FUNCTIONAL PREDICTION OF BIOACTIVE TOXINS IN SCORPION VENOM THROUGH BIOINFORMATICS

TAN THIAM JOO, PAUL
(*B. Appl. Sc. (Hons.), NUS*)

A THESIS SUBMITTED

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

DEPARTMENT OF BIOCHEMISTRY

NATIONAL UNIVERSITY OF SINGAPORE

2005

Acknowledgements

Throughout my Ph.D. candidature, I have been accompanied and supported by friends and family members to complete this thesis. So it is with deep gratitude that I express my heartfelt appreciation to the following:

- Almighty God who has blessed me with gifts and talents to share with others.
- Professor Vladimir Brusic, my supervisor and mentor, whom I owe lots of gratitude. Through his guidance and advice, I have improved on my writing skill and learnt to be an independent researcher. It is also through his faith in me that I have realised my potential.
- Professor Shoba Ranganathan, my co-supervisor, for her valuable advices and support which motivated me to pursue Ph.D.
- Seng Hong, Fahad, ZongHong, Anitha and XuanLinh for their computing assistance in my research.
- Asif, Heiny, Stephanie and Wilson for their critique of my dissertation and companionship during lunch and at I^2R.
- Judice, Chris, Yew Kwang and Lynn for their listening ears and encouragement during difficult times.
- Bernett, Lesheng, Vivek, Victor and Justin, my fellow post-graduate friends for their comradeship.
- My mother, Madam Soong Kim Song, for her perseverance in the face of adversity.
- My family especially my eldest sister, Anna, for their love, encouragement, prayers and support.

My deepest and sincere gratitude,

Paul Tan Thiam Joo

November, 2005

Table of Contents

Acknowledgements. I

Table of Contents… II

Summary……. VI

List of Tables…….. VIII

List of Figures……. IX

Part I: Chapter 1 Introduction 1

1.1 Research issues investigated in this thesis 6

1.2 Contribution of this thesis 8

1.3 A summary of the thesis 9

Part I: Chapter 2 Literature review 11

2.1 Use of bioinformatics to complement experimental studies 12

2.2 Genome sequencing of venomous animals 13

2.3 Sources of toxin data and related information 14

2.3.1 GenBank and GenPept databases 14

2.3.2 Swiss-Prot and TrEMBL databases 15

2.3.3 Protein Data Bank (PDB) 15

2.3.4 PubMed literature database 16

2.3.5 Issues on data collection, cleaning, annotation 16

2.4 Data warehouses of toxins 18

2.5 Bioinformatic tools 18

2.6 Bioinformatic applications 19

2.7 Prediction of structure and function of toxins 20

Chapter summary 22

Part II: Chapter 3 Classification of scorpion toxin data 24

3.1 Classification of scorpion toxins 27

3.2 Data classification of scorpion toxin sequences 29

3.3 Materials and Methods 30

3.3.1 Classification of sequences into groups by BLAST 31

3.3.2 Data classification into subgroups by Clustal W 34

3.3.2 Verification of groups and subgroups by MEGA 3.0 34

3.4 Results – Classification of scorpion toxin sequences 36

3.5 Discussion and conclusions 46

Chapter summary 48

Part II: Chapter 4 Extraction of functional peptide motifs in scorpion toxins 49

4.1 Materials and Methods 51

4.1.1 Scaling of binding affinities to a common scale in mutant toxin data 52

4.1.2 Data analysis 52

4.2 Results and discussion 53

4.2.1 Chloride channel motif 56

4.2.2 Sodium channels – β-excitatory motif 58

4.2.3 Sodium channels – β-mammal motif 60

4.2.4 Sodium channels – α-motif 62

4.2.5 Sodium channels – α-like motif 64

4.2.6 Potassium channel subtype – Ether-a-go-go-related K^+ channel motif 66

4.2.7 Potassium channel subtype – Small conductance Ca^{2+}-activated K^+ channel motif 67

4.2.8 Potassium channel subtypes – Large conductance Ca^{2+}-activated K^+ channel and voltage-dependent K^+ channel motifs 68

4.3 Conclusion 70

Chapter summary 72

Part II: Chapter 5 Functional prediction of bioactive toxins in scorpion venom 73

5.1 Prediction of functional properties of novel scorpion toxins by nearest neighbour analysis, sequence comparison and decision rules 75

5.2 Materials and Methods 76

5.2.1 Scorpion toxin data 76

5.2.2 Algorithm – nearest neighbour and rule-based 77

5.3 Results – Accurate prediction of functional properties of novel scorpion toxins. 79

5.4 Discussion and conclusions 89

Chapter summary 91

Part III: Chapter 6 Implementation of scorpion toxin data warehouse..... 93

6.1 Data warehouse for information usage and knowledge discovery 95

6.2 Implementation of the data warehouse of scorpion toxins, SCORPION2..... 97

6.3 Materials and methods 97

6.3.1 Data collection of native and mutant scorpion toxin sequences and their 3D structures 98

6.3.2 Generation of homology models of scorpion toxins 99

6.3.3 Data cleaning..... 100

6.3.4 Data annotation 100

6.4 Results..... 101

6.4.1 Database description 106

6.4.2 Description of the SCORPION2 records 110

6.5 Discussion and conclusion 114

Chapter summary 116

Part III: Chapter 7 Exploring bioinformatic approaches for functional prediction of bioactive scorpion toxins 118

7.1 Materials and Methods 119

7.1.1 Algorithm for predicting strength of binding affinity of scorpion toxins 121

7.2 Results 121

7.3 Discussion and conclusion 124

Chapter summary 125

Part IV: Chapter 8 General discussion 126

Chapter summary 130

Part IV: Chapter 9 Conclusion 132

9.1 Large-scale classification 133

9.2 Large-scale analysis 134

9.3 Development of functional prediction tool 135

9.4 Data warehouse of scorpion toxins 136

9.5 Evaluation of application of bioinformatics in venom research 137

Conclusion summary 138

9.6 Future works 139

References 142

Author's Publications 171

Appendix 1 172

Appendix 2 193

Summary

Scorpions are venomous animals that produce a myriad of important bioactive toxins that are used in ion channel studies, drug discovery, and even formulation of insecticides. Determining their structure-function relationships are of great interest for scientific, medical and industrial applications. This thesis presents a systematic bioinformatics approach to a large-scale study of structure-function relationships in scorpion toxin sequences. Systematic characterisation of their structural features and functional properties 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 improve the efficacy of research by assisting in selection of critical experiments. Bioinformatic approaches involve access to toxin data across multiple databases, inspection for errors, analysis and classification of toxin sequences and their structures, and the design and use of predictive models for simulation of laboratory experiments.

Several novel aspects are presented in this thesis. This is, to the author's knowledge, the first large-scale classification of currently known scorpion toxins based on ion channel specificity and primary sequence similarity. This classification is important for identification of the general patterns in their structure-function relationships. The author proposed a classification that has defined several new groups of scorpion toxins.

A new approach to extract functionally relevant motifs from scorpion toxins based on analyses of multiple sequence alignment of native scorpion toxin sequences, 3D structures and mutated scorpion toxin data was developed in this work. This approach identified critical functional residues at key positions in the toxin sequences which lack conserved residues in the multiple sequence alignment. The first report of

eight functionally relevant binding motifs to sodium and potassium channels facilitates the determination of specificity of newly identified scorpion toxins to various channel subtypes.

The most important contribution to scorpion venom research is a new bioinformatic tool for accurate identification of functional properties in newly identified scorpion toxins. It was developed from the large-scale analysis of scorpion toxin sequences. The prediction algorithm includes sequence comparison, nearest neighbour analysis and decision rules. High prediction accuracy of ion channel specificity, toxin subtype, toxicity action and cellular specificity was validated by experimental data.

The first database of native and mutant scorpion toxin sequences, developed as part of this work, is a major resource for efficient searching of scorpion toxin-related information. The records were cleaned of errors and contain highly enriched structural and functional information extracted from the literature. The 548 new homology models contribute to three-dimensional analyses of scorpion toxins. Integration of search, extraction, prediction and three-dimensional visualisation tools allows researchers to analyse scorpion toxin sequences efficiently.

The bioinformatics approach employed in this study is novel, generic and applicable for the studies of structure-function relationships of bioactive toxins from other venomous organisms. Because 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.

List of Tables

Table 1 Examples of venomous animals living on land and at sea.. 2

Table 2 Different criteria can be used to classify scorpion toxins. 28

Table 3 Summary of the classified groups for 393 scorpion toxins.. 37

Table 4 Classification of 135 K^+ scorpion toxin sequences.. 40

Table 5 Classification of 222 Na^+ scorpion toxin sequences.. 44

Table 6 Motifs of scorpion toxins extracted for Na^+, K^+ and Cl^- channels.. 55

Table 7 Functional properties predicted for the first test set of 52 new toxin sequences.. 81

Table 8 Functional properties predicted for the second test set of 127 new toxin sequences.. 86

Table 9 A summary of 82 scorpion toxin PDB structures in SCORPION2 104

Table 10 Description of fields in a SCORPION2 record. 111

Table 11 Four categories of strength of binding affinity. 120

Table 12 Predicted ion channel specificity and strength of binding affinity for 26 newly identified scorpion toxins. 122

Table 13 Physical properties of the 20 L-α-amino acids. 195

List of Figures

Figure 1 The 3D structures of scorpion toxins. 21

Figure 2 Flowchart of the large-scale classification of scorpion toxin data. 31

Figure 3 Classification of scorpion toxin sequences into groups using BLAST 33

Figure 4 Classification into subgroups using Clustal W 35

Figure 5 Verification of groups and subgroups by phylogenetic analysis.. 36

Figure 6 Phylogenetic tree of representative scorpion toxins. 38

Figure 7 Representative scorpion toxins from K^+ subfamilies. 41

Figure 8 Multiple sequence alignment of γ-KTx toxins.. 42

Figure 9 Multiple sequence alignment of CsEv1, Cn5 and CssII 43

Figure 10 Representative scorpion toxins from Na^+ toxin groups 1 – 18. 45

Figure 11 Representative scorpion toxins from Ca^{2+} toxin groups 1 – 4. 45

Figure 12 Scaling bindng affinities of Agitoxin 2 and its mutant sequences 54

Figure 13 Conserved residues of 18 Cl^- specific scorpion toxins 57

Figure 14 Cl^- specific scorpion toxins adopt the cysteine-stabilised α-helix fold 57

Figure 15 Conserved residues of 19 Na^+ β-excitatory toxins 59

Figure 16 Functional motif of β-excitatory toxins 60

Figure 17 Conserved residues of 13 experimentally determined β toxins. 61

Figure 18 Spatial organisation of the functional residues of Css 4. 62

Figure 19 Conserved residues of 14 experimentally determined α-toxins. 63

Figure 20 Functional and structural residues of Lqh αIT. 63

Figure 21 Functional and structural residues of BmK M1 64

Figure 22 Conserved residues of eight experimentally determined α-like toxins 65

Figure 23 Functional residues of BeKm-1 66

Figure 24 Functional residues of scorpion toxins targeting small conductance Ca^{2+}-activated K^+ channels.. 67

Figure 25 Functional residues of charybdotoxin 69

Figure 26 Mutiple sequence alignment of scorpion toxins targeting voltage-dependent K^+, large and small conductance Ca^{2+}-activated K^+ channels 69

Figure 27 Accuracy of functional prediction of *Annotate Scorpion* module. 80

Figure 28 Statistics of SCORPION2 database as of November 2005. 103

Figure 29 Number of records having errors or discrepancies. 103

Figure 30 Site map of SCORPION2 database.. 106

Figure 31 The web interface of the SCORPION2 database. 107

Figure 32 BLAST result upon submission of maurotoxin. 108

Figure 33 Visualisation of scorpion toxin 3D structures using Jmol.. 109

Figure 34 Flowchart of predicting ion channel specificity and strength of binding affinity. 120

Figure 35 Predicted binding affinity of KTX3 from *Buthus occitanus tunetanus*..... 123

Figure 36 Predicted binding affinity of AmmVIII from *Androctonus mauretinicus mauretinicus*.. 124

Figure 37 Venn diagram of the 20 naturally occurring amino acids based on their physicochemical properties.. 194

Part I: Chapter 1 Introduction

‘Man's mind stretched to a new idea never
goes back to its original dimensions.’

Sri da Avabhas (Adi Da Samraj)

1. Introduction

Scorpions are among the first land animals. They appeared some 450 million years ago (Briggs, 1987). There are more than 1,500 distinct species world-wide, living in every continent except Antarctica (Lourenco, 1994). All scorpion species produce venom which they use for hunting prey and defense against predators. Venom is a complex mixture of toxins – proteins, amines, lipids and other components (Martin-Eauclaire and Couraud, 1995). Venom-derived protein toxins are highly bioactive molecules belonging to a relatively small number of structural families. They display a variety of functional properties which include interaction with cellular receptors, ion channels, and assisting in prey digestion (Maslennikov *et al*., 1999; Kini, 2002; Zhu *et al*., 2003; Zhu *et al*., 2004a). The likely ancestral function of venoms was enzymatic activity involved in prey digestion, however, in some venomous animals including scorpions, their venom glands have evolved to produce potent toxins (Valentin and Lambeau, 2000) (**Table 1**).

Table 1 Examples of venomous animals living on land and at sea. Unlike poisonous animals (e.g. toads, puffer fish) which have toxins but have no method of delivery, venomous animals have specialised organs to deliver their venoms.

Terrestrial	Organs	Marine	Organs
Ant	Stinger/bite	Blue ring octopus	Bite
Assassin bug	Proboscis	Cone snail	Proboscis
Bee	Stinger	Coral	Tentacle
Black widow spider	Fang	Crown-of-thorns starfish	Spine
Centipede	Fang	Cuttlefish	Bite
Duck-billed platypus (male)	Spike	Jellyfish	Tentacle/Stinger
Gila monster	Bite	Sea anemone	Tentacle/Stinger
King cobra	Fang	Sea urchin	Spine/pedicellaria
Komodo dragon	Bite	Scorpion fish	Spine/Stinger
Rattlesnake	Fang	Stingray	Spine/Stinger
Scorpion	Stinger	Stonefish	Spine/Stinger
Wasp	Stinger	Yellow-lipped sea krait	Fang

Mortality and morbidity from animal envenomation remains a serious health issue (Theakston *et al*., 2003), accounting for more than 150,000 deaths per year (White, 2000). However, venoms also contain highly bioactive compounds for discovery of molecules with interesting pharmacological properties and potential therapeutics for an array of medical disorders (Alonso *et al*., 2003; Bradbury, 2003; Lewis and Garcia, 2003; Rajendra *et al*., 2004). An assortment of highly bioactive toxins characterised by high specificity and selectivity are used as research tools to characterise different ion channels subtypes and molecular isoforms of receptors (Grant *et al*., 2004; Li and Tomaselli, 2004; Rodriguez de la Vega and Possani, 2004; Lewis, 2004; Tsetlin and Hucho, 2004). Analyses of the interfaces between toxins and their channels/receptors facilitate design of synthetic equivalents of toxins without toxic properties which can be developed as potential therapeutics.

Rapidly emerging knowledge from studies of the molecular mechanism of the ion channels is used in the development of novel therapeutics for ion channel-related diseases such as epilepsy, cardiac arrhythmia and persistent pain syndromes (Curran, 1998; Catterall, 2002; Kohling, 2002; Wickenden, 2002a; Wickenden, 2002b; Wulff *et al*., 2003; Gottlieb *et al*., 2004). Therapeutics successfully developed from studies of animal venoms include Ancrod and Captopril that were developed from snake venom for treatment of hypertension and cardiac failure (von Segesser *et al*., 2001; Smith and Vane, 2003). Another example is Ziconotide, developed from marine cone snail venom, for treating severe chronic pain (Miljanich, 2004). Antivenoms are currently developed from animal antisera to treat envenomation (Harrison, 2004; Gazarian *et al*., 2005). Animal venoms are promising alternatives to chemical pesticides in agricultural pest management. The increased pest resistance to chemical pesticides, coupled with heightened awareness of the potential environmental, human and animal health

impacts of these chemicals, have prompted the search for development of bio-pesticide from animal venoms (e.g. Sun *et al*., 2002; Gilles *et al*., 2003; Szolajska *et al*., 2004). Identification of new toxin sequences and determination of their functional sites and structural properties is therefore of great interest and value for scientific, medical and commercial applications.

The number of different venom components in an individual scorpion consists of approximately 100 different toxins (Lourenco, 1994). Given 1500 scorpion species exist, the natural library of scorpion toxins is therefore estimated to contain some 100,000 different toxins (Lourenco, 1994). However, toxin entries in public protein and DNA databases represent only a tiny fraction, less than 1% of the estimated natural venom library (as of November 2005).

Sequence and three-dimensional (3D) structure data on these toxins are usually deposited in public repositories such as GenBank (Benson *et al*., 2005), Swiss-Prot (Bairoch *et al*., 2004) and Protein Data Bank (Deshpande *et al*., 2005). Functional and structural properties of toxins are reported mainly in published articles, while such annotations of entries in public sequence databases are very limited (Brusic *et al*., 2000). Advances in sequencing projects involving cDNAs and mass fingerprinting by mass spectrometry resulted in exponential accumulation of toxin data (e.g. Batista *et al*., 2004; Davies *et al*., 2004; He *et al*., 2004). For instance, a set of 170 conotoxin sequences were deposited into GenBank in 2001 (Conticello *et al*., 2001) which almost doubled the number of public conotoxin entries at that time. However, none of these sequences had any structural or functional annotations, only sequences were reported. Experimental characterisation [1] of structure-function relationships for the many

[1] Throughout this thesis, term 'characterised' describes procedures of laboratory-bench work or wet-lab experimentation.

individual toxin sequences is laborious, expensive and time-consuming. Increasingly, researchers are exploiting bioinformatics to expedite characterisation of the growing number of newly identified toxin sequences through information gathered from toxin data scattered in public repositories and the literature.

Bioinformatics is an interdisciplinary field incorporating computer science, mathematics and biology, for management and analysis of biological data. The main branches of bioinformatics are: 1) biological databases, 2) analysis and interpretation of biological data, and 3) development of analysis tools and algorithms. The biological databases, tools and algorithms are important methodologies in scientific research especially in genomics and proteomics, which generate huge amounts of data. These data are stored in biological databases which continue to grow in size and complexity where more than 700 biological databases are publicly available (Galperin, 2005). Insights gained from analyses and interpretations of the data are used for the development of new analysis tools and algorithms for analyses of data, and planning and minimisation of the number of further experiments.

This thesis describes original findings from application of bioinformatic-based approach to the large-scale study of structure-function relationships of scorpion toxins. In this thesis, the word 'structure' encompasses primary, secondary and tertiary structures of proteins unless stated otherwise. The current number of scorpion toxins that are structurally and functionally characterised is small and measures only in the hundreds, in contrast to the natural library of toxins that is estimated to be 100 times larger. However, with the expected rapid growth of toxin data through large-scale sequencing, experimental approach will need to be complemented with bioinformatic analyses for facilitating characterisation of the large number of newly identified toxin sequences.

1.1 Research issues investigated in this thesis

Large-scale analysis of scorpion toxins provides a global view of the general pattern of their structure-function relationships. This analysis in turns supports experimental studies by assisting in planning of critical experiments and, when properly used, it significantly improves the efficiency of experimental studies of structure-function relationships. However, such large-scale analyses are hindered by inadequate data management where scorpion toxin data are scattered across public databases. Records in the databases typically contain sequence information, while structure-function information is available in the literature. Thus, consolidating the scattered data into a centralised database and enriching the toxin data with structure-function information is a prerequisite for a systematic large-scale analysis. Information gained from such analysis is useful for developing new analytical tools for study of novel toxin sequences and prediction of their structural and functional properties.

The author of this work was earlier involved in building the SCORPION database (Srinivasan *et al*., 2002a) which contained 277 native scorpion toxin sequences. Mutation studies (such as site-mutagenesis) of scorpion toxins, which provide biologically relevant information on critical residues and their positions, are available in the literature and are normally not used for extraction of functional motifs in scorpion toxins.

The original contribution of this work is the systematic application of bioinformatic-based approach to: **1**) build the first comprehensive molecular database of scorpion toxins which includes native and synthetic variants, SCORPION2 for improved data management and detailed analysis, **2**) classify 393 native scorpion toxin sequences into functional groups based on ion channel specificity and sequence similarity using BLAST, multiple sequence comparison and phylogenetic analysis, and

3) develop a prediction tool for predicting the functional properties of uncharacterised scorpion toxins. This database was built using molecular data warehousing principles. It contains subject-orientated (scorpion toxins), integrated (comprehensive information), non-volatile (validated), and expert-interpreted (annotated) collection of biological data (e.g. a family of proteins that have similar structures and functions) (Schonbach *et al*., 2000). To the author's knowledge, data warehousing has not been applied neither to the large-scale management of scorpion toxin data nor the systematic study of their structure-function relationships.

The systematic application of bioinformatics to the study of venoms – venominformatics – is a combination of bioinformatics and venom research which has the potential to revolutionise the way that researchers manage toxin data and information. For example, currently there is no tool available for accurate prediction of functional properties of toxins. This research area is important for prediction of function in newly identified toxins. In general, toxins display an array of diverse functions where detailed examination of their molecular functional sites allows alterations of their pharmacological specificity, selectivity and potency especially in the field of drug design and discovery. Therefore, the specific objectives of this thesis were to focus on scorpion toxins and include the following sub-projects:

1) build a data warehouse of scorpion native and mutant toxin sequences with integrated query, extraction and prediction tools,
2) enrich records with structure and function information extracted from the literature and public information repositories,
3) predict tertiary structures of scorpion toxins by homology modeling for toxins without experimentally determined 3D structures,
4) analyse the toxin dataset (primary, secondary and tertiary structures) for

identification of functional motifs and,

5) develop a tool to predict specific functional properties of newly identified scorpion toxins.

1.2 Contribution of this thesis

The author's original contributions to the field of venom research include:

1) Organised a large and unique data set of 819 entries of scorpion toxin data from public databases and literature, inclusive of 426 scorpion mutant toxin sequences extracted solely from literature to develop the SCORPION2 database. This data warehouse of scorpion toxins is a major resource for researchers to identify scorpion toxins and analyse their sequence which otherwise would involve multiple querying of other databases.
2) Extracted functional information of binding affinity and toxicity data from approximately 500 scientific articles and deposited them into SCORPION2.
3) Classified currently known scorpion toxin sequences into functional groups for a broad view on the general pattern of structure-function relationships. The groupings contain scorpion toxin sequence groups that have not been previously defined and classified.
4) Developed a new approach to extract functionally relevant motifs from scorpion toxins based on analyses of multiple sequence alignment of scorpion toxins, 3D structures and scorpion mutant data. This approach also helped in the identification of critical functional residues at key positions in toxin sequences which lack conserved residues.

5) Developed the first prediction tool, *Annotate Scorpion* which accurately predicts the functional properties of newly identified scorpion toxins. The accuracy was validated by new experimental data. This tool helps reduce the number of experiments needed to characterise their functional properties.
6) Generated 548 homology models of scorpion toxins not available previously and made them publicly accessible for 3D analysis.

1.3 A summary of the thesis

This thesis consists of four parts. Part I provides an introduction to the importance and issues of venom research, and how bioinformatics can facilitate venom research (Chapter 1). A review on venominformatics applications and related information in major public databases, and bioinformatics applications available for analysing large number of toxin data is discussed (Chapter 2).

Part II presents the original findings of the research undertaken in this dissertation which includes a large-scale classification of scorpion toxins by functional properties and primary sequence similarity, for a global view on their general pattern of structure-function relationships (Chapter 3). Functional motifs were extracted from analyses of multiple sequence alignment of scorpion toxins, 3D structures and information from scorpion mutant data (Chapter 4). A new algorithm, based on sequence comparison, nearest neighbour analysis and decision rules to predict the functional properties of novel scorpion toxins, was implemented (Chapter 5). High prediction accuracy was achieved as validated by experimentally characterised scorpion toxin sequences.

Part III describes the implementation of specialised data warehouse of scorpion toxins – SCORPION2 – integrated with bioinformatics tools (Chapter 6). The current limitations of bioinformatics for functional prediction of scorpion toxins was also explored (Chapter 7).

Part IV (Chapters 8 and 9) draws conclusions from the bioinformatic-based approach to large-scale analysis of scorpion toxins and also discusses future directions.

The work presented in this thesis has been published in a series of journal articles. These include: the review on bioinformatics for venom science, Tan *et al*. (2003) – Chapter 2; Tan *et al*. (2006a) – Chapter 4 where functionally relevant scorpion toxin motifs were extracted from the approach of including scorpion mutant toxin data in the analysis; Tan *et al*. (2005) – Chapter 5 where the first functional prediction tool was developed for scorpion toxin research; Tan *et al*. (2006b) – Chapter 6 discussed the data warehouse of scorpion native and mutant toxin data with integrated bioinformatic tools for data analysis.

Part I: Chapter 2 Literature review

‘I do not fear computers. I fear the lack of them.’

Isaac Asimov

Researchers currently spend significant time and effort in searching for all available information on animal toxins because a centralised repository is lacking. Toxin sequences are scattered across numerous public databases. Most of the structural and functional information that can improve our understanding of bioactive toxins is stored in the literature. The scattered toxin data and literature information has created a need for improved data management in the field of toxin research. A data warehouse of toxins serves as a major repository for analysis and interpretation of consolidated, cleaned and enriched toxin data. The data warehouse with integrated bioinformatic tools also facilitates characterisation of the increasing number of newly identified toxin sequences with unknown function.

Venominformatics is a field combining venom biology and bioinformatics. Venom biology generates large quantities of biological data, while bioinformatics provides an effective means to store and analyse large volumes of complex biological data. Combining the two fields provides the potential for great strides in understanding and increasing the effectiveness of venom research. The main goal is the extraction of new knowledge from large-scale analysis of toxin data. The bioinformatic approach provides a means for the systematic study of a large number of toxins, and facilitates experimental design and selection of key experiments. This chapter focuses on resources containing toxin data, bioinformatic applications for analysis of toxin data, and prediction of their structure-function relationships.

2.1 Use of bioinformatics to complement experimental studies

Animal venoms contain a diverse array of bioactive toxins that have a variety of biochemical and pharmacological functions (Kordis and Gubensek, 2000; Fry *et al*.,

2003). Established methods for determining specific functions of toxins are based on experimental studies of naturally occurring peptides (e.g. Inceoglu *et al*., 2002; Zhu *et al*., 2004b), site-directed mutagenesis (e.g. Everhart *et al*., 2004; Ivanovski *et al*., 2004), or use of chemically modified variants (e.g. Chang *et al*., 2004; Chang *et al*., 2005). The pharmacological properties of toxins are tested in animal models such as mice, rats, crustaceans or insects. The experimentation is often supported by computational algorithms for sequence comparison (Wang *et al*., 2003; Cohen *et al*., 2004; Cohen *et al*., 2005) or for modelling of toxin 3D structures (Mourier *et al*., 2003; Benkhadir *et al*., 2004). 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 toxin data scattered across multiple databases, inspection for errors, classification and analysis of toxin sequences and their structures, and the design and use of predictive models for simulation of laboratory experiments.

2.2 Genome sequencing of venomous animals

Currently, genomes of honey bee, sea urchin and duck-billed platypus are being sequenced (http://www.genome.gov/10002154). Large-scale studies of toxins from identification of expressed sequences generate large amount of unannotated sequence data deposited in public databases. For example, 8966 unique putative sequences were assembled from the honey bee brain expressed sequence tag project (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=7460). The cDNA libraries constructed with mRNA isolated from venom glands have been used for

sequencing toxins in scorpions, snakes, and cone snails (e.g. Peng *et al*., 2003; Jiqun *et al*., 2004; Santos *et al*., 2004). Some of these projects have resulted in identification of hundreds of new toxin sequences. For example, 170 novel conotoxins were identified from cone snail expressed-sequence tag assemblage (Conticello *et al*., 2001). Bioinformatics aids in the large-scale studies of toxins where putative function can be assigned efficiently for large number of toxin sequences.

