JOURNAL OF PROTEOMICS 73 (2010) 2354–2364

available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/jprot

Efficient identification of proteins from ovaries and hepatopancreas of the unsequenced edible crab, *Cancer pagurus*, by mass spectrometry and homology-based, cross-species searching

Deborah A. Ward[a], *Elaine M. Sefton*[a], *Mark C. Prescott*[a], *Simon G. Webster*[b], *Geoff Wainwright*[c], *Huw H. Rees*[a], *Michael J. Fisher*[a,*]

[a]*School of Biological Sciences, The University of Liverpool, Biosciences Building, Crown St., Liverpool, L69 7ZB, United Kingdom*
[b]*School of Biological Sciences, University of Wales, Bangor, Gwynedd, LL57 2UW, United Kingdom*
[c]*2Bio Ltd., Crown St., Liverpool, L69 7ZB, United Kingdom*

ARTICLE INFO

Article history:
Received 26 April 2010
Accepted 15 July 2010

Keywords:
Crustacean proteomics
Unsequenced genome
Homology-based searching
Similarity-based proteomics
De novo sequencing
LC-MS/MS

ABSTRACT

Proteome maps of hepatopancreas (midgut gland) and ovarian tissues of the crustacean, *Cancer pagurus* (Decapoda; edible crab) have been produced by 2D-PAGE and identification of proteins, following trypsin proteolysis, by electrospray MS/MS and database searching. Owing to the lack of sequence information on proteins and fully sequenced genomes amongst the decapod crustaceans and given the evolutionary distance to the nearest full genome database (*Daphnia*), it was necessary to adopt a non-conventional identification approach. Thus, a strategy was developed for effective identification of decapod proteins by sequence similarity, homology-based cross-species database searching, using various algorithms and a combination of NCBI Crustacea and Arthropoda databases, together with the Arthropoda PartiGene database (Blaxter, University of Edinburgh). In both hepatopancreas and ovary tissues, the largest group of proteins identified were a variety of enzymes, followed by a smaller number of storage/transport proteins [including vitellogenin (yolk protein), several subunits of hemocyanin, cryptocyanin, ferritin and calreticulin], with fewer structural proteins (actin, tubulin) and heat-shock proteins, in addition to a number of proteins of miscellaneous functions. Such protein identifications allow the development of tools, such as antibodies and RNA/DNA probes, to investigate the functions of the proteins in specific tissues during development.

1. Introduction

Cancer pagurus is a commercially important animal with approximately 60,000 tonnes landed globally in 2008 (www.fao.org/fishery/species/2627/en). Currently, all *C. pagurus* are wild-caught, since farming of these animals has not been possible, due to their slow growth and exacting culture requirements. In those crustaceans which are successfully farmed (e.g. shrimp and lobster), difficulties with breeding are often experienced in captivity. Hence, the aquaculture industry is particularly interested in the process of crustacean ovarian development and its hormonal regulation.

Ovarian development in decapod crustaceans is hormonally controlled. There is evidence for regulation by the steroidal ecdysteroids, the terpenoid methyl farnesoate, and neuropeptides [1–3]. Vitellogenin synthesis is an important component of ovarian development and occurs in both the ovary and hepatopancreas [3–5]. This storage protein acts as a source of

* *Corresponding author*. Fax: +44 151 795 4406.
E-mail address: fishermj@liv.ac.uk (M.J. Fisher).

1874-3919/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jprot.2010.07.008

nutrients in embryogenesis. Therefore, to allow investigation of the mechanisms of hormonal control of ovarian development in the edible crab, *Cancer pagurus*, a prerequisite is to produce comprehensive proteome maps of these tissues.

There is a paucity of genome/protein sequence information in *C. pagurus*. Most of the completely sequenced genomes of arthropods are from insects, with only the water flea, *Daphnia pulex*, genome being completely available for Crustacea (http://wfleabase.org/genomics/). Within the decapod crustaceans, the sequences of approximately 15,000 gene/proteins have been submitted to the NCBI (taxonomy) website, with approximately 37 gene/protein entries from *C. pagurus*. Numerous protein/nucleotide sequences, including ESTs from sixty different arthropod species have been recently made available via the 'Arthropoda' database (http://www.nematodes.org/downloads_area/databases/ARTHROPODA/). The protein sequences from *C. pagurus* are likely to be quite divergent from the orthologous ones from the non-decapod sequences, owing to the evolutionary distances between the organisms. Thus, it is essential to employ an appropriate strategy to use short peptide *de novo* sequences derived from MS/MS data to identify the parent proteins from *C. pagurus*. Error-tolerant algorithms (e.g. MS-BLAST and FASTS) are available that permit the matching of peptides where some of the amino acids vary from the sequence held in the database.

