BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 7 2014, pages 931–940
doi:10.1093/bioinformatics/btt725

Sequence analysis

Advance Access publication December 13, 2013

NeuroPID: a predictor for identifying neuropeptide precursors from metazoan proteomes

Dan Ofer and Michal Linial*

Department of Biological Chemistry, Institute of Life Sciences, The Edmond J. Safra Campus, The Hebrew University of Jerusalem, Givat Ram 91904, Israel

Associate Editor: John Hancock

ABSTRACT

Motivation: The evolution of multicellular organisms is associated with increasing variability of molecules governing behavioral and physiological states. This is often achieved by neuropeptides (NPs) that are produced in neurons from a longer protein, named neuropeptide precursor (NPP). The maturation of NPs occurs through a sequence of proteolytic cleavages. The difficulty in identifying NPPs is a consequence of their diversity and the lack of applicable sequence similarity among the short functionally related NPs.

Results: Herein, we describe Neuropeptide Precursor Identifier (NeuroPID), a machine learning scheme that predicts metazoan NPPs. NeuroPID was trained on hundreds of identified NPPs from the UniProtKB database. Some 600 features were extracted from the primary sequences and processed using support vector machines (SVM) and ensemble decision tree classifiers. These features combined biophysical, chemical and informational–statistical properties of NPs and NPPs. Other features were guided by the defining characteristics of the dibasic cleavage sites motif. NeuroPID reached 89–94% accuracy and 90–93% precision in cross-validation blind tests against known NPPs (with an emphasis on Chordata and Arthropoda). NeuroPID also identified NPP-like proteins from extensively studied model organisms as well as from poorly annotated proteomes. We then focused on the most significant sets of features that contribute to the success of the classifiers. We propose that NPPs are attractive targets for investigating and modulating behavior, metabolism and homeostasis and that a rich repertoire of NPs remains to be identified.

Availability: NeuroPID source code is freely available at http://www.protonet.cs.huji.ac.il/neuropid

Contact: michall@cc.huji.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 16, 2013; revised on December 6, 2013; accepted on December 10, 2013

1 INTRODUCTION

Peptides are known to be key modulators in behavior, sensation and homeostasis (Brain and Cox, 2006). Biologically active peptides that are produced and secreted from neurons and act to modulate their function are collectively called peptide modulators or neuropeptides (NPs). NPs represent a widespread mode of communication that is found from Cnidarians to Bilaterians, including mammals. NP precursors (NPPs) are subjected to regulated cleavages that result in the production of functionally active NPs. The processing of NPs occurs along the secretory pathway (Gelman and Fricker, 2010). In most instances, NPs locally modulate presynaptic or postsynaptic cell activity (Southey *et al.*, 2008). From a functional perspective, known effects of NPs include stress control, pain perception, social behaviors and sleep–wake cycle (Nassel, 2002). NPs also regulate food uptake, maintaining appetite and body weight. The peripheral NPs that act outside the central nervous system regulate gastrointestinal and immunological functions (Insel and Young, 2000). For example, some NPs induce chemotactic response in attracting immature dendritic cells to the site of inflammation. In this context, substances P and K, two extensively studied NPs, induce the production and release of inflammatory cytokines from blood monocytes (Gonzalez-Rey *et al.*, 2007). 'Social' NPs, such as oxytocin and arginine vasopressin, regulate complex social cognition and behavior (including pair-bonding, social recognition and maternal behavior) (Insel and Young, 2000). The immense diversity among NPs contributes to the wide range of behavioral tasks that are carried out.

A common feature for the majority of NPs is their production from a larger precursor (NPP). The production of short bioactive peptides is a result of a series of cleavages and maturation events (Mirabeau *et al.*, 2007). NPP can produce multiple copies of different NPs (Mentlein and Dahms, 1994). Notably, a cluster of basic residues specifies these cleavage sites. The occurrence of dibasic residues specifies the canonical sites for intracellular endopeptidases such as Furin (Veenstra, 2000). However, some NPs that act in cell–cell communication (Funkelstein *et al.*, 2010) (e.g. cathepsin L) do not obey the dibasic residues specificity rule. The identity and regulation of the key peptidases that are responsible for the processing of the NPP remains an active research field.

NPs typically bind cell surface G-proteins coupled receptors (GPCRs) that initiate a signaling cascade. An evolutionary analysis of NPPs and their cognate GPCRs suggested a diversification that occurred in certain taxonomical branches (Jekely, 2013). The NPs and their receptors are attractive targets for drug development and translational medicine based on their role in feeding, sexual behavior and cellular homeostasis (Brain and Cox, 2006).

The goal of this research is to enable systematic identification of NPPs (and NPs) at a genome-wide scale. The difficulty in identifying NPs and classifying genes as potential NPs stems from the following: (i) Current gene annotation tools mostly

*To whom correspondence should be addressed.

rely on sequence conservation traits (Loewenstein *et al*., 2009). However, NPs that exhibit the same function may share minimal sequence similarity (Clynen *et al*., 2010). Additionally, homologous NPPs may still produce NPs that are species-specific. (ii) Structural inference tools (Lobley *et al*., 2009) fail when applied to short peptides. Consequently, assigning functions to known NPPs and identifying previously overlooked related genes call for developing an alternative strategy. Furthermore, the shortage of experimental validated sequences is large and growing. Thus, methods that are primarily based on rules extracted from the limited number of known examples lack the ability to generalize to unseen instances.

In this research, we applied a supervised machine learning (ML) model based on extracting features directly from the primary sequence. Importantly, our method for identifying NPP candidates is 'alignment free'. Statistical support vector machine (SVM) models and decision tree-based classifiers (e.g. Random forests) were chosen as the preferred strategy (Nielsen *et al*., 1999). We present Neuropeptide Precursor Identifier (NeuroPID), a predicting machine that was trained on a curated set of NPPs. The high accuracy of NeuroPID was confirmed based on cross-validation (CV) tests. Furthermore, we selected a sparse set of features that contributed maximally to the successful classification by NeuroPID. Finally, we provide a candidate list of NPP-related proteins. A list of filtered NPP predictions from the fruit fly (*Drosophila melanogaster*), the worm (*Caenorhabditis elegans)*, the silkworm (*Bombyx mori*) and the red imported fire ant (*Solenopsis invicta*) is available at www.protonet.cs.huji.ac.il/neuropid/results.

In addition, we analyzed the proteome of the honeybee (*Apis mellifera*) and the Monarch butterfly (*Danaus plexippus*). The NeuroPID code is available at www.protonet.cs.huji.ac.il/neuropid.

