Volume 309, number 3, 219–224 FEBS 11449 September 1992

Molecular cloning of a cDNA that encodes a serine protease with chymotryptic and collagenolytic activities in the hepatopancreas of the shrimp *Penaeus vanameii* (Crustacea, Decapoda)

Daniel Sellos and Alain Van Wormhoudt

Laboratoire de Biologie Marine, Collège de France, BP 225, 29182 Concarneau cedex, France

Received 15 May 1992; revised version received 3 August 1992

Two clones were isolated by screening a shrimp hepatopancreas cDNA library with a DNA fragment obtained by PCR amplification using two oligonucleotides based on the partial protein sequence of *Penaeus vannamei* chymotrypsin purified earlier. One of these clones, PVC 7 contains a complete cDNA coding for a serine protease. The deduced amino acid sequence shows the existence of a 270 residue-long preproenzyme containing a highly hydrophobic signal peptide of 14 amino acids. This suggests the existence of a putative zymogen form of the enzyme containing a 30 amino acid-long peptide which is cleaved to give a mature protein of 226 residues. A highly preferred codon usage is observed for this protein. The other obtained cDNA was found to encode the less predominant variant of the protein. Sequence alignments show that shrimp chymotrypsin is highly homologous with crab collagenase (77% homology taking into account the same amino acid at the same position, and 83% homology taking into account amino acids with conserved function) and that it is more similar to mouse trypsin (41% homology of strictly conserved amino acids) than to hornet chymotrypsin (35% homology).

Chymotrypsin; Collagenase; cDNA nucleotide sequence; Invertebrate; Crustacea

1. INTRODUCTION

Chymotrypsin isolated from the hepatopancreas of the shrimp *Penaeus vannamei*, is a serine protease homologous to vertebrate chymotrypsin [1] and to the pancreatic serine protease family. However, in common with Crustacean chymotrypsins [1,2], it possesses the unique property of being able to cleave collagenolytic substrates and native collagen.

The complete sequence of this enzyme is known for the crab [3], and the N-terminal primary structure has been determined in one penaeide species [4]. However, nothing is known concerning putative precursors or maturation proteins.

Comparison of complete mRNA sequences can be used to determine the evolutionary relationships between enzymes belonging to the serine protease family. The deduced amino acid sequences will also provide more information on the existence and nature of a preproprotein. Recently, in lobster, mRNAs coding for three digestive cathepsins have been sequenced [5], and the results of this study suggest the existence of a proenzyme, thereby confirming the earlier hypotheses of Maugle et al. [6], referring to the increase of total protease activity in vitro, due to the action of enterokinases, and of Al-Mohanna et al. [7] based on histological studies of the Penaeides. The problem of whether zymogens also exist for other digestive enzymes, including trypsin [8], the major protease in Penaeides, and chymotrypsin has not been solved.

In this paper, we report the molecular cloning of two different chymotrypsin cDNAs, the determination of the structure of their encoded proteins and an analysis of the chymotrypsin family. This will give information about the early evolution of this family of proteases which also possess collagenolytic activities.

Correspondence address: D. Sellos, Laboratoire de Biologie Marine, Collège de France, BP 225, 29182 Concarneau cedex, France. Fax: (33) 98 97 81 24.

EMBL sequence accession number: X66415

2. MATERIALS AND METHODS

Shrimps (*Penaeus vannamei*) were obtained from IFREMER (Brest). The hepatopancreas was removed by dissection, immediately frozen in liquid nitrogen and stored at −80°C until used.

2.1. *RNA isolation*

Total RNA extractions were made following the guanidine thiocyanate method [9]. Frozen hepatopancreases were disrupted in liquid nitrogen with a grinder. The powdered tissue was immediatly dissolved in a 4 M guanidine thiocyanate solution in 12.5 mM Tris-HCl, pH 7.6, with additional 12.5 mM EDTA and 0.1 M mercaptoethanol (5 ml per g of tissue). After complete dissolution, the solution was centrifuged at 12,000 × *g* for 10 min and one tenth of the volume of 20% Sarkosyl was added to the supernatant. After heating the solution to 65°C for 2 min, 0.1 g of caesium chloride was added per ml of solution. This was then layered onto a cushion of 5.7 M caesium chloride in 0.1 M EDTA, pH 8, and centrifuged at 20,000 rpm for 22 h at 20°C in a Beckman centrifuge with a SW 25 rotor. The obtained pellet was dissolved overnight at 4°C in a solution of 5 mM EDTA, pH 8, with additional 5% Sarkosyl and 5% mercaptoethanol (1 ml/g).

