Chapter 3

Results

1. Characterization of purified vitellin

In this study, Vt from the ovaries of vitellogenic *P. merguiensis* females was purified and characterized for its chemical properties. Vt was purified by DEAE-Sephacel column and Sephadex G-200 column, respectively. The electrophoretic patterns of fractions from purification process were run on 4-10% nondenaturing PAGE as shown in Fig. 20. Subunits of purified Vt were characterized by 5-15% SDS-PAGE as shown in Fig. 21. The N-terminal amino acid sequences of the 78, 87 and 104 kDa subunits were determined by Edman degradation. The N-terminal amino acid sequence of the 78 kDa and 87 kDa subunits were found to be identical and have sequence APWGADVPR whereas that of 104 kDa subunit is SIDSSVIADF.

The isoelectric point of purified Vt is 5.3, determined by isoelectro focusing, Fig. 22, indicating that *P. merguiensis* Vt is an acidic protein. The elution profile of purified standard and Vt when determining the amino acid composition by using an amino acid analyzer is shown in Fig. 23. The amino acid composition of the purified *P. merguiensis* Vt is similar to Vt from other penaeid shrimps and *M. rosenbergii* (Table 8). In addition, the amino acid composition of the primary structure of *P. merguiensis* Vt obtained from Vg cDNA was almost identical to purified Vt with the exception of the cysteine content. There are some cysteines in the Vg primary structure, but cysteine was undetectable in purified Vt (Table 8). Vt does not have protease activity when tested by using skim milk as substrate (Fig. 24).

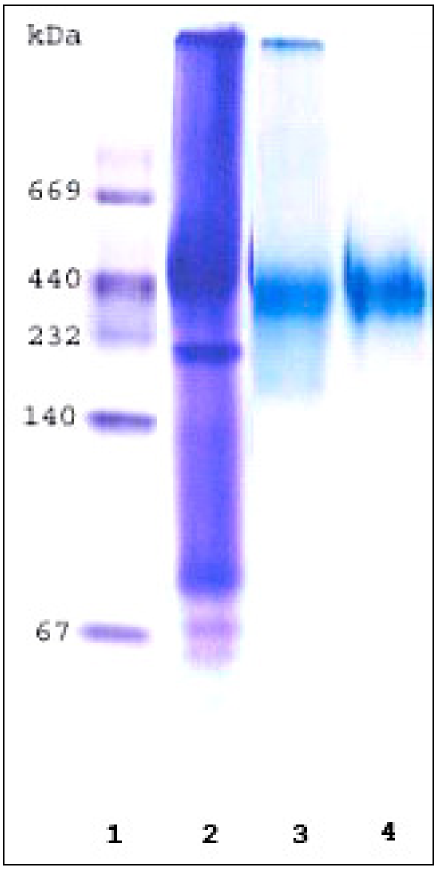


Fig. 20 4-10% Nondenaturing PAGE of Vt at various steps of purification.

The gel was stained with Coomassie Blue. Lane 1, molecular weight markers; lane 2, crude ovarian extract; lane 3, fractions containing Vt from DEAE-Sephacel column; lane 4, purified Vt from Sephadex G-200 column.

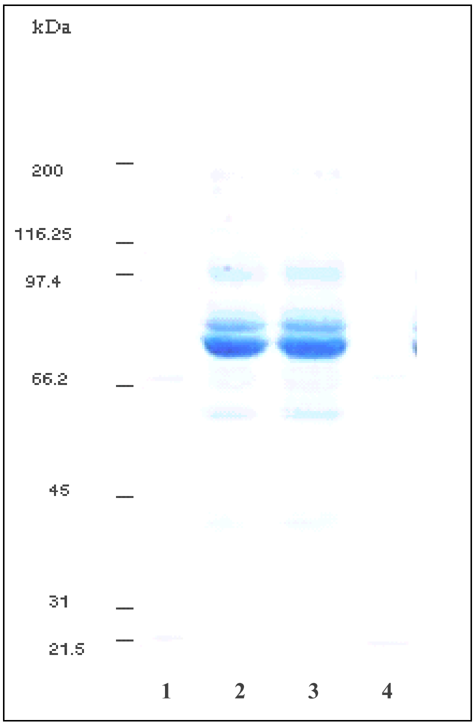


Fig. 21 5-15% SDS-PAGE of purified Vt.

The protein bands were stained with Coomassie Blue. Lane 1 and 4, molecular weight markers. Lane 2, purified Vt in the presence of β-mercaptoethanol. Lane 3, purified Vt in the absence of β-mercaptoethanol.

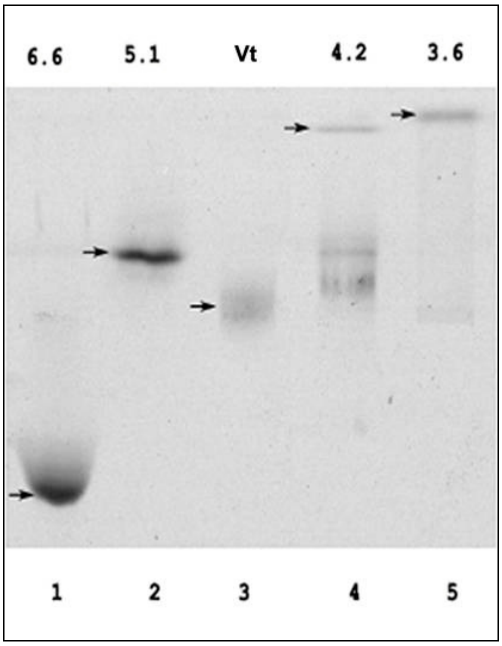


Fig. 22 Isoelectric focusing of purified Vt.

The protein markers used were: carbonic anhydrase, pI 6.6 (lane 1), β-galactosidase, pI 5.1 (lane 2), glucose oxidase, pI 4.2 (lane 4), and amyloglucosidase, pI 3.6 (lane 5). Lane 3 is purified Vt. The arrows represent protein bands with corresponding pI values.

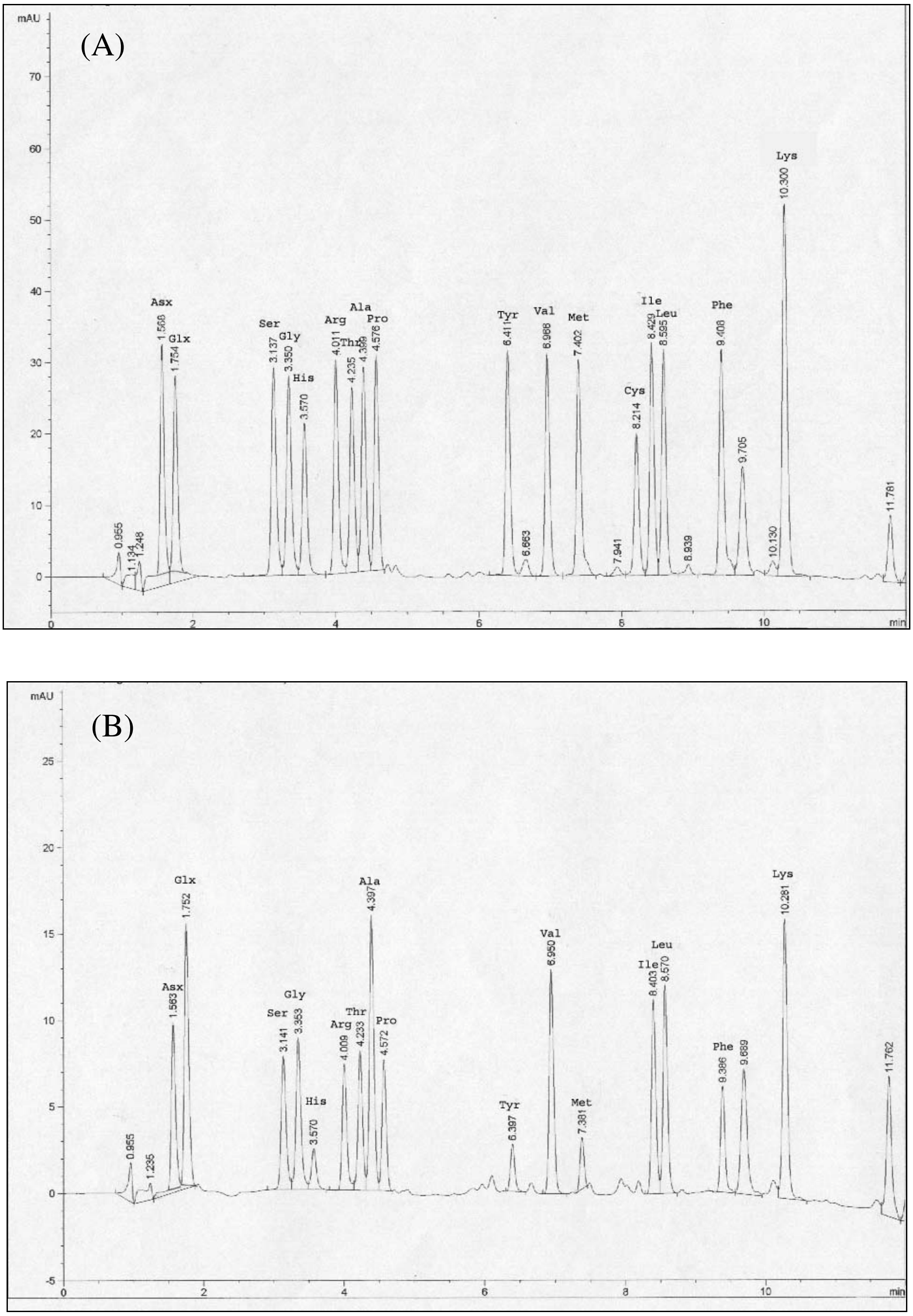


Fig. 23 Elution profiles of standard (A) and purified Vt (B) when determining the amino acid composition.

Table 8 Amino acid composition of purified Vt from *Penaeus merguiensis* and other shrimps.

Species	*Penaeus Merguiensis*		*Penaeus chinensis*	*Marsupenaeus Japonicus*	*Penaeus Monodon*		*Penaeus Semisulcatus*	*Litopenaeus Vannamei*	*Metapenaeus ensis*	*Macrobrachium rosenbergii*
Amino Acid	Vt[a]	Vt[b]	Vt[c]	Vt[d]	Vt[e]	Vt[f]	Vt[g]	Vt[g]	Vt[h]	Vt[i]
Lys	6.70	5.84	6.3	7.1	6.1	6.12	5.33	6.26	8.15	7.56
Arg	5.04	4.91	5.2	5.2	5.5	4.71	4.8	4.65	0.96	6.35
His	2.40	2.40	2.1	2	2.4	2.52	1.89	2.5	2.08	2.67
Asx	6.72	6.50	7.3	9.2	7.7	7.28	6.67	6.24	9.03	8.26
Glx	12.84	11.21	12.8	18.8	13.4	13.06	12.11	11.39	13.77	12.07
Tyr	1.93	3.13	2.7	2	2.4	2.94	3.33	3.16	2.01	2.54
Cys	UD	1.31	0.9	0	0.6	1	ND	1.31	ND	UD
Gly	6.73	6.23	6	9.7	7.2	6.04	8.3	7.05	6.62	8.38
Ala	10.99	11.76	11	9	11.7	10.83	10.62	10.78	13.21	6.99
Val	8.59	8.55	8.4	6.3	7.1	8.27	7.64	7.8	9.84	6.92
Leu	8.05	6.84	7.7	7.3	7.1	7.16	7.27	7.35	8.25	8.76
Ile	7.31	7.15	7	4.5	6.2	7.06	6.69	6.67	6.82	4.89
Phe	3.95	3.63	3.9	3.7	3.5	3.56	3.7	3.81	4.07	3.37
Try	ND	0.66	0	0	ND	ND	ND	0	ND	ND
Met	1.71	2.63	2.8	3.6	1.4	2.59	2.45	2.24	2	0.76
Ser	5.89	6.03	5	6.8	6.4	5.96	7.27	7.37	5.07	8.83
Thr	6.27	6.19	5.6	5.9	6.3	6.08	5.63	6.08	0.14	6.35
Pro	4.86	5.03	5.3	0	5.6	4.83	6.27	5.36	7.98	5.4

[a]From amino acid composition (this study); [b]From Vg cDNA (this study); [c]From Chang *et al*., 1996; [d]From Vazquez-Boucard and Ceccaldi, 1986; [e]From Chang *et al*., 1993; [f]From Quinitio, *et al*., 1990; [g]From Tom *et al*., 1992; [h]From Qui *et al*., 1997; [i]From Lee *et al*., 1997b. Each value was expressed as % mole of total amino acids. ND : not determined, UD : undetectable.

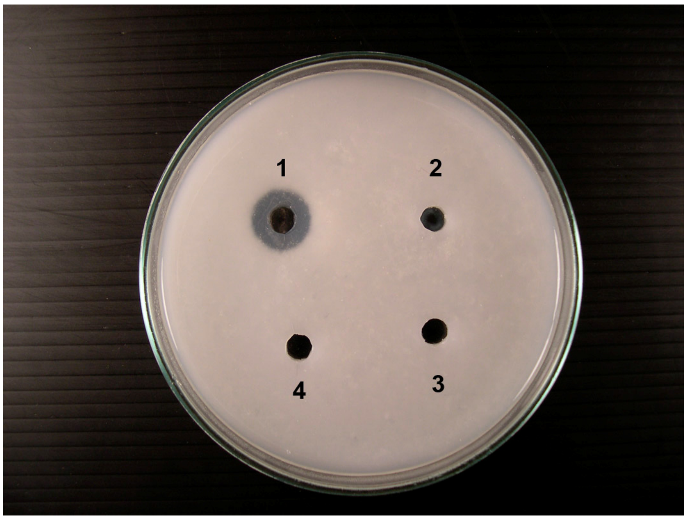


Fig. 24 Determination of protease activity of purified Vt.

Well 1, positive control (proteinase K, 100 μg); well 2, purified Vt (130 μg); well 3, purified Vt (260 μg); well 4, negative control (distilled water).

2. Characterization of cDNA encoding Vg from the ovary

2.1 Isolation of a cDNA fragment encoding the N-terminal region and full length of *P. merguiensis* Vg

A degenerate primer (FVg97) was designed on the basis of the N-terminal amino acid sequence of purified Vt found earlier by Edman degradation. The reverse primer (RVg1201) was constructed based on the conserved sequences of Vt reported for other shrimps. By using RT-PCR and this pair of primers (FVg97 and RVg1201), a PCR product of about 1,100 bp was obtained from the ovary of vitellogenic *P. merguiensis* females. The cDNA was cloned and sequenced. Sequence analysis showed that the product encoded the 9 N-terminal amino acids that matched exactly the experimentally determined 9 N-terminal amino acid sequence (APWGADVPR) of the 78 and 87 kDa subunits of purified Vt, and included a peptide sequence that was not encoded by any of the PCR primers. These initial data indicated that cDNA from a Vg gene encoding the N-terminal region had been cloned. The 5' and 3' fragments of Vg cDNA were cloned using the RACE technique. A full-length cDNA of *P. merguiensis* Vg was reconstructed by overlapping nucleotide sequences from the results of walking RT-PCR and RACE methods (Fig. 25, 26 and 27). To eliminate errors in cDNA reconstruction and to confirm that all cDNA fragments obtained from previous experiments were amplified from the same full-length Vg cDNA, one pair of specific primers (FVg24 and RVg 7786) was designed to amplify a large RT-PCR fragment. Its PCR product was 7.7 kb and contained ORF sequence (Fig. 28). Its nucleotide sequence was 99% identical to that of the full-length Vg cDNA previously determined.

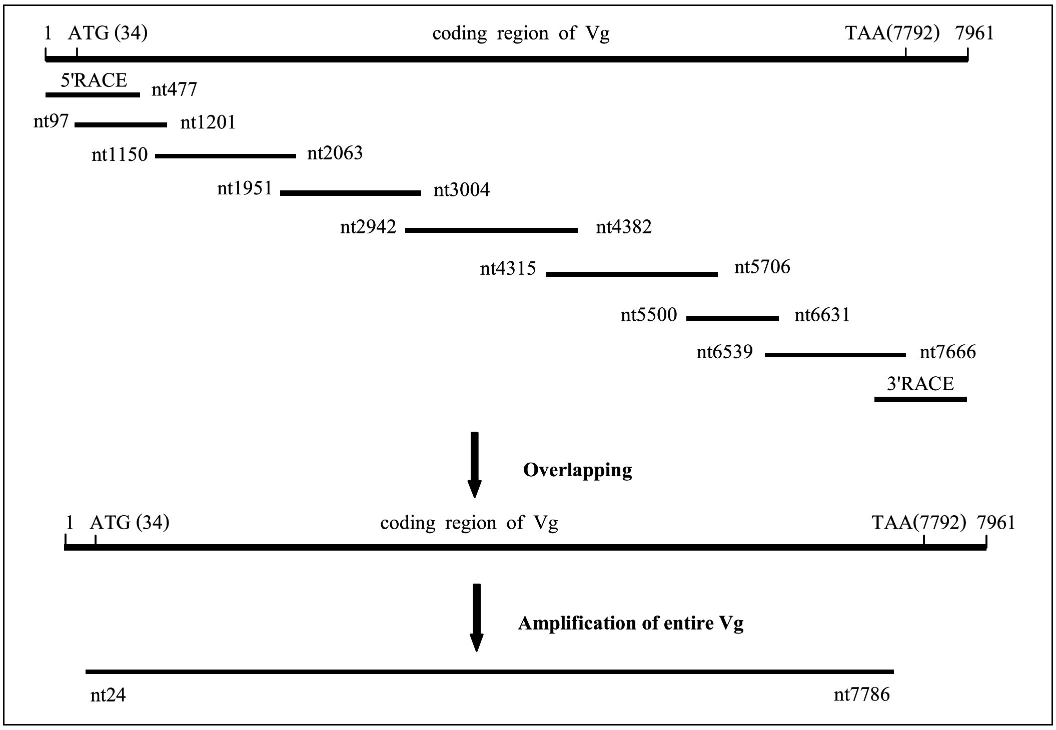


Fig. 25 RT-PCR cloning strategy and schematic view of Vg cDNA.