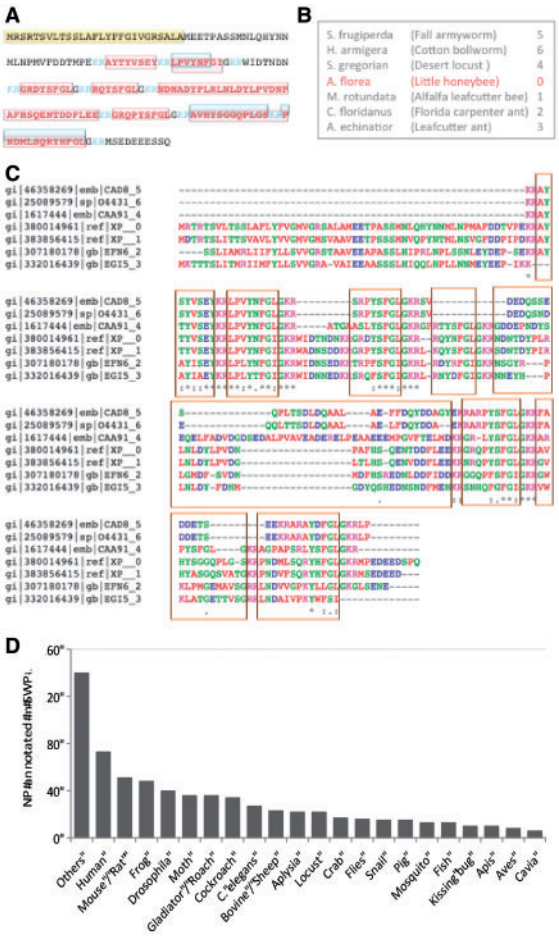

Fig. 1. Allatostatin family from insects. (**A**) Ten active NPs that are produced through cleavage of P85797 (UniProt) from *A. mellifera*. All the peptides were identified by MS technology, five of them were modified by amidation on Ile/Leu. (**B**) Sequences from the allatostatin family from Lepidoptera (seqs 5–6), Orthoptera (seq 4) and Hymenoptera (seqs 0–3) were used to assess their relatedness using ClustalW2 (Larkin *et al*., 2007). The sequence from *Apis florea* was used as a query (seq 0). It is similar to that of the *A. mellifera* (95% identical amino acids). (**C**) Multiple sequence alignment of sequences from the listed organisms. The set of candidate bioactive peptides are framed (as in A). (**D**) The list of organisms sorted by the number of annotated NPPs. Additional organisms (named as 'Others') are represented by a relatively small number of NPPs. Most sequences belong to mammals (human, mouse, rat, bovine and pig) and arthropods (crab, cockroach, flies and mosquitoes)

2 METHODS

2.1 Protein databases and proteomes

The main database resource used in this study is UniProtKB (Dimmer *et al*., 2012). The UniProtKB is composed of high-quality SW collection (called 'reviewed') and the TrEMBL database that archives the protein sequences from translated genomes. We collected all annotated NPPs by using the keyword 'Neuropeptide' [KW-0527]. In the list of NPs, we excluded NP receptors and sequences that are partial or the result of proteolysis (i.e. annotated 'Fragment'). Additional complete proteomes that were included in this study are the honeybee (*A.mellifera*), the Monarch butterfly (*D.plexippus*), the silkworm (*B.mori*) as well as the poorly annotated genome from the red imported fire ant (*S.invicta*). Most annotations were retrieved from UniProtKB. UniRef90 is a version of UniProtKB where proteins having >90% sequence similarity are clustered. Only one representative is considered from each cluster.

2.2 Training set

We collected the identified NPP sequences based on the annotation provided by UniProtKB (Dimmer *et al*., 2012). According to the notion of supervised classification task, the training phase calls for extracting features from 'positive' and 'negative' sets. Known NPPs comprise the 'positive' set, whereas the 'negative' sequences have functions unrelated to NPs (these are disjoint groups), wherein each sequence is a point in the high-dimensional space and each is transformed into a collection of binary and quantitative features.

The number of NP precursors in Metazoa is unknown. We demonstrate the difficulty using a prototype of allatostatins from *A. mellifera* (Fig. 1A–C). These NPs act to inhibit juvenile hormone biosynthesis and reduce the food intake (Stay and Tobe, 2007). Figure 1A shows the 10 active NPs that are produced from the precursor (197 amino acids). All 10 NPs were identified using a direct mass spectrometry (MS) (Hummon *et al*., 2006). Based on a comparative study of several insects, it is likely that most of the identified peptides are active as inhibitors of juvenile hormone production (Stay and Tobe, 2007).

Several observations should be noted from the example in Figure 1A–C. (i) The N'-terminal is occupied by signal sequences (signal peptide [SP], yellow background) (Petersen *et al*., 2011). (ii) The

length of the identified NPs ranges from 7 to 35 amino acids. (iii) Peptides may overlap. (iv) The flanking dibasic cleavage site motif [K|R] governs a cleavage signal for most identified peptides. (v) Glycine residue often precedes the cleavage site and is not included in the final peptide. The C-terminal glycine amidation is known to be a common post-translation modification for NPs (Merkler, 1994). Despite the complex processing of the allatostatin precursor, the abundance of dibasic motifs and their conservation is rather unique and seems to be a strong corollary for NP candidates (Veenstra, 2000). We assess the sequence similarity among insects that express allatostatin (Fig. 1B). Although the sequences of the NPs and the linkers between NPs show a low sequence similarity, the dibasic signals that are used as cleavage sites are conserved (Fig. 1C).

For the training phase, two non-redundant 'positive' sets were used: (i) 675 reviewed sequences from SwissProt (SW) according to their UniRef90 clusters and (ii) 2587 sequences from UniProtKB (combined non-reviewed and reviewed sets from TrEMBL and SW, respectively) according to their UniRef100 clusters. The organisms that dominate the list of known NPs from SW-UniRef90 are shown in Figure 1D. Several non-overlapping 'negative' sets were compiled based on the same length distributions and an identical composition of organisms with respect to the NP-curated set (i.e. 'positive'). We also kept the randomly selected negative set's proteins to within 100–150% of the size of the positive set.

2.3 Binary and quantitative features

Our goal was to define a set of features that would be instrumental in identifying NPPs. Importantly, all calculated features relied solely on the pre-pro-protein's primary sequence. The properties that rely on external predictors (e.g. 3D structural fold, secondary structure) or search engines (e.g. sequence similarity, evolutionary distance scores) were excluded. In the same line, we ignored known functional attributes (e.g. GO annotation enrichment). Short motifs that were extracted are guided by the proteolytic cleavage preferences. We also included some signatures that cover additional post-translational modifications such as amidation.

