Characterization of a cDNA encoding RP43, a CUB-domain-containing protein from the tube of *Riftia pachyptila* (Vestimentifera), and distribution of its transcript

Luc CHAMOY^{*1}, Maryse NICOLAÏ^{*1}, Brigitte QUENNEDEY^{*}, Françoise GAILL[†] and Jean DELACHAMBRE^{*2}

^{*}UMR CNRS 5548, Développement-Communication Chimique, Université de Bourgogne, 6 boulevard Gabriel, 21000 Dijon, France, and [†]UMR CNRS 7622 Biologie Moléculaire et Cellulaire du Développement, UPMC, 9 Quai Saint Bernard, 75005 Paris, France

A major 43 kDa protein from the protective tube of *Riftia pachyptila* (Vestimentifera), named RP43, was partly micro-sequenced after isolation by SDS/PAGE from the protein fraction of tubes collected around the hydrothermal vents at the East Pacific Rise. On the basis of the partial peptide sequences obtained, experiments using reverse-transcriptase-mediated PCR and rapid amplification of cDNA ends led to the complete cDNA sequence. Analysis of deduced amino acid sequence of RP43 showed the presence of CUB domains (100–110-residue-spanning domains first reported in the complement sub-components C1r/C1s, epidermal-growth-factor-related sea urchin protein and bone morphogenetic protein 1) that seem to be involved in protein–protein and glycosaminoglycan–protein inter-

actions. This peculiarity strongly suggests that RP43 might have a crucial role in tightening the different elements of the worm tube. However, the absence of chitin-binding motifs inclines us to favour a role in protein–protein interactions during assembly of the tube components. The *RP43* mRNA was found to be present in specific epidermal cells from the worm body wall but never in the chitin-synthesizing gland cells. This unexpected result clearly indicates that the major tube protein is synthesized in specialized areas of the outer epithelium and that at least two different tissues are involved in the synthesis of the exoskeleton.

Key words: binding protein, β -chitin, deep sea, hydrothermal vents, marine biopolymers.

INTRODUCTION

Most invertebrates possess a non-collagenous extracellular matrix (ECM) surrounding their epidermis. In this ECM, chitin microfibrils are embedded in a proteinaceous matrix. Several studies have shown that the tube worm *Riftia pachyptila* was an accurate model to determine the sequence steps of such ECM morphogenesis [1,2]. This giant organism (up to 2 m in length) is found in an extreme environment, around deep-sea hydrothermal vents [3]. This gutless worm, housing endosymbiotic bacteria, inhabits a tube made of chitin associated with proteins [4].

Previous studies have shown that chitin-secreting glands (also called pyriform glands) are involved in the production of the tube wall [5,6]. The efficiency of such a huge cellular system explained the tremendously high growth rate that these exoskeletons can reach *in situ* [7] and *in vivo* [8,9]. These results were obtained with chitin as a molecular marker for tube wall formation but very little is known today about its ECM proteins, although it is generally admitted that chitin-secreting glands are the site of tube components secretion [3,5,10].

Results obtained on the composition of the *Riftia pachyptila* tube wall have shown that the protein fraction of this exoskeleton contains several peculiar proteins such as RP43, a 43 kDa polypeptide [11], which was detected through the whole length and thickness of the tube wall. Such a distribution suggests that RP43 might be a protein involved in binding chitin or in stabilization of the tube structure.

The present paper reports the molecular cloning of *RP43* cDNA and its deduced amino acid sequence. To specify the

origin of this tube component, Northern blot and hybridization analyses *in situ* were performed on tissues from different regions of this deep-sea worm.

MATERIALS AND METHODS

Materials

Riftia pachyptila samples were collected on the East Pacific Rise depth by the *Nautille* and *Alvin* submersibles (HERO 94, LAR-VAE 95, HOT 96 and HOPE 99 cruises, 2600 m depth). The tube worm is made of four successive regions; from the upper to the most basal part these are the plume (also called the obturaculum), the vestimentum, the trunk and the opisthosome (reviewed in [3]). After dissection on board ship, the tissues were frozen in liquid nitrogen and stored at -80°C or fixed in Bouin solution for histological study [12].

Protein extraction

Riftia pachyptila tubes were homogenized in liquid nitrogen, then extracted for 12 h at 4°C in one of three different extraction solutions containing a protease inhibitor mixture (10 mM EDTA/10 $\mu\text{g}/\text{ml}$ leupeptin/10 $\mu\text{g}/\text{ml}$ pepstatin/0.02 % PMSF): (1) 10 mM Tris/HCl, pH 7.5, containing 0.5 % SDS, (2) 10 mM Tris/HCl, pH 7.5, containing 2 % (w/v) SDS or (3) 10 mM Tris/HCl, pH 7.5, containing 8 M urea. Homogenates were centrifuged at 10000 *g* for 10 min at 4°C ; the supernatants were then extracted with 3 vol. of acetone and the samples were stored at -20°C overnight. The proteins were harvested by centrifugation

Abbreviations used: C1r/C1s, subcomponents of C1, the first component of complement; CUB, 100–110-residue-spanning domain first reported in C1r/C1s, epidermal-growth-factor-related sea urchin protein and bone morphogenetic protein 1; ECM, extracellular matrix; RACE, rapid amplification of cDNA ends; RP43, 43 kDa *Riftia pachyptila* polypeptide; RT-PCR, reverse-transcriptase-mediated PCR; 1 $\times$ NaCl/sodium citrate ('1 $\times$ SSC') is 0.15 M NaCl/0.015 M sodium citrate.

¹ These authors contributed equally to this paper.

² To whom correspondence should be addressed (e-mail jdelach@u-bourgogne.fr).

The nucleotide sequence data reported will appear in DDBJ, EMBL and GenBank® Nucleotide Sequence Databases under the accession number AF233595.

at 10000 *g* for 20 min at 4 °C, washed twice with acetone and centrifuged for 10 min at 10000 *g*. The pellets were dried and used for further analysis. Protein concentrations were determined by the method of Bradford [13] with BSA as a standard.

