

Limulus Factor C

AN ENDOTOXIN-SENSITIVE SERINE PROTEASE ZYMOGEN WITH A MOSAIC STRUCTURE OF
COMPLEMENT-LIKE, EPIDERMAL GROWTH FACTOR-LIKE, AND LECTIN-LIKE DOMAINS*

(Received for publication, October 4, 1990)

Tatsushi Muta†, Toshiyuki Miyata‡§, Yoshio Misumi¶, Fuminori Tokunaga§, Takanori Nakamura§||, Yoshihiro Toh§, Yukio Ikehara¶, and Sadaaki Iwanaga‡**

From the ‡Department of Molecular Biology, Graduate School of Medical Science, and §Department of Biology, Faculty of Science, Kyushu University, Fukuoka 812, and ¶Department of Biochemistry, Fukuoka University School of Medicine, Fukuoka 814-01, Japan

Factor C is an endotoxin-sensitive, intracellular serine protease zymogen which initiates the coagulation cascade system in the limulus hemolymph. We have determined the entire amino acid sequence of factor C using recombinant DNA technique. The zymogen consisted of 994 amino acid residues with a calculated molecular mass of 109,648 Da. Most interestingly, factor C has five repeating units ("Sushi" domain or short consensus repeat) of about 60 amino acid residues each, which have been found in many proteins participating in the mammalian complement system. In addition to a typical serine protease domain in the carboxyl-terminal portion, characteristic segments with an epidermal growth factor-like, a lectin-like, a cysteine-rich, and a proline-rich domain were also found, revealing a unique mosaic protein structure. The serine protease domain was most analogous to human thrombin. Factor C was identified to localize in large granules in the cell, indicating that it is released from the cell by lipopolysaccharide stimulation. Furthermore, we identified a transcript possibly derived by alternative splicing of factor C mRNA, which encodes a protein sharing the amino-terminal portion of factor C. We suggest that factor C, a newly discovered type of serine protease zymogen, is a "coagulation-complement factor" which may play important roles in both hemostasis and host defense mechanisms.

Invertebrates have characteristic host defense systems which differ from mammalian immune systems (1). In *Limulus* (horseshoe crab), an *Arthropoda* species, this defense system is carried by the hemolymph, which contains a type of cell called an amebocyte. The amebocytes are extremely sensitive to endotoxin (LPS)¹, which is a major component of the cell surface of Gram-negative bacteria. When hemolymph

comes into contact with Gram-negative bacteria or LPS, the amebocytes begin to degranulate and hemolymph coagulation is initiated by the granule components (2, 3). This response is thought to be very important for *Limulus* not only in preventing the leakage of hemolymph, but also in engulfing invading microorganisms.

During the past decade we have studied the molecular mechanism of hemolymph coagulation in *Limulus* and have established a protease cascade. This cascade is based on three kinds of serine protease zymogens, factor C (4), factor B (5), proclotting enzyme (6), and one clottable protein, coagulogen (7). Endotoxin (LPS) activates the zymogen factor C to the active form, factor $\bar{C}$. Factor $\bar{C}$ then activates factor B to factor $\bar{B}$, which in turn converts the proclotting enzyme to the clotting enzyme. Each activation proceeds by limited proteolysis. The resulting clotting enzyme cleaves two bonds in coagulogen, which is a fibrinogen-like molecule in arthropods, to yield an insoluble coagulin gel. The coagulation cascade is also activated by (1,3)- β -D-glucan. In the presence of (1,3)- β -D-glucan, a protease that is tentatively named factor G is activated, leading to activation of the proclotting enzyme (8).

Factor C that binds LPS exists as a single-chain glycoprotein with an estimated molecular mass of 123 kDa or as a two-chain form, composed of one heavy (H) and one light (L) chain (4, 9). When this protein binds to LPS, it is autocatalytically activated to the active form (factor $\bar{C}$). Upon activation, a Phe-Ile bond in the L chain is cleaved, generating an A and a B chain. The resulting factor $\bar{C}$ is composed of the disulfide-linked H, A, and B chains (mentioned in order of appearance from the amino terminus in the single-chain form of factor C). The activation of factor C by Phe-Ile cleavage was quite unexpected. This is the first known activation mechanism in the coagulation and complement cascade systems involving a chymotrypsin-like specificity.

Our previous results have shown that the B chain of factor $\bar{C}$ has a serine protease domain, while the LPS-binding site is located in the H chain (10, 11). In the A chain we found a complement-like structure composed of about 60 amino acid residues (9). These characteristics of factor C suggest that this protein is a mosaic protein with a structure similar to that of complement component C1r, which is an initiator of the mammalian complement cascade.

To obtain more information about the domain-structure of this molecule, the structure of the LPS-binding site(s) and the mechanism of autocatalytic activation, we have established the cDNA sequence for factor C mRNA. Structural comparison of factor C with other proteins in various systems

* This work was supported by a Grant-in-Aid for Scientific Research from the Ministry of Education, Science and Culture of Japan. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EMBL Data Bank with accession number(s) D90271 and D90272.

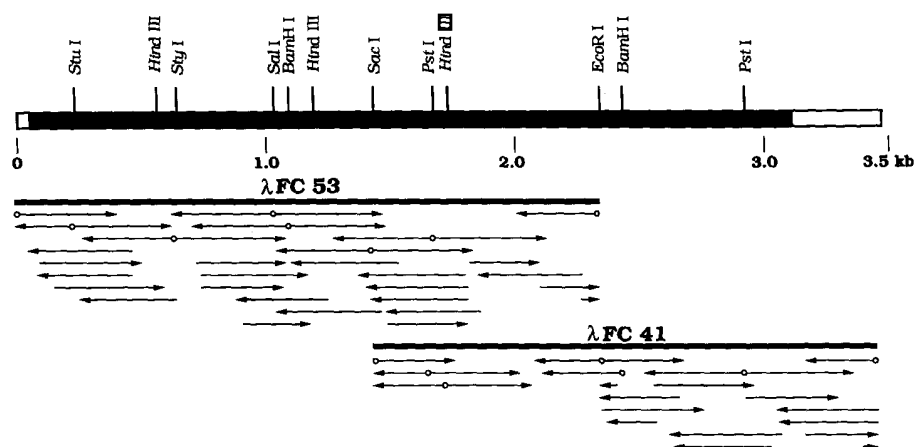
This article is dedicated to the memory of Staffan Magnusson.

|| Present address: Frontier Research Program, Institute of Physiological and Chemical Research (RIKEN), Wako, Saitama, 350-01, Japan.

** To whom correspondence should be addressed.

¹ The abbreviations used are: LPS, lipopolysaccharide; bp, base pairs; kb, kilobases; EGF, epidermal growth factor.

FIG. 1. Restriction map and sequence strategy. A black box at the top shows the open reading frame. Arrows with circles show sequencing utilizing restriction sites and arrows without circles were by deletions.



may offer insights into the evolution of host defense systems in the animal kingdom.