The features that were extracted from the 'positive' sets of NPPs are partitioned to several types:

(A) Biophysical quantitative properties:

(i) Molecular weight.
(ii) Charged amino acids occurrence.
(iii) Sequence length.
(iv) Isoelectric point (PI).
(v) Aromaticity (The relative frequency of Phe, Trp, Tyr).
(vi) Amino acid usage (20 features).
(vii) Bigrams pair from the calculated dipeptide frequencies (400 features).
(viii) The instability index—an estimate for the stability of a protein *in vitro* (Wilkins *et al.*, 1999).
(ix) GRAVY (Grand Average of Hydropathy)—the sum of hydropathy values of all amino acids, divided by the number of residues in the analyzed sequence (Kyte and Doolittle, 1982).
(x) Aliphatic index—the relative volume occupied by aliphatic side chains Ala, Val, Ile and Leu (Wilkins *et al.*, 1999). These properties have been shown to be informative and valuable for the classification tasks.

Most properties were derived from EXPASY prediction tools (Artimo *et al.*, 2012). We complemented the prediction tools from Biopython (Cock *et al.*, 2009) and Python (version 2.7). The power of these global features to predict function has been previously validated (Varshavsky *et al.*, 2007). The features from this section that exhibited substantial significance (P-value of <0.01) were included in the ML training phase.

(B) Binary features:

This group of features aims to capture the non-randomized appearance of certain amino acids within short windows. The binary features were successfully used for identifying secreted short toxin-like proteins (Tirosh *et al.*, 2012) in a large set of genomes (Naamati *et al.*, 2010). Specifically, it is designed as a 5-mers (k-mer) sliding window (2^5 = 32 features). A sequence is divided into 5-mers. The binary 5-mer was used for the following groups:

(i) Charged group—Asp, Glu, Lys, Arg, His (D,E,R,H,K).
(ii) Charged-polar group—Asp, Glu, Lys, Arg, Asn, Gln (D,E,R,K,N,Q).
(iii) GKR (G = Gly).
(iv) Basic residues—Lys, Arg (K,R).

In each case, the 'relevant group' of amino acids was signed as 1, whereas the other amino acids were signed as 0. Most solid information on NPs along their evolutionary tree concerns their precursors' cleavage model (Southey *et al.*, 2006).

The specificity of cleavage sites was defined according to the appearance of the following:

- *-*-Lys-Lys#
- *-*-Lys-Arg#
- *-*-Arg-Arg#
- Arg-*-*-Lys#
- Arg-*-*- Arg#

Where the # and * denote a cleavage site and other amino acids, respectively. We generalized the code to adapt it for a more relaxed definition. For example, the feature for the suspected cleavage site was formulated by the motif using the '**[R or K] [R or K]' and 'R**[R or K]; (lysine = K, arginine = R)' [known motif model (Southey *et al.*, 2006; Veenstra, 2000)].

(C) Information-based statistics:

C.1. Amino acid entropy (20 features): This set of features captures a property that demonstrates how non-randomly distributed each amino acid is in the sequence, based on the concept of molecular information entropy (Schneider, 2010). For a given sequence S amino acid type C, we mark its position in the sequence $p_1, \ldots, p_k$

The length of S and the number of C in the S are marked k and m, respectively.

We define that $p_0 = 0$ and $p_{k+1} = m + 1$

We define the entropy of C to be

$$entropy(c, s) = \sum_{i=1}^{k+1} \left(\frac{p_i - p_{i-1}}{m}\right) \log_2 \left(\frac{p_i - p_{i-1}}{m}\right)$$

C.2. Autocorrelation: This measure was calculated for certain amino acids that were calculated from the Bigram frequencies. The intuition stems from the observation of a consistent spacing of certain residues or short liner motifs (e.g. KK).

C.3. Over- and underrepresentation of motifs along the sequences: We marked the density of 'interesting' motifs in the sequence (normalized to the sequence length). Such motif 'counts' included the following:

(i) Canonic cleavage sites based on the known motif model (Southey *et al.*, 2006).
(ii) Potential site for N-glycosylation, formulated as [N]-[not P]-[S/T/C].
(iii) Potential aspartic acid and asparagine hydroxylation sites.

(iv) The high preference of certain amino acids in the vicinity of suspect proteolytic dibasic cleavage sites. Specifically, we count the tandem pair R/K and the adjacent amino acids (separated into amino acid equivalence groups. Note that several amino acids appear in several groups—'FYW', 'P', 'C', 'RHK', 'DE', 'CSTMNQ', 'RK', 'ST' and 'EKRDNQH'). We count the appearance in 1–2 positions that preceded or succeeded the [KR][KR] cleavage motif.

All classical NPs (as well as some of the non-classical ones) are secreted proteins (i.e. with SP but not transmembrane domain) that undergo processing in the endoplasmic reticulum (ER). Note that SP in all secretory proteins must be removed before the production of the precursor protein (Nielsen *et al.*, 1999). We did not include the SP as a feature in the training phase.

3 IMPLEMENTATION

3.1 ML tools

Proteins were labeled 'positive' or 'negative' according to the database in SwissProt (SW, 'reviewed' sequences). A two-sample Kolmogorov–Smirnov (KS) test was used to compare the distributions of values with respect to the background frequencies of all SW proteins. In addition, the variance for the 'positive' and 'negative' sets was calculated and used for defining the statistical significance of the distributions. A two-sample *t*-test was applied to compare the means of each distribution and to identify inherent differences.

There are several commonly used kernels that can be used in the ML scheme (Amari and Wu, 1999). Most notably are (i) the linear kernel: [u'*v], (ii) polynomial kernel [(gamma*u'*v + coef0)^degree] and (iii) non-linear RBF (radial basis function) kernel: e(-gamma*ju-vj^2) (Amari and Wu, 1999).

A specific kernel or overall approach may be more or less suitable, depending on the case at hand (Lewis *et al.*, 2006; Seeger, 2004). We experimented with a few of the routinely used kernel functions including the linear kernel, 2D and 3D degree polynomial kernel and the (non-linear) Gaussian RBF kernel in combination with a support vector classification (SVC) machine. These methods were initially chosen for their robustness, good programmatic support and resistance to overfitting. The resistance of the different classification and boosting methods to the 'curse of dimensionality' (given a small number of samples and numerous features) was extensively studied (Har-Peled *et al.*, 2012).

We used LibSVM, a general library for SVC and regression (http://www.csie.ntu.edu.tw/~cjlin/libsvm). ML programs are based on SVM and are implemented in the Matlab software toolkit (Zhao *et al.*, 2009). The scikit-learn toolkit implements it in the Python programming language (Pedregosa *et al.*, 2011). The performance of the SVM for parameter selection was tested. We used the LibSVM toolkit for exploring the hyper-parameter space. The selected parameter includes the penalty parameter C and the kernel parameter γ (gamma) of the RBF function (Fan *et al.*, 2008). The dependency of the ML predictor on the selected toolkit was noted but the impact on the results was found to be negligible and will not be discussed further.

