Chapter 8
The Little Known Universe of Short Proteins in Insects: A Machine Learning Approach

Dan Ofer, Nadav Rappoport, and Michal Linial

Abstract Modern genomics and proteomics technologies are turning out immense quantities of sequenced proteins. The only feasible way to assign functions to this flood of sequences is by applying state-of-the-art computational methods for automated functional annotation. We illustrate the significance of machine learning tools in identifying and annotating short bioactive proteins and peptides from insect genomes. Over 500,000 full-length proteins from insects are currently archived in databases, of which ~15 % are short proteins. Currently, most short sequences remain uncharacterized. We developed a platform to systematically identify the functional class of short toxin-like peptides in metazoa. We present data from eight representative genomes (140,000 proteins) that cover the main phylogenetic branches of Hexapoda. The platform is a trained machine-predictor that successfully identified ~800 toxin-like candidates, 250 of them predicted with high confidence. These proteins' functions include ion channel inhibition, protease inhibitors, antimicrobial peptides, and components of the innate immune system. Our systematic approach can be expanded to new genomes and other biological classes of proteins. Using similar methodologies, we illustrate the success of identifying overlooked neuropeptide precursors. The systematic discovery of insect neuropeptides and short toxin-like proteins allows developing new strategies for pest control and manipulating insects' behavior. The overlooked secreted short peptides are discussed with respect to their evolution and potential applications in biotechnology.

D. Ofer • M. Linial (✉)
Department of Biological Chemistry, Life Sciences Institute, The Hebrew University of Jerusalem, Jerusalem, Israel
e-mail: ddofer@gmail.com; michall@cc.huji.ac.il

N. Rappoport
School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel
e-mail: nadavrap@cs.huji.ac.il

C. Raman et al. (eds.), *Short Views on Insect Genomics and Proteomics*,
Entomology in Focus 3, DOI 10.1007/978-3-319-24235-4_8

Abbreviations

ANN	Artificial neural network
AUC	Area under ROC curve
CAFA	Critical automatic functional annotation
ClanTox	Classifier of animal toxins
CNS	Central nervous system
CV	Cross validation
ETH	Ecdysis-triggering hormone
HMM	Hidden Markov model
ICI	Ion channel inhibitor
MFS	Major facilitator superfamily
ML	Machine learning
MS	Mass spectrometry
nAChR	Nicotinic acetylcholine receptors
NGF	Nerve growth factor
NP	Neuropeptide
NPP	Neuropeptide precursor
OCLP	ω-Conotoxin-like protein
PSSM	Position-specific scoring matrix
SP	Signal peptide
SVM	Support vector machine
TIL	Trypsin inhibitor like
TOLIPs	Toxin-like proteins
TLS	Toxin-like stability

8.1 Automated Functional Classification of Proteins: Sequence Similarity

In recent years, there has been an exponential increase in biological data, particularly of gene and protein sequences. The rapid growth rate of protein sequence data cannot be handled by performing individual experimental studies to determine the function(s) of every single protein, as was traditionally the case. Therefore, computational prediction is currently the only feasible approach for high-throughput identification of protein function [1].

Generally, functional classification is performed using a supervised approach, i.e., inferring functional classification for a sequence according to existing sequences whose functions are known. The most naïve, supervised approach is the *nearest-neighbor* search [2]. In practical terms, a database of sequences is searched for a query sequence with the goal of identifying similar sequences. The most common algorithms and search engines that perform this task are BLAST and FASTA [3]. If

a significantly similar sequence is found, the query sequence will be considered to possess a similar function; this concept is often considered as "guilt by association." The "rule of the thumb" for this inference has been defined as the *twilight zone* concept [4]: for a sequence at least 100 amino acids long, it is most likely to be a homologue if at least 30 % of the amino acids are identical. Below this value, the sequence is in the "twilight zone," where the similarity cannot be separated from randomly occurring similarity.

Although this direct inference approach is useful for many sequences, it suffers from critical caveats:

1. In order to learn about a sequence, there must exist a significantly similar sequence whose function is known, essentially precluding function prediction for unknown protein families.
2. Many proteins with similar sequences have different functions and would therefore be mistakenly classified as having the same function. Such cases are common for *paralogs*.
3. Many proteins exist that share functionalities and active sites or domains but possess significantly different sequences, despite having similar functions.

8.2 Functional Classification of Proteins: An Ill-Defined Term

There is an obvious connection between the "granularity" of a function (general or specific) and the evolutionary diversity of the proteins that share it [5]. Typically, groups of proteins that share a high-level functionality (i.e., enzymes) are much more diverse than low-level (e.g., urease enzymes) functionality groups [6]. This simple notion serves to define a scoring method for functional similarities [7]. The critical automatic functional annotation (abbreviated CAFA) initiative serves to set a measure for the success and failure in functional assignment. Open competitions for functional assignment over thousands of proteins show that there is considerable room for improvement [8].

There is an obvious interest in having classification machines successfully learn *high-level functionality*. If the training set consists of proteins that share a low-level functionality, the classifier would only be able to detect proteins that belong to the narrow function that was learned. Essentially, this would reproduce the main caveat of the nearest-neighbor search, i.e., the need to have a known, near-identical representative of every possible biological functional group. However, if the training set consists of proteins that share a high-level function, the classifier will be able to detect any protein that belongs to a very broad class, even if "close" representatives are unknown.

The point can best be demonstrated via an example: Consider the case of the major facilitator superfamily (MFS). This superfamily includes over 300,000 proteins

capable of transporting small solutes in response to ion gradients [9]. In general terms, proteins of the family belong to the transmembrane transport system. However, if we use classifiers of low-level functionality, we would have a set of classifiers that classify the different types of MSFs (including polyol permease, nitrate transporter, multidrug resistance protein, sialic acid transporter, and many more). If a sequence of a novel subtype of MSF were found, we would not be able to identify its function at the lower level as it would not belong to any known transporter family. However, if we had access to an MSF classifier, we could identify the sequence as a novel type of MSF transporter. An illustration of such an instance is shown in Fig. 8.1.

It is difficult to learn high-level functionality computationally, particularly when we might barely understand them on the theoretical level. While we might expect nitrate transporters from different organisms to share similar sequences due to evolutionary homology, we would not expect this of a high-level group such as all transmembrane transporters. "Statistical modeling" methodologies can then be applied [10]. Central to the success of these methods is construction of multiple sequence alignments [11]. It is plausible to characterize a new sequence with direct sequence-based techniques such as position-specific scoring matrices (PSSMs) and hidden Markov models (HMMs) [12]. These methods are widely used for sequence classification and identification. A large resource for families and domains in proteins is structural knowledge (often reliant on experimental sources), together with statistical modeling of each family [12]. A resource that relies entirely on automatic learning schemes provides a complementary view [13]. These methods were successfully applied to characterize features such as conserved positions and active

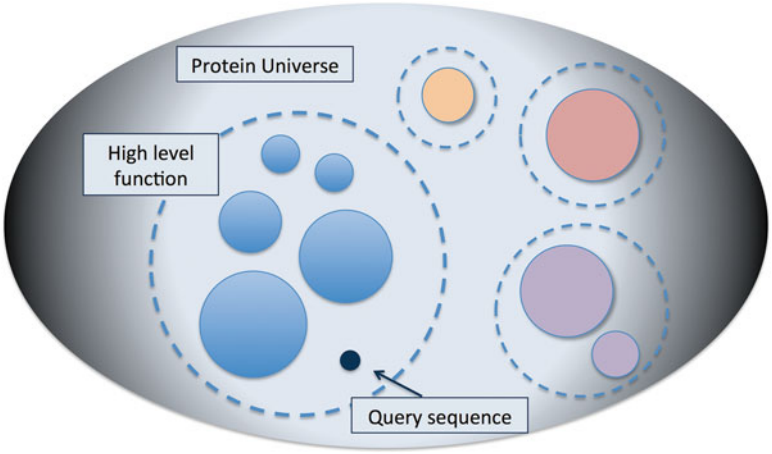

Fig. 8.1 Assigning functional annotation to a novel protein. Functional classes are represented as *colored circles* (i.e., the protein sequences that are similar and functionally related appear next to each other in this space). The level of functionality is relative to the diameter of the *dashed circles*. The *black dot* represents a novel sequence of unknown function. In this example, the sequence does not belong to any low-level functional classes (marked as *circles*) but does belong to a high-level function class (the *dashed circle*)

sites in enzymes and other binding sites in proteins [14]. However, there are cases in which multiple sequence alignment is unfeasible or simply uninformative.

