Songklanakarin J. Sci. Technol.
40 (5), 1017-1025, Sep. - Oct. 2018

http://www.sjst.psu.ac.th

Original Article

In silico analysis of cysteine cathepsins identified from the transcriptome profile of blunt snout bream (*Megalobrama amblycephala*)

Ngoc Tuan Tran[1, 2, 3] and Wei-Min Wang[1, 2*]

[1] *College of Fisheries, Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction, Ministry of Education, Huazhong Agricultural University, Wuhan, Hubei, 430070 China*

[2] *Key Lab of Freshwater Animal Breeding, Ministry of Agriculture, Huazhong Agricultural University, Wuhan, Hubei, 430070 China*

[3] *Center for Fish Biology and Fishery Biotechnology, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, Hubei, 430072 China*

Received: 1 March 2016; Revised: 5 June 2017; Accepted: 7 June 2017

Abstract

Cysteine cathepsins are described as lysosomal proteases with housekeeping and highly specialised functions in many organisms. In this study, cysteine cathepsins were identified from the transcriptome profile of blunt snout bream (*Megalobrama amblycephala*). A total of 41 cysteine cathepsin-like sequences were found and divided into groups of cathepsin (Cts): B, C, F, H, K, L, S, and Z. Twenty-nine of these sequences contained coding sequences, encoding 92-473 amino acids, which exhibited the highest (78%-95%) homology to the counterparts from zebrafish (*Danio rerio*). Multiple sequence alignment of the amino acids showed highly conserved domains among the cysteine cathepsins of *M. amblycephala* to its homologs. The putative proteins, including maCtsb1, maCtsc1, maCtsf2, maCtsh1, maCtsk5, maCtsl5, maCtss1, and maCtsz5 were characterised and structured using *in silico* methods. Additionally, codon usage of full length open reading frame sequences of these cathepsins was analysed. This study provided basic information for further studies on the specific functions of cysteine cathepsins of *M. amblycephala* in the future.

Keywords: cysteine cathepsin, blunt snout bream, *in silico*, homology modelling, codon usage

1. Introduction

The term “cathepsin”, that is derived from the Greek word “kathepsein” for “digesting”, describes the acidic proteases in the stomach mucosa (Brömme & Wilson, 2011; Turk *et al.*, 2012). Cathepsins were reported to involve the innate and adaptive immunity of organisms, cell death regulation, lysosome-mediated apoptosis, toll-like receptor signalling, activating or inhibiting certain cytokines, activation of granule serine proteases, and antigen processing and presentation (Conus & Simon, 2010). In humans, cathepsins include three different mechanistic classes: serine proteases (including cathepsins A and G), aspartic proteases (cathepsins D and E), and cysteine cathepsins (cathepsins B, C, F, H, K, L, O, S, V, X and W) (Dickinson, 2002; Rossi *et al.*, 2004; Brömme and Wilson, 2011; Turk *et al.*, 2012). Among these available cathepsins, cysteine cathepsins have been reported as lysosomal proteases with housekeeping and highly specialised functions (Brömme & Wilson, 2011). They belong to the clan CA of cysteine peptidases (Turk *et al.*, 2012) and most of them are members of the papain family cysteine

*Corresponding author
Email address: wangwm@mail.hzau.edu.cn

protease and C1 peptidase family (Colbert *et al.,* 2009). Initially, cysteine cathepsins were functionally characterised as intracellular enzymes responsible for the non-specific and bulk proteolysis in endosomal/lysosomal compartment (Turk *et al.,* 2012). Moreover, these proteins have been introduced as targets for the development and application of drug and vaccination against infections (e.g., parasitic infections) (Chen & Sun, 2012; Maldonado-Aguayo *et al.,* 2015).

Multiple types of cathepsin (B, D, F, H, K, L, S, and X) have been previously identified and studied in fish species: *Paralichthys olivaceus* (Cha *et al.,* 2012; Park *et al.,* 2009; Zhang *et al.,* 2008), *Cynoglossus semilaevis* (Chen & Sun, 2012), *Scophthalmus maximus* (Jia & Zhang, 2009; Zhou *et al.,* 2015), *Pseudosciaena crocea* (Li *et al.,* 2014), *Sparus aurata* (Carnevali *et al.,* 1999), *Carassius auratus* (Harikrish nan *et al.,* 2010), *Ictalurus punctatus* (Feng *et al.,* 2011; Yeh & Klesius, 2009), and *Lates calcarifer* (Azizan *et al.,* 2014). These studies confirmed that cathepsins are important in the physiological processes of a developing embryo and innate immunity against pathogens of fish, as well as in the immune-related activities of crustacean (Garcia-Carreño *et al.,* 2014) and mussel (*Cristaria plicata*) (Hu *et al.,* 2014).

Blunt snout bream (BSB) (*Megalobrama amblycephala*) is one of the economically significant aquaculture species in China. Many studies associated with the molecular techniques and analysis of transcriptome and microRNAs have been performed to contribute to the information on genetic improvement of this fish species (Gao *et al.,* 2012; Tran *et al.,* 2015; Yi *et al.,* 2013). Several important growth- and immune-related genes have been cloned and sequenced in BSB (Ding *et al.,* 2013; Huang *et al.,* 2014; Li *et al.,* 2013; Liu *et al.,* 2014; Luo *et al.,* 2014; Tian *et al.,* 2012; Tran *et al.,* 2015). However, cysteine cathepsins have not been studied in BSB so far. In this study, the sequences of cysteine cathepsins were identified from the previously obtained transcriptome profile of BSB (Tran *et al.,* 2015). The entire coding regions of the proteins were then characterised and structured using *in silico* methods. Codon usage based on the full length open reading frame (ORF) sequences was also analysed.