An initial PCR amplification was obtained by using a primer encoding the N-terminal sequence of Vt 78 kDa subunit. Contiguous sequence data with individual clones were obtained by primer-walking. Both 5' and 3' RACE were used to clone the 5' and 3' ends of the cDNA. Overlapping sequences were used for reconstruction of a full-length Vg cDNA.

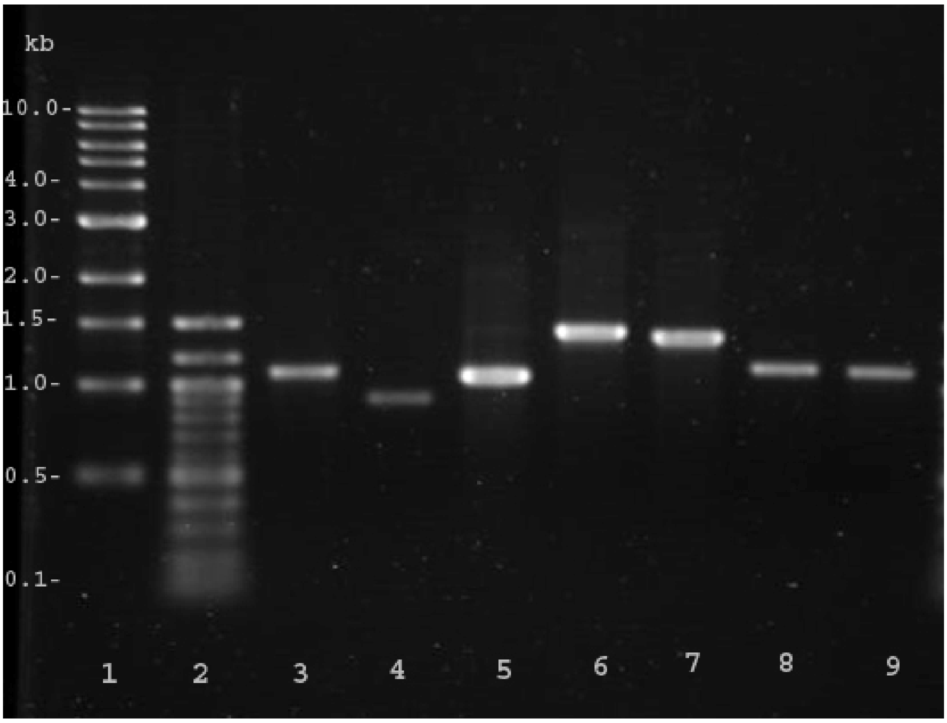
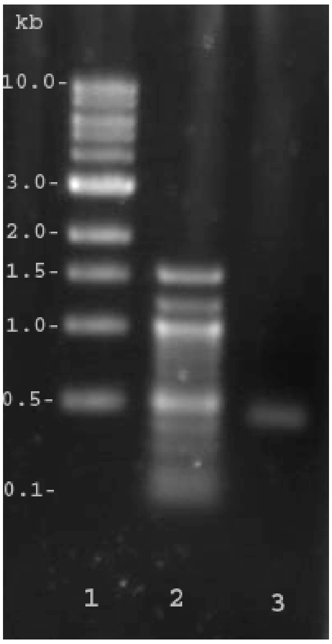
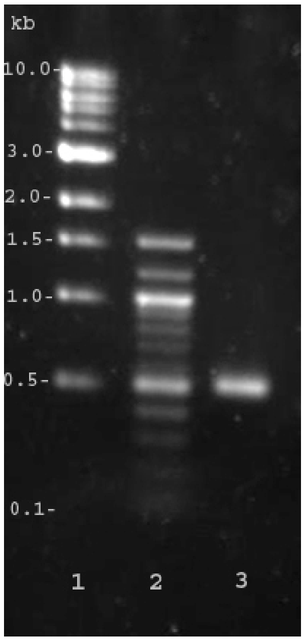


Fig. 26 Amplification of cDNA encoding the Vg cDNA fragments.

The PCR product was analyzed on 1% agarose gel electrophoresis. Lane 1, 1 kb DNA ladder marker; lane 2, 100 bp DNA ladder markers; lane 3-9, OneStep RT-PCR product amplified of cDNA encoding Vg of *P. merguiensis.*



(A)



(B)

Fig. 27 Vg cDNA electrophoresis patterns of 5' RACE and 3' RACE fragments.

The PCR product was analyzed on 1% agarose gel electrophoresis. Lane 1, 1 kb DNA ladder markers; lane 2, 100 bp DNA ladder markers; lane 3 (A), DNA fragments of 3' RACE; lane 3 (B), DNA fragments of 5' RACE.

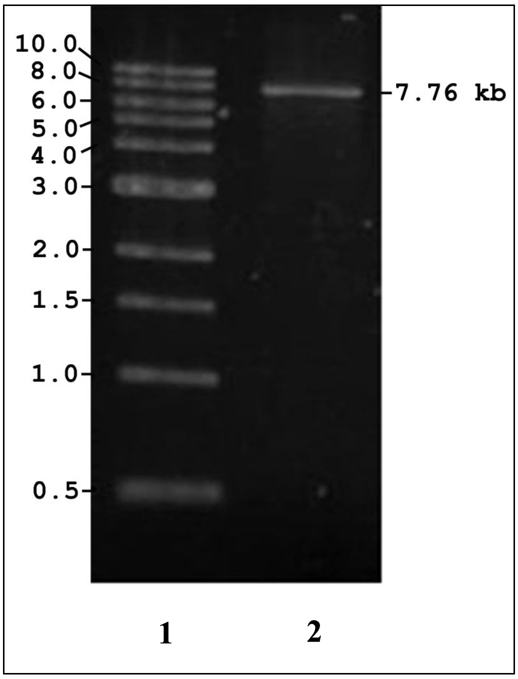


Fig. 28 Vg cDNA electrophoresis pattern of Vg open reading frame.

The PCR product was analyzed on 1% agarose gel electrophoresis. Lane 1, 1 kb DNA ladder markers; lane 2, PCR producted of a large fragment of Vg open reading frame.

2.2 Characterization of cDNA and deduced amino acid sequence of *P. merguiensis* Vg

The full-length cDNA of *P. merguiensis* Vg, obtained from walking RT-PCR cloning, RACE and confirmed by amplification of the entire Vg cDNA sequence was deposited in GenBank with accession number AY499620 and consists of 7,961 nucleotides (Fig. 29). Sequence analysis revealed the presence of a single large open reading frame of 7,758 bp, a 5' untranslated region (33 bp), a stop codon (TAA), a polyadenylation signal AATAAA 14 nucleotides upstream of the site of poly (A) addition and a 3' untranslated region (167 bp) (Fig. 30). The open reading frame encodes a peptide that comprises 2,586 amino acids. A signal peptide containing 18 contiguous highly hydrophobic amino acid residues was predicted by the eukaryotic signal peptide prediction program SignalP 3.0 (Bendtsen *et al*., 2004). This indicates that the protein is cleaved between amino acid residues 18 and 19 and exported from the cell. The mature protein has a molecular mass of 283,029 Da and its theoretical pI value is 6.36 as estimated by ProtParam (Gasteiger *et al*. 2005). Based on the N-terminal sequence of the 78 and 87 kDa subunits of purified Vt and the predicted signal peptide sequence, the molecular mass of the pro-Vg in *P. merguiensis* was predicted to be 281,235 Da.

1 GAAAAGTAGT CGTGGTGTGC TCGTCCAACC ACCATGACGA CCTCAACACT CCTCTTCGTT CTCGCTTTCG 70

71 TGGCAGGTGG TCTGGCAGCC CCCTGGGGAG CGGACGTGCC AAGATGCTCC ACCGAATGCC CCGTCACCGG 140

141 ATCCCCCAAA CTGGCCTACC AACCTGACAA GACCTACGCC TACGCCTACT CCGGCAAGTC CACAGTCCAG 210

211 CTCAAGGGCG TGGACAACGG CGACACTGAG ACCGAGTGGA CGGCAGGAGT TGATCTCACC TGGATCAGCC 280

281 CTTGCGACAT GGCCATCTCC TTCAGGAACA CCAAGATGGA TGGCGCCCGA GGTCCCACCG CTGCCAGGAC 350

351 GCTGGAGAGA TATCCACTGG TGGTGGCCGT TGTCGACGGG AGGGTGCAGC ATGTGTGCGC TCACCCAGAG 420

421 GACGAAGCAT GGGCCATCAA CCTGAAAAAG GGCGTTGCTT CGGCTTTCCA GAACTCCATC CCTTCTCTGT 490

491 CTGCTGTCAG CTCAGGCATC ACAGTAACTG AGACCGATGT CGTTGGAAAA TGCCCGACAA AGTATGAAAT 560

561 TGAGACCGAA GGAGAGAAGG TCATTGTTGT CAAGGAGAAG AACCACCGCC ACTGTCAAGA ACGTTACCCA 630

631 ACACCTGCTG CATTACCTGC ACCATGGCTG AAGGCTCCCC TGCCGATCGA GGAATCCAAA TCGCAGTGCA 700

701 GGCAGGAAAT CGCCAATGGC ATTTACACCG CCATCACGTG TCAGGACAAG AACATCGTTC GACCTGCCAT 770

771 TGGAATCTAC AAGTATGTGG AGGCCAGTCA GTATTCAACA CTTCGCTTCA TCTCCGAGTC CTCCGACACT 840

841 TCAGCCATCA GCGGCATCCC TTCAGGAGAA CTGAACATTG AAAGCCTGTT GTACAACCAC GAAACAATGA 910

911 AGGACCCACA GCTGGCACCT GAGTTGGATG AGCTCATGAA GGAGATCTGT GACAAGACCA AGGACACAGT 980

981 TGAGGCTGAA GCTGGTGCTT TGGTTGCCAA GGCTCTCCAT GTGTTACGTC GCATTCCTGA CACGGTTGTG 1050

1051 GTGGAGACTG CACAGAAAGT GAGACAAGGA CATTACTGCA GTGACTCTGC CAGGCTGGAG AGCATCTTCT 1120

1121 TGGACGCAGT TGCTTTCATC CACGAGTCTG GTGCAGTCAA GGTCATGGTC AACGAAATCG AGAATGGACG 1190

1191 AGCAACAGGG GGACGTCTCG CTCTGTACAC GGCAGCGCTC TACCTCATCC CACGACCCAA TATTGAGGCA 1260

1261 GTCAAGGCCC TCACACCACT GTTTGAAAGC CCTCGCCCAA TGCCCTCGTT GCTGCTGGCA GCTGCTACCA 1330

1331 TGGTAAACCA TTACTGCCGC CATACTCCAG CTTGCTATGA GAAAGCCCCA GTTGCGAGAA TTGCCGAGAT 1400

1401 TCTGGCCAAC AGAGTCCAGA CTCACTGCTC TCCTTCTGCT GGTGTTGAGG ACAACGAAGT AGCCCTTGCA 1470

1471 ATTTTCAAGA CAATTGGGAA TATGGGTGTA GCTACACCTG CCGTGACAAG GGCAGCCGTT CACTGCATTG 1540

1541 AGGTAGAAGG ACTGGAAACC AGCGTTCGGG TAGCTGCAGC TGAAGCCTTC AGACAAGCCA ACTGCTTCCG 1610

1611 TCCAGCAGTT GAGAAGCTAG TTGACATTGC TGTCCGACCA GCCTTTGAGA CCGAAGTCCG CATTGCTTCA 1680

1681 TATCTGGCTG CTGTTCGATG TGCTGAGCAG GAACATCTGG AAACAATTAT TGAAAAGATT TCAAAGGAAG 1750

1751 AGAATACTCA AGTTCGTGGA TTTGTTTTGG GTCACCTGAT TAACATCCAA GAGTCTACCT GCCCTGCCAA 1820

1821 GGAAAATCTC AGGTACCTCC TTGCCAACGT CATCCCTACC GACTTCGAGA GGGACTTCAG GAAATTCTCT 1890

1891 CGAAATATAG ATGTGGCTTA CCATGCCCCT GCCTTTGGCA TGGGTGCCGG CCTCGAATCG AACATCATCT 1960

1961 ATGCTCCCGG ATCTTTCGTT CCTCGTGCTG TTAACTTGAA AATGAAAGCA GATGTGGATG AGACGCACAT 2030

2031 GGACATTGCA GAGATTGGTG CGCGCTTTGT AGGAATTGAG TCCATCATTG AAGAACCCTT GGGCCCCCAG 2100

2101 GGATACCTAC GCACAGCAAC ATTTGGAAAG ATTATGGAGG ACATCACAGG TTTTGCAGGA GAGAAAGGCT 2170

2171 ACAAGGTCAT GGAGCAGCTC AAGCACACTC TGAGGACCAG GCGATCCATC GATTCTTCCG TCATCGCGGA 2240

Fig. 29 Full length Vg cDNA nucleotide sequence from *Penaeus merguiensis* was deposited in GenBank with accession number AY499620.

Nucleotide position is shown on left and right hand sides. Start codon (ATG), stop codon (TAA) and polyadenylation signal (AATAAA) are underlined.

