GENE CHARACTERISATION, ISOFORMS AND RECOMBINANT EXPRESSION *IN VITRO* OF CARCININ, AN ANTIBACTERIAL PROTEIN FROM THE SHORE CRAB, *CARCINUS MAENAS*

Virginia Brockton[1], John A. Hammond[2], Valerie J. Smith[3*]

Comparative Immunology Group,
Gatty Marine Laboratory,
University of St. Andrews,
Scotland, UK.

[1] Present address: Department of Biological Sciences, George Washington University, 2023 G St NW, 340 Lisner Hall, Washington DC 20052, USA

[2] Present address: Department of Structural Biology, Sherman-Fairchild Building, D159, 299 Campus Drive West, Stanford University School of Medicine, Stanford CA 94305, USA

[3*] Address for correspondence: Comparative Immunology Group, Gatty Marine Laboratory, University of St. Andrews, Scotland, UK. Tel.: +44 1334 463474; Fax: +44 1334 463443, vjs1@st-andrews.ac.uk

ABSTRACT

Carcinin is a whey acidic protein (WAP) domain-containing antimicrobial protein, produced by the circulating haemocytes of the shore crab, *Carcinus maenas*. Cloning of its full coding cDNA reveals that it shows some similarity to invertebrate defensins and has a valine-rich signal sequence followed by a defined cleavage site but no obvious acidic anionic 'pro' sequence. The C-terminus exhibits a unique cysteine array that is predicted to form 6 disulphide bonds in the tertiary structure. This 12 cysteine array arrangement is conserved in ESTs from related genera and seems to represent a novel tertiary structure amongst AMPs, unique to the Crustacea. There are at least 5 putative isoforms of the mature protein that may arise through the transcription of one or two alleles of a multi-exon gene. Several of the same transcripts have been found in different animals. These isoforms do not arise as a result of alternate splicing of the exons, but by the transcription of different alleles and /or single point mutation of the transcript at up to four loci in the gene. The most commonly expressed transcript of the protein was recombinantly expressed in bacterial fusion system to a yield of ca 2-3 μg ml^{-1} of culture. *In vitro* expression with or without the leader sequence confirms the bioinformatic prediction that the stability of the mature protein is reduced when the leader sequence is removed. Carcinin is one of very few invertebrate AMPs characterised at the gene, transcript and protein level and to be recombinantly expressed *in vitro* in a bacterial system.

KEYWORDS: Carcinin, crustin, crab, crustacean, *Carcinus maenas*, antibacterial protein, 11.5 kDa protein, WAP domain, recombinant bacterial expression

INTRODUCTION

Low molecular weight, non-specific antibacterial proteins have been known to be part of the innate immune repertoire of both vertebrates and invertebrates since the 1980's. However, such peptides were not isolated from a crustacean until 1995. The first of these peptides to be purified and characterised, was isolated from the circulating haemocytes of the shore crab, *Carcinus maenas* by Schnapp *et al.* (1996). It is a 6.5 kDa proline-rich cationic protein, with homology to bovine bactenecin 7 at the N-terminus (Schnapp *et al.*, 1996). In the same paper, an 11.5 kDa antibacterial protein (GenBank accession no. **AJ237947**) was also observed in the haemocytes of this crab (Schnapp *et al.*, 1996). The protein was subsequently partially purified and characterised by Relf *et al.* (1999) and found to be a cysteine rich, hydrophobic molecule. It is present in the granular haemocytes (Chisholm and Smith, 1992) and is active against Gram positive marine bacteria, including the crustacean pathogen, *Aerococcus viridans* var *homari* (Relf *et al.*, 1999). The partial amino acid sequence information revealed that it has a 4-disulphide core whey acidic protein (WAP) domain. This gives it some similarity to human anti-leukocyte proteinase and indicates that it may have a role as a serine protease (Relf *et al.*, 1999). It was subsequently designated 'carcinin' by Smith and Chisholm (2001).

Since the purification of carcinin, several expressed sequence tag (EST) sequences, with at least one WAP domain and similarity to carcinin, have been published and /or submitted to GenBank from a variety of other crustacean species including the shrimps, *Penaeus monodon* (GenBank accession nos. **BI018072**, **BI018073**, **BI018074** and **CF415873**) (Supungul *et al.*, 2002; Supungul *et al.,* 2004), *Marsupenaeus japonicus* (GenBank accession nos. **AB121740**, **AB121741, AB121742, AB121743** and **AB121744**) (Rattanachai *et al.*, 2004), *Litopenaeus vannamei* and *Litopenaeus setiferus*

(GenBank accession nos. **AF430071**, **AF430072**, **AF430073**, **AF430074**, **AF430074**, **AF430075**, **AF430076**, **AF430077**, **AF430078**, **AF430079**, **AY488492**, **AY488493**, **AY488494**, **AY488495**, **AY488496**, and **AY488497**) (Gross *et al.*, 2001; Bartlett *et al.*, 2002; Vargas-Albores *et al.*, 2004), the lobsters, *Homarus americanus* (GenBank accession no. **CN853187**) (Towle and Smith, unpublished) and *Homarus gammarus* (GenBank accession no. **AJ786653**) (Hauton *et al.*, 2005, in press), as well as the spiny lobster, *Panulirus argus* (GenBank accession no. **AY340636**) (Stoss *et al.*, 2004). Recently, ESTs from the blue crab, *Callinectes sapidus,* the signal crayfish, *Pacifastacus leniusculus,* and the Atlantic sand fiddler crab, *Celuca pugilator,* have also been cloned and sequences with homology to both carcinin (GenBank accession nos. **CV223924**, **CV224031**, **CV224051**, **CV224608**, **CV224766**, **CV225152**, **CV225167**, **CV225185**, **CV303224**, **CV434160**, **CV434250**, **CV463096**, **CV479075**, **CV479982**, **CV527864, CV022159, CV022069, CF542655, CF542655, CF542483, CF542463,** and **CF542304**), and crustins (GenBank accession nos. **CV006490**, **CV022228**, **CV070843**, **CV071080**, **CV071347**, **CV086486**, **CV086696**, **CV086957**, **CV161819**, **CV161936**, **CV223806**, **CV224132**, **CV224245**, **CV303048**, **CV303073**, **CV303122**, **CV462984**, **CV463498**, **CV479051**, **CV479302**, **CV479718**, **CV527758**, **CV527767** and **DW176897**) have been submitted to GenBank (Shafer, *et al*. 2004, unpublished, Söderhäll and Söderhäll, 2003, unpublished and Tang *et al*. 2005, unpublished).