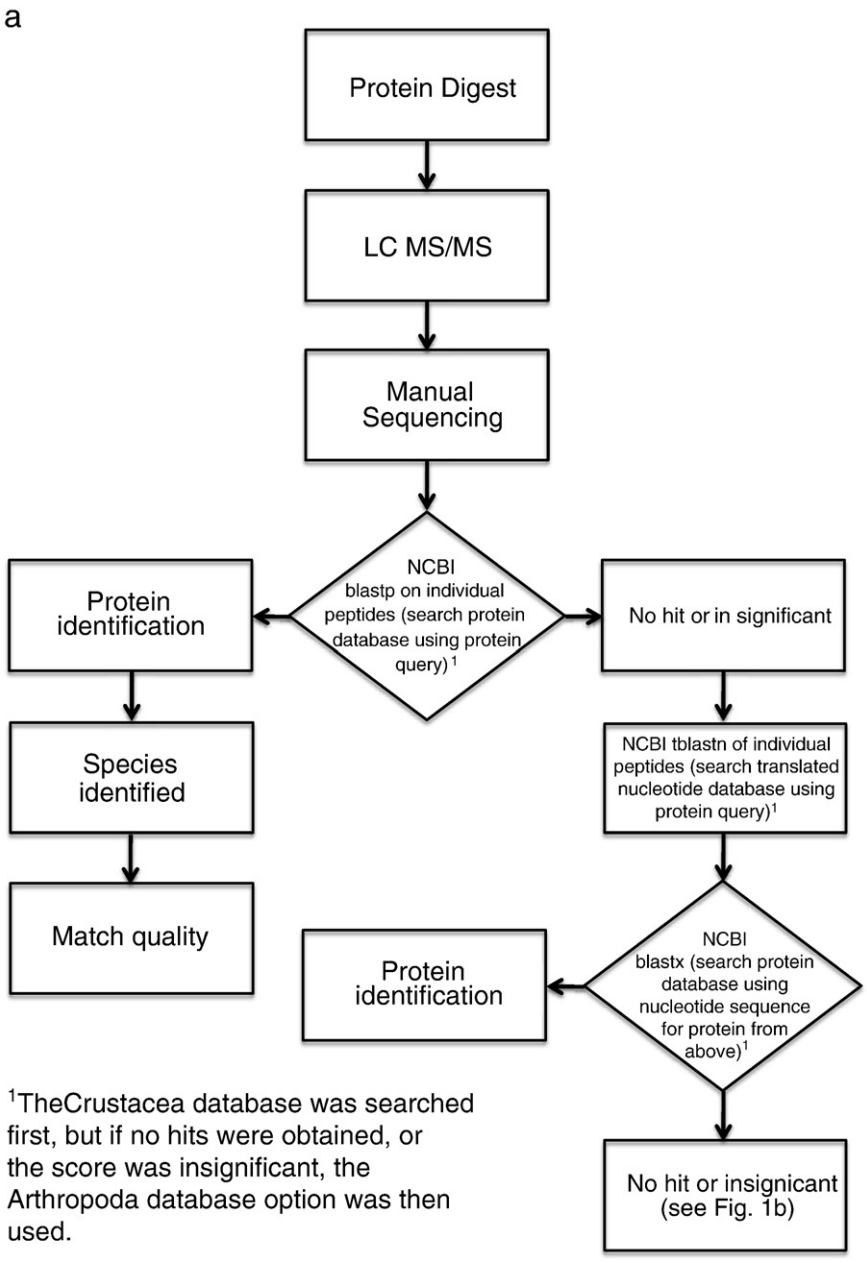

Fig. 1 – Workflow of strategies used for database searches of MS/MS data for protein identification. Initially, for all searches, strategy A (a) was used for searches against NCBI databases. Where this was not successful, strategy B (b) was employed, where initial MS–MS-derived peptide sequence searches were made against an in-house database constructed from FASTA EST sequences downloaded from the Arthropoda PartiGene database (see Materials and methods).

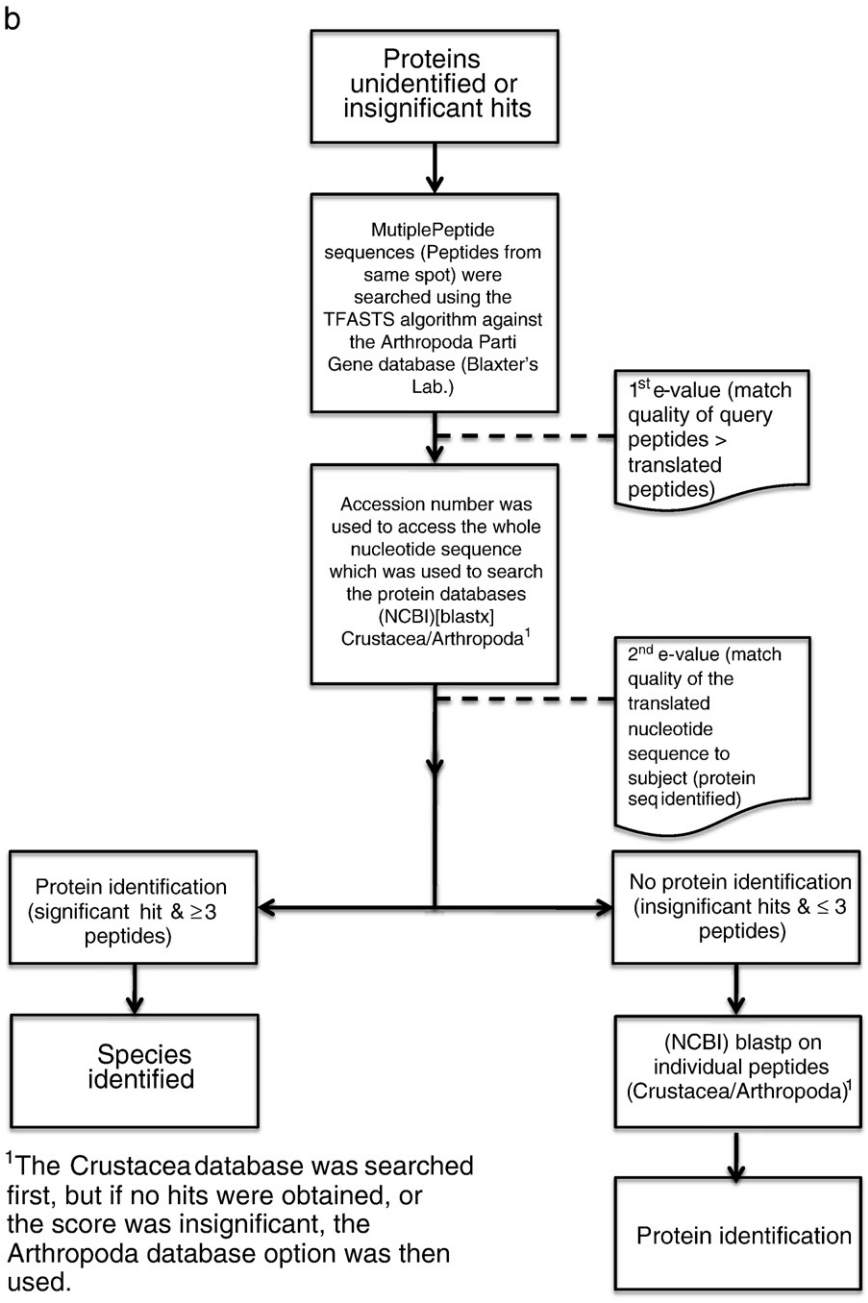

Fig. 1 (*continued*).

The conventional BLAST (Basic Local Alignment Sequence Tool) algorithm [6] has been optimised to identify database proteins from short peptide query sequences. MS-BLAST was developed by modifying the parameters of the BLAST algorithm to enable it to find imperfect matches using groups of short peptide sequences, as generated by MS/MS of trypsin-cleaved spots derived from 2D-PAGE fractionation of proteins [7]. The effectiveness of this algorithm for cross-species, sequence-similarity searching of databases has been evaluated [8].

Like MS-BLAST, FASTS permits peptides (e.g. short MS/MS-generated peptide sequences) of unknown order to be used for identification of proteins, including those from unsequenced organisms [9]. FASTS uses alignment probability as optimality criteria, rather than similarity scores. The algorithm used is a modified form of the FASTA algorithm (FAST-All i.e. a FAST search of either DNA or protein; [10]). As with MS-BLAST, the scoring matrix has been modified to account for the isoleucine/leucine and lysine/glutamine isobaric amino acids [9].

