Volume 292, number 1,2, 115–120 FEBS 10325 November 1991
© 1991 Federation of European Biochemical Societies 00145793/91/$3.50
ADONIS 001457939101035T

Molecular cloning of three cDNAs that encode cysteine proteinases in the digestive gland of the American lobster (*Homarus americanus*)

Maurice V. Laycock, Ron M. MacKay, Marco Di Fruscio and Jeffrey W. Gallant

Institute for Marine Biosciences, National Research Council of Canada, 1411 Oxford Street, Halifax, Nova Scotia B3H 3Z1, Canada

Received 10 June 1991

Three clones were isolated from a lobster digestive gland cDNA library, using oligonucleotide probes based on the partial amino terminal sequence of a digestive cysteine proteinase. The cDNAs, LCP1, LCP2 and LCP3 encode preproenzymes of 322, 323 and 321 amino acid residues, and putative mature enzymes of 217, 216 and 215 residues, respectively. Calculated mature protein molecular masses are 23386 (LCP1), 23093 (LCP2) and 23255 (LCP3) Sequence alignments show that the lobster enzymes are more similar to L (55–62% identity) than H (42–44%) or B (22–24%) cathepsins. Southern analysis indicated as many as eleven genes related to the three cDNAs.

Cysteine proteinase; cDNA cloning; Amino acid sequence; Lobster

1. INTRODUCTION

Cysteine proteinases account for 80% of the proteolytic activity in the digestive fluid of the American lobster and one of these enzymes has been isolated and characterized [1]. This enzyme is surprisingly similar to papain, a cysteine proteinase of the papaya fruit, not only in its kinetic properties, but, of the first 29 residues of the amino-terminal sequence, only 7 are different from papain. The lobster enzyme, however, was reported to have a mass of 28 kDa compared to 23 kDa for papain by denaturing gel electrophoresis, and an isoelectric point of 4.5, compared to 8.8. Although papain is the most well-characterized cysteine proteinase, more than 30 complete and partial sequences of this group of enzymes are known. Despite the wide diversity of sources, amino-terminal sequences, including the active site with the essential cysteine residue, all show a high degree of similarity to each other [2,3]. Digestive cysteine proteinases in the lobster may be related to cathepsins, the acidic cysteine proteinases of vertebrate lysosomes. Comparisons of complete sequences should indicate the evolutionary relationships of the digestive enzymes. The deduced amino acid sequences will also provide more accurate molecular weights and show whether or not a preproenzyme is produced, as for many other digestive proteinases and transported enzymes [4,5]. In this work, we report on the molecular cloning of three lobster cysteine proteinase mRNA species, the structure of the encoded proteins and an analysis of the lobster cysteine proteinase gene family.

2. MATERIALS AND METHODS

2.1. *Procedures for molecular cloning of lobster cysteine proteinase cDNA*

Routine molecular cloning techniques were used [6]. Lobsters were obtained locally and killed by severing the cerebral ganglia. The hepatopancreas was removed, frozen in liquid nitrogen and 10 g of tissue ground thoroughly with a mortar and pestle. RNA was prepared by the method of Turpen and Griffith [7] with the modification that the RNA product from the centrifugation step was extracted with phenol/chloroform before precipitation with ethanol (to yield 2.7 mg). Poly(A)$^+$ RNA (40 μg) was selected by oligo(dT)-cellulose column chromatography [6]. cDNA was prepared from 2.5 μg of poly(A)$^+$ RNA (cDNA Synthesis System, Bethesda Research Laboratories), treated with T4 DNA polymerase (Bethesda Research Laboratories) and ligated into λgt10 [8]. A library of 5.7 × 10^6 PFU was generated. Various oligodeoxynucleotide probes, whose sequences were derived from the partial amino acid sequence of the mature digestive proteinase (see LCP1 in Fig. 2), were phosphorylated with [γ-^{32}P]ATP (3000 Ci/mmol) and tested for specificity in Northern analyses [6] of poly (A)$^+$ RNA (data not shown). Two oligonucleotide mixtures:

