Proc. Natl. Acad. Sci. USA
Vol. 83, pp. 546–550, February 1986
Biochemistry

Evolution and organization of the human protein C gene

(thrombosis/factor IX/coagulation)

JORGE PLUTZKY*†, JO ANN HOSKINS‡, GEORGE L. LONG‡, AND GERALD R. CRABTREE*

*Department of Pathology, Stanford University School of Medicine, Stanford, CA 94305; †Department of Pathology, University of North Carolina Medical School, Chapel Hill, NC 37514; and ‡Division of Molecular and Cellular Biology, Lilly Research Laboratories, Indianapolis, IN 46285

Communicated by K. M. Brinkhous, August 26, 1985

ABSTRACT We have isolated overlapping phage genomic clones covering an area of 21 kilobases that encodes the human protein C gene. The gene is at least 11.2 kilobases long and is made up of nine exons and eight introns. Two regions homologous to epidermal growth factor and transforming growth factor are encoded by amino acids 46–91 and 92–136 and are precisely delimited by introns, as is a similar sequence in the genes for coagulation factor IX and tissue plasminogen activator. When homologous amino acids of factor IX and protein C are aligned, the positions of all eight introns correspond precisely, suggesting that these genes are the product of a relatively recent gene duplication. Nevertheless, the two genes are sufficiently distantly related that no nucleic acid homology remains in the intronic regions and that the size of the introns varies dramatically between the two genes. The similarity of the genes for factor IX and protein C suggests that they may be the most closely related members of the serine protease gene family involved in coagulation and fibrinolysis.

Protein C is a two-chain vitamin K-dependent serine protease that plays a fundamental role in hemostasis by preventing coagulation and promoting fibrinolysis (1). The anticoagulant effects of protein C are achieved through cleavage of factors Va and VIIIa of the intrinsic pathway (2), while clot lysis appears to result from interaction of protein C with the inhibitor of plasminogen activator (3). First isolated from bovine plasma in 1976 by Stenflo (4), the human protein (5) has since been shown to be a 62-kDa dimer with a heavy chain that contains the active serine site and a light chain that contains, at its amino terminus, γ-carboxylglutamic acid residues, which are highly conserved among the vitamin K-dependent factors (6). The protein circulates as a zymogen (7), is activated by thrombin coupled with an endothelial cofactor, thrombomodulin (8), and is inactivated by a specific plasma protease inhibitor (9). Heterozygous deficiency of protein C was first identified by Griffin *et al.* (10) and has been shown by several workers to be an autosomally dominant disorder manifested by a markedly increased tendency to clot (11). Homozygous protein C-deficient patients, with no detectable antigenic levels of protein C, suffer from massive venous thrombosis as neonates (12). A deficiency of protein S, the cofactor for protein C-mediated inactivation of factor Va, has been reported as also causing increased thrombosis (13). Partial cDNAs for protein C were characterized by Foster and Davie (14), and a full length cDNA giving the complete amino acid sequence of the human protein has been isolated in our laboratory (36). We report the isolation of overlapping genomic clones of the protein C gene, analysis of the gene's organization, and relationships to other serine protease genes.

MATERIALS AND METHODS

Isolation and Mapping of Genomic Clones. Two human genomic libraries constructed at the *Eco*RI site of Charon 4a and Charon 28 were screened by the Benton and Davis technique (15) using a protein C cDNA (36) corresponding to amino acids 225–419 plus the 3′ nontranslated region. Approximately 6×10^5 phage plaques were screened and three overlapping fragments of human genomic DNA, each of ≈20 kilobase pairs (kbp), were mapped using the restriction enzymes *Bam*HI, *Eco*RI, *Hin*dIII, *Pst* I, and *Sma* I.

DNA Sequencing. Genomic subfragments (see Fig. 1) from isolated λ phage were ligated into the appropriate restriction sites of plasmid pBR322, and the resulting chimeric plasmids were grown in *Escherichia coli* strain RR1 using standard procedures. Plasmid DNA was purified by CsCl banding as described elsewhere (36).

The strategy utilized to locate the desired regions for sequencing (intron-exon junctions and all exonic segments) was detailed Southern mapping (16) with specific radiolabeled protein C cDNA subfragments which represented the entire cDNA. By digesting with several different enzymes, small (≤600 bp) hybridizing fragments were identified as suitable for direct sequencing. For the purpose of identifying fragments for sequencing, the modified bi-directional Southern transfer described by Smith and Summers (17) was used. Isolated DNA fragments were sequenced by the chemical modification method of Maxam and Gilbert (18) as outlined (36).

Southern Blotting. Total human genomic DNA was isolated as described (19) and subjected to restriction endonuclease digestion. The DNA fragments were then separated on agarose gels, transferred to nitrocellulose, and hybridized as described by Southern (16), using cDNA probes corresponding to several regions of the protein C precursor.

Primer Extension. One microgram of a synthetic oligonucleotide [5′ d(GGGGCGGGTCGTGGAGATACTCG)] corresponding to nucleotides 30–53 as shown in Fig. 2 was hybridized with 3 μg of human liver poly(A)-selected RNA in 60% (vol/vol) formamide, 0.4 M NaCl, 20 mM Pipes (pH 6.4), and 2 mM EDTA, by heating to 85°C for 5 min and allowing the water bath (≈1 liter) to come to room temperature over a period of about 4 hr. The hybrids were recovered and primer extension reactions carried out and analyzed on 8% sequencing gels as described (20).

RESULTS

Organization of the Protein C Gene. From the 6×10^5 phage genomic clones screened, three phage clones, designated λpc4, λpc14, and λpc17, hybridized to our full length cDNA probe. By restriction endonuclease mapping, these clones were found to overlap as illustrated in Fig. 1. The region mapped includes the entire gene of 11.2 kbp, as well as 6 kbp of 5′ flanking DNA, and 4 kbp of 3′ flanking DNA.

