Data in Brief 38 (2021) 107301

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

The dataset of predicted trypsin serine peptidases and their inactive homologs in *Tenebrio molitor* transcriptomes

Nikita I. Zhiganov[a,1], Valeriia F. Tereshchenkova[b,1], Brenda Oppert[c], Irina Y. Filippova[b], Nataliya V. Belyaeva[a], Yakov E. Dunaevsky[d], Mikhail A. Belozersky[d], Elena N. Elpidina[d,*]

[a] *Division of Entomology, Faculty of Biology, Lomonosov Moscow State University, Moscow, Russia*
[b] *Division of Natural Compounds, Department of Chemistry, Lomonosov Moscow State University, Moscow, Russia*
[c] *USDA Agricultural Research Service, Center for Grain and Animal Health Research, Manhattan, KS 66502, USA*
[d] *Department of Plant Proteins, A.N. Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow, Russia*

ARTICLE INFO

Article history:
Received 6 May 2021
Revised 12 August 2021
Accepted 13 August 2021
Available online 16 August 2021

Keywords:
Serine peptidases
Pseudopeptidases
Inactive peptidase homologs
Trypsin
Tenebrio molitor
Tenebrionidae

ABSTRACT

Tenebrio molitor is an important coleopteran model insect and agricultural pest from the Tenebrionidae family. We used RNA-Seq transcriptome data from *T. molitor* to annotate trypsin-like sequences from the chymotrypsin S1 family of serine peptidases, including sequences of active serine peptidases (SerP) and their inactive homologs (SerPH) in *T. molitor* transcriptomes. A total of 63 S1 family tryspin-like serine peptidase sequences were *de novo* assembled. Among the sequences, 58 were predicted to be active trypsins and five inactive SerPH. The length of preproenzyme and mature form of the predicted enzyme, position of signal peptide and proenzyme cleavage sites, molecular mass, active site and S1 substrate binding subsite residues, and transmembrane and regulatory domains were analyzed using bioinformatic tools. The data can be used for further physiological, biochemical, and phylogenetic study of tenebrionid pests and other animal systems.

[*] Corresponding author.
E-mail address: elp@belozersky.msu.ru (E.N. Elpidina).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.dib.2021.107301

Specifications Table

Subject	Bioinformatics/Omics: Transcriptomics
Specific subject area	Assembly, annotation and structural analysis of trypsin-like SerP and SerPH sequences from the *Tenebrio molitor* transcriptome using bioinformatics tools.
Type of data	Tables and Supplementary text file, containing assembled sequences in fasta format.
How data were acquired	RNA sequencing by Illumina HiSeq 2000, structural analysis with appropriate software
Data format	Raw
Parameters for data collection	Laboratory colonies of *T. molitor* were maintained on a diet consisting of either 50–100% oat flakes or 85% stabilized wheat germ. RNA was isolated and sequenced from the larval midgut and from eggs, I and IV instar larvae, pupae, young males and females with non-chitinized integuments and two week old adult males and females of *T. molitor*. Parameters for sequencing, contigs assembly and analysis programs were default.
Description of data collection	mRNA-isolation, high throughput RNA-sequencing using Illumina HiSeq 2000, automated and manual assembly of contigs to full-length sequences, annotation and structural analysis using bioinformatic tools.
Data source location	Institution: USDA Agricultural Research Service, Center for Grain and Animal Health Research City/Town/Region: Manhattan, KS 66502 Country: USA Latitude and longitude: 27.2046° N, 77.4977° E Institution: Lomonosov Moscow State University, A.N. Belozersky Institute of Physico-Chemical Biology City/Town/Region: Moscow, 119992 Country: Russia Latitude and longitude: 55.705836° N, 37.521456° E
Data accessibility	Repository name: NCBI Data identification number: Accession numbers unique to each sequence are listed in Tables 1 and 2. Direct URL to data: Direct URLs to data are presented as hyperlinks in Tables 1 and 2, which can be opened by clicking on the corresponding accession number. Data are also available with the article in the Supplementary file.