no. 6: ($^{G}_{A}\mathrm{AA}^{T}_{G}\mathrm{GCCCAGCA}^{TGAT}_{GCTG}\mathrm{CCGCATTG}^{T}_{G}\mathrm{CCTTGGTC}^{T}_{C}\mathrm{TT}$;

corresponding to positions 16–27 of the mature protein) and no. 7

($\mathrm{GC}^{T}_{G}\mathrm{CC}^{T}_{G}\mathrm{TT}^{T}_{G}\mathrm{GT}^{T}_{G}\mathrm{CCAGTC}^{T}_{G}\mathrm{ACTTC}^{T}_{G}\mathrm{GT}$;

positions 2–11) were chosen to screen the library and to confirm clone identity. Plaque lifts (C/P Lift, Bio-Rad) were incubated for 16 h at 40°C with probe no. 6 (4 ng/ml) in 3 × SSC, 20 mM NaH_2PO_4, pH 7.0, 1% SDS, 10 × Denhardt's solution, 100 μg of sheared and denatured salmon sperm DNA per ml [6]. Before autoradiography, membranes were washed in 3 × SSC, 0.1% SDS at 40°C. Insert DNAs, detected in Southern hybridizations by both no. 6 and no. 7 probes, were subcloned in M13mp18 and nucleotide sequences were determined by the dideoxy method [6].

Correspondence address: M.V. Laycock, Institute for Marine Biosciences, National Research Council of Canada, 1411 Oxford Street, Halifax, Nova Scotia B3H 3Z1, Canada. Fax: (1) (902) 426 9413.

The truncated nature of the initially identified cDNAs (see Results and Discussion) required re-screening of the library. A 0.9 kb insert was purified by electrophoresis through low melting temperature agarose and radiolabelled with [α-^{32}P]dCTP (3000 Ci/mmol) by the random priming method [6]. Hybridization conditions were as described above, except that 6× replaced 3× SSC, the incubation temperature was 68°C, and the final washes were in 0.2 × SSC. The completion of the LCP1 sequence required direct sequencing of poly(A)$^+$ RNA (9,10) with an LCP1-specific primer (see Fig. 2).

2.2. *Isolation of DNA and Southern analysis*

Genomic DNA was prepared from the CsCl gradient that was used to purify RNA (see above). A viscous fraction located at the top of the high density cushion was removed, dialyzed and concentrated by butanol extractions. DNA was precipitated from 2.5 M ammonium acetate with ethanol, redissolved in 10 mM Tris-HCl, pH 8.0, 1 mM EDTA and stored at 2°C over chloroform. DNA was dialyzed on floating Millipore (0.025 μm) membranes, digested with the restriction enzyme *Pst*I (25 units per μg, 16 h), electrophoresed through a 0.8% agarose gel and transferred to a Zeta-Probe (Bio-Rad) membrane according to the manufacturer's alkaline blotting protocol. Preparation of probe DNAs and hybridization conditions were described above for library re-screening. Membrane washing conditions are described in the legend to Fig. 4.

3. RESULTS AND DISCUSSION

3.1. *Isolation and identification of cDNA clones*

The majority of the bacteriophage clones isolated by the screening of the cDNA library with oligonucleotides no. 6 and no. 7 carried insert DNAs of 0.9 kb. When sequenced, one of these was found to encode the entire mature digestive enzyme sequence from the carboxy-terminus to 25 amino acids beyond the amino-terminus of the mature protein. Partial sequences of several other 0.9 kb inserts were identical to the one taken to completion. An *Eco*RI* site was found at the 5' end of this cDNA (see LCP1 in Fig. 1). The deduced amino acid sequence matched that of the purified enzyme obtained by the Edman method, except at residue 9, where lysine was predicted instead of glutamate as in the published sequence [1].

