Proc. Natl. Acad. Sci. USA
Vol. 81, pp. 4766–4770, August 1984
Biochemistry

Characterization of a cDNA coding for human protein C

(DNA sequence analysis/vitamin K-dependent proteins/blood coagulation)

DONALD FOSTER AND EARL W. DAVIE

Department of Biochemistry, University of Washington, Seattle, WA 98195

Contributed by Earl W. Davie, May 2, 1984

ABSTRACT **Protein C is a precursor to a serine protease that is present in mammalian plasma. In its activated form, it readily inactivates factor V_a and factor $VIII_a$, two proteins that participate as cofactors in the blood coagulation cascade. In the present studies, a λgt11 library containing cDNA inserts prepared from human liver mRNA has been screened with an antibody to human protein C. Seven positive clones were isolated from 2×10^6 phage and were plaque-purified. The cDNA inserts of two of these phage were sequenced and shown to code for human protein C. Each cDNA insert coded for a portion of the light chain of the molecule, a connecting region, the heavy chain, a stop codon, a 3′-noncoding region, and a poly(A) tail. The length of the noncoding sequence on the 3′ end differed in the two clones, but each contained a processing or polyadenylylation signal followed by a poly(A) tail. The amino acid sequence as determined from the cDNA indicates that protein C is synthesized as a single-chain polypeptide containing the light chain and the heavy chain connected by a dipeptide of Lys-Arg. The single-chain molecule is then converted to the light and heavy chains by cleavage of two or more internal peptide bonds. In plasma, the heavy and light chains of protein C are linked together by a disulfide bond. The amino acid sequence of human protein C shows a high degree of homology with that of the bovine molecule. The DNA sequence coding for the catalytic region near the active site serine in human protein C also showed a high degree of DNA and amino acid sequence identity with prothrombin, factor IX, and factor X, three of the other vitamin K-dependent serine proteases that are present in plasma.**

Protein C (M_r 62,000) is one of several vitamin K-dependent glycoproteins present in plasma. It has been well-characterized from both human (1, 2) and bovine (3–5) sources, and the complete amino acid sequence for the bovine molecule has been established (6, 7). Protein C is composed of a heavy chain (M_r 41,000) and a light chain (M_r 21,000), and these two chains are held together by a disulfide bond. The light chain contains 11 γ-carboxyglutamic acid residues (6, 8) and one residue of β-hydroxyaspartic acid (9, 10).

Protein C is a precursor to a serine protease called "activated protein C." It is converted to an activated form by minor proteolysis in a reaction catalyzed by thrombin, trypsin, or a protease from Russell's viper venom (5). The activation of protein C by thrombin is greatly accelerated by a cofactor called thrombomodulin, a protein present in endothelial cells (11–13).

Activated protein C has strong anticoagulant activity. This is due to its inactivation of factor V_a (2, 14, 15) and factor $VIII_a$ (2, 16). Individuals with low plasma levels of protein C often have a history of thrombotic episodes (17–19). Virtual absence of protein C characterized by the homozygous state is associated with a fatal thrombosis in the neonatal period (20). These data indicate that protein C functions as an extremely important regulator of thrombin generation.

Recently, Miletich *et al.* have shown the presence of 5–15% of protein C in human plasma as a single-chain molecule, suggesting that it is synthesized initially as a single-chain molecule (21). In the present studies, we describe the isolation and characterization of two cDNAs coding for a portion of the light chain in addition to the heavy chain of human protein C. These data also indicate that protein C is synthesized as a single-chain molecule that undergoes processing into a two-chain molecule held together by a disulfide bond.

MATERIALS AND METHODS

Screening of the λgt11 cDNA Library. A λgt11 cDNA library containing cDNA inserts prepared from human liver mRNA was kindly provided by Savio L. C. Woo. Approximately 2×10^6 phage were screened by a modification of the method of Young and Davis (22). Antibody to human protein C was a gift from Walter Kisiel. It was purified by affinity chromatography from sheep serum as described by Canfield and Kisiel (23). The purified antibody was labeled with ^{125}I to a specific activity of 6×10^6 cpm/μg and was used to screen filters containing phage plated at a density of 1.5×10^5 plaques per 150-mm plate. Positive clones were isolated and plaque-purified.