Some non-crustacean species are also known to produce antibacterial proteins containing WAP domains (Hagiwara *et al.*, 2003). Indeed, the WAP domain has now become recognised as central to the function of molecules with serine protease activity and this activity is thought to be a fundamental aspect of the antibacterial properties of these molecules (Ganz, 2003; Chen *et al.*, 2004). Thus, at least for crustaceans, which lack adaptive immune capability based on clonal selection of lymphocyte subsets, and

often show weak phagocytic activity (Smith and Ratcliffe, 1978), WAP-containing antimicrobial proteins may be important defence effectors. They may also be significant in other physiological processes as WAP-containing proteins have been noted in other species in a variety of tissues not usually associated with antibacterial defence (Trexler *et al.*, 2001; Clauss *et al.*, 2002; Stoss *et al.*, 2004).

Given the need for crustaceans, like most invertebrates, to exercise some economy in the repertoire of proteins synthesised because of constraints of the genome size (Klein, 1989), it is likely that WAP-containing antibacterial proteins might participate in several biological activities in these animals. Such multifunctional ability of immune-associated genes is not unknown in crustaceans (Destoumieux-Garzon *et al.*, 2001; Lee *et al.*, 2003; Lee *et al.*, 2004) and is widely observed in other species (Gallo *et al.*, 1994; Bals and Wilson, 2003; Boman, 2003; Farnaud and Evans, 2003; Kamysz *et al.*, 2003; Koczulla *et al.*, 2003). A useful step towards investigating the possible biological roles of this WAP-containing protein in crustaceans, as well as investigating the role of the putative signal sequence, is to express the protein recombinantly. The present work was therefore undertaken to determine the full cDNA coding sequence for carcinin, to investigate the occurrence of isoforms and to establish a protocol for recombinant expression *in vitro.*

METHODS AND MATERIALS

ANIMALS

Specimens of *C. maenas* were collected from St. Andrews Bay, Scotland and acclimatised for 2 weeks in flow-through, seawater aquarium at 15 ºC, ± 2 °C before use. Using only healthy adult intermoult male crabs for experiments, ca 0.3 ml of

haemolymph was withdrawn undiluted from the unsclerotised part of a cheliped and immediately added to 1 ml aliquots of Trizol ® LS (Invitrogen Ltd., Paisley, UK). Total RNA was then isolated separately for each crab following the manufacturer's protocol with the quality and quantity of the RNA assessed spectrophotometrically.

RACE

The 5' and 3' ends of the carcinin mRNA (GenBank accession no. **AJ237947**) were amplified using the SMART™ RACE cDNA amplification kit (BD Biosciences, Oxfordshire, UK), following the manufacturers protocol using the previously isolated total RNA as starting template. Gene specific primers (Carc1F and Carc1R; Figure 1) were designed to the known (GenBank accession no. **AJ237947**) sequence and synthesised by MWG Biotech SA, Ebersberg, Germany. These gene-specific primers were used in conjunction with the RACE supplied primers to amplify the 5' and 3' sequences. PCR products of the expected size were gel purified using the QIAQuick® gel extraction kit (Qiagen Ltd, West Sussex, UK) and then cloned using the TOPO TA (Invitrogen) cloning kit following the manufacturer's protocols. Colonies that produced an amplicon of the expected size after PCR with M13 primers were re-grown in Luria-Bertani broth and the plasmids extracted using the Wizard™ plus SV miniprep DNA purification system (Promega UK, Southampton, UK). The Sequencing Service, University of Dundee, Scotland, performed the sequencing.

RT-PCR

First strand synthesis was conducted on 2 μg of RNA collected from each of 10 crabs using M-MLV reverse transcriptase (Sigma Aldrich, Poole, Dorset, UK), and oligo dT (Promega, UK) as the primer, following the manufacturer's protocol. Gene specific primers (Carc2F and Carc2R; Figure 1) were designed to the full-length sequence

(GenBank accession no. **AJ427538**) obtained using RACE. The cDNA was amplified (1 x 94 °C for 5 min, 30 x [94 °C for 30 s, 57 °C for 30 s and 72 °C for 2 min], plus 1 x 72 °C for 7 min) using proof-reading Advantage® Taq polymerase mix as per the supplier's protocol. The resulting RT-PCR amplicons of the expected size were gel purified, cloned and an isolated plasmid from one colony from each animal was sequenced as previously described for the RACE generated plasmids. The most common full length transcript sequences obtained were subjected to bioinformatic analyses.

GENOMIC DNA

Genomic DNA (gDNA) was isolated from two animals using the DNAeasy kit (Qiagen Ltd., West Sussex, UK) as per the proprietary protocol. The carcinin gDNA was amplified using the Carc2 primers (Figure 1), the amplicons purified, cloned and sequenced as described above for RACE and RT-PCR sequences.