Rescreening the library with the 0.9 kb insert yielded 67 clones with inserts of 0.9 kb, four of 1.1 kb and two of 1.2 kb. Sequencing a 1.1 kb cDNA showed it to be identical to the 0.9 kb cDNA, but extending 171 bases beyond the *Eco*RI* site to a natural *Eco*RI site that had escaped methylation during the preparation of the library (see LCP1 in Fig. 1). The remainder of the upstream sequence of LCP1 was obtained by direct sequencing of mRNA. A different sequence from that of LCP1 was obtained from one of the 1.2 kb cDNAs (LCP2). No internal *Eco*RI sites were encountered, allowing sequencing beyond an initiation codon. A third cDNA (LCP3), with a sequence distinct from both LCP1 and LCP2, was found among the 1.1 kb inserts. This also lacked any internal *Eco*RI sites and was sequenced up to a presumptive initiation codon, as shown in Fig. 2. Restriction maps of the three cDNAs are presented in Fig. 1.

3.2. *Sequence alignments and possible relationships of the lobster enzymes*

The complete amino acid sequences for the three lobster cysteine proteinases were aligned with published sequences of papain [11] and rat cathepsins L, H and B [12] as shown in Fig. 3. Each of the lobster polypeptides has a signal peptide consisting of a consecutive sequence of 14 hydrophobic amino acids near the amino-terminus, immediately preceded by lysine and followed by a region with both positive and negative charges, characteristic of secreted proteins. The leader sequences of the lobster enzymes show considerable similarity to papain and the cathepsins between amino acid positions

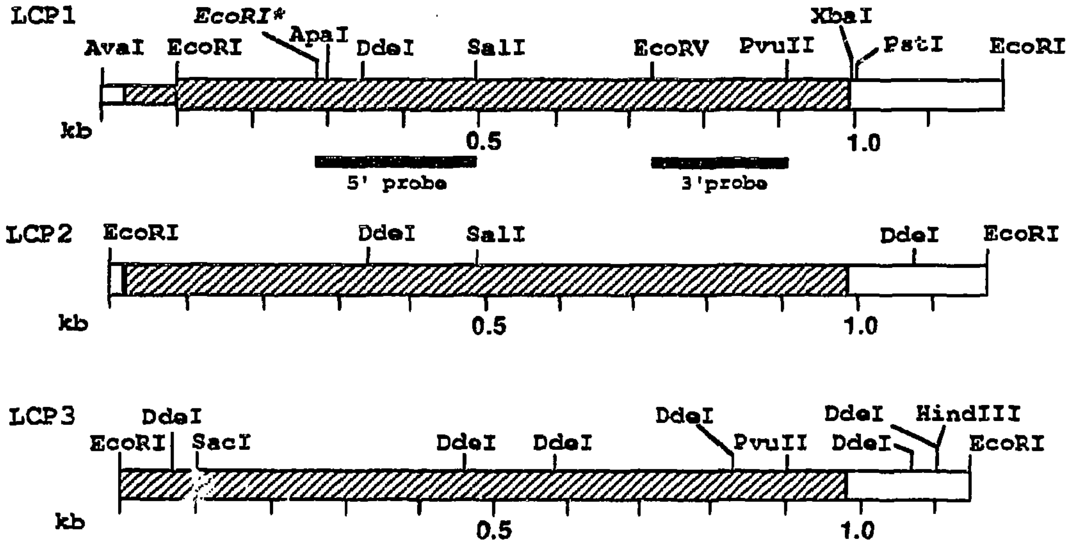

Fig. 1. Maps of the lobster cysteine proteinase cDNAs. Restriction enzyme recognition sites are those predicted by the nucleotide sequences of Fig. 2. Only the *Eco*RI* site found at the terminus of the 0.9 kb cDNA is indicated (see Results and Discussion). The 5' termini of the sense strands are at the left. Open and shaded bars represent non-coding and coding regions, respectively. The solid bars indicate restriction fragments used as probes in the Southern hybridization analysis (Fig. 4). Completion of the LCP1 sequence (narrow bar) required direct sequencing of poly (A)$^+$ RNA.

