doi:10.1016/j.jmb.2007.02.106 J. Mol. Biol. (2007) 369, 553–566

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Novel Families of Toxin-like Peptides in Insects and Mammals: A Computational Approach

Noam Kaplan*, Noa Morpurgo and Michal Linial

Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University, Jerusalem, Israel

Most animal toxins are short proteins that appear in venom and vary in sequence, structure and function. A common characteristic of many such toxins is their apparent structural stability. Sporadic instances of endogenous toxin-like proteins that function in non-venom context have been reported. We have utilized machine learning methodology, based on sequence-derived features and guided by the notion of structural stability, in order to conduct a large-scale search for toxin and toxin-like proteins. Application of the method to insect and mammalian sequences revealed novel families of toxin-like proteins. One of these proteins shows significant similarity to ion channel inhibitors that are expressed in cone snail and assassin bug venom, and is surprisingly expressed in the bee brain. A toxicity assay in which the protein was injected to fish induced a strong yet reversible paralytic effect. We suggest that the protein may function as an endogenous modulator of voltage-gated Ca^{2+} channels. Additionally, we have identified a novel mammalian cluster of toxin-like proteins that are expressed in the testis. We suggest that these proteins might be involved in regulation of nicotinic acetylcholine receptors that affect the acrosome reaction and sperm motility. Finally, we highlight a possible evolutionary link between venom toxins and antibacterial proteins. We expect our methodology to enhance the discovery of additional novel protein families.

*Corresponding author

Keywords: ANLP; OCLP; short proteins; genome annotation; Raalin

Introduction

Animal peptide toxins (APT) are short proteins that appear in animal venom and are aimed at inflicting harm to the organism on which the venom acts. APTs are extremely varied in terms of function and include ion channel inhibitors (ICIs), phospholipases, protease inhibitors, disintegrins, defensins and other biological groups. Even specific groups of ICIs, which inhibit the same target channels, often vary in sequence and structural fold.[1]

In light of the sequential, structural and functional diversity of APTs, it seems unfeasible to find a global characterization of APTs by standard automatic classification methods. Although the most obvious demand for APTs is their sorting to the secretory pathway, where they must mature to be secreted to the extracellular mellieux, this property is clearly not unique to these proteins. However, in spite of their diversity, many APTs do share a common structural feature: multiple disulfide bridges help maintain a rigid backbone, conferring high stability.[2] This property, in conjunction with multiple post-translational modifications,[3] is hypothesized to help maintain the toxin's functionality while traveling through the recipient's hostile bloodstream.

Many of the functions and structures of APTs are not exclusive to APTs. Instances of APT and APT-like proteins that act in non-venom contexts have been reported.[4–10] One of the most striking examples is that of Lynx1 and SLURP-1.[4,5,11] These are human proteins that not only possess similarity to snake α-neurotoxins, but also modulate nicotinic acetylcholine receptors (nAChRs) as do α-neurotoxins. Mutation in the gene of SLURP-1 causes Mal de Meleda disease, a skin disease that results from an improper activation of TNF-α.[4] Lynx1 has

Present address: N. Kaplan, Department of Computer Science and Applied Mathematics, Weizmann Institute, Rehovot, Israel.

Abbreviations used: ANLP, α-neurotoxin-like protein; AUC, area under curve; APT, animal peptide toxins; nAChR, nicotinic acetylcholine receptor; ICI, ion channel inhibitors; OCL, ω-conotoxin-like.

E-mail addresses of the corresponding author: kaplann@cc.huji.ac.il; noam.kaplan@weizmann.ac.il

recently been shown to affect neuronal activity and survival in the CNS.[12] These reported instances suggest that, in evolutionary terms, many toxins are homologs of endogenous non-venom proteins and may have been recruited to act in a venom context[13] or *vice versa*. Considering these findings, it is conceivable that there exist additional APT-like proteins, which adopt structural and functional principles that are similar to those of APTs. Specifically, we conjecture that there exist undiscovered endogenous APT-like ion channel modulator peptides.

In the first part of our study we constructed a computational classifier that attempts to find a general characterization of APTs, and thus given a sequence predicts whether the protein is APT-like. Next, we applied the classifier to identify putative novel APT and APT-like proteins. When applied to 10,157 predicted protein sequences from the recently sequenced honey bee (*Apis Mellifera*) genome,[14] novel APT-like sequences were identified, including homologs from non-venomous species. This prompted us to search a non-venomous organism for APT-like proteins. In a recent milestone paper,[15] the mouse FANTOM consortium produced a large set of full-length cDNA sequences, producing an extensive representation of the mouse transcriptome. Amongst these were 5154 putative novel proteins to which we have applied our classifier, resulting in an additional novel family of mammalian APT-like proteins.

Following our computational discoveries, we have conducted experiments in order to validate and elucidate the functions of the novel APT-like proteins.

Results

Classifier evaluation

A computational classifier was trained on a set of known ICIs as described in Materials and Methods. ICIs are only a subset of all APTs. The reason ICIs were used for training rather than APTs is that the definition of structurally stable APTs (or APT-like proteins) is often confusing. For example, many proteins annotated as toxins (bacterial toxins, for example) may not naturally belong to this category. Furthermore, we wish to avoid bias that may be introduced from manual selection of the instances in the training set. Thus, we train the classifier on the set of annotated ICIs with the hope that the classifier will generalize to include additional groups of APTs. This expectation is reasonable, since ICIs by themselves are extremely variable in sequence and structure.

Most state-of-the-art functional classification methods use position-specific information (e.g. evolutionary conserved positions) in order to find sequence motifs that are common to functional groups. Due to the large variation of APTs in sequence and structure, this commonly used approach is unsuitable in the case of APTs. Our classifier uses 545 general sequence-derived features that we had speculated to possibly be related to APT structural stability. The features were constructed so that they would reflect the frequency, distribution, packing and crude localization of cysteine residues within the sequence. However, the features were not restricted to cysteine-related features and were applied to all 20 amino acids. See Materials and Methods for a full description of the features.

The classifier was evaluated by a threefold cross-validation classification test. Area Under Curve (AUC) is an established measure of performance in this test, with AUC = 1 indicating perfect success. The classifier obtained a mean AUC of 0.9934 (standard deviation = 0.0026). The high performance in the cross-validation tests suggests that the classifier is indeed able to capture a robust phenomenon.

Although the classifier performs well on the cross-validation test, it is important to characterize what exactly the classifier has learned. For example, since the training set contained only ICIs as positive instances, we would like to assess whether the classifier will be able to detect only ICIs or other unrelated APT or APT-like groups as well. Generally, it would be a mistake to interpret the classifier's hypothesis as an explanation of an observed phenomenon. This is due to the fact that there is no preliminary reason that the characterization that the classifier has produced will be related in any way to a specific phenomenon. However, there is some indication that our classifier's hypothesis is related to cysteine-mediated structural stability: Amongst all 545 sequence-derived features, the classifier repeatedly identified the most dominant feature to be the frequency of cysteine residues within the sequence.

In order to assess the predictions made by the classifier, we applied the classifier to a non-redundant set of all 29,554 SwissProt proteins shorter than or equal to 150 aa (excluding the ICIs that were present in the training set). A histogram of the predictions is shown in Figure 1(a). A total of 997 proteins (3.37%) were predicted positive by the classifier (Supplementary Data Table 1). In order to assess whether these are false positive predictions, we tested the set of positive predictions for enrichment in biological functional categories. For biological functional categories we used the manually validated UniProt keyword annotations and the predicted InterPro motif groups associated with the proteins. The results (Table 1) show that the most highly over-represented groups are APT-related. Considering that the training process was performed only on ICIs, it is remarkable to note that several different APT-related functional categories are detected (ICIs, phospholipases, disintegrins, protease inhibitors, etc.). Note that although secreted proteins are enriched, only 13.4% of all secreted proteins are predicted positive, indicating that the classifier does not simply predict all short secreted proteins to be positive. From the score