2.3 Sources of toxin data and related information

Toxin data and information are scattered across multiple resources. The data include nucleotide and amino acid sequences, secondary structures and 3D structures deposited in public databases such as GenBank (Benson *et al*., 2005), Swiss-Prot (Bairoch *et al*., 2004) and PDB (Deshpande *et al*., 2005). Structure-function information, particularly mutation studies (such as site-directed mutagenesis), is available in the literature. The advantages and disadvantages of these databases for the creation of data warehouses of toxins would be reviewed in the sub-chapters. The issues of data collection, cleaning and annotation when consolidating the scattered data would also be described.

2.3.1 GenBank and GenPept databases

Toxin data are extracted from GenBank (Benson *et al*., 2005) database because it contains a comprehensive collection of publicly available nucleotide sequences. GenBank encourages direct submissions of new data and batch submissions from large-scale sequencing projects to help maintain accuracy, relevance and comprehensiveness of the database. GenPept protein database contains translated

nucleotide sequences found in GenBank. However, records in these databases contain only basic information such as the toxin sequence, its name, taxonomy of the source organism, and when available, a list of basic sequence features and references. The records need to be enriched with structural and functional information (such as residues important for folding, binding affinity and toxicity information) which is available in the literature (see section 2.3.4).

2.3.2 Swiss-Prot and TrEMBL databases

Toxin data are also extracted from Swiss-Prot and TrEMBL (Bairoch *et al*., 2004) databases because they have a comprehensive collection of protein sequences. Swiss-Prot contains a high level of curated structural and functional information that may include disulfide connectivity, secondary structure information, ion channel specificity and protein family classification, among others. TrEMBL contains computationally annotated translations of all EMBL (Cochrane *et al*., 2006) nucleotide sequence entries not yet integrated in Swiss-Prot. The information in Swiss-Prot and TrEMBL records expedites subsequent annotation when new structure-function information is available.

2.3.3 Protein Data Bank (PDB)

Analysing toxin 3D structures are important because toxin function is related to its structural folding. Inclusion of 3D structural information to toxin sequence analysis facilitates identification of residues that are important for structure and function. As of May 2005, the structural database PDB (Deshpande *et al*., 2005) contains only 82 3D structures of scorpion toxins in contrast to the estimated 100,000 different toxins in the

natural venom library. Because the growth of new scorpion toxin sequences outpaces that of experimentally solved 3D structures, toxin structure prediction is necessary to overcome this disparity. The experimentally solved 3D structures serve as templates for generating homology models of toxin sequences because a majority of scorpion toxins share a common scaffold (Kobayashi *et al*., 1991; Rodriguez de la vega and Possani, 2004, 2005). The generated homology models do not replace, but serve as an alternative to, experimentally determined structures because homology models may not be as accurate as the latter.

2.3.4 PubMed literature database

The wealth of information from the literature is important for interpretation of experiments and predictions. Most structural and functional information of toxin sequences is reported in published literature where abstracts of the published literature can be searched in PubMed (http://www.pubmed.gov/) or similar data sources. Extraction of structure-function information is important for enriching toxin data records, particularly those which have limited or no annotation. The enriched records enable a more detailed analysis in contrast to records with only sequence information.

2.3.5 Issues on data collection, cleaning, annotation

The collection of toxin data from different databases is hampered by different database formats and variations in fieldnames that describe the same information. For example, a toxin primary sequence is described in the 'translation' field of a GenBank record but in Swiss-Prot, it is described in the 'sequence' field. The differences in fieldnames describing the same information need to be standardised to a uniform data

representation. For example, a standard field such as 'translation' can be used to describe toxin primary sequence regardless of data sources. The uniform data representation is critical because consistency is required for efficiency of subsequent analyses.

When consolidating records from different databases, the same data may be duplicated in another database, resulting in data redundancy. Data cleaning involves removing these redundant records to improve on data quality. For example, of all snake venom phospholipase A_2 toxin entries in the GenBank and Swiss-Prot databases, 55% were redundant and needed to be filtered out prior to data analysis (Tan *et al*., 2003). Data cleaning also involves detecting discrepancies in data information, highlighting, and subsequently correcting the conflicts. Some examples include detecting discrepancy in the toxin primary sequence between literature and database, different names for the same sequence and missing links between databases (Srinivasan *et al*., 2002a).

Records in the public databases typically contain basic information. Data annotation, also known as data enrichment or enhancement, is the process of furnishing critical commentary or explanatory notes[1]. Data annotation enriches the data for extrapolation of meaningful insights from multi-source bits of information. Correlating the relevant information from multiple sources is critical for increasing the overall knowledge and for improving the understanding of a specific subject in the data warehouse (Karasavvas *et al*., 2004). It is important to differentiate experimentally determined function from those that have been predicted computationally (Karp *et al*., 2001) because the latter require subsequent validation. This would allow researchers to verify and decrease the propagation of incorrect predicted function during data

[1] http://dictionary.reference.com/search?q=annotation

annotation.

2.4 Data warehouses of toxins

To the author's knowledge, only four toxin data warehouses are currently available as major resources for the study of toxins. The databases contain entries collected from different sources, cleaned, organised, analysed and classified according to their structure-function relationships. The SCORPION (Srinivasan *et al*., 2002a) had 277 entries of native scorpion toxin sequences, annotated and classified according to their structural and functional properties. The SCORPION2 database has 819 entries of native and mutant scorpion toxin sequences annotated with functional information extracted from literature and 624 3D structures. The MOLLUSK[2] database contains 457 peptides from the cone snail venoms where each entry has a unique field to facilitate comparison of conotoxin entries. Functionally annotated entries of snake venom phospholipase A_2 (svPLA$_2$) and neurotoxins (svNTXs) are found in the svPLA$_2$ (Tan *et al*., 2003) and svNTXs (Siew *et al*., 2004) databases, respectively.

2.5 Bioinformatic tools

General bioinformatic tools commonly used in analyses of toxin data include but are not limited to BLAST (Altschul *et al*., 1997) and Clustal W (Thompson *et al*., 1994). The BLAST search tool finds regions of local similarity between query sequences and database sequences by calculating the statistical significance of matches. Uses of BLAST include inferring functional and evolutionary relationships between sequences as well as help identify members of gene families. Clustal W is a

[2] Mollusk database of cone snail toxins. http://research.i2r.a-star.edu.sg/MOLLUSK/

general purpose multiple sequence alignment program for nucleotide or protein sequences. It involves the optimal alignment of the greatest number of identical or similar residues into columns across many nucleotide or protein sequences. Patterns of aligned sequences can be used in the analysis of function, structure and phylogeny relationship between sequences. Phylogenetic tools such as Mega 3.0 (Kumar *et al*., 2004) have been developed as easy-to-use computer programs for inference of evolutionary relationship between sequences which provides a guide to their structure-function relationships. Different homology modeling servers e.g. SDPMOD (Kong *et al*., 2004) and Swiss-Model (Schwede *et al*., 2003) are available to generate homology models of toxins lacking experimental structures.

Specialised tools for analysis of toxins is currently lacking since toxin data needs to be analysed prior to development of analysis tool and such detailed analyses have been of limited scope.

2.6 Bioinformatic applications

Commonly used bioinformatic methods for analysing toxin data are:

- phylogenetic analysis,
- multiple sequence alignments,
- 3D structure analysis, and
- homology modeling.

Phylogenetic analysis has been used to study diversification of scorpion toxins, snake toxins and conotoxins (Conticello *et al*., 2001; Fry and Wuster, 2004; Zhu *et al*., 2004a), and classification of scorpion (Rodriguez de la Vega and Possani, 2004) and

snake toxins (Fry, 2005). Multiple sequence alignment and analysis of their 3D structures provide a complementary approach to site-directed mutagenesis for identification of functional residues in toxins (Chioato and Ward, 2003; Everhart *et al*., 2004; Karbat *et al*., 2004b). Homology modeling has been used in designing mutants to determine function of toxins and computational simulations of ligand-channel/receptor interactions to determine interacting residues and structure guided drug development (Chen and Pellequer, 2004; Dutertre *et al*., 2004; Liu and Lin, 2004; Yu *et al*., 2004). Researchers are increasingly using a combination of these bioinformatic methods to establish structure-function relationships (Bagdany *et al*., 2004; Giangiacomo *et al*., 2004; Karbat *et al*., 2004a; Ramos and Selistre-de-Araujo, 2004; Siew *et al*., 2004).

2.7 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 (McPherson, 1999). Therefore, prediction of 3D structures of proteins from primary structures by comparative analysis and homology modeling techniques is an attractive alternative for studying structure-function relationships in large number of toxins. The comparison of homology models with experimentally solved 3D structures of toxins enabled identification of putative functional residues involved in binding and catalytic site, which were subsequently experimentally validated (Hains *et al*., 1999; Church and Hodgson, 2002; Moreno-Murciano *et al*., 2003). 3D molecular simulations of toxin-receptor complexes have been used for determination of critical interacting residues on the surface of toxins (Grant *et al*., 2004; Wu *et al*., 2004; Yu *et al*., 2004).

The majority of scorpion toxin 3D structures determined share a common structural motif, called the cysteine-stabilised α-helix (CSH) fold (**Figure 1A**). The CSH fold comprises an α-helix cross-linked by three to four disulfide bridges to an extended β-sheets (Kobayashi *et al*., 1991). 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. A new fold, consisting of two short helixes cross-linked with two disulfide bridges, was recently characterised in a new family of weak K^+ toxins (Srinivasan *et al*. 2002b, Nirthanan *et al*., 2005) (**Figure 1B**).

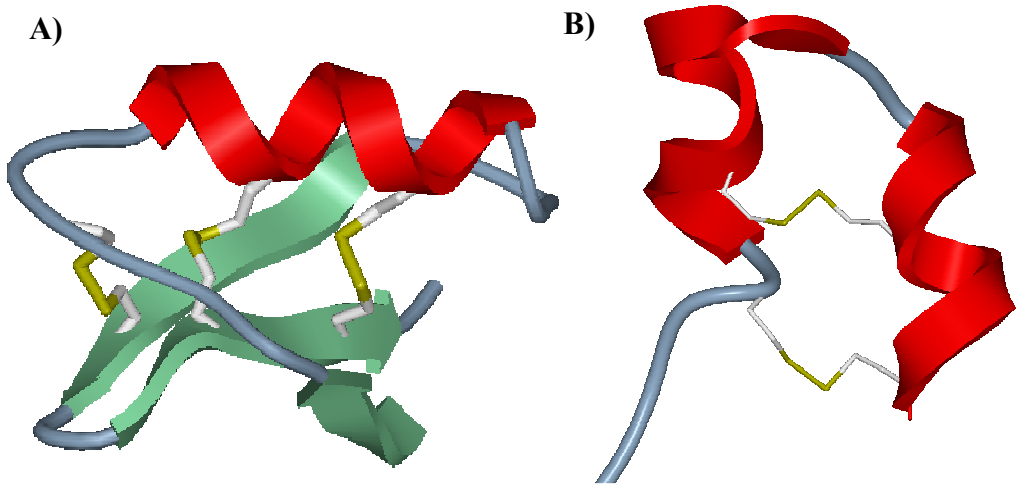

Figure 1 The 3D structures of scorpion toxins. **A**) Cysteine-stabilised α-helix fold was shared by majority of scorpion toxins. The fold consisted of an α-helix and two – three β-sheets cross-linked by three – four disulfide bonds. Represented by charybdotoxin (PDB ID: 2CRD). **B**) A new fold, consisting of two parallel helices linked by two disulfide bridges, was determined in a group of new family of weak K^+ scorpion toxins. Represented by hefutoxin (PDB ID: 1HP9).

To the best of the author's knowledge, a specialised bioinformatic tool for functional prediction of toxins does not exist, other than the tool presented in this thesis. Function of uncharacterised toxins is inferred from identification of characterised similar sequences using BLAST (Altschul *et al*., 1997) or FASTA (Pearson, 2000) programs. Alternatively, function is assigned by searches in pattern databases such as PROSITE (Hulo *et al*., 2004). Generally, all pattern databases use statistical approaches to assign confidence levels to query matches to the motifs but statistical significance does not necessarily equate to biological proof (Attwood, 2000). For example, 'Protein kinase C' (accession ID: PDOC00005) and 'Casein kinase II' (accession ID: PDOC0006) phosphorylation sites were found in a sodium specific toxin, AaHIT2 (Loret *et al*., 1990) upon submission in PROSITE. Both phosphorylation sites which are irrelevant for the function of sodium toxins have a high probability of occurrence in most protein sequences.

Conversely, mutation studies of toxins (such as site-directed mutagenesis and chemical modification) have identified critical residues important for both structural and functional properties. However, this information has not been used in large-scale analysis of toxin data for identifying critical structural and functional residues. Insights gained from the analysis can be used to develop functional prediction tools that include biological information from mutant data.

Chapter summary

- Large amount of toxin data with limited or no annotation is generated from increased automation of experimental techniques (e.g. large-scale sequencing of the three venomous animal genomes). These toxin data are scattered across general databases whose aim is to contain as many genomic or protein

sequences as possible.

- Structure-function information, in particular that of mutation studies of toxins, is available in the literature but is usually not used to enrich the toxin records in the general databases or extraction of functionally motifs.
- The scattered toxin data and structure-function information requires an improved data management in the field of toxin research. Venominformatics, a field combining toxin research and bioinformatics, allows systematic large-scale studies of toxin data where it facilitates experimental design and selection of key experiments by development of functional prediction tools.
- Specialised data warehouses of toxins are dedicated repositories of toxin data extracted from public databases, literature or other public repositories, and experimental measurements. Data warehouses have integrated bioinformatic tools for detailed data analysis and mining.
- The bioinformatic tools commonly used include BLAST for inference of functional and evolutionary relationships and Clustal W for analyses of structure, function and phylogeny relationships. Comparative homology modeling servers help predict tertiary structures of toxins which lack experimentally solved 3D structures.

Part II: Chapter 3 Classification of scorpion toxin data

'The great challenge in biological research today is how to turn data into knowledge. I have met people who think data is knowledge but these people are then striving for a means of turning knowledge into understanding.'

Sydney Brenner

Classification of all currently known scorpion toxins according to their function is necessary for clarifying the global perspective, including an overview of the functional repertoire of the toxins. Such knowledge will facilitate functional assignment of newly identified scorpion toxins. Classification also provides an effective means to retrieve relevant biological information from vast amounts of toxin data. Advances in genomics and proteomics have identified new scorpion toxin sequences at an ever-increasing rate. For example, more than 100 different components were identified from the proteome analysis of *Tityus cambridgei* scorpion venom, of which 26 have been partially sequenced (Batista *et al*., 2004). Many of these toxin sequences have yet unknown function. Consequently, there is a need to analyse and organise these sequences with currently known and annotated scorpion toxin data (nearly 1000 sequences as of November 2005) for a broad view on their general patterns in structure and function.

However, the available large-scale classification of scorpion toxin sequences is limited and is based mainly on the analysis of evolutionary properties of scorpion toxins currently known. These classifications were performed on toxins isolated from distinct scorpion species (Corona *et al*., 2002, Goudet *et al*., 2002) and also by specificity to different ion channels. For example, Possani *et al*. (1999) classified 36 sodium specific scorpion toxins into 10 groups based on animal species specificity and pharmacological effects on sodium channels. Since then, the number of known sodium specific scorpion toxins has increased to 213 toxins. For potassium specific toxins, Tytgat *et al*. (1999) classified them into three families (α-, β- and γ-scorpion toxins). The α- and β-toxins were classified based on peptide length and alignment of cysteines and other conserved residues while γ-toxins were based on specificity to ether-a-go-go

potassium channel subtype. The α-toxin family by that time contained 49 different toxins, comprising 12 subfamilies (Tytgat *et al*., 1999). This classification has expanded to 18 subfamilies as new scorpion toxin sequences were identified but did not fit into the former 12 subfamilies (Rodriguez de la Vega and Possani, 2004). In protein classification databases such as Pfam (Bateman *et al*., 2004) and ProDom (Servant *et al*., 2002), protein families are obtained from multiple sequence alignments of similar proteins. These groups however are based on sequence similarity and are not necessarily functionally relevant. For example, Toxin 3 family (accession ID: PF00537) in Pfam release 18.0 classified scorpion toxins along with plant defensins.

Here, the author describes a systematic large-scale classification of scorpion toxin sequences into groups based on ion channel specificity and primary sequence similarity, combined with multiple sequence alignments and phylogenetic analyses (Tan *et al*., 2005). This classification is based on Tytgat's approach (1999) of primary sequence similarity and multiple sequence alignment but refined using a larger number of scorpion toxin sequences. The toxin sequences were classified with reference to their structural and functional properties. This large-scale classification of currently known scorpion toxins reflects the underlying toxin families for a global view of their structure-function relationships. This is a dynamic field where classified groups can be defined and redefined as the number of known toxin sequences grows. Many groups contain scorpion toxin sequences that have not been classified. Highly accurate functional predictions of novel scorpion toxin sequences have been obtained by comparison with the classified groups (Chapter 5).

3.1 Classification of scorpion toxins

Scorpion toxins are important physiological probes for characterising ion channels. They have been classified into four broad groups, namely those that interact with sodium (Na^+), potassium (K^+), calcium (Ca^{2+}), or chloride (Cl^-) ion channels (Gordon and Gurevitz, 2003; Fuller *et al*., 2004; Giangiacomo *et al*., 2004; Lacinova, 2004) (**Table 2**). Scorpion toxins are also classified as long-chain toxins containing 60 – 70 amino acid residues with four disulfide bridges or short-chain toxins containing 30 – 40 amino acid residues with three or four disulfide bridges (Goudet *et al*., 2002). Na^+ toxins belong to the long-chain toxin family while K^+, Ca^{2+} or Cl^- toxins belong to the short-chain toxin family. Additionally, scorpion toxins can be classified according to species-specificity of toxicity (insect, crustacean or mammal). Some toxins show cross-specificity; for example, BmK M1 from *Buthus martensii* Karsch targets both insect and mammalian cells (Liu *et al*., 2005).

Based on electrophysiological studies, scorpion toxins that interact with Na^+ channels have been classified into three types: α, α-like and β toxins. The α toxins (e.g. AaHII from *Androctonus australis* Hector) slow or block the inactivation of Na^+ channel in a voltage-dependent mechanism whereas β toxins (e.g. Cn2 from *Centruroides noxius*) affect the Na^+ channel activation independently of membrane potential (Couraud *et al*., 1982). The third type is α-like toxins (e.g. LqhIII from *Leiurus quinquestriatus* Hebraeus) that induce sodium current in neuronal preparation but do not compete for AaHII binding (Gordon *et al*., 1996; Gordon and Gurevitz, 2003). The α toxins and α-like toxins bind to site 3, while β toxins bind to site 4 on the Na^+ channel (Jover *et al*., 1984). β toxins are further classified into depressant and excitatory toxins. Depressant toxins induce a block of action potentials whereas excitatory toxins cause a repetitive activity on Na^+ axonal membrane (Zlotkin *et al*.,

1985).

The subtypes of K^+ channels targeted by scorpion toxins include voltage-gated K^+ channels (Pragl *et al*., 2002), inward rectifier K^+ channels (Lu and MacKinnon, 1997), ether-a-go-go-related gene K^+ channels (Frenal *et al*., 2004; Korolkova *et al*., 2004) and Ca^{2+}-activated K^+ channels that include large, intermediate and small conductance Ca^{2+}-activated K^+ channels (Rodriguez de la Vega *et al*., 2003; Jouirou *et al*., 2004; Xu *et al*., 2004a). Two Ca^{2+}-channels subtypes were reported to be targeted by scorpion toxins: Type 1 ryanodine (Zamudio *et al*., 1997; Fajloun *et al*., 2000; Zhu *et al*., 2004b) and T-type voltage-gated Ca^{2+}-channels (Chuang *et al*., 1998; Lopez-Gonzalez *et al*., 2003). The ability of scorpion toxins to block Cl^- channels is controversial (Maertens *et al*., 2000; Dalton *et al*., 2003; Fuller *et al*., 2004).

Table 2 Different criteria can be used to classify scorpion toxins: peptide length, ion channel specificity and electrophysiology.

Peptide length	**Ion channel specificity**	**Electrophysiology**
Long chain 60 – 70 residues	Na^+	- α
		- α-like
		- β (depressant, excitatory)
Short chain 30 – 40 residues	K^+	- Voltage-gated
		- Inward rectifier
		- Ca^{2+}-activated
		- Ether-a-go-go
	Ca^{2+}	- Type 1 ryanodine
		- T-type voltage-gated
	Cl^-	- Cystic fibrosis transmembrane conductance regulator

3.2 Data classification of scorpion toxin sequences

Data classification is an important step for effective information management as it provides an overview of the categories of related biological sequences. It also describes the key relationships between particular characteristics and the corresponding data. An adequate classification of related biological sequences can be used to predict the function of an unknown sequence based on inference of homology between the unknown and the characterised sequences in a class. Homologous sequences are assumed to descend from a common evolutionary ancestor and thus likely share similar function.

Large-scale classification of scorpion toxin sequence data was achieved in this work by a combination of bioinformatic approaches through pairwise and multiple sequence alignments, and phylogenetic analyses. This is necessary because each individual approach has its limitations. The pairwise alignment does not give a clear indication of the domain structure of proteins (Bateman *et al*., 2000) as compared to multiple sequence alignment which gives a better picture of most conserved residues in a protein family. For multiple sequence alignment, because of large number of possible alignments, alignment methods often produce mistakes which compromise the quality of results (Cline *et al*., 2002). Different algorithms have been designed for assembly of multiple sequence alignments. However, none of these algorithms performs consistently better than the others (Poirot *et al*., 2003). An individual algorithm maybe better than others for certain types of problems, but none of these is the best across a broad range of alignment problems (Lassmann and Sonnhammer, 2002). In phylogeny, the accuracies of multiple sequence alignment affect the estimation of phylogenetic relationship among sequence data analysed. If the sequences align well, they are likely to be derived from a common ancestral sequence. On the other hand, a group of poorly

aligned sequences may share a complex and distant evolutionary relationship. Inaccurate multiple alignments will result in incorrect trees and erroneous interpretation of these trees. Thus, by combining multiple approaches, the limitations can be minimised and more accurate analyses of structure-function of scorpion toxins can result from these analyses.

3.3 Materials and Methods

Scorpion toxin sequences were collected from public databases and literature, and deposited into the SCOPRION2 database (discussed in Section 6.3). Of 393 currently known scorpion toxins, 383 sequences were classified into four broad families, namely sodium (Na^+), potassium (K^+), calcium (Ca^{2+}) and chloride (Cl^-) specific toxins. Potassium family was further divided into three subfamilies: α, β and γ according to Tytgat *et al*. (1999). One sequence belonged to the scorpion defensin family while the molecular targets of nine sequences were not reported and were called 'orphans'. The orphans and the defensin were not included in the classification. Within each broad group, sequences were further classified into groups based on primary structure similarity using BLAST (Altschul *et al*., 1997) and Clustal W (Thompson *et al*., 1994), and verified by phylogenetic analysis using MEGA 3.0 (Kumar *et al*., 2004) (**Figure 2**). The classification process has been illustrated using sodium group 1 (Na01) as an example.

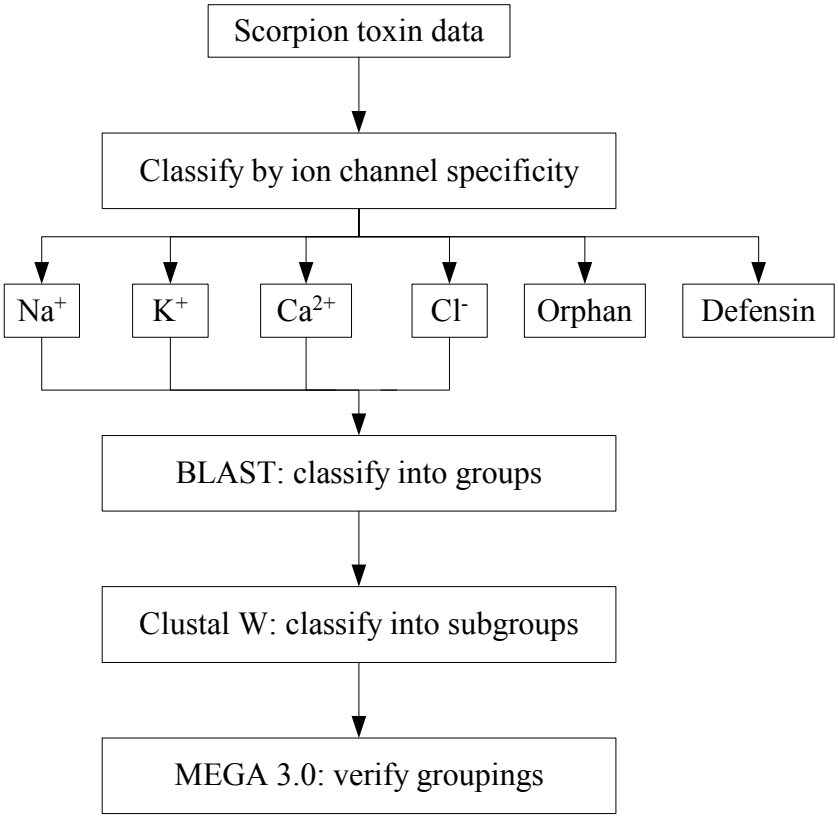

Figure 2 Flowchart of the large-scale classification of scorpion toxin data.

3.3.1 Classification of sequences into groups by BLAST

Within each broad ion channel family, a representative sequence toxin from a list of unclassified toxin sequences was submitted to the blastp program against the non-redundant (nr) database at NCBI (http://ww.ncbi.nlm.nih.gov/BLAST/). BLAST result obtained after every submission ranked protein sequences from the most similar to the least similar to the query. By looking at the expectation (*E*) value, a cut-off was determined by manual inspection for each group of sequences. The *E* value is a parameter that describes the probability of matches against database sequences. The cut-off value would vary depending on the similarity of the top scoring sequences, but usually occurred where there was a large change in *E* values between two consecutive sequences. The higher *E* value would serve as the cut-off value. Sequences which had

E values lower than this cut-off were clustered into a group. The sequences scoring higher than the cut-off were added to a new group. For example, in **Figure 3,** the query sequence was Na^+ toxin, BmKIT1 isolated from *Buthus martensii* Karsch. A large difference in the *E* values from $2\times e^{-20}$ to $1\times e^{-7}$ between sequences 17 and 18 (Bjxtr-It and Cse-V5, respectively) serves as a cut-off (**Figure 3A**). The higher *E* value of $1\times e^{-7}$ was set as the cut-off value to cluster sequences 1 – 17 into a preliminary group i.e. neurotoxin (KIT) to Scorpion toxin Bjxtr-IT.

The preliminary grouping of the 17 sequences was confirmed by submitting the last sequence (Bjxtr-IT) before the cut-off value $1\times e^{-7}$ for a second BLAST search (**Figure 3B**). In the result returned by the second BLAST search, the cut-off value of $8\times e^{-8}$ was set because a large difference in the *E* values occurred between sequences Toxin 1 and TbIT-I. Sequences with *E* values lower than $8\times e^{-8}$ were clustered and compared with the list of toxin sequences in the preliminary group. The grouping was finalised if both BLAST results clustered the same toxin sequences together. The confirmed group was named according to ion channel specificity followed by a number e.g. Na01. The classified sequences were removed from the list of unclassified toxin sequences and the process was repeated with subsequent sequences until the list was exhausted.