2241 CTTCTTTGGT AAACTGTATG GTAAGAGCAG ATCACATACC CACGCCGAAC TATTCGCCCG ATTCATGGGA 2310
2311 CACGAGATTA CTTACGCAGA TGTTGCCGGA AGCCTCAAAG GCGTCACAGC TGACACACTC ATTGAGACCT 2380
2381 TCTTCTCTTT CTTCGAGAAT TCTCTGGAGC ATATGAAGGA TCTCAACCTG AACACAGCAA GAACTGCTCA 2450
2451 GCTTTCCATG GATTACTCAC TGCCGACCAT TCAGGGAACA CCGCTCAGAC TGAGATTAGC TGGAACTGCT 2520
2521 GTTGCTGGCC TGAAGATGGA AGGCAACGTC AACATTGCCC AGATCCTGTC CGACCTCGGC AACTCCCAGA 2590
2591 CGGGCGTTAA ATTCTTCCCA GGCCTTTCTG TACACGCCAC TGGTTTTGTT GGCTTTGACT GGCTCCTTGC 2660
2661 CAGGGTGGGA ATCGAGATGC AGAACACCAT CTCTAGTGCC ACTGGAGCTG CCATCAAGAT CAGAACAACT 2730
2731 GAAAACAAGA AGATCGAGAT GGAATTGGAG GTCCCTGAGA AGATGGAACT CCTCAACATC AAGGCCGAGA 2800
2801 CTTACCTTGT CAAGGCTGTG GGCAAGAAGA TGACCGAGAT TTCTCCTTCC TCCATGAGAG ATGTCAGGAT 2870
2871 TCAGCGCAAC TCCTGCATTG GTGCTTTGGA ACCAGTATTT GGCCTCAAGG TGTGCTACGA CATGAACATC 2940
2941 CCTGACGTCT TCCGTGCTAA TGCCCTGCCA CTTGGTGAAC CAGCCATTGC CAAGCTGTAT GTTGAGAAGG 3010
3011 CAGATCCTTC CATGAGAGGT TACTTAGTGA CTGCTGCTAT CAAGAACAAG AGGGGTAACA AGGTCATTAA 3080
3081 GATGAATGTC GAAGCAGCTG GTGCCTCAAC ACCAAGAAGA GCAGAAATGA CCCTGTCCTA CACCAAGGAA 3150
3151 GAAGGAAGCC ACATTGTTTC TGCCAAGCTT GATTCCTCCA GCATTGCTGC AGGAGTGTGG ACCACTCTTA 3220
3221 CCAACGAGCA AGGACACAAG GCAATGGAGA CTTATGTCAA CTTCAATTAC GGTCAGATTG CTATTTCTCG 3290
3291 AGGCATTAAG CTGGAGGCGA TTGCAAGGGA AGCAAGTGTG GGAGAGGAGT TCCAAGTTAA TGTGTTCAGC 3360
3361 AGCGGCACCA GGAGCTTCCC CTCTGAATCA CACATTGTGG AGGCTAGATT CATCAAGAAA ACTAGTGGAC 3430
3431 CTGAATTCAA TGTGGATGTG ATTTGCAGGA CCAAAAATGC TCTAGCTGAA TTATTCGACT TAAATATTGA 3500
3501 AGTTGGAGCT GATTTCATGA AATTTTCTCC TAAGAATCTG TATCCTGCAA GATACATCCC CAAGACCCGC 3570
3571 ATTGTCCTAC CCGTAAACCT GCGAAAGATG GAAATCAATG CCGCCACTGC AGCCTGGAAA CTGATATCGT 3640
3641 ATATCCGTGC AGGAAGTCAA TCTGGCGGAA GCCGTGAGTT CATTTCTGCT CTCAAGCTTG CCAAAGGAAG 3710
3711 GAAGGATTTC ATCTCTGTAC AGGCTACTCA TACGATAGAA GGAACTTTCC CACAAAACAT CATCATCAAA 3780
3781 AATGTAGCAA CAGCCGAAGT TGGCAGATCA TCATACAAGG CAATGTATGA TCTGTTCTAT CACTCTGAGA 3850
3851 AAATGGGTGC TTCTCTTGAG GTTTTGCAAG CAGCTGGTAA TGAGAAGGTT GCCCATTTGG AAGCAATTTA 3920
3921 CGAACTTTCA GGATCGAAGT ACTGCACTAA ATTCTTGGCG GAGATTCCTG GCTACATTCA GCCAGTTAAA 3990
3991 GTTGAAGCCG GGATTGAACA AGGAGCGGAA GGTCGCTATA CACTAGAATC CGCCATCACA TATGGACAAC 4060
4061 GTACAGTACT TGAAGCTAGC GGACCAATCA TGGCTCGTTT CAGCTCCAAG ATTGCCAAGC TGCAAGCCAA 4130
4131 CATCAAAGTC AGGGCAATGG CAAGTGAACC CTATATCATT GGTGCCAATG TTGCGTTTGG CAGCAAGAAG 4200
4201 CAGATGATAG CCATGGAAAT CAAAGGTCGA TCAGAAGCTG TCATTGGTCT TGAATGGAAA ATGGTCCGAG 4270
4271 AAAGTTCCGA GAAAACCACT GTCGGCATAG TGTTCGTGCT CCCTGCCCTT ATCGAGAACA AGATCGACGC 4340
4341 TGAAATTACT GATGGACTTA TCCATGTTAG CTTCAACAAC CTGGTTCTGC CCAAGACCTC ATCCCGCCGT 4410
4411 CGGGTCAAGG GATTCGCTGA TGTTCATATT GCAGAGAAAA AGGCAAATGT GGAATTTTCT TGGGACGCCG 4480
4481 ATAATGCTCC TGAAAAGAAG TTGGTGTTGG ATGCAAGTCT GATCAGCAGC TCTGCCAATC CTGGACATGC 4550
4551 TGAGATCCAC GGGAATATTG TCATTGCTGG AGAGCCTTTC CACGCCAAAC TGGTTCTGAC TGCCGCAAAT 4620
4621 CTGGTAGAGC ACATGGAAGG GGAAAATGGA TTTAAGTTGA TCCTGACAAC GCCTAGCCAG AAGACGATTG 4690

Fig. 29 (Continued)

4691 TCTTGGGTGC CTCCTGTGAT GTCCAGGTGG CAGGAGCCAC GACTAAAGTC ATTTCCACCG TCGAATACAA 4760

4761 GAACATGAAG GACAGGAAAT ACAAATATGC AAGTGTGATT GCCTATGAGA AGCTTGGTGG CCCATTTGAT 4830

4831 TATGCAGTAG AAGCCAAGGT AACTTACAAA CAACCTGGAA CAGCAGAAAT AAAGCTCGGA ACAGCAGTAA 4900

4901 AACACCACTG GACACCAGAA GAACATGTTG TGGCAGTCAG GGTAGAAGCT GATGCTCCAA TTCTGAAGAC 4970

4971 ACCTGCCACA ATTGAATTCT CCATTCACAA TGCACCAAAC GCATTTGTTG GATTCTGCAA GATCGAAAGA 5040

5041 AATGCTCCAG CCACTGTGTT TGAATGGAAT GTGCAGATCA CACCTGAAGG AGGAATTGAG GCTGTTGAAG 5110

5111 CTGGTGTAGA CATGAAAGCC ATCATTGAAG TTCTGAAGAT TGTTCATGCC GCTGCTACTC TCCAGGAAGA 5180

5181 AAGTTATGAA ACCTATGGTC CACACACATC CCAATACCAG TACCGTTTCA CAAGGCCATC ACCCACCTCT 5250

5251 TACACCATGC AGATGAGGAC TCCGACCCGT ACCATGGAAG GAAGAGCCAA ACTATCACCA AGGGAATCTG 5320

5321 GAATCAAGTT CTACCCCAAT AAGGGCAAAA CTGAGTCCAA ATACGAAATT GGATACAAGG CCAACCACGA 5390

5391 GGGAAGGTGG GGAGGACATG CGTCCAAGTT GGAAGTGAGA ATGAACCATC CAGTGCTTCC TAAGCCCATC 5460

5461 ATGGCCGCTG TTCAGTACAC AGTAGCTGAA GAAACCACCA AGGGAACAAT TGAACTGGAC ATTTTCCCAG 5530

5531 AAGAAGCCAA CAAAATTACT GGTTCATTGG AAACTCAGAG AATTTCAGAA AATGCTATCA GAGCAGAAGC 5600

5601 CTTCTTGACT AGCAGGATGT TGAAAGTGAA CCCTAAGGCA ATCATCACTG CTGCTTATGC ACCAGAGACA 5670

5671 GTTGCTTTTG ATGTAGTGTT CCACAAGACT CCATCTGCAG CACCAATCTT TGCCATTGCT GCCAAGTATG 5740

5741 ACAAGACTGC AGCTCACAAT GCAGCTGCCA CATTCACAGT GAAGATGGAA GAGCGACCTG TCTTTGAAAT 5810

5811 CACTGCAGTG ACAGAACCCG AGGAAGCAGC CACCTGCAAT GGCATCAGAA TGAGGGCTGT TGCTTATGCC 5880

5881 GCAACTTTTG GAAAGTACAA CGTTATCTCC AAGATGTGCA GGCCTGCCTT CATTGAGGTG ACCGCAATGC 5950

5951 GGCCTGGTGG AGCAAAGGAA TACATTGCCA AGCTTGGCCT CCGATACCCT GACGCTGCTG AAGCAGGTGT 6020

6021 ATATGTGGCT AGCGGCAGAG CTGGAGAGAC TCACGGTGTT GCTGTTGCTG CTGTGAAGTT GGCTTCACCC 6090

6091 ACAATGCTTA AGGTCGAGAT GGCTTATGAA CCGCAAGAAG CAGAGGCAAT AATCAACGAA ATGACTGAAG 6160

6161 AATTTGAGAA GATCGCCGCA TCATTCACGT CTGTTGAAAT GGAGGTTGTC GAGTTCCTCA AGCAAGAAGC 6230

6231 TGCTGCAAAG GGTATTCAGT TCCCTTCGTC TCAGTTAGTC AACCTACTGG GAGTTGCAAA GGAGGAAATT 6300

6301 GCAGAGATCT ATCGAGATAT TGTCTCTGAG GCAATAATTT TTGATACCGA AATCCTTGGT GATATCCTGG 6370

6371 GAAGTCCTGT GGTGTCCTTC ATATCACGGG TCTACTTCGG TGTGTGGTCG GAGATAATTC ATCTTCAACA 6440

6441 CCATCTTTCC GTAAGCCTCA TCCAGACGAT AGAGAGATTC CAGCAGGAAT TAGGAAGCAT TTCTGAAATC 6510

6511 TTAATGGAAG TTGTGATGAC AGCAGCACGA ATGGCAGAAA CTGGAGAAGT TCCCGGAGTA GTGTTTGATG 6580

6581 CACTTGAAGA AATCAAAGCC ACCAAAGTCT TCAGGATTGT AAGGAGGGAA GTGGACGCAA TTTTAGAAGA 6650

6651 ATATCCGGAG GAGTATGAGG CTGTAAAACA CATTGTTCAT AATGTGGTTG CAATTCTCAA GCGAGATGTT 6720

6721 GGCATTGTTC GTGAGAGGCT CATGGAGATT CCAGCTGTTC TGAAAATCAT CGACTACACT ATGTATCACT 6790

6791 TCCATTCGGA ACGTGCATTT GCTGCAGAAG CGGAAAAGCT CGTCAGCCTG ATGCTCAACG AACTTCTTTT 6860

6861 TGTTTCAATG GAAAGAGAAG GCAACGGTGT TGCAGTCCGA ATTCCCCTCC ACCGACCCTT GTATTCACTG 6930

6931 ACGCAAGTGG CACAAGAAGC AGTGCCCAAC CCTGTGACAA TGCTTGAGAA CCTGATATTT GCATACGTTG 7000

7001 ATTACATTCC CATCCCTGTG AGCGACGCAA TCTGGGCCTT CTACAACTTG GTACCACGCT ACATTACGGA 7070

7071 CGTGCTGCCA CCCTACCCTC GAACAGCCAC GGTGGTTGGC GGCAGCGAGA TCCTGACCTT CAGCGGCCTT 7140

Fig. 29 (Continued)

7141 GTTGTGCGAG CACCTCGGTC GCCGTGCAAG GTTGTTCTAG CTGCTCACGG GTCCCACCGC CTCATGATGT 7210

7211 CCCACCCACA AGCCTCAGCC CCGGCACAGT TTGAATTGAA GACACCAGCA GCCACCGTGA TGATCAAGCC 7280

7281 TGACTTTGAG GTGGTGGTTA ATGGGCAACC CCTCGCGGGA TCCCAGCAAA CCATCGGAAA CGTTAGGATT 7350

7351 GTGAACACAG CCGAGCACAT TGAGGTGGGA TGTCCCCTAA TGAAGGTCGT GGTTGCCAAG GCAGGCGAGG 7420

7421 CCGTAGCCGT TGAGGCTTCA GGTTGGGTGT TTGGACGCGT GGCAGGCCTA TTGGGTCCCA ACAATGGAGA 7490

7491 AATCGCCCAT GACCGTCTCA TGCCCAGTGG TGCAGCAGCC TCCAACCCCC GCGATTTGGT AGCTGCTTGG 7560

7561 CAGGAGGACC GCCAGTGCTC CACCCCTGAG GTTCCTCGTG CAGAGACCAC AGTAGCTCGT CTGATTCAGT 7630

7631 GTGAAGCATT GCTGGGGATT CGGTCCAGGT GTAACCCAGT GGTTCACCCT CAGCCATTCA TTAAGATGTG 7700

7701 TCACGCTGCC CACAAGGCTT GTGATGCTGC CCAAGCTTAC AGAACCGTTT GTTCTCTGAG AGGAGTGGGA 7770

7771 GAAGTTTTCC CTTTGGGATG TTAACAACAC CTGTTGACAT GTTTACATTG CCTAAAAAGT TGACGACTCG 7840

7841 TTCAGATTTT GGATATGGCT ATAGTTGTTA ATGTATTTTT GTTACGTGAA AGCATCATAA AATATACAAA 7910

7911 AAAAAATAGA TCTCCAATAA AATTGCAACC AACTAAAAAA AAAAAAAAAA A 7961

Fig. 29 (Continued)

(A)

1	GAA	AAG	TAG	TCG	TGG	TGT	GCT	CGT	CCA	ACC	ACC	**ATG**	ACG	ACC	TCA	45
1												**M**	T	T	S	4
46	ACA	CTC	CTC	TTC	GTT	CTC	GCT	TTC	GTG	GCA	GGT	GGT	CTG	GCA	GCC	90
5	T	L	L	F	V	L	A	F	V	A	G	G	L	A	A	19

(B)

7741	AGA	ACC	GTT	TGT	TCT	CTG	AGA	GGA	GTG	GGA	GAA	GTT	TTC	CCT	TTG	7785
2570	R	T	V	C	S	L	R	G	V	G	E	V	F	P	L	2584
7786	GGA	TGT	**TAA**	CAA	CAC	CTG	TTG	ACA	TGT	TTA	CAT	TGC	CTA	AAA	AGT	7830
2585	G	C														2586
7831	TGA	CGA	CTC	GTT	CAG	ATT	TTG	GAT	ATG	GCT	ATA	GTT	GTT	AAT	GTA	7875
7876	TTT	TTG	TTA	CGT	GAA	AGC	ATC	ATA	AAA	TAT	ACA	AAA	AAA	AAT	AGA	7920
7921	TCT	CC***A***	***ATA***	***AAA***	TTG	CAA	CCA	ACT	AAA	AAA	AAA	AAA	AAA			7959

Fig. 30 5' Ends (A) and 3' ends (B) of nucleotide sequence of *Penaeus merguiensis* Vg cDNA.

This shows the untranslated regions and includes amino acid translation of the translated regions. The start codon (ATG) and stop codon (TAA) are in bold. A consensus polyadenylation signal (AATAAA) is shown in bold italics. Nucleotide or amino acid position is shown on left and right hand sides.

The deduced amino acid sequence of *P. merguiensis* Vg contains multiple copies of a consensus dibasic cleavage site, R-X-(K/R)-R (Arg724 to Arg727) or R-X-X-R (Arg613 to Arg616, Arg724 to Arg727, Arg942 to Arg945, Arg945 to Arg948, Arg1731 to Arg1734, Arg2194 to Arg2197 and Arg2372 to Arg2375) potentially cut by the subtilisin family of serine endoproteases (Barr 1991; Chen *et al*., 1997) as shown in Fig. 31. Cleavage at the RTRR site that starts at amino acid 724 would produce two subunits of molecular weights 78 and 203 kDa. The N-terminal amino acid sequence of 78 and 87 kDa subunits are identical as APWGADVPR sequence and located at residue 19 onwards of the Vg deduced amino acid sequence just after signal peptide processing. The N-terminal amino acid sequence of the 104 kDa subunit of Vt is SIDSSVIADF which corresponds to amino acid position 728 onwards, and is just after a consensus cleavage site from endoproteases of the subtilisin family.

In contrast to the known Vgs of insects and vertebrates, the *P. merguiensis* Vg possessed neither a polyserine domain nor a potential N-linked glycosylation site and this is also the case for other shrimps. The DictyOGlyc 1.1 server, producing neural network predictions for GlcNAc O-glycosylation sites, identified 3 possible glycosylation sites in the Vg primary structure at positions Ser37, Thr1750 and Ser1891 (Gupta *et al*., 1999). Possible phosphorylation and sulfated tyrosine sites were predicted by NetPhos 2.0 (Blom *et al*., 1999) and Sulfinator (Monigatti *et al*., 2002), respectively. The deduced Vg sequence has 75, 43 and 28 possible phosphorylation sites for serine, threonine and tyrosine, respectively, while five sites of sulfated tyrosines were predicted. Secondary structure was predicted by Porter software (Pollastri and McLysaght, 2005). Primary sequence, secondary structure and post-translation sites are shown in Fig. 32. Fig. 33 shows topology of deduced Vg of *P. merguiensis*.

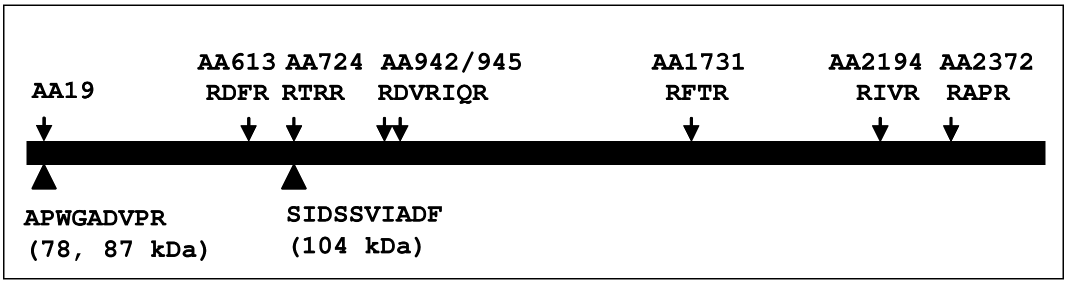


Fig.31 Schematic view of *Peneaus merguiensis* Vg cleavage sites and N-terminal amino acid sequence positions of Vt subunits.

The first arrow shows signal peptide cleavage site. The downward vertical arrows indicate the location of the predicted cleavage sites that have the RXXR consensus motif for subtilisin endoprotease. The upward vertical arrows represent the N-terminal amino acid sequence of Vt subunits.