This solution was then extracted with phenol/chloroform followed by chloroform and then precipitated in the presence of 0.3 M sodium acetate and ethanol (3 vols.). Poly(A) RNAs were selected using the usual method of chromatography on an oligo dT-cellulose column.

2.2. *PCR amplification*

Two oligonucleotides were synthesized, based on partial amino acid sequencing of the purified protein [1], and used with RNA from the shrimp hepatopancreas for PCR amplification.

The first oligonucleotide was based on the first nine residues of the N-terminal amino acid sequence of the protein. The second oligonucleotide was based on ten residues from the end of the amino acid sequence of a selected tryptic peptide of the protein.

Hepatopancreas total RNA (10 mg in 10 μl) and hepatopancreas poly(A)$^+$ RNA (200 ng in 10 μl) were heat denatured (70°C for 3 min), then cooled in ice. In a final volume of 20 μl of 1 × *Taq* buffer (Promega), the following were assembled: denatured RNA, 1 mM of each of the four dNTPs, 1 U per μl of RNasin, 100 ng of oligo-dT 12–18 and 200 U of BRL MuLV reverse transcriptase. The incubation time was 10 min at room temperature followed by 30 min at 42°C.

To the 20 μl of reverse transcription reaction, 80 μl of 1 × PCR buffer containing 50 pmol each of upstream and downstream primers and 4.5 U of Taq DNA polymerase (Promega) were added. After overlaying 100 μl of mineral oil on top of the solution, 32 PCR cycles were run as follows. After a first step of denaturation at 94°C for 4 min, the first PCR cycle was run with a low annealing temperature of 37°C for 1.5 min, then extension was conducted at 72°C for 3 min and denaturation at 94°C for 1.5 min. The second cycle was run with the same conditions except that the annealing temperature was 45°C. The thirty following cycles were run in the same conditions except that the annealing temperature was raised to 55°C.

Amplified products were analysed on 1.5% agarose gels with *Pst*I-digested lambda DNA as size markers.

2.3. *cDNA library*

A lambda ZAP cDNA library for shrimp hepatopancreas was established following the Stratagene protocol. The unamplified cDNA library, containing 5.8×10^6 independent phages was screened with the 520 bp amplified cloned fragment. Plaques were transferred to Hybond-N membranes (Amersham) and screened with the cloned probe labelled with the random priming kit from Biolabs using [^{32}P]dATP. Prehybridization of the duplicate membranes was achieved in a 50% formamide solution containing 1% SDS, 1 M NaCl and 100 μg yeast RNA per ml for 4 h at 42°C. For hybridization, the denatured probe was added (10^6 cpm/ml) for 16 h at the same temperature. The filters were washed twice in 2 × SSC for 5 min at room temperature, then twice in 2 × SSC containing 1% SDS at 65°C for 30 min, and then twice in 0.1 × SSC for 30 min at room temperature. The membranes were autoradiographed for two days using Hyperfilm-MP (Amersham) with an enhancer screen.

2.4. *Plasmid subcloning and DNA sequencing*

The recombinant clones which hybridized with the 520 bp cloned fragment were selected and isolated with successive cycles of purification. The phagemids contained in the selected phages were excised following the manufacturers protocol, digested with *Eco*RI and *Xho*I to determine the size of the inserts. Nine clones were selected and sequenced. The complete cDNA insert (clone PVC 7) and the insert that encodes a variant of the main chymotrypsin (clone PVC 5) extracted from low-melting-agarose gels and purified with the Gene Clean kit (Bio-Rad) were subcloned in Bluescript SK$^-$ (Stratagene) in order to allow sequencing on both strands. To confirm the nucleotide order at the end of the anti-sense strand, sequencing was performed using a synthetic oligonucleotide (probe no. 3) hybridizing to the region 400–383 and with PCR oligonucleotide no. 2. Single-stranded DNA was produced using M 13 KO7 helper phage and sequenced following the dideoxy nucleotide method with modification for extended DNA sequencing with the large fragment of DNA polymerase I. Electrophoresis of the extended products was performed with two or three successive loadings on a 5% acrylamide/bis-acrylamide (30:0.8) gels.