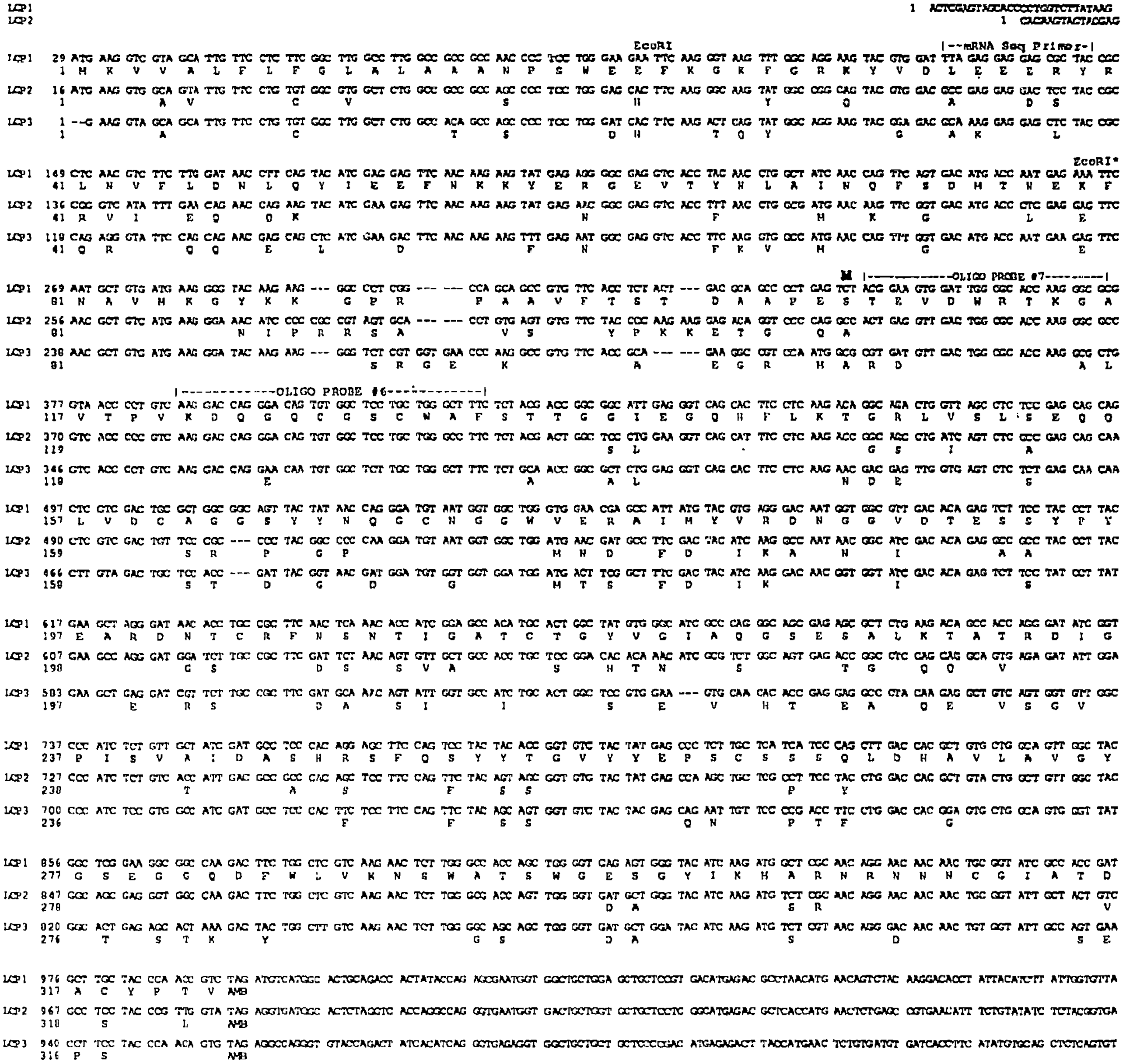
LCP1
LCP2
LCP3
EcoRI
|--mRNA Seq Primer-|
EcoRI*
M |------------OLIGO PROBE #7----------|
|------------OLIGO PROBE #6------------|

Fig. 2. Nucleotide and deduced amino acid sequences of cDNAs that encode lobster cysteine proteinases. The sense strand sequences are numbered from their 5' termini. The amino acid sequences, shown under the nucleotide sequences (completely for LCP1 and only where they differ for LCP2 and LCP3), are numbered from the putative initiating methionine residues. LCP1 nucleotide sequence 5' to the *Eco*RI site at position 92 was determined by sequencing poly(A)$^+$ RNA with a primer complementary to positions 128–145. The amino-terminus of the mature digestive proteinase is indicated by the letter M. Probes no. 6 and no. 7 (see Materials and Methods) were based on the mature enzyme amino acid sequence corresponding to the bracketed positions of LCP1.

