Insect Biochemistry and Molecular Biology 43 (2013) 189–196

Contents lists available at SciVerse ScienceDirect

Insect Biochemistry and Molecular Biology

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

LepChorionDB, a database of Lepidopteran chorion proteins and a set of tools useful for the identification of chorion proteins in Lepidopteran proteomes

Nikolaos G. Giannopoulos [a,b], Ioannis Michalopoulos [a], Nikos C. Papandreou [b], Apostolos Malatras [a,b], Vassiliki A. Iconomidou [b], Stavros J. Hamodrakas [b,*]

[a] Centre of Immunology and Transplantation, Biomedical Research Foundation, Academy of Athens, Athens 11527, Greece
[b] Department of Cell Biology and Biophysics, Faculty of Biology, University of Athens, Panepistimiopolis, Athens 15701, Greece

ARTICLE INFO

Article history:
Received 1 September 2012
Received in revised form
3 December 2012
Accepted 7 December 2012

Keywords:
Lepidoptera
Chorion protein database
A and B classes
Natural protective amyloids
Profile Hidden Markov Models (pHMMs)

ABSTRACT

Chorion proteins of Lepidoptera have a tripartite structure, which consists of a central domain and two, more variable, flanking arms. The central domain is highly conserved and it is used for the classification of chorion proteins into two major classes, A and B. Annotated and unreviewed Lepidopteran chorion protein sequences are available in various databases. A database, named LepChorionDB, was constructed by searching 5 different protein databases using class A and B central domain-specific profile Hidden Markov Models (pHMMs), developed in this work. A total of 413 Lepidopteran chorion proteins from 9 moths and 1 butterfly species were retrieved. These data were enriched and organised in order to populate LepChorionDB, the first relational database, available on the web, containing Lepidopteran chorion proteins grouped in A and B classes. LepChorionDB may provide insights in future functional and evolutionary studies of Lepidopteran chorion proteins and thus, it will be a useful tool for the Lepidopteran scientific community and Lepidopteran genome annotators, since it also provides access to the two pHMMs developed in this work, which may be used to discriminate A and B class chorion proteins. LepChorionDB is freely available at http://bioinformatics.biol.uoa.gr/LepChorionDB.

1. Introduction

Lepidoptera play an important role in ecology (pollination of plants), economy (silk production, damage to agricultural production) and health (transmission of infectious diseases, biotechnology) (Diaz, 2005; Goldsmith et al., 2005). During embryogenesis Lepidopteran eggs may be exposed to low/high temperatures, mechanical pressures, dessication, bacteria, viruses, etc., but thanks to the exceptional mechanical, physical and biological properties of the eggshell, the developing embryo is protected (Hamodrakas, 1992). The eggshell consists of a set of functional layers. Chorion, its outermost layer, constitutes 90–95% of the eggshell. About 95% of the chorion dry mass is proteinaceous (Iconomidou and Hamodrakas, 2008).

Chorion is a biological analogue of a cholesteric liquid crystal due to its helicoidal architecture (a lamellar ultrastructure of closely packed fibrils) (Bouligand, 1972; Hamodrakas, 1992; Mazur et al., 1982). The dominant secondary structure of the central region of the proteins that constitute these fibrils is a β-sheet type of structure, reminiscent in many ways of the cross-β type of structure, characteristic for amyloid fibrils. The propensity of formation of this extraordinary, natural protective amyloid, is inherent into the amino acid sequences of its constituent proteins, after millions of years of molecular evolution (Iconomidou et al., 2000, 2006).

Chorion structure is biochemically complex because it must serve many distinct functions and needs (Kafatos et al., 1977). For this reason it contains hundreds of different proteins, many of which have been isolated from various Lepidopteran species, such as *Bombyx mori* (Regier and Pacholski, 1985; Rodakis et al., 1982), *Antheraea polyphemus* or *Antheraea pernyi* (Moschonas et al., 1988). The similarity of amino acid sequences and prediction of secondary structure indicate that chorion proteins have a tripartite structure, consisting of a central region/domain and two flanking arms (Hamodrakas et al., 1982; Regier et al., 1983) (Supplementary Fig. 1). In *A. polyphemus*, the central regions of chorion proteins cover 42–48% of the total length of the sequence, with over 77% similarity

* Corresponding author. Tel.: +30 210 727 4931; fax: +30 210 727 4254.
E-mail address: shamodr@biol.uoa.gr (S.J. Hamodrakas).

0965-1748/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ibmb.2012.12.001

among them (Hamodrakas et al., 1982). Moreover, an invariant, tandem periodic repetition of glycine every 6 (six) amino acid residues is a characteristic feature of this region (Hamodrakas et al., 1985; Iconomidou and Hamodrakas, 2008). According to the degree of central region conservation and based on observed clustering in SDS gels, chorion proteins have been classified into five classes (A–E), most common of which are A and B i.e. in *A. polyphemus*, A and B class proteins represent 38% and 50% respectively of the dry mass of the chorion (Kafatos et al., 1977). The molecular weight range of A and B class chorion proteins is 9–12 kDa and 12–14 kDa, respectively (Hamodrakas, 1992). Different members of the same family are expressed at different stages during the period of choriogenesis (early, middle, late and very late stage) (Rodakis et al., 1982). In particular, in the most widely studied Lepidopteran species, *B. mori*, the chorion gene superfamily has two symmetrical branches, each consisting of three families: the α branch (A, CA, HcA families) and the β branch (B, CB, HcB families) (Lecanidou et al., 1986).

The flanking N- and C-arms show greater variability and are distinguished by the presence of characteristic peptide repeats, which are not found in the central domain. Apart from the similarities between protein sequences of the same class, similarities among the central regions/domains of proteins from different classes suggest a distant evolutionary relationship, thus strengthening the idea of a common origin from a single ancestral gene (Lecanidou et al., 1986).

Experimentally verified chorion protein sequences revealed unique motifs enabling the classification of sequences which are derived from genome or transcriptome sequencing as class A or class B chorion proteins. In this work, all such sequences were organised in a relational database, named LepChorionDB, the first database of Lepidopteran chorion proteins.