One-dimensional and two-dimensional PAGE

The protein pellets were dissolved in sample buffer [125 mM Tris/HCl (pH 6.8)/4% (w/v) SDS/0.02% Bromophenol Blue/10% (v/v) glycerol/4% (v/v) 2-mercaptoethanol], then heated for 3 min at 90 °C and separated on SDS-slab gels as described by Laemmli [14], with a 12.5% (w/v) polyacrylamide running gel. Polypeptides were either stained with Coomassie Brilliant Blue R-250 or blotted to PVDF sheets [15]. Protein molecular mass standards from 14 to 94 kDa (Bio-Rad) were used to estimate the apparent molecular masses of the separated polypeptides.

Two-dimensional electrophoresis was performed by the method of O'Farrell et al. [16]. Dried pellets of sample proteins were solubilized in sample buffer [9.5 M urea/4% ampholites (pH 3.5–10; Pharmacia LKB)/2% (w/v) CHAPS/5% (v/v) 2-mercaptoethanol]. The gel for the first dimension consisted of 9 M urea, 4% ampholites, 2% (w/v) CHAPS and a 13% (w/v) solution of polyacrylamide [30% (w/v) acrylamide/1.6% (w/v) bisacrylamide]. After migration (16 V/h per cm²), gels in the first dimension were placed for 15 min in equilibration buffer [62.5 mM Tris/HCl (pH 6.8)/2.3% (w/v) SDS/10% (v/v) glycerol/12.5% (v/v) 2-mercaptoethanol]. For electrophoresis in the second dimension, 10% (w/v) polyacrylamide gels were used [14]. Gels were then stained with silver [17].

Protein purification, enzymic cleavage *in situ* and protein microsequencing

Samples were separated on SDS-slab gels, blotted to PVDF sheets with a semi-dry electroblotter (Bio-Rad) and stained in a solution of 0.1% Ponceau S dye.

The 43 kDa polypeptide bands were pooled and digested *in situ* with bovine trypsin (0.28 µg/µg of protein; Boehringer) for 24 h at 37 °C in 100 mM Tris/HCl (pH 8)/1% (v/v) Triton X-100/10% (v/v) acetonitrile/4.5 mM dithiothreitol/10 mM iodoacetamide. The cleavage fragments were separated by reverse-phase HPLC by the method of Aebersold et al. [18] on an RP-18 LiChrospher column (250 mm × 4 mm; Merck) using water and HPLC-grade acetonitrile, each containing 0.1% (v/v) trifluoroacetic acid. All experiments were performed at a flow rate of 1 ml/min, with a linear gradient of 1.6–95% (v/v) acetonitrile over 70 or 98 min. The *A*₂₂₀ was checked with a Waters 996 photodiode detector. A second purification was performed for three fractions that gave poor resolution, with a non-linear gradient [10–24% (v/v) acetonitrile in 10 min, 24–74% in 65 min and 74–95% in 5 min] and heptafluorobutyric acid instead of trifluoroacetic acid. Peptide-containing fractions were dehydrated with a SpeedVac concentrator, then dissolved in 5 µl of water, acidified with 5 µl of formic acid and spotted on fibreglass coated with BioBreneTM. Sequencing was performed at the Institut de Biologie et de Chimie des Protéines (Université Lyon 1, Lyon, France) on an Applied Biosystems 470A gas-phase sequencer in accordance with the manufacturer's recommendations.

RNA isolation

Total RNA from different tissues was extracted as described previously [19]. Poly(A)-rich RNA was purified by affinity with oligo(dT)-cellulose [20]. The amount and purity of each RNA

sample were estimated from spectrophotometric measurements at 260 and 280 nm.

Reverse-transcriptase-mediated PCR (RT-PCR)

Poly(A)-rich RNA was used as a template for reverse transcription with an oligo(dT)-anchor primer [21] and Superscript II reverse transcriptase (Life Technologies). The resulting cDNA was used as a template for PCR; degenerate primers were designed from the amino acid sequences of purified peptide fragments.

The product was amplified by PCR. Each cycle consisted of 1 min at 94 °C, 1 min at different annealing temperatures and 2 min at 72 °C, repeated for 30 cycles, then 17 min at 72 °C for the last cycle extension step. The annealing temperatures tested were: 48 or 50 °C for 30 cycles, or 5 cycles at 42 °C and 25 cycles at 52 °C. The reaction was performed in 20 µl containing 50 pmol of each primer, 0.3 mM deoxynucleotides and 0.1 unit of *Taq* polymerase (Life Technologies) in the appropriate enzyme buffer. Separation of PCR products was performed on a 1% (w/v) agarose gel and DNA purification was performed on an anion-exchange column after solubilization of the gel-containing band (Smapur). Fragments were subcloned in a T/A cloning plasmid (pGEM-T Easy; Promega) and sequenced by the dideoxy chain-termination method [22].

Rapid amplification of cDNA ends (RACE) PCR

On the basis of the sequence of the fragment obtained by RT-PCR, specific primers were synthesized and used for RACE. Poly(A)-rich RNA (2 µg) was reverse transcribed with an oligo(dT)-anchor primer [21]. A nested PCR for 30 cycles was performed with the anchor primer in combination with a specific primer, then with a second specific primer. Each of the 30 cycles consisted of 1 min at 94 °C, 1 min at the annealing temperature (50 °C for the first 5 cycles and 60 °C for the 25 subsequent cycles) and 3 min at 72 °C, the extension step at 72 °C for the last cycle lasted for 17 min. The amplified product was purified and cloned as described above for RT-PCR.

For 5' RACE, the first-strand cDNA, 3'-tailed with terminal deoxynucleotidyl transferase and dCTP, was used as a matrix for second-strand cDNA synthesis, with 1 µM poly(G) adapter-primer, 50 µM dNTPs in one PCR cycle with 0.1 unit of *Taq* polymerase. The PCR conditions were as described above for 3' RACE. The amplified products were purified and cloned as for RT-PCR.

Sequence analysis

The signal peptide was located with the SignalP program [23]. Multiple sequence alignments were performed with the CLUSTAL program [24].