A)

Sequences producing significant alignments:	Score (Bits)	E Value
gi|3036821|emb|CAA76604.1| neurotoxin (KIT) [Mesobuthus martensi	152	3e-36
gi|3063655|gb|AAC14130.1| neurotoxin BmK IT precursor [Buthus...	150	1e-35
gi|3649606|gb|AAC61256.1| insect neurotoxin precursor [Buthus...	126	2e-28
gi|58176732|pdb|1T0Z|B Chain B, Structure Of An Excitatory In...	126	2e-28
gi|161147|gb|AAA29950.1| neurotoxin AaH IT1	121	7e-27
gi|69545|pir||XISR1A insect toxin 1 - Sahara scorpion >gi|223...	121	7e-27
gi|232628|gb|AAA03882.1| insect-specific neurotoxin precursor...	121	7e-27
gi|134372|sp|P01497|SIX1_ANDAU Insect toxin 1 precursor (AaH ...	121	7e-27
gi|56404741|sp|P68722|SIXB_LEIQH Insect neurotoxin 1b precursor	118	5e-26
gi|102777|pir||G34444 insect toxin 2 precursor - Sahara scorp...	118	6e-26
gi|56404740|sp|P68721|SIXA_LEIQH Insect neurotoxin 1a precursor	115	3e-25
gi|84690|pir||S08267 toxin 1 - scorpion (Leiurus quinquestria...	114	9e-25
gi|56404742|sp|P68723|SIXC_LEIQH Insect neurotoxin 1c precursor	108	5e-23
gi|56404743|sp|P68724|SIXD_LEIQH Insect neurotoxin 1d precursor	104	9e-22
gi|6094296|sp|P56637|SIXE_BUTJU Excitatory insect toxin Bjxtr...	100	2e-20
gi|3858953|emb|CAA09988.1| neurotoxin, variant 38K [Hottentotta	100	2e-20
gi|4140001|pdb|1BCG| Scorpion Toxin Bjxtr-It	100	2e-20
gi|102790|pir||C23727 neurotoxin V-5 - bark scorpion >gi|2052...	57.8	1e-07
gi|15825922|pdb|1I6G|A Chain A, Nmr Solution Structure Of The...	57.8	1e-07

B)

Sequences producing significant alignments:	Score (Bits)	E Value
gi|6094296|sp|P56637|SIXE_BUTJU Excitatory insect toxin Bjxtr...	175	3e-43
gi|4140001|pdb|1BCG| Scorpion Toxin Bjxtr-It	175	3e-43
gi|3858953|emb|CAA09988.1| neurotoxin, variant 38K [Hottentotta	174	1e-42
gi|3063655|gb|AAC14130.1| neurotoxin BmK IT precursor [Buthus...	101	6e-21
gi|3649606|gb|AAC61256.1| insect neurotoxin precursor [Buthus...	101	8e-21
gi|58176732|pdb|1T0Z|B Chain B, Structure Of An Excitatory In...	101	8e-21
gi|3036821|emb|CAA76604.1| neurotoxin (KIT) [Mesobuthus martensi	100	2e-20
gi|56404741|sp|P68722|SIXB_LEIQH Insect neurotoxin 1b precursor	95.9	3e-19
gi|56404743|sp|P68724|SIXD_LEIQH Insect neurotoxin 1d precursor	93.6	2e-18
gi|161147|gb|AAA29950.1| neurotoxin AaH IT1	93.2	2e-18
gi|56404740|sp|P68721|SIXA_LEIQH Insect neurotoxin 1a precursor	93.2	2e-18
gi|134372|sp|P01497|SIX1_ANDAU Insect toxin 1 precursor (AaH ...	93.2	2e-18
gi|69545|pir||XISR1A insect toxin 1 - Sahara scorpion >gi|223...	92.0	5e-18
gi|232628|gb|AAA03882.1| insect-specific neurotoxin precursor...	92.0	5e-18
gi|102777|pir||G34444 insect toxin 2 precursor - Sahara scorp...	90.1	2e-17
gi|56404742|sp|P68723|SIXC_LEIQH Insect neurotoxin 1c precursor	89.7	2e-17
gi|84690|pir||S08267 toxin 1 - scorpion (Leiurus quinquestria...	86.7	2e-16
gi|41017886|sp|P60275|SCXI_TITBA Insect toxin TbIT-I	58.2	8e-08
gi|58379083|gb|AAW72463.1| putative long-chain toxin precursor	57.4	1e-07

Figure 3 BLAST search results were used for classification of scorpion toxin sequences into groups based on high primary sequence similarity. First column provides the database accession numbers and toxin names. Second and third columns represent the score and *E* values. Sequences were classified into groups by analysing their *E* values. **A**) Result of submitting BmKIT1, a Na^+ toxin from *Buthus martensii* Karsch, against the nr database at NCBI. A marked increase in the *E* values from $2\times e^{-20}$ to $1\times e^{-07}$ would serve as a borderline where $1\times e^{-07}$ was designated as the cut-off value. Sequences having *E* values < $1\times e^{-07}$ were clustered into a preliminary group. **B**) Result of submitting Bjxtr-IT 1 from *Hottentotta judaica*, which was the last sequence before the cut-off value, $1\times e^{-07}$ into nr database. The cut-off value for this second BLAST search was $8\times e^{-08}$, where sequences having *E* values lower than $8\times e^{-08}$ were clustered into the group.

3.3.2 Data classification into subgroups by Clustal W

Within each group, multiple sequence alignment was performed using Clustal W. Sequences with distinct primary structure patterns were further classified into subgroups. Subgroups were given an alphabet after the group name e.g. Na01a. For example, the two Bjxtr-IT sequences had five distinct residues at positions 16 to 21 from the rest of the sequences in Na01 and was classified into a subgroup, Na01c (**Figure 4**). The C-terminal residues of BmK IT-AP, BmKIT1, Bm32-VI and BmK IT were distinct, sharing pattern of Na01c. These four were classified into a second subgroup while the rest of the sequences form the last subgroup.

3.3.2 Verification of groups and subgroups by MEGA 3.0

The author verified these groupings by phylogenetic analysis using MEGA 3.0. Phylogenetic trees were reconstructed using the neighbour-joining algorithm (Saitou and Nei, 1987). Sequences in Na01b were clustered and separated from Na01a and Na01c (**Figure 5**). The group and subgroup classification of each scorpion toxin sequence can be found in the SCORPION2 database entries (See Appendix 1).

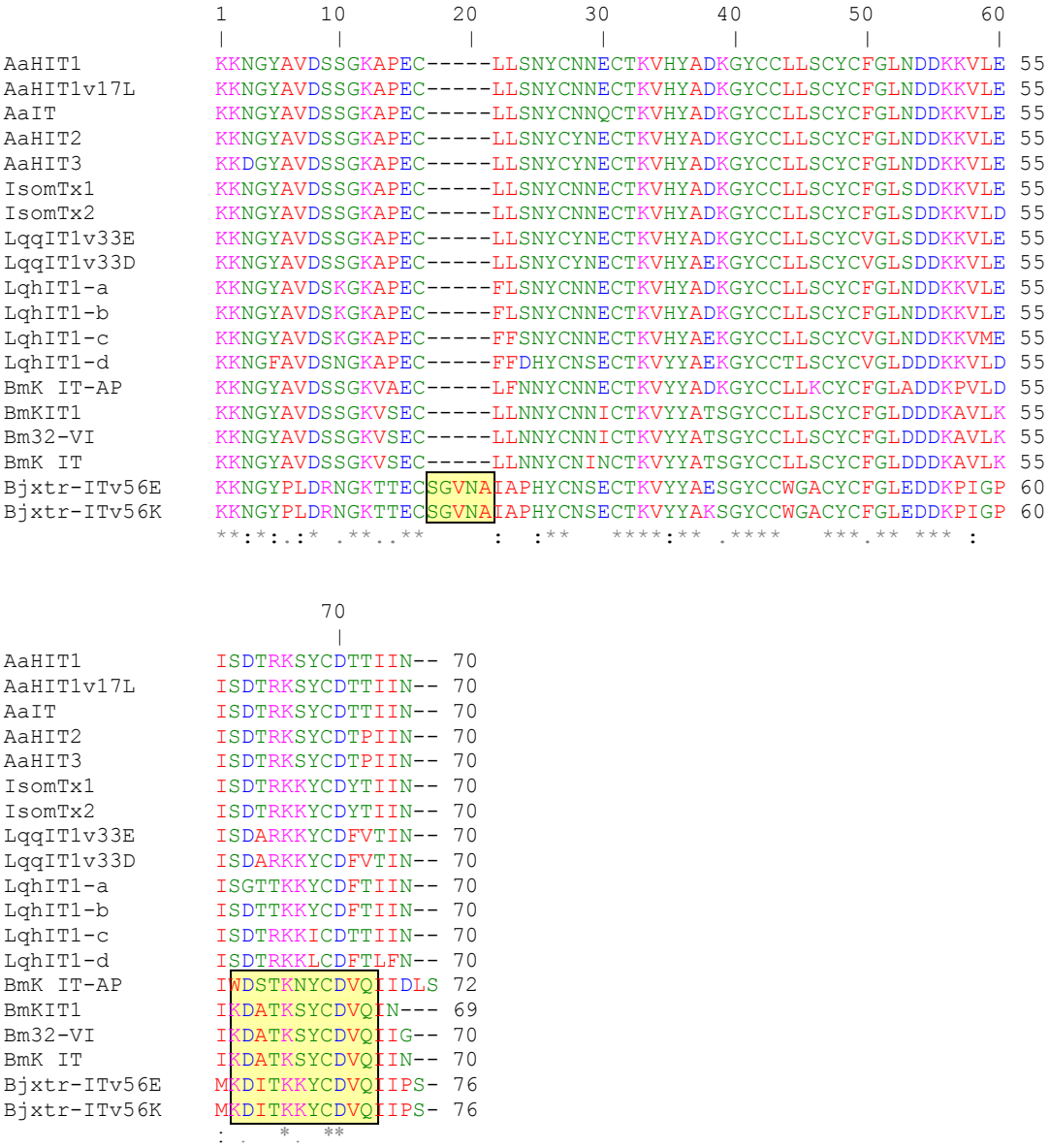

Figure 4 Further classification of toxin sequences in Na01 into subgroups based on distinct primary structure patterns in their multiple sequence alignment. The first column shows toxin names; second column shows the toxin primary sequences; third column displays the peptide length. The shaded boxes highlight distinct patterns among sequences.

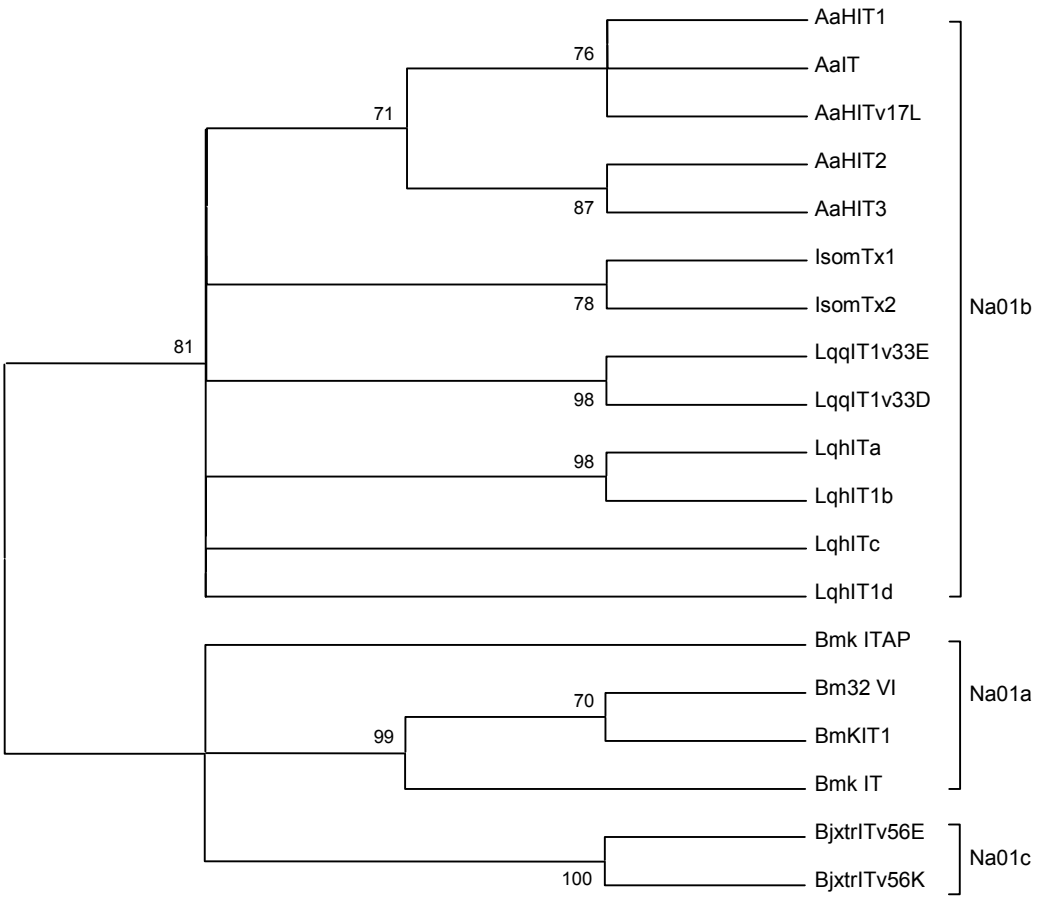

Figure 5 Verification of groups and subgroups by phylogenetic analysis using MEGA 3.0. Each bootstrap number represents the percentage of 1000-bootstrap value. Subgroup Na01b was separated from Na01a and Na01c. Within the branch of Na01a and Na01c, the two BjxtrIT sequences were classified into Na01c, while BmK IT, BmKIT1, Bm 32 VI and Bmk ITAP were classified as Na01a.

3.4 Results – Classification of scorpion toxin sequences

The 393 scorpion toxin sequences were classified into 62 groups (**Figure 6**). The general distribution of scorpion toxins in the 62 groups is not significantly different from previous classification in the literature (Possani *et al*., 1999; Rodriguez de la Vega and Possani, 2004) except for the larger number of toxin sequences analysed. Some toxin sequences had been reclassified and new groups were introduced

for sequences which did not fit into the previous classification. As new scorpion toxin sequences are identified and new information is available, revision of the classification will be necessary. The detailed description of the 62 groups is available in Appendix 1.

The long-chain Na^+ toxins were separated from the short-chain scorpion toxins namely, K^+, Ca^{2+} and Cl^- except toxins in BmP09 and BeI2 groups. BmP09 shares high homology with Na^+ toxins but has a sulfoxide at the C terminus which resulted in a dramatic switch from a Na^+ channel blocker to a K^+ channel blocker (Yao *et al*., 2005). BeI2 was misaligned with short K^+ toxins. Of 393 scorpion toxins, 135 K^+ toxins were classified into 32 groups, 222 Na^+ toxins were classified into 18 groups, while eight Ca^{2+} toxins were classified into four groups (**Table 3**). K^+ group 6 was further classified into three subgroups. Na^+ groups 2 and 9 were classified into four subgroups each, Na^+ group 1 into three subgroups and Na^+ group 12 into two subgroups. 19 Cl^- toxins were classified into two subgroups. Scorpine and nine scorpion toxin sequences with no annotated molecular target were assigned to the 'defensin' and 'orphan' groups, respectively.

Table 3 Summary of the classified groups for 393 scorpion toxins. Kurtoxin has been classified into Na^+ and Ca^{2+} since it targets both ion channels (Olamendi-Portugal *et al*., 2002).

	K^+	**Na^+**	**Ca^{2+}**	**Cl^-**	**Defensin**	**Orphan**
No. of groups	32	18	4	1	1	6
No. of toxins	135	222	8	19	1	9

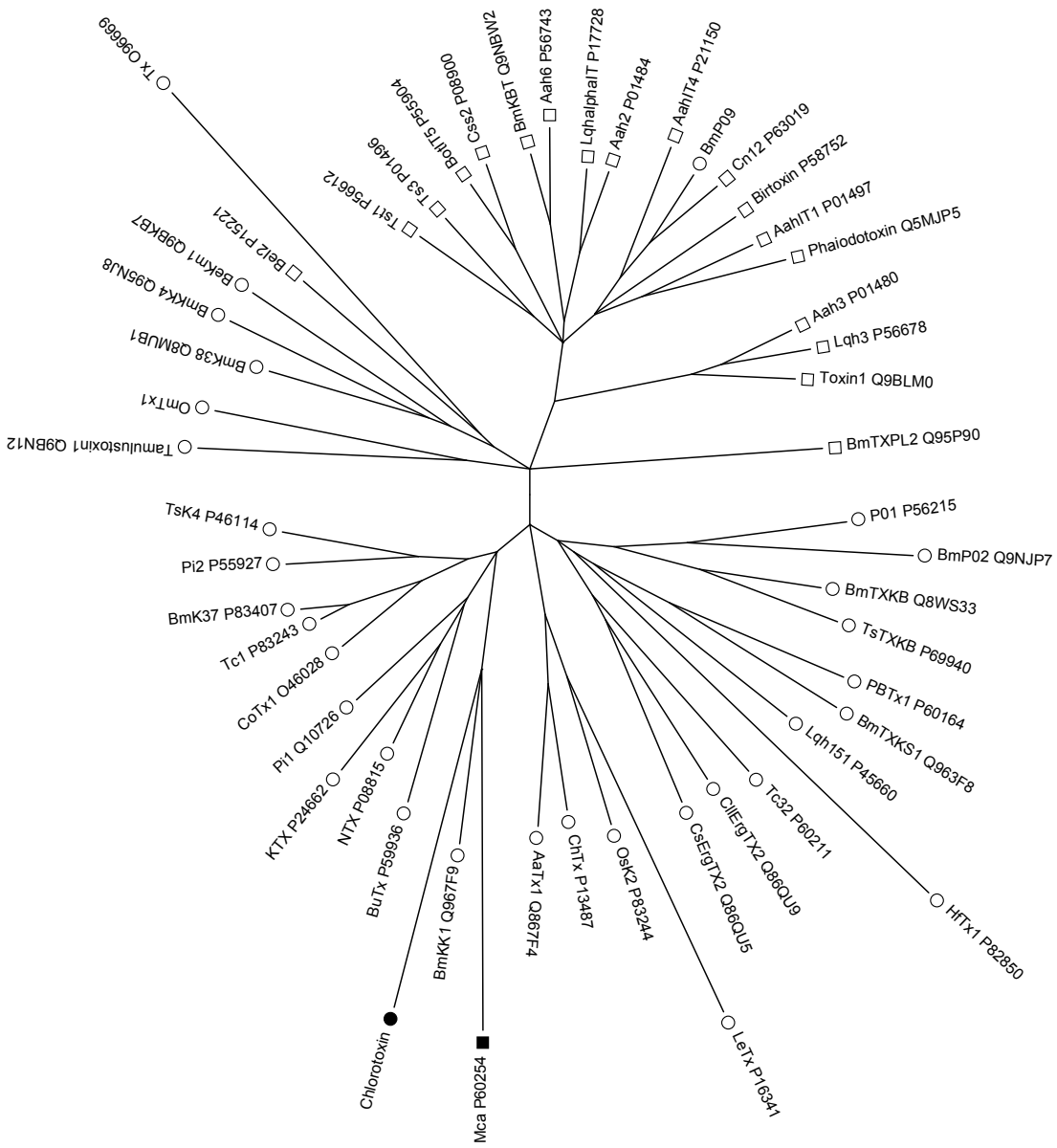

Figure 6 Phylogenetic tree of representative scorpion toxins for Na^+, K^+, Ca^{2+} and Cl^- channels. For clarity, only 52 branches are displayed. The orphan and defensin groups were not included. The tree was calculated using the neighbour-joining algorithm. □ Na^+, ○ K^+, ■ Ca^{2+} and ♧ Cl^-. The branches are labelled with toxin names and Swiss-Prot accession numbers except OmTx1 (Chagot *et al*., 2005) and BmP09 (Yao *et al*., 2005). Tree was generated using MEGA 3.0.

For K^+ toxins, a fourth subfamily, δ-KTx was introduced for κ-toxins following the nomenclature for α-, β- and γ-KTx subfamilies proposed by Tytgat *et al*. (1999) (**Table 4**). The hairpin structure of two short helices cross-linked with two disulfides in the δ-KTx subfamily (Srinivasan *et al*., 2002b; Nirthanan *et al*., 2005) differs from the conserved 3D structures of all known scorpion toxins, comprising of double or triple stranded β-sheets and an α-helix maintained by two – four disulfide bridges (Possani *et al*., 1999; Possani *et al*., 2000). Toxins in the δ-KTx subfamily are non- or weak ligands of voltage-dependent potassium channels (Srinivasan *et al*., 2002; Nirthanan *et al*., 2005).

Five new α-KTx subfamilies were proposed which did not fit into the expanded 18 α-KTx subfamilies (Rodriguez de la Vega and Possani, 2004), namely BmTXKS1 (α-KTx21), Tamulustoxins (α-KTx22), OmTX (α-KTx23), K^+ channel inhibiting toxin (α-KTx24) and BmK38 (α-KTx25) (**Figure 7**). The author had reclassified Lqh15-1 (α1.7) and PbTx3 (α1.10) in α-KTx1 subfamily into α-KTx16 and α-KTx4 subfamilies, respectively. Lqh15-1 shared higher sequence identity (~80%) with toxins in α-KTx16 than α-KTx1 (<50%) and was clustered within the same clade as α-KTx16 toxins in the phylogenetic tree. PbTx3 shared higher sequence identity (average 50%) with α-KTx4 toxins than those in α-KTx1 (average 40%). α-KT6 subfamily was divided into three subfamilies because they were separated into three clades. Tc1 (α-13.1) and OsK2 (α-13.2) were split into α-KTx13 and α-KTx20 because submission of each toxin into BLAST searches at NCBI (http://www.ncbi.nlm.nih.gov/BLAST/) did not return the other toxin and they were present on different clades.

The β-KTx subfamily consists of K^+ long chain toxins of 60 – 66 residues. The author classified BmTXKβ (β3) into β-KTx2 as it had only 21% identity with TsTXKβ (β1), AaTXKβ (β2) and BmTXKβ2 (β4) which form the β-KTx1 subfamily (average

70%). A new long chain potassium toxin, BmP09 (Yao *et al*., 2005) forms the β-KTx3.

Table 4 Classification of 135 K^+ scorpion toxin sequences. The first column shows subfamilies from this work. δ-KTx subfamily was introduced after α-, β- and γ-KTx subfamilies proposed by Tytgat *et al*. (1999). Second column lists the number of peptides in each subfamily. Third column shows the distribution of classified toxins as compared to that of the 18 α-KTx, 4 β-KTx and 5 γ-KTx subfamilies by Rodriguez de la Vega and Possani (2004). 'New toxin' represents the number of new toxins not previously included in the subfamily of Rodriguez de la Vega and Possani (2004).

K^+	**Peptide no.**	**Rodriguez de la Vega and Possani (2004)**	
		Subfamily	**Sequence increment**
α-KTx1	8	α-KTx1	0
α-KTx2	9	α-KTx2	3
α-KTx3	9	α-KTx3	0
α-KTx4	8	α-KTx4	4
α-KTx5	5	α-KTx5	0
α-KTx6 (a, b, c)	12 (4, 7, 1)	α-KTx6	2
α-KTx7	2	α-KTx7	0
α-KTx8	4	α-KTx8	0
α-KTx9	6	α-KTx9	2
α-KTx10	2	α-KTx10	0
α-KTx11	3	α-KTx11	0
α-KTx12	2	α-KTx12	1
α-KTx13	1	α-KTx13	1
α-KTx14	4	α-KTx14	0
α-KTx15	7	α-KTx15	0
α-KTx16	5	α-KTx16	2
α-KTx17	1	α-KTx17	0
α-KTx18	1	α-KTx18	0
α-KTx19	1	Cai *et al*., 2004	0
α-KTx20	1	α-KTx13	0
α-KTx21	**1**	**New**	**1**
α-KTx22	**2**	**New**	**2**
α-KTx23	**3**	**New**	**3**
α-KTx24	**1**	**New**	**1**
α-KTx25	**1**	**New**	**1**
β-KTx1	3	β1, β2, β4	0
β-KTx2	1	β3	0
β-KTx3	**1**	**New**	**1**
γ-KTx1	2	γ-KTx2	1
γ-KTx2	12	γ-KTx4	0
γ-KTx3	14	γ-KTx1, 3, 4, 5	1
δ-KTx1	**3**	**κ-hefutoxins**	**1**

```
Gp Name (SP ID)                Sequence

Alpha-KTx
 1 ChTx (P13487)               ----------------------ZFTNVSCT-TSKECWSVCQR---LHN-----TSRGKCMN----------KKCRCYS------
 2 NTX (P08815)                ----------------------TIINVKCT-SPKQCSKPCKE---LYGS----SAGAKCMN----------GKCKCYNN-----
 3 KTX (P24662)                ----------------------GVEINVKCS-GSPQCLKPCKD---A-G-----MRFGKCMN----------RKCHCTPK-----
 4 TsK4 (P46114)               ----------------------VFINAKCR-GSPECLPKCKE---AIG-----KAAGKCMN----------GKCKCYP------
 5 LeTx (P16341)               -------------------------AFCN--LRMCQLSCR------SL----GLLGKCIG-DKC-------ECVKH-------
 6 Pi1 (Q10726)                ------------------------LVKCR-GTSDCGRPCQQ---QTG-----CPNSKCIN----------RMCKCYGC-----
 7 Pi2 (P55927)                ------------------------TISCT-NPKQCYPHCKK---ETG-----YPNAKCMN----------RKCKCFGR-----
 8 P01 (P56215)                ------------------------VSC----EDCPEHCSTQ----------KAQAKCDN----------DKCVCEPI-----
 9 BmP02 (Q9NJP7)              ------------------------VGC----EECPMHCKGK----------NAKPTCDD----------GVCNCN-V-----
10 CoTx1 (O46028)              ------------------------AVCV-YR-TCDKDCKR---R-G-----YRSGKCIN----------NACKCYPY-----
11 PBTx1 (P60164)              ----------------------DEEPKESCS-D-EMCVIYCK----GEE-----YSTGVCDGP---------QKCKCSD------
12 BuTx (P59936)               -------------------WCSTCLDLACG-ASRECYDPCFK---AFG-----RAHGKCMN----------NKCRCYT------
13 Tc1 (P83243)                --------------------------AC----GSCRKKCK-------------GSGKCIN----------GRCKCY-------
14 BmKK1 (Q967F9)              ---------------------TPFAIKCA-TDADCSRKCPG-------------NPSCRN----------GFCACT-------
15 AaTx1 (Q867F4)              ----------------------ZIETNKKCQ-G-GSCASVCRR---VIG-----VAAGKCIN----------GRCVCYP------
16 Lqh15-1 (P45660)            ----------------------GLIDVRCY-DSRQCWIACKK---VTG-----STQGKCQN----------KQCRCY-------
17 BmKK4 (Q95NJ8)              ------------------------QTQCQ-SVRDCQQYCLT-------------PDRCSY----------GTCYCKTT-----
18 Tc32 (P60211)               ----------------------TGPQTTCQ-A-AMCEAGCK----GLG-----KSMESCQG----------DTCKCKA------
19 BmK37 (P83407)              -------------------------AACY-SS-DCRVKCVA---M-G-----FSSGKCIN----------SKCKCYK------
20 OsK2 (P83244)               ----------------------ACGPGCS---GSCRQKGD-------------RIKCIN----------GSCHCYP------
21 BmTXKS1 (Q963F8)            ---------------------------GIV-C-KVCKIICG----MQG-----KKVNICKAP---------IKCKCKK------
22 Tamulustoxin 1 (Q9BN12)     --------------------------RCH--FVVCTTDCRR---N-SP----GTYGECVKKEKGK------ECVCKS------
23 OmTx1                       -------------------------DPCY---EVCLQQHGN-------------VKECEE----------ACKHPVE-----
24 Tx (O96669)                 ----------------------CQNECCGISSLRERNYCAN-----LV----CINCFCQG-------RTYKICRCFFSIHAIR
25 BmK38 (Q8MUB1)              ----------------------KTATFCT--QSICQESCK------RQ----NKNGRCVIEAEGSLIY--HLCKCY-------

Beta-KTx
 1 TsTXKB (P69940)             -----------KLVALIPNDQLRSILKAVV-H-KVAKTQFGCP-AYEG-----YCNDHCNDIERKDGECHGFKCKCAKD-----
 2 BmTXKB (Q8WS33)             KNIKEKLTEVKDKMKHSWNKLTSMSEYACPVIEKWCEDHCAAK----------KAIGKCED----------TECKCLKLRK---
 3 BmP09                       ------------DNGYLLNKYTGCKIWCVINNESCNSECKLRRGNYGYCYFWKLACYCEGAPKSE-LWAYETNKCNGKM----

Gamma-KTx
 1 BeKm-1 (Q9BKB7)             ----------------------RPTDIKCS-ESYQCFPVCKSRF----G----KTNGRCVN----------GFCDCF-------
 2 CllErgTX2 (Q86QU9)          ------------------------DRDSCV-D-KSKCSKYG----YYG-----QCDECCKKAGDRAGNCVYFKCKCNP------
 3 CsErgTX2 (Q86QU5)           ------------------------DRDSCV-D-KSRCAKYG----YYG-----QCEVCCKKAGHRGGTCDFFKCKCV-------

Delta-KTx
 1 HfTx1 (P82850)              ------------------------GHACY---RNCWREGND-------------EETCKE----------RC----------
```

Figure 7 Representative scorpion toxins from four K^+ subfamilies, namely α, β, γ and δ. The group number is followed by the toxin name and the accession number from Swiss-Prot, except OmTx1 (Chagot *et al*., 2005) and BmP09 (Yao *et al*., 2005) which were extracted from the literature.