2. Methods

2.1. Lepidopteran protein sequence collection

All Lepidopteran protein sequences were downloaded from various sources, as follows: in the Entrez Protein Database (Sayers et al., 2012) and UniProt KnowledgeBase (UniProt Consortium, 2010), we searched for protein sequences of Lepidoptera (NCBI Taxonomy ID: 7088). In InsectaCentral (Papanicolaou et al., 2008), the proteins were downloaded by selecting the Lepidoptera taxon from the phylogenetic tree. In ButterflyBase (Papanicolaou et al., 2008), we downloaded the compressed files of protein sequences of Lepidopteran species. Finally, a consensus gene set created by merging all the gene sets using GLEAN and genes predicted by the BGI Gene Finder (BGF) were retrieved from SilkDB (Wang et al., 2005). All data from the databases above were downloaded in November 2011.

2.2. Hidden Markov model building

In order to identify Lepidopteran chorion proteins in our protein sequence collection, we used already existing and newly built Lepidopteran chorion protein profile Hidden Markov Models (pHMMs):

We first searched for existing Lepidopteran chorion protein pHMMs, in Pfam (Finn et al., 2010). By doing a keyword search for "chorion", Pfam returned 21 unique entries. After manual examination, we found that the only pHMM that could recognise Lepidopteran chorion proteins was "Chorion_1" (Pfam ID: PF01723) (Kravariti et al., 1995; Lecanidou et al., 1986) (Supplementary Fig. 2).

Then, to distinguish the two major Lepidopteran chorion protein classes, A and B, we constructed two pHMMs using hmmbuild from the HMMER 3.0 suite (Eddy, 2009), which were based on the conserved central regions of well characterised class A and B chorion proteins, as follows:

Class A pHMM was based on the multiple sequence alignment of *A. polyphemus* proteins pc18 (ButterflyBase ID: ALP00009_1), pc609, pc292 (UniProt ID: P02846) (Iconomidou and Hamodrakas, 2008; Jones and Kafatos, 1982) which were selected as the most appropriate and representative of this family (Fig. 1A).

Class B pHMM was based on the multiple sequence alignment of *B. mori* proteins Be2G12 (UniProt ID: Q99237), Bm2807 (UniProt ID: P08914), Bm1768 (UniProt ID: P08916) and *A. polyphemus* pc401 protein (UniProt ID: P02847) (Iconomidou and Hamodrakas, 2008; Jones and Kafatos, 1982) (Fig. 1B). Due to limited information in the public repositories, parameterisation of the models may be unreliable. To counteract this we utilised a parameter exploration technique that involved weighted counts conversion to mean posterior

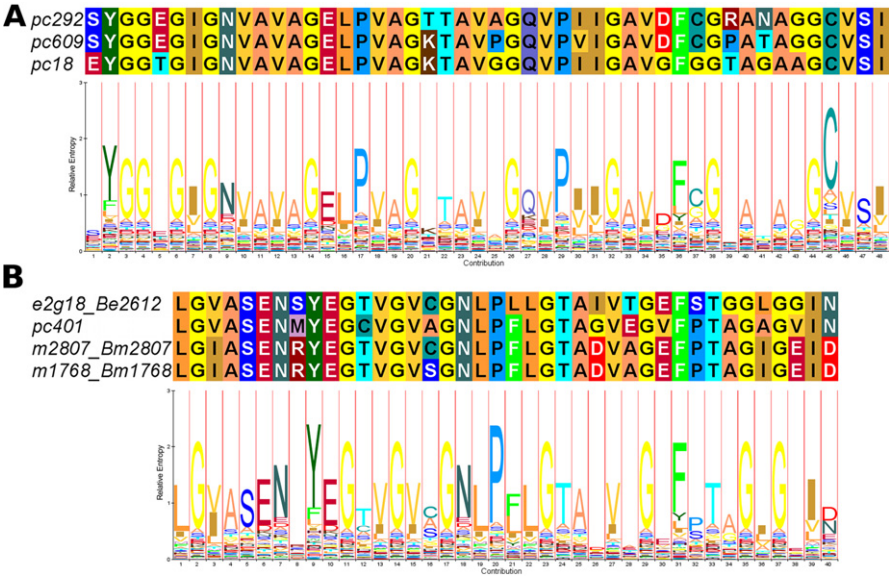

Fig. 1. Class A and class B pHMMs visualisations. (A) Multiple sequence alignment of the central regions of the A family. (B) Multiple sequence alignment of the central regions of the B family.

probability parameter estimates using mixture Dirichlet priors (Brown et al., 1993; Sjolander et al., 1996).

Since Lepidopteran chorion proteins belonging to the C, D and E Lepidopteran chorion protein classes are severely under-represented and some of them are specific for certain Lepidopteran species (e.g. Class E chorion proteins are limited to *A. polyphemus*), no HMM was produced for them.

To visualise all of the central aspects of pHMMs graphically in an intuitively understandable way, we produced HMM-Logos with LogoMat-M (Schuster-Bockler et al., 2004).

2.3. Chorion protein search

Chorion_1, classA and classB pHMMs were used to search for chorion proteins against the protein sequence collection, using the hmmsearch program of HMMER 3.0, choosing relaxed parameters in order not to exclude any potential chorion proteins.

A PHP-CLI script which eliminates protein duplicates also gives each individual potential chorion protein a unique LepChorionDB ID which consists of two parts: The first part is either "CLASSA" or "CLASSB" showing the corresponding chorion class of the protein, followed by a four digit number (e.g. CLASSA0012 or CLASSB0123). For each LepChorionDB entry, the algorithm calculates the molecular weight (MW) (Dawson et al., 1986) and a theoretical isoelectric point (pI), implementing ProMoST's binary search method (Halligan, 2009; Halligan et al., 2004).

To group potential chorion protein sequences from different databases which bear small discrepancies in their termini due to translation of near-identical DNA sequences, we developed a fragment identifying script which detects sequence overlaps between the N-terminus of one protein and the C-terminus of another and groups similar proteins using the MCL cluster algorithm (Enright et al., 2002). Further, protein sequences of each group were aligned using Clustal Omega (Sievers et al., 2011), providing a quick overview of these congener sequences. All Lepidopteran chorion proteins were also aligned in a single multiple sequence alignment. RAxML (Stamatakis, 2006) was implemented for the production of a maximum likelihood phylogenetic tree from the multiple sequence alignment of all chorion proteins produced by Clustal Omega.

