Proc. Natl. Acad. Sci. USA
Vol. 82, pp. 3591–3595, June 1985
Biochemistry

Characterization of an almost full-length cDNA coding for human blood coagulation factor X

(DNA sequence analysis/protein processing/amino acid sequence identity)

MARION R. FUNG, COLIN W. HAY, AND ROSS T. A. MACGILLIVRAY

Department of Biochemistry, University of British Columbia, Vancouver, BC V6T 1W5, Canada

Communicated by Harry B. Gray, February 1, 1985

ABSTRACT **A human liver cDNA library was screened by colony hybridization with a bovine factor X cDNA probe. Three of the positive plasmids contained overlapping DNA that coded for most of human factor X mRNA. DNA sequence analysis of these three clones allowed the prediction of the complete amino acid sequence of plasma factor X. From these studies, we predict that human factor X is synthesized as a single polypeptide chain precursor in which the light and heavy chains of plasma factor X are linked by the tripeptide Arg-Lys-Arg. The cDNA sequence also predicts that human factor X is synthesized as a preproprotein having an amino-terminal leader peptide of at least 28 amino acid residues. A comparison of the amino acid sequences of human and bovine factor X shows high sequence identity around the calcium-binding regions and catalytic regions but low sequence identity around the nonfunctional regions.**

Factor X (Stuart factor) is a plasma glycoprotein that is involved in both the intrinsic and extrinsic pathways of the blood coagulation cascade (1). During the clotting process, factor X is converted from an inactive zymogen to an active protease (factor X_a) by limited proteolysis (2). Factor X has been purified to homogeneity from both bovine (3) and human plasma (4) and consists of a light chain and a heavy chain linked by a disulfide bond. The complete amino acid sequences of the light and heavy chains of bovine factor X have been reported (5, 6), as well as the complete amino acid sequence of the light chain of human factor X (7). The light chain of human factor X contains 11 residues of γ-carboxyglutamic acid, which function in the binding of calcium ions (8), and a single residue of β-hydroxyaspartic acid (7, 9), the function of which is unclear. The heavy chain contains the peptide bond that is cleaved during the activation of factor X (2) and also contains the catalytic region that is essential for the proteolytic activity of factor X_a. The amino acid sequence of this catalytic region is homologous with the catalytic regions of other serine proteases, including many clotting factors (see ref. 1).

Studies from two laboratories have shown that factor X is synthesized by rat and human hepatoma cells as a precursor consisting of a single polypeptide chain (10, 11). After secretion into the tissue culture medium, the single-chain form is converted to the two-chain form found in plasma, but the nature of this conversion was not established in these studies. The isolation and characterization of cDNA clones coding for factor X has allowed the structure of the precursor to be predicted from the cDNA sequence. Fung *et al.* (12) characterized five overlapping cDNA clones that coded for most of bovine factor X mRNA. These studies showed that bovine factor X mRNA encodes a single polypeptide in which the light and heavy chains are joined by the dipeptide Arg-Arg. The cDNA sequence also predicted that bovine factor X is synthesized as a precursor containing a leader peptide of 40 amino acid residues. This leader peptide consists of both a putative signal peptide and a "pro" region. Conversion of the profactor X to plasma factor X occurs by cleavage of a peptide bond in the sequence Arg-Arg-Ala, where Ala represents the amino-terminal residue of the light chain of plasma factor X. Thus, factor X appears to be synthesized as a preproprotein similar to other plasma proteins, including albumin (13) and apolipoprotein A-II (14). Leytus *et al.* (15) have reported the characterization of a partial cDNA coding for human factor X. This clone codes for part of the light chain of factor X, a linking tripeptide Arg-Lys-Arg, the complete heavy chain, a short 3′ untranslated region, and a poly(A) region. Thus, human factor X appears to be synthesized as a single polypeptide chain precursor in which the light and heavy chains are linked by a basic tripeptide.

We now report the isolation and characterization of three cDNA clones that code for most of human factor X mRNA, including regions coding for a leader peptide of 28 amino acid residues, the complete light chain, the linking tripeptide, the complete heavy chain, a 3′ untranslated region, and a poly(A) region.

MATERIALS AND METHODS

Materials. All enzymes were obtained from Bethesda Research Laboratories except for *Bam*HI, which came from New England Biolabs, and *Escherichia coli* DNA polymerase I and Klenow fragment, which were purchased from Boehringer Mannheim.

