CHARACTERIZATION OF A GENE FAMILY ASSOCIATED WITH CALCIFIED STRUCTURES IN THE BLUE CRAB, *CALLINECTES SAPIDUS*

Patrick J. Kennedy

A Thesis Submitted to the
University of North Carolina at Wilmington in Partial Fulfillment
Of the Requirements for the Degree of
Master of Science

Department of Biological Sciences

University of North Carolina at Wilmington

2004

Approved by

Advisory Committee

Dr. Michael McCartney	Dr. Ann Stapleton

Dr. Thomas Shafer, Chair

Accepted by

Dean, Graduate School

This thesis has been prepared in the style and format

Consistent with the journal

Journal of Experimental Biology

TABLE OF CONTENTS

ABSTRACT

Eleven cDNAs from a family of genes were cloned from the calcified exoskeleton of the decapod crustacean *Callinectes sapidus*. Multiple, variant copies of a conserved 18-residue motif (xLxGPSGφφxxDGxxxQφ), unique to calcified crustacean exoskeleton, accounts for ~70% of the total amino acid residues. The proteins appear to be post-translationally cleaved by a trypsin-like serine protease at conserved recognition sites (RxKR). Two to six peptides, each containing either two or four copies of the 18-residue motif, are expected, depending on which pro-protein is cleaved.

Expression of the *CsproCP* gene family begins at the onset of calcification in the hypodermis of post-ecdysial, calcified cuticle, as shown by Northern analysis. The genes are never expressed in the hypodermis of the noncalcified arthrodial membrane. Western analysis, using an antibody against the 18-residue motif, shows that accumulation of peptides with this motif begins in the calcified cuticle several hours post-ecdysis and continues to anecdysis. The size of the detected peptides agrees with the presumed post-translational cleavage. The strong antibody binding to calcified cuticle proteins and the lack of binding to arthrodial membrane proteins from anecdysial crabs is consistent with immunohistochemical staining performed by Hequembourg (2002). Interestingly, the antibody also weakly binds to proteins from the tendon, another calcified structure in the crab. These results confirm that the proteins encoded by the *CsproCP* gene family are associated with calcification in *Callinectes sapidus*.

ACKNOWLEDGEMENTS

I thank Dr. Shafer for his guidance and for the opportunity to work on a very interesting phenomenon found in many biological organisms. I thank Dr. McCartney for his laboratory use and his patience in teaching me some complicated techniques. I thank Dr. Stapleton for answering many technical questions, but more importantly leading me in directions that allowed me to answer many of the questions on my own. I thank Dr. Coblentz for letting me bounce ideas off of her and the engaging conversations that have resulted. I thank my fellow lab mates, Anna Wynn and Liz Buda, for their help around lab and specifically for optimizing some protocols that have been essential to my research. I thank Mark Gay, the department's problem solver, for his technical guidance and for making me smile when least expected. I thank Dr. Roer for the use his lab and office space. This research was made possible by Murray Bridges at Endurance Seafood in Kill Devil Hills, NC, and support for this research was provided by the National Science Foundation.

I would like thank all of the faculty and staff in the Biological Sciences department at UNCW, my fellow graduate students, and the friends that I have made here in Wilmington, NC. Special thanks goes to Lisa Johnson for love and support over the last year. Lastly, I would like to thank my family and my parents, Tom Kennedy and Ruth Stoltz, for instilling the importance of education in my life. Always remember, "You belong to the Earth, the Earth does not belong to you!"

ENJOY!

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Numerous organisms utilize minerals in support, protection, mastication, and gravity perception (Lowenstam and Weiner, 1989). The mineral calcium carbonate ($CaCO_3$) exists in the amorphous form or in the crystal forms calcite, aragonite, and vaterite and is found in the skeletal structures of many invertebrates. In these structures, matrix macromolecules provide the proper environment for calcification by controlling the nucleation, orientation, and cessation of crystal growth. The soluble matrix is frequently composed of anionic proteins and glycoproteins, whereas a chitin-protein complex makes up the majority of the insoluble matrix (Mann, et al., 1989).

The crustacean exoskeleton is an excellent system to study the regulation of calcification (reviewed in Roer and Dillaman, 1984, 1993). Crustaceans grow by periodically shedding (molting) their old exoskeleton and replacing it with a new, slightly larger one. The molting process begins with enzymatic separation of the cuticle-secreting hypodermis from the exoskeleton and partial resorption of Ca^{2+} and CO_3^{2-}. Ecdysis marks the time when the animal emerges from its old exoskeleton. The outermost layers of the new exoskeleton, the epicuticle and exocuticle, are deposited before ecdysis but do not calcify until post-ecdysis. Secretion of the endocuticle, proximal to the exocuticle, begins shortly after the initial formation of the mineral in the pre-ecdysial layers. The endocuticle is immediately calcified as it is deposited and becomes the thickest and most heavily calcified layer. The molting process ends with the deposition of the innermost membranous layer. Anecdysis, when no more cuticle is produced, is the stage at which the animal spends the majority of its life.

In addition to temporal control, cuticle calcification is also spatially controlled since both calcified and noncalcified regions exist in the crustacean exoskeleton. The calcified cuticle covers most of the animal; however, to allow for movement of the appendages, the joints are covered by flexible, noncalcified cuticle, called the arthrodial membrane. The timing of arthrodial membrane deposition is similar to calcified cuticle (Williams et al., 2003), but is divided into two layers, the outer, lipid-rich epicuticle and the inner, chitinous procuticle (Neville, 1975).

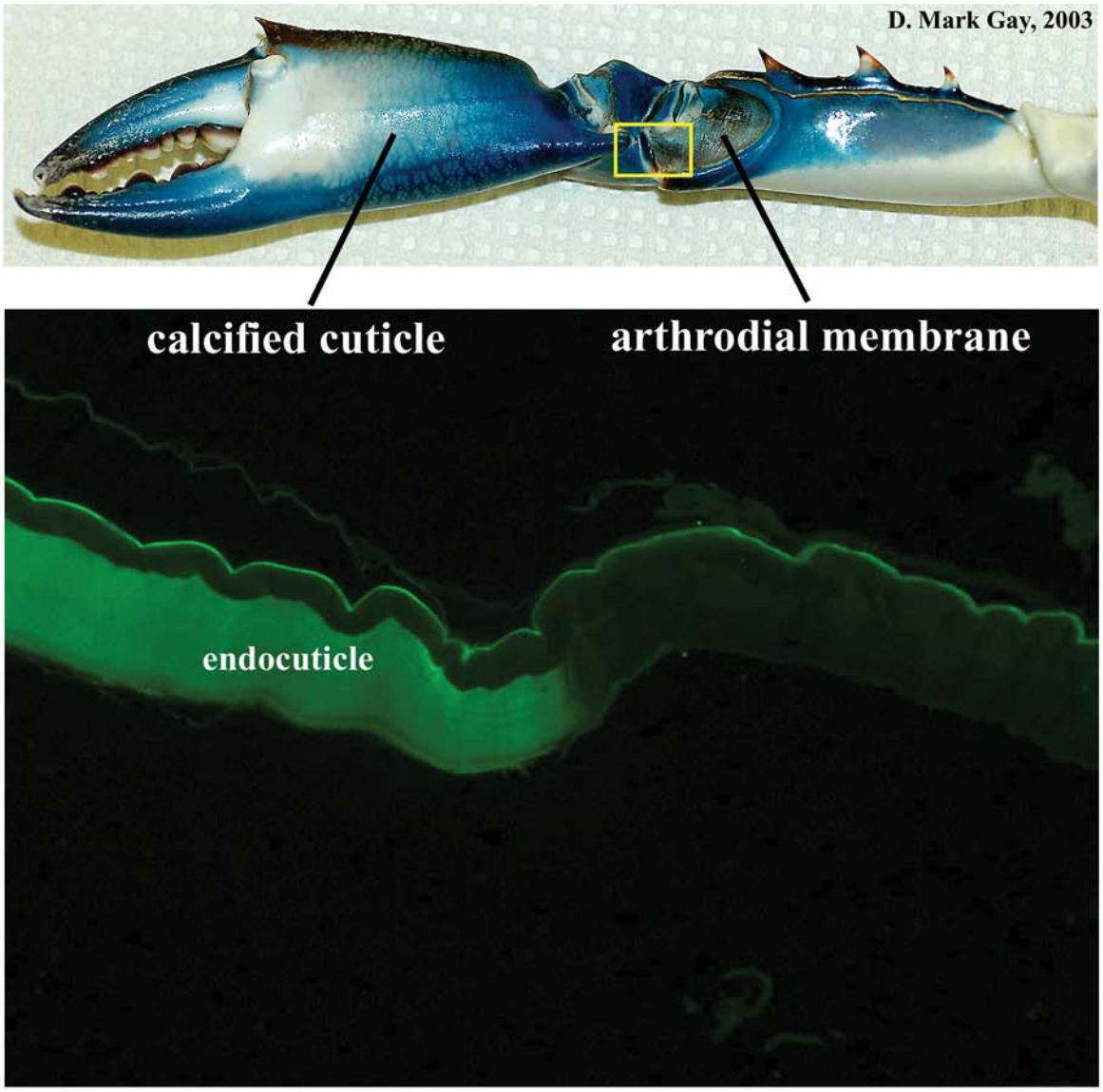
Roer et al. (1988) determined that the critically important timing of calcification is controlled by the cuticle and not the environment. Studies on the blue crab, *Callinectes sapidus*, exoskeleton have shown that initial mineral deposition coincides with significant changes in the biochemistry of the cuticle 1 to 3 h post-ecdysis. Specifically, the loss of two glycoproteins (Shafer et al., 1995) coincides with increased glycosidase activity (Roer et al., 2001). This is consistent with a model proposed by Coblentz et al. (1998) where a large shielding protein is enzymatically degraded or deglycosylated to expose crystal nucleation sites.

Limited information exists on the primary structure of matrix-associated proteins associated with calcification of the crustacean exoskeleton. Calcification-associated peptide (CAP)-1 from *Procambarus clarkii* has anti-calcification activity and chitin-binding ability (Inoue et al., 2001), and a recombinant protein based on the cDNA DD4 (crustocalcin) from *Marsupenaeus japonicus* binds Ca^{2+} (Endo et al., 2000). These two similar proteins are expressed only in the post-ecdysial cuticle and have been suggested to play roles in calcification. Additionally, several matrix proteins from the calcified cuticle and arthrodial membrane of *Homarus americanus* (Kragh et al., 1997; Andersen, 1998; Nousiainen et al., 1998) and *Cancer pagurus* (Andersen, 1999) have been purified and directly sequenced. A conserved 18-residue

motif (xLxGPSGφφxxDGxxxQφ; x=any residue; φ=hydrophobic residue) is present in many of the proteins from the calcified cuticle. Kumari et al. (1995) obtained similar sequences in the N-terminal fragments of some *Gecarcinus lateralis* cuticle proteins. This motif is not present in any of the arthrodial membrane proteins (Andersen, 1998; 1999) or the large number of insect cuticle proteins (Andersen et al., 1995). Andersen (1999) suggested that proteins with this motif may act as nucleators of crystal formation or regulators of crystal growth and size after nucleation.

Hequembourg (2002) used a chicken antibody made against a peptide with the sequence VLVGPSGIVTSDGQNVQF in immunohistochemical staining of the anecdysial cuticle of *Callinectes sapidus*. As expected, this antibody recognized calcified cuticle proteins but not arthrodial membrane proteins (Fig. 1; Hequembourg, 2002). This study reports the cloning and expression analysis of several cDNAs belonging to a family of genes from *C. sapidus* encoding post-translationally cleaved proteins with multiple, variant copies of the 18-residue motif (xLxGPSGφφxxDGxxxQφ).

Figure 1. Immunohistochemical staining of claw cuticle, including both the calcified cuticle and the arthrodial membrane, with an antibody made against the 18-residue motif (xLxGPSGφφxxDGxxxQφ) from anecdysial *C. sapidus* (Hequembourg, 2002). The yellow box in the picture of a blue crab claw represents a region where cuticle pieces were excised for staining.


D. Mark Gay, 2003
calcified cuticle
arthrodial membrane
endocuticle

MATERIALS AND METHODS

Animals and tissue isolation

Adult anecdysial, stage C_4 (Drach, 1939) blue crabs, *Callinectes sapidus*, were obtained from a local seafood market, and pre-ecdysial, stage D_2, and 0, 3, 6, 12, 24, and 48 h post-ecdysial crabs were obtained from a "shedding" operation in Kill Devil Hills, NC. Cuticle was excised from the mid-dorsal area covering the cardiac chamber and the dorso-branchial area covering the gill chamber. Pieces of both the arthrodial membrane and the adjacent calcified cuticle were removed from the carpus joint of the cheliped. Cuticle pieces used for protein extraction were rinsed with distilled water and the underlying hypodermis was scraped free. The hypodermis used for RNA extraction was removed from the mid-dorsal and arthrodial membrane areas where no muscle attachment sites and no other tissue types were present. The white levator muscle, which originates at the median plate and inserts at the fifth periopod, and its heavy tendon were excised as needed (Cochran, 1935). All tissues and cuticle pieces were stored either in 4°C or frozen in liquid nitrogen and stored at -80°C.

DNA and RNA isolation

Genomic DNA was isolated from muscle tissue of one crab according to Sambrook and Russell (2001) except the lysis buffer was modified to contain 10 mM Tris-HCl (pH 8), 0.1 M EDTA (pH 8), 0.5% (w/v) SDS, and 20 μg ml^{-1} RNase A and the proteinase K was added directly to the homogenate (Barreto, 2003). Resulting DNA was stored in 1X TE at -20°C.

Total RNA was isolated from multiple crabs from either fresh or frozen hypodermis using an RNeasy® Protect Mini kit (Qiagen) with the following modifications to the

manufacturer's instructions (Buda, 2004). Frozen tissue was stored overnight at -20°C in RNA*later*®-ICE (Ambion). One piece of mid-dorsal hypodermis was homogenized and centrifuged in 1 ml RLT buffer containing 10 µl β-mercaptoethanol, and the supernatant was bound to the column after addition of 50% ethanol. The arthrodial membrane hypodermis from two claw joints was subjected to the same protocol, except 1 ml Trizol (Invitrogen) and 70% ethanol were used instead. RNA was stored in nuclease-free H_2O at 4°C temporarily or at -80°C for long-term storage.

3' and 5' rapid amplification of cDNA ends (RACE)

All RACE experiments were performed on early post-ecdysial RNA from mid-dorsal hypodermis using the First-Choice™ RLM-RACE kit (RNA ligase mediated-rapid amplification of cDNA ends; Ambion) according to manufacturer's instructions; however, BD Advantage™ 2 Polymerase Mix (BD Biosciences) was used in PCR amplification. Gene-specific, 3' RACE sense primers (see Table 1 for primer sequences) were designed based on the amino acid sequences corresponding to the positions 2-8, 4-10, and 12-18 of the 18-residue motif following kit specifications and incorporating *Callinectes sapidus* codon usage preferences to minimize primer degeneracy. Gene-specific, 5' RACE antisense primers were designed from previously obtained 3' RACE sequences following kit specifications. cDNA for both 3' and 5' RACE was amplified by two rounds of PCR. In 3' RACE, the first round of PCR was primed with the 3' RACE primers (3RC1-4) and the adapter-specific primer (3RCOUT). The resulting products

Table 1 Primers used in RACE, RT-PCR, and Northern probe amplification

Primer	Sequence (5' ---> 3')
3RC1	YTSGTGGGYCCYTCCGGYAT
3RC2	GGYCCYTCCGGYGCYATYYT
3RC3	GAYGGYACYAACGTSCAGTTC
3RC4	GAYGGYACYCMNGTSCAGTTC
3RCOUT	GCGAGCACAGAATTAATACGACT
3RCIN	CGCGGATCCGAATTAATACGACTCACTATAGG
5RC1IN	AGACTCTCCGGTGTTGA
5RC1OUT	GCATCAACATATGAAACCAT
5RC2IN	CATGGGGAAGCTGAAAC
5RC2OUT	CAACACCCTACTGCTAACA
5RC3IN	CAGGCACTCCACTAGGA
5RC3OUT	ATGATGCCGCTCTCTCCGA
5RCOUT	GCTGATGGCGATGAATGAACACTG
5RCIN	CGCGGATCCGAACACTGCGTTTGCTGGCTTTGATG
F1SE	AAAAGTCATTTCAGTAGCC
F1AS	TGGATCTGATCTCATAATGT
F2SE	CATCTTGGAGTGCTGAGGAAGA
F2AS	TCGGCTCTCATAAAGTTTCTGA
F3SE	TCAGCGCCCAGTACGGAGAAT
F3AS	CGTGGGGGAACTGAATCAGA
1GNSE	GTATCATCTTTCCTGACGGGA
1GNAS	TGTTTGCCGTTGGAGAGAG
1SPSE	CCCTCTACGGCAGTGATTG
1SPAS	TGCATCAACATATGAAACCATAA
3SPSE	GTGGTTGGCTCCGCTGGTTT
3SPAS	TTAGCAACGACTGCCTCGCC

were then reamplified using the same 3' RACE primers but used a nested adapter-specific primer (3RCIN) instead. The first round of PCR in 5' RACE was primed with the primers (5RC1-3OUT) and the adapter-specific primer (5RCOUT). The second round was primed with the nested primers (5RC1-3IN) and the adapter-specific nested primer (5RCIN). The general PCR program for RACE was: 3 min at 94°C; 35 cycles of 30 s at 94°C, 30 s at 50°C (5' RACE) or 60°C (3' RACE), 30 s at 72°C; and 7 min at 72°C. The sizes of all RACE products were determined on 1% agarose gels according to Sambrook and Russell (2001).