Various strategies have been used for error-tolerant, sequence-similarity (homology)-based identification of proteins from MS/MS data [11–17]. In the case of Crustacea, proteomic studies have been limited, with MS/MS spectral data searching against either the NCBI protein database [18–21] or specialised EST databases for the relevant species [19,22]; these include two decapod crustaceans *Penaeus vannamei* and *Litopenaeus vannamei* [19,21].

We now report work on the production of proteome maps of hepatopancreas and ovary tissues of *Cancer pagurus* by 2D-

PAGE and the identification of proteins, following trypsin proteolysis, by electospray MS/MS and database searching. Owing to the paucity of sequence information on proteins from decapod crustaceans, the lack of a fully sequenced genome amongst the decapods, and the evolutionary distance to the nearest full genome database (*Daphnia*, Crustacea), it was necessary to combine the use of several algorithms in our protein identification approach. Thus, we have used a strategy for effective identification of decapod proteins by optimisation of homology-based cross-species searching, using a combination of NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and Arthropoda (http://www.nematodes.org/downloads_area/databases/ARTHROPODA/index.shtml) databases.

2. Materials and methods

2.1. Animals

Adult females of *Cancer pagurus* (Linnaeus), the edible crab, were obtained from fishermen in Anglesey, North Wales, and maintained in a recirculating seawater system at ambient light and temperature. Ovaries and hepatopancreas were dissected after cold-anaesthesia and rinsed in crustacean saline [22], before freezing in liquid nitrogen. The stage of ovarian development was determined according to established criteria [23,24]. In brief, crabs were assigned a stage of ovarian development from 0 to 4. Stages 1 to 4 ovaries are vitellogenic, with steadily increasing organ size, oocyte diameter, and quantity of accumulated yolk protein (orange colour).

2.2. Sample preparation

Hepatopancreas or ovary tissue (100 mg) from a Stage 0 female *C. pagurus* was homogenised in Tris buffer (40 mM) including a protease inhibitor cocktail (Sigma) used at 10× recommended concentration. The homogenate was centrifuged (15000 *g*, 4 °C, 30 min), the supernatant containing the proteins was carefully removed using a fine gauge needle, taking care not to dislodge the fat layer on the surface. The supernatant was placed into a clean tube and re-centrifuged as above. The protein content of the supernatant was determined by Bradford assay (Bio-Rad).

2.3. 2D-PAGE

Supernatant proteins (100 μg for analytical gels and 400 μg for preparative gels) were precipitated overnight at –20 °C following the addition of 5 sample volumes of 10% (v/v) TCA in acetone. The samples were then centrifuged for 5 min at 8000 g, the resulting supernatant was discarded and the precipitated protein pellets were re-solublised in 320 μl of rehydration buffer [8 M urea, 2 M thiourea, 4% (w/v) CHAPS [(3-[(3-cholamidopropyl) dimethylammonio]-1-propanesulfonate], 0.2% (w/v) Biolytes 3–10, 0.05% (w/v) ASB14 detergent] by shaking at room temperature for 2 h. The solutions were then centrifuged at 8,000 g for 5 min to remove any insoluble material and then applied to the immobilised pH gradient (IPG) strip holder (Bio-Rad). IPG strips (17 cm), pH 3–10 non-linear (Bio-Rad), were then placed over the protein solutions and overlaid with mineral oil. Isoelectric focusing was then performed using a Protean IEF cell (Bio-Rad), 250 v for 15 min (linear ramp), 10000 v for 3 h, followed by 10000 v to 60000 v/h (linear ramp). Following isoelectric focusing, the IPG strips were equilibrated, in equilibration buffer [50 mM Tris, pH 6.8, 6 M urea, 2% (w/v) SDS, 30% (v/v) glycerol, bromophenol blue] containing 20 mM DTT for 15 min, then in equilibration buffer containing 25 mM iodoacetamide for 15 min. IPG strips were then loaded onto 10% (w/v) SDS-PAGE gels and run at 15 mA/gel until the dye from the agarose sealing gel had migrated into the resolving gel and then at 30 mA/gel until the dye front had run off the gel. Electrophoresis was performed in Tris-glycine running buffer [25 mM Tris, 250 mM glycine, 0.1% (w/v) SDS]. Analytical gels were silver-stained [25] and the preparative gels stained with colloidal Coomassie blue [26]. The 2D-PAGE gels were scanned using a GS-710 Calibrated Imaging Densitometer (Bio-Rad) and the images analysed by PD Quest software (Bio-Rad).

2.4. Trypsin digestion and mass spectrometric identification of proteins

Spots selectively chosen for identification by mass spectrometry were excised from a preparative Coomassie Blue-stained gel and transferred to microcentrifuge tubes. The gel plugs were destained by washing in 50% (v/v) ammonium bicarbonate (50 mM), acetonitrile solution for 15 min at 37 °C. The plugs were then dehydrated using 100% acetonitrile at 37 °C and rehydrated overnight in 10 μl 50 mM ammonium bicarbonate containing 1 μl of trypsin solution (0.1 μg/μl trypsin in 50 mM acetic acid; Promega). The peptide mixture was removed and made up to a volume of 20 μl with 0.1% (v/v) formic acid in water, before injection into the LC-MS/MS system. The system consisted of an UltiMate nano-liquid chromatograph (LC Packings, Dionex, Surrey, UK) connected to a Q-Tof Micro™ tandem mass spectrometer (Micromass) operated in positive ion mode. Chromatography was on a C_{18} pre-column connected to a PepMap C_{18} column (3 μm 100 Å packing;15 cm × 75 μm i.d.; LC Packings) using a linear gradient of 5% (v/v) solvent B [0.1% (v/v) formic acid, 80% (v/v) acetonitrile in water] in solvent A [0.1% (v/v) formic acid, 2% (v/v) acetonitrile in water] to 100% (v/v) solvent B over 60 min at a flow rate of 200 nl/min. The eluted peptides were ionised by electrospray ionisation (ESI) and were monitored at 214 nm. The mass spectrometer was operated in Data Directed Analysis (DDA) mode, where a survey scan was acquired from 400–1500 (m/z), searching for doubly or triply charged ions, since these are most likely to be peptides. When such ions were detected, the spectrometer switched to MS/MS mode, and the ion of interest was fragmented in the collision cell and an MS/MS mass spectrum acquired over the mass range 80–2000 Da.

2.5. Database searching

MS/MS spectra of the peptides from *C. pagurus* were analysed using both NCBI-nr and EST databases. Since *C. pagurus* is still largely unsequenced, a limited number of gene/protein sequences are publically available (37 protein and 38 nucleotide entries currently available in the NCBI-nr database). Conventional database searches were made using MASCOT

Ion searches (Matrix Science; http://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=MIS), using default settings, allowing for carbamidomethyl modification of cysteine and oxidation of methionine. As a consequence of the poor representation, such database searches failed to identify even the most abundant proteins, therefore MS/MS combined with similarity searches was used to mine for orthologous proteins.

