Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 4673–4677, July 1985
Biochemistry

The nucleotide sequence of the gene for human protein C

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

Donald C. Foster, Shinji Yoshitake, and Earl W. Davie

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

Contributed by Earl W. Davie, April 9, 1985

ABSTRACT **A human genomic DNA library was screened for the gene for protein C by using a cDNA probe coding for the human protein. Three different overlapping λ Charon 4A phage were isolated that contain inserts for the gene for protein C. The complete sequence of the gene was determined by the dideoxy method and shown to span about 11 kilobases of DNA. The coding and 3′ noncoding portion of the gene consists of eight exons and seven introns. The eight exons code for a preproleader sequence of 42 amino acids, a light chain of 155 amino acids, a connecting dipeptide of Lys-Arg, and a heavy chain of 262 amino acids. The preproleader sequence and the connecting dipeptide are removed during processing, resulting in the mature protein composed of a heavy and a light chain held together by a disulfide bond. The heavy chain also contains the catalytic region for the serine protease. Two *Alu* sequences and two homologous repeats of about 160 nucleotides were found in intron E. The seven introns in the gene for protein C are located in essentially the same positions in the amino acid sequence as the seven introns in the gene for human factor IX, while the first three introns in protein C are located in the same positions as the first three in the gene for human prothrombin.**

Protein C is a precursor to a serine protease present in plasma that plays an important physiological role in the regulation of blood coagulation (1, 2). Human protein C is a vitamin K-dependent glycoprotein containing nine residues of γ-carboxyglutamic acid and one equivalent of β-hydroxyaspartic acid. Protein C shows considerable structural homology with the other vitamin K-dependent plasma proteins involved in blood coagulation, including prothrombin, factor VII, factor IX, and factor X. Protein C is synthesized as a single-chain polypeptide that undergoes considerable processing to give rise to a two-chain molecule held together by a disulfide bond. The two-chain form is converted to activated protein C by thrombin by the cleavage of a 12-residue peptide from the amino terminus of the heavy chain (2). This reaction is greatly accelerated by the presence of thrombomodulin (3). Activated protein C regulates the coagulation process by the inactivation of factor V_a (4, 5) and factor $VIII_a$ (4, 6) by minor proteolysis. Consequently, individuals lacking protein C often have a history of thrombotic disease (7, 8).

Studies from our laboratory (9) and that of others (10) have led to the isolation and characterization of the cDNA coding for human and bovine protein C. In the present investigation, the cDNA for human protein C has been used for the isolation of overlapping genomic clones from a λ Charon 4A phage library. The nucleotide sequence of the gene was then determined and compared with the genes for human factor IX (11, 12) and prothrombin (13).

MATERIALS AND METHODS

Screening of the Genomic Library. A human genomic library in λ Charon 4A phage (14) was screened for genomic clones of human protein C by the plaque hybridization procedure of Benton and Davis as modified by Woo (15) using a cDNA for human protein C (9) as the hybridization probe. The cDNA started at amino acid 64 of human protein C and extended to the second polyadenylylation signal (9). It was radiolabeled by nick-translation to a specific activity of 8×10^8 cpm/μg with all four radioactive ([α-^{32}P]dNTP) deoxynucleotides. The probe was denatured and hybridized to the filters at a concentration of 1×10^6 cpm/ml in a hybridization solution containing 6× NaCl/P_i (1× NaCl/P_i = 0.15 M NaCl/0.015 M sodium citrate, pH 7.0), 5× Denhardt's solution (1× = 0.02% polyvinylpyrrolidone/0.02% Ficoll/0.02% bovine serum albumin), 0.1% sodium dodecyl sulfate, 100 μg of yeast tRNA per ml, and 50% formamide at 42°C for 60 hr. The filters were washed in 1× NaCl/P_i containing 0.1% sodium dodecyl sulfate at 68°C for 1 hr and exposed to x-ray film for 16 hr. Positive clones were then isolated and plaque-purified.

DNA Sequence Analysis. Phage DNA was prepared from positive clones by the liquid culture lysis method as described by Silhavy *et al.* (16). The genomic DNA inserts in the purified phage were removed by digestion with *Eco*RI and then subcloned into pUC9 for subsequent restriction mapping and sequencing. In order to obtain overlapping DNA fragments, the DNA inserts were digested also with *Bgl* II, and the fragments corresponding to the gene for protein C were subcloned into the *Bam*HI site of pUC9.

The sequence of genomic fragments containing the gene for protein C was determined both by direct cloning of specific restriction fragments into the M13 phage cloning vectors mp10, mp11, mp18, and mp19, as well as by the BAL-31 exonuclease method described by Guo *et al.* (17) and Yoshitake *et al.* (12).

Dideoxy chain termination sequencing reactions were carried out with ^{35}S-substituted deoxyadenosine 5′-[α-thio]triphosphate (dATP[α-^{35}S]; Amersham) essentially as described in the sequencing manual provided by Amersham and run on buffer gradient gels as described by Biggin *et al.* (18). More than 90% of the sequence was determined two or more times, and ≈50% was determined on both strands. DNA sequences were stored and analyzed by the computer programs of Larson and Messing (19).

M13 vectors mp10, mp11, mp18, and mp19, deoxynucleotide triphosphates, and dideoxynucleotide triphosphates were purchased from P-L Biochemicals. 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.

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: kb, kilobase(s).

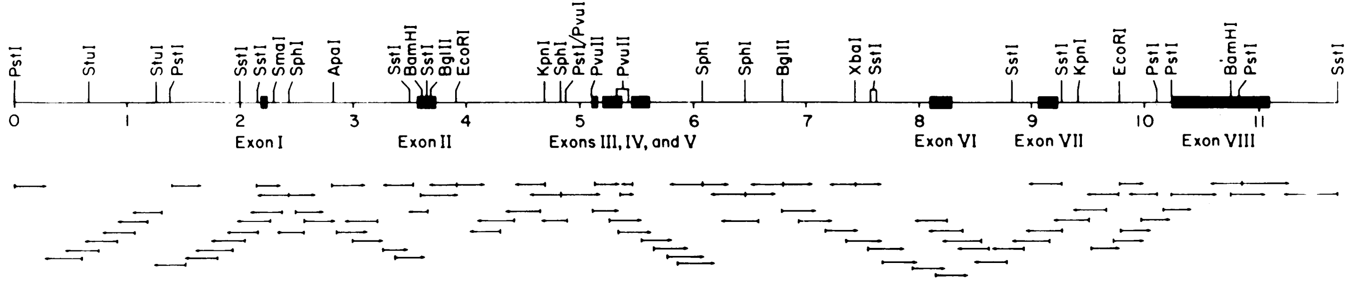

FIG. 1. Detailed restriction map and sequencing strategy for the gene for human protein C. The locations of each of the eight exons are shown with solid bars. The length and direction of each sequencing reaction are shown by thin arrows.