2.4. Evaluating the methods

We evaluated our method, as previously described (Magkrioti et al., 2004). We searched classA and classB pHMMs against the entire SwissProt database (training set) and we sorted all the hits according to their score in descending order. For each list, we identified the class of all Lepidopteran chorion proteins, based on their annotations. The training datasets produced were parsed by a PHP script to calculate Sensitivity = TP/(TP + FN) and Specificity = TN/(TN + FP), for a range of 5 score units, where TP is the number of true positives, TN the number of true negatives, FN the number of false negatives and FP the number of false positives. Sensitivity and specificity were plotted against the different cutoffs to identify the cutoff range where sensitivity and specificity meet. To test the validity of our cutoffs, we compared classA and classB pHMMs against a test set which consists of the class A and B chorion protein sequence entries of Entrez database which are not in the SwissProt. We manually classified those entries, as class A or B, according to their annotations. Finally, we removed protein sequences which had lower scores than their corresponding cutoffs.

3. Results

3.1. HMM search

Chorion_1 pHMM search identified 37 chorion proteins in SwissProt and 13 proteins from TrEMBL (UniProt release 2011_11) and the results were sorted by score (Supplementary Table 1). From the characterised proteins, class B proteins were found first, followed by a mixture of chorion protein classes, while at the bottom mostly class A proteins appear.

Using the class A-specific pHMM against UniProt, all 18 class A chorion proteins (12 A, 5 CA and 1 HCA) were identified at the top of the statistically significant results (Table 1). Among them, two uncharacterised proteins (UniProt IDs: Q6LD31 and Q17187) from TrEMBL were also found. The fact that they were between identified class A chorion proteins was a basic criterion in order to include them in class A. At the end of the list, two class B proteins were found; however, the gap of 14 orders of magnitude in their *p*-values compared to that of the last class A protein, led us to reject them.

Table 1
HMM search results of classA pHMM against UniProt: a list of UniProt proteins identified by classA pHMM compared with Chorion_1 (Pfam pHMM) and classB pHMM search results. Non-statistically significant results are marked in grey.

UniProt ID	Score	Source	Class	Chorion_1	classA	classB
P02846	107.1	UniProtKB/Swiss-Prot	A	✓	✓	
Q17214	78.6	UniProtKB/Swiss-Prot	CA	✓	✓	
Q17212	77.9	UniProtKB/Swiss-Prot	CA	✓	✓	
P13531	77.5	UniProtKB/Swiss-Prot	CA	✓	✓	
P08829	77.0	UniProtKB/Swiss-Prot	CA	✓	✓	
Q6LD31	76.9	UniProtKB/TrEMBL		✓	✓	✓
P50603	76.1	UniProtKB/Swiss-Prot	A	✓	✓	✓
P50602	75.8	UniProtKB/Swiss-Prot	A	✓	✓	✓
P0C0U2	75.6	UniProtKB/Swiss-Prot	A	✓	✓	✓
P0C0U3	75.6	UniProtKB/Swiss-Prot	A	✓	✓	✓
P43516	75.6	UniProtKB/Swiss-Prot	A	✓	✓	✓
P43517	75.6	UniProtKB/Swiss-Prot	A	✓	✓	✓
P43514	75.6	UniProtKB/Swiss-Prot	A	✓	✓	✓
P43513	75.5	UniProtKB/Swiss-Prot	A	✓	✓	✓
Q17213	75.3	UniProtKB/Swiss-Prot	CA	✓	✓	
Q17187	74.5	UniProtKB/TrEMBL		✓	✓	
P08929	71.1	UniProtKB/Swiss-Prot	A	✓	✓	
P08825	70.4	UniProtKB/Swiss-Prot	A	✓	✓	
P08826	67.3	UniProtKB/Swiss-Prot	A	✓	✓	
P05687	61.4	UniProtKB/Swiss-Prot	HCA	✓	✓	
P50604	13.8	UniProtKB/Swiss-Prot	B	✓	✓	✓
Q25261	12.2	UniProtKB/TrEMBL		✓	✓	✓

Table 2
HMM search results of classB pHMM against UniProt: a list of UniProt proteins identified by classB pHMM compared with Chorion_1 (Pfam pHMM) and classA pHMM search results. Non-statistically significant results are marked in grey.

UniProt ID	Score	Source	Class	Chorion_1	classA	classB
P08828	100.7	UniProtKB/Swiss-Prot	B	✓		✓
P08914	99.2	UniProtKB/Swiss-Prot	B	✓		✓
P08916	99.0	UniProtKB/Swiss-Prot	B	✓		✓
P08827	98.1	UniProtKB/Swiss-Prot	B	✓		✓
P08917	96.3	UniProtKB/Swiss-Prot	B	✓		✓
P05685	92.0	UniProtKB/Swiss-Prot	B	✓		✓
P08915	89.7	UniProtKB/Swiss-Prot	B	✓		✓
Q99238	89.5	UniProtKB/TrEMBL		✓		✓
Q17182	89.4	UniProtKB/TrEMBL		✓		✓
Q17184	89.4	UniProtKB/TrEMBL		✓		✓
Q17217	89.1	UniProtKB/Swiss-Prot	B	✓		✓
Q17216	88.2	UniProtKB/TrEMBL		✓		✓
P02847	87.2	UniProtKB/Swiss-Prot	B	✓		✓
Q17183	85.7	UniProtKB/TrEMBL		✓		✓
Q99237	85.4	UniProtKB/TrEMBL		✓		✓
Q7M3X8	84.5	UniProtKB/TrEMBL		✓		✓
Q17215	82.0	UniProtKB/TrEMBL		✓		✓
P02848	81.2	UniProtKB/Swiss-Prot	B	✓		✓
P13532	70.7	UniProtKB/Swiss-Prot		✓		✓
P05688	52.0	UniProtKB/Swiss-Prot	HCB	✓		✓
Q9N2N0	51.6	UniProtKB/TrEMBL		✓		✓
P20730	51.5	UniProtKB/Swiss-Prot	HCB	✓		✓
Q17218	50.5	UniProtKB/TrEMBL		✓		✓
P60607	50.4	UniProtKB/Swiss-Prot	B	✓		✓
P50604	49.6	UniProtKB/Swiss-Prot	B	✓	✓	✓
P43515	49.5	UniProtKB/Swiss-Prot	B	✓		✓
Q25261	48.8	UniProtKB/TrEMBL		✓	✓	✓
P08830	42.4	UniProtKB/Swiss-Prot	CB	✓		✓
P08930	41.5	UniProtKB/Swiss-Prot	CB	✓		✓
Q6LD31	19.8	UniProtKB/TrEMBL		✓	✓	✓
P43513	19.6	UniProtKB/Swiss-Prot	A	✓	✓	✓
P43514	19.6	UniProtKB/Swiss-Prot	A	✓	✓	✓
P50602	19.4	UniProtKB/Swiss-Prot	A	✓	✓	✓
P43516	19.3	UniProtKB/Swiss-Prot	A	✓	✓	✓
P0C0U2	19.2	UniProtKB/Swiss-Prot	A	✓	✓	✓
P0C0U3	19.2	UniProtKB/Swiss-Prot	A	✓	✓	✓
P50603	19.1	UniProtKB/Swiss-Prot	A	✓	✓	✓
P43517	19.0	UniProtKB/Swiss-Prot	A	✓	✓	✓