Reverse transcriptase-polymerase chain reaction (RT-PCR) and DNA amplification

RT-PCR verified that the 3' and 5' RACE sequence overlap was not a chimeric product. RT-PCR utilized the same PCR program as RACE. Intronic regions were determined by amplification of genomic DNA using the same primers used in RT-PCR but with the following PCR program: 3 min at 92°C; 10 min at 50-60°C; 30 cycles of 1 min at 92°C, 1 min at 50-60°C, 2 min at 72°C; and 5 min at 72°C.

Molecular cloning and DNA sequencing of PCR products

PCR products, whether from RACE, RT-PCR, or genomic PCR, were cloned into pGEM®-T Easy Vector (Promega) according to manufacturer's instructions. Colony PCR used SP6 and T7 vector primers and the program used in genomic DNA amplification. Cloned PCR inserts were sequenced using ABI PRISM® BigDye® Terminator v3.1 Cycle Sequencing kit (Applied Biosystems) according to manufacturer's instructions, and the sequences were read on an ABI 3100 DNA sequencer (Applied Biosystems). Cultured clones were stored at -20°C in 5 ml TB containing 10% glycerol.

Northern blotting

Plasmids were isolated from probe-containing, cultured clones using the Perfectprep® Plasmid Mini kit (Eppendorf) according to manufacturer's instructions. Inserts were *Eco*RI digested for 4 h at 37 °C, separated on a gel, and purified using QIAquick gel extraction kit (Qiagen) according to manufacturer's instructions. Double-stranded DNA probes (25 ng) were labeled with [α-^{32}P]-dCTP using a Random Prime Labeling kit (Roche) according to the manufacturer's instructions. Radiolabeled probes were purified using a QIAquick Free Nucleotide Removal kit (Qiagen) according to manufacturer's instructions and stored in 100 µl nuclease-free H_2O at -20°C.

Two µg RNA Millennium™ Markers (Ambion) and 5 µg (≤10 µl) total RNA (both in 20 µl RNA loading mix (GenHunter)) were denatured at 65°C and separated on a 1% agarose gel containing formaldehyde according to Sambrook and Russell (2001). RNA was transferred to a positively charged nylon membrane (Immobilon™-Ny+; Millipore) and UV cross-linked at 20,000 µJ cm^{-2} according to the manufacturer's instructions.

The blots were hybridized with 30 µl (of 100 µl) denatured, radiolabeled probe diluted in 500 µl hybridization buffer. All hybridizations and washes were performed in a 30 cm glass tube in a rotisserie oven at 68°C. After hybridization, blots were wrapped in cellophane and exposed to X-OMAT x-ray film (Kodak). Three different radiolabeled probes were hybridized to the same RNA blots with stripping in between each hybridization according to Millipore.

Protein extraction

Proteins were extracted according to Andersen (1999). Briefly, calcified cuticle, arthrodial membrane, tendon, and muscle were decalcified in 10% acetic acid at 4°C until no

more bubbles developed, and the residues were briefly rinsed in distilled H_2O. The proteins were extracted in 6 M urea in 0.1% trifluoroacetic acid (TFA) overnight at 4°C with continuous agitation and quantified at OD_{280}.

SDS-polyacrylamide gel electrophoresis (PAGE) and Western blotting

Ten-µg protein extracts were brought to a final volume of 20 µl containing a modified 1X LDS buffer (10% glycerol, 294 mM LDS, 278 mM Tris, 0.51 mM EDTA, 22 µM bromophenol blue, 3.5 µM phenol red, pH 8.5) and 1X NuPAGE™ sample reducing agent (Invitrogen). Electrophoresis was performed using the NuPAGE™ system (Invitrogen) according to manufacturer's instructions. Extracts and low range (14.4, 21.5, 31.0, 45.0, 66.2, and 97.4 kDa) standards (BioRad) were denatured at 95°C for 10 min and separated on a 4-12% Bis-Tris gel (Invitrogen) in 1X MES-SDS buffer at 200 V for 40 min. The resulting gels were either stained with Colloidal Coomassie Blue (Invitrogen) for 3 h and destained in H_2O overnight or immediately transferred for Western blotting.

Fractionated proteins were electroblotted to Invitrolon™ PVDF membrane (Invitrogen) at 25 V for 2 h using the NuPAGE™ blotting system (Invitrogen). The portion of the blots containing MW markers was removed, stained in amido black (0.1% (w/v) amido black, 10% methanol, 7% acetic acid) for 20 min, and destained (10% methanol, 10% acetic acid) for 2 h. The blots containing the extracts were rinsed in TBST (20 mM Tris, 0.5 M NaCl, 0.05% Tween 20, pH 7.5) then blocked in TBST containing 2% (w/v) dry milk (Carnation) at room temperature for 1 h with continuous agitation. The blots were probed overnight with primary antibody (Hequembourg, 2002) diluted to 1:1000 in blocking buffer. After primary incubation, the blots were washed in TBST 3 times for 10 min each wash and incubated in alkaline

phosphatase-labeled, goat anti-chicken IgY secondary antibody diluted 1:5000 in blocking buffer for 1 h. The blots were rinsed and then washed in TBST 3 times for 10 min each wash. Bound antibodies were colorimetrically detected with NBT/BCIP reagent (Roche) according to manufacturer's instructions.

Bioinformatics

Primer design and sequence analysis was performed using Vector NTI (v8.0) (InforMax, Invitrogen), which utilizes ContigExpress to configure RACE and EST sequences, BLAST for sequence homologies, ClustalW for alignments, and SIM4 for cDNA and genomic DNA sequence comparisons. Nucleotide and protein sequences for alignments and homology determination were downloaded from GenBank at http://www.ncbi.nlm.nih.gov.

An expressed sequence tag (EST) database containing ~5000 sequences was developed from a normalized, cDNA library of the hypodermal RNA of pre- and post-ecdysial, mid-dorsal cuticle and arthrodial membrane of *Callinectes sapidus* (Invitrogen). The optimizing of the sequencing protocols was performed by Ms. Amy McElhinney, and the sequencing was performed by Dr. Francie Coblentz and MWG Biotech. The principle investigator of the EST project is Dr. Thomas Shafer. To confirm identity of cDNA sequences from RACE and RT-PCR and to identify additional transcripts, the sequence $Tx_4GGxCCx_4GGx_{13}GAxGGx_{10}CA$ where x = any nucleotide was used to find all the 18-residue motif-containing sequences located in the database.

Nomenclature

Nomenclature developed in this research is based on the presumption that the translated products of the transcripts described in the Results are "pro-proteins" that are enzymatically cleaved to produce cuticular peptides similar to those directly sequenced from *C. pagurus* and *H. americanus* (Kragh et al., 1997; Nousiainen et al., 1998; Andersen, 1999). The genes and cDNAs are named *CsproCP#.#* (*italics=cDNA*; normal=pro-protein), where "*Cs*" refers to *Callinectes sapidus*, "*pro*" indicates that the open reading frame is an uncleaved pro-protein, "*CP*" refers to cuticle protein, the first # is the number of peptides produced by the presumed cleavage, and the second # designates which cDNA is indicated when there are more than one that produce the same number of cleaved products. The presumed *C. sapidus* peptides that are incorporated into the cuticle after cleavage are referred to as CsCP#.#n'# (note normal font), where n is a letter (a-e) beginning at the N-terminus corresponding to the particular peptide cleaved from CsCP#.#.

RESULTS

Cloning and characterization of multiple cDNAs

The presence of the 18-residue motif (xLxGPSGφφxxDGxxxQφ) in many calcified cuticle proteins of the decapod crustaceans *C. pagurus*, *H. americanus*, and *G. lateralis* (Kumari et al., 1995; Kragh et al., 1997; Nousiainen et al., 1998; Andersen, 1999) and the immunohistochemical staining seen in *C. sapidus* itself (Fig. 1; Hequembourg, 2002) suggested that transcripts that encode this motif could be identified in *C. sapidus*. In total, 11 transcripts from the post-ecdysial, mid-dorsal hypodermis were obtained whose inferred translations contain multiple, variant copies of the 18-residue motif. The motif accounts for ~70% of the total amino acid residues in each open reading frame (ORF) (Table 2; Figs. 2 and 3). RACE and RT-PCR were initially performed to identify several of the transcript sequences. Additional sequences were acquired by *in silico* analysis of an EST database developed from a cDNA library of the *C. sapidus* hypodermis (Table 2; Figs. 2 and 3).

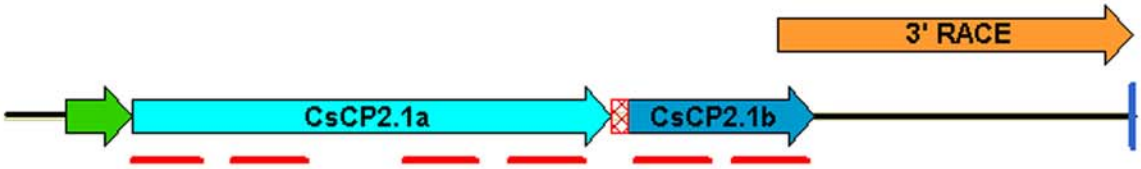
Gene-specific 3' RACE primers (3RC1-4 in Table 1), based on the N- and C-terminal halves of the 18-residue motif, amplified multiple products. Of these products, seven contain unique sequences with different lengths, each with a full or partial variant of the 18-residue motif in the inferred translation (Table 2; Figs. 2 and 3). In 5' RACE, gene-specific primers were based on the 3' RACE sequences (5RC1-3 in Table 1). These primers amplified multiple products resulting in three 5' RACE sequences that correspond to 3' RACE sequences. The three cDNAs obtained using RACE, *CsproCP3.1*, *CsproCP3.2*, and *CsproCP5.1*, were verified by RT-PCR (Table 2; Figs. 2

Table 2 The sequence details for the 11 cDNAs that code for the 18-residue motif (xLxGPSGφφxxDGxxxQφ). "*" indicates that a sequence does not have a complete open-reading frame in its translation; therefore, some sequence information is unknown (?)

cDNA	total length (bp)	5' UTR (bp)	3' UTR (bp)	ORF bp(AA)	total # of 18-residue motifs	# of motifs in "a" pep.	RACE product	EST product	# of ESTs in contig	primers
CsproCP2.1	830	44	234	552(183)	6	4	X	X	37	3RC4
CsproCP2.2	742	45	145	552(183)	6	4		X	6	
CsproCP2.3	661	43	267	351(116)	4	2		X	8	
CsproCP2.4	649	42	256	351(116)	4	2		X	8	
CsproCP3.1	933	56	169	705(234)	8	4	X			3RC3, 5RC3, F1SE, F1AS
CsproCP3.2	930	55	167	705(234)	8	4	X			3RC1, 5RC2, F2SE, F2AS
CsproCP3.3	666	43	116	507(169)	6	2	X	X	9	3RC2
*CsproCP4**	>919	?	>178	>741(247)	>9	4		X	1	
CsproCP5.1	1227	39	177	1011(337)	12	4	X	X	24	3RC2, 5RC1, F3SE, F3AS
CsproCP5.2	1189	40	144	1005(335)	12	4	X	X	8	3RC1
*CsproCP6**	>1018	?	?	>1018(339)	>13	4		RT-PCR product		F3SE, F3AS

Figure 2. Graphical representations of the 11 cDNA sequences from *C. sapidus* that code for proteins containing multiple, variant copies of the 18-residue motif (xLxGPSGφφxxDGxxxQφ): A) *CsproCP2.1*, B) *CsproCP2.2*, C) *CsproCP2.3*, D) *CsproCP2.4*, E) *CsproCP3.1*, F) *CSproCP3.2*, G) *CsproCP3.3*, H) *CsproCP4*, I) *CsproCP5.1*, J) *CsproCP5.2*, and K) *CsproCP6*. The representations are not to scale, and all virtually cleaved peptides with the same color denotes sequence similarity.

A) *CsproCP2.1*



B) *CsproCP2.2*

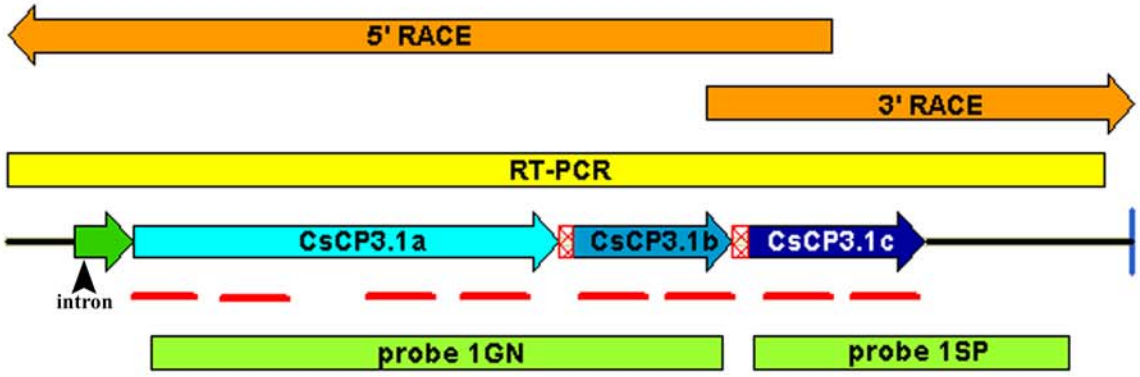
C) *CsproCP2.3*

D) *CsproCP2.4*

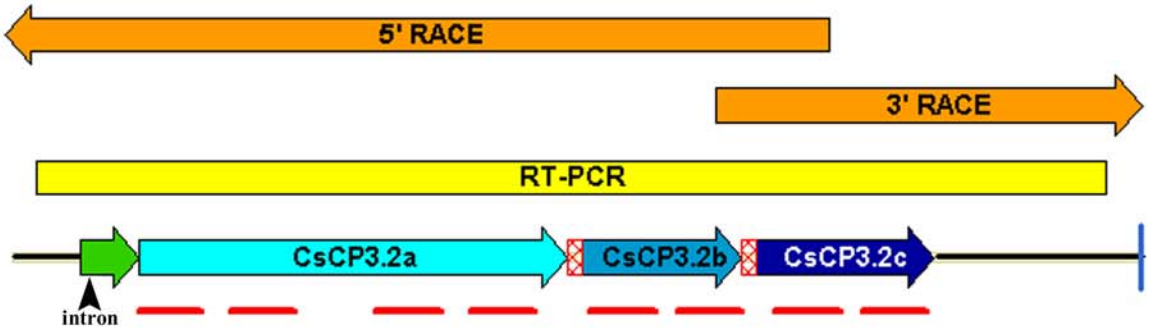
Key to symbols:

- signal peptide
- 18-residue motif
- 3' and 5' untranslated region (UTR)
- poly(A) site
- 4-residue cleavage site