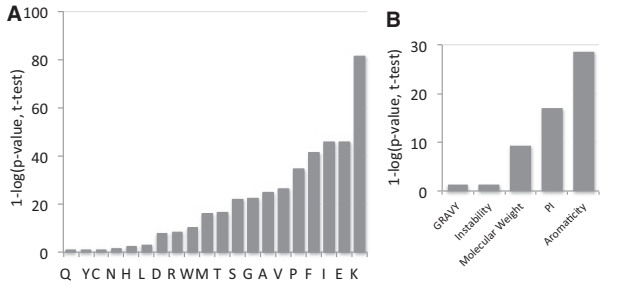

Fig. 2. Statistical significance for individual features extracts from known NPs. (**A**) The statistical significance for the occurrence of 20 amino acids from NPs with respect to background. The strong statistical value associated with K is a reflection of the preferable cleavage sites in NPPs. (**B**) The statistical significance for several selected biophysical properties. The aromaticity of NPs shows the most informative value. The y-axis is measured as 1-log(*P*-value)

3.2 Transforming sequences into features and statistics

We extracted features directly from the primary sequences and then used them to train an SVM. Recall that the performance of NeuroPID depends strongly on the set of features that were extracted from each of the previously identified NPPs as well as the negative sets that were used for training. SVM methods are resistant to possible overfitting or dimensionality, allowing a framework to examine the feature space's predictive power, without the need to stringently remove 'noisy' or potentially weak features in advance. We transformed various biochemical traits of the protein sequences into quantitative features (as described in Section 2.3).

The features that were found to be significant in the statistical tests are listed (Fig. 2). We examined each of the biochemical features before training the ML. The quantitative traits' statistical significance was a condition of their inclusion as features (as described in Section 2.3). Most of the features were statistically significant (2-sided *t*-test, $\alpha = 0.01$ or KS test <0.01). Exceptions to the latter were the instability index ($P = 0.36$), the GRAVY ($P = 0.41$) and amino acid frequencies of Cys ($P = 0.72$), His ($P = 0.04$), Asp ($P = 0.25$) and Asn ($P = 0.88$). A complete list of features and their statistical significant appearance in NPs versus the general proteins' background is available in Supplementary Data S1.

Figure 2 shows the statistical information for 20 amino acids occurrence (Fig. 2A) and several biophysical properties (Fig. 2B). The most individual significant feature is the occurrence of lysine (K, $P < 1.0e\text{-}83$).

3.3 NeuroPID—a discovery tool

A schematic flow of the research is shown in Figure 3. Briefly, the input to NeuroPID is a set of sequences, each described by hundreds of features. Some of the ML methods when applied to complete proteomes, containing thousands of proteins, resulted in hundreds of predictions (not shown). To reduce false positives (FPs), we applied SP prediction [SignalP 4.0 (Petersen *et al.*, 2011)]. Recall that the occurrence of SP was not used as a feature, but only as a filter to refine the final list of predicted 'candidate' NPPs.

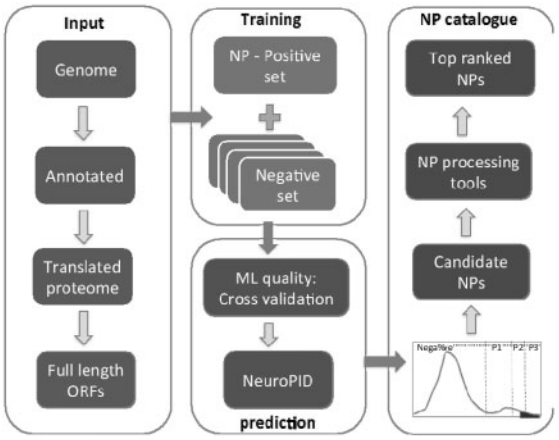

Fig. 3. A protocol for NPP prediction using NeuroPID. The working protocol is composed of three sections: (i) collecting the proteome-scale full-length sequences (ORFs); (ii) training and testing the performance of the ML; (iii) assessment of the predictions and the candidate NPPs. A refined list of predicted NPPs is created by filtering out sequences that lack SP at their N'-terminal segments

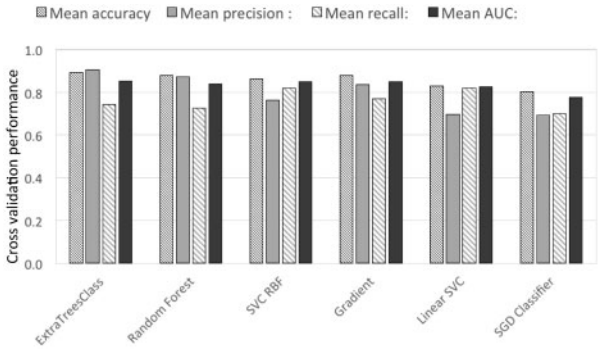

Fig. 4. CV tests for NPs using the complete non-redundant set from UniProtKB. Numerous classification models support the high accuracy and precision in the CV tests. The variations of the results in repeated tests (measured as mean accuracy, precision, recall and AUC) were negligible. SVC, support vector classifier; RBF, radial basis function kernel; SGD, stochastic gradient descent

An additional filter that was used to reduce the false-positive prediction takes advantage of the existing tools for predicting the cleavage sites that lead to the production of active NP peptides (Southey *et al.*, 2006, 2008). Thus, we tested the output of NeuroPID as input to cleavage site prediction tools.

The acquisition of the dataset from UniProtKB (Dimmer *et al.*, 2012) and the generation of the feature data from FASTA format sequences were done using the BioPython library (Cock *et al.*, 2009). Python (www.python.org) and Matlab code for the implementation are available on request.

3.4 Cross-validations

We randomly split the sequences and carried out 6-fold cross-validation (CV). A substantial fraction of the data (i.e. 10–40%) was used as a disjoint set and was not included at the training phase. The results of the CV tests for each of the NPP candidates were summed up to estimate the accuracy, sensitivity, precision and area under ROC curve (AUC). We used the default statistical definitions:

$$\text{Accuracy} = \text{TP} + \text{TN}/\text{TP} + \text{FP} + \text{TN} + \text{FN}$$

$$\text{Precision} = \text{TP}/\text{TP} + \text{FP}$$

$$\text{Recall} = \text{TP}/\text{TP} + \text{FN}.$$

After having generated the features data from the various datasets, we set up the ML. We initially applied a linear SVM machine. We estimated the performance and success of NP prediction in a 5-fold CV test. The accuracy rate for NP identification (out of a given set of proteins) in the CV tests reached a level of 82–89% accuracy. The range captures the variation on the machine method, the kernels and the parameters used. The CV tests for precision and recall yielded 73–91% and 77–82%, respectively. The AUC for all methods is a robust measure that ranges from 82 to 87%. To prevent overfitting, we only used 30–40% of the data for training and for testing the aforementioned parameters. Additionally, the reported results were based on balanced 'positive' and 'negative' training sets (1166 sequences in each set, Fig. 4).