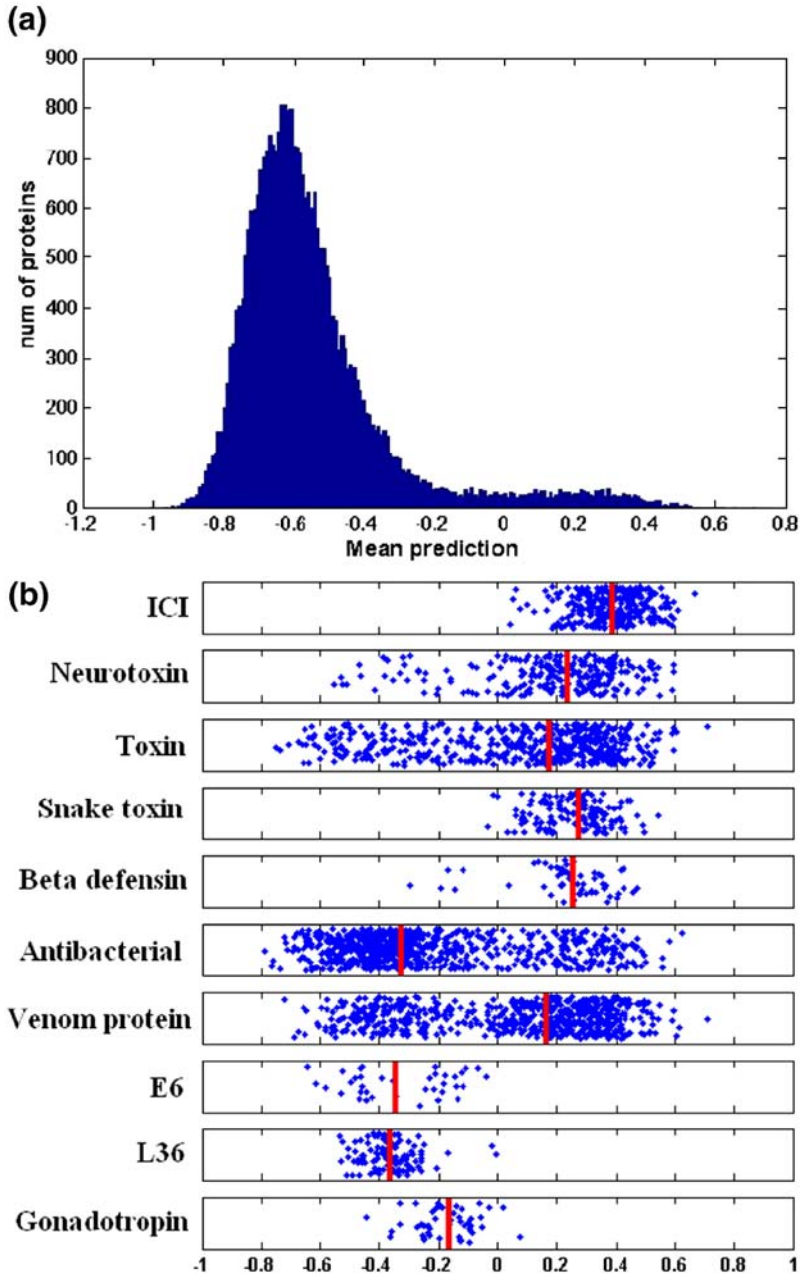

Figure 1. Distribution of prediction scores. (a) Score distribution on a non-redundant set of all 29,554 SwissProt proteins shorter than or equal to 150 aa. (b) Scores of selected biological groups. The horizontal axis represents the mean prediction score. Thick red vertical lines represent median values of each group. The groups ICI, Toxin, Neurotoxin and Antibacterial are based on UniProt keywords. All groups except the top one (ICI) include only proteins that were not part of the training set. The groups ICI, Snake toxin, Neurotoxin and Beta defensin receive mostly positive scores. The Toxin and Venom protein groups tend to be positive but the separation is weaker. The Antibacterial group is mostly negative, but there is clearly a significant portion of positive instances (note that Beta defensin is a subset of this group). The E6 (E6 early regulatory protein), L36 (ribosomal protein L36) and Gonadotropin groups are known to be cysteine-rich but are clearly predicted negative.

distributions of selected biological groups (Figure 1(b)), it is apparent that although most toxins obtain positive scores, many do not. This corresponds with the fact that many toxins (as defined by UniProt) do not belong to the class of structurally stable APTs discussed. Reassuringly, there are specific groups of toxins, such as neurotoxins and snake toxins, which obtain high scores. We have noticed that many false negative predictions occurred in cases where the APT is composed of an extremely long (>60 aa) preprotein with an extremely short (<10 aa) active peptide. In addition to toxins, it is apparent that various antibacterial groups are over-represented. Figure 1(b) shows that although antibacterial proteins mostly receive negative prediction scores, certain groups such as β-defensins are generally predicted positive. This corresponds with previous observations on structural and functional similarities between certain classes of antibacterial proteins and APTs.[9,16] One over-represented biological group that was suspected initially as false positives is that of the metallothioneins. Metallothioneins are ubiquitous cysteine-rich proteins that have been suggested to possess a variety of functions including zinc homeostasis and anti-oxidative effects. The full range of functions of these proteins remains unknown. There is no evidence of metallothionein-like toxins, and the high number of cysteine residues is used in the coordination of heavy metals rather than in the forming of disulfide bonds. However, antibacterial activity of a metallothionein protein expressed in housefly larvae has been reported recently,[17] possibly suggesting that the classification of metallothioneins as incorrect predictions may need to be reconsidered. Figure 1(b)

Table 1. Statistically enriched groups amongst the positive predictions

Biological group[a]	Positive	Total	*P*-value
Toxin	299	541	6.72E-303
Neurotoxin	172	242	2.896E-197
Snake toxin	119	137	1.016E-154
Signal peptide	379	2824	3.996E-134
Postsynaptic neurotoxin	76	99	3.356E-89
Phospholipase A2	83	171	1.196E-72
Knottin	62	81	3.832E-72
Serine protease inhibitor	105	324	1.128E-70
Acetylcholine receptor inhibitor	60	78	5.64E-70
Defensin	77	149	5.76E-70
Protease inhibitor	112	405	3.004E-67
Beta defensin	50	57	8.68E-64
Plant defense	69	132	6.88E-63
Antimicrobial	142	759	3.228E-62
Metal-thiolate cluster	64	123	5.52E-58
Antibiotic	125	656	3.44E-55
Snake cytotoxin	38	39	8.4E-53
Lipid degradation	71	188	3.084E-52
Gamma thionin	39	46	5.44E-48
Metallothionein superfamily	41	53	2.608E-47
S locus-related glycoprotein 1 binding pollen coat	34	35	7.4E-47
Whey acidic protein, core region	44	71	6.4E-44
Cardiotoxin	29	29	3.752E-40
Cyclotide	27	29	3.06E-35
Cadmium	28	34	3.784E-33
Gamma purothionin	26	29	9.32E-33
Vertebrate metallothionein	29	40	1.98E-31
Proteinase inhibitor I2, Kunitz metazoa	33	61	1.132E-29
Disintegrin	23	25	2.136E-29
Cell adhesion	32	62	7.8E-28
Calcium	70	409	1.28E-26
Cyclotide, bracelet	19	19	2.832E-25
Proteinase inhibitor I12, Bowman-Birk	23	33	7.16E-24
Fungicide	45	194	4.56E-22
Mammalian defensin	23	40	5.8E-21

[a] Biological groups were defined either according to UniProt keywords or InterPro entries.

shows the prediction results of three groups of short cysteine-rich proteins that do not function as APTs or as APT-like: gonadotropin, L36 ribosomal protein and E6 early regulatory protein families. These groups generally receive negative scores, suggesting that a large amount of cysteine residues is not sufficient for differentiating between APTs and non-APTs. In summary, the classifier is apparently able to correctly produce a non-trivial characterization of APT and APT-like proteins.

Prediction on honey bee proteins

Recently the honey bee genome has been assembled and annotated.[14] We have applied our classifier to all 10,157 protein sequences that were predicted from the honey bee genome. A total of 19 honey bee proteins were predicted to be APT-like proteins by our classifier (Supplementary Data Table 2). Of these, eight are predicted to possess a signal peptide, as expected of APTs. We discuss in detail the four highest scoring sequences.

Apamin and MCDP

Two of the proteins are well-known bee venom toxins, apamin and MCDP, both of which function as K^+ ICIs[18,19] (note that MCDP performs additional functions). State-of-the-art methods for motif finding and fold recognition, such as InterProScan[20] and PHYRE,[21] respectively, failed to detect both of these sequences as toxins. These two predictions suggest that the classifier is able to assign function to proteins beyond the capacity of structure-based or motif-based similarity tools.

OCLP1

The two remaining protein sequences are putative proteins, which we refer to as OCLP1 (ω-conotoxin-like protein 1) and Raalin (after *ra'alan*, the Hebrew word for toxin), respectively. OCLP1 is a 74 amino acid residue sequence that possesses a signal peptide followed by a cysteine-rich domain of ~30 amino acid residues and an unstructured tail (Figure 2(a)). An InterProScan search[20] for known sequence motifs indicates that this protein is related to the assassin bug toxins Ptu1, Ado1 and Iob1. These three proteins were isolated from the saliva of the assassin bug (*Reduviid*) species, and were shown to function as voltage-gated Ca^{2+} ICIs and to possess a fold similar to that of ω-conotoxins.[22–24] Multiple sequence alignment of OCLP1 with these assassin bug toxins (Figure 3(b)) strengthens the notion of homology of these proteins. The multiple sequence alignment shows conservation of the six cysteine residues and of positions G5, T20, Y25, A26, N27 and R28. It has been suggested in the case of the assassin bug toxin that positions K13, Y25 and R28 are functionally important.[22,23,25] However, K13 is replaced by an aspartic acid in OCLP1, raising the possibility for interaction with an alternative ion channel as a target. We constructed a model of the tertiary structure of OCLP1, modeled after the solved structure of Ado1 (PDB 1LMR) (Figure 2(b)). The side-chains of the amino acids in positions 25–28, which are fully conserved in OCLP1, and the three assassin bug ICIs are exposed at the tip of the protein structure, possibly constituting part of the interface with the ion channel. The PHYRE fold recognition server predicts OCLP1 to possess a fold similar to that of ω-conotoxins and the assassin bug toxin.