The two different protein identification strategies carried out are outlined in Fig. 1(a) and (b). Initially, for all proteins, strategy A was used, but where this was not successful, strategy B was employed.

2.5.1. *Strategy A*

In this strategy, for searches of the NCBI database, all peptide and nucleotide sequences were searched against the Arthropoda/Crustacea entries. Initially, *de novo* sequencing of short peptides was undertaken. For this, subtraction of background noise from the MS/MS spectra was performed using MassLynx (version 4) with the following parameters: polynomial order 15, 50 % below curve, tolerance 0.01. Smoothing of the spectrum was performed using the Savitzky Golay method with 2 smooths and smooth window ±3. The data were centred using the centroid top method at 80%. Any doubly-charged ions were converted to their singly-charged mass using MaxEnt. Partial peptide sequences were then determined by semi-automatic interpretation of these processed MS/MS data using the PepSeq software within the MassLynx package (Waters; default parameters were used: MW tolerance, 0.5 Da; intensity threshold, 0.75; fragment ion tolerance, 0.3 Da). Allowance was made for oxidised methionine and for carbamidomethyl modification of cysteine. Generally, we chose the highest scoring sequence. The resulting short peptide sequences were searched independently (Fig. 1a), using the NCBI-nr database (http://blast.ncbi.nlm.nih.gov/Blast.cgi) using BLASTp (default parameters), with the option to search for short, nearly exact matches. If there was no protein candidate or the quality match was poor, the peptides were then searched using tBLASTn (default parameters). The full translated nucleotide sequence was then searched using BLASTx (default parameters). If there was no protein candidate or the quality match was poor, identification strategy B (Fig. 1b) was carried out.

2.5.2. *Strategy B*

An in-house database was constructed with the FASTA EST sequences downloaded from the Arthropoda PartiGene database (http://www.nematodes.org/downloads_area/databases/ARTHROPODA/index.shtml) comprising Crustacea, Hexapoda and Chelicerata (courtesy of Professor Mark Blaxter, University of Edinburgh). All peptide and nucleotide sequences were searched against the Arthropoda/Crustacea entries in the database. The searches against the in-house database were performed on a local Unix server using the TFASTS algorithm (default parameters). This allowed multiple peptide sequences from the same protein to be batch-searched, enhancing protein identification specificity. After processing, the nucleotide accession number was used to identify the whole nucleotide sequence in the ARTHROPODA database. The full nucleotide sequence was then searched against the NCBI protein database (Crustacea/Arthropoda) using BLASTx (default parameters). If there was no protein candidate, or the quality match was poor, or there were ≤3 peptides matched, each peptide was searched independently using BLASTp (default parameters), with the option to search for short nearly exact matches.

3. Results and discussion

3.1. *2D-PAGE separation of hepatopancreas and ovarian proteins*

In preliminary tests using hepatopancreas and ovary of stage 0 *C. pagurus*, approximately 100 μg of soluble protein produced gels of high resolution after silver staining. To enable spots of varying intensities to be excised from each gel and to obtain protein identifications with high confidence, a higher concentration of protein was required. When gels were run with different protein concentrations, 400 μg of soluble protein produced a gel where the spots were well resolved (Figs. 2 and 3). Increasing the concentrations, particularly of the ovarian protein supernatant, resulted in a decrease in resolution and streaking of the acidic proteins, that can be attributed to the high fat content in the ovary tissue. Although lipid extraction [27] reduced the amount of streaking, the decrease was not appreciable and, therefore, it was not carried out in this study. Three replicate gels were run for each tissue and stained with colloidal Coomassie blue.

A number of random spots of varying intensities were excised from the gels from each tissue, trypsin-digested and subjected to LC-MS/MS, to allow us to evaluate the adopted protein identification strategy and to identify some of the proteins. It was predicted that the more abundant spots would, in the main, be enzymes or, in the case of proteins extracted from the hepatopancreas, storage proteins. The hypothesis is that both enzymes and storage proteins should be relatively well conserved across phylogenetically distant species.

3.2. *Identification of proteins*

3.2.1. *Database searching and interpretation of data*

As a consequence of the poor representation of genome/protein sequence information for decapod crustaceans, and the paucity of sequence information in *C. pagurus*, it became apparent from conventional database searches, that a different strategy was required to allow the characterisation of crustacean proteomes. Specifically, in some cases, identification was hampered, since orthologous genes and their corresponding proteins retain a lower percentage identity, as organisms become more phylogenetically distant. The protein identification strategies used in this study are outlined in Fig. 1a and b with an example of each strategy illustrated in the Supplementary Information (Figs. S1a and S1b).

Firstly, the MS/MS spectra were processed using the PepSeq software within the MassLynx package (Waters). A selected number of the peptide sequences processed from the most abundant spots on the gels were searched independently against the NCBI-nr protein database using the BLASTp algorithm. The percentage of sequences assigned to any one protein was low, with many sequences unidentified. FASTS and MS-BLAST homology searching, which permit the

matching of multiple peptides, where some of the amino acids vary from the sequence held in the database, also had limited use in this study, resulting from the under-representation of crustacean protein sequence information within the NCBI-nr database. By incorporating various NCBI databases (Crustacea or Arthropoda) with different algorithm options (tBLASTn or BLASTx; Fig. 1a), an increasingly confident protein identity was achieved. However, there were still a large percentage of quality peptide sequences unmatched. The second strategy used (Fig. 1b), incorporated not only the NCBI databases/algorithm options but also the Arthropoda PartiGene database developed in Professor Mark Blaxter's laboratory, University of Edinburgh.

An in-house developed database, based on the Arthropoda PartiGene database, was searched, using the TFASTS algorithm, allowing multiple peptides to be used simultaneously. Searching of multiple peptides maximises the search potential of the queries, increasing the chance of protein identity. TFASTS is a modified form of the FASTA algorithm and compares linked peptides to a translated DNA database [9]. An example of such a TFASTS search (Fig. 1b) is given in the Supplementary Information (Fig. S2) for ovary spot 1 (ultimately identified as protein disulphide isomerase).

The differences between the query sequences and subject and the unmatched MS/MS-derived peptides can be explained by the cross-species nature of the searches. In some cases, identification is hampered, since orthologous genes and their corresponding proteins retain a lower percentage identity, as organisms become more phylogenetically distant. Genetic diversity and geographic location have also been reported as contributory factors [28]. Distinct sequences with no database matches were detected, representing either completely novel proteins or genes with sequences that are too divergent from those of known sequence and similar function in other organisms to enable database matching.

