Insect Molecular Biology (2007) **16**(4), 455–468

Sequence analysis and molecular characterization of larval midgut cDNA transcripts encoding peptidases from the yellow mealworm, *Tenebrio molitor* L.

S. Prabhakar*, M.-S. Chen†, E. N. Elpidina‡, K. S. Vinokurov‡, C. M. Smith*, J. Marshall* and B. Oppert†

**Department of Entomology, Kansas State University, Manhattan, KS, USA;* †*USDA ARS Grain Marketing and Production Research Center, Manhattan, KS, USA; and* ‡*A. N. Belozersky Institute of Physico-Chemical Biology, Moscow State University, Leninskie Gory, Moscow, Russia*

Abstract

Peptidase sequences were analysed in randomly picked clones from cDNA libraries of the anterior or posterior midgut or whole larvae of the yellow mealworm, *Tenebrio molitor* Linnaeus. Of a total of 1528 sequences, 92 encoded potential peptidases, from which 50 full-length cDNA sequences were obtained, including serine and cysteine proteinases and metallopeptidases. Serine proteinase transcripts were predominant in the posterior midgut, whereas transcripts encoding cysteine and metallopeptidases were mainly found in the anterior midgut. Alignments with other proteinases indicated that 40% of the serine proteinase sequences were serine proteinase homologues, and the remaining ones were identified as either trypsin, chymotrypsin or other serine proteinases. Cysteine proteinase sequences included cathepsin B- and L-like proteinases, and metallopeptidase transcripts were similar to carboxypeptidase A. Northern blot analysis of representative sequences demonstrated the differential expression profile of selected transcripts across five developmental stages of *Te. molitor*. These sequences provide insights into peptidases in coleopteran insects as a basis to study the response of coleopteran larvae to external stimuli and to evaluate regulatory features of the response.

Keywords: *Tenebrio molitor*, peptidases, insect digestion, coleopteran pests, molecular entomology, serine proteinases, cathepsins L and B, carboxypeptidase A, peptidase homolog.

Received 26 October 2006; accepted following revision 28 February 2007.
Correspondence: Brenda Oppert, USDA ARS Grain Marketing and Production Research Center, 1515 College Ave., Manhattan, KS 66502, USA. Tel.: + 1 785 776 780; fax: + 1 785 537 5584; e-mail: bso@ksu.edu

Introduction

Peptidases have important functions in all living systems. They include exopeptidases, such as amino- and carboxypeptidases, and endopeptidases, also referred to as proteinases, which are grouped as serine, cysteine, metallo- or aspartic, as defined by the active site and catalytic mechanism (Neurath, 1989). Peptidase gene families have been subdivided into clans based on sequence similarity and catalogued in the MEROPS database (Rawlings *et al*., 2006). Insect digestive peptidases catalyse the breakdown of protein during digestion and provide amino acids for growth and development. Therefore, peptidases are attractive candidates for the development of new pest control proteins based on synthetic or natural inhibitors.

Genome studies are providing information on the complexity and diversity of peptidase genes in insects. In a comparison of genome sequences, there were 305 serine proteinase (chymotrypsin or trypsin family) genes predicted in *Anopheles gambiae*, 199–210 in *Drosophila melanogaster* and 118 in *Homo sapiens*, but only 13 in *Caenorhabditis elegans* and one in *Saccharomyces cerevisiae* (Rubin *et al*., 2000; International Human Genome Sequencing Consortium, 2001; Holt *et al*., 2002; Zdobnov *et al*., 2002; Ross *et al*., 2003). Preliminary annotations from the genome of another tenebrionid, *Tribolium castaneum*, indicate that there are 165 serine, 25 cysteine and 12 carboxypeptidase genes (B. Oppert, unpublished data). Transcriptome analysis of the insect midgut has provided functional and physiological information related to digestion (Pedra *et al*., 2003; Campbell *et al*., 2005; Xu *et al*., 2005). Multigene families of trypsins have been found in *Lucilia cuprina* and *Haematobia irritans exigua* (Elvin *et al*., 1993; Casu *et al*., 1994), with 125–220 different serine peptidase genes in *L. cuprina* (Elvin *et al*., 1994).

Insects can respond to diets by the differential expression of peptidase genes. For example, a cluster of seven trypsin and two chymotrypsin genes in *Anopheles gambiae* was induced by feeding (Müller *et al*., 1993), while the expression

of 28 different serine peptidase genes in *Helicoverpa armigera* fluctuated in response to dietary inhibitors (Bown *et al*., 1997). Inclusion of soybean trypsin inhibitor in the diet of *H. armigera* stimulated an initial up-regulation of peptidase genes and a longer term down-regulation of inhibitor sensitive peptidase genes (Bown *et al*., 2004). Similar responses to cysteine proteinase inhibitors also occur in coleopteran insects (Zhu-Salzman *et al*., 2003).

The earliest reports of insect larval digestive peptidases were of the yellow mealworm, *Tenebrio molitor*, a pest of stored grain and grain products (Applebaum *et al*., 1964). Larvae of *Te. molitor* have a digestive physiology that incorporates a complex of digestive peptidases, including those from serine and cysteine proteinase classes, operating in a midgut with a sharp pH gradient (Terra *et al*., 1985). Cysteine proteinase activity is compartmentalized to the anterior region of the larval midgut of *Te. molitor*, whereas serine proteinase activity is found in the posterior (Thie & Houseman, 1990; Terra & Ferreira, 1994; Vinokurov *et al*., 2006a). At least 20 different peptidase activities have been reported in the larval midgut of *Te. molitor*: four aminopeptidases, two carboxypeptidases, six cysteine proteinases, and the rest serine proteinases (Applebaum *et al*., 1964; Zwilling, 1968; Zwilling *et al*., 1972; Levinsky *et al*., 1977; Garty, 1979; Golan, 1981; Urieli, 1982; Ferriera *et al*., 1990; Thie & Houseman, 1990; Terra & Cristofoletti, 1996; Cristofoletti & Terra, 1999, 2000; Cristofoletti *et al*., 2005; Elpidina *et al*., 2005; Tsybina *et al*., 2005). A comprehensive biochemical analysis suggested that at least 15 different proteinase activities are expressed simultaneously in *Te. molitor* larvae under normal dietary conditions, including six cysteine and nine serine proteinases (Vinokurov *et al*., 2006a,b). The N-terminal sequences of the major trypsin and chymotrypsin digestive proteinases are available (Elpidina *et al*., 2005; Tsybina *et al*., 2005), and five procathepsin L-like proteinases from a cDNA library included the predicted sequence of a major digestive proenzyme (Cristofoletti *et al*., 2005).

To develop the molecular tools needed to understand coleopteran responses to external stimuli, we analysed cDNAs obtained from randomly picked clones of libraries of different stages or different sections of midgut from *Te. molitor* larvae. Sequences were grouped based on structural motifs and phylogeny, and expression patterns of representative genes were compared in different sections of the gut as well as in five developmental stages of *Te. molitor*. Genetic sequences encoding the major dietary trypsin and chymotrypsin were identified.

Results

Te. molitor *transcripts*

cDNA libraries were constructed using mRNA from the anterior (AM) and posterior midgut (PM) of late instar larvae, and entire midguts of early (S) and mid (L) instar *Te. molitor* larvae. Larger larvae were used to compare AM and PM gene expression patterns to similar studies with protein. Midguts were pooled from earlier instars (S and L) to identify transcripts that may be specific to developmental stages. Sequences were obtained for a total of 301, 373, 272 and 330 clones selected randomly from the AM, PM, S and L libraries, respectively. When compared to other insect genomes, 29.2% of *Te. molitor* expressed sequence tag (EST) sequences had homologues in the *D. melanogaster* genome (http://flybase.bio.indiana.edu), but approximately half were found in the *Tr. castaneum* genome (http://www.hgsc.bcm.tmc.edu/projects/tribolium). In the non-redundant GenBank database, 26% of *Te. molitor* EST sequences had no significant match ($E < 10^{-5}$), and 67% encoded proteins that were not peptidases (Fig. 1A). Of the remaining 7% encoding potential peptidases, 73% encoded proteins similar to serine proteinases, 13% were similar to cysteine proteinases and 14% were similar to metallopeptidases. No sequences were obtained that encoded aspartic proteinases.

The distribution of sequences in the AM and PM were similar to what has been reported previously in a biochemical

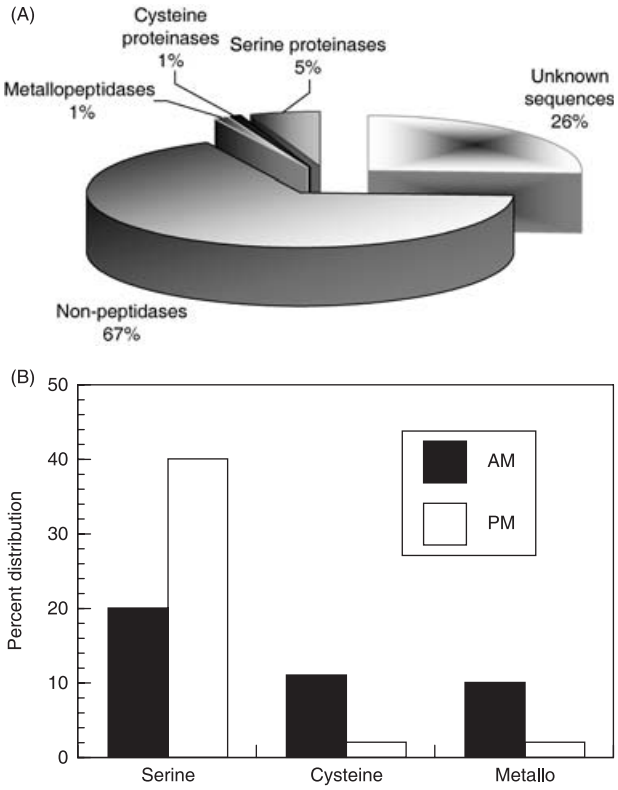

Figure 1. Analysis of expressed sequence tag sequences from *Tenebrio molitor.* (A) Categories of sequences from randomly picked clones of cDNA libraries from *Te. molitor* larvae (BlastX matches with $E < 10$). (B) Distribution of sequences encoding serine, cysteine, and metallopeptidases in the anterior and posterior midgut cDNA libraries of *Te. molitor* larvae.

analysis (Vinokurov *et al.*, 2006a). Thirty-seven cDNAs predicted to encode serine proteinases were identified in the PM library, whereas only 18 serine proteinase-encoding cDNAs were found in the AM library (Fig. 1B). In contrast, 10 sequences encoding cysteine proteinases were from the AM library and only two were from the PM library. A similar trend was observed for metallopeptidases, with more sequences obtained from the AM than from the PM library.