RECOMBINANT EXPRESSION

Two distinct carcinin coding sequences (RecCarc3 and RecCarc4) were expressed as GST (glutathione-S-transferase) fusion proteins using the pGEX-4T-1X bacterial vector (Amersham Pharmacia Biotech Inc., New Jersey, USA). RecCarc3, comprised the full coding region from the first methionine (M_1) codon (Figure 1) to the first stop (TAG) codon. RecCarc4 comprised the sequence from the glycine (G_{22}) codon, immediately following the putative TEA↓GL cleavage site, to the first stop (TAG) codon. These inserts (RecCarc3 and RecCarc4) were modified by PCR at the 5' and 3' ends (using Carc3 and Carc4 primers; Figure 1) to make them compatible with the linearised vector restriction sites BamH1 (G/GATCC) and Xho1 (C/TCGAG) respectively.

One microlitre of cDNA, from an animal identified as having the most commonly occurring transcript sequence, diluted 1:20, was amplified using proof-reading Advantage® Taq polymerase (94 °C for 7 min, then 30 x [94 °C for 30 s, 55 °C for 30 s and 72 °C for 2 min] and 1 x 72 °C for 2 min). Resulting amplicons of the appropriate size were purified using a QIAQuick ® Gel Extraction Kit (Qiagen Ltd.) as per the proprietary protocol.

Restriction digests were prepared for both the vector (using ~1000 ng of pGEX-4T-1) and modified insert sequences (~200 ng of either RecCarc3 or RecCarc4) following the manufacturers protocol. Reactions were cleaned up using a QIAQuick® PCR purification protocol (Qiagen Ltd. Crawlesy, UK) and the linearised vector and insert DNAs were ligated using T4 DNA Ligase (New England Biolabs (UK) Ltd., Hertfordshire, UK), as per the proprietary protocol creating two constructs (pGEX4T-13 and pGEX4T-14).

Competent cells of a protease deficient *E.coli* strain, BL21 (DE3) (Amersham Pharmacia Biosciences, Buckinghamshire, UK), were prepared and transformed with the expression constructs (pGEX-4T-1, pGEX4T-13 and pGEX4T-14) as described in the GST fusion system protocols (Amersham Pharmacia Biotech, 2002).

Two litre cultures of each transformed strain were grown at 20 ºC for ca 19 h (OD_{600} ~0.5-1.0), were induced with 0.1 M IPTG (isopropyl-β-D-thiogalactopyranoside) and incubated at 20 ºC for a further 3.5 h. The induced cultures were centrifuged (7,700 x g, 10 min, 4 ºC) and the resulting cell pellet resuspended (190 ml of 1 x phosphate buffered saline (PBS)) and sonicated (60 % power for 30 s at 5 s pulses) using a MS72 probe (Philip Harris Scientific, Staffordshire, UK). Triton X-100 (20 % in 1 x PBS) was added to a final concentration of 1 % and mixed on ice for 30 min. The sample was then centrifuged (21,000 x g, 10 min, 4 ºC) and the supernatant was filtered through a

0.45 μm nitrocellulose membrane (Whatman, Kent, UK). A 2 ml suspension of 50 % equilibrated (1 x PBS) GS-4B beads (Amersham Biosciences UK Ltd.), was added to the filtrate and mixed at room temperature for 30 min. After centrifugation (500 x g for 5 min, RT), the beads were washed 10 times in ice cold 1 x PBS. The fusion proteins were eluted from the beads using 20 mM glutathione elution buffer in one bed volume of 1x PBS after mixing for 5 min at RT. The centrifugation, washing and elution process was repeated 3 times; each time the supernatant was collected and a further bed volume of elution buffer was added.

The GST tag was cleaved from carcinin using 80 units of thrombin protease (Amersham Biosciences, Buckinghamshire, UK) at RT for 16 h. This digested sample was dialysed against 2000 volumes of cold (4 ºC) stirred binding buffer (50 mM Tris-HCl; 0.5 mM NaCl; pH 8.0) for ~18-24 h at 4 ºC to remove the glutathione. This was followed by thrombin removal by passing the sample through a Benzamidine-Sepharose 6B affinity column (Amersham Biosciences, Uppsala Sweden). After further dialysis against 1x PBS, the sample was applied to a fresh GS-4B column to remove the GST tag from the sample. At each stage of purification sample quality was checked on 12 % discontinuous SDS PAGE gel. Bands at ~11 kDa staining strongly with Coomassie Blue were excised from the gel and analysed by micromass TofSpec-2E MALDI-TOF MS (matrix assisted laser desorption ionization-time of flight mass spectrometer) after tryptic digest (Investigator ProGest Protein Digestion Station) using methods adapted from Shevchenko *et al*. (1996).

RESULTS

RACE

RACE extended the previously submitted sequence (GenBank accession no. **AJ237947**), by 141 bp at the 5' end and the by 237 bp at the 3' end including the poly A^+ tail (Figure 1). Translation of the extended sequence shows the N-terminus to be a valine rich sequence with two methionine residues (Figure 1). The translated C-terminal end was extended by a single tyrosine residue before the TAG stop codon. The resulting full length carcinin transcript sequence (Figure 1) was submitted to GenBank (GenBank accession no. **AJ427538**).

RT-PCR

Four novel full length transcripts (GenBank accession nos. **AJ821886**, **AJ821887**, **AJ821888** and **AJ821889**) were identified from seven crabs. All the isolated transcripts exhibit an identical signal/leader sequence.

Transcript 1 (GenBank accession no **AJ821886**) was found to be the most common and was identified in three crabs. Transcript 2 (GenBank accession no. **AJ821887**) was identified in two crabs, while transcripts 3 and 4 (GenBank accession nos. **AJ821888** and **AJ821889**) were identified once each in two crabs. The transcripts differ from each other at discrete nucleotide loci resulting in residue changes (Figure 1).

The consensus inferred protein sequence comprises a coding region with two ATG translation start sites (ORF Finder at www.ncbi.nlm.nih.gov/gorf/orfig.cgi). The first methionine (M_1) was selected as the most probable start codon as it is located immediately preceding an exon-intron boundary with a purine (A) situated three nucleotides upstream. This pattern of methionine location in a sequence has been used to identify the start methionine codon in other species (Farrell, 1996).