Screening a Human Liver cDNA Library. An adult human liver cDNA library (16) was generously provided by S. H. Orkin (Children's Hospital Medical Center, Boston). This library consists of human liver cDNA >500 base pairs (bp) long inserted into the *Pst* I site of pKT218 by homopolymeric dG·dC tailing. The cDNA library was screened by colony hybridization (17) with the 770-bp *Pst* I fragment of pBX2 (12), previously labeled by nick-translation (18), as a probe. Conditions for hybridization and washing were as described (19) to allow for possible mismatches between the bovine and human sequences. The library was later rescreened, using the 350-bp *Pst* I fragment of pcHX5 (see Fig. 1) cloned in M13mp8 as the probe (20).

Restriction Endonuclease Mapping. Plasmid DNA from positive colonies was isolated as described (21). The relationships between different plasmid isolates were determined by restriction endonuclease mapping and Southern blot analysis (22).

DNA Sequence Analysis. DNA sequence analysis was performed essentially as described by Deininger (23). Plasmid DNA was randomly sheared with a sonicator and the

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).

resulting DNA fragments were separated by electrophoresis in a 5% polyacrylamide gel. Fragments (300–500 bp) were recovered by electroelution, and the ends were repaired with T4 DNA polymerase. The sonicated fragments were then ligated into the *Sma* I site of M13mp9, and this DNA was used to transform *E. coli* strain JM103 (24). Single-stranded phage DNA was prepared as described (24). Sequence analysis was performed by the chain-termination method (25) as modified by Biggin *et al.* (26), with a synthetic heptadecanucleotide (P-L Biochemicals) as primer. The sequences of the 5′ and 3′ ends of the cDNA were confirmed by using the chemical cleavage method (27). All data were analyzed by using the DBUTIL program of Staden (28).

RESULTS AND DISCUSSION

Isolation of Human Factor X cDNAs. Bacterial colonies (240,000) of the human cDNA library were screened at high colony density with the 770-bp *Pst* I fragment of the bovine factor X cDNA pBX2 (12) as probe. Nine colonies hybridized specifically with the probe and were rescreened at lower colony density. Two positive clones from the second screen (designated pcHX5 and pcHX8) were studied further. Plasmid DNA was prepared from each of the clones and cleaved with *Pst* I. The resulting fragments were analyzed by Southern blotting, using the *Pst* I fragments of pcHX8 as hybridization probes. The analysis showed that the plasmids contained overlapping cDNA inserts (Fig. 1).

Subsequent sequence analysis showed that although pcHX8 extended to the poly(A) tail of factor X mRNA, pcHX5 lacked the extreme 5′ end of the coding region of factor X mRNA. Therefore, the human cDNA library was rescreened by using the 350-bp *Pst* I fragment of pcHX5, inserted into the vector M13mp8, as a hybridization probe. A longer clone was isolated (pcHX14; see Fig. 1); however, pcHX14 still lacked the extreme 5′ end of factor X mRNA (see below). Thus, we conclude that either the cDNA library used does not contain a full-length factor X cDNA clone or

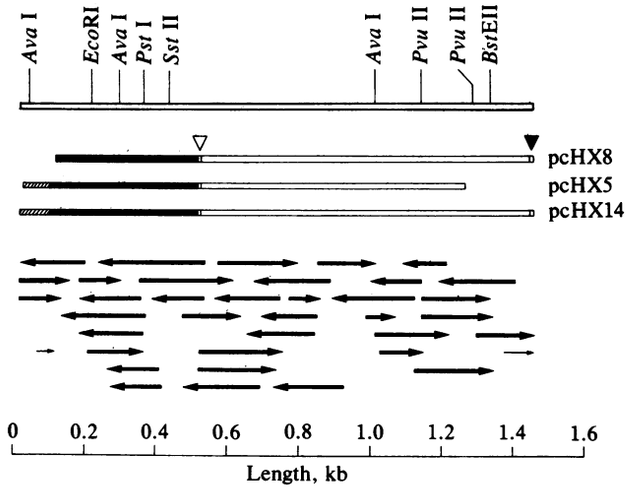

FIG. 1. Restriction map and sequencing strategy for human factor X cDNA. The bars below the restriction map represent the clones pcHX5, pcHX8, and pcHX14 and include regions coding for the leader peptide (hatched bar), the light chain of plasma factor X (solid bar), the heavy chain (open bar), and the 3′ untranslated sequence (▼). The region encoding the linker tripeptide (▽) is demarcated at the left of each open bar. The extent of sequencing is shown by the length of the arrows. DNA sequence determined on the coding strand is shown by an arrow pointing right; sequence determined on the noncoding strand is shown by an arrow pointing left. See text for details. kb, Kilobases.

such a clone is under-represented in the library compared to other factor X clones.