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.

Abbreviation: bp, base pair(s).

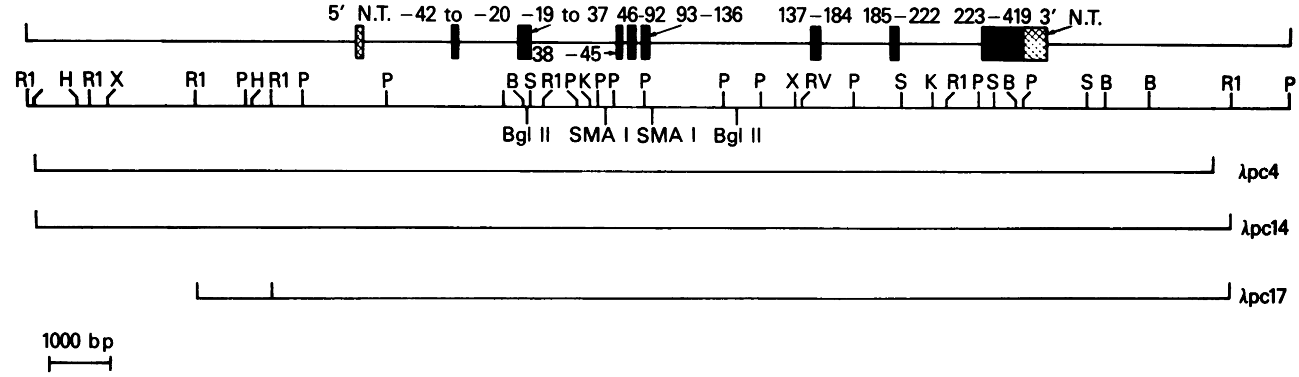
5' N.T.
-42 to -20
-19 to 37
38
-45
46-92
93-136
137-184
185-222
223-419
3' N.T.
R1 H R1 X R1 P H R1 P P B S R1 P K P P P P P X R V P S K R1 P S B P S B B R1 P
Bgl II SMA I SMA I Bgl II
λpc4
λpc14
λpc17
1000 bp

FIG. 1. Organization of the human protein C gene. The first line shows the positions of exons as rectangles on the chromosomal DNA. Numbers above the line indicate the amino acids at which intron/exon junctions occur. Nontranslated regions are shown as cross-hatched areas and coding areas are in black. The second line gives the positions of restriction endonuclease recognition sites. The straight lines below indicate the regions of overlap of the three phage clones: λpc4, λpc14, and λpc17. For the purpose of DNA sequencing, the 4.6-kbp *HindIII/BamHI*, 8.3-kbp *BamHI* and 1.3-kb *BamHI* fragments were inserted into plasmid pBR322. The abbreviations used are as follows: X, *Xba* I; H, *HindIII*; R_1, *EcoRI*; P, *Pst* I; B, *BamHI*; S, *Sst* I; K, *Kpn* I; R, *EcoRv*. Note from Fig. 2 that a small 5' noncoding region is present in the second exon.

The locus mapped in Fig. 1 represents the only locus closely homologous to protein C in the human genome since analysis of total human genomic DNA by Southern blotting of DNA digested with several restriction enzymes reveals only those restriction fragments expected from the locus shown in Fig. 1 (data not shown).