Experimental expression evidence is found for OCLP1 in dbEST.[26] Remarkably, the EST originates from the bee brain rather than the venom sac, which is located at the bottom of the abdomen. In order to validate expression of OCLP1 in the brain, we performed RT-PCR on RNA extracted from honey bee head and brain. OCLP1 showed a strong expression in the brain (Figure 5(a)). Searching for additional cDNA evidence, homologs were found in several insects and in *S. mediterranea*, a flatworm (Figure 3(b)). The cDNA were obtained from head, whole adult, whole larvae, wing disc and antennae tissues. Of special interest are the *Anopheles gambiae* and *Aedes aegypti* homologs, which both possess

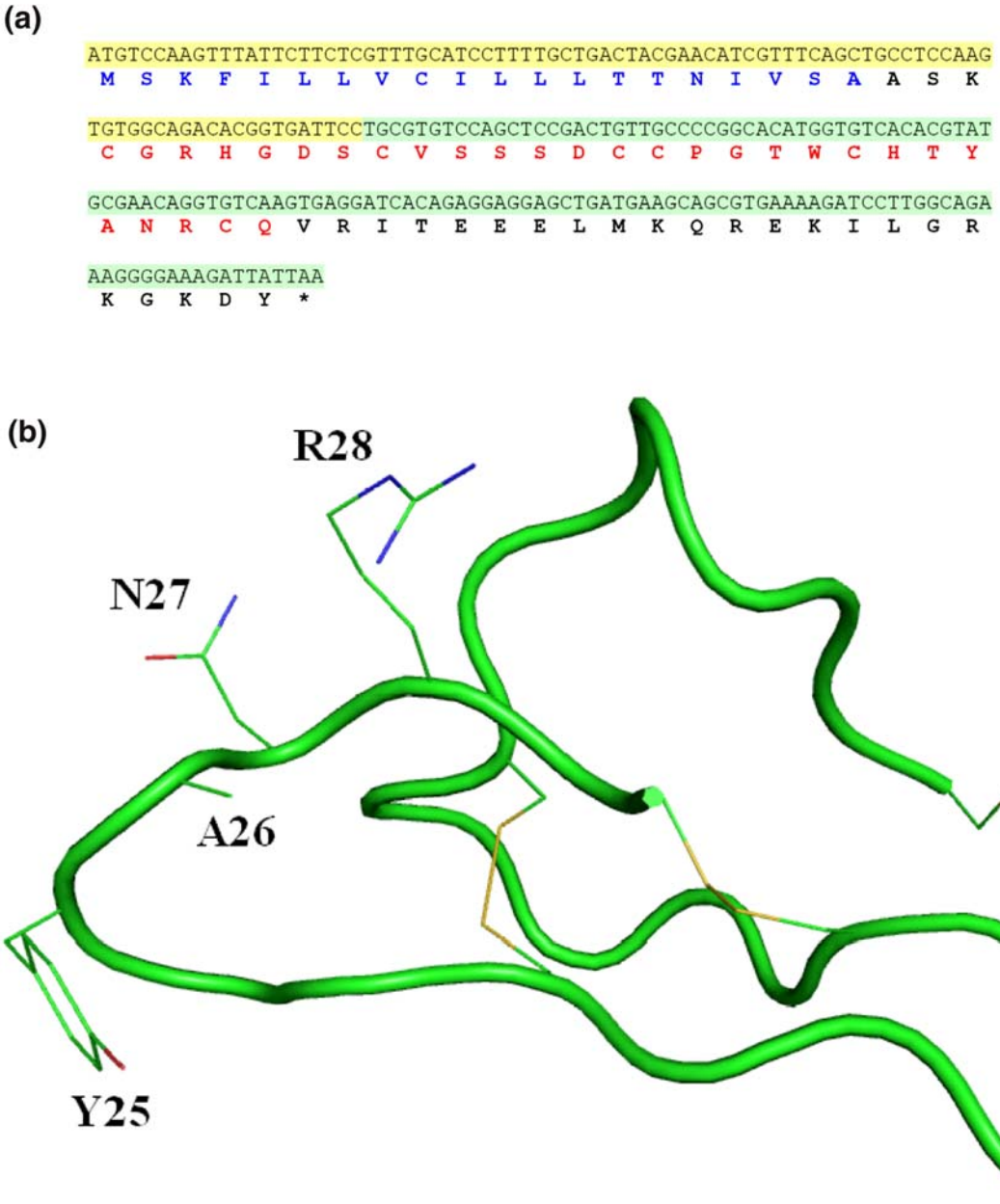

Figure 2. Sequence and predicted structure of OCLP1. (a) Nucleotide and amino acid sequence of OCLP1. Yellow and green backgrounds represent the first and second exons. Blue amino acids represent the putative location of the signal peptide (predicted by SignalP). Red amino acids represent the mature peptide and black letters represent an extended unstructured tail. Note the exon positioning in which the first exon ends just before the second cysteine of the putative mature peptide. (b) Structural model of OCLP1. Side-chains are shown for the six conserved cysteine residues (disulfide bonds appear in yellow) and for the conserved positions 25–28 that are unique to OCLP1 and the assassin bug toxins. Model was created using SDPMOD (homology modeled after 1LMR).[54]

signal peptides and are suspiciously long (335 and 372 aa, respectively). Interestingly, both homologs contain multiple repeated occurrences of ω-conotoxin-like (OCL) sequences (five in *A. gambiae* and four in *A. aegypti*). Remarkably, in those species for which genomic data are available, we observed that the locations of the exons were identical relative to the position of the putative OCL peptides, with a splice site located just before the second cysteine of the OCL repeat (compare Figures 2(a) and 3(a)). Multiple sequence alignment of OCLP1, its homologs and various other OCL proteins shows that apart from the six conserved cysteine residues, some positions show partial conservation, but only positions G5, Y/F25 and R/K28 are highly conserved (Figure 3(b)). InterProScan and PHYRE predict all repeats to possess an ω-conotoxin fold.

Raalin

Raalin is a short sequence of 29 aa. Since the predicted open reading frame (ORF) does not start with a methionine, it was suspected to be a truncated protein sequence. We have identified several

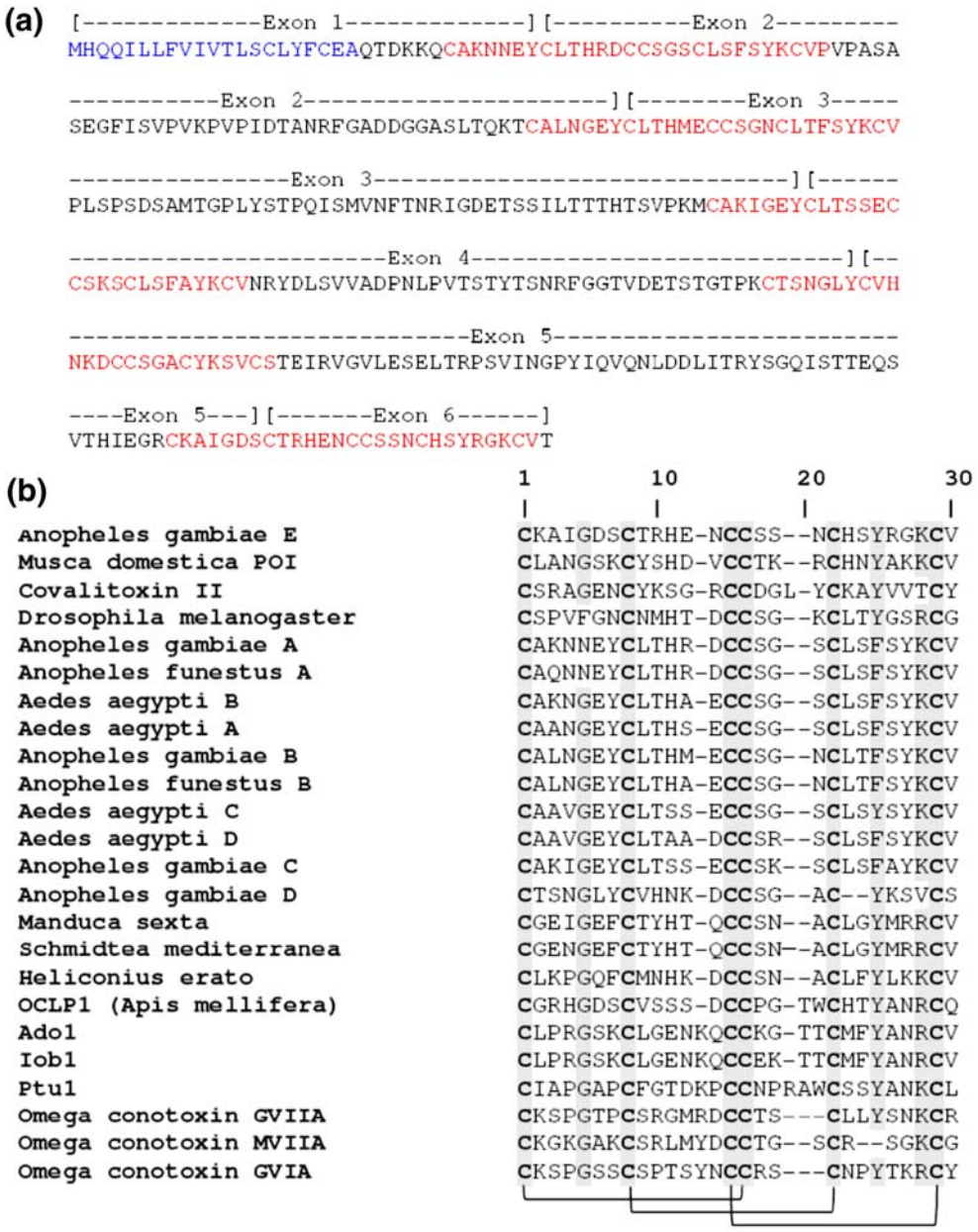

Figure 3. OCLP1 homologs. (a) Amino acid sequence of the *A. gambiae* OCLP1 homolog. Blue amino acids represent the putative location of the signal peptide (predicted by SignalP). Red amino acids represent the locations of the OCL repeats. Note that the exons are positioned similarly relative to the OCL repeats, with each of the exons ending before the second cysteine of an OCL repeat (see Figure 2(a)). (b) Multiple sequence alignment of OCL proteins. A–E indicates repeats within the OCLP1 protein homologs. Highly conserved positions are highlighted. Cysteine residues appear in bold. Disulfide connectivity is shown beneath the alignment. OCLP1 homologs are noted in species names only. A–E indicates OCL repeats. Only the OCL region is shown. Note the YANRC sequence, which is shared only by OCLP1, Ado1, Ptu1 and Iob1.