Northern blot analysis

Total RNA was separated on a formaldehyde/agarose (0.8%) gel and transferred on to a Hybond-N⁺ nylon membrane (Amersham). Equal RNA loading (10 µg) on gel lanes was checked by UV after staining of ribosomal RNA with ethidium bromide. Prehybridization and hybridization were performed at 65 °C in 5 × NaCl/sodium citrate/5 × Denhardt's solution/10% (v/v) dextran sulphate/0.5% SDS/250 µg/ml salmon sperm DNA with [α -³²P]dCTP random-primed DNA probe. After hybridization, membranes were successively washed twice for 20 min in 2 × NaCl/sodium citrate/0.1% SDS at room temperature and in 0.2 × NaCl/sodium citrate/0.1% SDS at 65 °C. Autoradiography was performed with Kodak Biomax MR film.

Hybridization *in situ*

The [α - 32 P]dCTP random-primed DNA was used as a probe for hybridization *in situ* to mRNA on *Riftia pachyptila* cryosections deposited on poly-(L-lysine)-coated slides. After three washes, sections were incubated in 0.2 M HCl for 10 min and acetylated with anhydrous acetic acid, then dehydrated in an ethanol series [70%, 95% and 100% (v/v)]. Dehydrated sections were incubated for 1 h at 25 °C in prehybridization buffer (4 × NaCl/sodium citrate/1 × Denhardt's solution) and the probe, dissolved in hybridization buffer [50% (v/v) formamide/10% (v/v) dextran sulphate/1 × Denhardt's solution/4 × NaCl/sodium citrate/500 µg/ml salmon sperm DNA], was deposited on each section for 16 h at 42 °C. After six washes (20 min each) in 4 × NaCl/sodium citrate, slides were dehydrated, coated with Ilford K5 emulsion and exposed for 4 days, then counterstained with light-green dye [12] before examination. Control sections fixed in Bouin solution were stained with Azan [12].

RESULTS

RP43 protein purification and microsequencing

Tube protein extracts from *Riftia pachyptila* were analysed by SDS/PAGE. Three different extraction solutions were tested and optimal results, i.e. a wide range of polypeptides sorted with a good extraction yield, were obtained with extraction buffer consisting of 10 mM Tris/HCl, pH 7.5, and 8 M urea. Under these conditions, several main bands were revealed after staining with Coomassie Blue (Figure 1A), the band with an apparent molecular mass of 43 kDa seeming to be a major one. Analysis of spots obtained after two-dimensional electrophoresis (Figure 1B) indicated that the 43 kDa band corresponded to a single polypeptide, called RP43.

This polypeptide was therefore chosen for further studies. Tube protein extract (500 µg) was separated by SDS/PAGE, the RP43 band was transferred to PVDF sheets and submitted to enzymic *in situ* cleavage. Tryptic fragments were separated by HPLC (Figure 2A). Of the five peaks collected (1, C, E1, E2 and P), three of them (C, E1 and E2), which were obtained with a poor resolution, were purified again as illustrated in Figure 2(B) for the E1 peak. Finally, five fragments were submitted to microsequencing and used for the design of oligonucleotides. A PIR/SwissProt/GenBank® database search for sequence identities showed that the different fragments share no identity with known proteins.

Isolation and sequence analysis of RP43 full-length cDNA

A PCR approach with cDNA from obturaculum, vestimentum, trunk and opisthosome was chosen to isolate the RP43 cDNA. Degenerate amplification primers were designed on the basis of the amino acid sequence of the five different tryptic fragments. Different PCR programmes were used and three fragments of around 500, 650 and 930 bp from the vestimentum cDNA were subcloned in pGEM-T Easy (Promega) and sequenced. An alignment of their sequences showed that they belonged to the same cDNA. The sequences of the four tryptic fragments E1, C, E2 and P were found in the deduced amino acid sequence from the three cDNA fragments obtained, allowing us to conclude that these fragments belonged to the RP43 cDNA.

The full-length cDNA (GenBank® accession number AF233595) was obtained by 5' and 3' RACE with specific primers deduced from the 930 bp cDNA fragment (Figure 3). Its 1708 bp sequence contained an open reading frame encoding a putative 421-residue protein with a predicted molecular mass of 44.206 kDa and a pI of 8.97. The five tryptic fragments and a 21-

Figure 1 Electrophoretic analysis of *Riftia pachyptila* tube proteins

(A) One-dimensional protein profile of a mature tube extract (45 µg loaded). (B) Two-dimensional electrophoretogram of proteins extracted from mature tube (100 µg). Arrows indicate the RP43 band; arrowheads indicate the positions of molecular mass markers of the sizes shown. pH values are indicated at the top of (B).

Figure 2 Reverse-phase HPLC fractionation of tryptic fragments from digested 43 kDa band

