Articles

Organization of the Gene for Human Factor XI†

Rei Asakai, Earl W. Davie, and Dominic W. Chung*

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

Received April 27, 1987; Revised Manuscript Received June 29, 1987

ABSTRACT: Factor XI (plasma thromboplastin antecedent) is a plasma glycoprotein that participates in the early phase of blood coagulation. The gene for the human protein has been isolated from two different λ phage genomic libraries. Four independent recombinant λ phage carrying overlapping DNA inserts that coded for the entire gene for factor XI were isolated and characterized by restriction mapping, Southern blotting, and selective DNA sequencing to establish the number and location of the intron–exon boundaries. The gene for human factor XI was 23 kilobases in length and consisted of 15 exons (I–XV) and 14 introns (A–N). Exon I coded for the 5′ untranslated region, and exon II coded for the signal peptide. The next eight exons (III–X) coded for the four tandem repeats of 90 or 91 amino acids that were present in the amino-terminal region of the mature protein. Each of these tandem repeats was coded by two exons that were interrupted by a single intron, and these introns were located in essentially the same position within each of the four tandem repeats. The carboxyl-terminal region of the protein, which contained the catalytic chain, was coded by five exons (XI–XV) that were interrupted by four introns. The last four introns were located in the same positions as those in the genes for human tissue plasminogen activator and human urokinase.

Factor XI is a plasma glycoprotein which participates in the contact phase of intrinsic blood coagulation (Davie et al., 1979). It is present in a zymogen form in plasma at a concentration of 4–6 μg/mL (Saito & Goldsmith, 1977). Factor XI has been extensively purified from human and bovine plasma (Bouma & Griffin, 1977; Kurachi & Davie, 1977; Koide et al., 1976; Kurachi et al., 1980). It has a molecular weight estimated between 125 000 and 160 000 and is a homodimer composed of two identical polypeptide chains linked by a disulfide bond(s). It is converted to an active protease, factor XI_a, by factor XII_a (or factor XII) in the presence of high molecular weight kininogen (HMWK) and a polyanionic surface. The activation reaction involves the cleavage of a single internal arginyl–isoleucine bond in each of the two polypeptide chains in factor XI and results in the formation of an active serine protease. Factor XI_a is composed of two heavy and two light chains, and these four chains are held together by disulfide bonds. The heavy chains are derived from the amino termini of the zymogen and responsible for the binding of factor XI to high molecular weight kininogen (van der Graaf et al., 1983) and for the calcium-dependent activation of factor IX (Sinha et al., 1985). The light chain contains the catalytic portion of the enzyme and is homologous to the trypsin family of serine proteases. Factor XI has also been purified from rabbit plasma (Wiggins et al., 1979a,b). This protein differs from human and bovine factor XI in that it is a monomeric protein with an apparent molecular weight of 83 000.

Factor XI circulates in plasma as a noncovalent complex with high molecular weight kininogen (Mandle et al., 1976). Factor XI_a participates in blood coagulation as a catalyst in the conversion of factor IX to factor IX_a in the presence of calcium ions (Fujikawa et al., 1974; DiScipio et al., 1978; Osterud et al., 1978). Individuals with factor XI deficiencies have varying clinical manifestations that range from a complete lack of symptoms to a severe hemorrhagic disorder resembling, but different from, factor IX deficiency. In a number of cases, a marked bleeding tendency is associated with the homozygous state (Saito et al., 1985). We have recently reported the isolation and sequence of a cDNA coding for human factor XI, which enabled us to predict the primary sequence of the protein (Fujikawa et al., 1986). These data indicated that factor XI is synthesized in the liver as a single polypeptide chain of 607 amino acids and a signal peptide of 18 amino acids. Two polypeptide chains are then linked by a disulfide bond(s) to form the zymogen that circulates in plasma. The two heavy chains of factor XI_a are derived from the amino-terminal region of the homodimer, and each chain contains four tandem repeats. Each of these repeats contains 90 or 91 amino acids. These tandem repeats are 58% identical with the four tandem repeats present in human plasma prekallikrein (Chung et al., 1986). In this paper, we report the isolation and characterization of the gene for human factor XI and compare its structural organization to the genes of several other serine proteases.

Experimental Procedures

Screening of Genomic Libraries. Overlapping recombinant λ phage containing DNA inserts coding for human factor XI were isolated from two human genomic libraries (Lawn et al., 1978; Yoshitake et al., 1985). These recombinant phage were identified by the plaque hybridization technique of Benton and Davis (1977) as modified by Woo (1979). Recombinant phage DNA was mapped according to the method of Rackwitz et al. (1984), in which partially digested phage DNA was hybridized separately to radiolabeled oligonucleotides complementary to the left and right cohesive ends of λ phage. DNA fragments containing regions of the factor XI gene were excised from the recombinant phage and subcloned into appro-

†This work was supported by Research Grant HL 16919 from the National Institutes of Health. D.W.C. is an Established Investigator of the American Heart Association.

0006-2960/87/0426-7221$01.50/0

priate restriction sites in the plasmid vectors pUC18 and pUC19 (Pharmacia), as well as pTZ18R and pTZ19R (Amersham).

DNA Sequencing. Recombinant phage and their subclones were digested with restriction enzymes, and the resulting fragments were cloned into M13mp18 and M13mp19. Plaques containing exons and their flanking sequences were identified by hybridization with radiolabeled factor XI cDNA. These clones were then sequenced by the chain terminator method of Sanger et al. (1977), utilizing ^{35}S-labeled deoxyadenosine 5′-triphosphate (dATP) and buffer gradient gels (Biggin et al., 1983). In other experiments, DNA fragments were cloned into pTZ18R and propagated in *Escherichia coli* strain JM103. Transformed *E. coli* isolates carrying recombinant plasmids to be sequenced were superinfected with a helper phage M13K07, and the packaged and released single-stranded forms of pTZ18R were sequenced by standard procedures.

Materials. All restriction enzymes were obtained from New England Biolabs or Bethesda Research Laboratories and used according to the manufacturer's instructions. Bacterial alkaline phosphatase, T4 DNA ligase, nuclease Bal 31, nuclease S1, *E. coli* DNA polymerase I, and the Klenow fragment of DNA polymerase I were purchased from Bethesda Research Laboratories. [^{35}S]dATPαS was obtained from Amersham. One of the human genomic libraries kindly provided by Dr. Tom Maniatis was constructed in λ Charon 4A phage. The second genomic library was constructed from DNA derived from a human fibroblast cell line with 5X chromosomes (Yoshitake et al., 1985).

Results and Discussion

Isolation of the Gene for Human Factor XI. Recombinant phage carrying overlapping fragments of the gene for human factor XI were isolated from two different human genomic libraries by hybridization with a factor XI cDNA. A single phage, designated λC23, was isolated from a partial *Alu*I/*Hae*III human fetal liver library (Lawn et al., 1978) (Figure 1). In restriction digests, Southern transfers, and hybridization experiments employing various fragments of the factor XI cDNA as probes, the DNA insert in this phage was shown to contain only the 5′ portion of the gene. Subsequently, three additional phage (λE10, λE36, and λE29) were isolated from a second genomic library constructed with DNA derived from a human fibroblast cell line (Yoshitake et al., 1985). An analysis of the genomic inserts in these phage indicated that λE10 contained the 3′ half of the gene and λE29 overlapped with λE10 and λC23. However, the 3′ portion of the insert in λE29 contains sequences [~6 kilobases (kb)] that were not collinear with λE10 and did not hybridize to the factor XI cDNA. These sequences apparently resulted from a cloning artifact in which unlinked sequences that originated from unrelated regions of the genome were inadvertently coinserted into the same recombinant phage.