The first 21 residues of the sequence are predicted to be highly hydrophobic and comprise a signal/leader sequence with a clear cleavage site identified between residues 21 and 22 (Figure 1).

The ProtParam tool (http://www.expasy.org/tools/protparam.html) predicts that the inferred putative full sequence of carcinin has a molecular mass of ~12.260 kDa, is cationic, valine rich (11.8 %) and has an instability index of 35.02. Based on consensus identification of a 4-disulphide core by PROSITE (E value = $8E^{-5}$), PRINTS and SMART databases, all sequences obtained in this study were all confirmed to have a WAP domain as described by Relf *et al.* (1999) for the native protein.

No N-glycosylation (N-β-GlcNAc) sites were predicted in the sequences but several sites are predicted as possible O-glycosylation (O-β-GlcNAc) sites on three threonines in the mature sequence at T_{49}, T_{54} and T_{69} (Figure 1). Andreu and Rivas (1998) have suggested that O-glycosylation is essential to the activity of some proteins by influencing their mode of action. In addition, three possible phosphorylation sites are predicted at positions 31 for tyrosine (Y_{31}), and 54 (T_{54}) and 67 (T_{67}) for threonine, but no serine modifications are identifiable. T_{54} is predicted to be both phosphorylated and glycosylated but only one of these modifications is possible at any one time and phosphorylation is less common than glycosylation (Andreu and Rivas, 1998).

PSI-BLAST and BLASTP analyses did not return any significant alignments (>40 % sequence identity) with isolated proteins from the databases searched (CDS translated, PDB, SWISSPROT, PIR & PRF released as of November 2005). The closest BLASTP matches (E < 0.003, ~33 % identity) were based on domain similarity with vertebrate whey acidic proteins (GenBank accession no. **AJ0053561.1**) and anti-leukoproteinase proteins (GenBank accession no. **X04470**) as previously reported (Relf *et al.*, 1999). This domain was based on the identification of the 4-disulphide core motif

Nucleotide sequence alignment using TBLASTN identified TrEMBL entries for ESTs from *Marsupenaeus japonicus, Litopenaeus vannamei*, *Litopenaeus setiferus*, *Homarus gammarus* and *Panulirus argus*. Sequence identity was only observed for these sequences from residue 25 onwards (P_{25}) of the carcinin sequence.

The sequence identities were between 32 % and 46 %, depending on the length sequence, with the highest identity share with the *Marsupenaeus japonicus* and *Pacifastacus leniusculus* ESTs. Now that the full N-terminal sequence has been deduced for carcinin, its overall identity to *Litopenaeus vannamei* ESTs drops to 23-26%. This indicates that sequences from different species may have high sequence identity in the C-terminal fragment, which confers activity, but not in the signal sequence portion as described for other antimicrobial proteins (Zanetti *et al.*, 1995). Similarly, identity of carcinin and a WAP-containing putative protein in *Litopenaeus setiferus,* which was initially thought to be 40-43% based on the C-terminal sequence, drops to 20-30% when the full coding sequence of carcinin is considered. As the Lv2 and Lv3 sequences do not exhibit a glycine repeat region, identified in the other published 'crustin' sequences, the identity between sequences of carcinin and *L. setiferus* may imply a closer evolutionary link than between *C. maenas* and *L. vannamei*.

The predicted secondary structure of carcinin indicates a random coiled structure with two possible β-sheets but no helices. This would support the classification of carcinin as a β-sheet or possibly a loop protein according to recent classification trends (Powers and Hancock, 2003). The absence of a helix would further distinguish this type of crustacean antibacterial protein from the insect defensins described by Ganz and Lehrer, (1994), and points to carcinin having greater structural similarity to the horseshoe crab defensins.

No suitable templates were identified for 3D modelling of the carcinin sequence, as the 25 % identity threshold was not achieved. Pairwise alignment of carcinin with known structures demonstrate that although cysteine conservation between some sequences is conserved, there is insufficient conservation in the remainder of the residues to find a match.

GENOMIC SEQUENCES

Two genomic sequences were obtained from two animals. The gDNA sequences are identical in overall structure and both have a conserved signal sequence. The sequences differ from each other by two non-synonymous nucleotide substitutions at positions 565 and 725. Alignment of the genomic and transcript sequences reveals that carcinin transcript comprises 4 exons and 3 introns.

RECOMBINANT EXPRESSION

Recombinant expression of carcinin using an *E.coli* system was successful for both recombinant sequences (GST-RecCar3 and GST-RecCar4) (Figure 2) with a yield of 2-3 $\mu g\ ml^{-1}$ of culture. A ~10 kDa protein band was confirmed as carcinin after the excised band (Figure 3) was subjected to peptide fingerprinting by mass spectrometry. Mascot analysis indicates that the highest Mowse scores are 114/140 and 138/140 with 8 and 9 peptides matched, respectively, and with 86 % sequence coverage.

After cleavage of the GST tag from the CarcRec4 fusion protein there was a significantly lower yield of the carcinin protein (~10 kDa) compared to the GST (~26 kDa) (Figure 3). This difference in yield of product was not observed for cleavage of the GST-RecCar4 fusion protein where the yield appeared equal for the two proteins when visualised by SDS PAGE gel. Unfortunately, neither affinity chromatography nor ion

exchange chromatography were successful in completely removing the cleaved GST tag from the recombinant product.

DISCUSSION

The present study has extended the known sequence for carcinin at both the transcript and genomic level and, in doing so, has described both the gene structure and five new putative isoforms of the carcinin protein. In addition, the protein has been recombinantly expressed *in vitro* with and without its leader sequence. From these expressed proteins, it appears that the leader sequence confers stability on the mature protein.

In silico analysis of the full length sequence shows that although cysteine residues are highly represented (10.9 %), the full sequence is actually dominated by valine residues (~11.8 %), particularly at the N-terminus. The leader sequence exhibits no similarity to leader sequences found in other crustacean ESTs which have some identity to carcinin in the mature sequence portion (Bartlett *et al.*, 2002; Supungul *et al.*, 2002).