Value of the Data

- Serine peptidases of the S1 chymotrypsin family and their inactive homologs were identified in transcriptome data from *Tenebrio molitor*, a serious pest of stored products and a biochemical and coleopteran model. This information is important for clarification of the organization of the digestive process in Tenebrionidae insects as well as the role of serine peptidases in life functions of the insect.
- The data obtained can help agricultural industries and researchers in developing measures to control these pests, as well as food industries incorporating insects as supplemental protein.
- The dataset revealed the presence of inactive serine peptidases homologs (pseudoenzymes) as have been observed in several other insects. The study can contribute to understanding their functional significance and, in particular, their role in the regulation of the action of active peptidases.
- These data form the basis for further bioinformatic and biochemical analyzes to clarify the relationship between the structure and function of serine peptidases and their inactive homologs.

1. Data Description

The sequences of trypsins and their inactive homologs from the S1 chymotrypsin family, according to MEROPS classification [1], are attached to this article in the Supplementary file with the reference accession numbers unique to each sequence received during submission to the open public repository NCBI database (https://www.ncbi.nlm.nih.gov) [2] and listed in Tables 1 and 2. Direct URLs to data are presented as hyperlinks in Tables 1 and 2, which can be opened by clicking on the corresponding accession number. The data represent assembled transcriptome sequences of serine peptidases and their inactive homologs for which structural analysis was carried out using bioinformatic tools. The data include 58 sequences of trypsins and five trypsin-like SerPH that were found in the *T. molitor* transcriptome obtained via a combination of data from the larval midgut, eggs, I and IV instar whole larvae, pupae, adult males and females. Of these, the sequences of 34 active trypsins and five SerPH were found in the larval midgut transcriptome (Table 1) and 24 active trypsins were absent in the larval midgut transcriptome and are considered non-gut peptidases (Table 2). The custom ID and corresponding NCBI accession numbers are listed in Tables 1 and 2. Sequences SerP1 [3], SerP183 [4], SerPH223 [5] and SerPH415 [6] were previously submitted by other research groups and coincided with those assembled by us, so we used existing NCBI accession numbers with relevant references.

From each S1 peptidase sequence, we analyzed the length of the predicted preproenzyme and mature enzyme form, molecular mass of the mature enzyme, active site and S1 substrate binding subsite residues, position of signal peptide and proenzyme cleavage sites, and the presence of a transmembrane domain. We also identified regulatory domains: clip_1 and clip_2 [7], carbohydrate-binding module family 14 (CBM_14) domain, low density lipoprotein receptor gene family (ldl_recept_a), scavenger receptor cysteine-rich (SRCR) domain, PAN_1 - (1) the N-terminal N domains of members of the plasminogen/hepatocyte growth factor family, (2) the apple domains of the plasma prekallikrein/coagulation factor XI family, and (3) domains of various nematode proteins referred to as the PAN module, thrombospondin type 1 repeats, Epstein-Barr virus nuclear antigen 3C (EBNA-3C), Family of unknown function (DUF5585), 104 kDa microneme/rhoptry antigen.

2. Experimental Design, Materials and Methods

2.1. Preparation of biological material and RNA isolation

The midgut transcriptome data from *T. molitor* larvae were obtained at The Center for Grain and Animal Health Research (CGAHR, Manhattan, KS USA), where laboratory colonies of *T. molitor* are maintained on a diet of 50% oat flakes, 2.5% brewer's yeast, and 47.5% wheat flour at 28 °C, 75% R.H., in darkness. Approximately five-week old larvae with an average weight of 5.1 mg from three independent biological replicates were fasted overnight and were placed on a diet consisting of 85% stabilized wheat germ, 10% wheat flour, and 5% brewer's yeast for 12 h. For each replicate, the midgut was extracted from 4 to 7 larvae and placed in room temperature RNAlater (Ambion, Austin TX USA). For RNA isolation, excess RNAlater was blotted, and pooled midguts were ground with a plastic pestle in 1.5 ml microfuge tube containing liquid nitrogen. Total RNA was isolated using the Absolutely RNA Kit with DNase on-column treatment (Agilent, La Jolla, CA USA).