DNA Sequence Analysis. Phage DNA was prepared from positive clones by a plate-lysate method (24), followed by banding on a cesium chloride step gradient essentially as described by Degen *et al.* (25). The cDNA insert was isolated by digestion with *Eco*RI and subcloned into the plasmid pUC9 (26). Appropriate restriction fragments from the insert were subcloned into M13mp10 and M13mp11 for sequencing by the dideoxy method (27). Sequencing reactions were carried out with deoxyadenosine 5′-(α-[^{35}S]thio)triphosphate (dATP[α-^{35}S]; Amersham) and run on gradient gels containing 0.5–2.5 × buffer TBE (0.089 M Tris·HCl/0.089 M boric acid/0.002 M EDTA buffer, pH 8.3) (28). Over 90% of each strand of the cDNA insert was sequenced two or more times. M13mp10 and M13mp11 were purchased from Amersham. Restriction enzymes, T4 DNA ligase, bacterial alkaline phosphatase, and the *Escherichia coli* DNA polymerase I (Klenow fragment) were purchased from New England Biolabs or from Bethesda Research Laboratories. Deoxynucleotide triphosphates and dideoxynucleotide triphosphates were purchased from P-L Biochemicals. $Na^{125}I$ was purchased from New England Nuclear. DNA sequences were stored and analyzed by the computer programs of Staden (29, 30).

RESULTS AND DISCUSSION

A human liver cDNA library cloned into a λgt11 phage vector was screened for cDNAs coding for human protein C. In these studies, an ^{125}I-labeled affinity-purified sheep antibody was used to detect phage plaques directing the synthe-

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

sis of a fusion protein of β-galactosidase and protein C. Seven positive clones were isolated by screening 2×10^6 phage, and each was plaque-purified. One clone, called λHC1026, gave a strong signal with the antibody probe and was found to contain two *Eco*RI fragments of approximately 1000 base pairs. Both DNA fragments were then cloned into M13mp11, and the nucleotide sequence of both ends was determined. One of these fragments was found to contain DNA sequences corresponding to the amino acid sequence of the amino-terminal end of the heavy chain of human protein C (1). This fragment also contained the entire 3′ end of the cDNA, including a poly(A) tail. The sequence of the other fragment did not correspond to any known portion of protein C and represents an unrelated cDNA cloned into the same vector. Another clone, called λHC1375, contained a single *Eco*RI insert of approximately 1400 base pairs and cross-hybridized to λHC1026.

The cDNA fragments coding for protein C were then subcloned into pUC9 and further characterized by restriction mapping (Fig. 1). The nucleotide sequences for the two inserts were then determined by using the strategy shown in Fig. 1. The cDNA insert in λHC1026 was found to be composed of 1026 base pairs coding for a portion of the light chain starting with amino acid 112, a connecting dipeptide, and the heavy chain of human protein C. Following the stop codon of TAG, there were 68 base pairs of 3′ noncoding sequence and a poly(A) tail of 29 base pairs. The polyadenylylation or processing sequence of A-T-T-A-A-A (31) was present 16 base pairs upstream from the poly(A) tail. The cDNA insert in λHC1375 was found to be composed of 1375 base pairs starting with amino acid 64 in the light chain. This clone contained 294 base pairs of 3′ noncoding sequence and a poly(A) tail of 9 base pairs. The processing or polyadenylylation sequence of A-A-T-A-A-A (31) was present 13 base pairs upstream from the poly(A) tail in this cDNA insert. The DNA sequence along with the predicted amino acid sequence for the two inserts is shown in Fig. 2.

By alignment with the bovine light chain, it appears probable that the carboxyl-terminal end of the light chain of human protein C ends with leucine-155. The cDNA sequence indicates that this carboxyl-terminal leucine and the amino-terminal aspartic acid of the heavy chain (1) are connected by the dipeptide Lys-Arg. These data indicate that protein C is initially synthesized as a single-chain molecule with a connecting dipeptide and that this dipeptide is removed during processing by proteolytic cleavage. These results are consistent with that recently published by Miletich *et al.* (21), who identified small amounts of single-chain protein C in human plasma. Furthermore, this mechanism of biosynthesis and processing is analogous to that for factor X, which is also synthesized as a single-chain molecule and cleaved into a two-chain molecule that circulates in plasma (32–34).

