upon homology with bovine protein C light chain and other vitamin K-dependent proteins, the human molecule is also presumed to undergo β -hydroxylation of an aspartyl residue (17). In addition to the above-mentioned limited proteolyses, it has been assumed that processing of nascent protein C requires cleavage of a signal peptide-containing precursor, to permit transport of the protein out of the cell (18).

Human plasma levels of protein C are significantly lower than those in the cow (14), which has prevented complete structural analysis of the human molecule. Only 14 and 18 amino terminal residues for the light and heavy chains, based upon amino acid sequencing, have been reported (14). Consequently, we have chosen to determine the polypeptide structure of human protein C and its nascent precursor by synthesizing, cloning, and sequencing human liver cDNA encoding protein C. Foster and Davie, employing a similar strategy, have reported a partial cDNA encoding human protein C which begins at amino acid #64 of the light chain (19). We report here the construction and screening of a human liver cDNA library in plasmid pBR322. Overlapping clones have been identified and characterized which include 73 5'-noncoding bases; coding sequence for a 42 amino acid leader peptide, a 155 amino acid light chain, a lys-arg linking dipeptide, a 262 amino acid heavy chain, and a stop codon; 296 3'-noncoding bases, and a 38 base polyadenylation segment.

MATERIALS AND METHODS

Construction of the human liver cDNA library. Human liver from a healthy 47 year old male organ donor was removed at the time of death and infused with cold neutral saline buffer. Total RNA was extracted by the guanidinium thiocyanate-ethanol procedure of Chirgwin, et al. (20). Polyadenylated RNA was isolated by two passages over an oligo dT-cellulose chromatographic column (21). One hundred grams tissue yielded 50 mg total RNA and two mg polyadenylated RNA. Reticulocyte translation (Bethesda Res. Labs) of the isolated message revealed a spectrum of protein products extending in size

up to about 200,000 daltons. The above polyadenylated RNA (13.6 μgm) was then used to synthesize cDNA by the method of Gubler and Hoffman (23). In this procedure, second strand synthesis is carried out in the presence of RNase H and DNA polymerase I, thereby avoiding the necessity of S_1 nuclease digestion of the 5'-end hairpin loop, and resulting in more complete synthesis at the 5'-end. Following dC tailing by a standard procedure (24) the cDNA was either purified on a 5% polyacrylamide gel or used directly for ligation into the dG tailed Pst I restriction endonuclease site of plasmid pBR322. Resulting chimeric plasmids were then used to transform CaCl_2 -treated (25) Escherichia coli strain RR1, and transformants were selected by tetracycline resistance. Gel fractionated (>600 bp) cDNA transformed at a level of $\sim 3 \times 10^6$ transformants/ μgm cDNA, which was approximately twenty times that of the unfractionated cDNA. The size of PstI inserts was also significantly increased for the gel fractionated cDNA. As reported here for protein C and for two other plasma proteins (data not presented), cDNA clones were generated which begin upstream from the ATG (Met) start site. The library contains a spectrum of insert sizes to over 2000 bp with the majority of them being 900 ± 300 bp.

High-Density Colony Screening. Transformed cells were plated in duplicate at an approximate density of 5,000 colonies per nitrocellulose filter (Millipore HATF082, 82 mm diameter) on L broth/1.5% agar plates containing 12.5 μg of tetracycline per ml, and amplified on plates containing 250 μg of chloramphenicol per ml, as outlined elsewhere (26). The final nitrocellulose filters containing cellular DNA were prehybridized (10 filters per Seal-a-Meal® bag) for 12 hrs at 60°C in 1M NaCl; 0.1M Tris, pH 7.2; 10X Denhardt's solution; 0.1% SDS; 0.1% Na-pyrophosphate; and 200 $\mu\text{gm}/\text{ml}$ sheared, heat-denatured salmon sperm DNA. Each bag was opened and the radiolabeled probe was added, dissolved in a small volume of prehybridizing solution, and the bags resealed. The two probes utilized for screening consisted of bovine protein C cDNA fragments corresponding to amino acids #-19 to +27 and #+119 to +336 of the nascent protein (see ref. 27). The two DNA fragments were generated by restriction enzyme digestion, purified from polyacrylamide gels and labeled by nick-translation using a standard protocol (28). Following hybridization (12-16 hrs, 60°C), the filters were briefly rinsed at room temperature with 1M NaCl; 1X Denhardt's solution; 0.05M Tris, pH 8.0; 0.1% SDS and 0.1% Na-pyrophosphate. Two 60 minute washes, each with the above solution at 60°C were performed; followed by three 30 minute washes in 0.15M NaCl, 150 mM Na-citrate, 0.1%

SDS, 0.1% Na-pyrophosphate, pH 7.0. The filters were then drained, covered with Saran Wrap® and exposed to x-ray film. Positive duplicate colonies were replated in duplicate and screened at a density of approximately 100-200 colonies per filter. At this point the hybridization signal is much more intense and individual colonies can be picked and grown in liquid media for plasmid preparation and analysis.

