Published online 11 October 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* ***D293–D297***
doi:10.1093/nar/gkm832

ATDB: a uni-database platform for animal toxins

Quan-Yuan He, Quan-Ze He, Xing-Can Deng, Lei Yao, Er Meng, Zhong-Hua Liu and Song-Ping Liang*

The Key Laboratory of Protein Chemistry and Developmental Biology of Ministry of Education, College of Life Sciences, Hunan Normal University, Changsha 410081, P. R. China

Received July 25, 2007; Revised September 20, 2007; Accepted September 21, 2007

ABSTRACT

Venomous animals possess an arsenal of toxins for predation and defense. These toxins have great diversity in function and structure as well as evolution and therefore are of value in both basic and applied research. Recently, toxinomics researches using cDNA library sequencing and proteomics profiling have revealed a large number of new toxins. Although several previous groups have attempted to manage these data, most of them are restricted to certain taxonomic groups and/or lack effective systems for data query and access. In addition, the description of the function and the classification of toxins is rather inconsistent resulting in a barrier against exchanging and comparing the data. Here, we report the ATDB database and website which contains more than 3235 animal toxins from UniProtKB/Swiss-Prot and TrEMBL and related toxin databases as well as published literature. A new ontology (Toxin Ontology) was constructed to standardize the toxin annotations, which includes 745 distinct terms within four term spaces. Furthermore, more than 8423 TO terms have been manually assigned to 2132 toxins by trained biologists. Queries to the database can be conducted via a user-friendly web interface at http://protchem.hunnu.edu.cn/toxin.

INTRODUCTION

Toxins are, according to the Oxford Dictionary of Biochemistry and Molecular Biology, any of various specific poisonous substances that are formed biologically. Conventionally, we also term any molecule from various animal venoms as a toxin, although it may not necessarily be poisonous. No one knows exactly the number of venomous species or the number of toxins they have developed for predation and defense, but it is obvious these numbers will be large. Taking spider as an example, it has ~38 000 described species with an even greater number awaiting characterization. Spider venoms are incredibly complex chemical cocktails, which may contain ~1.5–1.9 million polypeptide toxins based on even very conservative estimates (1,2,3). Exploring the natural treasury of toxins is of value to both basic research and drug design.

A large amount of data has come from fruitful researches on animal toxins in the last 40 years. Recently the accelerating output of data from cDNA library sequencing (4,5) and proteomics profiling researches (6) about toxinomes makes it urgent for biologists to build databases and develop standards to collect, store and classify toxin information. Some attempts have been made, such as the International Venom and Toxin Database (http://www.kingsnake.com/toxinology/), the Tox-Prot program (7), the snake neurotoxin database (http://sdmc.i2r.a-star.edu.sg/Templar/DB/snake_neurotoxin/), the scorpion toxin database (8), the MOLLUSK toxin database (http://research.i2r.a-star.edu.sg/MOLLUSK/) and so on. However, most of them are based on unformatted text, restricted to certain taxonomic groups and/or lack effective systems for data query and access.

ATDB, a uni-database platform, is designed to store chemical structures and annotation data of all animal toxins and presents a new conserved, structural terms system (Toxin Ontology) to standardize toxin functional annotations. It may be the most comprehensive toxin database and now contains 3235 peptide toxins and five small molecules from 379 species (Table 1). Most of them are annotated manually using more than 8423 TO (Toxin Ontology) terms by trained biologists. All data can be accessed and downloaded from a user-friendly web interface at http://protchem.hunnu.edu.cn/toxin.

DATA COLLECTION PIPELINE

Most protein and nucleic acid sequences of toxins were retrieved from general public databases such as the

*To whom correspondence should be addressed. Tel: +86 731 8872556; Fax: +86 731 8861304; Email: liangsp@hunnu.edu.cn

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Table 1. Number of entries for each species groups in ATDB (release 1.1, 14 June 2007)

Group of species	Taxonomy name	Number of toxins	Number of species
Snakes	Serpentes	998	114
Scorpions	Scorpiones	699	51
Spiders	Araneae	267	48
Cone snails	Conus	506	57
Sea anemones	Actiniaria	87	26
Insects	Hexapoda	38	22
Fish	Teleostei	19	8
Mammals	Mammalias	4	1
Lizards	Heloderma	5	2
Jellyfish1	Cubomedusae/ Scyphozoa	4	4
Worms	Cerebratulus	3	1
Sea stars	Asteroidea	3	1
Hydra	Hydroida	4	4
Toad	Amphibia	3	1
Others	–	172	39
All	Metazoa	2812	379

Note: The entries for recombinant expressed and chemically synthesized toxins are not included in the table.