The author reclassified toxins targeting ether-a-go-go potassium channel subtype into three subfamilies: γ-KTx1 – 3 instead of five subfamilies as proposed by Rodriguez de la Vega and Possani (2004). Three distinct patterns were observed from their multiple sequence alignment (**Figure 8**). The γ-KTx1 consists of BeKm-1 (γ-2.1) and BmKK7 (Swiss-Prot id: P59938), γ-KTx2 consists of all the γ-KTx4 toxins and γ-KTx3 consists of γ-KTx1, 3, 4 and 5.

```
γ4.13          γ-KTx2    DRDSCVDKSKCGKYGYYGQCDECCK-KAGDRAGTCVYYKCKCNP----
γ4.12                    ERDSCVEKSKCGKYGYYGQCDECCK-KAGDRAGTCVYYKCKCNP----
γ4.09                    DRDSCVDKSRCGKYGYYGQCDDCCK-KAGDRAGTCVYYKCKCNP----
γ4.10                    DRDSCVDKSRCGKYGYYGQCDECCK-KAGDRAGTCVYYKCKCNP----
γ4.11                    DRDSCVDKSQCGKYGYYGQCDECCK-KAGERVGTCVYYKCKCNP----
γ4.03                    DRDSCVDKSKCGKYGYYGQCDECCK-KAGDRAGICEYYKCKCNP----
γ4.01                    DRDSCVDKSKCSKYGYYGQCDECCK-KAGDRAGNCVYFKCKCNP----
γ4.06                    DRDSCVDKSKCSKYGYYGQCDKCCK-KAGDRAGNCVYFKCKCNQ----
γ4.07                    DRDSCVDKSKCAKYGYYGQCDECCK-KAGDRAGNCVYLKCKCNQ----
γ4.08                    DRDSCVDKSKCGKYGYYHQCDECCK-KAGDRAGNCVYYKCKCNP----
γ4.04                    DRDSCVDKSKCAKYGYYYQCDECCK-KAGDRAGTCEYFKCKCNP----
γ4.05                    DRDSCVDKSQCAKYGYYYQCDECCK-KAGDRAGTCEYFKCKCNP----
γ4.02          γ-KTx3    DRDSCVDKSKCGKYGYYQECQDCCK-NAGHNGGTCVYYKCKCNP----
γ1.01                    DRDSCVDKSRCAKYGYYQECQDCCK-NAGHNGGTCMFFKCKCA-----
γ1.01 isoform            DRDSCVDKSRCAKYGYYQECQDCCK-NAGHNGGTQMFFKCKAP-----
γ1.04                    DRDSCVDKSRCAKYGYYQECQDCCK-KAGHNGGTCMFFKCKCA-----
γ1.05                    DRDSCVDKSRCSKYGYYQECQDCCK-KAGHNGGTCMFFKCKCA-----
γ1.06                    DRDSCVDKSRCAKYGYYQECQDCCK-KAGHSGGTCMFFKCKCA-----
γ1.02                    DRDSCVDKSRCAKYGYYQECTDCCK-KYGHNGGTCMFFKCKCA-----
γ1.03                    DRDSCVDKSRCAKYGHYQECTDCCK-KYGHNGGTCMFFKCKCA-----
γ5.01                    DRDSCVDKSRCAKYGYYGQCEVCCK-KAGHNGGTCMFFKCMCVNSKMN
γ3.01                    GRDSCVNKSRCAKYGYYSQCEVCCK-KAGHKGGTCDFFKCKCKV----
γ3.03                    DRDSCVDKSRCAKYGYYGQCEVCCK-KAGHRGGTCDFFKCKCKV----
γ3.02                    DRDSCVDKSRCAKYGYYQQCEICCK-KAGHRGGTCEFFKCKCKV----
γ3.04                    DRDSCVDKSRCQKYGNYAQCTACCK-KAGHNKGTCDFFKCKCT-----
γ5.02                    DRDSCVDKSRCQKYGPYGQCTDCCK-KAGHTGGTCIYFKCKCGAESGR
γ2.01          γ-KTx1    -RP---TDIKCSES--Y-QCFPVCKSRFGKTNGRCVNGFCDCF-----
BmKK7                    -RP---TDIKCSAS--Y-QCFPVCKSRFGKTNGRCVNGLCDCF-----
                       *     . :*     * :*   **:. *.  *       * .
```

Figure 8 Multiple sequence alignment of γ-KTx toxins targeting ether-a-go-go K^+ channel subtype. The first column represents the classification in Rodriguez de la Vega and Possani (2004); second column represents the classification in this work; third column represents the toxin primary sequences. Multiple sequence alignment was generated by Clustal W.

In this work, the author classified 222 Na^+ scorpion toxins into 18 groups based on sequence similarity and ion channel specificity (**Table 5**). In contrast, Possani *et al*. (1999) classified 36 sequences into 10 groups based on animal species specificity and pharmacological properties. Information on toxicity to animal models such as insects (cockroach, locust), mammals (mice, rat) and crustaceans (crayfish, prawn) is limited in the literature. Toxin sequences in different groups proposed by Possani *et al*. (1999) shared high sequence similarity. For example, CssII, CsEv1 and Cn5 in Possani *et al*.'s groups 2, 8 and 9 shared 60 – 86% identity (**Figure 9**). Na^+ toxins were thus classified based on primary structure similarity and broad ion channel specificity. This work's groups 1, 2, 10 and 11 corresponded to Possani *et al*.'s groups 10, 3, 7 and 4 except for the larger number of toxin members analysed. Groups 3 – 7 and 9 contained sequences distributed in Possani *et al*.'s groups 1, 2, 5, 6, 8 and 9. Finally, new groups 8, 12 – 18 containing scorpion toxin sequences that were dissimilar to other groups were introduced. Representative sequences of the 18 Na^+ groups are shown in **Figure 10**.

```
Toxin    Possani's   Sequence
CsEv1    2           KEGYLVKKSDGCKYDCFWLGKNEHCNTECKAKNQGGSYGYCYAFACWCEGLPESTPTYPLPNKSC-
Cn5      9           KEGYLVNKSTGCKYGCLLLGKNEGCDKECKAKNQGGSYGYCYAFGCWCEGLPESTPTYPLPNKSCS
CssII    8           KEGYLVSKSTGCKYECLKLGDNDYCLRECKQQYGKSSGGYCYAFACWCTHLYEQAVVWPLPNKTCN
                     ******.** **** *: **.*: *  *** :   .* ******.***  * *.: .:*****:*.
```

Figure 9 Multiple sequence alignment of CsEv1, Cn5 and CssII in Possani *et al*.'s groups 2, 8 and 9. In this work, these sequences had been classified into group 9.

Scorpion toxins targeting Ca^{2+} channels were classified into four groups (**Figure 11**) as they shared less than 28% sequence identity among themselves. Imperatoxin I from *Pandinus imperator* is the only scorpion toxin isolated thus far that exists as a heterodimer (Zamudio *et al*., 1997).

Table 5 Classification of 222 Na^+ scorpion toxin sequences. The first column shows groups and subgroups determined in this work. Second column lists the number of peptides in each group or subgroup. Third column shows the distribution of classified toxins as compared to that of the 10 subfamilies by Possani *et al*. (1999). 'New toxin' represents the number of new toxins not previously included in the subfamily of Possani *et al*. (1999).

Na^+ Group (Subgroup)	**Peptide no.**	**Possani *et al.* (1999)**	
		Subfamily	**New Toxin**
1 (a, b, c)	18 (4, 12, 2)	10	15
2 (a, b, c, d)	19 (3, 4, 7, 5)	3	18
3	38	5, 6, 8	31
4	5	1	3
5	7	6	5
6	19	1, 5	15
7	6	1, 8	3
8	**2**	**New**	**2**
9 (a, b, c, d)	52 (6, 20, 24, 2)	2, 8, 9	42
10	5	7	4
11	33	4	27
12 (a, b)	**4 (3, 1)**	**New**	**4**
13	**1**	**New**	**1**
14	**3**	**New**	**3**
15	**5**	**New**	**5**
16	**3**	**New**	**3**
17	**1**	**New**	**1**
18	**1**	**New**	**1**

```
Gp Name (SP ID)             Sequence

 1 AahIT1 (P01497)          ----KKNGYAVDS-SGKAPECL----LSNYCNNECTK-VHYADKGYCC-----LLS--CYCFGLNDDKKVLEISDTRKSYCDTTIIN
 2 Tst1 (P56612)            -----KEGYLM-DHEGCKLSCF-IR-PSGYCGRECTL-KKGS-SGYC----AWP-A--CYCYGLPNWVKVWDRATN---KC------
 3 LqhalphaIT (P17728)      ----VRDAYIAKN-YNCVYECF----RDAYCNELCTK--NGASSGYCQWAGKYGNA--CWCYALPDNVPIR-VPG----KCR-----
 4 Aah3 (P01480)            ----VRDGYIVDS-KNCVYHCVP---P---CDGLCKK--NGAKSGSCGFLIPSGLA--CWCVALPDNVPIK-DPSY---KCHS----
 5 Lqh3 (P56678)            ----VRDGYIAQP-ENCVYHCFP---GSSGCDTLCKE--KGGTSGHCGFKVGHGLA--CWCNALPDNVGII-VEGE---KCHS----
 6 Aah2 (P01484)            ----VKDGYIVDD-VNCTYFCG----RNAYCNEECTK--LKGESGYCQWASPYGNA--CYCYKLPDHVRTK-GPG----RCH-----
 7 Ts3 (P01496)             ----KKDGYPVEY-DNCAYICWNY--DNAYCDKLCKD--KKADSGYCYWVHIL-----CYCYGLPDSEPTK-TNG----KCKS----
 8 Aah6 (P56743)            ----GRDGYVVKNGTNCKYSCEIGS-EYEYCGPLCKR--KNAKTGYC-----YAFA--CWCIDVPDDVKLYGDDGT---YCSS----
 9 Css2 (P08900)            -----KEGYLVSKSTGCKYECLKLG-DNDYCLRECKQQYGKSSGGYCY-----AFA--CWCTHLYEQAVVWPLPNK---TCN-----
10 AahIT4 (P21150)          -----EHGYLLNKYTGCKVWCV-IN-NEE-CGYLCNK-RRGGYYGYCY---FWKLA--CYCQGARKSE-LWNYKTN---KCDL----
11 BotIT5 (P55904)          ------DGYIR-KRDGCKVSCL-FG-NEG-CDKECKA-YGGSYG-YCWT---WGLA--CWCEGLPDDK-TWKSETN---TCG-----
12 BeI2 (P15221)            -----ADGYVK-GKSGCKISCF-LD-NDL-CNADCKY-YGGKLNSWCIP--DKSGY--CWCPNK-GWN-SIKSETN---TC------
13 BmKBT (Q9NBW2)           ----KKSGYPT-DHEGCKNWCV-LN-HS--CGILCEG-YGG--SGYCY---FWKLA--CWCDDIHNWVPTWSRATN---KCRA----
14 Toxin1 (Q9BLM0)          ----VRDGYFVEP-DNCVVHCMP---SSEMCDRGCKH--NGATSGSCKAFSKGGNA--CWCKGLR----------------------
15 Birtoxin (P58752)        --ADVPGNYPLDK-DGNTYKCFLLG-GNEECLNVC-K-LHGVQYGYCY-----ASK--CWCEYLED------DKDS---V-------
16 Phaiodotoxin (Q5MJP5)    -------KFIRHK-DESFYECGQLIGYQQYCVDACQA-HGSKEKGYCKGMAPFGLPGGCYCPKLPSNRVKMCFGALES-KCA-----
17 BmTXPL2 (Q95P90)         CSMVYGDSLSPWNEGDTYYGCQRQ--TDEFCNKICKL--HLASGGSCQQPAPFVKL--CTCQGIDYDNSFFFGALEK--QCPKLRE-
18 Cn12 (P63019)            -----RDGYPLAS-NGCKFGCSGLGENNPTCNHVCEK-KAGSDYGYCY-----AWT--CYCEHVAEGTVLWGDSGT---GPCRS---
```

Figure 10 Representative scorpion toxins from Na^+ toxin groups 1 – 18. The group number is followed by the toxin name and the accession number from Swiss-Prot.

```
Gp Name (SP ID)          Sequence

1 Mca (P60254)          --------------------G------DCLPHLKLCKE--NKDCCSKKCKRRGTN----IEKRCR-----------------------------------
2 IpTxI (P59888)        TMWGTKWCGSGNEATDISELGYWSNLDSCCRTHDHCDNIPSGQTKYGLT-NEGKYT--MMNCKCETAFEQCLRNVTGGMEGPAAGFVRKTYFDLYGNGCY
3 Kurtoxin (P58910)     -------------KIDGYPVDYWNCKRICWYNNKYCN-----DLCKGLKADSGYCWGWTLSCYCQGLPDNARIKRSGRCRA------------------
4 BjTx1                 -------------------VG-------CEECPAHCK----GKNAKPTC-DDG-------VCNCNV---------------------------------

1 Mca (P60254)          ---------------------------------
2 IpTxI (P59888)        NVQCPSQSEECPDGVATYTGEAGYGAWAINKLNG
3 Kurtoxin (P58910)     ---------------------------------
4 BjTx1                 ---------------------------------
```

Figure 11 Representative scorpion toxins from Ca^{2+} toxin groups 1 – 4. The group number is followed by the toxin name and the accession number from Swiss-Prot except BjTx1 which was extracted from Zhu *et al*. (2004b).

3.5 Discussion and conclusions

The author produced the first large-scale classification of 393 currently known scorpion toxins which include all toxins specific to Na^+, K^+, Ca^{2+} and Cl^- channels. With a large repository of toxin sequences, a broad perspective of the general patterns in their structure and function can be observed. This classification eliminates a number of errors that resulted from analysing small sample sizes. For example, prior to isolation of K^+ long-chain toxins of 60 – 64 residues (Legros *et al*., 1998), K^+ toxins were classified as short-chain toxins of 30 – 40 residues. With this new information, the primary sequence structure, but not peptide length, can be used to discriminate between long-chain Na^+ and K^+ toxins.

This classification was based on broad ion channel specificity and primary structure similarity because of three important reasons. First, the pharmacology of many of these toxins to ion channels remains to be determined. For example, to the author's knowledge, the pharmacological properties of BmK 41-2 (GenBank ID: AF327643), AamTx (Chen *et al*., 2005) and AamH2 (Chen *et al*., 2003) have not yet been validated. Second, for K^+ toxins, few of the peptides are selective for a given K^+ channel subtypes, for example, charybdotoxin (Gao and Garcia, 2003), maurotoxin (Vissan *et al*., 2004) and several others block both voltage-gated and Ca^{2+}-activated K^+ channels. However, iberiotoxin is a selective blocker of large-conductance Ca^{2+}-activated K^+ channel (Galvez *et al*., 1990) and P05 is specific to small-conductance Ca^{2+}-activated K^+ channel (Wu *et al*., 2002). Third, for Na^+ toxins, information on toxicity to different animal models is lacking. For example, toxicity to animal models was not determined for the 16 peptides identified from the cloning of *Centruroides sculpturatus* Ewing venomous gland (Corona *et al*., 2001). Further, toxins known as

mammal-specific can be toxic to insects and vice versa (Gordon *et al*., 1996). The animal group specificity is only relative and definite cross-reactivity exists (Selisko *et al*., 1996). Information on cross-reactivity remains to be determined. As a result, the primary structure similarity and broad ion channel specificity appear to be the best criteria available for classification purposes in scorpion toxins. The phylogenetic analyses and highly accurate predictions of scorpion toxin functional properties (see Chapter 5) indicate that current groupings are suitable for scorpion toxin characterisation.

Classification and nomenclature of bioactive toxins in animal venoms are still being developed – they are important pre-requisites for structural and functional studies. A lack of consistency in the nomenclature may led to confusion in scorpion toxin classification with possibility of multiple names ascribed to the same toxins (Goudet *et al*., 2002) or the same name used for different toxins (Becerril *et al*., 1997). Toxins have been named according to the order of fractions purified from the venom (e.g. Batista *et al*., 2004; Murgia *et al*., 2004). Currently, classification of two scorpion species *Centruroides exilicauda* Wood and *Centruroides sculpturatus* Ewing remain to be resolved (Valdez-Cruz *et al*., 2004b) where definition of species, genera, families and others is often subjective. A detailed classification based on structure-function relationships will provide a convenient solution to the naming of the toxins.

The author proposed new classification groups for scorpion toxin sequences that have not been classified, reflecting the diversity of different toxins available in the natural library of scorpion venoms. With the growth of number of newly identified scorpion toxins, the view of the structure-function relationships is constantly changing. It is important to note that this present large-scale classification of 393 different toxins in this work relates to current knowledge of the field. In the future, it is expected that

revision of classification will be necessary as more information becomes available. This classification approach, as a generic tool, can be applied to large-scale classification of other bioactive toxins and families of bioactive peptides.

Chapter summary

- Large-scale classification of all known scorpion toxins provides a global overview on their structure-function relationships, aids functional inference of novel scorpion toxins and facilitates retrieval of relevant biological information from large number of toxin data.
- In this work, 393 currently known scorpion toxins have been classified into 62 groups based on ion channel specificity and primary sequence similarity, combined with multiple sequence alignments and phylogenetic analyses. 14 new classification groups were proposed for scorpion toxin sequences that have not been classified.
- In the future, the present classification is expected to be revised to incorporate novel scorpion toxin sequences identified and new information made available.

Part II: Chapter 4 Extraction of functional peptide motifs in scorpion toxins

‘Many of life's failures are people who did not realise how close they were to success when they gave up.’

Thomas Edison

Newly identified scorpion toxins are deposited in public databases which provide limited functional annotation. To aid experimental characterisation of these toxins, function is inferred from identification of similar characterised sequences by pairwise alignment, such as using BLAST (Altschul *et al*., 1997) or FASTA (Pearson, 2000) tools. However, there are known cases where similar structures do not necessary imply functional similarity. The highly similar sequences may differ in function from the characterised sequences where even variation at a critical functional residue can markedly affects the activity of a protein. For example, scorpion toxins Ikitoxin and Birtoxin differ by only one residue at position 23, where Ikitoxin has glutamate instead of glycine, resulting in Ikitoxin having 1000-fold reduced activity than Birtoxin (Inceoglu *et al*., 2002). Pairwise alignment does not differentiate critical functional residues (e.g. an active site) from residues with no critical role (Hulo *et al*., 2004). Another approach to the inference of function involves searching sequence signature databases such as PROSITE (Hulo *et al*., 2004) or PRINTS (Attwood *et al*., 2003) for conserved motifs which usually relate to structural or functional meaning. These motifs were derived statistically from multiple sequence alignments of protein sequences that form a family. For example, the scorpion short toxins signature, C-x(3)-C-x(6,9)-[GAS]-K-C-[IMQT]-x(3)-C-x-C, has been deposited in PROSITE (accession ID: PS01138). However, the motifs that were derived statistically are not necessarily biologically relevant.

Mutation studies of scorpion toxins (such as site-directed mutagenesis and chemical modification) have identified critical residues important for both structural and functional properties (e.g. Cohen *et al*., 2005; Legros *et al*., 2005). The mutation data which provide biologically relevant information on critical residue and position

are available in the literature and normally are not used for extraction of functional motifs in scorpion toxins. Additionally, analyses of 3D structures of toxins can aid identification of the motif where non-contiguous residues in the primary sequence cluster spatially. Herewith, a set of scorpion toxin motifs extracted from analyses of multiple sequence alignment of classified scorpion native toxin sequences, 3D structures and information from mutation studies was described (Tan *et al*., 2006a).

This is the first report of eight functionally relevant binding motifs to Na^+ and K^+ channels which can facilitate the determination of specificity of newly identified scorpion toxins to various ion channel subtypes. Also, these motifs can help in detection of distant relationships between toxin sequences which may be overlooked in pairwise alignment analysis.

4.1 Materials and Methods

Literature on mutation studies of scorpion toxins were searched in the PubMed database (http://www.pubmed.gov/) using keywords such as 'scorpion toxin' and 'mutation'. Data of scorpion mutant toxins and information of mutation on toxin function were extracted from the literature and deposited into the SCORPION2 database. The positions and residue identities of mutation were noted. Effects of mutation on binding affinity were scaled to a common scale for comparison. Mutations that affected toxin function by more than 10-fold were defined as important residues. These residues were then compared to multiple sequence alignment of classified groups of native toxins (discussed in Section 3.3) to verify possible conservation of residues. The spatial organisation of the residues important for structure and function was determined from 3D structure analysis and mapped onto the native toxin primary sequences for extraction of motifs.

4.1.1 Scaling of binding affinities to a common scale in mutant toxin data

Binding affinity data from mutation studies were extracted from the literature and scaled to a common scale for comparison (Equation 1). Only K^+ and Na^+ binding affinity data were mapped because no binding affinity data was available for Cl^- toxins and only a set of five binding affinity measurements was available for Ca^{2+} toxins. Within K^+ and Na^+ groups, eight binding categories were defined for each group, namely of 1) Non-binding 2) Very low binding 3) Low binding 4) Moderately low binding 5) Moderate binding 6) Moderately high binding 7) High binding, and 8) Very high binding. The highest binding affinity data observed for K^+ and Na^+ were 0.08 pM (Koschak *et al*., 1998) and 4 pM (Hassani *et al*., 1999), respectively. The lowest limit for both K^+ and Na^+ was set at 10,000 nM because any concentration higher than this would mean the toxin is a non-binder to the ion channel.

Equation 1 Formula for scaling binding affinities of independent experiments to a common scale. $\log_{10}$ represents the common base 10 logarithm, x is the value of the binding affinity to be scaled. The highest binding affinity values were observed for each ion channel from independent binding experiments. The lowest binding affinity for each ion channel was set at 10,000 nM.

$$\text{Scaled X} = 1 + \frac{\log_{10}\frac{1}{x} - \log_{10}\frac{1}{\text{Lowest Binding Affinity}}}{\log_{10}\frac{1}{\text{Highest Binding Affinity}} - \log_{10}\frac{1}{\text{Lowest Binding Affinity}}} * 7$$

4.1.2 Data analysis

Mutations that affected structural folding, as detected by circular dichroism spectroscopy, were annotated as important to structural integrity. Mutations that affected binding affinity and toxicity by more than 10- and 100-fold as compared to native toxins were termed as ‘influential’ and ‘critical’, respectively. Multiple

sequence alignments of native toxins targeting the same ion channel subtype as the mutant toxins were performed using Clustal W (Thompson *et al*., 1994). These 'influential' and 'critical' residues were then compared to the multiple sequence alignment of native toxins to verify possible conservation of residues. The spatial organisation of the residues important for structure and function was visualised using Molsoft (http://www.molsoft.com/) and mapped onto the native toxin primary sequences for extraction of motifs. The motifs are represented as used in PROSITE database (Hulo *et al*., 2004). To test the relevance of these motifs, each was searched in Swiss-Prot (release 48.0) and TrEMBL (release 31.0) databases using ScanProsite (Gattiker *et al*., 2002) to scan for protein sequences containing the motif.

4.2 Results and discussion

In SCORPION2, a total of 426 scorpion mutant toxin records were extracted from 72 literature sources on mutation studies of scorpion toxins (as of November 2005). Scaling binding affinities of scorpion mutant toxins to a common logarithmic scale is useful for comparing effects of mutation on function. An example is shown in **Figure 12**, where Agitoxin 2 targets Shaker K^+ channel at moderate range after mapping to the common scale (K_d = 0.741 nM) (Ranganathan *et al*., 1996). Mutation study on Agitoxin 2 suggested that K27 and N30 are critical for binding affinity towards this channel (>600-fold reduced binding affinity as compared to that of native Agitoxin 2), while S11, R24, F25, M29 and T36 influence the ligand-channel interaction (>10-fold but <100-fold reduction), and T9, G10, R31, K32, H34 and P37 can be substituted without significant effect on binding affinity (<10-fold reduction) (Ranganathan *et al*., 1996).

The author extracted eight new motifs from analyses of native scorpion toxin sequences, 3D structures and mutation studies of Na^+ (four motifs) and K^+ (four motifs) toxins (**Table 6**). These motifs are highly specific to scorpion toxins where all the sequences returned by ScanProsite (Gattiker *et al*., 2002) were of scorpions i.e. no false positives were extracted. Further, these motifs can be used to search uncharacterised scorpion toxin sequences for inference of their specificities to different ion channel subtypes.

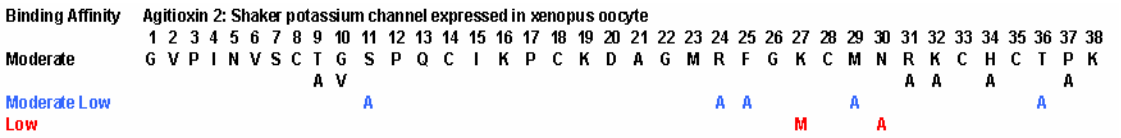

Figure 12 Scaling binding affinities of Agitoxin 2 and its mutant sequences for analysis of residues important for interaction with Shaker K^+ channel. Native Agitoxin 2 binds moderately to Shaker K^+ channel. Mutations at positions 9, 10, 31, 32, 34 and 37 did not affect binding affinity significantly (<7-fold). However, mutations at positions 11, 24, 25, 29 and 36 resulted in moderately low affinities. Further, mutations of lysine and asparagine to methionine and alanine at positions 27 and 30 respectively resulted in low affinities towards the channel. This suggests that lysine and asparagine at these positions are important for binding to Shaker K^+ channel.

Table 6 Motifs of scorpion toxins extracted for Na^+, K^+ and Cl^- channels. K_V – voltage-dependent K^+ channels, BK_{Ca} – large conductance Ca^{2+}-activated channel, SK_{Ca} – small conductance Ca^{2+}-activated channel, ERG – Ether-a-go-go K^+ channel. 'Hits' – number of sequences that contain the submitted motif upon search in Swiss-Prot and TrEMBL databases using ScanProsite. 'Expt.' – experimentally determined function. 'No info.' – Functional information is unavailable.

Na^+		**Hits**	**Expt.**	**No info.**
α	R-D-x-Y-I-x(4)-N-C-x-Y-x-C-x(5,7)-C-N-x-x-C-T-x-x-G-A-x(3,4)-Y-C-x(6)-G-N-x-C-x-C-x-x-L-P-x(4)-I-x(4,5)-[KR]-C-[HR]	22	4	15
α-like	R-D-x-Y-I-A-x(3)-N-C-x(3)-C-x(3,6)-C-x-x-L-C-x(3)-G-x(3)-G-x-C-x(6)-G-x-x-C-W-C-x-x-L-P-x-x-V-x-I-x(3)-G-K-C-H	16	4	9
β-excitatory	G-x(3)-D-x-x-G-K-x-x-E-C-x(4,9)-Y-C-x-x-E-C-x-K-V-x-Y-A-x-x-G-Y-C-C-x(3)-C-Y-C-x-G-L-x(16)-C	9	4	5
β-mammal	K-x-G-Y-x-V-x(4)-G-C-x(3)-C-x-x-L-G-x-N-x-x-C-x-x-E-C-x(9)-G-Y-C-Y-x-F-x-C[WY]-C-x-x-L-x(8)-L-x-x-K-x-C	25	9	14
K^+				
Kv	C-x-x-[SP]-x(1,2)-C-[YWIDLG]-x-x-C-x(8,10)-K-C-[MI]-N-x-x-C-[KRH]-C	27	19	8
BK_{Ca}	C-x-x-[SP]-x(1,2)-C-[YWIDLG]-x-x-C-x(8,10)-K-C-[MI]-[NG]-x-x-C-[KRH]-C	28	7	19
SK_{Ca}	C-x-x-[RK]-[RM]-C-x(3)-C-[RK]-x(7)-C-x(4)-C-x-C	5	5	-
ERG	C-x(3)-Y-x-C-x(3)-C-K-x-R-F-x-K-x(3)-R-C-x(4)-C-x-C	2	1	1
Cl^-	C-x-P-C-F-T-x(8)-C-x(2)-C-C-x(5,7)-C-x(2,3)-Q-C-[LI]-C	14	-	14

4.2.1 Chloride channel motif

In PROSITE document of scorpion short toxins signature (PDOC00875), Cl^- toxins was annotated as having the pattern, C-x(3)-C-x(6,9)-[GAS]-K-C-[IMQT]-x(3)-C-x-C. However, the motif that the author extracted based on conservation of residues among 18 Cl^- toxins was C-x-P-C-F-T-x(8)-C-x(2)-C-C-x(5,7)-C-x(2,3)-Q-C-[LI]-C (**Figure 13**). The conserved cysteine pattern in Cl^- toxins is different from that reported in the PROSITE document. This conserved cysteine pattern is distinct from those observed in Na^+ and K^+ toxins and can be used to differentiate Cl^- toxins from the other toxins.