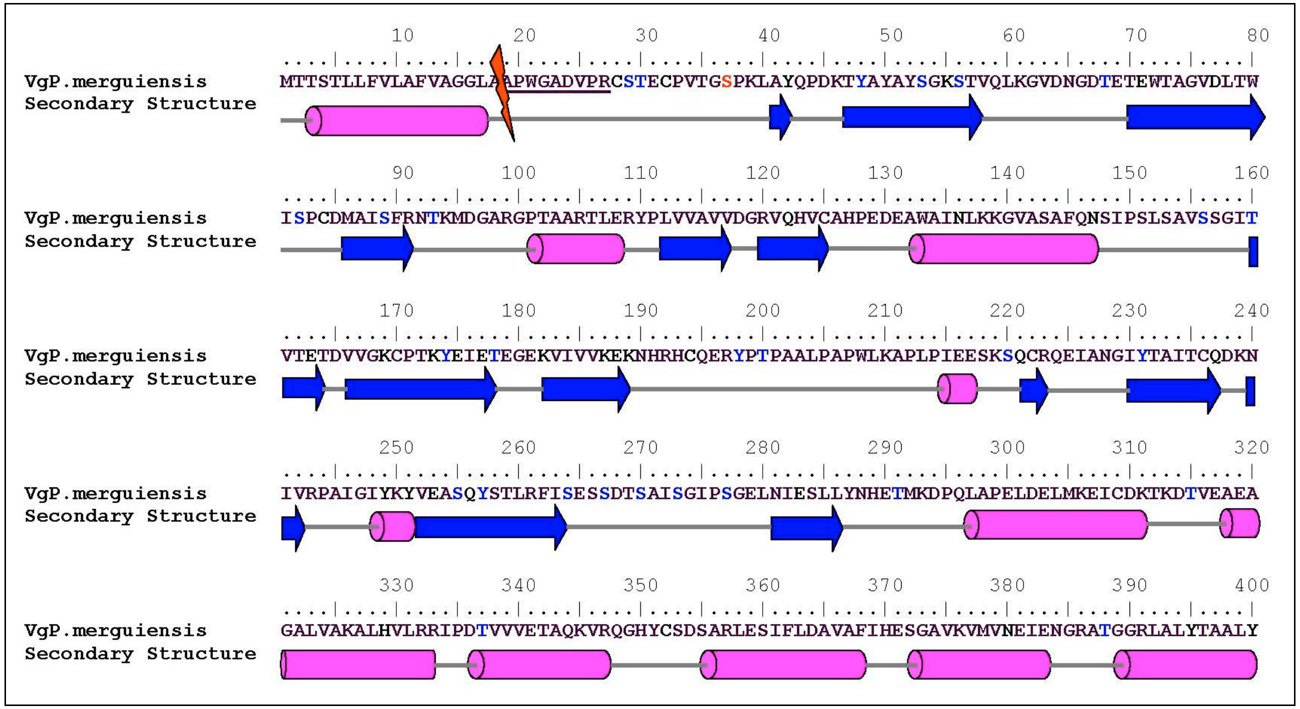


Fig. 32 Deduced amino acid sequence of the *Penaeus merguiensis* Vg with secondary structure and possible post-translation sites.

The N-terminal amino acids determined in this study are underlined. Amino acid sequences for a possible cleavage site with a consensus R-X-X-R are in white letters on a black background. Possible O-Linked GlcNAc glycosylation sites are in red letters. Possible phosphorylation sites are in blue letters and possible sulfated tyrosine sites have yellow background. For secondary structure, arrows (blue) indicate β-strands, and cylinders (pink) depict α-helices. The position of the observed signal peptide cleavage site and observed endoprotease cleavage sites are indicated by spark shape (red). This sequence was deposited in GenBank with protein ID number AAR88442.

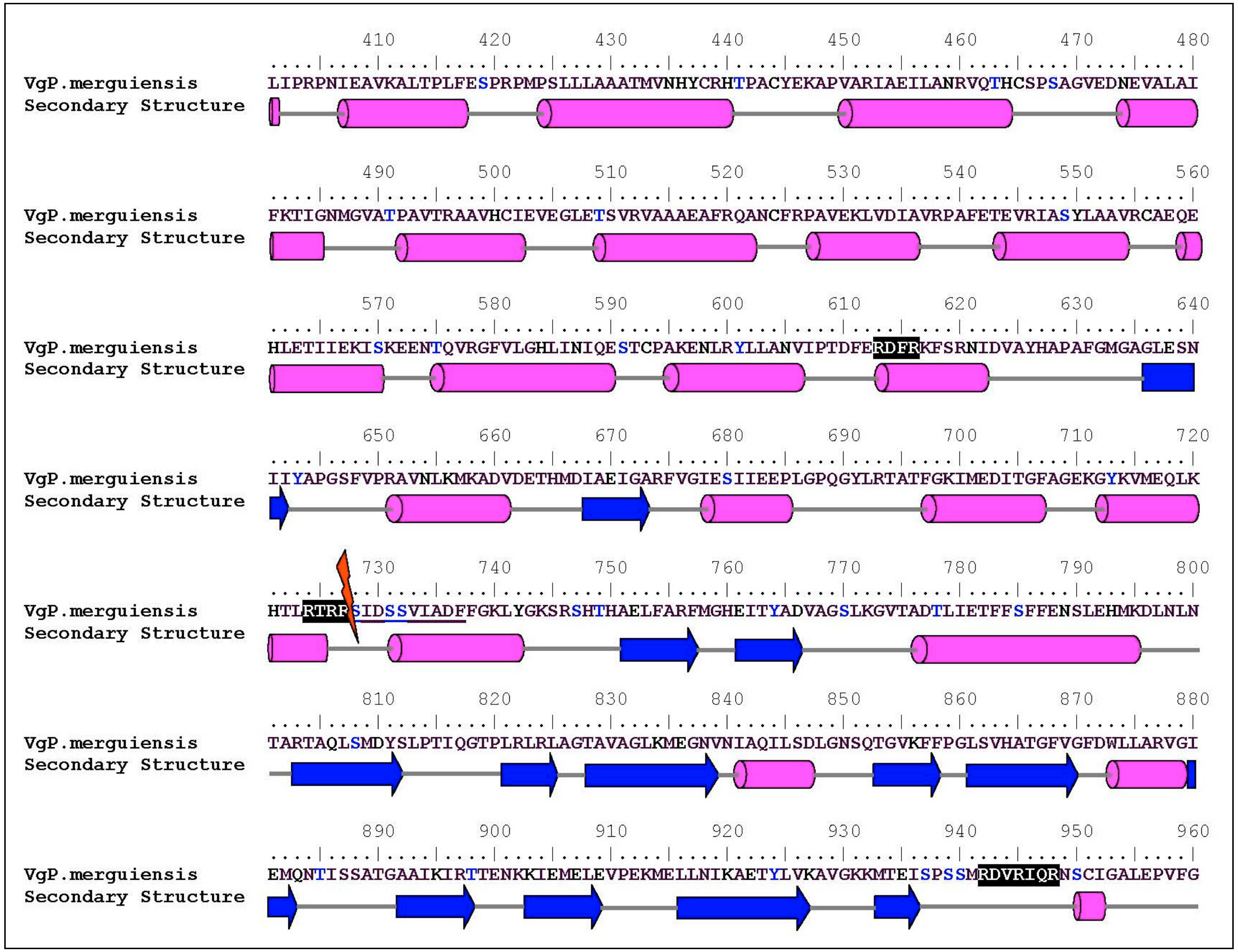


Fig. 32 (Continued)

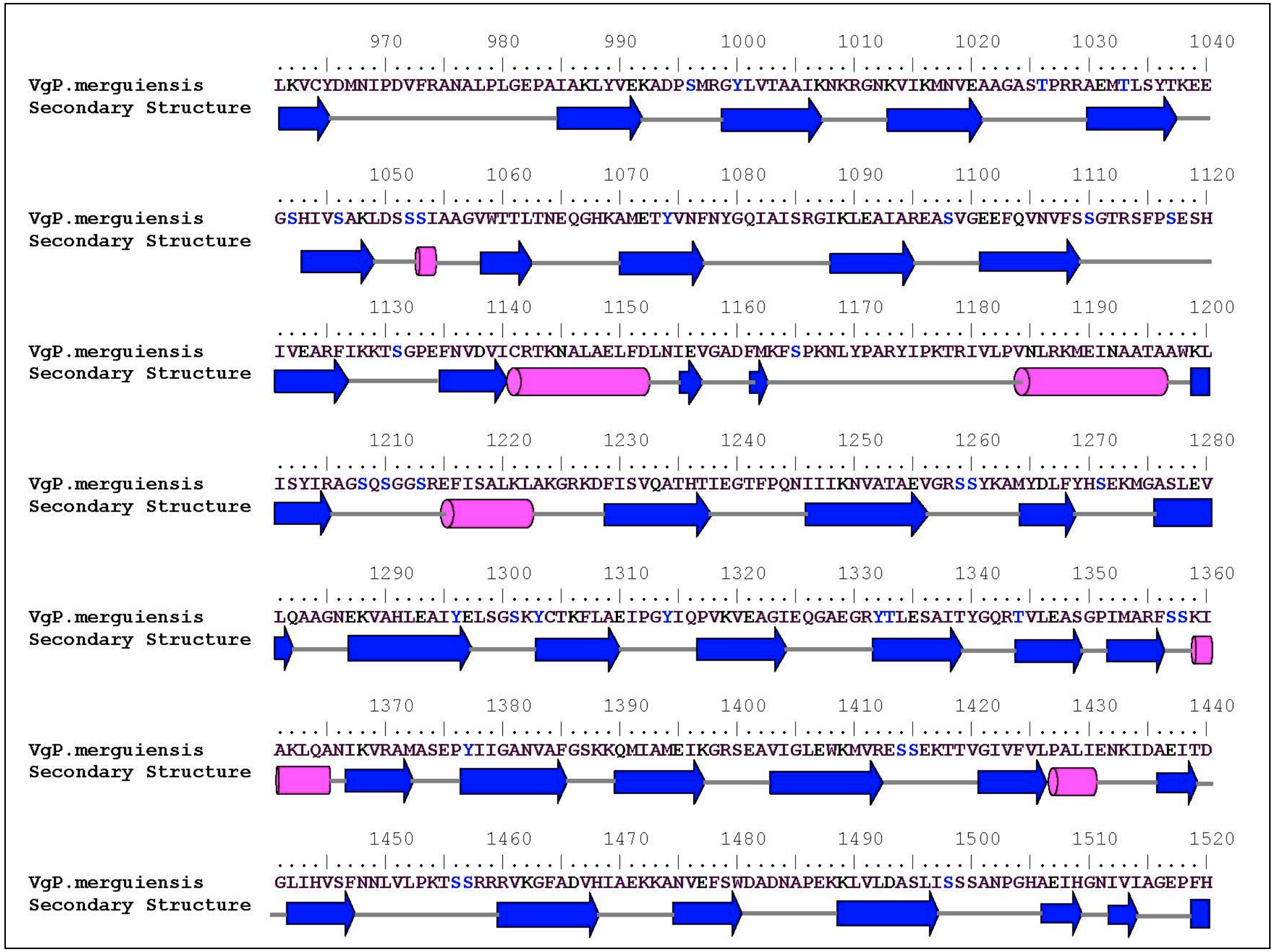


Fig. 32 (Continued)

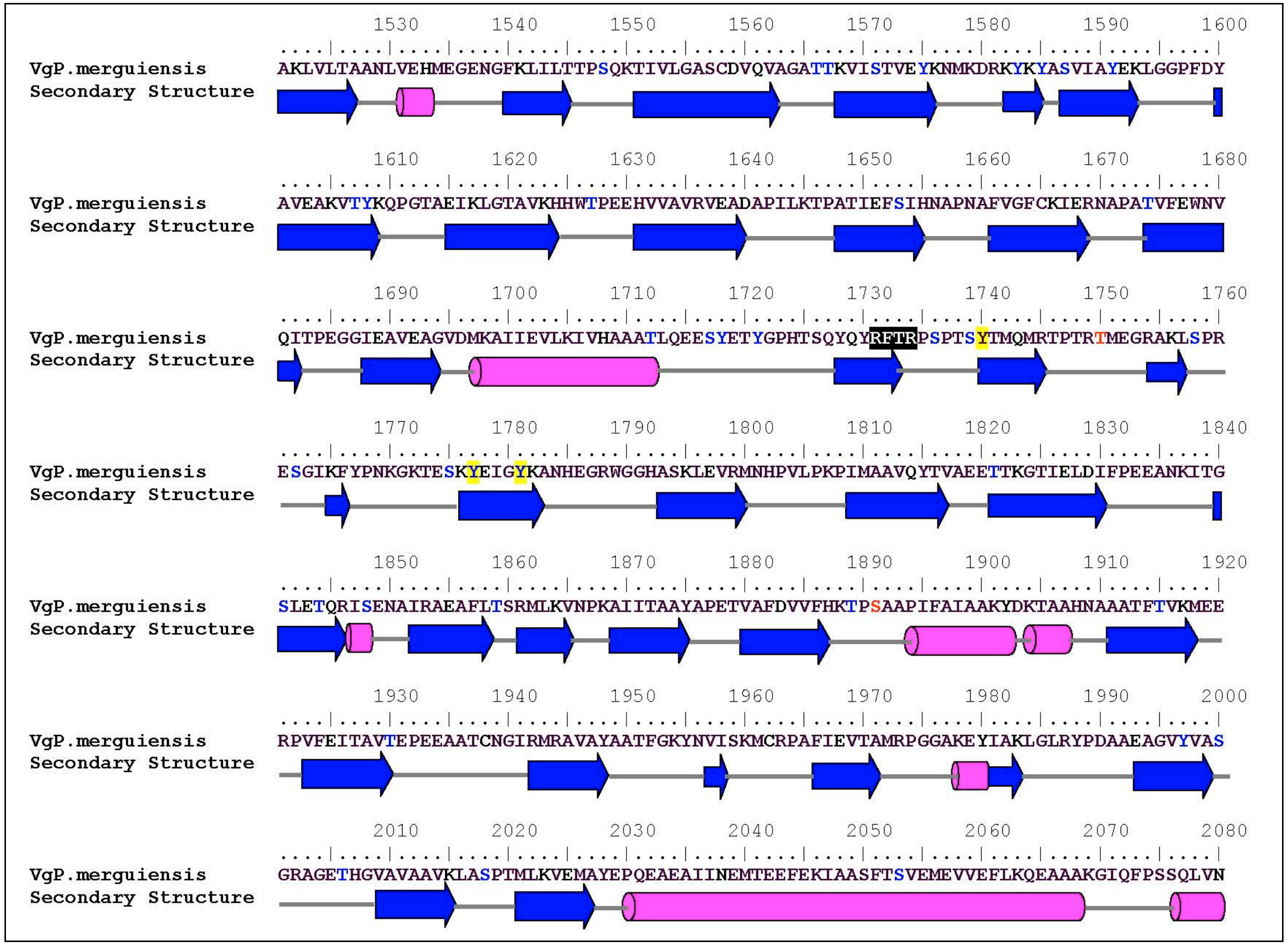


Fig. 32 (Continued)

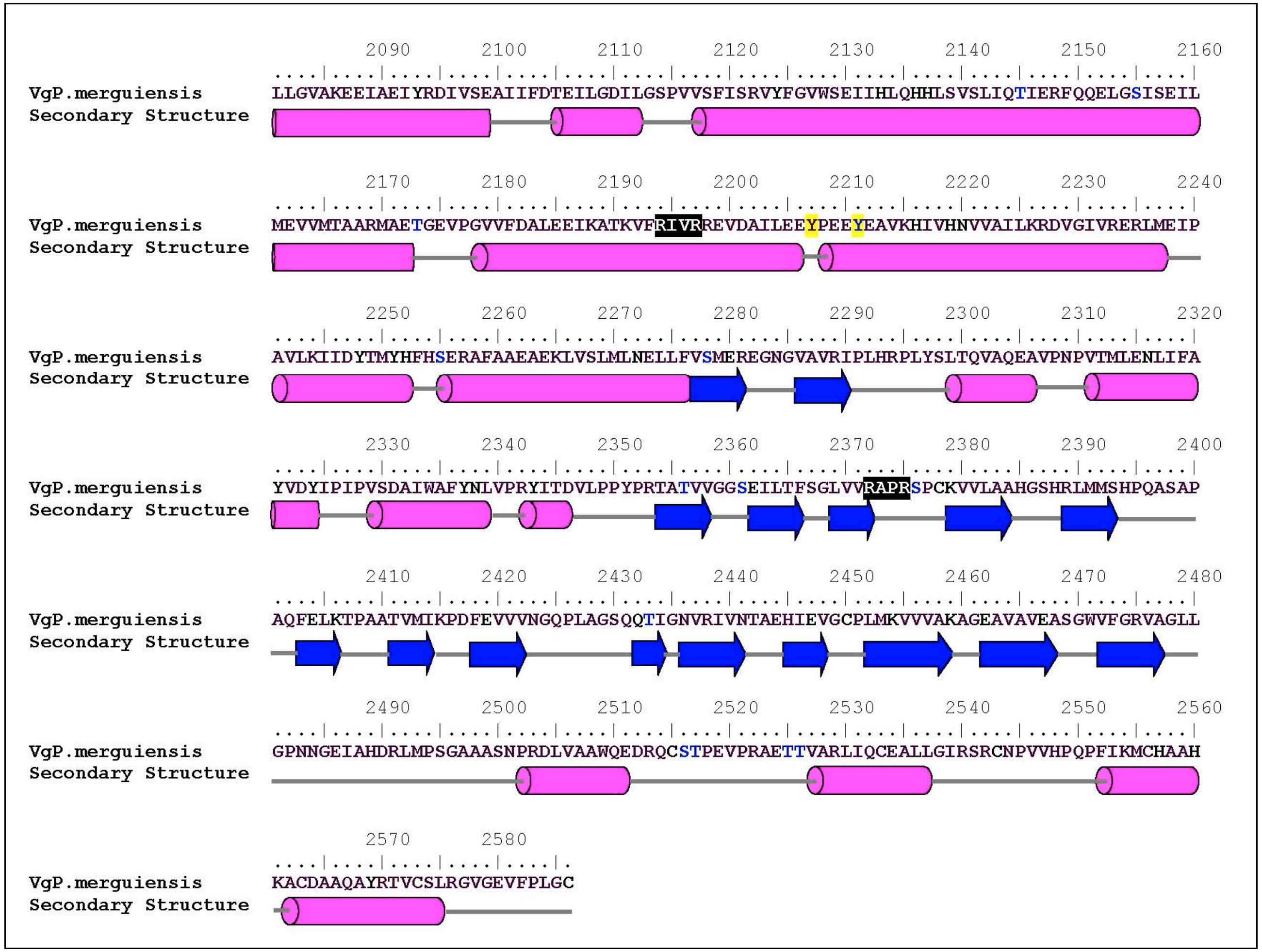


Fig. 32 (Continued)