3. RESULTS AND DISCUSSION

3.1. *PCR amplification*

The first degenerate oligonucleotide used for PCR amplification was composed of 26 bases (ATC GTG/C GGI GGI GTG/C GAA/G GCT/C ACC CC) based on the N-terminal amino acid sequence of the purified protein [1]: Ile-Val-Gly-Gly-Val-Glu-Ala-Thr-Pro. Inosine and limited degeneracy were introduced to limit the number of oligonucleotides in the mixture. The second oligonucleotide was composed of 29 bases (CCI CCG GTI CCA/G TCG ATA/G CAC/G ACC/G ACC/G

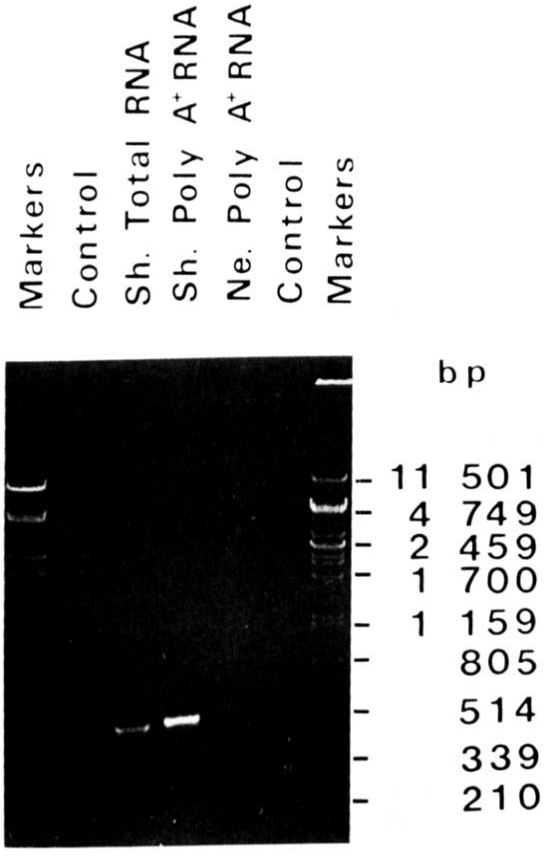

Fig. 1. Electrophoresis of *Eco*RI-digested PCR products on a 1.5% agarose gel. Size markers are *Pst*I-digested lambda DNA fragments. The control consists of all the different components minus mRNA. Heterologous mollusc digestive gland mRNA (Ne. Poly A$^+$ RNA) was used for comparison. 15 μl of amplification products from a total of 60 were loaded on the gel. The sense primer (oligonucleotide no. 1) was composed of 26 bases (ATC GTG/C GGI GGI ATA/C GGA/G GCT/C ACC CC) based on the amino acid sequence of the N-terminal part of the purified protein. The anti-sense primer (oligonucleotide no. 2) was 29 bases long (CC ICC GGT ICC A/GTC GAT A/GCA C/GAC C/GAC C/GCC) based on a part of the amino acid sequence of a peptide obtained after tryptic hydrolysis of the purified protein. 14 bases were added at the 5' end of these two nucleotides to generate protected *Sal*I and *Eco*RI restriction sites. PCR amplification was run with increasing annealing temperatures from 37°C for the first cycle to 45°C for the second, and 55°C for the thirty following cycles.