RESULTS AND DISCUSSION

A human genomic DNA library (2 × 10⁶ phage) in λ Charon 4A phage was screened with a radiolabeled cDNA probe for human protein C. Three different positive clones were isolated, and each was plaque-purified. These three clones exhibited unique patterns of *Eco*RI fragments upon electrophoresis in 0.7% agarose but also contained fragments in common with each other. Southern blot hybridization of digests of these clones with probes made from the 5′ and 3′ ends of the cDNA established that one of the clones (PCλ1) corresponded to the 5′ region of the gene for protein C, another (PCλ8) to the 3′ region, and the third (PCλ6) was positive to both sets of probes.

The genomic DNA inserts in PCλ6 and PCλ8 were mapped by single- and double-restriction-enzyme digestion followed by agarose gel electrophoresis, Southern blotting, and hybridization to radiolabeled 5′ and 3′ probes derived from the cDNA for human protein C. This analysis suggested that the gene for protein C was present in three *Eco*RI fragments of 4.4, 6.2, and 6.9 kilobases (kb) oriented 5′ to 3′ in the genome. The 4.4-kb fragment was isolated from phage PCλ6, and the 6.2-kb and 6.9-kb fragments were isolated from phage PCλ8; each was subcloned into the *Eco*RI site of pUC9. To provide

```
AGTGAATCTG GGCGAGTAAC ACAAAACTTG AGTGTCCTTA CCTGAAAAAT AGAGGTTAGA GGGATGCTAT GTGCCATTGT GTGTGTGTGT TGGGGGTGGG GATTGGGGGT GATTTGTGAG CAATTGGAGG -2001
TGAGGGTGGA GCCCAGTGCC CAGCACCTAT GCACTGGGGA CCCAAAAAGG AGCATCTTCT CATGATTTTA TGTATCAGAA ATTGGGATGG CATGTCATTG GGACAGCGTC TTTTTTCTTG TATGGTGGCA -1871
CATAAATACA TGTGTCTTAT AATTAATGGT ATTTTAGATT TGACGAAATA TGGAATATTA CCTGTTGTGC TGATCTTGGG CAAACTATAA TATCTCTGGG CAAAAATGTC CCCATCTGAA AAACAGGGAC -1741
AACGTTCCTC CCTCAGCCAG CCACTATGGG GCTAAAATGA GACCACATCT GTCAAGGGTT TTGCCCTCAC CTCCCTCCCT GCTGGATGGC ATCCTTGGTA GGCAGAGGTG GGCTTCGGGC AGAACAAGCC -1611
GTGCTGAGCT AGGACCAGGA GTGCTAGTGC CACTGTTTGT CTATGGAGAG GGAGGCCTCA GTGCTGAGGG CCAAGCAAAT ATTTGTGGTT ATGGATTAAC TCGAACTCCA GGCTGTCATG GCGGCAGGAC -1481
GGCGAACTTG CAGTATCTCC ACGACCCGCC CCTGTGAGTC CCCCTCCAGG CAGGTCTATG AGGGGTGTGG AGGGAGGGCT GCCCCCGGGA GAAGAGAGCT AGGTGGTGAT GAGGGCTGAA TCCTCCAGCC -1351
AGGGTGCTCA ACAAGCCTGA GCTTGGGGTA AAAGGACACA AGGCCCTCCA CAGGCCAGGC CTGGCAGCCA CAGTCTCAGG TCCCTTTGCC ATGCGCCTCC CTCTTTCCAG GCCAAGGGTC CCCAGGCCCA -1221
GGGCCATTCC AACAGACAGT TTGGAGCCCA GGACCCTCCA TTCTCCCCAC CCCACTTCCA CCTTTGGGGG TGTCGGATTT GAACAAATCT CAGAAGCGGC CTCAGAGGGA GTCGGCAAGA ATGGAGAGCA -1091
GGGTCCGGTA GGGTGTGCAG AGGCCACGTG GCCTATCCAC TGGGGAGGGT TCCTTGATCT CTGGCCACCA GGGCTATCTC TGTGGCCTTT TGGAGCAACC TGGTGGTTTG GGGCAGGGGT TGAATTTCCA  -961
GGCCTAAAAC CACACAGGCC TGGCCTTGAG TCCTGGCTCT GCGAGTAATG CATGGATGTA AACATGGAGA CCCAGGACCT TGCCTCAGTC TTCCGAGTCT GGTGCCTGCA GTGTACTGAT GGTGTGAGAC  -831
CCTACTCCTG GAGGATGGGG GACAGAATCT GATCGATCCC CTGGGTTGGT GACTTCCCTG TGCAATCAAC GGAGACCAGC AAGGGTTGGA TTTTTAATAA ACCACTTAAC TCCTCCGAGT CTCAGTTTCC  -701
CCCTCTATGA AATGGGGTTG ACAGCATTAA TAACTACCTC TTGGGTGGTT GTGAGCCTTA ACTGAAGTCA TAATATCTCA TGTTTACTGA GCATGAGCTA TGTGCAAAGC CTGTTTTGAG AGCTTTATGT  -571
GGACTAACTC CTTTAATTCT CACAACACCC TTTAAGGCAC AGATACACCA CGTTATTCCA TCCATTTTAC AAATGAGGAA ACTGAGGCAT GGAGCAGTTA AGCATCTTGC CCAACATTGC CCTCCAGTAA  -441
GTGCTGGAGC TGGAATTTGC ACCGTGCAGT CTGGCTTCAT GGCCTGCCCT GTGAATCCTG TAAAAATTGT TTGAAAGACA CCATGAGTGT CCAATCAACG TTAGCTAATA TTCTCAGCCC AGTCATCAGA  -311
CCGGCAGAGG CAGCCACCCC ACTGTCCCCA GGGAGGACAC AAACATCCTG GCACCCTCTC CACTGCATTC TGGAGCTGCT TTCTAGGCAG GCAGTGTGAG CTCAGCCCCA CGTAGAGCGG GCAGCCGAGG  -181
CCTTCTGAGG CTATGTCTCT AGCGAACAAG GACCCTCAAT TCCAGCTTCC GCCTGACGGC CAGCACACAG GGACAGCCCT TTCATTCCGC TTCCACCTGG GGGTGCAGGC AGAGCAGCAG CGGGGGTAGC   -51
                                                       -42
                                                       Met Trp Gln Leu Thr Ser Leu Leu Leu Phe Val Ala Thr Trp Gly Ile Ser Gly Thr Pro Ala
ACTGCCCGGA GCTCAGAAGT CCTCCTCAGA CAGGTGCCAG TGCCTCCAGA ATG TGG CAG CTC ACA AGC CTC CTG CTG TTC GTG GCC ACC TGG GGA ATT TCC GGC ACA CCA GCT     63
 -20
Pro Leu
CCT CTT G▼GTAAGGCCAC CCCACCCCTA CCCCGGGACC CTTGTGGCCT CTACAAGGCC CTGGTGGCAT CTGCCCAGGC CTTCACAGCT TCCACCATCT CTCTGAGCCC TGGGTGAGGT GAGGGGCAGA   190

TGGGAATGGC AGGAATCAAC TGACAAGTCC CAGGTAGGCC AGCTGCCAGA GTGCCACACA GGGGCTGCCA GGGCAGGCAT GCGTGATGGC AGGGAGCCCC GCGATGACCT CCTAAAGCTC CCTCCTCCAC   320
ACGGGGATGG TCACAGAGTC CCCTGGGCCT TCCCTCTCCA CCCACTCACT CCCTCAACTG TGAAGACCCC AGGCCCAGGC TACCGTCCAC ACTATCCAGC ACAGCCTCCC CTACTCAAAT GCACACTGGC   450
CTCATGGCTG CCCTGCCCCA ACCCCTTTCC TGGTCTCCAC AGCCAACGGG AGGAGGCCAT GATTCTTGGG GAGGTCCGCA GGCACATGGG CCCCTAAAGC CACACCAGGC TGTTGGTTTC ATTTGTGCCT   580