Although the full length carcinin sequence follows the overall α-defensin tripartite peptide pattern, as described by Ganz (2003), the sequence shows greater similarity to the β- or insect defensin patterns and the emerging crustacean defensin cysteine patterns (Kawabata *et al.*, 1996; Bartlett *et al.*, 2002; Supungul *et al.*, 2002). In addition, unlike the defensins described by Ganz and Lehrer (1994), the prepropiece and the mature sections of carcinin do not seem to be encoded by separate exons.

Using *in silico* prediction algorithms which describe protein stability (Guruprasad *et al.*, 1990) and half life (Baschmair *et al.*, 1986), carcinin is likely to be stable (instability index of 35.02), although it may only have a relatively short half-life (~10-30 h).

Cleavage of the leader sequence, as commonly observed for other AMPS (Ganz and Lehrer, 1994; Ganz, 2003), may result in a 'mature' sequence (89 residues), with an instability index ~ 40.03. Therefore we hypothesize that after cleavage, the mature carcinin protein is rendered unstable and quickly degrades. Support for this hypothesis is shown by the visible reduction in yield of the cleaved recombinant carcinin, compared to the GST tag, when analysed by SDS PAGE. The accelerated degradation of active carcinin may safeguard against possible deleterious effects of the rapid release (by granular cell degranulation and exocytosis) of potent AMPs into the haemolymph ensuring such effects to be short lived and cause minimum damage to the animal.

A WAP domain, also observed in the isolated carcinin protein (Relf *et al.* 1999), was found in each of the new full length carcinin sequences obtained in the present study. No additional functional domains or motifs were identified in the full length sequences. However, the single WAP domain appears to form part of a larger pattern of 12 cysteine residues within the carcinin sequence which is conserved across ESTs from several crustacean species. Based mainly on the conservation of these residues, ProDom clustered carcinin with ESTs from *Litopenaeus vannamei* (GenBank accession nos. **AF430071**, **AF430072, AF430073, AF430074, AF430075** and **AF430076**), *Litopenaeus setiferus* (GenBank accession nos. **AF430077**, **AF430078** and **AF430079**), *Panulirus argus* (GenBank accession no**. AY340636**), *Marsupenaeus japonicus* (GenBank accession no**. AB121741**) and *Homarus gammarus* (GenBank accession no**. AJ786653**) into a family of related proteins (ProDom accession no. **PD523494**). Further analysis showed that this pattern of 12 cysteine residues is also found in additional carcinin like transcripts from *Marsupenaeus japonicus* (GenBank accession no**. AB121740**, **AB121742**, **AB121743** and **AB121744**), as well as *Panulirus leniusculus* (GenBank accession no**. AF522504**), *Homarus americanus* (GenBank

accession no**. CN853187**), *Callinectes sapidus* (GenBank accession no**. CV022228** and **CV006490**) and *P. monodon* (GenBank accession no**. BI018072**, **BI018073**, **BI018074**), among others. We therefore propose that the clustering of these sequences, based on the 12 cysteines, represents a new structural family among defensin-like proteins that so far appear to be unique to the crustaceans. Should all 12 cysteines remain conserved in these sequences and the overall sequence identity remain above 25%, then these putative proteins should have similar 3D structures (Westhead *et al.*, 2002). If all six disulphide bonds are formed in the tertiary structure, as predicted by CYSPRED for carcinin in *C. maenas,* then carcinin and its orthologues would be the first family of 6 disulphide bonded AMPs proteins.

Five new distinct putative carcinin isoform transcipts (GenBank accession nos. **AJ427538**, **AJ821886**, **AJ821887**, **AJ821888** and **AJ821889**) were isolated from seven animals with >95 % identity between the sequences. Although transcribed from several exons, the diversity between the transcript sequences may arise from changes at discrete nucleotide loci both at the genomic and/or at the transcript level leading to non synonomous changes at discrete positions in the amino acid sequence. Two of the transcripts isolated were found in more than one animal. Alignment of these carcinin sequences reveals 15 nucleotide positions where the nucleotides are variable. Only 10 of these are non-synonymous, leading to changes in 6 residues (at residue positions 17, 22, 26, 39, 56 and 63; Figure 4). Of these 6 residue changes, only 5 lead to putative isoforms of carcinin, as the sixth may be a residue assignment error. Furthermore, *in silico* transcription and translation of the two genomic sequences shows an additional non-synonymous change at residue position 87, which could result in either a leucine (L) or a phenylalanine (F) amino acid at this position (Figure 4).

The putative isoform sequences all have a highly conserved leader sequence (residues 1-21) apart from one residue (residue 17; Figure 1) where a change in nucleotide 49 results in either an alanine or a threonine residue at this position (Figure 4). At position 22, however, the isoforms vary with arginine or glutamate substituting for glycine. This variation does not seem to affect the cleavage site at position 21. Glycine is the most common residue identified at this position in carcinin, and is often the first residue in the mature fragment of several AMPs (Tossi *et al.*, 2000), particularly in arthropods (Dimarcq *et al.*, 1998). The carcinin protein fragment sequenced by Relf *et al.* (1999) identified an asparagine (N) residue at position 26. However, all transcript and gDNA sequences in the present study identify the residue at position 26 as a proline (P) (Figure 4). As asparagine and proline rarely substitute for each other (PAM250 = -1), the C-terminal residues of carcinin obtained by the original N-terminus sequencing (Relf *et al*., 1999) may have been incorrectly assigned. Therefore, we now believe that the residue at position 26 should be a proline (Figures 1 and 4). The residue at position 39 (Figure 1) can be either a leucine (L), an isoleucine (I) or a valine (V) (Figure 4). These three residues do commonly substitute for each other (Westhead *et al.*, 2002), (PAM250 >2) without necessarily affecting on tertiary conformation as they have similar physicochemical properties and sizes. Accordingly, the different isoforms of carcinin may have arisen by random substitutions at this position. A change at nucleotide 244 would lead to a change in the residue at this position resulting in a serine (S), a lysine (K) or an asparagine (N) (Figures 1) at amino acid position 56 (Figure 4). The proline at position 63 (P_{63}) (Figure 1) appears to be substituted with equal frequency by alanine (A) in translated cDNA and gDNA sequences. Both residues are hydrophobic and are often substituted for each other (PAM250 = 1). Changes at these residues changes might thus contribute to the diversity of isoforms of carcinin (Figure 4). Possibly, further