The accession numbers of protein hits obtained were then used to retrieve the complete nucleotide sequence which was then searched in the NCBI database using BLASTx. The identified proteins, together with some supporting data are summarised in Table 1. The full supporting data are tabulated in the Supplementary Information (Tables S1 and S2).

Of 43 protein spots that were excised from the hepatopancreas gels (Fig. 2), 30 (70%) were confidently identified by the strategies employed. Similarly, 29 proteins (76%) of the 38 spots excised from the ovarian gels (Fig. 3) were identified.

To date, there have been a limited number of proteomic studies on Crustacea. Where these have been carried out, custom databases have been used in conjunction with those publically available. A study on *Artemia franciscana* (brine shrimp) used the MASCOT search engine to search a custom database containing 30,000 high-quality ESTs generated in-house, together with ESTs/proteins from crustaceans and insects that were publically available. 59 out of the 75 (78%) excised 2-D PAGE spots were confidently identified [21]. Another study on *Fenneropenaeus chinensis* (prawn) similarly used MASCOT to initially search both the NCBI-nr and EST databases online. Then, an in-house custom constructed database containing potential translation products of an *F. chinensis* EST database, was searched using SEQUEST. In total 51 out of the 67 (76%) excised 2-DE spots were confidently identified [29]. A third study on *Armadillidium vulgare* used MASCOT as in the previous studies, to search the NCBI database. The latter study did not construct an in-house database and, therefore, resulting from the paucity of crustacean genome/protein sequence

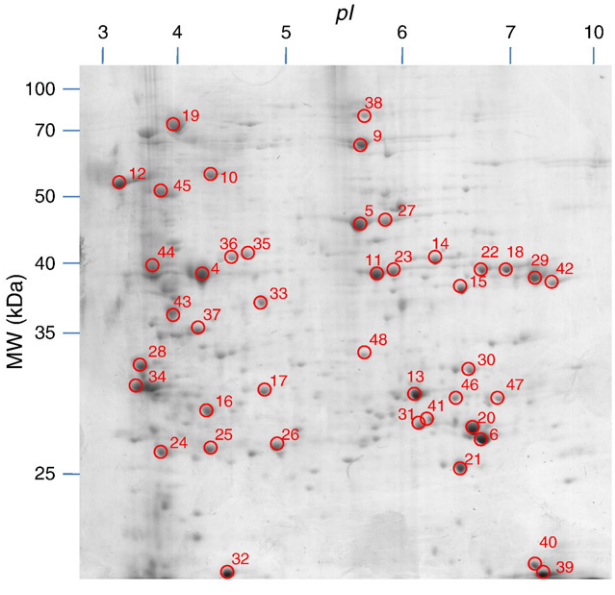

Fig. 2 – Colloidal Coomassie blue-stained 2D-PAGE gel image of supernatant from hepatopancreas tissue. The sample was run on a non-linear pH 3–10 isoelectric focusing strip, followed by fractionation on a 10% SDS-gel. Spots removed for identification by MS–MS mass spectrometry are circled.

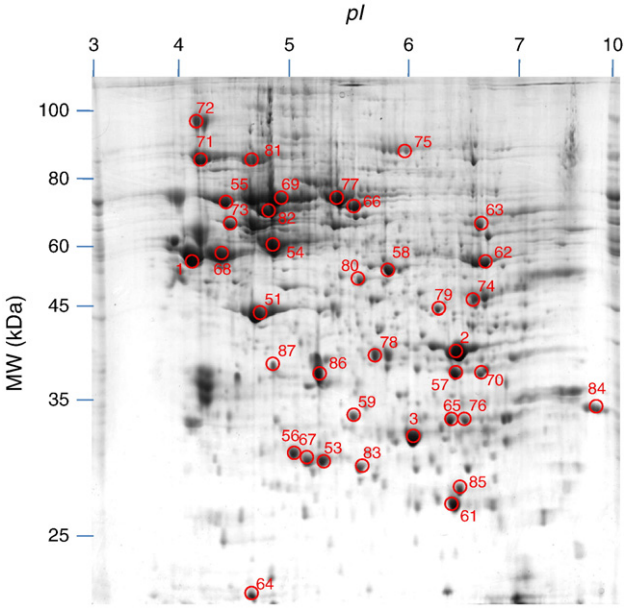

Fig. 3 – Colloidal Coomassie blue-stained 2D-PAGE gel image of supernatant from ovary tissue. The sample was run on a non-linear pH 3–10 isoelectric focusing strip, followed by fractionation on a 10% SDS-gel. Spots removed for identification by MS/MS mass spectrometry are circled.

Table 1 – Identification of proteins from *C. pagurus* hepatopancreas and ovary tissues.

Function	Spot no (o = ovary, h = hepatopancreas)	Protein ID	Protein accession no	No peptides hit	e-value (peptides vs nucleotide sequence)
Enzymes	1o	Disulphide isomerase	ACN89260.1	7	5.20E–42
	43 h		NP_524079.1	1	
	45 h		NP_730033.1	1	
			XP_001950073.1	1	
			NP_001037171	1	
	54o		EAT47483	4	1.20E–09
	2o	Arginine kinase	ABI98020	12	2.3E–52
	15 h		ACB46940.1	3	
			AF233355_1	1	
			ACB46942.1	3	
	6 h	Superoxide dismutase	AAF74771	5	4.00E–27
			CAR85665.1	3	
	13 h		CAR85665.1	5	5.00E–27
			AAF74771	1	
	3o		AAF74771	4	1.30E–41
	56o [a]	Glutathione peroxidase	ACJ53746.1	2	
	56o	Peroxiredoxin	ACO10673.1	2	
			ACF35639.1	2	
			XP_001849994.1	1	
	57o	Glyceraldehyde 3-phosphate dehydrogenase	AAS02313	2	9.00E–149
	58o	Quinone oxidoreductase	XP_001867838.1	2	5.90E–05
	62o	Catalase	ABW82155.1	2	2.90E–32
			AAR99908	2	4.60E–14
	63o	Bifunctional purine biosynthesis protein	XP_001945671	2	
			XP_001662952.1	2	
	67o [b]	Carbon-nitrogen hydrolase /(GE12820 gene product)/Vesicle transport protein	XP_002090411.1	1	
			XP_001657368.1	1	
			XP_002083725.1	1	
	75o	NADH dehydrogenase	ABD85303	1	0.39
	79o	Fumarylaceto-acetate hydrolase	ACO15559.1	3	0.03
	5 h	Enolase	AAC78141	4	
			AAS02303.1	1	
	27 h		AAC78141	3	
	9 h	Pyrroline-5-carboxylate dehydrogenase	XP_969408.1	1	
			EAT43137.1	1	
			XP_001849188	1	
	11h [b]	CG18473-PA	NP_731339.1	1	
		(Phospho-triesterase-like enzyme)	EAT37734	1	
		OR			
		Protein kinase catalytic domain	XP_001866186.1	1	
	17 h	Glyoxalase	ACO11529.1	3	
	20 h	Triose phosphate isomerase	ABB81879.1	4	
	31 h		ABB81879.1	2	
	46 h	Malate dehydrogenase (malic enzyme)	NP_609394	1	
			XP_001969473	1	
			NP_609394	3	
	47 h		XP_001969473	1	
	21 h	(GF22580) Putative short-chain alcohol dehydrogenase	XP_001965602.1	1	
	22 h	Fructose 1,6-bisphosphate aldolase	ACH81781.1	2	
	29 h		AAT01078	1	
			NP_001091766	1	
			AAU95197	1	
	26 h	Chymotrypsin	CAA71672.1	2	
			AAL67441.1	1	
			ACC68669.1	1	
Storage/ transport proteins	61o	Vitellogenin	gb|AAP76571.2	2	
	65o		ABX89617	1	
	30 h		ABC41925	1	
			ABX89617	1	
	66o [c]	Hemocyanin/Cryptocyanin	AAW57889	2	3.00E–17
			AAL27460.1	4	7.60E–13