2. Materials and Methods

2.1. Identification of cysteine cathepsins

Sequences of cysteine cathepsin genes obtained from the BSB transcriptome profile (Accession number: SRX731259) (Tran *et al.,* 2015) were used for the analysis in this study. The genes were identified by a BLAST homology search against the GenBank database (http://blast.ncbi.nlm.nih.gov/Blast). The putative amino acid sequences were predicted by performing the NCBI's Open Reading Frame Finder (http://www.ncbi.nlm.nih.gov/gorf/orfig.cgi), signal peptide employing the SignalP server (http://www.cbs.dtu.dk), and domain structures applying SMART program (http://smart.embl-heidelberg.de/). Multiple-sequence alignment of homologous cathepsin amino acid sequences from species was performed using ClustalW2 (http://www.ebi.ac.uk/ Tools/msa/clustalw2/). Neighbour-joining phylogenetic tree with 1000 bootstrap replications was constructed by MEGA 6.0 (Tamura *et al.,* 2013).

2.2. Protein structure prediction

The secondary structure of the proteins was predicted by the SOPMA server (https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html). Three-dimensional (3-D) models were constructed by the SWISS-MODEL (http://swissmodel.expasy.org/). The suitable templates used to predict the 3-D models were selected based on sequence identity (Fiser, 2010). The stereochemical quality and accuracy of the models were tested based on an analysis of a Ramachandran plot, Verify3D (http://services.mbi.ucla.edu/SAVES/), ProQ (http://www.sbc.su.se/~bjornw/ProQ/ProQ.html), and ProSA (https://prosa.services.came.sbg.ac.at/prosa.php).

2.3. Protein physicochemical and functional characterisation

Physicochemical properties of polypeptide chains were determined using ProtParam (http://web.expasy.org/protparam/). The types of protein were predicted using SOSUI (http://harrier.nagahama-i-bio.ac.jp/sosui/) and the presence of disulphide bonds in the proteins employed CYS_REC (http://linux1.softberry.com/).

2.4. Codon usage analysis

The genes and proteins of cysteine cathepsins selected from the RNA-Seq database were assembled and annotated as in Tran *et al.* (2015). ORFs were determined by NCBI's Open Reading Frame Finder (http://www.ncbi.nlm.nih.gov/gorf/orfig.cgi). Full lengths of the coding sequences were identified based on the start codon (AUG) and stop codons (UAA, UAG or UGA). To reduce errors, low quality sequences with ≤100 codons or sequences that contained internal stop codons were excluded (Yu *et al.,* 2012). Relative synonymous codon usage (RSCU) was calculated using CodonW 1.4.2 (http://codonw.sourceforge.net), following the formula reported by Sharp and Li, (1986). The codons with RSCU >1.0 were indicated as high frequency. GCi is defined as the fraction of cytosines (C) and guanines (G) in the "I" position of the codon, expressed as GCi = 3(Ci+Gi)/L for the ORF of length L (Duan *et al.,* 2015). The GCi index was calculated using Microsoft Excel 2010.

3. Results and Discussion

3.1. Identification of cysteine cathepsins from BSB

A total of 41 cysteine cathepsin-like sequences were determined from the transcriptome profile. These sequences were divided into groups of cathepsin (Cts): B (consisting of 10 sequences), C (2), F (3), H (2), K (5), L (13), S (1), and Z (5). Twenty-nine of these sequences contained coding sequences and encoding proteins (at least one domain in each protein) of between 92 and 473 amino acids. The BLASTX analysis indicated that the cysteine cathepsins in BSB were 58%-97% identical to their homologs in other fish species, the highest (78%-95%) similarity to the cathepsins of zebrafish

(63.4%) (data not shown). Most of the identified sequences that encoded putative proteins consisted of the expected structure that included a signal peptide and a cathepsin specific domain. *N*-terminal signal peptide and *N*-linked glycosylation sites were found in the putative proteins. The SMART analysis revealed that maCtsb, maCtsc, maCtsf, maCtsh, maCtsk, maCtsl, maCtsz consisted of the Papain family cysteine protease (Pept_C1) in their structures. Additionally, maCtsf, maCtsh, maCtsk, and maCtsl contained an inhibition motif (Inhibitor I29). Only maCtsf had cystatin-like (CY) domain. No other domain was found in the sequence of maCtss. These results were consistent with Brömme and Wilson (2011) who reported that all papain-like cysteine proteases contained a signal peptide, a propeptide, and a catalytic domain in their molecules.

Multiple sequence alignment revealed that the amino acids in cysteine cathepsins of BSB were highly conserved. For example, the sequences of maCtsl5 (Figure 1) had identical sequences in most of the regions and matched with its homologs in grass carp, common carp, and zebra fish that showed identities of 97%, 95%, and 93%, respectively, which suggested similar functions in these proteins. The motifs "GCNGG" and "CGSCWAFS" were found in the sequence of maCtsl5 and conserved with those in other fishes. The fact that "GCNGG" had a cysteine residue that may form a disulphuric bridge in the protein structure indicated its structural functions. The "CGSCWAFS" resembled the active cysteine site and the catalytic triad CHN (Maldonado-Aguayo *et al.*, 2015). Additionally, the start amino acids of the cysteine cathepsins in BSB were similar to their counterparts from other fishes, which indicated that the cysteine cathepsins of BSB were identified.

3.2. Phylogenetic analysis

A condensed phylogenetic tree was constructed based on the amino acid sequences of cysteine cathepsins of BSB with representative cathepsins in fishes. The overall topology of the tree displayed different cysteine cathepsins formed distinct paraphyletic clusters, differentiated by colours (Figure 2). The cysteine cathepsins of BSB have a close evolutionary relationship with their homologs of the cyprinids, especially with zebrafish. This suggests highly similar functions with their counterparts from different fishes.