MATERIALS AND METHODS²

RESULTS

Cloning of Factor C cDNA—Four-hundred thousand independent clones of the λ gt 11 cDNA library constructed from amebocyte poly(A)⁺ RNA were screened with anti-factor C IgG. Of 16 positive clones purified, the longest positive clone, λ FC53, was sequenced. This clone was shown to encode the 5'-noncoding region, the entire H and A chain, and an amino-terminal portion of the B chain (nucleotides 1–2467). To obtain clones encoding the remaining part of the B chain, we prepared a second cDNA library and screened it with the [³²P] λ FC53 insert as a hybridization probe. The clone λ FC41 encoding a cDNA sequence covering the carboxyl-terminal region of factor C starting in the middle part of the H chain (nucleotides 1438–3474), was thus obtained and sequenced (Fig. 1). The composite nucleotide sequence obtained from these two clones spans 3,474 bp (Fig. 2). The first ATG codon at nucleotide 45 begins an open reading frame of 3,057 bp followed by a TGA termination codon and a 365-bp 3'-noncoding sequence before a poly(A) tail. There is one potential polyadenylation or processing signal with the sequence AATAAA in the 3'-noncoding sequence which begins 30 bp before the poly(A) tail. The open reading frame encodes a putative signal sequence of 25 amino acid residues and a mature protein of 994 residues. The molecular mass of the mature protein without carbohydrate chains was calculated to be 109,648 Da. There are six potential N-glycosylation sites within the sequence Asn-Xaa-Ser/Thr. The potential glycosylation site (Asn⁷¹⁵) has been shown not to be glycosylated.

Partial Amino Acid Sequences of the Peptides Obtained from Factor C—To confirm the sequences of these clones, we determined partial amino acid sequences of factor C. The S-alkylated H and L chains were digested with lysyl endopeptidase or trypsin (Fig. M1). Purified peptides, whose purity was supported by their amino acid compositions, were sequenced. In this way, we obtained the sequences of 347 amino acid residues. Together with the peptide sequences published previously (9), amino acid sequences covering a total of 493 residues (underlined in Fig. 2), which comprises 49.6% of the

mature protein, have been confirmed by protein sequencing. The amino acid sequences agree completely with the sequence deduced from the cDNA sequence.

Northern Blotting—Northern blot analysis of amebocyte poly(A)⁺ RNA showed a band at about 3.9-kb hybridizing with λ FC53, in addition to a faint 2.0-kb band, which indicates that the composite sequence of 3.5 kb contains almost the full-length factor C cDNA when taking the length of the poly(A) tail into consideration (Fig. 3, lane 1).

The faint band at 2.0 kb observed in Northern blotting was thought to be a transcript with similar but not identical sequence to factor C mRNA. To investigate the sequence difference between the two bands, we prepared two kinds of probes from λ FC53. One of them was an *Eco*RI-*Sty*I fragment of λ FC53, encoding the 5'-terminal part of the cDNA, and the other was the 3'-terminal part of λ FC53, a *Pst*I-*Eco*RI fragment, corresponding to the carboxyl-terminal part of the H and A chain. The 2.0-kb band was also observed in the 5'-terminal probe (data not shown). However, it was not detectable in the 3'-terminal probe (Fig. 3, lane 2).

Cloning of the Transcript Corresponding to the 2.0-kb Band—By using the difference in reactivity to these two probes, we searched for clones corresponding to the 2.0-kb bands from the positive clones previously obtained. One clone, λ FC42, with a length of 1,915 bp, was isolated as a clone that hybridized with the 5'-terminal fragment but not with the 3'-terminal fragment. To deny the possibility that this clone was derived from cloning artifacts, a screening was newly carried out against one million clones. Of 194 clones that hybridized with λ FC53, two independent clones, λ AS1 and λ AS2, did not hybridize with the 3'-terminal fragment. Both of them had the same sequence as λ FC42 (Fig. M2).

The sequence analysis revealed that λ FC42 contains 55 bp of 5'-noncoding region, 1,494 bp of an open reading frame, 357 bp of 3'-noncoding region with a polyadenylation signal, and a short poly(A) tail. The downstream sequence from nucleotide 1528 was distinct from that of factor C, whereas the portion involving nucleotides 12–1527 was identical to that of nucleotides 1–1516 of factor C (Fig. 4). A protein encoded by these clones has 473 amino acid residues with a molecular mass of 51,652 Da, of which the amino-terminal 466 amino acids are identical to those of factor C. This clone seemed to have arisen by alternative splicing (see "Discussion").

Northern blotting using a probe from the 3'-terminal *Eco*RV-*Eco*RI fragment of λ FC42, that is unique to this clone, gave a band at the position corresponding to the 2.0-kb band (Fig. 3, lane 3).

² Portions of this paper (including "Materials and Methods," Figs. M1 and M2, and Table M1) are presented in miniprint at the end of this paper. Miniprint is easily read with the aid of a standard magnifying glass. Full size photocopies are included in the microfilm edition of the Journal that is available from Waverly Press.

FIG. 2. The composite nucleotide sequence (upper) and the deduced amino acid sequence (lower) of factor C. A polyadenylation signal (AA-TAAA) is double-underlined. The amino acid sequence is numbered beginning at the amino terminus of the mature factor C molecule. The amino acid residues that have been confirmed by sequencing of purified peptides are underlined and include data from Tokunaga *et al.* (9). Amino-terminal amino acid sequencing of each chain identified the three cleavage sites marked by arrowheads. The potential carbohydrate attachment sites for (Asn-Xaa-Ser/Thr) are indicated with closed diamonds and the putative attachment site confirmed not to carry carbohydrate is shown as an open diamond. The amino acid residues suspected to form the catalytic triad of the serine protease domain are circled.

