In silico identification of new secretory peptide genes in *Drosophila melanogaster*

Feng Liu*¹
Geert Baggerman¹
Wannes D’hertog
Liliane Schoofs
Geert Wets*


* Data Analysis & Modeling Group, Faculty of Applied Economics, Limburg University,
University Campus - Building D, 3590 Diepenbeek, Belgium.

¹ These authors contributed equally to this work.

Corresponding author: geert.baggerman@bio.kuleuven.ac.be
Abstract

Bioactive peptides play critical roles in regulating most biological processes in animals. The elucidation of the amino acid sequence of these regulatory peptides is crucial for our understanding of animal physiology. Most of the (neuro)peptides currently known were identified by purification and subsequent amino acid sequencing. With the entire genome sequence of some animals now being available, it has become possible to predict novel putative peptides. In this way, BLAST analysis of the *Drosophila melanogaster* genome has allowed to annotate 36 secretory peptide genes so far. Peptide precursor genes are, however, very difficult to predict using BLAST, prompting us to search for an alternative approach described here. With the described program we have scanned the *Drosophila* genome for predicted proteins with the structural hallmarks of neuropeptide precursors. As a result, 119 putative secretory peptide genes were predicted, including the 43 annotated (neuro)peptides. These putative (neuro)peptide genes contain conserved motifs reminiscent of known neuropeptides from other animal species. Peptides that display sequence similarities to the mammalian vasopressin, atrial natriuretic peptide and prolactin precursors and the invertebrate peptides orcokin, PTTH, TMOF, DIMs among others were discovered. A peptidomic analysis of fruit fly hemolymph revealed the presence of two of the predicted peptides in this tissue confirming the viability of the genome screening method used.

Our data provides further evidence that many neuropeptide genes were already present in the ancestor of Protostomia and Deuterostomia, prior to their divergence. This bioinformatic study opens perspectives for the genome-wide analysis of peptide genes in other eukaryotic model organisms.

Key words: BLAST – peptide precursor gene – *Drosophila melanogaster* – *in silico* peptide identification – peptidomics - neuropeptide
Introduction

Peptides occur in the whole animal kingdom, from the least evolved phyla with a very simple nervous system (coelenterates) to the highest vertebrates. They play a key role in many if not all physiological processes as neurotransmitter, neuromodulator or neurohormone, and are therefore of considerable biological interest. Many dozens of peptides have already been discovered usually based on their biological activity. Peptides are synthesized in the cell in the form of large preproproteins, which are then cleaved and modified to generate biologically active peptides. Because of their critical signalling role, naturally occurring peptides play important roles in pathogenesis. Although they are of considerable biological, medical and industrial importance, a predictive method for the systematic identification of all candidate bioactive peptides in an organism is lacking so far. Expanding our knowledge on the structural level of peptides is, however, crucial in studying their role and interactions.

Computational methods have become especially important since the advent of genome projects. By means of BLAST (Basic Local Alignment Searching Tool) analysis, an organism genome can be screened for peptide encoding genes, based on sequence similarity to known peptide genes from other organisms. Using BLAST, 36 peptide genes have already been found in *Drosophila melanogaster* (1,2,3,4). Likewise in *Anopheles gambiae*, 35 peptide-encoding genes were discovered using the same sequence similarity-based mining approach (5). While certain of these peptides have been studied in detail (6,7,8), more data on the entire peptidome of insects are needed for integrated functional analyses.

However, for *in silico* prediction of peptide precursor genes in large sequence datasets, the performance of the BLAST tool is limited because putative peptide sequences for which no orthologous (similar) biologically active peptide has been identified as yet (for instance, because of lack of suitable detection methods) will not be revealed. BLAST programs are based on similarity between sequences and do not take into account the existence of other structural hallmarks, in this case those of
peptide precursors. In addition, while BLAST programs are very suitable to scan databases for conserved proteins (9), they are far less efficient at finding similarity to short peptides when they are scanned against the whole genome sequence. Indeed, in most cases, only a short conserved motif is responsible for the function of a particular peptide and often only this short sequence motif, which can be 5 amino acids or less in length, is conserved. For instance, members of the invertebrate FMRFamide peptide family can share the carboxyterminal tetramer FMRFamide or the MRFamide tripeptide motif or only the RFamide motif.

In recent studies of Baggerman et al. (10,11,12), the peptidome (the battery of all present peptides) of the larval *Drosophila* central nervous system was analysed at the amino acid sequence level by means of nanoscale liquid chromatography combined by tandem mass spectrometry and database mining. These results provided biochemical evidence for the expression of not less than 40 peptides in the *Drosophila* brain at a specific time point. Interestingly, not only known or predicted peptides were identified in this study but also eight additional peptides that are encoded in 5 novel peptide genes that could not be identified by BLAST and hence were not predicted as being peptide precursor genes. Unfortunately, several as yet unknown peptide ions observed in the same study could not be fully sequenced. The lack of an appropriate database of putative peptides could be the reason. Sequencing of peptides is aided a great deal if an appropriate database resource is available. Naturally occurring (neuro)peptides are synthesized in the cell as protein precursors from which the peptides are cleaved by specific enzymes. The problem with the identification of these peptides by mass spectrometry combined with database searching (peptidomics) is that many of the cleaving enzymes are not known or that their specificity is not fully understood. This means that in a Mascot search the cleaving enzyme or method cannot be specified. In addition, peptides can have a fairly large number of PTMs (amidation, pyroglutamic acid, …). All this puts extra strains on the database searching. Therefore, reducing the size of the protein
database by including only putative peptide precursors makes the identification by peptidomics easier and reduces the computing time.

The bioinformatic approach described here will complement direct peptidome experiments. Peptidome experiments do not take into account the structural attributes of peptide precursors, such as the presence of a signal peptide, mono- and dibasic cleavage sites, etc.. Therefore, it is highly likely that with a peptidome experiment many false positives are picked up (especially when using inappropriate extraction methods), as reported recently (13). Another disadvantage of peptidome experiments is that (i) not all peptides are equally well extracted (ii) not all peptides ionise with the same efficiency and (iii) not all peptides are present in the same concentration. All prompted us to construct a peptide database in order to provide a strong support for future peptide research. In this study we describe a new in silico searching program that employs typical hallmarks of biological peptides and their precursors, which are not used by the currently available predictive algorithms.

Experimental procedures

Rationale
First, we examined the structural hallmarks of known peptide precursor proteins. Regulatory peptides are synthesized as part of larger precursor proteins that are subsequently processed into smaller active substances. All peptide precursor proteins are usually less than 500 amino acids in length and contain an N-terminal signal peptide (corresponding to a transmembrane domain) that directs them into the secretory pathway of the cell. After cleavage of the signal peptide, further processing by endoproteases occurs predominantly at processing sites, typically mono- and dibasic amino acid residues (14). For example, insect antimicrobial peptides (like defensins) contain multiple RK, RR, KK repeats within their sequences (15). Many peptide precursors contain multiple bioactive peptides that are often highly related,
for example the tachykinin precursor, the allatostatin precursors and the neuropeptide F precursor in *Drosophila* and other insects (16,17,18,19,20). However, in *Drosophila* as in other insects, peptide genes may also encode multiple, unrelated bioactive peptides or just a single bioactive peptide (21). Based on the common structural characteristics of known invertebrate peptide precursors we build a sensitive searching procedure to identify peptide genes in the *Drosophila* database. Two types of programs were constructed. The first one was built to find those peptide precursors that encode multiple highly related peptides. The second one searches for precursors containing a single peptide or multiple unrelated peptides that share conserved motifs with known peptide precursor proteins from other animal species. In both cases, the putative peptide sequence was defined by the presence of characteristic proteolytic cleavage sites, flanking the peptide sequences. In order to avoid the shortcomings of BLAST programs in searching long sequences for short similarity, we split the protein sequences from *D. melanogaster* into short subsequences (in silico cleavage at their putative processing sites), and then applied BLAST to compare the subsequences within each protein sequence to fish for those precursor sequences that have at least 2 similar subsequences. Additionally, the subsequences are also compared with subsequences derived from known peptide precursors in the Swissprot database. These subsequences are obtained by in silico processing the known peptide precursors from other animal species obtained from the Swissprot database. Because each *Drosophila* protein sequence is split into a number of subsequences, and because all of these subsequences are subsequently compared with all known peptide precursor subsequences, a very large number of alignments with a high score are obtained. Because similarity does not imply homology, only the alignments with sequence motifs from actual bioactive peptides are considered significant and the obtained subsequences as possible peptides. A fasta protein database containing all identified putative peptide precursors was constructed. This database was loaded on an in house Mascot server and used for
the identification of peptides in a peptidomic analysis of *Drosophila* hemolymph. Our results show that this technique is very efficient to find novel peptide genes.

**The program.**

The aim of the program is to mine for putative peptide precursors according to the rules and the techniques described above. The program is implemented in SAS, a powerful integrated software to access, manage, analyse and present data. External tools such as SignalP and BLAST need to be run independently. They communicate with the program by text files. The program includes a few sub-programs listed below.

**Protein.SAS.** The first part of the program, named Protein.SAS, serves to pick up all proteins from a specific species, in this case *Drosophila melanogaster*. The input of the sub-program consists of the Swissprot protein database files, and additional *Drosophila* genes at GenBank identified by Hild et al. (2003) (21). The relevant information for each of the *Drosophila* proteins, such as accession number, protein name, gene name, protein sequence, signal peptide information, length and mass is written into a SAS dataset. The first 70 amino acids of every protein sequence serve as output to a text-file in FASTA format, which is used as the input of SignalP. SignalP (www.cbs.dtu.dk/services/SignalP) for eukaryotes is then run to predict the presence and location of a signal peptide in each protein sequence. Next, the sub-program reads the output file by SignalP, and another SAS dataset is created which includes the predicted signal information of every *Drosophila* protein. The dataset is compared with the dataset of all *Drosophila* proteins, and the proteins are retained if they are either annotated to have signal peptides in the Swissprot protein database files or predicted to have signal peptides by SignalP. The comparison result is a dataset of *Drosophila* proteins having N-terminal signal peptides. From this dataset, only the proteins which are less than 500 amino acids in length are retained. In total,
5096 proteins make up the final *Drosophila* protein dataset which will be analysed further. The logic of Protein.SAS is illustrated in Figure 3.

**Peptide.SAS.** The sub-program serves to filter all the peptides or their precursors that are known in Metazoa today from Swissprot protein databases. The peptide precursors are identified by the keywords in each protein datafile. If a protein is picked up, its relevant information, like the information collected in *Drosophila* proteins, is written into a SAS dataset. Figure 4 describes the process.

**Cleavage.SAS.** The objective of the sub-program is to split protein sequences into subsequences after removal of the signal peptide sequence. The protein sequences are split *in silico* at cleavage sites, typical for peptide precursors. Conventional amino acid motifs that are required for cleavage of neuropeptides from their protein precursors in insects have been described as: GKR, GRR, GR, GRK, GKK, KR, RR, GK, RK, KK, R (22). From our statistical analyses on all known peptide precursors in all organisms (data not shown), it is clear that the processing of peptide precursors does not occur at every conventional cleavage site in the precursor. Cleavage also depends on the amino acids that are at the proximity of the cleavage site. For example, proteolytic processing at GKK followed by R always occurs. However, if GKK is followed by A, N, S or K, the processing may or may not occur. For other amino acids at this position, it has not been demonstrated whether processing occurs. Second example: proteolytic processing at a single R residue only occurs when there is a basic amino acid residue in position -4, -6 or -8 with respect to the single R. The basic amino acid is usually an R, but K or H residues work as well [19].

**BLAST analysis.** The output of Cleavage.SAS consists of two database files: a database of “*Drosophila* subsequences” and a ‘peptide’ database of “known metazoan peptides”. BLAST analysis is then conducted on these two databases.
The score matrix **PAM30** is used, and the expectation value (e-value) as well as the parameter **word size** are set to 6 and 2 respectively in order to find short but strong similarities. Figure 5 explains the process.