Sequence Analysis. The sequence of the exons and their flanking DNA is shown in Fig. 2. The coding sequence and

```
          -190          -176    -168    -160             -146    -138    -130            -116    -108
  C CAC ATC TGT CAA GGG TTT TGC CCT CAC CTC CCT CCC TGC TGG ATG GCA TCC TTG GTA GGC AGA GGT GGG CTT CGG GCA GAA CAA GCC GTG CTG
   -100           -86     -78     -70                 -56     -48     -40                -26     -18     -10
AGC TAG GAC CAG GAG TGC TAG TGC CAC TGT TTG TCT ATG GAG AGG GAG GCC TCA GTG CTG AGG GCC AAG GAA ATA TTT GTG GTT ATG GAT TAA CTC
     +1         10        20          30         40         50          60         70         80
GAA CTC CAG GCT GTC ATG GCG GCA GGA CGG CGA ACT TGC AGT ATC TCC ACG ACC CGC CCC TGT GAG TCC CCC TCC AGG CAG GTC TAT GAG GGG TGT
90          100          INTRON A                       1420      1430      1440      1450      1460
GGA GGG AGG GCT GC- --- --- --- 1310 bp --- --- --- --- --- --- CCC TTT CAT TCC GCT TCC ACC TGG GGG TGC AGG CAG AGC AGC AGC GGG
         1470       1480      1490      1500      1510      1520      1530      1540      1550
GGT AGG CAC TGC CCG GAG CTC AGA AGT CCT CCT CAG ACA GGT GCC AGT GCC TCC AGA ATG TGG CAG CTC ACA AGC CTC CTG CTG TTC GTG GCC ACC
                                                                     MET TRP GLN LEU THR SER LEU LEU LEU PHE VAL ALA THR
                                                                             -40                 -35                 -30
  1560       1570      1580      1590      1600      1610      1620      1630      1640      1650
TGG GGA ATT TCC GGC ACA CCA GCT CCT CTT GGT AAG GCC ACC CCA CCC CTA CCC GGG ACC CTT GTG GCC TCT ACA AGG CCC TGG TGG GCA TCT GCC
TRP GLY ILE SER GLY THR PRO ALA PRO LEU
                -25             -20
      1660      1670      1680      1690      1700      1710      1720      1730      1740
TAG GCC TTC ACA GCT TCC ACC ATC TCT CTG AGC CTT GGG TGA GGT GAG GGG CAG ATG GGA ATG GCA GGA ATC AAC TGA CAA GTC CCA GGT AGG CCA
1750      1760      1770      1780      1790      1800      1810      1820      1830      1840
GCT GCC AGA GTG CCA CAC AGG GGC TGC TAG GGC AGG CAT GCG TGA TGG CAG GGA GCC CCG CGA TGA CCT CCT AAA GCT CCC TCC TCC ACA CGG G
                                                                    INTRON B
--- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- 650 bp --- --- --- --- --- --- --- --- --- --- ---
          2500      2510      2520      2530      2540      2550      2560      2570      2580
--C TGT ATT GGA TGT TTT ACA TGA CGG TTT CAT CCC CAT GTT TTT GGA TGA GTA AAT TGA ACC TTA GAA AGG TAA AGA CAC TGG CTC AAG GTC ACA
   2590      2600      2610      2620      2630      2640      2650      2660      2670      2680
CAG AGA TCG GGG TGG GGT TCA CAG GGA GGC CTG TTC ATC TCA GAG CAA GGC TTC GTC CTC CAA CTG CCA TCT GCT TCC TGG GAG GAA AAG AGC AGA
       2690      2700      2710      2720      2730      2740      2750      2760      2770
GGA CCC CTG CGC CAA GCC ATG ACC TAG AAT TAG AAT GAG TCT TGA GGG GGC GGT GAC AAG ACC TTC CCA GGT TCT CCC AGC TCT GCT TCC TCA GAC
 2780      2790      2800      2810      2820      2830      2840      2850      2860      2870
CCC CTC ATG GCC CCA GCC CCT CTT AGG CCC CTC CAC CAA GGT GAG CTC CCC CTC CCT CCA AAA CCA GAC TCA GTG TTC TCC AGC AGC GAG CGT GCC
                                                                                 ASP SER VAL PHE SER SER SER GLU ARG ALA
                                                                                                 -15                 -10
   2880      2890      2900      2910      2920      2930      2940      2950      2960      2970
CAC CAG GTG CTG CGG ATC CGC AAA CGT GCC AAC TCC TTC CTG GAG GAG CTC CGT CAC AGC AGC CTG GAG CGG GAG TGC ATA GAG GAG ATC TGT GAC
HIS GLN VAL LEU ARG ILE ARG LYS ARG ALA ASN SER PHE LEU GLU GLU LEU ARG HIS SER SER LEU GLU ARG GLU CYS ILE GLU GLU ILE CYS ASP
                -5                              5                  10                  15                  20
        2980      2990      3000      3010      3020      3030      3040      3050      3060
TTC GAG GAG GCC AAG GAA ATT TTC CAA AAT GTG GAT GAC ACA GTA AGG CCA CCA TGG GTC CAG AGG ATG AGG CTC AGG GGC GAG CTG GTA ACC AGC
PHE GLU GLU ALA LYS GLU ILE PHE GLN ASN VAL ASP ASP THR
  25                 30                  35
   3070      3080      3090      3100      3110      3120      3130      3140      3150      3160
AGG GGC CTC GAG GAG CAG GTG GGA ACT CAA TGC TGA GGC CCT CTT AGG AGT TGT GGG GGT GGC TGA GTG GAG CGA TTA GGA TGC TGG CCC TAT GAT
       3170      3180      3190      3200                                          INTRON C
GTC GAC AGG CAC ATG TGA CTG CAA GAA CAG AAT TCA GGA A-- --- --- --- --- --- --- ---- -- --- 1300 bp --- --- --- --- --- --- ---

          4510      4520      4530      4540      4550      4560      4570      4580      4590
--G CGC GCC CCC TCC GCA CAC CGG CTG CAG GAG CCT GAC GCT GCC CGC TC TCT CCG CAG CTG GCC TTC TGG TCC AAG CAC GTC GGT GAG TGC GTTC
                                                                  LEU ALA PHE TRP SER LYS HIS VAL
                                                                          40                45
   4600      4610      4620      4630      4640      4650      4660      4670      4680      4690
TAG ATC CCC GGC TGG ACT ACC GGC GCC CGC GCC CTC GGG ATC TCT GGC CGC TGA CCC CCT ACC CCG CCT TGT GTC GCA GAC GGT GAC CAG TGC TTG
                       INTRON D                                                       ASP GLY ASP GLN CYS LEU
                                                                                                       50
       4700      4710      4720      4730      4740      4750      4760      4770      4780
GTC TTG CCC TTG GAG CAC CCG TGC GCC AGC CTG TGC TGC GGG CAC GGC ACG TGC ATC GAC GGC ATC GGC AGC TTC AGC TGC GAC TGC CGC AGC GGC
VAL LEU PRO LEU GLU HIS PRO CYS ALA SER LEU CYS CYS GLY HIS GLY THR CYS ILE ASP GLY ILE GLY SER PHE SER CYS ASP CYS ARG SER GLY
            55                  60                  65                  70                  75                  80
 4790      4800      4810      4820      4830      4840      4850      4860      4870      4880
TGG GAG GGC CGC TTC TGC CAG CGC GGT GAG GGG GAG AGG TGG ATG CTG GCG GGC GAC GGG GCG GGG CTG GGG CCG GGT TGG GGG CGC GGC ACC AGC
TRP GLU GLY ARG PHE CYS GLN ARG                                              INTRON E
  85               90
```