Table 1 (*continued*)

Function	Spot no (o = ovary, h = hepatopancreas)	Protein ID	Protein accession no	No peptides hit	e-value (peptides vs nucleotide sequence)
	69o		AAW57891.1	3	
			AAL27460.1	3	
	73o		AAW57891	4	
	77o		AAW57889	5	6.90E–07
			P83175.1|HCYB_CANPG	1	
	37 h		CAB38043.1	2	
	38 h		AAD09762.1	3	
			CAB38043.1	3	0.0035
	64o	Ferritin	AAX55641.1	2	
			ABB05537.1	2	
	32 h		AAX55641	1	
			CAA62186.1	1	
	12 h	Calreticulin	BAC57964	1	
	85o	Intraflagellar transport 140 homologue	XP_001946444.1	2	1.2
	18 h	NEFA-interacting nuclear protein NIP-30	XP_001606256.1	2	1.6
	28 h	GA18769-PA transporter	EAL29862	1	
Miscellaneous proteins	4 h/82o[b]	HSP70 OR Hemocyanin	AAT46566	1	
			ABF83606.1		1.30E–13
			CAL68995.1	2	
			AAW57893.1	2	
	71o	HSP90	ACO83357.1	4	4E–32
	55o	78 kDa glucose-regulated protein (HSP70 family)	ABM92447.1	4	1.20E–09
	81o	CDC48 cell division protein	XP_001982088.1	1	
			XP_002089877.1	1	
	19 h	Tetratricopeptide repeat protein (APC)	XP_001949024.1	2	1.2
	48 h	AGAP006729 (DUF1907)	XP_309016	1?	
	78o[b]	GD21488 gene product (ribonuclease inhibitor family) OR similar to CG3814-PA (Bax apoptosis inhibitor family)	XP_002105438.1	1	
			XP_391854	1	
Structural proteins	51o	Actin	AAG16253	4	9.7E–38
	24 h		ACI23575	2	
	36 h		AAY85815.1	2	1.00E–10
	68o	Tubulin	AAC47306.1	7	5.50E–33

[a] It is not possible to differentiate between these two glutathione dependent enzymes.
[b] It is not possible to distinguish between these functionally distinct proteins.
[c] See the main text for a discussion on the relationship between hemocyanin and cryptocyanin.

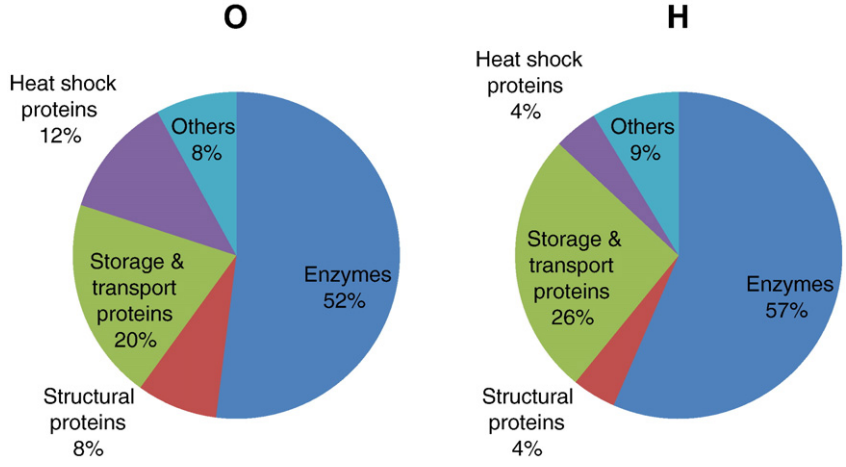

Fig. 4 – Distribution of proteins identified amongst different functional groups. O, ovary; H, hepatopancreas.

information in the database searched, the total percentage of proteins identified was considerably lower: 56 out of the 110 (50%) excised 2-D PAGE spots were confidently identified [20].

It is well recognised that conventional techniques for MS-based protein identification are heavily reliant upon exact matching of masses of peptides, to corresponding masses from sequences from database entries. As a consequence, many sequences searched from species whose genomes which are largely unsequenced, remain unidentified. There are, however, certain categories of proteins such as enzymes which are highly conserved in different species, enabling proteins to be identified by their sequence similarity to known orthologues from phylogenetically related species. Of course, this has limited use, because in the majority of studies, there is a need to identify proteins that are not well conserved between species.

It is evident from Supplementary Tables S1 and S2 that most of the gel-estimated molecular weight (Mr) and pI values matched well with the corresponding theoretical values of orthologous protein hits from other species. Appreciable differences in some of these values may be explained by amino acid sequence differences, including insertions or gaps, in some of the less conserved regions compared with other species. These factors will also affect the percentage coverage observed during database searches of MS data.

The strategy we present here, which integrates the conventional techniques for MS-based protein identification with a custom database (Arthropoda PartiGene), searched using the FASTS algorithm, has resulted in a 29% increase in protein identification in the hepatopancreas and 68% in the ovary when using strategy 2 compared to strategy 1. Confidence in our identifications is provided not only by the combined data in Tables S1 and S2 (numbers of peptides hit, peptide coverage, e-values at each stage, and comparison of theoretical and observed Mr and pI values), but by the fact that in many cases, hits were obtained against the same orthologous proteins from several species. However, in the case of spot 78 (ovary), more than one protein identification was possible.