Restriction sites in the four overlapping genomic fragments were determined by the partial digestion and cohesive end-mapping technique of Rackwitz et al. (1984) (Figure 1). The restriction map for 13 different restriction enzymes in these overlapping clones is also shown. *Hin*dIII, *Pst*I, and *Acc*I sites established the overlap among the various isolates, and altogether the DNA inserts in the four different phage covered a distance of 33 kb. Overlapping fragments derived from these four clones were then subcloned into the appropriate sites in plasmid vectors pUC18 or pTZ18R, as shown in the lower portion of Figure 1.

Localization of Intron and Exon Junctions. Intron and exon boundaries in the gene for factor XI were determined by

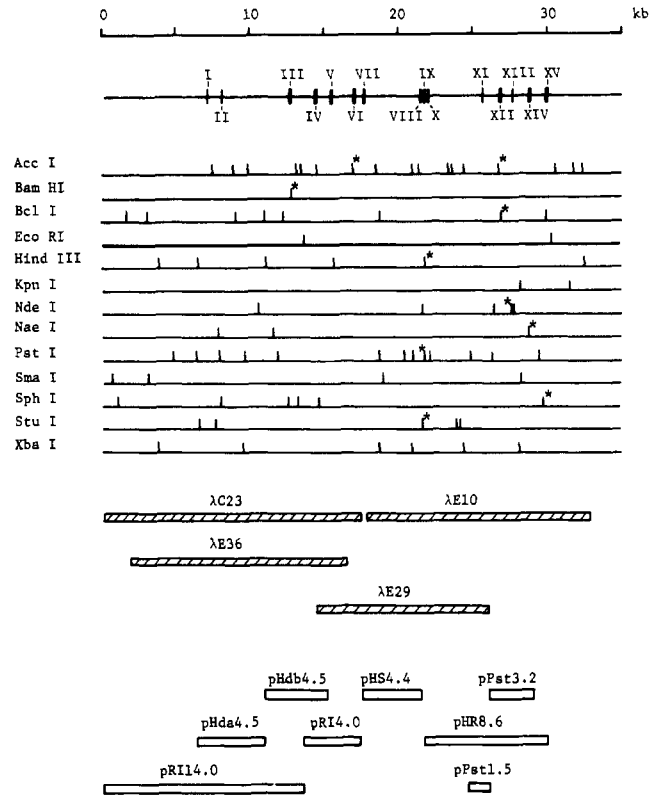

FIGURE 1: Restriction map of recombinant λ phage containing genomic sequences coding for human factor XI. The top line shows the size of the gene in kilobases of DNA, while the second line shows the location of the 15 exons (vertical bars). The λ phage inserts and the plasmid subclones are indicated by hatched and open bars in the lower half of the figure, respectively.

selective DNA sequencing. First, recombinant phage DNA was digested with three restriction enzymes (*Alu*I, *Rsa*I, and *Hae*III) that recognize nucleotide sequences of four bases. These digestions were carried out individually and in various combinations to generate a collection of fragments ranging from 200 to 400 base pairs in length. These fragments were then cloned into the *Sma*I site of M13mp18, and recombinant plaques carrying exon and adjacent intron sequences were identified by hybridization to radiolabeled factor XI cDNA. The M13 recombinant plaques were then sequenced by the dideoxy chain termination method. This approach made it possible to estimate rapidly the number of introns in the gene for factor XI. Second, the precise location of the exons in the gene was determined by selective DNA sequencing of the genomic fragments that were subcloned. Restriction sites with sequences of six bases in the factor XI cDNA were compared with those in the gene for factor XI, and selective sequencing was performed from these unique sites in the appropriate subcloned genomic fragments (identified in Figure 1 with an asterisk). Thus, sequencing from the *Bam*HI site in subclone pHdb4.5 led to the placement of exon III; sequencing from the *Acc*I site in subclone pRI4.0 led to the placement of exon VI; sequencing from the *Pst*I, *Stu*I, and *Hin*dIII sites in subclone pHS4.4 led to the placement of exons VIII, IX, and X, respectively. Exons XII and XIV were placed by sequencing from the *Acc*I and *Nae*I sites in the subclone pPst3.2, and exon XV was placed from the *Sph*I site in subclone pHR8.6. Third, the remaining exons, which do not contain appropriate restriction sites, were mapped by Southern blotting, and sequencing was performed from appropriate restriction sites closest to the respective exons (these sites were not shown in the restriction map of the entire gene). In these experiments, exons IV and V were placed by sequencing from

Table I: Intron and Exon Boundaries in the Gene for Human Factor XI

intron	approx size (kb)	exon	intron	exon	type
A	0.9	TTCAG	GTACAGTTTC.....CTCATTGTAG	GATGA	noncoding
B	4.5	TGGTG	GTAAGTAGAG.....TACATCACAG	AATGT	I
C	1.4	CGATG	GTAAATGCTT.....AAAAAAACAG	GTTTA	II
D	1.2	AAGCG	GTAAGATATG.....TTATTTCCAG	CTTGC	I
E	1.4	CATCG	GTGAGTGAGT.....TTTTATTCAG	TAACA	II
F	0.6	TCTGG	GTAATTATCG.....CGTCGCGCAG	CTTGT	I
G	3.4	CAAAG	GTAAGGAGTT.....TTTTTGTTAG	AAATC	II
H	0.1	CCCAG	GTAAACTGAG.....TGCTGTCTAG	TGTTC	I
I	0.09	GGGAA	GTAAGCCATA.....CCAACTGCAG	GGGCA	II
J	3.7	TAATG	GTGAGTATAA.....TCTGTTGCAG	AGTGT	I
K	1.4	TATGG	GTCAGTACCA.....TGCTCCTTAG	GGTAG	II
L	0.85	CACAG	GTACGGAGAA.....TTAAATTTAG	ATTCT	I
M	1.2	AAGAG	GTAAAAATGA.....TTTTTTTCAG	ACAAA	I
N	0.7	GCAAG	GTAACAGAGT.....CTCGTTCTAG	GGAGA	0
	consensus[a]	---AG	GTUAGT.....YYYYYYNYAG	G----	

[a] Taken from Mount (1982); U represents purines; Y represents pyrimidines.

nearby *Pvu*II and *Hin*dIII sites in subclone pRI4.0; exon VI was placed by sequencing from a *Sal*I site in subclone pHS4.4; exon XIII was placed by sequencing from an *Eco*RV site in pPst3.2. Exon XI was placed by shotgun sequencing of λE10 and by sequencing from a *Ssp*I site in subclone pPst1.5. Exons I and II were shown by restriction mapping to be located in subclone pHda4.5, and the intron–exon junction sequences were determined by specific priming with synthetic oligonucleotides with sequences derived from the 5′ noncoding region of the cDNA. In addition, sequences extending to the 5′ flanking region that contains the probable promoter site and sequences to the 3′ end beyond the polyadenylation site were also determined.

The selective DNA sequencing experiments showed that the gene for factor XI was 23 kb in length and was divided into 15 exons by 14 introns (Figure 2). The DNA sequence of each of the 15 exons was determined and found to be in complete agreement with the previously published cDNA sequences (Fujikawa et al., 1986). These results resolved the differences noted previously between the amino acid sequence predicted from the cDNA and that determined by Edman degradation (Kurachi et al., 1980).