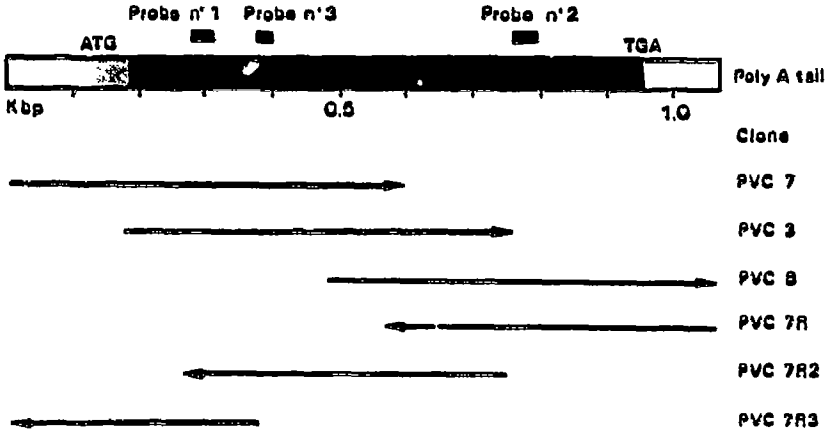

Fig. 2. Map of the shrimp chymotrypsin cDNA and sequencing strategy. Localization of the hybridizing site for the different oligonucleotide probes is indicated. Open and shaded bars represent non-coding and coding regions, respectively. The different clones used for the sequencing are indicated. PVC 7R is the insert from PVC 7, subcloned in Bluescript SK⁺ and initiated with the reverse primer. PVC 7R2 is the sequence determined from PVC 7R DNA and initiated with probe no. 2. PVC 7R3 is the sequence obtained with PVC 7R and initiated with probe no. 3.