FIG. 2. (*Figure continues on the opposite page.*)

```
       4890      4900      4910      4920      4930      4940      4950      4960      4970
ACC AGC TGC CCG CGC CCT CCC CTG CCC GCA GAG GTG AGC TTC CTC AAT TGC TCT CTG GAC AAC GGC GGC TGC ACG CAT TAC TGC CTA GAG GAG GTG
                                        GLU VAL SER PHE LEU ASN CYS SER LEU ASP ASN GLY GLY CYS THR HIS TYR CYS LEU GLU GLU VAL
                                                    95                 100                 105                 110
4980      4990      5000      5010      5020      5030      5040      5050      5060      5070
GGC TGG CGG CGC TGT AGC TGT GCG CCT GGC TAC AAG CTG GGG GAC GAC CTC CTG CAG TGT CAC CCC GCA GGT GAG AAG CCC CCA ATA CAT CGC CCA
GLY TRP ARG ARG CYS SER CYS ALA PRO GLY TYR LYS LEU GLY ASP ASP LEU LEU GLN CYS HIS PRO ALA
 115                 120                 125                 130                 135
  5080      5090      5100      5110      5120      5130      5140      5150      5160      5170
AGA ATC ACG CTG GGT GCG GGG TGG GCA GGC CCC CTG ACG GGG CGA CGG CGC GGG GGG CTC AGG AGG GTT TCT AGG GAG GGA GCG AGG AAC AGA GTᵀ
        5180      5190      5200      5210      5220      5230                INTRON F
GAG CCT TGG GGC AGC GGC AGA CGC GCC CCA ACA CCG GGG CCA CTG TTA GCG CAA TTC AGC CCG --- --- --- --- 2480 bp --- --- --- --- ---
         7720      7730      7740      7750      7760      7770      7780      7790      7800
--G GAG GAG TGC CTG GCA GGC CCC TCA CCA CCT CTG CCT ACC TCA GTG AAG TTC CCT TGT GGG AGG CCC TGG AAG CGG ATG GAG AAG AAG CGC AGT
                                                            VAL LYS PHE PRO CYS GLY ARG PRO TRP LYS ARG MET GLU LYS LYS ARG SER
                                                                        140                 145                 150
  7810      7820      7830      7840      7850      7860      7870      7880      7890      7900
CAC CTG AAA CGA GAC ACA GAA GAC CAA GAA GAC CAA GTA GAT CCG CGG CTC ATT GAT GGG AAG ATG ACC AGG CGG GGA GAC AGC CCC TGG CAG GTG
HIS LEU LYS ARG ASP THR GLU ASP GLN GLU ASP GLN VAL ASP PRO ARG LEU ILE ASP GLY LYS MET THR ARG ARG GLY ASP SER PRO TRP GLN
 155                 160                 165                 170                 175                 180
      7910      7920      7930      7940      7950      7960      7970      7980      7990
GGA GGC GAG GCA GCA CCG GCT GCT CAC GTG CTG GGT CCG GAA TCA CTG AGT CCA TCC TGG CAG CTA TGC TCA GGG TGC AGA AAC CGA GAG GGA AGC
 8000      8010      8020      8030      8040      8050      8060      8070      8080      8090
GCT GCC ATT GCG TTT GGG GGA TGA TGA AGG TGG GGG ATG CTT CAG GGA AAG ATG GAC GCA ACC TGA GGG GAG AGG AGC AGC CAG GGT GGG TGA GGG
     8100      8110      8120        INTRON G            8660      8670      8680      8690      8700
GAG GGG CAT GGG GGC ATG GAG GGG TCT GC- --- --- 530 bp --- --- CCC AGT GGG ACC ACA GCC AGG ACG GCC CTT CAA GAT AGG GGC TGA GGG
         8710      8720      8730      8740      8750      8760      8770      8780      8790
AGG CCC AAG GGG AAC ATC CAG GCA GCC TGG GGG CCA CAA AGT CTT CCT GGA AGA CAC AAG GCC TGG CCA AGC CTC TAA GGA TGA GAG GAG CTC GCT
  8800      8810      8820      8830      8840      8850      8860      8870      8880      8890
GGG CGA TGT TGG GTG TGG CTG AGG GTG ACC GAA ACA GTA TGA ACA GTG CAG GAA CAG CAT GGG CAA AGG CAG GAA GAC ACC CTG GGA CAG GCT GAC
        8900      8910      8920      8930      8940      8950      8960      8970      8980
ACT GTA AAA TGG GCA AAA ATA GAA AAC GCC AGA AAG GGC CTA AGC CTA TGC CCA TAT GAC CAG GGA ACC CAG GAA AGT GCA TAT GAA ACC CAG GTG
 8990      9000      9010      9020      9030      9040      9050      9060      9070      9080
CCC TGG ACT GGA GGC TGT CAG GAG GCA GCC CTG TGA TGT CAT CAT CCC ACC CCA TTC CAG GTG GTC CTG CTG GAC TCA AAG AAG AAG CTG GCC TGC
                                                                                VAL VAL LEU LEU ASP SER LYS LYS LYS LEU ALA CYS
                                                                                185                 190                 195
    9090      9100      9110      9120      9130      9140      9150      9160      9170      9180
GGG GCA GTG CTC ATC CAC CCC TCC TGG GTG CTG ACA GCG GCC CAC TGC ATG GAT GAG TCC AAG AAG CTC CTT GTC AGG CTT GGT ATG GGC TGG AGC
GLY ALA VAL LEU ILE HIS PRO SER TRP VAL LEU THR ALA ALA HIS CYS MET ASP GLU SER LYS LYS LEU LEU VAL ARG LEU
            200                 205                 210                 215                 220
            9190      9200      9210      9220      9230      9240      9250      9260      9270
CAG GCA GAA GGG GGC TGC CAG AGG CTT GGG TAG GGG GAC TAG GCA GGC TGT TCA GGT TTG GGG GAC CCC GCT CCC CAG GTG CTT AAG CAA GAG GCT
   9280      9290      9300      9310      9320      9330      9340      9350      9360      9370
TCT TGA GCT CCA CAG AAG GTG TTT GGG GGG AAG AGG CCT ATG TGC CCC CAC CCT GCC CAC CCA TGT ACA CCC AGT ATT TTG CAG TAG GGG GTT CTC
       9380      9390      9400      9410      9420      9430      9440      9450      9460
TGG TGC CCT CTT CGA ATC TGG GCA CGG TAC CTG CAC ACA CAC ATG TTT GTG AGG GGC TAC ACA GAC CTT CAC CTC TCC ACT CCC ACT CAT GAG GAG
 9470      9480      9490      9500      9510      9520      9530      9540      9550      9560