Classification of peptidase transcripts

PCR amplification of *Te. molitor* larval gut cDNAs identified in the EST screen as potential peptidases yielded 50 full-length cDNAs that grouped into 26 nonredundant clusters (Tables 1 and 2). Most of the sequences were obtained from the AM and PM libraries, but sequences unique to the S and L libraries also were found in the serine proteinase group. Most sequences had predicted signal peptides ranging from 16 to 41 amino acids and activation peptides. The majority of the sequences were serine proteinases, and predicted molecular masses of mature enzymes were from 21 227 to 26 398 Da, with acidic to basic isoelectric points (Table 1). Predicted sequences of mature cathepsins were from 22 582 to 40 500 Da, although the greater molecular mass was for AM4-72, and this transcript lacked a typical signal peptide (Table 2). The predicted carboxypeptidase A (CPA) transcripts encoded proteins with molecular masses ranging from 33 516 to 34 994 Da. Peptidases with a loss of sequence conservation in the catalytic site were predicted to be catalytically inactive and were classified as proteinase homologues (Ross *et al.*, 2003).

Tm serine proteinases and homologues

To characterize peptidases further by their specificity, residues were identified that determine primary specificity (Rawlings & Barrett, 1993; Perona & Craik, 1995; Ross *et al.*, 2003; Jiang *et al.*, 2005) (Table 1). All of the predicted serine proteinase sequences were compared to the S1 serine peptidase family, clan SA, with conserved sequence motifs TAAHC, DIAL and GDSGGP, containing the catalytic residues His57, Asp102 and Ser195 (chymotrypsin numbering, Rawlings & Barrett, 1993). Trypsin activation was indicated for all sequences, with an Arg and mostly Ile (and in a few cases, Val) bridging the activation site, conserving the hydrophobic residue at the N-terminus predicted to hydrogen bond to the Asp194 residue preceding the active site Ser (Kraut, 1971). In addition, the molecular masses of the proenzymes and mature peptides were within the range of other digestive serine proteinases, and sequences contained a single catalytic or similar domain, with a few exceptions. Sequences with a substitution of Gln for the catalytic His57, and with Gly, Val or Leu substitutions for the active site serine (Ser195) residue, were predicted to be catalytically inactive and were classified as serine proteinase homologues (SPHs). Six cysteine residues that form three disulphide bonds in S1 serine peptidases were found at conserved positions in most sequences (Kraut, 1971). The exceptions were AM1-11, PM4-06 and S3-80, each with one additional cysteine residue, but all were predicted to be functional serine proteinases based on catalytic residues. Most sequences missing one conserved cysteine were predicted to be SPHs, with the exception of PM4-60.

The TAAHC conserved motif contains the catalytic His57, which together with Ser195 or Asp102 form catalytic diads that determine rate acceleration and catalytic efficiency of peptide bond hydrolysis in the substrate (Perona & Craik, 1995). *Te. molitor* serine proteinases contained the conserved His57, whereas most SPHs contained a Gln substitution. Although 78% of serine proteinases in *D. melanogaster* contain the TAAHC region (Ross *et al.*, 2003), only 54% of the *Te. molitor* sequences had this motif, with 25% containing TAGHC and the remainder having TSGHC or TSAHC. All *Te. molitor* serine proteinases had the conserved Asp102, but there was less conservation in the DIAL motif. While this motif was found in 37% of serine proteinases from *D. melanogaster* (Ross *et al.*, 2003), it was in only one of the 24 *Te. molitor* serine proteinase sequences (PM4-06, chymotrypsin). The remainder of the serine proteinases contained the motifs DISV (all of the trypsins), DVAL, DVGL, DVGM, DIGM or DIGL. These motifs are also in *D. melanogaster* serine proteinases and SPHs. The exception was clone L4-25, with an Asn in the position of Asp102, a substitution that has been demonstrated to reduce enzyme activity 10^4-fold (Craik *et al.*, 1987).

Similarly to *D. melanogaster*, most *Te. molitor* S1 family proteinases contained the active site Ser residue (85 and 79%, respectively; Ross *et al.*, 2003). However, five *Te. molitor* S1 family proteinases had a Thr substitution for the active site Ser. The Ser to Thr change in evolution would have occurred with only a single base change, and it has been hypothesized that an intermediate peptidase may have contained Thr as a nucleophile (Barrett *et al.*, 1998). In *D. melanogaster*, there are five potential sequences (AAF58664, 53414, 22154, 49207, 44895) with either the GDTGGP or GDTGSP motif (http://www.ento.okstate.edu/labs/jiang/table1.htm). Serine proteinases from *L. cuprina* contain the Ser to Thr substitution, but the activity of these peptidases was not documented (Elvin *et al.*, 1994). We have tentatively classified the *Te. molitor* sequences as serine proteinase analogues (SPA), but their functional activity needs to be evaluated.

Residues in the S1 binding pocket and two loops that connect the walls of the binding pocket determine substrate specificity (Hedstrom *et al.*, 1992). Adjacent to loop 1, trypsins have a conserved Asp189, whereas chymotrypsins have a Ser at that position. The functional predictions for *Te. molitor* sequences correlate to these conserved amino

Table 1. cDNAs encoding serine proteinases from *Tenebrio molitor* larvae, with characteristics of the predicted peptides

Cluster	Clone ID	Accession no.	Mature enzyme Mm (Da)	pI*	Signal peptide (aa)	Conserved sequences around active site residues†			Binding pocket residues‡	Putative ID§
						TAAHC	DIAL	GDSGGP		
1	PM1-75¶	DQ356014	22 714	6.87	16		DISV		DG-	trypsin
	PM1-83	DQ356015	22 748	6.87	16		DISV		DG-	trypsin
	PM2-70	AY845177	22 715	6.87	16		DISV		DG-	trypsin
	PM2-03	DQ356016	24 227	7.74	16		DISV		DG-	trypsin
	PM1-95	DQ356017	22 716	6.87	16		DISV		DG-	trypsin
2	**AM1-11**	DQ356022	24 908	5.28	16		DVAL		GG-	SP
	AM1-21**	DQ356023	–	–	16		missing	missing	–	SPH
3	**AM1-62**	DQ356018	21 227	5.90	17		DVAV		GG-	SPH
	AM2-58	DQ356019	24 973	6.20	17		DVAV		GS-	SP
	AM3-01	DQ356020	24 660	6.45	17		DVAV		GS-	SP
	AM4-75	DQ356021	24 700	6.55	17		DVAV		GS-	SP
4	**PM4-86**	DQ356024	24 702	3.90	16	TAGHC	DIAV		SGP	chymotrypsin
	PM5-92	DQ356025	24 702	3.89	16	TAGHC	DIAV		SGP	chymotrypsin
5	S3-80	DQ356026	25 979	4.26	16	TAGHC	DVGL		SGL	chymotrypsin
6	AM1-01	DQ356027	23 857	3.98	16	TAAQC	DIGL	GDGGAA	NGS	SPH
	AM3-26	DQ356028	24 169	3.94	16	TAAQC	DIGL	GDGGAP	NGS	SPH
7	**PM5-90**	DQ356029	24 348	3.98	16	TAAQC	DIGL	GDGGDP	NGS	SPH
8	AM4-49	DQ356030	23 935	4.10	16	TVAQC	DIGL	GDGGSP	SGS	SPH
9	**PM4-06**††	DQ356031	22 899	8.95	16				SGK	chymotrypsin
10	**AM2-68**	DQ356033	24 294	3.79	16	TAGHC	DIGL		GGT	SP
	AM4-68	DQ356032	24 463	4.22	16	TAGHC	DIGL		GGT	SP
	L3-22	DQ356034	24 362	4.22	16	TAGHC	DIGL		GGT	SP
11	PM4-60	DQ356038	25 435	4.46	41		DIGL	GDTGGP	GGS	SPA
	PM5-80	DQ356039	25 458	4.54	41		DIGL	GTDGGP	GGS	SPH
	PM4-36**	DQ356040	–	–	41		DIGL	–	E –	SPH
12	PM4-08	DQ356037	26 152	4.31	16	TAGQC	DVGM	GDVGGA	GGS	SPH
13	**PM4-54**	DQ356035	26 398	4.27	16	TAGQC	DIGM	GDVGGA	GGS	SPH
	PM4-63	DQ356036	25 517	4.37	16	TAGQC	DIGM	GDVGGA	GGS	SPH
14	L4-25	DQ356041	25 703	4.54	19		NIGL	GDTGSP	GGQ	SPH
15	**AM4-47**	DQ356042	25 155	4.06	17	TAGQC	DIGL	GDLGSP	GGS	SPH
	PM2-57	DQ356043	25 164	4.06	17	TAGQC	DIGL	GDLGSP	GGS	SPH
16	**PM1-93**	DQ356044	24 522	4.12	16	TSGSC	DIGV	GDVGSP	GGS	SPH
17	**PM3-37**	DQ356045	25 105	4.51	19		DIGL	GDSGSP	RGS	SP
18	**L3-34**	DQ356046	25 446	4.34	18	TSAHC	DIGL	GDSGSP	RGT	SP
19	**PM2-01**	DQ356047	25 424	4.25	21	TSGHC	DIGL	GDTGIP	GG-	SPA
20	PM4-31	DQ356048	25 145	4.20	21	TSGHC	DIGL	GDTGSP	GGS	SPA
21	L4-24	DQ356049	25 143	4.20	21	TSGHC	DIGL	GDTGSP	GGS	SPA
	S3-72	DQ356050	25 173	4.20	21	TSGHC	DIGL	GDTGSP	GGS	SPA

Bold sequences were used in expression studies (see Fig. 4).
*Isoelectric point is for the mature enzyme.
†Only sequences differing from conserved regions are given (active site residues are underlined).
‡Binding pocket residues are found in the specificity substrate-binding pocket of the enzyme; '–' indicates lack of similarity in region.
§Identification based on active site residues, as per Perona & Craik (1995) and Ross *et al.* (2003). SP, serine proteinase; SPH, serine proteinase homologue; SPA,serine proteinase analogue.
¶All sequences in this cluster have a predicted N-terminus identical to TmT1 (Tsybina *et al.*, 2005).
**Truncated sequence; no prediction on mature enzyme.
††Sequence encodes TmC1 (Elpidina *et al.*, 2005).

acids, as only the *Te. molitor* trypsins in cluster 1 have this conserved Asp. Correspondingly, the *Te. molitor* sequences from clusters 4, 5 and 9 had a Ser at that position, and therefore were predicted to encode chymotrypsins. AM4-49 contained Ser189, but it was predicted to be a SPH because of the lack of conservation in active site residues.