CAGGTTTAAACGGAAGCTGGAAGCTCTGAAGTAACTTAACTATGCTCTAGCTGCTTTTGGTCTGGTTAGTTCTAGGACTAGCCCAACAAATGGCTGAGTCTAGTCCA 120
 M V L A S F L V S G L V L G I L A Q Q M R P V Q S R
 -25
 GAGGAGTAGATCTGGCTTGTGTGATGAAGAGGCTTGAAGTGTGAGATCCAGGCTATGTTTCAAGCTCCATGAACAACTGACGACTTCTATGATGAGGCTTATT 240
 G V D L G L C D E T R F E C K C G D P G Y V F N V P M K Q C T Y F Y R W R P Y C
 -40
 GTAACCATGTGATGACCTGGAGGCTAAGGACATTTGTCCAAAGTACAAAGATGTCAAGATGTAAAGCTGCTTGTATGATGTGTACTTGTCCACCTAACAAATATGCTACTTGGT 360
 K P C D D L E A K D I C P K Y K R C Q E C K A G L D S C V T C P P N K Y G T W C
 -60
 GTAGCGGTGAATGTCAATGTAAAGATGAGGATCTGTGACGAGGACAGGAGCTTGTACCTGTGTCGACAGATATGAAGGAGGCACTGTGAATCTCAAGAGGTGTCTCTCTTC 480
 S G E C Q C K N G G I C D Q R L P G A C T C R D R Y E G A H C E I L K G C P L D P
 -80
 CATGGATTCTCAAGTTCAGGAAGTCAAGAACCCAGAGATATCCCAAACTATTGACACAGCTGTTCACAGGTTCAAGCTTAAAGGCTGGCAAGATAGCTGTCTTCCAAATG 600
 S D S Q V Q E V R N P P D N P Q T I D Y S C S P G F K L K G V A R I S C L P N G
 -100
 GACAGTGGATGCTTCAACCAATGATTTGAGATGTGCAAGGTTTCATCTCCAGAACAGGAAAGTGAATGCTCTAGTGGCAATATGATGAAGGGGCTACTTGTAGGTTCT 720
 Q W S S F P P K C I R E C A K V S S P E H G K V N A P S G N M I E G A T L R F S
 -120
 CATGTGATGTGCTACTTGTATGTTGTCAGAAACATTAACCTGCGAGGTAATGTGTCAGTGGAGTGGCAAAATACCAAGTGAAGTGTGCTGTCTGCTGACCTGTATCTGT 840
 C D S P Y Y L I G Q E T L T C Q G N G Q W S G Q I P Q C K K L V F C P D L D P V
 -140
 TAAACATGCTGAACACAGGTTTAAATTTGTGTGGAACAAATATGTTGCTTCTTCAAGGCACTGAAGTGAACCTATAGTGTGTGGGTGAACCTTCTGATGGTGTTCACACT 960
 N H A E H Q V K I G V E Q K Y G Q F P Q G T E V T Y T C S G N Y F L M G F N T L
 -160
 TAAATGTAAACCTGAGTGGTCTGTCAGGATCAGACCATCTCTGTAAAGTGGACAGACAGAGGTGACCTGTGACAGTAAAGCTGTAGACTTCTTGGATGATGTGTGAACCTG 1080
 K C N P D G S W S G S Q P S C V K V A D R E V D C D S K A V D F L D D V G E P V
 -180
 TCAGATCCACTGTCTGCTGCTGCTCTTGTGACAGCTGTACTGTGTGGGTACAGCCATATACCAAGCTTCTCAGTGTGTGTCGACCATATGATGAGGAGTGTGACACT 1200
 R I H C P A G C S L T A G T V W G T A I Y H E L S S V C R A A I H A G K L P N S
 -200
 CTGAGGGGGGCTGATGTGATGAACATGGCCCTACTGCGACTTCTGCTGATGACCTGATGGGATAAATGCGAAGGTTCGATGCTGTGATGCTGTGATGCTGTGATG 1320
 G G A V H V V N N G P Y S D F L G S D L N G I K S E E L K S L A R S F R F D Y V
 -220
 TCAGTTCACACAGAGTAGATCAGATGCTCTGATGATGCTTGTGAGTGAAGAACTGTGTGTAGTACATCAAAACAGAGCTGGGAAAGAGCTCAAGGTGTGTACCA 1440
 S S S T A G T A G C P D G W F E V E E N C V Y V T S K R A W E R A G Q V C T N
 -240
 ATATGCTGCTGCTGCTGCTGCTGCTAGACAAAGATTAATTCAGGTTCTGCTGACCTGACCTGACCTGACCTGACCTGACCTGACCTGACCTGACCTGACCTGACCT 1560
 M A A R L A V L D K D L I P S S L T E T L R G K G L T T T W I G L H R L D A E K
 -260
 AGCCCTTGTGTGGAGCTAATGATGTAGTATGTGTGTGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGAT 1680
 P F V W E L M D R S N V V L N D N L T F W A S G E P G N E T N C V Y L D I R D Q
 -280
 AGCTGACCTGTGTGAGGAAACAGTCAAGTCAAGCTCAAGCTTCTGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGAT 1800
 L Q P V W K T K S C F Q P S S F A C M M D L S D R N K A C C D D P G L E N H
 -300
 AGCCACACTTCATGACAAAGTATGATGGTCTTATGCTGCTTCTTATAGGTACAGCTGTGAGGTTCCTCACTACTGAGTGAACCTGAGACCTGACCTGACCTGACCT 1920
 A T L H G Q S I D G F Y A G S S I R Y S C E V L H Y L S G T E T V T C T T N G T
 -320
 CATGGATGCTTAAACCTGATGATCAAGTCAATCACTGCAACCTGCAACCTGCAACCTGCAACCTGCAACCTGCAACCTGCAACCTGCAACCTGCAACCTGCAACCT 2040
 W S A P K P R C I K V I T C Q N P P V P S Y G S V E I K P P S R T N S I S R V G
 -340
 GGTCACTTCTTGAAGTGTGCAAGGTTACCCCTGCAATGACGAGAGAGCAACCTCTCCAAACCTGATGCTCAGAACCTCTACTGCTGAGTGTGCTGCTGATGATGATGAT 2160
 S P F L R L P R L P L P L A R A A K P P P K P R S S Q P S T V D L A S K V K L P
 -360
 CTGAAGTCAATTACCGGTTAGGTTCTGAGCCATTACAGCTGCGAGTGTGAGTACTGAGTACTGAGTACTGAGTACTGAGTACTGAGTACTGAGTACTGAGTACTGAGT 2280
 E G H Y R V G S R A I Y T C E S R Y E L L G S Q G R R C D S N G N W S G R P A
 -380
 CTAGCTGATTCAGCTTGTGAGGCTGAGCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT 2400
 S C I P G C R S D S P R S P F I W N G N S T E I G Q W P W Q A G I S R W L A D
 -400
 ACCAATATGTGTTTCTCAAGTGTGAGGATGCTTATGATGAAGTGTGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGAT 2520
 H N M W F L Q C G G S L L N E K W I V T A A C C V T Y S A T A E I I D P S Q F K
 -420
 AAATCTATCTGGCAAGTACTACCTGATGACAGTACAGAGTACAGTACAGTACAGTACAGTACAGTACAGTACAGTACAGTACAGTACAGTACAGTACAGTACAGT 2640
 I Y L G K Y Y R D D S R D D D Y V Q V R E A L E I H V N P N Y D P G N L N F I
 -440
 TAGCCATATCACTGAAACCTCTGTTTGTGACCAAGAGTGTCAACCAATCTGTCTGCTACTGACATCACAACAGGACATCTGAAGGAGGAGGATACAGCTGACAGT 2760
 A L I Q L K T P V I C L P T D I T T R E H L K E G T L A V V T G
 -460
 GTTGGGTTTGAATGAAACACATATTCAGAGATGATTCACCAAGCTGTCTGCTGTTGTGTCAGCAAGCACTGTGAAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAG 2880
 W G L N E N T Y S E M I Q Q A V L P V V A A S T C E E G Y K E A D L P L T V T
 -480
 CAGAACATGTCTGTGAGGTTACAGAGGAGGAGTATGATGCTGCTGAGTGTGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAGGAG 3000
 E N M F C A G Y K K G R Y D A C S G D G G P L V F A D D S R T E R R W V L E G
 -500
 GGATGTGCTGAGTGGGAGTCCAGTGTGAGGAGGAGTAAAGGCTTAAAGGCTTAAAGGCTTAAAGGCTTAAAGGCTTAAAGGCTTAAAGGCTTAAAGGCTTAAAGGCT 3120
 I V S W G S P S G C G K A N Q Y G G F T K V N V F L S W I R Q F I
 -520
 TTAATCATGTTTAAAGCTTGTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT 3240
 GTCAATGATGAGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGATGAT 3360
 TTCAATGAAGTGGTGTCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCT 3474

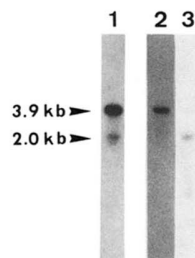


FIG. 3. Northern blot analysis of factor C mRNA. Five μ g of poly(A)⁺ RNA from amebocytes were separated by 1.0% agarose gel electrophoresis, transferred to the membrane, and hybridized with ³²P-labeled probes. Hybridization probes used were the λ FC53 (lane 1), a *Pst*I-*Eco*RI fragment of λ FC53 (lane 2), or an *Eco*RV-*Eco*RI fragment of λ FC42 (lane 3), respectively. See text for details.

1 12
 GTAAGATATCATC-
 1527
 -GTACTGATAATGTTACTGCAACCT 1550
 T D N V T A T *
 473
 AAGTAGACTTTATTTAGTCTAAGATATCTGTGCTCAATTGTGACTTGC 1600
 TACTCTTACTTTATGTAATAAAAGCTGTATATAAGTATTAAGTCTTGTT 1650
 ATTTTTACTGTTTAAAGTATACTACAAGGTTTACACTTAAAGGAAAT 1700
 TTCTATCTGACAAATTTCAATAAAAGTGGATGTTACTTTTGTATTTCTGTTG 1750
 GTTGTAAATGCACATCCGGTTTGTAAACATCTAAGTTCCCAACACAGG 1800
 CTCTGTGGTCTTGTAAAAATCTAAACTAATACCACCTTTTAATATGTTT 1850
 AACCTTTTATTTATGTATGCTTATCAGAGACTAAATAAAGGTTTAAAT 1900
 AACTGTAAAAA 1915

FIG. 4. The nucleotide and deduced amino acid sequences of λ FC42. The sequences of the regions which differ from factor C mRNA are shown. Nucleotides 12–1527 are identical with nucleotides 1–1516 of factor C (Fig. 2). The possible 5' splice junction consensus sequence, GT, is shown by an underline. The potential N-glycosylation site is shown as a closed diamond. A polyadenylation signal (AATAAA) is double-underlined.