TTATAGAGCT GTTTATCTGC TTGGGACCTG CACCTCCACC CTTTCCCAAG GTGCCCTCAG CTCAGGCATA CCCTCCTCTA GGATGCCTTT TCCCCCATCC CTTCTTGCTC ACACCCCCAA CTTGATCTCT   710
CCCTCCTAAC TGTGCCCTGC ACCAAGACAG ACACTTCACA GAGCCCAGGA CACACCTGGG GACCCTTCCT GGGTGATAGG TCTGTCTATC CTCCAGGTGT CCCTGCCCAA GGGGAGAAGC ATGGGGAATA   840
CTTGGTTGGG GGAGGAAAGG AAGACTGGGG GGATGTGTCA AGATGGGGCT GCATGTGGTG TACTGGCAGA AGAGTGAGAG GATTTAACTT GGCAGCCTTT ACAGCAGCAG CCAGGGCTTG AGTACTTATC   970
TCTGGGCCAG GCTGTATTGG ATGTTTTACA TGACGGTCTC ATCCCCATGT TTTTGGATGA GTAAATTGAA CCTTAGAAAG GTAAAGACAC TGGCTCAAGG TCACACAGAG ATCGGGGTGG GGTTCACAGG  1100
GAGGCCTGTC CATCTCAGAG CAAGGCTTCG TCCTCCAACT GCCATCTGCT TCCTGGGGAG GAAAAGAGCA GAGGACCCCT GCGCCAAGCC ATGACCTAGA ATTAGAATGA GTCTTGAGGG GGCGGAGACA  1230
                                                                                                                           -19
                                                                                                                         ▼Asp Ser Val Phe Ser Ser Ser
AGACCTTCCC AGGCTCTCCC AGCTCTGCTT CCTCAGACCC CCTCATGGCC CCAGCCCCTC TTAGGCCCCT CACCAAGGTG AGCTCCCCTC CCTCCAAAAC CAG▼ AC TCA GTG TTC TCC AGC AGC  1353
                                                 -1  +1
Glu Arg Ala 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
GAG CGT GCC 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  1458
                                                         37
    Phe Glu Glu Ala Lys Glu Ile Phe Gln Asn Val Asp Asp Thr
    TTC GAG GAG GCC AAG GAA ATT TTC CAA AAT GTG GAT GAC ACA▼GTAAGGCCAC CATGGGTCCA GAGGATGAGG CTCAGGGGCG AGCTGGTAAC CAGCAGGGGC CTCGAGGAGC  1570

AGGTGGGGAC TCAATGCTGA GGCCCTCTTA GGAGTTGTGG GGGTGGCTGA GTGGAGCGAT TAGGATGCTG GCCCTATGAT GTCGGCCAGG CACATGTGAC TGCAAGAAAC AGAATTCAGG AAGAAGCTCC  1700
AGGAAAGAGT GTGGGGTGAC CCTAGGTGGG GACTCCCACA GCCACAGTGT AGGTGGTTCA GTCCACCCTC CAGCCACTGC TGAGCACCAC TGCCTCCCCG TCCCACCTCA CAAAGAGGGG ACCTAAAGAC  1830
CACCCTGCTT CCACCCATGC CTCTGCTGAT CAGGGTGTGT GTGTGACCGA AACTCACTTC TGTCCACATA AAATCGCTCA CTCTGTGCCT CACATCAAAG GGAGAAAATC TGATTGTTCA GGGGGTCGGA  1960
AGACAGGGTC TGTGTCCTAT TTGTCTAAGG GTCAGAGTCC TTTGGAGCCC CCAGAGTCCT GTGGACGTGG CCCTAGGTAG TAGGGTGAGC TTGGTAACGG GGCTGGCTTC CTGAGACAAG GCTCAGACCC  2090
GCTCTGTCCC TGGGGATCGC TTCAGCCACC AGGACCTGAA AATTGTGCAC GCCTGGGCCC CCTTCCAAGG CATCCAGGGA TGCTTTCCAG TGGAGGCTTT CAGGGCAGGA GACCCTCTGG CCTGCACCCT  2220
CTCTTGCCCT CAGCCTCCAC CTCCTTGACT GGACCCCCAT CTGGACCTCC ATCCCCACCA CCTCTTTCCC CAGTGGCCTC CCTGGCAGAC ACCACAGTGA CTTTCTGCAG GCACATATCT GATCACATCA  2350
AGTCCCCACC GTGCTCCCAC CTCACCCATG GTCTCTCAGC CCCAGCAGCC TTGGCTGGCC TCTCTGATGG AGCAGGCATC AGGCACAGGC CGTGGGTCTC AACGTGGGCT GGGTGGTCCT GGACCAGCAG  2480
CAGCCGCCGC AGCAGCAACC CTGGTACCTG GTTAGGAACG CAGACCCTCT GCCCCCATCC TCCCAACTCT GAAAAACACT GGCTTAGGGA AAGGCGCGAT GCTCAGGGGT CCCCCAAAGC CCGCAGGCAG  2610
AGGGAGTGAT GGGACTGGAA GGAGGCCGAG TGACTTGGTG AGGGATTCGG GTCCCTTGCA TGCAGAGGCT GCTGTGGGAG CGGACAGTCG CGAGAGCAGC ACTGCAGCTG CATGGGGAGA GGGTGTTGCT  2740
CCAGGGACGT GGGATGGAGG CTGGGCGCGG GCGGGTGGCG CTGGAGGGCG GGGGAGGGGC AGGGAGCACC AGCTCCTAGC AGCCAACGAC CATCGGGCGT CGATCCCTGT TTGTCTGGAA GCCCTCCCCT  2870
                                                                                                         38                        45
                                                                                                        ▼Leu Ala Phe Trp Ser Lys His Val ▼
CCCCTGCCCG CTCACCCGCT GCCCTGCCCC ACCCGGGCGC GCCCCTCCGC ACACCGGCTG CAGGAGCCTG ACGCTGCCCG CTCTCTCCGC AG▼CTG GCC TTC TGG TCC AAG CAC GTC G▼GTGAGT  2993
                                                                                       46
                                                                                     ▼Asp Gly Asp Gln Cys Leu Val Leu Pro Leu Glu
GCGTTCTAGA TCCCCGGCTG GACTACCGGC GCCCGCGCCC CTCGGGATCT CTGGCCGCTG ACCCCCTACC CCGCCTTGTG TCGCAG▼ AC GGT GAC CAG TGC TTG GTC TTG CCC TTG GAG  3111
                                                                                                                                       91
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 Trp Glu Gly Arg Phe Cys Gln Arg
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 TGG GAG GGC CGC TTC TGC CAG CGC  3216
                                                                                                                      92                 ◆
                                                                                                                    ▼Glu Val Ser Phe Leu Asn
G▼GTGAGGG GGAGAGGTGG ATGCTGGCGG GCGGCGGGGC GGGGCTGGGG CCGGGTTGGG GGCGCGGCAC CAGCACCAGC TGCCCGCGCC CTCCCCTGCC CGCAG▼ AG GTG AGC TTC CTC AAT  3336
```