Whole-body transcriptomes from different stages of the life cycle of *T. molitor* - eggs, larvae of the 1st instar (the first week after hatching), larvae of the 4th instar (in the fifth week after hatching), pupae, young males and females with non-chitinized integuments, and two weeks old adult males and females were obtained from the laboratory colony at the Lomonosov Moscow State University (Moscow, Russia). All stages were maintained on milled oat flakes at 26 ± 0.5 °C and 75% relative humidity. Two replicates were obtained for eggs, 1st and 4th instar larvae, pupae, adult males and young females; one replicate was obtained for young males and adult fe-

Table 1
Sequences of trypsins (SerP) and tripsin-like homologs (SerPH) found in the larval gut transcriptome of *Tenebrio molitor*. All active sites of active serine peptidase sequences consist of H D S, except for SerPH70 (H D D) and other SerPH (H D G), and all S1 subsites are D G G.

Identification	NCBI Nucleotide Accession	NCBI Protein Accession	Preproenzyme/Mature Enzyme (aa)	Mature Enzyme (Da)	Signal Peptide (aa)	Proenzyme Cleavage Site	Regulatory Domains(aa position)
SerP1[1]	DQ356014	ABC88729	258/227	22,742	16	R|IVGG	-
SerP2	MW603455	QWS65012	252/227	23,618	16	R|IVGG	-
SerP4	MW603456	QWS65013	250/225	24,140	15	R|IVGG	-
SerP6	MW603457	QWS65014	258/226	23,414	17	R|IVGG	-
SerP11	MW603486	QWS65043	447/231	26,306	19	K|IIGG	Thrombospondin type 1 repeats (132-170)
SerP30	MW603458	QWS65015	249/226	24,862	16	K|IIGG	-
SerP40	MW603459	QWS65016	392/241	26,535	21[*]	G|NPGG	Clip_1[5] (68-112)
SerP48	MW603460	QWS65017	321/295	32,018	22	R|IVGG	-
SerP55	MW603461	QWS65018	1,672/245	26,947	21	R|VVRG	CBM_14 (138-189, 212-263, 303-354), Ldl_recept_a (969-1003, 1089-1122, 1240-1279), SRCR (1134-1236, 1288-1345), PAN_1 (1003-1084)
SerP76	MW603462	QWS65019	387/362	39,417	-	K|IIGG	-
SerP84	MW603463	QWS65020	332/308	33,286	18	K|VVGG	-
SerP86	MW603464	QWS65021	458/258	28,226	22	R|ILDG	Clip_2[5] (28-78)
SerP113	MW603465	QWS65022	386/255	28,255	23	R|IING	Clip_2 (32-87)
SerP119	MW603466	QWS65023	387/253	28,333	26	L|IVGG	Clip_1 (34-80)
SerP125	MW603467	QWS65024	278/254	27,535	19	R|IVGG	-
SerP127	MW603468	QWS65025	376/247	27,075	22	R|IVNG	Clip_2 (27-80)
SerP131	MW603469	QWS65026	375/247	26,799	22	R|VVNG	Clip_2 (30-83)
SerP135	MW603470	QWS65027	292/246	26,850	22	G|IIGG	-
SerP141	MW603471	QWS65028	444/259	28,844	-	R|IFGG	Clip_2 (102-152)
SerP145	MW603472	QWS65029	370/241	26,781	22	H|IVGG	Clip_1 (31-76)
SerP163	MW603473	QWS65030	354/254	28,041	21[*]	V|IAFG	Clip_1 (28-73)
SerP173	MW603474	QWS65031	347/249	27,495	19	F|VFGG	Clip_1 (26-71)
SerP183[2]	AB363980	BAG14262	383/265	29,203	18	R|IYGG	Clip_2 (21-74)
SerP193	MW603475	QWS65032	375/247	27,564	22[*]	R|ILGG	Clip_2 (48-97)
SerP209	MW603476	QWS65033	258/227	22,943	16	R|IIGG	-
SerP227	MW603477	QWS65034	376/251	27,969	23	L|IVGG	Clip_1 (31-76)
SerP228	MW603478	QWS65035	374/250	27,849	20	L|IVGG	Clip_1 (28-74)
SerP247	MW603479	QWS65036	379/257	28,343	18	T|IISM	Clip_1 (26-71)
SerP266	MW603480	QWS65037	281/256	27,895	18	K|IVGG	-
SerP272	MW603481	QWS65038	404/297	32,710	17[*]	K|IYGG	Clip_2 (21-74)