3.3. Protein structure prediction and model validation

The secondary structure predicted that random coil dominated, followed by alpha helix, extended strand, and beta turn in all proteins, except maCtsh1 with alpha helix was predominant. The remaining structural elements of proteins were not predicted (Table 1). The 3-D structures of proteins were modelled based on the available templates based on the sequence structural identity (15.6-80.0%, Table 2 and Figure 3). The quality of predicted models was evaluated using Ramachandran plots, Verify3D, ProQ, and ProSA (Table 2). Ramachandran plot analysis revealed 78.9%-86.9% of residues were in the most favoured regions, while 9.6%-18.5% in the additional favoured regions for all models. The overall average G-factor arranged from -0.24 to 0.08 (cut-off value: -0.5), suggested the acceptable quality for the predicted models (Ramachandran *et al.*, 1963). ProQ values were high for the models of maCtsb1 (LGscore: 3.666; MaxSub: 0.345),

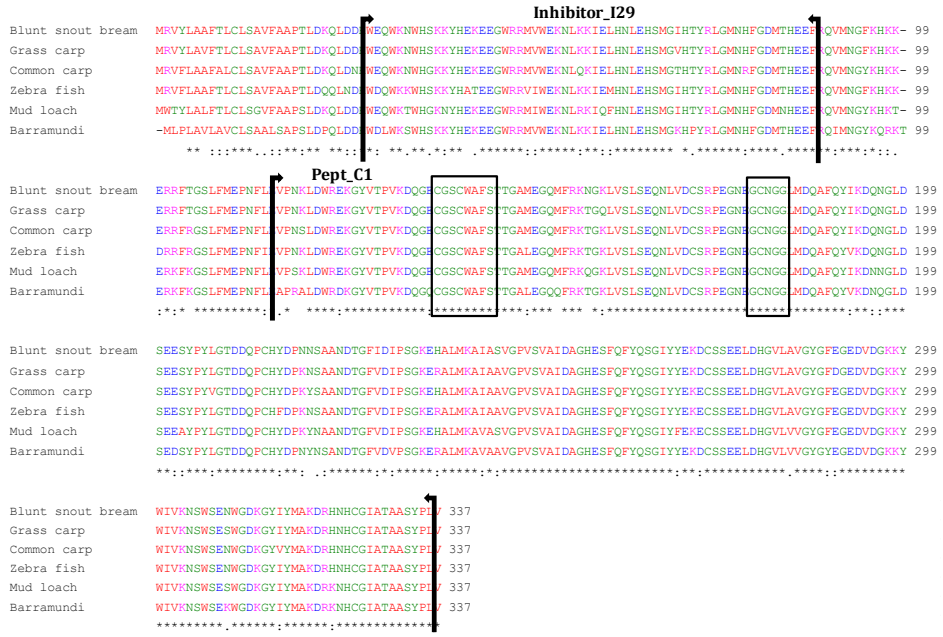

Figure 1. Multiple sequence alignment of maCtsl5 with its homologs in fish species: *Ctenopharyngodon idella* (GenBank accession number: AJY58107.1), *Cyprinus carpio* (BAD08618.1), *Danio rerio* (NP_997749.1), *Misgurnus mizolepis* (ABQ08058.1), and *Lates calcarifer* (ABV59078.1). Asterisk marks (*) indicates identical amino acids. Sequences are numbered on the right, conserved substitutions are indicated by (:), semi-conserved by (.) and deletions by dashes. Inhibitor_I29 and Pept_C1 domains are indicated by arrows. The two motifs "GCNGG" and "CGSCWAFS" are boxed.

