Published online 8 May 2009 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* ***W363–W368***
doi:10.1093/nar/gkp299

ClanTox: a classifier of short animal toxins

Guy Naamati[1], Manor Askenazi[2,3] and Michal Linial[2,*]

[1]Computer Science and Engineering, [2]Department of Biological Chemistry, Institute of Life Sciences, The Sudarsky Center for Computational Biology, The Hebrew University of Jerusalem, Israel and [3]Blais Proteomics Center, Dana-Farber Cancer Institute, Boston, MA 02115, USA

Received January 18, 2009; Revised March 22, 2009; Accepted April 15, 2009

ABSTRACT

Toxins are detected in sporadic species along the evolutionary tree of the animal kingdom. Venomous animals include scorpions, snakes, bees, wasps, frogs and numerous animals living in the sea such as the stonefish, snail, jellyfish, hydra and more. Interestingly, proteins that share a common scaffold with animal toxins also exist in non-venomous species. However, due to their short length and primary sequence diversity, these, toxin-like proteins remain undetected by classical search engines and genome annotation tools. We construct a toxin classification machine and web server called ClanTox (Classifier of Animal Toxins) that is based on the extraction of sequence-driven features from the primary protein sequence followed by the application of a classification system trained on known animal toxins. For a given input list of sequences, from venomous or non-venomous settings, the ClanTox system predicts whether each sequence is toxin-like. ClanTox provides a ranked list of positively predicted candidates according to statistical confidence. For each protein, additional information is presented including the presence of a signal peptide, the number of cysteine residues and the associated functional annotations. ClanTox is a discovery-prediction tool for a relatively overlooked niche of toxin-like cell modulators, many of which are therapeutic agent candidates. The ClanTox web server is freely accessible at http://www.clantox.cs.huji.ac.il.

INTRODUCTION

Animal toxins are one of the most highly over-represented functional groups among the short proteins. Toxins are proteins that appear in animal venom and are aimed at inflicting harm to the organism on which the venom acts (1,2). They are extremely varied in terms of function and include ion channel inhibitors, phospholipases, protease inhibitors, disintegrins, membrane pore inducers and more (3,4).

In recent years, several proteins that resemble animal toxins were identified in non-venomous contexts and shown to act as cell activity modulators. These toxin-like proteins include proteases, protease inhibitors, as well as secreted proteins, which resemble cell antigens, growth factors and more (5). A strong evolutionary relation exists between animal toxins and ancestral cysteine cross-linked proteins (6–8). One of the most striking example is the set of human and rodent proteins exhibiting a strong similarity to snake α-neurotoxins (9,10).

Animal toxins are very diverse set of proteins. Some snake toxin metalloproteinases are >400 amino acids and the length of the black widow spider latrotoxin is >1500 amino acids. Other potent toxins like the conotoxins may be very short (10–20 amino acids). Nevertheless, many of the animal toxins that appear in venom are rather short (30–120 amino acids) and vary in sequence, structure and function. Due to their short length and high sequence variation, toxin-like proteins are unclassifiable by standard motif-detection or fold-recognition methods (4). Based on these facts, we suspect that toxin-like proteins are understudied. To remedy this situation, we have developed a classifier for ranking protein sequences according to their toxin-like properties. We have extracted sequence-driven properties to create a robust characterization of toxin-like proteins. These properties include features of structural stability, distribution and frequency of amino acids, protein length and more (11). Application of the classifier to the $\sim$10 000 predicted protein sequences from the sequenced honeybee (*Apis mellifera*) genome identified several toxin-like sequences, including some that were previously unknown. For one of these sequences, we confirmed an overlooked function of a voltage-gated channel inhibitor (11).

Herein, we present the web-based classifier ClanTox (Classifier of Animal Toxins), which can be used to conduct a large-scale search for toxin-like protein sequences from a wide range of sequencing and transcriptomic data. In addition to detecting animal toxins the ClanTox server identifies additional short proteins that function as cell modulators and are characterized by multiple

*To whom correspondence should be addressed. Tel: +972 2 6585425; Fax: +972 2 6586448; Email: michall@cc.huji.ac.il

© 2009 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

scaffold-stabilizing disulfide-bridges. Among these are metallothioneins, proteases and their inhibitors, defensins, growth factors and a variety of proteins from the innate immune systems (11). Many of these proteins are secreted proteins that act in the extracellular milieu through binding to membrane receptors and ion channels. However, other cysteine-rich proteins that do not adopt toxin-like properties are predicted negative by the classifier (e.g. ribosomal proteins). ClanTox performs best with short sequences (<200 amino acids) or for longer proteins that were separated by their individual domains. The rapid expansion in mass spectrometry (MS) proteomics and the depth of transcriptome coverage by new generation sequencing suggests that ClanTox will become increasingly useful as a discovery tool focusing attention on short toxin-related and disulfide-containing proteins.