−24 and −86. Ishidoh et al. have suggested that positions −35 to −55 of cysteine proteinase propeptides may play a role in their processing [12]. The lobster mature protein amino acid sequences retain most of the primary structural features characteristic of other cysteine proteinases, including the highly conserved region surrounding the active site cysteine residue (at position 25 in Fig. 3). However, the proline residue next to the mature amino-terminus of all other known cysteine proteinases is absent from the lobster sequences, although proline does occur at the −2 position of all three lobster polypeptides. The lobster mature proteins have more sequence identity with the L cathepsin sequence than with either H or B cathepsins from rat (Fig. 3). If

```
                     -130       -120       -110         -100        -90        -80
PAP                MAM IPSISKLLFV AICLFVYMGL SFG--DFSIVGY SQNDLTSTER LIQLFESWML KHNKIYKNID
LCP1                          MKV VALFLFGLAL AAA--NPS---- ---------- ----WEEFKG KFGRKYVDLE
LCP2                          MKV AVLFLCGVAL AAA--SPS---- ---------- ----WEHFKG KYGRQYVDAE
LCP3                           KV AALFLCGLAL ATA--SPS---- ---------- ----WDHFKT QYGRKYGDAK
RCL                          MTPL LLLAVLCLGT ALA--TPK---- ------FDQT FNAQWHQWKS THRRLYGTNE
RCH                      MWTALPLL CAGAWLLSAG ATA--ELT---- ------VNAI EKFHFTSWMK QHQKTYSSRE
RCB                          MWWS LIPLSCLLAL TSAHDKPS---- ---------- ---------- -----FHPLS

      -70        -60        -50            -40           -30        -20        -10
PAP    EKIYRFEIFK DNLKYIDETN KK----NNSYWLGL NVF--ADMSNDE FKEKYTGSIA GNYTTTELSY EEVLNDGDVN
LCP1   EERYRLNVFL DNLQYIEEFN KKYERGEVTYNLAI NQF--SDMTNEK FNAVMKGYKK GPRPAAVFTS T----DAAPE
LCP2   EDSYRRVIFE QNQKYIEEFN KKYENGEVTFNLAM NKF--GDMTLEE FNAVMKGNIP -RRSAPVSVF YPKK-ETGPQ
LCP3   EELYRQRVFQ QNEQLIEDFN KKFENGEVTFKVAM NQF--GDMTNEE FNAVMKGYKK GSRGEPKAVF TA---EGRPM
RCL    EE-WRRAVWE KNMRMIQLHN GEYSNGKHGFTMEM NAF--GDMTNEE FRQIVNGYRH QKHKKGRLFQ E----PLMLQ
RCH    YS-HRLQVFA NNWRKIQAHN QR----NHTFKMGL NQF--SDMSFAE IKHKYLWSEP QNCSATKSNY L----RGTGP
RCB    DD-------- ----MINYIN KQ----NTTWQAGR N-FYNVDIS--- YLKKPCGTVL GGPKLPERVG F----SEDIN

       1        10              20         30         40            50         60            70
PAP    IPEYVDWRQK G----AVTPVKNQG SCGSCWAFSA VVTIEGIIKI RT-GNL-NEYSE QELLDCDRRS Y---GCNGGYPWS
