Bioinformatics for venom and toxin sciences

Paul T. J. Tan, Asif M. Khan and Vladimir Brusic

Date received (in revised form): 23rd January 2003

Abstract
Venomous animals produce a myriad of important pharmacological components. The individual components, or venoms (toxins), are used in ion channel and receptor studies, drug discovery, and formulation of insecticides. The toxin data are scattered across public databases which provide sequence and structural descriptions, but very limited functional annotation. The exponential growth of newly identified toxin data has created a need for better data management. Venominformatics is a systematic bioinformatics approach in which classified, consolidated and cleaned venom data are stored into repositories and integrated with advanced bioinformatics tools for the analysis of structure and function of toxins. Venominformatics complements experimental studies and helps reduce the number of essential experiments.

INTRODUCTION
The focus of venominformatics is the bioinformatics-driven acquisition, manipulation and analysis of venom data. Various animals produce venom in specialised glands, for efficient hunting of prey and for defence against predators. Venomous animals are diverse and include, among others, species of jellyfish, cone snails, bees and wasps, spiders, scorpions, fish, snakes and even platypuses. They deliver venom through biting or stinging.1 This review will focus mainly on venom-derived protein toxins (venom-toxins) from cone snails, scorpions and snakes. Venom is a complex concoction of toxins – proteins (most of which exhibit enzymatic activity), amines, lipids and other components.2,3 Toxins are highly active molecules that target various cellular receptors and assist in prey digestion.4 The probable ancestral function of venoms was enzymatic activity involved in prey digestion; however, in some animals including marine stingers, cone snails, spiders, scorpions and snakes, venom glands have evolved to produce potent toxins.4 Mortality and morbidity due to envenomation are significant problems in developing countries.5 For example, more than 10,000 deaths per year due to snake bites have been reported in Nigeria.6 Venom-toxins display a wealth of pharmacological properties and exert their effects through interaction with a diverse range of targets, such as various cellular receptors,4 membranes7 and ion channels.8 Venom-toxins are highly active and have multiple uses. For example, they have been used as research tools for characterisation of various ion channels and receptors,3,4,14 development of vaccines5 and antitoxins,6,17 therapeutic agents and drugs,1,18–24 allergy treatment,25 investigation of blood coagulation mechanisms,26 formulation of insecticides27,28 and in evolutionary studies.29–31 An effective drug for treatment of severe chronic pain, the Ziconotide, was developed from an omega-conotoxin found in the venom of the cone snail Conus magus.23

Established methods for determining specific functions of venom-toxins are based on the experimental studies of naturally occurring peptides,32 site-directed mutagenesis33 or use of chemically modified variants.34 The pharmacological properties of venom-toxins are tested in animal models such as mouse, rat or insects. The
experimentation is often supported by computational algorithms for sequence comparison or for modelling of venom-toxin three-dimensional (3D) structure. Systematic functional study of even one individual toxin requires a significant experimental effort. Consequently, most research groups focus on determining functional properties of individual toxins or small groups of toxins. Bioinformatic analyses can improve the efficacy of research by assisting in selection of critical experiments. Bioinformatic approaches involve access to venom data across multiple databases, inspection for errors, analysis and classification of venom-toxin sequences and their structures, and the design and use of predictive models for simulation of laboratory experiments.

Rough estimates of the number of different venom components in an individual snake, scorpion or cone snail is in the range of 50–200 different toxins. Given the variety of venomous species, the natural library of venom-toxins is therefore estimated to contain millions of different toxins. Venom-toxins belong to a relatively small number of structural families that display a variety of functional properties. Public protein and DNA databases contain toxin entries representing only a tiny fraction, which is currently (November 2002) less than 1 per cent of the estimated natural library.

With sequencing projects involving expressed sequence tags, the accumulation of venom-toxin data entered the exponential growth phase. Recently, a set of 170 conotoxin sequences were deposited into GenBank, which almost doubled the number of public conotoxin entries. However, none of these sequences had any functional or structural annotation.

Sequence and structure data on these toxins are usually deposited in public repositories such as GenBank, EMBL, DDBJ, SWISS-PROT, PIR and PDB (published annually in the Nucleic Acids Research database issue). Functional properties of venom-toxins are found mainly in published articles, while functional annotations of entries in public sequence databases are very limited. The growing number of well-characterised toxins and the high complexity of related information, such as their structural and functional properties, have created a need for improved data management. Studies using cDNA libraries extracted from venom glands have produced large numbers of sequences for which functional properties are not known.