A phylogenetic tree based on maximum parsimony analysis of an alignment of *Te. molitor* serine proteinase ESTs and related sequences in *Tr. castaneum* (Tc#), *Choristoneura fumiferana* (CfTryp, Wang *et al.*, 1993), *Spodoptera frugiperda* (SfTryp), *D. melanogaster* (Dm-Tryp, DmChym, Adams *et al.*, 2000), *Homo sapiens* (HsTryp, HsChym, Emi *et al.*, 1986) and *Bos taurus* (BtTryp, Le Huerou, 1990) resulted in distinct groupings of proteinase sequences (Supplementary Material Fig. S1A). PM2-70, encoding the *Te. molitor* trypsin TmT1 (Tsybina *et al.*, 2005),

Table 2. cDNAs encoding cysteine and metallopeptidases from *Tenebrio molitor* larvae, with characteristics of the predicted peptides

Cluster	Clone ID	Accession no.	Mature enzyme Mm (Da)	pI*	Signal peptide (aa)	Active site residues	Binding pocket residues	Putative ID
22	**AM4-18**	DQ356051	23 602	4.61	19	QCSN	HH	cathepsin B-like
23	**AM3-87**	DQ356052	26 336	4.74	19	QCHN	AH	cathepsin B-like
24	**AM3-32**	DQ356053	22 582	3.83	16	QCHN		cathepsin L
	AM4-22	DQ356054	22 617	4.01	16	QCHN		cathepsin L
25	AM4-72	DQ356055	40 500	3.92	–	QCHN		cathepsin L
26	AM2-60	DQ356060	34 873	5.30	19	RRYE	HEH	carboxypeptidase A
	AM1-30	DQ356058	34 900	5.08	19	RRYY	HEH	carboxypeptidase A homolog
	AM2-51	DQ356061	33 516	6.20	19	RRYE	HEH	carboxypeptidase A
	AM1-02	DQ356056	34 650	5.08	19	RRYE	HEH	carboxypeptidase A
	AM1-72	DQ356059	34 861	5.30	19	RTYE	HEH	carboxypeptidase A homolog
	AM3-75	DQ356057	34 994	5.08	19	RRYE	HEH	carboxypeptidase A
	L4-60	DQ356062	34 994	5.08	19	RRYE	HEH	carboxypeptidase A

Bold sequences were used in expression studies (see Fig. 4).
*Isoelectric point is for the mature enzyme.

was found in a clade with 10 *Tr. castaneum* serine proteinases and *D. melanogaster* trypsin. Another clade contained S3-80 and PM4-86, *Te. molitor* enzymes that were predicted to be chymotrypsins by sequence analysis using *Tr. castaneum*, *S. frugiperda* and *D. melanogaster* serine proteinases. A smaller, closely related branch contained the sequence encoding the *Te. molitor* chymotrypsin TmC1 (Elpidina *et al.*, 2005), PM4-06 and two *Tr. castaneum* sequences. Combined with the previous clade, these data support the chymotrypsin clade of the tree. Clade 1 consisted of mostly SPHs, whereas clades 2 and 3 were mostly serine proteinases.

Tm cysteine proteinases

Predicted *Te. molitor* cathepsin L enzymes had the conserved diad residues Cys25 and His159 (papain numbering), and Gln19 and Asn/Asp175 that stabilize the molecule (Rawlings & Barrett, 1993), both found in members of the C1 cysteine peptidase family, clan CA (Table 2). All contained the conserved trio of cystine residues found in mammalian homologues, Cys22/63, Cys56/95 and Cys153/Cys200, but AM3-32 had two additional Cys residues and AM4-72 had a total of 13 Cys residues. Although AM4-72 was similar to these cathepsin L sequences, the sequence was highly divergent in the N-terminus and lacked a typical activation site as well as signal peptide. There were no N-glycosylation sites in any of the *Te. molitor* cathepsin L sequences.

In the predicted *Te. molitor* cathepsin B sequences, belonging to the same family as cathepsin L, AM4-18 lacked the conserved His159 as well as the residue's C-terminal to the conserved Asn175 (Table 2). However, AM4-18 contained the two His residues (His110/111) in the occluding loop region of cathepsin B that block the C-terminal end of the active site cleft and cause the enzyme to act as a dipeptidase (Musil *et al.*, 1991). There were 12 Cys residues in AM3-87, similar to typical cathepsin B proteinases, but there were 18 Cys residues in AM4-18. There was one potential glycosylation site in AM4-18 (at residue 297), but none in AM3-87. As both sequences lack some of the conserved residues found in other cathepsin B enzymes, it is unknown if these represent clones of active enzymes and they were therefore classified as 'cathepsin B-like'. An alignment of cathepsin-like proteinases from the *Te. molitor* EST library was made with cathepsins from *Aedes aegypti* (AaCathB), *Anthonomus grandis* (AgCP, De Olivera Neto *et al.*, 2004), *B. taurus* (BtCathB), *Carica papaya* (Papain, Cohen *et al.*, 1986), *Callosobruchus maculates* (CmCathL, Zhu-Salzman *et al.*, 2003), *Diabrotica virgifera virgifera* (DvCAL1, Koiwa *et al.*, 2000; DvCathL1-4, DvCathB1,2, Bown *et al.*, 2004), *D. melanogaster* (DmCathB, Adams *et al.*, 2000), *Giardia lamblia* (GlCathB, McArthur *et al.*, 2000), *H. sapiens* (HsCathB, Gilmore *et al.*, 2005), *Te. molitor* (TmCAL1a-c, 2 and 3, Cristofoletti *et al.*, 2005), and *Triatoma infestans* (TiCathB), and this alignment was used to construct a phylogenetic tree (Supplementary Material Fig. S1B). There were two major clades in this tree, with the upper clade consisting of cathepsin L proteinases and the lower clade of cathepsin B. Both AM4-72 and AM3-32 were found in the cathepsin L clade, although AM4-72 had no discernible signal peptide and aligned to the other sequences poorly. Of the cathepsin L sequences, the only digestive enzyme supported with biochemical data is TmCAL2. However, there is some evidence that TmCAL3 also may be digestive (Cristofoletti *et al.*, 2005), and AM3-32 and AM4-72 are found in this clade. Further support for this hypothesis is the inclusion of CmCathL in this clade, a putative gut digestive proteinase (Zhu-Salzman *et al.*, 2003). In the cathepsin B clade, AM4-18 was found in a clade with lysosomal enzymes (AaCathB and BtCathB), but was more closely related to *Tr. castaneum* sequences found in tandem on

chromosome 3 (Tc02952 and Tc02953) and a sequence from an insect digestive tract (TiCathB). AM3-87 was closely related to a *Tr. castaneum* sequence also located on chromosome 3 (Tc02954). It is likely that the *Tr. castaneum* sequences are a result of gene duplication, but this remains to be determined in the *Te. molitor* sequences.

Tm carboxypeptidases

Carboxypeptidase sequences predicted from *Te. molitor* ESTs were similar to CPA members of the M14 metallopeptidase family, clan MC, with a conserved zinc-binding motif (Table 2). As with all other midgut peptidase cDNAs, signal peptide and activation sites were identified in *Te. molitor* CPAs. The residues in the catalytic zinc binding site, His69, Glu72 and His196 (*B. taurus* CPA numbering, Titani *et al*., 1975; Christianson & Lipscomb, 1986), were conserved in all *Te. molitor* CPA sequences. Residues involved in substrate binding, Arg127, Arg145 and Tyr248, were not conserved in AM1-72, as there was a Thr instead of Arg145. Lack of conservation also was found in the catalytic nucleophile, Glu270, with a Tyr residue at that position in AM1-30. Therefore, it is unknown if these clones represent functionally active CPAs and were thus classified as CPA homologues.

Alignments were made with *Te. molitor* CPAs and CPA from *A. aegypti* (AaCPA, Edwards *et al*., 1997), *Culicoides sonorensis* (CsCPA, Campbell *et al*., 2000), *D. melanogaster* (DmCPA, Adams *et al*., 2005), *Helicoverpa armigera* (HaCPA1 and 2, Bown & Gatehouse, 2004), *Mayetiola destructor* (MdCPA1-5, Liu *et al*., 2006), *Sitodiplosis mosellana* (SmCPA, Mittapalli *et al*., 2006), *Simulium vittatum* (ScCPA, Ramos *et al*., 1993), *Trichoplusia ni* (TnCPA1-5, Wang *et al*., 2004), *H. sapiens* (HsCPA, Strausberg *et al*., 2002) and *B. taurus* (BtCPA, Le Huerou *et al*., 1991), and the alignment was used in a phylogenetic analysis (Supplementary Material Fig. S1C). *Te. molitor* CPAs were found in a clade with two *Tr. castaneum* sequences and a CPA from *C. sonorensis*. These sequences were more distantly related from the clade containing lepidopteran CPAs from *H. armigera* and *T. ni*, although *Tr. castaneum* sequences were also found in this clade.