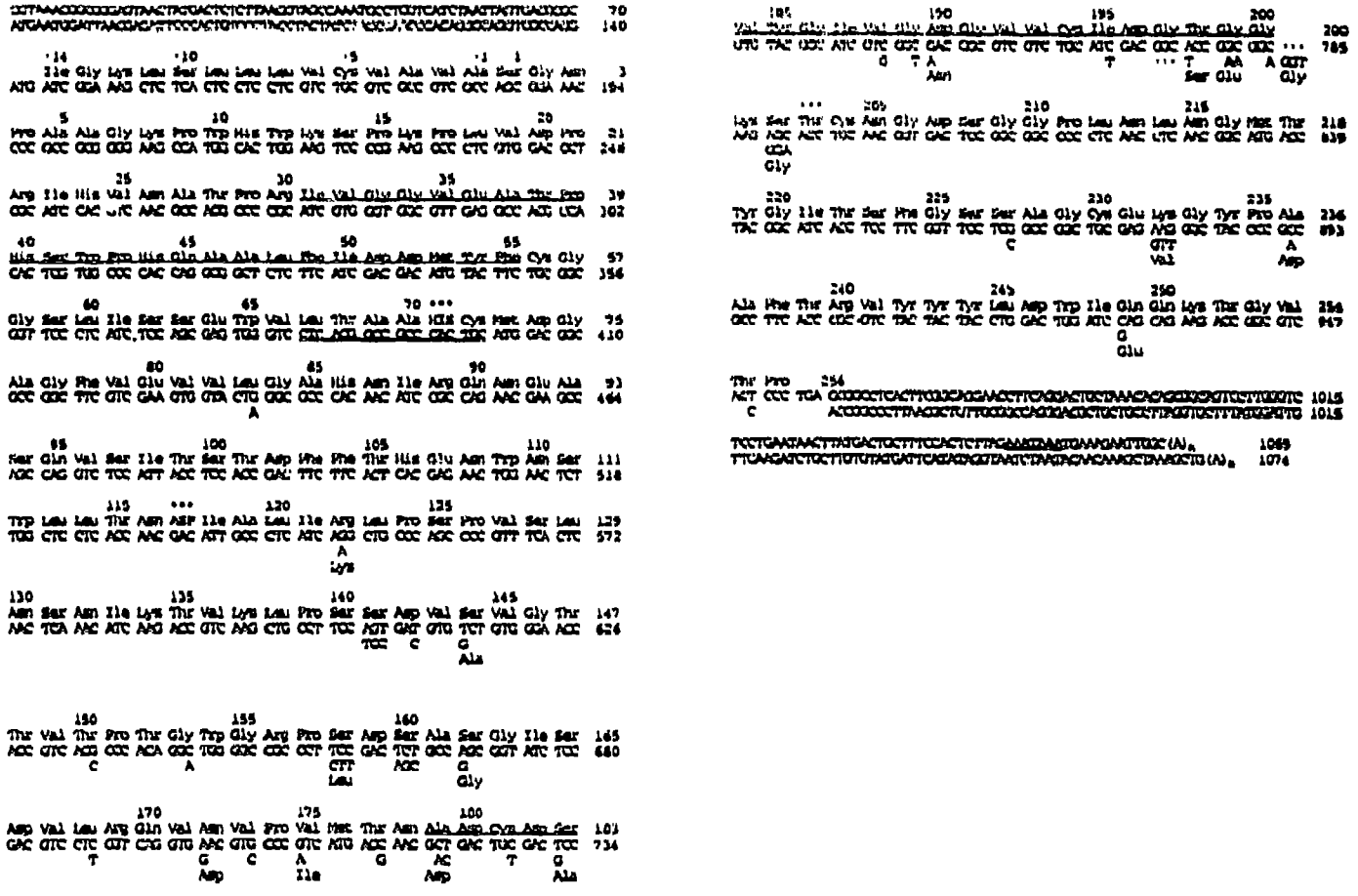

Fig. 3. Nucleotide sequence of shrimp hepatopancreas chymotrypsin cDNA and deduced amino acid sequence. The amino acid sequence is numbered sequentially from the first amino acid of the putative zymogen. The presumed activation peptide comprises amino acids 1 to 30, and the proposed signal peptide is represented by amino acids −14 to −1. The underlined residues were shown to be identical with those determined by partial amino acid sequencing. The amino acids that comprise the charge relay system (His-57, Asp-102 and Ser-195) by the chymotrypsin numbering system [19] are indicated by capital letters and asterisks (residues His-71, Asp-117 and Thr-203, respectively). The proposed polyadenylation site and the site for the hybridization of sequencing probe no. 3 for the anti-sense strand (nucleotides 384-401) are underlined. The nucleotide changes observed in clone PVC 5 are shown under the sequence. When this nucleotide modification gives way to an amino acid change, the residue is indicated. The nucleotide fragment obtained by PCR amplification is located between bases 275 and 786.

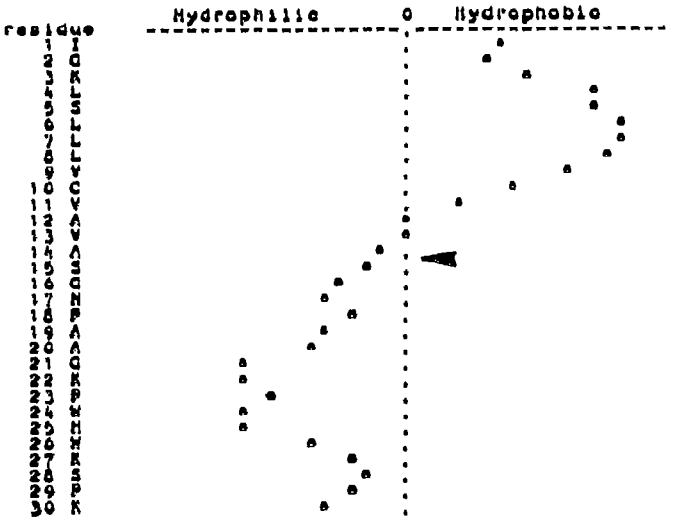

Fig. 4. Hydropathy plot of the N-terminal segment of the prechymotrypsinogen. The dotted vertical line indicates hydropathic neutrality. Hydrophobic scores appear to the right, hydrophilic scores to the left. The site of a putative proteolytic cleavage of the transport peptide is indicated by the arrowhead.