3.2.2. *Proteins identified*

In both the hepatopancreas and ovaries, as expected, the largest group of proteins identified were enzymes (Fig. 4), followed by a lower number of storage and transport proteins, with fewer structural and heat-shock proteins. Furthermore, we did not detect transcription factors in the spots taken, since they presumably occur at low abundance. Additionally, the methods utilized in extracting and fractionating proteins from tissue samples favours extraction of more soluble cellular components. However, it is also the case that enzymes and storage proteins are relatively well conserved across phylogenetically similar species. Amongst the enzymes identified, is a particularly interesting and unusual manganese superoxide dismutase (MnSOD). All decapod crustaceans depend on the copper protein, haemocyanin, for oxygen transport. It has been shown that the blue crab, *Callinectes sapidus*, lacks the ubiquitous cytosolic copper-dependent enzyme copper/zinc superoxide dismutase, but has an unusual cytosolic MnSOD that lacks a mitochondrial transit peptide and is, thus, retained in the cytosol; a second MnSOD occurs in the mitochondria [30]. Such a unique phenomenon occurs in all Crustacea that use haemocyanin for oxygen transport. Apparently, the MnSOD gene duplication is reputed to be as old as the origin of the arthropod phylum [30].

Storage/transport proteins identified included vitellogenin (yolk protein), several subunits of hemocyanin, cryptocyanin, ferritin and calreticulin. Hemocyanin, which is synthesised in hepatopancreas, is the protein responsible for the transport of oxygen in the hemolymph. Members of the arthropod hemocyanin gene family, viz. hemocyanin, cryptocyanin, and phenoloxidase, consist of multisubunits that assemble into hexamers and higher aggregates [31]. In the case of hemocyanin, the hexamers exhibit species-specific subunit heterogeneity. Firstly, family members in arthropods are involved in at least three major functions. In many crustaceans, oxygen transport occurs through the copper-containing hemocyanin. Secondly, the enzymic activity of a sequence-related copper protein, phenoloxidase, is involved in the innate immune response [32]. Thirdly, phenoloxidase and perhaps hemocyanin are involved in cross-linking (sclerotization) of proteins in the new exoskeleton after moulting [33]. Cryptocyanin is another member of the hemocyanin gene family, but has no oxygen transport or oxidase function, since it has a reduced number of histidine residues that are critical for copper-binding. It is produced in high concentrations during premoult and is involved in forming the new exoskeleton [34,35]. Hemocyanin consists of multisubunits, since the subunits self-assemble into hexamers, 2-hexamers, 4-hexamers, 6-hexamers, and 8-hexamers [36]. Similarly, cryptocyanin subunits also assemble into hexamers [37]. Hemocyanin has been reported previously in ovaries [38–40]. In the crab, *Cancer magister*, hemocyanin is present in oocytes at all stages of development, but increased during the vitellogenic stage [41] which is probably coincident with patency of ovaries [42]. This hemocyanin is indistinguishable from maternal hemocyanin and is believed to be derived via endocytosis from the maternal hemolymph [41,43]. Ferritin and calreticulin are iron- and calcium-binding proteins, respectively. Other groups of proteins detected include heat-shock proteins and structural proteins (actin and tubulin), in addition to a number of miscellaneous function.

The current work has demonstrated how *de novo* peptide sequencing of peptide fragments derived from 2D PAGE experiments can yield information relating to the identity, and by inference putative function, of proteins expressed in tissues of a species with very poor representation within protein and DNA sequence databases. This new knowledge now opens the possibility of enabling new tools such as antibodies and RNA/DNA probes to be developed that can be used to investigate the function of these proteins within their specific tissues and at various stages of reproductive development. For example, an ability to investigate the changes in expression, location and function of the structural proteins actin and tubulin in ovary tissues during oocyte development and, in particular, around the time of developmental arrest at prophase-I of meiosis, would likely enable a deeper understanding of this delicately balanced developmental process, which currently stalls oocyte maturation and leads to pauses in ovarian maturation.

Acknowledgments

We thank the BBSRC for a Studentship to ES and for purchase of proteomic equipment. We are most grateful to Professor Mark Blaxter, University of Edinburgh, for making the Arthropoda PartiGene database sequences available to us for downloading. We thank Mr Ashley Tweedale, University of Wales Bangor and Mr Steve Corrigan for the supply and maintenance of *Cancer pagurus*.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at 10.1016/j.jprot.2010.07.008.