CAG GCT GTG TGG GCC TCA GCA CCC TTG GGT GCA GAG ACC AGC AAG GCC TAG CCT CAG GGC TGT GCC TCC CAC AGA CTG ACA GGG ATG GAG CTG TAC
    9570      9580      9590      9600      9610      9620      9630      9640      9650      9660
AGA GGG AGC CCT AGC ATC TGC CAA AGC CAC AAG CTG CTT CCC TAG CAG GCT GGG GGC ACC TAT GCA TTG GCC CCG ATC TAT GGC AAT TTC TGG AGG
            9670      9680      9690      9700      9710      9720      9730      9740      9750
GGG GGT CTG GCT CAA CTC TTT ATG CCA AAA AGA AGG CAA GCA TAT TGA GAA AGG CCA AAT TCA CAT TTC CTA CAG CAT AAT CTA TGG CCA GTG GCC
   9760      9770      9780      9790      9800      9810      9820      9830      9840    INTRON H
CCC CGT GGG GCT TGG CTT AGA ATT CCC AGG TGC TCT TCC CAG GGA ACC ATC AGT CTG GAC TGA GAG GAC CTT CTC TCT CAG GTG GG- -- 240 bp ---
       10090     10100     10110     10120     10130     10140     10150     10160     10170
-CT CAC GAC TCC GTG ACT CCT GAA AAC CAA CCA GCA TCC TAC CCC TTT GGG ATT GAC ACC TGT TGG CCA CTC CTT CTG GCA GGA AAA GTC ACC GTT
 10180     10190     10200     10210     10220     10230     10240     10250     10260     10270
GAT AGG GTT CCA CGG CAT AGA CAG GTG GCT CCG CGC CAG TGC CTG GGA CGT GTG GGT GCA CAG TCT CCG GGT GAA CCT TCT TCA GGC CCT CTG CCC
      10280     10290     10300     10310     10320     10330     10340     10350     10360
AGG CCT GCT GCA GGA GAG TAT GAC CTG CGG CGC TGG GAG AAG TGG GAG CTG GAC CTG GAC ATC AAG GAG GTC TTC GTC CAC CCC AAC TAC AGC AAG
                GLY GLU TYR ASP LEU ARG ARG TRP GLU LYS TRP GLU LEU ASP LEU ASP ILE LYS GLU VAL PHE VAL HIS PRO ASN TYR SER LYS
                    225                 230                 235                 240                 245                 250
10370     10380     10390     10400     10410     10420     10430     10440     10450     10460
AGC ACC ACC GAC AAT GAC ATC GCA CTG CTG CAC CTG GCC CAG CCC GCC ACC CTC TCG CAG ACC ATA GTG CCC ATC TGC CTC CCG GAC AGC GGC CTT
SER THR THR ASP ASN ASP ILE ALA LEU LEU HIS LEU ALA GLN PRO ALA THR LEU SER GLN THR ILE VAL PRO ILE CYS LEU PRO ASP SER GLY LEU
            255                 260                 265                 270                 275                 280
   10470     10480     10490     10500     10510     10520     10530     10540     10550     10560
GCA GAG CGC GAG CTC AAT CAG GCC GGC CAG GAG ACC CTC GTG ACG GGC TGG GGC TAC CAC AGC AGC CGA GAG AAG GAG GCC AAG AGA AAC CGC ACC
ALA GLU ARG GLU LEU ASN GLN ALA GLY GLN GLU THR LEU VAL THR GLY TRP GLY TYR HIS SER SER ARG GLU LYS GLU ALA LYS ARG ASN ARG THR
 285                 290                 295                 300                 305                 310                 315
      10570     10580     10590     10600     10610     10620     10630     10640     10650
TTC GTC CTC AAC TTC ATC AAG ATT CCC GTG GTC CCG CAC AAT GAG TGC AGC GAG GTC ATG AGC AAC ATG GTG TCT GAG AAC ATG CTG TGT GCG GGC
PHE VAL LEU ASN PHE ILE LYS ILE PRO VAL VAL PRO HIS ASN GLU CYS SER GLU VAL MET SER ASN MET VAL SER GLU ASN MET LEU CYS ALA GLY
                320                 325                 330                 335                 340                 345
  10660     10670     10680     10690     10700     10710     10720     10730     10740     10750
ATC CTC GGG GAC CGG CAG GAT GCC TGC GAG GGC GAC AGT GGG GGG CCC ATG GTC GCC TCC TTC CAC GGC ACC TGG TTC CTG GTG GGC CTG GTG AGC
ILE LEU GLY ASP ARG GLN ASP ALA CYS GLU GLY ASP SER GLY GLY PRO MET VAL ALA SER PHE HIS GLY THR TRP PHE LEU VAL GLY LEU VAL SER
        350                 355                 360                 365                 370                 375
      10760     10770     10780     10790     10800     10810     10820     10830     10840
TGG GGT GAG GGC TGT GGG CTC CTT CAC AAC TAC GGC GTT TAC ACC AAA GTC AGC CGC TAC CTC GAC TGG ATC CTC CAT GGG CAC ATC AGA GAC AAG
TRP GLY GLU GLY CYS GLY LEU LEU HIS ASN TYR GLY VAL TYR THR LYS VAL SER ARG TYR LEU ASP TRP ILE LEU HIS GLY HIS ILE ARG ASP LYS
380                 385                 390                 395                 400                 405                 410
10850     10860     10870     10880     10890     10900     10910     10920     10930     10940
GAA GCC CCC CAG AAG AGC TGG GCA CCT TAG CGA CCC TCC CTG CAG GGC TGG GCT TTT GCA TGG CAA TGG ATG GGA CAT TAA AGG GAC ATG TAA CAA
GLU ALA PRO GLN LYS SER TRP ALA PRO STOP
            415                 420
    10950     10960     10970     10980     10990     11000     11010     11020     11030     11040
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
          11050     11060     11070     11080     11090     11100     11110     11120     11130
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
  11140     11150     11160     11170     11180     11190     11200     11210     11220     11230
CTC TGT TTA TGA AAA AGA ATA AAA AAC ACA ACC ACG AAG CCA CTA GAG CCT TTT CCA GGG CTT TGG GAA GAG CCT GTG CAA GCC GGG GAT GCT GAA
        11240     11250     11260     11270     11280     11290     11300     11310     11320
GGT GAG GCT TGA CCA GCT TTC CAG CTA GCC CAG CTA TGA GGT AGA CAT GTT TAG CTC ATA TCA CAG AGG AGG AAA CTG AGG GGT CTG AAA GGT TTA
11330     11340
CAT GGT GGA GTT --- --- --- --- --- --- --- --- --- --- --- ---
```