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

```
Cys Ser Leu Asp Asn Gly Gly Cys Thr His Tyr Cys Leu Glu Glu Val Gly Trp Arg Arg Cys Ser Cys Ala Pro Gly Tyr Lys Leu Gly Asp Asp Leu Leu Gln
TGC TCT CTG GAC AAC GGC GGC TGC ACG CAT TAC TGC CTA GAG GAG GTG GGC TGG CGG CGC TGT AGC TGT GCG CCT GGC TAC AAG CTG GGG GAC GAC CTC CTG CAG      3440
              136
  Cys His Pro Ala
  TGT CAC CCC GCA G▼GTGAGAAGCC CCCAATACAT CGCCCAGGAA TCACGCTGGG TGCGGGGTGG GCAGGCCCCT GACGGGCGCG GCGCGGGGGG CTCAGGAGGG TTTCTAGGGA GGGAGCGAGG  3564
AACAGAGTTG AGCCTTGGGG CAGCGGCAGA CGCGCCCAAC ACCGGGGCCA CTGTTAGCGC AATCAGCCCG GGAGCTGGGC GCGCCCTCCG CTTTCCCTGC TTCCTTTCTT CCTGGCGTCC CCGCTTCCTC  3694
CGGGCGCCCC TGCGACCTGG GGCCACCTCC TGGAGCGCAA GCCCAGTGGT GGCTCCGCTC CCCAGTCTGA GCGTATCTGG GGCGAGGCGT GCAGCGTCCT CCTCCATGTA GCCTGGCTGC GTTTTTCTCT  3824
GACGTTGTCC GGCGTGCATC GCATTTCCCT CTTTACCCCC TTGCTTCCTT GAGGAGAGAA CAGAATCCCG ATTCTGCCTT CTTCTATATT TTCCTTTTTA TGCATTTTAA TCAAATTTAT ATATGTATGA  3954
AACTTTAAAA ATCAGAGTTT TACAACTCTT ACACTTTCAG CATGCTGTTC CTTGGCATGG GTCCTTTTTT CATTCATTTT CATAAAAGGT GGACCCTTTT AATGTGGAAA TTCCTATCTT CTGCCTCTAG  4084
GGCATTTATC ACTTATTTCT TCTACAATCT CCCCTTTACT TCCTCTATTT TCTCTTTCTG GACCTCCCAT TATTCAGACC TCTTTCCTCT AGTTTTATTG TCTCTTCTAT TTCCCATCTC TTTGACTTTG  4214
TGTTTTCTTT CAGGGAACTT TCTTTTTTTT CTTTTTTTTT GAGATGGAGT TTCACTCTTG TTGTCCCAGG CTGGAGTGCA ATGACGTGAT CTCAGCTCAC CACAACCTCC GCCTCCTGGA TTCAAGCGAT  4344
TCTCCTGCCG CAGCCTCCCG AGTAGCTGGG ATTACAGGCA TGCGCCACCA CGCCCAGCTA ATTTTGTGTT TTTAGTAGAG AAGGGGTTTC TCCGTGTTGG TCAAGCTGGT CTTGAACTCC TGACCTCAGG  4474
TGATCCACCT GCCTTGGCCT CCTAAAGTGC TGGGATTACA GGCGTGAGCC ACCGCGCCCA GCCTCTTTCA GGGAACTTTC TACAACTTTA TAATTCAATT CTTCTGCAGA AAAAAATTTT TGGCCAGGCT  4604
CAGTAGCTCA GACCAATAAT TCCAGCACTT TGAGAGGCTG AGGTGGGAGG ATTGCTTGAG CTTGGGAGTT TGAGACTAGC CTGGGCAACA CAGTGAGACC CTGTCTCTAT TTTTAAAAAA AGTAAAAAAA  4734
GATCTAAAAA TTTAACTTTT TATTTTGAAA TAATTAGATA TTTCCAGGAA GCTGCAAAGA AATGCCTGGT GGGCCTGTTG GCTGTGGGTT TCCTGCAAGG CCGTGGGAAG GCCCTGTCAT TGGCAGAACC  4864
CCAGATCGTG AGGGCTTTCC TTTTAGGCTG CTTTCTAAGA GGACTCCTCC AAGCTCTTGG AGGATGGAAG ACGCTCACCC ATGGTGTTCG GCCCCTCAGA GCAGGGTGGG GCAGGGGAGC TGGTGCCTGT  4994
GCAGGCTGTG GACATTTGCA TGACTCCCTG TGGTCAGCTA AGAGCACCAC TCCTTCCTGA AGCGGGGCCT GAAGTCCCTA GTCAGAGCCT CTGGTTCACC TTCTGCAGGC AGGGAGAGGG GAGTCAAGTC  5124
AGTGAGGAGG GCTTTCGCAG TTTCTCTTAC AAACTCTCAA CATGCCCTCC CACCTGCACT GCCTTCCTGG AAGCCCCACA GCCTCCTATG GTTCCGTGGT CCAGTCCTTC AGCTTCTGGG CGCCCCCATC  5254
ACGGGCTGAG ATTTTTGCTT TCCAGTCTGC CAAGTCAGTT ACTGTGTCCA TCCATCTGCT GTCAGCTTCT GGAATTGTTG CTGTTGTGCC CTTTCCATTC TTTTGTTATG ATGCAGCTCC CCTGCTGACG  5384
ACGTCCCATT GCTCTTTTAA GTCTAGATAT CTGGACTGGG CATTCAAGGC CCATTTTGAG CAGAGTCGGG CTGACCTTTC AGCCCTCAGT TCTCCATGGA GTATGCGCTC TCTTCTTGGC AGGGAGGCCT  5514
CACAAACATG CCATGCCTAT TGTAGCAGCT CTCCAAGAAT GCTCACCTCC TTCTCCCTGT AATTCCTTTC CTCTGTGAGG AGCTCAGCAG CATCCCATTA TGAGACCTTA CTAATCCCAG GGATCACCCC  5644
CAACAGCCCT GGGGTACAAT GAGCTTTTAA GAAGTTTAAC CACCTATGTA AGGAGACACA GGCAGTGGGC GATGCTGCCT GGCCTGACTC TTGCCATTGG GTGGTACTGT TTGTTGACTG ACTGACTGAC  5774
TGACTGGAGG GGGTTTGTAA TTTGTATCTC AGGGATTACC CCCAACAGCC CTGGGGTACA ATGAGCCTTC AAGAAGTTTA ACAACCTATG TAAGGACACA CAGCCAGTGG GTGATGCTGC CTGGTCTGAC  5904
TCTTGCCATT CAGTGGCACT GTTTGTTGAC TGACTCACTG ACTGACTGGC TGACTGGAGG GGGTTCATAG CTAATATTAA TGGAGTGGTC TAAGTATCAT TGGTTCCTTG AACCCTGCAC TGTGGCAAAG  6034
                                                                                     137
                                                                                     Val Lys Phe Pro Cys Gly Arg Pro Trp Lys Arg
TGGCCCACAG GCTGGAGGAG GACCAAGACA GGAGGGCAGT CTCGGGAGGA GTGCCTGGCA GGCCCCTCAC CACCTCTGCC TACCTCAG▼TG AAG TTC CCT TGT GGG AGG CCC TGG AAG CGG        6154

Met Glu Lys Lys Arg Ser 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
ATG GAG AAG AAG CGC AGT 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        6259
      184
  Trp Gln
  TGG CAG▼GTGGGAGGCG AGGCAGCACC GGCTCGTCAC GTGCTGGGTC CGGGATCACT GAGTCCATCC TGGCAGCTAT GCTCAGGGTG CAGAAACCGA GAGGGAAGCG CTGCCATTGC GTTTGGGGGA  6385
TGATGAAGGT GGGGGATGCT TCAGGGAAAG ATGGACGCAA CCTGAGGGGA GAGGAGCAGC CAGGGTGGGT GAGGGGAGGG GCATGGGGGC ATGGAGGGGT CTGCAGGAGG GAGGGTTACA GTTTCTAAAA  6515
AGAGCTGGAA AGACACTGCT CTGCTGGCGG GATTTTAGGC AGAAGCCCTG CTGATGGGAG AGGGCTAGGA GGGAGGGCCG GGCCTGAGTA CCCCTCCAGC CTCCACATGG GAACTGACAC TTACTGGGTT  6645
CCCCTCTCTG CCAGGCATGG GGGAGATAGG AACCAACAAG TGGGAGTATT TGCCCTGGGG ACTCAGACTC TGCAAGGGTC AGGACCCCAA AGACCCGGCA GCCCAGTGGG ACCACAGCCA GGACGGCCCT  6775
TCAAGATAGG GGCTGAGGGA GGCCAAGGGG AACATCCAGG CAGCCTGGGG GCCACAAAGT CTTCCTGGAA GACACAAGGC CTGCCAAGCC TCTAAGGATG AGAGGAGCTC GCTGGGCGAT GTTGGTGTGG  6905