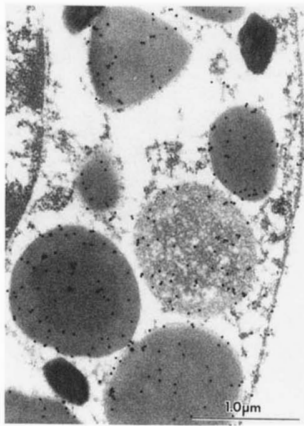


FIG. 5. Subcellular localization of Factor C. Hemocytes were fixed, stained with anti-factor C IgG and gold-particle-conjugated second antibody, and then scanned. Numerous gold particles are present in the large granules. The bar is 1 μ m.

Subcellular Localization of Factor C—Using anti-factor C polyclonal antibody, the subcellular localization of factor C was examined. Sections of the amebocytes were treated with anti-factor C antibody followed by gold particle-conjugated second antibody. Many gold particle conjugates were observed mainly in large granules in the cell, indicating that factor C antigen is localized in the large granules and exocytosed by stimulation with LPS (Fig. 5).

DISCUSSION

A computer-assisted homology search indicated that several interesting folding structures exist in the sequence of factor C. First, five internally repeated units with about 60 amino acid residues each are present, four in the H chain in addition to one existing in the A chain (Fig. 6A). These units are called "Sushi" domains (16) because their disulfide folding structure, which was first determined in β_2 -glycoprotein I (17), resembles Sushi, in traditional Japanese cuisine. They are also called short consensus repeat. This kind of domain has been found as tandem repeats in more than 20 proteins, mainly in proteins participating in the mammalian complement system. Most of the complement factors shown to have this domain interact with C3b or C4b (18). Actually, some of the Sushi domains in complement receptor type 1 (CR1) have been shown to be binding sites for C3b or C4b (19). They seem to provide sites for protein-protein interactions in other proteins. Sushi domains in factor C may interact with other proteins such as *Limulus* factor B or other components. No protein containing the Sushi domain has, until now, been identified in invertebrates. Factor C is the first example of a protein that has a complement-like structure in invertebrates.

Second, an epidermal growth factor (EGF)-like domain is present in the H chain (Fig. 6B). This EGF-like domain displays most similarity to that of the laminin B2 chain. Proteins containing this type of domain involve several blood coagulation factors, complement factors, some receptors, including the low density lipoprotein receptor. Cys⁸¹ and Cys⁹³, Cys⁸⁷ and Cys¹⁰⁰, and Cys¹⁰² and Cys¹¹¹ are thought to form disulfide bonds by analogy with the folding structure determined in bovine factor X (20). Asp⁹⁴, which is a site for β -hydroxylation in some EGF-like domains, was shown to be unmodified by amino acid and sequence analyses.

Third, a lectin-like domain was unexpectedly found between the third and fourth Sushi domains. This region shows similarity to so called C-type lectins (21) (Fig. 6C). No protease has, until now, been known to include a lectin-like

domain. This type of domain is also found in proteins whose functions are not limited to binding of carbohydrates, such as the IgE receptor, and pulmonary surfactant-associated protein. Moreover, this domain shows similarity to cartilage link protein and proteoglycan core protein.

There are two potential glycosylation sites in this domain (Asn⁴⁹⁸ and Asn⁵⁰⁹). Both of these asparagines could not be identified in amino acid sequencing, suggesting that these residues are modified. Cys⁴¹¹ and Cys⁴²², Cys⁴³⁹ and Cys⁵³⁹, and Cys⁵¹³ and Cys⁵³¹ are thought to be linked to one another based on the disulfide folding structure determined in the lectin from acorn barnacles (22).

Two other characteristic sequences were found in the H chain. One is a Cys-rich region, which is located in the amino-terminal portion of the H chain. Eleven cysteine residues out of 56 are clustered in this region. The other is in the carboxyl-terminal portion of the H chain, in which 13 of 48 residues are proline. This region shows similarity to connecting region of mammalian coagulation factor XII (23) (Fig. 6D).

The B chain portion of factor C is composed of a typical serine protease, with a catalytic triad, His⁷⁸⁴-Asp⁸⁴⁰-Ser⁹⁴¹, judged from similarity with other proteases. Six cysteines out of seven in the B chain are believed to form disulfide bonds between Cys⁷⁶⁹ and Cys⁷⁸⁵, Cys⁹⁰⁷ and Cys⁹²⁶, and Cys⁹³⁷ and Cys⁹⁷¹ by analogy with those of trypsin (24). The remaining Cys⁸⁶⁰ is considered to pair with Cys⁷²⁷ in the A chain. The B chain shows the strongest similarity to human α -thrombin (36.7%). This is consistent with the result that factor C has a similar specificity toward synthetic fluorogenic peptide substrates as α -thrombin (4). The substrate-binding site, Asp⁹³⁵, is conserved, which indicates that factor C is a trypsin type enzyme, although the cleavage site associated with zymogen activation is Phe⁷³⁷-Ile⁷³⁸ (9). The proline residue at the P₂ site could be important for cleavage. As a matter of fact, the peptide bond cleaved during conversion to the two-chain form, Arg⁶⁶⁵-Ser⁶⁶⁶, is also preceded by a proline residue.

The mosaic structure of factor C is summarized in Fig. 7. Three membrane proteins, endothelial leukocyte adhesion molecule (25), lymph node homing receptor (26), and granule membrane protein 140 (27), have been reported to contain a lectin-like, an EGF-like domain and several Sushi-like domains. They are thought to play roles in cell recognition and the name "Selectin" has been proposed for them. The similarity in domain composition between factor C and these proteins is of interest because these three proteins all seem to be associated with biological defense systems. Factor C may play another role, for example, cell adhesion after degranulation, which seems to be important for the removal of clots and foreign material.

There is a report indicating that all components necessary for hemolymph coagulation exist in the granules of the amebocytes (28). Our immunohistochemical observation with electron microscopy demonstrates that factor C is localized in the large granules in the cell. When LPS comes in contact with amebocytes, these granules are secreted and hemolymph coagulation is initiated by the components of the granules.

We also cloned a minor transcript found on Northern blotting which shares the 5'-terminal sequence with factor C mRNA. Since the boundary in λ FC42 contains the sequence GT (nucleotides 1527 and 1528), a conserved sequence at 5' splice sites, this region appears to be a 5' splicing junction site and alternative splicing results in the transcription of this clone. The protein encoded by this transcript has the same amino-terminal region as factor C, including the Cys-rich, the EGF-like, and three Sushi domains and the first half of the lectin-like domain (Fig. 7). Only 7 amino acid residues with

FIG. 6. Sequence similarities between factor C and other proteins: Sushi domains (A); EGF-like domains (B); lectin-like domains (C); proline-rich regions (D). The sequence references are as follows: A, human factor XIII b subunit, mouse factor H, human factor B, human β_2 -glycoprotein I, and human C4b-binding protein (29). B, human laminin B2 chain, (30), bovine protein C, factor X, factor IX, protein Z, and protein S, and human C1r and low density lipoprotein receptor (31), bovine factor VII (32), and human C1s (33). C, lectin from acorn barnacles (22), human IgE receptor, and chicken hepatic lectin (21) and human asialoglycoprotein receptor 1 (34). D, human factor XII (23). Amino acid residues identical with those of factor C are boxed. Conserved cysteine and tryptophan residues are shadowed.