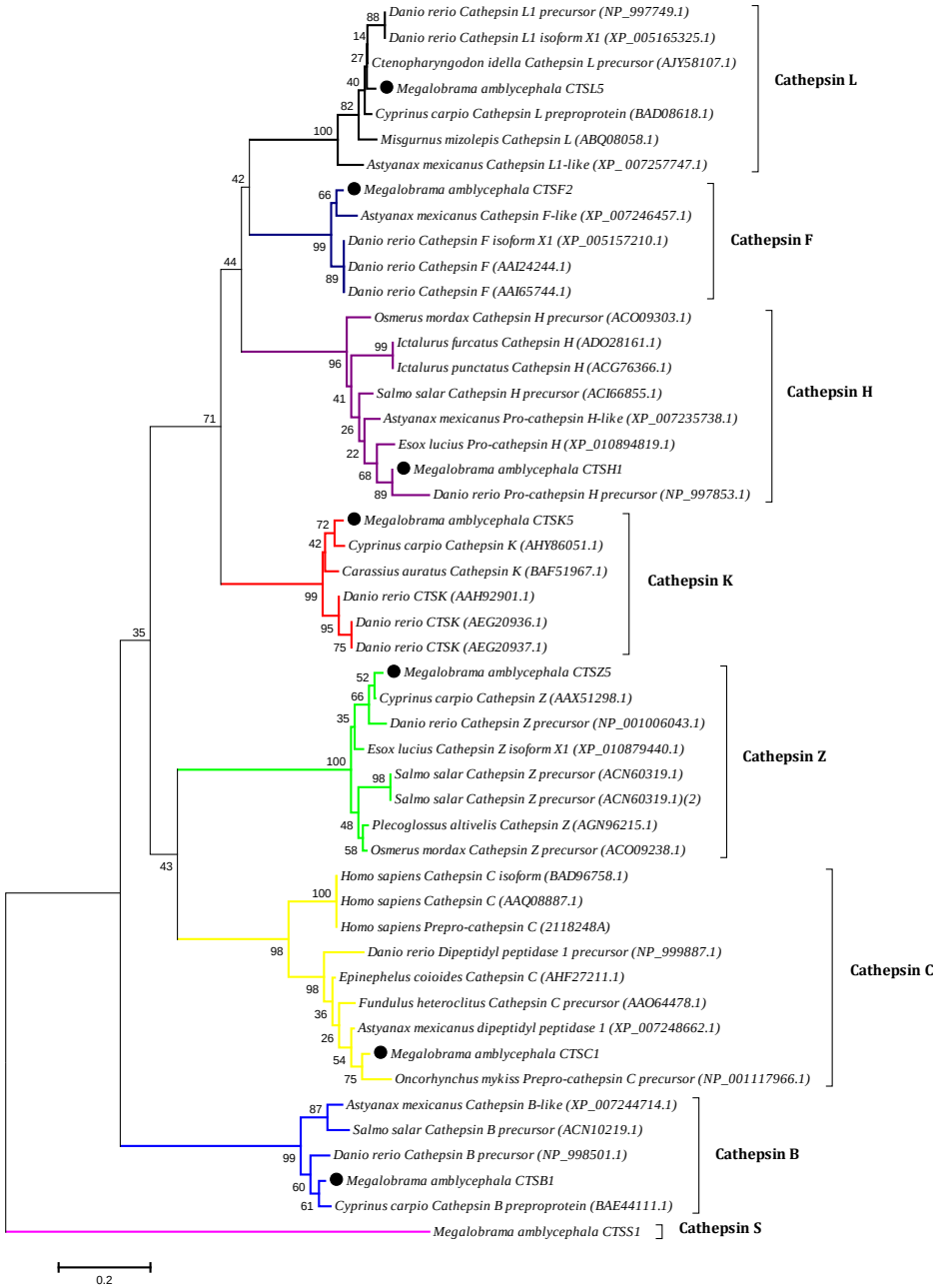

Figure 2. Neighbour-joining phylogenetic tree constructed based on the sequences of cysteine cathepsins of BSB (black dot) and their homologs in selected animals (GenBank accession numbers in the parentheses). The numbers at the branches indicate 1000 replication bootstrap and the bar (0.1) indicates the genetic distance.

maCts1 (2.246; 0.154), maCtsf2 (4.129; 0.309), maCtsk5 (4.804; 0.359), maCtsl5 (4.807; 0.312), and maCtsz5 (4.498; 0.396), while a little low for maCtsh1 (0.789; 0.101) and maCtss1 (0.410; 0.049). This indicated a very good quality of the former six models, fairly good quality of the model for maCtsh1, and bad quality for maCtss1 based on the recommendation of Wallner and Elofsson (2003). Z-Scores (ProSA) of all models were within a plot of typical scores for all available proteins in the PDB database (Figure 3 *Z) and values of single residue energies (window 40) were mostly negative (Figure 3 *E) (Wiederstein and Sippl, 2007). Additionally, Verify3D analysis revealed 50%-100% of the residues had an average 3D-1D score ≥0.2, and 87.4%-100% had positive scores (cut-off score is zero), suggesting these models were valid (Liithy *et al.*, 1992). The combined results indicated that the predicted models for maCtsb1, maCts1, maCtsf2, maCtsk5, maCtsl5, maCtsz5, and maCtsh1 were good and reliable, while the model for maCtss1 was not plausible for use possibly due to its shorter putative amino acid sequence.

Table 1. Calculated secondary structure elements using SOPMA.

Element	MaCtsb1	MaCtsc1	MaCtsf2	MaCtsh1	MaCtsk5	MaCtsl5	MaCtss1	MaCtsz5
Alpha helix	23.33	26.5	37.21	36.16	30.03	29.67	16.67	17.94
3_{10} helix	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Pi helix	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Beta bridge	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Extended strand	19.09	24.84	17.55	16.52	27.63	19.58	33.33	23.92
Beta turn	15.76	10.99	10.99	12.05	11.71	10.98	16.67	12.29
Bend region	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Random coil	41.82	38.02	34.25	35.27	30.63	39.76	33.33	45.85
Ambiguous states	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Other states	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 2. Homology modelling of three dimensional models for proteins using the SWISS-MODEL.

Parameter	MaCtsb1	MaCtsc1	MaCtsf2	MaCtsh1	MaCtsk5	MaCtsl5	MaCtss1	MaCtsz5
Template (PDB ID)	1qdq.1.A	1jqp.1.A	1cs8.1.A	2o6x.1.A	7pck.1.A	1cjl.1.A	4d9v.1.A	1deu.1.A
Residue range	Leu79-Pro329	Asp19-Leu455	Glu170-Val472	Leu29-Lys223	Leu26-Met333	Asp22-Val337	Ser5-Val36	Glu23-Leu296
Resolution (A)	2.18	2.4	1.8	1.4	3.2	2.2	2.5	1.7
Sequence identity (%)	80	62.7	36.2	46.2	64.7	68.0	15.6	70.6
Procheck								
Total number of residues	251	437	303	195	308	316	32	274
Residues in most favoured regions (%)	82.8	86.9	78.9	84.8	80.6	83.9	85.2	82.4
Residues in additional allowed regions (%)	15.7	9.6	18.5	12.3	17.5	14.7	11.1	17.6
Residues in generously allowed regions (%)	0.5	2.9	1.5	1.8	1.1	0.7	3.7	0
Residues in disallowed regions (%)	1	0.5	1.1	1.2	0.8	0.7	0	0
G-factors	-0.01	-0.24	0	-0.02	0.06	0.08	-0.16	0.02
ProQ								
Lgscore	3.666	2.246	4.129	0.789	4.804	4.807	0.41	4.498
MaxSub	0.345	0.154	0.309	0.101	0.359	0.312	0.049	0.396
ProSA (Z-Score)	-7.26	-7.35	-6.66	-4.01	-8.71	-8.64	0.06	-6.33
Verify3D								
3D-1D score ≥0.2 (%)	100	75.5	89.1	67.2	86.7	82.0	50	90.15
Residues with positive scores (%)	100	87.4	96.7	100	98.7	97.2	100	99.6