CTGAGGGTGA CTGAAACAGT ATGAACAGTG CAGGAACAGC ATGGGCAAAG GCAGGAAGAC ACCCTGGGAC AGGCTGACAC TGTAAAATGG GCAAAAATAG AAAACGCCAG AAAGGCCTAA GCCTATGCCC  7035
                                                                                                                  185
                                                                                                                  Val Val Leu Leu Asp Ser Lys
ATATGACCAG GGAACCCAGG AAAGTGCATA TGAAACCCAG GTGCCCTGGA CTGGAGGCTG TCAGGAGGCA GCCCTGTGAT GTCATCATCC CACCCCATTC CAG▼GTG GTC CTG CTG GAC TCA AAG   7159
                                                            ○                                           223
Lys Lys Leu Ala Cys 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
AAG AAG CTG GCC TGC 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 G ▼ GTATGGGCTG  7266
GAGCCAGGCA GAAGGGGGCT GCCAGAGGCC TGGGTAGGGG GACCAGGCAG GCTGTTCAGG TTTGGGGGAC CCCGCTCCCC AGGTGCTTAA GCAAGAGGCT TCTTGAGCTC CACAGAAGGT GTTTGGGGGG  7396

AAGAGGCCTA TGTGCCCCCA CCCTGCCCAC CCATGTACAC CCAGTATTTT GCAGTAGGGG GTTCTCTGGT GCCCTCTTCG AATCTGGGCA CAGGTACCTG CACACACATG TTTGTGAGGG GCTACACAGA  7526
CCTTCACCTC TCCACTCCCA CTCATGAGGA GCAGGCTGTG TGGGCCTCAG CACCCTTGGG TGCAGAGACC AGCAAGGCCT GGCCTCAGGG CTGTGCCTCC CACAGACTGA CAGGGATGGA GCTGTACAGA  7656
GGGAGCCCTA GCATCTGCCA AAGCCACAAG CTGCTTCCCT AGCAGGCTGG GGGCTCCTAT GCATTGGCCC CGATCTATGG CAATTTCTGG AGGGGGGGTC TGGCTCAACT CTTTATGCCA AAAAGAAGGC  7786
AAAGCATATT GAGAAAGGCC AAATTCACAT TTCCTACAGC ATAATCTATG CCAGTGGCCC CGTGGGGCTT GGCTTAGAAT TCCCAGGTGC TCTTCCCAGG GAACCATCAG TCTGGACTGA GAGGACCTTC  7916
TCTCTCAGGT GGGACCCGGC CCTGTCCTCC CTGGCAGTGC CGTGTTCTGG GGGTCCTCCT CTCTGGGTCT CACTGCCCCT GGGGTCTCTC CAGCTACCTT TGCTCCATGT TCCTTTGTGG CTCTGGTCTG  8046
TGTCTGGGGT TTCCAGGGGT CTCGGGCTTC CCTGCTGCCC ATTCCTTCTC TGGTCTCACG GCTCCGTGAC TCCTGAAAAC CAACCAGCAT CCTACCCCTT TGGATTGACA CCTGTTGGCC ACTCCTTCTG  8176
GCAGGAAAAG TCACCGTTGA TAGGGTTCCA CGGCATAGAC AGGTGGCTCC GCGCCAGTGC CTGGGACGTG TGGGTGCACA GTCTCCGGGT GAACCTTCTT CAGGCCCTCT CCCAGGCCTG CAGGGGCACA  8306
                                                                                      224
                                                                                      Gly Glu Tyr Asp Leu Arg Arg Trp Glu Lys Trp Glu Leu Asp
GCAGTGGGTG GGCCTCAGGA AAGTGCCACT GGGGAGAGGC TCCCCGCAGC CCACTCTGAC TGTGCCCTCT GCCCTGCAG▼GA GAG TAT GAC CTG CGG CGC TGG GAG AAG TGG GAG CTG GAC   8426
                                        ◆                                   ○
Leu Asp Ile Lys Glu Val Phe Val His Pro Asn Tyr Ser Lys Ser Thr Thr Asp Asn Asp Ile Ala Leu Leu His Leu Ala Gln Pro Ala Thr Leu Ser Gln Thr
CTG GAC ATC AAG GAG GTC TTC GTC CAC CCC AAC TAC AGC AAG AGC ACC ACC GAC AAT GAC ATC GCA CTG CTG CAC CTG GCC CAG CCC GCC ACC CTC TCG CAG ACC      8531

Ile Val Pro Ile Cys Leu Pro Asp Ser Gly Leu Ala Glu Arg Glu Leu Asn Gln Ala Gly Gln Glu Thr Leu Val Thr Gly Trp Gly Tyr His Ser Ser Arg Glu
ATA GTG CCC ATC TGC CTC CCG GAC AGC GGC CTT GCA GAG CGC GAG CTC AAT CAG GCC GGC CAG GAG ACC CTC GTG ACG GGC TGG GGC TAC CAC AGC AGC CGA GAG      8636
                    ◆                                                           ◆
Lys Glu Ala Lys Arg Asn Arg Thr 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
AAG GAG GCC AAG AGA AAC CGC ACC 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      8741
                                                                    ○
Met Leu Cys Ala Gly 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
ATG CTG TGT GCG GGC 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      8846

Val Ser Trp Gly Glu Gly Cys Gly Leu Leu His Asn Tyr Gly Val Tyr Thr Lys Val Ser Arg Tyr Leu Asp Trp Ile His Gly His Ile Arg Asp Lys Glu Ala
GTG AGC TGG GGT GAG GGC TGT GGG CTC CTT CAC AAC TAC GGC GTT TAC ACC AAA GTC AGC CGC TAC CTC GAC TGG ATC CAT GGG CAC ATC AGA GAC AAG GAA GCC      8951
                        419
Pro Gln Lys Ser Trp Ala Pro STOP                                                                          ●
CCC CAG AAG AGC TGG GCA CCT TAG  CGACCCTCCC TGCAGGGCTG GGCTTTTGCA TGGCAATGGA TGGGACATTA AAGGGACATG TAACAAGCAC ACCGGCCTGC TGTTCTGTCC TTCCATCCCT  9075
CTTTTGGGCT CTTCTGGAGG GAAGTAACAT TTACTGAGCA CCTGTTGTAT GTCACATGCC TTATGAATAG AATCTTAACT CCTAGAGCAA CTCTGTGGGG TGGGGAGGAG CAGATCCAAG TTTTGCGGGG  9205
                                                                      ●
TCTAAAGCTG TGTGTGTTGA GGGGGATACT CTGTTTATGA AAAAGAATAA AAAACACAAC CACGAAGCCA CTAGAGCCTT TTCCAGGGCT TTGGGAAGAG CCTGTGCAAG CCGGGGATGC TGAAGGTGAG  9335
GCTTGACCAG CTTTCCAGCT AGCCCAGCTA TGAGGTAGAC ATGTTTAGCT CATATCACAG AGGAGGAAAC TGAGGGGTCT GAAAGGTTTA CATGGTGGAG CCAGGATTCA AATCTAGGTC TGACTCCAAA  9465
ACCCAGGTGC TTTTTTCTGT TCTCCACTGT CCTGGAGGAC AGCTGTTTCG ACGGTGCTCA GTGTGGAGGC CACTATTAGC TCTGTAGGGA AGCAGCCAGA GACCCAGAAA GTGTTGGTTC AGCCCAGAAT  9595
```