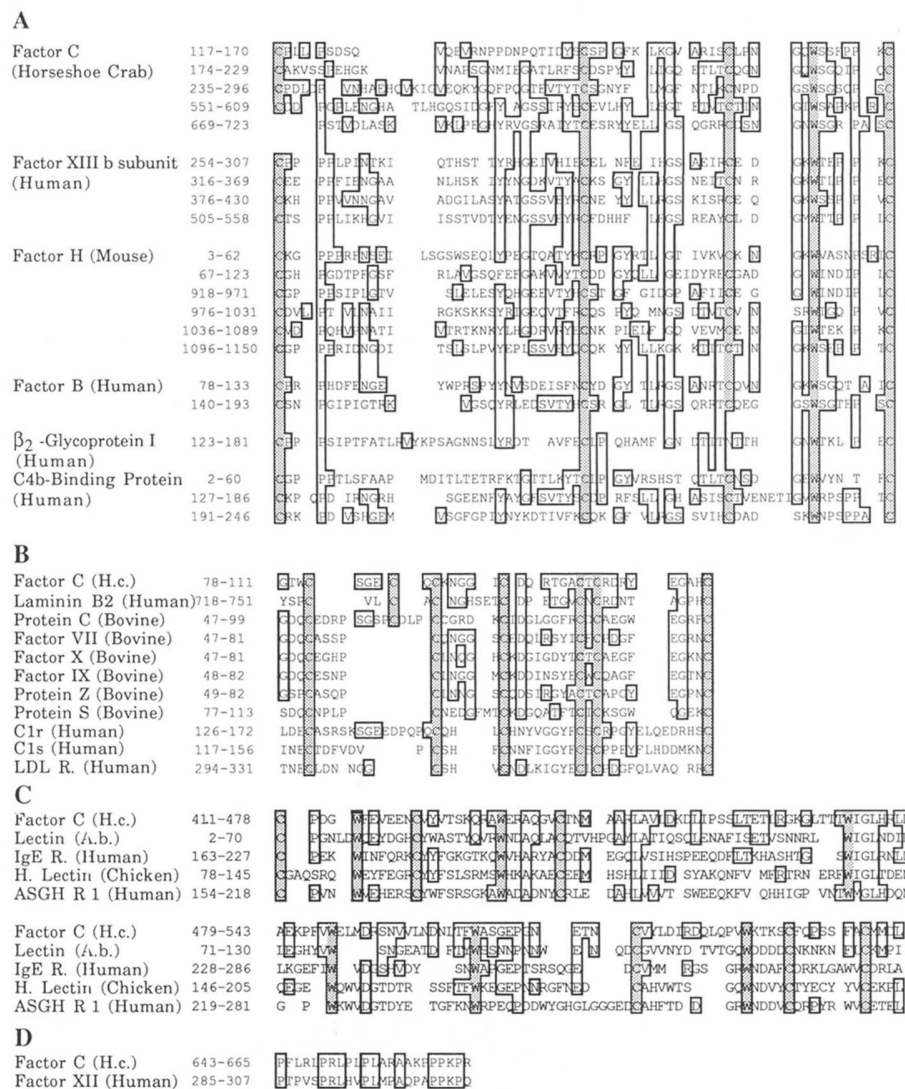
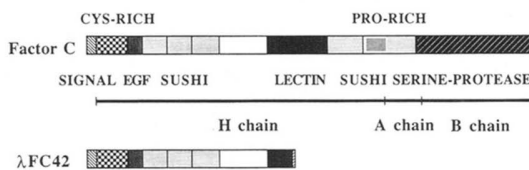


FIG. 7. The domain structure of factor C. Factor C has a mosaic structure of several domains including a signal peptide (–25 to –1), one Cys-rich region (1–76), one EGF-like domain (77–111), five Sushi domains (117–170, 174–229, 235–296, 551–609, and 615–723), one lectin-like domain (411–543), one Pro-rich region (618–665, located within the fifth Sushi domain) and one serine protease domain (738–994). The H, A, and B chain correspond to each domain as illustrated. A tentative alternative splicing product encoded by λ FC42 contains the signal peptide, the Cys-rich region, the EGF-like domain, three Sushi domains, and the first half of the lectin-like domain.

one potential N-linked glycosylation site (Asn-Val-Thr) at its carboxyl-terminal end are different from mature factor C.

The mosaic structure of serine proteases containing Sushi domains is shown in Fig. 8. These proteases, except for factor C, have been found only in mammalian complement systems. Proteases shown to have both Sushi and EGF-like domain include only factor C, C1r and C1s, each of which is associated with the initiation of cascade reactions. This structural similarity is suggestive because factor C and C1r can be thought

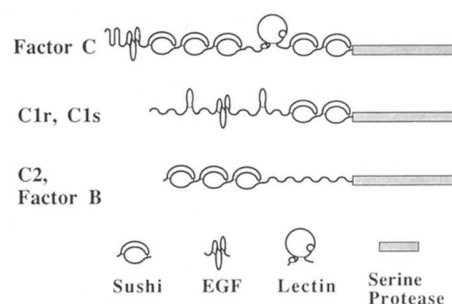


FIG. 8. Comparison of the structures of serine proteases containing Sushi domains. The domain structures of serine proteases with Sushi domains including *Limulus* factor C, C1r, C1s, C2, and complement factor B are illustrated.

to have similar activation mechanisms. LPS for factor C may correspond to C1q bound to an immune complex for C1r, and both activations occur autocatalytically.

In conclusion, factor C, the putative initiator of the *Limulus* coagulation cascade, is structurally related to mammalian complement factors. The coagulation cascade functions in preventing not only leakage of hemolymph but also invasion of foreign substances into the hemolymph. The initiator of this cascade, factor C, is a newly discovered type of serine protease zymogen, a "coagulation-complement factor," which

may play an important role both in hemostasis and defense mechanisms. In considering the structural similarity of the initiators of the two systems, the *Limulus* coagulation and mammalian complement systems may have evolved from a common origin. With regard to the evolution of host defense systems in the animal kingdom, this is consistent with a possible common evolutionary origin of the coagulation and complement cascades.

Acknowledgments—We would like to express our thanks to Dr. Akio Ito for his technical advice, Dr. Hideo Igarashi for preparing antiserum, Chizuko Takabayashi-Sueyoshi for her technical assistance in the protein-based study, and Drs. Kazuo Fujikawa, Lawrence J. Berliner, Daniel A. Walz, and Staffan Magnusson for manuscript preparation. We also thank Satsuki Kajiyama for her technical assistance and Nobuko Ueno for her secretarial assistance.