A recent analysis of scorpion toxin data from public databases revealed the presence of numerous errors in the sequences, incomplete data, poor annotation and discrepancies of information for the same entry from different database sources. The examples of errors that were found in the scorpion entries in major sequence databases are: wrong links between databases, different names for the same sequence, different sequences for the same toxin, missing links between databases, toxin names from journals not used in the database, and SWISS-PROT links to PDB structures of poor homology. We observed similar problems with entries of other venom-toxins from cone snails and snakes.

This review focuses on bioinformatics applications for management of venom-toxin data and the resources for systematic bioinformatics study of their structure–function relationship. The field of venominformatics, a marriage between computer science and venom biology, is in its infancy, but it has the potential to revolutionise the way that researchers manage venom-related data and information. The number of venom-toxins that are structurally and functionally characterised is currently small and measures only in hundreds. However, with advances in the field and the expected rapid growth of toxin data, experimental approach will need to be combined with theoretical analyses. The bioinformatics-based approach facilitates experimental design and selection of key
experiments, and provides means for systematic study of large number of venom-toxins.

**DATA AND INFORMATION ON VENOM-TOXINS**

Information describing venom-toxin sequences and the features, structure and function of venom-toxins is scattered across multiple sources. Primary databases such as GenBank or SWISS-PROT contain only basic sequence information. Each entry contains the toxin sequence, its name, taxonomy of the source organism, and when available, a list of features and references. Public databases contain highly redundant data entries. For example, we found that, of all snake venom phospholipase A2 (svPLA2) toxin entries in the GenBank and SWISS-PROT databases, 55 per cent were redundant and needed to be filtered out prior to the analysis.

The structural database PDB contains 3D information of approximately 200 venom toxins (as of November 2002). Secondary databases, such as Pfam, SCOP and BLOCKS, collectively contain very limited, if any, functional information on venom toxins. Most functional data are found only in the main text of published journal articles. A centralised well-annotated repository of venom-toxins, which would be useful for study of their structure–function relationship, does not exist. Specialised databases attempt to collate structural and functional information of venom-toxins in a single repository and enable advanced data analyses. Nomenclature of venom-toxins is generally not well developed. Sequence comparisons, phylogenetic studies and functional groupings were used for defining nomenclature and classification of scorpion toxins and conotoxins. These nomenclatures attempt to provide a unique informative name to each characterised venom-toxin which indicates species of venomous animal and cysteine pairing. However, with the growing number of characterised toxins, these nomenclatures will require updating.

Venom-toxins often display multiple functional effects. Several snake venom phospholipase A2 (svPLA2) toxins, for example, have multiple molecular targets and were reported to induce more than one pharmacological effect (shown in Table 1). Standard bioinformatics tools such as sequence comparison, profile analysis or motif identification are not sufficient for accurate analysis of functional properties of venom-toxins. Consequently, advanced tools are required for detailed structure–function analysis of venom-toxins. A clean and comprehensive collection of venom data enriched with structural and functional information provides means for more detailed structure–function analysis.

**MOLECULAR DATABASES OF VENOM-TOXINS**

To our knowledge, only three specialised venom databases are currently available as major resources for the study of venom-toxins. The databases contain entries collected from different sources, cleaned, organised, analysed and classified according to their structure–function relationship. The SCORPION database of nearly 300 entries of scorpion toxin sequences are annotated and classified according to their structural and functional properties. The MOLLUSK database contains more than 450 peptides from the cone snail venoms where each entry has a unique field to facilitate comparison of conotoxin entries. Functionally annotated entries of svPLA2 toxins are found in the svPLA2 database. The following general steps were carried out to create these specialised databases:

- **Data classification step:** the existing classification system of a particular group of venom-toxins was used as a basis for definition of new, more detailed classification and definition of representative groups of venom-toxins based on their primary structures,
molecular targets and functional properties.

Data extraction step 1: extraction of entries from primary databases using keywords and forming the initial data set.

Data extraction step 2: sequence similarity searches to extract entries that were missed by keyword search (approximately 1 per cent of the size of initial data set for svPLA2).

Data cleaning step: elimination of redundant and duplicate entries.

Data enrichment step 1: adding functional annotations from literature (e.g., molecular target, binding affinity, pharmacological properties).

Data enrichment step 2: adding structural data such as PDB structure models, homology models or cysteine pairing to data entries.