CLASSIFER DESCRIPTION

The classifier underlying ClanTox was derived by machine learning on a set of true and false instances obtained from a manually reviewed set of ion channel toxin inhibitors. The ClanTox server takes a given set of proteins sequences (~10 000) and sends them to the underlying classifier after each sequence is translated into a vector of sequence-derived global features (545 features). Specifically, the vector is independently sent to 10 boosted-stump classifiers, each of which produces a numerical result. During the learning phase, for a given set of true instances, we randomly generated 10 sets of false instances [as described in (11)]. For each set of false instances we trained a parameter-tuned boosted stump classifier. The output of the 10 classifiers was normalized to the highest positive prediction of each classifier. The final prediction of the meta-classifier is the mean prediction of the underlying 10 classifiers. The standard deviation of the score indicates how much the 10 subclassifiers agree with one another. We consider a prediction to be a positive prediction (i.e. the protein is toxin like) if the mean is far greater than the standard deviation, suggesting a robust hypothesis, which is not biased by any specific set of false instances. In a 3-fold cross-validation test, the performance of the classifier was measured by the area under curve (AUC; maximal success is translated to AUC = 1.0). The classifier showed a high level of success, with a mean AUC of 0.9934 (SD = 0.0026). Two of the most dominant features found to improve the quality of the classifier were associated with the frequency and distribution of cysteine residues within the primary sequence, which are indeed, crucial structural factors underlying toxins stability. Validation tests for the functional enrichment of keywords for positively predicted sequences and a formal description of the classifier and training set construction are shown in 'Details' section of the ClanTox website.

Proteome-wide screening for toxin-like prediction

An attractive application for ClanTox is the screening of cDNA libraries, EST collections and RNA-Seq of complete genomes for toxin-like proteins. At present, there are hundreds of multi-cellular genomes that are fully or partially sequenced. Any set of protein sequences can be retrieved from PIR (12), SRS (13) or UniProtKB (14) and then tested by ClanTox. Testing the level of prediction in several non-venomous model organisms (mammals, fruit fly and worm) indicated that for short sequences (protein length 20–160 amino acids), positive prediction (marked as P1–P3) accounts for 3–9% of the sequences. The positive predictions drop to only ~1–3% when the entire range for protein length is included (data not shown).

An interesting test case is the analysis of snakes (Taxonomy ID: 8570). Among the snakes, the largest two snake families are Elapids and Vipers. Elapids comprise the biggest family of land (found in tropical and subtropical regions) and sea snakes (mostly found in Indian and the Pacific oceans). Vipers comprise a family of venomous snakes found all over the world (except in Australia and Madagascar). There are 797 and 354 short, full-length sequences (defined as above) from Elapids (Taxonomy ID: 8602) and Vipers (Taxonomy ID: 8690), respectively. ClanTox positively predicted most of these sequences (86% in Elapids and 76% in Vipers) as toxins and toxin-like (Table 1). Positive predictions are somewhat arbitrarily partitioned into three levels from P3 (most significant) to P1 (less significant).

Investigating the prediction performance for major classes of toxins revealed that each toxin type (i.e. PLA2) shows a tendency for a unique prediction level. For example, the Elapid PLA2 is mostly predicted at a P1 level, while toxins from Vipers including Ammodytin, a myotoxic phospholipase-like protein, are mostly detected at a P2 level. Other toxin classes like Elapid Cardiotoxin and Cytotoxin are mostly predicted at a P3 level. More importantly, the number of toxins that were miss, is negligible (Table 1, N prediction for only 1 out of 449 sequences).

USER PERSPECTIVE AND WEB INTERFACE

Input

ClanTox receives a set of ~10 000 protein sequences as free text in an input box or as an uploaded file in FASTA text format. Following transformation of the sequences to a numerical vector and the activation of the classifier, the results are presented as a ranked list of all input proteins, sorted by the predictive score.

Prediction

The classifier presents four labels—N for negative prediction and P1–P3, reflecting three levels of positive predictions for toxin-like sequences. The most significant predictions (labeled P3) accounts for proteins with a mean score >0.2 as well as having a coefficient of variation (CV) <0.5. The negative predictions (i.e. predicted as non-toxin) account for all sequences with a mean score <–0.2 (Figure 1). For definitions of the prediction labels see 'Details' section.