Comparison of EST sequences and purified TmT1 and TmC1 proteins

An N-terminal sequence was obtained from the major trypsin, TmT1, from the *Te. molitor* larval midgut (Tsybina *et al*., 2005). This N-terminal sequence was identical to the predicted N-termini of the mature enzymes of sequences from cluster 1 (Table 1). The N-terminus of the major *Te. molitor* larval midgut chymotrypsin, TmC1 (Elpidina *et al*., 2005), corresponded to only one predicted mature peptidase, PM 4-06 (Fig. 2B). For further comparison of the corresponding sequences, mass spectra of tryptic peptides of each purified proteinase were obtained (data not shown). The peaks in the mass range of 600–3000 Da from both spectra were subjected to tandem mass spectrometry (MS/MS) fragmentation in order to compare with peptide sequences predicted from the cDNA.

The masses of four peptides from the tryptic digest of TmT1 and of their MS/MS fragments were identical to those of predicted peptides 92–102, 103–112, 236–249 and 250–258 of PM1-75, PM2-70 and PM1-95 (PM2-70 is shown in Fig. 2A). The predicted C-terminal peptides of PM1-83 and PM2-03 differed from that of TmT1 (data not shown). However, PM1-75, PM2-70 and PM1-95 had predicted amino acid sequences, molecular masses, and pI values similar to those calculated for TmT1 (Tsybina *et al*., 2005). Therefore, TmT1 may be the protein product of one of these trypsin genes, or alternatively, the purified protein may represent a mixture of trypsin isoforms. The differences in amino acid sequences of PM1-75, PM2-70 and PM1-95 are located in long tryptic peptides and were not resolved by the mass spectral analysis.

One *Tr. castaneum* trypsin gene, Tc07017, was closely related to PM2-70, with conservation in the signal peptide and activation sequence (Fig. 2A). Tenebrionid trypsin sequences were compared to trypsins from *D. melanogaster* (DmTryp, Adams *et al*., 2000), *H. sapiens* (HsTryp, Emi *et al*., 1986), *B. taurus* (BtTryp, Le Huerou, 1990) and chymotrypsin from *H. sapiens* (Emi *et al*., 1986). The conserved sequence around the active site His57 was identical in insects, but differed slightly in mammals. The sequence around Asp102 was identical in mammalian trypsins, but the sequences around Ser195 were identical in all genes.

The TmC1 tryptic digest in the analysed mass range also consisted of four peptides (Fig. 2B). The MS/MS of the tryptic peptides from TmC1 were identical to 112–118, 150–171, 253–260 and the C-terminal 266–274 predicted peptides of PM4-06. Therefore, PM4-06 was predicted to be the gene encoding the previously purified digestive chymotrypsin, TmC1. PM4-06 was most closely related to another *Te. molitor* chymotrypsin, PM4-86, and also to a sequence from *Tr. castaneum*, Tc15780. These sequences were more related to the chymotrypsins from *D. melanogaster* and *H. sapiens* than another chymotrypsin sequence obtained from early instar *Te. molitor* larvae, clone S3-80. Clone S3-80 had two homologous sequences in *Tr. castaneum*, Tc11824 and Tc11825, which are located in tandem on chromosome 9 and share 57 and 65% identity with S3-80, respectively.

Expression analyses

To investigate the expression patterns of individual peptidases and homologues representative of selected clusters, mRNA levels were compared by Northern analysis (Table 3; Fig. 3). The individual peptidase transcripts had variable expression patterns in five developmental stages (first, mid and late instar larva, pupa and adult) of *Te. molitor*.

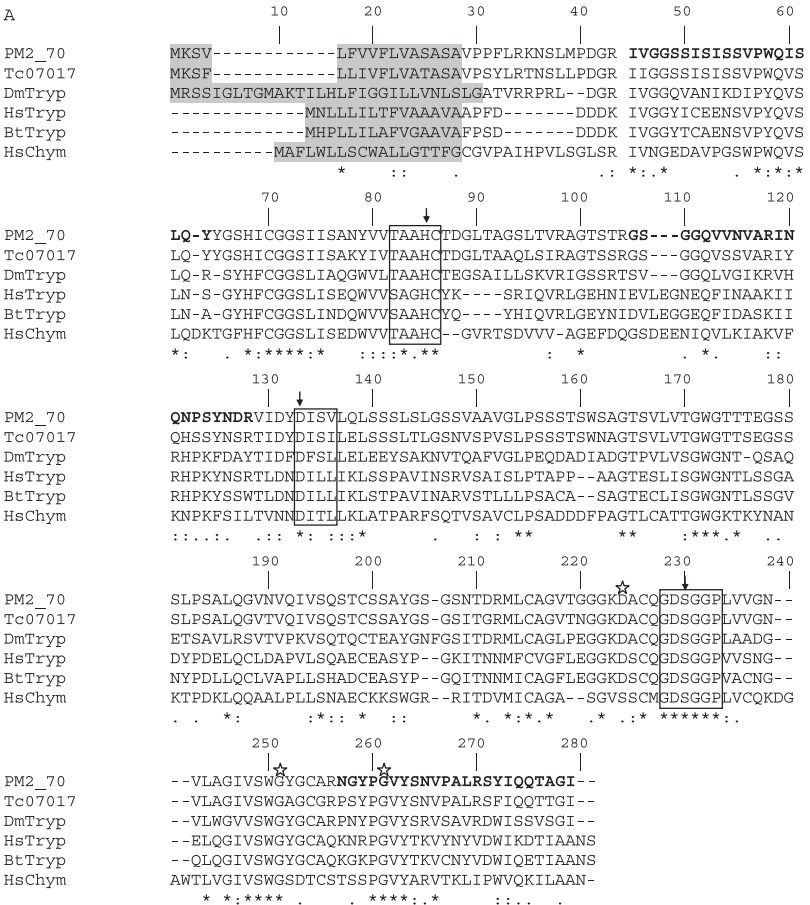

Figure 2. Alignment of peptide sequences from *Tenebrio molitor* digestive proteinases (in bold) to *Te. molitor* cDNA, *Tribolium castaneum* genome sequences, and other related sequences. (A) TmT1 peptide sequences aligned with trypsins from *Te. molitor* (PM2-70), *Tr. castaneum* (Tc07017), *Drosophila melanogaster* (DmTryp, AAF52738), *Homo sapiens* (HsTryp, AAA61232), *Bos taurus* (BtTryp, CAA38513) and HsChym (NP_001898). (B) TmC1 peptide sequences aligned with chymotrypsins from *Te. molitor* (S3-80, PM4-86 and PM4-06), *Tr. castaneum* (Tc11824, Tc11825 and Tc15780), *D. melanogaster* (DmChym, AAF49326) and HsChym (NP_001898). Shaded regions at the N-terminus are predicted signal peptides; space indicates activation site; conserved sequence motifs are boxed; catalytic residues are indicated by arrows; stars indicate residues that determine specificity or bind to substrate.

Table 3. Relative abundance of mRNA transcripts in different developmental stages of *Tenebrio molitor* larvae

Clone ID	Tentative ID*	L1†	L2	L3	Pupa	Adult
PM2-70	trypsin (tmt1b)	1	0.82	0.42	–	0.23
PM1-93	SPH	1	0.98	0.54	–	0.47
PM4-86	SP	1	0.25	–	–	–
PM5-80	SP	1	0.80	0.67	–	0.37
PM4-54	SPH	1	0.84	0.72	–	1.01
AM1-62‡	SPH	–	–	1	–	0.46
L3-34	SP	1	–	–	0.78	1.02
PM5-90	SPH	1	–	–	–	–
AM4-47	SPH	1	0.55	0.46	–	–
AM1-11	SP	1	0.21	0.76	–	0.95
PM2-01	SPA	1	–	–	–	0.32
PM4-06	chymotrypsin (tmc1a)	1	0.99	0.95	–	0.74
AM2-68	SP	1	1.06	0.91	–	0.80
PM3-37	SP	1	1.04	0.90	–	1.11
AM3-87	cathepsin B-like	1	0.87	1.04	1.03	1.42
AM4-18	cathepsin B-like	1	1.10	0.94	–	1.03
AM3-32	cathepsin L	1	0.81	1.04	–	1.19
PM2-22	ubiquitin	1	1.06	1.06	1.10	1.20

*Tentative identification of the gene product. SP, serine proteinase; SPH, serine proteinase homologue.
†Relative intensity (pixels per unit area) of band in each developmental stage divided by L1 (except AM1-62). No detectable expression of RNA is indicated by (–). L1, L2, and L3 refer to the larval stage, as defined in the Experimental procedures.
‡Expressed only in third instar and adult (value is relative to L3).