Final step: creation of a searchable database.

Present classifications of venom-toxins do not facilitate correlation of structure to function as most schemes focus on addressing only a single feature of a toxin. For example, there are multiple classifications of svPLA2s, each using a different criterion: mechanism of catalysis and requirement of certain ions,54 pharmacological effects,55 peptide length,54 disulphide pattern,56,57 molecular mass,54 cellular location (secreted or cytoplasmic),54,58 or primary sequence similarity.54,59 Therefore, we defined representative groups of venom-toxins to facilitate structure–function studies.

Venom data were collected from major public databases using relevant keyword and sequence similarity searches for comprehensive data collection (initial data set). For example, keyword expression ‘serpentes AND phospholipase OR PLA2’ was used for extraction of svPLA2. Next, data were filtered to remove redundant entries (duplicates and partial entries that are earlier versions of another entry) from the initial data set, which are common in public databases. The entries were cross-referenced with original literature and databases to check for discrepancies in sequences, their names and structural and functional properties. The data-cleaning step is essential to prevent propagation of errors. Functional information from experimental studies was extracted from literature to enrich the entries.

These three toxin databases have several search or data extraction tools to facilitate analysis: Keyword search, BLAST search using BLAST 2.0 algorithm, Structure viewer, Download entries and Annotate scorpion tool. The Keyword search feature allows users to search the databases using keywords, while the BLAST Search enables users to perform sequence similarity search against the individual database. The 3D structures of venom toxins can be viewed through Structure viewer and their sequences can be downloaded in a FastA file format, using the Download feature. The novel features of these databases include:

- 3D structure viewer, which is uncommon in non-structural database;
- functional annotation of entries;
- intelligent structure and function.
prediction tools (SCORPION and MOLLUSK databases).

The SCORPION database is an example of a platform where several standard and advanced bioinformatics applications have been integrated in a single data warehouse for analysis of scorpion toxins. It contains fully referenced data, including peptide sequences, nucleotide sequences, 3D structures, and related structural and functional annotations of scorpion toxins. The database has a standard sequence alignment BLAST search tool, and an intelligent prediction tool that is able to ascribe putative functional annotation to a query sequence.

BIOINFORMATICS APPLICATIONS

Commonly used bioinformatics methods for analysing venom-toxin data are:

- phylogenetics analysis;
- multiple sequence alignments;
- 3D structure analysis; and
- homology modelling.

**Phylogenetic analysis**

Phylogenetic analysis has been used to determine diversification of conotoxins, and classification of scorpion and snake toxins. The analysis of the 3D structures provides a complementary approach to site-directed mutagenesis for identification of functional residues in venom toxins.

**Multiple sequence alignment**

Multiple sequence alignment was used to identify related function of two novel defensins from scorpion venom. The analysis of the 3D structures provides a complementary approach to site-directed mutagenesis for identification of functional residues in venom toxins.

**Analysis of 3D structures**

Further, the analysis of 3D structures of venom toxins have been used as molecular probes for acetylcholine (ACh) receptors: nAChR (nicotinic), mAChR (muscarinic) and ion channels: potassium, sodium and calcium. The homology modelling has been used for determining the potential binding site to sodium channels.

Researchers are increasingly using combination of these bioinformatics tools to establish structure–function relationship.

**PREDICTION OF STRUCTURE AND FUNCTION OF TOXINS**

Crystallisation of macromolecules is a slow and complex process, which requires optimisation of various interdependent physical, chemical and biological parameters. Therefore, the prediction of 3D structures of proteins from primary structures by comparative protein modelling techniques is an attractive alternative for studying structure–function relationship in large number of toxins. The comparison of homology models with experimentally solved 3D structures of venom-toxins enabled identification of putative functional residues involved in binding and catalytic site, which were subsequently experimentally validated. 3D molecular simulations of toxin–receptor complexes have been used for determination of critical interacting residues on the surface of toxins.