The approximate size of the introns, the sequences at the intron and exon junctions, and the splice junction type for the factor XI gene are shown in Table I. The introns ranged in size from approximately 4.5 kb (intron B) to 0.09 kb (intron I). The sequences at the splice junctions follow the GT-AG rule of Breathnach and Chambon (1981) and are similar to the consensus sequence described by Mount (1982). The 3′ splice acceptor of intron C is a noted exception to the consensus in that the acceptor dinucleotide sequence of AG is preceded by a sequence of AAAAAAAC. Usually, pyrimidine residues occur in this location (Table I).

The organization of the gene shows a strong correlation of the exon distribution with recognizable structural domains within the protein (Figure 3). The first intervening sequence (intron A) occurs in the 5′ noncoding region immediately preceding the initiator methionine codon, while the second intervening sequence (intron B) occurs within the first residue of the mature protein (Glu at position +1). Each of the four tandem repeats at the amino terminus of the protein is separated by an intron that is located in essentially the same position. Each tandem repeat also contains one intron within the repeat that is also located in essentially the same position. As shown in Table I, the splice junction types are strictly conserved among the four repeats. Thus, introns B, D, F, H, and J separate the four repeats from each other and from the signal peptide, and all the splice junctions for these introns are type I. Within each of the four repeats, there is a conserved intron (C, E, G, and I), and each of these introns contains a type II splice junction. These results show a highly conserved organization among the 4 amino acid repeats of 90 or 91 residues in factor XI and provide additional evidence that each repeat was duplicated as a discrete entity.

5′ Promoter and Flanking Sequences. Genomic sequences on the 5′ flanking region of the gene for factor XI have also been determined. A single potential promoter sequence of GTATAT (underlined in Figure 2) that matches the TATA consensus sequences (Breathnach & Chambon, 1981) is identified. No obvious candidate for a CAAT consensus sequence was evident upstream from the apparent promoter sequence. Experiments to define the initiation site for transcription by the primer extension method employing oligonucleotides and poly(A) RNA from human liver (Luse et al., 1981) were inconclusive. This was probably due to the extremely low level of factor XI mRNA in human liver. An examination of the sequences in the first exon showed the presence of three possible mini-cistrons (identified by initiator codons prior to three stop codons) that are 5′ to the major coding sequence (shown in Figure 2 in brackets). According to the scanning model for the initiation of translation, the sequence context of these initiator codons and their ability to support the initiation of translation were almost identical (Kozak, 1986). These short upstream cistrons occur frequently in genes that require carefully regulated expression (Kozak, 1987). Although these mini-cistrons are probably not functional, they may play a regulatory role by modulating the efficiency of translation.