**Extract.SAS, Shift.SAS, Motif.SAS.** These programs are used to screen the result output by BLAST and determine the biologically significant matches. The sub-program Extract.SAS extracts the *Drosophila* proteins which have at least two similar subsequences within the protein. The sub-program Shift.SAS reads the comparison result from the BLAST analysis and computes the shift value. The shift value is the minimal distance between the N- or C- terminal of a subsequence and the matching amino acids in the subsequence. From the statistical analysis of the known peptide precursors, these shift values should be low. This means that the motif should be close to a cleavage site. The shift value is set to be no larger than 3 in the program. The sub-program Motif.SAS reads the comparison results between *Drosophila* subsequences and known metazoan peptide subsequences, as well as the comparison results among known metazoan peptide sequences themselves, and identifies the *Drosophila* subsequences that contain conserved peptide motifs.

**Tm pred and SOSUI.** Finally, online software TMpred at http://www.ch.embnet.org/software/TMPRED_form.html (23) and SOSUI http://sosui.proteome.bio.tuat.ac.jp/sosuimenu().html are used to determine whether a protein has a single transmembrane region at its N-terminus (Fig. 6). The minimum and maximum length of the hydrophobic part of the transmembrane region was respectively set at 17 residues and 33 residues respectively. For the TMpred program a score above 500 for both inside to outside as well as outside to inside helices was considered to be significant for the presence of the N-terminal transmembrane region. A score of 250 was considered to be significant for the presence of an inside to outside helix of any second or third transmembrane region.
A putative peptide precursor was retained if one program predicts a single transmembrane region at the N-terminus. When both programs predict the absence of an N-terminal transmembrane region, the protein was deleted from the list. The cut-off of the start of the transmembrane region was set at the 20th residue; transmembrane regions that started at or after this point were not considered to be at the N-terminal side and corresponding proteins were deleted from the database.

**Mass spectrometry**

Animals: *D. melanogaster* are kept in 250 ml bottles and bread on a standard diet that consists of 70 ml water; 17 g sucrose; 0.45 g yeast; 0.9 g agar; 0.5 ml 8 % Nipagin and 0.36 ml propionic acid. Wandering stage larvae were collected. The larvae were washed in water and their cuticle was punctured with a fine stainless steel needle. Hemolymph leaking from the wound was collected using glass microcapillaries and transferred to a 0.5 µl micro tube containing 200 µl of ice cold methanol/water/acetic acid (90/9/1, v/v/v). In total 30 µl of hemolymph was collected. After the extraction the sample was centrifuged, the supernatant was dried, dissolved in 15 µl of 2 % acetonitrile in MQ-water with 0.1% formic acid and filtered. Ten µl of this sample was injected on the LC-Qtof MS/MS system.

Capillary LC-tandem MS experiments were conducted using an Ultimate HPLC pump, a column-switching device (Switchos) and a Famos autosampler (all LC Packings, The Netherlands) coupled to a Q-Tof mass spectrometer (Micromass, UK). Chromatography was performed using a guard column (µ-guard column MGU-30 C18, LC-Packings, The Netherlands) acting as a reverse phase support to trap the peptides. Ten µl of the sample (corresponding to 50 *Drosophila* CNS equivalents) was loaded on the pre-column with an isocratic flow of 2 % acetonitrile in MQ-water with 0.1% formic acid at a flow rate of 10 µl/min. After 2 min, the column-switching valve was switched, placing the pre-column online with the analytical capillary column, a Pepmap C18, 3µm 75µm x 150mm nano column (LC Packings, The
Netherlands). Separation was conducted using a linear gradient from 95% solvent A, 5% solvent B to 80% A, 20% B in 90 minutes, followed by a linear gradient from 80% A, 20% B to 50% A, 50% B in 60 minutes (solvent A: water, formic acid; 99.9/0.1 (v/v); solvent B: acetonitrile, formic acid; 99.9, 0.1 (v,v) ). The flow rate was set at 150 nl/min.

The LC system was connected in series to the electrospray interface of the Q-Tof device. The column eluent was directed through a stainless steel emitter (Proteon, Denmark). Needle voltage was set at 1,650 Volts, cone voltage at 35 Volts. Nitrogen was used as nebulising gas. Parent ions with 2, 3 or 4 charges of sufficient ion intensity (threshold was set at 15 counts/second) were automatically recognized by the charge state recognition software (MassLynx 3.5, Micromass, UK) and selected for fragmentation as they elute from the column. Argon was used as a collision gas; collision energy was set at 25 to 40 eV depending on the mass and charge state of the selected ion. The detection window in the survey scan was set from 400 to 1400 mass to charge (m/z). Fragmentation spectra were acquired from m/z 50 to 2000.

Identification of peptides.

Peptides were identified using Mascot (Matrixscience, UK). The database used for these searches was the database of putative secretory peptide precursors from Drosophila. The setting were as follows: peptide tolerance was set at ± 0.3 Da, MS/MS tolerance at ± 0.2. Enzyme was set to none and variable modifications amide (C-terminal), pyroglutamine (Q) and oxidation (M) were selected. Only hits with significance level of 95% or higher were retained.

Results

1. Construction of two databases.

First, we constructed two databases. The first database is generated as follows: Drosophila proteins that are less than 500 amino acids in length and that start with a
signal peptide are assembled from SWISS and TrEMBL databases, as well as from a collection of additional *Drosophila* genes identified by Hild et al. (22). The program SignalP for eukaryotes was used to predict the occurrence of a signal peptide for a protein sequence (24). As a result, 5096 *Drosophila* protein sequences are retained.

Then, all these protein sequences were split into short subsequences at the conventional cleavage sites, taken into account the nature of the amino acids in the proximity of each cleavage site. These subsequences form the first database, which we named ‘*Drosophila* subsequence’.

The second database is a ‘peptide’ database that comprises the subsequences - obtained by in silico cleavage at mono- or dibasic processing sites- of all known peptide precursor proteins known in metazoans to date. These annotated peptides or peptide precursors were filtered from the SWISS_PROT (release 42.11) and TrEMBL (release 25.11) databases as follows: A protein is retained when it is annotated as (neuro)peptide precursor or when its name contains the word ‘neuropeptide’. Proteins of which the corresponding protein file contains keywords such as peptide, neuropeptide, hormone or neurotransmitter are also retained. But if these proteins have a subcellular location as membrane protein (as indicated in the protein file) or if they are characterised by key words such as receptor, signal-anchor, transmembrane, binding protein, DNA binding, nuclear protein, nuclear transport, enzyme or words ending in ‘ase’, they are excluded. In total, 2858 proteins meet these criteria. These peptide precursors are subsequently split into short subsequences at the conventional cleavage sites, also taken into account the character of the amino acids in the proximity of each cleavage site. This collection of peptide precursor subsequences constitutes the ‘peptide’ database.

### 2. Setup of datamining analysis

Standalone BLAST is used to compare the two above-mentioned databases. Interpretation of the results generated by BLAST involves evaluation of the matches
to determine whether they are significant. Therefore, genuine and biologically meaningful similarities need to be distinguished from the irrelevant and essentially random ones. If the alignment is similar to a motif, it is considered significant, and the subsequence is considered a putative peptide. In order to find the conserved motifs, all known peptide precursor subsequences were compared by BLAST.

Four types of analysis were performed:

1. The ‘Drosophila subsequence’ database is compared with itself and those protein sequences, which have at least two similar subsequences within the same protein sequence, are retained (first screening method).

2. The peptide precursor subsequences in the ‘peptide’ database are compared with each other and the obtained similar amino acid sequence tags are considered as possible motifs.

3. The ‘Drosophila subsequence’ database is compared with the ‘peptide’ database and those Drosophila subsequences that display sequence similarities to a conserved motif within a known peptide precursor subsequence in another metazoan organism are retained (second screening method).

4. The retained proteins from 1 and 3 are then analysed by a transmembrane prediction method. Our analysis of the annotated Drosophila neuropeptide precursors indicates that almost all have a single transmembrane region, which is located at the N-terminus and which corresponds to the signal peptide. Therefore, in a final step we fine-tuned the generated list of putative peptide precursors based on this hallmark. The list was curated by the deletion of (i) all soluble proteins (lacking membrane-spanning regions), (ii) of proteins having more than one transmembrane region and (iii) of proteins having one transmembrane region that is not located in the N-terminal region.
Screening method 1

The first screening method is based on the principle that multiple peptides encoded by a single invertebrate peptide precursor gene are often highly related. Therefore, proteins were only selected if they have at least 2 similar subsequences and if the matching amino acid sequence is at or close to the N or C terminus of at least one subsequence. Therefore, the structural pattern of a putative peptide precursor is:

……[cleavage1]-x1(3,60)-[cleavage2]-……-[cleavage3]-x2(3,60)-[cleavage4]……

x1(3,60) and x2(3,60) are two similar subsequences which are between 3 and 60 amino acids long. [cleavage1(-4)] can be any conventional cleavage site listed above. The subsequences do not need to be adjacent within the precursor.

Using this screening method we found 58 peptide precursors in *Drosophila*, 10 of which are well known peptide precursor genes that encode at least 2 related bioactive peptides, drosulfakinin (dsk), FMRFamide, shortNPF, tachykinin, capa or mt-cap2b, diuretic hormone.

For example, the protein identified by accession number Q9V808 is a putative peptide precursor. By comparing database ‘*Drosophila* subsequence’ with itself, we obtained three similar subsequences in Q9V808 (see example in Fig.5). All the putative *Drosophila* peptide precursors, mined by this screening method are depicted in Table 1.

Screening method 2.

The fact that only 9 of the 44 known neuropeptide precursors as well as one known immune induced peptide in *Drosophila* were listed by the first screening method, indicated that the catalogue of putative regulatory peptide precursors obtained by the first screening method is doubtless incomplete. Therefore, we set out for a second screening method that screens for *Drosophila* proteins having a signal peptide and of which at least one subsequence has at least 3/5 amino acids at or close to the N or C
terminus identical to a known peptide. In addition, the identical 3/5 amino acids should be similar to a conserved motif present in known peptides. The retained proteins are then further filtered by the transmembrane prediction analysis as in the first method.

By means of the second method we found 70 Drosophila peptide precursor genes in total, 42 of which are known peptide precursors and 28 are novel. Each of these putative peptide precursor genes encodes multiple non-related peptides or only a single putative peptide. For example, protein Q8MS86 was identified as a putative peptide precursor. The similar subsequence is Q8MS86_2: WKILTAGSHFRWL. The similar known peptides are P11885_2: YVMSHFRWNKF from Rana catesbeiana, and P06298_8: NGNYRMHHFRWGPPKD from Xenopus laevis. The total output of putative peptide precursors mined by this screening method is shown in Table 2.

The combined computational methods generated in total 75 novel putative peptide precursors in D. melanogaster, in addition to the 43 known ones.

**Peptidomic analysis**

The nanoLC-tandem MS method allows us to select and fragment the peptide ions as they elute from the column, even when co-eluting with other peptides. Peptides were identified by subjecting their fragmentation spectra to a Mascot search on an in-house server. This bioinformatics tool (http://www.matrixscience.com) allows the identification of proteins and peptides by matching MS data against any FASTA format protein or (translated) nucleic acid sequence database. In a typical MS/MS ion search, we combined all MS/MS data of every peptide selected for fragmentation during a LC-MS run, in a comprehensive peak list. This type of file contains the centroided mass values and associated intensity values of all the parent ions selected and corresponding fragmentation peaks, and can be submitted to Mascot for fully automated identification of several tens of peptides at the same time.
In total more than 500 ions were automatically selected for fragmentation. Twenty peptides were identified most of which are known to occur in the hemolymph such as the Attacins and DIMs (14). In addition we identified 2 novel peptides: LDDSENNDQVVGLLDVAQGANHANDGAREA and a truncated form of this peptide, LLDVADQGANHANDGAREA (Fig. 7). These peptides originate from protein CG7738 which was identified as a putative peptide precursor by screening method 1 (Fig. 8).