LCP1   STE-VDWRTK G----AVTPVKDQG QCGSCWAFST TGGIEGQHFL KT-GRL-VSLSE QQLVDCAGGS YYNQGCNGGWVER
LCP2   ATE-VDWRTK G----AVTPVKDQG QCGSCWAFST TGSLEGQHFL KT-GSL-ISLAE QQLVDC-SRP YGPQGCNGGWMND
LCP3   ARD-VDWRTK A----LVTPVKDQF QCGSCWAFSA TGALEGQHFL KN-DEL-VSLSE QQLVDC-STD YGNDGCGGGWMTS
RCL    IPKTVDWREK G----CVTPVKNQG QCGSCWAFSA SGCLEGQMFL KT-GKL-ISLSE QNLVDC-SHD QGNQGCNGGLMDF
RCH    YPSSMDWRKK GN---VVSPVKNQG ACGSCWTFST TGALESAVAI AS-GKM-MTLAE QQLVDC-AQN FNNHGCQGGLPSQ
RCB    LPESFDAREQ WSNCPTIAQIRDQG SCGSCWAFGA VEAMSDRICI HTNGRVNVEVSA EDLLTC-CGI QCGDGCNGGYPSG

                 80                        90        100                       110        120
PAP    ALQLVAQY-GI HYRNTYPY---------------EG VQRYCRSREK GPYAAKT-----------------DGV RQVQPYNEGA
LCP1   AIMYVRDNGGV DTESSYPY---------------EA RDNTCRFNSN TIGATCT-----------------GYV GIAQGS-ESA
LCP2   AFDYIKANNGI DTEAAYPY---------------EA RDGSCRFDSN SVAATCS-----------------GHT NIASGS-ETG
LCP3   AFDYIKDNGGI DTESSYPY---------------EA EDRSCRFDAN SIGAICT-----------------GSV EVQHT--EEA
RCL    AFQYIKENGGL DSEESYPY---------------EA KDGSCKYRAE YAVANDT-----------------GFV DIPQQ--EKA
RCH    AFEYILYNKGI MGEDSYPY---------------IG KNGQCKFNPE KAVAFVK-----------------NVV NITLND-EAA
RCB    AWNFWTRKGLV SGGVYNSHIGCLPYTIPPCEHHVNG SRPPCTGEGD TPKCNKMCEAGYSTSYKEDKHYGYTS YSVSDS-EKE

                130        140         150            160                170         180        190
PAP    LLYSIAN-QPV SVVLEAAGKD FQLYRGGIFV GP-C---GNKVDHA VAAVGYG---PN-----Y ILIKNSWGTG WGENGYIRIK
LCP1   LKTATRDIGPI SVAIDASHRS FQSYYTGVYY EPSC--SSSQLDHA VLAVGYG---SEGGQD-F WLVKNSWATS WGESGYIKMA
LCP2   LQQAVRDIGPI SVTIDAAHSS FQFYSSGVYY EPSC--SPSYLDHA VLAVGYG---SEGGQD-F WLVKNSWATS WGDAGYIKMS
LCP3   LQEAVSGVGPI SVAIDASHFS FQFYSSGVYY EQNC--SPTFLDHG VLAVGYG---TESTKD-Y WLVKNSWGSS WGDAGYIKMS
RCL    LMKPVATVGPI SVAMDASHPS LQFYSSGIYY EPNC--SSKDLDHG VLVVGYGYEGTDSNKDKY WLVKNSWGKE WGMDGYIKIA
RCH    MVEAVALYNPV SFAFEVT-ED FMMYKSGVYS SNSCHKTPDKVNHA VLAVGYG---EQNGLL-Y WIVKNSWGSN WGNNGYFLIE
RCB    IMAEIYKNGPV EGAFTVF-SD FLTYKSGVYK HEA---GDVMGGHA IRILGWG---IENGVP-Y WLVANSWNVD WGDNGFFKIL

               200      210
PAP    RGTGNSYGVC GLYTSSFYPV KN
LCP1   RNRNNN---C GIATDACYPT V
LCP2   RNRNNN---C GIATVASYPL V
LCP3   RNRDNN---C GIASEPSYPT V
RCL    KDRNNH---C GLATAASYPI VN
RCH    RGKNM----C GLAACASYPI PQV
RCB    RGENH----C GIESEIVAGI PRTQQYWGRF
```