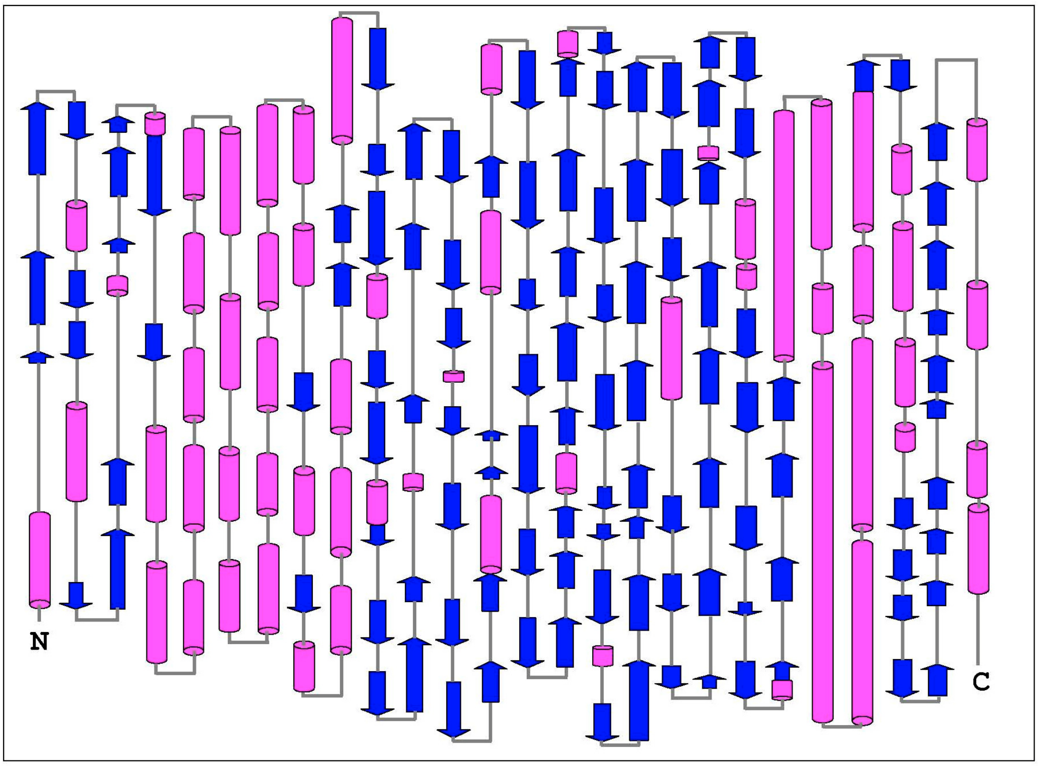


Fig. 33 Topology of deduced Vg in *Penaeus merguiensis* shows the secondary structure from N-terminal to C-terminal end.

Arrows (blue) indicate β-strands and direction from N- to C-terminal end. Cylinders (pink) depict α-helices.

3. Bioinformatics and computer analysis

3.1 BLAST analysis of deduced Vg from *P. merguiensis*

PSI and PHI-BLAST (position-specific iterated and pattern-hit initiated BLAST) (Altschul *et al*., 1997) network server from NCBI was used to search for similar sequences and conserved domains that have homology with the primary structure of the deduced amino acid sequence of Vg from *P. merguiensis* (Protein ID; AAR88442). Sequence analysis using BLAST revealed three putative conserved domains in the Vg sequence of *P. merguiensis*. The amino acid sequence could be divided into three segments; positions 1-1090 contain the lipoprotein N-terminal domain and DUF1081; positions 1091-2328 do not contain any conserved domains; and the C-terminal segment from residue 2329 to the end contains a von Willebrand factor (vWF) type D domain as shown in Fig. 34A. The N-terminal region of the Vg in this shrimp thus has similarities with other serum lipid binding proteins such as mammalian apolipoprotein B (ApoB), apolipophorin (Apo), retinoid-fatty acid binding glycoprotein (Retin) in insects, microsomal triglyceride transfer protein (MTP) and also Vg in other species. Sequences that showed similarity to banana shrimp Vg were separated into 3 analyses depending on similarity and the extent of the region of similarity to deduced Vg from banana shrimp as shown in Fig.34B. The entire length of deduced Vg sequence of *P. merguiensis* and all known other decapod crustacean full-length sequences which were available at NCBI on 1 May 2006 were used to constructed one phylogenetic tree. The N-terminal amino acid region and C-terminal amino acid region were analyzed separately. The middle part of Vg from banana shrimp does not show conservation to other proteins outside decapod crustaceans, and may have diverged rapidly.

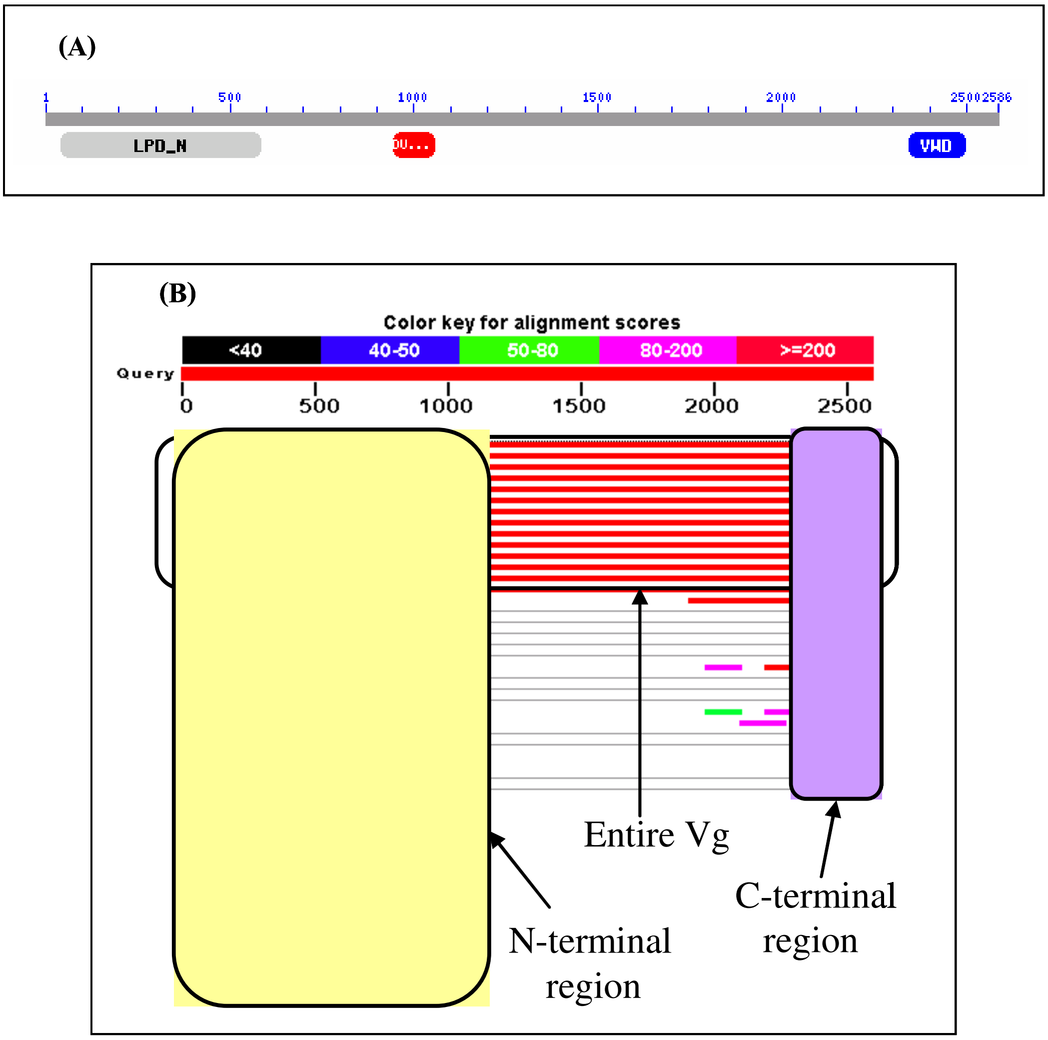


Fig. 34 BLAST analysis of deduced Vg.

The BLAST results showing (A) the three putative conserved domains of Vg sequences and (B) sequences with some similarity at the N- or C-terminal region. The arrows and blocks indicate the selected regions which were used in further analysis that each phylogenetic tree presented is based on.

3.2 Comparison of the primary structure of *P. merguiensis* Vg to Vgs of other decapod crustaceans

The deduced amino acid sequence including the signal peptide of *P. merguiensis* Vg and other 10 identified decapod crustaceans were multiple-aligned using the Clustal W computer program (Thompson *et al*., 1994). The alignment of the deduced amino acid sequences of these Vgs revealed a similarity among these 11 species (13 sequences) along the full sequence, including the signal peptide and N-terminal region (Fig. 35A) as well as for the C-terminal region (Fig. 35C). The structure of the consensus cleavage site for the subtilisin family of endoproteases, R-X-(K/R)-R, was observed in all Vgs, and their amino acid positions were nearly identical (Fig. 35B). Table 15 shows the overall amino acid identity ratios (%) in pairwise comparisons as calculated by BioEdit version 7.0.1 (Hall, 1999). The primary sequence had the highest similarity (91.4%) to the corresponding Vg sequence from *P. semisulcatus* (Genbank AY051318) and 86.9% similarity to that of *L. vannamei* (Genbank AY321153). It also has high similarity (above 63.5%) to the Vg from *P. monodon* (Genbank DQ288843), *M. japonicus* (Genbank AB033719) and Vg1 and Vg2 from *M. ensis* (Genbank AY103478 and AF548364, respectively). There was a weaker similarity to Vg3 from *M. ensis* (GenBank AY530205) (50.3%). The percentages of identical residues for nonpenaeid Vg are low, ranging 42.6%-34.9% for other decapod crustaceans. These include *C. quadricarinatus* (Genbank AF306784), *P. hypsinotus* (Genbank AB117524), *M. rosenbergii* (Genbank AB056458) and *P. trituberculatus* (GenBank DQ000638), while still retaining sequence similarity (34.6%) along the entire length of Vg in *C. feriatus* (GenBank AY724676). A phylogenetic tree of decapod crustacean Vgs was generated using the neighbor-joining method (Fig. 36). This unrooted tree shows that the currently known decapod Vgs group into four distinct lineages, which agree with taxonomic classification. The results from the phylogenetic tree and BLAST also show that Vg from *P. merguiensis* is most closely related to the Vg from *P. semisulcatus*, followed by the Vg of *L. vannamei*, *P. monodon*, *M. japonicus* and *M. ensis*, respectively in Penaeidea. The results from the phylogenetic tree and % identity ratio correspond well.

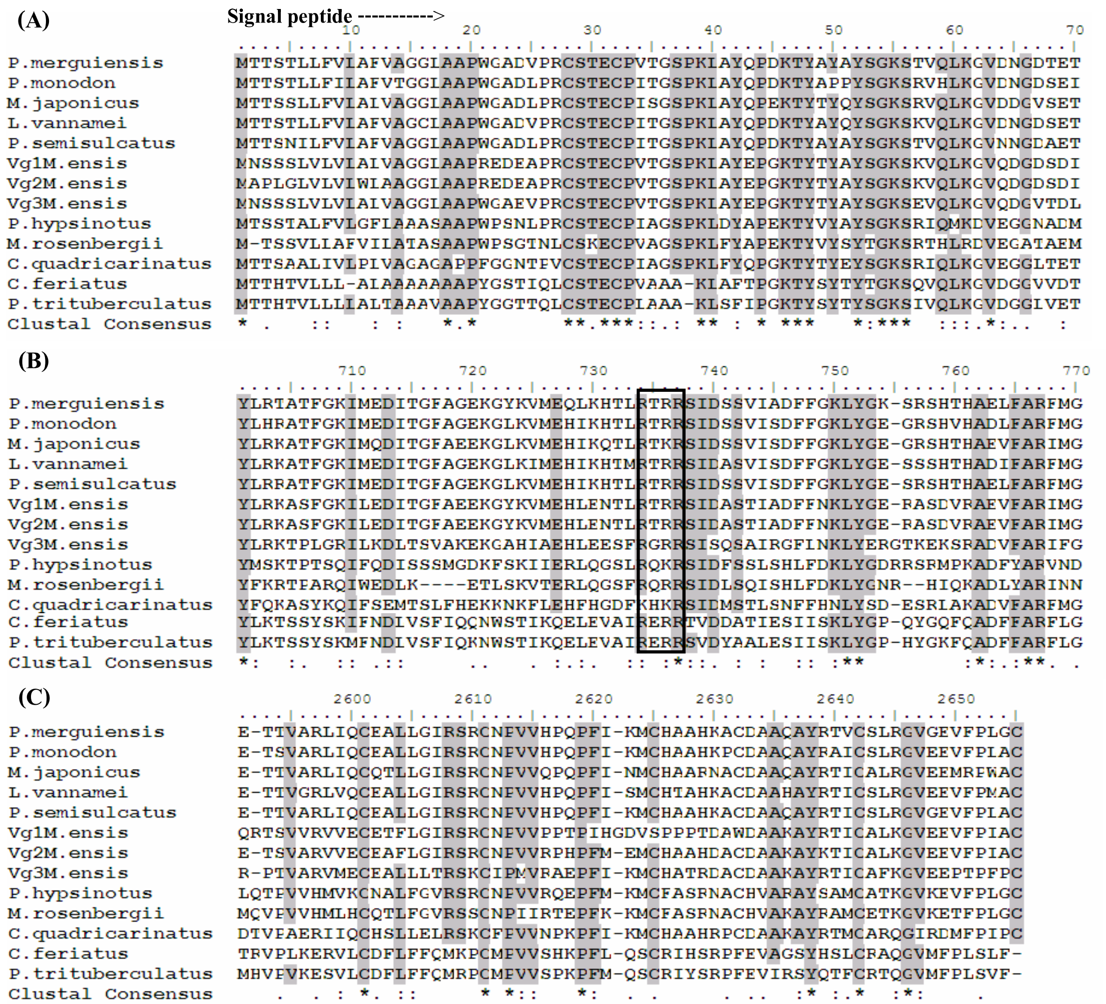


Fig. 35 Partial alignment of the deduced amino acid sequences of *Penaeus merguiensis* Vg with other decapod crustaceans.

(A) Sequences for the signal peptide and N-terminal region of Vg. (B) Consensus cleavage sequences (R-X-K/R-R) for processing by subtilisin-like endoproteases are shown in the black bordered box. (C) C-terminal region of Vg. Identical amino acid residues are highlighted in gray. The numbers above the amino acid sequences indicate the position in the alignment of the amino acid residue. The (-) indicates a gap introduced into the amino acid sequence to allow for the maximum degree of identity in the alignment. Identical, conserved and semi-conserved amino acid residue positions are indicated by asterisks, double dots and single dots, respectively.

Table 9 The percentage ratios of the overall amino acid identity between the full-length Vg amino acid sequences including signal peptide sequences of 11 decapod crustacean species (13 sequences).