Among the serine proteinases, only transcript L3-34 was expressed in the pupal (nonfeeding) stage. All transcripts, except AM1-62, were expressed in first instar larvae. The genes encoding PM2-70, PM4-06, AM2-68, PM5-80, PM4-54 and PM3-37 transcripts shared a similar expression pattern: these genes were expressed at the highest level in the first instar larvae; their expression level was decreased slightly in subsequent larval stages; no expression was detected in pupae; and a moderate to high level of expression was observed in adults. This expression pattern suggests that these proteinases may be involved in protein digestion. Transcripts for PM4-86, PM5-90, AM4-47 and PM2-01 were only expressed in early instars and were either at low levels or not detected in later instar larvae, pupae or adults. Other transcripts had variable expression patterns: there was a high level of AM1-11 expression in first and late instar larvae as well as in the adults; AM1-62 was expressed only in late instar larvae; PM1-93 was detected at higher levels in first and mid instar larvae, but at very low levels in later instar and adults; and L3-34 was only expressed in first instar larvae, pupae and adults. Transient expression in specific developmental stages of selected transcripts suggests a role of developmental regulation for these genes. Of interest is that, although

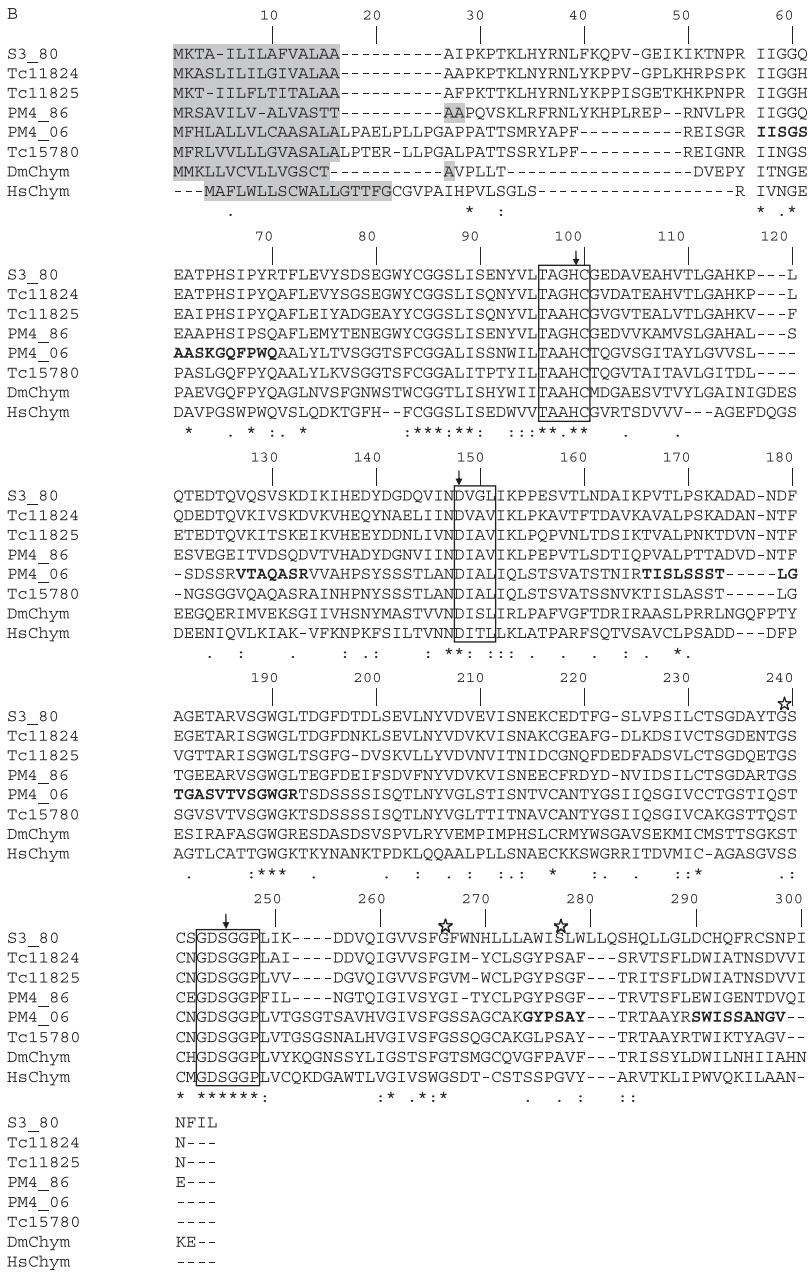

Figure 2. *(Continued)*

L3-34 was isolated from L3 larvae, it was expressed at very low levels in this larval stage.

Among the three cysteine proteinases analysed, AM3-87 (a predicted cathepsin B-like enzyme) was expressed in all developmental stages, including the pupa. AM4-18 (cathepsin B-like) and AM3-32 (cathepsin L) were expressed at increased levels in first instar larvae and adults and at moderate levels in other larval instars.

Discussion

Our results indicate that midgut transcripts encoding cysteine proteinases were isolated from the AM, whereas the majority of serine proteinases were found in the PM. These data agree with previous observations that protein digestion in *Te. molitor* larvae is compartmentalized (Thie & Houseman, 1990; Terra & Cristofoletti, 1996; Vinokurov *et al.*, 2006a). Trypsin and chymotrypsin transcripts from the PM were related to previously purified and biochemically characterized proteinases from the PM, TmT1 and TmC1 (Elpidina *et al.*, 2005; Tsybina *et al.*, 2005). Sequences in cluster 1 may constitute individually or simultaneously the major digestive trypsin(s) TmT1, responsible for 84% of the total trypsin activity in the PM. The transcript PM4-06 had a predicted protein with an N-terminus identical to that of the purified protein TmC1, and internal sequence that

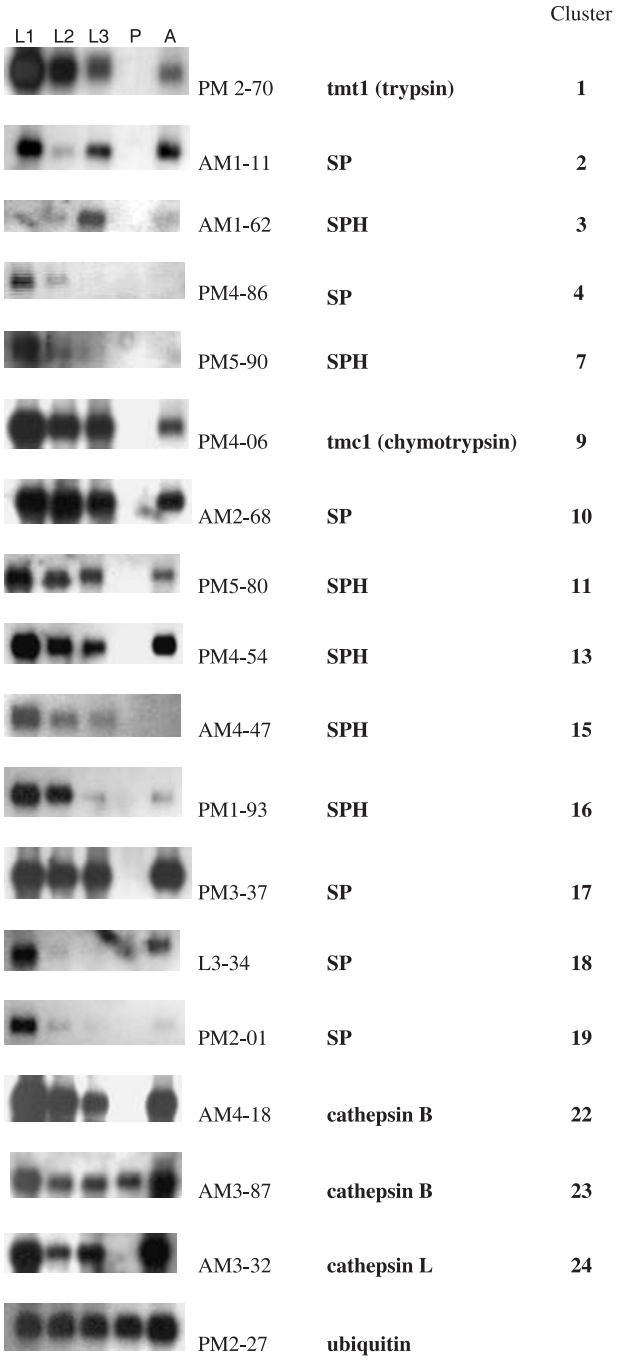

Figure 3. Northern blot analysis of RNA extracted from first instar larvae (L1), mid instar larvae (L2) and late instar larvae (L3), pupae and adults. Probes specific to selected cDNAs from representative clusters were used for analysis. SP, serine proteinase; SPH, serine proteinase homologue.

correlated to MS/MS spectra. Biochemical data indicate that this protein contributes about 42% of the total chymotrypsin activity in the PM (Elpidina *et al*., 2005). Our study complements others involving cDNAs encoding cathepsin L enzymes as well as the biochemical characterization of these enzymes in *Te. molitor* (Cristofoletti *et al*., 2005; Vinokurov *et al*., 2006b). All cathepsin transcripts were localized to the AM and correspond to biochemical characteristics of cysteine proteinases in the AM (Vinokurov *et al*., 2006a).

There were no cDNAs corresponding to the AM 'heavy' trypsin-like proteinases, with apparent molecular masses of 59 kDa, observed in our previous biochemical studies (Vinokurov *et al*., 2006b). The lack of cDNAs encoding heavy trypsins supports the hypothesis that these enzymes are complexes resulting from the association of trypsin monomers under certain gut conditions (Brito *et al*., 2001; Wagner *et al*., 2002). However, it is still possible that transcripts for these enzymes may be of low abundance. Alternatively, these heavy trypsins may have sequences that are not similar to known proteinases.

Although this is the first report of CPA cDNA transcripts in *Te. molitor*, several studies have characterized digestive CPA from other insect midguts (Ramos *et al*., 1993; Bown *et al*., 1998; Edwards *et al*., 2000; Bown & Gatehouse, 2004; Wang *et al*., 2004). All cDNAs for carboxypeptidases and their homologues were from the AM, which suggests that these enzymes may be found in the AM of *Te. molitor* larvae. However, there are likely to be other CPAs in *Te. molitor* that were not represented in our EST sequences.

SPHs are apparently a common phenomenon in vertebrates and invertebrates, but their function is not well characterized. In invertebrates, SPHs have been proposed to participate in immune and antimicrobial responses (Kawabata *et al*., 1996; Dimopoulos *et al*., 1997). Almost half of the serine proteinase cDNAs were categorized as SPHs, and we speculate that they may be involved in compensation to plant proteinase inhibitors or toxins by providing a 'sink' of molecules that sequester these molecules. Alternatively, they may be involved in mounting immune responses to pathogen invasion.

Our results emphasize the importance of serine and cysteine proteinases in the growth, physiology and development of *Te. molitor*. Northern blot analysis revealed that, in most instances, peptidase genes are expressed at higher levels in first instar larvae than other developmental stages, corresponding to the active phase of rapid growth.