Cl^- toxins adopt the cysteine-stabilised α-helix (CSH) fold commonly observed in scorpion toxins and the conserved residues are clustered spatially at the β-sheets and loop surfaces (**Figure 14**). Since these 18 toxins shared 48 – 100% identity and their 3D foldings are conserved, this cluster could form the putative interaction site in Cl^- toxins. Mutation performed on these conserved residues would clarify their roles in function and structure.

```
                1        10         20         30
                |        |          |          |
D000104         MCMPCFTTDPNMAKKCRDCCGGNGK--CFGPQCLCNR-- 35
D000206         MCMPCFTTDHNMAKKCRDCCGGNGK--CFGPQCLCNR-- 35
D000207         MCMPCFTTDPNMANKCRDCCGGGKK--CFGPQCLCNR-- 35
D000154         -CGPCFTTDANMARKCRECCGGIGK--CFGPQCLCNRI- 35
D000615         -CGPCFTTDANMARKCRECCGGNGK--CFGPQCLCNRE- 35
D000106         MCMPCFTTDHQMARKCDDCCGGKGRGKCYGPQCLCR--- 36
D000237         MCMPCFTTDHQMARKCDDCCGGKGRGKCYGPQCLCRG-- 37
D000205         MCMPCFTTDHQTARRCRDCCGGRGR-KCFG-QCLCGYD- 36
D000110         MCMPCFTTRPDMAQQCRACCKGRGK--CFGPQCLCGYD- 36
D000111         -CGPCFTTDPYTESKCATCCGGRGK--CVGPQCLCNRI- 35
D000171         -CGPCFTKDPETEKKCATCCGGIGR--CFGPQCLCNRGY 36
D000238         -CGPCFTTDHQTEQKCAECCGGIGK--CYGPQCLC-RG- 34
D000239         -CGPCFTTDRQMEQKCAECCGGIGK--CYGPQCLC-RG- 34
D000240         -CGPCFTTDHQTEQKCAECCGGIGK--CYGPQCLCNRG- 35
D000105         RCSPCFTTDQQMTKKCYDCCGGKGKGKCYGPQCICAPY- 38
D000109         RCKPCFTTDPQMSKKCADCCGGKGKGKCYGPQCLC---- 35
D000140         RCGPCFTTDPQTQAKCSECCGRKG-GVCKGPQCICGIQ- 37
D000629         RCPPCFTTNPNMEADCRKCCGGRGY--CASYQCICPGG- 36
                 * ****.       *  **       * . **:*
```

Figure 13 Conserved residues of 18 Cl^- specific scorpion toxins are **C-x-P-C-F-T-x(8)-C-x(2)-C-C-x(5-7)-C-x(2-3)-Q-C-[LI]-C**. The first column represents the record accession number found in the SCORPION2 database; second column represents the toxin primary sequences; third column represents the peptide length. Multiple sequence alignment was generated by Clustal W.

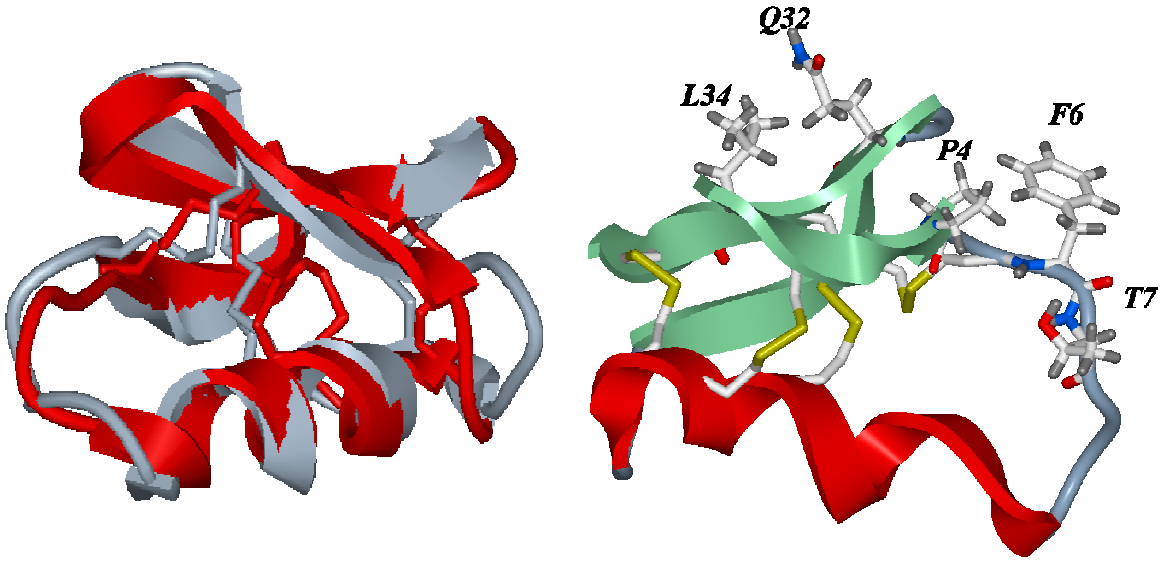

Figure 14 Scorpion toxins specific for Cl^- channel adopt the cysteine-stabilised α-helix fold where the β-sheets and α-helix are connected by three disulfide bonds. The fourth disulfide bond is formed between the β-sheet and the loop. **A)** Superimposition of chlorotoxin (PDB: 1CHL, grey) and insectotoxin I5A (PDB: 1SIS, red) with root-mean-square distance (rmsd) = 2.00 Å demonstrated that 3D foldings are conserved. **B)** The conserved residues among 18 Cl^- toxins are P4, F6, T7, Q32 and L34 (numbered as in chlorotoxin) where they are clustered spatially at the β-sheets and the preceding loop.

4.2.2 Sodium channels – β-excitatory motif

Inclusion of mutation studies in the analyses of conserved residues helps determine their importance for structure and function of toxin peptides. For example, in Na^+ β-excitatory toxins, the conserved residues are: **KK**xGxxxDxxGKxxECx(4,9)YCxxxC**T**KVxYAxxGYCCxxxCYCxGLx**DDK**x(9)**K**xxC**D** (**Figure 15**). Mutation studies of Bjxtr-IT (Cohen *et al*., 2004; Karbat *et al*., 2004a) demonstrated that toxicity to insects and binding affinity to insect Na^+ channels as compared to that of wild type Bjxtr-IT were mildly affected (<7-fold) for K1, K2, T32, D54, D55, K56, K66 and D70 (shown in bold). However, mutations at D8, K12, Y26, K33 and V34 strongly affected binding affinity, ranging from 12 to >10,000-fold reduction in binding affinity. Thus, different conserved residues play different roles in function and structure of scorpion toxins. Further, when negative-charged E15 was mutated to positive-charged arginine, toxicity to insects was severely affected (>10,000-fold) but not binding affinity (<7-fold). This suggests that negative-charged E15 is involved in toxic action.

Another residue important for toxicity and binding affinity is glutamate at position 30 in the multiple sequence alignment of β-excitatory toxins (**Figure 15**). Residues are not conserved at position 30 because glutamine, isoleucine and asparagine were also observed. Mutation of E30 to glutamine, leucine, arginine and even conservative substitution to aspartate caused reduction of more than 8-fold in toxicity and 43-fold in binding affinity. Without mutant data, this position and residue identity (negative-charged and hydrophilic), which is important for function, would be overlooked. Thus, by incorporation of mutation information, the functional motif of β-excitatory toxins to insect Na^+ channels can be summarised as G-x(3)-D-x-x-G-K-x-x-E-C-x(4,9)-Y-C-x-x-E-C-x-K-V-x-Y-A-x-x-G-Y-C-C-x(3)-C-Y-C-x-G-L-x(16)-C

where E is important for toxicity to insects. These functional residues are clustered spatially at one surface of Bjxtr-IT (**Figure 16**).

```
                1       10        20        30        40        50        60
                |       |         |         |         |         |         |
D000004         KKNGYAVDSSGKAPEC-----LLSNYCNNECTKVHYADKGYCCLLSCYCFGLNDDKKVLE 55
D000137         KKNGYAVDSSGKAPEC-----LLSNYCNNECTKVHYADKGYCCLLSCYCFGLNDDKKVLE 55
D000125         KKNGYAVDSSGKAPEC-----LLSNYCNNQCTKVHYADKGYCCLLSCYCFGLNDDKKVLE 55
D000005         KKNGYAVDSSGKAPEC-----LLSNYCYNECTKVHYADKGYCCLLSCYCFGLNDDKKVLE 55
D000139         KKDGYAVDSSGKAPEC-----LLSNYCYNECTKVHYADKGYCCLLSCYCFGLNDDKKVLE 55
D000824         KKNGYAVDSSGKAPEC-----LLSNYCNNECTKVHYADKGYCCLLSCYCFGLSDDKKVLE 55
D000825         KKNGYAVDSSGKAPEC-----LLSNYCNNECTKVHYADKGYCCLLSCYCFGLSDDKKVLD 55
D000006         KKNGYAVDSSGKAPEC-----LLSNYCYNECTKVHYADKGYCCLLSCYCVGLSDDKKVLE 55
D000007         KKNGYAVDSSGKAPEC-----LLSNYCYNECTKVHYAEKGYCCLLSCYCVGLSDDKKVLE 55
D000233         KKNGYAVDSKGKAPEC-----FLSNYCNNECTKVHYADKGYCCLLSCYCFGLNDDKKVLE 55
D000234         KKNGYAVDSKGKAPEC-----FLSNYCNNECTKVHYADKGYCCLLSCYCFGLNDDKKVLE 55
D000235         KKNGYAVDSKGKAPEC-----FFSNYCNNECTKVHYAEKGYCCLLSCYCVGLNDDKKVME 55
D000236         KKNGFAVDSNGKAPEC-----FFDHYCNSECTKVYYAEKGYCCTLSCYCVGLDDDKKVLD 55
D000165         KKNGYAVDSSGKVAEC-----LFNNYCNNECTKVYYADKGYCCLLKCYCFGLADDKPVLD 55
D000003         KKNGYAVDSSGKVSEC-----LLNNYCNNICTKVYYATSGYCCLLSCYCFGLDDDKAVLK 55
D000136         KKNGYAVDSSGKVSEC-----LLNNYCNNICTKVYYATSGYCCLLSCYCFGLDDDKAVLK 55
D000184         KKNGYAVDSSGKVSEC-----LLNNYCNINCTKVYYATSGYCCLLSCYCFGLDDDKAVLK 55
D000001         KKNGYPLDRNGKTTECSGVNAIAPHYCNSECTKVYYAESGYCCWGACYCFGLEDDKPIGP 60
D000002         KKNGYPLDRNGKTTECSGVNAIAPHYCNSECTKVYYAKSGYCCWGACYCFGLEDDKPIGP 60
                **:*:.:* .**..**     :  :**   ****:** .****   ***.** *** :

                        70
                        |
D000004         ISDTRKSYCDTTIIN-- 70
D000137         ISDTRKSYCDTTIIN-- 70
D000125         ISDTRKSYCDTTIIN-- 70
D000005         ISDTRKSYCDTPIIN-- 70
D000139         ISDTRKSYCDTPIIN-- 70
D000824         ISDTRKKYCDYTIIN-- 70
D000825         ISDTRKKYCDYTIIN-- 70
D000006         ISDARKKYCDFVTIN-- 70
D000007         ISDARKKYCDFVTIN-- 70
D000233         ISGTTKKYCDFTIIN-- 70
D000234         ISDTTKKYCDFTIIN-- 70
D000235         ISDTRKKICDTTIIN-- 70
D000236         ISDTRKKLCDFTLFN-- 70
D000165         IWDSTKNYCDVQIIDLS 72
D000003         IKDATKSYCDVQIN--- 69
D000136         IKDATKSYCDVQIIG-- 70
D000184         IKDATKSYCDVQIIN-- 70
D000001         MKDITKKYCDVQIIPS- 76
D000002         MKDITKKYCDVQIIPS- 76
                : .  *. **
```

Figure 15 Conserved residues of 19 Na^+ β-excitatory toxins are KKxGxxxDxxGKxxECx(4,9)YCxxxCTKVxYAxxGYCCxxxCYCxGLxDDKx(9)KxxCD where x represents any residue and number(s) in parenthesis represents the number of intervening residues. Multiple sequence alignment was generated by Clustal W. The first column represents the record accession number found in the SCORPION2 database; second column represents the toxin primary sequences; third column represents the peptide length. Mutation studies were performed on Bjxtr-IT (D000001).

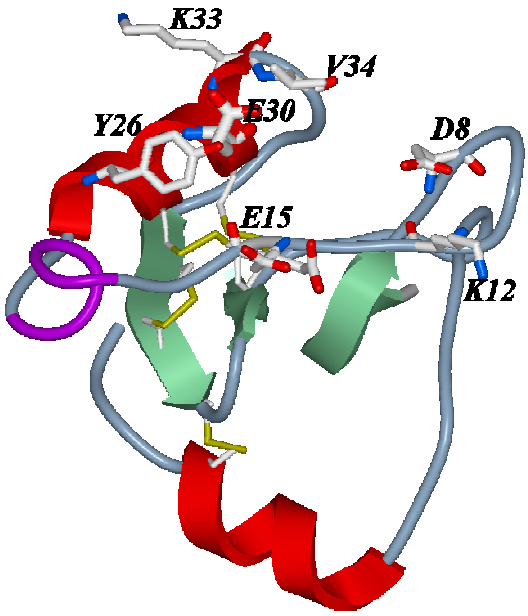

Figure 16 Functional motif of β-excitatory toxins represented by Bjxtr-IT (PDB: 1BCG). Only residues that affected binding affinity and toxicity determined in mutation studies are displayed: D8, K12, E15, Y26, E30, K33 and V34. These are clustered spatially at one surface of Bjxtr-IT.

4.2.3 Sodium channels – β-mammal motif

In β-mammal specific toxins, the conserved residues in this group are KxGYxVx(4)GC**K**xxCxxLGxNxxCxxECx(9)GYCYxFxCxCxxLx(7)**P**LxxKxC (**Figure 17**). Mutation studies of Css 4 (Cohen *et al*., 2005) demonstrated that mutation of K13 and P61 (shown in bold) did not affect binding affinity to mammal sodium channels significantly (≤2-fold). Y4, W47 and K63 are involved in structural integrity – their mutations to alanine disrupted secondary structures. Substitution of N22, Y40, Y42 and F44 to alanine resulted in 600-fold reduction of binding affinity while L19 to alanine resulted in 150-fold reduction in binding affinity to mammal sodium channels. This suggests different positions in the toxin sequence are more important in binding affinity than others. Information such as the physicochemical properties of functional residues can also be obtained from mutation studies. For example, a stronger reduction

on binding affinity was obtained upon charge inversion of E28 to arginine (900-fold) than charge neutralising substitutions to alanine (600-fold), glutamine (400-fold) and leucine (50-fold). The functional motif for β-mammal specific toxins can be summarised as K-x-G-Y-x-V-x(4)-G-C-x(3)-C-x-x-L-G-x-N-x-x-C-x-x-E-C-x(9)-G-Y-C-Y-x-F-x-C[WY]-C-x-x-L-x(8)-L-x-x-K-x-C. The 3D structure of Css 4 was modeled with template 1CN2 where the sequence identity between target and template is 83%. The functional residues are clustered spatially at the surface (**Figure 18**).

```
                1       10        20        30        40        50        60
                |       |         |         |         |         |         |
D000181         KEGYLVNSYTGCKFECFKLGDNDYCLRECRQQYGKGSGGYCYAFGCWCTHLYEQAVVWPL 60
D000182         KEGYLVNSYTGCKFECFKLGDNDYCKRECKQQYGKSSGGYCYAFGCWCTHLYEQAVVWPL 60
D000176         KEGYLVNHSTGCKYECYKLGDNDYCLRECKQQYGKGAGGYCYAFGCWCTHLYEQAVVWPL 60
D000177         KEGYLVNHSTGCKYECFKLGDNDYCLRECKQQYGKGAGGYCYAFGCWCNHLYEQAVVWPL 60
D000057         KEGYIVNLSTGCKYECYKLGDNDYCLRECKQQYGKGAGGYCYAFGCWCTHLYEQAVVWPL 60
D000056         KEGYLVNHSTGCKYECFKLGDNDYCLRECRQQYGKGAGGYCYAFGCWCTHLYEQAVVWPL 60
D000041         KEGYLVDKNTGCKYECLKLGDNDYCLRECKQQYGKGAGGYCYAFACWCTHLYEQAIVWPL 60
D000051         KEGYLVSKSTGCKYECLKLGDNDYCLRECKQQYGKSSGGYCYAFACWCTHLYEQAVVWPL 60
D000054         KEGYLVELGTGCKYECFKLGDNDYCLRECKARYGKGAGGYCYAFGCWCTQLYEQAVVWPL 60
D000052         KEGYLVKKSDGCKYGCLKLGENEGCDTECKAKNQGGSYGYCYAFACWCEGLPESTPTYPL 60
D000058         KEGYLVNKSTGCKYGCLKLGENEGCDKECKAKNQGGSYGYCYAFACWCEGLPESTPTYPL 60
D000053         KEGYLVNKSTGCKYGCFWLGKNENCDKECKAKNQGGSYGYCYSFACWCEGLPESTPTYPL 60
D000035         KDGYLVEK-TGCKKTCYKLGENDFCNRECKWKHIGGSYGYCYGFGCYCEGLPDSTQTWPL 59
                *:**:*.   ***  *  **.*: *  **: :   .: ****.*.*:*  * :.: .:**

D000181         PNKTCN 66
D000182         PNKTCN 66
D000176         PKKTCN 66
D000177         PKKTCN 66
D000057         PKKTCT 66
D000056         PNKTCS 66
D000041         PNKRCS 66
D000051         PNKTCN 66
D000054         KNKTCR 66
D000052         PNKSC- 65
D000058         PNKSCS 66
D000053         PNKSCS 66
D000035         PNKTC- 64
                 :* *
```

Figure 17 Conserved residues of 13 experimentally determined β-mammal toxins are KxGYxVx(4)GCKxxCxxLGxNxxCxxECx(9)GYCYxFxCxCxxLx(7)PLxxKxC where x represents any residue and number(s) in parenthesis represents the number of intervening residues. Multiple sequence alignment was generated by Clustal W. The first column represents the record accession number found in the SCORPION2 database; second column represents the toxin primary sequences; third column represents the peptide length. Mutation studies were performed on Css 4 (D000181).

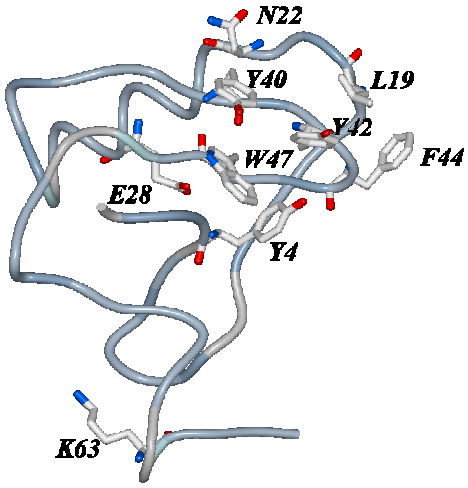

Figure 18 The functional residues of Css 4, a β-mammal toxin, are clustered spatially at one surface of the molecule. Only residues that had mutant information are shown. This is a homology model built from the template, 1CN2 which have sequence identity of 83% with Css 4.

4.2.4 Sodium channels – α-motif

In α-toxins, the motif extracted from 14 experimentally determined α-toxins is R-D-x-Y-I-x(4)-N-C-x-Y-x-C-x(5,7)-C-N-x-x-C-T-x-x-G-A-x(3,4)-Y-C-x(6)-G-N-x-C-x-C-x-x-L-P-x(4)-I-x(4,5)-[KR]-C-[HR]. Mutation studies of Lqh αIT (Zilberberg *et al*., 1997; Karbat *et al*., 2004b) showed that residues at positions 15, 25, 28, 56 and 54 (as in Lqh αIT) can be mutated to alanine without significant effect on binding affinity to insect Na^+ channels (<2-fold reduction), which correlated well with the observation that there is no conservation of residues at these positions (**Figure 19**). Conserved N44 is involved in structural and functional integrity because mutation to alanine disrupted the secondary structures and reduced binding affinity by 31-fold (**Figure 20**). Mutation of I57 to alanine and threonine caused more than 92-fold reduction in binding affinity. Positively charged residues at position 62 and 64 were essential to binding affinity as mutation to the neutral alanine resulted in more than 50-fold reduction of peptide binding.

```
                1       10        20        30        40        50        60
                |       |         |         |         |         |         |
D000025         VRDAYIAKNYNCVYECFRD--AYCNELCTKNGASSG-YCQWAGKYGNACWCYALPDNVPI 57
D000029         VRDAYIAKNYNCVYECFRD--SYCNDLCTKNGASSG-YCQWAGKYGNACWCYALPDNVPI 57
D000030         VRDAYIAQNYNCVYFCMKD--DYCNDLCTKNGASSG-YCQWAGKYGNACWCYALPDNVPI 57
D000040         VRDAYIAKPENCVYHCATN--EGCNKLCTDNGAESG-YCQWGGRYGNACWCIKLPDRVPI 57
D000042         VRDAYIAKPENCVYECATN--EYCNKLCTDNGAESG-YCQWVGRYGNACXCIKLPDRVPI 57
D000037         VRDAYIAKPENCVYECGIT--QDCNKLCTENGAESG-YCQWGGKYGNACWCIKLPDSVPI 57
D000038         VRDAYIAKPHNCVYECARN--EYCNDLCTKNGAKSG-YCQWVGKYGNGCWCIELPDNVPI 57
D000634         VRDGYIALPHNCAYGCLNN--EYCNNLCTKDGAKIG-YCNIVGKYGNACWCIQLPDNVPI 57
D000045         ARDAYIAKPHNCVYECYNPKGSYCNDLCTENGAESG-YCQILGKYGNACWCIQLPDNVPI 59
D000028         GRDAYIAQPENCVYECAKN--SYCNDLCTKNGAKSG-YCQWLGRWGNACYCIDLPDKVPI 57
D000068         GRDAYIAQPENCVYECAQN--SYCNDLCTKNGATSG-YCQWLGKYGNACWCKDLPDNVPI 57
D000031         GRDAYIADSENCTYTCALN--PYCNDLCTKNGAKSG-YCQWAGRYGNACWCIDLPDKVPI 57
D000047         GRDAYIADSENCTYFCGSN--PYCNDVCTENGAKSG-YCQWAGRYGNACYCIDLPASERI 57
D000027         ERDGYIVQLHNCVYHCGLN--PYCNGLCTKNGATSGSYCQWMTKWGNACYCYALPDKVPI 58
                 **.**.   **.* *       ** :**.:**  * **:   ::**.* *  **    *

D000025         R-VPGKCHRK 66
D000029         R-VPGKCH-- 64
D000030         R-IPGKCHS- 65
D000040         R-VPGKCHR- 65
D000042         R-VWGKCHG- 65
D000037         R-VPGKCQR- 65
D000038         R-VPGKCHR- 65
D000634         R-VPGRCHPA 66
D000045         R-IPGKCH-- 66
D000028         R-IEGKCHF- 65
D000068         R-IPGKCHF- 65
D000031         R-ISGSCR-- 64
D000047         K-EPGKCG-- 64
D000027         KWLDPKCY-- 66
                :     *
```

Figure 19 Conserved residues of 14 experimentally determined α-toxins are RDxYIx(4)NCxYxCx(5,7)CNxxCTxxGAxxGYCx(6)GNxCxCxxLPx(4)Ix(5,6)C where x represents any residue and number(s) in parenthesis represents the number of intervening residues. Multiple sequence alignment was generated by Clustal W. The first column represents the record accession number found in the SCORPION2 database; second column represents the toxin primary sequences; third column represents the peptide length. Mutation studies were performed on Lqh αIT (D000025).

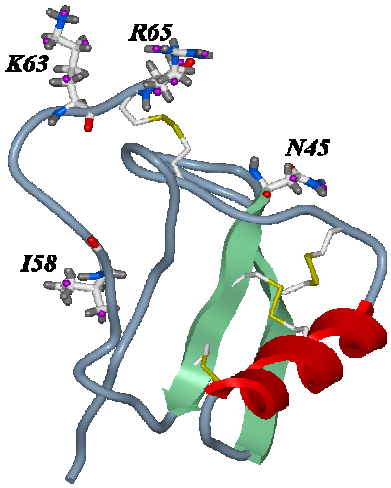

Figure 20 Functional and structural residues of Lqh αIT (PDB: 1LQH), an α-toxin. N45 is involved in structural and functional integrity. I58, K63 and R65 are important for binding affinity towards insect Na^+ channel.

4.2.5 Sodium channels – α-like motif

The motif for α-like toxins is R-D-x-Y-I-A-x(3)-N-C-x(3)-C-x(3,6)-C-x-x-L-C-x(3)-G-x(3)-G-x-C-x(6)-G-x-x-C-W-C-x-x-L-P-x-x-V-x-I-x(3)-G-K-C-H. The site-directed mutagenesis of BmK M1 (Sun *et al*., 2003; Wang *et al*., 2003) highlighted the importance of four residues important for function and structure in α-like toxins (**Figure 21**). Mutations at Y5 and N11 disrupted the secondary structures as measured by circular dichroism spectroscopy, suggesting their role in maintaining structural integrity. This is corroborated by the crystal structure of BmK M1 where N11 forms hydrogen bonds with residues 58 and 59 (He *et al*., 1999). Positive charged at positions 62 and 64 is important for binding affinity to insect Na^+ channel where negative-charged residues (D and E) resulted in more than 167-fold reduction. Mutation of conserved P9 did not significantly affect function (two-fold reduction in toxicity and binding affinity) (**Figure 22**).