FIG. 2. Nucleic acid sequence of the protein C gene. Bases are numbered relative to the proposed transcription initiation site. Exons are underlined and amino acids are numbered from the amino-terminal residue in the plasma protein. Gaps in the sequence are shown as dashed lines, with the approximate length of the gap noted. Regions corresponding to a TATA box (−34 to −25), a transcriptional start site (−2 to +6) and a polyadenylylation recognition site (11,155 to 11,160) are boxed.

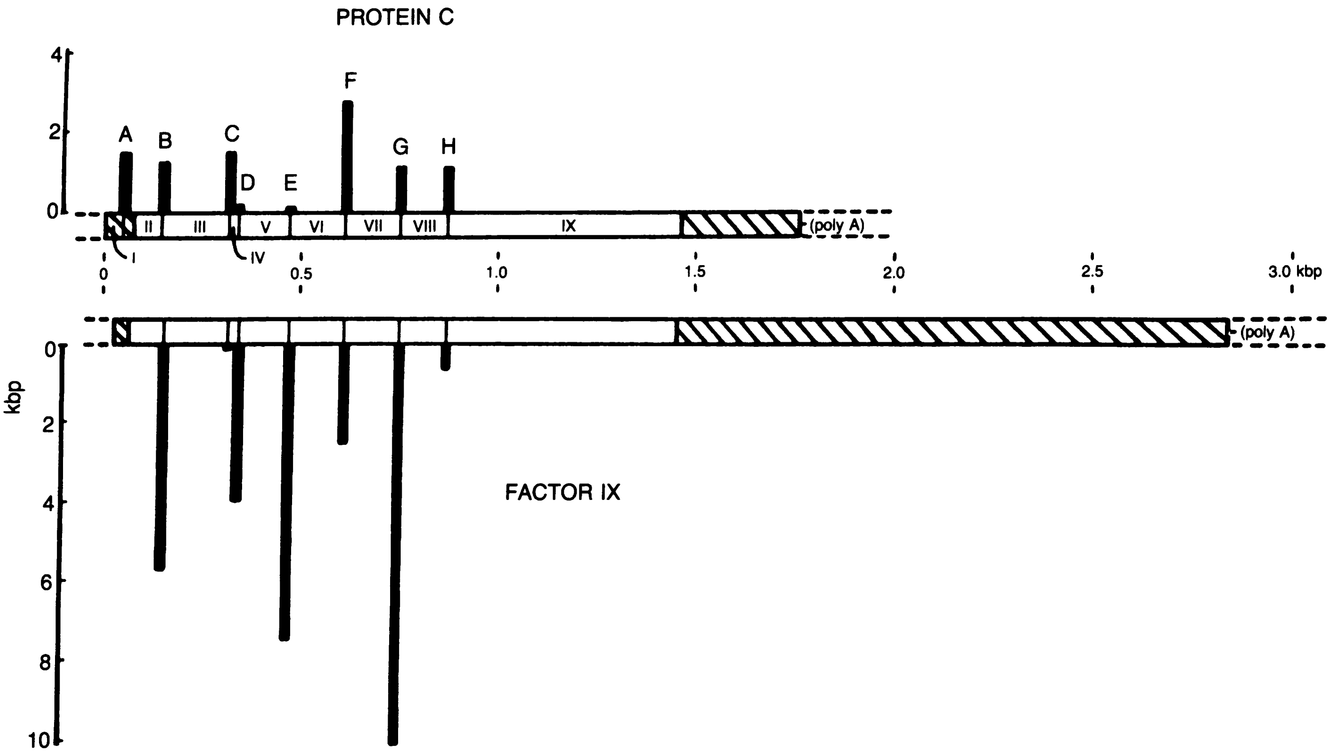

FIG. 3. Comparison of the size of introns and exons in the protein C gene. The length in bp of exons is shown by the horizontal bars while the length of each intron is proportional to the height of each vertical bar. The horizontal bars have been aligned on the basis of amino acid sequence homology. Hatched horizontal regions represent noncoding regions of the mRNAs.