(A) Peptide peaks from digestion *in situ* of the 43 kDa band. Peaks 1 and P (in black) were microsequenced. Peaks C, E1 and E2 (in grey) were subjected to a second HPLC separation before microsequencing. (B) Second HPLC separation of E1 peak (in black).

```

1          ACGTCGGCTGATACTCGGTCTCGACGCAACTAGACCACAGCAGCACTGTTGAGGTACACAAGAAAA
67  ATG CGT GTC ATC TTT GTC ATA TCG CTG GTG TCT TTC ATG TTT GTG ACA TGG CAG ACG AAC CCT
   Met Arg Val Ile Phe Val Ile Ser Leu Val Ser Phe Met Phe Val Thr Trp Gln Thr Asn Pro
130  GTC CAC TGC CGA CAA CAC AAG ACA CCG CAA GTG GTC GAT CTT GGA TAT TCT GGA GCG AAA GTC
   Val His Cys Arg Gln His Lys Thr Pro Gln Val Val Asp Leu Gly Tyr Ser Gly Ala Lys Val
22  Val His Cys Arg Gln His Lys Thr Pro Gln Val Val Asp Leu Gly Tyr Ser Gly Ala Lys Val
193  GAA ACG GTG ATG AAA ACA TCT GAT GTT GTA TAT GCG AAA CCA ATG TGT CAC GGA GGA GCG TGC
   Glu Thr Val Met Lys Thr Ser Asp Val Val Tyr Ala Lys Pro Met Cys His Gly Gly Arg Cys
43  Glu Thr Val Met Lys Thr Ser Asp Val Val Tyr Ala Lys Pro Met Cys His Gly Gly Arg Cys
256  CCT CCT CCT CCG CCA ACG AAA ACA TGT AGC AAG CCA TCG TCA CTA AAG GCA GTG TCA AGA GGG
   Pro Pro Pro Pro Pro Thr Lys Thr Cys Ser Lys Pro Ser Ser Leu Lys Ala Val Ser Arg Ala
64  Pro Pro Pro Pro Pro Thr Lys Thr Cys Ser Lys Pro Ser Ser Leu Lys Ala Val Ser Arg Ala
319  GGA CAT GCA GGA TAC TTT GGA ACA ATT GGT AGT GGA AAA TAC AAG AAC AAT GAG AAT TGT TGG
   Gly His Ala Gly Tyr Phe Gly Thr Ile Gly Ser Gly Lys Tyr Lys Asn Asn Glu Asn Cys Trp
85  Gly His Ala Gly Tyr Phe Gly Thr Ile Gly Ser Gly Lys Tyr Lys Asn Asn Glu Asn Cys Trp
382  TGG AGT ATC CAG GCA CCG GCA GAT CAG CGA CTG CGT CTA CAC TTT GAG TCA TTA TAC CTG GAG
   Trp Ser Ile Gln Ala Pro Ala Asp Gln Arg Leu Arg Leu His Phe Glu Ser Leu Tyr Leu Glu
106  Trp Ser Ile Gln Ala Pro Ala Asp Gln Arg Leu Arg Leu His Phe Glu Ser Leu Tyr Leu Glu
445  TAC GGA AAC AGA TAC TGT CCA TAC GAT TAC CTG AAC ATT CAC AAC GGA AAG AAT GAG AAC GGA
   Tyr Gly Asn Arg Tyr Cys Pro Tyr Asp Tyr Val Asn Ile His Asn Gly Lys Asn Glu Asn Gly
127  Tyr Gly Asn Arg Tyr Cys Pro Tyr Asp Tyr Val Asn Ile His Asn Gly Lys Asn Glu Asn Gly
508  CAA CGT TTA GGA AAA TAT TGT GGA ATA AAG GCG CCT GAC GAT ATA ATA ACT ACT GGA AAT TCT
   Ala Arg Leu Gly Lys Tyr Cys Gly Ile Lys Ala Pro Asp Asp Ile Ile Thr Thr Gly Asn Ser
148  Ala Arg Leu Gly Lys Tyr Cys Gly Ile Lys Ala Pro Asp Asp Ile Ile Thr Thr Gly Asn Ser
571  GGT TAT GTC AGA TTT TCG ACT GAT AGC TCT GTA ACA AAA TAC GCG TTC AGG TTG CGT TAT TCT
   Ala Tyr Val Arg Phe Ser Thr Asp Ser Ser Val Thr Lys Tyr Gly Phe Arg Leu Arg Tyr Ser
169  Ala Tyr Val Arg Phe Ser Thr Asp Ser Ser Val Thr Lys Tyr Gly Phe Arg Leu Arg Tyr Ser
634  ATT GTG TCT TGC AAG AGT TTA CAT ACT ATA TTC GGT TCA CAC GGT GAA TTT GGA ACG GTG GGG
   Ile Val Ser Cys Asn Ser Leu His Thr Ile Phe Gly Ser His Gly Glu Phe Gly Thr Val Gly
190  Ile Val Ser Cys Asn Ser Leu His Thr Ile Phe Gly Ser His Gly Glu Phe Gly Thr Val Gly
697  CCG GCA TAC GCG AAC TTC GAG GAA TGT TCT TGG AAG ATT GAG GTT CCA AAT CGA AAA CGA GCG
   Pro Ala Tyr Gly Asn Phe Glu Glu Cys Ser Trp Lys Ile Glu Val Pro Asn Gly Lys Arg Val
211  Pro Ala Tyr Gly Asn Phe Glu Glu Cys Ser Trp Lys Ile Glu Val Pro Asn Gly Lys Arg Val
760  CGT CTG CAC TTC AAA CGT TTC CAC CTG GAG AAG AAC AAC AAA TAT TGC CCG TAC GAT AAA CTG
   Arg Leu His Phe Lys Arg Phe His Leu Glu Asn Asn Asn Lys Tyr Cys Pro Tyr Asp Lys Leu
232  Arg Leu His Phe Lys Arg Phe His Leu Glu Asn Asn Asn Lys Tyr Cys Pro Tyr Asp Lys Leu
823  AAA ATC TAC GAT GGA TCG AGC GCC ACC GCG TCA CTG AAA GGC ACT TTA TGT GGG AAA AGA CGT
   Lys Ile Tyr Asp Gly Ser Ser Ala Ser Ala Ser Leu Lys Gly Thr Leu Cys Gly Lys Arg Arg
253  Lys Ile Tyr Asp Gly Ser Ser Ala Ser Ala Ser Leu Lys Gly Thr Leu Cys Gly Lys Arg Arg
886  CCG AAA GAT ATC GAG TCG ACA GGA AAG ACT ATG TTC GTG ACG TTC TCG TCA GAC ATC TCG TTG
   Pro Lys Asp Ile Glu Ser Thr Gly Lys Thr Met Phe Val Thr Phe Ser Ser Asp Ile Ser Leu
274  Pro Lys Asp Ile Glu Ser Thr Gly Lys Thr Met Phe Val Thr Phe Ser Ser Asp Ile Ser Leu
949  ACA TTC CCG GGT TTC CGT ATT CAG TAT TCA GTT CCA GCC TCG TGT AGC GTC GTG AAT GAG TTG
   Thr Phe Pro Gly Phe Arg Ile Gln Tyr Ser Val Pro Ala Ser Cys Ser Val Val Asp Glu Leu
295  Thr Phe Pro Gly Phe Arg Ile Gln Tyr Ser Val Pro Ala Ser Cys Ser Val Val Asp Glu Leu
1012 ACT GGA CCG TCG CGA ACC TTC GGC ACC ACA GGT TCC CAG TAT GAG AAT AAC GAG CTC TGT TCC
   Thr Gly Pro Ser Gly Thr Phe Gly Thr Thr Gly Ser Gln Tyr Glu Asn Asn Glu Val Cys Ser
316  Thr Gly Pro Ser Gly Thr Phe Gly Thr Thr Gly Ser Gln Tyr Glu Asn Asn Glu Val Cys Ser
1075 TGG AAG ATT GAA GTC GCC CAG AAT AAA CGA GTG CTT CTT CAC ATC TAC CGA TTC AAC ATA GAG
   Trp Lys Ile Glu Val Ala Gln Asn Lys Arg Val Leu Leu His Ile Tyr Arg Phe Asn Ile Glu
337  Trp Lys Ile Glu Val Ala Gln Asn Lys Arg Val Leu Leu His Ile Tyr Arg Phe Asn Ile Glu
1138 GTT GAG GGC AAT TGC GAA TAC GAC AGC TTG ACA CTC TAC AAA GGG CCT AAC GAC AGC GCA CGA
   Val Glu Ala Asn Cys Glu Tyr Asp Ser Leu Thr Val Tyr Lys Gly Pro Asn Asp Ser Ala Pro
358  Val Glu Ala Asn Cys Glu Tyr Asp Ser Leu Thr Val Tyr Lys Gly Pro Asn Asp Ser Ala Pro
1201 ATG TTG GGG AAG TAT TGC GGA GAA ACA ATC CCA ACG TTC ATG GTG TCC TCC GGG AAT ACT ATG
   Met Leu Gly Lys Tyr Cys Gly Glu Thr Ile Pro Thr Phe Met Val Ser Ser Gly Asn Thr Met
379  Met Leu Gly Lys Tyr Cys Gly Glu Thr Ile Pro Thr Phe Met Val Ser Ser Gly Asn Thr Met
1264 TTC ATT AAG TTC CAA ACT GAT GGC TCG GCG AGA AAA CCC GGC TTC TGG ATC GGC TAC ACA ACA
   Phe Ile Lys Phe Gln Thr Asp Gly Ser Ala Arg Lys Pro Gly Phe Trp Ile Gly Tyr Thr Thr
400  Phe Ile Lys Phe Gln Thr Asp Gly Ser Ala Arg Lys Pro Gly Phe Trp Ile Gly Tyr Thr Thr
1327 GGC TAA TGAAGCTCAAAGGGAATCGGTGTCGGTGGCAGTTGCGTTTGCCATGGCAGTCCAGTAGTCATGCTCGTCGGA
   Gly *
421  Gly *
1407 TGGCCACCGTTATTGCTGCCAGCTACTGATGTAGCCATCGGAAGCGTGCTCAGCCGACGCGAACTAACTGTTGCAACTGTC
1490 GTAATATACTAAATAATTCATATCTGCGCAACAGTACACCACTTTAATGACATGTGCAACGATAGCAGAAATGCATGCAA
1573 TACCAGTATAACCATACGTACCAATCCTCGTCTTGACAAATGAAACGGCTTTTACGTCGGAACGCTGACTGTGGTGGCGCC
1656 CGATTCTCAAAAATAAAATTGATTGAAAGTCTCAAAAAAAAAAAAAAAAAAAAAA
1708