REFERENCES

- Brehélin, M. (ed) (1986) *Immunity in Invertebrates*, Springer-Verlag, Berlin
- Levin, J., and Bang, F. B. (1964) *Bull. Johns Hopkins Hosp.* **115**, 265–274
- Ornberg, R. L., and Reese, T. S. (1979) *Prog. Clin. Biol. Res.* **29**, 125–130
- Nakamura, T., Morita, T., and Iwanaga, S. (1986) *Eur. J. Biochem.* **154**, 511–521
- Nakamura, T., Horiuchi, T., Morita, T., and Iwanaga, S. (1986) *J. Biochem. (Tokyo)* **99**, 847–857
- Nakamura, T., Morita, T., and Iwanaga, S. (1985) *J. Biochem. (Tokyo)* **97**, 1561–1574
- Miyata, T., Hiranaga, M., Umez, M., and Iwanaga, S. (1984) *J. Biol. Chem.* **259**, 8924–8933
- Morita, T., Tanaka, S., Nakamura, T., and Iwanaga, S. (1981) *FEBS Lett.* **129**, 318–321
- Tokunaga, F., Miyata, T., Nakamura, T., Morita, T., Kuma, K., Miyata, T., and Iwanaga, S. (1987) *Eur. J. Biochem.* **167**, 405–416
- Nakamura, T., Tokunaga, F., Morita, T., Iwanaga, S., Kusumoto, S., Shiba, T., Kobayashi, T., and Inoue, K. (1988) *Eur. J. Biochem.* **176**, 89–94
- Nakamura, T., Tokunaga, F., Morita, T., and Iwanaga, S. (1988) *J. Biochem. (Tokyo)* **103**, 370–374
- Chirgwin, J. M., Przybyla, A. E., MacDonald, R. J., and Rutter, W. J. (1979) *Biochemistry* **18**, 5294–5299
- Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B. H., and Hood, L. E. (1986) *Nature* **321**, 674–679
- Vieira, J., and Messing, J. (1987) *Methods Enzymol.* **153**, 3–11
- Lipman, D. J., and Pearson, W. R. (1985) *Science* **227**, 1435–1441
- Ichinose, A., Bottenus, R. E., and Davie, E. W. (1990) *J. Biol. Chem.* **265**, 13411–13414
- Lozier, J., Takahashi, N., and Putnam, F. W. (1984) *Proc. Natl. Acad. Sci. U. S. A.* **81**, 3640–3644
- Müller-Eberhard, H. J. (1988) *Annu. Rev. Biochem.* **57**, 321–347
- Klickstein, L. B., Bartow, T. J., Miletic, V., Rabson, L. D., Smith, J. A., and Fearon, D. T. (1988) *J. Exp. Med.* **168**, 1699–1717
- Højrup, P., and Magnusson, S. (1987) *Biochem. J.* **245**, 887–892
- Drickamer, K. (1988) *J. Biol. Chem.* **263**, 9557–9560
- Muramoto, K., and Kamiya, H. (1986) *Biochim. Biophys. Acta* **874**, 285–295
- McMullen, B. A., and Fujikawa, K. (1985) *J. Biol. Chem.* **260**, 5328–5341
- Kauffman, D. L. (1965) *J. Mol. Biol.* **12**, 929–932
- Bevilacqua, M. P., Stengelin, S., Gimbrone, M. A., Jr., and Seed, B. (1989) *Science* **243**, 1160–1165
- Siegelman, M. H., van de Rijn, M., and Weissman, I. L. (1989) *Science* **243**, 1165–1172
- Johnston, G. I., Cook, R. G., and McEver, R. P. (1989) *Cell* **56**, 1033–1044
- Mürer, E. H., Levin, J., and Holme, R. (1975) *J. Cell. Physiol.* **86**, 533–542
- Perkins, S. J., Haris, P. I., Sim, R. B., and Chapman, D. (1988) *Biochemistry* **27**, 4004–4012
- Pikkarainen, T., Kallunki, T., and Tryggvason, K. (1988) *J. Biol. Chem.* **263**, 6751–6758
- Rees, D. J. G., Jones, I. M., Handford, P. A., Walter, S. J., Esnouf, M. P., Smith, K. J., and Brownlee, G. G. (1988) *EMBO J.* **7**, 2053–2061
- Takeya, H., Kawabata, S., Nakagawa, K., Yamamichi, Y., Miyata, T., Iwanaga, S., Takao, T., and Shimonishi, Y. (1988) *J. Biol. Chem.* **263**, 14868–14877
- Mackinnon, C. M., Carter, P. E., Smyth, S. J., Dunbar, B., and Fothergill, J. E. (1987) *Eur. J. Biochem.* **169**, 547–553
- Spies, M., Schwartz, A. L., and Lodish, H. F. (1985) *J. Biol. Chem.* **260**, 1979–1982

SUPPLEMENTAL MATERIAL

to

Limulus Factor C: AN ENDOTOXIN-SENSITIVE SERINE PROTEASE ZYMOGEN WITH A MOSAIC STRUCTURE OF COMPLEMENT-LIKE, EPIDERMAL GROWTH FACTOR-LIKE, AND LECTIN-LIKE DOMAINS.

by

Tatsushi Muta, Toshiyuki Miyata, Yoshio Misumi, Fuminori Tokunaga, Takanori Nakamura, Yoshihiro Toh, Yukio Ikehara, and Sadaaki Iwanaga

MATERIALS AND METHODS

Preparation of amebocyte cDNA libraries—A total RNA from the amebocytes (*Tachyplesus tridentatus*) was extracted by the method of Chirgwin *et al.* (12). Poly (A)⁺ RNA was then fractionated by oligo dT cellulose (Sigma Chemical Co. St. Louis, MO) column chromatography. λ gt 11 cDNA libraries were constructed from poly A⁺ RNA using cDNA Synthesis System (Amersham-Japan) with EcoRI linker and cDNA Synthesis kit (Pharmacia LKB Biotechnology, Uppsala) with EcoRI adapter.

cDNA cloning—A first screening was carried out by using anti-factor C IgG. Rabbit antiserum raised against factor C was kindly prepared by Dr. H. Igarashi, Tokyo Metropolitan Research Laboratory of Public Health. IgG was purified by Affigel-Protein A (Nippon Bio-Rad laboratories, Tokyo) column chromatography. Goat anti-rabbit IgG horseradish peroxidase conjugate (Nippon Bio-Rad Laboratories) was used as a second antibody and positive clones were developed by 4-chloro-1-naphthol and H₂O₂. The second preparation of libraries constructed with EcoRI adapter was screened with [³²P]-λPC53 insert labelled with multiprime DNA-labelling system (Amersham-Japan). The inserts of positive clones were transferred into pUC118 (Takara Shuzo Co., Kyoto). Subcloning was carried out utilizing restriction sites and deletions with Deletion Kit from Takara Shuzo Co. Nucleotide sequence determination was performed on Applied Biosystems 370 A DNA sequencer using fluorescence labelled nucleotide primer (13). Single stranded DNA templates were prepared using MV1184 as host cell and M13K07 as helper phage (14). All of the sequences were confirmed from both strands with overlapping.

Northern blotting—Five μ g of poly A⁺ RNA was electrophoretically separated on 1.0 % agarose gel, transferred to GVHP membrane (Nippon Millipore Kogyo K.K., Yonezawa). The filter was hybridized with ³²P-labelled probes and washed finally in 0.1 x SSC (0.15 M NaCl, 15

mM sodium citrate), 0.1 % sodium laurylsulfate at 65°C. The washed filter was applied to autoradiography.

Determination of partial amino acid sequence of factor C—Factor C was purified as previously described (4). The purified material was reduced and S-pyridylethylated and the resulting H and L chains were separated as previously described (9). The S-pyridylethylated H chain was digested with lysyl endopeptidase (Wako Pure Chemical Industries, Ltd., Osaka) and the L chain with cyanogen bromide followed by L-(toeylamido 2-phenyl) ethyl chloromethyl ketone treated trypsin (Worthington Biochemical Corp., Freehold, NJ) (9). The resulting peptides were first separated by gel filtration with Sephadex G-50 fine (1.6 x 141 cm) (Pharmacia Fine Chemicals, Uppsala) and then reversed-phase high performance liquid chromatography using a VYDAC 214TP5415 (4.6 x 150 mm) (The Separation Group, CA) or a TSKgel ODS-120T (4.6 x 150 mm) (TOSOH Co. Ltd., Tokyo) column (Fig. M1). The purified peptides were subjected to amino acid analysis on PICO-TAG system (Waters, Millipore Corp., Milford, MA) or Hitachi 835 automatic analyzer and to amino-terminal sequence analysis on Applied Biosystems 477 A Protein Sequencer or Beckman System 890 D Sequencer (Table M1).