Table 1. ClanTox predictions of Elapids and Vipers short proteins

Taxonomy toxin family	P3	P2	P1	N	Total
Elapids	291	147	260	99	797
Cardiotoxin	**40**	4	1	0	45
Cytotoxin	**44**	0	1	0	45
PLA2	2	20	**115**	1	137
Vipers	79	142	47	86	354
Disintegrin	**32**	14	3	0	49
Ammodytin	24	**71**	15	0	110
PLA2	14	**36**	13	0	63

The dominating prediction level for each toxin family type is marked in bold. PLA2, phospholipase A2.

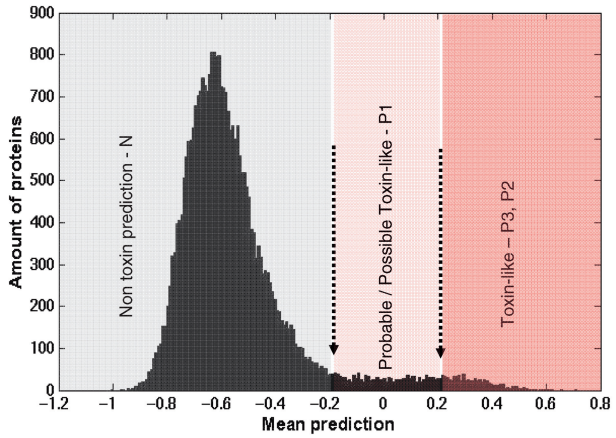

Figure 1. Score distribution for the non-redundant set of all ~30 000 SwissProt proteins shorter than 150 amino acids. The proteins that were used for the classifier positive training set were excluded. Positive and secured prediction scores are (P2–P3) assigned to the tail of the distribution with a mean score >0.2. The intermediate levels of score ranging between –0.2 and 0.2 are considered probable/possible toxin-like. More refined positive prediction confidence (marked as P3, P2 and P1) is defined according to the scale of the SD relative to the mean score. The negative prediction (N) is associated with a mean score <–0.2 that includes the large Gaussian-like distribution for the vast majority of the proteins.

Display of prediction results

The results of the classifier are displayed in a series of presentations (Figure 2):

(i) A summary and a pie chart of the predictions for all the submitted sequences −P3 to P1; most confident to least confident prediction and N; high confidence for negative prediction (Figure 2A).
(ii) A detailed table of predictions (Figure 2B) including quantitative details and information from external prediction tools.
(iii) A histogram of the mean scores (ranging from –1.0 to 1.0) for the entire set of input sequences (Figure 2C).
(iv) A graphical representation of the protein, focused on cysteine's appearance along the sequence (Figure 2D).

Extended data for ClanTox prediction

A detailed table is associated with each sequence showing: (i) the accession number and ID based on UniProtKB identifiers; (ii) the protein name; (iii) the protein length in amino acids; (iv) the number of cysteines in the sequence; (v) the prediction confidence level, color-coded; (vi) the mean score (ranging from –1.0 to 1.0) and (vii) the standard deviation (SD) reported by the meta-classifier. Additional information is provided (if available in the sequence header) including (viii) whether the sequence is annotated as consisting of a protein fragment (Frag) and (ix) the organism's official name. In addition to textual output, each sequence can be viewed in the table as a graphical representation, showing the sequence as a bar (not to scale) with indication of the cysteine residue locations.

Additional information for each of the sequence is obtained through an active link to the UniProtKB page and to the ProtoNet (15) protein card. Users can activate, for each sequence from the ClanTox candidate list, the following tools: (i) a NCBI BLAST search to detect related sequences; (ii) a ProtoNet search (15) providing a cluster that best fits the sequence selected (the view through the ProtoNet hierarchical tree presents the most relevant protein family from the UniProtKB sequence database. ProtoNet clusters typically provide a rich set of annotations associated with the members of the relevant cluster) and (iii) detection of signal- peptides based on the SignalP 3.0 program (17). SignalP predicts the presence of a signal peptide at the amino terminal of the sequence (tuned for eukaryotes). The detailed table created by ClanTox can be sorted by any of its columns (Figure 2B). Positive predictions that fail to be predicted as secreted proteins should be considered suspicious. This is extended also to proteins that include known subcellular signatures such as the ER retention signal or the Glycosylphosphatidylinositol (GPI) anchor.