substitutions or combinations of substitutions could give rise to many more isoforms of carcinin than those observed. However, a total of six isoforms (GenBank accession nos. **AJ237947**, **AJ427538**, **AJ821886**, **AJ821887, AJ821888** and **AJ821889**) have been identified from protein (Relf *et al.*, 1999) and cDNA in the present study as well as two further putative sequences from gDNA sequencing (Figure 4).

Carcinin is transcribed from a multi-exon encoding gene that comprises of at least 4 exons and 3 introns. The full 3' UTR of the transcript has been identified on the gDNA sequence but not the 5' UTR. This may indicate that there is an extra exon further upstream encoding the 5'UTR that has yet to be identified. In addition, multiple alleles may also be present the same animal. The diversity of cDNA and gDNA sequences identified in the present study indicates that the isoforms of carcinin may arise either through transcription of different alleles and/or that individual transcripts undergo post transcriptional modification leading to both synonymous and non-synonymous changes in nucleotides. Although each animal may express a complete repertoire of transcripts from more than one gene, due to the overall conservation of sequence, it is more likely that the diversity occurs at the population level.

The experimentally obtained mass of 11.534 kDa reported for carcinin by Relf, *et al.* (1999) is markedly smaller than the mass predicted for the translated full transcript (12.226 kDa) and larger than the predicted mass of the "mature" protein (10.162 kDa). Predicted PTMs could not account for these differences. Thus we suggest that, based on the evidence presented in the current study, this discrepancy is probably due to rapid autoproteolytic degradation of the full length sequence and short half life of the purified protein between mass spectroscopy and Edman N-terminal analyses performed by Relf *et al*. (1999).

ACKNOWLEDGEMENTS

This work was funded by a Natural Environment Research Council studentship (NER/S/A/2000/03633) to VB.

REFERENCES

Bals, R. and Wilson, J.M. 2003. Cathelicidins--a family of multifunctional antimicrobial peptides. Cellular and Molecular Life Sciences 60, 711-720.

Bartlett T. C., Cuthbertson B. J., Shepard E. F., Chapman R. W., Gross P. S. and Warr G. W. (2002) Crustins, homologues of an 11.5-kDa antibacterial peptide, from two species of penaeid shrimp, *Litopenaeus vannamei* and *Litopenaeus setiferus*. Marine Biotechnology 4, 278-293.

Baschmair A., Finley D. and Varshavsky A. (1986) *In vivo* half-life of a protein is a function of its amino-terminal residue. Science 234, 179-186.

Boman H. G. (2003) Antibacterial peptides: basic facts and emerging concepts. Journal of Internal Medicine 254, 197-215.

Chen J. Y., Pan C. Y. and Kuo C. M. (2004) cDNA sequence encoding an 11.5-kDa antibacterial peptide of the shrimp *Penaeus monodon*. Fish and Shellfish Immunology 16, 659-664.

Chisholm J. R. S. and Smith V. J. (1992) Antibacterial activity in the hemocytes of the shore crab, *Carcinus maenas*. Journal of the Marine Biological Association of the United Kingdom 72, 529-542.

Clauss A., Lilja H. and Lundwall A. (2002) A locus on human chromosome 20 contains several genes expressing protease inhibitor domains with homology to whey acidic protein. Biochemical Journal 368, 233-242.

Destoumieux-Garzon D., Saulnier D., Garnier J., Jouffrey C., Bulet P. and Bachère E. (2001) Crustacean immunity. Antifungal peptides are generated from the C terminus of shrimp hemocyanin in response to microbial challenge. Journal of Biological Chemistry 276, 47070-47077.

Dimarcq J. L., Bulet P., Hetru C. and Hoffmann J. (1998) Cysteine-rich antimicrobial peptides in invertebrates. Biopolymers 47, 465-477.

Farrell, R.E., Jr (1996) RNA Methodologies: A Laboratory Guide for Isolation and Characterization, Second Edition, Farrell, R.E., Jr., ed., Academic Press, London.

Farnaud S. and Evans R. W. (2003) Lactoferrin-a multifunctional protein with antimicrobial properties. Molecular Immunology 40, 395-405.

Gallo R. L., Ono M., Povsic T., Page C., Eriksson E., Klagsbrun M. and Bernfield M. (1994) Syndecans, cell surface heparan sulfate proteoglycans, are induced by a proline-rich antimicrobial peptide from wounds. Proceedings of the National Academy of Sciences of the USA 91, 11035-11039.

Ganz T. (2003) Defensins: Antimicrobial peptides of innate immunity. Nature Reviews Immunology 3, 710-720.

Ganz T. and Lehrer R. I. (1994) Defensins. Current Opinion in Immunology 6, 584-589.

Gross P. S., Bartlett T. C., Browdy C. L., Chapman R. W. and Warr G. W. (2001) Immune gene discovery by expressed sequence tag analysis of hemocytes and hepatopancreas in the Pacific White Shrimp, *Litopenaeus vannamei*, and the Atlantic White Shrimp, *L. setiferus*. Developmental and Comparative Immunology 25, 565-577.

Guruprasad K., Reddy B. V. and Pandit M. W. (1990) Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Engineering 4, 15-161.