This kind of *p*-value use in the results lists was a basic criterion for discarding proteins during the whole process of protein homology confirmation. Similarly, when we compared our database against class B pHMM, all 18 proteins (13 B, 3 CB and 2 HCB) were found at the top statistical significant results (Table 2). Unreviewed proteins (UniProt IDs: Q99238, Q17182, Q17184, Q17216, Q17183, Q99237, Q7M3X8, Q17215, P13532, Q9N2N0, Q17218 and Q25261) produced comparable *p*-values with those of characterised class B proteins. Thus we included them to the B family chorion proteins. The last 9 proteins had *p*-values at least 7 orders of magnitude far off, from those of the characterised class B proteins and therefore we did not consider them as Class B proteins.

The total number of collected Lepidopteran proteins from 5 different databases was 317,082. Using HMM-based searches we initially identified 255 class A and 421 class B chorion proteins. Pairwise sequence comparison detecting identical proteins reducing these numbers to 164 and 271 unique entries, respectively (Fig. 2). Then, by using the fragment identifier algorithm, we suggested a further reduction of the number of class A and B chorion proteins to 120 and 236, respectively. The total of 435 Lepidopteran chorion proteins that were identified originate from 9 moths and 1 butterfly species (Supplementary Fig. 3).

The average MW of UniProt Class A and B LepChorionDB proteins is 12.4 kDa and 15.5 kDa, respectively, while the average MW of Class A and B LepChorionDB proteins is 12.1 kDa and 15.0 kDa, respectively (Fig. 3). The average pI for UniProt Class A and B LepChorionDB proteins is 4.7 and 4.2 pH respectively, while the average pI of Class A and B LepChorionDB proteins is 5.8 and 4.7, respectively (Fig. 4).

3.2. Alignment characteristics

Visualisation of the multiple sequence alignment of all LepChorionDB proteins (Supplementary Text 1), demonstrates the conservation of the central region/domain. Areas with higher rates of C or P in the "arms" of some LepChorionDB proteins, representing

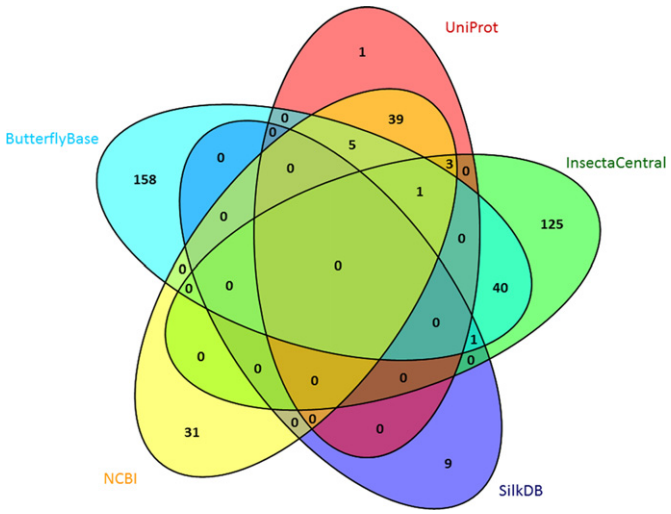

Fig. 2. Venn diagram of the sources of LepChorionDB entries. Venn diagram showing the intersections of UniProt, InsectaCentral, SilkDB, NCBI and ButterflyBase LepChorionDB entries.

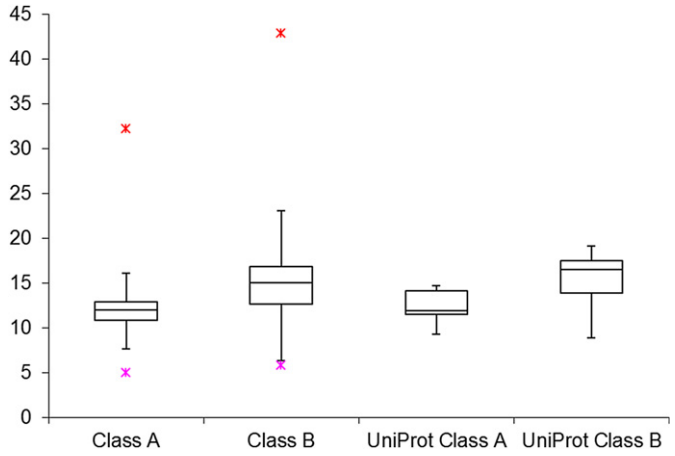

Fig. 3. BoxPlot of molecular weight in kDa from class A and class B chorion protein sequences, comparing LepChorionDB to UniProt proteins.