The amino acid sequence of human protein C as predicted from the cDNA is shown in Fig. 3 along with the amino acid sequence of the bovine molecule as determined by amino acid sequence analysis (6, 7). The sequence of the human protein shows a high degree of homology with major portions of bovine protein C, including the active site region, location of the cysteine residues, and conservation of the apparent carbohydrate attachment sites involving asparagine residues. Human protein C also shares with bovine protein C an aspartic acid residue in position 71 that is converted to β-hydroxyaspartic acid (9, 10) as well as the unusual sequence of Asn-X-Cys as a probable carbohydrate attachment site at asparagine-329. This differs from the typical carbohydrate binding site to asparagine in the sequence of Asn-X-Thr or Asn-X-Ser (7). This suggests that carbohydrate addition has occurred at asparagine-329 prior to disulfide bond formation at cysteine-331. In this situation, cysteine-331 in the Asn-X-Cys sequence has considerable structural similarity to serine in the Asn-X-Ser sequence. Except for a difference of four amino acids after arginine-149 in human protein C, the human and bovine heavy chains can be aligned without insertions or gaps. The alignment of the four amino acids in the bovine sequence with the eight amino acids in the human sequence in this region is not obvious, making precise localization of the gap (or insertion) difficult. Overall, about 75% of the amino acids in the human heavy chain are conserved in the bovine molecule, with the highest degree of divergence being found in the amino-terminal end, the gap or insertion region, and the carboxyl-terminal end.

Upon activation by thrombin, the heavy chain of human protein C is cleaved between arginine-12 and leucine-13. This releases an activation peptide of 12 amino acids and generates activated protein C with a new amino-terminal leucine in the heavy chain (1). As noted previously (1), this ap-

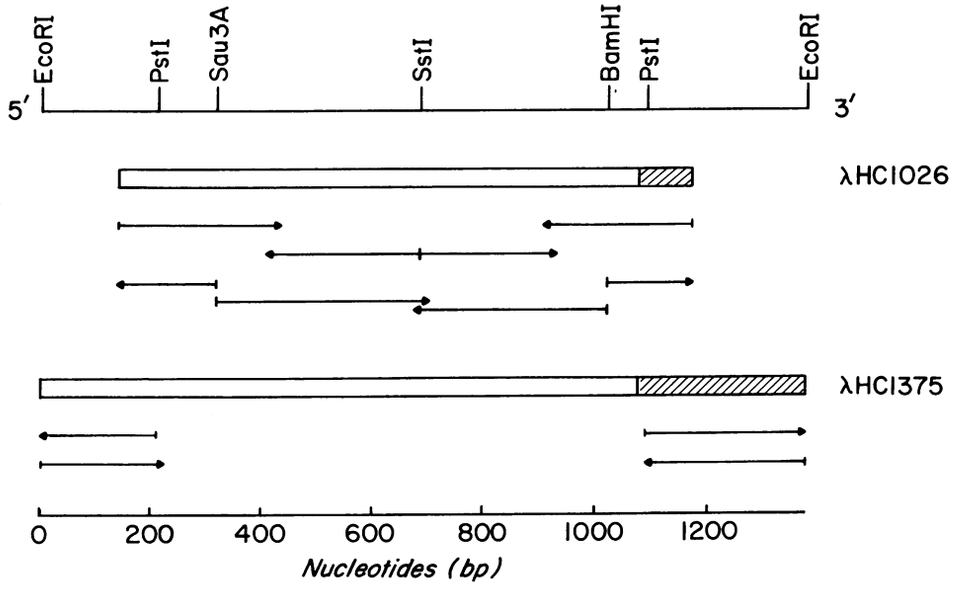

FIG. 1. Partial restriction map and sequencing strategy for the cDNA inserts in clones λHC1026 and λHC1375. The extent of sequencing is shown by the length of each arrow, and the direction of the arrow indicates the strand that was sequenced. The hatched bars indicate the 3′ noncoding region in each cDNA insert.