Discussion

Because of the availability of its complete genome sequence, Drosophila becomes a model insect for peptide research. We have identified in total 118 putative peptide precursor genes in D. melanogaster by applying the here presented database searching programs. 43 of them are annotated peptide precursors. All predicted peptide precursors meet with following criteria: (i) each putative peptide precursor is less than 500 amino acids in length and has a signal peptide; (ii) each precursor contains one or several putative peptides that are flanked by conventional cleavage sites. Here are two possibilities: the precursor contains two or more peptides that share sequence similarities or alternatively, the precursor contains a single peptide that shares conserved motifs with known peptide precursor subsequences from other organisms; (iii) all predicted peptide precursors have one N-terminal transmembrane region.

Several of the genes mined by our method encode peptides that display significant sequence similarities to known vertebrate or invertebrate neuropeptides. These similarities have not been discovered by BLAST scanning of the whole Drosophila genome. We will discuss a few examples. A putative peptide encoded by CG3868, mined by the first method, displays sequence similarities with a antifreeze glycopeptide precursor identified in Antarctic fish (26). The salivary gland glue protein (CG18087) contains a putative peptide sequence that displays significant
similarities to vertebrate neurophysins. Neurophysins are a group of small, soluble proteins secreted by the hypothalamus. They serve as binding proteins for oxytocin and vasopressin during their transport to the posterior pituitary. They are secreted with the hormones but have no known functions other than serving as a carrier. In vertebrates neurophysins originate from the vasopressin peptide precursor. The salivary gland glue protein (CG18087) does not contain a vasopressin/oxytocine-like peptide.

Putative peptides from two genes, CG9358 and BK003312, display sequence similarities to conserved parts of the prolactin precursor. Prolactin and growth hormone are two distinct neuropeptide hormones that have been found in all vertebrate groups but not in cyclostomes (27), although prolactinergic neurons that were detected immunochemically occur in a protochordate (28). The GH/PRL superfamily is likely to have a prevertebrate origin but a putative invertebrate member was so far not found, in contrast to other neuropeptide superfamilies that are highly conserved in vertebrates and invertebrates. Examples are tachykinins, gastrin, insulin, neuropeptide Y, corticotropin releasing factor, calcitonin-gene related peptide (29).

*Drosophila* BK002187 encodes a peptide with sequence similarities to atrial natriuretic peptide (ANP). Natriuretic peptides are vertebrate hormones that play a pivotal role in cardiovascular and body fluid homeostasis in vertebrates (30). Although a novel natriuretic peptide has recently been found in the heart and brain of the hagfish, the most primitive vertebrate (31), no member of this family was has as yet been described in invertebrates. Finally, a peptide encoded by the *Drosophila* LP04693 displays sequence similarities to γ-MSH, a pituitary hormone, derived from the pro-opiomelanocortin precursor, the function of which has remained elusive (32) (Fig. 7)

When we consider similarities to invertebrate neuropeptides that are known (annotated) at this moment, the CG1565 protein contains a putative peptide that has
an N-terminal hexamer, contained within orcokinin, a myotropic neuropeptide discovered in crustaceans (33,34), but which has so far not been identified in insects (Fig. 8). A putative peptide sequence within the trunk protein precursor displays striking similarities with prothoracicotropic hormone, a neuropeptide that has so far only been identified in lepidopteran species, in which it stimulates ecdyson biosynthesis in the prothoracic glands (35). Interestingly, the sex-specific gene, MSOPA, as identified by Jin et al. (36) encodes a putative peptide that shows sequence similarities to a male accessory gland-specific 57kDa peptide precursor. Next, the putative *Drosophila* peptide encoded by CG8087 displays more than 60% sequence identities with a neuropeptide derived from a neurospecific peptide precursor in the terrestrial snail, *Helix lucorum* (37). Finally, some mined genes (CG16882, CG11131, CG7465, CG1221, Argos, Trunk) have been predicted or shown to encode for ligands of membrane receptors, such as EGF, Toll or Torso receptors (38), a function in line with the peptidergic nature of their products. Since the publication of the *Drosophila* genome sequence, several microarray studies have been performed and we observed that some of the mined peptide precursor genes are upregulated by ecdyson (CG7350, CG7608, CG1807, CG 7350) (39,40). Ecdyson is an ecdysteroid involved insect metamorphosis and reproduction. It is the precursor of 20-OH-ecdysone, the functional counterpart of vertebrate estrogen (41). In this way, our data are in accordance with the reported interactions of peptide and steroid hormone signalling cascades in vertebrates (42,43). Other mined genes are upregulated after infection (44) and encode for peptides that are secreted into the haemolymph such as attacin, diptericin, drosocin and various *Drosophila* immune induced peptides or DIMs (15). With the currently established program, several additional putative peptide precursor genes that display sequence similarities to known DIMs, were found. Three of them (CG32851, CG5791, CG15065) form part of the Toll pathway (45) and one (CG18107) is rhythmically expressed in the head (46).
Former microarray studies revealed that regulation of transcription of known neuropeptide genes as well as other putative peptide precursor genes established in this study is circadian clock dependent (i.e. capa, corazonin, CG4784, CG1807; (46,47), nutrient-dependent (CG10918, CG15225) (48) or sex-specific (accessory gland peptides, CG7738, CG11458) (36,49).

Our program also picked up the drosocrystallin gene (50) as well as other annotated cuticular proteins. In *Tenebrio molitor*, biologically active peptides display strong sequence similarities to parts of cuticle proteins and therefore they might be processed from them (51).

Given the fact that proteolytic processing does not always occur at every conventional cleavage site (52), our established catalogue of predicted peptide precursors is doubtless incomplete and it will be a difficult challenge to consider the existence of these unconventional cleavage sites in the further refinement of our method.

Only two of the characterized peptide precursors were not mined by our method, i.e. the diuretic hormone precursor or CG8348 because it has 4 transmembrane regions, and the proctolin precursor ‘Q8MMJ7’ because its sequence is too short (5 amino acids) to be filtered by the program. Inherent to datamining methods, a few cases could be false positives: CG5559 has been annotated to encode a conserved protein involved in synaptic vesicle fusion, CG6409 has been predicted to be a component of the endoplasmatic reticulum, CG11577 has been predicted to permanently reside in the lumen based on its C-terminal sequence (53) and CG6357 encodes a putative cysteine protease.

The database of predicted and known peptide precursors in *Drosophila* as established in this paper will serve several applications in experimental research. Many unassigned masses observed in peptidomic experiments could not be identified, which could be attributed to the lack of an appropriate peptide database (10). Mass spectrometric data will become much easier to read and interpret if the
database against which they are scanned is much smaller than the SWISS-PROT database. In a peptidomic analysis of the hemolymph of the fruit fly, we were able to identify 2 novel peptides originating from the CG7738. The first peptide, LDDSENNDQVVGLLDVADQGANHANDGAREA is 31 AA in length and is flanked at the amino-terminal side by the cleavage site of the signal peptide (Fig. 8). At the carboxyterminus, the peptide is flanked by an arginine residue that could act as a monobasic cleavage site. The second peptide, LLDVADQGANHANDGAREA, is a truncated homologue of the first one. This example clearly demonstrates that the peptide database identified in this study facilitates the mass spectrometric identification of peptides in *Drosophila*.

Also in mammalian models, genome-wide analysis of peptides by mass spectrometry has recently boosted (54). Construction of a peptide database, like the one presented here for *Drosophila*, will be of high value to support these studies. As the structural hallmarks of peptide precursor sequences are highly conserved across phyla, we foresee that the established search program can be adapted for the genome-wide analysis for peptide precursor genes in other animal model systems that have a sequenced genome.

**Acknowledgements.**

This research was sponsored by the FWO grant G0146.03. G. Baggerman is a post-doc fellow of the FWO.
References


Legends to the figures.

Table 1: Putative Peptide Precursors: Each putative peptide precursor in *Drosophila melanogaster* encodes at least 2 similar subsequences. The similar subsequences are in bold. The similar amino acids are underlined.

Table 2: Putative Peptide Precursors: Each putative peptide precursor in *Drosophila melanogaster* encodes a similar subsequence to a known peptide. Each putative peptide precursor sequence is accompanied by the known peptide precursor that displays similarities. Both the similar subsequence and the peptide are in bold. The conserved motifs are underlined.

Table 3: Comparison of γ-MSH sequences of different metazoan species to the putative γ-MSH-like sequence in *D. melanogaster*

Table 4: Comparison of Orcokinin sequences of different invertebrate species to the predicted Orcokinin in *D. melanogaster*

Table 5: Comparison of Immune induced peptide sequences (DIM) from *D. melanogaster*

Fig. 1. Construction of database containing all *Drosophila* proteins that are less than 500 residues in length and that have a signal peptide.

Fig. 2. Construction of neuropeptide precursor database across metazoan species.

Fig. 3. Cleavage of protein sequences in the two datasets and BLAST analysis of the obtained subsequences.

Fig. 4. Strategy for the final filtering of the BLAST comparison results.

Fig. 5. Screening method 1. The protein identified by accession number Q9V808 was retained as a putative peptide precursor. By comparing database ‘*Drosophila* subsequence’ with itself, we obtained three similar subsequences in Q9V808 (underlined). The number following the accession number represents a different subsequence within a protein sequence by its position.

Q9V808_6: IPYEVKVDVPQPYIVE
Q9V808_8: IPYEVKVPVDKPYEVKVPVPQPYEVI
Q9V808_9: IPYEVKVPVPQPYEVI

Fig. 6. Screening method 2. Q8MS86 was identified as a putative peptide precursor. The subsequence iQ8MS86_2: WKILTAGSHFRWL. is similar to the subsequences P11885_2 (YVMSHFRWNKF) and P06298_8 (NGNYRMHHFRWGSPPKD) derived the corticotropin-lipotropin precursors of *Rana catesbeiana* and *Xenopus laevis* respectively.

Fig. 7: Fragmentation spectra and annotation of 2 peptides identified in the hemolymph a) LDDSENNDOVCGGVL DLVDADQQGANHANDGAREA b) LLLDOADQQGANHANDGAREA Both peptides originate from the CG7734 that was picked up by screening method 1. The most important fragment ions (mainly b an y – type ions are indicated on the spectra)