Homology search—The *Limulus* factor C protein sequence was compared with all entries in the NBRF Protein Sequence Database (release 23.0, December, 1989) with GENETYX system (Software Development Co., LTD., Tokyo) on the basis of the algorithm of the computer program FASTP (15).

Subcellular Localization—The amebocytes were collected into a sterilized tube and immediately fixed with 4 % paraformaldehyde and 1.2 % glutaraldehyde. The sections mounted on a nickel grid were incubated with factor C IgG after a blocking with 1 % bovine serum albumin. The sections were washed and then incubated with goat anti-rabbit IgG gold colloidal particles (20 nm) (E-Y Labs, Inc., San Mateo, CA). After extensive washing, the sections were doubly stained with 2 % uranyl acetate and lead tartrate and examined under an electron microscope (Hitachi H-500).

Table M1 Amino acid sequences of proteolytic fragments of factor C^a

Cycle No.	Factor C ^b Amino acid (pmol)	8-8 Amino acid (nmol)	7-18 Amino acid (nmol)	5-3 Amino acid (pmol)	5-10 Amino acid (pmol)	7-9 Amino acid (nmol)	8-1 Amino acid (nmol)	5-21 Amino acid (pmol)	5-21-13 ^c Amino acid (pmol)	5-21-21 ^f Amino acid (pmol)
1	Arg 22	Pro 6.57	Ala 1.91	X	X	X	Val 2.40	Val 231	Val 290	X
2	Gly 112	X	Gly 2.65	Gly -	Pro 417	Ile 4.68	Ser -	Asn 251	Asn 82	X
3	Val 150	Asp 1.50	Leu 2.61	Gly 87	Leu 374	Arg 3.01	Ser -	Ala 338	Ala 196	X
4	Asp 12	Asp 2.67	Asp 2.30	Ile 55	Leu 448	Glu 3.61	Pro 2.25	Pro 270	Pro 202	Asp 49
5	Leu 105	Leu 1.95	Ser -	Pro 25	Pro 342	X	Glu 2.18	X	Ser 35	Ser 28
6	Gly 96	Glu 1.72	X	Asp 25	Pro 342	Ala 1.73	His 1.72	Gly 177	Gly 131	Pro 126
7	Leu 105	Ala 2.36	Val 1.33	Gln 41	Ser 179	Lys -	Gly 1.20	Asn 151	Asn 94	Tyr 101
8	ProC ^d -	X	Thr 0.70	Arg 5	Asp 126	X	X	Met 177	Met 135	Tyr 116
9	Asp 40	X	X	Thr 33	Ser 129	X	X	Ile 119	Ile 132	Leu 105
10	Glu 51	Pro 0.88	Gly 36	Gln 137	X	X	X	Glu 110	Glu 80	Ile 101
11	Thr 16	Pro 0.96	Ala 38	Val 128	X	X	X	Gly 104	Gly 112	Gly 104
12	Arg 14	Asn 0.61	ProC ^d -	Gln 155	X	X	X	Ala 119	Ala 101	Gln 54
13	Phe 63	X	His -	Glu 89	X	X	X	Thr 93	Thr 47	Glu 55
14	Glu 52	X	ProC ^d -	Val 114	X	X	X	Leu 101	Leu 77	Thr 43
15	ProC ^d -	X	Arg 5	Arg 32	X	X	X	Arg 48	Arg 43	Leu 62
16	Lys 26	X	Asp 13	Asn 64	X	X	X	Phe 77	X	X
17	ProC ^d -	X	Arg 9	Pro 101	X	X	X	Ser 42	X	X
18	Gly 87	X	Tyr 17	Pro 159	X	X	X	ProC ^d -	Gln 36	X
19	Asp 29	X	Glu 15	Asp 44	X	X	X	Asp 36	Gly 44	X
20	Pro 82	X	Gly 16	Asn 71	X	X	X	Ser 44	Asn 10	X
21	Gly 105	X	Ala 19	Pro 96	X	X	X	Pro 31	Gly 23	X
22	Tyr 44	X	His -	Gln 47	X	X	X	Tyr 27	Gln 10	X
23	Val 83	X	ProC ^d -	Thr -	X	X	X	Tyr 52	X	X
24	Phe 76	X	Glu 10	Ile 21	X	X	X	Leu 32	X	X
25	Asn 49	X	Ile 17	Asp 33	X	X	X	Ile 19	X	X
26	Val 100	X	Leu 13	Tyr 28	X	X	X	Gly 31	X	X
27	Pro 71	X	Lys 5	Ser 21	X	X	X	Gln 17	X	X
28	Met 27	X	X	X	X	X	X	Glu 14	X	X
29	X ^e	X	X	X	X	X	X	Thr -	X	X
30	Gln 30	X	X	X	X	X	X	Leu 16	X	X
31	ProC ^d -	X	X	X	X	X	X	ProC ^d -	X	X
32	X	X	X	X	X	X	X	Gln -	X	X
33	Tyr 41	X	X	X	X	X	X	X	X	X
34	Phe 39	X	X	X	X	X	X	X	X	X
35	X	X	X	X	X	X	X	X	X	X

Positions 1-34 43-49 64-75 90-115 116-143 170-176 177-183 185-217 185-199 200-221

Repetitive Yield(%) 98.4 94.8 92.2 90.3 92.9 91.6
Yield(%) Gly2-Gly18 Ile3-Ile25 Pro3-Pro21 Asn2-Asn7 Ala3-Ala12 Leu9-Leu15

^aThe peptides whose yields are shown in pmol are sequenced by Applied Biosystems 477A Protein Sequencer and those shown in nmol are by Beckman 890 Sequencer.

^bIntact factor C.

^c5-pyridylethylcysteine.

^dNot identified.

^eNot quantitatively determined.

^fPeptides obtained by trypsin digestion of peptide 5-21.

(continued)