Polyadenylation Site. As previously noted in the cDNA sequence for human factor XI, a sequence of AACAAA may serve as the polyadenylation signal (Fujikawa et al., 1986). This is unusual because a mutation of the polyadenylation signal sequence of AATAAA to AAGAAA in the gene for the adenovirus E1A inhibits the endonucleolytic cleavage of the transcript that precedes polyadenylation (Montell et al., 1983). The nucleotide sequence of the 3′ flanking region of the gene for factor XI indicates that the AACAAA sequences can potentially form a stem–loop structure with downstream sequences (Figure 4) similar to the potential stem–loop structures found at the polyadenylation sites of the histone H2A transcript (Birchmeier et al., 1983), the adenovirus E2A transcript (McDevitt et al., 1984), and the γ chain of human fibrinogen (Rixon et al., 1985). Studies on the correct formation of 3′ ends of sea urchin histone H2A mRNA show that

```
TACGAAAT AAAATTAAAA AAATAAATTC AGTGTATTGA GAAAGCAAGC AATTCTCTCA AGGTATATTT CTGACATACT AAGATTTTAA CGACTTTCAC AAAT

(M)         stop        (M)                                  stop
ATG CTG TAC TGA GAG AGA ATG TTA CAT AAC ATT GAG AAC TAG TAC AAG TAA ATA TTA AAG TGA AGT GAC CAT TTC CTA CAC AAG CTC

                 (M)                                                  stop
ATT CAG AGG AGG ATG AAG ACC ATT TTG GAG GAA GAA AAG CAC CCT TAT TAA GAA TTG CAG CAA GTA AGC CAA CAA GGT CTT TTC AG
|[GTACAGTTTC AGAACTTACT ATTTAACATT CCTCTCAAGC AAATACGCCT TGAAATGCTT TTTTTAAATC ATAGGAATTT AAAAACACTT TACAATAGAG
AATGATTGAT TTTTAAAATG TGTCTGATTT AGCTTTGTAG AGATGTTCCG CTAATATCCA TAACTAATCT GAGAGGAAAT GTGGAACAAC AGAAGAGTAA
CAGTGTCTAC TCAGTAACAA GCGTTTTACG AGTT..... Intron A .....CCTCAAGG AAAGAAAGAA AGGAAAAAAA TTGGGAAAGG AAACAAAGAT
GAAAAATTGG GGTGGGGAGA GCGGTCAGAT GGTGGCCATG AGAAGGATCT GAACACAGAG AGCGGCGGGG CCGGCGGGGA AGGAGGGAGG AGGGGAGAGC
GCTGCTTCCC CGTGGGTTCC GGCTTCTGCA GAGCTGTAAG AGTTGAATGC CACACACAGT CACACTAAGG AATGCTCCAG GATTGGGAAA GATAAAATTC
AACATTATAA TGAGAACACT GTGAATGCTA TTGAATTAAC TACTCCCCTC TCTCCCTATT TCTTGTAAGT CTTAGTGTCA GTAAACTAAT TATAAATTTA
                                                  -18
                                                    M  I  F  L  Y  Q  V  V  H  F  I  L  F  T  S
CATTTTATGT TCTAAAAGCA TGCACCTTTT TCTCATTGTA G]| G ATG ATT TTC TTA TAT CAA GTG GTA CAT TTC ATT TTA TTT ACT TCA
          -1
 V  S  G
GTT TCT GGT G |[GTAAGTAGAG TGTTATCTTA ACTATGGGCT GGGAGAGGGA AATCACACTG CAATCTCCAC ACATGTGGGA GAATCCCACA CCATTTATGC
CGGGAAGGAA ATAAAATGTT TTTATTAACT TCCTGCCTGA GGCTCCAGAG GTTTTCAAAG CAGGGTAGGA ATTGAGGTGA AAAAATTGTT TGTAC.....
........Intron B......TTTCTTACAT ACATATATTA TGCGGCGGTG AAGTGCCGGC AAAAGTGTAC CCATCACCCA AGTAGTGAAC ACAGCCTCCT
TCAGGTAATT TTTCAACGCT CACCCCACTC CCATCCTCCC ATCTAGTGGC ATATTGAAAA TCAATTTGTC TTGTAAAATT AAATAATCCA ATTAGGAGGG
GGATATATTC TAAGGAAATT AGTGCATGAT GCACACACAC ACACACACAC ACAGAACACG TGTGTGCGCA TGTGCACATG AGAGAGAGTG AGAGAGAAAC
TGGGTCTTGC TCTGTCGCCC AGGCTGGATT GCAGTGGTAA AATCACAGCT CACTGCAGCC TCAAACTCCC AGGACTCAGG AGATCCTCCT ACCTCAGCCT
CCCGAGTAGC TGGGATTACA GGTGGAAACA ACCATGCCCA GCTAGTATTT TTTTTTTTTT TTGTATTTTT TATAGAGACA GGGTCTTGCC ATGTTGCCCA
GGCTGGTCTT GAACTCCTCA GCTCAAGCAA TCTACCTACC TTAGCCTCCC AAAGTGCTGG GATTACACGC ATGAGCCACT GCGCCCACTC CGCATTATTA
AATATAGAAC ATTTATTTGA TTCATCAGTT AATATTCTTC TTAAA .....CCTTTAT GAGATTACCA CCTAACTAGA TGTATGCCCA GTAAAATCCA
                                  1
                                   E  C  V  T  Q  L  L  K  D  T  C  F  E  G  G  D  I  T  T  V
ACATAACGCA TGCCATGTAC TACATCACAG]| AA TGT GTG ACT CAG TTG TTG AAG GAC ACC TGC TTT GAA GGA GGG GAC ATT ACT ACG GTC

 F  T  P  S  A  K  Y  C  Q  V  V  C  T  Y  H  P  R  C  L  L  F  T  F  T  A  E  S  P
TTC ACA CCA AGC GCC AAG TAC TGC CAG GTA GTC TGC ACT TAC CAC CCA AGA TGT TTA CTC TTC ACT TTC ACG GCG GAA TCA CCA
                     55
 S  E  D  P  T  R  W
TCT GAG GAT CCC ACC CGA TG |[GTAAATGCTT ATGTTTCTAC ATCGAGGAGA CAGATTTTTA AAGG.....Intron C.....GGCATGAG ATAAAGTAGT
TTGTTTCCTT CTTTTTGGCT TTCTGTGTGC TGACTTTTAA GATCCATTAT TTTAAAAACA TAAATTCCTA TTCATTAATA TGTATTTTTT AAAAAAACAG]|
 56
   F  T  C  V  L  K  D  S  V  T  E  T  L  P  R  V  N  R  T  A  A  I  S  G  Y  S  F
G TTT ACT TGT GTC CTG AAA GAC AGT GTT ACA GAA ACA CTG CCA AGA GTG AAT AGG ACA GCA GCG ATT TCT GGG TAT TCT TTC
                                    90
 K  Q  C  S  H  Q  I  S
AAG CAA TGC TCA CAC CAA ATA AGC G  |[GTAAGATATG TTCTCAGAAT CAACAAATA CCAGCTG.....Intron D......GCCCCTA GAATCTGGAA
                                   91
                                     A  C  N  K  D  I  Y  V  D  L  D  M  K  G  I  N  Y  N  S
GGTACTCATG TCTTCTGCTT TTATTTCCAG]| CT TGC AAC AAA GAC ATT TAT GTG GAC CTA GAC ATG AAG GGC ATA AAC TAT AAC AGC

 S  V  A  K  S  A  Q  E  C  Q  E  R  C  T  D  D  V  H  C  H  F  F  T  Y  A  T  R
TCA GTT GCC AAG AGT GCT CAA GAA TGC CAA GAA AGA TGC ACG GAT GAC GTC CAC TGC CAC TTT TTC ACG TAC GCC ACA AGG
                                   144
 Q  F  P  S  L  E  H  R
CAG TTT CCC AGC CTG GAG CAT CG |[GTGAGTGAGT CCCAGGACAT TCGAGTGGTC GATGAAAAAC AGAATCGTGA TTTACTAAAA AGCTTTTGCC
ATCAACTTTA TGCCAGAATT TATTTTGAAC CCCTAAAAGA CATTTCTATA AAAGTACTCC TAGTTTTCTT CATGAAAAAT ACACTTAAAG CCTAATTTGG
ATGCATTTCA TTTATGGTAA GGAGTCTATC TTTTAATAAC ACTGTCAGAA AAATATATAT ACTTGGCTAA TTTCAAAAGC GCTACACTTT TAAATTGGCA
CTTTTGAAAC AGCTGCAATT GGTATGATTG TCAGTGCC.. ...Intron E......TGCT TAGCAACACT GCTGGGACCA TGCCCAGCCA TTCAGCCTCC
CAGATGGATG CTTCGGGGTC TCGCAGGTCC TCTCTCCAAA GGGGACTTTC TTAATATCTC ATGTTTTTTC CTCCTTGCAG TTGGAAGAAT AAGACACTTT
                         145
                           N  I  C  L  L  K  H  T  Q  T  G  T  P  T  R  I  T  K  L  D  K  V
TCCTTTTTCT TTTTATTCAG]| T AAC ATT TGT CTA CTG AAG CAC ACC CAA ACA GGG ACA CCA ACC AGA ATA ACG AAG CTC GAT AAA GTG
                                                  180
 V  S  G  F  S  L  K  S  C  A  L  S  N  L
GTG TCT GGA TTT TCA CTG AAA TCC TGT GCA CTT TCT AAT CTG G |[GTAATTATCG ACTTCTTGAT GATGTAATTC AACCATTAAA TATGCTGATG
ATTACAGTAG ATCTCACTCA GGATACCAGC TTATGCTCAC GATGAAACGG ACCCAAAGAT CTTTACCTTC TTCATGTGAT AGATTTCATC ATGTCCTATA
CAGTTAGATC CTCTATTTAA.....Intron F.....GATCTTGGG ATACACTTAA ATTTTTTAAT ATGGAATTTA CACATATGTG ACCGGAATTT TCCTGATAGC
                                      181
                                        A  C  I  R  D  I  F  P  N  T  V  F  A  D  S  N  I
TGGTGAATTG AGTCCCTGAC ATAGTTCTTC CGTCGCGCAG]| CT TGT ATT AGG GAC ATT TTC CCT AAT ACG GTG TTT GCA GAC AGC AAC ATC

 D  S  V  M  A  P  D  A  F  V  C  G  R  I  C  T  H  H  P  G  C  L  F  F  T  F  F  S