Preliminary results from the *Tr. castaneum* gene annotation project indicate that there are many peptidase genes (B. Oppert, unpublished data). However, information on tissue and expression profiles for most genes is not available. Although protein digestion in *Tr. castaneum* larvae is primarily by cysteine proteinases, higher levels of serine proteinases are induced when larvae are fed a cysteine proteinase inhibitor (Oppert *et al*., 1993, 2003, 2005). A combination of inhibitors that target both serine and cysteine proteinases is necessary to prevent this adaptive response (Oppert *et al*., 2004, 2005). Based on the digestion profile in *Te. molitor*, we predict that inhibitors

that target serine (trypsin and chymotrypsin) and cysteine proteinases (cathepsin L) also will be required to control feeding damage by *Te. molitor* larvae. However, given that cysteine proteinases are more important in the initial stages of food digestion in *Te. molitor* larvae (Vinokurov *et al.*, 2006b), cathepsin inhibitors may play a critical role in the inhibition of digestion. Studies of the expression and localization of *Te. molitor* midgut peptidase genes and the proteins they encode will provide new insights into the compensatory response of *Te. molitor* to dietary proteinase inhibitors and to other antinutritional compounds.

Experimental procedures

Insect rearing and dissection

Tenebrio molitor larvae used in this research were derived from a laboratory colony. Insects were reared at a relative humidity of 60–70% at 25 °C. Late instar larvae of both sexes, weighing 1.13 ± 0.03 g (n = 15), were used for AM and PM midgut cDNA library construction and isolation of digestive trypsin and chymotrypsin. 'Small' (0.89 ± 0.24 mg, n = 27) and 'large' (168 ± 0.25 mg, n = 10) instar larvae were used for whole larvae cDNA library construction (S and L, respectively). Larvae from three different instars, pupae and adults were used for RNA isolation for Northern blots.

Actively feeding larvae were immobilized on ice and were dissected by excising the anterior and posterior ends, removing the gut with forceps and placing the gut into a solution of diethylpyrocarbonate (DEPC)-treated water. The entire midgut was divided into AM and PM sections of identical length by cutting midway between the most anterior region and the insertion of the Malpighian tubules. The contents of the gut were gently forced out using a pair of forceps and the tissue was rinsed twice in DEPC-treated water, dried on filter paper, and immediately transferred to TRIReagent™ (Molecular Research Center, Inc., Cincinnati, OH, USA).

cDNA library construction and sequencing

Total cellular RNA was extracted from dissected gut tissue using TRIReagent, according to the procedure provided by the manufacturer. Four cDNA libraries (AM, PM, S and L) were constructed from RNA samples using a SMART™ cDNA library construction kit (Clontech, Palo Alto, CA, USA), which synthesizes cDNA using oligo-dT primers, according to the protocol provided by the manufacturer with one modification. Instead of using the original phage vector, PCR fragments were cloned directly into a plasmid using a TOPO TA cloning kit (Invitrogen, Carlsbad, CA, USA). Approximately 500 clones from each of the four libraries were picked randomly and unidirectional sequences were obtained commercially (MegaBACE 4000, Amersham Biosciences, Piscataway, NJ, USA; Rexagen DNA Sequencing Service, Seattle, WA, USA; ABI Prism™ 3730xl, SeqWright DNA Techonology Services, Houston, TX, USA). Universal primers M13 F (-20) and M13 R were used for sequencing of predicted peptidase clones. Specific synthetic primers were used to confirm the sequences from both directions. Some sequence analysis also was performed using an ABI 3700 DNA sequencer at the DNA Sequencing and Genotyping Facility, Department of Plant Pathology, Kansas State University, Manhattan, KS, USA.

Computer-based sequence analysis

Vector sequences were trimmed using Sequencher (Gene Codes Corporation, Ann Arbor, MI, USA). To group into clusters and to identify redundant sequences, cDNAs were analysed using a customized BLASTN program that produced outputs with sequence assembly parameters similar to those of the CAP3 assembly program (Huang & Madan, 1999). Sequences also were analysed using the BLASTX algorithm (http://www.ncbi.nlm.nih.gov/BLAST). Sequences were grouped by their similarity, based on predicted open reading frame (ORF), to sequences in the database at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov; Bethesda, MD, USA). Full length sequences were obtained with synthetic primers for nonredundant peptidase-encoding clones.

Secretion signal peptides were predicted using PSORT or SignalP (http://psort.nibb.ac.jp/form2.html; http://www.cbs.dtu.dk/services/SignalP). The ExPASy Proteomics tools on the website of the Swiss Institute of Bioinformatics (http://www.expasy.ch) were used to calculate molecular mass and isoeletric point of deduced protein sequences. Alignments were made with MULTALIN (PBIL, Lyon, France) or CLUSTALW in pair-wise comparisons (gap weight = 8, gap length weight = 2; Thompson *et al.*, 1994).

Isolation and purification of Te. molitor *digestive trypsin and chymotrypsin*

Isolation and purification of digestive trypsin (TmT1) and chymotrypsin (TmC1) was performed as described earlier (Tsybina *et al.*, 2005 and Elpidina *et al.*, 2005, respectively). Two hundred PM of late instar larvae were homogenized in 0.75% NaCl and clarified with centrifugation. The extract was dialysed against 20 mM potassium sodium phosphate buffer, pH 6.9, containing 0.02% sodium azide. Dialysed extracts were subjected to batch chromatography on DEAE-Sephadex A-50 in the same buffer, followed by gel filtration on a Superdex-75 column fast protein liquid chromatography (FPLC) in the same buffer (G.E. Healthcare Europe GmbH, Freiburg, Germany). Aliquots of resolved and purified enzymes were subjected to SDS-PAGE according to Laemmli (1970), followed by in-gel tryptic hydrolysis.

In-gel tryptic hydrolysis of purified trypsin and chymotrypsin

Tryptic peptides of purified digestive trypsin and chymotrypsin were obtained after hydrolysis of proteins in a 0.15% Coomassie Brilliant Blue R-250-stained gel. A 1 mm^2 gel piece was excised and washed twice in 150 µl 40% (v/v) acetonitrile in 0.1 M NH_4HCO_3 for 20 min at 56 °C. The gel was further dehydrated in 150 µl of the same buffer, dried and supplemented with 3 µl modified sequencing-grade trypsin (Promega, Madison, WI, USA) dissolved in 0.05 M NH_4HCO_3 to a final concentration of 10 µg/ml. The hydrolysis was performed for 15 h at 37 °C and stopped by the addition of 5 µl of 0.1% trifluoroacetic acid in a 10% solution (v/v) of acetonitrile and water, followed by thorough mixing. The gel solution was used for MALDI-TOF MS and MS/MS analysis of peptides.

MALDI-TOF MS and MS/MS analysis

Aliquots (1 µl) of the sample were mixed on a steel target with an equal volume of 2,5-dihydroxybenzoic acid (Sigma-Aldrich, St. Louis, MO, USA) solution (10 mg/ml in 30% acetonitrile/0.5% trifluoroacetic acid), and the droplet was left to dry at room temperature. Mass spectra were recorded on an Ultraflex MALDI-TOF-TOF

mass spectrometer (Bruker Daltonik, Bremen, Germany) equipped with a 337 nm laser with positive ion detection. Each mass spectrum was obtained as the sum of a minimum of 200 laser shots. Fragment ion spectra were generated by laser-induced dissociation slightly accelerated by low-energy collision-induced dissociation using helium as a collision gas. The MH^+ molecular ions of the tryptic digest were measured in the reflector mode, and the accuracy of mass peak measurement was 0.02%. Mass peak accuracy of the measurement of MS/MS fragmentation of peptides was 0.05%. Correspondence of the masses and MS/MS peptide fragments to predicted protein peptides was manually interpreted with GPMAW 4.04 (Lighthouse Data, Odense, M Denmark).

RNA isolation and Northern blot analysis

Total cellular RNA from five different developmental stages (first-instar, mid-instar, late-instar larvae, pupae and adults) of *Te. molitor* was extracted as previously described. For Northern blots, equal amounts (5 μg) of total RNA were separated in a 1.2% agarose gel containing formaldehyde and transferred to a nylon membrane. The membrane was baked at 80 °C for 2 h to fix RNA on to the membrane. Membranes were hybridized separately to 20 individual cDNA probes, each representing a different peptidase cluster. The cDNA probes were labelled with ^{32}P-dCTP using a random labelling kit from Stratagene (La Jolla, CA, USA). Hybridization was carried out overnight at 42 °C in a plastic bag containing 15 ml hybridization solution (10% dextran sulphate/1% sodium dodecyl sulfate (SDS)/1 M NaCl, pH 8.0). After hybridization, the membranes were washed twice with 2 × sodium chloride/sodium citrate (SSC) at room temperature for 30 min, twice with 2 × SSC, 1% SDS at 65 °C for 30 min, and twice with 0.1 × SSC plus 1% SDS at room temperature for 30 min. The membranes were exposed to Kodak SR-5 X-ray film overnight (Kodak, Rochester, NY). The expression of mRNA in developmental stages of *Te. molitor* larvae was measured as band intensity per unit (NucleoTech, San Mateo, CA, USA).

Phylogenetic analysis

For each type of peptidase, amino acid sequences were aligned with the program CLC Free Workbench (www.CLCbio.com). The resulting alignments were then imported into PAUP 4.0b 10 for phylogenetic analysis (Swofford, 2002). Specifically, we conducted maximum parsimony analysis using a heuristic search with gaps counted as missing data, 10 random taxon addition replicates and tree-bisection and reconnection (TBR) branch swapping. All characters were equally weighted. Bootstrap analyses on the shortest length trees via heuristic (100 replicates) or fast-sequence addition (10 000 replicates) approaches were also performed to determine the robustness of nodes.

Acknowledgements

We wish to thank Haobo Jiang and Zhen Zou for sharing information related to *Tr. castaneum* genome annotations, and Haobo Jiang and Mike Kanost for critical comments on the manuscript. This is contribution #06–299-J from the Kansas Agricultural Experiment Station. This work was supported by USDA-CSREES-RAMP (Agreement no. 2004-51001-02226), Civil Research and Development Foundation (Grant No RB2-2396-MO-02), International Science and Technology Center project #3455 and Russian Foundation for Basic Research (Grant no. 06-04-49147a). Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G. *et al*. (2000) The genome sequence of *Drosophila melanogaster*. *Science* **287**: 2185–2195.