8.3 Functional Classification of Proteins: Machine Learning

When sequence alignment is meaningless (e.g., only a small number of sequences can be aligned, or the information content of the multiple alignment is minimal), other methodologies have to be adapted. Some methods produce function classifiers that are not directly sequence based, but rather rely on "global" sequence-derived quantitative features that are extracted from the sequence and typically do not take amino acid sequential positions into consideration (e.g., the length of the protein, the amino acid frequencies, the weight of the protein) [15]. In such instances, the sequences are transformed into multidimensional space where each protein is represented by vectors of features in that space. Then, a classifier is learned by a statistical approach such as a support vector machine (SVM) [16] or a decision tree classifier method such as random forests [17]. Such methods which avoid the requirement for sequence alignments have shown success in learning high-level functional traits (such as the high level of protein family structural folds) while often being far more computationally efficient. While both the direct sequence-based approaches and the sequence-derived feature approaches may use the same information as input, namely, the sequence itself, they can perform very differently due to the manner in which they exploit the data and the information they extract from it. There are cases in which an intelligent choice of numerical features (i.e., those that can best capture the characteristics of the relevant sequences) can significantly outperform popular alignment models (e.g., HMM and PSSM) and can be extended for a variety of other data structures such as gene co-expression data.

We have now set the stage and background on the difficulties and solutions for functional inference of novel protein sequences. In this chapter, we show how for large sets of insect proteins we applied statistical learning methods to annotate unknown sequences as belonging to distinct, high-level functional classes. We considered sequences that are impossible to characterize by existing direct sequence-based methods (because the sequences are not alignable), but for which global, sequence-derived features can successfully characterize them exceptionally well.

The statistical learning technique framework is based on the notion of supervised machine learning methods, in which a group of known, annotated sequences serve for the *learning*/*training* phase. This group of sequences is referred to as the *training set*. Once the learning stage is complete, the characterization that was computationally learned, known as the *hypothesis*, can be used to classify unidentified sequences. The advantage of machine learning methods over the sequence similarity, nearest-neighbor approach (described above) is in the fact that the learning methods can identify the minimal conserved set of characteristics in each family of proteins and focus on searching only for these characteristics. This makes the learning methods much more powerful than the naïve sequence similarity approach.

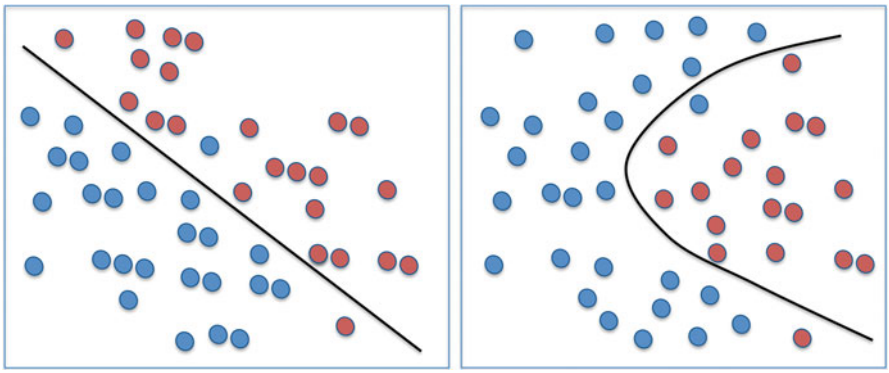

Fig. 8.2 Support vector machine (SVM) classification between labeled instances (*red* and *blue*). The *black line* represents the separator. The scenarios show a linear and a nonlinear SVM classification

The principal goal in the machine learning approach is to regard the problem as one of supervised classification or prediction in a binary (i.e., "Yes/No," "True/False," Positive/Negative) or a multi-class prediction problem. In a binary classification problem setting (such as SVM classification) [18], the goal is to classify two classes of points (indicated as positives and negatives) by constructing an optimal separator according to distinguishing features of items belonging to those sets so as to distinguish optimally between them and to classify new instances as belonging to one class or the other. The separator is set to ensure a maximal margin to the points (Fig. 8.2). The separator does not have to be linear and derives from a class of similarity measures (so-called kernel functions). The use of such a technique provides a set of "classifiers" that can then be tested and then used to predict "unlabeled" new instances [19].

The field of machine learning is immense with a strong impact on predictive biology [20, 21]. Machine learning technologies cover supervised methods of binary classification including various SVMs and kernels, decision trees, artificial neural networks (ANNs), ensembles of classifiers (such as random forests and AdaBoost) [17], and unsupervised methods for clustering [22].

8.4 Short Proteins: An Overlooked Niche

The ability to learn about a protein by comparing it to its (inferred) homologues has been used in functional prediction, secondary structure prediction, three-dimensional fold prediction, and several other applications. However, the power of sequence similarity-based tools is greatly diminished for short protein sequences. This is because when comparing short sequences it is difficult to distinguish genuine homology from mere evolutionary noise/coincidence. For example, submitting a short amino acid sequence to a sequence similarity search server such as BLAST

will usually result in matches with barely significant or insignificant e-values (a statistical measure of significance for the expectation value), even for sequences with high percentages of identity. Therefore, the detection of homologues for short proteins by using sequence similarity tends to fail.

The difficulty in the identification of homologues is only one of the problems associated with short proteins [23]. Let us consider newly sequenced genomes. The first task is identifying potential (putative) gene products. The main steps for identifying the encoded proteins include:

1. Sequence similarity: While this method is the most powerful computational approach, as indicated, it fails to detect short proteins.
2. Comparative genomics: This method requires the aligned genomes of related species. Additionally, this method is likely to fail to detect short proteins for similar reasons to the sequence similarity approach.
3. Ab initio gene prediction: The default parameters require a minimal length for potential ORFs (open reading frames), which may further hinder the detection of short proteins.
4. High coverage of the transcriptome and proteome by high-throughput experimental technologies.

Some experimental methods focus on detection of mRNA expression and others on detection of protein expression [24]. High-throughput experiments for the detection of evidence for mRNA expression are perhaps the best source of data for detecting new proteins but are often far from comprehensive due to the fact that many genes are only expressed under certain conditions and technical reasons [25]. Deep sequencing technologies (e.g., RNA-Seq of the transcriptome) can overcome the problem of low coverage. However, a flood of short noncoding and fragmented transcripts are also detected with the potential to mask the transcripts of short proteins. The most direct approach is mass spectrometry (MS) proteomics. Nevertheless, this high-throughput protein expression technology requires special tweaking in order to detect short proteins and is limited to the detection of highly expressed proteins. Furthermore, if a short protein is not already a known candidate, it will not be found [26]. As a consequence of these computational and experimental difficulties, it is conceivable that short proteins represent a relatively understudied and neglected niche [23].

8.5 Short Proteins: Why Do We Care?

Considering the identification of short proteins as a possibly underrepresented group, one might pause to ask how many short proteins are there and what kinds of functions can be associated with them. Examination of the SwissProt database [27] shows that 2 % of the registered sequences (excluding fragments) are less than 50 amino acids in length and 10 % are shorter than 100 amino acids. Clearly, the cellular machinery is capable of producing many functioning small proteins, and these proteins are involved in biological activities.

What biological functions do these proteins fulfill? To answer this question, we performed a simple statistical enrichment test for biological functions on the group of all SwissProt proteins shorter than 100 amino acids [28]. The enrichment test is aimed at finding biological groups that appear significantly more often than expected for a random selection of proteins. Several biological groups and functions are highly overrepresented among short proteins. The significance of being short for neuropeptides (NPs) is evident; of 1430 annotated proteins, 1170 were shorter than 100 amino acids [29]. Other enriched keywords include "signal peptide," "secretion," and more. The functional groups that had the highest statistical enrichment value include "toxin," "neurotoxin," "ion channel inhibitor" (ICI), "sodium channel inhibitor," and "scorpion long-chain toxin." For example, from the 2080 proteins that are annotated as being ICIs, 1900 are less than 100 amino acids long. This group includes most animal toxins.

8.5.1 The ID of Animal Toxins

Toxins are animal venom proteins aimed at inflicting harm to the organism on which the venom acts. They are extremely varied in terms of function and effect and include ion channel inhibitors (ICIs), phospholipases, protease inhibitors, disintegrins, membrane pore inducers, and more [30].

ICIs constitute the most widely studied group of toxins. Even specific groups of ICIs which inhibit the same channel type are varied in sequence and structural folds [31]. One group of ICIs whose evolution has been previously studied is the potassium ion channel inhibitors (K^+ ICIs). K^+ ICIs are found in a wide variety of venomous species and possess at least ten different structural folds [32]. In spite of this, all K^+ ICIs possess two residues that are critical for function, a Lys and a Tyr or Phe, which are known as the functional dyad [33]. Surprisingly, even though these residues appear in very different positions in the sequences of K^+ ICIs, the solved structures show they are closely aligned in space relative to each other.