FIG. 2. Nucleotide sequence for the gene for human protein C. The first base of the methionine codon where translation is initiated is numbered +1. Arrowheads indicate intron–exon splice junctions. The two *Alu* sequences in intron E have been underlined with a solid line; the 18-base repeats flanking the first *Alu* sequence and the 8-base repeats flanking the second *Alu* sequence have been underscored with dots. The highly conserved sequences of C-C-A-G-C-C-T-G-G have been underlined with a heavy solid line, contrasting with the two homologous 160-bp repeats in intron E which have been lightly underlined. The polyadenylylation or processing sequences of A-T-T-A-A-A and A-A-T-A-A-A at the 3′ end are boxed. The consensus of C-T-T-T-G, which also may be involved in polyadenylylation or cleavage of mRNA at the 3′ end, is underlined with a wavy line. ◆, Potential carbohydrate binding sites to asparagine residues; ↙, apparent cleavage sites for processing of the connecting dipeptide; ↓, site of cleavage in the heavy chain when protein C is converted to activated protein C; ○, active site aspartic acid, histidine, and serine residues; ●, sites of polyadenylylation.

DNA sequence overlapping the two *Eco*RI junctions between the three fragments, two *Bgl* II fragments of 3.3 and 7.0 kb were isolated and subcloned into the *Bam*HI site of pUC9. These two clones span the *Eco*RI sites.

A detailed restriction map as well as approximate placement of the exon regions within the subcloned fragments were established by further restriction analysis and Southern blotting (Fig. 1). When the 5′ and 3′ ends of the gene were established, the nucleotide sequence of the gene was determined by the dideoxy chain-termination method using nuclease *BAL*-31 to provide overlapping sequences between the ends of large restriction fragments.

The nucleotide sequence for the gene for human protein C spans ≈11 kb of DNA (Fig. 2). Comparison of the genomic sequence with that of the cDNA (9) revealed that the gene consists of eight exons ranging in size from 25 to 885 nucleotides and seven introns ranging in size from 92 to 2668 nucleotides. An additional intron(s) in the 5′ noncoding region cannot be ruled out because a cDNA covering this region was not available for comparison with the gene. Also,