We further examined the robustness of the findings in terms of the number of FPs and false negatives (FN) by repeating the CV phase with different protocols. We varied the fraction of the trained set to account for only 60–90% of the input, and repeated the test multiple times, leaving out different random subsamples each time. The input includes all NPs (at 90 and 100% identity, non-redundant set). In all tests, the results were stable as confirmed by repeating the protocols several times with independent 'negative' sets. Figure 4 shows the results from different classifiers and boosting models for 6-fold CV tests.

We repeated the CV tests using alternative kernels (polynomial ^2 and ^3 kernel), a stochastic gradient descent SVM and a non-linear model (called SVC–RBF). All these changes resulted in similar results (Fig. 4). The use of more complex polynomial kernel models increased the run-time substantially with a minimal improvement of the results (as measured by AUC). Thus, we continued to analyze models using Random Forest, Gradient Boosting and Extra Trees ensemble decision trees, as well as the 'best' classification SVM models (SVC-RBF, and linear SVC).

The CV tests for numerous classification models for the annotated NPs reported a high level of accuracy. An AUC value of 0.82–0.87 was reached for all NPPs in UniProtKB and used as a baseline for training of the NeuroPID (Fig. 4). Recall that the performance of the NeuroPID depends strongly on the set of features that were included in building the model (the 'Feature Space'). Such information was determined and measured during the ML training phase. Inclusion of physiochemical traits was reliant on statistical analysis for significance. Features that showed no significant differences in the statistical tests (KS-test/*t*-test, Fig. 2) were not included in the ML scheme.

Before activation of 'new prediction' scheme, additional CV tests were performed while the training phase was repeated varying 'negative' sets, alternative kernels function and extensive

parameter tuning. Results were normalized to provide measurement for the accuracy (defined by the precision and the recall) for the various models.

To assess the robustness of NeuroPID, we repeated the procedure of tuning a predicting machine by selecting alternative 'negative' sets. We performed it for the 'negative' set of secretory proteome (SP or TMD comtaining proteins) and the proteins composed of nuclear proteins. The rationale behind the selection of the later set was to provide a 'challenging' set in which the inherent appearance of basic residues prevails (e.g. histons, transcription factors). The performance in a 6-fold CV test remained high, with AUC for the different models ranging from 0.86 to 0.88.

3.5 Taxonomy-based CV

Once the ML was trained and tested, we examined unseen examples derived from complete proteomes. We defined the task as identifying unseen NPPs for sequences that belong to specific taxa (e.g. Chordata and arthropods). We repeated the tests by training the ML from SW or UniProtKB (see Section 2). In all cases, only representative sequences from UniRef90 clusters were used. The results are summarized in Table 1.

The high performance of the CV for the Arthropoda and Chordata shows that restricting the prediction to any specific taxa increased the performance in almost all instances. A careful analysis shows that the improved performance for arthropods is mainly in the Recall and Precision (improvement of >10% when compared with the analysis for all the NP sequences to the taxa-based analysis). Importantly, all classification models (three methods, Table 1) show a similar trend in the CV test in view of the performance of the general analysis (Fig. 4). Note that as can be anticipated, the CV tests outperformed for the SW collection in the taxa-based analysis.

A similar performance was recorded by using a larger set of the redundant NPP collection. The rest of the predictions were performed on the basis of the non-redundant UniProt collection.

3.6. Post-training feature selection in NeuroPID

Several methods for supervised feature selection can be applied to identify the minimal set of features that contribute maximally to the rich model underlying the NeuroPID performance.

The initial set (see Section 2.3) included 561 features. We applied a 10–20% step reduction with a L1 loss function. The results from analysis via RFECV function (Pedregosa *et al.*, 2011). The function is based on feature ranking with recursive feature elimination and 3-fold stratified cross-validated selection for the best number of features. The RFECV function, implemented with 10% feature elimination per step, identified 13 'strong' features. Repeating the protocol and setting a step reduction for 30 features in each step, resulted in 21 selected features. A similar procedure, recursive feature elimination proposed 15 features. Checking the identity of these features revealed an overlapping core set of features, shared by most of the aforementioned feature extraction methods. A degree of randomality is expected and unavoidable, given a robust set of non-sparse features, combined with decision trees. Hence, the variability. Applying principle component analysis for identifying the most informative features is guided by geometrical considerations for compression into a sparse feature space. It led to 22 features that are both sparse and informative. A unified list of 23 highly informative features that are consistent between the feature extraction methods and the principle component analysis method includes molecular weight, counts of potential cleavage sites, binary appearance in 5-mer window for the specific appearance of KR (5 of 32 combinations), GKR (3/32 combinations), charged residues (7/32 combinations) and the entropy level for G, N and P. Interestingly, none of the 400 features of the bigram (dipeptide frequency) had been selected in the feature extraction method. For a detailed analysis of informative features see Supplementary Data S2.

3.7 NeuroPID models for insect proteomes

The NeuroPID was tested with several organisms on complete proteomes, ranging from 10K for *A. mellifera* to almost 40K for *D. melanogaster*. Several ML methods (Random forest, Extra

Table 1. CV test for correct assignment of NPs according to taxonomical partition

	Random forest	SVC RBF	Gradient boosting	Linear SVC	Random forest	SVC RBF	Gradient boosting	Linear SVC
	SW Arthropods				UniProt Arthropods			
Mean accuracy	0.944	0.923	0.948	0.942	0.916	0.900	0.923	0.860
Mean precision	0.937	0.892	0.946	0.931	0.931	0.968	0.941	0.950
Mean recall	0.924	0.920	0.924	0.924	0.954	0.889	0.952	0.850
Mean AUC	0.941	0.922	0.945	0.939	0.887	0.908	0.901	0.868
	SW Chordata				UniProt Chrodata			
Mean accuracy	0.963	0.936	0.965	0.951	0.898	0.746	0.907	0.845
Mean precision	0.938	0.835	0.943	0.881	0.914	0.906	0.923	0.893
Mean recall	0.912	0.924	0.915	0.928	0.906	0.621	0.914	0.828
Mean AUC	0.946	0.932	0.948	0.943	0.897	0.768	0.906	0.848

Table 2. NeuroPID prediction performance for individual species

Organism	Number of UniProtKB	Number of full-length UniProtKB	Number of SP annotated	Number of NP and SP annotations	Number of NeuroPID predictions, All methods	Representative GO function
Bombyx mori	17 908	17 069	138	6	69	Innate immunity;Insulin-like; Chorion, Hormne (NP)
Solenopsis invicta	14 356	84	12	2	4	Innate immunity
Drosophila melanogaster	39 961	31 091	475	21	120	Innate immunity;Developmental; Channel ligand; Receptors, Hormne (NP)
Caenorhabditis elegans	26 005	25 534	464	21	89	Hormne (NP), Channel ligand; Receptor, Protease

SP, Signal peptide sequence according to UniProtKB annotation.