Fig. 8. Amino acid sequence of CG7734. The peptides that were identified in the hemolymph are indicated in bold. The signal peptide predicted by SignalP is underlined.
<table>
<thead>
<tr>
<th>Table 1</th>
<th>The known peptide precursors (10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>PD00940; Drosulfakinin precursor (DSK) OR CG18909</td>
</tr>
<tr>
<td></td>
<td>MPLMALACFC LVLVLIPAPQ TSLQNAKKDR RLQLESKIG</td>
</tr>
<tr>
<td></td>
<td>GEIDQP1ANL VGSFLGQDG RNQKTMFSF GVVRPLSRPI</td>
</tr>
<tr>
<td></td>
<td>IFIELDLMD DDEERTKAKR PDYDGMRFPG KRGGDQFDD</td>
</tr>
<tr>
<td></td>
<td>YGRMF</td>
</tr>
<tr>
<td>(2)</td>
<td>P10552; FMRFamide precursor (FMRF) OR CG2346</td>
</tr>
<tr>
<td></td>
<td>MGIALMFALL LYGQSAIHS EIIDEFTPAG NLSDLQADSEV</td>
</tr>
<tr>
<td></td>
<td>SPQSNNDLVI ALLGDQXETE AELFRRHRIS VIGFDYSFVA</td>
</tr>
<tr>
<td></td>
<td>VVVLFPKXGR KPRVYDPEL EARRSQVQDN FMRRFQYQAE</td>
</tr>
<tr>
<td></td>
<td>QLPPFERGYS ADDEGMMK APRMDYWEPD QKDGFMRF</td>
</tr>
<tr>
<td></td>
<td>PKQDFSRFR ERDPQDFMRF GRDPQDFMRF GRRDFQDFMR</td>
</tr>
<tr>
<td></td>
<td>FPPMTPAFEM DM FPPMTPAFEM MRPPFQSNFH FQRSNPEHEL</td>
</tr>
<tr>
<td></td>
<td>RSPPQPMDRE RPDPNDFMRF RPADQCVVRK GMDSNFLGR</td>
</tr>
<tr>
<td></td>
<td>GKKLQDPAPE SKPVSNGQN PGERSPVKDA MLTEMKQR</td>
</tr>
<tr>
<td></td>
<td>QDQVQKNAGQ ATTTQDSVE QDQFQGR</td>
</tr>
<tr>
<td>(3)</td>
<td>QVGG55; Hugin protein precursor (HUG) OR CG6371</td>
</tr>
<tr>
<td></td>
<td>MCGPSYCTLL LIAASCYILV CSHAKSLQGT SKLQDGNSH</td>
</tr>
<tr>
<td></td>
<td>AGSARGLSSP ASPALSEARQ KRAMDSYKEL TDIDESLEEN</td>
</tr>
<tr>
<td></td>
<td>SLAQASATM QVAMPPQGQ EFPPDTPMPL YYLLQLKLR</td>
</tr>
<tr>
<td></td>
<td>QLQSMGEPAY RVRTPRLGRS IDSRLWLLA GATGMASEE</td>
</tr>
<tr>
<td></td>
<td>AIDQPMQWM VKSVFVMFPR LKRAQCVCGG D</td>
</tr>
<tr>
<td>(4)</td>
<td>Q8SZ21; Tachykinin precursor (TK) OR CG14734</td>
</tr>
<tr>
<td></td>
<td>MRPLGLIAL ALLLLLLLA PSSAADTETE SSQSGSLPQGA</td>
</tr>
<tr>
<td></td>
<td>EKPRVKVKA PSTQFQGMK KGGEKDQTS GNMQLSSQPD</td>
</tr>
<tr>
<td></td>
<td>LDYAEADDE SYAENGRLK KRAMVLVPGL GKKFPFRRN</td>
</tr>
<tr>
<td></td>
<td>LSDLQVLLEE ERLRDLSSLQ FFDVAGRDRG SAVGQRAPR</td>
</tr>
<tr>
<td></td>
<td>FQPGRKRPFA LLADDDDEAE DETERQKQ APRNSVFQGR</td>
</tr>
<tr>
<td></td>
<td>GRKVDQHYYH KRAALDDSYR LQGRQDFQD FRKSVFVAG</td>
</tr>
<tr>
<td></td>
<td>KEDDQLGNOV DGGHEQAL VHPWLYLWSG APFAPFDM</td>
</tr>
<tr>
<td></td>
<td>RCRFPALFE</td>
</tr>
<tr>
<td>(5)</td>
<td>Q9MIF6; Cardio accelerator peptide 2b (CAPA OR MT-CAP2B) OR CG15520</td>
</tr>
<tr>
<td></td>
<td>MKSMVWLVTL WVIPETFH AETSTHKQKR AANWGLYAFF</td>
</tr>
<tr>
<td></td>
<td>RVGRPSLPSL ASRDLQGKAR IQEYLQIDY QSEDYYAEFQ</td>
</tr>
<tr>
<td></td>
<td>KRASGLYAPP RVGDRARL WKAHLRGLQ VLDRKTPSR</td>
</tr>
<tr>
<td></td>
<td>SSSLWFUGPRL GKRSDAKSAF ADISKQKEL N</td>
</tr>
<tr>
<td>(6)</td>
<td>Q9VIQO; Short neuropeptide F precursor (SNPF) OR CG13968</td>
</tr>
<tr>
<td></td>
<td>MFHILKRELQ GCLALALCLVL SLQMQOPPAQ EVSSAQEGHL</td>
</tr>
<tr>
<td></td>
<td>VQPPVERQSS KDSFLGQLTS LYNLDLQQR YGSPGRQQPE</td>
</tr>
<tr>
<td></td>
<td>QVERKAORSP SRSLRFPSRD PDLMSNIVLW RarfDvNQKPE</td>
</tr>
<tr>
<td></td>
<td>IEPSRSLRFSPG RDPRSFQPM RRATYDQDL RELNINQQQ</td>
</tr>
<tr>
<td></td>
<td>QQLSTPEGQ DLADYDQGLY EKVRAVRQ POLR NWGSRVQPE</td>
</tr>
<tr>
<td></td>
<td>ANNADNQEIEQ SQRLWYQQLL SGKDRMLLSA LQDQIEYPN</td>
</tr>
<tr>
<td></td>
<td>VASYANQED DDLMDNQTF PQEVRKFRM RLRPRSTSLA</td>
</tr>
<tr>
<td></td>
<td>PSEQQHEPEE TSSIPKFTQO</td>
</tr>
<tr>
<td>(7)</td>
<td>Q9VF7; Allatostatin/MIP precursor (MIP) OR CG6456</td>
</tr>
<tr>
<td></td>
<td>MAHTKTRRTY GFLMVLILGL SAGNLVWASG SGAPOGSNPE</td>
</tr>
<tr>
<td></td>
<td>GGGDSLEQVV LQDLQDSDLV GNNKRAMQSL QSSWRKRRS</td>
</tr>
<tr>
<td></td>
<td>GDVSDFPDMY TGGVFPVIT DGNNTNWDK FERLASSQA</td>
</tr>
<tr>
<td></td>
<td>QQQQQPLQQ SQSGSEFDFF LAGEROVRK AHKSMVRHAGF</td>
</tr>
<tr>
<td></td>
<td>KRQPAQWNKR FKAGMRQRRF PMNNLWKMNR NDQWQLEQ</td>
</tr>
<tr>
<td></td>
<td>GKKRSQLS PS N</td>
</tr>
<tr>
<td>(8)</td>
<td>Q9WON6; Neuropeptide like 1 precursor (NPLP1) OR CG3441</td>
</tr>
<tr>
<td></td>
<td>MQAVIQSANS SRLKMLSSM LLNAAIQPSR IIVSATOVDV</td>
</tr>
<tr>
<td></td>
<td>NVSPCEMELS INQLMSFSPSE YQHLASALRN QKLNLRLERVQ</td>
</tr>
<tr>
<td></td>
<td>LVAGVEEPQLG EYPDFEEKRD RSVAAALQHG LLNAPKRSLS</td>
</tr>
<tr>
<td></td>
<td>TLANNQGQER AAPDQGRDSQ DSGQEPQXK RYGLSRLAAG</td>
</tr>
<tr>
<td></td>
<td>LNTQGKVVG TLQARDFPLG PWKNXKTMN ARLQSAFSP</td>
</tr>
<tr>
<td></td>
<td>RPDRKNVAAV ARTNQGEHG QRAGAEKPNL GALKSSPUPG</td>
</tr>
<tr>
<td></td>
<td>VQPKREDERR RLPAADDYA DMQSWYNYW SYGAYADDLW</td>
</tr>
<tr>
<td></td>
<td>NDYRRAEKRK LRVPVTPRA TASTHRSLR</td>
</tr>
<tr>
<td>(9)</td>
<td>Allatostatin precursor (AST) OR BCDWA:RE16553 OR CG13633</td>
</tr>
<tr>
<td></td>
<td>MNSLHALHLL LAVCCVGIA SYBVIQAPQSR GQDLDDADV</td>
</tr>
<tr>
<td></td>
<td>AADZADNNG DNIDQVERY APFLRPRARM YTNQGPRMQ</td>
</tr>
<tr>
<td></td>
<td>LPPVNPLGK RSPYFPLGK RQKDYQDOQ KEPIDYVAPF</td>
</tr>
<tr>
<td></td>
<td>NYLDAREAVR GFQGRNRTRR PQFPNFLGK R</td>
</tr>
<tr>
<td>(10)</td>
<td>QML70; Immune induced protein 10 (IM10) OR CG18279.</td>
</tr>
<tr>
<td></td>
<td>MKSFGLIALA ICVGCVIAAE PQRYDGMRNG HFGVGFSGPQ</td>
</tr>
<tr>
<td></td>
<td>VYRQCONEGR YSVGFVQGQQ QPAQPHQGRY TVEADYGFV</td>
</tr>
<tr>
<td></td>
<td>NRKRAACPPA HTISQPPFSA GMRQKNQAKS VQIPOGRAR</td>
</tr>
<tr>
<td></td>
<td>PQDFHWRPR TVDQGNYQFY IQQGRSHPQL HVAPDRTSV</td>
</tr>
<tr>
<td></td>
<td>INGGQVYQRR SRSIPQHHQ RVDFYQDQGN GFFSGAQFR</td>
</tr>
<tr>
<td></td>
<td>GINARQVEQ NVBRQDDQGQ IWNWNSVSRK RFDGRVTID</td>
</tr>
<tr>
<td></td>
<td>RQHTIVSGQ GRPAQHY</td>
</tr>
</tbody>
</table>
The putative peptide precursors (47)

1. Q8IP68; CG31813.
2. Q8IPI9; CG32829.
3. Q8MS63; BCDNA: LP08232.
4. Q8SZG8; LP03261p or CG11470.
5. Q960I1; LP07079p or CG12164.
6. Q9NEG3; EG: BACR43E12.5 OR CG14418.
7. Q9V7U4; CG15615.
8. Q9V080; CG31001 OR CG6564 OR CG15901.
9. Q8V401; EG: 96G10.8 OR CG14265.
10. Q9V5U4; CG13227.
11. Q9V5U4; CG15615.
12. Q9V7U4; CG15615.
13. Q9V7U4; CG15615.
14. Q9V7U4; CG15615.
15. Q9V7U4; CG15615.
16. Q9V7U4; CG15615.
17. Q9V7U4; CG15615.
18. Q9V7U4; CG15615.
19. Q9V7U4; CG15615.
20. Q9V7U4; CG15615.

The putative peptide precursors (47)
Table 2
The known peptide precursors (42)

<table>
<thead>
<tr>
<th>Case #</th>
<th>Accession</th>
<th>Name (Abbr.)</th>
<th>Species/Gene</th>
<th>Peptide Sequence</th>
</tr>
</thead>
</table>
| 1      | O96690     | Pigment-dispersing factor (PDF) | Pigment-dispersing factor | MARW6TVPAL VLAGCCQGW YGAGMAMDK ERVYKVKKR DILDFMVNGVQQSFQGQG VTCRQLFLIS NSLGSPPVR GKNRSLNLS PSLIFRMNMA G
| 2      | P08040     | Drosulfakinin (DSK) | Procambarus clarkii | MRLR16FRCV LVLVLPILFGD TSLQGKAQDR RLESlEKSIG VLVHLKFKGRK KPRYKQDEL EAKFSOQVGN FMHF3KRQAE QPLPFFSEYG SDELGCMK AAMRDQGDP RQGREGMDRD FRQDPRGFRMPR RBPDQDFMRPR GDPRDFQDMR PRFSTPRAFPR FRRTFQDFMR PRNRFUPRM FMRSTSFPR FGBPQDFVR G
| 3      | P10552     | FMRFamide (FMRF) | Neosulfakinin-I. Sarcophaga bullata | MTLLYQVGLL LLVAATYKVS AECCTPGATS DFCTVFSMLR TMEQNPFLRF SRQMPFLRFK AGSDPFLRFLR GRSPDNPFLRF QKAAADNPFL RGFQDRADYN LFRPGRPS FNFSRPNF RF L
| 4      | P17975     | Adipokine precursor (AKH) | FMRFamide | MNLROKLLVIA AVFLKACLQ VCLQTFSPDW GIDFVQCVL NNVCQGTKMLD YGIDFQTCAT CQLKGKFSK AI POCEDIASIA PFLNAM
| 5      | P81829     | Leucokinin precursor (LK) | Leucokinin | MQCPLQVQTV LFTLSMCG YTFQYSRGW T
| 6      | Q02978     | Eclosion hormone precursor (EH) | Eclosion hormone | MNSKLYLCT FVAWAVMLV VMGNAPHSM YTHKRFDSMG GIDFQCVL NNVCQGTKMLD YGIDFQTCAT CQLKGKFSK AI POCEDIASIA PFLNAM
| 7      | P25331     | Eclosion hormone precursor. Bombyx mori. | MANKLTAIV VALAVPMVN LDYANCSFAP ASYDAMEIC IENCAQCKKM FGFWFGGLC AECICRAGKR DIPCESFAS ISPPFLNLK
| 8      | Q24049     | Amnesiac neuropeptide precursor (AMN) | MRRCTAYVC FTLFPFLFAA SLARRRVVSG SNGASSALAC RQFEQSLASR RPERAECRGT QLYVYVYNG AGQSLCAAV LCCRXYLET FNSFCSELYF VGQFTFAAAAR TRQFPVTLS WPLCNDSEK VLTKKPSGSL 1GRSVPQKQ PKFRQENPFA LSPLSLEMMR
| 9      | P08090     | Molluscan insulin-related peptide 3 precursor. | Helisoma trivolvis. | MNLROKLLVIA AVFLKACLQ VCLQTFSPDW GIDFVQCVL NNVCQGTKMLD YGIDFQTCAT CQLKGKFSK AI POCEDIASIA PFLNAM
| 10     | Q26377     | Crz precursor (CR2) | Helicoverpa armigera. | MNLROKLLVIA AVFLKACLQ VCLQTFSPDW GIDFVQCVL NNVCQGTKMLD YGIDFQTCAT CQLKGKFSK AI POCEDIASIA PFLNAM
| 11     | Q8IA34     | Neuropeptide IFamide precursor | Leucophaea maderae. | MNLROKLLVIA AVFLKACLQ VCLQTFSPDW GIDFVQCVL NNVCQGTKMLD YGIDFQTCAT CQLKGKFSK AI POCEDIASIA PFLNAM
| 12     | P25331     | Eclosion hormone precursor. Bombyx mori. | MANKLTAIV VALAVPMVN LDYANCSFAP ASYDAMEIC IENCAQCKKM FGFWFGGLC AECICRAGKR DIPCESFAS ISPPFLNLK

P01162; FMRFamide. Helisoma trivolvis.