Applebaum, S.W., Birk, Y., Harpaz, I. and Bondi, A. (1964) Comparative studies on proteolytic enzymes of *Tenebrio molitor* L. *Comp Biochem Physiol* **11**: 85–103.

Barrett, A.J., Rawlings, N.D. and Woessner, J.F. (1998) Introduction: family S1 of trypsin (clan SA). In *Handbook of Proteolytic Enzymes* (Barrett, A.J., Rawlings, N.D. and Woessner, J.F., eds), pp. 3–12. Academic Press, New York, NY.

Bown, D.P. and Gatehouse, J.A. (2004) Characterization of a digestive carboxypeptidase from the insect pest corn earworm (*Helicoverpa armigera*) with novel specificity towards C-terminal glutamate residues. *Eur J Biochem* **271**: 2000–2011.

Bown, D.P., Wilkinson, H.S. and Gatehouse, J.A. (1997) Differentially regulated inhibitor-sensitive and insensitive protease genes from the phytophagous insect pest, *Helicoverpa armigera*, are members of complex multigene families. *Insect Biochem Mol Biol* **27**: 625–638.

Bown, D.P., Wilkinson, H.S. and Gatehouse, J.A. (1998) Midgut carboxypeptidase from *Helicoverpa armigera* (Lepidoptera: Noctuidae) larvae: enzyme characterisation, cDNA cloning and expression. *Insect Biochem Mol Biol* **28**: 739–749.

Bown, D.P., Wilkinson, H.S. and Gatehouse, J.A. (2004) Regulation of expression of genes encoding digestive proteases in the gut of a polyphagous lepidopteran larva in response to dietary protease inhibitors. *Physiol Entomol* **29**: 278–290.

Bown, D.P., Wilkinson, H.S., Jongsma, M.A. and Gatehouse, J.A. (2004) Characterisation of cysteine proteinases responsible for digestive proteolysis in guts of larval western corn rootworm (*Diabrotica virgifera*) by expression in the yeast *Pichia pastoris*. *Insect Biochem Mol Biol* **34**: 305–320.

Brito, L.O., Lopes, A.R., Parra, J.R.P., Terra, W.R. and Silva-Filho, M.C. (2001) Adaptation of tobacco budworm *Heloithis virescens* to proteinase inhibitors may be mediated by the synthesis of new proteinases. *Comp Biochem Physiol* **128B**: 365–375.

Campbell, C.L., Vandyke, K.A., Letchworth, G.J., Drolet, B.S., Hanekamp, T. and Wilson, W.C. (2005) Midgut and salivary gland transcriptomes of the arbovirus vector *Culicoides sonorensis* (Diptera: Ceratopogonidae). *Insect Mol Biol* **14**: 121–136.

Casu, R.E., Jarmey, J.M., Elvin, C.M. and Eisemann, C.H. (1994) Isolation of a trypsin-like serine protease gene family from the sheep blowfly *Lucilia cuprina*. *Insect Mol Biol* **3**: 159–170.

Christianson, D.W. and Lipscomb, W.N. (1986) X-ray crystallographic investigation of substrate binding to carboxypeptidase A at subzero temperature. *Proc Natl Acad Sci USA* **83**: 7568–7572.

Cohen, L.W., Coghlan, V.M. and Dihel, L.C. (1986) Cloning and sequencing of papain-encoding cDNA. *Gene* **48**: 219–227.

Craik, C.S., Roczniak, S., Largman, C. and Rutter, W.J. (1987) The catalytic role of the active site aspartic acid in serine proteases. *Science* **237**: 909–913.

Cristofoletti, P.T. and Terra, W.R. (1999) Specificity, anchoring, and subsites in the active center of a microvillar aminopeptidase purified from *Tenebrio molitor* (Coleoptera) midgut cells. *Insect Biochem Mol Biol* **29**: 807–819.

Cristofoletti, P.T. and Terra, W.R. (2000) The role of amino acid residues in the active site of a midgut microvillar aminopeptidase from the beetle *Tenebrio molitor. Biochim Biophys Acta* **1479**: 185–195.

Cristofoletti, P.T., Ribeiro, A.F. and Terra, W.R. (2005) The cathepsin L-like proteinases from the midgut of *Tenebrio molitor* larvae: Sequence, properties, immunocytochemical localization and function. *Insect Biochem Mol Biol* **35**: 883–901.

De Oliveira Neto, O.B., Batista, J.A., Rigden, D.J., Franco, O.L., Fragoso, R.R., Monteiro, A.C., Monnerat, R.G., Grossi-De, Sa and M.F. (2004) Molecular cloning of a cysteine proteinase cDNA from the cotton boll weevil *Anthonomus grandis* (Coleoptera: Curculionidae). *Biosci Biotechnol Biochem* **68**: 1235–1242.

Dimopoulos, G., Richman, A., Muller, H.M. and Kafatos, F.C. (1997) Molecular immune responses of the mosquito *Anopheles gambiae* to bacteria and malaria parasites. *Proc Natl Acad Sci USA* **94**: 11508–11513.

Edwards, M.J., Lemos, F.J., Donnelly-Doman, M. and Jacobs-Lorena, M. (1997) Rapid induction by a blood meal of a carboxypeptidase gene in the gut of the mosquito *Anopheles gambiae*. *Insect Biochem Mol Biol* **27**: 1063–1072.

Edwards, M.J., Moskalyk, L.A., Donelly-Doman, M., Vlaskova, M., Noriega, F.G., Walker, V.K. and Jacobs-Lorena, M. (2000) Characterization of a carboxypeptidase A gene from the mosquito, *Aedes aegypti. Insect Mol Biol* **9**: 33–38.

Elpidina, E.N., Tsybina, T.A., Dunaevsky, Y.E., Belozersky, M.A., Zhuzhikov, D.P. and Oppert, B. (2005) A chymotrypsin-like proteinase from the midgut of *Tenebrio molitor* larvae. *Biochimie* **87**: 771–779.

Elvin, C.M., Vuocolo, T., Smith, W.J., Eisemann, C.H. and Riddles, P.W. (1994) An estimate of the number of serine protease genes expressed in sheep blowfly larvae (*Lucilia cuprina*). *Insect Mol Biol* **3**: 105–115.

Elvin, C.M., Whan, V. and Riddles, P.W. (1993) A family of serine protease genes expressed in adult buffalo fly *Haematobia irritans exigua. Mol Gen Genet* **240**: 132–139.

Emi, M., Nakamura, Y., Ogawa, M., Yamamoto, T., Nishide, T., Mori, T. and Matsubara, K. (1986) Cloning, characterization and nucleotide sequences of two cDNAs encoding human pancreatic trypsinogens. *Gene* **41**: 305–310.

Ferriera, C., Bellinello, G.L. and Ribeiro, A.F. and Terra, W.R. (1990) Digestive enzymes associated with glycocalyx, microvillar membranes and secretory vesicles from midgut cells of *Tenebrio molitor* larvae. *Insect Biochem* **20**: 839–847.

Garty, N. (1979) Isolation and characterization of a chymotrypsin-like enzyme from *Tenebrio molitor* larvae. MSc Thesis, Faculty of Agriculture, The Hebrew University, Rehovot, Israel.

Gilmore, B.F., Harriott, P. and Walker, B. (2005) The inactivation of bovine cathepsin B by novel N-chloro-acetyl-dipeptides: application of the Houghten 'tea bag' methodology to inhibitor synthesis. *Biochem Biophys Res Commun* **333**: 1284–1288.

Golan, R. (1981) Isolation, characterization and comparative study of proteolytic enzymes from the midguts of *Tenebrio molitor* adults and larvae as a basis of possible biological pest control with naturally occurring protease inhibitors from plant sources. MSc Thesis, Faculty of Agriculture, The Hebrew University, Rehovot, Israel.

Hedstrom, L., Szilagyi, L. and Rutter, W.J. (1992) Converting trypsin to chymotrypsin: the role of surface loops. *Science* **255**: 1249–1253.

Holt, R.A., Subramanian, G.M., Halpern, A., Sutton, G.G., Charlab, R., Nusskern, D.R. *et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**: 129–149.

Huang, X. and Madan, A. (1999) CAP3: a DNA sequence assembly program. *Genome Res* **9**: 868–877.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Jiang, H., Wang, Y., Gu, Y., Guo, X., Zou, Z., Scholz, F., Trenczek, T.E. and Kanost, M.R. (2005) Molecular identification of a bevy of serine proteinases in *Manduca sexta* hemolymph. *Insect Biochem Mol Biol* **35**: 931–943.

Kawabata, S., Tokunaga, F., Kugi, Y., Motoyama, S., Miura, Y., Hirata, M. and Iwanaga, S. (1996) Limulus factor D, a 43 kDa protein isolated from horseshoe crab hemocytes, is a serine proteinase homologue with antimicrobial activity. *FEBS Lett* **298**: 146–150.

Koiwa, H., Shade, R.E., Zhu-Salzman, K., D'Urzo, M.P., Murdock, L.L., Bressan, R.A. and Hasegawa, P.M. (2000) A plant defensive cystatin (soyacystatin) targets cathepsin L-like digestive cysteine proteinases (DvCALs) in the larval midgut of western corn rootworm (*Diabrotica virgifera virgifera*). *FEBS Lett* **471**: 67–70.

Kraut, J. (1971) Chymotrypsinogen: X-ray structure. In *The Enzymes*, Vol. 3 (Boyer, P.D., ed.), pp. 165–183, Academic Press, New York, NY.

Laemmli, U.K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**: 680–685.

Le Huerou, I., Wicker, C., Guilloteau, P., Toullec, R. and Puigserver, A. (1990) Isolation and nucleotide sequence of cDNA clone for bovine pancreatic anionic trypsinogen. Structural identity within the trypsin family. *Eur J Biochem* **193**: 767–773.

Le Huerou, I., Guilloteau, P., Toullec, R., Puigserver, A. and Wicker, C. (1991) Cloning and nucleotide sequence of a bovine pancreatic preprocarboxypeptidase A cDNA. *Biochem Biophys Res Commun* **175**: 110–116.