DNA Sequence Analysis. Most of the sequence analysis was performed using pcHX5 and pcHX8. However, in determining the sequence of the 5′ and 3′ ends, pcHX14 was also used. Plasmid DNA from pcHX5 and pcHX8 was randomly sheared and ligated into the *Sma* I site of M13mp9. Subclones containing factor X cDNA inserts were identified by plaque hybridization (24) with the *Pst* I inserts of pcHX5 and pcHX8 as probes; a total of 35 different M13 templates were isolated and their sequences were determined. This allowed the reconstruction of most of the factor X cDNA sequence (Fig. 1, thick arrows). The sequence was completed by the chemical cleavage method (27) (Fig. 1, thin arrows). The complete nucleotide sequence of human factor X cDNA and the predicted amino acid sequence for the protein are shown in Fig. 2. The position of each nucleotide was determined an average of 4.9 times, and 84% of the sequence was determined on both strands. Much of the cDNA sequence was determined for both pcHX5 and pcHX8. Only a single nucleotide difference was found between these two cDNAs; the codon for amino acid residue 344 was TTC (phenylalanine) in pcHX5 and TAC (tyrosine) in pcHX8. The clone described by Leytus *et al.* (15) contained the TAC codon in this position. This difference represents either a cloning artifact or a polymorphism in the factor X alleles of the individual whose liver mRNA was used in the construction of the cDNA library.

The sequence agrees well with those regions of factor X that had been sequenced directly by using protein chemistry techniques. Nucleotides 85–501 encode the complete light chain of factor X. The predicted amino acid sequence is in complete agreement with that determined by McMullen *et al.* (7). Nucleotides 511–1428 encode the heavy chain of factor X including three regions whose amino acid sequences have been determined previously. Nucleotides 511–558 code for the amino-terminal sequence of the heavy chain of factor X reported by DiScipio *et al.* (4), except that the cDNA sequence predicts serine residues at positions 150 and 157 (Fig. 2) whereas DiScipio *et al.* reported an unidentified residue and a threonine residue for these two positions, respectively. Nucleotides 667–717 encode the same amino-terminal sequence of the heavy chain of factor X_a reported by DiScipio *et al.* (29), except that the cDNA sequence predicts that residue 208 is a tryptophan rather than a threonine. The reason for these differences is unclear but may be the result of reverse transcriptase errors during cDNA synthesis, of polymorphisms, or of incorrect amino acid assignments during the later stages of the automatic Sequenator analyses. Nucleotides 1171–1245 encode the active-site region of factor X_a; the predicted sequence agrees with the amino acid sequence reported by DiScipio *et al.* (29). During the conversion of factor X to factor X_a, a glycopeptide of 52 amino acid residues (residues 143–194, Fig. 2) is released (29). There are two potential N-glycosylation sites in the activation peptide, at positions 181 and 191 (Fig. 2). By homology with other serine proteases (see ref. 1), the catalytic triad in factor X_a probably consists of His-236, Asp-282, and Ser-379 (Fig. 2).

As reported by Leytus *et al.* (15), the cDNA sequence predicts that the light and heavy chains of factor X are joined by the tripeptide Arg-Lys-Arg (encoded by nucleotides 502–510 in Fig. 2). McMullen *et al.* (7) reported that the carboxyl-terminal sequence of the light chain was Leu-Glu-Arg, whereas the amino-terminal sequence of the heavy chain of plasma factor X is Ser-Val-Ala (4). Thus, the basic tripeptide must be eliminated during the conversion from a single chain to the two-chain form of factor X. Similar basic peptide linkages have been found in other plasma protein precursors, including bovine factor X (12) and bovine and