P21143; Ceacamakinin. Leucophaea maderae.

P08090; Molluscan insulin-related peptide 3 precursor. Lymnaea stagnalis.

P5975; Adipokine hormone II precursor. Leucophaea maderae.

P1146; FMRFamide. Helisoma trivolvis.


P08090; Molluscan insulin-related peptide 3 precursor. Lymnaea stagnalis.

P21143; Ceacamakinin. Leucophaea maderae.

P08090; Molluscan insulin-related peptide 3 precursor. Lymnaea stagnalis.

P21143; Ceacamakinin. Leucophaea maderae.
(12) Q95NV8; Allatostatin preprohormone (AST2) OR BCDNA:RH3507 OR CG14919
M98PFCYLLC YLLL4LFFPA LSEARPGSAEG TQD3GDDGLGQV
QVARVGGYGGYDMPAQIPA YIHYPMRFL QMFLAQYRTP
YSAYLSPTY GWNL4LYRFL ESFQQYVRBQ CFY4NPSFCR
K
P42559; Allatostatin. Manuda Sexta.
QYFRQCVQYF QYFC

(13) Q9NIP6; Cardiac accelerator peptide 2b precursor (CAPA OR MT-CAP2B) OR CG15520
M98G7L441VL VVIFFA9PST ASTQ9O96QK9 QANG9L4YFAP
RR9GVS9GDPLS9A NSL9RDLQ9RAG VLD9GIYD9GAS QYDE6NA9QF
K9K9Q8V9RQ CFY9NP9ISCF PK
P42559; Allatostatin. Manduca Sexta.
QVRFRQCYFN PISCF

(14) Q9U4J0; Edysis-TRIGGERING hormone (ETH) OR CG18105
M98PFCYLLC YLLL4LFFPA LSEARPGSAEG TQD3GDDGLGQV
QVARVGGYGGYDMPAQIPA YIHYPMRFL QMFLAQYRTP
YSAYLSPTY GWNL4LYRFL ESFQQYVRBQ CFY4NPSFCR
K
P42559; Allatostatin. Manuda Sexta.
QYFRQCVQYF QYFC

(15) Q9VC44; Allatostatin (AST) OR BCDNA:RE16553 OR CG13633
M98PFCYLLC YLLL4LFFPA LSEARPGSAEG TQD3GDDGLGQV
QVARVGGYGGYDMPAQIPA YIHYPMRFL QMFLAQYRTP
YSAYLSPTY GWNL4LYRFL ESFQQYVRBQ CFY4NPSFCR
K
P42559; Allatostatin. Manuda Sexta.
QYFRQCVQYF QYFC

(16) Q9VCW0; FLP-1 OR F23B2.5 (CCAP) OR CG4910
M98PFCYLLC YLLL4LFFPA LSEARPGSAEG TQD3GDDGLGQV
QVARVGGYGGYDMPAQIPA YIHYPMRFL QMFLAQYRTP
YSAYLSPTY GWNL4LYRFL ESFQQYVRBQ CFY4NPSFCR
K
P42559; Allatostatin. Manuda Sexta.
QYFRQCVQYF QYFC

(17) Q9VG55; Hugin protein precursor (HUG) OR CG6371
M98PFCYLLC YLLL4LFFPA LSEARPGSAEG TQD3GDDGLGQV
QVARVGGYGGYDMPAQIPA YIHYPMRFL QMFLAQYRTP
YSAYLSPTY GWNL4LYRFL ESFQQYVRBQ CFY4NPSFCR
K
P42559; Allatostatin. Manuda Sexta.
QYFRQCVQYF QYFC

(18) Q9VIQ0; Short neuropeptide F (SNPF) OR CG13968
M98PFCYLLC YLLL4LFFPA LSEARPGSAEG TQD3GDDGLGQV
QVARVGGYGGYDMPAQIPA YIHYPMRFL QMFLAQYRTP
YSAYLSPTY GWNL4LYRFL ESFQQYVRBQ CFY4NPSFCR
K
P42559; Allatostatin. Manuda Sexta.
QYFRQCVQYF QYFC

(19) Q95SV5; CG13586 OR BCDNA:SD05282.
M98PFCYLLC YLLL4LFFPA LSEARPGSAEG TQD3GDDGLGQV
QVARVGGYGGYDMPAQIPA YIHYPMRFL QMFLAQYRTP
YSAYLSPTY GWNL4LYRFL ESFQQYVRBQ CFY4NPSFCR
K
P42559; Allatostatin. Manuda Sexta.
QYFRQCVQYF QYFC

(20) Q9VLK4; Diuretic hormone class-II precursor (DH31) OR CG13094.
M98PFCYLLC YLLL4LFFPA LSEARPGSAEG TQD3GDDGLGQV
QVARVGGYGGYDMPAQIPA YIHYPMRFL QMFLAQYRTP
YSAYLSPTY GWNL4LYRFL ESFQQYVRBQ CFY4NPSFCR
K
P42559; Allatostatin. Manuda Sexta.
QYFRQCVQYF QYFC
(23) Q9VV77: Allatostatin/MIP OR CG6456

MAHKKRTRTY FLMVVLILLG SAGCNLVASG SAGSPPSNEP
GGGGLSEQV QLRLQSESLD GNQRMASQL GQQNSRKRSSS
GVDDPQDYM TGSHFVPLDV DTINNWDT PERLASQQA
QCOQTPQQQO SQQOQDDDFD LADDVIPVER AKQMNYYVAGC
KRRQACQWNK FRAGKWKRP TWNLKGMGM KRDQQPKLQG
GKIGFRQPSL N

Q85YP7: Prohormoncystic peptide (PTSP). Bombyx mori.

QMRCLFLALV FQGVATVTAA EPEHDAAQQP TDENVLiLTD
DGRANSLLIS IQNMQPSEQE YQSLARALPN LNQMLRRERQ
LAVGEEQPLG EYPDYLEE NVSPEMESL LAVGEEQPLG EYPDYLEE
LPPGNHKLCG PALSDAMDVV CPHGFNTLPR KRESLLGNSD
MFSQHNGAAV HGLRLQSLLI AAMLTAAMAM VTPTGSGHQL
1 precursor (ILP1) OR CG141 (26) Q9VT50; Probable insulin-like peptide 3 precursor (ILP3) OR CG14167). MG1R5MRQOR RILLPSLLL 111MIGQVQAT MKLGRKLPF
TLKLKCVQVF NAMTERRFLDP VNFQIGDGF BRSELLLDS
DQSVQMKTTR LQGQVPPEDC CLAECTHMD EV LAYCAAKRRP VTCNKL.
P26773; Bombyxin B-1 precursor (BBXBL). Bombyx mori.

MKT5VSNFLV IVISLMCQSG AQVRVTYGC HRAHTLRLAD
CPEVKERARGA QAPYFYKRTQ YLGSRKQRV VDCCRFPRCT
LQVLSYCG

(24) Q9WO6; Neuropeptide like 1 precursor (NPL1)

MQAQVLSAHS SRRILMMLS LNAAJQPPRS IIVSATDODA
NVPSCEMSL IQNMQPSEQE YQSLARALPN LNQMLRRERQ
LAVGEEQPLG EYPDYLEE NVSPEMESL LAVGEEQPLG EYPDYLEE
LPPGNHKLCG PALSDAMDVV CPHGFNTLPR KRESLLGNSD
MFSQHNGAAV HGLRLQSLLI AAMLTAAMAM VTPTGSGHQL
1 precursor (ILP1) OR CG141 (26) Q9VT50; Probable insulin-like peptide 3 precursor (ILP3) OR CG14167). MG1R5MRQOR RILLPSLLL 111MIGQVQAT MKLGRKLPF
TLKLKCVQVF NAMTERRFLDP VNFQIGDGF BRSELLLDS
DQSVQMKTTR LQGQVPPEDC CLAECTHMD EV LAYCAAKRRP VTCNKL.
P26773; Bombyxin B-1 precursor (BBXBL). Bombyx mori.

MKT5VSNFLV IVISLMCQSG AQVRVTYGC HRAHTLRLAD
CPEVKERARGA QAPYFYKRTQ YLGSRKQRV VDCCRFPRCT
LQVLSYCG

(25) Q5YIK3; Neuropeptide F (NPF) OR CG10342

MQOQRMCLVL ACVALALLAA GCRVSAASNR PPRKNDVNMT
ADAYQKFLQL DITYGDARV RFGRKGSILME ILRNHEMDI
NLQKNAGNFG EFVSIE
P23442; Islet amyloid polypeptide precursor (IAPP). Rattus norvegicus.

MESR5KQQPL AQ1LTLSQVF YQSLARALPN LNQMLRRERQ
LAVGEEQPLG EYPDYLEE NVSPEMESL LAVGEEQPLG EYPDYLEE
LPPGNHKLCG PALSDAMDVV CPHGFNTLPR KRESLLGNSD
MFSQHNGAAV HGLRLQSLLI AAMLTAAMAM VTPTGSGHQL
1 precursor (ILP1) OR CG141 (26) Q9VT50; Probable insulin-like peptide 3 precursor (ILP3) OR CG14167). MG1R5MRQOR RILLPSLLL 111MIGQVQAT MKLGRKLPF
TLKLKCVQVF NAMTERRFLDP VNFQIGDGF BRSELLLDS
DQSVQMKTTR LQGQVPPEDC CLAECTHMD EV LAYCAAKRRP VTCNKL.
P26773; Bombyxin B-1 precursor (BBXBL). Bombyx mori.

MKT5VSNFLV IVISLMCQSG AQVRVTYGC HRAHTLRLAD
CPEVKERARGA QAPYFYKRTQ YLGSRKQRV VDCCRFPRCT
LQVLSYCG

(26) Q9VT50; Probable insulin-like peptide 1 precursor (ILP1) OR CG141

MFSQHNGAAV HGLRLQSLLI AAMLTAAMAM VTPTGSGHQL
LPGFNNKLCLG PALSADAVV CFHGRFQFLFR KRESELQND
KDQETDQEYQ QDSSQWQDLT GAGSFSPLL TNYGLSEVL
KMRHRHLHT GQYDECCVR TCSYLEALTY CLFK
(27) Q9VT51; Probable insulin-like peptide 2 precursor (IRP) OR CG8982

MS9LPSFSM VMVIALLSST KQALOQCTS KLNOEMLSTY
CEEYNVPVH KRAMPSGD LDDNLPLFQV QFEEEDENDI
SEPLRSALFP YQSLVGFLVS LAVRRTTRQ RQVGIVERCC
KSCKMRALEL YCVSSRNN
P01325; Insulin 1 precursor (INS1 OR INS-1). Mus musculus.