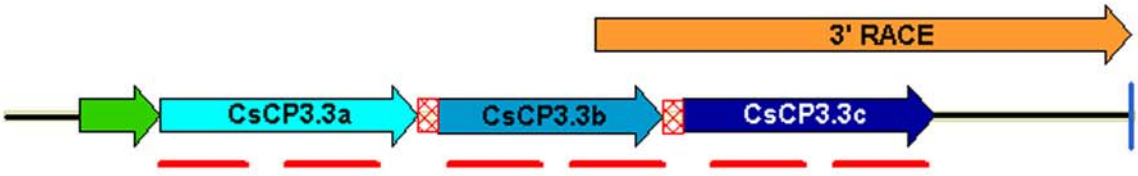
E) *CsproCP3.1*



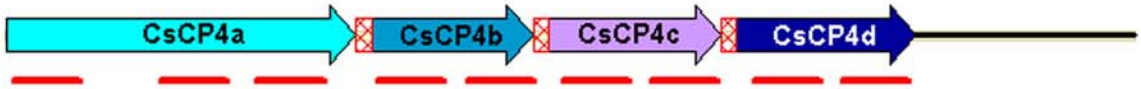
F) *CsproCP3.2*



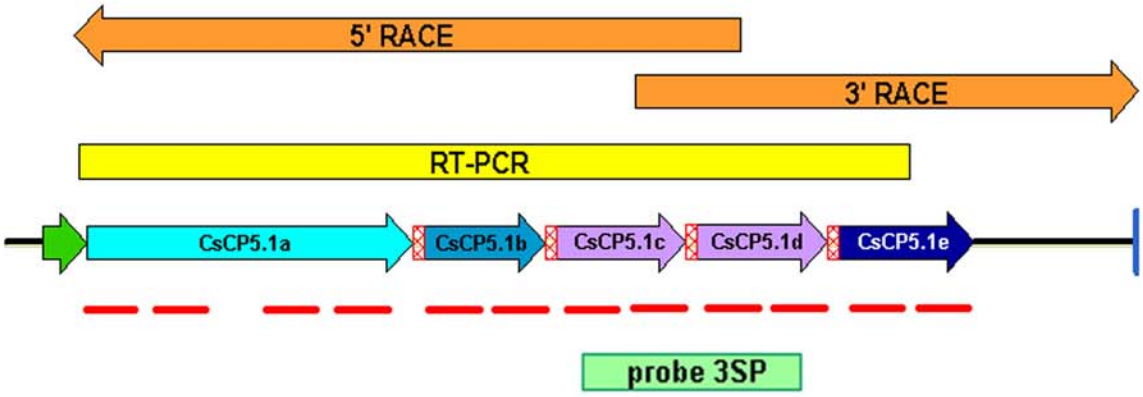
G) *CsproCP3.3*



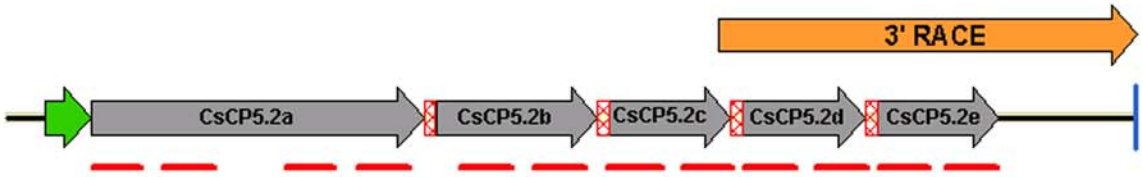
H) *CsproCP4*



I) *CsproCP5.1*



J) *CsproCP5.2*



K) *CsproCP6*

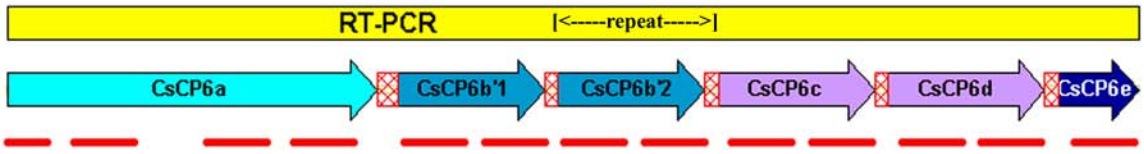


Figure 3. The nucleotide sequences and the inferred translations of the 11 cDNAs from *C. sapidus* whose inferred translated products contain multiple, variant copies of the 18-residue motif (xLxGPSGφφxxDGxxxQφ): A) *CsproCP2.1*, B) *CsproCP2.2*, C) *CsproCP2.3*, D) *CsproCP2.4*, E) *CsproCP3.1*, F) *CsproCP3.2*, G) *CsproCP3.3*, H) *CsproCP3.4*, I) *CsproCP4*, J) *CsproCP5.1*, K) *CsproCP5.2*, L) and *CsproCP6*. The nucleotide sequences in bold and underlined encode the 18-residue motif. The green regions in italics represent the signal peptide, and the boxed red regions represent the apparent cleavage sites (RxKR). Upper case letters represent translated regions, and lower case letters represent the 3' and 5' untranslated regions (UTRs). Regions in the open reading frames with the same color denotes sequence similarity.

A) *CsproCP2.1*

```
1   agtagccttatcttggagtgctgaggaagaccaggacagcagcc

45  ATGAAGCTTTTGGTAGCGATGTGTCTGATGGCGGTGGGCGTCAGCGCC
-16 M  K  L  L  V  A  M  C  L  M  A  V  G  V  S  A
93     TCATATGGGCCAGCCGGAATCGTCCATCCTGACGGGACACTGCAGCAGCTCACCCGCGAGGAGGCTGAAAATATC
1      S  Y  G  P  A  G  I  V  H  P  D  G  T  L  Q  Q  L  T  R  E  E  A  E  N  I
168 GCCGTTGTCGGGGCCTCTGGAGTGACATTCCACGATGGGTCACACATTCAGTTTAACCGCGACGCAGCGCTGCACCATGCTGGTATCGTACCACCCGTGCCCGTGCCAGTCATGCTGGAAACTCCA
26  A  V  V  G  A  S  G  V  T  F  H  D  G  S  H  I  Q  F  N  R  D  A  A  L  H  H  A  G  I  V  P  P  V  P  V  P  V  M  L  E  T  P
294 GGCCTTTATGGTGCTACTGGCATTGTCATGCCCGACGGAAACAATGTGCAGTTCACCGCTGATCAAGCTGCAAACATC
68  G  L  Y  G  A  T  G  I  V  M  P  D  G  N  N  V  Q  F  T  A  D  Q  A  A  N  I
372 GCCGTAATCGGGCCCTCTGGCGTCGTCATGGCTGATGGCAAGAACATACAGTTGAACGATGAAGGCGTCCCTTCCCGCAAGAAACGCGCAGTA
92  A  V  I  G  P  S  G  V  V  M  A  D  G  K  N  I  Q  L  N  D  E  G  V  P  S  R  K  K  R  A  V
465 CTCCTTGAGGGTCCCTCCGGTCTGATCTTCAATGATGGGCAAGTCAGGCATCTTCCTCCTGGTGTACAGATC
125 L  L  E  G  P  S  G  L  I  F  N  D  G  Q  V  R  H  L  P  P  G  V  Q  I
537 GTTCTTTTGACTACTTCTGGTGCCGTCCTTTCTAACGGTGACAACGTCCAGTTCAGCAAGTAA
149 V  L  L  T  T  S  G  A  V  L  S  N  G  D  N  V  Q  F  S  K  *   168
600 ttatcactttcataaaaatggacacaattattggcctgtaattttcgccatctgtacttcatagcttttctgtcttcttaccaccctattggtttaggaaatatgaaaagcctttcttgttcaacatgga

730 agctatccagttatatgtgtgacgagtttcctgcagtccacaaaaaacaatggatgtaaacagatatcttttataatactctttctaaatgttaaaactgc(a)N  830
```

B) *CsproCP2.2*

```
1   cagcagccttatcttggagtgctgaggaagaccaggacagcagct

46  ATGAAGCTTTTGGTAGCGATGTGTCTGATGGCGGCGGGCGTCAGCGCC
-16 M  K  L  L  V  A  M  C  L  M  A  A  G  V  S  A
94     TCATATGGGCCAGCCGGAATCGTCCATCCTGACGGGACACTGCAGCAGCTCACCCGCGAGGAGGCTGAAAATATC
1      S  Y  G  P  A  G  I  V  H  P  D  G  T  L  Q  Q  L  T  R  E  E  A  E  N  I
169 GCCGTTGTCGGTGACTCTGGAGTGACATTCCACGATGGGTCACACATTCAGTTTAACCGCGACGCAGCGCTGCACCATGCTGGTATCGTGCCACCCGTGCCCGTGCCAGTCATGCTTGAAACTCCA
26  A  V  V  G  D  S  G  V  T  F  H  D  G  S  H  I  Q  F  N  R  D  A  A  L  H  H  A  G  I  V  P  P  V  P  V  P  V  M  L  E  T  P
295 GGCCTTTATGGTGCTACTGGCATTGTCATGCCCGACGGAAACAATGTGCAGTTCACCGCTGATCAAGCTGCAAACGTC
68  G  L  Y  G  A  T  G  I  V  M  P  D  G  N  N  V  Q  F  T  A  D  Q  A  A  N  V
373 GCCGTAATCGGGCCCTCTGGCGTCGTCATGGCTGATGGCAAGAACGTACAGTTGAACGATGAAGGCCTCCCTTCCCGCAAGAAACGTTCCAAG
94  A  V  I  G  P  S  G  V  V  M  A  D  G  K  N  V  Q  L  N  D  E  G  L  P  S  R  K  K  R  S  K
466 CCTGTCGTTGGCGATTCAGGATACATCACCGCAAGTGGAAGGCCAGTCCAGCTTCCTCACGGTGTCACGATC
125 P  V  V  G  D  S  G  Y  I  T  A  S  G  R  P  V  Q  L  P  H  G  V  T  I
538 CTAATTGCTGGTGACACTGGGCTGCTGCTCTCCAACGGAGAAGCTGTGCAGCTTTACGAATAG
149 L  I  A  G  D  T  G  L  L  L  S  N  G  E  A  V  Q  L  Y  E  *   168
601 atatttattcaaaacacctgaaattgatttgtgctctattttccgaggcagtgatacacgtatttcttatagaggaaaagcacgtactctccattgccaagcactactaatctctccctataaataaata

732 tatatattttt(a)N  742
```

C) *CsproCP2.3*

```
1   cagtagccatcaaggagttctgcaagaccttacattagccacc

44  ATGAAGTTTCTGGTAGTGTTGTGTTTGATGGCGGTGGGTGCCAACGCT
-16 M  K  F  L  V  V  L  C  L  M  A  V  G  A  N  A
92     AAATTTGGAAAACATGGCATTGTTATGCCCGACGGCGTCAATGTCCAGTTTACTCATGACCAGGCCGAAAACATC
1      K  F  G  K  H  G  I  V  M  P  D  G  V  N  V  Q  F  T  H  D  Q  A  E  N  I
167 CTTATGATCGGCCCCTCTGGCGCCATCACTGCTGACGGCAAGCACGTGCAGCTGGACCGAGATGGACTTCCTGTAGTCCGCGCCAAGAGAGAGGTG
26  L  M  I  G  P  S  G  A  I  T  A  D  G  K  H  V  Q  L  D  R  D  G  L  P  V  V  R  A  K  R  E  V
263 CTGCTGCAGGGACCCTCCAGTGTTCTGTTCAAGGACGGACAGAGCAGGTCTCTTTCTGGTGGTGTAGAAATT
58  L  L  Q  G  P  S  S  V  L  F  K  D  G  Q  S  R  S  L  S  G  G  V  E  I
335 GTCGAAATCACTGAGACTGGAGCCGTCTTGTCCAACGGTGACAATGTTCAGTTCCTTGTCTAG
82  V  E  I  T  E  T  G  A  V  L  S  N  G  D  N  V  Q  F  L  V  *  101
398 tgtttctttttccatcacttatgctaccaatggtcgcttggactacatcaccagcaaactacttgacaacttcactctgtcatcacagtcgccggcagtatgcatgttggcaatcttctttctggcgaa

528 atgaaatccaactactaaactaaatctatttatccacgaccctgtcctatctatctattttcttctacacctctatcagcacatcaactgataggagcttcatggaattatatataaattttcaataaat

658 atgg(a)N  661
```

D) *CsproCP2.4*

```
1   agtagccaacaaggagttctgcaagaccttacattagccacc

43  ATGAAGTTTCTGGTAGTGTTGTGTTTGATGGCGGTGGGTGCCAACGCT
-16 M  K  F  L  V  V  L  C  L  M  A  V  G  A  N  A
91     AAATTTGGAAAACATGGCATTGTTATGCCCGACGGCGTCAATGTCCAGTTTACTCATGACCAGGCCGAAAACATC
1      K  F  G  K  H  G  I  V  M  P  D  G  V  N  V  Q  F  T  H  D  Q  A  E  N  I
166 CTTATGATCGGCCCCTCTGGCGCCATCACTGCTGACGGCAAGCACGTGCAGCTGGACCGAGATGGACTTCCTGTAGTCCGCGCCAAGAGAGAGGTG
26  L  M  I  G  P  S  G  A  I  T  A  D  G  K  H  V  Q  L  D  R  D  G  L  P  V  V  R  A  K  R  E  V
262 CTGCTGCAGGGACCCTCCAGTGTTCTGTTCAAGGACGGACAGAGCAGGTCTCTTTCTGGTGGTGTAGAAATT
58  L  L  Q  G  P  S  S  V  L  F  K  D  G  Q  S  R  S  L  S  G  G  V  E  I
334 GTCCAAATTACCAACACCGGAGCCATATTGTCCAACGGTGACAATGTTCAGTTCCGTGTCTAG
82  V  Q  I  T  N  T  G  A  I  L  S  N  G  D  N  V  Q  F  R  V  *  101
397 tgtttctttttcccatcacttatgctaccgagggtcgcttggtctacatcaccaacaaacaactttaacaacttcactgtctctgacaatatgcttgctggcgctcttctttccggcgaaatgaaatcca

527 gctgctaaattaaaaatatttatccacgaccctgtcctatctatctattttcttctacacctccatcagcacatctactgattggagcttcatggaattatatataaattttcaataaacatg(a)N 649
```