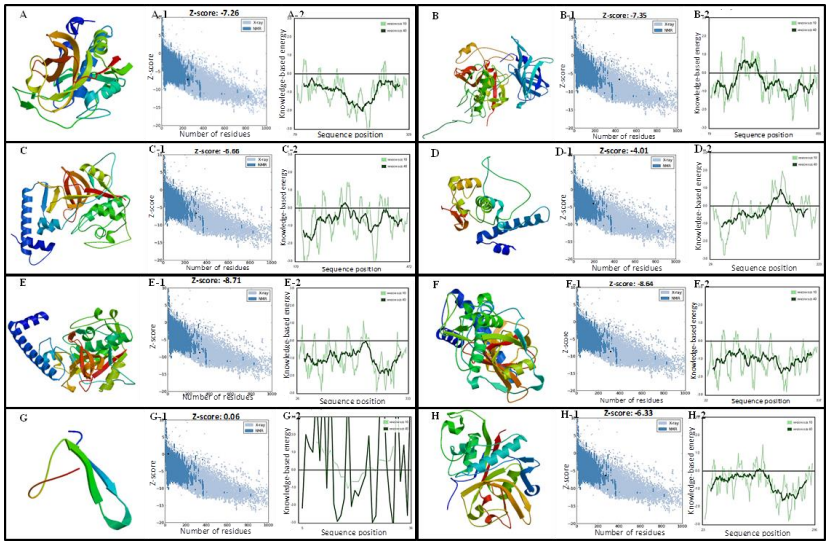

Figure 3. Three-dimensional structures of cysteine cathepsins generated by SWISS-MODEL: A) maCtsb1 B) maCtsc1 C) maCtsf2 D) maCtsh1 E) maCtsk5, F) maCtsl5, G) maCtss1, and H) maCtsz5. Results of validation for each model performed by ProSA server: *-Z: Z-score analyses (black dot) of the predicted models. *-E: Plots of single residue energies for each models, where window sizes of 40 and 10 residues are distinguished by dark- and light-green lines, respectively. Positive values indicate problematic or erroneous parts of the structure.

3.4. Physicochemical and functional characterisation

The physicochemical characterisation of BSB's cathepsins was analysed using the Protparam tool (Table 3). The theoretical isoelectric point (pI) of the proteins ranged from 5.52 to 6.53 (pI <7), which indicated that these proteins were acidic in character. The results of the extinction coefficient (EC) of proteins, except maCtss1, measured at 280 nm were from 35,130 to 92,540 M^{-1} cm^{-1} and 34,380 to 91,790 M^{-1} cm^{-1}, when it is assumed all pairs of cysteine residues form cysteines and are reduced, respectively. Basically, the high EC values indicate the high proportion of cysteine, tryptophan, and tyrosine in the proteins. The EC of maCtss1 measured at 280 nm was 1490 M^{-1} cm^{-1}, which was due to the lack of tryptophan residues in its sequence. The instability index (II) value revealed that, except for maCtsf2 (II=40.74) and maCtss1 (II=61.83), all proteins were stable (II=28.86~ 37.73) (II <40, Guruprasad *et al.* (1990)). This study resulted in high aliphatic index (AI) values, that ranged from 58.75 to 86.48 and implied high thermostability of these proteins (Ikai, 1980). Low grand average hydropathicity (GRAVY) values indicated the hydrophilic character of the proteins in natural conditions.

Most proteins were rich in leucine (5.3-8.2%), serine (5.4-18.5%), and glycine (7.1-11.3%), while pyrrolysine and selenocysteine were absent (data not shown). Cysteine, phenylalanine, and tryptophan lacked in maCtss1, which suggested that maCtss1 was possibly identified as a partial protein. The maCtsb1, maCtsc1, maCtsh1, maCtss1, and maCtsz5 were classified as soluble proteins, the others as membrane proteins (Table 4). The positions of the cysteines were predicted to form disulphide bonds of all proteins. The most probable pattern of cysteine pairing was predicted (Table 5). The pattern of pairs of cysteines found in all proteins indicated the presence of disulphide bonds in the proteins, which is crucial in folding and stabilising protein structures (Hogg, 2003).

3.5. Codon usage in cysteine cathepsins of BSB

Codon usage analysis in the cysteine cathepsins was based on 11 full length ORF sequences after filtering and removing low quality sequences from 41 transcripts assembled and annotated by Tran *et al.* (2015). The overall codon usage was calculated from 3,594 codons of 11 full length ORF sequences (Figure 4). The results found that, except for the three stop codons, GCG (RSCU: 1.37) presented as the most frequent (occurrence: 39.0‰ and average frequency: 2.5 times), followed by AAC (1.26), and AUG (1.00) (with an occurrence of >30‰). These codons were named as high-frequency codons. However, UGA (0.55) was the lowest frequency codon (0.56‰) and another 17 codons also had low frequency (<10‰). The codons: AAC (1.26, encoding asparagine) and GCG (1.37, glutamic acid) were used more frequently than the synonymous codons for corresponding amino acids (1.7 and 2.2 times, respectively). The UAA stop codon was used most frequently (1.39), followed by UAG (1.09), and UGA (0.55). The CUG (2.40) and UUA (0.22) codons, that encode leucine, were the highest and lowest RSCU values among the 64 calculated codons, respectively.

Table 3. Parameters computed using the ProtParam tool.

Protein	No. of aa	Mol. Wt. (Da)	pI	-R	+R	EC*	II	AI	GRAVY
MaCtsb1	330	36358.9	5.56	35	26	82360-81360	33.32	66.79	-0.312
MaCtsc1	455	51030.7	6.31	46	42	90145-89270	35.69	70.66	-0.274
MaCtsf2	473	53068.6	6.43	54	52	89560-88810	40.74	74.42	-0.357
MaCtsh1	224	25518.9	6.35	22	19	35130-34380	36.75	62.72	-0.443
MaCtsk5	333	37014	5.92	37	33	75790-75290	28.86	72.88	-0.346
MaCtsl5	337	38522	5.52	50	36	79340-78840	31.42	58.75	-0.697
MaCtss1	54	5716.3	6.53	7	7	1490	61.83	86.48	-0.363
MaCtsz5	301	33433.4	5.84	29	26	92540-91790	37.73	65.08	-0.439

**First value is based on the assumption both cysteine form cysteines and the second assumes that both cysteine residues are reduced*

Table 4. Transmembrane regions identified using the SOSUI server.