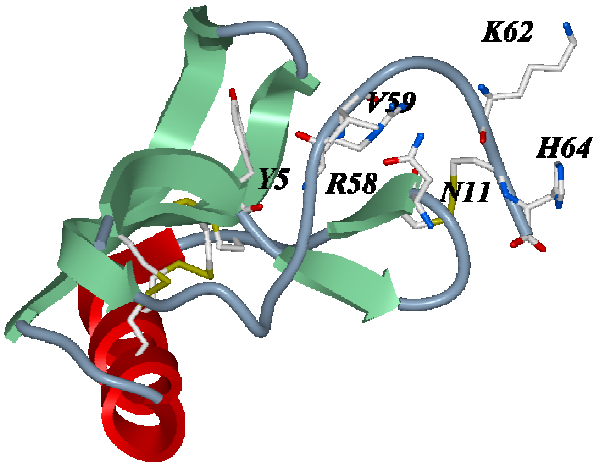

Figure 21 Functional and structural residues of BmK M1 (PDB: 1DJT), an α-like toxin as determined by mutation studies. Y5 and N11 are involved in structural integrity where N11 forms hydrogen bonds with R58 and V59. K62 and H64 are important for binding affinity towards insect Na^+ channel.

```
                1       10        20        30        40        50        60
                |       |         |         |         |         |         |
D000621         VRDGYIAQPENCVYHCIP---DCDTLCKDNGGTGGHCGFKLGHGIACWCNALPDNVGIIV 57
D000622         VRDGYIAKPENCAHHCFPGSSGCDTLCKENGGTGGHCGFKVGHGTACWCNALPDKVGIIV 60
D000073         VRDGYIAQPENCVYHCFPGSSGCDTLCKEKGGTSGHCGFKVGHGLACWCNALPDNVGIIV 60
D000077         GRDGYIAQPENCVYHCFPGSSGCDTLCKEKGATSGHCGFLPGSGVACWCDNLPNKVPIVV 60
D000038         VRDAYIAKPHNCVYECAR-NEYCNDLCTKNGAKSGYCQWVGKYGNGCWCIELPDNVPIRV 59
D000143         GRDAYIAQPENCVYECAK-NSYCNDLCTKNGAKSGYCQWLGKYGNACWCEDLPDNVPIRI 59
D000634         VRDGYIALPHNCAYGCLN-NEYCNNLCTKDGAKIGYCNIVGKYGNACWCIQLPDNVPIRV 59
D000126         VRDAYIAKPENCVYHCAG-NEGCNKLCTDNGAESGYCQWGGRYGNACWCIKLPDDVPIRV 59
                 **.*** *.**.: *       *: **...*.  *:*       * .***  **:.* * :

D000621         DGVKCHK- 64
D000622         DGVKCH-- 66
D000073         EGEKCHS- 67
D000077         GGEKCH-- 66
D000038         PG-KCHR- 65
D000143         PG-KCHF- 65
D000634         PG-RCHPA 66
D000126         PG-KCHR- 65
                 * :**
```

Figure 22 Conserved residues of eight experimentally determined α-like toxins are RDxYIAxPxNCxxxCx(3,5,6)CxxLCxxxGxxxGxCx(6)GxxCWCxxLPxxVxIx(3)GKC H, where x represents any residue and the number in parenthesis represents the number of intervening residues. Multiple sequence alignment was generated by Clustal W. The first column represents the record accession number found in the SCORPION2 database; second column represents the toxin primary sequences; third column represents the peptide length. Mutation studies were performed on BmK M1 (D000038).

From the analysis of 222 aligned Na^+ native toxin sequences, a conserved tyrosine was observed at the N-terminal (data not shown) where mutation studies suggest its involvement in maintaining the structural folds of the toxins (Wang *et al*., 2003; Cohen *et al*., 2005). Also, a tyrosine is conserved at one position before the fourth cysteine in excitatory toxins and the fifth cysteine in the rest of Na^+ toxins which is involved in structure and function of Na^+ toxins (Sun *et al*., 2003). In binding studies, the α and α-like toxins compete to bind at receptor site 3 of insect Na^+ channels (Gilles *et al*., 1999), which could explain the similarity of positive residues observed in these toxins at the C-terminal that are involved in binding affinity. In β toxins, the presence of glutamate before the third cysteine in excitatory toxins and fourth cysteine in the rest is conserved which is important for binding affinity (Cohen *et al*., 2005). The difference observed between both α (include α-like toxins) and β

toxins is that α toxins have an asparagine before the first cysteine but glycine is observed at this position instead in β toxins. Mutation in BmK M1 suggested asparagine was involved in folding (Wang *et al*., 2003). There is a need for more mutation studies performed on other Na^+ toxins and other positions in the toxin sequences to determine their structure-function relationships.

4.2.6 Potassium channel subtype – Ether-a-go-go-related K^+ channel motif

Scorpion toxins target various K^+ channel subtypes with different specificities. In ether-a-go-go-related gene K^+ (ERG) channel subtype, residues that are important for binding of BeKm-1 to the ERG channel had been identified by mutagenesis (Korolkova *et al*., 2002). Y11, K18, R20 and K23 were critical to binding while F21 and R27 were involved in structural folding. The functional residues were located on the α-helix and the following loop (**Figure 23**). The recognition motif for ERG subtype can be expressed as C-x(3)-Y-x-C-x(3)-C-K-x-R-F-x-K-x(3)-R-C-x(4)-C-x-C.

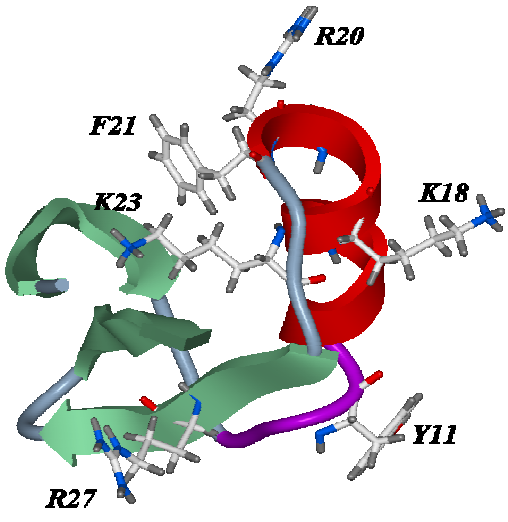

Figure 23 Mutagenesis study of BeKm-1 identified Y11, K18, R20 and K23 (in bold) were important for binding to human ERG channel. F21 and R27 were involved in structural folding.

4.2.7 Potassium channel subtype – Small conductance Ca^{2+}-activated K^+ channel motif

The ERG functional surface is similarly shared by scorpion toxins that target small conductance Ca^{2+}-activated K^+ channels (SK_{Ca}). Mutagenesis of leiurotoxin and P05 identified two positions, 6 and 7, as important for binding to this subtype (Sabatier *et al*., 1993; Sabatier *et al*., 1994; Shakkottai *et al*., 2001; Wu *et al*., 2002) (**Figure 24A**). However, the Ts-κ which also targets this subtype has two known functional residues, R6 and R9, located outside the α-helix structure (Lecomte *et al*., 1999) (**Figure 24B**). Though the positions of the functional residues are not spatially conserved, the residue identity required to bind to SK_{Ca} was positive-charged arginine where mutation to hydrophobic leucine and even conservative substitution to lysine affected binding affinity.

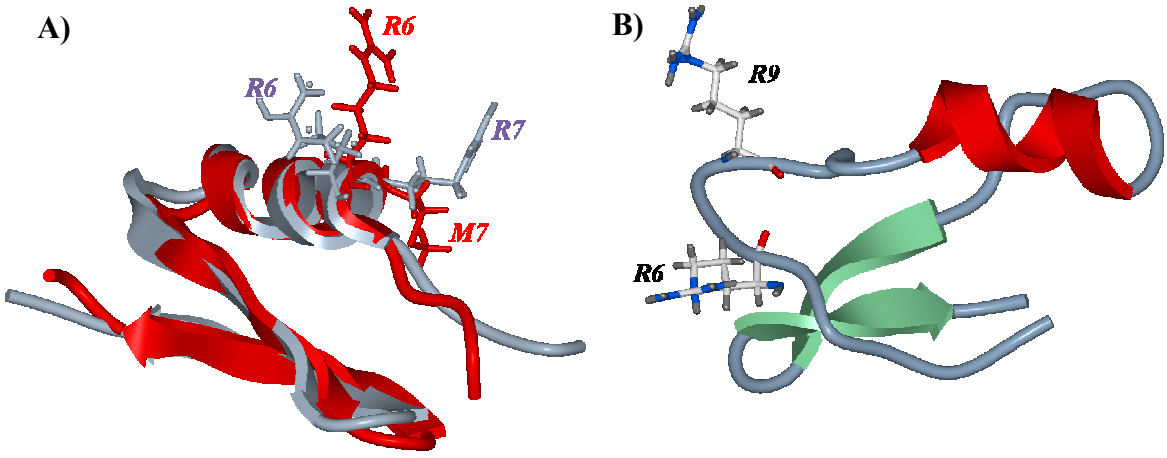

Figure 24 Functional residues of scorpion toxins targeting small conductance Ca^{2+}-activated K^+ channels. **A)** Superimposition of leiurotoxin (PDB: 1SCY, red) and P05 (PDB: 1PNH, grey) with rmsd = 1.57 Å. The two functional residues, R and (R/M) extend from the α helix. **B)** The two functional residues, R6 and R9 in Ts-κ (PDB: 1TSK) were located outside the α helical structure.

4.2.8 Potassium channel subtypes – Large conductance Ca^{2+}-activated K^+ channel and voltage-dependent K^+ channel motifs

For scorpion toxins that bind to voltage-dependent K^+ channels (K_V), Goldstein and Miller (1993) had demonstrated that lysine at position 27 in charybdotoxin from *Leiurus quinquestriatus hebraeus* physically occlude the pore of this subtype, thus preventing the flow of K^+ ions. This functional residue, lysine and an aromatic residue (tyrosine or phenylalanine) separated by 6.6 ± 1.0 Å formed the functional dyad that target K_V channels (Dauplais *et al*., 1997). In addition, mutagenesis of charybdotoxin on K_V channels and large conductance Ca^{2+}-activated channels (BK_{Ca}) highlighted several positions which are important for binding (Park and Miller, 1992; Goldstein and Miller, 1993; Stampe *et al*., 1994). These include positions 10, 14, 25, 29 and 34 where mutations led to drastic reduction in binding affinity. The functional residues for K_V and BK_{Ca} channels are located on the β-sheets in contrast to that for ERG and SK_{Ca} channels which are located at the α-helix. These residues also fall within the functional dyad radius of 6.6 ± 1.0 Å, which may explain their involvement in toxin-channel interaction due to their close proximity to the functional dyad (**Figure 25**). Only a single residue difference in the functional surface distinguishes K_V and BK_{Ca} channels where asparagine is specific to K_V channels. This was determined from analyses of multiple sequence alignment of scorpion toxins that target different channel subtypes and mutant data (**Figure 26**). Iberiotoxin that is highly specific for BK_{Ca} channel (Galvez *et al*., 1990) has glycine instead of asparagine at position 30. P05 targets SK_{Ca} channel only and mutation at positions 22-24 from IGD to MNG resulted in recognition of K_V channels (Wu *et al*., 2002). This is corroborated by Schroeder *et al*. (2002) where they mutated glycine to asparagine at position 30 in iberiotoxin, which caused the mutant toxin to target both K^+ channel subtypes.

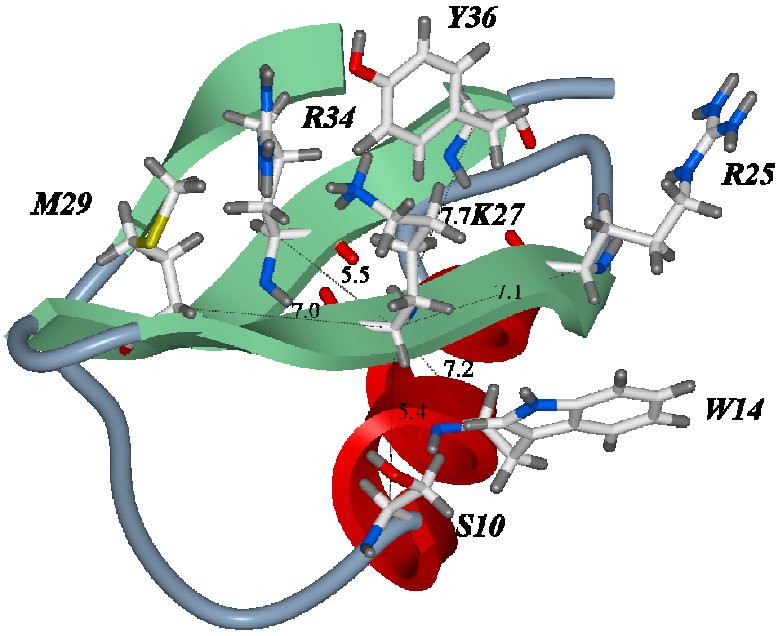

Figure 25 Functional residues of charybdotoxin (PDB: 2CRD) determined by mutagenesis studies, which are important for binding to voltage-dependent K^+ and large conductance Ca^{2+}-activated K^+ channels. Most of the functional residues (R25, K27, M29, R34 and Y36), except S10 and W14 are located on the flat surface of the β-sheets. S10 and W14 reside at the α-helix. Spatial distances between Cα atoms of the functional residues with that of critical K27 demonstrated that they are within 6.6 ± 1.0 Å. This may explain their involvement in interaction due to their close proximity to the functional dyad.

Toxin	Sequence	Channel	Ref.
Charybdotoxin	-QFTNVSCTTSKECWSVCQRLHNT-SRGKCMNKKCRCYS--	BK_{Ca},K_V	1,2
Iberiotoxin	-QFTDVDCSVSKECWSVCKDLFGV-DRGKCMGKKCRCYQ--	BK_{Ca}	3
Pi4	--IEAIRCGGSRDCYRPCQKRTGC-PNAKCINKTCKCYGCS	Kv	4
Pi2	----TISCTNPKQCYPHCKKETGY-PNAKCMNRKCKCFGR-	Kv	5
PbTx3	--EVDMRCKSSKECLVKCKQATGR-PNGKCMNRKCKCYPR-	Kv	6
HgTx1	-TVIDVKCTSPKQCLPPCKAQFGIRAGAKCMNGKCKCYPH-	Kv	7
AgTx1	GVPINVKCTGSPQCLKPCKDAGMR--FGKCINGKCHCTPK-	Kv	8
P05	-----TVCN-LRRCQLSCR-SLGL--LGKCIGVKCECVKH-	SK_{Ca}	9

Figure 26 Multiple sequence alignment of representative scorpion toxins which target voltage-dependent K^+ (Kv), large and small conductance Ca^{2+}-activated K^+ channels (BK_{Ca}, SK_{Ca}). The box highlights the position and residue identity involved in differentiating K_V and BK_{Ca}. References: 1 (Vazquez *et al*., 1989), 2 (Deutsch *et al*., 1991), 3 (Galvez *et al*., 1990), 4 (Olamendi-Portugal *et al*., 1998), 5 (Rogowski *et al*., 1996), 6 (Huys and Tytgat, 2003), 7 (Koschak *et al*., 1998), 8 (Garcia *et al*., 1994), 9 (Wu *et al*., 2002).

Scorpion toxins use different residues for targeting various K^+ channel subtypes. The functional dyad of lysine and an aromatic residue residing at the β-sheets interact with voltage-dependent K^+ channels and large conductance Ca^{2+}-activated channels. Toxins which target small conductance Ca^{2+}-activated channels and ERG channels have functional residues located at the α-helix.

4.3 Conclusion

Given the large diversity of ion channels (Zuo and Ji, 2004; Korn and Trapani, 2005), a number of different binding motifs in scorpion toxins can be defined. Dauplais *et al*. (1997) reported a conservation of a functional dyad motif in K^+ toxins with unrelated structures while none has been reported for Na^+, Cl^- and Ca^{2+} toxins. This is the first report of binding motifs for four K^+ ion channel subtypes (voltage-dependent K^+ channels, large- and small-conductance Ca^{2+}-activated channels, and ether-a-go-go channel), four binding site motifs for Na^+ channels and a conserved motif for Cl^- channels.

The motifs reported here included information from mutation studies of scorpion toxins and 3D structure analyses except for Cl^- channel for which mutation study is not available. The motifs reported in this work have biological and functional relevance which complement motifs obtained statistically. Mutation studies help determine if conserved residues within groups of scorpion toxins are important for integrity of molecular structure and for function. The mutant data provide information on residues and positions which are important for structure and function, and those that are not. Further, critical information on physicochemical properties of key residues can be obtained from mutant data where there may lack conservation of residues at key positions in multiple sequence alignments. This is important because even semi-

conservative substitution of residues can affect activity in scorpion toxins where a substitution from arginine to lysine resulted in a 80-fold decreased activity in Lqh αIT (Karbat *et al*., 2004b). Scaling the effects of mutation on binding affinity to a common scale provides a semi-quantitative comparison to differentiate critical residues from residues that play no role in binding affinity.

One major goal of analysing 3D protein structure is to understand the relationship between the primary and tertiary structures. If this relationship were known, then the structure and function of a protein could be reliably predicted from its amino acid sequence. In fact, Mouhat *et al*. (2004a) proposed that the spatial distribution of key functional residues is more important than its fold when considering the interaction of a toxin with its ion channel target. Spatial proximity to the key functional residues is likely to influence interaction, especially in toxin-channel complexes which involves a combination of electrostatic, hydrophobic and hydrogen bonding interactions (Xu *et al*., 2003; Mouhat *et al*., 2004b). 3D structure analyses aid identification of motifs where non-contiguous residues are clustered spatially and help extraction of functional motifs where their spatial positions are mapped onto the primary sequence. These linearised motifs can then be used to compare novel sequences for prediction of function. The information of spatial proximity on residues near to critical functional residue facilitates the design of mutation studies.

This systematic approach of including mutant data of scorpion toxins and 3D structure analyses for extraction of motifs can serve as a model for other proteins where mutation studies and 3D structures are available. This approach complements motifs that are derived statistically by including biological information for a more accurate inference of function for newly identified scorpion toxin sequences.

Chapter summary

- Two main approaches are available for functional inference of newly identified scorpion toxins. First is the identification of similar characterised sequences by pairwise alignment tools such as BLAST or FASTA. However, pairwise alignment is unable to differentiate functional residues from those with no critical role. Second, function can be inferred from matches to patterns in signature databases such as PROSITE. The patterns are usually derived statistically but are not necessarily biologically relevant.
- However, biologically relevant information on critical residue and position is available in mutation studies but is usually not used for extraction of functional motifs in scorpion toxins.
- This is the first report on eight functionally relevant binding motifs to Na^+ and K^+ channels which were extracted from the approach of analysing multiple sequence alignments of native scorpion toxins, 3D structures and information from mutation studies. The motifs reported in this work have biological and functional relevance. Multiple sequence alignment of native toxins targeting the same ion channel subtype allows conserved residues to be determined. 3D structure analyses aid identification of motifs where non-contiguous residues are clustered spatially. Mutation studies provide information on residues, positions and the physicochemical properties of key residues important for structure and function.
- This systematic approach can serve as a model for other proteins where mutation studies and 3D structures are available.

Part II: Chapter 5 Functional prediction of bioactive toxins in scorpion venom

‘Once a new technology rolls over you, if you are not part of the steamroller, you're part of the road.’

Stewart Brand

Scorpion toxins are important experimental tools for studies of biochemical and pharmacological properties of ion channels which have significance in the development of novel therapeutics for ion channel diseases (Mouhat *et al.,* 2005; Devaux *et al*., 2004; Fuller *et al*., 2004). Identification of new toxin sequences from scorpion venoms is thus critical for scientific research and medical applications. The number of functionally characterised scorpion toxins is steadily growing, but the number of newly identified toxin sequences is increasing at much faster pace. These newly identified sequences usually have primary structure information with limited or no functional feature characterised. With an estimated 100,000 different variants, there is a pressing need to accurately predict functions of these sequences where bioinformatic analysis of scorpion toxins is becoming a necessary tool for their systematic functional analysis.

Here, the author reports a bioinformatics-driven approach involving scorpion toxin structural classification, functional annotation, sequence comparison, nearest neighbour analysis and decision rules, which produces highly accurate predictions of scorpion toxin functional properties. The methodology reported is of importance because it provides a bioinformatics basis for systematic functional analysis of large sets of toxin sequences. The prediction tool, termed *Annotate Scorpion*, is the first novel tool in the field of scorpion toxin research which predicts functional properties of uncharacterised scorpion toxins. It facilitates selection of critical experiments where comprehensive characterisation of even a single toxin involves valuable time and resources.

5.1 Prediction of functional properties of novel scorpion toxins by nearest neighbour analysis, sequence comparison and decision rules

The classified groups and subgroups of scorpion toxins, proposed in Chapter 3, were used as a basis for the development of a bioinformatics-driven approach for prediction of functional properties of scorpion toxins including ion channel specificity (Na^+, K^+, Ca^{2+} or Cl^-), toxin subfamily (α, α-like, β, K^+, Ca^{2+} or Cl^-), toxin potency (neurotoxic or nontoxic), and target cell specificity (insect-, crustacean-, or mammal-specific). This new hierarchical classification scheme, underpinned by sequence, structural and functional similarity of sequences helps understanding of their structural and functional properties. This method termed "*Annotate Scorpion*" combines sequence comparison, nearest neighbour analysis and decision rules. This module is data-driven and thus statistical in its nature. Testing of the *Annotate Scorpion* showed it could predict functional properties of scorpion toxins with high accuracy. Here the author describes the method and the algorithm, and presents the results of the assessment of prediction accuracy.

The scorpion toxin functional prediction module, *Annotate Scorpion*, automatically generates putative functional annotation for the query toxin and places query sequence into an appropriate structural group. It combines sequence comparison, nearest neighbour analysis and decision rules 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%), toxin action (83.3%) and toxin type (68.9%) were achieved on a test set of 52 newly characterised scorpion toxins (Tan *et al*., 2005). Because it uses nearest neighbour analysis, the *Annotate Scorpion* tool has high specificity. Similar bioinformatic-based

approach of predicting structure and specific function can be applied to other toxins as demonstrated by the *Analysis* module in the MOLLUSK database (http://research.i2r.a-star.edu.sg/MOLLUSK/).

5.2 Materials and Methods

Two separate testings were performed on the accuracy of the module where the two test sets consisted of 52 and 127 toxin sequences collected from public databases and literature. The two separate test sets, containing newly characterised scorpion toxins, were used for evaluating the prediction accuracy of *Annotate Scorpion*. The primary toxin set was analysed and classified into groups based on primary sequence similarity and ion channel specificity (as discussed in Chapter 3). These groups were used as comparison targets by the "*Annotate Scorpion*" method for prediction of anonymous protein sequences.

5.2.1 Scorpion toxin data

The initial primary data set of 220 sequences was extracted from the SCORPION database. Of these, 196 complete sequences representing mature toxins were compared and used for defining toxin groups (see Section 3.3.1), while 24 partially sequenced toxins were not analysed, but were later added to the groups. The first test set of 52 sequences was obtained from the secondary collection of scorpion toxin sequences where 18 were functionally characterised for their receptor and target cell specificities, toxin action and toxin type by experimentation, whereas 34 sequences were partially characterised.

Upon prediction of their functional properties, these 52 sequences were subsequently added to the initial primary data set of 220 sequences, totaling 272 sequences. The second primary data set consisted of the enlarged primary data set of 272 sequences and 426 scorpion mutant toxin data for prediction of functional features in the second test set. The second test set of 127 sequences was obtained from a separate collection where 18 were functionally characterised for their receptor and target cell specificities, and toxin action and toxin type by experimentation.

5.2.2 Algorithm – nearest neighbour and rule-based

The aim of *Annotate Scorpion* prediction module is to generate accurate functional annotations for a query scorpion toxin. The predicted functional properties are toxin type, toxin action, ion channel specificity, and cellular target specificity. This module combines sequence comparison, nearest neighbour analysis and decision rules for assigning a putative classification of a query sequence to a structure-function group. The logic involved performing a similarity search of a query sequence against the primary data set in the SCORPION2 database by BLAST program. A large change in score values separates sequences dissimilar to the query from similar sequences. This permits grouping of the query to similar sequences whereby nearest neighbour analysis will assign the functional and structural properties of the sequences in the group to the query. The nearest neighbour rule states that a test instance is classified based on the classification of nearest training instances (Cover and Hart, 1967; Dasarathy, 1991). The algorithm uses 10 rules for grouping and prediction of specific function of scorpion toxins.

1) The length of query sequence is maximum 200 amino acids.
2) If the bit score of the nearest neighbour is < 20, then NOT SCORPION TOXIN.

3) If the query sequence is 100% identical to an existing sequence in the primary data set, than IDENTICAL to known toxin.
4) If the query sequence is 100% identical to a portion of an existing sequence in the primary data set, then PARTIAL SEQUENCE.
5) If the identity to the nearest neighbour is < 50%, then NEW GROUP.
6) If the difference of bit scores between two consecutive neighbours is ≥ 30, this serves as a cut-off point. The number of nearest neighbours, N = 5.
7) If the N neighbours belong to same subgroup, then EXISTING SUBGROUP.
8) If the N neighbour sequences belong to the same group but are from different subgroups, then NEW SUBGROUP.
9) If the N neighbours belong to different groups, then NEW GROUP.
10) If the group which the nearest neighbour belongs to consists of M sequences and M < N, then only the top M neighbours are considered in rules 7, 8 and 9.

The threshold of < 20 bit score was selected for differentiating sequences that are not scorpion toxins because the range of sequence identities between 393 native scorpion toxins was 30 – 100%. A new group is proposed for a query if no nearest neighbour is found (<50% identity) or the nearest neighbours (>50% identity) belong to different groups. A new subgroup is proposed if the nearest neighbours are in the same group but belong to different subgroups. For example, submission of TbIT-1 (Pimenta *et al*., 2001) into *Annotate Scorpion* returned the top five nearest neighbours of >50% identity where Tst1, Tb1 and Ts1 (SCORPION2 ID: D000008, D000009, D000012) belong to Na^+ subgroup 2a while TsNTXP and Ts4 (D000162, D000010) belong to Na^+ subgroup 2c. A new subgroup was proposed for TbIT-1.

5.3 Results – Accurate prediction of functional properties of novel scorpion toxins

The functional properties predicted by the *Annotate Scorpion* module are the toxin subfamily, toxin action, ion channel specificity and target cell type. The accuracy of *Annotate Scorpion* module upon submission of the two test sets has been summarised in **Figure 27** where this work's predictions were compared with the features of functionally characterised scorpion toxin sequences. For the first test set, 91%, 69%, 83% and 44% were correctly predicted for ion channel specificity, toxin subfamily, toxin action and target cell type, respectively. In the second test set, 99%, 83%, 50% and 40% correct predictions were obtained for ion channel specificity, toxin subfamily, toxin action and target cell type, respectively.

In the first test set, only two sequences could not be annotated (**Table 7**). No similar sequences were found for κ-hefutoxins 1 and 2, thus *Annotate Scorpion* could assign neither the group, nor functional features to them. These two sequences constitute a new scorpion toxin fold as determined by NMR which consists of two parallel helices linked by two disulfide bridges without any β-sheets (Srinivasan *et al*., 2002b). 12 sequences were predicted as members of new groups, of which nine (75%) were correctly predicted as members of new groups, and three (25%) belong to existing groups (*Be*Km-1, BKTx, and *Bm*Tx3 belong to K^+ group 1, Na^+ group 7, and K^+ group 4). New subgroups were predicted for seven sequences. The remaining 31 toxin sequences were classified into the well-defined groups.

Of 52 sequences, 47 had known ion channel specificity, 18 for cellular target specificity and toxin action, and 45 for toxin type. Ion channel specificity was correctly predicted for 43 (91.5%) toxins. Tamulustoxin 1 and 2 were wrongly classified as Na^+ toxins instead of K^+ toxins (4.25% misclassification).