MALLVHPLFL LALLALWQFR PTQAFVQHDL CGPHLEALV
LVCGERFFFY TPRKREVED PQQEQLLESGG DQLQGTLQ
ETVQAAAQIV Gir DOTCTCGCQ GLQLENYCN
(30) Q9H4Q9; CG13117 protein (ILP7) OR CG13137

AAQ89696; Insulin-like peptide 5 precursor. Anopheles gambiae.

MNLPLACLVL LEPADFIVAS GGLLEELDIY ETQVETTAAE YLRATRKRK
AAMIENKRTS TEFQGDSWTW HYFNNHFLR DRRSDGPTS ISEMCTTQS CTWYAVEACF PSE
AAQ89696; Insulin-like peptide 5 precursor. Anopheles gambiae.

MNLPLACLVL LEPADFIVAS GGLLEELDIY ETQVETTAAE YLRATRKRK
AAMIENKRTS TEFQGDSWTW HYFNNHFLR DRRSDGPTS ISEMCTTQS CTWYAVEACF PSE
AAQ89696; Insulin-like peptide 5 precursor. Anopheles gambiae.

MNLPLACLVL LEPADFIVAS GGLLEELDIY ETQVETTAAE YLRATRKRK
AAMIENKRTS TEFQGDSWTW HYFNNHFLR DRRSDGPTS ISEMCTTQS CTWYAVEACF PSE
AAQ89696; Insulin-like peptide 5 precursor. Anopheles gambiae.

MNLPLACLVL LEPADFIVAS GGLLEELDIY ETQVETTAAE YLRATRKRK
AAMIENKRTS TEFQGDSWTW HYFNNHFLR DRRSDGPTS ISEMCTTQS CTWYAVEACF PSE
AAQ89696; Insulin-like peptide 5 precursor. Anopheles gambiae.

MNLPLACLVL LEPADFIVAS GGLLEELDIY ETQVETTAAE YLRATRKRK
AAMIENKRTS TEFQGDSWTW HYFNNHFLR DRRSDGPTS ISEMCTTQS CTWYAVEACF PSE
AAQ89696; Insulin-like peptide 5 precursor. Anopheles gambiae.

MNLPLACLVL LEPADFIVAS GGLLEELDIY ETQVETTAAE YLRATRKRK
AAMIENKRTS TEFQGDSWTW HYFNNHFLR DRRSDGPTS ISEMCTTQS CTWYAVEACF PSE
AAQ89696; Insulin-like peptide 5 precursor. Anopheles gambiae.

MNLPLACLVL LEPADFIVAS GGLLEELDIY ETQVETTAAE YLRATRKRK
AAMIENKRTS TEFQGDSWTW HYFNNHFLR DRRSDGPTS ISEMCTTQS CTWYAVEACF PSE
AAQ89696; Insulin-like peptide 5 precursor. Anopheles gambiae.

MNLPLACLVL LEPADFIVAS GGLLEELDIY ETQVETTAAE YLRATRKRK
AAMIENKRTS TEFQGDSWTW HYFNNHFLR DRRSDGPTS ISEMCTTQS CTWYAVEACF PSE
AAQ89696; Insulin-like peptide 5 precursor. Anopheles gambiae.

MNLPLACLVL LEPADFIVAS GGLLEELDIY ETQVETTAAE YLRATRKRK
AAMIENKRTS TEFQGDSWTW HYFNNHFLR DRRSDGPTS ISEMCTTQS CTWYAVEACF PSE
AAQ89696; Insulin-like peptide 5 precursor. Anopheles gambiae.

MNLPLACLVL LEPADFIVAS GGLLEELDIY ETQVETTAAE YLRATRKRK
AAMIENKRTS TEFQGDSWTW HYFNNHFLR DRRSDGPTS ISEMCTTQS CTWYAVEACF PSE
AAQ89696; Insulin-like peptide 5 precursor. Anopheles gambiae.
(33) P10334; Accessory gland-specific peptide 26Ab precursor (ACP26AB OR MST26AB OR MST355B OR CG9024) OR MST26AB OR MST355B). Drosophila mauritiana.

(34) P16548; Accessory gland-specific peptide 26Ab precursor (ACP26AB OR MST26AB OR MST355B). Drosophila mauritiana.

(35) P33738; Accessory gland-specific peptide 26Ab precursor (ACP26AB OR MST26AB OR MST355B). Drosophila mauritiana.

(36) P36193; Drosocin precursor (DRO) OR CG10816.

(37) P14954; Cecropin A1/A2 precursor (CECA1) OR CG1367. Drosophila melanogaster.

(38) P14954; Cecropin A1/A2 precursor (CECA1) OR CG1367. Drosophila melanogaster.

(39) P45884; Attacin A precursor (ATTA) OR CG10146.

(40) P45884; Attacin A precursor (ATTA) OR CG10146.

(41) P82706; Immune-induced protein 1 precursor (IM1) OR CG18108.

(42) O77150; Immune-induced protein 2 precursor (IM2) OR BCDNA:RH08291 OR CG18106.

(43) Q9VLV9; PROCT OR CG7105. CG7105 protein MGVRPRSHGTG IGCGSGHRWL LVWMTVLLVV VPPHLVDGR YLPTRSHGDDL DKLREIAEDH LQAASAIRPP ILPA

(44) P10419; Antho-RFamide neuropeptides type 1 precursor. Anthopleura elegantissima.

(45) P14954; Cecropin A1/A2 precursor (CECA1) OR CG1367. Drosophila melanogaster.


(47) P14954; Cecropin A1/A2 precursor (CECA1) OR CG1367. Drosophila melanogaster.

(48) P14954; Cecropin A1/A2 precursor (CECA1) OR CG1367. Drosophila melanogaster.

(49) P14954; Cecropin A1/A2 precursor (CECA1) OR CG1367. Drosophila melanogaster.

(50) P14954; Cecropin A1/A2 precursor (CECA1) OR CG1367. Drosophila melanogaster.

(51) P14954; Cecropin A1/A2 precursor (CECA1) OR CG1367. Drosophila melanogaster.

(52) P14954; Cecropin A1/A2 precursor (CECA1) OR CG1367. Drosophila melanogaster.

(53) P14954; Cecropin A1/A2 precursor (CECA1) OR CG1367. Drosophila melanogaster.

(54) P14954; Cecropin A1/A2 precursor (CECA1) OR CG1367. Drosophila melanogaster.

(55) P14954; Cecropin A1/A2 precursor (CECA1) OR CG1367. Drosophila melanogaster.

(56) P14954; Cecropin A1/A2 precursor (CECA1) OR CG1367. Drosophila melanogaster.

(57) P14954; Cecropin A1/A2 precursor (CECA1) OR CG1367. Drosophila melanogaster.

(58) P14954; Cecropin A1/A2 precursor (CECA1) OR CG1367. Drosophila melanogaster.
The putative peptide precursors (28)

(1) P02841; Salivary glue protein Sgs-7 precursor (SGS7) or CG18087

MIAVPTIAC CILLIGFSQL ALOGACACQP CGPGKACCTG
CPEKPOLCQQ LIDSRNILQQ KTTRCVGCQEP QMMI

P35455; Vasopressin-neurophysin 2-copeptin precursor (AVP). Mus musculus.

MLAMMAJMTL SAPCFLLSAF SACYPQNPC RGGEKAISSDM
ELRQLCPCQG KGGRCGFQPS ICACDELGF CVTFAELQRQ
EENYLPSPCQ SQGPKCSGSG RCAAVCIGS DESCAVEACED HDGFFRULTRA REPQSNALD GFARALLRQ VQLAGTRESV DSARKF

(2) P07701; Salivary glue protein Sgs-5 precursor (SGS5) OR CG7596.

MFNIKLILLLL LAVSFHHGQ AVQETKEEK PVSEPEIIBSE

(3) Q27241; EIG71ED protein (EIG71ED) OR L71-4 OR CG7350

MHTTAVTFLF SVLTVLVVAQ QRNRCDELR TRCERCETLV
NAADRNLPVL QUQRCRTKRN NRRWGNVRC ETLRLNCGLS

(4) Q00805; Giant-lens protein precursor (ARGOS) OR AOS OR GIL

MPTTLMILPC MLLLLITAAA VAVGOSTRLPV EVFETTPTTS

(5) Q24155; Trunk protein precursor (TRK) OR CG5619

MKQGQSLAEV LTWAVLGTTA QDDADCAEL STOSLAKILMG QAFNPRLMIS DFPGEPEEKS YHLGKRYGKS ELFYADSADA
ISVSHFPMTE TNNHFAVEKL KPEAKESKLTR TPLSNMDFVRG

(6) Q8IME0; CG32851

MLKLSITFFL GLLALANSAP LSPDNPGNV IINGDCVNGC

(7) Q9V8F7; CG18107

MRFAPITTVF VGLLALANAI PLSPDNPVGN IIENCGCWNC

(8) Q9V8G2; CG15065

MRWNSLFLVC CLLAHAVAFLN LSPGNIVGNC DRCHVCVRG

(9) Q9VD48; CG5791 OR BCDNA:RH16331

MKLLSITFLF GLLALASA

(10) Q8IMR6; CG31081

MLGIVFLTLL AGSSA

Accession:Q16992; LWamide neuropeptides precursor. Anthopleura elegantissima.

MLGIVFLTVLF MLKLSITFFL GLLALANSAP LSPDNPGNV IINGDCVNGC

(11) Q8IMB6; CG31081

MLGIVFLTVLF MLKLSITFFL GLLALANSAP LSPDNPGNV IINGDCVNGC

(12) Q9VD48; CG5791 OR BCDNA:RH16331

MKLLSITFLF GLLALASA

Accession:Q16992; LWamide neuropeptides precursor. Anthopleura elegantissima.
(11) Q8MS86; BCDNA:LP04693
MSCSAWTQTP THTHKHRAIQ IVTISVILII ECSALVACSL TPTSSLFALH RWK1L1TACS HFWK
P11885; Corticotropin-lipotropin precursor. Rana catesbeiana.
MLQPVWHACI LAILGVFVHR GVEGRSCQNE SNKCTDLSES DGLIECIAK MRLDSAE5PV FP4FGHNIQPYL SE4NIKYMVS
HEP6NNKHGR NTSNDDN2NN NNGKREDIA NVKYY4P4K TPTSSLFALH RR
WKILTAGS HFRW L

(12) Q8M6O6; Odor binding protein b(OBP-B)
MRVLLAFVLL LGLSVLATK E PELEVKIVSM C AKENNVH RKK ALDLLMSYRL KKKTHNVMCF INCIFERTNI LQKVKEKVVK ENHNCDSIKD ADKCAESFQK FQCLVKIEMK VRGIDRG P01272; Glucagon precursor(GCG).Bos Taurus.
MKSLYFVAGL FVMLVQGSWQ RSLQNTEEKS SSFPAPQTDP LGDPDQINED KRHSQGTFTS DYSKYLDSRR AQDFVQWLMN TKRNKNNIAK RHDEFERHAE GTFTSDVSSY LEGQAAKEFI AWLVKGRGRR DFPEEVNIVE E LRRR0HADGS FSDEMNTVLD SLATRDFINW LLQTKITDRK

(13) Q8M5U18; GH13848p(CG14995).
MGLRSGVWGL GDLPTLLLTS VAPAAFTRPN SRGNLLTSLE VSTLERVTIS WTGKTRAPRE KRGSDLSDVS IIKRMRGVEV LALSVNKIST LSTFEDCTKL QELYLRKNSI SDINEIAYLQ NLPSLRNLKL ENEFCAHCAG LPLNLKLDNV EVTPQVQEDA LGQOVAAVAPE DEVEDYAEQD QQORSRFSQQ PQTQC000QY PQSQPQQQO QQOGQ000QQ QQQGCTTPLK EEYYQSDRPA YPAHYRHSQT DLTEWEEHQQ VPQVHHNPYG SQKQLHQPQR RSAGPEMTYP RNGSARENGG EWDPEDRSRA RRPEGRYSQD TSSLSASVNN HYSGYHRPPR NRNSILSAD LCLVKELDYA SLEVLEHAVR CRIDELANE