Download options

ClanTox provides several download options. The results can be downloaded in tabular format, as a flat file of selected FASTA sequences and as a list of UniProtKB ID. Several optional filters are used to limit the downloaded list according to the prediction labels (P3, P2, P1 and N).

Interface with external servers

We selected a small set of tools that provide crucial and complementary knowledge for the effective analysis of toxin-like proteins. These include: the SDPMOD, a homology modeling tool that specializes in structures of small disulfide-rich proteins (18); PANDORA (19), which provides an annotation-driven analysis and visualization for the candidate sequences. PANDORA is compatible with the FASTA format of UniProtKB ID as made available by the ClanTox download functionality.

ClanTox helps new users by providing answers to frequently asked questions (FAQ). Furthermore, sequence retrieval servers are listed to support the user in compiling

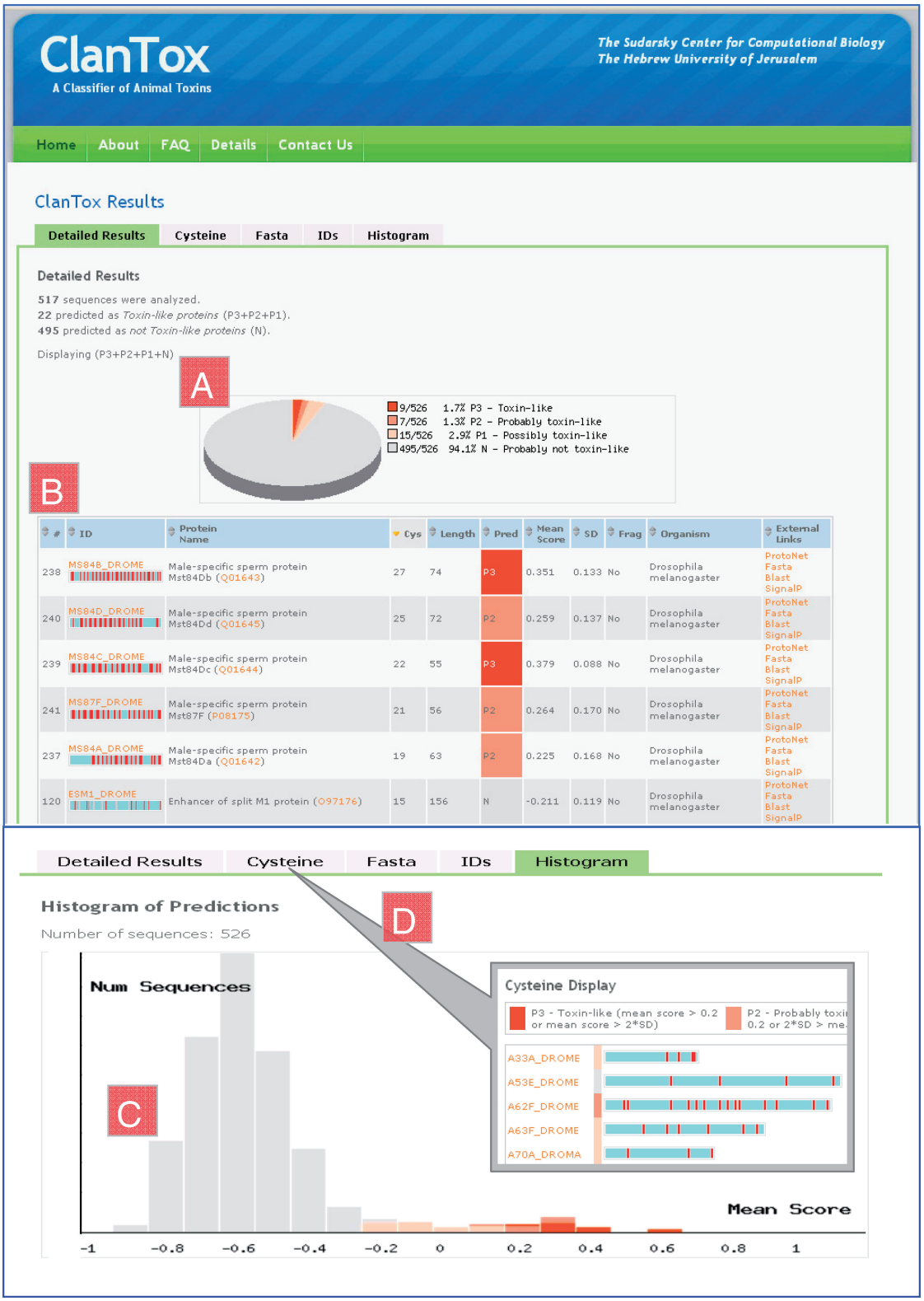

Figure 2. Screenshots of the ClanTox result page. (**A**) A pie chart displaying the distribution of prediction classes for 526 short sequences (20–160 amino acids) retrieved from the Drosophila proteomes following filtration of all sequences that contain fragments. The positive predictions (P1–P3) constitute 5.9% of the sequences and are marked in shades of red (P3) to light pink (P1). (**B**) A detailed table of toxin-like proteins predictions sorted by the number of cysteines (Cys). Few examples labeled as P3 (red), P2 (light red) and N (gray) are shown. Together with the ID column, a graphical scheme is used to represent protein sequences (not to scale, light blue bar) marked by cysteine residue distribution (red vertical bars). In addition, the table shows the protein names, UniProtKB accession, the number of cysteines, sequence length, the mean score and standard deviation (SD). Several links to external tools are presented for each predicted protein. (**C**) Histogram of mean scores for the 526 Drosophila sequences that are used as input. The positive predictions are color coded as in A. Most sequences were predicted negative and are shown in gray. (**D**) A cysteine centric view, the sequences shown by their relative length. Cysteines are marked as red vertical lines.

sequence sets to be analyzed by ClanTox. Using the UniProtKB search engine, the user can retrieve sequences starting with a list of IDs or accession numbers for obtaining associated annotations, cross-references and additional functional information. The user can alternatively use batch retrieval systems such as PIR (12) and SRS (13) to obtain a sequence sets for analysis. PICR (20) allows cross-referencing between alternative identifiers and can be used as a preprocessing step that leverages additional identifiers used by other major proteome resources.