On the other hand, some scorpion toxins, while sharing the same structural fold, act to inhibit different ion channels, including Ca^{2+}, K^+, Na^+, and Cl^-. This surprising observation shows that although there are many toxin folds, none is definitively associated with any particular ion channel selectivity [31]. This raises interesting questions regarding evolutionary convergence, divergence, and functional conservation.

8.5.2 TOLIPs: Endogenous Toxin-Like Proteins

Toxins appear only in very specific branches of the evolutionary tree. However, these branches are widely dispersed, including insects, snakes, sea anemones, spiders, the marine cone snail, and even mammals. Still, many toxins possessing similar functions (e.g., ICIs) appear in several unrelated venomous species. The

possibility to detect endogenous toxin-like proteins (called TOLIPs) is attractive for a number of reasons. For example, maintaining their function as ion channel blockers provides a new layer of regulation at the protein activity level (i.e., blocking ion flux through the channel). Additionally, an extensive search through the literature shows that toxin-like proteins exist in multiple species and are expressed in a wide range of nonvenomous organisms and tissues. Two striking examples are LYNX1-Ly6 [34] and SLURP-1 [35] from mammals. These are human proteins that not only possess similarity to snake α-neurotoxins but also modulate nicotinic acetylcholine receptors (nAChR) as do α-neurotoxins. The identification of SLURP-1 as an epidermal neuromodulator has helped explain the phenotype of the Mal de Meleda disease, a skin disease that results from improper activation of TNF-α [35]. A crucial question that arises in the field of insect genomics and proteomics is how many of the potential TOLIPs have actually been discovered.

8.5.3 Neuropeptides: Master Regulators of Insect Life

NPs regulate most aspects of insect life, from growth to behavior. The effect of NPs can only be fully appreciated by taking in the complementary view of their receptors and the underlying signaling network [36]. Due to the central role of insect NPs in mating behavior, growth, and reproduction, they are attractive targets for management of pests in agriculture. In our study, we addressed NPs as short neuromodulatory peptides that possess fundamental physiological roles [29, 37].

NPs are key modulators in behavior, sensation, and homeostasis [38]. Similar to endogenous TOLIPs, these peptides function in biological communication for a wide range of metazoans, from cnidarians to bilaterians, including mammals. The NPs are very short active peptides (5–30 amino acids) produced from parts of longer precursor molecules that are subjected to multiple cleavages. The posttranslational end products are subsequently modified and secreted. It is estimated that there are tens of NP precursor (NPP) genes in *Drosophila* and the honeybee. This rough estimation is based on annotations derived from only a small number of model organisms [39]. Similar to TOLIPs, the NP sequences are mostly non-alignable. Sequence similarity methods fail to predict or provide a comprehensive catalogue of NP bioactive peptides or their precursors. Active NPs in insects are diverse in terms of the site of action, the pattern of modification, and the specificity [40]. For example, the same NP may act both in the central nervous system (CNS) and as a hormone in the hemolymph, leading to different physiological responses.

8.5.4 From Features to Predictors

The goal of this section is to present a systematic approach for identifying insect TOLIPs as well as candidate NPs. We provide the analysis for a large number of insect proteomes that are archived in insect genomic resources and in central

resources such as the UniProtKB protein database. Many NPs play roles in regulating the behavior and physiology of larger animals, notably in terms of metabolism, pain regulation, and social behavior. Generating a catalogue of the proteome's short bioactive peptides (i.e., functional peptidome) will benefit the biotechnological community that seeks new directions for pest management and manipulating insect behavior in general.

We set out to construct two machine learning classifiers that are trained on: (1) animal ICI toxins and (2) NPs and NPP genes. For both types of short protein active modulators, characterization by sequence alignment-based methods is ineffective. Hence, the main logic in our approach is to identify the features that capture the characteristics of the types of proteins we seek to identify.

The scheme we present is composed of three main parts: (1) data analysis from genomes to short proteins, (2) the (supervised) machine learning approach and prediction, and (3) the annotation and functional validation phase.

The workflow for data mining and prediction using machine learning is composed of several steps: (1) acquiring the appropriate data in the form of protein sequences of the desired class and selecting "negative" sequences; (2) extracting features derived from the selected sequences; (3) constructing the training sets, training the classifier(s) according to the data and the generated features, and validating the predictions; (4) testing the classifier and comprehending its predictions; and (5) applying the classifiers to new proteins and selecting top predictions for further validation.

The result of all this is the discovery and subsequent annotation of new TOLIPs and NPPs (Fig. 8.3). While we discuss here the application in the context of insects, the protocol is applicable to any genome or proteome.

We will illustrate the protocol for the case of the animal toxin classifier. The ICIs share a general characteristic that can be described as structural stability (in short, toxin-like stability, TLS). Importantly, in most instances, the TLS is governed by the presence of disulfide bridges that are formed from cysteine residues along the sequence in question. The apparent rigidity of the scaffold of the proteins, together with posttranslational modifications (e.g., glycosylation), imposes rigid structural constraints.

One of the keys for a successful prediction when using machine learning is the selection of those features that may best characterize the targets we wish to predict (namely, to best separate TOLIPs from non-TOLIPs). The choice of features here was guided by the notion of stability, which is known to be associated with a large number of disulfide bridges. Therefore, the features were constructed so that they could reflect cysteine-mediated stability by encoding properties such as the frequency and the spread of cysteine residues within the sequence. In the predictor, called ClanTox (classifier of animal toxins), we used 545 features, many of which captured the TLS [41]. However, these were not restricted to cysteine-related features and were applied to all amino acids and other structural and sequence-derived qualities. The method used here represents each sequence as a vector that contains various numerical sequence features.

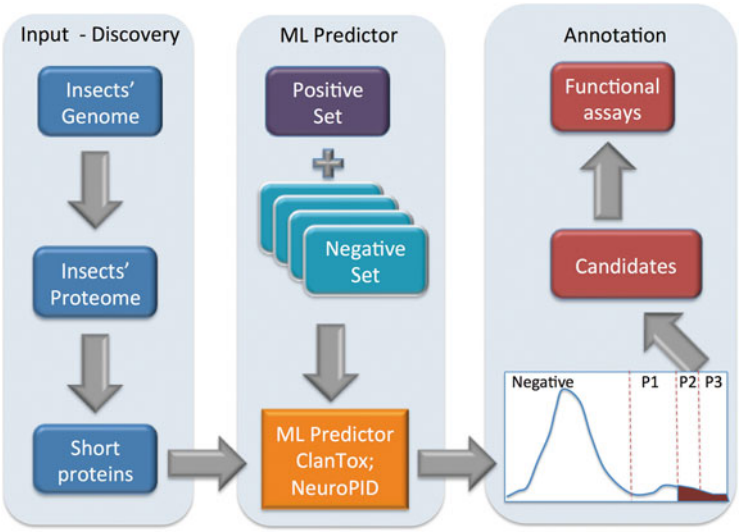

Fig. 8.3 The flow from an unannotated insect genome to the discovery of functional bioactive peptides. The scheme shows the training of the machine learning tools ClanTox and NeuroPID. These are two prediction platforms for the discovery of TOLIPs and NPs + NPPs

The general properties we examined that applied to all proteins included amino acid frequencies (20 features), amino acid pair frequencies (400 features), and sequence length (1 feature), among others. These features were shared between the two predictors (for TOLIPs and for NPPs). We will not discuss the selection and removal of redundancy from the training set, compiling alternative negative sets, tuning of the machine learning model parameters, or the cross-validation protocol.

UniProtKB and specifically the SwissProt database remain reliable sources of annotated sequences of complete proteomes and also for collecting information on toxins [42]. Eight thousand seven hundred insect proteins from UniProtKB were used as a background for testing the ClanTox predictor [41]. The results of testing short insect proteins (length <120 amino acids) from the SwissProt database (used as the input) are summarized in Table 8.1.