(14) Q8MVS0; Neuropeptide FF-related PQRF(PQRF).
Brachydanio rerio.
MNGLLEDRLL VEMLR SLLHG SQRQERNPSV LHQPQR FGSGLSTEER IQSRDWETVP GQIWSMAVPQ RFGKK A

(15) Q8JG0U; Neuropeptide FF-related PQRF(PQRF).
Brachydanio rerio.
MGRVRLCLCV SLLVISCLAQV GMFVPIILDV DLDFDTNPLL KVADYGELRM TDNLTVPS RSR1TLFPLW LGTQSNQNL TNVPHJMNRR FRNNVIVUA HS

(16) Q9VIQ6; CG16772
MPKLLLISQ LLLASYQALS LEAAKQIDIA SLYES85TE TDUTLHFPLHP PQYQPQHPHHR KVTFTTPPEE TTTTPFRPETT KPA1EQEGERA TAAFPDQLGG LQDGVQQQNN4 GDALSPETTP ERETPPTSPT TITTTPKPKHK RHPHRFHDHP HPHFYPVPYIQ YYYPHPGVIFK SAAGKP1LTPP STTPPTAKDA KRESPELSGY PSYTPFRSPY SPQVIPVPFPQ FRNPWSFGXY GPPSGCFQGYPY HNNHRHHEDD SEGKDPRKDS GEDKEDDEVA LKQFGPFSYYY VYLIFRFVPI SVYSPSFPQRG QGSYSPYGL Y
P81819; Carcinustatin 16. Carcinus maenas. QGPPSYGL

(17) Q9W1F8; CG13565
MLNVYLLAVV SVFIPHHAB PGDIN5DE LDGKYLEAQ SKYDGGPFTV LRSAANQVTV YYECVOCQSE FKTSVYQKPC AAGKISSGHG RDLVYLVRLM DFLYKDWTSS KLKRNFDEID KASASFIINL QLV P37086; Orcocinin. Orconectes limosus. NFDEIDRSGF GFN

(18) Q9VWY7; CG11577
MLLKRLPGFV TLWVVLQLAG ADSPEEEQGV RYANRCEACK ILATEEARL GET0GSHDV TIGYSSVDDVK PPKTTEYRRS ELRLLE5EN CVCERYN5L HERSSTRDP AG0SSSTQFO LHLGVDVPLW VDG1P1YELW DKPVEVTMQ TKCENLLE YE10SE6WY H1QDE5LLK HIC6EDVLRK KAREC5LEQ LEAPPAER REAKGK0EKE L EYQTSQ5; Granin-GAse secreting. Arvicathania anserigei. QAPFRFPAAP STAGGAAGTVL ARMYPRGSHK AVGHLMGKKS TOB0YALADDR DLGK0LQGAL VHKEAARANL LGLLE6ATGN W6QFPGWQGL 9FLTTDNSF DYPFDHFVQAK AKL0V0YVQ AKLGKEGTAS

(19) BK002023; HDC09365.
MMFSVVFVPL LFL1LVLLSSA QAANSILRAC PGDMML0RVA CPNGPSMFA KRTGLGFLDV EDHLADL0SS EHHNH0LS5 RIFREDGVDV SCDKEMCS0SF5 TLRAYCDS P15411; Bombyxin A-2 precursor (BBX2A). Bomyx mori. MIIHAL1MAL TLOMVW0STQ PQG0VTVYCS HLA5TMADLM CNEEG5D55RS QAOFASY5GA W3MPSAGGR IVDECLLRPC SVDVLSSYC

(20) BK002187; HDC10589.
MIIHAL1MAL TLOMVW0STQ PQG0VTVYCS HLA5TMADLM CNEEG5D55RS QAOFASY5GA W3MPSAGGR IVDECLLRPC SVDVLSSYC
(21) BK002297; HDC11617.
MVSQEFTHLV SIFWLSYLPK SLLSYGHDGH GLIQIDSSFI
NKKQLRKRFT DRKPRQAYSA SQLERLENEF NLDKYLSVSK
RVELSKSLLL TEYQRUVSLL SFPQVRULTV TSCQDHPTTV
SRSLAPYLSG LPLCSFKVRL ITVRPLCRLE SKVHRGPGNL
RQVGDAPGRL KKGKGDYWA NMNFYIVAVS ILIASAYARR EENNIQSLSQ
RDVLEEESLR EIRGIGASIL SAGKSALKGF AKGLAEHFAN GKRTAEDHEM
MKRLEAAVRL LDSLEHPFEEA SEKEFQFGNQ EEKREKD
MVSQEFTHLV SIFWLSYLPK SLLSYGHDGH GLIQIDSSFI
NKKQLRKRFT DRKPRQAYSA SQLERLENEF NLDKYLSVSK
RVELSKSLLL TEYQRUVSLL SFPQVRULTV TSCQDHPTTV
SRSLAPYLSG LPLCSFKVRL ITVRPLCRLE SKVHRGPGNL
RQVGDAPGRL KKGKGDYWA NMNFYIVAVS ILIASAYARR EENNIQSLSQ
RDVLEEESLR EIRGIGASIL SAGKSALKGF AKGLAEHFAN GKRTAEDHEM
MKRLEAAVRL LDSLEHPFEEA SEKEFQFGNQ EEKREKD

(22) BK002714; HDC14730.
MLLLLLVLLC LLLLLLLLLV LRFCVFAALA TQGPNQTRTV
DELCVQQLLL PGVIRAIAQL AQAARPQAQW AIVVQFTHTR
RMNVQNAGE PSERPFSRNN LTTQRASNLL LTSQDTATMA
SQPDGRKQDG HPACICYTADQ QRPPQFQQRE IPEARSQMDK
TQFGRKSGRD ILLSCCKFL LSFFGNTKGE SQDEVFTSIL
VELLRLRESL ALRKPPFATM INDQY
P55098; Neurokinin B precursor(TAC3 OR NKNB OR TAC2). Mus musculus.
MNFKYIVAVS ILIASAYARR EENNIQSLSQ RDVLEEESLR EIRGIGASIL SAGKSALKGF AKGLAEHFAN GKRTAEDHEM
MKRLEAAVRL LDSLEHPFEEA SEKEFQFGNQ EEKREKD
MLLLLLVLLC LLLLLLLLLV LRFCVFAALA TQGPNQTRTV
DELCVQQLLL PGVIRAIAQL AQAARPQAQW AIVVQFTHTR
RMNVQNAGE PSERPFSRNN LTTQRASNLL LTSQDTATMA
SQPDGRKQDG HPACICYTADQ QRPPQFQQRE IPEARSQMDK
TQFGRKSGRD ILLSCCKFL LSFFGNTKGE SQDEVFTSIL
VELLRLRESL ALRKPPFATM INDQY

(23) BK002770; HDC15079.
MQAIFILCAL LVCLLVLLRL SWSGGSESQ DVTTTNPKFL
VRKHDSSHYY WSDNIRMAHR EWLNSKSFAR QNNISRQDNL PMNDSYIDL PQFPDDTSDP SNAEDTTFIC
NFMLDFCYPD SGISFGLSAT IILILA
035417; Beta-neoendorphin-dynorphin precursor(MST138).Mus musculus.
MAWRRMLAAM CLVWNSMMV ADDCLSLCLC AVRIQDDQPRP
INPLICSLEC QDOVPFSCNE ETORQFSSFL TLTVSGGLAK
DDELDEVALE EGLSHAWEKL EPVLKELKEK LSLTSPERK
GRFLLSSESN GKESELAGAD MDNCAAQLGR TVQNFQEDLR
QKAPRKYSGFL KYKPRFRSMEE ARDEDODQGG DQVHEQELIK
RYGGFLRARL KLLKWQNDKR YGGFLLRQFK VVTRQGEPFN
TYSELDV

(24) BK003312; HDC00783.
MDTSIVIVIV IVIAIAIDFD IDIPGQLQEL ILLSRILGKL RQNGSQNAAT RTATHRTTFK DDDK
Q28318; Prolactin precursor(PRL). Capra hircus.
MKTSVIVIVV IVIAIAIDFD IDIPGQLQEL ILLSRILGKL RQNGSQNAAT RTATHRTTFK DDDK
MDTSIVIVIV IVIAIAIDFD IDIPGQLQEL ILLSRILGKL RQNGSQNAAT RTATHRTTFK DDDK
Q28318; Prolactin precursor(PRL). Capra hircus.
MKTSVIVIVV IVIAIAIDFD IDIPGQLQEL ILLSRILGKL RQNGSQNAAT RTATHRTTFK DDDK
Q28318; Prolactin precursor(PRL). Capra hircus.

(25) BK003730; HDC05827.
MDRLISITFL CWCIPVMISG ASLRAWVFNV EKCHFGDSTC
LVRSINALIK HYKPGIEPIG LRLPLAYNPC DSVIYEPSPR
Q28318; Prolactin precursor(PRL). Capra hircus.
MKTSVIVIVV IVIAIAIDFD IDIPGQLQEL ILLSRILGKL RQNGSQNAAT RTATHRTTFK DDDK
Q28318; Prolactin precursor(PRL). Capra hircus.
MKTSVIVIVV IVIAIAIDFD IDIPGQLQEL ILLSRILGKL RQNGSQNAAT RTATHRTTFK DDDK
Q28318; Prolactin precursor(PRL). Capra hircus.

(26) Q9VNV8; MSOPA OR CG14560
MNFIQAVLVL VLVAVLAPR QEDPANLPAP EAAAPPPAAA AAPPAAAAAP FAPPAPFAPAA PQAAPAGGSSR KKNVHNI
TIG

(27) Q9VDF8; CG3119
MKRKLSTFLY CMCIPVMSG ALRAVWNANX EUKCHFGDSTC
LVRSINALIK HYKPGIEPIG LRLPLAYNPC DSVIYEPSPR
Q28318; Prolactin precursor(PRL). Capra hircus.
MKTSVIVIVV IVIAIAIDFD IDIPGQLQEL ILLSRILGKL RQNGSQNAAT RTATHRTTFK DDDK
Q28318; Prolactin precursor(PRL). Capra hircus.
MKTSVIVIVV IVIAIAIDFD IDIPGQLQEL ILLSRILGKL RQNGSQNAAT RTATHRTTFK DDDK
Q28318; Prolactin precursor(PRL). Capra hircus.

(28) Q9W0X2; CG9358 OR BCDNA:RE09339
MRKASLAVFCC VCVGAAAPP EKTRYNKIDS VNMDEVLGN
Q28318; Prolactin precursor(PRL). Capra hircus.
MKTSVIVIVV IVIAIAIDFD IDIPGQLQEL ILLSRILGKL RQNGSQNAAT RTATHRTTFK DDDK
Q28318; Prolactin precursor(PRL). Capra hircus.
MKTSVIVIVV IVIAIAIDFD IDIPGQLQEL ILLSRILGKL RQNGSQNAAT RTATHRTTFK DDDK
Q28318; Prolactin precursor(PRL). Capra hircus.

(29) Q9W0X2; CG9358 OR BCDNA:RE09339
MRKASLAVFCC VCVGAAAPP EKTRYNKIDS VNMDEVLGN
Q28318; Prolactin precursor(PRL). Capra hircus.
MKTSVIVIVV IVIAIAIDFD IDIPGQLQEL ILLSRILGKL RQNGSQNAAT RTATHRTTFK DDDK
Q28318; Prolactin precursor(PRL). Capra hircus.
MKTSVIVIVV IVIAIAIDFD IDIPGQLQEL ILLSRILGKL RQNGSQNAAT RTATHRTTFK DDDK
Q28318; Prolactin precursor(PRL). Capra hircus.