Large-Scale Plasmid Preparation. Large preparations of plasmid suitable for DNA sequencing were grown in M-9 medium in the presence of chloramphenicol and were isolated by centrifugation in CsCl/ethidium bromide gradients after cell lysis (29).

Restriction Mapping and Southern Hybridization. Plasmid DNA was prepared from 1 ml overnight liquid cultures by the method of Birnboim and Doly (30), cleaved with restriction endonucleases according to the supplier's recommendations, and Southern blotted from 1% agarose gels by the modified method of Smith and Summers (31) and hybridized as described above.

DNA Sequencing. Cesium chloride-purified plasmid or restriction fragments from mini-prepped DNA (30) suitable for DNA radiolabeling and sequencing were obtained by electrophoresis on 3.5% polyacrylamide gels and crush-elution (32). Blunt-end or 5'-overhanging DNA fragments were treated with bacterial alkaline phosphatase and T4 polynucleotide kinase according to the method of Maxam and Gilbert (32). 3'-overhanging fragments were labeled with [³²P]-cordycepin according to the directions provided by the supplier (New England Nuclear). Labeled ends were separated on polyacrylamide gels after restriction endonuclease digestion and sequenced by the chemical modification technique of Maxam and Gilbert (32). The G reaction was performed for 7 min at 20°C, and the G+A reaction for 10 min at 37°C. Polyacrylamide sequencing gels (6% and 20%) were run according to Sanger and Coulson (33), fixed with 7% (wt/vol) acetic acid (6% gels dried), and exposed to x-ray film with the aid of intensifying screens. All DNA sequences were stored and analyzed with the aid of the Lilly Research Laboratories DNA computing environment.

Determination of protein C mRNA size by Northern hybridization. Isolated polyadenylated RNA from different cell types or tissues was separated by electrophoresis on an agarose gel in the presence of formaldehyde, transferred to nitrocellulose and hybridized with nick-translated probe as described elsewhere (34). The probe utilized was the cDNA insert from pHc-8, shown in Figure 2.

Figure 1. Northern blot of liver mRNA for protein C. Total poly-A selected mRNA (5 μ g) was separated by electrophoresis, transferred to nitrocellulose and hybridized with a [32 P]-radiolabeled human cDNA fragment (1100 bp) from clone pHc-8, which includes DNA encoding amino acids #225-419 of the nascent protein (see Fig. 3). Total mRNA run in each of the lanes was from the following sources: (A) human T-cells; (B) human liver; (C) rat liver; (D) cultured human fibroblasts; and (E) cultured human HepG-2 transformed cells.

RESULTS

Northern hybridization reveals that the major species of protein C mRNA both in human and rat liver cells is approximately 1800-1850 bases in length (Figure 1). Hybridization also reveals the presence of a distinct minor species approximately 200 bases shorter, which may represent an alternative 3'-polyadenylation site (see Discussion).

A **B** **C** **D**

Figure 2. Schematic presentation of overlapping human liver cDNA clones for protein C. Part A: The size and degree of overlap for three independent protein C clones. Also included are the homopolymeric tails utilized in insertion of the cDNA into the *Pst*I site of the plasmid vector, pBR322. The *Pst*I sites at the ends of each insertion are generated by the cloning technique. The clones are oriented in the figure so that they correspond to the mRNA coding sequence and the direction (left to right) of translation. Clone pHC-8 was isolated from a separate liver library generated by methods similar to those described in this paper. Parts B, C and D: Sequencing strategies for clones pBHC-12, pHC-1 and pHC-8, respectively. Horizontal arrows show the site of [32 P] labelling (vertical line) and length of sequence obtained for individual fragments. In some cases fragments were sequenced more than once. Thick and thin arrows represent 3'-cordycepin- and 5'-ATP-labeled fragments, respectively. All sites within the inserts for the enzymes listed at the right are shown except for *Hgi*AI at positions 242, 389 and 1082, and *Hha* I at positions 399, 493, 571, 584 and 676. Numbering refers to nucleotide position as shown in Figure 3.