Protein	Types of protein	Type of transmembrane	Length	Position	Transmembrane region
MaCtsb1	Soluble	-	-	-	-
MaCtsc1	Soluble	-	-	-	-
MaCtsf2	Membrane	Primary	23	5-27	RGYPVIACYVLVALVLGIEDGPD
MaCtsh1	Soluble	-	-	-	-
MaCtsk5	Membrane	Primary	23	1-23	MYTFGGIPLIVVVWCGLAHTLDN
MaCtsl5	Membrane	Primary	24	1-24	MRVYLAAFTLCLSAVFAAPTLDK
MaCtss1	Soluble	-	-	-	-
MaCtsz5	Soluble	-	-	-	-

Table 5. Disulphide (SS) bond pattern of pairs predicted using CYS_REC

Protein	CYS_REC	Protein	CYS_REC
MaCtsb1	Cys92-Cys121	MaCtsh1	Cys97-Cys215
	Cys104-Cys148		Cys133-Cys176
	Cys140-Cys206		Cys136-Cys219
	Cys141-Cys144		Cys144-Cys212
	Cys177-Cys210		Cys167-Cys213
	Cys185-Cys196	MaCtsk5	Cys139-Cys180
MaCtsc1	Cys24-Cys112		Cys173-Cys213
	Cys48-Cys130		Cys272-Cys322
	Cys248-Cys291	MaCtsl5	Cys137-Cys180
	Cys284-Cys323		Cys170-Cys214
	Cys313-Cys329		Cys273-Cys326
MaCtsf2	Cys97-Cys461	MaCtsz5	Cys27-Cys165
	Cys109-Cys123		Cys81-Cys124
	Cys281-Cys322		Cys118-Cys156
	Cys315-Cys355		Cys146-Cys162
			Cys171-Cys206

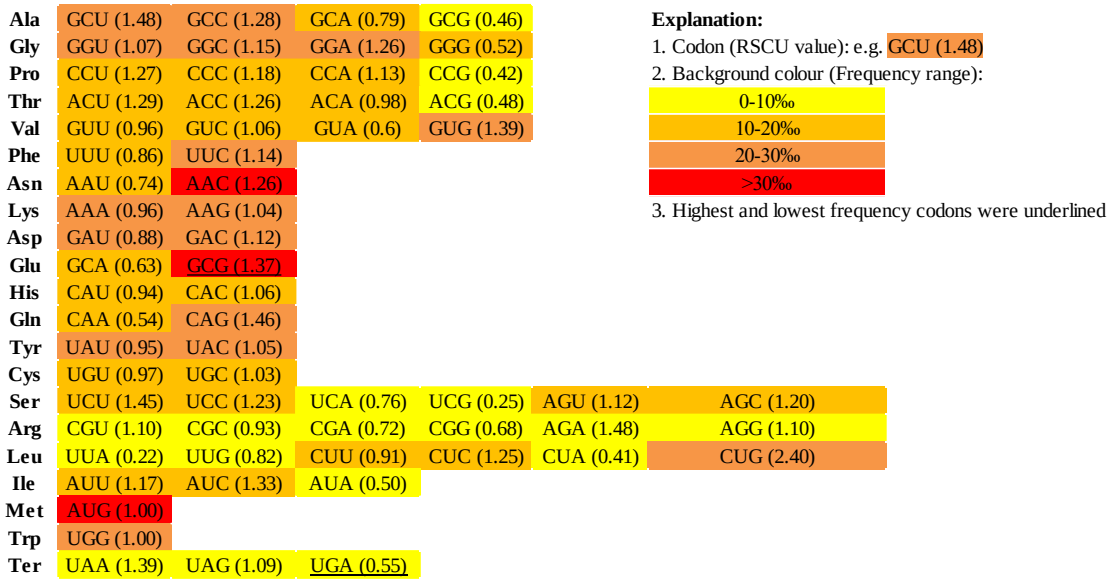

Figure 4. Unequal use of 64 codons in the cysteine cathepsins of BSB. The codons for the same amino acids are listed on the left and are coloured yellow, orange yellow, orange, and red to show the occurrence frequencies of 0.0‰-10.0‰, 10.0‰-20.0‰, 20.0‰-30.0‰, and >30.0‰, respectively. The data are shown as a triplet codon and the highest and lowest frequency codons are underlined.

Figure 4 shows four NUA codons: GUA (encoding valine), UUA (leucine), CUA (leucine), and AUA (isoleucine) had quite low RSCU values (0.6, 0.22, 0.41, and 0.50, respectively), suggesting the increase of protein production by inhibition of mRNA degradation (Al-Saif and Khabar, 2012; Feng *et al.*, 2013; Duan *et al.*, 2015). Four NCG codons: GCG (alanine), CCG (proline), ACG (threonine), and UCG (serine) also showed low RSCU values (0.46, 0.42, 0.48 and 0.25, respectively). This may facilitate the prevention of mutations caused by DNA methylation. Methylated cytosine in the CG dinucleotide is more easily deaminated into thymine, which leads to the G in the 3rd codon position and tends to be wobbly. The species with a high DNA methylation level tends to avoid NCG codons to produce mutations (Gonzalez-Ibeas *et al.*, 2007; Sterky *et al.*, 2004). The low RSCU values of NCG codons indicate relatively high methylation in these cathepsins. The overall GC content was calculated as 0.502, which was different in different codon positions, the highest (0.562) in GC_3 (GC content of 3rd nucleotide in codons), lowest (0.419) in GC_2 (GC content of 2nd nucleotide in codons), and intermediate (0.524) in GC_1 (GC content of 1st nucleotide in codons). This was consistent with the observation in previous analysis from RNA-Seq database of BSB (Duan *et al.*, 2015).

4. Conclusions

The initial results of this study identified the presence of eight cysteine cathepsin-like sequences in BSB. These proteins had a closet phylogenetic relationship with cyprinids rather than other species. This study provides basic knowledge that is useful for future studies on the specific functions of these proteins in BSB. Additionally, this provides a reference database to identify the cysteine proteases in different aquatic animals.

Acknowledgements

The first author Tran Ngoc Tuan would like to thank the China Scholarship Council for providing the scholarship to attend the doctoral program in Huazhong Agricultural University, Wuhan, Hubei, China.