REFERENCES

[1] Quackenbush LS. Crustacean endocrinology, a review. Can J Fish Aquat Sci 1986;43:2271–82.
[2] Keller R. Crustacean neuropeptides: structures, functions and comparative aspects. Experientia 1992;48:439–48.
[3] Wainwright G, Rees HH. Hormonal regulation of reproductive development in crustaceans. In: Atkinson D, Thorndyke M, editors. Environment and Animal Development. Oxford: Bios; 2001. p. 71–84.
[4] Paulus JE, Laufer H. Vitellogenesis in the hepatopancreas and ovaries of *Carcinus maenas*. Biol Bull 1982;163:375–6.
[5] Tsang W-S, Quackenbush LS, Chow BKC, Tiu SHK, He J-G, Chan S-M. Organization of the shrimp vitellogenin gene: evidence of multiple genes and tissue specific expression by the ovary and hepatopancreas. Gene 2003;303:99–109.
[6] Altshul S, Maden T, Schaffer A, Zhang J, Zhang Z, Miller W, et al. Nucleic Acids Res 1997;25:436–40.
[7] Shevchenko A, Sunyaev S, Loboda A, Shevchenko A, Bork P, Ens W, et al. Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. Anal Chem 2001;73:1917–26.
[8] Habermann B, Oegema J, Sunyaev S, Shevchenko A. The power and the limitation of cross-species protein identification by mass spectrometry-driven sequence similarity searches. Mol Cell Proteomics 2004;3:238–49.
[9] Mackey AJ, Haystead AJ, Pearson WR. Getting more from less. Algorithms for rapid protein identification with multiple short peptide sequences. Mol Cell Proteomics 2002;1:139–47.
[10] Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988;85:2444–8.
[11] Shevchenko A, Valcu C-M, Junqueira M. Tools for exploring the proteomosphere. J Proteomics 2009;72:137–44.
[12] Liska AJ, Sunyaev S, Shilov IN, Schaeffer DA, Shevchenko A. Error-tolerant EST database searches by tandem mass spectrometry and MultiTag software. Proteomics 2005;5: 4118–22.
[13] Pandey A, Choudhary MK, Bhushan D, Chattopadhyay A, Chakraborty S, Datta A, et al. The nuclear proteome of Chickpea (*Cicer arietinium* L.) reveals predicted and unexpected proteins. J Proteome Res 2006;5:3301–11.
[14] Gowd KH, Krishnan KS, Balaram P. Identification of *Conus amadis* disulfide isomerase : minimum sequence length of peptide fragments necessary for protein annotation. Mol Biol Syst 2007;3:554–66.
[15] Grossman J, Fischer B, Baerenfaller K, Owiti J, Buhmann JM, Gruissem W, et al. A workflow to increase the detection rate of proteins from unsequenced organisms in high-throughput proteomics experiments. Proteomics 2007;7:4245–54.
[16] Waridel P, Frank A, Thomas H, Surendranath V, Sunyaev S, Pevzner P, et al. Sequence similarity-driven proteomics in organisms with unknown genomes by LC-MS/MS and automated *de novo* sequencing. Proteomics 2007;7:2318–29.
[17] Junqueira M, Spirin V, Balbuena TS, Thomas H, Adhubei I, Sunyaev S, et al. Protein identification pipeline for the homology-driven proteomics. J Proteomics 2008;71:346–56.
[18] O'Connell PA, Pinto DM, Chisholm KA, MacRae TH. Characterisation of the microtubule proteome during post-diapause development of *Artemia franciscana*. Biochim Biophys Acta 2006;1764:920–8.
[19] Chongsatja P, Bourchookarn A, Lo CF, Thongboonkerd V, Krittanai Ch. Proteomic analysis of differentially expressed proteins in *Penaeus vannamei* hemocytes upon Taura syndrome virus infection. Proteomics 2007;7:3592–601.
[20] Herbinière J, Grève P, Strub J-M, Thierse D, Raimond M, van Dorsselaer A, et al. Protein profiling of hemocytes from the terrestrial crustacean *Armadillidium vulgare*. Dev Comp Immunol 2008;32:875–82.
[21] Robalino J, Carnegie RB, O'Leary N, Owry-Patat SA, de la Vega E, Prior S, et al. Contributions of functional genomics and proteomics to the study of immune responses in the Pacific white leg shrimp *Litopenaeus vannamei*. Vet Immunol Immunopathol 2009;128:110–8.
[22] Wang W, Meng B, Chen W, Ge X, Liu S, Yu J. A proteomic study on postdiapaused embryonic development of brine shrimp (*Artemia franciscana*). Proteomics 2007;7:3580–91.
[23] Webster SG. Neurohormonal control of ecdysteroid biosynthesis by *Carcinus maenas in vitro*, and preliminary characterization of the putative molt-inhibiting hormone. Gen Comp Endocrinol 1986;61:237–47.
[24] Wainwright G. Hormonal control of ovarian development in the edible crab, *Cancer pagurus*. 1995 PhD Thesis, University of Liverpool, UK.
[25] Wainwright G, Prescott MC, Webster SG, Rees HH. Mass spectrometric determination of methyl farnesoate profiles and correlation with ovarian development in the edible crab, *Cancer pagurus*. J Mass Spectrom 1996;31:1338–44.
[26] Heukshoven J, Dernick R. Simplified method for silver staining of proteins in polyacrylamide gels and the mechanisms of silver staining. Electrophoresis 1985;6:103–12.
[27] Neuhoff V, Arold N, Taube D, Ehrhardt W. Improved staining of proteins in polyacrylamide gels including isoelectric-focusing gels with clear background at nanogram sensitivity using Coomassie brilliant Blue G-250 and R-250. Electrophoresis 1988;9:255–62.
[28] Wang W, Vignani R, Scali M, Sensi E, Tiberi P, Cresti M. Removal of lipid contaminants by organic solvents from oilseed protein extract prior to electrophoresis. Anal Biochem 2004;329:139–41.
[29] Khamnamtong B, Klinbunga S, Menasveta P. Genetic diversity and geographic differentiation of the giant tiger shrimp (*Penaeus monodon*) in Thailand analyzed by mitochondrial COI sequences. Biochem Genet 2009;47:42–55.
[30] Jiang H, Li F, Xie YS, Huang B, Zhang J, Zhang J, et al. Comparative proteomic profiles of the hepatopancreas in Fenneropenaeus chinensis response to hypoxic stress. Proteomics 2009;9:3353–67.
[31] Brouwer M, Brouwer H, Grater W, Brown-Paterson N. Replacement of a cytosolic copper/zinc superoxide dismutase by a novel cytosolic manganese superoxide dismutase in crustaceans that use copper (haemocyanin) for oxygen transport. Biochem J 2003;374:219–28.

[32] Terwilliger NB, Ryan M, Phillips MR. Crustacean hemocyanin gene family and microarrays studies of expression change during eco-physiological stress. Integr Comp Biol 2006;46:991–9.
[33] Söderhäll K, Cerenius L. Role of prophenoloxidase-activating system in invertebrate immunity. Curr Opin Immunol 1998;10:23–8.
[34] Sugumaran M. Unified mechanism for sclerotization of insect cuticle. Adv Insect Physiol 1998;27:220–334.
[35] Terwilliger NB. Hemolymph proteins and molting in crustaceans and insects. Am Zool 1999;39:589–99.
[36] Terwilliger NB, Ryan M, Towle D. Evolution of novel functions: cryptocyanin helps build new exoskeleton in *Cancer magister*. J Exp Biol 2005;208:2467–74.
[37] Van Holde KE, Miller KI. Hemocyanins. Adv Protein Chem 1995;47:1–81.
[38] Terwilliger NB, Dangott LD, Ryan MC. Cryptocyanin, a crustacean molting protein: evolutionary link with arthropod hemocyanins and insect hexamerins. Proc Natl Acad Sci USA 1999;96:2013–8.
[39] Fielder DK, Rao KR, Fingerman M. A female-limited lipoprotein and the diversity of hemocyanin components in the dimorphic variant of the fiddler crab *Uca pugilator*, as revealed by disc electrophoresis. Comp Biochem Physiol 1971;39B:291–7.
[40] Gilchrist RA, Lee WL. Carotenoid pigments and their possible role in reproduction in the sand crab *Emerita analoga*. Comp Biochem Physiol 1972;42B:263–94.
[41] Durliat M. Occurrence of plasma proteins in ovary and egg extracts from *Astacus leptodactylus*. Comp Biochem Physiol 1984;78B:745–53.
[42] Terwilliger NB, Dumler K. Ontogeny of decapod crustacean hemocyanin: effects of temperature and nutrition. J Exp Biol 2001;204:1013–20.
[43] Tsukimura B, Kamemoto FI. In vitro stimulation of oocytes by the presumptive mandibular organ secretions in the shrimp, *Penaeus vannamei*. Aquaculture 1991;92:59–66.