homologs from insect cDNA sequences (Figure 4). Amongst these is a 108 aa *Drosophila melanogaster* homolog. Reassuringly, the *Drosophila* homolog possesses a signal peptide, which is followed by a region of high similarity to Raalin, supporting the notion that the honey bee Raalin sequence is indeed a sequence missing its signal peptide. As for localization of expression, the *A. gambiae* homolog was found in the head and the *Bombyx mori* homolog was found in the brain. In all putative homologs, the region of similarity is exclusive to the short cysteine-rich region where the putative peptide is located. No evidence of functional or structural similarity to known APTs was found by structure and sequence prediction tools.

Prediction on mouse proteins

FANTOM is a newly available resource for the mouse transcriptome, with thousands of previously

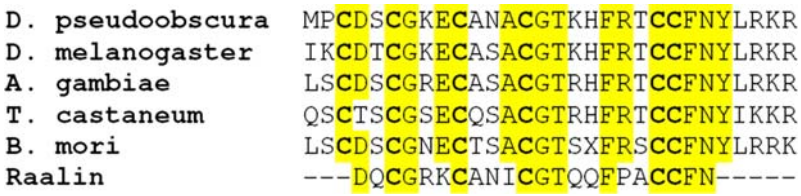

Figure 4. Multiple sequence alignment of Raalin and putative orthologs. Positions that are identical in at least five sequences are highlighted. Note that this alignment shows only the putative mature peptide region. Homologs are noted in species names only.

unreported transcripts.[15] Amongst these are 5154 sequences that have been identified as novel proteins. A total of 16 of the 5154 novel FANTOM sequences were predicted by our classifier to be APT-like (Supplementary Data Table 3). Of these, 14 possess a signal peptide. We elaborate about one of these sequences, a 111 aa sequence that we refer to as mANLP1 (mouse α-neurotoxin like protein 1). mANLP1 possesses a signal peptide and is identified by both InterProScan and PHYRE as "snake toxin-like" (also known as the three finger toxin fold). We have identified several mouse and human ANLP homologs that are clustered in the genome, constituting previously unknown gene clusters (Table 2). The gene cluster consists of several gene products that are related to the Ly6-uPAR family. All genes in the cluster posses a signal peptide but lack a GPI anchor that is characteristic for other members of the Ly6-uPAR family. Current expression evidence shows that ANLP genes are mostly expressed in the testis. Some gene transcripts were also detected in lung and brain tissue.

OCLP1 expression and toxicity assays

OCLP1 was expressed in Escherichia coli, purified and refolded (see Materials and Methods; Figure 5(b) and (c)). Next, in order to test whether OCLP1 possesses toxic effects, we conducted classical bioassays of toxicity.

First, to test the toxicity in an insect context, 4 μl of the purified compound was injected into laboratory-bred blowfly larvae (*Sarcophaga faculata*) as described.[27] More than ten larvae were injected with no significant effect on apparent movement or on hatching (24 h after injection). We concluded that the

Table 2. Proteins of the human and mouse ANLP loci

Genes	Name[a]	GenBank / gene symbol (no. of sequences)	Alt. transcript	Location	SP[c]	UniProt accession (aa)	Expression evidence
Mouse chromosome 9qA4							
1	mANLP1	Gm846 (12)	AK144787	chr9:35,319,955-35,439,989	Y	Q3UW31 (111) Q3UMN1 (149) Q6UY27 (113)	Epididymis, lung
2	Seminal vesicle protein 7	9530004K16Rik caltrin, *SVS7* (13)	AF134204	chr9:35,357,257-35,361,517	Y	Q9R018 (99) SVS7_MOUSE	Epididymis, brain
3	mANLP2	D730048I06Rik (15)	AK033813	chr9:35,537,721-35,539,783	Y	Q9CQB8 (106) Q3UW50 (83)	Mammary gland, epididymis
4	mANLP3	9230110F15Rik (11)	AK020329	chr9:35,588,000-35,593,821	Y	Q9D262 (117)	Epididymis
5	ANLP4	AK136639 (2)	AK033758	chr9:35,597,526-35,599,607	Y	Q8CC74 (168)	Epididymis
6	Pseudogene	ENSMUSP00000048154		chr9:36,136,495-36,183,148	N	Q6UY27 (99)	Predicted
7	mANLP5	LOC434396 (2)	AK136744	chr9:36,282,074-36,291,634	Y	Q3UW02 (101)	Epididymis sperm, testis
8	Secreted seminal-vesicle Ly-6 protein 1	Gm191, *SSLP1*, A630095E13Rik (13)	AK144443	chr9:36,385,426-36,388,273	Y	Q3UN54 (99)	Seminal vesicles
9	Acrosomal protein SP-10 (precursor)	*ACRV1*, Msa63 (14)	AK030129	chr9:36,442,921-36,448,515	Y	ASPX_MOUSE (261) Q9DAM6 P50289	Spermatid, testis epididymis
Human chromosome 11q23.4							
1	Acrosomal vesicle protein 1 isoform (precursor)[b]	*ACRV1* (12)	11 alt. splicing	chr11:125,047,440-125,056,152	Y	P26436 (265) ASPX_HUMAN	Acrosome, testis
2	hANLP1	*PATE* (3)	AF462605	chr11:125,121,398-125,124,952	Y	Q8WXA2 (126)	Prostate, testis, brain
3	hANLP2	LVLF3112 (10)	C11orf38	chr11:125,152,446-125,152,714	Y	Q6UY27 (113)	Secretion
4	hANLP3	AK123042 (11)	FLJ41047	chr11:125,208,421-125,215,174	Y	(95)	Prostate

[a] Mouse and human ANLPs are marked as mANLP and hANLP, respectively. Gene symbols according to HUGO are in italics.
[b] *ACRV1* is the only sequence in these loci that has an extended sequence and does not have a snake-toxin-like motif.
[c] Sequences predicted to possess a signal peptide (Y, yes; N, no).

OCLP1 in its expressed form is inactive towards the blowfly larvae, suggesting that OCLP1 is inactive in affecting movement in this biological setting.

Second, to test the biological toxic effect in vertebrate context, a standard assay is to monitor alteration in movement of freshwater aquarium fish. We used male Guppy fish (*Gambusia affinis*) (2–2.5 cm each). OCLP1 was injected to nine fish and their activity was monitored by video for 1 min every hour. For negative controls, four fish were injected with buffer and three fish were injected with unfolded reduced OCLP1. For a positive control, one fish was injected with hydralysin, a pore forming toxin from Hydra.[28] The results are summarized in Figure 5(d). We observed in the fish injected with OCLP1 a reversible reduction in locomotor ability: the fish moved their fins weakly and remained on the floor of the aquarium for a few hours (6–10 h), compared to the negative control fish, which swam constantly during this time. After 8 h most (6/7) of the OCLP1 injected fish completely recovered (for one fish the damage was irreversible and it died after 16 h). For two fish (out of the total of nine), the impairment in locomotion was not significant after 1 h. None of the negative control fish had any impairment in locomotion. The positive control fish was paralyzed 8 h after injection and finally died.