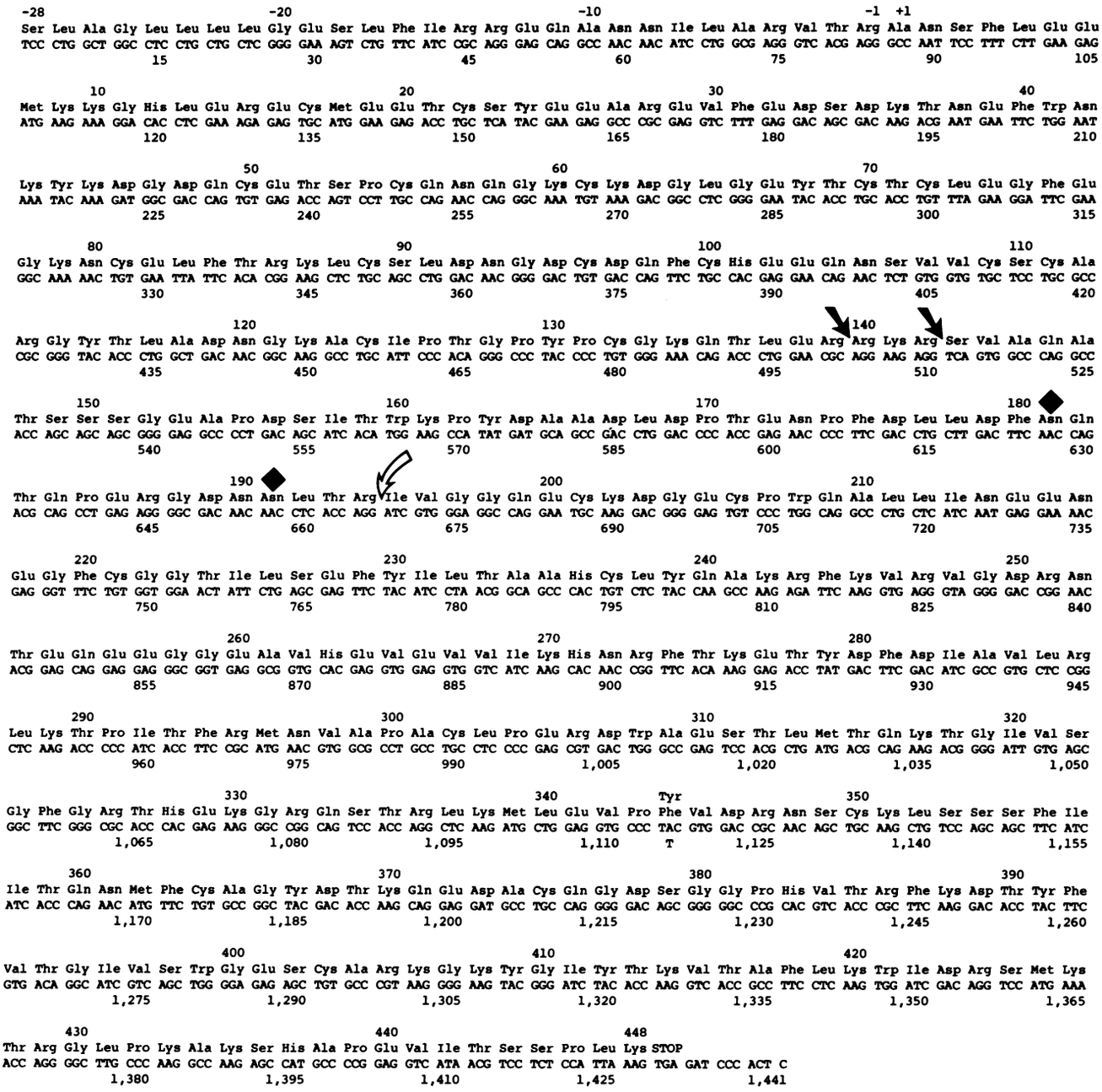

FIG. 2. Nucleotide sequence of human factor X cDNA. The sequence was determined by analysis of the overlapping clones shown in Fig. 1. The predicted amino acid sequence of human preprofactor X is shown above the DNA sequence. Putative cleavages to form two-chain factor X are shown by the solid arrows, the bond cleaved by factor IX_a is shown by the open arrow, and potential attachment sites for carbohydrate are indicated by solid diamonds. See text for details.

human protein C (30, 31). The identity of the protease(s) responsible for these cleavages is unknown.