```

Figure 3 *RP43* cDNA and deduced amino acid sequences

The predicted signal peptide sequence is underlined and the putative cleavage site is indicated by a downward arrow. The putative polyadenylation site is shown in bold. The sequences of the five tryptic fragments are doubly underlined. The positions of the three CUB domains are enclosed in shaded boxes.

residue predicted signal peptide were found. Three CUB domains (100–110-residue-spanning domains first reported in the complement subcomponents C1r/C1s and also in epidermal-growth-factor-related sea urchin protein and bone morphogenetic protein 1) were also localized at positions 84–188, 193–303 and 309–418 (Figure 3). A multiple sequence alignment (Figure 4) of these three *RP43* domains showed 71–73% of amino acids corresponding to the consensus CUB signature defined by Bork and Beckmann [25]. In addition, if reference is kept by the conserved 15 hydrophobic and four aromatic residues (in bold in Figure 4) forming the core of the domain structure [26], agreement reaches 80% for hydrophobic residues and 100% for aromatic ones. It is important to note that in both cases most of non-corresponding amino acids are similar to exceptions known in other CUB domains. If only the three *RP43* CUB domains are considered in the alignment, 36–45% pairwise amino acid identity can be

observed, which is consistent with results obtained into other multi-CUB proteins [25]. When all CUB domains present in Figure 4 are subjected to pairwise identity alignments with the three domains of *RP43*, identities obtained range from 23% to 36% with an average of 29%. Similar percentages are observed in overall CUB domains pairwise identity alignments [25].

Distribution of the *RP43* transcript

The expression pattern of *RP43* was examined in different anatomical regions. The obturacular region, with its branchial plume, is the primary exchange surface of the animal. The vestimentum serves to hold the plume out of the tube opening in the external seawater. The trunk houses endosymbiotic bacteria and the opisthosome acts as a holdfast, against which the worm can move into the tube. Northern blot hybridizations were

Figure 4 Multiple alignment of RP43 CUB domains with other proteins

The three CUB domains of RP43 (RP43/Rifpa-1 to 3) were aligned with: A5p/Xenla-1 (SwissProt accession number P28824), the first CUB of *Xenopus laevis* neuronal cell surface type I membrane protein; BMP1/Human-1 (P13497), the first CUB of human bone morphogenetic protein 1; BMP10/Parli-1 (P42674), the first CUB of sea urchin blastula Zn-metalloprotease 10; C1r/Human-2 (P00736), the second CUB of human complement C1r component; TLD/Drome-5 (P25723), the fifth CUB of *Drosophila melanogaster* dorso-ventral patterning Tollid metalloprotease; and TSG6/Human (P98066), the only CUB of human tumour-necrosis-factor-inducible protein. The accession numbers in parentheses were used for consultation about the structures and functions of these proteins as well as for primary references. The two disulphide bonds (—S—S—) between nearest-neighbour cysteine residues (Cys-1—Cys-2 and Cys-3—Cys-4) are indicated by broken lines. Amino acids present in all proteins, or in all except one, are boxed in black and those with no exception are highlighted by an asterisk. Amino acids with the same characteristics and present in all proteins are indicated in grey and also with a white circle. Characteristics are defined as the following: a, aromatic (F, Y, W); h, hydrophobic (L, I, V, M, A); t, turn-like or polar (G, P, N, Q, H, S, T); —, negatively charged (E, D); +, positively charged (R, K). Conserved hydrophobic, aromatic, turn-like, negatively and positively charged positions defined in [25] as a signature for CUB domains are in the 'CUB sign' line. Particular positions forming the core of the domain as described in [26] are shown in bold. Identical amino acids in the three CUB domains of RP43 are indicated with an x.