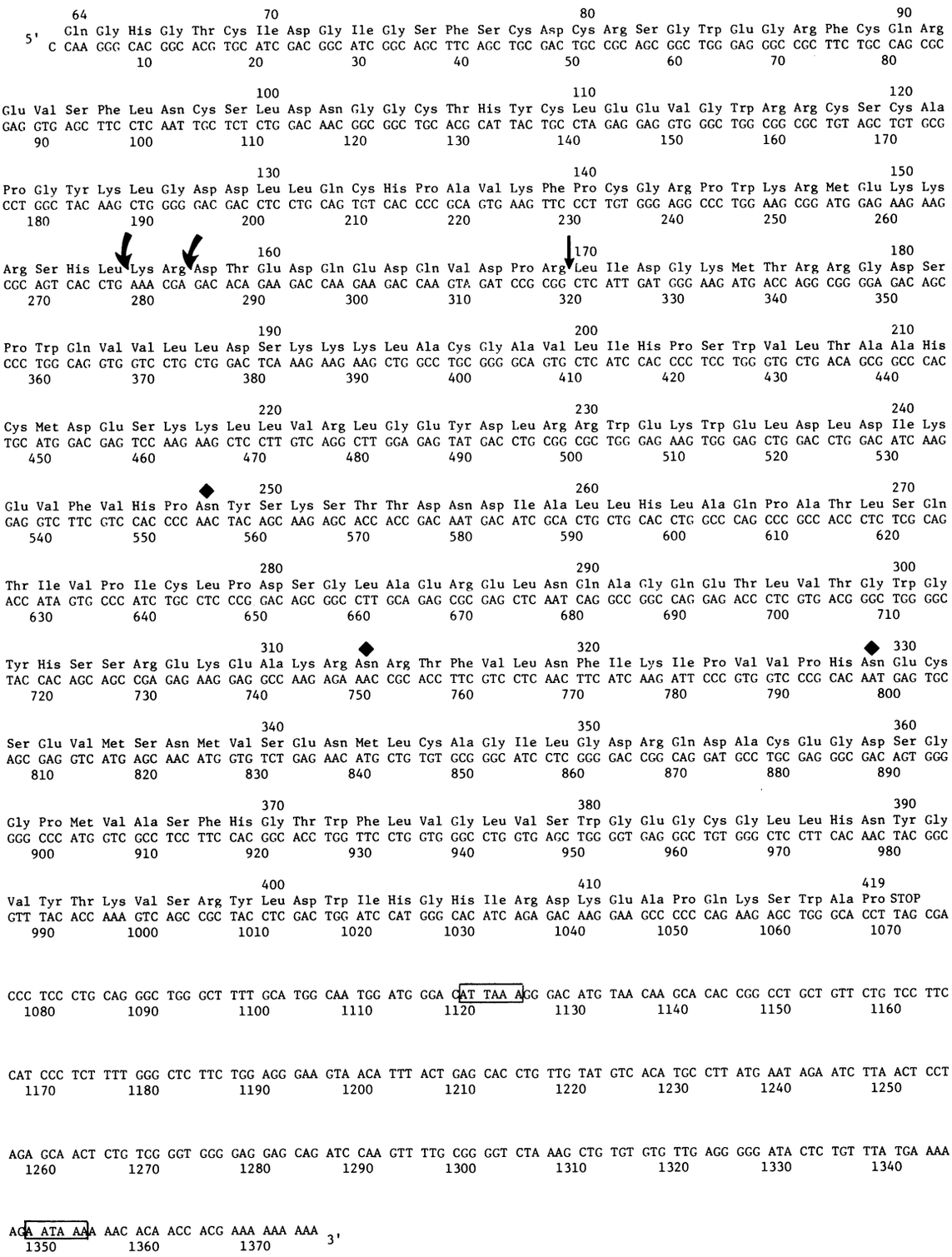

FIG. 2. Nucleotide sequence of the cDNA inserts in λHC1026 and λHC1375 that code for human protein C. The predicted amino acid sequence is shown starting with residue 64 in the light chain. This numbering assumes that the length of the light chain of bovine and human protein C is identical. ◆, Potential carbohydrate binding sites to asparagine residues; ↓, apparent cleavage sites for removal of the connecting dipeptide; ↓, site of cleavage in the heavy chain when protein C is converted to activated protein C; ●, site of polyadenylylation in λHC1026. The processing or polyadenylylation sequences are shown in boxes.

pears to be one of the few serine proteases that has a leucine rather than an isoleucine or valine in this position.