Using only the SwissProt database, we identified ~270 proteins ranked as putative TOLIPs, ~60 of them at a high level of predictive confidence. For example, the prediction from *Bombyx mori* includes a large number of bombyxins (types B, C, D, and G), chorion class high-cysteine proteins, fungal-chymotrypsin-trypsin inhibitors, and eclosion hormone. While each of these proteins is activated under different stimuli and at a specific developmental stage, there is only a limited set of functions that are enriched among the top-scoring predictions (Table 8.1). These functions include signaling of the innate immune system, serine protease inhibitors, and

Table 8.1 Statistically enriched keywords among the TOLIP predictions from SwissProt short proteins from insects

SwissProt keywords	Enrichment (Bonferroni)[a]
Disulfide bond	1.0E-55
Defensin	8.0E-16
Secreted	4.9E-08
Ion channel inhibitor	1.1E-07
Signal	1.1E-06
Cleavage on pair of basic residues	1.3E-06
Neurotoxin	2.4E-05
Protease inhibitor	9.0E-05
Serine protease inhibitor	9.0E-05
Toxin	6.7E-04
Knottin	7.7E-04
Hormone	3.2E-03
Zinc	1.0E-02
Antimicrobial	1.8E-02
Fungicide	2.1E-02
Calcium channel inhibitor	3.7E-02
Metal binding	3.9E-02

[a]Enrichment of keywords (p-value <0.05, Bonferroni correction), with all insect sequences of length <120 amino acids as the background set

antimicrobial functions. Considering that the training process was performed only on ICIs, it is remarkable to note that high-confidence TOLIPs share modulatory and signaling functions in development (e.g., bombyxins), the immune system (e.g., defensins, antimicrobial), and modulating tissues (e.g., protease inhibitors).

8.6 Test Case: TOLIPs in the Curated *D. melanogaster* Genome

From a set of thousands of sequences, we seek the predictor to announce for each sequence whether or not it is a toxin (or TOLIP). Discovery of overlooked TOLIPs calls for validating the top predictions (Fig. 8.3).

The most studied insect, *Drosophila melanogaster*, serves as a "testing ground." We applied the prediction platform to the complete proteome (almost 20,000 annotated genes in UniProtKB). One hundred sixty-one proteins were predicted to be TOLIPs by the ClanTox platform, with half of them ranked at the top confidence scale. Despite the high level of curation for *D. melanogaster*, about 60 % of the predictions were uncharacterized. However, most of these TOLIPs carry the signature of protease inhibitors (e.g., Kazal domain). Apparently, some TOLIPs with

Kazal domains have antibacterial and antifungal activities [43]. The rest are proteins belonging to drosomycins, sperm and seminal fluid, and metallothioneins [44].

The drosomycin family (seven sequences positively predicted) demonstrates the diversity of TOLIPs. Drosomycins are short, secreted proteins that possess antifungal activity. Similar to classical toxin ICIs, after removal of the signal peptide (SP), the mature peptides circulate in the hemolymph [45]. There, as part of the innate immune system, they act as ligands to alter intracellular signaling pathways.

8.7 Overlooked TOLIPs in Honeybee

The honeybee (*Apis mellifera*), the first sequenced genome of a venomous insect, was used for testing such a discovery phase, as its annotation level is only partial.

Among the 66 positive predictions, 26 suggest a higher level of confidence for being TOLIPs. Among them 73 % are named "uncharacterized." We observed that almost all possess a signal peptide (cleavable 20–25 amino acids at the N-terminal) [46]. About 50 % of the predicted proteins share a trypsin inhibitor-like (TIL) domain. The predictions with the highest scores are listed in Table 8.2. Structural representatives are shown in Fig. 8.4.

Once a top candidate TOLIP is detected, traditional state-of-the-art (albeit limited) sequence similarity methods can be activated. As illustrated for the sequence of H9KQJ7, applying sensitive tools for detecting remote homologues revealed a rich and a surprising resemblance to ω-conotoxins and to a set of related sequences. The multiple sequence alignment (Fig. 8.5) shows conservation of the several

Table 8.2 A sample of the top predictions of TOLIPs from *Apis mellifera*

Entry	Protein names	Len[a]	SP[b]	Family/PDB model	New discovery
P56587	Tertiapin (TPN)	21	Sec	PDB: 1TER	Tertiapin toxin like
H9K243	Uncharacterized	29	Frag	PDB: XU1	TNF and conotoxin
P01500	Apamin	46	Yes	PDB: 1TER	Tertiapin toxin like
B7UUK0	Apamin protein	46	Yes	PDB: 1TER	Tertiapin toxin like
H9KCD7	Uncharacterized	47	Yes	IPR: Zn-Fg	Zinc finger
P01499	Degranulating	50	Yes	PDB: 1TER	Tertiapin toxin like
H9K853	Uncharacterized	50	Yes	PDB: 1TER	Tertiapin toxin like
H9K3H8	Uncharacterized	58	Yes	PDB:1HP8	p8MTCP1 oncogene
P83563	Allergen Api m 6	71	Sec	PDB:1CCV	Trypsin inhibitor like
H9KEA0	Uncharacterized	71	Yes	TIL	Trypsin inhibitor like
H9KQJ7	Uncharacterized	74	Yes	PDB: 2WH9	ICI, OCLP, ω-conotoxin
Q27SJ8	Allergen Api m 6-1	92	Yes	TIL	Trypsin inhibitor like

Sec secreted, *Frag* fragment
[a]*Len* length in amino acids
[b]*SP* signal peptide

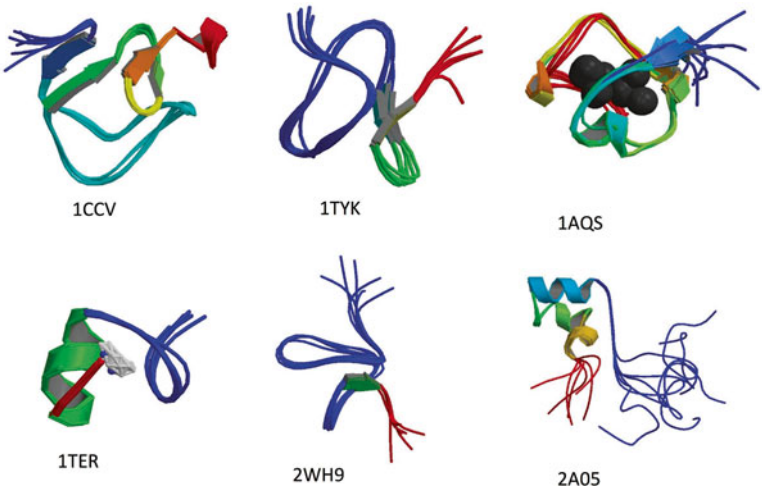

Fig. 8.4 Representatives of the top predictions for TOLIPs. The PDB accession ID is shown. The 3D structures were determined by NMR. *Connecting lines* indicate disulfide bonds; copper ions are shown as *black balls*. Sequences are colored (*rainbow*) from N- to C-terminals. For details, see text and Table 8.2

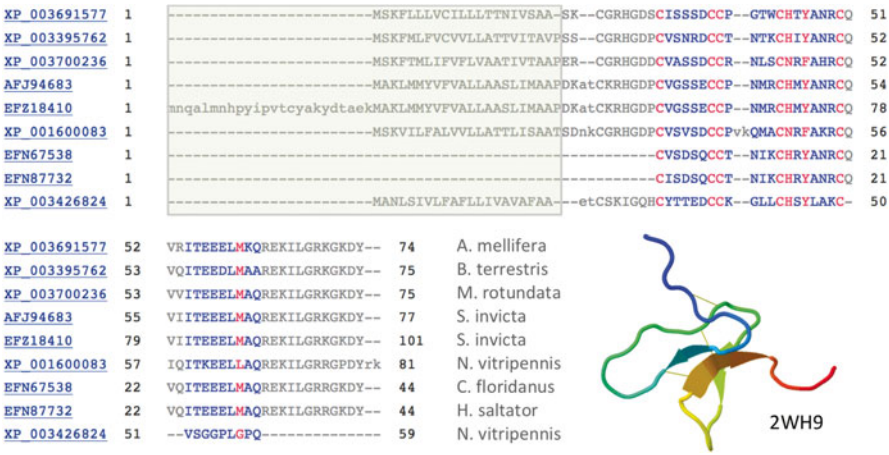

Fig. 8.5 Multiple sequence alignment of H9KQJ7 from *Apis mellifera*. The most similar sequences are from Hymenoptera including several bees, a wasp, and a number of ants. The most conserved amino acids are shown in *red*. These sequences are best modeled to PDB: 2WH9. This 3D prototype is a neurotoxin which was isolated from *Plesiophrictus guangxiensis* (tarantula) venom. The active peptide inhibits the Kv2.1 channel in human pancreatic β-cells

cysteines, which are critical in terms of structure. In addition, some positions are also conserved. Most notable is the Phe/Tyr in position 46 (F/Y numbered by the full-length H9KQJ7 sequence). It is known that the aromatic F/Y followed by N/A/R comprises the key amino acids for binding specificity to several ion channels

[47]. From the structural perspective, H9KQJ7 resembles a classical ICI from tarantula (PDB: 2WH9) as well as a large collection of structurally solved ICIs including omega-conotoxin, jingzhaotoxin, hanatoxin, huwentoxin-I, hainantoxin-I, and heteropodatoxin. Interestingly, a reversible effect of the honeybee H9KQJ7 protein on Ca^{2+} channel activity has been confirmed experimentally [48].