NUMBERS OF IDENTICAL POSITIONS (MATURE PROTEIN)

	PAP	LCP1	LCP2	LCP3	RCL	RCH	RCB
PAP	212	94	88	83	95	83	63
LCP1		217	166	136	127	91	61
LCP2			216	151	128	95	61
LCP3				215	138	91	55
RCL					218	101	56
RCH						219	62
RCB							253

Fig. 3. Comparison of amino acid sequences of the three lobster cysteine proteinases LCP1, LCP2 and LCP3, with papain (PAP) and rat cathepsins L, H and B (RCL, RCH and RCB, respectively). The numbers refer to positions in the papain sequence. Position 1 refers to the mature protein amino-terminus and the negative numbers identify positions in the propeptide and signal peptide. Sequences are aligned using gaps to achieve maximal position identity. Residues conserved in all sequences are in bold. Numbers of position identities between pairs of mature protein sequences are shown at the end of the aligned sequences.

the digestive enzymes evolved from a cathepsin it is more likely to have been from an L type rather than an H or B.

3.3. *Structural features of the mature enzymes*

The cysteine residues that form disulfide bonds in papain between positions 22–63, 56–95 and 153–200

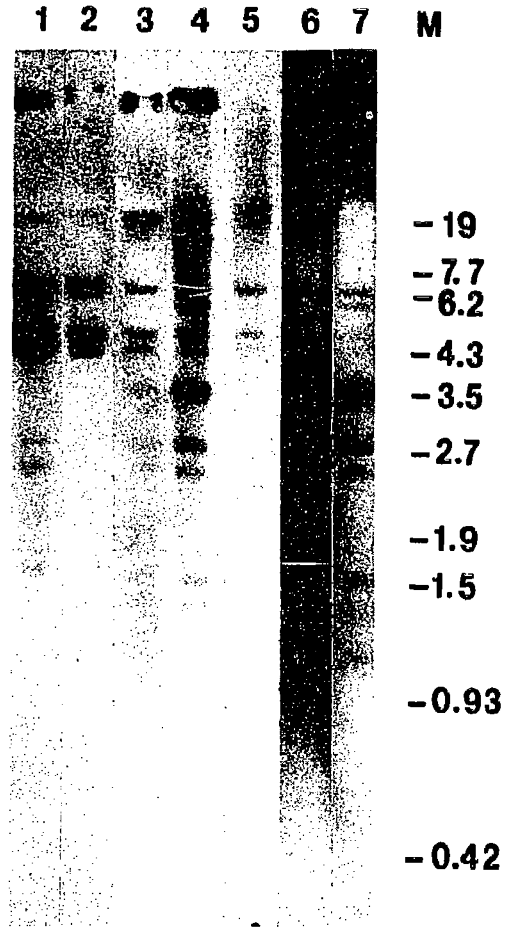

Fig. 4. Southern analysis of lobster cysteine proteinase genes. *Pst*I-digested DNA was hybridized to 5' (lane 1) and 3' (lane 2) coding region probes derived from LCP1 (see Fig. 1), and to the entire 0.9 kb LCP1 cDNA (lane 5), LCP2 (lanes 3 and 6) and LCP3 (lanes 4 and 7) cDNAs. Post-hybridization washes were at 68°C in 4 × SSC, 0.1% SDS (lanes 1–4) or 0.5 × SSC, 0.1% SDS (lanes 5–7). Lane M indicates the size (in kb) of DNA markers.

[13] are conserved in other cysteine proteinases and are also present in the lobster sequences (Fig. 3). LCP1 has two additional cysteine residues at positions 106 and 207. Assuming the tertiary structure of LCP1 is similar to that of papain, it is unlikely that a disulfide bridge could be formed between these additional cysteine residues because they would be too far apart [13]. It is expected, therefore, that the mature enzymic product of LCP1 has three free sulfhydryl groups, including that in the active site, and LCP2 and LCP3 mature enzymes will have two free sulfhydryl groups each.

Earlier experiments on denaturing gels indicated a molecular mass of about 28 kDa for a purified cysteine proteinase preparation from lobster digestive juice [1]. From the data presented here, the calculated molecular masses of LCP1 mature protein is 23 386 Da. If we assume that the LCP2 and LCP3 polypeptides are processed at the same site as LCP1, then the molecular masses of the LCP2 and LCP3 mature enzymes would be 23 093 and 23 255 Da, respectively. Confirmation of these molecular masses for LCP1 and LCP2 has been obtained by ion-spray mass spectrometric analysis of purified enzyme [14]. The mass spectral evidence indicates that the enzymes corresponding to LCP1 and LCP2 are secreted into the digestive fluid and the mature polypeptides undergo no post-translational modifications other than the formation of disulfide bridges.