(*continued on next page*)

Table 1 (*continued*)

Identification	NCBI Nucleotide Accession	NCBI Protein Accession	Preproenzyme/Mature Enzyme (aa)	Mature Enzyme (Da)	Signal Peptide (aa)	Proenzyme Cleavage Site	Regulatory Domains(aa position)
SerP282	MW603482	QWS65039	328/270	29,212	17*	G|ITGG	Clip_1 (27-53)
SerP345	MW603483	QWS65040	359/234	26,360	22	L|IVGG	Clip_1 (30-76)
SerP370	MW603484	QWS65041	407/257	28,048	23*	K|ISNG	Clip_2 (27-69, 78-121)
SerP409	MW603485	QWS65042	447/234	26,142	22*	K|IGKG	Clip_1? (58-92), Clip_2 (148-195)
SerPH70	MW419917	QRE01765	280/249	27,206	19	K|IVGG	-
SerPH216	MW419918	QRE01766	348/254	28,298	16	K|IGND	Clip PPAF-2 (21-66)
SerPH223[3]	AJ400904	CAC12696	400/264	29,412	21	R|ITGN	Clip PPAF-2 (64-104)
SerPH235	MW419919	QRE01767	407/263	28,787	16	K|ITGN	Clip PPAF-2 (53-96)
SerPH415[4]	AB084067	BAC15605	444/261	28,711	15	N|LIGG	Clip PPAF-2 (68-114)

[1] Prabhakar et al., 2007

[2] Kim et al., 2008

[3] Kwon et al., 2000

[4] Lee et al., 2002

[5] Clip domains were grouped manually (Clip_1 and Clip_2) and position curated according to [7].

* Signal peptide cleavage site probability is less than 0.5.

Table 2
Non-gut trypsin sequences found only in the combined transcriptome of *Tenebrio molitor* and absent in the larval gut transcriptome. All active sites consist of H D S, and S1 subsites are D G G.