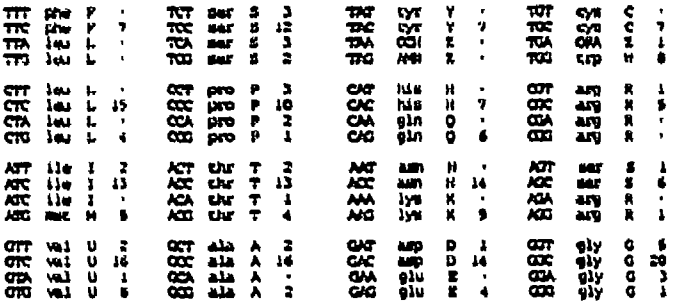

Fig. 5. Codon usage table for the shrimp preprochymotrypsin cDNA. The same result is obtained using only the sequence corresponding to the mature protein. The starting methionine and stop codons were included.

CC) based on a part of the amino acid sequence of a peptide obtained after tryptic hydrolysis of the protein: Gly-Val-Val-Cys-Ile-Asp-Gly-Thr-Gly-Gly. Bases generating *Sal*I and *Eco*RI restriction sites were added at the 5′ end of these two oligonucleotides to facilitate subsequent cloning. After 32 cycles of amplification, a strong band of DNA of around 520 bpa was specifically obtained for shrimp hepatopancreas total and poly(A) RNAs (Fig. 1). In contrast a weak band of around 300 nucleotides in length was observed using *Nephrops* (Crustacea Decapoda) poly(A) RNA and nothing was generated for the control.

After extraction, the DNA fragment was digested with *Eco*RI and cloned in the *Eco*RI restriction site of pBluescript SK+ (Stratagene). Eight clones (CP1–CP8) were sequenced and all of them appeared to contain the same insert with a few differences (Fig. 3, framed sequence). Two of these differences relate to the replacement of a single amino acid by another in the translated protein (Asp or Ala in position 172 and Asp, Ala or His in position 183; see Fig. 3). These two locations were effectively found with different amino acids in the deduced sequences corresponding to the two variants of the shrimp chymotrypsin. The amplified sequence was 510 bp long (without counting the nucleotides forming the *Eco*RI sites). The cloned fragment, CP8, encodes for the first 170 amino acids of the shrimp chymotrypsin and was used to probe Northern blots of shrimp RNAs. The hybridizing RNA band was specifically found in the hepatopancreas and the observed size was around 1,000 bases (data not shown).

3.2. *cDNA library screening*

The insert extracted from the clone, CP8, was also used to screen a shrimp hepatopancreas lambda ZAP cDNA library. With the highly stringent conditions used (50% formamide, 42°C), around 300 clones (on a total of 5×10^4) were found to be positive. Twelve of them were selected and eleven appeared to contain an insert. Nine clones were sequenced and all of them contained the whole or a part of the chymotrypsin cDNA (Fig. 2). The complete cDNA (clone PVC 7) is 1,069 bp (without the poly A tail). The complete nucleotide sequence was determined by overlapping the three clones. The complementary strand was determined on the insert subcloned in Bluescript SK⁺ (PVC 7R) using three different primers.The coding sequence is 810 bp long (Fig. 3). The deduced protein sequence is 270 residues in length (without the starting methionine). This protein is composed of a highly hydrophobic peptide covering residues −14 to −1. On a hydropathy plot of the first 30 residues of the N-terminal segment of the protein, as described by Kyte and Doolittle [10], the hydrophobic character of the signal peptide is followed by the hydrophilic character of the downstream peptide signal (Fig. 4).

The first 14 residues must be contained in the signal peptide and the remaining 30 residues may be present in the amino terminus of the zymogen. The proteolytic cleavage site must be located near the point where the hydrophobic index drops dramatically and enters the hydrophilic range. A striking homology is observed with trypsinogen 1 in the amino acid sequence [11] of the signal peptide and in the location of the cleavage site [12]. The protein should be processed as a zymogen and contains a peptide of 30 residues (numbered 1 to 30) which is cleaved to give the active enzyme that is formed of 226 amino acids (numbered 31 to 256). One of the nine sequenced clones (PVC 5) shows at least 46 changes in the nucleotide alignment, the deletion of the three bases coding for glycine in position 197 and the addition of a codon, GGT, for a glycine in position 200 in the variant. The 3′ untranslated domains of the two cDNAs are highly different and no attempt to align them was made. These 46 modifications give way to 15 replace-