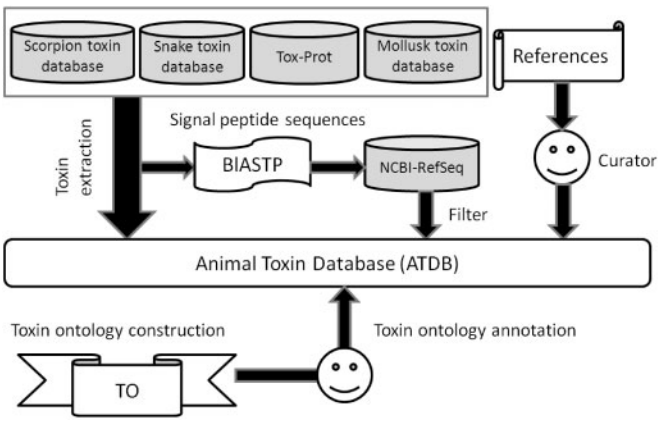

Figure 1. Schematic overview of the pipeline of data integration in ATDB. All sequence data were downloaded by December 2006. Signal peptide sequences were extracted by an in-house Perl script. Taking these sequences as probes, we searched the NCBI-RefSeq database by BLASTP and filtered by the key word 'venom gland' in tissue specificity annotations. Toxin ontology construction and annotation were mainly done manually by trained biologists.

UniProtKB/Swiss-Prot and TrEMBL (9), NCBI-RefSeq (10), NCBI-nucleotide (10) and from the specified toxin databases mentioned above. The pipeline of sequence extraction is shown in Figure 1. First, all protein sequences in UniProtKB/Swiss-Prot/TrEMBL and the toxin databases were downloaded. The redundant entries among them were eliminated based on sequence and species information. Second, to obtain many toxin sequences as possible, we applied the signal peptide sequence of known toxins to BLAST NCBI-RefSeq database, because the signal peptide is the most conserved component of toxin precursors. A total of 7354 sequences were retrieved and then filtered by the key word 'venom gland' in tissue specificity. After removing redundant entries, an additional 458 sequences were obtained. Finally, another 56 sequences were extracted from recent publications. All the related nucleotide sequences were retrieved from the NCBI-nucleotide database. In total, 3235 peptides and 1281 nucleic acid sequences were collected for further annotation.

It is obvious that comprehensive reference information is crucial for an understanding of the function and evolution of toxins. The most common question about toxins is which species produce them. To answer this, we constructed a taxonomic tree of venomous animals. It contains 22 taxonomic layers from domain to subspecies, 379 species and 808 nodes. Descriptions of more than 400 taxonomic categories (nodes) about ecology, distribution and evolution information have been extracted from the NCBI-Taxonomy database (10) and Google search results manually. Over 500 pictures for species/taxonomic groups were selected and downloaded from the Internet following automatic and manual photograph processing. Users can browse the species tree smoothly via the hyperlink http://protchem.hunnu.edu.cn/toxin/Browse/Species.htm.

Domain architecture of toxin sequences was predicted by HMMER software (11) based on the deposit of hmm models of the Pfam database version 22.0(12). Data about the linkage pattern of disulfide bridges and regular expression are calculated by an in-house Perl script. The pictures of sequence features for toxins were drawn by Perl script using the GDI model. The IC50 and ID50 values of toxins were downloaded mainly from the toxin databases mentioned above. Other annotation information such as GO annotations, PDB cross-links were extracted from the UniProtKB/Swiss-Prot and TrEMBL and GenBank databases.