Identification	NCBI Nucleotide Accession	NCBI Protein Accession	Preproenzyme/Mature Enzyme (aa)	Mature Enzyme (Da)	Signal Peptide (aa)	Proenzyme Cleavage Site	Regulatory Domains (aa position)
NGSerP3	MW628465	QWS65044	259/228	24,386	16	K|IVGG	-
NGSerP5	MW628466	QWS65045	333/236	26,035	24	R|IVGG	-
NGSerP14	MW628467	QWS65046	1293/286	31,448	-	R|IVGG	Ldl_recept_a (774–808, 842–878, 1260–1289), SRCR (900–995)
NGSerP15	MW628468	QWS65047	516/235	25,699	23	R|IVGG	EBNA-3C (201–246)
NGSerP20	MW628469	QWS65048	361/238	26,395	17*	R|IVGG	-
NGSerP21	MW628470	QWS65049	276/228	24,732	22	R|IVGG	-
NGSerP22	MW628471	QWS65050	290/242	25,975	17*	R|VVGG	-
NGSerP24	MW628472	QWS65051	810/243	27,555	19	R|IVGG	Family of unknown function (DUF5585) (259–544)
NGSerP26	MW628476	QWS65055	254/227	24,214	23	R|IVGG	-
NGSerP27	MW628473	QWS65052	369/242	26,704	19	R|IVGG	-
NGSerP28	MW628477	QWS65056	310/241	27,033	26	R|IVGG	-
NGSerP35	MW628478	QWS65057	260/231	24,884	21*	R|IVGG	-
NGSerP37	MW628479	QWS65058	298/251	27,327	19	R|IVNG	-
NGSerP65	MW628474	QWS65053	619/240	26,001	20	R|IVGG	-
NGSerP66	MW628480	QWS65059	523/245	27,560	-	R|VVNG	Clip_1[1] (164–207)
NGSerP77	MW628481	QWS65060	288/252	27,164	17	G|IIGG	-
NGSerP104	MW628482	QWS65061	332/300	32,646	18	R|IFGG	-
NGSerP109	MW628483	QWS65062	964/247	26,987	17	H|IVGG	104 kDa microneme/rhoptry antigen (291–529)
NGSerP116	MW628484	QWS65063	381/257	28,382	16	V|IAFG	Clip_2 (25–80)
NGSerP166	MW628485	QWS65064	376/259	28,449	15	F|VFGG	Clip_2 (23–78)
NGSerP275	MW628486	QWS65065	430/257	28,969	23	R|IYGG	Clip_2[1] (32–75, 80–135)
NGSerP297	MW628487	QWS65066	350/255	28,238	18	R|ILGG	Clip_1 (25–70)
NGSerP317	MW628475	QWS65054	389/246	27,195	16	R|IIGG	-
NGSerP347	MW628488	QWS65067	367/256	28,001	25*	R|IIGG	Clip_1 (36–81)

[1] Clip domains were grouped manually (Clip_1 and Clip_2) and position curated according to [7].

* Signal peptide cleavage site probability is less than 0.5.

males. RNA was extracted using the RNEasy Mini kit (Qiagen, Hilden, Germany). Immediately prior to isolation, the samples were homogenized by trituration in liquid nitrogen. The concentration of isolated RNA was measured on a Qubit (Thermofisher, Waltham, MA USA) fluorimeter using a set of reagents for high-sensitivity RNA analysis. The integrity of the RNA was checked by capillary electrophoresis on a Bioanalyzer 2100 (Agilent). The NEBNext RNA Library Prep Kit for Illumina (New England Biolabs, Ipswich, MA USA) was used to prepare the libraries according to recommended protocol with a fragmentation time of 5 min.

2.2. Sequencing of cDNA

The resulting midgut total RNA was sent to a sequencing facility (National Center for Genome Resources - NCGR, Santa Fe, NM, USA), where mRNA was isolated by polyA, standard libraries were made, and paired-end sequencing was performed on a Illumina HiSeq 2000 (San Diego, CA, USA) using standard protocols from the manufacturer. Approximately 240 million sequence reads were obtained, with an approximate 250 bp insert.

The libraries of different *T. molitor* developmental stages were sequenced on a Illumina HiSeq 2000 (Lomonosov Moscow State University, Moscow, Russia) using the TruSeq SBS Kit v3 reagent kit (200 cycles) with the following settings: read length 101, index read length 7, reverse reading length 101. The preprocessed samples contained from 7 million to 24 million reads.

2.3. Assembly of contigs and final sequences

Assembly of *T. molitor* larval midgut sequences was performed *de novo* with SeqManNGen (v. 4.0.1.4, DNAStar, Madison, WI USA) and described in [8]. It included NCGR assembly from all replicates, resulting in 197,800 contigs (N50 = 2232), previous databases of Sanger sequencing [3] and pyrosequencing [9] of mRNA from the larval gut.

The whole *T. molitor* transcriptome assembly was performed with SeqManNGen (v 15.0.0.160, default parameters) and included the midgut assembly described above combined with the Illumina sequencing data obtained for *T. molitor* developmental stages. There were 342,592,161 total reads assembled, with 143,807,206 reads not assembled and 382,435,025 removed during sampling due to read depth. Reads were assembled into 130,559 contigs, with 36,463 contigs >1 kb.