As also reported by Leytus *et al.* (15), the cDNA sequence predicts that the heavy chain sequence is followed by a TGA stop codon (nucleotides 1429–1431 in Fig. 2), a 3′ untranslated region of 10 nucleotides (nucleotides 1432–1441), and a poly(A) tail. The putative polyadenylylation signal (32) A-T-T-A-A-A (nucleotides 1422–1427) is located 15 nucleotides upstream of the poly(A) tail. Because of the unusually short 3′ untranslated region, the polyadenylylation signal is contained within the coding region of factor X mRNA. The mRNAs coding for the β subunit of human chorionic gonadotropin (33) and the abnormal α-globin Constant Spring (34) also have short 3′ untranslated regions (16 nucleotides). In these two mRNAs, the polyadenylylation signal is located 16 nucleotides upstream of the poly(A) tail and contains the UAA codon that is used as a stop codon.

Plasmids pcHX5 and pcHX14 also contain a region coding for an amino-terminal leader peptide of 28 residues. This leader peptide does not contain a methionyl residue in the same reading frame as the factor X protein sequence, suggesting that these two clones are lacking part of the leader peptide and the 5′ untranslated region of factor X mRNA. The sequence of the leader peptide of human factor X is homologous to those found in other vitamin K-dependent clotting factors (12, 19, 30, 35, 36), as shown in Fig. 3. The amino-terminal regions of the leader sequences contain many hydrophobic residues (residues −36 to −23 in Fig. 3) and probably constitute the signal sequence necessary for translocation of the nascent polypeptide chain across the

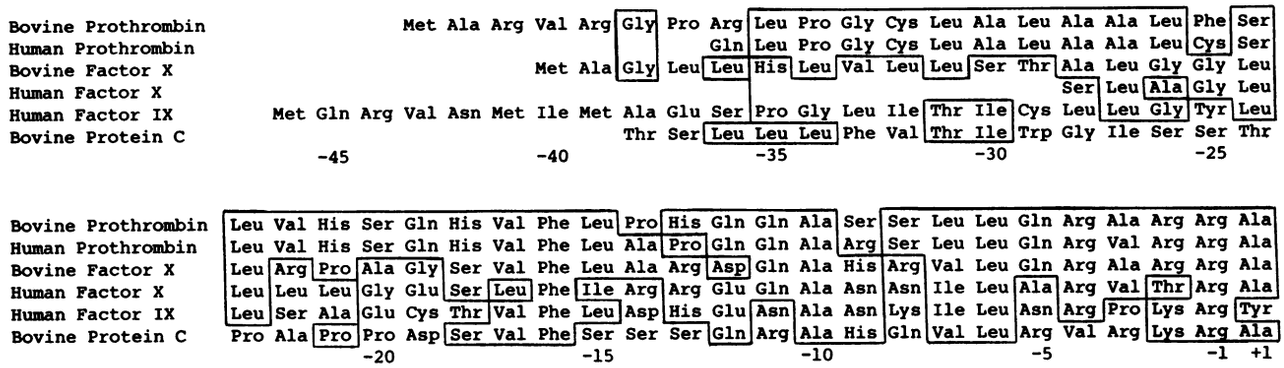

FIG. 3. Comparison of the leader sequences of human factor X, bovine prothrombin (35), human prothrombin (19), bovine factor X (12), human factor IX (36), and bovine protein C (30), as predicted from the cDNA sequences. Identical residues in corresponding positions in two or more of the protein sequences are boxed. The sequences are numbered backwards from the cleavage site that gives rise to the mature protein found in plasma. For bovine factor X and human factor IX, the 5′-most ATG codon has been assumed to code for the initiator methionyl residue. The leader sequences of human factor X, bovine protein C, and human prothrombin are incomplete, as they do not encode a possible initiator methionyl residue.

rough endoplasmic reticulum (37). This region is followed by a more hydrophilic region (residues −22 to −1) that shares greater sequence homology than the signal-peptide region. Conversion of these proteins to the form found in plasma occurs by cleavage of a bond that is carboxyl-terminal to an arginyl residue. Because this cleavage is not typical of signal peptidase (see refs. 13 and 14), it has been proposed that these clotting factors are synthesized as preproproteins. In that case, signal peptidase may cleave the preproprotein to give a pro fragment of nine residues, as an alanine residue is invariant at position −10 in the leader peptides for these proteins and human factor X (Fig. 3). The nature and the