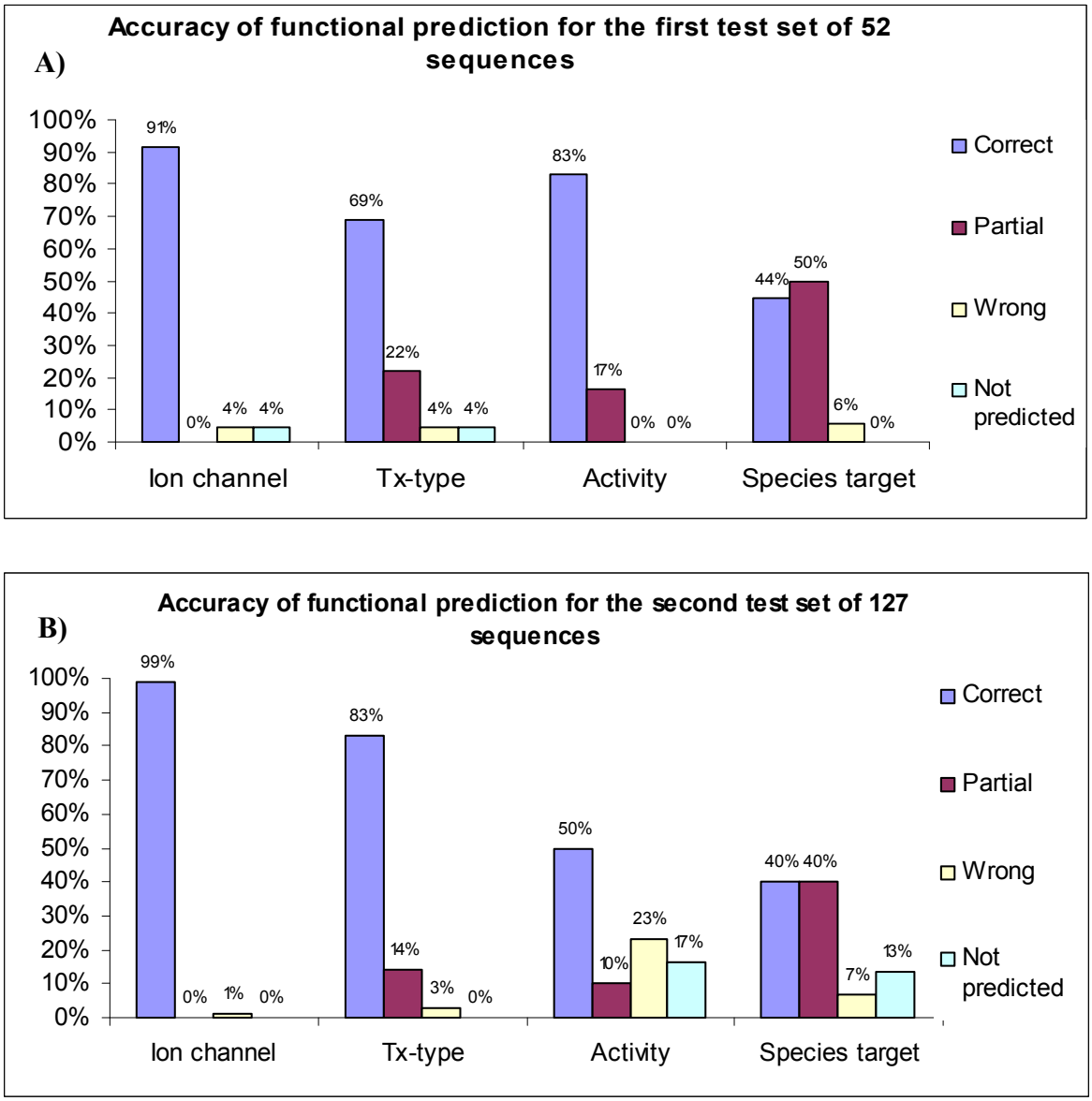

Figure 27 Accuracy of functional prediction of *Annotate Scorpion* module determined using two separate test sets. The functional properties predicted include ion channel specificity, toxin type, toxin action and cellular species target. This work's predictions of these properties were compared with features of functionally characterised scorpion toxin sequences. 'Correct' represents the prediction agrees with experimentally characterised feature. 'Partial' represents the prediction suggests at least two properties of which one agrees with experimentally characterised feature. 'Wrong' represents the prediction disagrees with characterised feature and 'Not predicted' represents no features predicted. **A)** Accuracy of prediction for first test set of 52 sequences. **B)** Accuracy of prediction for second test set of 127 sequences.

Table 7 Functional properties predicted for the first test set of 52 new toxin sequences. Abbreviations correspond to the scorpion species: *Aah*, *Androctonus australis* Hector; *Be*, *Buthus eupeus*; *Bm*, *Buthus martensii*; *Bom*, *Buthus occitanus mardochei*; *Bs*, *Buthus sindicus*; *Bt*, *Buthus tamulus*; *Cg*, *Centruroides gracilis*; *Cll*, *Centruroides limpidus limpidus*; *Hf*, *Heterometrus fulvipes*; *Lqh*, *Leiurus quinquestriatus hebreaus*; *Tb*, *Tityus bahiensis*; *Tst*, *Tityus stigmurus*; *Os*, *Orthochirus scrobiculosus*. The *Annotate Scorpion* module assigned each query sequence to a group, and predicted its ion channel type, toxin type, toxin action and species specificity. The Group indicates group or subgroup that contains toxin sequences similar to the query. A new subgroup is denoted by the group number and an asterisk. If no nearest neighbour is found, the query is assigned as a 'New' group. Toxin type describes the subfamily of the query sequence, where Na^+ toxins are classified into alpha (α), alpha-like (α') and beta (β) subfamilies and K^+, Ca^{2+} and Cl^- toxins have been classified into single subfamilies each. Toxin action describes the nature of the toxin: T, neurotoxic; N, nontoxic. Abbreviations correspond to species specificity: M, mammal; I, insect; C, crustacean; Pu: putative annotation. Ex: experimentally determined function. Ref.: functional annotation in Swiss-Prot. Dashes (-) represent no experimental characterisation or no functional annotation in references. Swiss-Prot (^{SP}) accession numbers refer to direct submission of toxin sequences. Question marks (?) represent sequences not functionally annotated by *Annotate Scorpion*.

Toxin name	Group	Ion Channel		Toxin Type		Toxin Action		Species Specificity		Ref.
		Pu	Ex/Ref.	Pu	Ex/Ref.	Pu	Ex	Pu	Ex	
Aah(Toxin 1)	New	Na^+	-	α, α'	-	T	-	I, M	-	1
Aah(Toxin 2)	New	Na^+	-	α, α'	-	T	-	I, M	-	1
Aah(Toxin 3)	New	Na^+	-	α, α'	-	T	-	I, M	-	1
Aah(Toxin 4)	3	Na^+	-	α, α'	-	T	-	I, M, C	-	1
Aah(Toxin 5)	6	Na^+	-	α	-	T	-	I, M	-	1
*Be*Km-1	New	K^+	K^+	K	K	T	T	M	M	2
Bm(ANEPII)	11	Na^+	Na^+	β	β	T	-	I, M	-	^{SP}Q9BKJ1
Bm(BKTx)	New	Na^+	Na^+	α	α	T	T	I, M	M	3
*Bm*32-VI	1a	Na^+	Na^+	β	β	T	T	I	I	4
*Bm*33-I	1a	Na^+	Na^+	β	β	T	T	I	I	4
*Bm*KdITAP3	11	Na^+	Na^+	β	β	T	T, N	I, M	I,M	5
*Bm*KK1	New	K^+	K^+	K	K	T	-	M	-	6
*Bm*P01	8	K^+	K^+	K	K	T, N	-	M	-	^{SP}Q9U522
*Bm*KK3	New	K^+	K^+	K	K	T	-	M	-	6
*Bm*Tx3	New	K^+	K^+	K	K	T	T	M	M	7
*Bm*TXKS1	New	K^+	K^+	K	K	T, N	-	?	-	8
Bom α6a	3	Na^+	Na^+	α, α'	α	T	-	I, M, C	-	9
Bom α6b	3	Na^+	Na^+	α, α'	α	T	-	I, M, C	-	9
Bom α6c	3	Na^+	Na^+	α, α'	α	T	-	I, M, C	-	9
Bom α6d	3	Na^+	Na^+	α, α'	α	T	-	I, M, C	-	9
Bom α6e	3	Na^+	Na^+	α, α'	α	T	-	I, M, C	-	9
*Bs*IT1	11	Na^+	Na^+	β	β	T	T	I, M	I	10
*Bs*IT2	11	Na^+	Na^+	β	β	T	T	I, M	I	10
*Bs*IT3	11	Na^+	Na^+	β	β	T	T	I, M	I	10
*Bs*IT4	11	Na^+	Na^+	β	β	T	T	I, M	I	10
Bt(Tamapin)	5	K^+	K^+	K	K	T	T	M	M	11
Bt(Tamapin 2)	5	K^+	K^+	K	K	T	T	M	M	11
Bt(Tamulustoxin 1)	New	Na^+	K^+	β	K	T	T	I, M	M	12
Bt(Tamulustoxin 2)	New	Na^+	K^+	β	K	T	T	I, M	M	12
*Cg*2	9c	Na^+	Na^+	β	-	T	-	I, C	-	13
*Cll*9	9*	Na^+	Na^+	β	-	T	-	I, C	-	13
*Hf*Tx 1	?	?	K^+	?	K	T	-	?	-	14
*Hf*Tx 2	?	?	K^+	?	K	T	-	?	-	14
Lqh α6a	3	Na^+	Na^+	α, α'	α'	T	-	I, M, C	-	9
Lqh α6b	3	Na^+	Na^+	α, α'	α'	T	-	I, M, C	-	9
Lqh α6c	3	Na^+	Na^+	α, α'	α'	T	-	I, M, C	-	9
*Lqh*ChTx-b	1	K^+	K^+	K	K	T	-	M	-	9
*Lqh*ChTx-c	1	K^+	K^+	K	K	T	-	M	-	9
*Lqh*ClTx-a	1*	Cl^-	Cl^-	Cl	Cl	T	-	I	-	9
*Lqh*ClTx-b	1a	Cl^-	Cl^-	Cl	Cl	T	-	I, M	-	9
*Lqh*ClTx-c	1a	Cl^-	Cl^-	Cl	Cl	T	-	I, M	-	9
*Lqh*ClTx-d	1a	Cl^-	Cl^-	Cl	Cl	T	-	I, M	-	9
*Lqh*IT1-a	1*	Na^+	Na^+	β	β	T	-	I	-	9
*Lqh*IT1-b	1*	Na^+	Na^+	β	β	T	-	I	-	9
*Lqh*IT1-c	1*	Na^+	Na^+	β	β	T	-	I	-	9
*Lqh*IT1-d	1*	Na^+	Na^+	β	β	T	-	I	-	9
*Lqh*IT2-13	11	Na^+	Na^+	β	β	T	-	I, M	-	9
*Lqh*IT2-53	11	Na^+	Na^+	β	β	T	-	I, M	-	9
*Os*K2	New	K^+	K^+	K	K	T	T	M	I	15
*Tb*2 II	2b	Na^+	Na^+	α, β	β	T, N	T	M	I, M	16
*Tb*IT-I	2*	Na^+	Na^+	α, β	β	T, N	T	I, M	I	16
Tst 2	2b	Na^+	Na^+	β	β	T	T	M	M	17

References: 1 (Ceard *et al*., 2001), 2 (Korolkova *et al*., 2001), 3 (Srinivasan *et al*., 2001), 4 (Escoubas *et al*., 2000), 5 (Guan *et al*., 2001), 6 (Zeng *et al*., 2001), 7 (Vacher *et al*., 2001), 8 (Zhu *et al*., 2001), 9 (Froy *et al*., 1999), 10 (Ali *et al*., 2001), 11 (Pedarzani *et al*., 2002), 12 (Strong *et al*., 2001), 13 (Possani *et al*., 2000), 14 (Srinivasan *et al*., 2002b), 15 (Dudina *et al*., 2001), 16 (Pimenta *et al*., 2001), 17 (Becerril *et al*., 1996), ^{SP} direct submission to Swiss-Prot.

Na^+ toxins belong to α, α-like and β subfamilies whereas K^+, Cl^- and Ca^{2+} specific toxins belong to respective single families. Cross-referencing with original literature or annotations in the databases showed that of 45 test sequences, 31 (68.9%) had correct, 10 (22.2%) partially correct, and 4 (8.9%) wrong predictions of toxin type. Partially correct predictions were those where *Annotate Scorpion* predicted two possible types, of which one was correct. For Na^+ toxins, 15 of 17 toxin sequences were correctly predicted as β subfamily while two were annotated as belonging to either α or β subfamily. Of six α toxins, BKTx was correctly predicted and another five toxins as either α or α-like toxins. Three α-like toxins, Lqh α6a, 6b and 6c were annotated as either α or α-like toxins.

Toxin action describes the potency of the toxin where 'neurotoxic' elicits toxic effect upon injection into target organisms while 'nontoxic' does not cause the effect. Comparison to experimental results showed that of 18 test sequences, 15 (83.3%) had correct predictions, and three (16.7%) had partially correct predictions of toxicity. Tb2 II and TbIT-I were experimentally determined to be neurotoxic but were predicted as being either neurotoxic or nontoxic. BmKdITAP3 was observed to be weakly toxic in insect and nontoxic in mammal but was predicted to be toxic.

Of 18 experimentally characterised sequences, eight (44.4%) predictions of target cell specificity were correct. Another eight peptides were predicted to interact with both insect and mammalian cells, but experimental toxicity was reported only for mammalian cells (three peptides) or insect cells (five peptides). One peptide was predicted to interact with mammalian cells, while experimental results showed specificity for both mammalian and insect cells. Only a single prediction was incorrect (*Os*K2). The species specificity of *Bm*TXKS1 was not assigned as there were no nearest neighbours and *Annotate Scorpion* predicted that it belongs to a new group.

In the second test set of 127 sequences, only one sequence could not be annotated (**Table 8**). No similar sequences were found for Toxin 6 from *Buthus martensii* Karsch (Swiss-Prot: Q95P85), thus *Annotate Scorpion* could assign neither the group, nor functional features to it. This sequence constitutes a new scorpion toxin group. 11 sequences, inclusive of Toxin 6, had primary sequence information only and thus predictions were not compared against them. 20 sequences were predicted as members of new groups, of which 14 (70%) were correctly predicted as members of new groups, and six (30%) belong to existing groups (KTX1, AamH1, AamH2, AamH3, Lqh4 and Lqhβ1 belong to K^+ group 1, Na^+ groups 4, 6, 5, 6 and 10, respectively). New subgroups were predicted for five sequences. The remaining 91 toxin sequences were classified into the well-defined groups and subgroups.

Of 127 sequences, 111 had known ion channel specificity, 30 for cellular target specificity and toxin action, and 71 for toxin type. Ion channel specificity was correctly predicted for 110 (99.1%) toxins. OmTx3 was wrongly classified as Na^+ toxins instead of K^+ toxins (0.9% misclassification).

Cross-referencing with original publications or annotations in the databases showed that of 71 test sequences, 59 (83.1%) had correct, 10 (14.1%) partially correct, and 2 (2.8%) wrong predictions of toxin type. For Na^+ toxins, five of six toxin sequences were correctly predicted as β subfamily while one was annotated as belonging to either α or β subfamily. Of 12 α toxins, AamH1 and AmmVIII were correctly predicted, another seven toxins as either α or α-like toxins, another two toxins as either α or β subfamily and Tc48b was wrongly annotated as β subfamily. BmKM7, an α-like toxin, was annotated as either α or α-like toxins. OmTx3 belonging to K^+ subfamily was wrongly annotated as α or β subfamily.

Comparison of toxin action with experimental results showed that of 30 test sequences, 15 (50.0%) had correct predictions, three (10.0%) had partially correct predictions of toxicity, seven wrong predictions (23.3%) and five (16.7%) could not be predicted. KTX3 was experimentally determined to be neurotoxic but was predicted as being either neurotoxic or nontoxic. Tspep1 and Kbot1 were nontoxic but were predicted to be neurotoxic or nontoxic. Tf4, Tspep2, Tspep3, Cn12, BjαIT, Bestoxin and Altitoxin were nontoxic but were wrongly predicted as neurotoxic. The toxin action for Cn11, BotIT6, BmKIM2, BmKITa1 and Lqhβ1 could not be predicted.

Of 30 experimentally characterised sequences, 12 (40.0%) predictions of target cell specificity were correct. Another six peptides were predicted to interact with both insect and mammalian cells, but experimental toxicity was reported only for mammalian cells (three peptides), insect cells (two peptides) or crustacean cells (one peptide). One peptide was predicted to target both insect and crustacean cells, but was reported for insect cells only. Three peptides target both insects and mammalian cells, but were predicted to interact with mammalian cells (one peptide) or insects cells (two peptides). Two peptides had a cellular target predicted correctly where experimental toxicity was reported for two cellular targets. Two predictions were incorrect (IsTx and BmK37). The cellular targets of Ikitoxin, Dortoxin, Bestoxin and Altitoxin were not assigned.

The author also tested *Annotate Scorpion* with sequences other than scorpion toxin and in all cases the results were ‘no similar record found’. This has shown that *Annotate Scorpion* is robust and highly accurate in assigning putative annotation to previously unseen scorpion toxins. The toxin sequences from the test sets have been incorporated into the prediction module as templates for functional annotation of new toxin sequences.

Table 8 Functional properties predicted for the second test set of 127 new toxin sequences. Abbreviations correspond to the scorpion species: Ap, *Anuroctonus phaiodactylus*; Aah, *Androctonus australis* Hector; Aam, *Androctonus amoreuxi*; Amm, *Androctonus mauretanicus mauretanicus*; Bt, *Buthus tamulus*; BmK, *Buthus martensii* Karsch; Bot, *Buthus occitanus tunetanus*; Ce, *Centruroides elegans*; Cg, *Centruroides gracilis*; Cn, *Centruroides noxius*; Cex, *Centruroides exilicauda*; Cll, *Centruroides limpidus limpidus*; CsE, *Centruroides sculpturatus* Ewing; Hs, *Heterometrus spinifer*; Iv, *Isometrus vittatus*; Lqh, *Leiurus quinquestriatus hebreaus*; Lqq, *Leiurus quinquestriatus quinquestriatus*; Oc, *Opistophthalmus carinatus*; Om, *Opisthacanthus madagascariensis*; Pg, *Parabuthus granulatus*; Pt, *Parabuthus transvaalicus*; Tc, *Tityus cambridgei*; Td, *Tityus discrepans*; Tf, *Tityus fasciolatus*; Ts, *Tityus serrulatus*; Tt, *Tityus trivittatus*; Tz, *Tityus zulianus*. The *Annotate Scorpion* module assigned each query sequence to a group, and predicted its ion channel type, toxin type, toxin action and species specificity. The Group indicates group or subgroup that contains toxin sequences similar to the query. A new subgroup is denoted by the group number and an asterisk. If no nearest neighbour is found, the query is assigned as a 'New' group. Toxin type describes the subfamily of the query sequence, where Na^+ toxins are classified into alpha (α), alpha-like (α') and beta (β) subfamilies and K^+, Ca^{2+} and Cl^- toxins have been classified into single subfamilies each. Toxin action describes the nature of the toxin: T, neurotoxic; N, nontoxic. Abbreviations correspond to species specificity: M, mammal; I, insect; C, crustacean; Pu: putative annotation. Ex: experimentally determined function. Ref.: functional annotation in Swiss-Prot. Dashes (-) represent no experimental characterisation or no functional annotation in references. Swiss-Prot (^{SP}) accession numbers refer to direct submission of toxin sequences. Question marks (?) represent sequences not functionally annotated by *Annotate Scorpion*.

Toxin name	Group	Ion Channel		Toxin Type		Toxin Action		Species Specificity		Ref.
		Pu	Ex/ Ref.	Pu	Ex/ Ref.	Pu	Ex	Pu	Ex	
AaHTX2	4	K	K	K	K	T	-	M	-	1
AamH1	New	Na	Na	α	α	T	-	M	-	2
AamH2	New	Na	Na	α,α'	α	T	-	M	-	2
AamH3	New	Na	Na	α,α'	α	T	-	M	-	2
AamTx	4	K	K	K	K	T	-	M	-	3
AmmTX3	4	K	K	K	K	T	T	M	M	4
AmmVIII	6	Na	Na	α	α	T	T	IM	M	5
Ap(Anuroctoxin)	6	K	K	K	K	T	-	M	-	6
Ap(Phaiodotoxin)	New	Na	Na	α,α'	-	N	N	M	M	7
Ap(Phaiodotoxin 2)	New	Na	-	α,α'	-	N	-	M	-	7
Ap(Phaiodotoxin 3)	New	Na	-	α,α'	-	NT	-	M	-	7
BjαIT	3	Na	Na	α,α'	α	T	N	M	M	8
BmK(ANEPIII)	11	Na	-	β	-	?	-	I	-	SPQ9BKJ0
BmK(KTX1)	New	K	K	K	K	T	-	IM	-	SPQ8MQL0
BmK(Toxin 6)	?	?	-	?	-	?	-	?	-	SPQ95P85
BmK(X-29S)	New	K, Ca	-	K, Ca	-	T	-	M	-	SPQ7Z0F1
BmK12b	1a	Cl	-	Cl	-	T	-	I	-	SPQ9BJW4
BmK37	10	K	K	K	K	T	T	M	I	9
BmK38	New	K	K	K	-	?	-	M	-	SPQ8MUB1
BmKAEP2	11	Na	-	β	-	?	-	IM	-	SPQ86M31
BmKIM2	11	Na	Na	β	β	?	T	I	IM	10
BmKITa	11	Na	-	β	-	?	-	IM	-	SPQ9XY87
BmKITa1	11	Na	Na	β	-	?	T	IM	I	11
BmKK4	4	K	-	K	-	T	-	IM	-	12
BmKK7	1	K	K	K	K	T	-	M	-	SPP59938
BmKM7	3	Na	Na	α,α'	α'	T	-	M	-	13
BmKSKTx1	19	K	K	K	K	T	-	?	-	14
BmKTXPL2	New	Na	-	α	-	?	-	I	-	15
BmKX	New	K, Ca, Na	-	K,Ca, α	-	T	-	M	-	16
BmKα3	6	Na	Na	α	-	T	-	M	-	SPQ9GUA7
Bot(Kbot1)	9	K	K	K	K	NT	N	M	M	17
Bot(KTX3)	3	K	K	K	K	NT	T	M	M	18
BotIT6	11	Na	Na	β	β	?	T	I	I	19
Bt(Neurotoxin)	3	Na	-	α,α'	-	T	-	M	-	SPP60277
BtITx3	1a	Cl	Cl	Cl	Cl	T	T	IC	I	20
BtK-2	9	K	K	K	K	NT	-	M	-	21
Buthus sp(Toxin)	New	K	-	K	-	?	-	M	-	SPP83108
CeErg1	16b	K	K	K	K	T	-	M	-	22
CeErg2	16b	K	K	K	K	T	-	M	-	22
CeErg3	16b	K	K	K	K	T	-	M	-	22
CexErg1	16b	K	K	K	K	T	-	M	-	22
CexErg2	16b	K	K	K	K	T	-	M	-	22
CexErg3	16b	K	K	K	K	T	-	M	-	22
CexErg4	16b	K	K	K	K	T	-	M	-	22
CgErg1	16b	K	K	K	K	T	-	M	-	22
CgErg2	16b	K	K	K	K	T	-	M	-	22
CgErg3	16b	K	K	K	K	T	-	M	-	22
Cll2b	9b	Na	Na	β	-	T	-	M	-	SPP59899
Cll3	9b	Na	Na	β	-	T	-	M	-	SPQ7Z1K9
Cll4	9b	Na	Na	β	-	T	-	M	-	SPQ7Z1K8
Cll5b	9c	Na	Na	β	-	T	-	IC	-	SPQ7Z1K7
Cll5c	9c	Na	Na	β	-	T	-	IC	-	SPQ7YT61
Cll5c*	9c	Na	Na	β	-	T	-	IC	-	SPQ7Z1K6
Cll6	9c	Na	Na	β	-	T	-	IC	-	SPQ7Z1K5
Cll7	9*	Na	Na	β	-	T	-	IMC	-	SPP59865
Cll8	9c	Na	Na	β	-	T	-	IC	-	SPQ7Z1K4
CllErg1	16b	K	K	K	K	T	-	M	-	22
CllErg2	16b	K	K	K	K	T	-	M	-	22
CllErg3	16b	K	K	K	K	T	-	M	-	22
CllErg4	16b	K	K	K	K	T	-	M	-	22
Cn11	11	Na	Na	β	β	?	T	I	IC	23
Cn12	New	Na	Na	β	-	T	N	IMC	IMC	24
CnErg2	16b	K	K	K	K	T	-	M	-	22
CnErg3	16b	K	K	K	K	T	-	M	-	22
CnErg4	16b	K	K	K	K	T	-	M	-	22
CnErg5	16b	K	K	K	K	T	-	M	-	22
CsE1x	9a	Na	Na	β	-	T	-	MC	-	25
CsE3	9b	Na	Na	β	-	T	-	M	-	25
CsE8	9c	Na	Na	β	-	T	-	IC	-	25
CsE9b	9d	Na	Na	?	-	?	-	?	-	25

CsEErg1	16b	K	K	K	K	T	-	M	-	22
CsEErg3	16b	K	K	K	K	T	-	M	-	22
CsEErg5	16b	K	K	K	K	T	-	M	-	22
CsEIa	9a	Na	Na	β	-	T	-	MC	-	25
CsEKerg1	16b	K	K	K	K	T	-	M	-	26
CsErg2	16b	K	K	K	K	T	-	M	-	22
CsErg4	16b	K	K	K	K	T	-	M	-	22
CsEv1b	9c	Na	Na	β	-	T	-	IC	-	25
CsEv1c	9c	Na	Na	β	-	T	-	IC	-	25
CsEv1d	9c	Na	Na	β	-	T	-	IC	-	25
CsEv1e	9c	Na	Na	β	-	T	-	IC	-	25
CsEv2a*	9c	Na	Na	β	-	T	-	I	-	25
CsEv2b	9c	Na	Na	β	-	T	-	IC	-	25
CsEv2c	9c	Na	Na	β	-	T	-	IC	-	25
CsEv2d	9c	Na	Na	β	-	T	-	IC	-	25
CsEv3b	9c	Na	Na	β	-	T	-	I	-	25
CsEv3b*	9c	Na	Na	β	-	T	-	IC	-	25
CsEv4	9c	Na	Na	β	-	T	-	IC	-	27
CsEv5	2*	Na	Na	α,β	α	T	T	M	IM	27
Hs(κ-KTX1.3)	18	K	K	K	K	?	-	?	-	28
Iv(IsomTx1)	1b	Na	Na	β	-	?	-	I	-	29
Iv(IsomTx2)	1b	Na	Na	β	-	?	-	I	-	29
Lqh4	New	Na	Na	α,α'	α	T	T	M	IM	30
Lqh6	5	Na	Na	α,α'	α	T	T	IM	IM	31
Lqh7	5	Na	Na	α,α'	α	T	T	IM	IM	31
Lqhβ1	New	Na	Na	α,β	β	?	T	IM	I	32
Lqq(Insect2 clone 6)	11	Na	Na	β	-	?	-	I	-	33
Lqq(Insect2 clone 8)	11	Na	Na	β	-	?	-	I	-	33
Oc(Opicalcine 1)	1	Ca	Ca	Ca	Ca	T	-	M	-	34
Oc(Opicalcine 2)	1	Ca	Ca	Ca	Ca	T	-	M	-	34
OcKTx1	6	K	K	K	K	T	-	M	-	35
OcKTx2	6	K	K	K	K	T	-	M	-	35
OcKTx3	6	K	K	K	K	T	-	M	-	35
OcKTx4	6	K	K	K	K	T	-	M	-	35
OcKTx5	6	K	K	K	K	T	-	M	-	35
Om(IsTx)	New	K	K	K	K	T	T	M	C	36
OmTx1	New	K	K	K	K	T	-	M	-	37
OmTx2	New	K	K	K	K	T	-	M	-	37
OmTx3	New	Na	K	α,β	K	?	-	IM	-	37
Pg(PBTx10)	11	K	K	K	K	?	-	?	-	38
Pt(Altitoxin)	15	Na	Na	β	-	T	N	?	M	39
Pt(Bestoxin)	15	Na	Na	β	-	T	N	?	M	39
Pt(Dortoxin)	15	Na	Na	β	-	T	T	?	M	39
Pt(Ikitoxin)	15	Na	Na	β	β	T	T	?	M	40
Tc30	4	K	K	K	K	T	-	IM	-	41
Tc32	New	K	K	K	K	T	-	M	-	41
Tc48a	2*	Na	Na	β	-	T	-	IM	-	42
Tc48b	2a	Na	Na	β	α	T	-	IM	-	43
Tc49b	2a	Na	Na	β	-	T	T	IM	M	44
Td(Ardiscretin)	2*	Na	Na	β	-	T	T	IM	IC	45
Td(Discrepin)	New	K	K	K	K	T	-	M	-	46
Tf4	2*	Na	Na	α,β	α	T	N	M	MC	47
TsPep1	New	K	-	K	-	NT	N	M	M	48
TsPep2	New	K	-	K	-	T	N	M	M	48
TsPep3	New	K	-	K	-	T	N	M	M	48
TtBut-toxin	12	K	K	K	K	T	-	M	-	49
Tz1	2a	Na	Na	β	β	T	T	IM	M	50

References: 1 (Legros *et al.*, 2003), 2 (Chen *et al.*, 2003), 3 (Chen *et al.*, 2005), 4 (Vacher *et al.*, 2002), 5 (Alami *et al.*, 2003), 6 (Bagdany *et al.*, 2005), 7 (Valdez-Cruz *et al.*, 2004a), 8 (Arnon *et al.*, 2005), 9 (Xu *et al.*, 2004a), 10 (Peng *et al.*, 2002), 11 (Liu *et al.*, 2003), 12 (Zhang *et al.*, 2004), 13 (Guan *et al.*, 2004), 14 (Xu *et al.*, 2004b), 15 (Zhu and Li, 2002), 16 (Wang *et al.*, 2005), 17 (Mahjoubi-Boubaker *et al.*, 2004), 18 (Meki *et al.*, 2000), 19 (Mejri *et al.*, 2003), 20 (Dhawan *et al.*, 2002), 21 (Dhawan *et al.*, 2003), 22 (Corona *et al.*, 2002), 23 (Ramirez-Dominguez *et al.*, 2002), 24 (del Rio-Portilla *et al.*, 2004), 25 (Corona *et al.*, 2001), 26 (Nastainczyk *et al.*, 2002), 27 (David *et al.*, 1991), 28 (Nirthanan *et al.*, 2005), 29 (Coronas *et al.*, 2003b), 30 (Corzo *et al.*, 2001), 31 (Hamon *et al.*, 2002), 32 (Gordon *et al.*, 2003), 33 (Zaki and Maruniak, 2003), 34 (Zhu *et al.*, 2003), 35 (Zhu *et al.*, 2004c), 36 (Yamaji *et al.*, 2004), 37 (Chagot *et al.*, 2005), 38 (Huys *et al.*, 2004), 39 (Inceoglu *et al.*, 2005), 40 (Inceoglu *et al.*, 2002), 41 (Batista *et al.*, 2002a), 42 (Batista *et al.*, 2004), 43 (Murgia *et al.*, 2004), 44 (Batista *et al.*, 2002b), 45 (D'Suze *et al.*, 2004b), 46 (D'Suze *et al.*, 2004a), 47 (Wagner *et al.*, 2003), 48 (Pimenta *et al.*, 2003), 49 (Coronas *et al.*, 2003a), 50 (Borges *et al.*, 2004), [SP] direct submission to Swiss-Prot.