The nucleotide sequences near the active-site serine for protein C and three other human vitamin K-dependent serine protease cDNAs (25, 34–36) are shown in Fig. 4. DNA and amino acid sequence homology in this region is highly conserved. The cDNAs isolated in this investigation lack the 5′ end. This region codes for approximately 63 additional amino acids present in the light chain and a leader sequence that is typical of secreted proteins (37). It should be of interest to compare the DNA sequence corresponding to this region of protein C with factor IX and factor X, since these three pro-

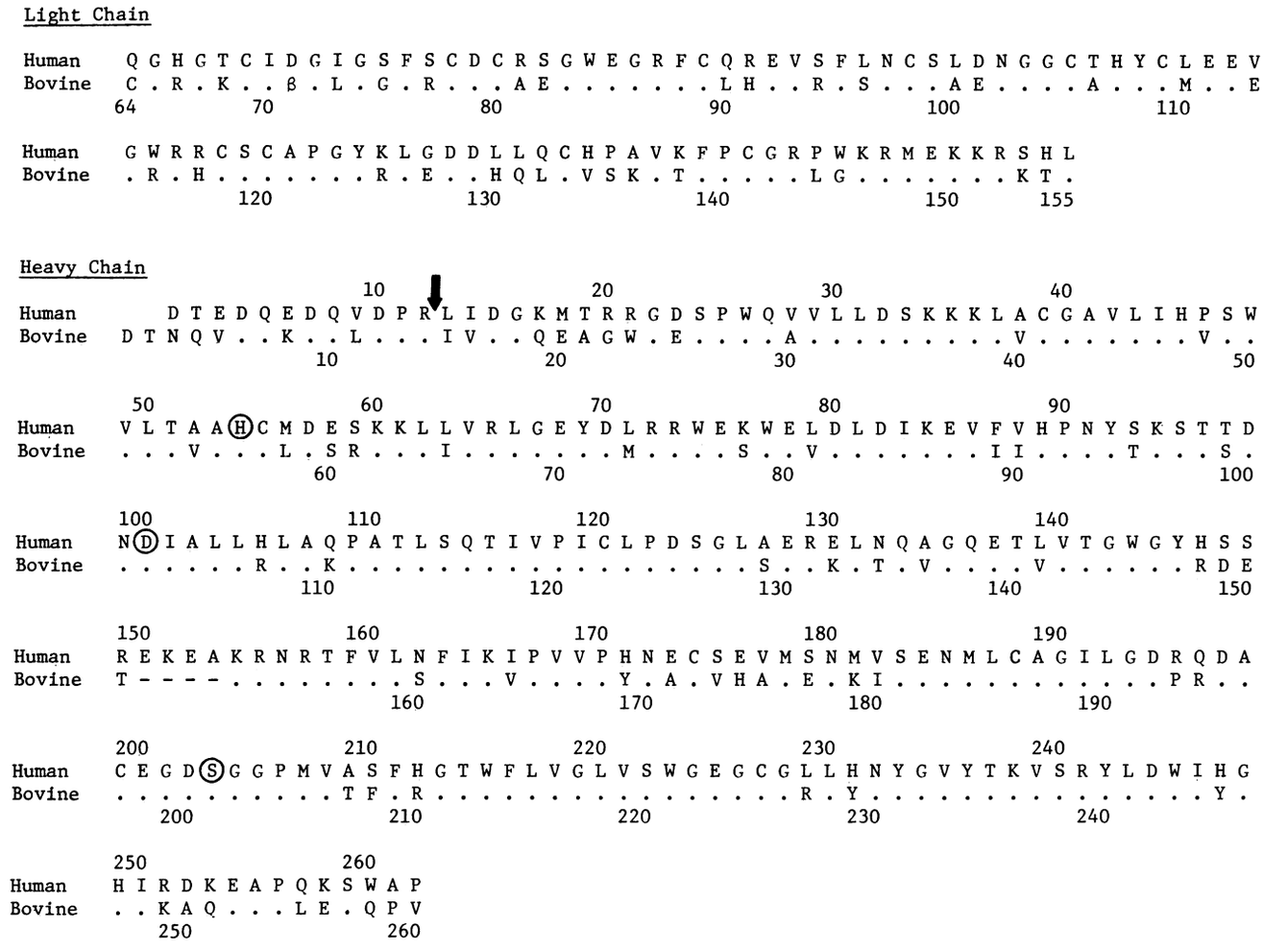

FIG. 3. Amino acid sequences for the heavy chains and part of the light chains of human and bovine protein C. Dots indicate identity between the two proteins. Dashes were inserted for maximal identity between the two proteins. ⬇, Site of cleavage in the heavy chain when protein C is converted to activated protein C. The active-site residues, including histidine-54, aspartic acid-100, and serine-203, are circled. The first amino acids in the light and heavy chains start with number one. The bovine sequences are taken from Fernlund and Stenflo (6, 7). The single-letter code for amino acids is: Ala, A; Arg, R; Asn, N; Asp, D; Cys, C; Gln, Q; Glu, E; Gly, G; His, H; Ile, I; Leu, L; Lys, K; Met, M; Phe, F; Pro, P; Ser, S; Thr, T; Trp, W; Tyr, Y; Val, V.