References

Al-Saif, M., & Khabar, K. S. (2012). UU/UA dinucleotide frequency reduction in coding regions results in increased mRNA stability and protein expression. *Molecular Therapy, 20*(5), 954-959.

Azizan, S., Wan, K.-L., & Mohd-Adnan, A. (2014). Molecular characterisation and expression analysis of cathepsin D from the Asian seabass *Lates calcarifer*. *Sains Malaysiana, 43*(8), 1139-1148.

Brömme, D., & Wilson, S. (2011). Role of cysteine cathepsins in extracellular proteolysis. In W. C. Parks & R. P. Mecham (Eds.), *Extracellular matrix degradation* (pp. 23-51). New York, NY: Springer-Verlag Berlin Heidelberg.

Carnevali, O., Centonze, F., Brooks, S., Marota, I., & Sumpter, J. (1999). Molecular cloning and expression of ovarian cathepsin D in seabream, *Sparus aurata*. *Biological of Reproduction, 61*(3), 785-791.

Cha, I. S., Kwon, J., Mun, J. Y., Park, S. B., Jang, H. B., Nho, S. W., . . . Jung, T. S. (2012). Cathepsins in the kidney of olive flounder, *Paralichthys olivaceus*, and their responses to bacterial infection. *Developmental and Comparative Immunology, 38*(4), 538-544.

Chen, L., & Sun, L. (2012). Cathepsin B of *Cynoglossus semilaevis*: Identification, expression, and activity analysis. *Comparative Biochemistry and Physiology - Part B: Biochemistry and Molecular Biology, 161*(1), 54-59.

Colbert, J. D., Matthews, S. P., Miller, G., & Watts, C. (2009). Diverse regulatory roles for lysosomal proteases in the immune response. *European Journal of Immunology, 39*(11), 2955-2965.

Conus, S., & Simon, H.-U. (2010). Cathepsins and their involvement in immune responses. *Swiss Medical Weekly, 140*(w13042).

Dickinson, D. (2002). Cysteine peptidases of mammals: their biological roles and potential effects in the oral cavity and other tissues in health and disease. *Critical Reviews in Oral Biology and Medicine, 13*(3), 238-275.

Ding, Z., Wu, J., Su, L., Zhou, F., Zhao, X., Deng, W., . . . Liu, H. (2013). Expression of heat shock protein 90 genes during early development and infection in *Megalobrama amblycephala* and evidence for adaptive evolution in teleost. *Developmental and Comparative Immunology, 41*(4), 683-693.

Duan, X., Yi, S., Guo, X., & Wang, W. (2015). A comprehensive analysis of codon usage patterns in blunt snout bream (*Megalobrama amblycephala*) based on RNA-Seq data. *International Journal of Molecular Sciences, 16*(6), 11996-12013.

Feng, C., Xu, C.-J., Wang, Y., Liu, W.-L., Yin, X.-R., Li, X., . . . Chen, K.-S. (2013). Codon usage patterns in Chinese bayberry (*Myrica rubra*) based on RNA-Seq data. *BMC Genomics, 14*(1), 732.

Feng, T., Zhang, H., Liu, H., Zhou, Z., Niu, D., Wong, L., . . . Liu, Z. (2011). Molecular characterization and expression analysis of the channel catfish cathepsin D genes. *Fish and Shellfish Immunology, 31*(1), 164-169.

Fiser A. (2010) Template-based protein structure modeling. In D. Fenyö (Ed.), *Computational Biology. Methods in Molecular Biology (Methods and Protocols, Vol. 673, pp. 17-43*). Totowa, NJ: Humana Press.

Gao, Z., Luo, W., Liu, H., Zeng, C., Liu, X., Yi, S., & Wang, W. (2012). Transcriptome analysis and SSR/SNP markers information of the blunt snout bream (*Megalobrama amblycephala*). *PloS One, 7*(8), e42637.

Garcia-Carreño, F., Navarrete del Toro, M., & Muhlia-Almazan, A. (2014). The role of lysosomal cysteine proteases in crustacean immune response. *Invertebrate Survival Journal, 11*, 109-118.

Gonzalez-Ibeas, D., Blanca, J., Roig, C., González-To, M., Picó, B., Truniger, V., . . . Arús, P. (2007). MELOGEN: An EST database for melon functional genomics. *BMC Genomics, 8*(1), 306.

Guruprasad, K., Reddy, B. B., & Pandit, M. W. (1990). Correlation between stability of a protein and its dipeptide composition: A novel approach for predicting *in vivo* stability of a protein from its primary sequence. *Protein Engineering, 4*(2), 155-161.

Harikrishnan, R., Kim, M.-C., Kim, J.-S., Han, Y.-J., Jang, I.-S., Balasundaram, C., & Heo, M.-S. (2010). Immune response and expression analysis of cathepsin K in goldfish during *Aeromonas hydrophila* infection. *Fish Shellfish Immunol, 28*(4), 511-516.

Hogg, P. J. (2003). Disulfide bonds as switches for protein function. *Trends in Biochemical Sciences, 28*(4), 210-214.

Hu, X., Hu, X., Hu, B., Wen, C., Xie, Y., Wu, D., . . . Gao, Q. 2014. Molecular cloning and characterization of cathepsin L from freshwater mussel, *Cristaria plicata*. *Fish and Shellfish Immunology, 40*(2), 446-454.

Huang, C. X., Wei, X. L., Chen, N., Zhang, J., Chen, L. P., Wang, W. M., . . . Wang, H.L. (2014). Growth differentiation factor 9 of *Megalobrama amblycephala*: Molecular characterization and expression analysis during the development of early embryos and growing ovaries. *Fish Physiology and Biochemistry, 40*(1), 193-203.

Ikai, A. (1980). Thermostability and aliphatic index of globular proteins. *Journal of Biochemistry, 88*(6), 1895-1898.