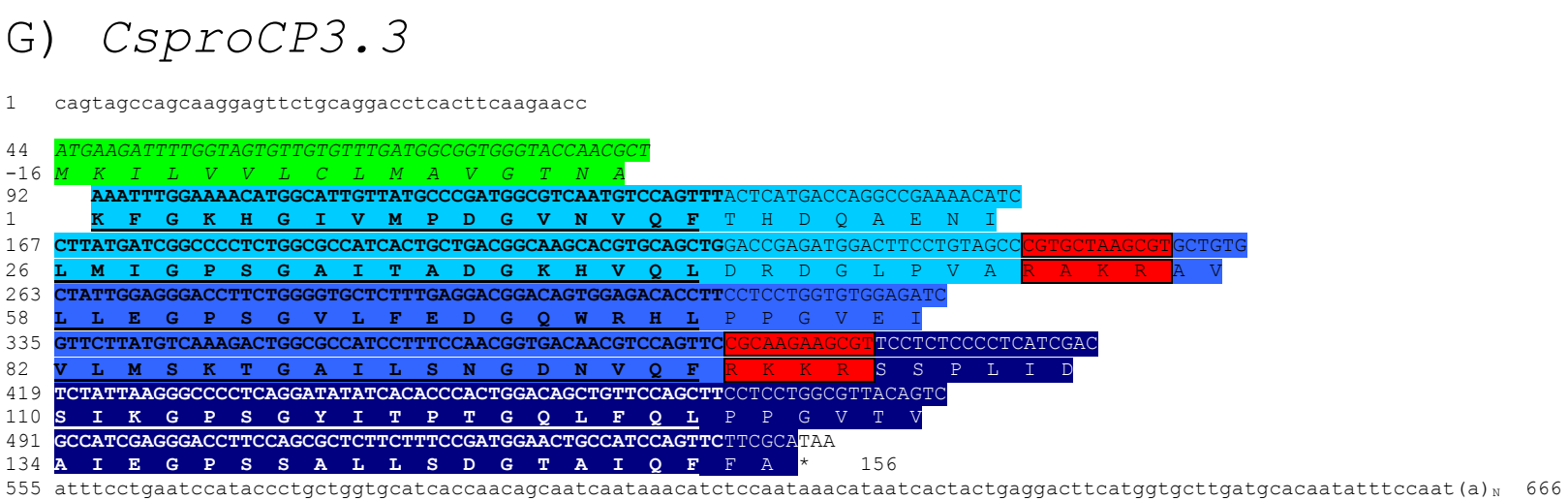
E) *CsproCP3.1*

```
1   aaaagtcatttcagtagccttatcttggagtgctgaggaagaccaagactgcagtc

57  ATGAAGCTTTTGGTAGCATTGTGCGTGATGGCAGTGGGTGTCAGCGCC
-16 M  K  L  L  V  A  L  C  V  M  A  V  G  V  S  A
105    CAGTACGGAGAATCTGGTATCATCTTTCCTGACGGGACGCTGAGGCAACTCACCCCAGAGGAGGCTGCCAACATC
1      Q  Y  G  E  S  G  I  I  F  P  D  G  T  L  R  Q  L  T  P  E  E  A  A  N  I
180 GCTGCTATCGGGCAGTCTGGAGTGGTCTTTAAGGATGGATCAAACAAGCAGTTTGACATGGATTTTGCCGCCCTGCACAACAACCTCCCCGCCCCAGCCAGGCCCGAGGAAGTGACCTTC
26  A  A  I  G  Q  S  G  V  V  F  K  D  G  S  N  K  Q  F  D  M  D  F  A  A  L  H  N  N  L  P  A  P  A  R  P  E  E  V  T  F
300 GGTCCCTACGGCTATCATGGCATCATAAAGCCCGACGGCAACAACGTGCAGTTCTCCCATGACCAGCACAGCAACGTT
66  G  P  Y  G  Y  H  G  I  I  K  P  D  G  N  N  V  Q  F  S  H  D  Q  H  S  N  V
378 GTCCTGGTCGGCCCCTCAGGTGTCATTACTGCTGACGGCAAGAACTTGCAGCTGGATCAGGATGGCCTCCCTCTCCCACTCCGCAGGAAGCGCGCCGTG
92  V  L  V  G  P  S  G  V  I  T  A  D  G  K  N  L  Q  L  D  Q  D  G  L  P  L  P  L  R  R  K  R  A  V
477 GCCCTCGAGGGTCCCTCCGGCGTATTGTTCGCAGACGGCCAGCTGAGACACCTCCCTGTGGGCGTGACTGTC
125 A  L  E  G  P  S  G  V  L  F  A  D  G  Q  L  R  H  L  P  V  G  V  T  V
549 GTCAGTGTGGGTCCCTCTGGCGCCACTCTCTCCAACGGCAAACACGTCCAGTTCCGTGAGAAGCGCTCAGCACCCTCTACG
149 V  S  V  G  P  S  G  A  T  L  S  N  G  K  H  V  Q  F  R  E  K  R  S  A  P  S  T
630 GCAGTGATTGATGAGAGTGGCATCATCACCCCAAGCGGACGGCCAATTCACCTTCCCCTGGGCACATACGTT
176 A  V  I  D  E  S  G  I  I  T  P  S  G  R  P  I  H  L  P  L  G  T  Y  V
702 GTTAATCACGGGCCTTCTGGAATTTTGCTCAACACCGGAGAGTCTATTCAATTCGAATTATAA
200 V  N  H  G  P  S  G  I  L  L  N  T  G  E  S  I  Q  F  E  L  *   219
765 gttcataacaacaatattatatgaaatggtgttagcaataagacctcatccttgtttctgaaggagatatattttagtacaatgacataagagattatggtttcatatgttgatgcactatcataaacat

895 tatgagatcagatccacaaatatatttcatatcatgact(a)N  933
```

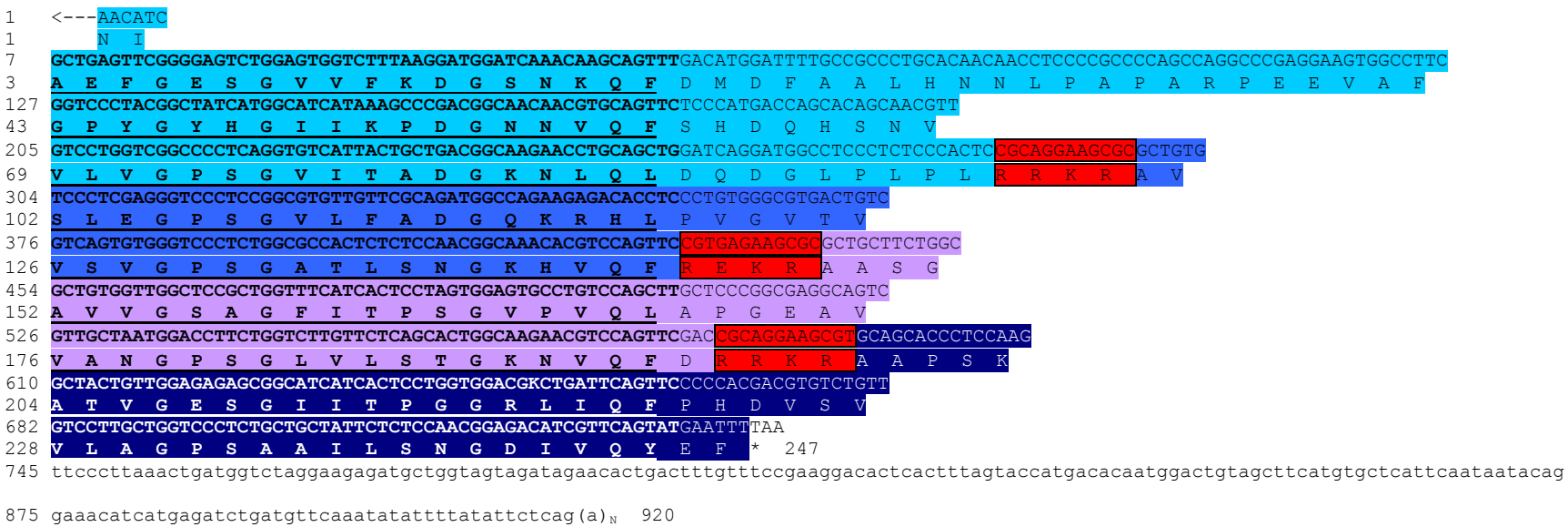
F) *CsproCP3.2*

```
1   aaaagtgtcccagtagccgcatcttggagtgctgaggaagaccaggatcgcagcc

56  ATGAAGTTTTTGGTAGCATTGTGCGTGATGGCAGTGGGTGTCAGCGCC
-16 M  K  F  L  V  A  L  C  V  M  A  V  G  V  S  A
104    CAGTACGGAGAATCTGGTATCATCTTTCCTGACGGGACGCTGAAGCAACTCGCCCCAGAGGAGGCTGCCAACATC
1      Q  Y  G  E  S  G  I  I  F  P  D  G  T  L  K  Q  L  A  P  E  E  A  A  N  I
179 GCTGAGCTCGGGGAGTCTGGAGTGGTCTTTAAGGATGGATCACACAAGCAGTTTGACATGGAGTTTACCGCCCTGCACAACAACCTCCCCGCCCCAGCCAGGCCCGAGGAAGTGACCTTC
26  A  E  L  G  E  S  G  V  V  F  K  D  G  S  H  K  Q  F  D  M  E  F  T  A  L  H  N  N  L  P  A  P  A  R  P  E  E  V  T  F
299 GGTCCCTACGGCTACCATGGCATCATAAAGCCCGACGGCAACAACGTGCAGTTCTCCCATGACCAGCACAGCAACGTT
66  G  P  Y  G  Y  H  G  I  I  K  P  D  G  N  N  V  Q  F  S  H  D  Q  H  S  N  V
377 GTCCTGGTCGGCCCCTCAGGTGTCATTACTGCTGACGGCAAGAACTTGCAGCTGGATCAGGATGGCCTCCCTCTCCCACTCCGCAGGAAGCGCGCTGTG
92  V  L  V  G  P  S  G  V  I  T  A  D  G  K  N  L  Q  L  D  Q  D  G  L  P  L  P  L  R  R  K  R  A  V
476 GCCCTCGAGGGTCCCTCCGGCGTGAAGTTCGCAGACGGCCAGCTGAGACACCTCCCTGTGGGCGTGACTGTC
125 A  L  E  G  P  S  G  V  K  F  A  D  G  Q  L  R  H  L  P  V  G  V  T  V
548 GTCAGTGTGGGTCCCTCTGGCGCCACTCTCTCCAACGGCGACAACGTCCAGTTCCGTGAGAAGCGCGCTGCACCCTCTCAG
149 V  S  V  G  P  S  G  A  T  L  S  N  G  D  N  V  Q  F  R  E  K  R  A  A  P  S  Q
629 GCCGTTGTAGGAGAAGGTGGCATCATCACCCCAGGTGGAGTCCAGTTTCAGCTTCCCCATGGTGTGTATATT
176 A  V  V  G  E  G  G  I  I  T  P  G  G  V  Q  F  Q  L  P  H  G  V  Y  I
701 GTCTCTAAAGGGCCTTCTGCAGCTCTGCTCTCCAACGGACAAGCTGTTCAGTATGAATTCTAG
200 V  S  K  G  P  S  A  A  L  L  S  N  G  Q  A  V  Q  Y  E  F  *  219
764 tttcaacacaacgatgttctaggatgagctgttagcagtagggtgttggctttttgaaagacaatgttttcattacagtgacacaatggattatagcttcatgtgctgacttactatcagaaactttatg

894 agagccgattttcaaatatatgtatatatatcacaat(a)N  930
```

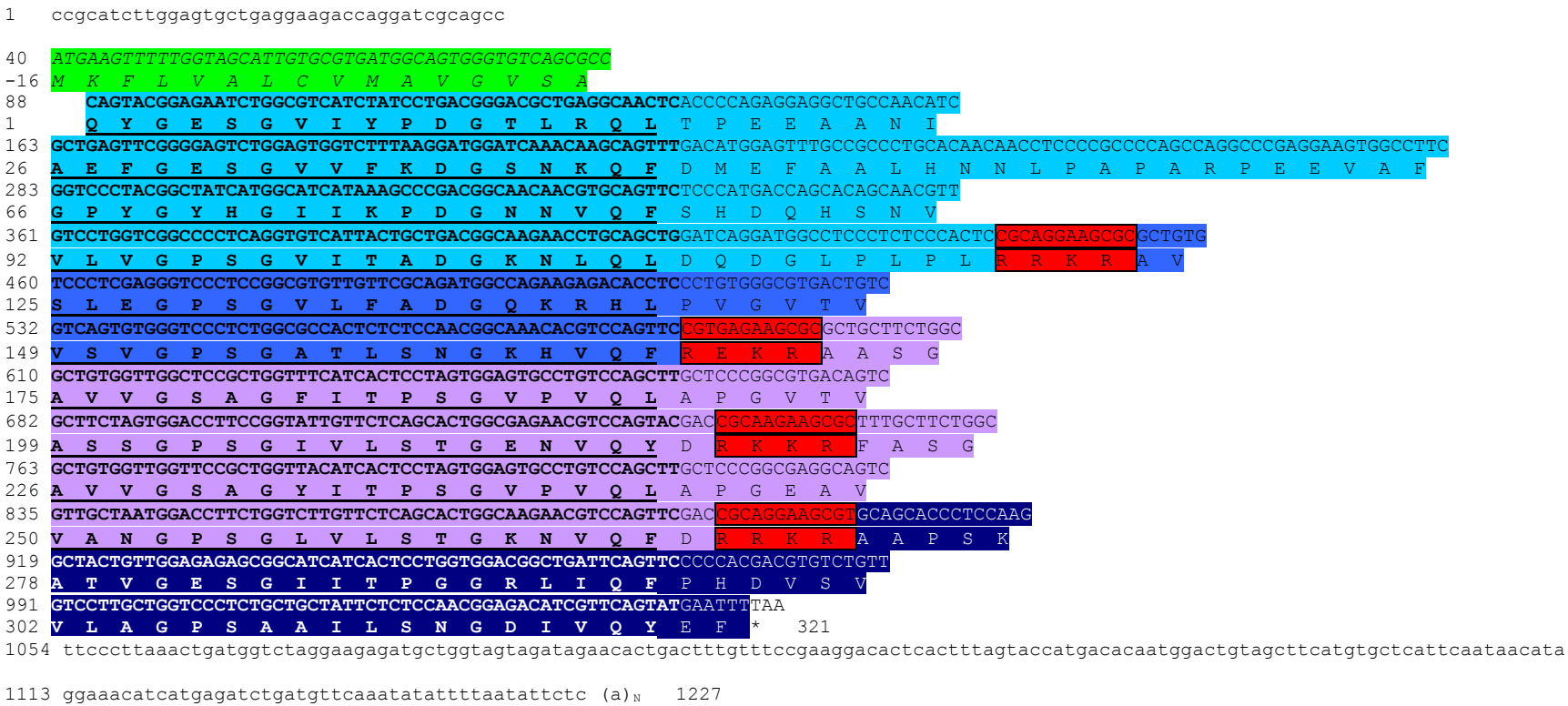
G) *CsproCP3.3*


1 cagtagccagcaaggagttctgcaggacctcacttcaagaacc
44 ATGAAGATTTTGGTAGTGTTGTGTTTGATGGCGGTGGGTACCAACGCT
-16 M K I L V V L C L M A V G T N A
92 AAATTTGGAAAACATGGCATTGTTATGCCCGATGGCGTCAATGTCCAGTTTACTCATGACCAGGCCGAAAACATC
1 K F G K H G I V M P D G V N V Q F T H D Q A E N I
167 CTTATGATCGGCCCCTCTGGCGCCATCACTGCTGACGGCAAGCACGTGCAGCTGGACCGAGATGGACTTCCTGTAGCCCGTGCTAAGCGTGCTGTG
26 L M I G P S G A I T A D G K H V Q L D R D G L P V A R A K R A V
263 CTATTGGAGGGACCTTCTGGGGTGCTCTTTGAGGACGGACAGTGGAGACACCTTCCTCCTGGTGTGGAGATC
58 L L E G P S G V L F E D G Q W R H L P P G V E I
335 GTTCTTATGTCAAAGACTGGCGCCATCCTTTCCAACGGTGACAACGTCCAGTTCCGCAAGAAGCGTTCCTCTCCCCTCATCGAC
82 V L M S K T G A I L S N G D N V Q F R K K R S S P L I D
419 TCTATTAAGGGCCCCTCAGGATATATCACACCCACTGGACAGCTGTTCCAGCTTCCTCCTGGCGTTACAGTC
110 S I K G P S G Y I T P T G Q L F Q L P P G V T V
491 GCCATCGAGGGACCTTCCAGCGCTCTTCTTTCCGATGGAACTGCCATCCAGTTCTTCGCATAA
134 A I E G P S S A L L S D G T A I Q F F A * 156
555 atttcctgaatccataccctgctggtgcatcaccaacagcaatcaataaacatctccaataaacataatcactactgaggacttcatggtgcttgatgcacaatatttccaat(a)N 666