Discussion

Computational characterization of APT and APT-like proteins

We have created a computational classifier that seems able to capture the non-trivial notion of being APT or APT-like. This was confirmed both by cross-validation and evaluation of predictions on a large test set. Reassuringly, we find that even though the classifier is trained only on ICIs, it is able to detect

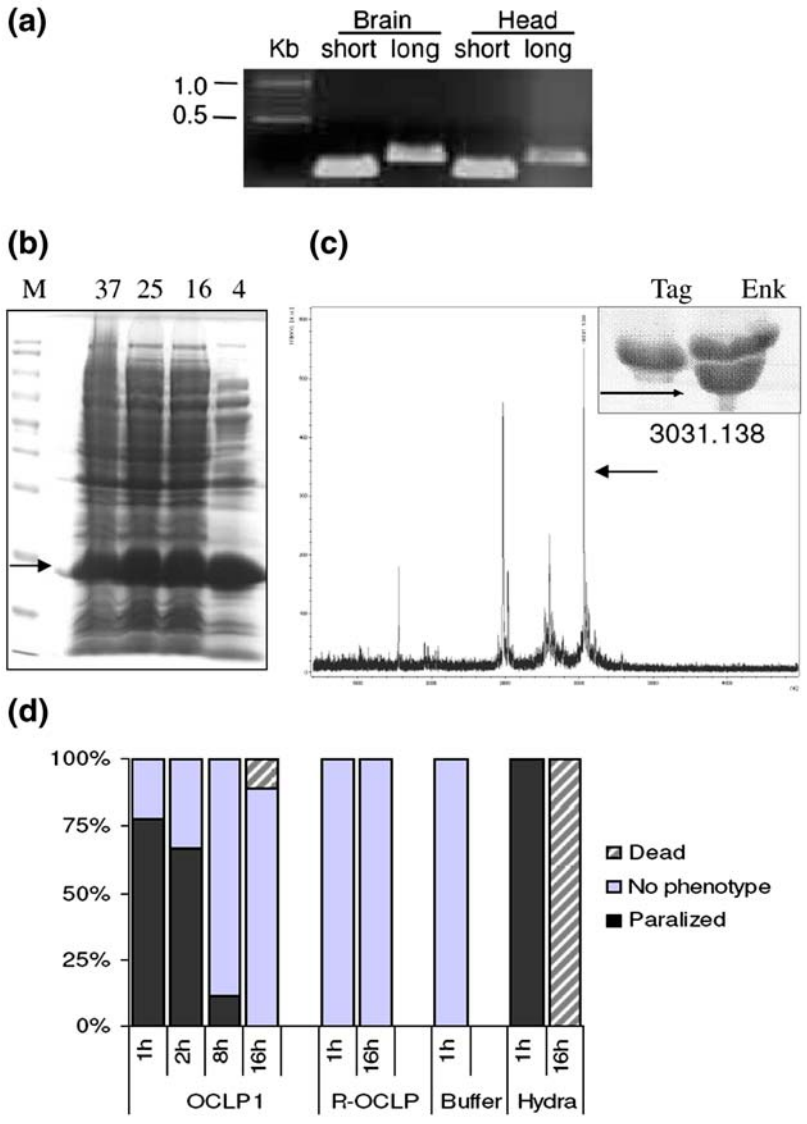

Figure 5. Expression, purification and functional assay of the honey bee OCLP1. (a) Expression of OCLP1 RNA. Products of RT-PCR using total RNA extracted from bee brain and head following separation on 1.5% agarose gel are shown. The short version (169 nt) is the OCLP1 mature form and the long version (240 nt) is the full-length transcript. The similar expression level in head and brain indicates that OCLP1 is expressed in the brain rather than tissues outside the brain, such as the salivary gland. (b) A 28 amino acid residue OCLP1 peptide was expressed as a fusion protein with a cellulose binding domain (CBD) tag. High levels of expression were achieved when cells were grown at varying temperatures (37, 25 or 16 °C, expressed for 12 h; and 4 °C for 30 h). Equivalent of 100 µl of growing culture from the soluble fractions was loaded on 12% SDS-gel. OCLP1 under these conditions is expressed at high levels (~10–20 mg/l). M, Molecular weight protein marker. Temperatures of growth are marked as 37, 25, 16 and 4. (c) Following refolding, 4 µg of the tagged protein were loaded on 15% SDS-tricine gel before (left) and after (right, marked by arrow, inset) cleavage with entrokinase. MALDI-TOF Mass spectrometry spectra of the OCLP1 peptide is shown. A major band of the expected size (3013.14 daltons) is identified from the low molecular weight peptide extracted from the gel. (d) Toxicity assay results. Summary of the swimming ability at time interval after the injection (1, 2, 8 and 16 h) is shown. OCLP1 label indicates injection of purified folded OCLP1 (nine fish). R-OCLP indicates injected OCLP1 without the refolding step (three fish). Additional controls used are buffer injection (four fish) and purified Hydra toxin (one fish).

other groups of non-related APT and APT-like proteins. This finding suggests that this functional super-category, of being APT or APT-like, is not an artificial category that is a union of various smaller functional categories, but rather a genuine biological group that possesses its own unique characteristics. The training of the classifier suggests that a high amount of cysteine residues is indeed crucial for most proteins of this category, but this feature is evidently not sufficient to define this group. The successful computational characterization of this group enables the detection of novel protein families that are APT or APT-like but do not share sequence or structural similarity with any known proteins.

Mutual evolution of APTs and antibacterial proteins

Many antibacterial proteins can be thought of as functional analogs of APTs, since they are also secreted and aimed at harming foreign organisms in a non-specific manner. Therefore, it is interesting to note the consistent prediction by the classifier of some groups of antibacterial proteins, such as β-defensins, γ-thionins and WAP proteins, to be APT-like. These predictions are supported by previous studies that have observed structural similarities between some of these classes of antibacterial proteins and APTs.[9,10,16,29–31] Remarkably, a recent article[32] reports that crotacetin, a rattlesnake venom protein homologous to convulxin, exhibits antibacterial activity. While convulxin is able to activate platelets and induce their aggregation by acting *via* receptor binding mode, crotacetin acts in an unknown manner against Gram-positive and Gram-negative bacteria.

Venomous animals are subject to an evolutionary pressure to maintain their venom effectiveness. In cone snails, snakes and spiders, extensive duplications and mutation events in venom proteins have been observed,[33,34] resulting in a rich mixture of peptide variants that cover a broad range of specificities towards their preys. Similarly, antibacterial peptides must cope with the fast evolution of the microbial world. The functional groups detected by our classifier are consistent with the concept of APT-like proteins as a shared evolutionary solution for a cellular defense mechanism. Altogether, we suggest these findings to indicate a close evolutionary relationship between APTs and antibacterial proteins, including possible evolutionary recruitment of antibacterial proteins to act as venom toxins or *vice versa*.

Discovery of novel putative toxin-like neuropeptides in insects

We have detected two putative APT-like bee sequences of hypothetical proteins, OCLP1 and Raalin. Several evidences provide strong support that OCLP1 is APT-like: it possesses a signal peptide, shares sequence similarity with voltage-gated Ca^{2+} ICIs and is predicted by independent methods to be OCL. Remarkably, this protein is expressed in the brain of the honey bee. Still, some venom toxins are known to be additionally expressed in non-venomous tissues, including the brain.[7] However, since the bee venom has been studied extensively, it seems unlikely that OCLP1 is a venom toxin. Significant evidence supporting this notion is found in the form of homologs in non-venomous organisms (Figure 3(b)). In two instances, the homologs contain multiple OCL repeats. This form of multiple repeats of a small peptide is a common form for preproteins of several neuropeptides and of APTs.[6] A strong validation for the homology of these proteins is an exact match of exon length and boundaries in these sequences. Although several of the homologs of OCLP1 function as voltage-gated Ca^{2+} ICIs, the *Anopheles gambiae* and *Musca domestica* homologs have been previously suggested to function as inhibitiors of melanization by inhibiting phenoloxidase.[35,36] These functionalities need not necessarily contradict, the biochemical mode of action of these proteins is yet unknown. Multiple sequence alignment suggests that OCLP1 is most similar to the assassin bug ICIs, sharing a unique five amino acid sequence (YANRC) with these proteins (Figures 2(b) and 3(b)), two of which have been suggested to be critical for ICI function. In order to further probe into the function of OCLP1, we expressed the protein and conducted standard toxicity assays in blowfly larvae and fish. While the larvae were not affected by the protein, the fish exhibited a significant yet reversible paralytic effect, which was expressed in seriously damaged locomotion ability. Although this assay cannot pinpoint a specific molecular mechanism, it does support our classification of OCLP1 as toxin-like. We are currently performing experiments to elucidate the molecular function of OCLP1.

Raalin is a 29 amino acid residue APT-like fragment with homologs in several insects. None of them show any similarity to proteins of known function. Although no known ESTs were found for the bee sequence, in homologs that have data on expression localization, the expression is localized to the brain and head. All full-length homologs possess signal peptides. All homologs share a short cysteine-rich region of similarity, while the sequence segments that are not included in the putative mature peptide are not conserved. This is typical for many secreted proteins that undergo post-translational cleavage. It is likely that Raalin does not function as a venom toxin, due to its existence in non-venomous insects and its EST localization to the head and brain.

Although further experimental research is needed in order to determine the exact biological functions of these two groups of peptides and their expression localization, we hypothesize that these might be only a few of several toxin-like neuropeptides that may exploit the same mechanisms as do APTs. We are currently performing experimental assays in order to study the localization and molecular functions of OCLP1 and Raalin.

Discovery of a novel mammalian APT-like gene cluster

We have identified eight mouse proteins that are predicted to possess a snake toxin-like fold (also referred to as the snake α-neurotoxin fold or three finger toxin fold) and whose genes are clustered in the genome to a single locus in chromosome 9. In addition, we identified an analogous human locus in chromosome 11 containing only four such genes. We collectively refer to these proteins as ANLPs (Table 2). Both ANLP loci contain a previously known gene coding for SP-10.[37] Additionally, the human locus contains the gene that codes for PATE.[38] The snake toxin-like fold is shared by both snake α-neurotoxins and several mammalian proteins. Mammalian proteins that possess this fold are generally divided into (a) extracellular domains of receptors such as activin and (b) short secreted proteins such as Lynx1 and SLURP-1, some of which have been shown to modulate nAChRs. Interestingly, the known nAChR modulators including SLURP-1 and Lynx1 also constitute a gene cluster, which is distinct from the ANLP cluster that we have identified. In terms of known expression, all ANLP proteins are derived from sequences identified in the testis, although some have also been found to be expressed elsewhere (including lung and brain) (Table 2).