GAC AGT GTC ATG GCT CCC GAT GCT TTT GTC TCT GGC CGA ATC TGC ACT CAT CAT CCC GGT TGC TTG TTT TTT ACC TTC TTT TCC
                                   234
 Q  E  W  P  K  E  S  Q  R
CAG GAA TGG CCC AAA GAA TCT CAA AG |[GTAAGGAGTT AACAAGTAAG GATAATTTGT TATCTTCTAA AAA.....Intron G.....CTGA
                                                                    235
                                                                      N  L  C  L  L
CTTTACTTTC TCTAGGTGCT GTAAAAATGT TTTTATGTGT TTGATATGAT ATATTTCTAC TTCCCTTTTG TTTTTGTTAG]| A AAT CTT TGT CTC CTT

 K  T  S  E  S  G  L  P  S  T  R  I  K  K  S  K  A  L  S  G  F  S  L  Q  S  C  R  H
AAA ACA TCT GAG AGT GGA TTG CCC AGT ACA CGC ATT AAA AAG AGC AAA GCT CTT TCT GGT TTC AGT CTA CAA AGC TGC AGG CAC
```

```
          270
 S   I   P
AGC ATC CCA G ][GTAAACTGAG AGTTCTGCAT TCTGGCTGAG AGTGACCAGC CCCGAGGAGG CTGATACATG CTGAGGGAGG GTCTCACTCT GACATGTGGT
           271
             V   F   C   H   S   S   F   Y   H   D   T   D   F   L   G   E   E   L   D   I   V   A   A   K   S
CTGCTGTCTA G]| TG TTC TGC CAT TCT TCA TTT TAC CAT GAC ACT GAT TTC TTG GGA GAA GAA CTG GAT ATT GTT GCT GCA AAA AGT

 H   E   A   C   Q   K   L   C   T   N   A   V   R   C   Q   F   F   T   Y   T   P   A   Q   A   S   C
CAC GAG GCC TGC CAG AAA CTG TGC ACC AAT GCC GTC CGC TGC CAG TTT TTT ACC TAT ACC CCA GCC CAA GCA TCC TGC
               325
 N   E   G   K
AAC GAA GGG AA ][GTAAGCCATA TGAAGGGTTA TGCAGACACC CTTGTCCCGT CTGCCTGTGA GGTGCATTAT GTTTATACCG TTTTGTTTCC
           326
             G   K   C   Y   L   K   L   S   S   N   G   S   P   T   K   I   L   H   G   R   G   G
AACTGCAG]| G GGC AAG TGT TAC TTA AAG CTT TCT TCA AAC GGA TCT CCA ACT AAA ATA CTT CAC GGG AGA GGA GGC
                                                   360
 I   S   G   Y   T   L   R   L   C   K   M   D   N
ATC TCT GGA TAC ACA TTA AGG TTG TGT AAA ATG GAT AAT G ][GTGAGTATAA TGTCACTTGA AAAAATATAG CTGAAGGAAT TATTCCATGC

TTCATACATC ACAATCAAGA CTGTCAGTTA TAGCCACAGA AGGGAGAACA TTCAGGAAAT AACAAATTTT .......Intron J.....AATGCT
           361
             E   C   T   T   K   I   K   P   R   I   V   G   G   T   A   S   V   R   G   E   W   P   W   Q   V
TCTGTTGCAG]| AG TGT ACC ACC AAA ATC AAG CCC AGG ATC GTT GGA GGA ACT GCG TCT GTT CGT GGT GAG TGG CCG TGG CAG GTG

 T   L   H   T   T   S   P   T   Q   R   H   L   C   G   G   S   I   I   G   N   Q   W   I   L   T   A   A   H
ACC CTG CAC ACA ACC TCA CCC ACT CAG AGA CAC CTG TGT GGA GGC TCC ATC ATT GGA AAC CAG TGG ATA TTA ACA GCC GCT CAC
               417
 C   F   Y   G
TGT TTC TAT GG ][GTCAGTACCA CGGCTGTTTT TATTAGTTCA TCTTCTTCAC ACATTTATAA AAAATATTAC TAGCATGTTA GGAAATAAAT ACTTTA...

......Intron K.....TG CTCATTCATT TTTTGTGTAT AATGGATTTT CTTTATAGGG TGAATATGTT TTTTATCCCG AAAAATCTTA GGATAAAATC

ACTTTTTTCT ACCTAAATGT CCATCATTGG CAGAAAATAT TAGTAATAAT TAAACAGCCA CACACTTCAC AATGTCTGGG AATTATTTTT AGTAAAGGAA
                                     418
                                       V   E   S   P   K   I   L   R   V   Y   S   G   I   L   N   Q   S   E   I
ATTTCTTTCC CTCTGTTGTT TGCTCCTTAG]| G GTA GAG TCA CCT AAG ATT TTG CGT GTC TAC AGT GGC ATT TTA AAT CAA TCT GAA ATA

 K   E   D   T   S   F   F   G   V   Q   E   I   I   I   H   D   Q   Y   K   M   A   E   S   G   Y   D   I   A
AAA GAG GAC ACA TCT TTC TTT GGG GTT CAA GAA ATA ATA ATC CAT GAT CAG TAT AAA ATG GCA GAA AGC GGG TAT GAT ATT GCC
                                         475
 L   L   K   L   E   T   T   V   N   Y   T
TTG TTG AAA CTG GAA ACC ACA GTG AAT TAC ACA G ][GTACGGAGAA TTTTATCCGG AAAGTTGTCT CCAATGGTGA ACTGGATAAA ATGTTTAACA

CTACTAGACT TACGGCCTGA CCCTGCCAAT CTCTCCATGC GTTATCATCA TGAAAGGGAG AGGGCCTGGA ATGCTAGTCA TTCACTCTGC TAAGGCTGAC

ACACTTTCCT GGCTATTGAA.....Intron L.....ATCGTGCTGA ACCTGAGGGA GGAAAATACA CGACAACAAG GCAAAAAATG AATATAGTAA

ACAAAGAAAA CACAGATAAT GTACAGTGGA AGAAGAGTCT CTTCTGGAAA AGAGGATATA TTTTGCGTCT CATATTTAAA CCACGATTTT TTAAATTTAG]|

 D   S   Q   R   P   I   C   L   P   S   K   G   D   R   N   V   I   Y   T   D   C   W   V   T   G   W   G   Y
AT TCT CAA CGA CCC ATA TGC CTG CCT TCC AAA GGA GAT AGA AAT GTA ATA TAC ACT GAT TGC TGG GTG ACT GGA TGG GGG TAC
               507
 R   K   L   R
AGA AAA CTA AGA G ][GTAAAAATGA TGTTGTTATA TGTGCTCCAT CCTAGAAATG AAGAGCGGAA CCTTTT.....Intron M.....CTGAG

ATTGCACCAC TGCACTCCAG CCTGGGCGAC AGAAAGAGAC TCCGTCTCAA TTAAAAATAT ATATATATAT ATATATTTAT ATGTATGCAT ATATGTTTAT
                                                   508
                                                     D   K   I   Q   N   T   L   Q   K   A   K   I   P   L
GTGTATTGTG TATGGTTATT CTACAAACGA ACCAAAAAAA TTTTTTTCAG]| AC AAA ATA CAA AAT ACT CTC CAG AAA GCC AAG ATA CCC TTA

 V   T   N   E   E   C   Q   K   R   Y   R   G   H   K   I   T   H   K   M   I   C   A   G   Y   R   E   G   G
GTG ACC AAC GAA GAG TGC CAG AAG AGA TAC AGA GGA CAT AAA ATA ACC CAT AAG ATG ATC TGT GCC GGC TAC AGG GAA GGA GGG
                 554
 K   D   A   C   K
AAG GAC GCT TGC AAG ][GTAACAGAGT GTTCTTAGCC AATGGAATAT ATGCAAATTG GAATGCTTAA TGCGTTGGGG TTTTTTTGTT TGTTTTGTTT

TTTTTGTTTG TTTTTTTTTG AGACAGAGTC TCGCTCTGTT GCCCAGGCTG GAGTGCAGTG GCTCGATCT.....Intron N.....C AAGACAACAT
                                                                                                   555
                                                                                                     G   D
TTTAGGCAAA ATCAGCCTGA GCAAGATGTG CTGAAGATGG GAAGCGTCTG AGTTGATCTG TGCACCTTTT CTTGTCTCCC CTCGTTCTAG]| GGA GAT

 S   G   G   P   L   S   C   K   H   N   E   V   W   H   L   V   G   I   T   S   W   G   E   G   C   A   Q   R
TCG GGA GGC CCT CTG TCC TGC AAA CAC AAT GAG GTC TGG CAT CTG GTA GGC ATC ACG AGC TGG GGC GAA GGC TGT GCT CAA AGG

 E   R   P   G   V   Y   T   N   V   V   E   Y   V   D   W   I   L   E   K   T   Q   A   V  stop
GAG CGG CCA GGT GTT TAC ACC AAC GTG GTC GAG TAC GTG GAC TGG ATT CTG GAG AAA ACT CAA GCA GTG TGA ATG GGT TCC CAG

GGG CCA TTG GAG TCC CTG AAG GAC CCA GGA TTT GCT GGG AGA GGG TGT TGA GTT CAC TGT GCC AGC ATG CTT CCT CCA CAG TAA

CAC GCT GAA GGG GCT TGG TGT TTG TAA GAA AAT GCT AGA AGA AAA CAA ACT GTC ACA AGT TGT TAT GTC CAA AA ▮ CTCCCGTTCT

ATGATCGTTG TAGTTTGTTT GAGCATTCAG TCTCTTTGTT TTTGATCACG CTTCTATGGA GTCCAAGAAT TACCATAAGG CAATGTTTCT GAAGATTACT

ATATAGGCAG ATATACCAGA AAATAACCAA GTAGTGGCAG TGGGGATCAG GCAGAAGAAC TGGTAAAAGA AGCCACCATA AATAGATTTG TTCGATGAAA

GATGAAAACT GGAAGAAAGG AGAACAAAGA CAGTCTTCAC CATTTTGCAG GAATCTACAC TCTGCCTATG TGAACACATT TCTTTTGTAA AGAAAGAATT

TGATT
```