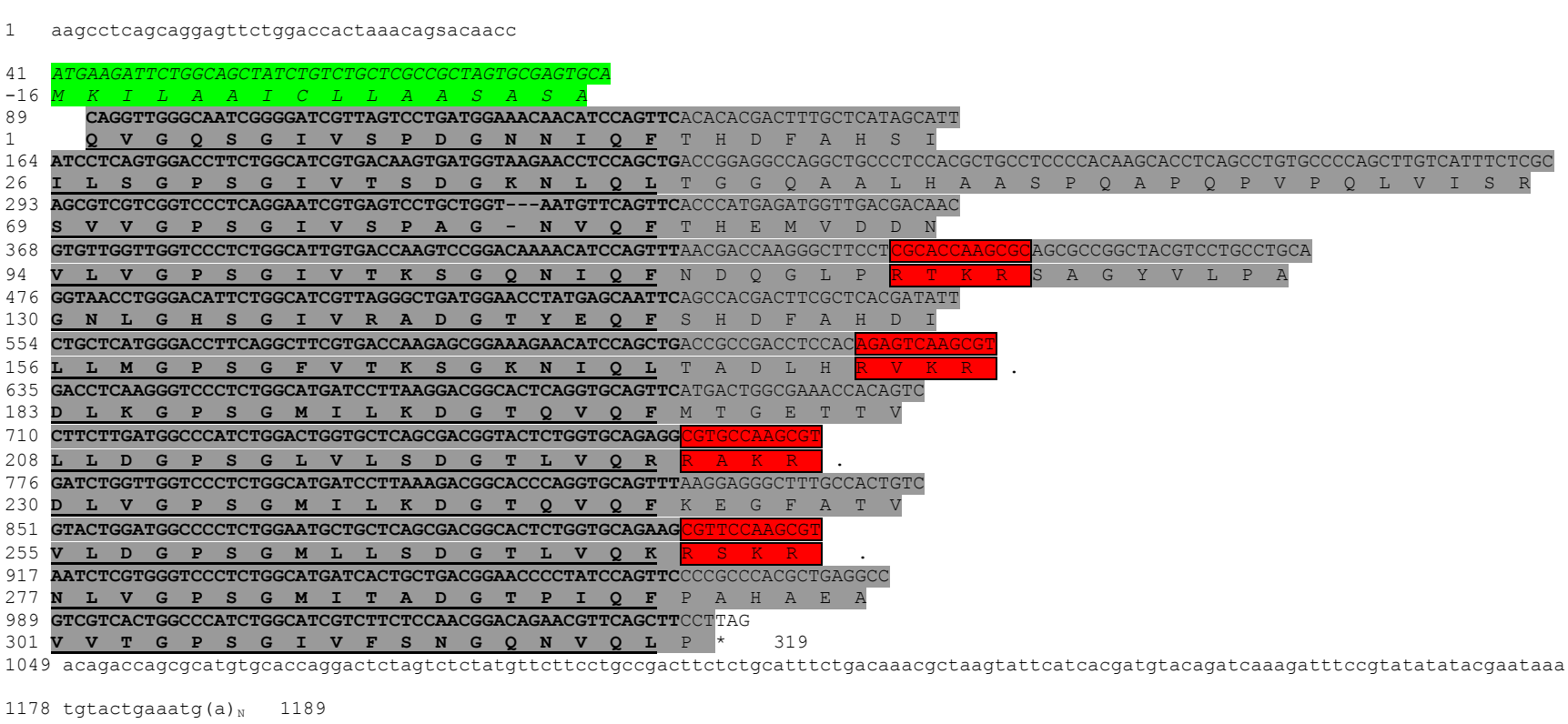
H) *CsproCP4*


1 <---AACATC
1 N I
7 GCTGAGTTCGGGGAGTCTGGAGTGGTCTTTAAGGATGGATCAAACAAGCAGTTTGACATGGATTTTGCCGCCCTGCACAACAACCTCCCCGCCCCAGCCAGGCCCGAGGAAGTGGCCTTC
3 A E F G E S G V V F K D G S N K Q F D M D F A A L H N N L P A P A R P E E V A F
127 GGTCCCTACGGCTATCATGGCATCATAAAGCCCGACGGCAACAACGTGCAGTTCTCCCATGACCAGCACAGCAACGTT
43 G P Y G Y H G I I K P D G N N V Q F S H D Q H S N V
205 GTCCTGGTCGGCCCCTCAGGTGTCATTACTGCTGACGGCAAGAACCTGCAGCTGGATCAGGATGGCCTCCCTCTCCCACTCCGCAGGAAGCGCGCTGTG
69 V L V G P S G V I T A D G K N L Q L D Q D G L P L P L R R K R A V
304 TCCCTCGAGGGTCCCTCCGGCGTGTTGTTCGCAGATGGCCAGAAGAGACACCTCCCTGTGGGCGTGACTGTC
102 S L E G P S G V L F A D G Q K R H L P V G V T V
376 GTCAGTGTGGGTCCCTCTGGCGCCACTCTCTCCAACGGCAAACACGTCCAGTTCCGTGAGAAGCGCGCTGCTTCTGGC
126 V S V G P S G A T L S N G K H V Q F R E K R A A S G
454 GCTGTGGTTGGCTCCGCTGGTTTCATCACTCCTAGTGGAGTGCCTGTCCAGCTTGCTCCCGGCGAGGCAGTC
152 A V V G S A G F I T P S G V P V Q L A P G E A V
526 GTTGCTAATGGACCTTCTGGTCTTGTTCTCAGCACTGGCAAGAACGTCCAGTTCGACCGCAGGAAGCGTGCAGCACCCTCCAAG
176 V A N G P S G L V L S T G K N V Q F D R R K R A A P S K
610 GCTACTGTTGGAGAGAGCGGCATCATCACTCCTGGTGGACGKCTGATTCAGTTCCCCCACGACGTGTCTGTT
204 A T V G E S G I I T P G G R L I Q F P H D V S V
682 GTCCTTGCTGGTCCCTCTGCTGCTATTCTCTCCAACGGAGACATCGTTCAGTATGAATTTTAA
228 V L A G P S A A I L S N G D I V Q Y E F * 247
745 ttcccttaaactgatggtctaggaagagatgctggtagtagatagaacactgactttgtttccgaaggacactcactttagtaccatgacacaatggactgtagcttcatgtgctcattcaataatacag
875 gaaacatcatgagatctgatgttcaaatatattttatattctcag(a)N 920

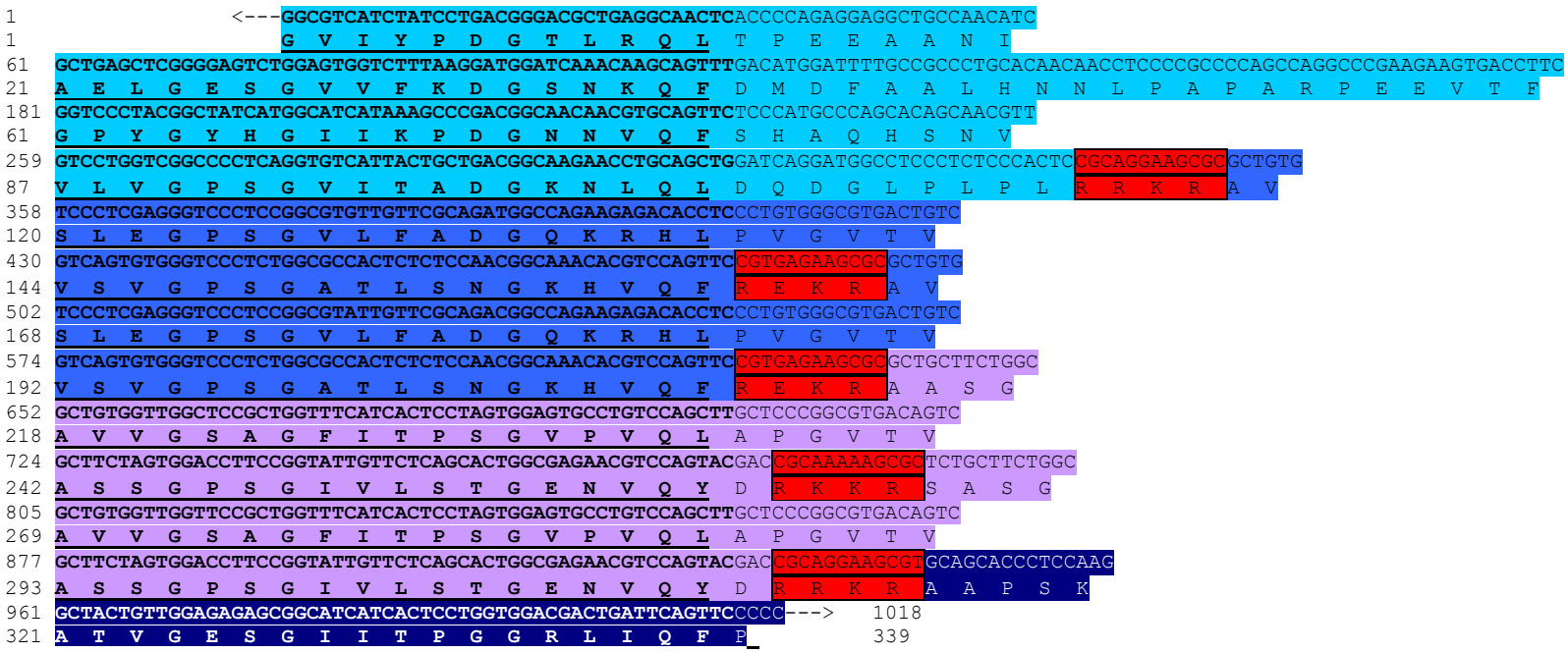
I) *CsproCP5.1*


1 ccgcatcttggagtgctgaggaagaccaggatcgcagcc
40 ATGAAGTTTTTGGTAGCATTGTGCGTGATGGCAGTGGGTGTCAGCGCC
-16 M K F L V A L C V M A V G V S A
88 CAGTACGGAGAATCTGGCGTCATCTATCCTGACGGGACGCTGAGGCAACTCACCCCAGAGGAGGCTGCCAACATC
1 Q Y G E S G V I Y P D G T L R Q L T P E E A A N I
163 GCTGAGTTCGGGGAGTCTGGAGTGGTCTTTAAGGATGGATCAAACAAGCAGTTTGACATGGAGTTTGCCGCCCTGCACAACAACCTCCCCGCCCCAGCCAGGCCCGAGGAAGTGGCCTTC
26 A E F G E S G V V F K D G S N K Q F D M E F A A L H N N L P A P A R P E E V A F
283 GGTCCCTACGGCTATCATGGCATCATAAAGCCCGACGGCAACAACGTGCAGTTCTCCCATGACCAGCACAGCAACGTT
66 G P Y G Y H G I I K P D G N N V Q F S H D Q H S N V
361 GTCCTGGTCGGCCCCTCAGGTGTCATTACTGCTGACGGCAAGAACCTGCAGCTGGATCAGGATGGCCTCCCTCTCCCACTCCGCAGGAAGCGCGCTGTG
92 V L V G P S G V I T A D G K N L Q L D Q D G L P L P L R R K R A V
460 TCCCTCGAGGGTCCCTCCGGCGTGTTGTTCGCAGATGGCCAGAAGAGACACCTCCCTGTGGGCGTGACTGTC
125 S L E G P S G V L F A D G Q K R H L P V G V T V
532 GTCAGTGTGGGTCCCTCTGGCGCCACTCTCTCCAACGGCAAACACGTCCAGTTCCGTGAGAAGCGCGCTGCTTCTGGC
149 V S V G P S G A T L S N G K H V Q F R E K R A A S G
610 GCTGTGGTTGGCTCCGCTGGTTTCATCACTCCTAGTGGAGTGCCTGTCCAGCTTGCTCCCGGCGTGACAGTC
175 A V V G S A G F I T P S G V P V Q L A P G V T V
682 GCTTCTAGTGGACCTTCCGGTATTGTTCTCAGCACTGGCGAGAACGTCCAGTACGACCGCAAGAAGCGCTTTGCTTCTGGC
199 A S S G P S G I V L S T G E N V Q Y D R K K R F A S G
763 GCTGTGGTTGGTTCCGCTGGTTACATCACTCCTAGTGGAGTGCCTGTCCAGCTTGCTCCCGGCGAGGCAGTC
226 A V V G S A G Y I T P S G V P V Q L A P G E A V
835 GTTGCTAATGGACCTTCTGGTCTTGTTCTCAGCACTGGCAAGAACGTCCAGTTCGACCGCAGGAAGCGTGCAGCACCCTCCAAG
250 V A N G P S G L V L S T G K N V Q F D R R K R A A P S K
919 GCTACTGTTGGAGAGAGCGGCATCATCACTCCTGGTGGACGGCTGATTCAGTTCCCCCACGACGTGTCTGTT
278 A T V G E S G I I T P G G R L I Q F P H D V S V
991 GTCCTTGCTGGTCCCTCTGCTGCTATTCTCTCCAACGGAGACATCGTTCAGTATGAATTTTAA
302 V L A G P S A A I L S N G D I V Q Y E F * 321
1054 ttcccttaaactgatggtctaggaagagatgctggtagtagatagaacactgactttgtttccgaaggacactcactttagtaccatgacacaatggactgtagcttcatgtgctcattcaataacata
1113 ggaaacatcatgagatctgatgttcaaatatattttaatattctc (a)N 1227

J) *CsproCP5.2*


1 aagcctcagcaggagttctggaccactaaacagsacaacc
41 ATGAAGATTCTGGCAGCTATCTGTCTGCTCGCCGCTAGTGCGAGTGCA
-16 M K I L A A I C L L A A S A S A
89 CAGGTTGGGCAATCGGGGATCGTTAGTCCTGATGGAAACAACATCCAGTTCACACACGACTTTGCTCATAGCATT
1 Q V G Q S G I V S P D G N N I Q F T H D F A H S I
164 ATCCTCAGTGGACCTTCTGGCATCGTGACAAGTGATGGTAAGAACCTCCAGCTGACCGGAGGCCAGGCTGCCCTCCACGCTGCCTCCCCACAAGCACCTCAGCCTGTGCCCCAGCTTGTCATTTCTCGC
26 I L S G P S G I V T S D G K N L Q L T G G Q A A L H A A S P Q A P Q P V P Q L V I S R
293 AGCGTCGTCGGTCCCTCAGGAATCGTGAGTCCTGCTGGT---AATGTTCAGTTCACCCATGAGATGGTTGACGACAAC
69 S V V G P S G I V S P A G - N V Q F T H E M V D D N
368 GTGTTGGTTGGTCCCTCTGGCATTGTGACCAAGTCCGGACAAAACATCCAGTTTAACGACCAAGGGCTTCCTCGCACCAAGCGCAGCGCCGGCTACGTCCTGCCTGCA
94 V L V G P S G I V T K S G Q N I Q F N D Q G L P R T K R S A G Y V L P A
476 GGTAACCTGGGACATTCTGGCATCGTTAGGGCTGATGGAACCTATGAGCAATTCAGCCACGACTTCGCTCACGATATT
130 G N L G H S G I V R A D G T Y E Q F S H D F A H D I
554 CTGCTCATGGGACCTTCAGGCTTCGTGACCAAGAGCGGAAAGAACATCCAGCTGACCGCCGACCTCCACAGAGTCAAGCGT
156 L L M G P S G F V T K S G K N I Q L T A D L H R V K R .
635 GACCTCAAGGGTCCCTCTGGCATGATCCTTAAGGACGGCACTCAGGTGCAGTTCATGACTGGCGAAACCACAGTC
183 D L K G P S G M I L K D G T Q V Q F M T G E T T V
710 CTTCTTGATGGCCCATCTGGACTGGTGCTCAGCGACGGTACTCTGGTGCAGAGGCGTGCCAAGCGT
208 L L D G P S G L V L S D G T L V Q R R A K R .
776 GATCTGGTTGGTCCCTCTGGCATGATCCTTAAAGACGGCACCCAGGTGCAGTTTAAGGAGGGCTTTGCCACTGTC
230 D L V G P S G M I L K D G T Q V Q F K E G F A T V
851 GTACTGGATGGCCCCTCTGGAATGCTGCTCAGCGACGGCACTCTGGTGCAGAAGCGTTCCAAGCGT
255 V L D G P S G M L L S D G T L V Q K R S K R .
917 AATCTCGTGGGTCCCTCTGGCATGATCACTGCTGACGGAACCCCTATCCAGTTCCCCGCCCACGCTGAGGCC
277 N L V G P S G M I T A D G T P I Q F P A H A E A
989 GTCGTCACTGGCCCATCTGGCATCGTCTTCTCCAACGGACAGAACGTTCAGCTTCCTTAG
301 V V T G P S G I V F S N G Q N V Q L P * 319
1049 acagaccagcgcatgtgcaccaggactctagtctctatgttcttcctgccgacttctctgcatttctgacaaacgctaagtattcatcacgatgtacagatcaaagatttccgtatatatacgaataaa
1178 tgtactgaaatg(a)N 1189