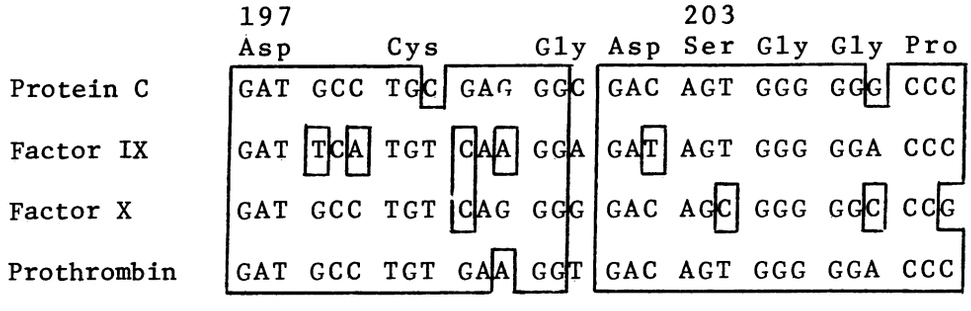

FIG. 4. Alignment of the active-site nucleotide sequences for the human vitamin K-dependent serine proteases. Conserved nucleotides are included in boxes. The numbering of the amino acids is from Fig. 3 for human protein C. Data for factor IX, factor X, and prothrombin are from ref. 35, 34, and 25, respectively.

teins show considerable amino acid homology in their first 150 amino acids (6).

The availability of a cDNA as a probe for the isolation of the gene for human protein C will now make it possible to compare the structure of this gene with that of the other vitamin K-dependent serine proteases. The genes for factor IX and prothrombin have already been shown to have considerable similarity in their coding regions and intron/exon boundaries in the 5′ regions (25, 36).

The authors thank Drs. Savio L. C. Woo and Vincent Kidd for helpful discussions and kindly providing the λgt11 cDNA library. We also wish to thank Drs. Kotoku Kurachi, Dominic Chung, Walter Kisiel, Shinji Yoshitake, Evan Sadler, Steven Leytus, and Mark Rixon for their help and advice. This work was supported in part by a research grant (HL 16919) from the National Institutes of Health.

1. Kisiel, W. (1979) *J. Clin. Invest.* **64**, 761–769.
2. Marlar, R. A., Kleiss, A. J. & Griffin, J. (1982) *Blood* **59**, 1067–1072.