Three independently isolated cDNA clones overlap with one another, and together contain the coding region for nascent human protein C, as well as 5'- and 3'-non-coding segments. Figure 2 schematically presents the relationship of these three clones and the strategy used to determine the cDNA sequences. The size of the total cloned cDNA (1,800 bp) is consistent with the predicted mRNA length determined from Figure 1. Figure 3 presents the cDNA (mRNA coding strand) sequence and the corresponding nascent polypeptide amino acid sequence for human protein C. The proposed nascent protein contains 461 amino acids, consisting of a 42 amino acid "leader"

```

      10      20      30      40
C TGC AGG GGG GGG GGG GGG GGG CTG TCA TGG CGG CAG GAC GGC

      50      60      70      80      90
GAA CTT GCA GTA TCT CCA CGA CCC GCC CCT ACA GGT GCC AGT GCC TCC

      100      110      120      130      140
AGA ATG TGG CAG CTC ACA AGC CTC CTG CTG TTC GTG GCC ACC TGG GGA
    MET TRP GLN LEU THR SER LEU LEU PHE VAL ALA THR TRP GLY
      -40      -35      -30

      150      160      170      180      190
ATT TCC GGC ACA CCA GCT CCT CTT GAC TCA GTG TTC TCC AGC AGC GAG
ILE SER GLY THR PRO ALA PRO LEU ASP SER VAL PHE SER SER SER GLU
      -25      -20      -15

      200      210      220      230
CGT GCC CAC CAG GTG CTG CGG ATC CGC AAA CGT GCC AAC TCC TTC CTG
ARG ALA HIS GLN VAL LEU ARG ILE ARG LYS ARG ALA ASN SER PHE LEU
      -10      -5      5
      ↑      ↑

      240      250      260      270      280
GAG GAG CTC CGT CAC AGC AGC CTG GAG CGG GAG TGC ATA GAG GAG ATC
GLU GLU LEU ARG HIS SER SER LEU GLU ARG GLU CYS ILE GLU GLU ILE
    Y      Y      10      V      15      V      Y      20V

      290      300      310      320      330
TGT GAC TTC GAG GAG GCC AAG GAA ATT TTC CAA AAT GTG GAT GAC ACA
CYS ASP PHE GLU GLU ALA LYS GLU ILE PHE GLN ASN VAL ASP ASP THR
      25V      V      V      30      35

      340      350      360      370      380
CTG GCC TTC TGG TCC AAG CAC GTC GAC GGT GAC CAG TGC TTG GTC TTG
LEU ALA PHE TRP SER LYS HIS VAL ASP GLY ASP GLN CYS LEU VAL LEU
      40      45      50

      390      400      410      420      430
CCC TTG GAG CAC CCG TGC GCC AGC CTG TGC TGC GGG CAC GGC ACG TGC
PRO LEU GLU HIS PRO CYS ALA SER LEU CYS CYS GLY HIS GLY THR CYS
      55      60      65

      440      450      460      470
ATC GAC GGC ATC GGC AGC TTC AGC TGC GAC TGC CGC AGC GGC TGG GAG
ILE ASP GLY ILE GLY SER PHE SER CYS ASP CYS ARG SER GLY TRP GLU
      70      β      75      80      85

      480      490      500      510      520
GGC CGC TTC TGC CAG CGC GAG GTG AGC TTC CTC AAT TGC TCG CTG GAC
GLY ARG PHE CYS GLN ARG GLU VAL SER PHE LEU ASN * CYS SER LEU ASP
      90      95      100

      530      540      550      560      570
AAC GGC GGC TGC ACG CAT TAC TGC CTA GAG GAG GTG GGC TGG CGG CGC
ASN GLY GLY CYS THR HIS TYR CYS LEU GLU GLU VAL GLY TRP ARG ARG
      105      110      115

      580      590      600      610      620
TGT AGC TGT GCG CCT GGC TAC AAG CTG GGG GAC GAC CTC CTG CAG TGT
CYS SER CYS ALA PRO GLY TYR LYS LEU GLY ASP ASP LEU LEU GLN CYS
      120      125      130

      630      640      650      660      670
CAC CCC GCA GTG AAG TTC CCT TGT GGG AGG CCC TGG AAG CGG ATG GAG
HIS PRO ALA VAL LYS PHE PRO CYS GLY ARG PRO TRP LYS ARG MET GLU
      135      140      145