K) *CsproCP6*


1 <---GGCGTCATCTATCCTGACGGGACGCTGAGGCAACTCACCCCAGAGGAGGCTGCCAACATC
1 G V I Y P D G T L R Q L T P E E A A N I
61 GCTGAGCTCGGGGAGTCTGGAGTGGTCTTTAAGGATGGATCAAACAAGCAGTTTGACATGGATTTTGCCGCCCTGCACAACAACCTCCCCGCCCCAGCCAGGCCCGAAGAAGTGACCTTC
21 A E L G E S G V V F K D G S N K Q F D M D F A A L H N N L P A P A R P E E V T F
181 GGTCCCTACGGCTATCATGGCATCATAAAGCCCGACGGCAACAACGTGCAGTTCTCCCATGCCCAGCACAGCAACGTT
61 G P Y G Y H G I I K P D G N N V Q F S H A Q H S N V
259 GTCCTGGTCGGCCCCTCAGGTGTCATTACTGCTGACGGCAAGAACCTGCAGCTGGATCAGGATGGCCTCCCTCTCCCACTCCGCAGGAAGCGCGCTGTG
87 V L V G P S G V I T A D G K N L Q L D Q D G L P L P L R R K R A V
358 TCCCTCGAGGGTCCCTCCGGCGTGTTGTTCGCAGATGGCCAGAAGAGACACCTCCCTGTGGGCGTGACTGTC
120 S L E G P S G V L F A D G Q K R H L P V G V T V
430 GTCAGTGTGGGTCCCTCTGGCGCCACTCTCTCCAACGGCAAACACGTCCAGTTCCGTGAGAAGCGCGCTGTG
144 V S V G P S G A T L S N G K H V Q F R E K R A V
502 TCCCTCGAGGGTCCCTCCGGCGTATTGTTCGCAGACGGCCAGAAGAGACACCTCCCTGTGGGCGTGACTGTC
168 S L E G P S G V L F A D G Q K R H L P V G V T V
574 GTCAGTGTGGGTCCCTCTGGCGCCACTCTCTCCAACGGCAAACACGTCCAGTTCCGTGAGAAGCGCGCTGCTTCTGGC
192 V S V G P S G A T L S N G K H V Q F R E K R A A S G
652 GCTGTGGTTGGCTCCGCTGGTTTCATCACTCCTAGTGGAGTGCCTGTCCAGCTTGCTCCCGGCGTGACAGTC
218 A V V G S A G F I T P S G V P V Q L A P G V T V
724 GCTTCTAGTGGACCTTCCGGTATTGTTCTCAGCACTGGCGAGAACGTCCAGTACGACCGCAAAAAGCGCTCTGCTTCTGGC
242 A S S G P S G I V L S T G E N V Q Y D R K K R S A S G
805 GCTGTGGTTGGTTCCGCTGGTTTCATCACTCCTAGTGGAGTGCCTGTCCAGCTTGCTCCCGGCGTGACAGTC
269 A V V G S A G F I T P S G V P V Q L A P G V T V
877 GCTTCTAGTGGACCTTCCGGTATTGTTCTCAGCACTGGCGAGAACGTCCAGTACGACCGCAGGAAGCGTGCAGCACCCTCCAAG
293 A S S G P S G I V L S T G E N V Q Y D R R K R A A P S K
961 GCTACTGTTGGAGAGAGCGGCATCATCACTCCTGGTGGACGACTGATTCAGTTCCCCC---> 1018
321 A T V G E S G I I T P G G R L I Q F P 339

and 3). RT-PCR primers (F3SE and F3AS in Table 1) for *CsproCP5.1* additionally amplified the partial transcript *CsproCP6*, which contains a repeat of a 132 bp region present only once in *CsproCP5.1* (Table 2; Figs. 2 and 3). Although the size of *CsproCP6* is unknown, it is nearly identical to *CsproCP5.1* except for the additional repeat; therefore, the predicted size of *CsproCP6* is ~132 nucleotides longer than *CsproCP5.1*.

A search of the EST database using a degenerate sequence based on the 18-residue motif was performed to further verify the RACE contigs and to obtain additional sequences. This search generated 169 sequences. Contigs were determined, and the multiple ESTs for each contig helped to account for PCR errors. The cDNAs *CsproCP2.1*, *CsproCP2.2*, *CsproCP2.3*, *CsproCP2.4*, *CsproCP3.3*, and *CsproCP5.2* each contain an intact ORF and are contigs of five or more sequences (Table 2; Figs. 2 and 3). *CsproCP4* is from a single clone in the library and does not contain an intact ORF (Table 2; Figs. 2 and 3). The sequences obtained by RACE and RT-PCR were also located in the EST database. In most cases, the database did not contain more sequence information; however, it did contribute to an additional 80 bps to the 5' end of *CsproCP5.1*.

A 4-residue motif (RxKR) exists in each of the translations one to five times and appears to be a recognition site for a trypsin-like serine protease (Figs. 2 and 3). Assuming cleavage occurs at all of the recognition sites, the 11 pro-proteins could be cleaved into 37 peptides. Thus, the nomenclature for the different *CsproCP* cDNAs reflects the putative cleavage of the CsproCP pro-proteins into their respective CsCP peptides. The N-terminally cleaved, "a" peptides contain either two or four copies of the

18-residue motif, but all other putatively cleaved peptides contain only two copies of the motif (Table 2; Figs. 2 and 3).

Genomic DNA amplification and cDNA sequence comparison

To examine the possibility of alternate splicing, amplification of genomic DNA was attempted using all nine combinations of the three sets of RT-PCR primers (Table 1). Primers F2SE and F1AS amplified genomic DNA corresponding to *CsproCP3.1*, and primers F2SE and F2AS amplified DNA corresponding to *CsproCP3.2* (Fig. 4). The two genomic sequences contain introns that interrupt the signal peptide in each pro-protein (Fig. 4). The different length (118 bp and 103 bp) and sequence of each intron indicates that *CsproCP3.1* and *CsproCP3.2* are coded by different genes; however, it cannot yet be determined if alternate splicing is responsible for any of the other transcripts.

Tissue- and stage-specific gene expression

The expression of the some of the *CsproCP* genes was determined by Northern analysis. The probes were designed according to specificity of transcript hybridization. Probe 1GN (primers 1GNSE and 1GNAS; Table 1) was designed to detect the transcripts *CsproCP3.1*, *CsproCP3.2*, *CsproCP4*, *CsproCP5.1*, and *CsproCP6*; therefore, acting as a general probe. It encompasses the nucleotide sequence 121 to 593 in cDNA *CsproCP3.1*, which encodes the majority of the cleaved peptides CsCP3.1a and CsCP3.1b (Figs. 2E and 3E). This region shares ~94% homology with the corresponding nucleotide regions in the four other cDNAs. Probe 1SP (primers 1SPSE and 1SPAS; Table 1) encompasses the nucleotide sequence 621-881 in *CsproCP3.1*, which encodes CsCP3.1c and part of

Figure 4. The intron regions from genomic DNA that interrupt the signal peptides of *CsproCP3.1* (A) and *CsproCP3.2* (B). The different lengths and sequences show that different genes encode these two transcripts.

A) *CsproCP3.1* intron

```
37  ATGAAGCTTTTGgtatgattcttcacttgctggtgttcaatatttttttccttttttgcaagtcaatgtgaggactacttttc
-16 M   K   L   L
122 agctgttgacgtgcctttctcatgccctccgtgtgactccaacagGTAGCATTGTGC  179
-12                                              V   A   L   C
```

B) *CsproCP3.2* intron

```
37  ATGAAGTTTTTGgtatgattcttcattgatggtgttcagtgatctttacagtcaatatgataaagtctttacgctggtgtcacgc
-16 M   K   F   L
122 ctttctcatgccctccgtgtgactccaacagGTAGCATTGTGC  164
-12                                 V   A   L   C
```

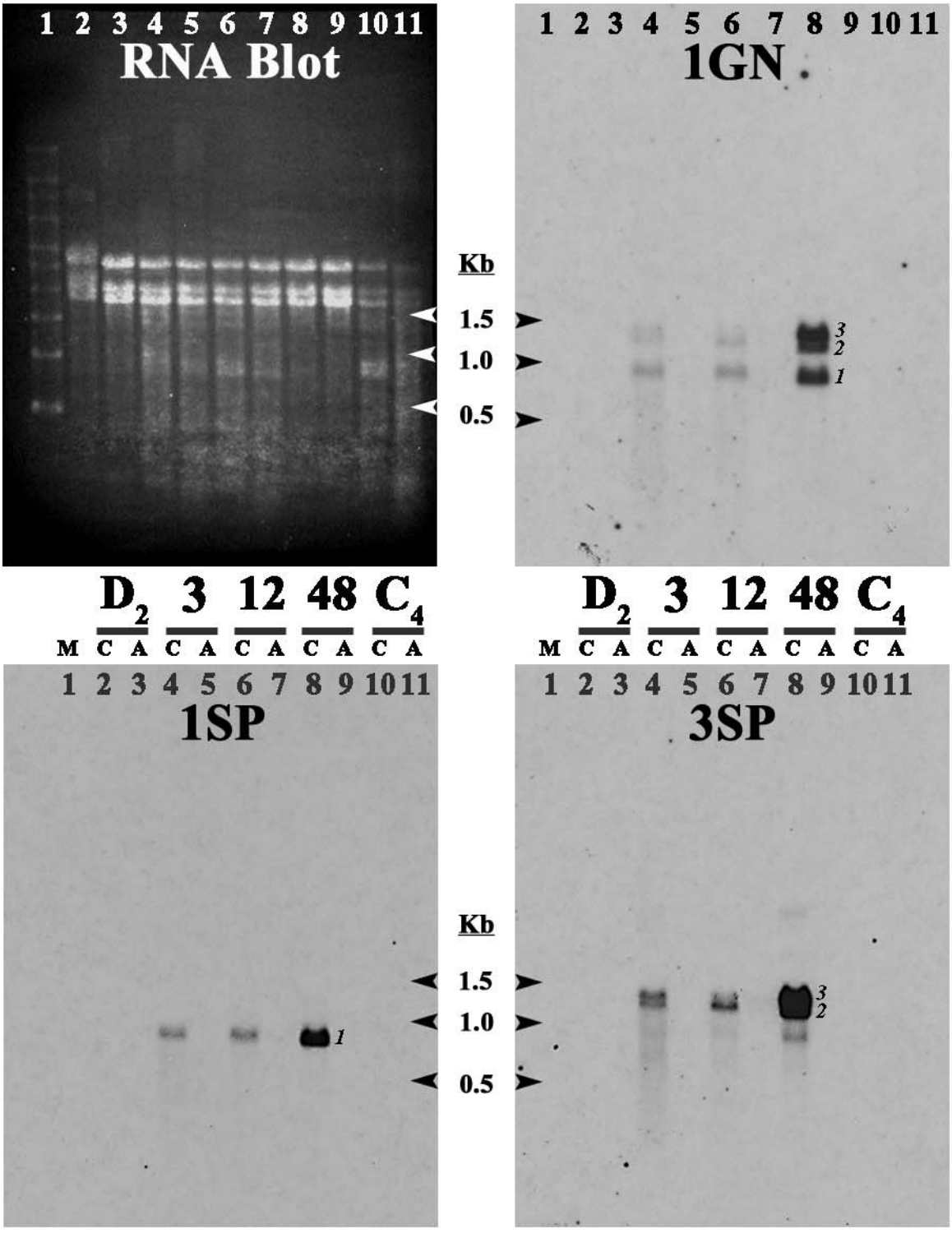
the 3' untranslated region (Fig. 2E and 3E). This region does not share significant homology with any of the other cDNAs; therefore, 1SP should only detect *CsproCP3.1* (Fig. 2E and 3E). Probe 3SP (primers 3SPSE and 3SPAS; Table 1) was designed to detect the transcripts *CsproCP5.1* and *CsproCP6*. It encompasses the nucleotide sequence 613-842 in *CsproCP5.1*, which encodes CsCP5.1c and CsCP5.1d (Fig. 2I and 3I). This region is shared with the *CsproCP6* region encoding CsCP6c and CsCP6d and is not found in any of the other cDNAs.

The three probes only detected transcripts in hypodermal RNA of cuticle that is in the process of calcification (Fig. 5B-D). No transcripts were detected in the RNA from pre-ecdysial and anecdysial mid-dorsal hypodermis, and none were detected at any stage in the arthrodial membrane hypodermis (Fig. 5B-D). The transcripts were weakly detected in mid-dorsal hypodermis at 3 h and 12 h post-ecdysis and strongly detected at 48-h post-ecdysis (Fig. 5B-D). Hybridization of probe 1GN yielded three bands whose sizes agreed well with that of *CsproCP3.1* and *CsproCP3.2* (~930 bps), *CsproCP5.1* (1227 bps), and *CsproCP6* (~1360 bps) (1, 2, and 3 respectively; Fig. 5B). Although probe 1GN should have recognized *CsproCP4*, whether it was detected or not is unknown because its size is unknown. The transcript detected by probe 1SP putatively corresponds to *CsproCP3.1* (1; Fig. 5B, C), and the transcripts detected by probe 3SP putatively correspond to *CsproCP5.1* and *CsproCP6* (2 and 3 respectively; Fig. 5B, D).

Tissue- and stage-specific protein expression

SDS-PAGE and Western blotting were performed to determine the protein content and the presence of the 18-residue motif in cuticle proteins extracted at various

Figure 5. Northern expression of the gene family encoding the 18-residue motif on calcified cuticle (C) vs. noncalcified arthrodial membrane (A) at various time points in the molting cycle. Lane 1 is the RNA marker. Lanes 2 and 3 are from a pre-ecdysial (D_2) crab. Lanes 4 and 5 are from a 3 h post-ecdysial crab. Lanes 6 and 7 are from a 12 h post-ecdysial crab. Lanes 8 and 9 are from a 48 h post-ecdysial crab. Lanes10 and 11 are from an anecdysial (C_4) crab. A) Blot containing 5 µg transferred RNA and 2 µg RNA markers and the gene expression determined by probes B) 1GN, C) 1SP, and D) 3SP. 1 putatively corresponds to both *CSproCP3.1* and *CSproCP3.2* in B but only to *CSproCP3.1* in C. 2 putatively corresponds to *CSproCP5.1*, and 3 putatively corresponds to *CSproCP6*.