Figure 5 Northern hybridization analysis of *Riftia pachyptila* RNA with RP43 cDNA

Total RNA (10 µg) from tissues from the four main regions of the worm were loaded on each lane. Lower panel: UV detection of ethidium bromide-stained total RNA in the formaldehyde/agarose gel before transfer, showing equal loading. The tissues chosen were the obturaculum region (Ob), vestimentum (V), trunk (T) and opisthosome (Op). The blot was hybridized with a ³²P-labelled RP43 cDNA probe (top panel). The approximate length of transcript detected, estimated with RNA molecular mass markers, is indicated at the right.

performed on 10 µg of total RNA isolated from obturaculum, vestimentum, trunk and opisthosomal regions (Figure 5). With the 500 bp PCR fragment as a probe, a single transcript of 1.8 kb (a size consistent with that of RP43 cDNA) was observed in vestimentum and trunk but no signal was detected in the opisthosome or in the obturaculum (Figure 5).

To determine the spatial distribution of the RP43 transcript, hybridization analysis was performed *in situ* on *Riftia pachyptila* sections with the 500 bp cDNA probe. RP43 mRNA was restricted to epithelial cells under apical cuticular plaques (Figure 6A) that are present only in the vestimentum and in the anterior part of the trunk body wall. No label was observed in other outer epidermal cells or in other tissues. Azan-stained sections (Figures 6B and 6C) permitted the identification of the different vestimentum tissues hybridized *in situ*. The specificity of this localization *in situ* was checked by using another epidermal cDNA probe that labelled epidermal cells but not the apical cuticular plaque cells (results not shown).

DISCUSSION

RP43, a CUB-domain-containing protein

The electrophoretic protein profiles of *Riftia pachyptila* tube were already analysed in a previous study [11]. Thirteen major electrophoretic bands, grouped in four triplets (approx. 23, 28, 35 and 43 kDa), were found in different tube extracts. The relative intensities of these bands differed from one tube to another, suggesting individual variability in the tube protein pattern. In any one tube, the pattern was also different between the inside layers and external ones, which were synthesized earlier. However, the major 43 kDa band was always present in these different extracts; from two-dimensional electrophoresis analysis it corresponds to a single polypeptide. The fact that the amino acid sequence deduced from the full-length cDNA obtained contains the five tryptic sequences confirms that this cDNA corresponds to the 43 kDa polypeptide. Moreover, the deduced molecular mass and pI are consistent with the values estimated from the electrophoretic study.

The CUB domains found in RP43 indicate that this protein is a new member of a large (more than 151) and functionally diverse protein group possessing one or several spanning domains of approx. 110 residues [25] that bind ligands. Structure analysis on crystallized CUB domains showed that they are built from ten β-strands arranged into a sandwich of two five-β-stranded sheets, each sheet of the sandwich containing one parallel and four anti-parallel β-strands [26]. The first CUB proteins were described in C1r and C1s [27,28], epidermal-growth-factor-related protein (fibropelline) of the sea urchin [29], the vertebrate bone morphogenetic protein 1, similar with *Drosophila* Tollid [30], and spermadhesins (reviewed in [31]).

All these proteins possess a signal peptide, implying extra-cellular localization. The presence of a signal peptide in the RP43 sequence confirms that this polypeptide extracted from the extracellular tube structure is actually an extracellular component. Sequence alignments showed that the RP43 CUB domains present 19 highly conserved positions forming the core of the domain [26] and that it is a pertinent member of the family of proteins containing CUB domains. Almost all CUB domains contain four cysteine residues, which probably form two disulphide bridges (Cys-1—Cys-2, Cys-3—Cys-4) as shown for C1r and C1s [32] and spermadhesins [33–35]. The fact that C1r, C1s and spermadhesins have the same arrangement of the disulphide bridges is a further indication of their structural similarity and of their functional relationships. However, several CUB-domain residues are not conserved, suggesting that the different domains might have distinct binding specificities and thus different ligands, which might be carbohydrates, glycoproteins or phospholipids [31].

No CUB proteins have been found in arthropod cuticles or in invertebrate exoskeletons but most of them are known to be involved in binding processes, facilitating cell–cell or matrix–

Figure 6 *In situ* hybridization analysis with *RP43* cDNA

Light micrographs of apical cuticular plate structures from the vestimentum of *Riftia pachyptila*. (A) Cross cryosection of vestimentum hybridized with a ^{32}P -labelled *RP43* cDNA probe. Arrows highlight labelling. (B) Azan-stained vestimentum section at same magnification as in (A). (C) Enlargement of an apical cuticular plate in vestimentum. Scale bars: 140 μm (A, B), 20 μm (C). Abbreviations: cp, cuticular plaque; ec, epidermal cells; ep, epithelium; lm, longitudinal muscle; um, underlying muscle; pg, pyriform gland (also called chitin-secreting gland).

matrix interactions. We can therefore propose that *RP43* might bind a ligand that could be either proteins or chitin, the two main tube components known so far. However, chitin binding seems unlikely because all other sugar-binding CUB proteins have acid polysaccharidic ligands [36], whereas chitin is neutral.

Spermadhesins (AWN, AQN-1 and AQN-3), each with a single CUB domain, bind mainly glycoproteins of the zona pellucida surrounding the oocyte or sperm surface [31]. Cubilin, possessing 27 CUB domains, binds the intrinsic factor–vitamin B12 complex by its CUB domains 5–8, and also a receptor-associated protein by its CUB domains 13–14 [37]. In *Drosophila*, Tollid binds the protein Decapentaplegic, a member of the transforming growth factor β family [30]. From these studies it is more likely that *RP43* is involved in protein–protein interactions than in direct chitin binding. However, the possibility cannot be dismissed that *RP43* might establish some bridges with chitin via an association with glycosylated chitin-binding proteins that are also present in the tube (results not shown).