Several cDNAs provide supporting evidence for the expression pattern of such ω-conotoxin-like proteins (called OCLP, omega-conotoxin-like protein). OCLP-related cDNAs are found in *Anopheles gambiae*, *A. funestus*, *Aedes aegypti*, *D. melanogaster*, *Manduca sexta*, and *Heliconius erato*. None of these sequences had been previously characterized as ICI.

8.8 Evolutionary Diversity of ω-Conotoxin-Like Proteins in Insects

OCLP from honeybee has strong support for being a TOLIP: (1) It possesses a signal peptide; (2) it shares sequence similarity with assassin bug voltage-gated Ca^{2+} ICIs; (3) and structural modeling assigned the sequence to ω-conotoxin and related toxins with very high confidence (see Fig. 8.5). The expression of OCLP is exclusive to the brain (Linial and Bloch, unpublished).

A remote homologue search identified proteins in *A. gambiae* and *Ae. aegypti* containing multiple units of the OCL (omega-conotoxin-like) domain. Such organization is actually the hallmark of neuropeptides but was also noted for toxins (e.g., sarafotoxin; [49]). Remarkably, other toxins and functional motifs share the core of the OCL motif, specifically, covalitoxin II from tarantula and POI (phenol oxidase inhibitor) from *Musca domestica* [48]. The OCLP in honeybee is similar to these toxins but also to Ptu1 and ADO1, two related toxins from the assassin bugs *Peirates turpis* and *Agriosphodrus dohrni*, respectively. The function of Ptu1 as an effective Ca^{2+} channel blocker has been confirmed [50]. The OCL domain is conserved also in the freshwater planarian (*Schmidtea mediterranea)* sequence [48]. Focusing on the expansion of OCL domain in insects reveals duplication events in distinct branches along the insects' phylogeny (Fig. 8.6). There are nine OCL domains from *Nasonia vitripennis* that appear in five proteins. The repeated nature of OCL domains occurs also in proteins from *A. gambiae* and *Ae. Aegypti* (Fig. 8.6).

The short proteins discussed here raise the question of the evolutionary origin of proteins that share the OCL domain. The evolutionary relatedness that combines insects, the cone snail, and the flatworm planaria strongly argues for the importance of this fold in a diverse ecological setting. The possibility of de novo evolution for short proteins has been presented [51] and was supported by tracing the recent expansion of short immune-related proteins in mammals. Although evolutionary divergence is the most plausible explanation, convergent evolution for toxin-like proteins cannot be excluded.

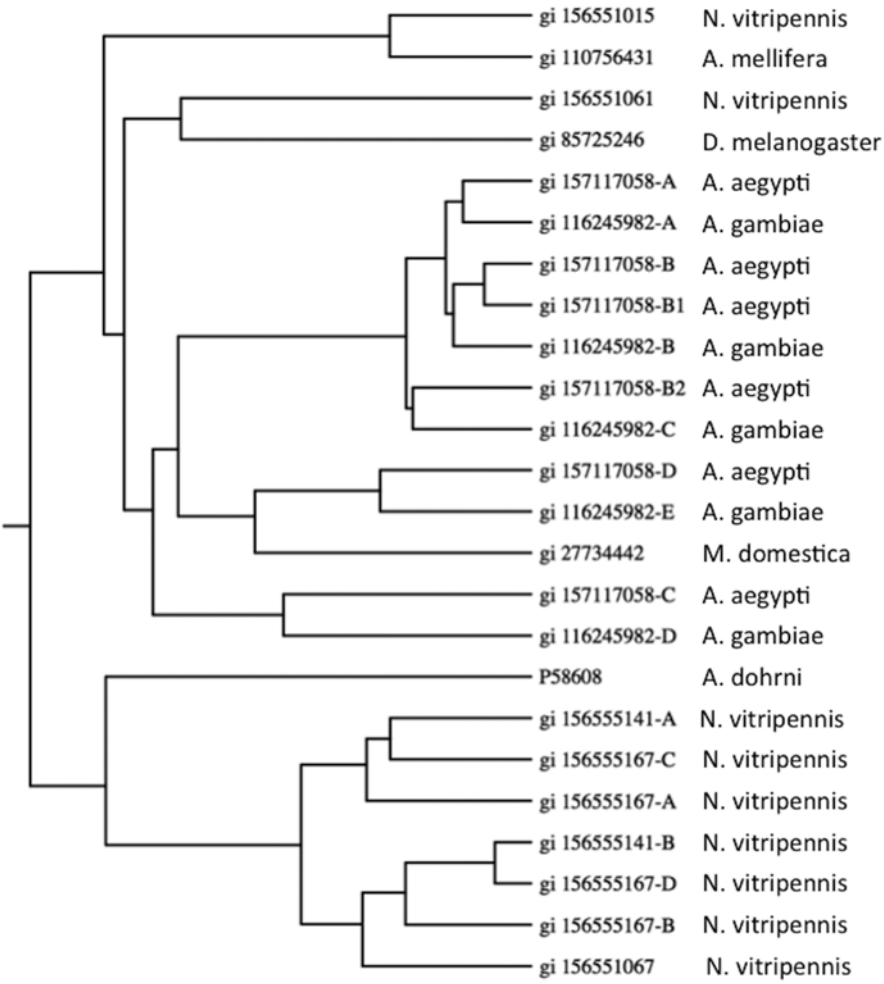

Fig. 8.6 Homology distance tree of insect proteins that contain the OCL domain. OCLP from the honeybee and a collection of other insects are shown. The OCL domains share an identical structure to ω-conotoxins with three cysteine bridges that govern the stable and compact structure of the OCL domains. The protein identifier is based on NCBI protein database

8.9 Overlooked TOLIPs in Fully Sequenced Insect Genomes

Application of the same protocol applied for the honeybee (Table 8.2) to all other insects whose genomes have been completed reveals that hundreds of overlooked TOLIPs can be traced across the entire phylogenetic tree. It is important to note that the sequences of the honeybee were not included in the training set for prediction of honeybee proteins. The same is true for the other recently sequenced genomes which were practically unavailable when the ClanTox predictor was trained. The

high-confidence predictions for these genomes reach a total of 235 (from 790 positive predictions). The most dominant functions associated with ClanTox predictions are the trypsin/chymotrypsin inhibitors, metallothioneins, TGF like, growth factor domains, defensin like, ICIs, and membrane-disrupting peptides.

Acyrthosiphon pisum (pea aphid) is represented by almost 40,000 protein sequences. The many short proteins (~9500) reflect the high number of fragmented sequences. When all the short proteins were tested using the ClanTox platform, only 18 sequences were predicted at a high level of confidence; another 112 sequences were predicted with only moderate confidence. One short sequence (J9KHE3, 63 amino acids) is a secreted protein that resembles a cysteine-rich secretory protein domain of Tpx-1 which is related to ion channel toxins and regulates ryanodine receptor Ca^{2+} signaling (PDB: 2A05, Fig. 8.4). The rest of the proteins from pea aphid are unlikely to act as secreted cell modulators.

The complete genome of the silk moth *B. mori* provides a glimpse of a different branch of the insect tree. There, several neurohormone proteins, including the insulin-like bombyxins, were positively predicted as being TOLIPs. The main functions of bombyxins are as growth factors for wing imaginal disks [52] and for general promotion and regulation of growth and metabolism [53]. The *B. mori* protein H9JHN8 is another example of a secreted protein which is uncharacterized and captures the characteristics of an overlooked TOLIP. A structural view identified numerous proteins that resemble the following functions: (1) spaetzle protein from *Drosophila*, which acts in development and in the immune system; (2) a protein from horseshoe crab involved in hemostasis and host defense; and (3) classical neurotrophins including β-nerve growth factor, brain-derived neurotrophic factor, and neurotrophin 3/4.

A surprisingly high number of positively predicted TOLIPs were associated with *A. gambiae*. There were 51 high-confidence predicted TOLIPs, many characterized by a trypsin inhibitor-like (TIL) domain (Table 8.2). Several other domains were also detected including EGF like, WAP (whey acidic protein), and elafin. A representative of the elafin family is a short secreted protein (Q7Q332) that resembles a large number of snake toxin proteins. The similarity to the 3D structure of nawaprin (PDB: 1UDK) from the venom of the spitting cobra, *Naja nigricollis*, is striking. The nonconventional circular structure is stabilized by the presence of four disulfide bonds. Interestingly, the nawaprin and elafin proteins (represented by the human leukocyte elastase-specific inhibitor) share several unique structural features but minimal sequence similarity. The functions of nawaprin or Q7Q332 from the mosquito are still not known.