```

```

      680      690      700      710
AAG AAG CGC AGT CAC CTG AAA CGA GAC ACA GAA GAC CAA GAA GAC CAA
LYS LYS ARG SER HIS LEU LYS ARG ASP THR GLU ASP GLN GLU ASP GLN
150      155      160      165

      720      730      740      750      760
GTA GAT CCG CGG CTC ATT GAT GGG AAG ATG ACC AGG CGG GGA GAC AGC
VAL ASP PRO ARG LEU ILE ASP GLY LYS MET THR ARG ARG GLY ASP SER
      170      175      180

      770      780      790      800      810
CCC TGG CAG GTG GTC CTG CTG GAC TCA AAG AAG AAG CTG GCC TGC GGG
PRO TRP GLN VAL VAL LEU LEU ASP SER LYS LYS LYS LEU ALA CYS GLY
      185      190      195

      820      830      840      850      860
GCA GTG CTC ATC CAC CCC TCC TGG GTG CTG ACA GCG GCC CAC TGC ATG
ALA VAL LEU ILE HIS PRO SER TRP VAL LEU THR ALA ALA HIS CYS MET
      200      205      210

      870      880      890      900      910
GAT GAG TCC AAG AAG CTC CTT GTC AGG CTT GGA GAG TAT GAC CTG CCG
ASP GLU SER LYS LYS LEU LEU VAL ARG LEU GLY GLU TYR ASP LEU ARG
      215      220      225

      920      930      940      950
CGC TGG GAG AAG TGG GAG CTG GAC CTG GAC ATC AAG GAG GTC TTC GTC
ARG TRP GLU LYS TRP GLU LEU ASP LEU ASP ILE LYS GLU VAL PHE VAL
      230      235      240      245

      960      970      980      990      1000
CAC CCC AAC TAC AGC AAG AGC ACC ACC GAC AAT GAC ATC GCA CTG CTG
HIS PRO ASN TYR SER LYS SER THR THR ASP ASN ASP ILE ALA LEU LEU
      *      250      255      260

      1010      1020      1030      1040      1050
CAC CTG GCC CAG CCC GCC ACC CTC TCG CAG ACC ATA GTG CCC ATC TGC
HIS LEU ALA GLN PRO ALA THR LEU SER GLN THR ILE VAL PRO ILE CYS
      265      270      275

      1060      1070      1080      1090      1100
CTC CCG GAC AGC GGC CTT GCA GAG CGC GAG CTC AAT CAG GCC GGC CAG
LEU PRO ASP SER GLY LEU ALA GLU ARG GLU LEU ASN GLN ALA GLY GLN
      280      285      290

      1110      1120      1130      1140      1150
GAG ACC CTC GTG ACG GGC TGG GGC TAC CAC AGC AGC CGA GAG AAG GAG
GLU THR LEU VAL THR GLY TRP GLY TYR HIS SER SER ARG GLU LYS GLU
      295      300      305

      1160      1170      1180      1190
GCC AAG AGA AAC CGC ACC TTC GTC CTC AAC TTC ATC AAG ATT CCC GTG
ALA LYS ARG ASN ARG THR PHE VAL LEU ASN PHE ILE LYS ILE PRO VAL
      310      315      320      325

      1200      1210      1220      1230      1240
GTC CCG CAC AAT GAG TGC AGC GAG GTC ATG AGC AAC ATG GTG TCT GAG
VAL PRO HIS ASN GLU CYS SER GLU VAL MET SER ASN MET VAL SER GLU
      *      330      335      340

      1250      1260      1270      1280      1290
AAC ATG CTG TGT GCG GGC ATC CTC GGG GAC CGG CAG GAT GCC TGC GAG
ASN MET LEU CYS ALA GLY ILE LEU GLY ASP ARG GLN ASP ALA CYS GLU
      345      350      355

      1300      1310      1320      1330      1340
GGC GAC AGT GGG GGC CCC ATG GTC GCC TCC TTC CAC GGC ACC TGG TTC
GLY ASP SER GLY GLY PRO MET VAL ALA SER PHE HIS GLY THR TRP PHE
      360      365      370