Tree classifier, SVC-RBF, linear SVC and Gradient Boosting) were applied on the complete proteomes in an unfiltered mode resulting in a large number of weak predictions (1–5K NPP prediction for a proteome). Including a probability threshold leads to a drastic reduction in the output, marking only the top ranked predictions. For example, in the Random Forest protocol, a 'certenty' threshold of 0.99 reduced the unfiltered predictions for *B. mori* from >4000 to only 16 positive results results. Similarly, the application of Gradient Boosting (GBR) protocol for the Monarch butterfly (16 183 sequences, 4856 predictions) reduced the number of predictions 10-fold at a 99.5% threshold.

NeuroPID was activated with SVM and decision tree classifier methodologies to identify novel candidate NPPs. The NeuroPID results for NPPs from the silkworm, fly, worm and fire ant are found in www.protonet.cs.huji.ac.il/neuropid/results/.

Table 2 summarizes the prediction results from NeuroPID for a filtered input. In all cases, the proteins that are annotated as NPs were eliminated from the training phase to avoid 'trivial' predictions. The filters include (i) consistency for all four of the prediction methods (Extra Trees, RBF SVM, linear SVM and Gradient Boosting); (ii) presence of SP in the sequence; (iii) the sequences are full-length (not fragment). Filters (ii) and (iii) are based on the annotations from UniProtKB. We illustrate the results for proteomes that are extremely well annotated (worm, fruit fly) and proteomes that are poorly annotated (fire ant, silkworm). A common finding across the proteomes is the functional enrichment of innate immunity, some structural proteins as well as hormones and NPs. However, FPs that include short ligands and some receptors remain. Adding the 'certenty threshold' (see above) and a confirmation for high probability cleavage of the sequence, using tools such as NeuroPred (Southey *et al.*, 2006) is the preferred protocol to reduce the number of FP predictions.

We further analyzed the results for additional proteomes. Figure 5 shows the results and the consistency level among the methodologies that were applied. The consistency among the decision tree classifiers was specifically high. Considering all four mentioned methods, we identified eight sequences that were supported by at least three of them for *A. mellifera*. A similar trend was found for the analysis of *Thaumeledone gunteri*) (Fig. 5). In such instances, the annotated NPP sequences were excluded from the training phase.

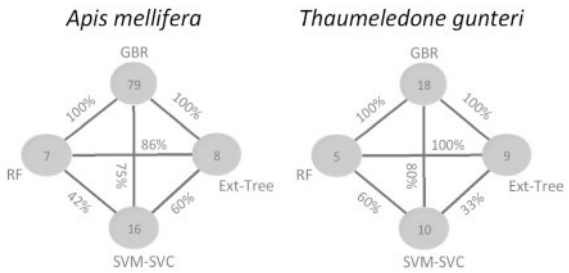

Fig. 5. NeuroPID prediction based on SVM and decision tree methods for the honeybee and octopus proteomes. The number of predictions is shown (nodes) along the percentage of the overlap for each the prediction pairs. The percentage is calculated from the smaller number of predictions in the pair. Random Forest (RF), Extra Trees classifier (Ext-Tree), SVM–SVC (polynomal kernel) and Gradient tree boosting (GBR)

One of the sequences that was consistently identified by the different ML methods is an uncharacterized sequence from *A. mellifera* (UniProt: H9K152, 183 amino acids). We propose that this sequence is an overlooked NPP. (i) It is a secreted protein. (ii) The most significant conservation motif of the dibasic sequences dominates the sequence. (iii) Using MS identification, a 52-amino acid peptide belonging to the full-length ORF was identified in the honeybee brain extract (Audsley and Weaver, 2006). (iv) Pfam resource (Punta *et al.*, 2012) identified an evolutionary domain that is associated with active peptides from frog, fish and mammals. The CRF domain (corticotropin-releasing factor) is postulated in stress and anxiety, osmolarity, thermoregulation, growth and metabolism in mammals.

4 DISCUSSION

4.1 NeuroPID performance

The set of NPs that are produced from a single NPP often share little sequence resemblance but a strong functional coherence. Thus, it is evident that the functionality of the NPs cannot be assessed without considering their cognate GPCRs that act as signaling receptors (Jekely, 2013).

In this study, we only considered the sequence features of the NPPs from UniProtKB. The goal of this research is to identify NPPs (and related proteins) in a genomic scale. We show that despite the fact that SP sequences were not included as a feature in the training phase, we were able to successfully identify many of the NPPs from an input of secreted proteome (not shown). Based on this finding, we confirm that the information captured by the selected features (~600 individual features) provides rich information for a high-performing predictor.

In ML-based methodologies, the properties and relative size of the 'negative' set is of utmost importance (Ben-Hur *et al.*, 2008). We repeated the training and learning phases using a 'negative' set that includes a matched length distribution from secretory proteins (i.e. contains SP/TMD that are not NPs). It resulted in an excellent performance in the CV test (AUC = 0.94).

To overcome any possible overfitting, we repeated the training and learning phases after removal of a large set of features, and with an unbiased negative training set. Specifically, we removed the bigram (400 features) and a number of amino acids prefix/suffix groups, leaving ~152 features. In this new setting, we used both a larger set for the training phase (>1165 'positive' instances) and a controlled set. We overcame the pitfall of overfitting by using an unbiased 'negative' training set. Improved training performance was achieved using the aforementioned larger more heterogenous UniProt 'positive' set (~2600 proteins, AUC = 0.82–0.91), and testing performance (identification of true negatives) was significantly improved as well. Based on the above observations, we concluded that pruning for a limited set of features (by different methodologies of feature ranking and elimination) results in high performance of the NeuroPID and an improved run time while not deteriorating the predictive power. The procedure and the machine performance remained robust for a wide range of tuned parameters.

Inspecting the results of the NPP predictions shows the need to remove FPs. Many of the predicted NPPs are named 'uncharacterized', but also receptor proteins, enzymes and even nuclear proteins are listed among the top predictions. Filtration of FPs takes advantage of the consistency among independent methods (Table 2, Fig. 5). For example, proteolytic products (Zhang and Zhu, 2012) that act as antimicrobial peptides were identified among the top predictions from *S. invicta* (e.g. I2E7P6, I2E7P4). These sequences were predicted by all the tested ML methods (GBR, RF, Ext-Tree and SVM-SVC). Submitting these sequences to NeuroPred (Southey *et al.*, 2006) further reduces the FPs. NeuroPred reports the probabilities for proteolytic cleavages that are essential for the formation of NPs.

We illustrate the analysis for 89 proteins from *C. elegans* (Table 2). We first tested the occurrence of genuinely known NPPs in the filtered list. There are 10 NPPs that belong to several families: insulin-like peptides (3), FMRF-neuropeptides (5) and neuropeptide-like (npl) proteins (2). In addition, there are nine proteins that are named 'uncharacterized'.