Potential coding sequences, starting at methionine and covering at least 20% of the mRNA sequence, were found in the *T. molitor* contigs using custom software. BLAST [10] and custom scripts were used to identify ORFs homologous to those encoding serine peptidases from the S1 chymotrypsin family and their inactive homologs. The sequence of human trypsin 2 (UniProt ID P07478) was used as a query. Multiple sequence alignment with BioEdit (v. 7.0.5) was used to refine and build consensus sequences, and in the case of SNPs, the amino acid chosen was the highest percentage and more than 50% of the total. ORFs that were grouped into blocks with identity of at least 95% and that overlapped with another block of at least 10 amino acid residues were considered as a unique peptidase.

2.4. Analysis of sequences

The molecular mass of the mature enzyme of the predicted protein was computed using ExPASy Server (https://web.expasy.org/compute_pi/). Signal peptide was predicted with SignalP 5.0 server (http://www.cbs.dtu.dk/services/SignalP/index.php) [11]. The start of the mature enzyme, positions of proenzyme cleavage site, active site and S1 substrate binding subsite residues were predicted by sequence homology through alignment with mature human trypsin 2 (UniProt ID P07478) using BioEdit (v. 7.0.5) and Clustal Omega multiple sequence alignment tool (https://www.ebi.ac.uk/Tools/msa/clustalo/) [12]. The presence of a transmembrane domain was predicted with TMHMM Server (v. 2.0) (http://www.cbs.dtu.dk/services/TMHMM/) [13].

Regulatory domains: clip_1 and clip_2, CBM_14, ldl_recept_a, SRCR, PAN_1, thrombospondin type 1 repeats, EBNA-3C, Family of unknown function (DUF5585), 104 kDa microneme/rhoptry antigen, were searched with HMMER web server (https://www.ebi.ac.uk/Tools/hmmer/) [14], and using NCBI Conserved Domains search service (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) [15]. Clip domains were manually curated and annotated according to [7].

Ethics statement

Institutional Review Board Statement: This research was performed in accordance with Kansas State University Research Compliance Office, Institutional Biosafety Committee registration number 1191, “Functional Genomics of Stored Product Insects”.

CRediT author statement

Nikita I. Zhiganov: Investigation, Software. **Valeriia F. Tereshchenkova:** Writing - Original Draft, Validation. **Brenda Oppert:** Supervision. **Irina Y. Filippova:** Writing - Review & Editing. **Nataliya V. Belyaeva:** Visualization. **Yakov E. Dunaevsky:** Data Curation. **Mikhail A. Belozersky:** Formal analysis. **Elena N. Elpidina:** Conceptualization, Methodology, Writing - Review & Editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Acknowledgments

This study was supported by Russian Foundation for Basic Research grant № 20-54-56044 Iran_t (collection of active serine peptidases data) and by Russian Science Foundation grant № 19-74-00080 (collection of inactive serine peptidases homologs data). 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. USDA is an equal opportunity provider and employer.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.107301.

References

[1] N.D. Rawlings, A.J. Barrett, X.Huang P.D.Thomas, A. Bateman, R.D. Finn, The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database, Nucleic Acids Res. 46 (2018) D624–D632, doi:10.1093/nar/gkx1134.

[2] NCBI Resource Coordinators, Database resources of the national center for biotechnology information, Nucleic Acids Res 46 (Database issue) (2018) D8–D13, doi:10.1093/nar/gkx1095.

[3] S. Prabhakar, M.S. Chen, E.N. Elpidina, K.S. Vinokurov, C.M. Smith, J. Marshall, B. Oppert, Sequence analysis and molecular characterization of larval midgut cDNA transcripts encoding peptidases from the yellow mealworm, *Tenebrio molitor* L, Insect Mol. Biol. 16 (4) (2007) 455–468, doi:10.1111/j.1365-2583.2007.00740.x.