FIGURE 2: Nucleotide sequence of the gene for human factor XI, including the flanking regions, exons, and intron–exon boundaries. Exon sequences are shown in nucleotide triplets, and the predicted amino acids in single-letter code are shown above each triplet. Potential promoter sequences in the 5′ flanking region are underlined. Potential Z DNA sequences (intron B and intron M) are also underlined. The polyadenylation signal sequence AACAAA is double underlined, and the poly(A) addition site is indicated by an arrow on the 3′ end of the gene.

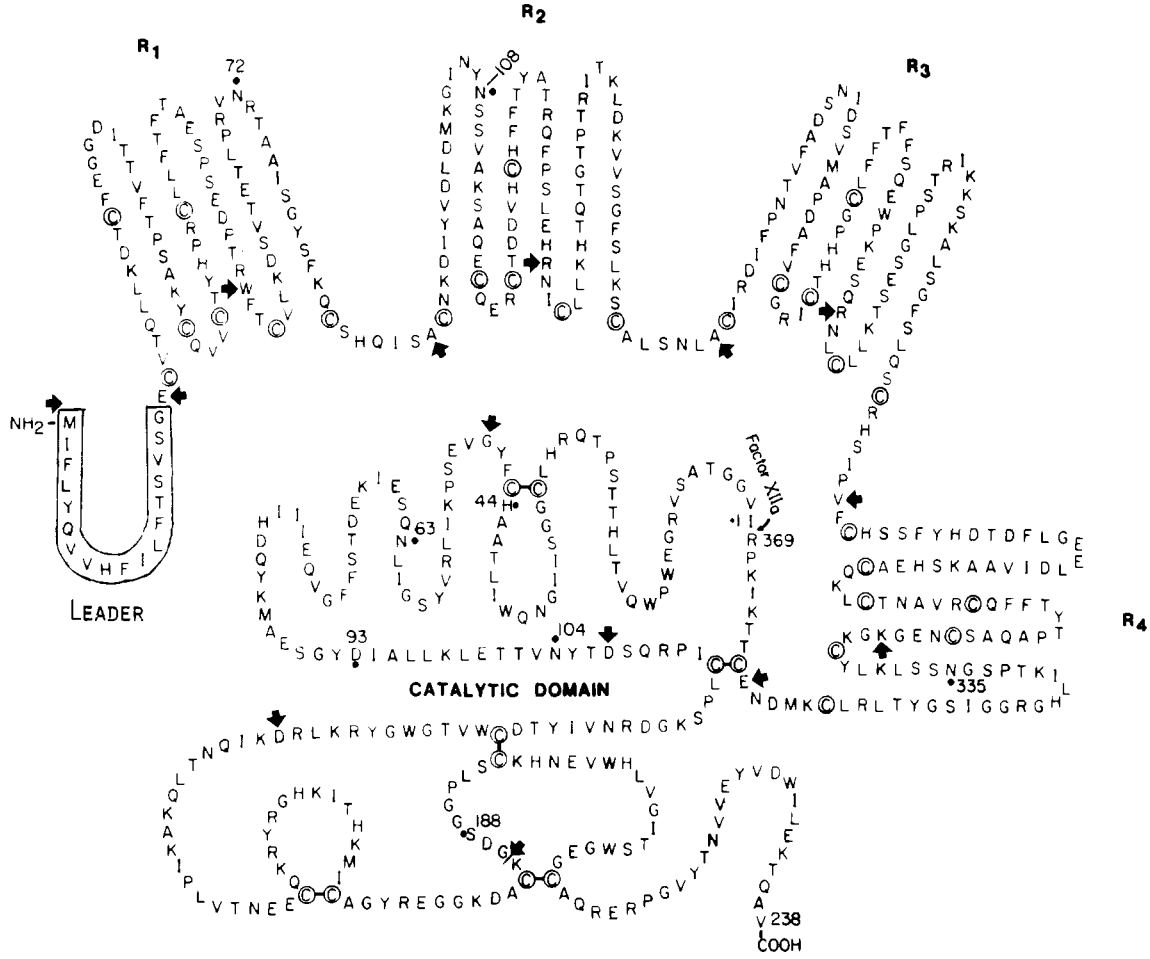

FIGURE 3: Location of the 14 introns in human factor XI. Solid arrows indicate the location of each intron. The leader sequence is shown in an open box. Cys residues are circled. Solid dots identify the three amino acids involved in catalysis. Potential N-linked carbohydrate binding sites are shown by solid diamonds.

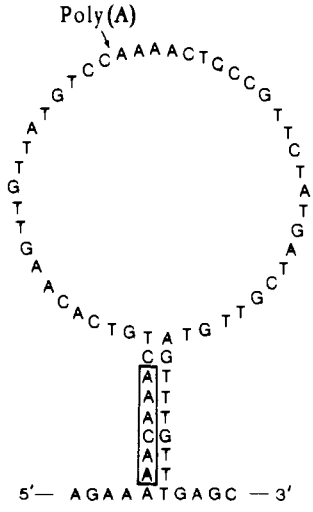

FIGURE 4: Potential secondary structure around the polyadenylation site of the gene for human factor XI. The poly(A) addition site is indicated by an arrow, and the polyadenylation signal sequence of AACAAA is boxed.

a single base point mutation in a potential stem–loop region prevents the correct formation of the 3′ termini. However, a second downstream complementary point mutation, which restores the stem–loop structure, restores the correct formation of the 3′ termini. These studies suggest that secondary structure is important in the recognition and correct processing in the formation of 3′ termini. The formation of a stem–loop structure of AACAAA with a reverse complementary sequence downstream may serve as an effective polyadenylation signal for the correct processing and maturation of the factor XI mRNA.