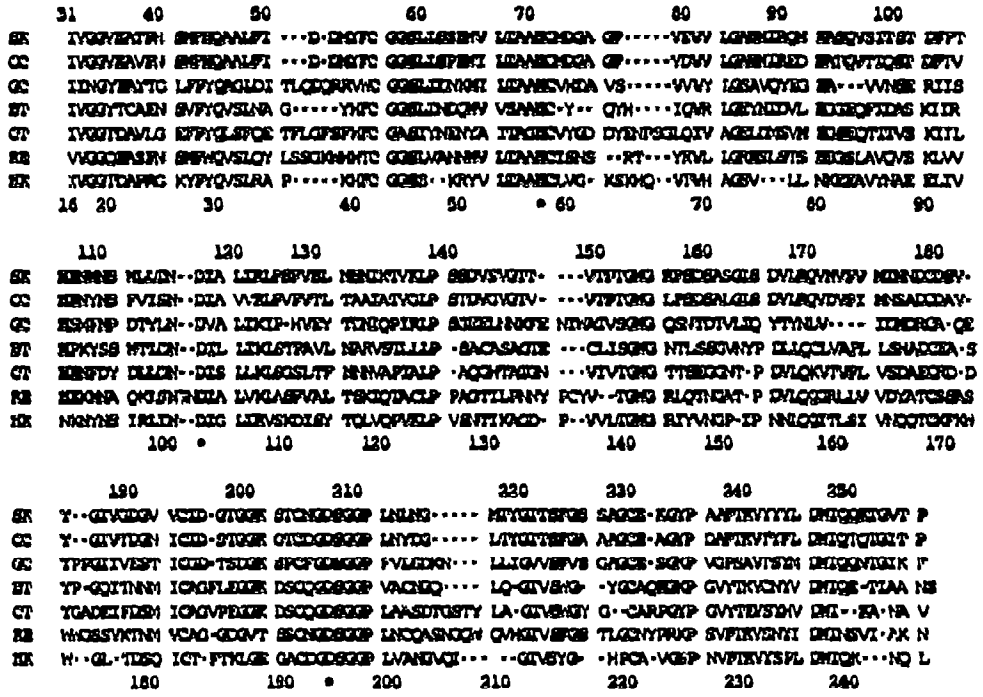

Fig. 6. Amino acid sequence alignments of collagenolytic proteinase, trypsin, elastase and chymotrypsin. Amino acid sequences are shown for shrimp chymotrypsin (SK), UCA crab collagenolytic proteinase (CC, [3]), cattle grub collagenase (GC, [14]), bovine trypsin (BT, [16]), rat elastase (RE, [17]), crayfish trypsin (CT, [8]) and hornet chymotrypsin (HK, [15]). Amino acid residues which are shared between shrimp chymotrypsin and at least one other enzyme are shown with bold letters. Gaps have been introduced in the sequences to facilitate aligments. The numbering above the alignments relates to the shrimp putative chymotrypsinogen as used in Fig. 3. The numbering under the alignments refers to the bovine chymotrypsinogen numbering system [8,19].

ments in the amino acid sequence. These differences were already observed in the amino acid compositions obtained for the two variants of the chymotrypsin [1]. We could deduce that clone PVC 7 encodes the variant, BI, and clone PVC 5 encodes the variant, BII. Globally, the BII is enriched in acidic residues and is slightly more hydrophobic. The interesting feature is the replacement of the serine residues with hydrophobic residues in the C-terminal domain of BII. A highly preferred codon usage is found for the preprotein as well as for the mature enzyme (Fig. 5). Most of the amino acids with two codon possibilities are found to be only coded by one specific triplet. Even in the case of amino acids with six codon possibilities, one is highly preferred: CTC for leucine, CGC for arginine. This highly oriented coding preference is not observed with other kinds of proteins from crustaceans ([8,13], and D. Sellos, unpublished results).