Rhodnius prolixus is the most important vector of the Chagas parasite in Africa [54]. The complete genome was determined, but it is poorly annotated. Overlooked TOLIP detection using ClanTox identified T1H9H6. The sequence resembles a three-fingered fold that is abundant in snake venoms including cardiotoxins, denmotoxin, and α-bungarotoxin. These sequences belong to the diverse family named Upar/Ly6 [55]; many of the proteins are membrane markers of cells that belong to the innate immune system. The protein resemblance to snake neurotoxin (e.g., cobra) has been reported [56].

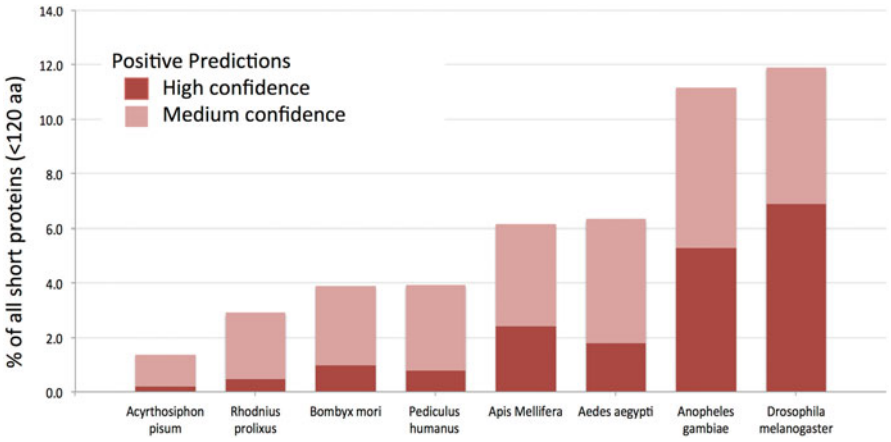

Fig. 8.7 A histogram showing the fraction (in %) of positive predictions with respect to short proteins in the indicated genomes. Several insect representatives from complete proteomes are listed. The number of high-confidence predictions ranges from 9 to 66 for *Pediculus humanus* and *Drosophila melanogaster*, respectively

In sum, the fraction of predicted TOLIPs among the short proteins varies drastically among insect genomes (Fig. 8.7). We attribute these large differences to the quality of the genome assembly. From a biological perspective, it may reflect varying complexities in the modulation of cell communication, the immune system, and/or neuronal functions.

The discovery of TOLIPs in insects led to an unexpected finding that showed the abundance of TOLIPs in viruses [57]. A cross talk of insects and their viruses was proposed. For example, protein B6S6X8 (113 aa) from *Betabaculovirus* is similar to many of the short peptides in *Drosophila* proteomes [57]. In another instance, a cysteine-rich encoding region was transferred from the endoparasitic wasp *Campoletis sonorensis* to a symbiotic polydnavirus (CsPDV) [58].

8.10 Neuropeptide Precursors in Insects

The ideas that exemplified TOLIPs and the methods used were successfully applied to identify neuropeptide precursor (NPP) genes. Neuropeptides are the products of a posttranslational regulated process of cleavage and modification from NPPs. The mature peptides are secreted from neurons and thus are collectively called neuropeptides (NPs). NPs act through their direct interaction with their receptors on presynaptic or postsynaptic cells [59]. In insects, NPs function in cell communication and affect social behaviors, including mating, food uptake, and metabolism [60].

Insects have evolved a large repertoire of NPs. Figure 8.8 (left) shows the number of annotated NPs from major taxonomical groups. We considered only the data

associated with "complete proteomes" (defined by UniProtKB). The proportion of NPs from all annotated protein sequences is maximal in insects when compared to worms and mammals (Fig. 8.8, right).

8.10.1 A Neuropeptide Precursor Prototype

Studying the molecular processing of NPs is essential for designing a collection of relevant features extracted from sequences. These features should capture the essence of discriminating properties of the "true" *vs.* "false" sets (see Sect. 8.3). The predictor can be used to classify new instances of unknown sequences.

A prototypical example of NPs is the allatostatin family from *A. mellifera* [61]. Approximately 500 neurons in the honeybee brain produce allatostatins [62] which act to inhibit juvenile hormone biosynthesis and reduce food intake. The precursor protein (UniProtKB: P85797, 197 amino acids) produces after cleavage ten active NPs which were identified by MS experiments [61]. The allatostatin NPP of the Pacific beetle cockroach (*Diploptera punctata*) contains 13 identified NPs (Fig. 8.9).

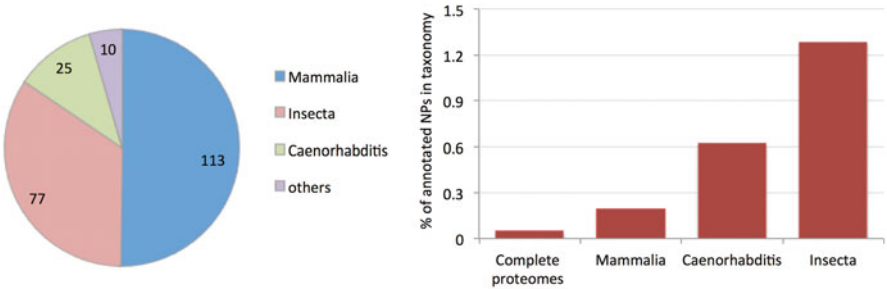

Fig. 8.8 Partition of annotated neuropeptides from major taxonomical groups. Number of annotated neuropeptides (NPs, *left*). Fraction (%) of SwissProt keyword "neuropeptide" from the sequences of "complete proteomes" (*right*). The fraction of NPs from insects is 6.7-fold relative to the fraction of NPs in mammals

```
66-  97   KRL....YDFG.........LG.............KRA..YsyvSEYKRL........pvYN..FGLGKR
98- 120   SKM....YGFG.........LG.............KR.......DG..RM..........YS..FGLGKR
121-164   DYD....Y.YGeededdqqaIGdedieesdvgdlmdKR..........DRL..........YS..FGLGKR
165-191   ARP....YSFG.........LG.............KRA..P...SGAQRL..........YG..FGLGKR
192-220   GGS...1YSFG.........LG.............KR.........GDGRL..........YA..FGLGKRPVNS
221-253   GRSsgsrFNFG.........LG.............KRS..D...DIDFRE..........LEekFAEDKR
254-316    YPqehrFSFG.........LG.............KREveP...SELEAVrne(25)slhYP..FGIRKL
346-367   RRP....FNFG.........LG.............KRI..P.........M..........YD..FGIGKR
```

Fig. 8.9 Sequence of neuropeptide precursor P12764 (ALLS_DIPPU, 370 aa) from *Diploptera punctata* (Pacific beetle cockroach). The repeated nature of the sequence is shown. The repeated segments account for 13 bioactive NPs, called allatostatin-1 to allatostatin-13. The NPs are consecutively colored *red* and *blue*. The dibasic residues (cleavage sites) are highlighted in *yellow*

Though each peptide has a unique sequence, all share the Tyr/Phe-Xaa-Phe-Gly-Leu/Ile-NH2 consensus sequence. Furthermore, each active NP is amidated on the terminal Leu/Ile. Evidently, the above properties cannot be captured by methods for remote homology detection that are based on sequence alignments.

8.10.2 Neuropeptide Precursors: Feature Extraction

Similar to the arguments raised for identification of TOLIPs (Sects. 8.6, 8.7, 8.8, and 8.9), Fig. 8.9 illustrates the difficulty in using sequence alignment for identifying NPPs. We thus set out to train a predictor using supervised machine learning. To this end, we compiled nonredundant "positive" and "negative" sets. The nonredundant "positives" included all annotated sequences from SwissProt as well as the automatically inferred sequences from UniProtKB. The "negative" sets included sequences that are basic in nature (i.e., enriched with basic residues) as well as randomly selected proteins from Metazoans with an identical length distribution.

A characteristic "feature" for the majority of NPs is their production from larger precursor proteins (NPP) [63]. In most cases, NPPs produce different NPs that may participate in executing a behavior [64]. A dominant feature is the presence of clusters of dibasic residues that specify these proteolytic cleavage sites. Nevertheless, some NPPs do not use dibasic residues as a cleavage signal.