We then submitted the combined list of NPP candidates (with agreement of all 4 ML methods, Table 2) to NeuroPred (Southey *et al.*, 2006). Inspecting the probability of cleavage according to NeuroPred (Southey *et al.*, 2006) shows only four proteins that have no predicted cleavage sites. Interestingly, two of them are NPPs (npl-28 and npl-30) that have non-canonical cleavage sites. For the rest of the proteins, many high probability cleavage sites (NeuroPred threshold >0.8) were identified (~11/protein). We observed that it is not the overall high probability of cleavage sites but the sites' density (i.e. number of sites/100 amino acids) that characterize known NPPs. The uncharacterized proteins: P34639 (ZK512.1) and P52881 (F46C5.2) have a high density of potential protolytic sites. Although a functional validation for these sequences is missing, the distinct developmental expression and a resemblance to proteins that function in intercellular interactions are evident. Additional protein from the *C. elegans* list that shares some characteristics with NPPs is P91573 (Warthog protein 6). The protein is subjected to a process of autocleavage. The resultant peptides are active in intercellular signaling.

4.2 MS-validated NPs

The experimental validation for NPPs is mostly based on the identification of NPs through a peptide identification scheme using MS (Svensson *et al.*, 2007). The impact of the MS technologies in the field of NP identification cannot be underestimated (Altstein and Nassel, 2010; Hummon *et al.*, 2006; Ons *et al.*, 2011). However, the short length of NPs, the complex set of partially cleaved products and the abundance of post-translation modifications result in low coverage of NPs. The matrix-assisted laser desorption/ionization (MALDI) MS has been successfully used to identify hundreds of NPs from a single or multiple neurons. Together with other MS technologies (e.g. electrospray ionization techniques), a large collection of short peptides becomes available (Skold *et al.*, 2007). The bulk of the peptide spectra may include cytokines, growth factors, antimicrobial peptides, toxin-like proteins (Tirosh *et al.*, 2013) and protein degradation intermediates. However, the challenge remains to successfully identify the subset of modulatory NP peptides from accurate tandem MS data.

To this end, peptide-centric databases were developed for accelerating the identification process from complex biological samples [SwePepe (Falth *et al.*, 2006); NeuroPedia (Kim *et al.*, 2011)]. These resources include thousands of potential peptides (and their modified versions). A complementary database archives annotated bioactive peptides according to their functional groups [PeptideDB (Liu *et al.*, 2008)]. Importantly, matching a short peptide to a complete proteome from a multicellular organism may not satisfy the minimal significant threshold (Noble and MacCoss, 2012). In this study, we propose NeuroPID as an initial layer of filtration. The user is encouraged to search for a peptide match among the filtered collection of NPP candidates using NP prediction tools and knowledge-based resources.

In this study, we mainly focused on identifying insect NPPs from model organisms (e.g. fruitfly) and from a number of poorly annotated proteomes (e.g. ant). At present, the most complete annotations for NPs rely on data from MS technology. The peptidome of brain and hemolymph from *D. malenogaster* revealed ~30 peptides, some of which are genuine NPs (Schoofs and Baggerman, 2003). When applied to *A. mellifera,* >200 NPs from 36 genes were reported, of which about half were confirmed experimentally. The metabolic states and the behavior of insects are dictated by the compositions of NPs. Consequently, high coverage identification of NPs is of a great importance to the biotechnology industry (Altstein, 2001).

Differently from most proteins for which homology search is a powerful technique for functional inference (Loewenstein *et al*., 2009), the inference of validated NPs is limited to closely related species. Recently, a collection of NPs from MS experiments in *C. elegans* was successfully used to reveal the conservation of NPs in *Caenorhabditis briggsae* (Husson *et al*., 2009). In view of the MS-based discovery for insects, the NeuroPID provides a robust method that is insensitive to sequence similarity. We suggest applying NeuroPID as a first step in the discovery workflow (Fig. 3). It presents a putative list of NPPs for newly sequenced genomes. The list of top candidates can be further refined using available prediction tools. For example, NeuroPred (Southey *et al*., 2006) provides the probability of the proteolytic site along the length of the precursor sequence.

4.3 NeuroPID annotations of unexplored genomes

We present NeuroPID, a robust method for identifying NPPs on a proteome-wide scale. Arthropods are a taxon in which NPPs expanded to provide a large collection of bioactive peptides that execute social and behavioral tasks.

The efficiency and quality of gene annotation and especially experimental validated proteins lags behind the explosion in sequencing. Identifying NPPs is important not only for the sake of increasing annotation coverage but also as a mean to regulate insect social behavior, metabolic state and communication. In this view, an exciting genomic initiative with the goal of sequencing 5000 social insects genomes was recently announced (Robinson *et al*., 2011). The expectation is that NeuroPID will be a valuable contribution in identifying components that govern the behavior and elements of insect communication. The predictive power of NeuroPID is expected to improve along with the increase in the coverage of NPPs, mainly from MS experimental data. We expect that the analysis presented in this study will be useful in leveraging future expansions of the protein space.

ACKNOWLEDGEMENTS

The authors thank Nadav Rappoport and Kerem Wainer for reading the manuscript and for useful comments and discussions throughout the project.

Funding: The work was partially supported by PROSPECT, EU FRVII (201648) and the ISF grant 592/07.

Conflict of Interest: none declared.

REFERENCES

Altstein,M. (2001) Insect neuropeptide antagonists. *Biopolymers*, **60**, 460–473.

Altstein,M. and Nassel,D.R. (2010) Neuropeptide signaling in insects. *Adv. Exp. Med. Biol.*, **692**, 155–165.

Amari,S. and Wu,S. (1999) Improving support vector machine classifiers by modifying kernel functions. *Neural Netw.*, **12**, 783–789.

Artimo,P. *et al.* (2012) ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.*, **40**, W597–W603.

Audsley,N. and Weaver,R.J. (2006) Analysis of peptides in the brain and corpora cardiaca-corpora allata of the honey bee, Apis mellifera using MALDI-TOF mass spectrometry. *Peptides*, **27**, 512–520.

Ben-Hur,A. *et al.* (2008) Support vector machines and kernels for computational biology. *PLoS Comput. Biol.*, **4**, e1000173.

Brain,S.D. and Cox,H.M. (2006) Neuropeptides and their receptors: innovative science providing novel therapeutic targets. *Br. J. Pharmacol.*, **147** (**Suppl. 1**), S202–S211.

Clynen,E. *et al.* (2010) Bioinformatic approaches to the identification of novel neuropeptide precursors. *Methods Mol. Biol.*, **615**, 357–374.

Cock,P.J. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Dimmer,E.C. *et al.* (2012) The UniProt-GO Annotation database in 2011. *Nucleic Acids Res.*, **40**, D565–D570.