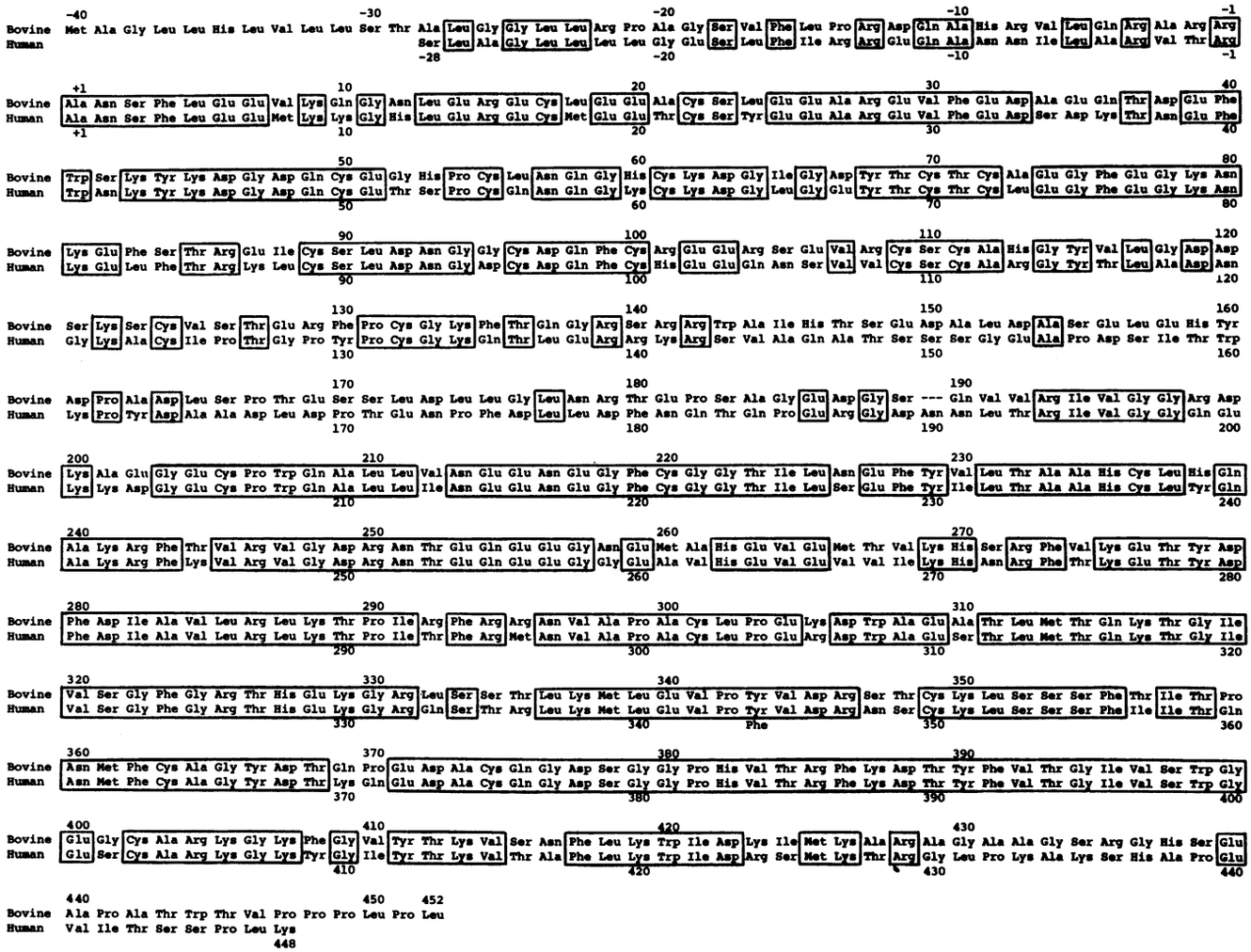

FIG. 4. Comparison of the amino acid sequences of bovine (12) and human factor X, as predicted from the cDNA sequences. Identical amino acids in corresponding positions are boxed. A single gap has been inserted in the bovine sequence (between residues 189 and 190) to maximize the homology. The carboxyl-terminal 5 residues of the bovine sequence were not encoded in the cDNA and have been taken from ref. 6.

location of the protease that converts the proprotein to the plasma form of the protein are unknown. However, human factor X differs from the other vitamin K-dependent proteins in that the proprotein protease cleaves a bond carboxyl-terminal to a Thr-Arg sequence rather than a double basic sequence.