3.3. *Sequence alignment*

Alignment of the matured protein was established with a series of serine proteases: collagenolytic proteinase from the crab [3], insect collagenase and chymotrypsin [14,15], bovine [16] and crayfish trypsin [8] and rat elastase [17] (Fig. 6). On searching the data bank [18] these proteins were found to contain the best scores of homology with shrimp chymotrypsin. To facilitate comparisons the bovine chymotrypsinogen numbering system is used. For each of these six proteins, all the cysteines are found at the same location, and this shows the importance of the secondary structure in the enzymatic activities of this family of proteases. Several domains (N-terminal sequence, segments 41–45, 51–58, 101–109, 139–143, 193–201 (chymotrypsinogen numbering system)) and the end of the protein show high levels of conservation. The shrimp chymotrypsin is highly homologous to the crab proteinase which was defined as a collagenase: 171 residues are found at the same location on a total of 226 amino acids (76%), 184 on a total of 226 (81%) if we take into consideration the conserved function of the amino acid.

This shows the difficulty of classifying the different enzymes in the serine protease family. For instance, invertebrate serine proteases are active against a wide range of different substrates and cannot be characterized on these alone. They have also to be characterized by amino acid and nucleotide sequence alignments.

Acknowledgments: This research was partialy supported by the 'Conseil Régional de Bretagne, France'. We are grateful to Professor B. Leadbeater (University of Birmingham) for the improvment of the English.

REFERENCES

[1] Van Wormhoudt, A., Le Chevalier, P. and Sellos, D. (1992) Comp. Biochem. Physiol. (in press).
[2] Tsai, I.H., Chuang, K.L. and Chuang, J.L. (1986) Comp. Biochem. Physiol. 85B, 235–239.
[3] Grant, G.A., Henderson, K.O., Eisen, A.Z. and Bradshaw, R.A. (1980) Biochemistry 19, 4653–4659.
[4] Tsai, I.H., Lu, P.J. and Chuang, J.L. (1991) Bioch. Biophys. Acta 1080, 59–67.
[5] Laycock, M.V., Mackaye, R.M., Di Fruscio, M. and Gallant, J.W. (1991) FEBS Lett. 292, 115–120.
[6] Maugle, P.D., Deshimaru, O., Katayama, T. Nagatani, T. and Simpson, K.L. (1983) Bull. Jpn. Soc. Sci. Fish 49, 1421–1427.
[7] Al-Mohanna, S.Y., Nott, J.A. and Lane, D.J.W. (1985) J. Mar. Biol. Ass. UK 65, 901–910.
[8] Titani, K., Sasagawa, T., Woodbury, R.G., Ericsson, L.H., Dorsam, H., Kraemer, M., Neurath, H. and Zwilling, R. (1983) Biochemistry 22, 1459–1465.
[9] Chirgwin, J.J., Przbyla, A.E., MacDonald, R.J. and Rutter, W.J. (1979) Biochemistry 18, 5294.
[10] Kyte, J. and Doolittle, R.F. (1982) J. Mol. Biol. 157, 105–132.
[11] MacDonald, R., Stary, S.J. and Swift, G.H. (1982) J. Biol. Chem. 257, 9724–9732.
[12] Scheele, G. (1986) in: Cellular Processing of Proteins in the Exocrine Pancreas: Biology, Pathology and Diseases (V.L.W. Go et al. eds.) Raven Press, New York.
[13] Aota, S.I., Gojobori, T., Ishibashi, F., Maruyama, T. and Ikemura, T. (1988) Nucleic Acids Res. 16, 315–402.
[14] Le Croissy, A., Gilles, A.M., De Wolf, A. and Keil, B. (1987) J. Biol. Chem. 262, 7546–7551.
[15] Jany, K.D. and Hang, H. (1983) FEBS Lett. 158, 98–102.
[16] Le, H., Wicker, C., Guilloteau, P., Toullec, R. and Puigserver, A. (1990) Eur. J. Biochem. 193, 767–773.
[17] MacDonald, R.J., Swift, G.H., Quinto, C., Swain, W., Pictet, R.L., Nikovits, W. and Rutter, W.J. (1982) Biochemistry 21, 1453–1463.
[18] Dessen, P., Fondrat, C., Valencien, C. and Mugnier, C. (1990) Bisance: French service for access to biomolecular sequence databases, Comp. Appl. Biosci. 6, 355–356.
[19] Hartley, B.S. (1970) Phil. Trans. R. Soc. London, B 257, 77–87.