Update and future development

ClanTox will be extended in two main directions: (i) adding analysis tools that will be locally inserted. Such tools will include disulfide-connectivity servers, 3D predicted structures and functional inference tools; (ii) incorporating the presence of the signal peptide and the properties of the cysteine disulfide connectivity predictions into the framework of the machine-learning scheme. Web server developers, which are interested in interfacing directly with ClanTox should contact the authors.

CONCLUSIONS

In recent years, a handful of toxins and toxin-like proteins have been identified. The Tox-Prot database (2) was developed to enable the systematic annotation of proteins, which act as animal toxins and are produced by venomous and poisonous animals. A related resource covering about 3000 manually annotated toxins from UniProtKB has even introduced a specialized toxin ontology (TO) (21). While all these resources depend on manual annotations and on expert contributions, ClanTox provides a systematic scheme for proteome-wide prediction of toxin-like proteins. Furthermore ClanTox acts also on non-venomous organisms to detect toxin-like proteins that, to date, are only sporadically identified. It is suggested that in evolutionary terms, toxins are homologs of toxin-like proteins that often act as modulators of channels and receptors at the cell membrane. By using ClanTox to rank the statistically significant sequences matching toxin-like criteria, the user can focus on a relatively small fraction of high-confidence candidates. These positively predicted sequences can then be further challenged by a set of tools and predictors directly from the ClanTox results table. Information on the validity of the signal peptide, the distribution and connectivity of cysteine residues from the primary sequence (22) and the functional class based on disulfide connectivity patterns (23) will fill the gap in knowledge for many of the overlooked short and cysteine-rich proteins.

ACKNOWLEDGEMENTS

We would like to thank Solange Kresanty for her excellent and creative work in design and managing the website. We specifically thank Noam Kaplan, who initiated the project and set the foundation for ClanTox. We thank Mooli Tayar for his technical support and the ProtoNet team for their instrumental support and valuable discussion.

FUNDING

Israel Science Foundation (ISF) and the US–Israel Binational Science Foundation (BSF); Sudarsky Center for Computational Biology (SCCB) (to G.N.).

Conflict of interest statement. None declared.