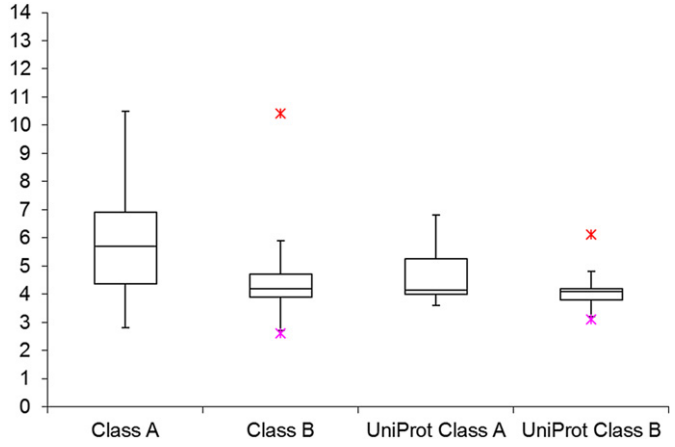

Fig. 4. BoxPlot of isoelectric point from class A and class B chorion protein sequences, comparing LepChorionDB to UniProt proteins.

mainly HCA/HCB (High Cysteine) or CA/CB (High Proline) chorion proteins, respectively, were also found. The rectangular phylogram (Supplementary Fig. 4) constructed by RAxML from the multiple sequence alignment of all LepChorionDB entries, shows a clear separation between chorion protein classes of α and β branches (A, CA, HCA and B, CB, HCB) validating their true relations.

3.3. Cutoff estimation

The score cutoff was estimated as the middle value of the cutoff range where Sensitivity and Specificity were equal to 1: 40.0 for classA and at 30 for classB (Fig. 5). Using the estimated cutoffs in the non-SwissProt Entrez proteins, we produced neither false positives nor false negatives (Supplementary Tables 2 and 3). We applied the cutoff to the entire set of LepChorionDB proteins and we filtered out 22 proteins which had lower scores (with the exception of those that their central region was truncated and located either in the N- or C-terminus, implying that those sequences were fragmented), leaving 413 Class A and B chorion proteins.

3.4. LepChorionDB website

The website of LepChorionDB includes the following menu options: Home, Search, Compare, Filter, Manual and Download

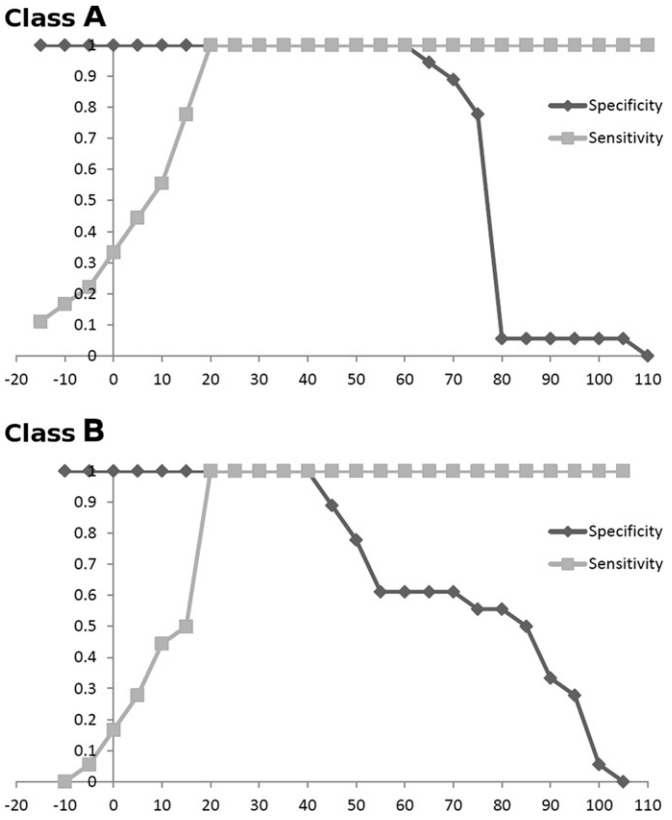

Fig. 5. Plots of sensitivity and specificity against the different cutoffs, in order to find the cutoff for class A HMM (top) and class B HMM (bottom).

(Fig. 6). The default option ("Home" button) offers a short description of the database and presents the authors who contributed to this work.

By clicking the "Search" button, the users can retrieve the protein entries of the database. The database can be searched either by parameters or by accession. In the parametric search, data can be obtained by filling in the available fields, as follows: In "Chorion Protein Class" field the users may choose between class A and B chorion proteins in order to limit the results to one of the chorion families. In the "Organism" field the users may choose between ten species, to obtain results for a specific Lepidopteran species. In the "Source" field a user may choose between the databases that the proteins were derived from. In the "*p*-value Cut-off" field the users may determine a *p*-value threshold. In the "Score Cut-off" field a user may determine an HMMER-based score threshold (the score matrix is based on the BLOSUM62). In the "Molecular Weight (kD)" and "Isoelectric Point" fields the users may determine the range of molecular weight and pI, respectively. In the second section of the search page, data can be obtained by submitting one or more protein names from LepChorionDB or other databases. Protein names can be separated by tabs, commas, spaces or new line characters.

After searching the database, either by parameters or by accession, the number of the results is displayed at the top of the page. Next to it, a FASTA link directs to the FASTA format of the results and a Newick link directs to a Newick formatted tree of the results, suitable for phylogenetic tree visualisation. Below, one or more LepChorionDB tables will appear, with 13 available fields for each LepChorionDB protein entry. The first field is the LepChorionDB Protein Name. The next field contains the sequence in FASTA format where the header line consists of the LepChorionDB Protein Name and the names from external databases. Furthermore, the conserved