Levinsky, H., Birk, Y. and Applebaum, S.W. (1977) Isolation and characterization of a new trypsin-like enzyme from *Tenebrio molitor* L. larvae. *Int J Peptide Protein Res* **10**: 252–264.

Liu, X., Fellers, J.P., Zhu, Y.C., Mutti, N.S., El-Bouhssini, M. and Chen, M.S. (2006) Cloning and characterization of cDNAS encoding carboxypeptidase-like proteins from the gut of Hessian fly larvae [*Mayetiola destructor* (Say)]. *Insect Biochem Mol Biol* **36**: 665–673.

McArthur, A.G., Morrison, H.G., Nixon, J.E., Passamaneck, N.Q., Kim, U., Hinkle, G. *et al.* (2000) The Giardia genome project database. *FEMS Microbiol Lett* **189**: 271–273.

Mittapalli, O., Wise, I.L. and Shukle, R.H. (2006) Characterization of a serine carboxypeptidase in the salivary glands and fat body of the orange wheat blossom midge, *Sitodiplosis mosellana* (Diptera: Cecidomyiidae). *Insect Biochem Mol Biol* **36**: 154–160.

Müller, H.-M., Crampton, J.M., Torre, A.D., Sinden, R. and Crisnti, A. (1993) Members of a trypsin gene family in *Anopheles gambiae* are induced in the gut by blood meal. *EMBO J* **12**: 2891–2900.

Musil, D., Zucic, D., Eng, R.A., Mayr, I., Huber, R., Popovič, T., Turk, V., Towatari, T., Katunuma, N. and Bode, W. (1991) The refined 2.15 Å x-ray crystal structure of human liver cathepsin B: the structural basis for its specificity. *EMBO J* **10**: 2321–2330.

Neurath, R. (1989) The diversity of proteolytic enzymes. In *Proteolytic Enzymes, a Practical Approach* (Beynon, R.J. and Bond, J.S., eds), pp. 1–23. Oxford University Press, Oxford, UK.

Oppert, B., Morgan, T.D., Culbertson, C. and Kramer, K.J. (1993) Dietary mixtures of cysteine proteinase and serine proteinase inhibitors exhibit increased toxicity toward the red flour beetle, *Tribolium castaneum*. *Comp Biochem Physiol* **105C**: 379–385.

Oppert, B., Morgan, T.D., Hartzer, K., Lenarcic, B., Galesa, K., Brzin, J., Turk, V., Yoza, K., Ohtsubo, K. and Kramer, K.J. (2003) Effects of proteinase inhibitors on growth and digestive proteolysis of the red flour beetle, *Tribolium castaneum* (Herbst) (Coleoptera: Tenebrionidae). *Comp Biochem Physiol* **134C**: 481–490.

Oppert, B., Morgan, T.D. and Kramer, K.J. (2004) Inhibitor strategies to control coleopteran pests. In *International Congress Series, Animals and Environments: Proceedings of the Third International Conference of Comparative Physiology and Biochemistry*, Vol. 1275 (Morris, S. and Vosloo, A., eds), pp. 149–156. Elsevier, Amsterdam, The Netherlands.

Oppert, B., Morgan, T.D., Hartzer, K. and Kramer, K.J. (2005) Compensatory proteolytic responses to dietary proteinase inhibitors in the red flour beetle, *Tribolium castaneum* (Herbst) (Coleoptera: Tenebrionidae). *Comp Biochem Physiol* **140C**: 53–58.

Pedra, J.H., Brandt, A., Westerman, R., Lobo, N., Li, H.-M., Romero-Severson, J., Murdock, L.L. and Pittendrigh, B.R. (2003) Transcriptome analysis of the cowpea weevil bruchid: Identification of putative proteinases and α-amylases associated with food breakdown. *Insect Mol Biol* **12**: 405–412.

Perona, J.J. and Craik, C.S. (1995) Structural basis of substrate specificity in the serine proteases. *Protein Sci* **4**: 337–360.

Ramos, A., Mahowald, A. and Jacobs-Lorena, M. (1993) Gut-specific genes from the black fly *Simulium vittatum* encoding trypsin-like and carboxypeptidase-like proteins. *Insect Mol Biol* **1**: 149–163.

Rawlings, N.D. and Barrett, A.J. (1993) Evolutionary families of peptidases. *Biochem J* **290**: 205–218.

Rawlings, N.D., Morton, F.R. and Barrett, A.J. (2006) MEROPS: the peptidase database. *Nucleic Acids Res* **34**: D270–D272.

Ross, J., Jiang, H., Kanost, M.R. and Wang, Y. (2003) SPs and their homologs in the D*rosophila melanogaster* genome: An initial analysis of sequence conservation and phylogenetic relationship. *Gene* **304**: 117–131.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Miklos, G.L.G., Nelson, C.R. and Hariharan, I.K. (2000) Comparative genomics of the eukaryotes. *Science* **287**: 2204–2215.

Strausberg, R.L., Feingold, E.A., Grouse, L.H., Derge, J.G., Klausner, R.D., Collins, F.S. *et al.* (2002) Generation and initial analysis of more than 15000 full-length human and mouse cDNA sequences. *Proc Natl Acad Sci USA* **99**: 16899–16903.

Swofford, D.L. (2002) PAUP* Phylogenetic analysis using parsimony (and other methods), version 4.0b10. Sinauer Associates, Sunderland, Massachusetts.

Terra, W.R. and Cristofoletti, P.T. (1996) Midgut proteinases in three divergent species of Coleoptera. *Comp Biochem Physiol* **113B**: 725–730.

Terra, W.R. and Ferreira, C. (1994) Insect digestive enzymes: properties, compartmentalization and function. *Comp Biochem Physiol* **109B**: 1–62.

Terra, W.R., Ferreira, C. and Bastos, F. (1985) Phylogenetic consideration of insect digestion. Disaccharidases and the spatial organization of digestion in the *Tenebrio molitor* larvae. *Insect Biochem* **15**: 443–449.

Thie, N.M.R. and Houseman, J.G. (1990) Cysteine and serine proteolytic activities in larval midgut of yellow mealworm, *Tenebrio molitor* L. (Coleoptera: Tenebrionidae). *Insect Biochem* **20**: 741–744.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acid Res* **22**: 4673–4680.

Titani, K., Ericsson, L.H., Walsh, K.A. and Neurath, H. (1975) Amino-acid sequence of bovine carboxypeptidase B. *Proc Natl Acad Sci USA* **72**: 1666–1670.

Tsybina, T.A., Dunaevsky, Y.E., Belozersky, M.A., Zhuzhikov, D.P., Oppert, B. and Elpidina, E.N. (2005) Digestive proteinases of yellow mealworm (*Tenebrio molitor*) larvae. Purification and characterization of a trypsin-like proteinase. *Biochem (Mosc)* **70**: 300–305.

Urieli, N. (1982) Isolation and characterization of a chymotrypsin-like enzyme from the digestive tract of *Tenebrio molitor and Locusta migratoria*. M. Sc. thesis, Faculty of Agriculture, The Hebrew University, Rehovot, Israel.

Vinokurov, K.S., Elpidina, E.N., Oppert, B., Prabhakar, S., Zhuzhikov, D.P., Dunaevsky, Y.E. and Belozersky, M.A. (2006a) Diversity of digestive proteinases in *Tenebrio molitor* (Coleoptera: Tenebrionidae) larvae. *Comp Biochem Physiol B* **145**: 126–137.

Vinokurov, K.S., Elpidina, E.N., Oppert, B., Prabhakar, S., Zhuzhikov, D.P., Dunaevsky, Y.E. and Belozersky, M.A. (2006b) Fractionation of digestive proteinases from *Tenebrio molitor* (Coleoptera: Tenebrionidae) larvae and role in protein digestion. *Comp Biochem Physiol B* **145**: 138–146.

Wagner, W., Möhrlen, F. and Schnetter, W. (2002) Characterization of the proteolytic enzymes in the midgut of the European Cockchafer, *Melolontha melolontha* (Coleoptera: Scarabaeidae). *Insect Biochem Mol Biol* **32**: 803–814.

Wang, P., Li, G. and Kain, W. (2004) Characterization and cDNA cloning of midgut carboxypeptidases from *Trichoplusia ni*. *Insect Biochem Mol Biol* **34**: 831–843.

Wang, S., Magoulas, C. and Hickey, D.A. (1993) Isolation and characterization of a full-length trypsin-encoding cDNA clone from the Lepidopteran insect, *Choristoneura fumiferana*. *Gene* **136**: 375–376.

Xu, X., Dong, Y., Abraham, E.G., Kocan, A., Srinivasan, P., Ghosh, A.K., Sinden, R.E., Ribeiro, J.M., Jacobs-Lorena, M., Kafatos, F.C. and Dimopoulos, G. (2005) Transcriptome analysis of *Anopheles stephensi–Plasmodium berghei* interactions. *Mol Biochem Parasitol* **142**: 76–87.

Zdobnov, E.M., von Mering, C., Letunic, I., Torrents, D., Suyama, M., Copley, R.R. *et al.* (2002) Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. *Science* **298**: 149–159.

Zhu-Salzman, K., Koiwa, H., Salzman, R.A., Shade, R.E. and Ahn, J.-E. (2003) Cowpea bruchid *Callosobruchus maculatus* uses a three-component strategy to overcome a plant defensive cysteine protease inhibitor. *Insect Mol Biol* **12**: 135–145.

Zwilling, R. (1968) Zur Evolution der Endopeptidasen-IV. α- and β-protease aus *Tenebrio molitor*. *Z Physiol Chem* **349**: 326–332.

Zwilling, R., Medugorac, I. and Mella, K. (1972) The evolution of endopeptidases-XIV. Non-tryptic cleavage specificity of a BAEE-hydrolyzing enzyme (β-protease) from *Tenebrio molitor*. *Comp Biochem Physiol* **43**: 419–424.

Supplementary material

The following supplementary material is available for this article online:

Figure S1. Phylogenetic trees comparing *Tenebrio molitor* and related sequences.

This material is available as part of the online article from http://www.blackwell-synergy.com.