TOXIN ONTOLOGY CONSTRUCTION AND ANNOTATIONS

Although previous databases/projects have collected a large number of descriptions about toxin function, most of them are stored in unformatted style, which prevents data search, exchange and comparison. The Toxin Ontology project provides a conserved, structural and controlled vocabulary to describe toxin function and attributes in any organism. It contains more than 745 terms with distinct definitions in four term spaces. As Gene Ontology (GO) (13), which has three term spaces to handle different perspectives of gene functions, Toxin Ontology (TO) contains four term spaces for answering four biological questions: (i) what is the category of the toxin(s)? (Category); (ii) what is the biological activity of the toxin(s) (Bio-Activity); (iii) what is the target of the toxin(s) (Target); (iv) what are the symptoms induced by toxin(s) (Symptom). Figure 2 shows the top level structure of TO:

Category

This term space describes the issue of toxin classification including two top branches: functional categories and species categories. The first one is based on molecular

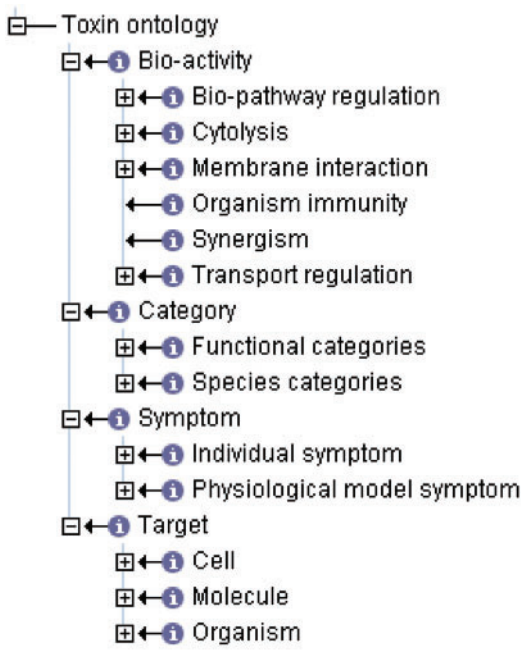

Figure 2. The top-level structure of Toxion Ontology. It contains four term spaces to handle different aspects of toxin functions. Detailed descriptions about it can be found in main text.

functions across species. The second one follows species classification at top level and then follows the characterization of structure or function, which are accepted by related communities.

Bio-activity

This term space covers most of the mechanisms by which the toxins take effect, such as cytolysis, membrane interaction, channel transport regulation, vesicle transport regulation.

Target

This term space has three branches to describe the targets of toxin. The Organism branch mentions the species or tissues affected by toxin. The 'Mammal' term (TX : 0000075) assignment to a toxin means the toxin can act on mammals. The Cell branch describes the type of cell and organelles affected by toxin. The Molecule branch contains detailed classification of the molecules, which interact with toxins such as enzymes, GPCRs (G protein-coupled receptor) or ion channels.

Symptom

This term space has two branches. The first one (individual symptom) describes the symptoms that appear in an individual animal. These effects are divided into two parts: local/regional effects and systemic effects. The other branch covers physiological model symptoms which records the symptoms of certain physiological preparations (such as nerve–muscle preparation) induced by a toxin.

TO annotation (TOA) for all toxins is time-consuming work. Up to now, more than 8423 TO terms have been manually assigned to 2132 toxins based on annotations of the Tox-Prot, GO annotations and related publications. Each term assignment was independently reviewed by at least two biologists to avoid artificial errors. Additionally, we defined five TO evidence codes to describe how these annotations were assigned and what is the type of the evidence to support an annotation.

THE ATDB WEB INTERFACE

ATDB provides a user-friendly web interface for data query, visualization and analysis. Users can query ATDB with toxin ID (ID of ATDB), protein name, gene name, nucleic acid ID and key words. A more complex query can be conducted by 'Advanced text search', which allows users to query the database by species group, taxonomic ID, domain name, the number of disulfide bridges, target and sequence length as well as molecular weight. When querying in this way, the search scope can be narrowed down or expanded using more features by choosing the relationship among diverse features in conjunction ('AND') or combination ('OR'). Users can also BLAST the ATDB by inputting sequences they are interested in to find toxin homologs.