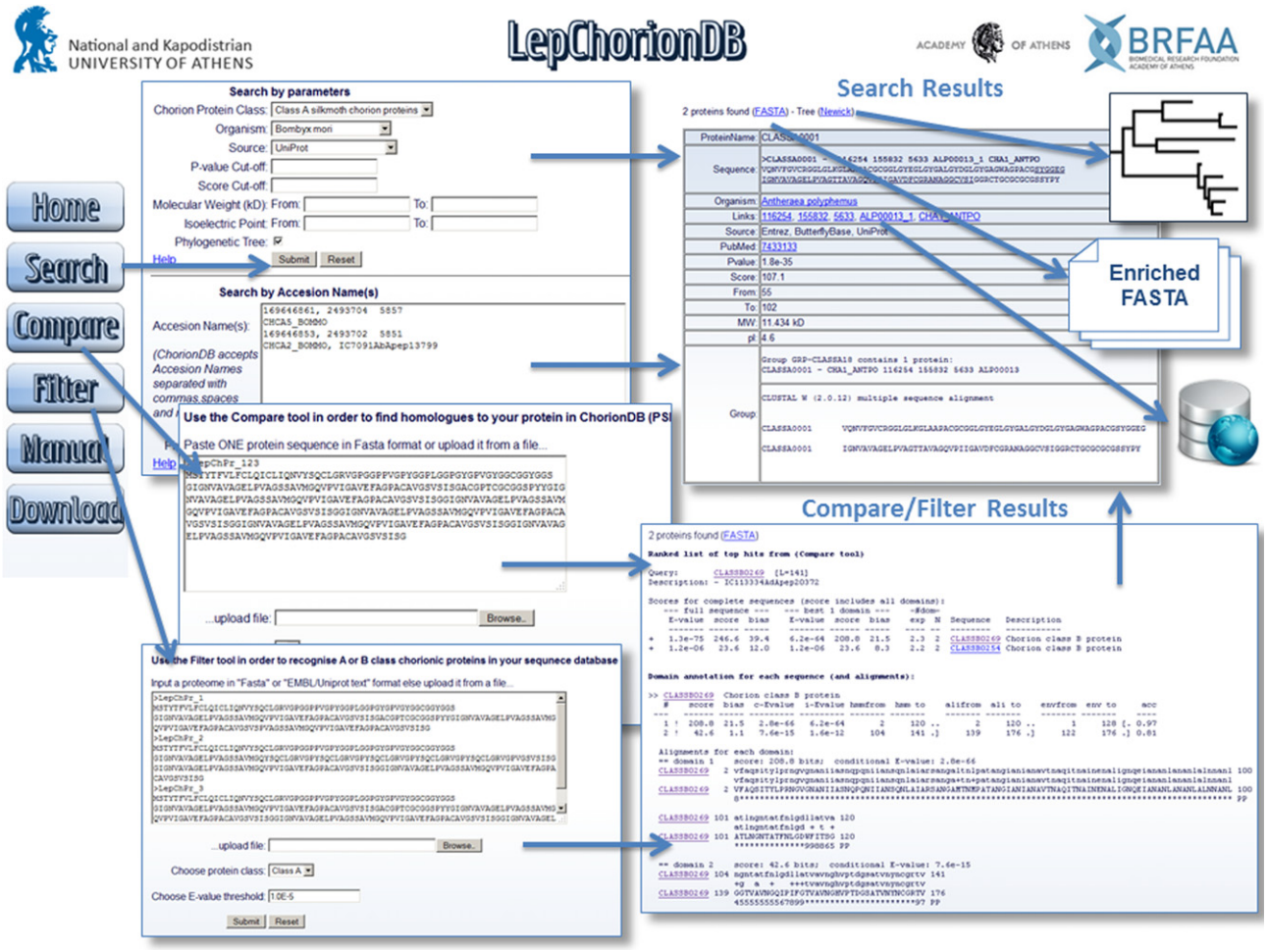

Fig. 6. Web shots of LepChorionDB Search, Compare and Filter tool work flow. Search by parameters or by accession name(s) produces of list of LepChorionDB entries. The results of Compare and Filter tool is an HMMER-like output. The output has links to LepChorionDB entries.

domain is underlined. The "Organism" field shows the species where the protein was derived from. The species name is linked to NCBI Taxonomy for detailed information on each Lepidopteran species. The "Links" field shows links to external databases that refer to the same protein. The "Source" field shows the database(s) where the protein has been downloaded from. The "PubMed" field shows the bibliographic reference where the protein has been described. Proteins characterised as chorion ones only by the HMM search can be distinguished from those which are described experimentally, by the lack of any bibliographic reference in their description. "*p*-value" and "Score" fields show the statistical significance and the score of each protein. The "From" and "To" fields indicate the first and last amino acid of the conserved domain of the protein sequence. "MW" field is the calculated molecular weight of the protein, while the "pI" field is its estimated isoelectric point. The "Group" field shows similar proteins, which are grouped together in multiple sequence alignments in CLUSTAL W (2.0.12) format.

The Compare tool is used in order to perform a protein sequence homology search of a single sequence query against all LepChorionDB entries. It is based on the Jackhmmer program of the HMMER3 suite. Compare input interface allows the users to submit or upload a single sequence query, choose the number of iterations (from one to five) and alter the default *E*-value cutoff. The first iteration performs a search of a protein sequence against LepChorionDB and it produces BLAST-like results. If two or more iterations are selected, a PSI-BLAST-like search is activated where all the matches that pass the inclusion thresholds of the previous round are put in a multiple alignment together with the original query sequence and a new pHMM is generated. The new pHMM is used in order to search the database again. This iterative process continues until no new sequences are detected or the maximum number of iterations is reached. The default number of iterations is one which is used for closer homologue searches and the maximum is five, which is better for distant homologues (PSI-BLAST like) (Eddy, 2009).

The Filter tool is used to identify A or B class chorion proteins from a protein sequence collection by performing HMM searches. In the Filter tool interface, the users may submit or upload a sequence database in FASTA or EMBL/UniProt text format against which the HMM search will be performed. The users choose to search for either A or B class candidate proteins and they may alter the default *E*-value threshold.

Both Compare and Filter have the same output format. At the top of the page, the number of the results is displayed and next to it, a FASTA link directs to the FASTA format of the results. In the first section of the output, users see a synopsis of the results in a BLAST-like list of top ranked hits sorted by ascending *E*-values. In the second section of the output, a per-domain alignment between each query model and target sequence is presented, together with annotations about the degree of conservation and the expected accuracy for each position of the alignment. Throughout the results, all LepChorionDB

protein names of each target sequence are linked to the related LepChorionDB entries. For more information refer to LepChorionDB manual and HMMER3 user guide.

LepChorionDB includes a "Download" link, providing the alignments of the central regions on which the HMMs were based, the HMMs which enabled the identification and classification of the chorion proteins, the sequences of all Lepidopteran chorion proteins, a multiple sequence alignment of them, and a phylogenetic tree.