Cycle No.	6-15 Amino acid (pmol)	7-30 Amino acid (nmol)	7-17 Amino acid (nmol)	7-6 Amino acid (nmol)	8-21 Amino acid (pmol)	5-19 Amino acid (pmol)	7-41 Amino acid (nmol)	3-12 Amino acid (pmol)	6-9 Amino acid (pmol)
1	Leu 326	Tyr 2.21	X	Val 7.11	Ala 337	X	X	Pro 180	Val 230
2	Val 314	Gly 2.27	Asn 1.06	Ala 3.46	Val 373	Leu 323	Leu 155	Phe 157	Ile 206
3	Phe 312	Gln -	Pro 2.46	Asp 6.47	Asp 470	Ala 295	Ile 315	Val 176	Thr 164
4	ProC ^d -	Phe 2.68	Asp 0.52	Arg 1.10	Phe 268	Arg 132	Pro 325	Trp 36	ProC ^d -
5	Pro 255	Pro 2.79	Gly 1.31	Glu 3.12	Leu 243	Ser 55	Ser 55	Glu 84	Gln 131
6	Asp 74	Gln -	X	Val 3.68	Asp 341	Phe 235	Ser 79	Leu 170	Asn 149
7	Leu 205	Gly 1.94	Trp 0.27	Asp 4.55	Asp 417	Arg 166	Met 115	Pro 179	X
8	Asp 84	Thr 1.52	X	X	Val 130	Phe 188	Asp 82	Pro 178	X
9	Pro 135	Glu 1.39	Gly 0.24	X	Gly 140	Asp 156	Arg 701	Val 111	X
10	Val 141	Val 1.64	X	X	Glu 323	Tyr 128	Ser 18	Pro 146	X
11	Asn 59	X	Gln -	X	Pro 97	Val 135	Asn 53	Ser 39	X
12	His -	Tyr 1.56	X	X	Val 99	Ser 27	Val 81	Tyr 85	X
13	Ala 113	Thr 1.10	X	X	Arg 41	Ser 29	Val 127	Gly 70	X
14	Glu 59	X	X	X	Ile 60	Ser 30	Leu 114	Ser 36	X
15	His -	X	X	X	His 66	Thr 7	Asn 40	Val 60	X
16	Gln 76	Gly 1.10	X	X	ProC ^d -	Ala 50	Asp 42	Glu 42	X
17	Val 65	X	X	X	Pro 57	Gly 52	X	Ile 34	X
18	Lys 33	Tyr 1.19	X	X	Ala 63	Arg 74	Leu -	Lys 39	X
19	X	X	X	X	Gly 61	Ser 7	Thr -	Pro 50	X
20	X	Leu 0.86	X	X	ProC ^d -	Gly 41	Phe -	Pro -	X
21	X	X	X	X	Ser 10	ProC ^d -	Trp -	Ser 19	X
22	X	X	X	X	Leu 33	Pro 29	Ala 41	Arg 9	X
23	X	X	X	X	Thr 14	Asp 25	Ser 6	Thr 14	X
24	X	X	X	X	Ala 29	Gly 26	Gly 38	Asn 7	X
25	X	X	X	X	Gly 35	X	Glu 25	X	X
26	X	X	X	X	X	Phe 14	Pro 30	X	X
27	X	X	X	X	X	Glu 24	Gly 37	X	X
28	X	X	X	X	X	Val 13	X	X	X
29	X	X	X	X	X	Glu 29	Glu 22	X	X
30	X	X	X	X	X	Glu 32	Thr 18	X	X
31	X	X	X	X	X	X	X	X	X
32	X	X	X	X	X	X	X	X	X
33	X	X	X	X	X	X	X	X	X
34	X	X	X	X	X	X	X	X	X
35	X	X	X	X	X	X	X	X	X

Positions 232-249 256-275 283-293 299-305 310-334 391-420 452-457 482-511 612-635

Repetitive Yield(%) 90.1 96.4 65.5 87.7 88.9 87.3 93.5 88.9
Yield(%) Val2-Val17 Tyr1-Tyr18 Gly5-Gly9 Val1-Val6 Ala1-Ala24 Ala3-Ala16 Pro1-Pro26 Ile2-Ile17

(continued)

Cycle No.	L-CN2 ^R		CN2-728 ^B		L-CN3 ^R	
	Amino acid	Yield (pmol)	Amino acid	Yield (pmol)	Amino acid	Yield (pmol)
1	Trp	29	Tyr	166	Ile	379
2	Phe	273	Tyr	110	Gln	515
3	Leu	246	Arg	264	Glc	480
4	Gln	169	Asp	136	Ala	553
5	PeC	-	Asp	151	Val	370
6	Gly	128	Ser	43	Leu	376
7	Gly	128	Arg	205	Pro	411
8	Ser	30	Asp	92	Val	369
9	Leu	95	Asp	146	Val	424
10	Leu	131	Asp	141	Ala	423
11			Tyr	78	Ala	536
12			Val	68	Ser	117
13			Gln	78	Thr	95
14			Val	61	PeC	-
15			Arg	172	Glu	118
16			Glu	37	Glu	130
17			Ala	63	Gly	150
18			Leu	49	Tyr	133
19			Glu	47	X	-
20			Ile	39	Glu	-
21			His	18	Ala	-
22			Val	31	Asp	-
23			Asn	27	Leu	-
24			Pro	30	Pro	-
25			Asn	34	Leu	-
26			Tyr	23		
27			Asp	27		
28			Pro	19		
29			Gly	18		
30			Asn	16		
31			Leu	12		
32						
33						
34						
35						

Positions 765-774 807-837 894-918

Repetitive 85.3 92.4 95.6
Yield (%) Leu3-leu9 Tyr1-Tyr26 Ala4-Ala10

^aPeptides obtained by CNBr digestion of S-pyridylethylated L-chain (Tokunaga et al., 1987)
^bPeptide obtained by trypsin digestion of peptide L-CN2.

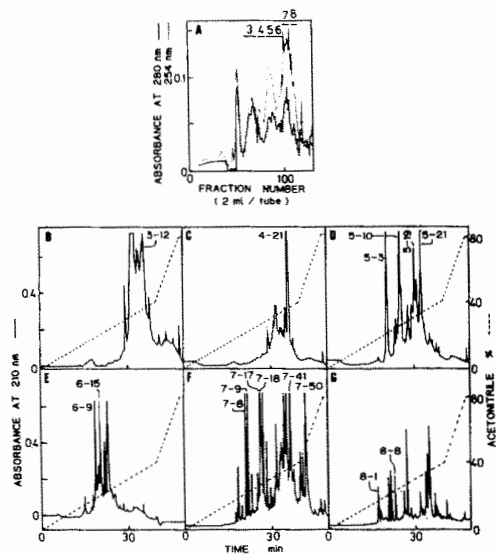


FIG. M1. Separation of proteolytic fragments from factor C. (A) Gel filtration on a Sephadex G-50 fine column. S-pyridylethylated H chain was digested with lysyl endopeptidase and the digest was separated on the column (1.6 x 141 cm) in 9% formic acid at the flow rate of 20 ml/h. The solid line shows absorbance at 280 nm and the dotted line shows the absorbance at 254 nm. (B)-(G) Each peak corresponding to a solid bar in Fig. 4A was collected and further purified by reversed phase HPLC employing the indicated gradients of acetonitrile in 0.1% trifluoroacetic acid. The columns used were Vydac 214TP5415 (4.6 x 150 mm) in (B) and TSKgel ODS-120T (4.6 x 150 mm) in (C)-(G). (B) peak 3, (C) peak 4 (D) peak 5, (E) peak 6, (F) peak 7, (G) peak 8. The solid lines show the absorbance at 210 nm, and the dotted lines show the gradients of acetonitrile in 0.1% trifluoroacetic acid. The flow rate was 1.0 ml/min. The indicated peptides were sequenced.

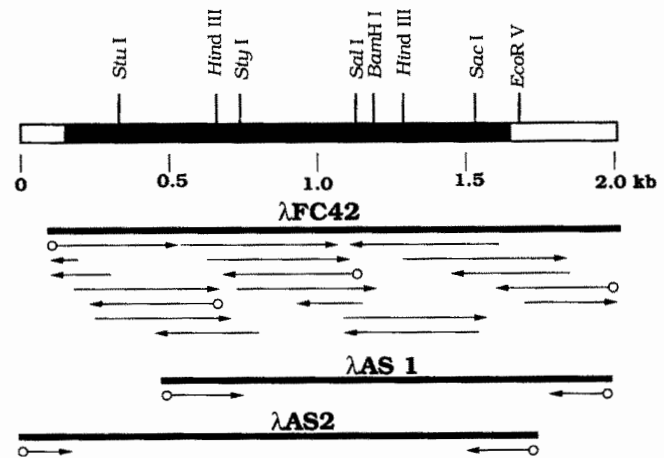


FIG. M2. Restriction map and sequence strategy of the tentative alternative splicing product of factor C. The black and shaded box at the top shows the open reading frame. The shaded box indicates the region encoding the different amino acid sequence from factor C. Arrows with circles show sequencing utilizing restriction sites and arrows without circle were by deletions.