```


CTG	GTG	GGC	CTG	GTG	AGC	TGG	GGT	GAG	GGC	TGT	GGG	CTC	CTT	CAC	AAC	
LEU	VAL	GLY	LEU	VAL	SER	TRP	GLY	GLU	GLY	CYS	GLY	LEU	LEU	HIS	ASN	
	375					380					385					
TAC	GGC	GTT	TAC	ACC	AAA	GTC	AGC	CGC	TAC	CTC	GAC	TGG	ATC	CAT	GGG	
TYR	GLY	VAL	TYR	THR	LYS	VAL	SER	ARG	TYR	LEU	ASP	TRP	ILE	HIS	GLY	
	390				395					400					405	
CAC	ATC	AGA	GAC	AAG	GAA	GCC	CCC	CAG	AAG	AGC	TGG	GCA	CCT	TAG	CGA	
HIS	ILE	ARG	ASP	LYS	GLU	ALA	PRO	GLN	LYS	SER	TRP	ALA	PRO	STOP		
			410						415					420		
CCC	TCC	CTG	CAG	GGC	TGG	GCT	TTT	GCA	TGG	CAA	TGG	ATG	GGA	CAT	TAA	
AGG	GAC	ATG	TAA	CAA	GCA	CAC	CGG	CCT	GCT	GTT	CTG	TCC	TTC	CAT	CCC	
TCT	TTT	GGG	CTC	TTC	TGG	AGG	GAA	GTA	ACA	TTT	ACT	GAG	CAC	CTG	TTG	
TAT	GTC	ACA	TGC	CTT	ATG	AAT	AGA	ATC	TTA	ACT	CCT	AGA	GCA	ACT	CTG	
TGG	GGT	GGG	GAG	GAG	CAG	ATC	CAA	GTT	TTG	CGG	GGT	CTA	AAG	CTG	TGT	
GTG	TTG	AGG	GGG	ATA	CTC	TGT	TTA	TGA	AAA	<u>AGA</u>	<u>ATA</u>	<u>AAA</u>	AAC	ACA	ACC	
ACG	AAA	AAA	AAA	AAA	AAA	AAA	AAA	AAA	AAA	AAA	AAA	AAA	AAA	ACC	CCC	
CCC	CGC	CCC	CCC	CCC	CTG	CAG										

Figure 3. Nucleotide and amino acid sequences derived from human liver cDNA coding for protein C. The amino acid sequence for nascent protein C is shown below the determined nucleotide sequence. Clone pBHC-12 possesses a T residue at position #520. Open arrows above the DNA sequence mark the ends of individual clones, excluding homopolymeric tails for each of the three overlapping clones (presented schematically in Figure 2). Proposed sites of limited proteolytic cleavage of the precursor to generate activated, mature heavy and light chain are shown with closed arrows below the amino acid sequence and are discussed in the text. Also shown are the predicted residues for vitamin K-dependent γ -carboxylation (γ), β -hydroxylation (β) and glycosylation(*). The enzymatic active site residues observed for all serine proteases are underlined. The polyadenylation recognition sequence, AATAAA, is boxed.

Figure 4. Comparison of nascent human protein C with other vitamin K-dependent clotting factors. Shown in one letter code are the amino acid sequences for several vitamin K-dependent clotting factors. (In the mature molecules, several of the amino terminal Glu residues exist as γ -carboxyglutamate.) Identical amino acids are boxed. The comparison of the entire human protein C sequence is shown only for human factor IX. Apparent gaps are shown by dashes (-). Numbering is based upon the human protein C precursor as presented in Figure 3. The above sequences are taken from the following works: bovine protein C (27), human factor IX (36), human factor X (42), and human prothrombin (43).

peptide and both the light (155 amino acids) and heavy (262 amino acids) chains which are joined by a Lys-Arg dipeptide. The amino acid composition of the precursor is: 35 Gly, 24 Ala, 29 Val, 49 Leu, 18 Ile, 15 Phe, 15 Trp, 8 Tyr, 18 Thr, 36 Ser, 24 Lys, 28 Arg, 18 His, 20 Pro, 30 Asp, 34 Glu, 13

Asn, 15 Gln, 8 Met and 24 Cys. Codon usage and the predominance of base C or G in the third position are similar to that for other mammalian mRNAs.