The goal of extracted features from the training sets is to capture the particular traits and variance between the "positive" and "negative" sets. The information collected to construct a predictor covers:

(A) Biophysical quantitative properties [65] including: (1) the length and molecular weight, (2) frequency of the amino acids or their grouping (e.g., charged amino acids) and dipeptide frequencies (400 features), and (3) quantitative indices, such as aromaticity, instability index, hydropathy, and PI (i.e., isoelectric point).
(B) Binary features that capture the nonrandomized appearance of certain amino acids in short, overlapping windows. This features grouping stems from the occurrence of certain residues near known cleavage sites such as G-KR (Gly, Lys, Arg), lack of flanking proline at cleavage sites, and structural considerations (e.g., disordered, accessible regions).
(C) Appearance and frequency of known sequence motifs. The most important motifs stem from conservation by the processing endopeptidases, such as flanking pairs of basic residues [66]. In addition, we considered potential amidation, hydroxylation, and N-glycosylation sites.
(D) Information-based statistics. The intuition is to trace the entropy, the autocorrelation of the potential cleavage sites, and the repeated nature of the sequences (see an example in Fig. 8.9).

8.10.3 Prediction of Insect Neuropeptide Precursors

Testing the performance of the machine learning approach for identifying known and novel NPPs was carried out using a cross-validation (CV) protocol. Accordingly, a substantial fraction of the data (i.e., 10–40 %) was removed and excluded during the training phase and used as a test set. The protocol is repeated multiple times using a different subset of the data each time. The results of the CV tests for each of the NPP candidates were summed up to estimate the accuracy, sensitivity, precision, and AUC (area under ROC curve). The accuracy rate for NPP identification reached a level of 82–89 % for insect NPPs. The class of random forest [17] ensemble decision tree method performed best. Slightly lower performance was recorded for gradient boosting decision trees and linear and nonlinear SVM models [29].

By increasing the thresholds of the prediction tools, we filtered the number of NPP candidates to a few tens. For example, the random forest protocol at a "certainty" threshold of 0.99 reduced the predictions for *B. mori* from ~4000 to only 16. All NPPs are secreted proteins, and thus each has a signal peptide sequence in the N-terminal which is removed prior to the production of the precursor protein. This is a strong feature that was used to remove many of the false positives of the prediction machine.

8.10.4 Identifying Candidate NPPs in Insect Proteomes

Two hundred ninety-seven proteins (total of 20,600 protein sequences) in *A. pisum* include a signal peptide (SP) and are thus candidates for being NPPs. A test of our NPP machine learning platform (NeuroPID) yielded about one-third as potential candidates and 13 as high-probability NPPs. Experimental information on these sequences is lacking. The ETH (ecdysis-triggering hormone) precursor was identified among these poorly characterized predicted proteins (Table 8.3). In most insect species, the ETH precursor produces two active peptides [67]. ETH genes and their receptors have also been identified in tick (Arachnida) and water flea (Crustacea). Notably, several of these sequences, while marked as uncharacterized, are highly expressed (Table 8.3). The task of validating these as NPPs calls for functional experimentation and independent evidence (e.g., using MS).

8.11 Insect Short Active Peptides for Human Health and Agriculture

The efficiency and quality of experimentally validated proteins lag behind the explosive growth in sequencing. We expect that the analysis presented in this study will be useful for leveraging the expansion of protein space. In this chapter, we

Table 8.3 Top predictions of neuropeptide precursors from *Acyrthosiphon pisum*

Protein name[a]	Function/expression	Domains[b]
Chemosensory protein-like prec	Pheromone-BP	
Ecdysis-triggering hormone prepro	ETH novel NPP	
Miple protein prec	GF, heparin binding	PTN/MK, C-ter.
Mitochondrial TIM14-like prec	Chaperone -HSP70	DNAJ
Odorant-binding protein 7 prec	GPCR BP	
Odorant-binding protein 8 prec	GPCR BP	
UC protein LOC100159063 prec	Expression—high	
UC protein LOC100161501 prec	Expression—low	
UC protein LOC100162497 prec	Expression—high	TMEMB_9
UC protein LOC100166422 prec	Expression—low	
UC protein LOC100169149 prec	Highly conserved	
UC protein LOC100302326 prec	Expression—medium	
UC protein LOC100574827 prec		

[a]*Prepro* preprotein, *prec* precursor, *UC* uncharacterized, *GF* growth factor
[b]Domains are listed according to the Pfam abbreviations

introduced machine learning approaches for large-scale protein classification of short peptides that resemble animal toxins as well as NPPs and cell modulators.

The therapeutic potential of toxins has been realized and has led to the development of toxin-based drugs, with ICI toxins being the lead for such development [68]. Insect peptides that act in the innate immune system (e.g., defensins) and antibacterial proteins [69] are additional classes of potential drugs which can also be developed as pesticides. Below we outline some of the benefits and applications of these molecules for human health and agriculture.

There are several benefits for the pharmaceutical industry to focus on biological active peptides in general and on toxins and TOLIPs in particular. For example, the 3D high stability of the backbones makes them appealing for drug design, and several toxin-based drugs are already available on the market in synthetic form. A well-known example takes advantage of the mimicry of the MVIIA ω-conotoxin from the marine cone snail *Conus magus* which acts as a blocker of the voltage-gated Ca^{2+} channel [70]. The clinical application of this drug is for chronic, uncontrollable pain. Utilization of the classifiers described earlier in this chapter such as ClanTox and NeuroPID may expand the range of known insect modulators. ProFET (Protein Feature Engineering Toolkit) is another such framework for the machine learning approach to protein function, offering an easy to use, universal platform as well as state of the art results in classification of high-level functions [72].

Another benefit for drug design is that most toxins and the TOLIPs are resistant to proteolysis. This is not only a by-product of their structural compactness but also because many TOLIPs are actually protease inhibitors [56]. This property ensures stability in use as a drug, which is reflected in the protein half-life. Posttranslational modifications on most of these toxins provide an additional layer of stability in the cell and most importantly in the extracellular space.

Insects populate many ecological niches and some of them are considered pest species. Management of pests has significant economic implications. With the increase in environmental awareness, new insecticidal compounds must be explored. The coevolution of insects and plants for millions of years argues that hundreds of TOLIPs are attractive candidates for screening novel targets in plants and animals. Potentially, different folds of TOLIPs can be used in a rational design for as yet unknown targets. Through mimetic approaches, the scaffold of these short proteins can be reduced and directed toward protecting diverse hosts from pests by disturbing and damaging selected membranes of pathogens and even altering mating behavior to control the balance in the ecosystem.

Novel NPPs from insects can play a biotechnological lead in regulating social behavior, metabolic status, and communication. In this view, an exciting genomic initiative with the goal of sequencing 5000 arthropod genomes was recently announced [71]. The expectation is that prediction methods for short proteins will make a valuable contribution for identifying unexplored modes of insect communication, among other features. The power of our method increases with the increase in sequenced genomes, transcriptomes, and proteomes of related species. We expect that the analysis presented in this study will be useful in leveraging future expansions of protein space.

Acknowledgments We thank Noam Kaplan for his seminal contribution to the work. Many of the ideas and formulation of the problem resulted from many stimulating discussions. N.K.'s beautiful thesis was used as a starting point for this chapter. The development and maintenance of ClanTox, NeuroPID, and the protein database have been a long-term effort of many people over the last 10 years. We thank Solange Karsenty for her contribution to the development of ClanTox and NeuroPID platforms and the system group in the School of Computer Science and Engineering for long-term support of our web servers.