5′ and 3′ nontranslated regions are divided into nine exons by eight introns. There is perfect agreement between the sequence of the exons and the full length cDNA sequence (36). This again indicates that only a single copy of the protein C gene is present in the human genome. Analysis of the sequences at the borders of introns is shown in Fig. 3 and reveals that in every case the GT/AG rule (21) is observed. The entire length of the gene is 11.2 kbp; hence 83% of the gene consists of intronic sequences. The 3′ nontranslated region contains a polyadenylylation signal (AAUAAA) 21 nucleotides from the termination codon and polyadenylylation occurs on either an adenosine or on the preceding guanosine.

Transcription Initiation Site. Comparison of the total length of the cDNA sequence with the mRNA size based upon RNA blot hybridization (36) suggests that the cDNA is very close to full length. Analysis of the genomic sequence immediately upstream from the region corresponding to the mRNA sequence (underlined in Fig. 2) reveals one potential transcriptional start site, one base upstream from the cloned cDNA sequence. This proposed start site is based upon sequence similarity with the transcriptional start site consensus sequence (PyCAPyPyPyPyPy) reported by Corden *et al.* (22), and its position 30 bases downstream from a potential "TATA" box (21). Alternatively, the adjacent upstream genomic sequence could be a part of a second intron in the 5′ untranslated region.

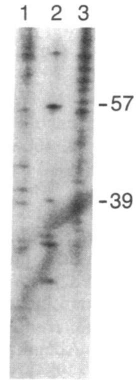

FIG. 4. The start-site for transcription of the protein C gene. Primer extension was carried out. Lanes 1 and 3, M13 sequence ladders used to obtain size markers. The figures shown on the right are the length, in bp, from the end of the oligonucleotide used for primer extension. The cluster of fainter bands at the top of the gel correspond to 72–85 nucleotides from the end of the oligonucleotide.

Analysis of transcripts produced by primer extension using an oligonucleotide (Fig. 4) corresponding to positions 30–53 of the protein C sequence indicated a start site 54–56 nucleotides from the end of the primer corresponding to the nucleotides about 1 bp upstream of the cDNA sequence shown in Fig. 2. In addition, several weaker bands corresponding to longer and shorter transcripts were present on the gel (Fig. 4), and it is unclear if these longer transcripts represent additional sites of initiation of transcription or crosshybridization to other mRNAs.

DISCUSSION

Comparison of the Gene Structures of the Serine Proteases. Protein C, like many of the blood coagulation proteins, is a serine protease and exerts its role in blood coagulation by virtue of an active-site serine in the heavy chain. Neurath *et al.* (23) have found that each of these genes is likely to be derived from the same primitive gene, and chymotrypsin, trypsin, and elastase are generally felt to be the archetypal serine proteases. The trypsin gene family can be traced to prokaryotes by virtue of sequence homology between mammalian, invertebrate, and prokaryotic trypsins. In this evolutionary pathway, invertebrate trypsin from crayfish represents the link, having homology to mammalian and prokaryotic trypsins (24).

Evolutionary relatedness between proteins is commonly established by comparing amino acid homology between the proteins. However, with the ability to study the structure of the genes encoding these proteins, it is possible to provide an independent test of these estimates of relatedness, in that the number and positions of introns are likely to be a relatively immutable feature and provide an independent estimate of the relatedness of two proteins. Several groups have explored the similarities of serine proteases (see ref. 25) and with the same objective we have studied the positions of introns in the genes of some of the members of the serine protease family. Remarkably, a comparison of the gene region corresponding to the serine protease domain of protein C with the rat genes (25) for trypsin (three introns), chymotrypsin (five introns), and elastase (six introns) reveals that the sole protein C intron position within this domain aligns only with the first intron of elastase. In contrast, when homologous amino acids are aligned, the position of all eight introns of factor IX (26) and protein C correspond (Fig. 3). This fact attests to the close evolutionary relationship between the two genes and suggests that they are products of a relatively recent duplication.

Despite the conservation of exon size and sequence between protein C and factor IX, the sizes of the introns have diverged remarkably (Fig. 3). Furthermore, no recognizable homology exists between intron sequences of the two genes among the 3 kbp of intronic sequences presented in this paper [compare Fig. 2 of this manuscript with Fig. 4 of Anson *et al.* (26)].

Banjai *et al.* (27) have noted a region within several coagulation factors with close homology to epidermal growth factor (28) and, more recently, to the transforming growth factors produced by retroviral-infected cells (29). Interestingly, these regions (amino acids 46–91 and 92–136 in the protein C sequence) is sharply delimited by exons in protein C, factor IX, factor X, and tissue plasminogen activator (30). The conservation of these regions and its presence on a single exon in such diverse proteins as tissue plasminogen activator and protein C suggests that these regions were incorporated into the structure of the genes before the original duplication and have maintained a remarkable degree of homology to transforming growth factor and epidermal growth factor through several hundred million years of evolution. In particular, the position of each half cysteine bond remains the same in this region, suggesting that each of these primary structures has a similar tertiary structure. Thus it is possible that these regions have some conserved function, possibly related to that of the growth factors.

Prothrombin, another vitamin K-dependent serine protease coagulation factor, does not contain the growth factor domain but has in its place two "kringle" structures, first described by Magnusson *et al.* (31). Prothrombin, however, does have clear homology with other vitamin K-dependent coagulation factors in the leader peptide, γ-carboxylglutamate or GLA region and serine protease domains (6, 32, 36).