Z DNA Sequences. An analysis of the sequence of the gene for factor XI shows that there are at least two regions consisting of alternating purines and pyrimidines that favor the formation of Z DNA (Wang et al., 1979). The first region occurs in intron B, approximately 3.5 kb from the exon II/intron B junction (Figure 2). This region is about 50 nucleotides in length and consists of a segment with 11 d(CA) dinucleotide repeats flanked on both sides by alternating purines and pyrimidines of other combinations. Although this region contains two bases that are out of the purine–pyrimidine alternation, this appears to be a common occurrence in natural Z DNA sequences and apparently does not interfere with Z DNA formation (Nordheim et al., 1982; Azorin et al., 1983). The repeat sequence $d(CA/GT)_n$ is widely distributed in eukaryotic genomes. It is present in approximately 50 000 copies per human genome and is the most prevalent form of Z DNA sequences (Hamada et al., 1982). This type of purine–pyrimidine alternation has been found in the genes for globin, immunoglobulins, actins, and factor IX (Miesfeld et al., 1981; Hamada et al., 1982; Hamada & Kakunaga, 1982; Yoshitake et al., 1985). A second region is about 60 nucleotides in length and occurs in intron M. This region consists of 10 consecutive d(AT) alternating repeats followed by purine–pyrimidine alternations of other combinations. The precise function of Z DNA is not clear, but it has been proposed that Z DNA may be involved in sequences with enhancer properties and may play a role in the regulation of transcription (Cereghini

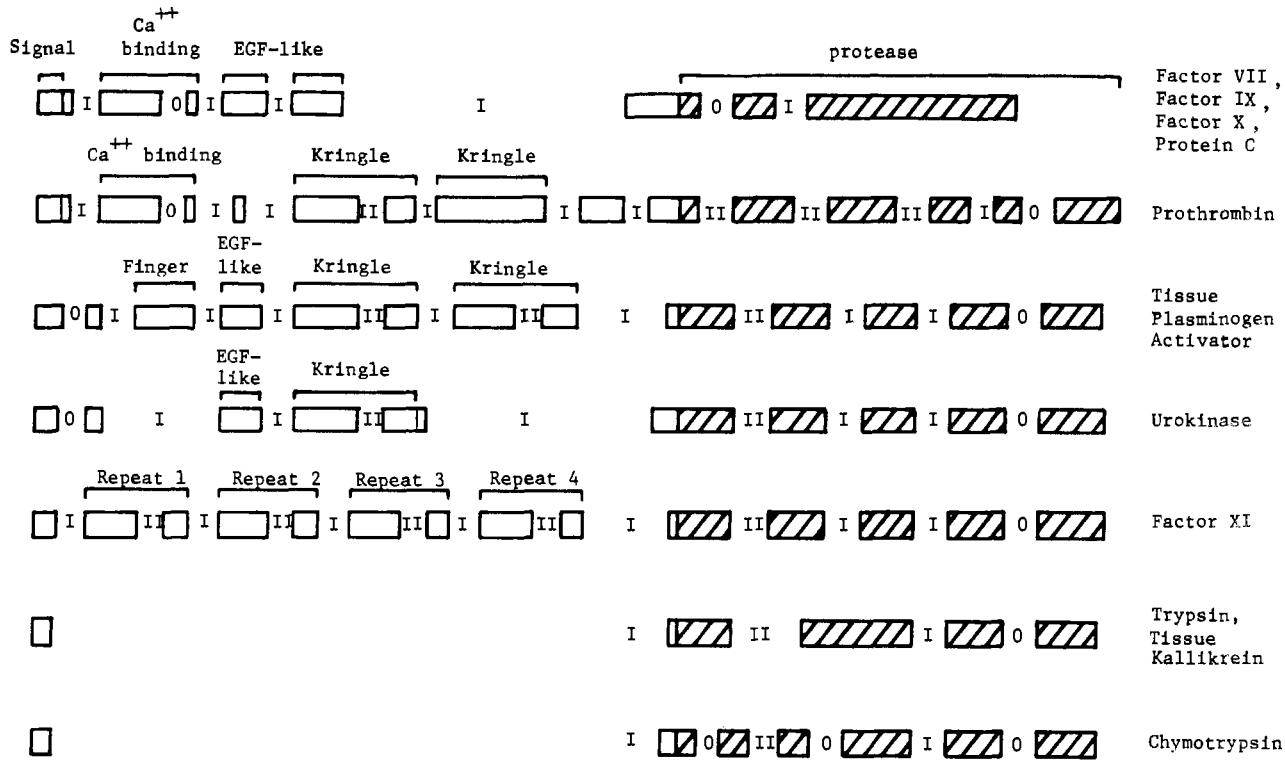

FIGURE 5: Comparison of the structural organization of the genes for several serine proteases. Exons in the trypsin-like catalytic portion of the molecule are represented by hatched bars. Exons coding for other structural domains at the amino termini of the molecules are represented by open bars. The introns are not shown. The splice junction types (0, I, or II) relative to the reading frame are indicated between the exons [modified from Rogers (1985)].

et al., 1983; Nordheim & Rich, 1983). Z DNA sequences have also been implicated to promote recombinational gene conversion events that lead to gene expansion in miltimember gene families (Slightom et al., 1980; Kmiec & Holloman, 1984). Additional studies are necessary to determine whether these sequences in the gene for factor XI possess enhancer properties or play a role in the regulation of the expression of the factor XI gene. In addition to Z DNA sequences, the gene for factor XI also contains at least one copy of *Alu* repeat sequence, which is located in intron N (data not shown). Alu repeat sequences are widely distributed in the human genome and are known to be present within introns of many genes.

Evolutionary Relatedness to Other Serine Proteases. A comparison of the protein sequence of factor XI with other serine proteases indicates that the catalytic light chain of the molecule is highly homologous both to factor IX and to tissue plasminogen activator (t-PA). Our results make it possible to extend the comparison to the level of gene organization and to genes of the pancreatic serine proteases (Craik et al., 1984; Bell et al., 1984; Mason et al., 1983). As shown in Figure 5, the gene for factor XI bears a closer resemblance to those of t-PA and urokinase than to factors VII, IX, and X and protein C. Thus, the catalytic chain of the factor XI molecule is coded by five exons. The locations and the splice junction types of the introns in this region of the gene for factor XI are identical with those in the genes for human t-PA (Ny et al., 1984; Degen et al., 1986) and human urokinase (u-PA) (Riccio et al., 1985). This subfamily of serine protease is charcteristically interrupted by an intron immediately preceding the conserved Gly-Asp-Ser-Gly-Gly-Pro sequence that contains the active-site serine (next to the last exon). In contrast, the genes for the vitamin K dependent subfamily of serine proteases, including prothrombin (Degen & Davie, 1987), factor VII (O'Hara et al., 1987), factor IX (Anson et al., 1984; Yoshitake et al., 1985), factor X (Leytus et al., 1986), and protein C (Foster et al., 1985; Plutzky et al., 1986), lack this intron. These results suggest that factor XI is derived from an ancestor which is more closely related to the t-PA subfamily than to the vitamin K dependent serine protease subfamily. The fact that there is no sequence similarity in the amino-terminal region of factor XI with any other serine protease supports that exon shuffling plays a significant role in the evolution of the gene for this protein.

ACKNOWLEDGMENTS

We express our gratitude to Dr. Kazuo Fujikawa for kindly providing the cDNA for human factor XI and for very stimulating discussions. We also thank Drs. Patrick Chou and Yim Foon Lee for providing the oligonucleotides and Drs. Donald Foster, Steven Leytus, Barbara Schach, and Akitada Ichinose for valuable help and advice. Thanks are also due to Jeff Harris and Keith Loeb for excellent technical assistance.