References

1. Loewenstein Y et al (2009) Protein function annotation by homology-based inference. Genome Biol 10:207
2. Sasson O, Kaplan N, Linial M (2006) Functional annotation prediction: all for one and one for all. Protein Sci 15:1557–1562
3. Altschul SF et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402
4. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12:85–94
5. Rost B, Sander C (1994) Conservation and prediction of solvent accessibility in protein families. Proteins 20:216–226
6. Shachar O, Linial M (2004) A robust method to detect structural and functional remote homologues. Proteins 57:531–538
7. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202
8. Radivojac P et al (2013) A large-scale evaluation of computational protein function prediction. Nat Methods 10:221–227

9. Schuldiner S, Shirvan A, Linial M (1995) Vesicular neurotransmitter transporters: from bacteria to humans. Physiol Rev 75:369–392
10. Biegert A, Soding J (2009) Sequence context-specific profiles for homology searching. Proc Natl Acad Sci U S A 106:3770–3775
11. Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. Genome Inform (International Conference on Genome Informatics) 23:205–211
12. Punta M et al (2012) The Pfam protein families database. Nucleic Acids Res 40:D290–D301
13. Portugaly E, Linial N, Linial M (2007) EVEREST: a collection of evolutionary conserved protein domains. Nucleic Acids Res 35:D241–D246
14. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol 5:e1000585
15. Rao HB, Zhu F, Yang GB, Li ZR, Chen YZ (2011) Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res 39:W385–W390
16. Chapelle O (2007) Training a support vector machine in the primal. Neural Comput 19:1155–1178
17. Breiman L (2001) Random forests. Mach Learn Cybern 45:5–32
18. Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ (2003) SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res 31:3692–3697
19. Leslie C, Eskin E, Noble WS (2002) The spectrum kernel: a string kernel for SVM protein classification. Pac Symp Biocomput 2002:564–575
20. Tarca AL, Carey VJ, Chen XW, Romero R, Draghici S (2007) Machine learning and its applications to biology. PLoS Comput Biol 3:e116
21. Ben-Hur A, Ong CS, Sonnenburg S, Scholkopf B, Ratsch G (2008) Support vector machines and kernels for computational biology. PLoS Comput Biol 4:e1000173
22. Rappoport N, Linial N, Linial M (2013) ProtoNet: charting the expanding universe of protein sequences. Nat Biotechnol 31:290–292
23. Frith MC et al (2006) The abundance of short proteins in the mammalian proteome. PLoS Genet 2:e52
24. Kondo T et al (2010) Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. Science 329:336–339
25. Ponting CP, Belgard TG (2010) Transcribed dark matter: meaning or myth? Hum Mol Genet 19:R162–R168
26. Lubec G, Afjehi-Sadat L (2007) Limitations and pitfalls in protein identification by mass spectrometry. Chem Rev 107:3568–3584
27. Wu CH (2006) Bioinformatics for proteomics at the Protein Information Resource (PIR). Mol Cell Proteomics 5:S341–S341
28. Rappoport N, Fromer M, Schweiger R, Linial M (2010) PANDORA: analysis of protein and peptide sets through the hierarchical integration of annotations. Nucleic Acids Res 38:W84–W89
29. Ofer D, Linial M (2013) NeuroPID: a predictor for identifying neuropeptide precursors from metazoan proteomes. Bioinformatics. 30(7):931–940.
30. Fry BG (2005) From genome to "venome": molecular origin and evolution of the snake venom proteome inferred from phylogenetic analysis of toxin sequences and related body proteins. Genome Res 15:403–420
31. Mouhat S, Jouirou B, Mosbah A, De Waard M, Sabatier JM (2004) Diversity of folds in animal toxins acting on ion channels. Biochem J 378:717–726
32. Norton RS, Pallaghy PK (1998) The cystine knot structure of ion channel toxins and related polypeptides. Toxicon 36:1573–1583
33. Terlau H, Olivera BM (2004) Conus venoms: a rich source of novel ion channel-targeted peptides. Physiol Rev 84:41–68
34. Ibanez-Tallon I et al (2002) Novel modulation of neuronal nicotinic acetylcholine receptors by association with the endogenous prototoxin lynx1. Neuron 33:893–903

35. Chimienti F et al (2003) Identification of SLURP-1 as an epidermal neuromodulator explains the clinical phenotype of Mal de Meleda. Hum Mol Genet 12:3017–3024
36. Schoofs L, Beets I (2013) Neuropeptides control life-phase transitions. Proc Natl Acad Sci U S A 110:7973–7974
37. Karsenty S, Rappoport N, Ofer D, Zair A, Linial M (2014) NeuroPID: a classifier of neuropeptide precursors. Nucleic Acids Res 42:W182–W186
38. Brain SD, Cox HM (2006) Neuropeptides and their receptors: innovative science providing novel therapeutic targets. Br J Pharmacol 147(Suppl 1):S202–S211
39. Nassel DR (2002) Neuropeptides in the nervous system of Drosophila and other insects: multiple roles as neuromodulators and neurohormones. Prog Neurobiol 68:1–84
40. Vanden Broeck J (2001) Neuropeptides and their precursors in the fruitfly, Drosophila melanogaster. Peptides 22:241–254
41. Naamati G, Askenazi M, Linial M (2009) ClanTox: a classifier of short animal toxins. Nucleic Acids Res 37:W363–W368
42. Dimmer EC et al (2012) The UniProt-GO Annotation database in 2011. Nucleic Acids Res 40:D565–D570
43. Kim BY et al (2013) Antimicrobial activity of a honeybee (*Apis cerana*) venom Kazal-type serine protease inhibitor. Toxicon 76:110–117
44. Palmiter RD (1998) The elusive function of metallothioneins. Proc Natl Acad Sci U S A 95:8428–8430
45. Tian C et al (2008) Gene expression, antiparasitic activity, and functional evolution of the drosomycin family. Mol Immunol 45:3909–3916
46. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8:785–786
47. Lipkind GM, Fozzard HA (1994) A structural model of the tetrodotoxin and saxitoxin binding site of the Na+channel. Biophys J 66:1–13
48. Kaplan N, Morpurgo N, Linial M (2007) Novel families of toxin-like peptides in insects and mammals: a computational approach. J Mol Biol 369:553–566
49. Kloog Y et al (1988) Sarafotoxin, a novel vasoconstrictor peptide: phosphoinositide hydrolysis in rat heart and brain. Science 242:268–270
50. Sousa SR, Vetter I, Lewis RJ (2013) Venom peptides as a rich source of cav2.2 channel blockers. Toxins 5:286–314
51. Su M, Ling Y, Yu J, Wu J, Xiao J (2013) Small proteins: untapped area of potential biological importance. Front Genet 4:286
52. Nijhout HF, Grunert LW (2010) The cellular and physiological mechanism of wing-body scaling in *Manduca sexta*. Science 330:1693–1695
53. Mizoguchi A et al (2013) Prothoracicotropic hormone acts as a neuroendocrine switch between pupal diapause and adult development. PLoS One 8:e60824
54. Schofield CJ, Jannin J, Salvatella R (2006) The future of Chagas disease control. Trends Parasitol 22:583–588
55. Lee PY, Wang JX, Parisini E, Dascher CC, Nigrovic PA (2013) Ly6 family proteins in neutrophil biology. J Leukoc Biol 94:585–594
56. Tirosh Y, Ofer D, Eliyahu T, Linial M (2013) Short toxin-like proteins attack the defense line of innate immunity. Toxins 5:1314–1331
57. Naamati G, Askenazi M, Linial M (2010) A predictor for toxin-like proteins exposes cell modulator candidates within viral genomes. Bioinformatics 26:i482–i488
58. Cui L, Webb BA (1996) Isolation and characterization of a member of the cysteine-rich gene family from Campoletis sonorensis polydnavirus. J Gen Virol 77(Pt 4):797–809
59. Jekely G (2013) Global view of the evolution and diversity of metazoan neuropeptide signaling. Proc Natl Acad Sci U S A 110:8702–8707
60. Insel TR, Young LJ (2000) Neuropeptides and the evolution of social behavior. Curr Opin Neurobiol 10:784–789
61. Hummon AB et al (2006) From the genome to the proteome: uncovering peptides in the Apis brain. Science 314:647–649

62. Kreissl S, Strasser C, Galizia CG (2010) Allatostatin immunoreactivity in the honeybee brain. J Comp Neurol 518:1391–1417
63. Mirabeau O et al (2007) Identification of novel peptide hormones in the human proteome by hidden Markov model screening. Genome Res 17:320–327
64. Mentlein R, Dahms P (1994) Endopeptidases 24.16 and 24.15 are responsible for the degradation of somatostatin, neurotensin, and other neuropeptides by cultivated rat cortical astrocytes. J Neurochem 62:27–36
65. Artimo P et al (2012) ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res 40:W597–W603
66. Southey BR, Sweedler JV, Rodriguez-Zas SL (2008) Prediction of neuropeptide cleavage sites in insects. Bioinformatics 24:815–825
67. Roller L et al (2010) Ecdysis triggering hormone signaling in arthropods. Peptides 31:429–441
68. Fox JW, Serrano SM (2007) Approaching the golden age of natural product pharmaceuticals from venom libraries: an overview of toxins and toxin-derivatives currently involved in therapeutic or diagnostic applications. Curr Pharm Des 13:2927–2934
69. Lai Y, Gallo RL (2009) AMPed up immunity: how antimicrobial peptides have multiple roles in immune defense. Trends Immunol 30:131–141
70. Brady RM, Baell JB, Norton RS (2013) Strategies for the development of conotoxins as new therapeutic leads. Mar Drugs 11:2293–2313
71. Consortium iK (2013) The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J Hered 104:595–600
72. Ofer, Dan, and Michal Linial. "ProFET: Feature engineering captures high-level protein functions." Bioinformatics (2015): btv345.