Species	*Penaeus merguien*	*Penaeus monodon*	*Marsup. japonicus*	*Litop. vannamei*	*Penaeus semisulcatus*	*M. ensis* Vg1	*M.ensis* Vg2	*M.ensis* Vg3	*Pandalus hypsinotus*	*Macrobra. rosenbergii*	*Cherax quadricarina*	*Charybdis feriatus*	*Portunus tritubercula*
P. merguiensis		0.841	0.791	0.869	0.914	0.635	0.645	0.503	0.400	0.365	0.426	0.346	0.349
P. monodon			0.764	0.812	0.837	0.626	0.636	0.497	0.396	0.362	0.419	0.337	0.342
M. japonicus				0.789	0.799	0.646	0.651	0.515	0.412	0.374	0.439	0.342	0.346
L. vannamei					0.869	0.628	0.639	0.507	0.399	0.362	0.431	0.342	0.346
P. semisulcatus						0.632	0.643	0.498	0.398	0.362	0.427	0.344	0.346
Vg1M. ensis							0.966	0.530	0.387	0.353	0.407	0.329	0.337
Vg2M. ensis								0.538	0.393	0.359	0.413	0.332	0.340
Vg3 M. ensis									0.371	0.350	0.401	0.318	0.327
P. hypsinotus										0.605	0.409	0.347	0.354
M. rosenbergii											0.387	0.316	0.325
C. quadricarinatus												0.381	0.378
C. feriatus													0.782
P. trituberculatus													

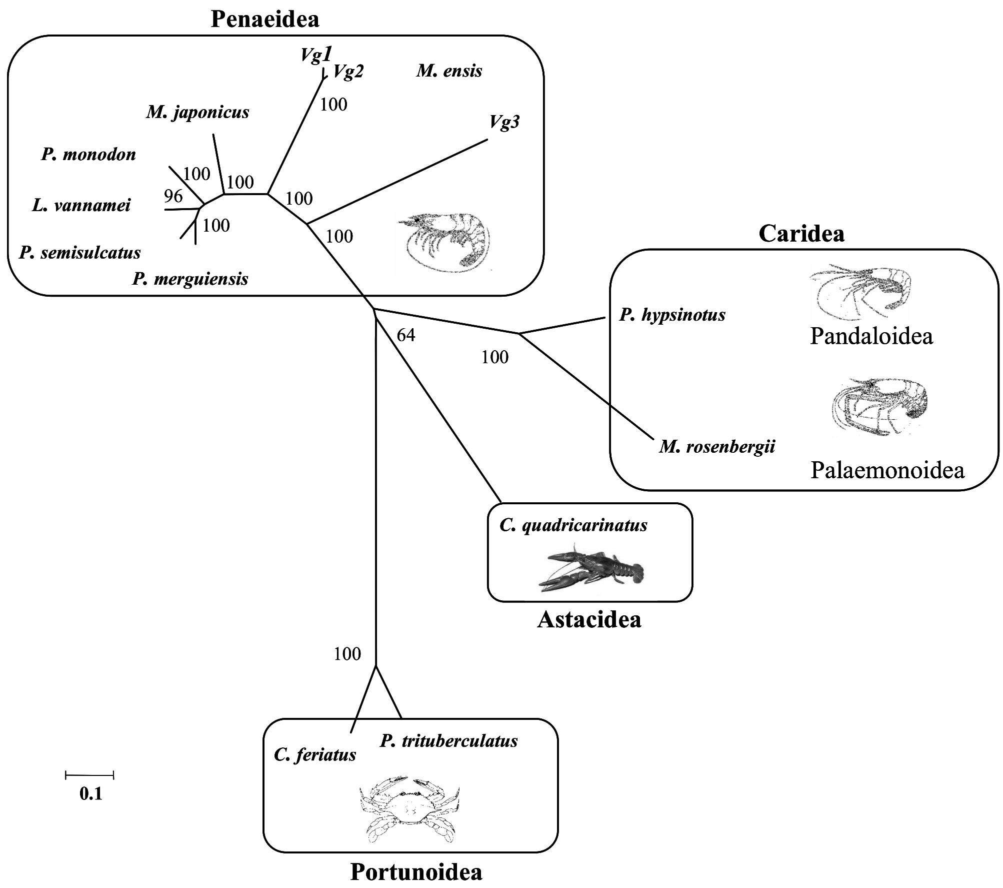


Fig. 36 Phylogenetic tree analysis of Vg from different crustacean.

The unrooted tree was constructed with MEGA (Kumar *et al*., 2004) using the neighbor-joining method (Saito and Nei, 1987) based on the alignment of amino acid sequences of ten crustacean Vgs using Clustal W (Thomson *et al*., 1994). The distance method used was the JTT Matrix (Jones *et al*., 1992); otherwise default parameters were used. Bootstrap values (%) based on 1000 replicate analyses are shown at branch points which indicated how reliable of that branch tree. The standard bar at the bottom shows the branch length corresponding to 0.1 amino acid changes per residue. The illustrations in the tree were obtain from De-Bruin (1995).

3.3 Phylogenetic tree of N-terminal regions of vitellogenins

Phylogenetic trees analysis based on sequences obtained from BLAST gave low bootstrap value, since the amino acid sequence is quite non-conserved even though they are from the same protein family. Instead I searched the Pfam (Protein Families) database using *P. merguiensis* Vg as query. The results showed significance at the N-terminal region. This lipoprotein amino terminal region family contains Vg from various species, apolipophorins (Apo), apolipoprotein B (ApoB), retinoid- and fatty acid-binding glycoprotein (Retin), a clottable protein (Clot) and the microsomal triglyceride transfer protein (MTP). These proteins are all involved in lipid transport except the clottable protein, which functions in coagulation systems. A phylogenetic tree based on the alignment of amino acid sequences from the lipoprotein N-terminal region family using Clustal W (Thomson *et al*., 1994) was reconstructed by MEGA version 3.1 using the neighbor-joining method. The distance method used was JTT Matrix (Jones *et al*., 1992); default parameters were used otherwise. The phylogenetic tree of the N-terminal region is shown in the Fig. 37A. MTP was chosen as the out group. Phylogenetic analysis of the N-terminal region showed high bootstrap values for the branch of Vgs from vertebrates, separated from Vgs and clottable proteins from invertebrates. In addition, this tree indicated that Vgs from insects are grouped with clottable proteins from crustaceans.

Surprisingly, the phylogenetic tree showed that Vgs from decapod crustaceans are located in the same cluster as Apo, ApoB and Retin; not in the same group as Vg from other species (vertebrates and invertebrates). This means that Vg in decapod crustaceans are paralogous to Vg of other species. In other words, Vg gene duplication occur before species speciation/separation in decapod crustacean. It is assumed in the literature that all Vgs are orthologous: that a single gene duplication of the ancestral lipoprotein gene give rise to Vg, and that this Vg gene has been inherited with the same function in all vitellogenic species. Instead, it appears decapod crustacean Vg is paralogous to other Vg, being descended from a different gene family member and having its function change to Vg after decapod crustaceans separated from other species. The event can be confined to decapod crustaceans, since *Daphnia magnia*, a non-decapod crustacean, has Vg that groups with insects and other species, not with decapod crustacean Vg. Because the branches that separated Vgs from vertebrate, invertebrate, insect, nematode into one group and Vg from decapod crustacean, Apo, ApoB and Retin into another group did not have significant bootstrap support, 55 and 50, respectively, we wanted to confirm this surprising result. Dr. Robert G. Beiko (Genomics and Computational Biology, Institute for Molecular Bioscience, University of Queensland and ARC Centre in Bioinformatics) helped me to establish confidence in this scenario by constructing phylogenetic tree using the MrBayes (Huelsenbeck and Ronquist, 2001; Huelsenbeck *et al*., 2001; Holder and Lewis, 2003) program based on a MUSCLE (Edgar, 2004) alignment of the lipoprotein N-terminal region family. MrBayes was run for 2.5 million generations, and the first 500,000 generations were discarded as burn-in (Fig. 37B). The phylogenetic tree reconstructed by MrBayes showed significant values, 92 and 99 at the branch that separated Vgs of decapod crustacean from Vgs of other species. We thus have more confidence that Vg in decapod is indeed paralogous with Vgs of other species.

(A)

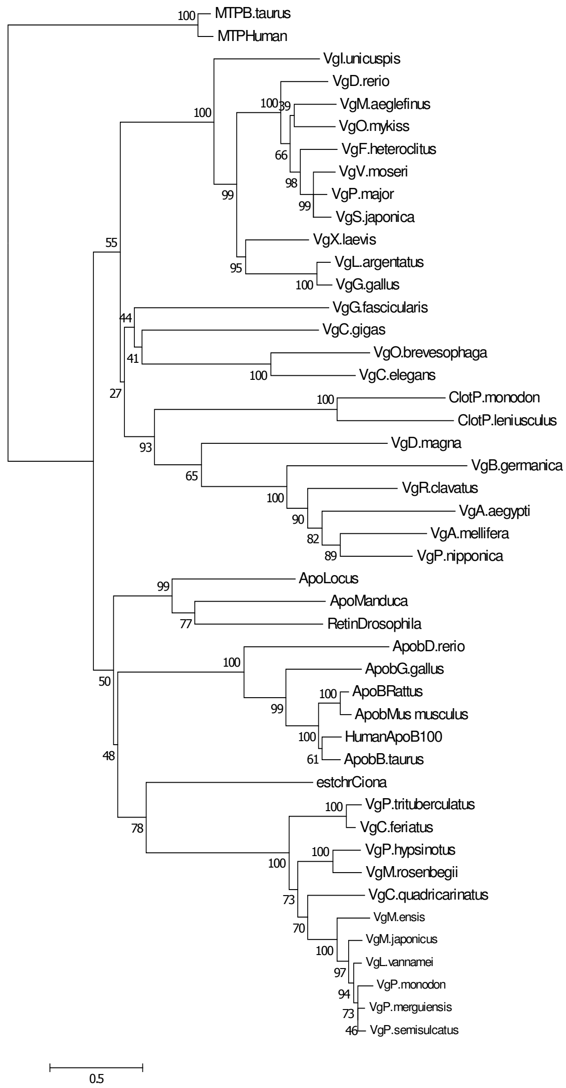


Fig. 37A Phylogenetic tree analysis of lipoprotein N-terminal region family.

(A) The tree was constructed with MEGA (Kumar *et al.*, 2004) using the neighbor-joining method (Saito and Nei, 1987) based on the alignment of amino acid sequences of lipoprotein N-terminal region family using Clustal W (Thomson *et al.*, 1994). The distance method used was the JTT Matrix (Jones *et al.*, 1992); otherwise default parameters were used. Percentage bootstrap values based on 1000 replicate analyses are shown at branch points which indicated how reliable of that branch tree. The standard bar at the bottom shows the branch length corresponding to 0.5 changes per residue.

(B)

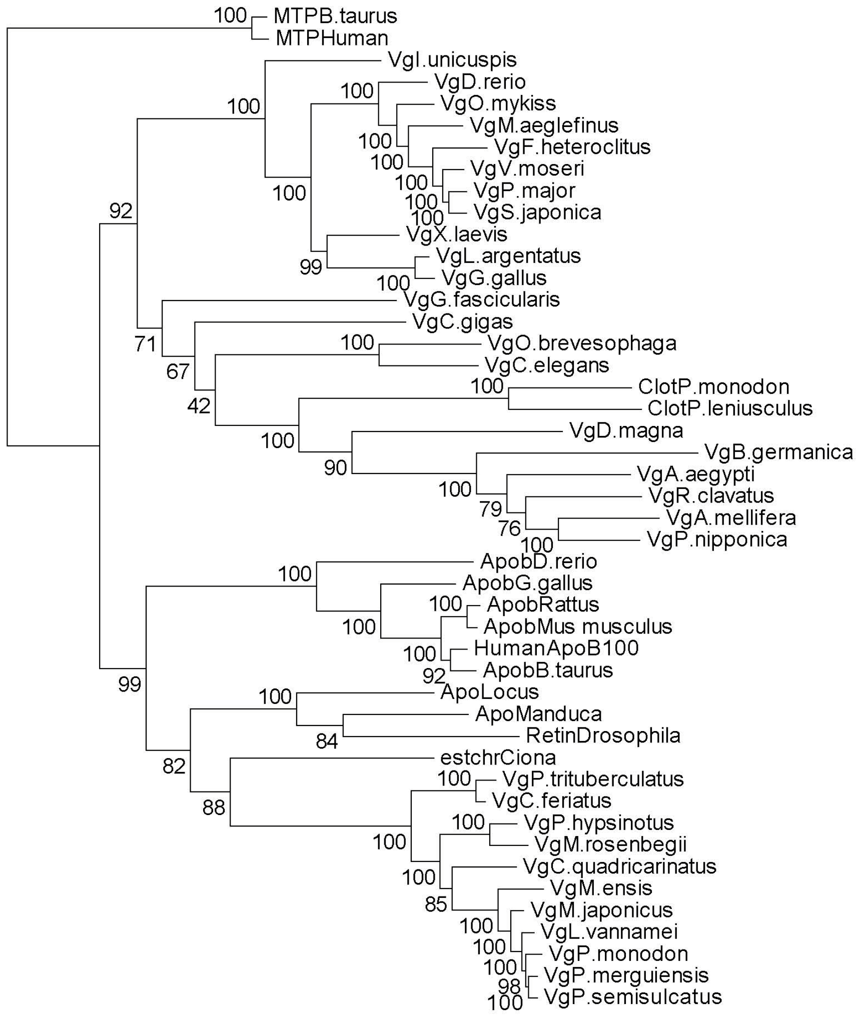


Fig. 37B Phylogenetic tree analysis of lipoprotein N-terminal region family.

(B) The tree was constructed with MrBayes (Huelsenbeck and Ronquist, 2001; Huelsenbeck *et al.*, 2001; Holder and Lewis, 2003) by Dr. Robert G. Beiko (Genomics and Computational Biology, Institute for Molecular Bioscience, University of Queensland) based on the alignment of amino acid sequences of lipoprotein N-terminal region family using MUSCLE (Edgar, 2004). MrBayes consensus values (percentage) are shown at branch points.

3.4 Phylogenetic tree of C-terminal regions of vitellogenin

The von Willebrand factor (vWF) type D domain at the C-terminal region was analysed by searching the Pfam (Protein Families) database using *P. merguiensis* Vg as query. The vWF type D domain is found in Vg of other species and in Apo, Retin, Clot, vWF and mucin but not in ApoB. The phylogenetic tree based on the alignment of amino acid sequences of vWF type D domain using Clustal W (Thomson *et al*., 1994) was reconstructed by MEGA version 3.1 using the neighbor-joining method and the tree is shown in Fig. 38. The distance method used was JTT Matrix (Jones *et al*., 1992); default parameters were used. Clusters consisting of Apo, Retin, vWF and mucin grouped together, and were clearly separated from Vgs group as expected. The Clot of *P. monodon* was located in the same branch as Vgs from insect and this group has a closer relationship with the cluster of Vgs from invertebrates and nematodes than the group of Vgs from decapod crustacean. When comparing only Vg sequences, the vWF type D domain at the C-terminal region were more conserved than the lipoprotein N-terminal domain. The vWF domain may be more highly conserved due to its function in protein-protein interaction which limits allowable amino acid variation.

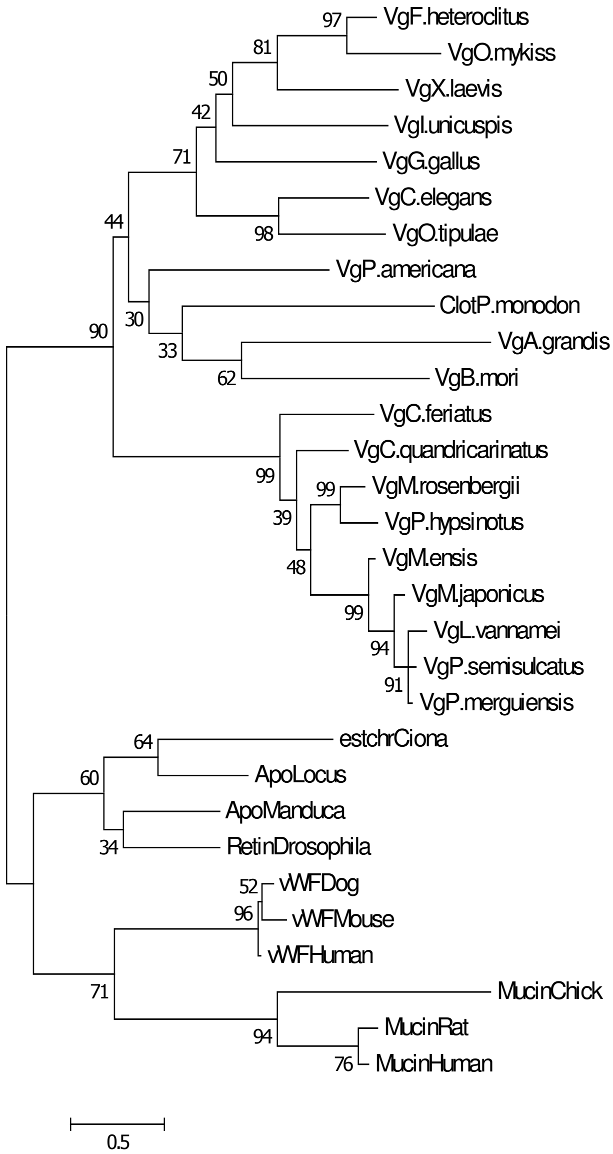


Fig. 38 Phylogenetic tree analysis of the von Willebrand factor (vWD) type D domain at the C-terminal region of Vg and other corresponding region in other related proteins.