To facilitate data access through the species and TO trees, four integrated tree views were designed with similar formats (Figure 3). They include a left tree and a right table. In the default setting, the tree is compact presenting only terms assigned to the toxin. To see the complete tree, the user can click the 'Complete' hyperlink on the top. One can select and expand the branches of the tree by clicking nodes (terms). At the same time, details about the term will be shown in the right table. All toxins related to the term can be accessed just by clicking the 'getSequence' button and an entry list will be displayed. Filtering and selecting can be done manually via a filter through keyword matches. The selected sequences can be downloaded smoothly in Excel or FASTA format by clicking the 'Excel download' and 'Fasta download' buttons, respectively.

The web interface also includes brief but comprehensive help material in a News/FAQ page. A short introduction about TO is presented on the 'TO introduction' page. All data stored in ATDB can be downloaded freely as flat file and SQL script for MySQL 5.0. The OBO file of TO and TOA file are also available from the 'Download' page. Researchers who intend to provide raw data or suggestions are encouraged to contact authors through the 'Submission' page in the web interface.

FUTURE DEVELOPMENT

There is a major release of ATDB every 3 months with incremental updates as appropriate. Current and future work includes populating the database with more data entries. The system of TO will be further examined and optimized to accommodate the development of toxinomics research. Additionally, it is planned to integrate the multi-alignment tool ClusterW (14) into ATDB for fast sequence comparison.