Hagiwara K., Kikuchi T., Endo Y., Huqun, Usui K., Takahashi M., Shibata N., Kusakabe T., Xin H., Hoshi S., Miki M., Inooka N., Tokue Y. and Nukiwa T. (2003) Mouse SWAM1 and SWAM2 are antibacterial proteins composed of a single whey acidic protein motif. Journal of Immunology 170, 1973-1979.

Hauton C., Brockton V. and Smith, V.J. (2005) Cloning of a crustin-like, single-whey-acidic-domain, antibacterial peptide from the haemocytes of the European Lobster, *Homarus gammarus*, and its response to infection with bacteria. Molecular Immunology. In press.

Kamysz W., Okroj M. and Lukasiak J. (2003) Novel properties of antimicrobial peptides. Acta Biochimica Polonia 50, 461-469.

Klein J. (1989) Are invertebrates capable of anticipatory immune responses? Scandinavian Journal of Immunology 29, 499-505.

Koczulla R., von Degenfeld G., Kupatt C., Krotz F., Zahler S., Gloe T., Issbrucker K., Unterberger P., Zaiou M., Lebherz C., Karl A., Raake P., Pfosser A., Boekstegers P., Welsch U., Hiemstra P. S., Vogelmeier C., Gallo R. L., Clauss M. and Bals R. (2003) An angiogenic role for the human peptide antibiotic LL-37/hCAP-18. Journal of Clinical Investigation 111, 1665-1672.

Lee S. Y., Lee B. L. and Söderhäll K. (2003) Processing of an antibacterial peptide from hemocyanin of the freshwater crayfish *Pacifastacus leniusculus*. Journal of Biological Chemistry 278, 7927-7933.

Lee S. Y., Lee B. L. and Söderhäll K. (2004) Processing of crayfish hemocyanin subunits into phenoloxidase. Biochemical and Biophysical Research Communications 322, 490-496.

Rattanachai A., Hirono I., Ohira T., Takahashi Y. and Aoki T. (2004) Cloning of kuruma prawn *Marsupenaeus japonicus* crustin-like peptide cDNA and analysis of its expression. Fisheries Science 70, 765-771.

Relf J. M., Chisholm J. R. S., Kemp G. D. and Smith V. J. (1999) Purification and characterisation of a cysteine-rich 11.5kDa antibacterial protein from the granular haemocytes of the shore crab, *Carcinus maenas*. European Journal of Biochemistry 264, 350-357.

Schnapp D., Kemp G. D. and Smith V. J. (1996) Purification and characterisation of a proline-rich antibacterial peptide, with sequence similarity to bactenecin-7, from the haemocytes of the shore crab, *Carcinus maenas*. European Journal of Biochemistry 240, 532-539.

Shevchenko A., Wilm M., Vorm O. and Mann M. (1996) Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. Analytical Chemistry 68, 850-858.

Smith V. J. and Chisholm J. R. S. (2001) Antimicrobial proteins in crustaceans. In:. Beck G., Sugurmaran M. and Cooper E. L (Eds.), Advances in Experimental Medicine and Biology, Kluwer Academic/Plenum Publications, 484, pp. 95-112.

Smith V. J. and Ratcliffe N. A. (1978) Host defence reactions of the shore crab, *Carcinus maenas* (L.) *in vitro*. Journal of the Marine Biological Association of the United Kingdom 58, 367-379.

Stoss T. D., Nickell M. D., Hardin D., Derby C. D. and McClintock T. S. (2004) Inducible transcript expressed by reactive epithelial cells at sites of olfactory sensory neuron proliferation. Journal of Neurobiology 58, 355-368.

Supungul P., Klinbunga S., Pichyangkura R., Hirono I., Aoki T. and Tassanakajon A. (2004) Antimicrobial peptides discovered in the black tiger shrimp *Penaeus monodon* using the EST approach. Diseases of Aquatic Organisms 61, 123-135.

Supungul P., Klinbunga S., Pichyangkura R., Jitrapakdee S., Hirono I., Aoki T. and Tassanakajon A. (2002) Identification of immune-related genes in hemocytes of black tiger shrimp (*Penaeus monodon*). Marine Biotechnology 4, 487-494.

Tossi A., Sandri L. and Giangaspero A. (2000) Amphipathic, A-helical antimicrobial peptides. Biopolymers 55, 4-30.

Trexler M., Banyai L. and Patthy L. (2001) A human protein containing multiple types of protease-inhibitory modules. Proceedings National Academy of Sciences 98, 3705-3709.

Vargas-Albores F., Yepiz-Plascencia G., Jimenez-Vega F. and Avila-Villa A. (2004) Structural and functional differences of *Litopenaeus vannamei* crustins. Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology 138, 415-422.

Westhead D. R., Parish J. H. and Twyman R. M. (2002) Bioinformatics. BIOS Scientific Publishing Ltd., Leeds.

GLOSSARY

AMP	antimicrobial proteins
cDNA	complementary DNA
CDS	chemical data service
EDTA	ethylenediaminetetraacetic acid
EST	expressed sequencing tags
gDNA	genomic DNA
GST	glutathione S transferase

IPTG	isopropyl-beta-D-thiogalactopyranoside
MALDI-TOF MS	matrix assisted laser desorption ionization-time of flight mass spectrometer
M-MLV	moloney murine leukemia virus
ORF	open reading frame
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PDB	protein database
PIR	protein information resource
PRF	protein research foundation
RACE	rapid amplification of cDNA ends
RNA	ribonucleic acid
RT-PCR	reverse transcriptase polymerase chain reaction
SDS PAGE	sodium dodecyl (lauryl) sulfate-polyacrylamide gel
TBE	tris borate ethylenediaminetetraacetic acid buffer
UTR	untranslated region
WAP	whey acidic protein

FIGURE LEGENDS

Figure 1: Full first length carcinin cDNA sequence (GenBank accession no. **AJ427538)** obtained using RACE. Indicated are the recombinantly expressed protein sequences RecCarc3 and RecCarc4 (upper case text), the 5' and 3' UTRs (lower case text), the non-synonymous nucleotide changes leading to isoforms identified in the current study (lower case bold text) or synonymous (lower case italic text), the stop codon (*), the signal/leader sequence (underlined) and the putative cleavage site of the leader sequence (black bold text). The gene specific primers (Carc1, Carc2; F and R) used to amplify carcinin cDNA (GenBank accession nos. **AJ237947** and **AJ427538** respectively) are shown (black boxes). Carc3 and Carc4 primer pairs, used to modify the full length carcinin cDNA sequence for ligation into the recombinant expression vector pGEX-4T-1X are also shown in black boxes.