The tree was constructed with MEGA (Kumar *et al.*, 2004) using the neighbor-joining method (Saito and Nei, 1987) based on the alignment of amino acid sequences of vWF type D domain using Clustal W (Thomson *et al.*, 1994). The distance method used was JTT Matrix (Jones *et al.*, 1992); otherwise default parameters were used. Percentage bootstrap values based on 1000 replicate analyses are shown at branch points which indicated how reliable of that branch tree. The standard bar at the bottom shows the branch length corresponding to 0.5 changes per residue.

3.5 Tertiary structure modelling of N-terminal region vitellogenin

The three-dimensional structure has been solved for only one member of the large lipoprotein family, lamprey lipovitellin (LV) (Anderson *et al*., 1998; Thompson and Banaszak, 2002). LV was purified from lamprey eggs, crystallized and X-ray diffraction used to solve the structure. X-ray crystallography data are not available from Protein Data Bank (PDB), only the finally determined three-dimensional (3-D) locations of atoms. The lamprey data, entry 1LSH were retrieved from Protein Data Bank (PDB). This consists of the deduced amino acid sequence, N-terminal sequence information and 3-D co-ordinates of each atom in the structure. LV from lamprey comprises A and B polypeptide chains and consists of a number of folding domains as shown in Fig. 39A. A tertiary structure model of N-terminal Vg from banana shrimp was built using the 3-D structure from lamprey LV chain A (PDB entry 1LSH_A) as template (Fig. 39B). 3D-Jigsaw (Bates *et al*., 2001), EasyPred (Lambert *et al*., 2002) and Modeller (Sali *et al*., 1995) servers returned comparable models, which is expected, since the 1LSH_A template shares approximately 19.4% identity with the N-terminal region of Vg sequence from *P. merguiensis*. For 3-D structure 20% identity is pretty good, we expect same 3-D structure. Each model was named according to the server used for modelling; 3D-Jigsaw_Vg, EasyPred_Vg and Modeller_Vg. Models are shown in different orientations in Fig. 40 and Fig. 41. Overall, all structure models shared similarity, with a main domain structure very similar to the LV template, except that the Vg model from EasyPred has a different orientation between the N-sheet and other domains. All models have predominantly antiparallel β-sheet domains including the N-sheet, C-sheet, A-sheet, and also a large helical domain. Our models are missing some parts of the A-sheet because the underlying sequence is located on chain B of lamprey LV which can not be aligned to the remaining Vg sequence from banana shrimp. Thus tertiary structure could be predicted only for the N-terminal region of Vg. Currently, von Willebrand factor (vWF) type D domain structure is not available from the PDB databank, and so there is no template for modeling this region. We conclude from the similarity between all models that A- and C-sheets form a lipid cavity and seem to be a boundary between the α-helix domain and the lipid cavity. The α-helix domain appeared as 2 layers of 18 helical structures that form an arch and then this arch forms a clasp around the β-sheet structure of A- and C-sheets in the middle of the models. The N-sheet domain forms a barrel-like conformation and includes 11 large β-strands wrapped around a helix of 13 residues.

The C-Sheet consists of mainly 7 β-strands and the A-sheet in Vg models comprise 8 large β-strands.

The relative prediction performance among the three programs was compared. The loops and some small secondary structure features are different between the different prediction models. Superimposition shown in Fig. 42 reveals that the model of Vg from the 3D-Jigsaw server gave the most similar structure prediction to the template LV. However, it could not be concluded that 3D-Jigsaw built the best model for the N-terminal region of Vg due to the low identity between the LV template and N-terminal region of Vg from *P. merguiensis*.

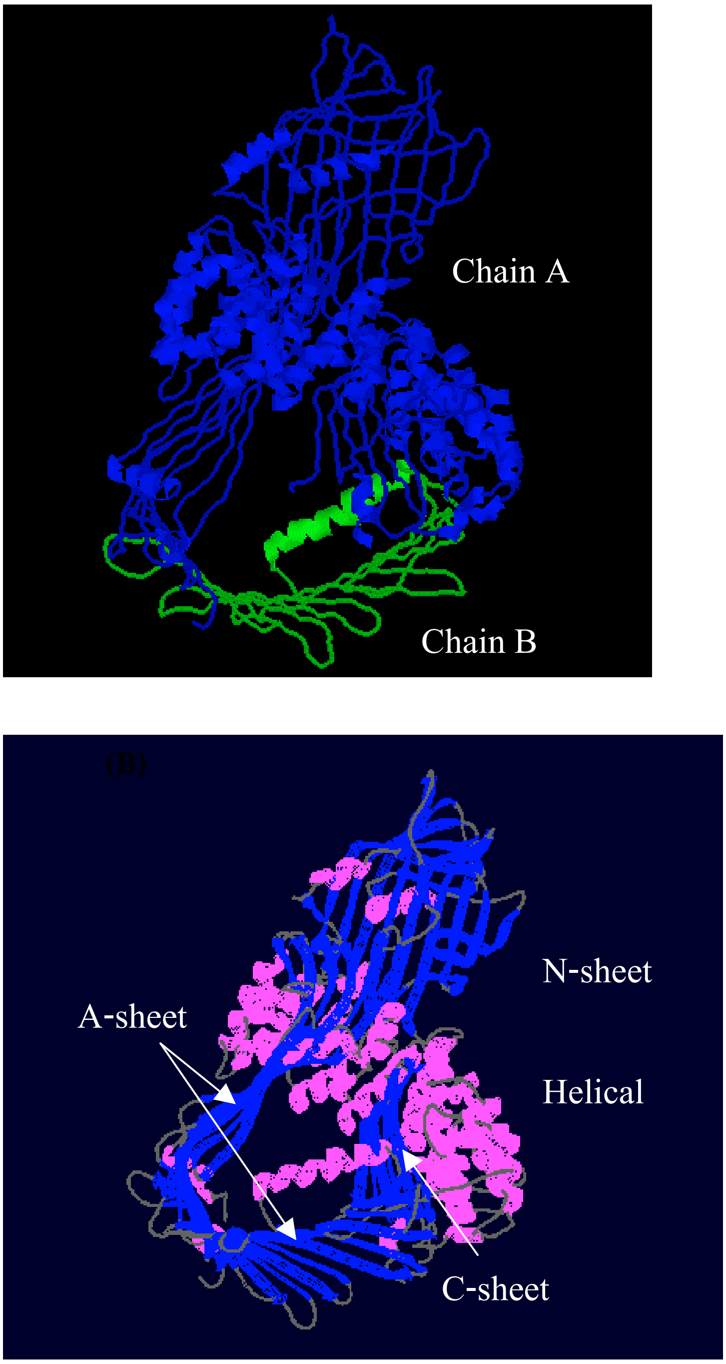


Fig. 39 Stereoview ribbon or cartoon diagrams of lamprey LV (Protein Data Bank entry 1LSH).

Chain A is colored blue whereas chain B is colored green (A). The main domains of LV are shown in (B), β-sheets are colored blue, α-helix structures are colored pink, loops and turns are colored gray.

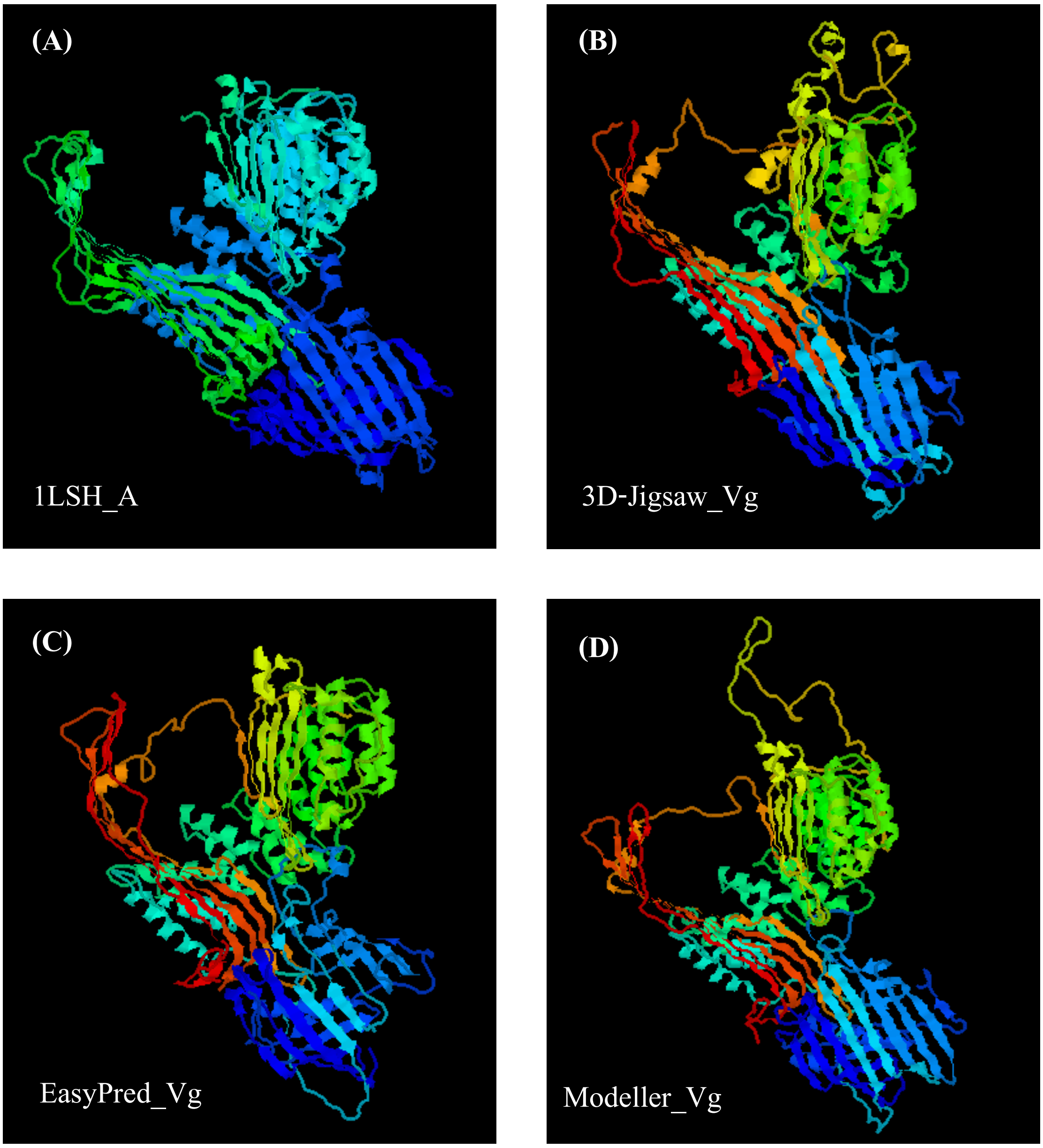


Fig. 40 Side views of three models of Vg and the 1LSH_A template.

The cartoon diagrams in the same orientation are shown as 1LSH_A (A), 3D-Jigsaw_Vg (B), EasyPred_Vg (C) and Modeller_Vg (D). The shading from blue to red shows the direction from N-terminal to C-terminal.

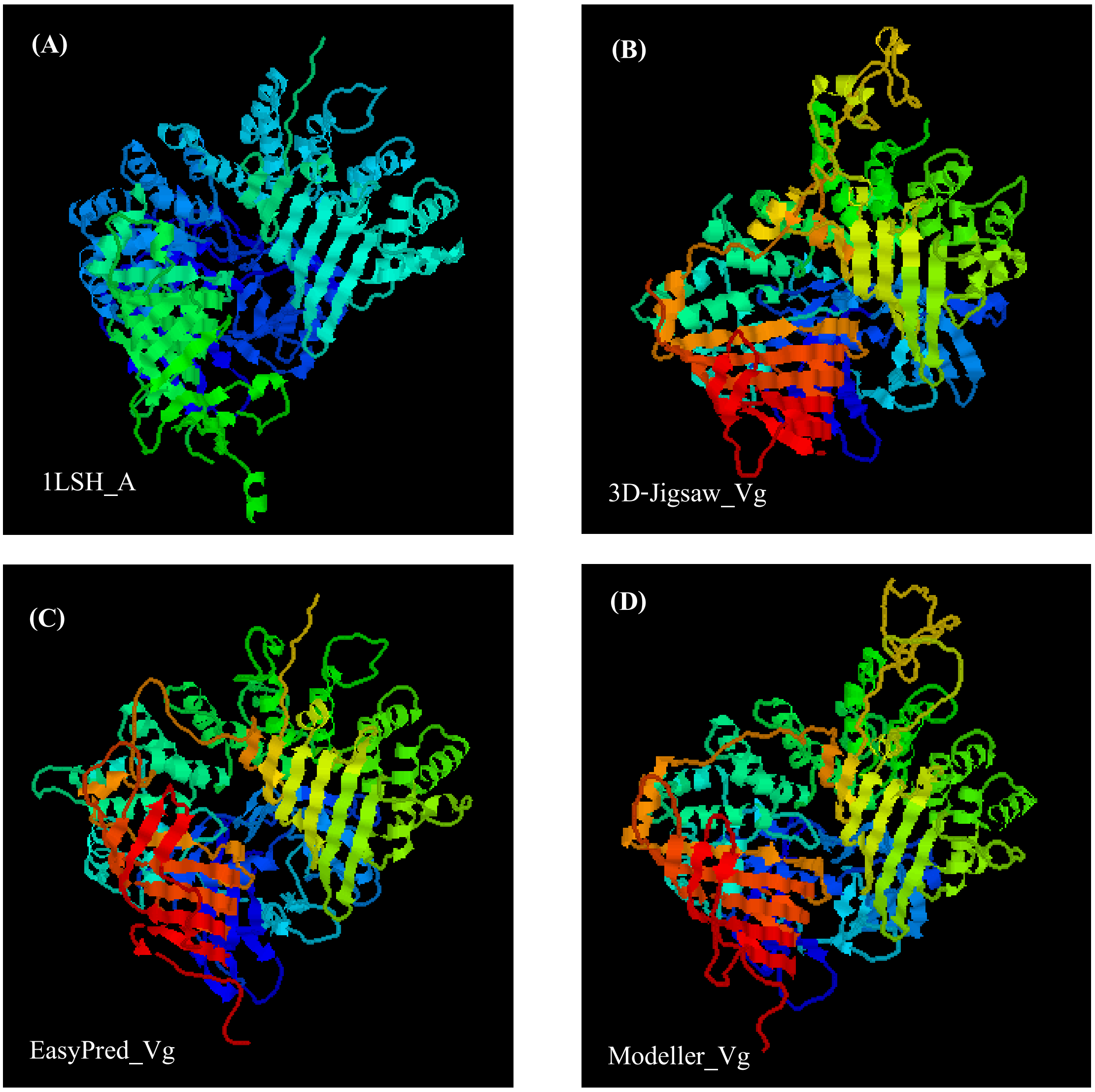


Fig. 41 Down views of three models of Vg and the 1LSH_A template.

The cartoon diagrams in the same orientation are shown as 1LSH_A (A), 3D-Jigsaw_Vg (B), EasyPred_Vg (C) and Modeller_Vg (D). The shading from blue to red shows the direction from N-terminal to C-terminal.