Distribution of the *RP43* transcript

From the Northern blot analysis, the *RP43* mRNA was localized only in vestimentum and trunk body wall. This distribution is in agreement with the finding that these two regions are involved in tube wall synthesis because they are rich in chitin-producing glands [3]. In contrast, the obturacular region, which is lacking in such glands and is therefore considered a non-secreting exoskeleton tissue [3], was found to be lacking also in *RP43* transcripts. However, the correlation between the distribution of *RP43* mRNA and chitin-producing gland regions cannot be

applied to the opisthosome. Indeed, this basal part presents two types of gland [3], including those that secrete chitin, and is thought to be involved in tube wall production. The second type of gland would provide additional proteins acting in the plugging of the tube distal part [3] but we failed to detect *RP43* mRNA in this region.

The hybridization *in situ* revealed that the *RP43* mRNA is restricted to cells underlying apical cuticular plaques associated with a set of muscles homologous with adductor and abductor muscles described in the vestimentiferan worm *Lamellibrachia barhami* [38]. Both Van der Land and Nørrevang [39] and Webb and Ganga [38] have suggested that this system assists in the movement of the animal within its tube. These plaques are numerous on the outer surface of the vestimentum and on the anterior part of the trunk epidermis [3]. This distribution could explain the greater abundance of *RP43* mRNA in the vestimentum wings than in the whole trunk body wall observed in Northern blot analysis. In the same way, its absence in the opisthosome could easily be related to the lack of plaques in this area.

Such a restricted distribution of the *RP43* mRNA is somewhat surprising because it is generally admitted that the chitin-secreting glands (the so-called pyriform glands) are the site of secretion of tube components [3,5,10]. Putative protein material linked to chitin microfibrils has been observed at the ultrastructural level in these tissues, i.e. before their secretion outside the organism [2,5]. Nevertheless, no defined results have been published on protein synthesis by these glands. Our results indicate clearly that a major tube protein is synthesized in specialized areas of the outer epithelium and that at least two different tissues are

involved in the exoskeleton synthesis. The chitin-secreting glands, which were previously thought to be the site of synthesis of tube wall components [10,40], seem to be mainly devoted to chitin production, whereas the outer epidermis would be specialized in the synthesis of major protein components. Tube wall formation would therefore be a two-step process: first, chitin crystallites are polymerized and bound to chitin-binding proteins in the gland lumen, and secondly, bridging proteins such as RP43 are involved in final embedding outside the glands, i.e. in the surrounding seawater, at the meeting point of chitin microfibrils with major proteins.

Conclusion

The present characterization of a new CUB-domain-containing protein, involved in the construction of the *Riftia pachyptila* tube, opens the way to further studies including the characterization of other major protein components, binding experiments and functional studies with recombinant mutated proteins. Such approaches could lead to a good understanding of protein–protein and protein–chitin interactions in this exoskeleton, which is probably a better model for the study of these interactions than arthropod cuticles, in which chitin fibres are very thin and the protein fraction is far more complex.

We thank the chief scientists Jim Childress, Horst Felbeck, Lauren Mullineaux and François Lallier for sampling tube worms, and Juliette Ravaux and Bruce Shillito for help with tissue collection. This work was supported by funding from INSU and Dorsales.