Some indication for the possible function of ANLP proteins is available from the SP-10 and PATE genes. SP-10 has been previously considered to be a promising contraceptive vaccine immunogen, but its molecular function remains unknown. SP-10 has been shown to participate in the acrosomal reaction and has been shown to be associated with the membrane either by a lipid anchor or by interaction with an unidentified membrane protein.[39] Similar evidence of membrane interaction was found for PATE, whose function is also unknown.[38] It is interesting to note that nAChRs, specifically the α7 subunit, have also been identified as crucial participants in the acrosome reaction.[40] Furthermore, this reaction is inhibited by α-bungarotoxin, a snake α-neurotoxin that inhibits nAChRs. In light of these findings, we hypothesize that the ANLP proteins, including SP-10 and PATE, may function as endogenous modulators of nAChRs that are involved in control of the acrosome reaction and sperm motility. We are currently testing this hypothesis experimentally.

Pharmaceutical implications

Lately, the therapeutic potential of toxins has been realized and has led to the development of toxin-based drugs, with ICI toxins being the most popular targets for development.[41] Defensins and other antibacterial proteins are also attracting attention as potential drugs and pesticides.[42] Furthermore, the high stability of the backbones of these molecules makes them appealing for drug-design. One example of a toxin-based drug that is already on the market is Ziconotide,[43–45] which is a synthetic form of MVIIA ω-conotoxin, a voltage-gated Ca^{2+} ICI from *Conus magus*, which is delivered directly to the patient's central nerve system. Ziconotide is used for treatment of chronic pain. In this respect, the bee protein OCLP1 and its homologs are attractive candidates for further research, as they show high similarity to voltage-gated Ca^{2+} ICIs from both assassin bugs and marine cone snails, and may be natively expressed in the brain. Although we focus in our analysis on a few insect and mammalian protein families, we provide a list of 997 short eukaryotic proteins (Supplementary Data Table 1) that are predicted to be APT-like, many of whom have little or no functional information. We expect our classifier to expand the range of known toxin, toxin-like and antibacterial proteins, thereby contributing potential targets for further pharmaceutical research.

Materials and Methods

Feature construction

The following 545 sequence-derived features were used to transform a given sequence into a vector:

(I) Amino acid frequencies (20 features).
(II) Amino acid pair frequencies (400 features).
(III) Sequence length. Hereby referred to as m (one feature).
(IV) Cysteine binary 5-mers (32 features). Sequence is divided into m-4 amino acid 5-mers. Each 5-mer is translated into a binary 5-mer. Cysteine residues are translated into 1, and the rest of the amino acids are translated into 0.
(V) Polarity binary 5-mers (32 features). Same as in (IV), except that Asp, Glu, Lys, Arg, Asn, Gln are translated into 1 and the rest of the amino acids are translated into 0.
(VI) Amino acid entropy (20 features). A quantitative measure of how each amino acid type is spread in the sequence. For a given amino acid type c, we mark $p_1,\ldots, p_k$ its positions in the sequence. We define $p_0=0$ and $p_{k+1}=m+1$. We define the entropy of c to be: $entropy(c) = -\sum_{i=1}^{k+1}\left(\frac{p_i-p_{i-1}}{m}\right)\log_2\left(\frac{p_i-p_{i-1}}{m}\right)$.
(VI) Circular mean (40 features). A quantitative measure that encodes the relative location and spread of each amino acid type in the sequence. For a given amino acid type c, we mark its positions in the sequence by $p_1,\ldots,p_k$. The feature formalizes the following notion: If the sequence is spread clockwise around the two-dimensional unit circle, we can calculate the mean of the points on the circle that match $p_1,\ldots,p_k$ and define it as the circular mean of c. Formally, we define: $CM(c)=(-2,2)$ if $k=0$ and $CM(c) = \left(\frac{1}{k}\sum_{i=1}^{k}\sin\left(\frac{2\pi(p_i-1)}{m}\right), \frac{1}{k}\sum_{i=1}^{k}\cos\left(\frac{2\pi(p_i-1)}{m}\right)\right)$ otherwise.

Training set

To construct the training set, all sequences of proteins annotated in UniProt[46] as "Ionic channel inhibitor" were obtained. Fragments and proteins longer than 100 amino

acid residues were excluded, leaving 534 ICI sequences. Note that this includes both mature peptides and preproteins. Next, clustering was performed in order to remove redundancy (necessary in order to avoid bias of the cross-validation results). Following this step, 289 proteins remained so that no two proteins share an identity of 80% or more. These proteins constitute the true training instances (the rationale for using only ICIs as true instances is discussed in Results). As for the false instances, these were randomly selected from UniProt. The false instances were generated in three sets: (I) random full-length proteins; (II) random fragments of random proteins, with lengths matching those of the true instances; (III) N-terminal fragments of random proteins, with lengths matching those of the true instances. The protein fragments are intended to avoid length bias, and the random fragments are intended to avoid N-terminal bias. Each of the three sets is twice the size of the set of true instances, a total of 1734 false instances. Following this, clustering is performed to remove redundancy (80% identity). The final training set consists of the union of the false and true non-redundant sets. Note that for each boosted stumps classifier, a separate false set is generated.

It is important to note that for prediction on the honey bee proteins, the sequences of apamin and MCDP (and their homologs) were not included in the training set.

Learning algorithm

The learning algorithm that was used is a meta-classifier based on the boosted stumps algorithm. A decision-stump is a decision-tree that has only one node. The stump classifier finds the best linear separation available by a single feature. In the boosted stumps method, the AdaBoost boosting algorithm[47] is applied to the stump classifier. In order to determine the optimal number of iterations, a parameter-tuning framework was constructed in which, for a given parameter value, the classifier is evaluated by its AUC performance in a threefold cross-validation test, and the parameter value that maximizes the AUC is chosen for the final classifier.

As mentioned earlier, the situation that we face in the APT classification scenario is slightly different from the classical classification problem in the sense that there exists label noise, i.e. we are trying to capture a property (structural stability) that is not well defined. Therefore, it is not clear that training the classifier to fit the training set well would translate into proper generalization, since some small portion of the labels is incorrect. Although some classifiers including AdaBoost are considered relatively resistant to label noise, we take an additional precaution by constructing a meta-classifier as follows: For a given set of true instances, we randomly generate ten sets of false instances (as described in Training set). Next, for each set of false instances we train a parameter-tuned boosted stump classifier. The outputs of all ten classifiers are normalized by the highest positive prediction of each classifier on the training set (respective to each classifier). The prediction of the meta-classifier is the mean average of the predictions of all ten classifiers. Additionally, the meta-classifier provides the standard deviation of the predictions on each sequence as a measure of robustness. We consider a prediction to be a positive prediction (i.e. the protein is APT) if the mean is greater than the standard deviation. By employing this meta-classifier approach we are able to provide a robust hypothesis, which is not biased by any specific set of false instances. Note that in contrast to a classical classification scenario in which we would fit the whole training set (which includes all false instances) as best as possible and therefore possibly err on mislabeled instances, in our method the chance of making a mistake on a specific mislabeled false instance is reduced, since that would require the false instance to be repeatedly chosen for the random false sets of the ten sub-classifiers. For a more formal description of the classifier and training set construction, see Technical notes in the Supplementary Data.

OCLP1 RNA expression assay

RT-PCR was performed on total RNA extracted from head and brain of young honey bees (kindly provided by G. Bloch of the Hebrew University). Oligonucleotide primers were designed to cross an intron/exon to ensure amplification of fully processed RNA. Two pairs were used for the mature OCLP1 (169 nt) and the full-length transcript (240 nt).

OCLP1 short forward: 5′TCATGTCCAAGTTTATTC-TTC3′
OCLP1 short reverse: 5′AGGAGCTCTTAACACCT-GTTCGCA3′
OCLP1 long forward: 5′CTTAATCTTTCCCCTTTC-TGC3′
OCLP1 long reverse: 5′AGGAGCTCTTAACACCTGT-TCGCA3′

Expression and purification of OCLP1

OCLP1 was cloned and expressed in a pET22 vector (Novagen). The OCLP1 was prepared by designed oligonucleotides following *E. coli* codon preference optimization (a total of 28 amino acid residues, mature peptide from the first cysteine). The OCLP1 was fused to CBP-tagged (cellulose binding domain) at the N-terminal (sequence identical to pET38 expression vector, Novagen). All chemicals were at a molecular biology analytical grade including restriction endonucleases and DNA ladders (Promega), isopropyl thio-β-D-galactopyranoside (IPTG) for *E. coli* expression induction (Sigma). Urea was purchased from Merck.

DNA sequencing (ABI automated DNA sequencer, Perkin–Elmer) confirmed the correct reading frame and the OCLP1 sequence. The confirmed pOCLP1 was introduced to *E. coli* BL21(DE3) strain. The cells were grown at varying times and temperatures for expression optimization in LB medium supplemented with ampicillin (100 μg/ml). At $A_{600\ nm}$ of 0.45 the cells were induced by 0.15 mM IPTG and expression was carried out for 6–12 h (for expression at 37, 25 and 16 °C) or 30–48 h (for 4 °C expression). Cells were lysed in 20 mM Tris-HCl (pH 7.5) following either sonication protocol (3–10 ml culture) or French press lysis (>0.5 l). The bacterial lysates were separated to soluble and insoluble fraction by centrifugation (14,000*g*, 20 min). The OCLP1 was recovered in the soluble fraction (>90%) following growth for 18–24 h at for 16 °C. Efficiency of the expression ranged from 10–30 mg/ml. An increased solubility[34] and effectiveness of the purification under strong denaturants[48] by the CBD tag has been previously reported.