Falth,M. *et al.* (2006) SwePep, a database designed for endogenous peptides and mass spectrometry. *Mol. Cell Proteomics*, **5**, 998–1005.

Fan,R.E. *et al.* (2008) LIBLINEAR: a library for large linear classification. *J. Mach. Learn. Res.*, **9**, 1871–1874.

Funkelstein,L. *et al.* (2010) Unique biological function of cathepsin L in secretory vesicles for biosynthesis of neuropeptides. *Neuropeptides*, **44**, 457–466.

Gelman,J.S. and Fricker,L.D. (2010) Hemopressin and other bioactive peptides from cytosolic proteins: are these non-classical neuropeptides? *AAPS J.*, **12**, 279–289.

Gonzalez-Rey,E. *et al.* (2007) Regulation of immune tolerance by anti-inflammatory neuropeptides. *Nat. Rev. Immunol.*, **7**, 52–63.

Har-Peled,S. *et al.* (2012) Approximate Nearest Neighbor: Towards Removing the Curse of Dimensionality. *Theory Comput.*, **8**, 321–350.

Hummon,A.B. *et al.* (2006a) Discovering new invertebrate neuropeptides using mass spectrometry. *Mass Spectrom. Rev.*, **25**, 77–98.

Hummon,A.B. *et al.* (2006b) From the genome to the proteome: uncovering peptides in the Apis brain. *Science*, **314**, 647–649.

Husson,S.J. *et al.* (2009) Comparative peptidomics of Caenorhabditis elegans versus C. briggsae by LC-MALDI-TOF MS. *Peptides*, **30**, 449–457.

Insel,T.R. and Young,L.J. (2000) Neuropeptides and the evolution of social behavior. *Curr. Opin. Neurobiol.*, **10**, 784–789.

Jekely,G. (2013) Global view of the evolution and diversity of metazoan neuropeptide signaling. *Proc. Natl Acad. Sci. USA*, **110**, 8702–8707.

Kim,Y. *et al.* (2011) NeuroPedia: neuropeptide database and spectral library. *Bioinformatics*, **27**, 2772–2773.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Larkin,M.A. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Lewis,D.P. *et al.* (2006) Support vector machine learning from heterogeneous data: an empirical analysis using protein sequence and structure. *Bioinformatics*, **22**, 2753–2760.

Liu,F. *et al.* (2008) The construction of a bioactive peptide database in Metazoa. *J. Proteome Res.*, **7**, 4119–4131.

Lobley,A. *et al.* (2009) pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics*, **25**, 1761–1767.

Loewenstein,Y. *et al.* (2009) Protein function annotation by homology-based inference. *Genome Biol.*, **10**, 207.

Mentlein,R. and 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.

Merkler,D.J. (1994) C-terminal amidated peptides: production by the in vitro enzymatic amidation of glycine-extended peptides and the importance of the amide to bioactivity. *Enzyme Microb. Technol.*, **16**, 450–456.

Mirabeau,O. *et al.* (2007) Identification of novel peptide hormones in the human proteome by hidden Markov model screening. *Genome Res.*, **17**, 320–327.

Naamati,G. *et al.* (2010) A predictor for toxin-like proteins exposes cell modulator candidates within viral genomes. *Bioinformatics*, **26**, i482–i488.

Nassel,D.R. (2002) Neuropeptides in the nervous system of Drosophila and other insects: multiple roles as neuromodulators and neurohormones. *Prog. Neurobiol.*, **68**, 1–84.

Nielsen,H. *et al.* (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.*, **12**, 3–9.

Noble,W.S. and MacCoss,M.J. (2012) Computational and statistical analysis of protein mass spectrometry data. *PLoS Comput. Biol.*, **8**, e1002296.

Ons,S. *et al.* (2011) Neuropeptide precursor gene discovery in the Chagas disease vector Rhodnius prolixus. *Insect Mol. Biol.*, **20**, 29–44.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res. Arch.*, **12**, 2825–2830.

Petersen,T.N. *et al.* (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, **8**, 785–786.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
Robinson,G.E. *et al.* (2011) Creating a buzz about insect genomes. *Science*, **331**, 1386.
Schneider,T.D. (2010) 70% efficiency of bistate molecular machines explained by information theory, high dimensional geometry and evolutionary convergence. *Nucleic Acids Res.*, **38**, 5995–6006.
Schoofs,L. and Baggerman,G. (2003) Peptidomics in Drosophila melanogaster. *Brief. Func. Genomics Proteomics*, **2**, 114–120.
Seeger,M. (2004) Gaussian processes for machine learning. *Int. J. Neural Syst.*, **14**, 69–106.
Skold,K. *et al.* (2007) The significance of biochemical and molecular sample integrity in brain proteomics and peptidomics: stathmin 2-20 and peptides as sample quality indicators. *Proteomics*, **7**, 4445–4456.
Southey,B.R. *et al.* (2006) NeuroPred: a tool to predict cleavage sites in neuropeptide precursors and provide the masses of the resulting peptides. *Nucleic Acids Res.*, **34**, W267–W272.
Southey,B.R. *et al.* (2008) Prediction of neuropeptide cleavage sites in insects. *Bioinformatics*, **24**, 815–825.
Stay,B. and Tobe,S.S. (2007) The role of allatostatins in juvenile hormone synthesis in insects and crustaceans. *Annu. Rev. Entomol.*, **52**, 277–299.
Svensson,M. *et al.* (2007) Neuropeptidomics: MS applied to the discovery of novel peptides from the brain. *Anal. Chem.*, **79**, 15–16; 18–21.
Tirosh,Y. *et al.* (2012) Short toxin-like proteins abound in Cnidaria genomes. *Toxins*, **4**, 1367–1384.
Tirosh,Y. *et al.* (2013) Short toxin-like proteins attack the defense line of innate immunity. *Toxins*, **5**, 1314–1331.
Varshavsky,R. *et al.* (ed.) (2007) *Algorithms in Bioinformatics*. Lecture Notes in Computer Science. When Less is More: Improving Classification of Protein Families with a Minimal Set of Global Features. Vol. 4645, Springer Berlin Heidelberg, pp. 12–24.
Veenstra,J.A. (2000) Mono- and dibasic proteolytic cleavage sites in insect neuroendocrine peptide precursors. *Arch Insect Biochem. Physiol.*, **43**, 49–63.
Wilkins,M.R. *et al.* (1999) Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol.*, **112**, 531–552.
Zhang,Z. and Zhu,S. (2012) Comparative genomics analysis of five families of antimicrobial peptide-like genes in seven ant species. *Dev. Comp. Immunol.*, **38**, 262–274.
Zhao,Z.D. *et al.* (2009) RBF-SVM and its application on reliability evaluation of electric power system communication network. *Mach. Learn. Cybern.*, **2**, 1188–1193.