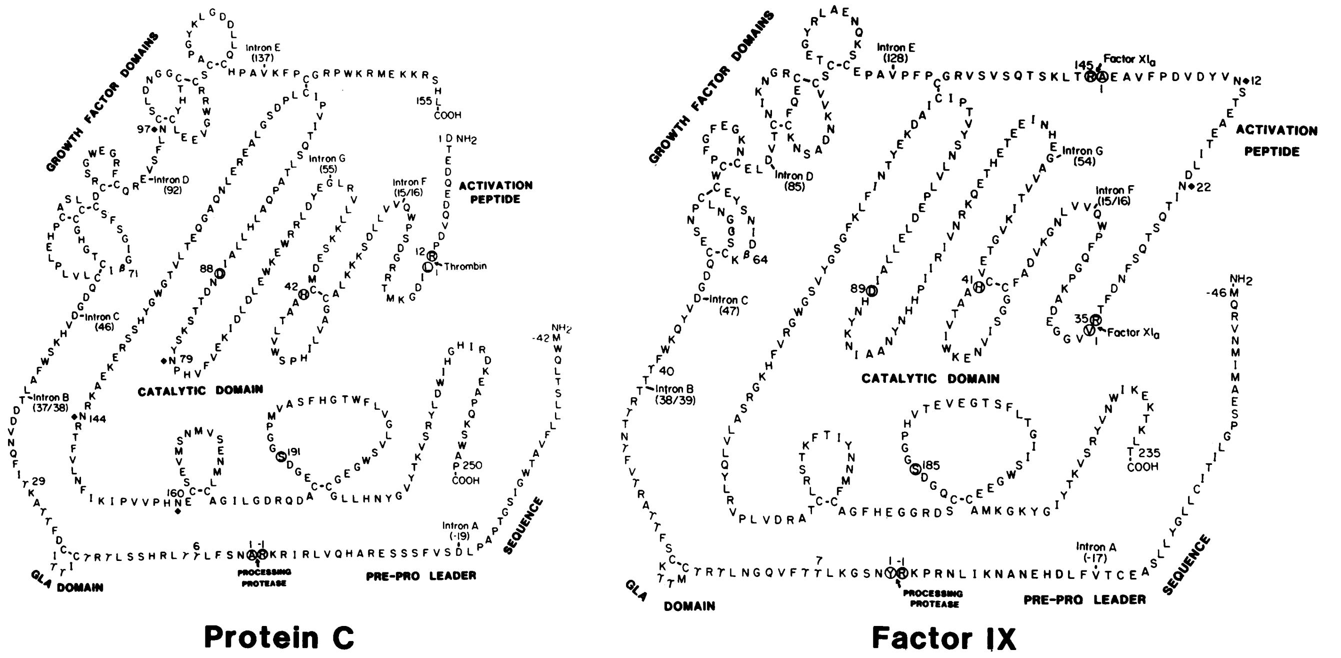

FIG. 3. Amino acid sequence and tentative structures for human prepro-protein C and preprofactor IX. Protein C is shown without the Lys-Arg dipeptide, which connects the light and heavy chains. Locations of the seven introns (A through G) for each gene are indicated by solid bars. Amino acids flanking known proteolytic cleavage sites are circled. The active-site histidine, aspartic acid, and serine residues are also circled. ◆, Potential carbohydrate binding sites. The proposed disulfide bonds have been placed by analogy to those in bovine prothrombin and epidermal growth factor. The first amino acids in the light chain, activation peptide, and heavy chain start with number 1 and differ from that shown in Fig. 2. The factor IX structure was that of Yoshitake *et al.* (12). γ, γ-carboxyglutamic acid; β, β-hydroxyaspartic acid.

several potential intron/exon splice donor and acceptor sequences were identified in the 5′ noncoding region. All the intron/exon splice junctions were similar to the consensus sequences recently summarized by Mount (20) and follow the G-T/A-G rule of Breathnach and Chambon (21).

Several potential "TATA" sequences were found upstream from the preproleader sequence in the gene for human protein C. The sequences of T-A-T-A-A-T-A (starting at position −1785) and T-A-T-A-A-T-T (starting at position −1853) show the strongest homology with the consensus sequence of T-A-T-A-^A_T-A-^A_T. Both, however, lack nearby "CAAT" sequences upstream. If either of these sequences is associated with initiation of transcription, then protein C would have either a very long 5′ noncoding sequence or an additional intron(s) in the 5′ noncoding region of the gene.

Two polyadenylylation or processing sequences of A-T-T-A-A-A and A-A-T-A-A-A (22) were found 47 and 276 nucleotides downstream from the translation stop codon (nucleotides starting at 9022 and 9251). The second of these also has a sequence of C-T-T-T-G starting 37 nucleotides downstream. This latter sequence corresponds to the C-A-^T_C-T-G consensus sequence and also may be involved in polyadenylylation or cleavage at the 3′ end of the mRNA (23). The DNA sequence of eight separate cDNAs at the 3′ end indicates that polyadenylylation occurs with about equal frequency downstream from the two polyadenylylation or processing sites (data not shown).

The gene for protein C contains two *Alu* sequences (24), and both are located in intron E (solid underline in Fig. 2). The first is a complete copy with an orientation of 3′ to 5′. It is flanked by the direct repeat sequence of T-C-T-T-T-C-A-G-G-G-A-A-C-T-T-T-C-T. The second *Alu* sequence is 30 nucleotides after the flanking repeat of the first and is a partial copy of an *Alu* sequence oriented 5′ to 3′. This *Alu* sequence lacks the right half of the *Alu* consensus sequence and is flanked by the direct repeat of A-A-A-A-A-T-T-T. Intron E also contains two direct repeats of about 160 nucleotides of unknown significance (dashed underline in Fig. 2). These repeats are about 93% homologous and start at nucleotides 5628 and 5800. They are separated by 10 nucleotides. A computer comparison of this sequence with the National Institutes of Health sequence data bank revealed no significant homology with published sequences.

The cDNA sequence (9), along with that of the gene, provides the entire amino acid sequence for human preproprotein C (Fig. 3 *Left*). These data indicate that human protein C, like the other vitamin K-dependent coagulation factors, is initially synthesized as a single-chain precursor with a preproleader sequence of 42 amino acids. This leader sequence shows considerable amino acid sequence homology with that recently described for bovine protein C (10). Based on homology with the leader sequence of bovine protein C and other γ-carboxylated coagulation proteases in the region from −1 to −20, it is likely that this leader sequence is cleaved by a signal peptidase after the alanine residue at position −10. This would yield a prozymogen form with a highly basic propeptide of nine residues. Processing to the mature protein that circulates in plasma involves additional proteolytic cleavage after residues at −1, 155, and 157 to remove the amino-terminal propeptide and the Lys-Arg dipeptide that connects the light and heavy chains (9). The processing of the single chain is not complete, however, because about 5–15% of the protein C in human plasma is present as a single-chain molecule (25).

The amino acid composition of the mature protein C circulating in plasma was calculated as follows: $\mathrm{Asp_{28}Asp}$ $(\beta\mathrm{OH})_1$ $\mathrm{Thr_{15}}$ $\mathrm{Ser_{30}}$ $\mathrm{Glu_{24}}$ $\mathrm{Gln_{13}}$ $\mathrm{Gla_9}$ $\mathrm{Pro_{18}}$ $\mathrm{Gly_{33}}$ $\mathrm{Ala_{21}}$ $\mathrm{Val_{26}}$ $\mathrm{Met_7Ile_{16}Leu_{43}Tyr_8Phe_{13}Lys_{22}His_{17}Arg_{23}Trp_{13}Cys_{24}}$, in which Gla is γ-carboxyglutamic acid and Asp(βOH) is β-hydroxyaspartic acid. The molecular weight for the protein was calculated to be 47,456 without carbohydrate and about 61,600 with the addition of 23% carbohydrate (26). Four of the potential carbohydrate chains bound to asparagine occur at residues 97 in the light chain and at residues 79, 144, and 160 in the heavy chain (Fig. 3).