Refolding of OCLP1

The expressed protein was purified on cellulose beads, eluted as purified free protein and was subjected to an *in vitro* refolding protocol. The folding is based on oxidation conditions and gradual removing of urea.[49]

Briefly, 1.5 mg of the fused protein (21.5 kDa) was dissolved in 5 ml of the solubilization buffer (6 M urea, 100 mM NaCl, 10 mM Tris–HCl (pH 8.4)) and 5 ml of reduction buffer (10 mM DTT in 50 mM Tris–HCl (pH 8.4)). Following 30 min incubation (room temperature), the mixture was diluted again with the oxidation buffer (3 M urea, 4 mM GSH, 0.4 mM GSSG, 0.1% (w/v) NaN3 in 50 mM Tris–HCl (pH 8.5 0). Dialysis was done against the same buffer (ten volumes) except that urea was decreased by half. Following a complete dialysis (four steps, total of 36 h), the material was purified by cellulose beads and washed in the dialysis buffer. The folded protein was cleaved by recombinant enterokinase (Novagen, molar ratio of 1:100 enzyme/substrate, 16 h, 20 °C). The cleavage occurs at the C-terminal of the recognition site (DDDK), allowing a complete removal of affinity tag sequences. The impurities following the cleavage of the tag were removed by Centriprep10 kDa apparatus (Amicon). The purified OCLP1 was stored at concentration of 20 μM in aliquots (−70 °C) for further use. The resulting peptide was eluted from a 15% (w/v) SDS-Tricine gel and the identity of OCLP1 was verified by MALDI-Tof (Bruker) according to a routine protocol.[50]

Toxicity assays

OCLP1 was diluted to a working solution of 1 μM in PBS and applied directly to the bioassay of larvae and fish. An identical amount of OCLP1 was used by omitting the oxidation step and maintaining the preparation in reducing condition (10 mM DTT) before application. 4 μl were injected to each of the fish and larvae. 2 μM of purified Hydralysin was used for the positive control. The fish were injected laterally at the base of their tail. A paralytic fish is defined as a fish with significantly diminished body movements, preventing the fish from changing its location at least 1 h after injection. The paralysis is characterized by the inability of the fish to cross the aquarium despite active fin movements. Furthermore, the paralyzed fish remain mostly at the bottom of the aquarium as opposed to normal fish that explore the entire aquarium volume.

Bioinformatic sources and tools

All training set proteins were obtained from the UniProt database. The set of 29,554 SwissProt proteins was obtained by taking all SwissProt proteins shorter than or equal to 150 aa and removing redundancy, so that following the process no two proteins are more than 90% identical. The set of 10,157 honey bee predicted protein sequences is the official GLEAN3 predicted gene set.[14] The set of 5154 novel mouse proteins was obtained from the website of the FANTOM project.[15] SignalP[51] was used for predicting signal peptides. ClustalW[52] was used for multiple sequence alignment and phylogenetic analysis. NCBI-BLAST[53] was used for local alignment searches. PHYRE[21] was used for fold recognition. InterProScan[20] was used for detection of sequence motifs. SDPMOD,[54] a homology modeling tool that specializes in structures of small disulfide-rich proteins, was used to construct a 3D model of OCLP1. The ENSEMBL[55] browser was used for genomic searches in *Apis mellifera*, *Drosophila melanogaster*, *A. gambiae* and *A. aegyptis*. CD-HIT[56] was used to cluster the sequences in order to construct non-redundant sets. All expression data were obtained from NCBI nucleotide and EST databases.[26] *Tribolium castaneum* genomic search was performed in the Harvard Genome Sequencing Center website†. The group designated "Antibacterial" contains proteins that have at least one of the following UniProt keywords: Antimicrobial, Fungicide and Antibiotic. The group designated "Venom proteins" contains proteins whose UniProt entries stated localized expression in venom under the TISSUE field. Snake toxin, Gonadotropin, Beta defensin, E6 and L36 represent InterPro[57] groups IPR003571, IPR001545, IPR001855, IPR001334 and IPR000473, respectively.

Accession numbers

GLEAN3 Accession numbers (honey bee predicted gene set): OCLP1: GB19297 (GenBank XM_001120252); Raalin: GB11222 (GenBank).

OCLP1 homologs: A. aegypti: TIGR predicted gene 25118.t00051, GenBank DW209856; *A. gambiae*: Gen Bank CR528986, BX627342; *A. funestus*: GenBank CD578321; *M. sexta*: GenBank BE015616; *S. mediterranea*: GenBank DN297299; *H.erato*: GenBank DT664910; *D. melanogaster*: GenBank CX309613.

Raalin homologs: D. melanogaster: GenBank AAY54877 (IP05928p), BT022461; *A. gambiae*: GenBank BX614868; *B. mori*: GenBank BP115546; *D. pseudoobscura*: GenBank EAL27268 (GA13311).

ANLP: See Table 2.

UniProt accession numbers: ω-conotoxin GVIIA: P05483; ω-conotoxin GVIA: P01522; Covalitoxin II: P82601; ω-conotoxin MVIIA: P05484; *M. domestica* POI: P81765; Apamin: P01500; MCDP: P01499; Iob1: P58609; Ado1: P58608; Ptu1: P58606; SLURP-1: P55000; Lynx1: Q9BZG9; PATE: Q8WXA2; human SP-10: ASPX_HUMAN; mouse SP-10: ASPX_MOUSE;

OMIM accession numbers: Mal de Meleda #248300.

Acknowledgements

We are grateful to Daniel Sher (Dr Zlotkin laboratory, the Hebrew University) for his advice and support. Injections in insects and fish were conducted by D. Sher. We thank Alex Inberg for mass spectrometry support and Alomone laboratories (Jerusalem) for support and advice. N.K. received a fellowship from the Sudarsky Center for Computational Biology. This study was supported by the EU Framework VI NoE BioSapiens and DIAMONDS.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2007.02.106

References

1. Mouhat, S., Jouirou, B., Mosbah, A., De Waard, M. & Sabatier, J. M. (2004). Diversity of folds in animal toxins acting on ion channels. *Biochem. J.* **378**, 717–726.