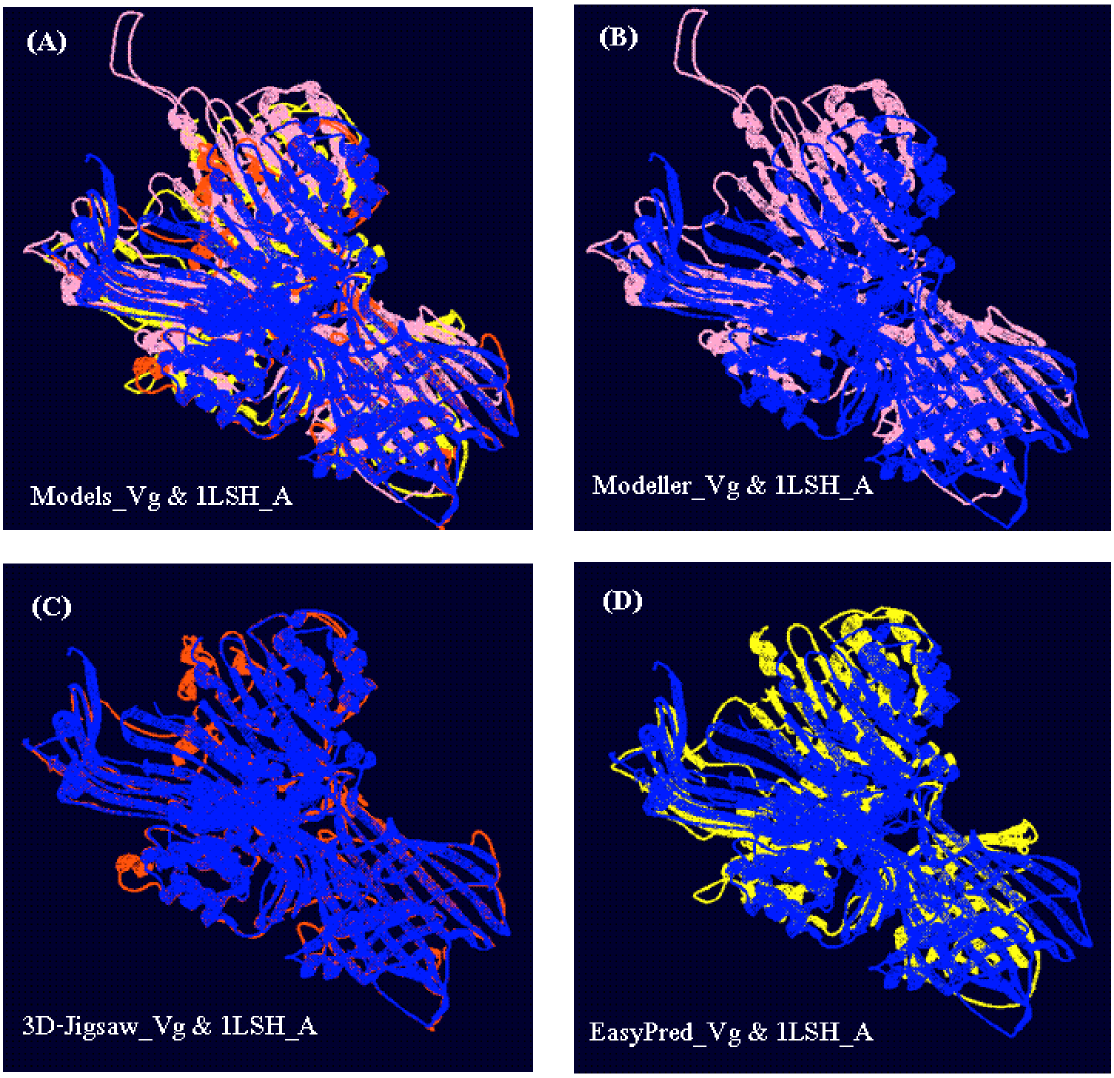


Fig. 42 Superimposition of all presented models.

(A) Superimposition of models from Modeller_Vg (pink), 3D-Jigsaw_Vg (red), EasyPred_Vg (yellow) and lamprey lipovitellin template (1LSH_A) (blue). (B) Superimposition of Modeller_Vg (pink) and 1LSH (blue). (C) Superimposition of 3D-Jigsaw_Vg (red) and 1LSH_A (blue). (D) Superimposition of EasyPred_Vg (yellow) and 1LSH_A (blue).

4. Expression of Vg gene in different tissues.

The site of Vg mRNA expression was determined, by RT-PCR analysis of multiple tissues. Total RNA samples were extracted from various tissues of vitellogenic females and hepatopancreas of mature males. Vg specific primers (FVg112 and RVg711) were used to amplify Vg cDNA fragments. The expected size of the specific Vg PCR product is 600 bp. The specific Vg PCR fragment was detected only in the ovary and hepatopancreas of vitellogenic females. Heart, muscle and intestine of vitellogenic females, and hepatopancreas of male shrimps did not have detectable levels of product (Fig. 43). 18s rRNA was used as an internal control for each tissue sample to confirm that the same amount of total RNAs was used. The expected size of the 18s rRNA PCR product is 298 bp. The specific band of 18s rRNA indicated that roughly the same amount of total RNAs were loaded in the RT-PCR reactions for each tissue sample.

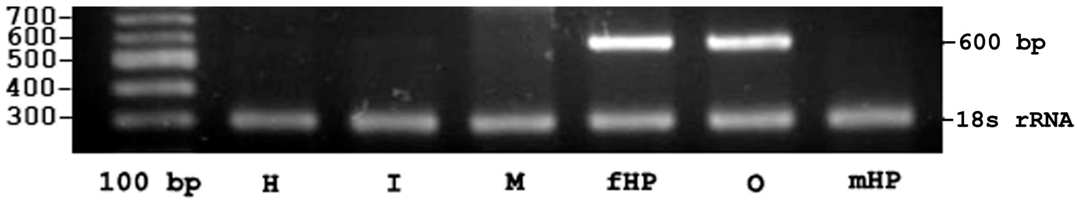


Fig. 43 RT-PCR analysis of the expression of Vg mRNA in different *Penaeus merguiensis* tissues.

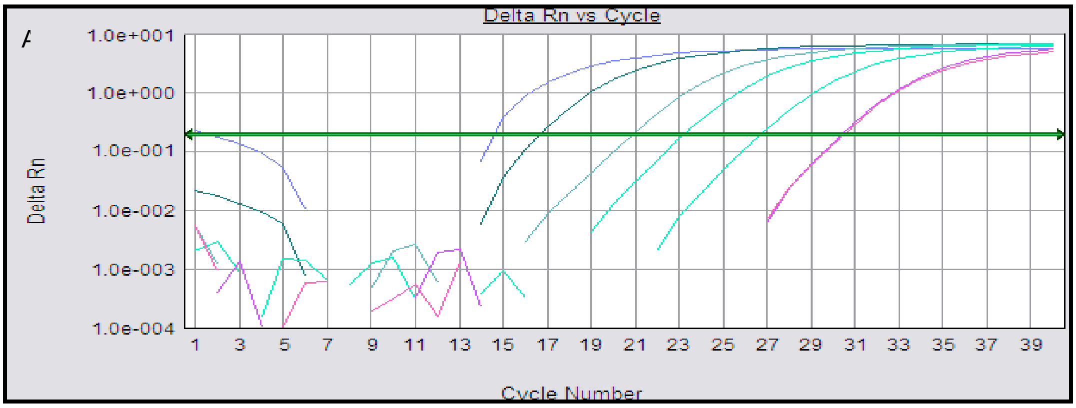
Total RNA was extracted from hepatopancreas (fHP), ovary (O), muscle (M), heart (H) and intestine (I) of vitellogenic females and the hepatopancreas (mHP) from a mature male. The RT-PCR of 18s rRNA transcript was an internal control.

5. Relative quantification of Vg mRNA expression during ovarian development by real-time PCR

Changes in Vg mRNA levels in the hepatopancreas and ovary at 4 stages of ovarian development were examined by real-time PCR. To ensure equal amounts of total RNA in samples, levels of 18s rRNA in the samples were also measured for standardization. Quantities of mRNA in samples were calculated from a standard curve derived from known amounts of the target gene (Vg mRNA or 18s rRNA). The standard curves for quantification of Vg gene and 18s rRNA were linear over six and five, respectively, orders of magnitude with the linear correlation (r), between threshold cycles (Ct) and the copy number of the target gene, being over 0.99 in each case. A representative amplification plot of standards and the corresponding standard curves are shown in Fig. 44 and Fig. 45.

Relative Vg mRNA levels are normalized by dividing copy number of Vg by copy number of 18s rRNA from the same sample. The changes in the relative levels of Vg mRNA expression in both tissues are shown in Fig. 46 and Table 10. Sixteen ovaries were divided equally into four groups based on GSI and developmental stages (1-4). GSI increased as vitellogenesis progressed. In contrast, the relative level ($x10^{-3}$) of Vg mRNA expression in the ovary increased from 0.69 ± 0.18 for shrimps in stage 1 to reach its maximum at 9.27 ± 0.62 in stage 2 and rapidly decreased in stage 3 (4.14 ± 0.31) and 4 (1.40 ± 0.22). In the hepatopancreas, the relative level ($x10^{-3}$) of Vg mRNA increased from 0.004 ± 0 to 0.57 ± 0.15 for shrimps in stage 1 and 2, respectively. The highest level was observed at stage 3 (1.08 ± 0.08) and thereafter declined in stage 4 (0.48 ± 0.06).

Relative values of Vg from stage 1 of ovarian development in each tissue were arbitrarily designated as the calibrator. In the ovarian tissue after calibration, the relative value of Vg at stage 2 of ovarian development is significantly increases relative to stage 1, 3, and 4 as 13.5, 6.0 and 2.0-fold, respectively ($P < 0.01$). In the hepatopancreas tissue after calibration, the relative value of Vg is significantly increased in the hepatopancreas at stage 3 as 270-fold ($P < 0.05$). Furthermore, the relative Vg mRNA expression in the ovary is greater than that in the hepatopancreas at all stages of ovarian development indicating that ovary is the major source of Vg synthesis. These results from real-time PCR show Vg gene expression in the ovary and hepatopancreas, which confirm the RT-PCR results that determined the sites of Vg synthesis.



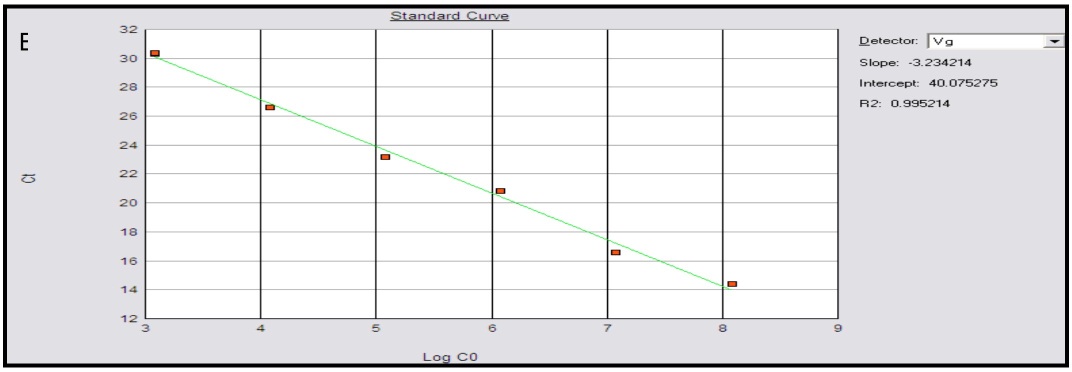


Fig. 44 Real-time PCR standard curve of *Penaeus merguiensis* Vg.

Amplification plot (A) and standard curve (B) for Vg standards were performed in ranging from $1.2x10^3$-$1.2x10^8$ copy number of target sequence. The standards are amplified for 40 cycles, and the standard curve is made by plotting the log of input amount (copy number) against the first cycle (Ct) that showed significant increase in fluorescence (Delta Rn).

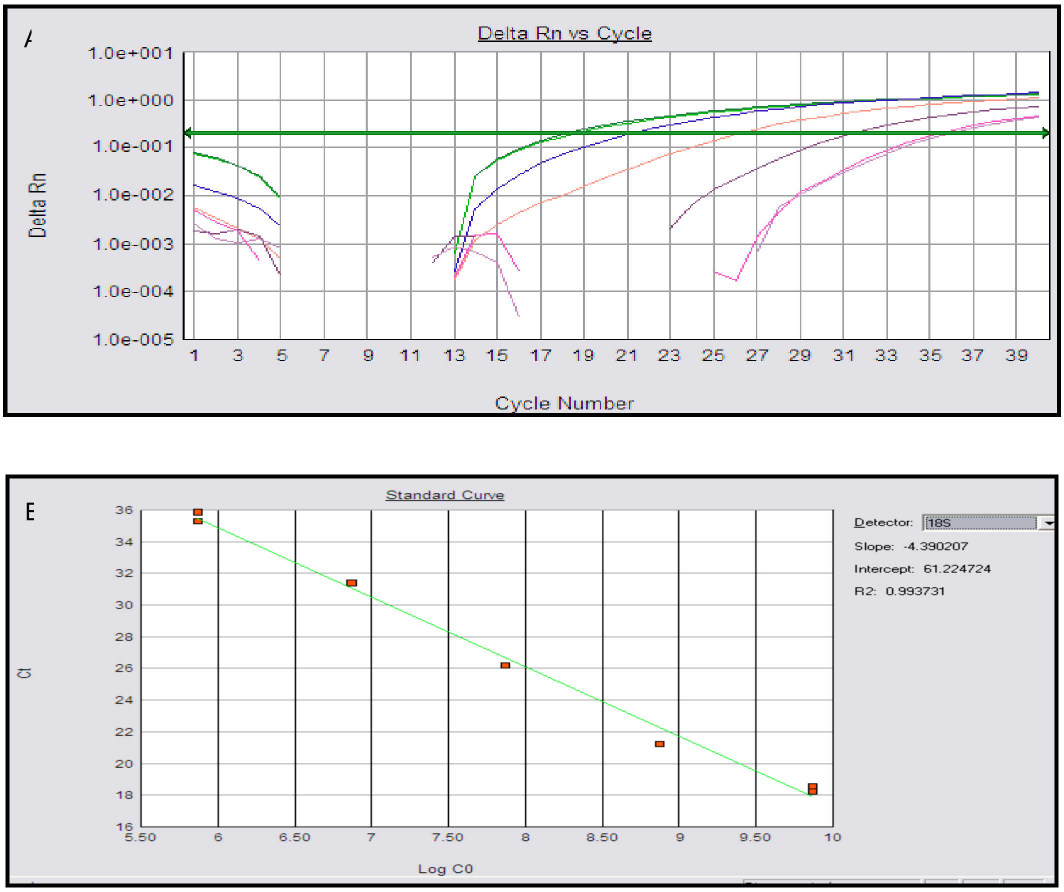


Fig. 45 Real-time PCR standard curve of 18s rRNA.

Amplification plot (A) and standard curve (B) for 18srRNA standards were monitered in ranging from $7.4x10^5$-$7.4x10^9$ copy number of target sequence. The standards are amplified for 40 cycles, and the standard curve is made by plotting the log of input amount (copy number) against the first cycle (Ct) that showed significant increase in fluorescence (Delta Rn).

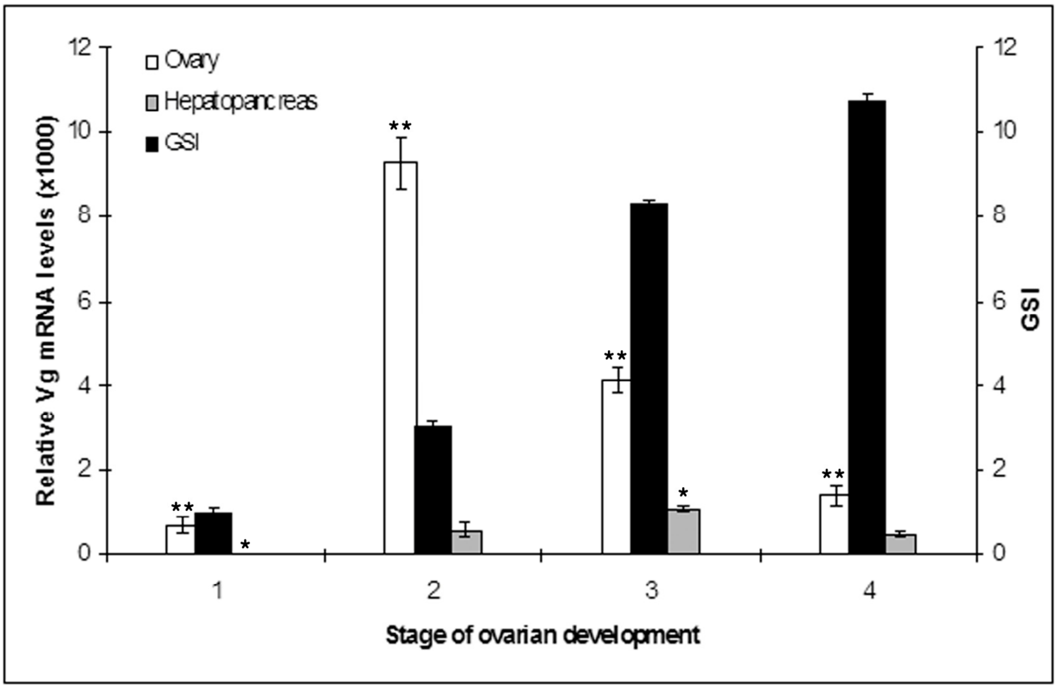


Fig. 46 Real-time PCR analysis of Vg cDNA.

Changes in the gonadosomatic index (GSI) and expression levels of Vg mRNA in the ovary and hepatopancreas of naturally maturing shrimps. Female shrimps were divided into four groups based on the developmental stages of the ovaries. Open columns, Vg mRNA level in the ovary relative to an 18s rRNA standard; shaded columns, Vg mRNA level in the hepatopancreas relative to an 18s rRNA standard; solid columns, GSI. Results are represented as the mean ± standard error of the mean of four shrimps in each stage. Significant differences between stages, $P < 0.05$ and $P < 0.01$, are indicated with * and **, respectively. Data were analyzed by ANOVA.