4. Discussion

Analysis of the results from the searches using Chorion_1 pHMM from Pfam, showed that Chorion_1 recognises Lepidopteran chorion proteins generally, without being able to distinguish between classes, although it has a preference for class B proteins. This bias is due to the composition of the multiple sequence alignment on which Chorion_1 was built: There are 24 class B proteins compared to 20 class A (Supplementary Fig. 2); the central region of 6 class A proteins is removed (Supplementary Fig. 2); the multiple alignment of GGXG consensus which is a class A "signature", is not optimal (Fig. 7). All this undermined the quality and quantity of protein sequences representing the central region of Class A in Chorion_1, shifting the balance towards class B proteins.

In contrast, classA and classB pHMMs of LepChorionDB developed in this work, displayed high ability to distinguish classes in any set of sequences. Each pHMM only identified proteins of its own class with high statistical significance, while the other class sequences were clearly separated at the end of the list, with obviously lower statistical significance (many orders of magnitude higher *p*-values). Comparing the classification of proteins based on the classA and classB pHMMs with the unambiguous classification from SwissProt, it is shown that classA and classB failed in no chorion protein sequence. Therefore, it is shown that the selection of the specific multiple alignments to create A and B family pHMMs and the use of the capabilities of LepChorionDB, is an appropriate solution for finding and grouping of chorion proteins of Lepidoptera.

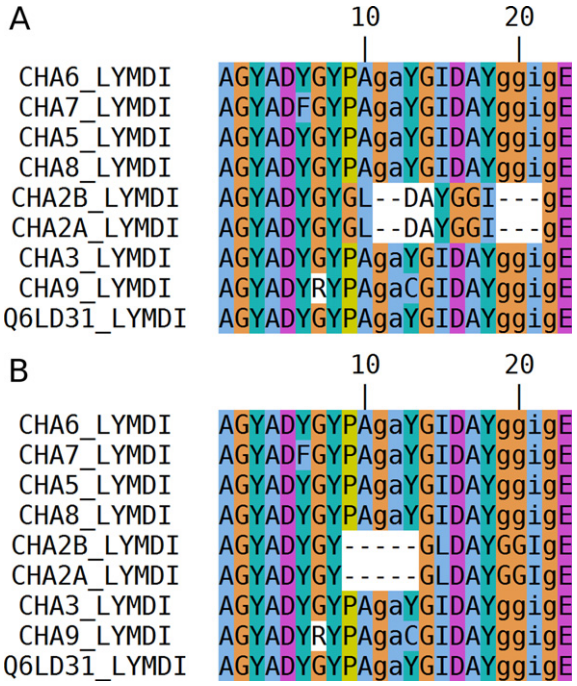

Fig. 7. Suboptimal local multiple sequence alignment in Chorion_1. (A) Original local multiple sequence alignment of class A chorion proteins in Chorion_1. (B) Optimal local multiple sequence alignments of the aforementioned area.

Many protein sequences were produced by automatic translation of transcript or genomic sequences. As the selection was based exclusively on the central region, the sequence of some N- and C-terminal arms might be the outcome of the translation of wrong ORFs or non-coding sequences. Although no attempt was made to eliminate any protein sequences, the fact that chorion proteins are small and acidic (with class B proteins slightly bigger and more acidic than those of class A), should make the user cautious in cases of alkaline or high molecular weight potential chorion proteins found in LepChorionDB. To filter out such pseudo-positives, LepChorionDB enables searches within pI and MW ranges which are set by the user.

Proteins with nearly identical sequences have been grouped, as we believe that they are fragments of the same protein. That grouping, as well as the multiple sequence alignment of all LepChorionDB proteins and the maximum likelihood phylogenetic tree can help the expert to decide whether similar protein sequences are just fragments of the same protein, alleles of the same gene or paralog genes.

5. Conclusions

LepChorionDB aims to provide a user-friendly online access to data related to the major Lepidopteran chorion protein classes A and B (minor classes, such as C, D, or E were not included). The database will be updated annually for novel Lepidopteran chorion proteins. We welcome feedback from users for further improvement of this database, such as additional features. We believe that LepChorionDB will be extremely useful for researchers working in the areas of ecology, evolution, functional and comparative proteomics, and biochemistry of insects.

Availability and requirements

The LepChorionDB is freely available at http://bioinformatics.biol.uoa.gr/LepChorionDB/. The website has been tested with Firefox, Chrome, Internet Explorer, Safari and Opera browsers.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank the University of Athens for financial support. NGG and IM contributed equal responsibility and effort to this paper. We should like to thank Prof. R. Feyereisen and the anonymous reviewers for useful criticism.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.ibmb.2012.12.001.

References

Bouligand, Y., 1972. Twisted fibrous arrangements in biological materials and cholesteric mesophases. Tissue Cell 4, 189–217.
Brown, M., Hughey, R., Krogh, A., Mian, I.S., Sjolander, K., Haussler, D., 1993. Using Dirichlet mixture priors to derive hidden Markov models for protein families. Proc. Int. Conf. Intell. Syst. Mol. Biol. 1, 47–55.
Dawson, R.M.C., Elliott, D.C., Elliott, W.H., Jones, K.M., 1986. Data for Biochemical Research, third ed. Oxford University Press, Oxford.
Diaz, J.H., 2005. The evolving global epidemiology, syndromic classification, management, and prevention of caterpillar envenoming. Am. J. Trop. Med. Hyg. 72, 347–357.
Eddy, S.R., 2009. A new generation of homology search tools based on probabilistic inference. Genome Inform. 23, 205–211.

Enright, A.J., Van Dongen, S., Ouzounis, C.A., 2002. An efficient algorithm for large-scale detection of protein families. Nucl. Acids Res. 30, 1575–1584.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E.L., Eddy, S.R., Bateman, A., 2010. The Pfam protein families database. Nucl. Acids Res. 38, D211–D222.

Goldsmith, M.R., Shimada, T., Abe, H., 2005. The genetics and genomics of the silkworm, Bombyx mori. Annu. Rev. Entomol. 50, 71–100.