† http://www.hgsc.bcm.tmc.edu/projects/tribolium/

2. Bastolla, U. & Demetrius, L. (2005). Stability constraints and protein evolution: the role of chain length, composition and disulfide bonds. *Protein Eng. Des. Sel.* **18**, 405–415.
3. Buczek, O., Bulaj, G. & Olivera, B. M. (2005). Conotoxins and the posttranslational modification of secreted gene products. *Cell Mol. Life Sci.* **62**, 3067–3079.
4. Chimienti, F., Hogg, R. C., Plantard, L., Lehmann, C., Brakch, N., Fischer, J. *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.
5. Ibanez-Tallon, I., Miwa, J. M., Wang, H. L., Adams, N. C., Crabtree, G. W., Sine, S. M. & Heintz, N. (2002). Novel modulation of neuronal nicotinic acetylcholine receptors by association with the endogenous prototoxin lynx1. *Neuron*, **33**, 893–903.
6. Kloog, Y., Ambar, I., Sokolovsky, M., Kochva, E., Wollberg, Z. & Bdolah, A. (1988). Sarafotoxin, a novel vasoconstrictor peptide: phosphoinositide hydrolysis in rat heart and brain. *Science*, **242**, 268–270.
7. Ma, D., Armugam, A. & Jeyaseelan, K. (2001). Expression of cardiotoxin-2 gene. Cloning, characterization and deletion analysis of the promoter. *Eur. J. Biochem.* **268**, 1844–1850.
8. Radis-Baptista, G., Kubo, T., Oguiura, N., Prieto da Silva, A. R., Hayashi, M. A., Oliveira, E. B. & Yamane, T. (2004). Identification of crotasin, a crotamine-related gene of *Crotalus durissus terrificus*. *Toxicon*, **43**, 751–759.
9. Torres, A. M., Wang, X., Fletcher, J. I., Alewood, D., Alewood, P. F., Smith, R. *et al.* (1999). Solution structure of a defensin-like peptide from platypus venom. *Biochem. J.* **341**, 785–794.
10. Torres, A. M., Wong, H. Y., Desai, M., Moochhala, S., Kuchel, P. W. & Kini, R. M. (2003). Identification of a novel family of proteins in snake venoms. Purification and structural characterization of nawaprin from *Naja nigricollis* snake venom. *J. Biol. Chem.* **278**, 40097–40104.
11. Miwa, J. M., Ibanez-Tallon, I., Crabtree, G. W., Sanchez, R., Sali, A., Role, L. W. & Heintz, N. (1999). lynx1, an endogenous toxin-like modulator of nicotinic acetylcholine receptors in the mammalian CNS. *Neuron*, **23**, 105–114.
12. Miwa, J. M., Stevens, T. R., King, S. L., Caldarone, B. J., Ibanez-Tallon, I., Xiao, C. *et al.* (2006). The prototoxin lynx1 acts on nicotinic acetylcholine receptors to balance neuronal activity and survival in vivo. *Neuron*, **51**, 587–600.
13. Fry, B. G. (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.
14. Sequencing Consortium, T. H. (2006). Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*, **444**, 512.
15. Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N. *et al.* (2005). The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.
16. Pelegrini, P. B. & Franco, O. L. (2005). Plant gamma-thionins: novel insights on the mechanism of action of a multi-functional class of defense proteins. *Int. J. Biochem. Cell Biol.* **37**, 2239–2253.
17. Jin, H. Y., Hu, Q., Jun, J. Y., Ju, A., Sen, L. D., Qian, D. R. & Lin, Q. R. (2005). Preliminary studies on the zinc-induced metallothionein protein with antibacterial activity in housefly larvae, *Musca domestica*. *Acta Biol. Hung.* **56**, 283–295.
18. Hugues, M., Romey, G., Duval, D., Vincent, J. P. & Lazdunski, M. (1982). Apamin as a selective blocker of the calcium-dependent potassium channel in neuroblastoma cells: voltage-clamp and biochemical characterization of the toxin receptor. *Proc. Natl Acad. Sci. USA*, **79**, 1308–1312.
19. Ziai, M. R., Russek, S., Wang, H. C., Beer, B. & Blume, A. J. (1990). Mast cell degranulating peptide: a multifunctional neurotoxin. *J. Pharm. Pharmacol.* **42**, 457–461.
20. Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. & Lopez, R. (2005). InterProScan: protein domains identifier. *Nucl. Acids Res.* **33**, W116–W120.
21. Kelley, L. A., MacCallum, R. M. & Sternberg, M. J. (2000). Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299**, 499–520.
22. Bernard, C., Corzo, G., Adachi-Akahane, S., Foures, G., Kanemaru, K., Furukawa, Y. *et al.* (2004). Solution structure of ADO1, a toxin extracted from the saliva of the assassin bug, *Agriosphodrus dohrni*. *Proteins: Struct. Funct. Genet.* **54**, 195–205.
23. Bernard, C., Corzo, G., Mosbah, A., Nakajima, T. & Darbon, H. (2001). Solution structure of Ptu1, a toxin from the assassin bug *Peirates turpis* that blocks the voltage-sensitive calcium channel N-type. *Biochemistry*, **40**, 12795–12800.
24. Corzo, G., Adachi-Akahane, S., Nagao, T., Kusui, Y. & Nakajima, T. (2001). Novel peptides from assassin bugs (Hemiptera: Reduviidae): isolation, chemical and biological characterization. *FEBS Letters*, **499**, 256–261.
25. Flinn, J. P., Pallaghy, P. K., Lew, M. J., Murphy, R., Angus, J. A. & Norton, R. S. (1999). Roles of key functional groups in omega-conotoxin GVIA synthesis, structure and functional assay of selected peptide analogues. *Eur. J. Biochem.* **262**, 447–455.
26. Boguski, M. S., Lowe, T. M. & Tolstoshev, C. M. (1993). dbEST–database for "expressed sequence tags". *Nature Genet.* **4**, 332–333.
27. Sher, D., Knebel, A., Bsor, T., Nesher, N., Tal, T., Morgenstern, D. *et al.* (2005). Toxic polypeptides of the hydra–a bioinformatic approach to cnidarian allomones. *Toxicon*, **45**, 865–879.
28. Zhang, M., Fishman, Y., Sher, D. & Zlotkin, E. (2003). Hydralysin, a novel animal group-selective paralytic and cytolytic protein from a noncnidocystic origin in hydra. *Biochemistry*, **42**, 8939–8944.
29. Hagiwara, K., Kikuchi, T., Endo, Y., Huqun, Usui, K., Takahashi, M. *et al.* (2003). Mouse SWAM1 and SWAM2 are antibacterial proteins composed of a single whey acidic protein motif. *J. Immunol.* **170**, 1973–1979.
30. Sallenave, J. M. (2002). Antimicrobial activity of antiproteinases. *Biochem. Soc. Trans.* **30**, 111–115.
31. Simpson, A. J., Maxwell, A. I., Govan, J. R., Haslett, C. & Sallenave, J. M. (1999). Elafin (elastase-specific inhibitor) has anti-microbial activity against Gram-positive and Gram-negative respiratory pathogens. *FEBS Letters*, **452**, 309–313.
32. Radis-Baptista, G., Moreno, F. B., de Lima Nogueira, L., Martins, A. M., de Oliveira Toyama, D., Toyama, M. H. *et al.* (2006). Crotacetin, a novel snake venom C-type lectin homolog of convulxin, exhibits an unpredictable antimicrobial activity. *Cell Biochem. Biophys.* **44**, 412–423.
33. Escoubas, P. & Rash, L. (2004). Tarantulas: eight-legged pharmacists and combinatorial chemists. *Toxicon*, **43**, 555–574.
34. Murashima, K., Kosugi, A. & Doi, R. H. (2003).

Solubilization of cellulosomal cellulases by fusion with cellulose-binding domain of noncellulosomal cellulase engd from *Clostridium cellulovorans*. *Proteins-Struct. Funct. Genet.* **50**, 620–628.
35. Daquinag, A. C., Sato, T., Koda, H., Takao, T., Fukuda, M., Shimonishi, Y. & Tsukamoto, T. (1999). A novel endogenous inhibitor of phenoloxidase from *Musca domestica* has a cystine motif commonly found in snail and spider toxins. *Biochemistry*, **38**, 2179–2188.
36. Shi, L., Li, B. & Paskewitz, S. M. (2006). Cloning and characterization of a putative inhibitor of melanization from *Anopheles gambiae*. *Insect Mol. Biol.* **15**, 313–320.
37. Herr, J. C., Flickinger, C. J., Homyk, M., Klotz, K. & John, E. (1990). Biochemical and morphological characterization of the intra-acrosomal antigen SP-10 from human sperm. *Biol. Reprod.* **42**, 181–193.
38. Bera, T. K., Maitra, R., Iavarone, C., Salvatore, G., Kumar, V., Vincent, J. J. *et al.* (2002). PATE, a gene expressed in prostate cancer, normal prostate, and testis, identified by a functional genomic approach. *Proc. Natl Acad. Sci. USA*, **99**, 3058–3063.
39. Foster, J. A. & Herr, J. C. (1992). Interactions of human sperm acrosomal protein SP-10 with the acrosomal membranes. *Biol. Reprod.* **46**, 981–990.
40. Bray, C., Son, J. H. & Meizel, S. (2002). A nicotinic acetylcholine receptor is involved in the arosome reaction of human sperm initiated by recombinant human ZP3. *Biol. Reprod.* **67**, 782–788.
41. Rajendra, W., Armugam, A. & Jeyaseelan, K. (2004). Neuroprotection and peptide toxins. *Brain Res. Brain Res. Rev.* **45**, 125–141.
42. Biragyn, A. (2005). Defensins–non-antibiotic use for vaccine development. *Curr. Protein Pept. Sci.* **6**, 53–60.
43. Bowersox, S. S. & Luther, R. (1998). Pharmacotherapeutic potential of omega-conotoxin MVIIA (SNX-111), an N-type neuronal calcium channel blocker found in the venom of *Conus magus*. *Toxicon*, **36**, 1651–1658.
44. Miljanich, G. P. (2004). Ziconotide: neuronal calcium channel blocker for treating severe chronic pain. *Curr. Med. Chem.* **11**, 3029–3040.
45. Wermeling, D. P. (2005). Ziconotide, an intrathecally administered N-type calcium channel antagonist for the treatment of chronic pain. *Pharmacotherapy*, **25**, 1084–1094.
46. Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B. *et al.* (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucl. Acids Res.* **34**, D187–D191.
47. Freund, Y. & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* **55**, 119–139.
48. Berdichevsky, Y., Lamed, R., Frenkel, D., Gophna, U., Bayer, E. A., Yaron, S. *et al.* (1999). Matrix-assisted refolding of single-chain Fv- cellulose binding domain fusion proteins. *Protein Expr. Purif.* **17**, 249–259.
49. Cao, P., Mei, J. J., Diao, Z. Y. & Zhang, S. (2005). Expression, refolding, and characterization of human soluble BAFF synthesized in *Escherichia coli*. *Protein Expr. Purif.* **41**, 199–206.
50. Inberg, A., Bogoch, Y., Bledi, Y. & Linial, M. (2007). Cellular processes underlying maturation of P19 neurons: changes in protein folding regimen and cytoskeleton organization. *Proteomics*, **7**, 910–920.
51. Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783–795.
52. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
53. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402.
54. Kong, L., Lee, B. T., Tong, J. C., Tan, T. W. & Ranganathan, S. (2004). SDPMOD: an automated comparative modeling server for small disulfide-bonded proteins. *Nucl. Acids Res.* **32**, W356–W359.
55. Birney, E., Andrews, D., Caccamo, M., Chen, Y., Clarke, L., Coates, G. *et al.* (2006). Ensembl 2006. *Nucl. Acids Res.* **34**, D556–D561.
56. Li, W., Jaroszewski, L. & Godzik, A. (2002). Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*, **18**, 77–82.
57. Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D. *et al.* (2005). InterPro, progress and status in 2005. *Nucl. Acids Re.* **33**, D201–D205.

Edited by B. Honig

(Received 17 December 2006; received in revised form 14 February 2007; accepted 21 February 2007)
Available online 15 March 2007