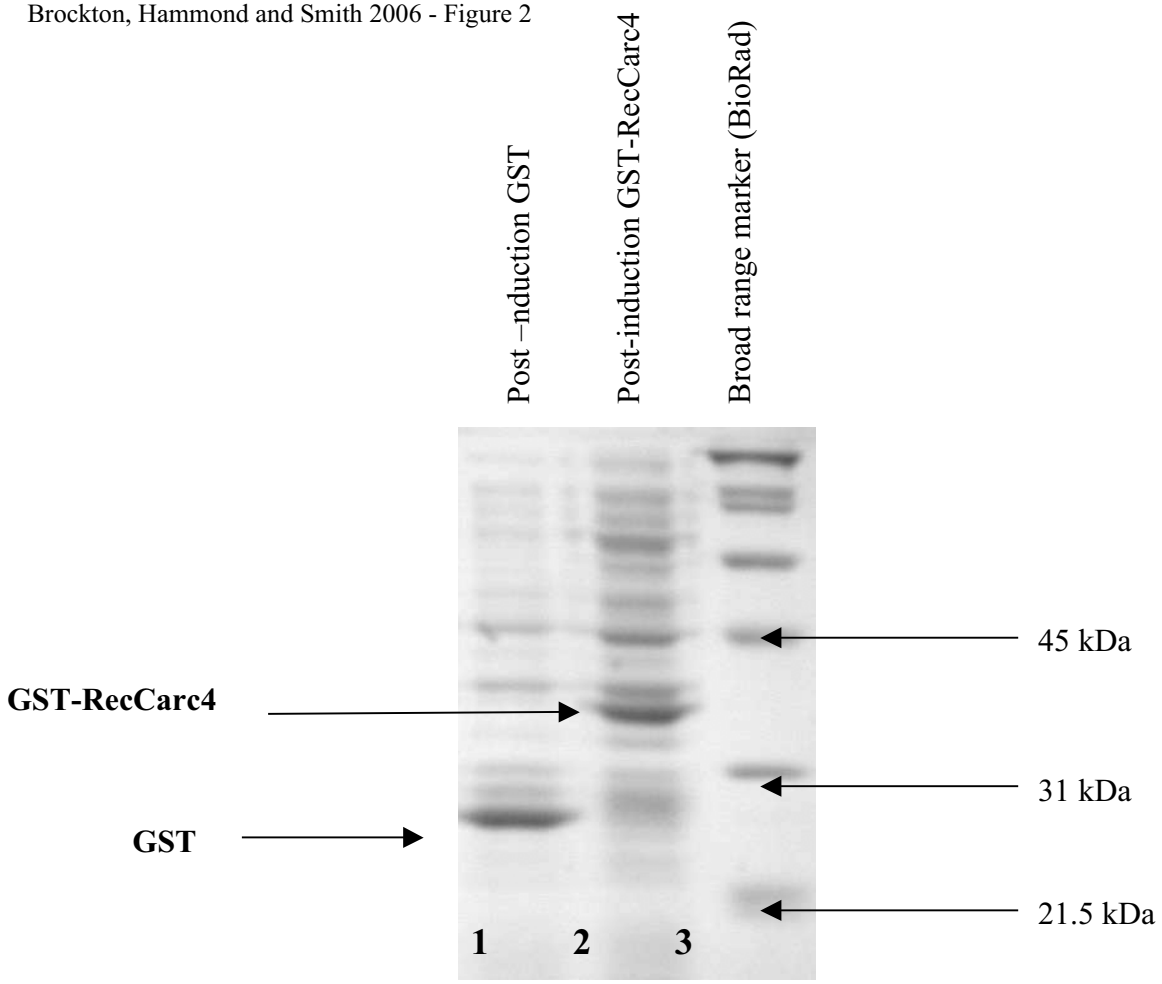
Figure 2: Image of SDS PAGE of post induction expression of the positive control 26 kDa GST protein (Lane 1) and the expression of the GST tagged recombinant carcinin fusion protein (RecCarc4) in Lane 2 (~33 kDa).

Figure 3: Image of the SDS PAGE gel of cleaved GST-RecCarc4 sample eluate. The lower band (~12 kDa) was confirmed as carcinin by protein mass fingerprinting. The lysate pellet sample clearly shows a substantial insoluble or bound carcinin portion. The digest times of the eluates were varied, 6 h (Lane 2), 10 h (Lane 3) and 20 h (Lane 4) using 80 units thrombin in each.

Figure 4: Alignment of all the known carcinin isoform sequences. The 2 predicted gDNA sequences, the original isolated carcinin protein fragment (GenBank accession no. **AJ237947**) (Relf *et al*. 1999), the first full length transcript sequence (GenBank accession no. **AJ427538**) as well the four additional carcinin sequences (GenBank accession nos. **AJ821886**, **AJ821887**, **AJ821888** and **AJ821889)** have been aligned to highlight locations of residue changes leading to isofom diversity of carcinin.

Brockton, Hammond and Smith 2006 - Figure 1

Brockton, Hammond and Smith 2006 - Figure 2



Brockton, Hammond and Smith 2006 - Figure 3