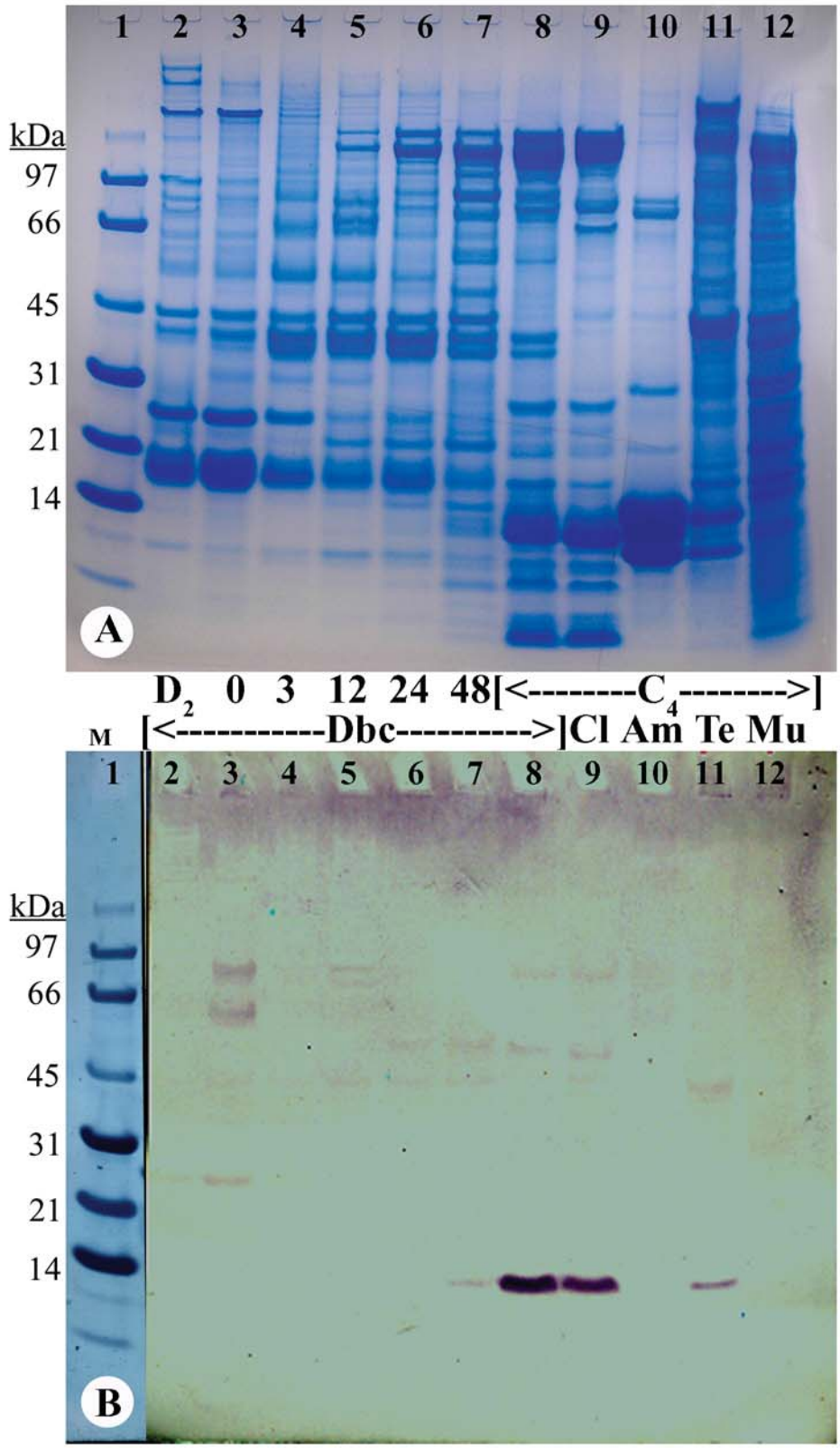

1 2 3 4 5 6 7 8 9 10 11
RNA Blot
1 2 3 4 5 6 7 8 9 10 11
1GN
Kb
1.5
1.0
0.5
3
2
1
D2 3 12 48 C4
M C A C A C A C A C A C A
1 2 3 4 5 6 7 8 9 10 11
1SP
Kb
1.5
1.0
0.5
1
D2 3 12 48 C4
M C A C A C A C A C A C A
1 2 3 4 5 6 7 8 9 10 11
3SP
Kb
1.5
1.0
0.5
3
2

times before and after the initiation of calcification and the deposition of the endocuticle. The proteins were extracted from the acetic acid-treated tissues with 6M urea in 0.1% TFA as in previous studies (Kragh et al., 1997; Andersen, 1998, 1999; Nousiainen et al., 1998). However, the present study used different tissues and cuticle types.

Coomassie staining of proteins indicated a loss of certain major protein bands and a gain of others from pre- to post- to anecdysis (Fig. 6A). For example, there is a decrease of 25 and 18 kDa bands and a high molecular weight (MW) band, while a 12 and 40 kDa band and several low MW bands increase. The arthrodial membrane and calcified cuticle proteins of anecdysial crabs show quite different banding patterns; whereas, the tendon and muscle proteins are more homogenous.

Western analysis showed strong binding of the 18-residue motif antibody to dorso-branchial and claw cuticle proteins at 12 kDa during anecdysis but no binding to arthrodial membrane proteins (Fig. 6B). None of the proteins from pre-ecdysial (D_2), and 0, 3, 12, and 24-h post-ecdysial cuticle were detected, but weak binding did occur in 48-h post-ecdysial cuticle. No binding occurred in muscle proteins, but interestingly, weak binding did occur in tendon proteins at the same molecular weight as the cuticle proteins.

Figure 6. Gel electrophoresis (A) of protein extracts from various cuticle and tissue types in *C. sapidus*, and the Western analysis (B) of the presence of the 18-residue motif in those proteins. Lane 1 is the molecular weight marker. Lane 2 is pre-ecdysial (D_2) proteins from the calcified, dorso-branchial cuticle (Dbc). Lane 3 is from the Dbc of a crab right after ecdysis (0). Lane 4 is from the Dbc of a 3 h post-ecdysis crab. Lane 5 is from the Dbc of a 12 h post-ecdysis crab. Lane 6 is from the Dbc of a 24 h post-ecdysis crab. Lane 7 is from the Dbc of a 48 h post-ecdysis crab. Lane 8 is from the Dbc of an anecdysial (C_4) crab. Lane 9 is from the claw cuticle (Cl) of an anecdysial crab. Lane 10 is from the arthrodial membrane (Am) of an anecdysial crab. Lane 11 is from the tendon (Te) of an anecdysial crab. Lane 12 is from the muscle (Mu) of an anecdysial crab.


1 2 3 4 5 6 7 8 9 10 11 12
kDa
97
66
45
31
21
14
A
D2 0 3 12 24 48[<--------C4-------->]
M [<----------Dbc---------->]Cl Am Te Mu
1 2 3 4 5 6 7 8 9 10 11 12
kDa
97
66
45
31
21
14
B

DISCUSSION

This study investigated the molecular mechanisms associated with the spatial and temporal control of calcification in the *Callinectes sapidus* exoskeleton. An 18-residue motif (xLxGPSGφφxxDGxxxQφ) was observed in proteins from the calcified cuticle of some decapod crustaceans, but this motif was not observed in proteins from the flexible, noncalcified arthrodial membrane found at the joints (Kumari et al., 1995; Kragh et al., 1997; Andersen, 1998, 1999; Nousiainen et al., 1998). An antibody made against the 18-residue motif showed similar patterns in immunohistochemical staining of the cuticle from *C. sapidus* (Fig. 1; Hequembourg, 2002). Thus, the 18-residue motif (xLxGPSGφφxxDGxxxQφ) was a prime candidate for the molecular analysis of calcification in *C. sapidus*.

Eleven cDNAs that code for multiple, variant copies of the 18-residue motif (xLxGPSGφφxxDGxxxQφ) have been cloned and characterized from the hypodermis of early post-ecdysial, calcifying cuticle. The copies of this motif account for ~70% of the total amino acid residues in each ORF. The inferred translations share strong homology with the previously identified, calcified cuticle proteins containing the 18-residue motif. However, the full-length proteins from *Homarus americanus* (Kragh et al., 1997; Nousiainen et al., 1998), and *Cancer pagurus* (Andersen, 1999) are much smaller than the translated products of the cDNAs reported here.

In addition to the highly repeated 18-residue motif, a 4-residue motif (RxKR) exists one to five times in each of the cDNA translations. These four residues are not present in any of the directly-sequenced *C. pagurus*, *H. americanus*, and *G. lateralis* proteins. Therefore, the working hypothesis is that all the open reading frames code for pro-proteins that are post-translationally

cleaved into smaller peptides. Each cleaved peptide contains either two or four copies of the 18-residue motif, and it is these peptides that are the *C. sapidus* homologs of the *C. pagurus*, *H. americanus*, and *G. lateralis* proteins. The cleavage appears to be mediated by a trypsin-like serine protease, because this protease is known to cleave proteins at the C-terminal side of arginine and lysine.

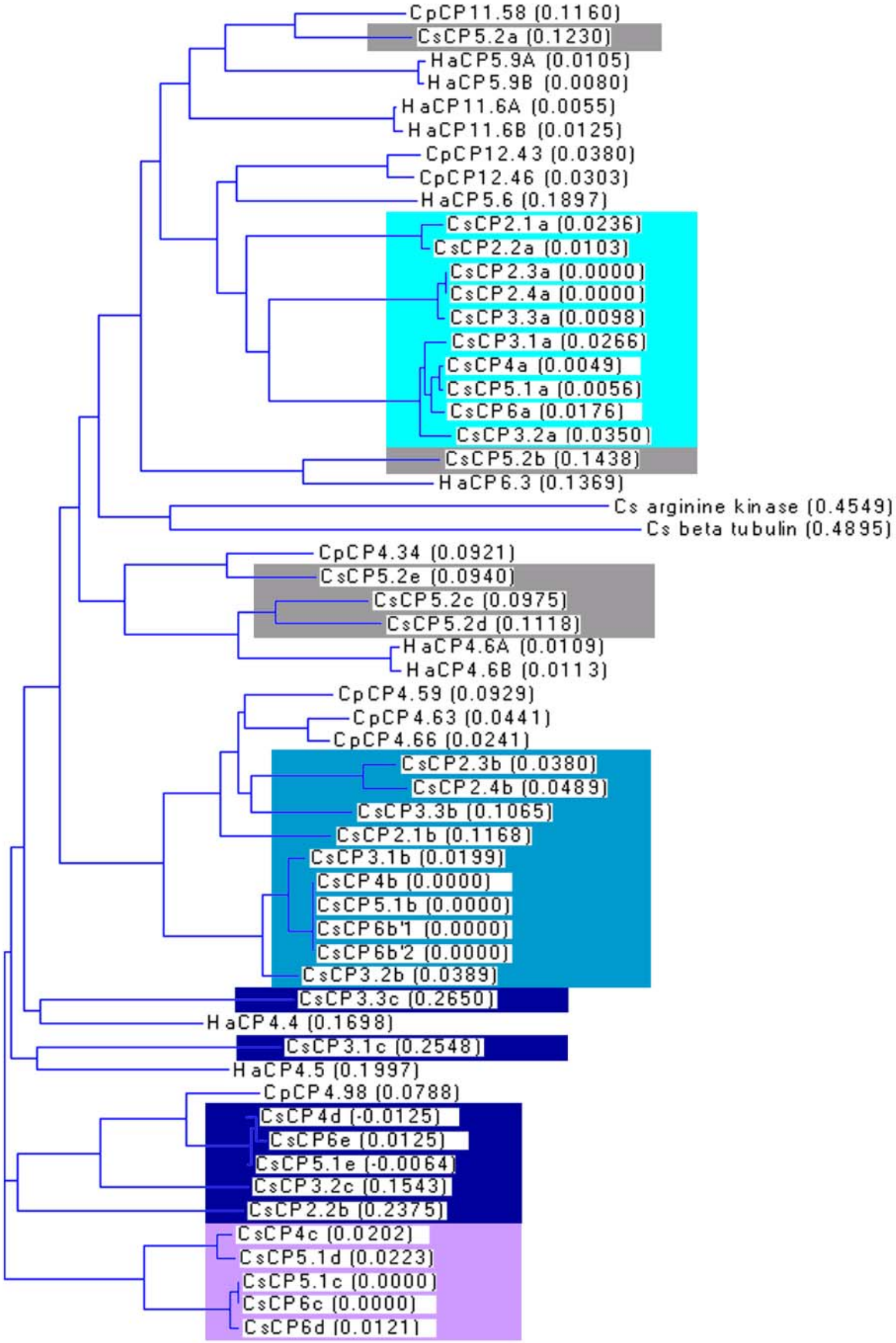
Several serine proteases have been characterized in crustaceans and function in digestion (Vanwormhoudt et al., 1995; Klein et al., 1996; Roy et al., 1996; Hernandez-Cortes et al., 1997), immunity (Soderhall and Cerenius, 1998; Kanost, 1999), or sensory perception (Levine et al., 2001). Specifically, serine proteases with trypsin-like and CUB domains are often secreted and can function in the processing of extracellular proteins (Bond and Beynon, 1995). Levine et al. (2001) cloned and characterized *csp* from the olfactory organ of the spiny lobster *Panulirus argus*, which encodes a hormonally regulated CUB-serine protease. They speculated that this protease could be involved in molting by aiding in the enzymatic breakdown of the old cuticle during pre-ecdysis or by degrading cuticular anchoring proteins.

Another 4-residue (RYRR) recognition site for enzymatic digestion was found in the open reading frame of the *Galaxin* transcript from the calcified exoskeleton of a reef coral (Fukuda et al, 2003). Additionally, the CAP-1 transcript codes for a C-terminal arginine and lysine that are not present in the mature CAP-1 protein (Inoue et al., 2003). A recombinant protein with the two, paired basic residues from the CAP-1 cDNA showed higher anti-calcification activity than a recombinant protein without them. It was suggested that these two residues are removed by carboxypeptidase B or E and might contribute to a change in protein conformation leading to easier association with the calcium carbonate microcrystal.

Assuming cleavage occurs, I have virtually cleaved the *C. sapidus* pro-proteins into their respective peptides and aligned them with the full-length *C. pagurus* and *H. americanus* cuticle proteins that contain the 18-residue motif. Due to the large number of sequences, the alignments are not shown, but rather the hierarchical clustering they produce. This clustering is meant to show sequence similarity and not imply phylogeny (Fig. 7). The *C. sapidus* sequences appear to group into five main clusters. The "a" and "b" peptides each group into one distinct cluster. Most of the C-terminal peptides cluster together, and some of the inner peptides from CsproCP4, CsproCP5.1, and CsproCP6 cluster together. The peptides from CsproCP5.2 cluster together and are distinct from the other *C. sapidus* sequences. In general, the *C. sapidus* peptides are more homologous with themselves than with the *C. pagurus* and *H. americanus* peptides, as expected, and most of the *C. sapidus* peptide groups have corresponding *C. pagurus* and/or *H. americanus* proteins.

The high number of homologous repeat regions and the repeated region present in *CSproCP6* suggested possible alternate splicing. Genomic DNA products corresponding to *CsproCP3.1* and *CsproCP3.2* were amplified, and both sequences contain different sized introns that interrupt the signal peptide. This result and the fact that no introns divide the open reading frames between the similar peptides imply that these two transcripts came from two separate genes. Although alternate splicing has not been determined for the other transcripts, it appears that the transcripts encoding proteins with the 18-residue motif (xLxGPSGφφxxDGxxxQφ) are from a family of genes. The cuticle

Figure 7. Hierarchical clustering of the virtually cleaved *C. sapidus* translations (CsCPs) and all 18-residue motif-containing proteins from *C. pagurus* (CpCP#) and *H. americanus* (HaCP#). *C. sapidus* sequences for arginine kinase and beta-tubulin were used as outgroups. Same color denotes sequence similarity.


CpCP11.58 (0.1160)
CsCP5.2a (0.1230)
HaCP5.9A (0.0105)
HaCP5.9B (0.0080)
HaCP11.6A (0.0055)
HaCP11.6B (0.0125)
CpCP12.43 (0.0380)
CpCP12.46 (0.0303)
HaCP5.6 (0.1897)
CsCP2.1a (0.0236)
CsCP2.2a (0.0103)
CsCP2.3a (0.0000)
CsCP2.4a (0.0000)
CsCP3.3a (0.0098)
CsCP3.1a (0.0266)
CsCP4a (0.0049)
CsCP5.1a (0.0056)
CsCP6a (0.0176)
CsCP3.2a (0.0350)
CsCP5.2b (0.1438)
HaCP6.3 (0.1369)
Cs arginine kinase (0.4549)
Cs beta tubulin (0.4895)
CpCP4.34 (0.0921)
CsCP5.2e (0.0940)
CsCP5.2c (0.0975)
CsCP5.2d (0.1118)
HaCP4.6A (0.0109)
HaCP4.6B (0.0113)
CpCP4.59 (0.0929)
CpCP4.63 (0.0441)
CpCP4.66 (0.0241)
CsCP2.3b (0.0380)
CsCP2.4b (0.0489)
CsCP3.3b (0.1065)
CsCP2.1b (0.1168)
CsCP3.1b (0.0199)
CsCP4b (0.0000)
CsCP5.1b (0.0000)
CsCP6b'1 (0.0000)
CsCP6b'2 (0.0000)
CsCP3.2b (0.0389)
CsCP3.3c (0.2650)
HaCP4.4 (0.1698)
CsCP3.1c (0.2548)
HaCP4.5 (0.1997)
CpCP4.98 (0.0788)
CsCP4d (-0.0125)
CsCP6e (0.0125)
CsCP5.1e (-0.0064)
CsCP3.2c (0.1543)
CsCP2.2b (0.2375)
CsCP4c (0.0202)
CsCP5.1d (0.0223)
CsCP5.1c (0.0000)
CsCP6c (0.0000)
CsCP6d (0.0121)

protein genes of insects often have their signal peptides interrupted by introns as well (reviewed in Willis, 1996). These insect genes also tend to be tandemly arranged on the chromosome (Willis, 1999). Therefore, because of the intron location and the equal gene expression, it is possible that the *CsproCP* genes are also found in clusters under the control of one regulator. As yet, I have no evidence to support this speculation.