REFERENCES

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.

Azorin, F., Nordheim, A., & Rich, A. (1983) *EMBO J. 2*, 649–655.

Bell, G. I., Quinto, C., Quiroga, M., Valenzuela, P., Craik, C. S., & Rutter, W. J. (1984) *J. Biol. Chem. 259*, 14265–14270.

Benton, W. D., & Davie, R. W. (1977) *Science (Washington, D.C.) 196*, 188–192.

Biggin, M. D., Gibson, T. T., & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. U.S.A. 80*, 3963–3965.

Birchmeier, C., Folk, W., & Birnstiel, M. L. (1983) *Cell (Cambridge, Mass.) 35*, 433–440.

Bouma, B. N., & Griffin, J. H. (1977) *J. Biol. Chem. 252*, 6432–6437.

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

Cereghini, S., Herbomel, P., Jounneau, J., Saragosti, S., & Katinka, M. (1983) *Cold Spring Harbor Symp. Quant. Biol. 47*, 935–944.
Chung, D. W., Fujikawa, K., McMullen, B. A., & Davie, E. W. (1986) *Biochemistry 25*, 2410–2417.
Craik, C. S., Choo, Q. L., Swift, G. H., Quinto, C., & MacDonald, R. J. (1984) *J. Biol. Chem. 259*, 14255–14264.
Davie, E. W., Fujikawa, F., Kurachi, K., & Kisiel, W. (1979) *Adv. Enzymol. Relat. Areas Mol. Biol. 48*, 277–318.
Degen, S. J. F., & Davie, E. W. (1987) *Biochemistry 26*, 6165–6177.
Degen, S. J. F., Rajput, B., & Reich, E. (1986) *J. Biol. Chem. 261*, 6972–6985.
DiScipio, R. G., Kurachi, K., & Davie, E. W. (1978) *J. Clin. Invest. 61*, 1528–1538.
Foster, D. C., Yoshitake, S., & Davie, E. W. (1985) *Proc. Natl. Acad. Sci. U.S.A. 82*, 5673–5677.
Fujikawa, K., Legaz, M. E., Kato, H., & Davie, E. W. (1974) *Biochemistry 13*, 4508–4516.
Fujikawa, K., Chung, D. W., Hendrickson, L. E., & Davie, E. W. (1986) *Biochemistry 25*, 2417–2424.
Hamada, H., & Kakunaga, T. (1982) *Nature (London) 298*, 396–398.
Hamada, H., Petrino, M. G., & Kakunaga, T. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 6465–6469.
Kmiec, E. B., & Holloman, W. K. (1984) *Cell (Cambridge, Mass.) 36*, 593–598.
Koide, T., Kato, H., & Davie, E. W. (1976) *Methods Enzymol. 45*, 65–73.
Kozak, M. (1986) *Cell (Cambridge, Mass.) 44*, 283–292.
Kozak, M. (1987) *Cell (Cambridge, Mass.) 47*, 481–483.
Kurachi, K., & Davie, E. W. (1977) *Biochemistry 16*, 5831–5839.
Kurachi, K., Fujikawa, K., & Davie, E. W. (1980) *Biochemistry 19*, 1330–1338.
Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G., & Maniatis, T. (1978) *Cell (Cambridge, Mass.) 15*, 1157–1174.
Leytus, S. P., Foster, D. C., Kurachi, K., & Davie, E. D. (1986) *Biochemistry 25*, 5098–5102.
Luse, D. S., Haynes, J. R., Van Leeuwen, D., Schon, E. A., Cleary, M. L., Shapiro, S. G., Lingrel, J. B., & Roeder, R. G. (1981) *Nucleic Acids Res. 9*, 4339–4354.
Mandle, R. J., Colman, R. W., & Kaplan, A. P. (1976) *Proc. Natl. Acad. Sci. U.S.A. 73*, 4179–4183.
Mason, A. J., Bronwyn, A. E., Cox, D. R., Shine, J., & Richards, R. I. (1983) *Nature (London) 303*, 300–307.
McDevitt, M. A., Imperiale, M. J., Ali, H., & Nevins, J. R. (1984) *Cell (Cambridge, Mass.) 37*, 993–999.
Miesfeld, R., Krystal, M., & Arnheim, N. (1981) *Nucleic Acids Res. 9*, 5931–5938.
Montell, G., Fisher, E. F., Caruthers, M. H., & Berk, A. J. (1983) *Nature (London) 305*, 600–605.
Mount S. M. (1982) *Nucleic Acids Res. 10*, 459–472.
Nordheim, A., & Rich, A. (1983) *Nature (London) 303*, 674–679.
Nordheim, A., Lafer, E. M., Peck, L. J., Wang, J. C., Stollar, B. D., & Rich, A. (1982) *Cell (Cambridge, Mass.) 31*, 309–318.
Ny, T., Elgh, F., & Lund, B. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 5355–5359.
O'Hara, P. J., Grant, F. J., Haldeman, B. A., Gray, C. L., Insley, M. Y., Hagen, F. S., & Murray, M. J. (1987) *Proc. Natl. Acad. Sci. U.S.A. 84*, 5158–5162.
Osterud, B., Bouma, B. N., & Griffin, J. H. (1978) *J. Biol. Chem. 253*, 5946–5951.
Plutzky, J., Hoskins, J. A., Long, G. L., & Crabtree, G. R. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83*, 546–550.
Rackwitz, H. R., Zehetner, G., Frischauf, A. M., & Lehrach, H. (1984) *Gene 30*, 195–200.
Riccio, A., Grimaldi, G., Verde, P., Sebastio, G., Boast, S., & Blasi, F. (1985) *Nucleic Acids Res. 13*, 2759–2771.
Rixon, M. W., Chung, D. W., & Davie, E. W. (1985) *Biochemistry 24*, 2077–2086.
Rogers, J. (1985) *Nature (London) 315*, 458–459.
Saito, H., Ratnoff, O. D., Bouma, B. N., & Seligsohn, U. (1985) *J. Lab. Clin. Med. 106*, 718–722.
Sanger, F., Nicklen, S., & Coulson, A. R. (1977) *Proc. Natl. Acad, Sci. U.S.A. 74*, 5463–5467.
Sinha, D., Koshy, A., Seaman, F. S., & Walsh, P. N. (1985) *J. Biol.Chem. 260*, 10714–10719.
Slightom, J. L., Blechl, A. E., & Smithies, O. (1980) *Cell (Cambridge, Mass.) 21*, 627–638.
van der Graaf, F., Greengard, J. S., Bouma, B. N., Kerbiriou, D. M., & Griffin, J. H. (1983) *J. Biol. Chem. 258*, 9669–9675.
Wang, A. H. J., Quigley, G. J., Kolpak, F. J., Crawford, J. L., van Boom, J. H., van der Marel, G., & Rich, A. (1979) *Nature (London) 282*, 680–686.
Wiggins, R. C., Cochrane, C. G., & Griffin, J. H. (1979a) *Thromb. Res. 15*, 487–495.
Wiggins, R. C., Cochrane, C. G., & Griffin, J. H. (1979b) *Thromb. Res. 15*, 475–486.
Woo, S. L. C. (1979) *Methods Enzymol. 68*, 381–395.
Yoshitake, S., Schach, B. G., Foster, D. C., & Davie, E. W. (1985) *Biochemistry 24*, 3736–3750.