Jia, A., & Zhang, X.-H. (2009). Molecular cloning, characterization and expression analysis of cathepsin D gene from turbot *Scophthalmus maximus*. *Fish and Shellfish Immunology, 26*(4), 606-613.

Li, M., Li, Q., Yang, Z., Hu, G., Li, T., Chen, X., & Ao, J. (2014). Identification of cathepsin B from large yellow croaker (*Pseudosciaena crocea*) and its role in the processing of MHC class II-associated invariant chain. *Developmental and Comparative Immunology, 45*(2), 313-320.

Li, S., Gul, Y., Wang, W., Qian, X., & Zhao, Y. (2013). PPARγ, an important gene related to lipid metabolism and immunity in *Megalobrama amblycephala*: Cloning, characterization and transcription analysis by GeNorm. *Gene, 512*(2), 321-330.

Liithy, R., Bowie, J., & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature, 356*(6364), 83-85.

Liu, X., Tang, X., Wang, L., Li, J., Wang, H., Wei, S., . . . Chen, N. (2014). Molecular cloning and expression analysis of mannose receptor in blunt snout bream (*Megalobrama amblycephala*). *Molecular Biology Reports, 41*(7), 4601-4611.

Luo, W., Zhang, J., Wen, J.-F., Liu, H., Wang, W.-M., & Gao, Z.-X. (2014). Molecular cloning and expression analysis of major histocompatibility complex class I, IIA and IIB genes of blunt snout bream (*Megalobrama amblycephala*). *Developmental and Comparative Immunology, 42*(2), 169-173.

Maldonado-Aguayo, W., Chávez-Mardones, J., Gonçalves, A. T., & Gallardo-Escárate, C. (2015). Cathepsin gene family reveals transcriptome patterns related to the infective stages of the salmon louse *Caligus rogercresseyi*. *PloS One, 10*(4),

Park, E.-M., Kim, Y.-O., Nam, B.-H., Kong, H. J., Kim, W.-J., Lee, S.-J., & Kim, K.-K. (2009). Cloning and expression analysis of cathepsin D in the olive flounder *Paralichthys olivaceus*. *Bioscience, Biotechnology, and Biochemistry, 73*(8), 1856-1859.

Ramachandran, G., Ramakrishnan, C., & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology, 7*(1), 95-99.

Rossi, A., Deveraux, Q., Turk, B., & Sali, A. (2004). Comprehensive search for cysteine cathepsins in the human genome. *Biological Chemistry, 385*(5), 363-372.

Sharp, P. M., & Li, W.-H. (1986). An evolutionary perspective on synonymous codon usage in unicellular organisms. *Journal of Molecular Evolution, 24*(1-2), 28-38.

Sterky, F., Bhalerao, R. R., Unneberg, P., Segerman, B., Nilsson, P., Brunner, A. M., . . . Strauss, S. H. (2004). A Populus EST resource for plant functional genomics. *Proceedings of the National Academy of Sciences of the United States of America, 101*(38), 13951-13956.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution, 30*(12), 2725-2729.

Tian, Y.-M., Chen, J., Tao, Y., Jiang, X.-Y., & Zou, S.-M. (2012). Molecular cloning and function analysis of insulin-like growth factor-binding protein 1a in blunt snout bream (*Megalobrama amblycephala*). *Zoological Research, 35*(4), 300-306.

Tran, N. T., Gao, Z.-X., Zhao, H.-H., Yi, S.-K., Chen, B.-X., Zhao, Y.-H., . . . Wang, W.-M. (2015). Transcriptome analysis and microsatellite discovery in the blunt snout bream (*Megalobrama amblycephala*) after challenge with *Aeromonas hydrophila*. *Fish and Shellfish Immunology, 45*(1), 72-82.

Tran, N. T., Liu, H., Jakovlić, I., & Wang, W.-M. (2015). Blunt snout bream (*Megalobrama amblycephala*) MyD88 and TRAF6: Characterisation, comparative homology modelling and expression. *International Journal of Molecular Sciences, 16*(4), 7077-7097.

Turk, V., Stoka, V., Vasiljeva, O., Renko, M., Sun, T., Turk, B., & Turk, D. (2012). Cysteine cathepsins: from structure, function and regulation to new frontiers. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 1824*(1), 68-88.

Wallner, B., & Elofsson, A. (2003). Can correct protein models be identified? *Protein Science,* 12(5), 1073-1086.

Wiederstein, M., & Sippl, M. J. (2007). ProSA-web: Interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research, 35*(Suppl. 2), W407-W410.

Yeh, H.-Y., & Klesius, P. H. (2009). Channel catfish, *Ictalurus punctatus*, cysteine proteinases: Cloning, characterisation and expression of cathepsin H and L. *Fish and Shellfish Immunology, 26*(2), 332-338.

Yi, S., Gao, Z.-X., Zhao, H., Zeng, C., Luo, W., Chen, B., & Wang, W.-M. (2013). Identification and characterization of microRNAs involved in growth of blunt snout bream (*Megalobrama amblycephala*) by Solexa sequencing. *BMC Genomics, 14*(1), 754.

Yu, T., Li, J., Yang, Y., Qi, L., Chen, B., Zhao, F., . . . Wu, J. (2012). Codon usage patterns and adaptive evolution of marine unicellular cyanobacteria *Synechococcus* and *Prochlorococcus*. *Molecular Phylogenetics and Evolution, 62*(1), 206-213.

Zhang, F.-T., Zhang, Y.-B., Chen, Y.-D., Zhu, R., Dong, C.-W., Li, Y.-Y., . . . Gui, J.-F. (2008). Expressional induction of *Paralichthys olivaceus* cathepsin B gene in response to virus, poly I: C and lipopolysaccharide. *Fish and Shellfish Immunology, 25*(5), 542-549.

Zhou, Z.-j., Qiu, R., & Zhang, J. (2015). Molecular characterization of the cathepsin B of turbot (*Scophthalmus maximus*). *Fish Physiology and Biochemistry, 41*(2), 473-483.