[4] C.H. Kim, S.J. Kim, H. Kan, H.M. Kwon, K.B. Roh, R. Jiang, Y. Yang, J.W. Park, H.H. Lee, N.C. Ha, H.J. Kang, M. Nonaka, K. Soderhall, B.L. Lee, A three-step proteolytic cascade mediates the activation of the peptidoglycan-induced toll pathway in an insect, J. Biol. Chem. 283 (12) (2008) 7599–7607, doi:10.1074/jbc.M710216200.

[5] T.H. Kwon, M.S. Kim, H.W. Choi, C.H. Joo, M.Y. Cho, B.L. Lee, A masquerade-like serine proteinase homologue is necessary for phenoloxidase activity in the coleopteran insect, *Holotrichia diomphalia* larvae, Eur. J. Biochem. 267 (20) (2000) 6188–6196, doi:10.1046/j.1432-1327.2000.01695.x.

[6] K.Y. Lee, R. Zhang, M.S. Kim, J.W. Park, H.Y. Park, S. Kawabata, B.L. Lee, A zymogen form of masquerade-like serine proteinase homologue is cleaved during pro-phenoloxidase activation by Ca^{2+} in coleopteran and *Tenebrio molitor* larvae, Eur. J. Biochem. 269 (17) (2002) 4375–4383, doi:10.1046/j.1432-1033.2002.03155.x.

[7] H. Jiang, M.R. Kanost, The clip-domain family of serine proteinases in arthropods, Insect Biochem. Mol. Biol. 30 (2) (2000) 95–105, doi:10.1016/s0965-1748(99)00113-7.

[8] A.G. Martynov, E.N. Elpidina, L. Perkin, B. Oppert, Functional analysis of C1 family cysteine peptidases in the larval gut of *Tenebrio molitor* and *Tribolium castaneum*, BMC Genom. 16 (1) (2015) 75, doi:10.1186/s12864-015-1306-x.

[9] B. Oppert, S.E. Dowd, P. Bouffard, L. Li, A. Conesa, M.D. Lorenzen, M. Toutges, J. Marshall, D.L. Huestis, J. Fabrick, C. Oppert, J.L. Jurat-Fuentes, Transcriptome profiling of the intoxication response of *Tenebrio molitor* larvae to *Bacillus thuringiensis* Cry3Aa protoxin, PLoS One 7 (4) (2012) e34624, doi:10.1371/journal.pone.0034624.

[10] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, Lipman D.J., Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410, doi:10.1016/S0022-2836(05)80360-2.

[11] J.J. Almagro Armenteros, K.D. Tsirigos, C.K. Sønderby, T.N. Petersen, O. Winther, S. Brunak, G. von Heijne, H. Nielsen, SignalP 5.0 improves signal peptide predictions using deep neural networks, Nat. Biotechnol. 37 (4) (2019) 420–423, doi:10.1038/s41587-019-0036-z.

[12] F. Sievers, A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J.D. Thompson, D.G. Higgins, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Mol. Syst. Biol. 7 (2011) 539, doi:10.1038/msb.2011.75.

[13] A. Krogh, B. Larsson, G. von Heijne, E.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, J. Mol. Biol. 305 (3) (2001) 567–580, doi:10.1006/jmbi.2000.4315.

[14] S.C. Potter, A. Luciani, S.R. Eddy, Y. Park, R. Lopez, R.D. Finn, HMMER web server: 2018 update, Nucleic Acids Res. 46 (W1) (2018) W200–W204, doi:10.1093/nar/gky448.

[15] S. Lu, J. Wang, F. Chitsaz, M.K. Derbyshire, R.C. Geer, N.R. Gonzales, M. Gwadz, D.I. Hurwitz, G.H. Marchler, J.S. Song, N. Thanki, R.A. Yamashita, M. Yang, D. Zhang, C. Zheng, C.J. Lanczycki, A. Marchler-Bauer, CDD/SPARCLE: the conserved domain database in 2020, Nucleic Acids Res. 48 (D1) (2020) 265–268, doi:10.1093/nar/gkz991.