REFERENCES

1. Tan,P.T., Veeramani,A., Srinivasan,K.N., Ranganathan,S. and Brusic,V. (2006) SCORPION2: a database for structure-function analysis of scorpion toxins. *Toxicon*, **47**, 356–363.
2. Jungo,F. and Bairoch,A. (2005) Tox-Prot, the toxin protein annotation program of the Swiss-Prot protein knowledgebase. *Toxicon*, **45**, 293–301.
3. Fry,B.G., Wuster,W., Kini,R.M., Brusic,V., Khan,A., Venkataraman,D. and Rooney,A.P. (2003) Molecular evolution and phylogeny of elapid snake venom three-finger toxins. *J. Mol. Evol.*, **57**, 110–129.
4. Tan,P.T., Khan,A.M. and Brusic,V. (2003) Bioinformatics for venom and toxin sciences. *Brief Bioinform.*, **4**, 53–62.
5. Whittington,C.M., Papenfuss,A.T., Bansal,P., Torres,A.M., Wong,E.S., Deakin,J.E., Graves,T., Alsop,A., Schatzkamer,K., Kremitzki,C. *et al.* (2008) Defensins and the convergent evolution of platypus and reptile venom genes. *Genome Res.*, **18**, 986–994.
6. 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.
7. Kini,R.M. (2002) Molecular moulds with multiple missions: functional sites in three-finger toxins. *Clin. Exp. Pharmacol. Physiol.*, **29**, 815–822.
8. Fry,B.G., Vidal,N., Norman,J.A., Vonk,F.J., Scheib,H., Ramjan,S.F., Kuruppu,S., Fung,K., Hedges,S.B., Richardson,M.K. *et al.* (2006) Early evolution of the venom system in lizards and snakes. *Nature*, **439**, 584–588.
9. Chimienti,F., Hogg,R.C., Plantard,L., Lehmann,C., Brakch,N., Fischer,J., Huber,M., Bertrand,D. and Hohl,D. (2003) Identification of SLURP-1 as an epidermal neuromodulator explains the clinical phenotype of Mal de Meleda. *Hum. Mol. Genet.*, **12**, 3017–3024.
10. Miwa,J.M., Ibanez-Tallon,I., Crabtree,G.W., Sanchez,R., Sali,A., Role,L.W. and Heintz,N. (1999) Lynx1, an endogenous toxin-like modulator of nicotinic acetylcholine receptors in the mammalian CNS. *Neuron*, **23**, 105–114.
11. Kaplan,N., Morpurgo,N. and Linial,M. (2007) Novel families of toxin-like peptides in insects and mammals: a computational approach. *J. Mol. Biol.*, **369**, 553–566.
12. Wu,C.H. (2006) Bioinformatics for proteomics at the Protein Information Resource (PIR). *Mol. Cell Proteomics*, **5**, S341–S341.
13. Zdobnov,E.M., Lopez,R., Apweiler,R. and Etzold,T. (2002) The EBI SRS server-new features. *Bioinformatics*, **18**, 1149–1150.
14. Leinonen,R., Nardone,F., Zhu,W. and Apweiler,R. (2006) UniSave: the UniProtKB sequence/annotation version database. *Bioinformatics*, **22**, 1284–1285.
15. Kaplan,N., Sasson,O., Inbar,U., Friedlich,M., Fromer,M., Fleischer,H., Portugaly,E., Linial,N. and Linial,M. (2005) ProtoNet 4.0: a hierarchical classification of one million protein sequences. *Nucleic Acids Res.*, **33**, D216–D218.
16. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
17. Bendtsen,J.D., Nielsen,H., von Heijne,G. and Brunak,S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.

18. Kong,L., Lee,B.T., Tong,J.C., Tan,T.W. and Ranganathan,S. (2004) SDPMOD: an automated comparative modeling server for small disulfide-bonded proteins. *Nucleic Acids Res.*, **32**, W356–359.
19. Kaplan,N., Vaaknin,A. and Linial,M. (2003) PANDORA: keyword-based analysis of protein sets by integration of annotation sources. *Nucleic Acids Res.*, **31**, 5617–5626.
20. Cote,R.G., Jones,P., Martens,L., Kerrien,S., Reisinger,F., Lin,Q., Leinonen,R., Apweiler,R. and Hermjakob,H. (2007) The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics*, **8**, 401.
21. He,Q.Y., He,Q.Z., Deng,X.C., Yao,L., Meng,E., Liu,Z.H. and Liang,S.P. (2008) ATDB: a uni-database platform for animal toxins. *Nucleic Acids Res.*, **36**, D293–297.
22. Martelli,P.L., Fariselli,P. and Casadio,R. (2004) Prediction of disulfide-bonded cysteines in proteomes with a hidden neural network. *Proteomics*, **4**, 1665–1671.
23. Lenffer,J., Lai,P., El Mejaber,W., Khan,A.M., Koh,J.L., Tan,P.T., Seah,S.H. and Brusic,V. (2004) CysView: protein classification based on cysteine pairing patterns. *Nucleic Acids Res.*, **32**, W350–W355.