Figure 4 compares the amino acid sequence of human protein C with several other vitamin K-dependent coagulation factors. Homology with other vitamin K-dependent coagulation factors extends throughout the protein, including that of the heavy chain which corresponds to the archetypal serine protease trypsin. (Degree of homology for the entire protein shown only for factor IX in Figure 4.) Comparison of human with bovine (data taken from ref 27) sequences reveals 82% identical nucleic acids and 74% identical amino acids. The bovine molecule contains one amino acid insertion after human residues #159 and 160 (Figure 4), and a four amino acid deletion (#305-308, Figure 4) relative to the human molecule.

Bovine protein C contains two additional glutamyl residues (positions #23 and 35 in Figure 4) which can serve as substrates for γ -carboxylation. As in bovine protein C, residue #71 is aspartic acid, and is presumed to exist in the mature molecule as β -hydroxyaspartic acid (17). On the basis of homology with bovine protein C (39,40) asparaginyl residues #97, 248, 313 and 329 in Figure 3 are tentatively assigned as glycosylation sites.

Alignment of human protein C with human factor IX (36) reveals 34% overall amino acid sequence identity, and necessitates several gaps as shown in Figure 4. The largest gap is that existing at the junction of heavy and light chains and the site of serine protease activation. Human factor IX contains all nine of the γ -carboxylated glutamyl residues found in human protein C as well as three additional sites (residues #32, 35 and 39, Figure 4).

DISCUSSION

Due to the low abundance of circulating human protein C and messenger RNA, it has not been possible to determine the amino acid structure of the molecule by direct amino acid sequencing or in vitro mRNA translation. Utilizing the tools of recombinant DNA technology, we have succeeded in determining the entire amino acid sequence for the nascent human protein C precursor. The 461 amino acid precursor is trimmed by limited proteolysis to yield both the light (155aa) and heavy (262aa) chains known to exist in the mature, circulating, inactive form of the molecule. Final removal of a highly polar 12 amino acid "activation" peptide by the action of thrombin subsequently generates activated protein C. Circulating human protein C levels in plasma are $\sim 4 \mu\text{g/ml}$, compared to $\sim 120 \mu\text{g/ml}$ for prothrombin. The

percentage of total liver mRNA coding for prothrombin is believed to be ~1%. Assuming equal rates of turnover and translation, one would predict the level of protein C mRNA to be ~0.03%. Screening of approximately 50,000 colonies with a bovine cDNA probe corresponding to the middle third of the mRNA resulted in nine confirmed (by colony and Southern hybridization) protein C clones. This represents a frequency of 0.02%, which agrees well with that expected on the basis of circulating protein concentrations.

According to Northern hybridizations, human protein C mRNA is 1800-1850 bases in length. This is in good agreement with the size (1798 bp) of the total cDNA sequence, excluding the homopolymer tails. Figure 1 reveals the presence in all liver cells of a minor hybridizing component which is ~200 bases shorter than the major component. The cDNA sequence contains at the 3' end an alternative polyadenylation recognition site (bases #1530-1535, Figure 3) which may account for a shortened mRNA species. Alternative polyadenylation recognition sites is a well documented phenomenon (37). Translation of this smaller mRNA would be expected to result in an identical protein product as from the larger species. The fact that the smaller form is a minor form of mRNA may explain why a corresponding cloned cDNA has not been observed. It is also possible that the shortened mRNA is the result of alternative or incorrect splicing of the transcribed gene product, in which case a different protein would be produced.

This communication reports the entire amino acid sequence for nascent human protein C precursor, including previously unknown sequence for a 42 amino acid prepropeptide as well as amino acid residues #15-63 of the mature "light" chain. On the basis of homology with other vitamin K-dependent clotting factors (see Fig. 4) and eukaryotic signal peptides (18), we propose that the prepropeptide consists of a typical 33 amino acid "signal" peptide and a linking nine amino acid propeptide. A similar scheme and the enzymes involved in the processing have been presented elsewhere (26). The propeptide is extremely basic in composition (3 Arg, 1 Lys) and may play a role in nascent protein folding and post-translational modification (e.g. γ -carboxylation, β -hydroxylation) prior to its removal.