(30) Q9W0X2; CG9358 OR BCDNA:RE09339
MRKASLAVFCC VCVGAAAPP EKTRYNKIDS VNMDEVLGN
Q28318; Prolactin precursor(PRL). Capra hircus.
MKTSVIVIVV IVIAIAIDFD IDIPGQLQEL ILLSRILGKL RQNGSQNAAT RTATHRTTFK DDDK
Q28318; Prolactin precursor(PRL). Capra hircus.
MKTSVIVIVV IVIAIAIDFD IDIPGQLQEL ILLSRILGKL RQNGSQNAAT RTATHRTTFK DDDK
Q28318; Prolactin precursor(PRL). Capra hircus.

(31) Q9W0X2; CG9358 OR BCDNA:RE09339
MRKASLAVFCC VCVGAAAPP EKTRYNKIDS VNMDEVLGN
Q28318; Prolactin precursor(PRL). Capra hircus.
MKTSVIVIVV IVIAIAIDFD IDIPGQLQEL ILLSRILGKL RQNGSQNAAT RTATHRTTFK DDDK
Q28318; Prolactin precursor(PRL). Capra hircus.
MKTSVIVIVV IVIAIAIDFD IDIPGQLQEL ILLSRILGKL RQNGSQNAAT RTATHRTTFK DDDK
Q28318; Prolactin precursor(PRL). Capra hircus.
<table>
<thead>
<tr>
<th>Species</th>
<th>Peptide name</th>
<th>Sequence</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mus musculus</td>
<td>γ-MSH</td>
<td>YVMGHFRWDRFamide</td>
<td>(55)</td>
</tr>
<tr>
<td><em>Rana catesbeiana</em></td>
<td>γ-MSH</td>
<td>YVMHFRWNNKFamide</td>
<td>(56)</td>
</tr>
<tr>
<td><em>Xenopus laevis</em> (African clawed frog)</td>
<td>γ-MSH</td>
<td>YVMTHFRWNNKFamide</td>
<td>(57)</td>
</tr>
<tr>
<td><em>Oncorhynchus keta</em> (chum salmon)</td>
<td>γ-MSH</td>
<td>HSYSMEHFRWamide</td>
<td>(58)</td>
</tr>
<tr>
<td>Theromyzon tessulatum</td>
<td>γ-MSH</td>
<td>YVMGHFRWDFK</td>
<td>(59)</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>γ-MSH</td>
<td>WKILTAGSHFRWL</td>
<td>This study</td>
</tr>
<tr>
<td>Species</td>
<td>Peptide name</td>
<td>sequence</td>
<td>Ref.</td>
</tr>
<tr>
<td>-------------------------</td>
<td>-----------------------</td>
<td>---------------------------</td>
<td>------</td>
</tr>
<tr>
<td><em>Orconectes limosus</em></td>
<td>Orcokinin</td>
<td>NFDEIDRSFGFN</td>
<td>(33)</td>
</tr>
<tr>
<td></td>
<td>[V&lt;sup&gt;13&lt;/sup&gt;]-Orcokinin</td>
<td>NFDEIDRSFGFNV</td>
<td></td>
</tr>
<tr>
<td><em>Carcinus maenest</em></td>
<td>[S&lt;sup&gt;1&lt;/sup&gt;]-Orcokinin</td>
<td>NFDEIDRRSGFN</td>
<td>(60)</td>
</tr>
<tr>
<td></td>
<td>[V&lt;sup&gt;15&lt;/sup&gt;]-Orcokinin</td>
<td>NFDEIDRRSGFNV</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[A&lt;sup&gt;15&lt;/sup&gt;]-Orcokinin</td>
<td>NFDEIDRRSGFNA</td>
<td></td>
</tr>
<tr>
<td><em>Procambarus clarkii</em></td>
<td>Orcokinin</td>
<td>NFDEIDRSSFGN</td>
<td>(61)</td>
</tr>
<tr>
<td></td>
<td>[V&lt;sup&gt;13&lt;/sup&gt;]-Orcokinin</td>
<td>NFDEIDRSSFGN</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[A&lt;sup&gt;15&lt;/sup&gt;]-Orcokinin</td>
<td>NFDEIDRSSFGNA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[T&lt;sup&gt;1&lt;/sup&gt;]-[H&lt;sup&gt;13&lt;/sup&gt;]-Orcokinin</td>
<td>NFDEIDRTGFGFNV</td>
<td></td>
</tr>
<tr>
<td><em>Cherax destructor</em></td>
<td>Orcokinin</td>
<td>NFDEIDRSSFGN</td>
<td>(62)</td>
</tr>
<tr>
<td></td>
<td>[V&lt;sup&gt;13&lt;/sup&gt;]-Orcokinin</td>
<td>NFDEIDRSSFGNV</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[A&lt;sup&gt;15&lt;/sup&gt;]-Orcokinin</td>
<td>NFDEIDRSSFGFNA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[T&lt;sup&gt;1&lt;/sup&gt;]-[H&lt;sup&gt;13&lt;/sup&gt;]-Orcokinin</td>
<td>NFDEIDRTGFGFNV</td>
<td></td>
</tr>
<tr>
<td><em>Cancer borealis</em></td>
<td>[V&lt;sup&gt;13&lt;/sup&gt;]-Orcokinin</td>
<td>NFDEIDRSSFGFNV</td>
<td>(34)</td>
</tr>
<tr>
<td></td>
<td>[A&lt;sup&gt;15&lt;/sup&gt;]-Orcokinin</td>
<td>NFDEIDRSSFGFNA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[S&lt;sup&gt;1&lt;/sup&gt;]-[V&lt;sup&gt;13&lt;/sup&gt;]-Orcokinin</td>
<td>NFDEIDRSSFGFNV</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NFDEIDRSGFA</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>Drm Orcokinin</td>
<td>NFDEIDKASASFSILN QLV</td>
<td></td>
</tr>
</tbody>
</table>

This study
<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>IM1</td>
<td>GNVIINGDCRVCNVHGamide</td>
<td>(15)</td>
</tr>
<tr>
<td>IM2</td>
<td>GNVINGDCKYCNVHGamide</td>
<td></td>
</tr>
<tr>
<td>IM3</td>
<td>GNVIINGDCRVCNVRA</td>
<td></td>
</tr>
<tr>
<td>IM4</td>
<td>GTVLIQTDNQYIRT</td>
<td></td>
</tr>
<tr>
<td>CG15065</td>
<td>GNVIINGDCRHCNVRGamide</td>
<td>This study</td>
</tr>
<tr>
<td>CG5791</td>
<td>GNTIVNGRCQHCNVDPY</td>
<td>This study</td>
</tr>
</tbody>
</table>
Protein Database

Select Drosophila proteins: type='Drosophila melanogaster'

Drosophila proteins

Predicted signal peptide information

combine

output

Drosophila proteins that have a signal peptides and that are less than 500 amino acids in length
Peptide.SAS

Protein Database

input

uniprot_sprot.dat
uniprot_trembl.dat
uniprot_trembl_new.dat

Select proteins with keywords:
(Neuro)peptide, hormone, neurotransmitter;
but without keywords: receptor, signal-anchor,
transmembrane, binding protein, DNA binding, nuclear
protein, nuclear transport, enzyme or words ending in ‘ase’

output

Known peptide precursor proteins from all metazoan

SAS Dataset

Fig 2
**Fig 3**

- **Drosophila** proteins that have a signal peptide and that are less than 500 amino acids in length
- Known peptide precursors from all metazoan organisms

**Cleavage.SAS**
- Cleave sequences using the conventional cleavage sites

**txt-file**
- Peptide subsequences in FASTA format

**BLAST**
- Run BLAST programs:
  - Query: *Drosophila* subsequences; target: *Drosophila* subsequences.
  - Query: Peptide subsequences; target: *Drosophila* subsequences.
  - Query: Peptide subsequences; target: Peptide subsequences.
BLAST comparisons

Subsequence comparisons among *Drosophila* subsequences

output

BLAST comparisons

Subsequence comparisons between *Drosophila* subsequences and peptide subsequences

output

BLAST comparisons

Subsequence comparisons among peptide subsequences

output

BLAST comparisons

txt-file

Extract.SAS

input

Extract the *Drosophila* proteins having at least 2 similar subsequences within the protein

output

Extract.SAS

Subsequence comparisons among *Drosophila* subsequences

input

Shift.SAS

Compute shift value

output

Shift.SAS

Compute shift value

input

Shift <= 3

Motif.SAS

Whether a *Drosophila* subsequence is similar to a motif

output

Motif.SAS

Whether a *Drosophila* protein has one single transmembrane region

input

TMpred or SOSUI online tools

Drosophila proteins, that contain at least one subsequence similar to a motif for which the shift value is no larger than 3

input

Drosophila proteins, that contain at least 2 similar subsequences, and for which the shift value of at least one subsequence is no larger than 3

input

The *Drosophila* putative peptide precursors

End

Fig 4
The protein sequence 'Q9V808' (The similar subsequences are in bold. The match (similar amino acid sequence tags) is underlined.):

Q9V808; CG6564 protein (CG30101 OR CG6564 OR CG15901). Drosophila melanogaster.
MRMFVLPCLA VCVALACGG A VE DE KAE GD G K T V EKRLH LGD Y H YQPH HEHIKTVIE KKIPVPYTVT
KHYPYTVKPK IPEYKVP KP YKPKVPK KHYPYTVKPK VPYKPKVPK KHYPYTVKPK VPYKPKVPK KHYPYTVKPK VPYKPKVPK
 The similar subsequences are in bold. The match (similar amino acid sequence tags) is underlined.:

Fig 5
The comparison between P11885_2 and Q8MS86_2:

Query=P11885_2
(11 letters)

>Q8MS86_2
Length = 13

Score = 22.3 bits (45), Expect = 1.1
Identities = 5/5 (100%), Positives = 5/5 (100%)

Query:  4 SHFRW 8
Sbjct:   8 SHFRW 12

The comparison between P11885_2 and P06298_8:

Query=P11885_2
(11 letters)

>P06298_8
Length = 17

Score = 22.7 bits (46), Expect = 0.32
Identities = 6/8 (75%), Positives = 6/8 (75%)

Query:  1 YVMSHFRW 8
Y M HFRW
Sbjct:   4 YRMHHFRW 11

The protein sequence 'Q8MS86', 'P11885' and 'P06298' (The similar subsequences are in bold. The match is underlined):

Q8MS86; LP04693p(BCDNA:LP04693). Drosophila melanogaster.
MSCSAMTWQTP THTKKHRAIQ IVTIISVLII ECSALVACSL TPTSSLPAH RWKILTAGS HFRWL

P11885; Corticotropin-lipotropin precursor. Rana catesbeiana.
MLQPVWHAC LAILGVFIFH VGEVRSQCWE SNACTDLSSE DGILLELCAC KMDLSAESPV FPGNGHIQPL
SENIRFVYMGS HFRWKPKF GRR NSTSNMDNNN NGGYKREDIA WYPILNIFLG SDMONTQEGI MEDDALOQD
SKRYSMHMF RWSHPVGKXR P1IKYFPDTA EERSSFTSPFI ELRRKLLELF DYPDTSSEE LONGELLEG
VKKGKRYMHI HFRWGPPKRD KRYGGFTMPTE RSQTP1MTLF KNAIKKNAHK KGQ

P06298; POMCA. Corticotropin-lipotropin A precursor. Xenopus laevis.
MFRLWGCFL AILGIFICFIH GEVQCSSQRES SRAQDSSES GVCCEIRACK TDLSAESPVF PGNHLPQLS
ESIRKHYMTH FRWNGQFRHN STQNGDSTQK YKREDISSYP VFLSFLPSQD NAPGNNMEE PLDRQENKRA
YSMHRFGK PVGKKRPPF VYVNGVEES AESYMELRR ELSLELYDE IDLDDEIEdN EVKSAITKRN
GNYRMHHFRW GSPPKDRKG YMPFRSQGT PLMTFLKNAI IRKSHKKGQ

Fig 6
Fig 7
Fig 8