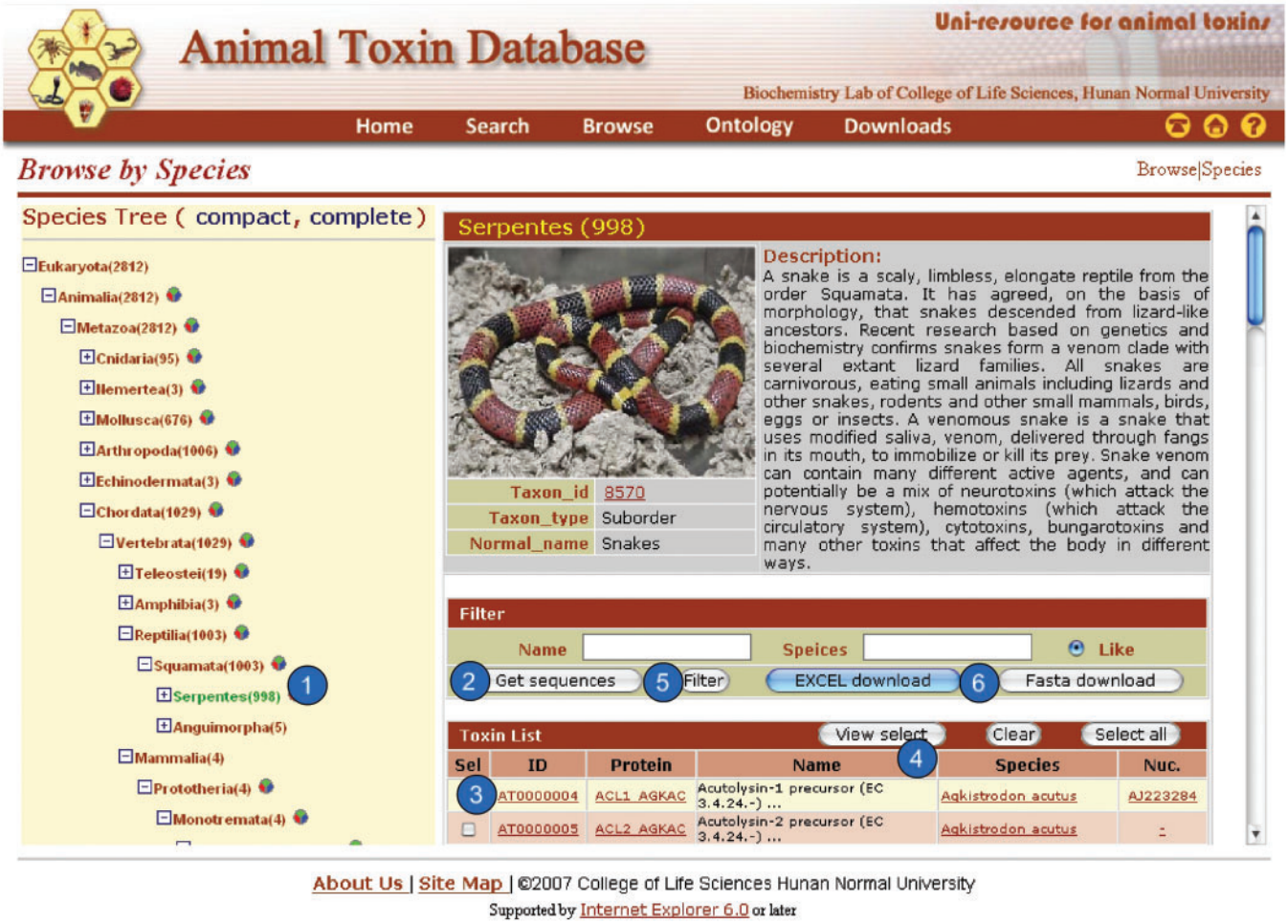

Figure 3. The tree view of species. It includes a left tree and a right table. (**1**) Users can focus and expand the branches of the tree by clicking leaves (*Serpents* suborder in this figure) and detailed information about the taxonomic group will be shown in the right table. (**2**) If you want to get all toxins related to the term, just click the 'getSequence' button to display the toxin list. (**3**, **4** and **5**) Users can select toxins manually and filter them by keywords via a filter. (**6**) The selected sequences can be downloaded smoothly as Excel file and FASTA file by clicking the 'Excel download' and 'Fasta download' buttons, respectively.

ACKNOWLEDGEMENTS

We thank Dr David J. Studholme (The Sainsbury Laboratory, UK), Dr Jing-Chu Luo (Peking University, China) and Dr Xiang-Hua Liu (The University of Texas, Health Science Centre at Houston, USA) for suggestions and comments on the manuscript. This work was partly supported by National Science Foundation of China (30430170 and 30670640) and National Laboratory of Protein Engineering and Plant Genetic Engineering at Peking University. Funding to pay the Open Access publication charges for this article was provided by National Science Foundation of China (30430170 and 30670640).

Conflict of interest statement. None declared.

REFERENCES

1. King,G.F. (2004) The wonderful world of spiders: preface to the special Toxicon issue on spider venoms. *Toxicon*, **43**, 471–475.
2. Escoubas,P. and Rash,L. (2004) Tarantulas: eight-legged pharmacists and combinatorial chemists. *Toxicon*, **43**, 555–574.
3. Tedford,H.W., Sollod,B.L., Maggio,F. and King,G.F. (2004) Australian funnel-web spiders: master insecticide chemists. *Toxicon*, **43**, 601–618.
4. Conticello,S.G., Gilad,Y., Avidan,N., Ben-Asher,E., Levy,Z. and Fainzilber,M. (2001) Mechanisms for evolving hypervariability: the case of conopeptides. *Mol. Biol. Evol.*, **18**, 120–131.
5. Kozlov,S., Malyavka,A., McCutchen,B., Lu,A., Schepers,E., Herrmann,R. and Grishin,E. (2005) A novel strategy for the identification of toxinlike structures in spider venom. *Proteins*, **59**, 131–140.
6. Liao,Z., Cao,J., Li,S., Yan,X., Hu,W., He,Q., Chen,J., Tang,J., Xie,J. *et al.* (2007) Proteomic and peptidomic analysis of the venom from Chinese tarantula Chilobrachys jingzhao. *Proteomics*, **7**, 1892–1907.
7. Jungo,F. and Bairoch,A. (2005) Tox-Prot, the toxin protein annotation program of the Swiss-Prot protein knowledgebase. *Toxicon*, **45**, 293–301.
8. 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.
9. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

10. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R. *et al*. (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **35**, D5–D12.
11. Karlin,S. and Altschul,S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl Acad. Sci. USA*, **87**, 2264–2268.
12. Finn,R.D., Mistry,J., Schuster-Bockler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A. *et al*. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res*., **34**, D247–D251.
13. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S. *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet*., **25**, 25–29.
14. Thompson,J.D., Higgins,D.G. and 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. *Nucleic Acids Res*., **22**, 4673–4680.