The nucleic acid and amino acid sequences presented in Figure 3 agree with the sequences reported by Foster and Davie (19) beginning at amino acid #62 except for bases #413-415 (TGC→CAA, Cys→Gln) and #865 (T→C). We believe that the triplet reported by Foster and Davie at position 413-415 may represent a cloning artifact, since they are at the extreme end of their cDNA. The single base difference (T→C) at position #865 may represent a

true genetic variant. The nucleic acid sequence (excluding the homopolymeric tails) presented in Figure 3 agrees for overlapping clones and corresponding genomic sequence (data presented elsewhere, see ref. 38) with the one exception of base #520 which appears in genomic and pBHC-12 clones as a T. Foster and Davie (19) report this base as a G in their cDNA. The resulting codon in either case codes for the amino acid Ser. The occurrence of each of the bases from two independent sources strongly suggests that the liver donor was heterozygous for protein C. Both our amino acid sequence and that of Foster and Davie (19), disagree with those of the heavy and light amino termini based upon protein sequencing (14). Residues #10, 12, 13, 159 and 175 (Figure 3) were originally reported to be Ile, Trp, Lys, Pro and Val, respectively (14). It is unlikely that the changes represent true genetic variants or cloning artifacts because the differences would require a minimum of eight base changes.

Human protein C consists of the following structural elements: a leader peptide (amino acids -42 to -1), a γ -carboxylation domain (aa #1-45), two epidermal growth factor domains (aa #46-91 and 92-137) (39), an activation peptide region (aa #144-169), and the serine protease (aa #170-419). Human protein C has a high degree of homology with its bovine counterpart as well as with other vitamin K-dependent coagulation factors (Figure 4, shown in its entirety only for Factor IX). For example, the γ -carboxyglutamate domain of human protein C has 59% identical amino acid residues with human factors IX and X and 50% with human prothrombin. The homology continues through the growth factor domains at a level of ~40% identical amino acids for factors IX and X but not for prothrombin. This is presumably due to the fact that prothrombin contains in the position of the growth factor domains two "kringle" domains (each approximately 100 amino acids), which are not found in protein C and factors IX and X (40). The degree of amino acid sequence identity for human protein C throughout the serine protease domain is 30-40% when compared to factors IX, X and prothrombin. One area of protein C which does not show good homology with other coagulation proteins is the site of nascent chain cleavage to heavy and light chains and activation peptide cleavage (amino acids #144-169, Figure 4). Since these processes are specific for each precursor, thereby conferring a desired biological specificity, there is good reason for this structural domain being different in otherwise homologous proteins.

By applying the homology of human protein C with other proteins and the disulfide bridge scheme presented in reference 39, the following cystine

bridges can be tentatively assigned: amino acids #17-22, 50-63, 59-78, 80-89, 98-109, 105-118, 120-133, 141-277, 196-212, 331-345, and 356-384. Two cysteine residues (#64 and 69) remain unassigned and may also exist as a disulfide bridge.

Human protein C and coagulation factor IX show an overall protein structural homology of 34%. Analysis of the human protein C gene structure (ref 38) and comparison with that of a factor IX (36) shows that both contain several introns in the protein coding portion of the mRNA precursor and, more surprisingly, that all of the introns occur at the same positions in the two genes. Furthermore, the protein structural domains discussed above correspond to the resulting gene's exonic regions (38). This is consistent with the hypothesis, first proposed by Gilbert (41), that introns may bound protein structural-functional domains and play a role in the genetic interchange of common protein structural building blocks.

ACKNOWLEDGEMENTS The authors wish to thank JoAnn Hoskins, John Shepherd and Joseph Modelevsky for technical assistance. We wish to also thank Ms. Sherry Pike for an excellent job typing the manuscript.

REFERENCES

1. Esmon, C.T. (1983) Blood 62, 1155-1158.
2. Gardiner, J.E. and Griffin, J.H. Progr in Hematol, Vol. XIII (E.B. Brown, ed.) pp. 265-278, Grune and Stratton, Inc. 1983.
3. Esmon, C.T. and Owen, W.G. (1981) Proc Natl Acad Sci USA 78, 2249-2252.
4. Esmon, N.L., Owen, W.G. and Esmon, C.T. (1982) J. Biol Chem 257, 859-864.
5. Kisiel, W., Canfield, W.M., Ericsson, L.H. and Davie, E.W. (1977) Biochemistry 16, 5824-5831.
6. Seegers, W.H., Martar, R.A. and Walz, D.A. (1978) Thromb. Res 13, 233-243.
7. Walker, F.J., Sexton, P.W. and Esmon, C.T. (1979) Biochim Biophys Acta 571, 333-342.
8. Vehar, G.A. and Davie, E.W. (1980) Biochemistry 19, 401-410.
9. Seegers, W.H., McCorry, L.E., Graben, H.D., Sakuragawa, N., and Agrawal, B.B.L. (1972) Throm Res 1, 443-460.
10. Comp, P.C., Jacocks, R.M., Fenell, G.L. and Esmon, C.T. (1982) J. Clin. Invest. 70, 127-134.
11. Griffin, J.H., Evatt, B., Zimmerman, T.S., Kleiss, A.J. and Wideman, C. (1981) J. Clin Invest 68, 1370-1373.
12. Seligsohn, U., Berger, A., Abend, M. Rubin, L., Attias, D., Zivelin, A. and Rapaport, S.I. (1984) NE J. Med. 310, 559-562.
13. Griffin, J.H., Mosher, D.F., Zimmerman, T.S. and Kleiss, A.J. (1982) Blood 60, 261-264.
14. Kisiel, W. (1979) J. Clin Invest 64, 761-769.
15. Miletich, J.P., Leykam, J.F. and Broze, G.J. (1983) Blood 62, Suppl 1, 306a.
16. DiScipio, R.G. and Davie, E.W. (1979) Biochemistry 18, 899-904.