Congenital deficiency of protein C is an autosomally dominant disease characterized by superficial and deep venous thrombosis (11) occurring during childhood or early adulthood. The heterozygous form of the disease has an incidence of 1 per 16,000 individuals (33). Individuals that inherit two defective protein C alleles have catastrophic thrombotic events as neonates that are invariably fatal without treatment (13). Because individuals with a defective protein C allele do not always manifest the disease or have abnormally low levels of the protein (34), a precise genetic test for the defective allele would be useful. With the organization of the normal protein C gene defined, the feasibility of genetic testing for hereditary protein C deficiency can now be determined.

Note Added in Proof. Since submission of this manuscript, an additional sequence for the human protein C gene has been published (35).

We wish to thank Dr. Philip Leder for the phage genomic libraries, Kathy Shaw for editorial assistance, and Rama M. Belagaje for synthesis of the oligonucleotide. We also thank John Shephard of Lilly Research Laboratories for technical assistance in DNA sequencing. These studies were supported in part by a grant from the Lilly Research Laboratories to G.R.C. and Grant HL33942-01 from the National Institutes of Health.

1. Stenflo, J. (1984) *Semin. Thromb. Hemostasis* **10,** 109–121.
2. Kisiel, W., Canfield, W. M., Ericsson, L. H. & Davie, E. W. (1977) *Biochemistry* **16,** 5824–5831.
3. Sakata, Y., Curriden, S., Lawrence, D., Griffin, J. H. & Loskutoff, D. J. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 1121–1125.
4. Stenflo, J. (1976) *J. Biol. Chem.* **251,** 355–363.
5. Kisiel, W. (1979) *J. Clin. Invest.* **64,** 761–769.
6. Fernlund, P. & Stenflo, P. (1982) *J. Biol. Chem.* **257,** 12170–12179.
7. Esmon, N. L., Owen, W. G. & Esmon, C. T. (1982) *J. Biol. Chem.* **257,** 859–864.
8. Esmon, C. T. & Owen, W. G. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 2249–2252.
9. Marlar, R. A. & Griffin, J. H. (1980) *J. Clin. Invest.* **66,** 1186–1189.
10. Griffin, J. H., Evatt, B., Zimmerman, T. S., Kleiss, A. J. & Wideman, C. (1981) *J. Clin. Invest.* **68,** 1370–1373.
11. Bertina, R. M., Broekmans, A. W., Van der Linden, I. K. & Mertens, K. (1982) *Throm. Hemostasis Gen. Inf.* **48,** 1–5.
12. Branson, H. E., Katz, J., Marble, R. & Griffin, J. H. (1983) *Lancet* **ii,** 1165–1168.
13. Comp, P. E. & Esmon, C. T. (1984) *N. Engl. J. Med.* **311,** 1525–1528.
14. Foster, D. & Davie, E. W. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 4766–4770.
15. Benton, W. D. & Davis, R. W. (1977) *Science* **196,** 180–182.
16. Southern, E. M. (1975) *J. Mol. Biol.* **98,** 503–517.
17. Smith, G. E. & Summers, M. D. (1980) *Anal. Biochem.* **109,** 123–129.
18. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560.
19. Crabtree, G. R. & Kant, J. K. (1982) *Cell* **31,** 159–166.
20. Crabtree, G. R., Comeau, C. M., Fowlkes, D. M., Fornace, A. J., Malley, J. D. & Kant, J. A. (1985) *J. Mol. Biol.*, in press.
21. Breathnach, R., Benoist, O., O'Hare, K., Gannon, F. & Chambon, P. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 4853–4857.
22. Corden, J., Wasylyk, B., Buchwalder, A., Sassone-Corsi, P., Kedinger, C. & Chambon, P. (1980) *Science* **209,** 1406–1414.
23. Neurath, H. (1984) *Science* **224,** 350–357.
24. Titani, K., Sasagawa, T., Woodbury, R. G., Ericsson, L. H., Dorsam, H., Kraemer, M., Neurath, H. & Zwilling, R. (1983) *Biochemistry* **22,** 1459–1464.
25. Craik, C. S., Rutter, W. J. & Fletterick, R. (1983) *Science* **220,** 1125–1129.
26. Anson, D. S., Choo, K. H., Rees, D. J. G., Giannelli, F., Gould, K., Huddleston, J. A. & Brownlee, G. G. (1984) *EMBO J.* **3,** 1053–1060.
27. Banjai, L., Varadi, A. & Patthy, L. (1983) *FEBS Lett.* **163,** 37–41.
28. Gregory, H. & Preston, B. M. (1977) *Int. J. Pept. Protein Res.* **9,** 107–118.
29. Marquardt, H., Hunkapiller, M. W., Hood, L. E., Twardzik, D. R., De Larco, J. E., Stephenson, J. R. & Todaro, G. J. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 4684–4688.
30. Ny, T., Elgh, F. & Lund, B. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 5355–5359.
31. Magnusson, S., Sottrup-Jensen, L. & Peterson, T. E. (1976) in *Proteolysis and Physiological Regulation*, eds. Ribbons, D. W. & Brew, K. (Academic, New York), pp. 203–238.
32. Stenflo, J. & Fernlund, P. (1982) *J. Biol. Chem.* **257,** 12180–12185.
33. Brockmans, A. N., Van der Linden, I. K., Veltkamp, J. J. & Bertina, R. M. (1983) *Thromb. Hemostasis Gen. Inf.* **50,** 1096.
34. Seligsohn, U., Berger, A., Abend, M., Rubin, L., Atties, D., Zivelin, A. & Rapaport, S. I. (1984) *N. Engl. J. Med.* **310,** 559–562.
35. Foster, D. C., Yoshitake, S. & Davie, E. W. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 4673–4677.
36. Beckmann, R. J., Schmidth, R. J., Santerre, R. F., Plutzky, J., Crabtree, G. R. & Long, G. L. (1985) *Nucleic Acids Res.* **13,** 5233–5247.