Northern analysis shows that many *CsproCP* genes are only expressed in the hypodermis of cuticle that is in the process of calcification. Expression does not occur in the hypodermis of cuticles that are not in the process of calcification, which include the pre-ecdysial and anecdysial cuticle, as well as the arthrodial membrane. Gene expression begins at ~3 h post-ecdysis when calcification is initiated. Expression is low at 3 and 12 h post-ecdysis, but increases dramatically at 48 h post-ecdysis. The expression during post-ecdysis and the lack of expression in the arthrodial membrane is consistent with the immunohistochemical staining performed by Hequembourg (2002).

Multiple probes for Northern blotting were used in this investigation to attempt to determine the total number and sizes of transcripts as well as whether or not there are any expression differences between those transcripts. Three probes, amplified from existing 3' RACE or RT-PCR cDNA, were initially designed to recognize one, a subset, or all of the transcripts encoding the 18-residue motif. Hybridization of probe 1GN, designed to recognize all of the transcripts, produced three bands whose sizes agreed well with the transcripts *CsproCP3.1* and *CsproCP3.2*, *CsproCP5.1*, and *CsproCP6*. *CsproCP3.1* and *CsproCP3.2* have very similar sizes; therefore, one band could correspond to two transcripts. Although probe 1GN was originally designed to recognize all the transcripts from the *CsproCP* gene family, the additional transcripts, *CsproCP2.1-2.4*, *CsproCP5.2*, and *CsproCP3.3*, were acquired after probe

1GN was designed and implemented. Sequence analysis suggests that the probe probably would not recognize these additional transcripts. Probe 1SP was designed to be specific for *CsproCP3.1*, and it produced one band of the same size as one band in the RNA blot probed with 1GN. The size agreed well with *CsproCP3.1* and *CsproCP3.2*; however, probe 1SP probably only recognized *CsproCP3.1*. Probe 3SP was designed to recognize *CsproCP5.1* and *CsproCP6*. 3SP detected two bands, also present in the 1GN Northern, whose sizes agreed well with *CsproCP5.1* and *CsproCP6*. Additional Northern blotting needs to be performed with a more general probe that will recognize all the transcripts. The equal expression of the detected transcripts supports the notion that the *CsproCP* genes are under the control of one regulator; a common feature in insects.

Previous studies on the electrophoretic patterns of the pre- and post-ecdysial, calcified cuticle proteins have concentrated on the soluble proteins from stage D_3 to 5 h post-ecdysis (Shafer et al., 1994, 1995; Coblentz et al., 1998; Tweedie, 2001). In this study, 10% acetic acid was used to decalcify the tissues and remove the soluble proteins, and the resulting residues were subjected to 6M urea in 0.1% TFA to extract the insoluble proteins. Additionally, the study examined a broader time course from pre- and post-ecdysial crabs that encompassed D_2 to 48 h post-ecdysis as well as anecdysial crabs. More tissues including the arthrodial membrane, tendon, and muscle were also examined. Coomassie staining of fractionated proteins showed some interesting patterns including a loss and gain of 12-25 kDa and high MW proteins and specifically, a gain in low MW ($\leq$12 kDa) proteins. The arthrodial membrane and calcified cuticle protein patterns of anecdysial crabs are extremely different as expected by the different functions of each cuticle type. The protein banding patterns are not the focus of this investigation, and any further analysis would be mere speculation.

Western blotting was used to compare the immunohistochemical staining of the proteins from anecdysial, claw cuticle and arthrodial membrane. Consistent with Hequembourg (2002), the antibody strongly detected proteins in the calcified cuticle but did not bind to arthrodial membrane proteins. Binding of the antibody to anecdysial claw and dorso-branchial cuticle showed that there were minimal differences in protein content from the different calcified cuticle regions of the crab. Therefore, it assumed that all calcified cuticle includes proteins containing the 18-residue motif. Although gene expression levels are high at 48 h post-ecdysis, only weak binding of the antibody occurred at 48 h post-ecdysis. Apparently the protein has not had time to accumulate within the cuticle. By anecdysis, the proteins have accumulated in the cuticle; therefore, it is not necessary for crabs to express these genes at this stage. The increase of the 12 kDa and low MW proteins seen in PAGE strongly correlates with the increase in signal of the 12 kDa band in Western blotting. Thus, it is speculated that the 12 kDa PAGE bands are proteins that contain the 18-residue motif. Given the predicted sizes of the cleaved peptides, 12 kDa for four 18-residue motif-containing peptides and 5 kDa for two motif-containing peptides, it was expected that the antibody would have detected both sizes. However, only the 12 kDa peptides were detected. The peptide with the sequence VLVGPSGIVTSDGQNVQF used as the antigen is more similar to the 5 kDa cleaved peptides with two 18-residue motifs than to the 12 kDa peptides. It is unclear at this time why the antibody only recognized the larger peptides.

Muscle and tendon tissues were intended to be negative controls. The antibody did not bind to muscle proteins as expected; however, the antibody weakly recognized proteins from the tendon at the same MW as the calcified cuticle. Yamaguchi et al. (2003) determined that crab tendon is composed of 42% calcium phosphate. The results suggest that proteins with the 18-

residue motif are involved in multiple aspects of decapod crustacean calcification and are not limited to calcium carbonate mineralization in the cuticle.

It should be noted that proteins containing the 18-residue motif were isolated from all four layers of the *G. lateralis* cuticle (Kumari et al., 1995), and electron microscopy studies used colloidal-gold labeled antibodies specific to the 18-residue motif that detected antigens in the epi- and exocuticle as well as the endocuticle (Hequembourg, 2002). These results are in contrast to both the immunohistochemical staining from light microscopy (Fig. 1; Hequembourg, 2002) and the results found here. Some genes encoding proteins with the 18-residue motif might be expressed at earlier times than those found in this study, and thus explaining the presence in the other cuticle layers. Alternatively, some proteins might be transported to the outer layers through the epithelial projections that are intertwined within the cuticle by way of the pore canals (Roer and Dillaman, 1984; 1993). Regardless, the elevated gene expression at 48 h post-ecdysis and the immunohistochemical staining (Fig. 1; Hequembourg, 2002) show that proteins with the 18-residue motif are at least concentrated in the endocuticle.

Various investigations in this laboratory have examined the proteins that are involved in the initial calcification of the exocuticle. The proteins encoded by the *CsproCP* genes might be involved in initial calcification because expression coincides with its onset. However, the post-ecdysis expression of the genes does not fit with our model of large, shielding proteins that are degraded to expose nucleation sites present in the pre-ecdysial layers (Coblentz et al., 1998). The genes investigated here coding for the 18-residue motif are not present in the new cuticle before ecdysis and so cannot be the proposed nucleation sites. The function of the CsproCP proteins may be initiated when, after their synthesis, a trypsin-like serine protease cleaves the pro-proteins at the last arginine of the 4-residue motif, producing their respective peptides. Then

the four residues (RxKR) are removed by a constitutive enzyme such as carboxypeptidase B or E, similar to CAP-1 (Inoue et al., 2003), to cause an advantageous conformational change in the peptide. The regular spacing of predicted β-sheets (data not shown), corresponding to the repeated 18-residue motif, may provide Ca^{2+} binding sites or interact with the crystal faces after nucleation to orient growth. In conclusion, it has been shown that the *CsproCP* gene family encodes proteins with a highly repeated structure that are involved in calcification of mineralized structures in the blue crab, *Callinectes sapidus*. Additional work, such as functional studies through RNAi and *in vitro* mineralization assays, needs to be performed on these genes and proteins to further understand their specific roles in calcification.

LITERATURE CITED

Andersen, S. O. (1999). Exoskeletal proteins from the crab, *Cancer pagurus*. *Comp. Biochem. Physiol.* **123A**, 203-211.

Andersen, S. O. (1998). Characterization of proteins from arthrodial membranes of the lobster, *Homarus americanus*. *Comp. Biochem. Physiol.* **99B,** 453-458.

Andersen, S. O., Højrup, P. and Roepstorff, P. (1995). Insect cuticular proteins. *Insect Biochem. Molec. Biol.* **25,** 153- 176.

Barreto, F. S. (2003). Assortative mating as a barrier to gene flow in a coral reef fish species flock. *M. S. Thesis*, University of North Carolina at Wilmington.

Bond, J. S. and Beynon, R. J. (1995). The astacin family of metalloendopeptidases. *Protein Sci.* **4**, 1247-1261.

Buda, E. S. (2004). Prophenoloxidase activating factor from the blue crab, *Callinectes sapidus*. *M. S. Thesis*, University of North Carolina at Wilmington.

Coblentz, F. E., Shafer, T. H. and Roer, R. D. (1998). Cuticular proteins from the blue crab alter *in vitro* calcium carbonate mineralization. *Comp. Biochem. Physiol.* **121B,** 349-360.

Cochran, D. M. (1935). The skeletal musculature of the blue crab, *Callinectes sapidus* Rathbun. *Smithsonian Miscellaneous Collections*. **92(9)**, 1-76.

Drach, P. (1939). Mue et cycle d'intermue chez les crustacéés decapodes. *Ann. Inst. Oceanogr., Monaco.* **19**, 103-391.

Endo, H., Persson, P. and Watanabe, T. (2000). Molecular cloning of the crustacean DD4 cDNA encoding a Ca^{2+}-binding protein. *Biochem. Biophys. Res. Com.* **276,** 286-291.

Fukuda, I., Ooki, S., Fujita, T., Murayama, E., Nagasawa, H., Isa, Y. and Watanabe, T. (2003). Molecular cloning of a cDNA encoding a soluble protein in the coral exoskeleton. *Biochem. Biophys. Res. Com.* **304**, 11-17.

Hequembourg, S. J. (2002). Early patterns of calcium and protein deposition in the carapace of the blue crab. *M.S. Thesis*, University of North Carolina at Wilmington.

Hernadez-Cortes, P., Whitaker, J. R. and Garcia-Carreno, F. L. (1997). Purification and characterization of chymotrypsin from *Penaeus vannamei* (Crustacea-Decapoda). *J. Food Biochem.* **21**, 497-514.

Inoue, H., Moriaki, O. and Nagasawa, H. (2001). Purification and structural determination of a phosphorylated peptide with anti-calcification and chitin-binding activities in the exoskeleton of the crayfish, *Procambarus clarkii*. *Biosci. Biotechnol. Biochem.* **65**, 1840-1848.

Inoue H., Ohira, T., Ozaki, N. and Nagasawa, H. (2003). Cloning and expression of a cDNA encoding a matrix peptide associated with calcification in the exoskeleton of the crayfish. *Comp. Biochem. Physiol.* **136B**, 755-765.

Kanost, M. R. (1999). Serine protease inhibitors in arthropod immunity. *Dev. Comp. Immuno.* **23**, 291-301.

Klein, B., LeMoullac G., Sellos, D. and Van Wormhoudt A. (1996). Molecular cloning and sequencing of trypsin cDNAs from *Peneaus vannamei* (Crustacea-Decapoda): use in assessing gene expression during the molt cycle. *Int. J. Biochem Cell Biol.* **28**, 551-563.

Kragh, M., Mølbak, L. and Andersen, S. O. (1997). Cuticular proteins from the lobster, *Homarus americanus*. *Comp. Biochem. Physiol.* **118B**, 147-154.

Kumari S. S., Willis, J. H. and Skinner, D. M. (1995). Proteins of crustacean exoskelton: IV. Partial amino acid sequences of exoskeletal proteins from the Bermuda land crab, *Gecarcinus lateralis*, and comparisons to certain insect proteins. *J. Exp. Zool.* **273**, 389-400.

Levine, M. Z., Harrison, P. J. H., Walthall, W. W., Tai, P. C. and Derby, C. D. (2001). A CUB-serine protease in the olfactory organ of the spiny lobster *Panulirus argus*. *J. Neurobiol.* **49**, 277-302.

Lowenstam, H. A. and Weiner, S. (1989). *On Biomineralization*. New York, NY: Oxford University Press.

Mann, S., Webb, J. and Williams, R. J. P. (1989). *Biomineralization*. New York, NY: VCH Publishers.

Neville, A. C. (1975). Biology of the arthropod cuticle. pp. 17-19. New York: Springer.

Nousiainen, M., Rafn, K., Shou, L., Røepstorff, P. and Andersen, S. O. (1998). Characterization of exoskeletal proteins from the American lobster, *Homarus americanus*. *Comp. Biochem. Physiol.* **119B,** 189-199.

Roer, R. D., Halbrook, K. E. and Shafer, T. H. (2001). Glycosidase activity in the post ecdysial cuticle of the blue crab, *Callinectes sapidus*. *Comp. Biochem. Physiol*. **128B**, 683-690.

Roer, R. D. and Dillaman, R. M. (1993). Molt-related change in integumental structure and function. In *The Crustacean Integument – Morphology and Biochemistry* (ed. M. N. Horst and J. A. Freeman), pp 1-37. Boca Raton, FL: CRC Press.

Roer, R. D., Burgess, S. K., Miller, C. G. and Dail, M. B. (1988). Control of calcium carbonate in nucleation in pre- and postecdysial crab cuticle. In *Chemical Aspects of Regulation of Mineralization* (ed. C. S. Sikes and A. P. Wheeler), pp. 21-24. Mobile, AL: University of South Alabama Publication Service.

Roer, R. D. and Dillaman, R. M. (1984). The structure and calcification of the crustacean cuticle. *Amer. Zool.* **24**, 893-909.

Roy, P., Colas, B. and Durand, P. (1996). Purification, kinetical, and molecular characterizations of a serine collagenolytic protease from green shore crab (*Carcinus maenas*) digestive gland. *Comp. Biochem. Physiol.* **115B**, 87-95.

Sambrook, J. and Russell, D. W. (2001). *Molecular cloning: a laboratory manual*. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Shafer, T. H., Roer, R. D., Midgette-Luther, C. and Brookins, T. A. (1995). Postecdysial cuticle alteration in the blue crab, *Callinectes sapidus*: synchronous changes in glycoproteins and mineral nucleation. *J. Exp. Zool*. **271**, 171-182.

Shafer, T. H., Roer, R. D., Miller, C. G. and Dillaman, R. M. (1994). Postecdysial changes in the protein and glycoprotein composition of the cuticle of the blue crab, *Callinectes sapidus*. *J. Crustacean Biol.* **14**, 210-219.

Soderhall, K. and Cerenius, L. (1998). Role of the prophenoloxidase-activating system in invertebrate immunity. *Curr. Opin. Immunol.* **10**, 23-28.

Tweedie, E. P., Coblentz, F. E. and Shafer, T. H. (2004). Purification of a soluble glycoprotein from the uncalcifiecd ecdsysial cuticle of the blue crab *Callinectes sapidus*. *J. Exp. Biol.* In press.

Van Wormhoudt, A., Sellos, D. and Donval A. (1995). Chymotrypsin gene-expression during the intermolt cycle in the shrimp *Peneaus vannamei* (Crustacea-Decapoda). *Experientia*. **51**, 159-163.

Williams, C. L., Dillaman, R. M., Elliott, E. A. and Gay, D. M. (2003). Formation of the arthrodial membrane in the blue crab, *Callinectes sapidus*. *J. Morph.* **256**, 260-269.

Willis, J. H. (1999). Cuticular proteins in insects and crustaceans. *Amer. Zool.* **39**, 600-609.

Willis, J. H. (1996). Metamorphosis of the cuticle, its proteins, and their genes. In *Metamorphosis/post-embryonic reprogramming of gene expression in amphibian and insect cells* (ed. L. I. Gilbert, B. G. Atkinson and Tata, J.), pp 253-282. Orlando, FL: Academic Press, Inc.

Yamaguchi, I., Itoh, S., Suzuki, M., Sakane, M., Osaka, A. and Tanaka, J. (2003). The chitosan prepared from crab tendon I: the characterization and the mechanical properties. *Biomaterials.* **24**, 2031, 2036.