Bioinformatic analyses using one or more computational tools are often performed for prediction of structure and function of a small number of toxin sequences. The majority of scorpion toxin 3D structures determined to date share a common structural motif, called the cysteine-stabilised ß-helix (CSH) fold, except Bjxtr-IT from *Buthotus judaicus* and k-hefutoxins from *Heterometrus fulvipes*. The CSH fold comprises an ß-helix on one surface, two to three ß-sheets on the opposite and the internal volume made up of two to three disulphide bonds. Thus, scorpion toxins are a good example of dissimilar proteins sharing similar structural scaffolds. The CSH-type scorpion toxins have different lengths of loops, and types of turns, resulting in a wide range of pharmacological properties. This makes the prediction of function from structure (primary, secondary and 3D) alone a difficult task.
The scorpion toxin functional prediction module, Annotate Scorpion, automatically generates putative functional annotation for the query toxin and places the query sequence into an appropriate structural group. The module predicts toxin type, toxin action, target receptor and target cells (eg insect-, crustacean- or mammal-specific). This module is data-driven and thus statistical in its nature. The Annotate Scorpion module combines sequence comparison, nearest neighbour analysis and heuristics to assign a putative membership of a query sequence to a functional or structural group. This module provides multiple sequence alignment of the test sequence along with the nearest neighbour sequences available in the database. High accuracy predictions of target receptor (91.5 per cent), toxin action (83.3 per cent) and toxin type (68.9 per cent) were achieved on a test set of newly characterised scorpion toxins (Tan et al., manuscript in preparation).

Because it uses nearest neighbour analysis, the Annotate Scorpion tool has high specificity and only predicts scorpion toxins. Similar bioinformatics-based approach of predicting structure and specific function can be applied to other toxins as demonstrated by the Analysis module in MOLLUSK database.

The accuracy of the prediction modules is limited by the availability of present data. Nevertheless, as more new venom sequences are characterised and included into the databases, the accuracy of the structure–function prediction tools is expected to improve.

LARGE-SCALE STUDIES
Genomes of venomous animals have not yet been sequenced. The honeybee genome project only started in 2002, while other venomous animals are not yet under consideration. Large-scale studies for identification of expressed sequences started producing a significantly large amount of unannotated sequence data. The cDNA libraries constructed with mRNAs isolated from venom glands have been used for sequencing venom-toxins in scorpions, snakes, spiders and cone snails. Some of these projects have resulted in hundreds of new venom-toxin sequences.

Venom-toxins are subject to accelerated mutation and hypervariability. The mutations appear to be directed towards modifying molecular surface, rather than random substitutions. Such great variation will require genomics studies aimed at better understanding of molecular diversity among venom-toxins, their structure and function. We anticipate that large-scale studies of venom-toxins will combine identification of expressed sequences with proteomic approaches.

CONCLUSION
The number of newly identified venom-toxin sequences is growing fast, while their functional characterisation is lagging. Laboratory experiments involving large numbers of sequences are extremely costly and time-consuming. Conventional experimental methods for structure–function analysis, such as site-directed mutagenesis, chemical modifications and functional study of naturally occurring variants, together with bioinformatics-driven structure–function study of venom-toxins enable the identification of key experiments and help optimisation of experimental design.

Specialised toxin databases have been developed to bridge the gap between a fast growth of data and the slower pace of experimental validation studies. The utilisation of knowledge, particularly of functional information, has been made more efficient by enriched descriptions of venom-toxin entries and integration of bioinformatics tools to facilitate comprehensive analysis of the data. With the growing number of toxin data being characterised, the bioinformatics analysis will become increasingly important for the management and analysis of venom-toxin data, both in the form of specialised databases and specialised tools for analysis.
of these pharmacologically important proteins.

The uses of 3D modelling and molecular threading have been major tools in the structure–function analysis of venom-toxins. More than 200 3D structures of toxins have been determined and deposited in the PDB database. Therefore, homology models of high reliability can be created for toxins that lack experimentally determined 3D structural information. 3D models of these toxins, coupled with multiple sequence alignment of each group in the venom-toxin classification, can be employed to investigate the functional residues of each pharmacological activity. These 3D homology models can be stored in the databases for use in visualisation, and analysis of active sites.

With the current major expansions in the types of biological data, bioinformatics is essential for data interpretation, understanding biological processes, and suggestion of key necessary experiments. Computational algorithms will play an increasingly important role in the formation and testing of hypotheses.

Though in its infancy, venominformatics offers a solution to data management in an emerging field, and bears the potential to expedite drug discovery process. Venominformatics can help fast characterisation of potential drug candidates, and accurate prediction of functional properties of novel toxins based on existing data. Because venom-toxins are functionally diverse, but belong to a limited number of structural families, they are ideal for application of data mining techniques for discovery of previously unknown relationships among data. The venominformatics lessons will be useful for study of diverse types of active peptides.

References


52. URL: http://sdmc.krdl.org.sg:8080/MOLLUSK/