Nucleotides 271–273 (Fig. 2) encode an aspartic acid residue that undergoes post-translational modification to form a β-hydroxyaspartic acid residue found in plasma factor X (7, 9). This is similar to the cDNAs for bovine factor X (12) and protein C (30), in which the β-hydroxyaspartic acid residue is also encoded by an aspartic acid codon.

In the positions where they overlap, the sequence for factor X cDNA agrees with that reported by Leytus *et al.* (15), with three exceptions. Leytus *et al.* reported that residues 450 and 973 (equivalent to positions 756 and 1288 in Fig. 2) were C and G, whereas our sequence contains T and A in these positions, respectively. These differences could be the result of cloning artifacts or polymorphisms in the factor X alleles studied. The third difference occurs at position 817 in Fig. 2, where both pcHX5 and pcHX8 contain the sequence A-A-G-G-T-G-A-G-G-G-T, whereas Leytus *et al.* report only -G-A- (nucleotides 511–512 in their sequence). The extra nine nucleotides are required to maintain the alignment between human and bovine factor X sequences (see Fig. 4), suggesting that the clone isolated by Leytus *et al.* may have undergone a small deletion during construction and amplification of the cDNA library.

Comparison with Bovine Factor X. A comparison of the amino acid sequence of bovine and human preprofactor X is shown in Fig. 4. Overall, the two sequences display 65% sequence identity when a single gap is inserted in the bovine activation peptide sequence (between residues 189 and 190, Fig. 4) to maximize the homology. The leader peptides exhibit only 39% sequence identity at the amino acid level but 63% identity at the nucleotide level. The light chains exhibit 70% homology at the amino acid level, and the amino acid homology is 84% for residues 194–429 of the heavy chain. Presumably, this homology reflects the functional importance of these two regions of factor X. In contrast, the activation peptides (residues 143–194 in the human sequence) and the carboxyl-terminal regions (residues 430–448 of the human sequence) exhibit 14% and 5% sequence identity, reflecting the lack of function associated with these regions. Indeed, a carboxyl-terminal peptide can be removed from the heavy chain of factor X_a without altering its activity (38).

The comparison shown in Fig. 4 differs from that reported by Leytus *et al.* (15). This is mainly the result of differences between the bovine factor X amino acid sequence determined by protein chemistry techniques (in refs. 5 and 6, used by Leytus *et al.*) and that predicted from the cDNA sequence and used in Fig. 4 (see ref. 12 for a full discussion). In every case, however, the sequence predicted from the bovine cDNA shares greater sequence identity with the human factor X sequence.

We thank Dr. Stuart Orkin for sending us the human liver cDNA library, Mark Boguski and Debbie Cool for helpful suggestions, and Dr. Grant Mauk for critically reading the manuscript. This work was supported in part by grants from the Medical Research Council of Canada, the British Columbia Health Care Research Foundation, and the Canadian Heart Foundation. M.R.F. was supported by a studentship from the Medical Research Council of Canada.