The DNA sequence of the coding region for the gene for human protein C agrees well with that of the cDNA for human protein C (9) except for the triplet coding for Asp-214. Both the genomic sequence (GAT) and the cDNA sequence (GAC) specify aspartic acid at this position. It is likely that the discrepancy is due to either polymorphism or a cloning artifact at nucleotide 7228. The genomic DNA sequence and the sequence of longer cDNA molecules have shown that the amino acid at residue 64 is cysteine rather than glutamine as previously reported (9). This discrepancy is likely to have resulted from an artifactual error introduced into the cDNA sequence adjacent to the *Eco*RI linker used in constructing the λgt11 cDNA library. This phenomenon has been observed in several other cDNAs characterized in this laboratory (unpublished results).

Protein C shows considerable amino acid sequence and structural homology with the other vitamin K-dependent coagulation factors including prothrombin, factor VII, factor IX, and factor X. Factor IX, factor X, and protein C are unusually similar in that they have common domain structures throughout their molecules including a γ-carboxyglutamic acid domain, two potential growth factor domains, an activation peptide or connecting region, and a catalytic domain (27). In prothrombin, the potential growth factor domains have been replaced by two kringle structures. The similarity between these proteins is also evident at the level of the gene where protein C and factor IX show unusual homology. This is illustrated in Fig. 3, which shows the proposed domain structures and the seven introns in the genes for these two proteins. In both genes, the introns occur in essentially the same positions throughout the amino acid sequence of the two proteins. The similarity between these two genes is further reflected in the conservation of splice junction type. All seven introns in the gene for protein C exhibit the same splice junction type as the intron in the corresponding location in the gene for factor IX (12). However, a computer search of the DNA sequences within the introns of the genes in protein C and factor IX showed no significant homology, indicating that the sequences of these regions of the genes are not conserved during evolution.

The locations of the introns in the genes for protein C and factor IX are primarily between various functional domains of the two proteins (Fig. 3). Exon II spans the highly conserved region of the leader sequence and the γ-carboxyglutamic acid domain. Exon III includes a stretch of eight amino acids which connect the γ-carboxyglutamic acid and growth factor domains. Exons IV and V each represent a potential growth factor domain, while exon VI covers a connecting region that includes the activation peptide. Exons VII and VIII cover the catalytic domain typical of all serine proteases.

The first three introns in the gene for human prothrombin (28) also occur in the same position in the amino acid sequence as those of protein C and factor IX. In prothrombin, however, the γ-carboxyglutamic acid region is followed by two kringle structures, which are unrelated in sequence to the potential growth factor domains of protein C and factor IX. After the first three introns, there appears to be no similarity in gene structure between that of prothrombin and those of factor IX and protein C.

The alignment of intron boundaries in the genes for protein C, factor IX, and prothrombin provides additional evidence for the evolution of these genes from a common ancestral precursor. This could have resulted from the joining of numerous fragments of similar DNA sequences by a translocation event(s) between chromosomes during evolution. This could lead to the formation of a gene coding for a serine protease containing additional domains such as the potential growth factor domains, kringle domains, and γ-carboxyglutamic acid domains (12).

The authors thank Drs. Dominic Chung, Steven Leytus, Takehiko Koide, and Kotoku Kurachi for helpful discussions and advice, and Dr. Tom Maniatis for kindly providing the human genomic library constructed in λ Charon 4A bacteriophage. This work was supported in part by a research grant (HL 16919) from the National Institutes of Health. D.C.F. was supported by National Institutes of Health Training Grant GM 07270.

1. Stenflo, J. (1976) *J. Biol. Chem.* **251**, 355–363.
2. Kisiel, W., Ericsson, L. H. & Davie, E. W. (1976) *Biochemistry* **15**, 4893–4900.
3. Esmon, C. T. & Owen, W. G. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 2249–2252.
4. Kisiel, W., Canfield, W. M., Ericsson, L. H. & Davie, E. W. (1977) *Biochemistry* **16**, 5824–5831.
5. Marlar, R. A., Kleiss, A. J. & Griffin, J. (1982) *Blood* **59**, 1067–1072.
6. Vehar, G. A. & Davie, E. W. (1980) *Biochemistry* **19**, 401–410.
7. Griffin, J. H., Evatt, B., Zimmerman, T. S., Kleiss, A. J. & Wideman, C. (1981) *J. Clin. Invest.* **68**, 1370–1373.
8. Griffin, J. H., Mosher, D. F., Zimmerman, T. S. & Kleiss, A. J. (1982) *Blood* **60**, 261–264.
9. Foster, D. & Davie, E. W. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 4766–4770.
10. Long, G. L., Belagaje, R. M. & MacGillivray, R. T. A. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 5653–5656.
11. Anson, D. S., Choo, K. H., Rees, D. J. G., Giannell, F., Gould, J. A., Huddleston, J. A. & Brownlee, G. G. (1984) *EMBO J.* **3**, 1053–1060.
12. Yoshitake, S., Schach, B. G., Foster, D. C., Davie, E. W. & Kurachi, K. (1985) *Biochemistry*, in press.
13. Degen, S. J. F., MacGillivray, R. T. A. & Davie, E. W. (1983) *Biochemistry* **22**, 2087–2097.
14. Maniatis, T., Hardison, R. C., Lacy, E., Lauer, J., O'Connell, C., Quon, D., Sim, G. K. & Efstratiadis, A. (1978) *Cell* **15**, 687–702.
15. Woo, S. L. C. (1979) *Methods Enzymol.* **68**, 381–395.
16. Silhavy, T. J., Berman, W. L. & Enquist, L. W. (1984) *Experiments with Gene Fusions* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY), pp. 140–141.
17. Guo, L. H., Yang, R. C. A. & Wu, R. (1983) *Nucleic Acids Res.* **11**, 5521–5540.
18. Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 3963–3965.
19. Larson, R. & Messing, J. (1982) *Nucleic Acids Res.* **10**, 39–50.
20. Mount, S. M. (1982) *Nucleic Acids Res.* **10**, 459–472.
21. Breathnach, R. & Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349–383.
22. Proudfoot, N. & Brownlee, G. (1981) *Nature (London)* **252**, 359–362.
23. Berget, S. M. (1984) *Nature (London)* **309**, 179–181.
24. Deininger, P. L., Jolly, D. J., Rubin, C. M., Freidmann, T. & Schmid, C. W. (1981) *J. Mol. Biol.* **151**, 17–33.
25. Miletich, J. P., Leykam, F. J. & Broze, G. J. (1983) *Blood* Suppl. 1, **62**, 306a.
26. Kisiel, W. & Davie, E. W. (1981) *Methods Enzymol.* **80**, 320–332.
27. Banyai, L., Varadi, A. & Patthy, L. (1983) *FEBS Lett.* **163**, 37–41.
28. Davie, E. W., Degen, S. J. F., Yoshitake, S. & Kurachi, K. (1983) *Dev. Biochem.* **25**, 45–52.