Halligan, B.D., Ruotti, V., Jin, W., Laffoon, S., Twigger, S.N., Dratz, E.A., 2004. ProMoST (Protein Modification Screening Tool): a web-based tool for mapping protein modifications on two-dimensional gels. Nucl. Acids Res. 32, W638–W644.

Halligan, B.D., 2009. ProMoST: a tool for calculating the pI and molecular mass of phosphorylated and modified proteins on two-dimensional gels. Meth. Mol. Biol. 527, 283–298. ix.

Hamodrakas, S.J., Asher, S.A., Mazur, G.D., Regier, J.C., Kafatos, F.C., 1982. Laser Raman studies of protein conformation in the silkmoth chorion. Biochim. Biophys. Acta 703, 216–222.

Hamodrakas, S.J., Etmektzoglou, T., Kafatos, F.C., 1985. Amino acid periodicities and their structural implications for the evolutionarily conservative central domain of some silkmoth chorion proteins. J. Mol. Biol. 186, 583–589.

Hamodrakas, S.J., 1992. Molecular architecture of helicoidal proteinaceous eggshells. Results Probl. Cell. Differ. 19, 115–186.

Iconomidou, V.A., Hamodrakas, S.J., 2008. Natural protective amyloids. Curr. Protein Pept. Sci. 9, 291–309.

Iconomidou, V.A., Vriend, G., Hamodrakas, S.J., 2000. Amyloids protect the silkmoth oocyte and embryo. FEBS Lett. 479, 141–145.

Iconomidou, V.A., Chryssikos, G.D., Gionis, V., Galanis, A.S., Cordopatis, P., Hoenger, A., Hamodrakas, S.J., 2006. Amyloid fibril formation propensity is inherent into the hexapeptide tandemly repeating sequence of the central domain of silkmoth chorion proteins of the A-family. J. Struct. Biol. 156, 480–488.

Jones, C.W., Kafatos, F.C., 1982. Accepted mutations in a gene family: evolutionary diversification of duplicated DNA. J. Mol. Evol. 19, 87–103.

Kafatos, F.C., Regier, J.C., Mazur, G.D., Nadel, M.R., Blau, H.M., Petri, W.H., Wyman, A.R., Gelinas, R.E., Moore, P.B., Paul, M., Efstratiadis, A., Vournakis, J.N., Goldsmith, M.R., Hunsley, J.R., Baker, B., Nardi, J., Koehler, M., 1977. The eggshell of insects: differentiation-specific proteins and the control of their synthesis and accumulation during development. Results Probl. Cell. Differ. 8, 45–145.

Kravariti, L., Lecanidou, R., Rodakis, G.C., 1995. Sequence analysis of a small early chorion gene subfamily interspersed within the late gene locus in Bombyx mori. J. Mol. Evol. 41, 24–33.

Lecanidou, R., Rodakis, G.C., Eickbush, T.H., Kafatos, F.C., 1986. Evolution of the silk moth chorion gene superfamily: gene families CA and CB. Proc. Natl. Acad. Sci. U S A 83, 6514–6518.

Magkrioti, C.K., Spyropoulos, I.C., Iconomidou, V.A., Willis, J.H., Hamodrakas, S.J., 2004. cuticleDB: a relational database of Arthropod cuticular proteins. BMC Bioinform. 5, 138.

Mazur, G., Regier, J., Kafatos, F., 1982. In: King, R.C., Akai, H. (Eds.), Insect Ultrastructure. Plenum Press, New York, pp. 150–183.

Moschonas, N.K., Thireos, G., Kafatos, F.C., 1988. Evolution of chorion structural genes and regulatory mechanisms in two wild silkmoths: a preliminary analysis. J. Mol. Evol. 27, 187–193.

Papanicolaou, A., Gebauer-Jung, S., Blaxter, M.L., Owen McMillan, W., Jiggins, C.D., 2008. ButterflyBase: a platform for lepidopteran genomics. Nucl. Acids Res. 36, D582–D587.

Regier, J.C., Pacholski, P., 1985. Nucleotide sequence of an unusual regionally expressed silkmoth chorion RNA: predicted primary and secondary structures of an architectural protein. Proc. Natl. Acad. Sci. U S A 82, 6035–6039.

Regier, J.C., Kafatos, F.C., Hamodrakas, S.J., 1983. Silkmoth chorion multigene families constitute a superfamily: comparison of C and B family sequences. Proc. Natl. Acad. Sci. U S A 80, 1043–1047.

Rodakis, G.C., Moschonas, N.K., Kafatos, F.C., 1982. Evolution of a multigene family of chorion proteins in silkmoths. Mol. Cell. Biol. 2, 554–563.

Sayers, E.W., Barrett, T., Benson, D.A., Bolton, E., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Dicuccio, M., Federhen, S., Feolo, M., Fingerman, I.M., Geer, L.Y., Helmberg, W., Kapustin, Y., Krasnov, S., Landsman, D., Lipman, D.J., Lu, Z., Madden, T.L., Madej, T., Maglott, D.R., Marchler-Bauer, A., Miller, V., Karsch-Mizrachi, I., Ostell, J., Panchenko, A., Phan, L., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Shumway, M., Sirotkin, K., Slotta, D., Souvorov, A., Starchenko, G., Tatusova, T.A., Wagner, L., Wang, Y., Wilbur, W.J., Yaschenko, E., Ye, J., 2012. Database resources of the National Center for Biotechnology Information. Nucl. Acids Res. 40, D13–D25.

Schuster-Bockler, B., Schultz, J., Rahmann, S., 2004. HMM Logos for visualization of protein families. BMC Bioinform. 5, 7.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J.D., Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539.

Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I.S., Haussler, D., 1996. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Comput. Appl. Biosci. 12, 327–345.

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690.

UniProt Consortium, 2010. The Universal protein Resource (UniProt) in 2010. Nucl. Acids Res. 38, D142–D148.

Wang, J., Xia, Q., He, X., Dai, M., Ruan, J., Chen, J., Yu, G., Yuan, H., Hu, Y., Li, R., Feng, T., Ye, C., Lu, C., Li, S., Wong, G.K., Yang, H., Xiang, Z., Zhou, Z., Yu, J., 2005. SilkDB: a knowledgebase for silkworm biology and genomics. Nucl. Acids Res. 33, D399–D402.