1. Jackson, C. M. & Nemerson, Y. (1980) *Annu. Rev. Biochem.* **49,** 765–811.
2. Fujikawa, K., Coan, M. H., Legaz, M. E. & Davie, E. W. (1974) *Biochemistry* **13,** 5290–5299.
3. Fujikawa, K., Legaz, M. E. & Davie, E. W. (1972) *Biochemistry* **11,** 4882–4891.
4. DiScipio, R. G., Hermodson, M. A., Yates, S. G. & Davie, E. W. (1977) *Biochemistry* **16,** 698–706.
5. Enfield, D. L., Ericsson, L. H., Fujikawa, K., Walsh, K. A., Neurath, H. & Titani, K. (1980) *Biochemistry* **19,** 659–667.
6. Titani, K., Fujikawa, K., Enfield, D. L., Ericsson, L. H., Walsh, K. A. & Neurath, H. (1975) *Proc. Natl. Acad. Sci. USA* **72,** 3082–3086.
7. McMullen, B. A., Fujikawa, K., Kisiel, W., Sasagawa, T., Howald, W. N., Kwa, E. Y. & Weinstein, B. (1983) *Biochemistry* **22,** 2875–2884.
8. Stenflo, J. & Suttie, J. W. (1977) *Annu. Rev. Biochem.* **46,** 157–172.
9. Fernlund, P. & Stenflo, J. (1983) *J. Biol. Chem.* **258,** 12509–12512.
10. Graves, C. B., Munns, T. W., Willingham, A. K. & Strauss, A. W. (1982) *J. Biol. Chem.* **257,** 13108–13113.
11. Fair, D. S. & Bahnak, B. R. (1984) *Blood* **64,** 194–204.
12. Fung, M. R., Campbell, R. M. & MacGillivray, R. T. A. (1984) *Nucleic Acids Res.* **12,** 4481–4492.
13. Strauss, A. W., Bennett, C. A., Donohue, A. M., Rodkey, J. A., Boime, I. & Alberts, A. W. (1978) *J. Biol. Chem.* **253,** 6270–6274.
14. Gordon, J. I., Budelier, K. A., Sims, H. F., Edelstein, C., Scanu, A. M. & Strauss, A. W. (1983) *J. Biol. Chem.* **258,** 14054–14059.
15. Leytus, S. P., Chung, D. W., Kisiel, W., Kurachi, K. & Davie, E. W. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 3699–3702.
16. Prochownik, E. V., Markham, A. F. & Orkin, S.H. (1983) *J. Biol. Chem.* **258,** 8389–8394.
17. Hanahan, D. & Meselson, M. (1980) *Gene* **10,** 63–67.
18. Maniatis, T., Jeffrey, A. & Kleid, D. G. (1975) *Proc. Natl. Acad. Sci. USA* **72,** 1184–1188.
19. Degen, S. J. F., MacGillivray, R. T. A. & Davie, E. W. (1983) *Biochemistry* **22,** 2087–2097.
20. Brown, D. M., Frampton, J., Goelet, P. & Karn, J. (1982) *Gene* **20,** 139–144.
21. Katz, L., Kingsbury, D. T. & Helinski, D. R. (1973) *J. Bacteriol.* **114,** 577–591.
22. Southern, E. M. (1975) *J. Mol. Biol.* **98,** 503–517.
23. Deininger, P. L. (1983) *Anal. Biochem.* **129,** 216–223.
24. Messing, J. (1983) *Methods Enzymol.* **101,** 20–78.
25. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 5463–5467.
26. Biggin, M. D., Gibson, T. J. & Hong, G. F. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 3963–3965.
27. Maxam, A. & Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560.
28. Staden, R. (1982) *Nucleic Acids Res.* **10,** 4731–4751.
29. DiScipio, R. G., Hermodson, M. A. & Davie, E. W. (1977) *Biochemistry* **16,** 5253–5260.
30. Long, G. L., Belagaje, R. & MacGillivray, R. T. A. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 5653–5656.
31. Foster, D. & Davie, E. W. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 4766–4770.
32. Proudfoot, N. J. & Brownlee, G. G. (1976) *Nature (London)* **263,** 211–214.
33. Fiddes, J. C. & Goodman, H. M. (1980) *Nature (London)* **286,** 684–687.
34. Proudfoot, N. J. & Longley, J. I. (1976) *Cell* **9,** 733–746.
35. MacGillivray, R. T. A. & Davie, E. W. (1984) *Biochemistry* **23,** 1626–1634.
36. Jaye, M., de la Salle, H., Schamber, F., Balland, A., Kohli, V., Findeli, A., Tolstoshev, P. & Lecocq, J.-P. (1983) *Nucleic Acids Res.* **11,** 2325–2335.
37. Blobel, G., Walter, P., Chang, C. N., Goldman, B. M., Erickson, A. H. & Lingappa, V. R. (1979) in *Secretory Mechanisms*, eds. Hopkins, C. R. & Duncan, C. J. (Cambridge Univ. Press, London), Vol. 33, pp. 9–36.
38. Fujikawa, K., Titani, K. & Davie, E. W. (1975) *Proc. Natl. Acad. Sci. USA* **72,** 3359–3363.