REFERENCES

- Gaill, F. (1993) Aspects of life development at deep sea hydrothermal vents. *FASEB J.* **7**, 558–565
- Shillito, B., Lechaire, J. P. and Gaill, F. (1993) Microvilli-like structures secreting chitin crystallites. *J. Struct. Biol.* **111**, 59–67
- Jones, M. L. (1985) Hydrothermal vents of the Eastern Pacific; an Overview, vol. 6, Biological Society of Washington, Washington DC
- Gaill, F. and Hunt, S. (1986) Tubes of deep sea hydrothermal vent worms *Riftia pachyptila* (Vestimentifera) and *Alvinella pompejana* (Annelida). *Mar. Ecol. Prog. Ser.* **34**, 267–274
- Gaill, F., Shillito, B., Lechaire, J. P., Chanzy, H. and Goffinet, G. (1992) The chitin secreting system from deep sea hydrothermal vent worms. *Biol. Cell* **76**, 201–204
- Shillito, B., Lübbering, B., Lechaire, J.-P., Childress, J. J. and Gaill, F. (1995) Chitin localization in the tube secretion system of a repressurized deep-sea tube worm. *J. Struct. Biol.* **114**, 67–75
- Lutz, R., Shank, T., Fornari, D., Haymon, R., Lilley, M., Von Dam, K. and Desbruyères, D. (1994) Rapid growth at deep-sea vents. *Nature (London)* **371**, 663–664
- Gaill, F., Shillito, B., Ménard, F., Goffinet, G. and Childress, J. J. (1997) Rate and process of tube production by the deep-sea hydrothermal vent tubeworm *Riftia pachyptila*. *Mar. Ecol. Prog. Ser.* **148**, 135–143
- Shillito, B., Ravaux, J., Gaill, F., Delachambre, J., Thiebaut, E. and Childress, J. J. (1999) Preliminary data on carbon production of deep-sea vent tubeworms. *Mar. Ecol. Prog. Ser.* **183**, 275–279
- Southward, E. C. (1988) Development of the gut and segmentation of newly settled stages of *Ridgeia* (Vestimentifera); implications for relationship between Vestimentiferan and Pogonophora. *J. Mar. Biol. Ass. U.K.* **74**, 265–278
- Ravaux, J., Chamoy, L. and Shillito, B. (2000) Synthesis and maturation processes of a vent worm exoskeleton (*Riftia pachyptila*). *Mar. Biol.* **136**, 505–512
- Martoja, R. and Martoja-Pierson, M. (1967) Initiation aux Techniques de l'histologie Animale, vol. 1, Masson, Paris
- Bradford, M. M. (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72**, 248–254
- Laemmli, U. K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T₄. *Nature (London)* **227**, 680–685
- Towbin, H., Staehelin, T. and Gordon, J. (1979) Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc. Natl. Acad. Sci. U.S.A.* **76**, 4350–4354
- O'Farrell, P. Z., Goodman, H. M. and O'Farrell, P. H. (1977) High resolution two-dimensional electrophoresis of basic as well as acidic proteins. *Cell* **12**, 1133–1142
- Wray, W., Boulakas, T., Wray, V. P. and Hancock, P. (1981) Silver staining of proteins in polyacrylamide gels. *Anal. Biochem.* **118**, 197–203
- Aebersold, R. H., Leavitt, J., Saavedra, R. A., Hood, L. E. and Kent, S. B. (1987) Internal amino acid sequence analysis of proteins separated by one- or two-dimensional gel electrophoresis after *in situ* protease digestion on nitrocellulose. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 6970–6974
- Bouhin, H., Charles, J. P., Quennedey, B., Courrent, A. and Delachambre, J. (1992) Characterization of a cDNA clone encoding a glycine-rich cuticular protein of *Tenebrio molitor*: developmental expression and effect of a juvenile hormone analogue. *Insect Mol. Biol.* **1**, 53–62
- Celano, P., Vertino, P. M. and Casero, Jr, R. A. (1993) Isolation of polyadenylated RNA from cultured cells and intact tissues. *Biotechniques* **15**, 26–28
- Frohman, M. A., Dush, M. K. and Martin, G. R. (1988) Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc. Natl. Acad. Sci. U.S.A.* **85**, 8998–9002
- Sanger, F., Nicklen, S. and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5463–5467
- Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**, 1–6
- Higgins, D. G. and Sharp, P. M. (1988) CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**, 237–244
- Bork, P. and Beckmann, G. (1993) The CUB domain. A widespread module in developmentally regulated proteins. *J. Mol. Biol.* **231**, 539–545
- Romero, A., Romao, M. J., Varela, P. F., Koln, I., Dias, J. M., Carvalho, A. L., Sanz, L., Topfer-Petersen, E. and Calvete, J. J. (1997) The crystal structures of two spermadhesins reveal the CUB domain fold. *Nat. Struct. Biol.* **4**, 783–788
- Leytus, S. P., Kurachi, K., Sakariassen, K. S. and Davie, E. W. (1986) Nucleotide sequence of the cDNA coding for human complement C1r. *Biochemistry* **25**, 4855–4863
- Tosi, M., Duponchel, C., Meo, T. and Julier, C. (1987) Complete cDNA sequence of human complement C1s and close physical linkage of the homologous genes C1s and C1r. *Biochemistry* **26**, 8516–8524
- Delgadillo-Reynoso, M. G., Rollo, D. R., Hursh, D. A. and Raff, R. A. (1989) Structural analysis of the uEGF gene in the sea urchin *Strongylocentrotus purpuratus* reveals more similarity to vertebrate than to invertebrate genes with EGF-like repeats. *J. Mol. Evol.* **29**, 314–327
- Shimell, M. J., Ferguson, E. L., Childs, S. R. and O'Connor, M. B. (1991) The *Drosophila* dorsal–ventral patterning gene *taloid* is related to human bone morphogenetic protein 1. *Cell* **67**, 469–481
- Topfer-Petersen, E., Romero, A., Varela, P. F., Ekhlasi-Hundrieser, M., Dostalova, Z., Sanz, L. and Calvete, J. J. (1998) Spermadhesins: a new protein family. Facts, hypotheses and perspectives. *Andrologia* **30**, 217–224
- Gagnon, J. and Arlaud, G. J. (1985) Primary structure of the A chain of human complement-classical-pathway enzyme C1r. N-terminal sequences and alignment of autolytic fragments and CNBr-cleavage peptides. *Biochem. J.* **225**, 135–142
- Sanz, L., Calvete, J. J., Mann, K., Schafer, W., Schmid, E. R. and Topfer-Petersen, E. (1991) The amino acid sequence of AQN-3, a carbohydrate-binding protein isolated from boar sperm. Location of disulphide bridges. *FEBS Lett.* **291**, 33–36
- Sanz, L., Calvete, J. J., Mann, K., Schafer, W., Schmid, E. R., Amselgruber, W., Sinowatz, F., Ehrhard, M. and Topfer-Petersen, E. (1992) The complete primary structure of the spermadhesin AWN, a zona pellucida-binding protein isolated from boar spermatozoa. *FEBS Lett.* **300**, 213–218
- Sanz, L., Calvete, J. J., Mann, K., Schafer, W., Schmid, E. R. and Topfer-Petersen, E. (1992) The complete primary structure of the boar spermadhesin AQN-1, a carbohydrate-binding protein involved in fertilization. *Eur. J. Biochem.* **205**, 645–652
- Manaskova, P., Meszarosova, A., Liberda, J., Voburka, Z., Ticha, M. and Jonakova, V. (1999) Aggregated forms of heparin-binding and non-heparin-binding proteins of boar seminal plasma and their binding properties. *Folia Biol. (Praha)* **45**, 193–201
- Kristiansen, M., Kozyraki, R., Jacobsen, C., Nexø, E., Verroust, P. J. and Moestrup, S. K. (1999) Molecular dissection of the intrinsic factor–vitamin B12 receptor, cubilin, discloses regions important for membrane association and ligand binding. *J. Biol. Chem.* **274**, 20540–20544
- Webb, M. and Ganga, K. S. (1980) Studies on *Lamelliibrachia barhami* (Pogonophora). III. Plaques, glands and epidermis. *Ann. Univ. Stellenbosch Ser. A*, **2**, 1–27
- van der Land, J. and Norrevang, A. (1977) Structure and relationships of *Lamelliibrachia* (Annelida, Vestimentifera). *K. Dan. Vidensk. Selsk. Biol. Skr.* **21**, 1–102
- Gardini, S. L. and Jones, M. L. (1993) Vestimentifera. In *Microscopic Anatomy of Invertebrates*, vol. 12 (Harrison, F. W. and Rice, M. E., eds), pp. 371–460, Wiley-Liss, New York