17. Fernlund, P. and Stenflo, J. (1983) J. Biol Chem **258**, 12509-12512.
18. Jackson, R.C. and Blobel, G. (1980) Ann NY Acad Sci **343**, 391-403.
19. Foster, D. and Davie, E.W. (1984) Proc Natl Acad Sci USA **81**, 4766-4770.
20. Chirgwin, J.M., Przybyla, A.E., MacDonald, R.J. and Rutter, W.J. (1979) Biochemistry **18**, 5294-5299.
21. Aviv, H. and Leder, P. (1972) Proc Natl Acad Sci USA **69**, 1408-1412.
22. Pelham, H.R.B. and Jackson, R.J. (1976) Eur J. Biochem **67**, 247-256.
23. Gubler, U. and Hoffman, B.J. (1983) Gene **25**, 263-269.
24. Peacock, S.L., McIver, C.M. and Monchan, J.J. (1981) Biochim Biophys Acta **655**, 243-250.
25. Dagert, M. and Ehrlich, S.D. (1979) Gene **6**, 23-28.
26. Fung, M.R., Campbell, R.A. and Macgillivray, R.T.A. (1984) Nucleic Acids Res. **12**, 4481-4492.
27. Long, G.L., Belagaje, R.M. and MacGillivray, R.T.A. (1984) Proc Natl Acad Sci USA **81**, 5653-5656.
28. Maniatis, T., Jeffrey, A. and Kleid, D.G. (1975) Proc Natl Acad Sci USA **72**, 1184-1188.
29. Katz, L., Williams, P.H., Sato, S., Leavill, R.W. and Helinski, D.R. (1973) J. Bacteriol. **114**, 577-591.
30. Birnboim, H.C. and Doly, J. (1979) Nucleic Acids Res. **7**, 1513-1523.
31. Smith, G.E. and Summers, M.D. (1980) Anal. Biochem **109**, 123-129.
32. Maxam, A.M. and Gilbert, W. (1980) Methods Enzymol **65**, 499-560.
33. Sanger, F. and Coulson, A.R. (1978) FEBS Lett. **87**, 107-110.
34. Crabtree, G.R. and Kant, J.A. (1981) J. Biol Chem **256**, 9718-9720.
35. Grantham, R., Gantier, C., Govey, M., Mercier, R. and Pave, R. (1980) Nucleic Acids Res. **8**, 49-62.
36. Anson, D.S., Choo, K.H., Rees, D.J.G., Giannelli, F., Gould, F., Huddleston, J.A. and Brownlee, G.G. (1984) EMBO Journal **3**, 1053-1060.
37. Chung, D.W., Que, B.G., Rixon, M.W., Mace, M. and Davie, E.W. (1983) Biochemistry **22**, 3244-3250.
38. Plutzky, J., Hoskins, J.A., Long, G. L. and Crabtree, G.R. (1985) manuscript submitted for publication.
39. Banyai, L., Varadi, A. and Patthy, L. (1983) FEBS Letters **163**, 37-41.
40. Stenflo, J. and Fernlund, P. (1982) J. Biol Chem **257**, 12180-12185.
41. Gilbert, W. (1987) Nature **271**, 501.
42. Fung, M.R., Hay, C.W. and MacGillivray, R.T.A. (1985) Proc Natl Acad Sci USA, in press.
43. Degen, S.J.F., MacGillivray, R.T.A. and Davie, E.W. (1983) Biochemistry **22**, 2087-2097.