3. Stenflo, J. (1976) *J. Biol. Chem.* **251,** 355–363.
4. Esmon, C. T., Stenflo, J., Suttie, J. W. & Jackson, C. M. (1976) *J. Biol. Chem.* **251,** 3052–3056.
5. Kisiel, W., Ericsson, L. H. & Davie, E. W. (1976) *Biochemistry* **15,** 4893–4900.
6. Fernlund, P. & Stenflo, J. (1982) *J. Biol. Chem.* **257,** 12170–12179.
7. Stenflo, J. & Fernlund, P. (1982) *J. Biol. Chem.* **257,** 12180–12190.
8. DiScipio, R. G. & Davie, E. W. (1979) *Biochemistry* **18,** 899–904.
9. Drakenberg, T., Fernlund, P., Roepstorff, P. & Stenflo, J. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 1802–1806.
10. McMullen, B., Fujikawa, K. & Kisiel, W. (1983) *Biochem. Biophys. Res. Commun.* **115,** 8–14.
11. Esmon, C. T. & Owen, W. G. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 2249–2252.
12. Owen, W. G. & Esmon, C. T. (1981) *J. Biol. Chem.* **256,** 5532–5535.
13. Esmon, N. L., Owen, W. G. & Esmon, C. T. (1982) *J. Biol. Chem.* **257,** 859–864.
14. Kisiel, W., Canfield, W. M., Ericsson, L. H. & Davie, E. W. (1977) *Biochemistry* **16,** 5824–5831.
15. Dahlbäck, B. & Stenflo, J. (1980) *Eur. J. Biochem.* **107,** 331–335.
16. Vehar, G. A. & Davie, E. W. (1980) *Biochemistry* **19,** 401–410.
17. Griffin, J. H., Evatt, B., Zimmerman, T. S., Kleiss, A. J. & Wideman, C. (1981) *J. Clin. Invest.* **68,** 1370–1373.
18. Griffin, J. H., Mosher, D. F., Zimmerman, T. S. & Kleiss, A. J. (1982) *Blood* **60,** 261–264.
19. Bertina, R. M., Broekmans, A. W., Van der Linden, I. K. & Mertens, K. (1982) *Thromb. Haemostasis* **48,** 1–5.
20. Seligsohn, U., Berger, A., Abend, M., Rubin, L., Attias, D., Zivelin, A. & Rapaport, S. I. (1984) *N. Engl. J. Med.* **310,** 559–562.
21. Miletich, J. P., Leykam, J. F. & Broze, G. J. (1983) *Blood* **62,** Suppl. 1, 306a.
22. Young, R. A. & Davis, R. W. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 1194–1198.
23. Canfield, W. M. & Kisiel, W. (1982) *J. Clin. Invest.* **70,** 1260–1272.
24. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) in *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 371–372.
25. Degen, S. J. Friezner, MacGillivray, R. T. A. & Davie, E. W. (1983) *Biochemistry* **22,** 2087–2097.
26. Vieira, J. & Messing, J. (1982) *Gene* **19,** 259–268.
27. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
28. Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 3963–3965.
29. Staden, R. (1977) *Nucleic Acids Res.* **4,** 4037–4051.
30. Staden, R. (1978) *Nucleic Acids Res.* **5,** 1013–1015.
31. Proudfoot, N. J. & Brownlee, G. G. (1976) *Nature (London)* **263,** 211–214.
32. Graves, C. B., Munns, T. W., Willingham, A. K. & Strauss, A. W. (1982) *J. Biol. Chem.* **257,** 13108–13113.
33. Fair, D. S. & Edgington, T. S. (1982) *Circulation* **66,** II-173 (abstr.).
34. Leytus, S. P., Chung, D. W., Kisiel, W., Kurachi, K. & Davie, E. W. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 3699–3702.
35. Kurachi, K. & Davie, E. W. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 6461–6464.
36. Davie, E. W., Degen, S. J. Friezner, Yoshitake, S. & Kurachi, K. (1983) in *Calcium-Binding Proteins 1983: Developments in Biochemistry*, eds. de Bernard, B., Sottocasa, G. L., Sandri, G., Carafoli, E., Taylor, A. N., Vanaman, T. C. & Williams, R. J. P. (Elsevier, Amsterdam), Vol. 25, pp. 45–52.
37. Blobel, G., Walter, P., Chang, C. N., Goldman, B. M., Erickson, A. H. & Lingappa, R. (1979) in *Secretory Mechanisms: Society for Experimental Biology Symposium*, eds. Hopkins, C. R. & Duncan, C. J. (Cambridge Univ. Press, London), Vol. 33, pp. 9–36.