3.4. *Analysis of cysteine proteinase gene family*

The results of Southern hybridization experiments suggest that an extensive family of genes encode cysteine proteinases in the lobster (Fig. 4). At least seven *Pst*I-generated DNA fragments are detected by the LCP1 5' coding region probe (lane 1) and four of these are also detected by the LCP1 3' coding region probe (lane 2). Three of these common bands hybridize quite strongly, indicating three distinct LCP1-related genes. In contrast, each of the probes prepared from the entire LCP2 and LCP3 cDNAs detects a single strongly hybridizing fragment, probably their respective genes (see the 20 kb band in lanes 3 and 6, and the 3.6 kb band in lanes 4 and 7), as well as other less strongly hybridizing fragments. These results demonstrate at least eight and as many as eleven distinct genomic elements related to the LCP1, LCP2 and LCP3 cDNAs. As noted above, each of the proteins encoded by these cDNAs is related to the L cathepsins of vertebrates. If crustaceans possess cysteine proteinases of the H or B types (and they may, given the presence of cathepsin B-related proteinases in more primitive animals, i.e. the nematode *Haemonchus contortus* [14] and the trematode *Schistosoma mansoni* [15]), then genes that encode these proteins may not have been detected in this analysis and the size of the lobster cysteine proteinase gene family may be even larger than estimated here.

Acknowledgement: This is NRC publication No. 32987.

REFERENCES

[1] Laycock, M.V., Hirama, T., Hasnain, S., Watson, D. and Storer, A.C. (1989) Biochem. J. 263, 439–444.

[2] Portnoy, D.A., Erikson, A.H., Kochan, J., Ravetch, J.V. and Unkeless, J.C. (1986) J. Biol. Chem. 261, 14697–14703.

[3] Blocklehurst, K., Willenbrock, F. and Salih, E. (1987) in: New Comprehensive Biochemistry, vol. 16 (Neuberber, A. and van Deeman, L.L.M. eds.) pp. 39–143, Elsevier, Amsterdam.

[4] Neurath, H. and Walsh, K.A. (1976) Proc. Natl. Acad. Sci. USA 73, 3825–3832.

[5] Nothwehr, S.F. and Gordon, J.I. (1990) Bioessays 12, 479–484.

[6] Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Molecular Cloning, A Laboratory Manual, 2nd edn. Cold Spring Harbour, NY.

[7] Turpen, T.H. and Griffith, O.M. (1986) BioTechniques 4, 11–13.

[8] Huynh, T.V., Young, R.A. and Davis, R.W. (1985) in: DNA Cloning, vol. I (Glober, D.M. ed.) pp. 49–78, IRL Press, Oxford.

[9] DeBorde, D.C., Naeve, C.W., Herlocher, M.L. and Maasab, H.F. (1986) Anal. Biochem. 157, 275–282.

[10] Hamby, R.K., Sims, L., Issel, L. and Zimmer, E. (1988) Plant Mol. Biol. Report 6, 175–192.

[11] Cohen, L.W., Coghlan, V.M. and Dihel, L.C. (1986) Gene 48, 219-227.

[12] Ishidoh, K., Imajoh, S., Emori, Y., Ohno, S., Kawasaki, H., Minami, Y., Kominami, E., Katunuma, N. and Suzuki, K. (1987) FEBS Lett. 226, 33-37.

[13] Kamphuis, I.G., Kalk, K.H., Swarte, M.B.A. and Drenth, J. (1984) J. Mol. Biol. 179, 233-257.

[14] Thibault, P., Pleasance, S., Laycock, M.V., MacKay, R.M. and Boyd, R.K. (1991) J. Mass Spectrom. Ion Processes (in press).

[15] Pratt, D., Cox, G.N., Milhausen, M.J. and Boisvenue, R.J. (1990) Mol. Biochem. Parisitol. 43, 181-192.

[16] Klinkert, M.-Q., Felleisen, R., Link, G., Ruppel, A. and Beck, E. (1989) Mol. Biochem. Parisitol. 33, 113-122.