5.4 Discussion and conclusions

There is a need for an accurate functional prediction tool for scorpion toxins as more than half of the data in the two test sets collected from public databases and the literature are molecular clones (52% and 69%, respectively), having only sequence information or partial functional information. The author has developed the first generic bioinformatic functional prediction tool for accurate functional annotation of scorpion toxins that can help reduce the number of experiments conducted for characterising novel scorpion toxin sequences. The bioinformatics-based approach of collecting, cleaning, annotating and classifying scorpion sequences into groups and subgroups allowed prediction of the functions of query scorpion sequences with high accuracy. The initial process of cleaning the data is critical for preventing the propagation of errors. For example, at the time of this study, Swiss-Prot database had annotated excitatory and depressant toxins as belonging to α-toxin subfamily (P01497, P15147, P19856, P55904, P80962, P19855, P24336, P81240, P15228, P55903, O61668 and Q9U7E5). These two groups of toxins, however, belong to β-toxin subfamily (Gordon *et al*., 1998; Oren *et al*., 1998). If these annotations were not corrected, the predictions would be less accurate.

The author had classified scorpion toxins using BLAST and Clustal W results, which are in agreement with the phylogenetic analysis (discussed in Chapter 3). Detailed classification of scorpion toxin sequences into well-organised groups allows better correlation of structure-function relationships and thereafter, classification of new sequences by the prediction tool. Sequences that are similar (in primary, secondary and tertiary structures) often perform similar function. Most scorpion toxins share the CSH fold (Bontems *et al*., 1991). By clustering a query sequence with its nearest neighbours in the well-defined groups, functional properties of the nearest

neighbours could be ascribed to the query sequence. The algorithm is robust even if the query sequence could not be classified into the defined groups as it ascribed the functional properties of the five nearest neighbours to the query. Novel scorpion toxin sequences that show relatively low primary sequence similarity are assigned into new groups. If nearest neighbours do not exist, *Annotate Scorpion* will not make predictions. This restriction will result in scorpion toxins that have new fold (e.g. κ-Hefutoxins) may not be annotated. However, once a representative toxin is entered in the database, it becomes a template for further predictions. The accuracy of the prediction module is limited by the availability of present data. Nevertheless, as more new venom sequences are characterised and included into the data set, the accuracy of the structure-function prediction tool is expected to improve.

Among the four functional properties predicted, higher accuracies were obtained for ion channel specificity and toxin type as compared to toxin action and cellular target specificity. The difference in accuracies could be that the biological aspect of toxic action and cellular species target is functionally more complex. The ion channel specificity and toxin subtype could be predicted based on sequence similarity whereas toxin action and cellular specificity are dependent on various factors besides sequence similarity. The route of toxin administration affects level of toxicity in animal models. For example, no apparent effect was observed in mice when scorpion toxin Cll9 was administered intraperitoneally but induced sleep upon intracerebroventricular route (Corona *et al*., 2003). Also, all α-toxins are toxic to mice to a similar extent when injected subcutaneously, but they differ prominently upon intracerebroventricular injection (Gordon and Gurevitz, 2003). Finally, the age, genetic background and gender of animal models determine toxicity of scorpion toxins (Padilla *et al*., 2003). For prediction of cellular specificity, scorpion toxins have been

tested on one or combination of animal models such as insects (cockroach, locust), mammals (mice, rat) and crustaceans (crayfish, prawn). The specificity to other organisms has not been fully determined. For example, binding studies correlated with toxicity towards mammals and insects have revealed that even toxins known as mammal-specific can be toxic to insects and vice versa (Gordon *et al*., 1996) The animal group specificity is only relative and definite cross-reactivity exists (Selisko *et al*., 1996). The lack of cellular specificity data prevents accurate prediction of this functional property. Many scorpion toxins are generally characterised on a limited number of pharmacological targets and animal models where many other true physiological targets remain to be discovered.

The detailed structural and functional classification of scorpion toxin sequences into groups and subgroups can be used for accurate prediction of ion channel specificity and toxin subtype, and to a lesser extent on cellular target and toxin action. The predictions can effectively reduce the number of critical experiments performed. This generic bioinformatic-driven approach serves as a model for functional prediction of novel toxins from other venomous animals as demonstrated in the MOLLUSK (http://research.i2r.a-star.edu.sg/MOLLUSK/) and snake venom neurotoxins databases (Siew *et al*., 2004).

Chapter summary

- Experimental characterisation of the large number of newly identified scorpion toxins is prohibitively expensive and time-consuming. Thus, there is a need to accurately predict functions of these sequences to expedite their characterisation by facilitating selection of critical experiments.

- This work reports on the first novel prediction tool, *Annotate Scorpion*, which predicts functional properties of uncharacterised scorpion toxins. The algorithm combines sequence comparison, nearest neighbour analysis and decision rules to predict ion channel specificity, toxin subfamily, toxin potency and cellular specificity.
- High accuracy of predictions, particularly ion channel specificity and toxin subfamily, was obtained by validation with two sets of experimentally characterised scorpion toxins. Lower accuracies were obtained for toxin potency and cellular specificity because of different biological and experimental factors involved in the functional properties other than sequence similarity.
- The methodology reported here demonstrates that bioinformatics can be applied to systematic functional analysis of large sets of scorpion toxin sequences. This generic approach serves as a model for prediction of novel toxins from other venomous animals.

Part III: Chapter 6 Implementation of scorpion toxin data warehouse

'[An] argument against the Human Genome Project was that it was trivial, it wasn't really science. It was referred to as a fishing expedition, or a mindless collecting of facts. What they did not realise is how these databases were going to transform how we think about biology and medicine.'

Leroy Hood

Large-scale analysis of scorpion toxin data provides a holistic view of all the currently available data which is important for better understanding of their structure-function relationships. However, scorpion toxin data are reported as long lists of sequences in literature reviews (e.g. Possani *et al*., 1999; Tytgat *et al*., 1999; Goudet *et al*., 2002) or are scattered across multiple public databases with very limited structural and functional annotation. Discrepancies between records representing the same toxin in various databases are also common. Mutation studies (such as site-directed mutagenesis) of scorpion toxins provide a large amount of data on their functional and structural features are available in the literature but are usually not used in large-scale analyses. Structural analyses of scorpion toxins are impeded by availability of 3D structures in PDB (Deshpande *et al*., 2005) (only 20% of 393 reported scorpion toxin sequences as of November 2005). The access to and analysis of the growing number of scorpion toxin data scattered across multiple data sources is becoming increasingly difficult. Thus, there is a need to create a centralised repository for improved data management to facilitate analyses in the field of scorpion toxins.

The author was involved in building the SCORPION database (Srinivasan *et al*., 2002a) which contained 277 native scorpion toxins. In SCORPION, mutant toxin data available in the literature had not been tapped in the bioinformatics study of scorpion toxins. Values of lethal assays in animal models and binding affinities towards various ion channels were not extracted from the literature.

SCORPION2 database is the first data warehouse of native scorpion toxins and artificial mutant toxins with integrated bioinformatics tools for systematic analysis of their structure-function relationships. It is publicly available at <http://sdmc.i2r.a-star.edu.sg/scorpion/> and supersedes the earlier SCORPION database. SCORPION2

serves as a one-stop repository of currently available scorpion toxin data where SCORPION2 records are cleaned of errors and highly enriched with structure-function information from literature. Query, extraction, prediction and structure visualisation tools have been integrated to facilitate manipulation of large number of scorpion toxin data for analysis. In this chapter, the general steps involved in the implementation of data warehouse of scorpion toxins, the SCORPION2 and advantages of data warehouse in the field of toxin research over general databases are discussed.

6.1 Data warehouse for information usage and knowledge discovery

The completion and ongoing sequencing of various genome projects including three venomous animals (honey bee, sea urchin and duck-billed platypus) marked the beginning of biological research into 'large-scale science' of venomous animals (Collins *et al*., 2003). The large volume of biological data has posed an exciting challenge for researchers on how to extract information from raw data and transform the information into knowledge. These voluminous data are scattered in diverse biological databases with limited annotation. In the period of one year, 171 new publicly accessible databases were created, totaling 719 biological databases in 2005 (Galperin, 2005). These databases play key roles in biological research.

Easy access to diverse biological databases and efficient analyses of these data are important for interpretation of experimental results, discovery of new knowledge and planning research (Schonbach *et al*., 2000). However, the heterogeneous formats of the geographically diverse biological databases impede access to and extraction of useful information. These databases have different data formats, database management systems and data manipulation languages, among others. Different databases also encode data at various levels of complexity ranging from low-level data to high-level

information. Examples of low-level data are those that report experimental values as compared to high-level information (e.g. interpretation of data by domain experts), which are derived from the low-level data. Depending on the data that they contain, these databases serve different functions.

General purpose sequence databases, such as GenBank (Benson *et al*., 2005), aim to expand and disseminate nucleotide sequence information. Another example is Swiss-Prot (Bairoch *et al*., 2004), a general database of protein sequences which focuses on building a high level of functional annotation of proteins, integrating with other databases and maintaining low-redundancy. In contrast to a general purpose database (e.g. Swiss-Prot), which contains all known protein sequences, a molecular biology data warehouse contains subject-orientated, integrated, non-volatile, expert-interpreted collection of biological data (e.g. a family of proteins that have similar structures and functions) (Schonbach *et al*., 2000). The data warehouse involves integrating information on a specific subject from different databases, systems, and locations into a central database for more accurate and in-depth analysis of biological data, improved research planning and knowledge discovery. Non-volatile data means that data is never removed even though more data are added i.e. data is stable in a data warehouse[1]. Expert curation and annotation is essential to ascertain the relevance and accuracy of the entries in a biological data warehouse since the types of errors and inconsistencies can be domain-specific (Schonbach *et al*., 2000). The advantages of data warehouses include providing an integrated environment for consistency in naming conventions, measurement of variables and encoding data attributes (Schonbach *et al*., 2000) and eliminate potential technical and semantic heterogeneity (Karasavvas *et al*., 2004).

[1] http://www.sdgcomputing.com/glossary.htm

These advantages motivated the author to build a data warehouse of scorpion toxins with the SCORPION2 database as an example of a bioinformatic platform for improved analysis of their complex structure-function relationships.

6.2 Implementation of the data warehouse of scorpion toxins, SCORPION2

Here the author describes the creation of SCORPION2 database which involve access to toxin data scattered across multiple databases, inspection for errors, analysis and classification of toxin sequences and their structures, and the design and use of predictive models for simulation of laboratory experiments (Tan *et al*., 2003). SCORPION2 serves as a valuable primary resource with integrated bioinformatic tools for analysis and prediction of structure-function relationships of scorpion toxins that may typically involve querying multiple resources.

6.3 Materials and methods

Public databases used for creation of the SCORPION2 database were: GenBank and Swiss-Prot for consolidation of scorpion toxin sequences, PDB for extraction of scorpion toxin PDB structures and PubMed (http://www.pubmed.gov/) for extraction of published scorpion toxin sequences not deposited in any databases, mutant scorpion toxin data, structure-function information and literature. The general steps of creating a biological data warehouse of scorpion toxins were implemented: data collection, data cleaning and data annotation. Homology models were also generated for scorpion toxins that lack 3D structures.

6.3.1 Data collection of native and mutant scorpion toxin sequences and their 3D structures

For data collection of scorpion native toxin sequences, two strategies were employed: keyword and sequence similarity searches. A keyword search identifies sequences by comparing words or strings of characters through the written descriptions or annotated section of a database record while a sequence similarity search compares the sequences themselves. Keyword search "scorpion toxin" was performed in Swiss-Prot, GenBank and PDB databases. The search results were screened for records containing scorpion toxins only which formed an initial dataset while irrelevant records (non-scorpion toxins such as snake, bacterial and plant toxins etc.) were removed. Each sequence in the initial dataset was subsequently submitted into the protein BLAST tool (blastp) with default parameters against the non-redundant (nr) database at NCBI (http://www.ncbi.nlm.nih.gov/BLAST/). The nr database removes redundant identical sequences to yield a collection of unique sequences which facilitates data cleaning. The aim of submitting scorpion toxin sequences into BLAST was to collect scorpion toxin sequences that were missed out by keyword searching because the latter approach is not exhaustive. For example, Opicalcines 1 and 2 from *Opistophthalmus carinatus* (Swiss-Prot ID: P60252 and P60253) were not picked up by the keyword search 'scorpion toxin' but by BLAST search using a query sequence of Imperatoxin from *Pandinus imperator*. The majority of biological records in general purpose databases have limited or no annotation, in particular the nucleotide data derived from sequencing of genomes where high throughput is the priority. Thus, sequence similarity search is needed to gather these records.

Sequences in the BLAST results were compared against the initial dataset where relevant omitted sequences were added to the initial dataset. The process was

repeated until the initial dataset was exhausted. Published novel sequences not deposited in any databases were obtained using a literature search in PubMed and subsequent extraction of sequences from publications. The purpose of performing keyword and sequence similarity searches, and extracting sequences from literature was to generate a comprehensive dataset, which served as a better representation of the complete sequence information.

Data of mutant scorpion toxins were extracted solely from literature by exhaustive keyword searches (e.g. 'scorpion toxin', 'chemical modification', 'mutation', 'mutant', 'alanine scanning', 'mutagenesis', 'analog' and 'chimera') in PubMed. In a literature, each mutant toxin sequence was represented as an individual record. For example, if there were 15 unique mutant sequences in a literature, 15 records were generated.

6.3.2 Generation of homology models of scorpion toxins

After the collection of scorpion toxin sequences and 3D structures, scorpion toxins that currently lack experimentally determined structures (e.g. x-ray crystallography or nuclear magnetic resonance) were inspected if homology models could be generated based on availability of template structures and pairwise alignment. The approach began with identification of the template sequence, which has an experimentally determined 3D structure in PDB and a high pairwise sequence identity to the query sequence also known as the target sequence. The cut-off sequence identity value between the template and target sequences was set at ≥30% where Rost (1999) demonstrated that 90% of the template-target pairs had similar structures. The alignment of the query sequence to the template structure was optimised and the homology model was generated using the automated homology modelling feature in

the SDPMOD server (Kong *et al*., 2004). The quality of each homology model was checked manually, facilitated by assessment by the PROCHECK program (Laskowski *et al*., 1993) and deposited into SCORPION2 database.

6.3.3 Data cleaning

Data errors were detected manually by comparison of records between public databases and referencing to original literature. Records were inspected primarily for errors in their primary sequences, disulfide linkages, functional annotation and toxin names. Inconsistencies in the reported primary sequences between original references were annotated in the toxin record field termed 'Conflict'. Discrepancies in structural or functional information were annotated in a field called 'Comment'. Duplicates or identical sequences belonging to the same scorpion species reported in different databases were detected by pairwise sequence comparison and combined as a single record with hyperlinks to each underlying database. Different names referring to the same toxin sequences were consolidated as synonyms.

6.3.4 Data annotation

Functional and structural information from original literature was annotated to each native toxin record. Examples of functional information included binding affinity towards specific ion channel subtypes, toxic symptoms observed upon administering to animal models or suggestion of possible interactive site. Structural information annotated included lengths of signal peptides and matured toxins, disulfide connectivity and post-translational modification (e.g. amidation). For records of mutant scorpion toxins, functional information extracted from literature included

changes in binding affinities towards specific ion channels and differences in lethal dosages in animal models compared with that of native toxin peptides. Structural information, such as disulfide connectivity and monitoring of changes in secondary structures of mutant sequences by circular dichroism spectroscopy, was also annotated in the mutant record. The annotated fields include citations for the articles in which the experimental data were reported.

After collection, cleaning and annotation of the dataset, database creation was published on the internet using Templar (http://research.i2r.a-star.edu.sg/Templar/). Templar is a sub-system of BioWare which consists of four data warehousing modules for the retrieval, annotation, publishing and data updating of specialist molecular databases (Koh *et al*., 2004a). The organisation of the compiled data was designed using data warehousing principles (Schonbach *et al*., 2000) where SCORPION2 records were maintained as flat-file format to facilitate easy data analysis, extraction of value-added interpretation of the data and database maintenance.

6.4 Results

The differences between SCORPION2 and SCORPION have been summarised in **Figure 28**. SCORPION2 contains more than 800 records of native and mutant scorpion toxins. Approximately 60% of records were extracted from literature. Of 624 3D structures available in SCORPION2, 82 were extracted from the PDB database (**Table 9**) and 542 homology models of native and mutant toxins were generated using the comparative modeling tool, SDPMOD (Kong *et al*., 2004). Ten scorpion species previously not present in SCORPION are: *Androctonus amoreuxi, Anuroctonus phaiodactylus, Babycurus centrurimorphus, Isometrus vittatus, Opisthacanthis madagascariensis, Opistophthalmus carinatus, Parabuthus granulatus, Tityus*

fasciolatus, Tityus trivittatus and *Tityus zulianus*. Some 50 scorpion species are represented in SCORPION2. More than 400 newly added records of mutant toxin records were extracted exclusively from the literature. The number of literature references has increased to ~500. The newly added functional information includes binding affinity (>500 records) and toxicity (~200 records).

The SCORPION2 interface provides hyperlinks to the corresponding entries of the relevant external databases. Errors were minimised to ensure high data quality by checking with original literature and cross-referencing across databases. 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. Examples of errors that were found in the scorpion toxin 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 (**Figure 29**).

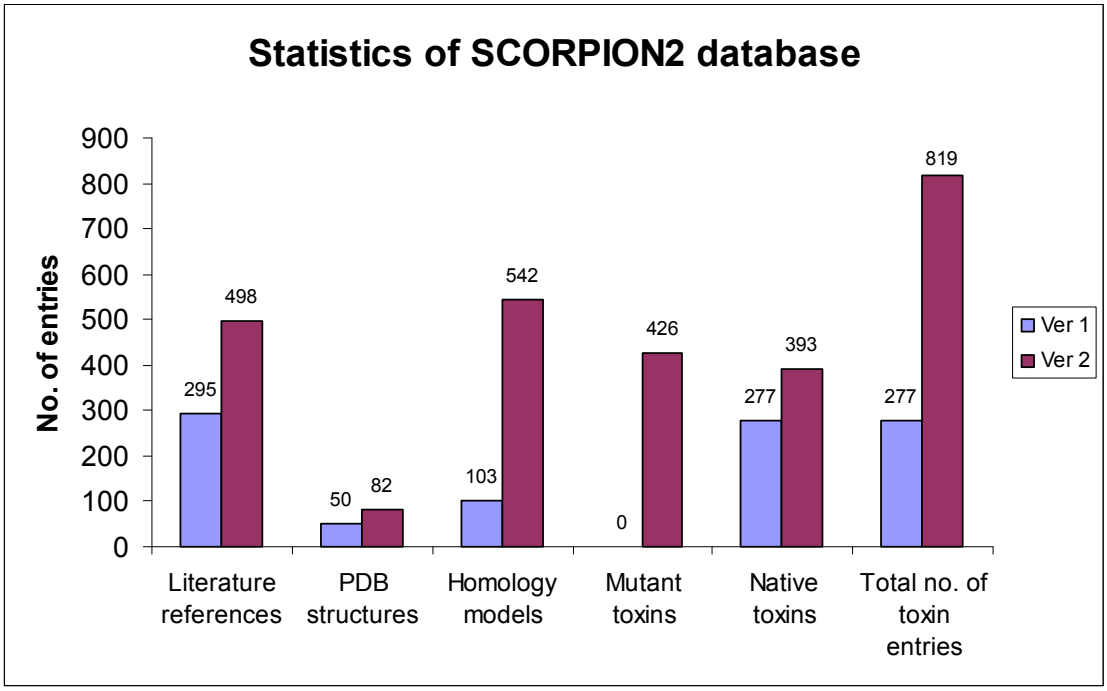

Figure 28 Statistics of SCORPION2 database on the number of records, 3D structures and literature available as of November 2005.

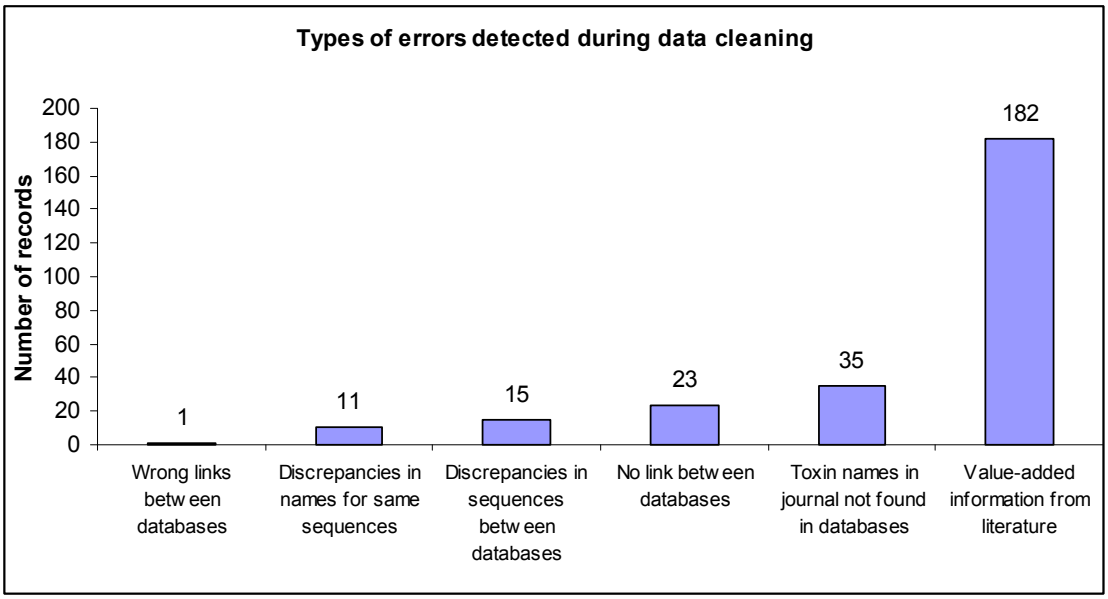

Figure 29 Number of records having errors or discrepancies summed by category of error during the initial collection of scorpion toxin records. The graph also shows the number of records that required additional annotations.

Table 9 A summary of 82 scorpion toxin PDB structures available in SCORPION2. The structures are listed in alphabetical order according to scorpion species and divided into four broad groups of ion channel specificity, namely K^+ (grey), Na^+ (yellow), Ca^{2+} (green) and Cl^- (cyan) ion channels as of November 2005. Multiple structures of the same toxin are available for some toxins.

Scorpion species	Toxin name	PDB identity
Androctonus mauretanicus mauretanicus	Kaliotoxin 1	2KTX, 1KTX
	P01	1ACW
	P05	1PNH
Buthus eupeus	BeKm-1	1J5J, 1LGL
Buthus martensi Karsch	BmKTX	1BKT
	BmP01	1WM7
	BmP02	1DU9
	BmP03	1WM8
	BmTx1	1BIG
	BmBKTx1	1Q2K, 1R1G
	BmKK2	1PVZ
	BmKK4	1S8K
	BmK X	1RJI
	BmTx2	2BMT
	BmTx3	1M2S
Centruroides limbatus	Hongotoxin 1	1HLY
Centruroides margaritatus	Margatoxin	1MTX
Centruroides noxius	Cobatoxin1	1PJV
	Noxiustoxin	1SXM
	Ergtoxin	1NE5, 1PX9
Heterometrus fulvipes	Hefutoxin 1	1HP9
Heterometrus spinnifer	HsTx1	1QUZ
Leiurus quinquestriatus hebraeus	Agitoxin 2	1AGT
	Charybdotoxin	2CRD
	Leiurotoxin I	1SCY
	LQH 18-2	1LIR
Opisthacanthus madagascariensis	IsTx	1WMT
	OmTx1	1WQC
	OmTx2	1WQD
	OmTx3	1WQE
Orthochirus scrobiculosus	Osk1	1SCO
Pandinus imperator	Pi2	2PTA

Pandinus imperator	Pi3	1C49
	Pi4	1N8M
	Pi7	1QKY
Scorpio maurus palmatus	Maurotoxin	1TXM
Tityus cambridgei	Tc1	1JLZ
Tityus serrulatus	Butantoxin	1C55, 1C56
	Ts κ	1TSK
	TsTx K α	1HP2

Androctonus australis Hector	AaH2	1AHO, 1PTX
Buthus martensii Karsch	Bm α TX12	1OMY
	BmK IT-AP	1T0Z
	BmK M1	1DJT, 1SN1
	BmKM2	1CHZ
	BmKM4	1SN4
	BmKM7	1KV0
	BmKM8	1SNB
Centruroides exilicauda	CsE V1	1VNA, 1VNB
	CsEI	1B3C, 2B3C
Centruroides noxius	Cn2	1CN2
	Cn12	1PE4
Centruroides sculpturatus	CsE V	1NRA, 1NRB
	CsE V2	1JZA, 1JZB
	CsE V3	2SN3
	CsE V5	1I6F, 1I6G, 1NH5
Hottentotta judaica	Bjxtr IT	1BCG
Leiurus quinquestriatus hebraeus	Lqh α IT	1LQH, 1LQI
	Lqh3	1BMR, 1FH3
Leiurus quinquestriatus quinquestriatus	Lqq3	1LQQ
Mesobuthus tamulus	A neurotoxin	1DQ7
Tityus serrulatus	TsTx-VII	1B7D, 1NPI

Scorpio maurus palmatus	Maurocalcine	1C6W
Pandinus imperator	Imperatoxin A	1IE6

Leiurus quinquestriatus hebraeus	Chlorotoxin	1CHL
Mesobuthus eupeus	Insectotoxin I5A	1SIS

6.4.1 Database description

Five bioinformatics features as available in the earlier SCORPION database (Srinivasan *et al*., 2002a) are: **Search Scorpion, BLAST Scorpion, Download FASTA, Scorpion Structure** and **Annotate Scorpion**. A new query tool, **Activity Scorpion** was developed in SCORPION2 database (**Figure 30**).

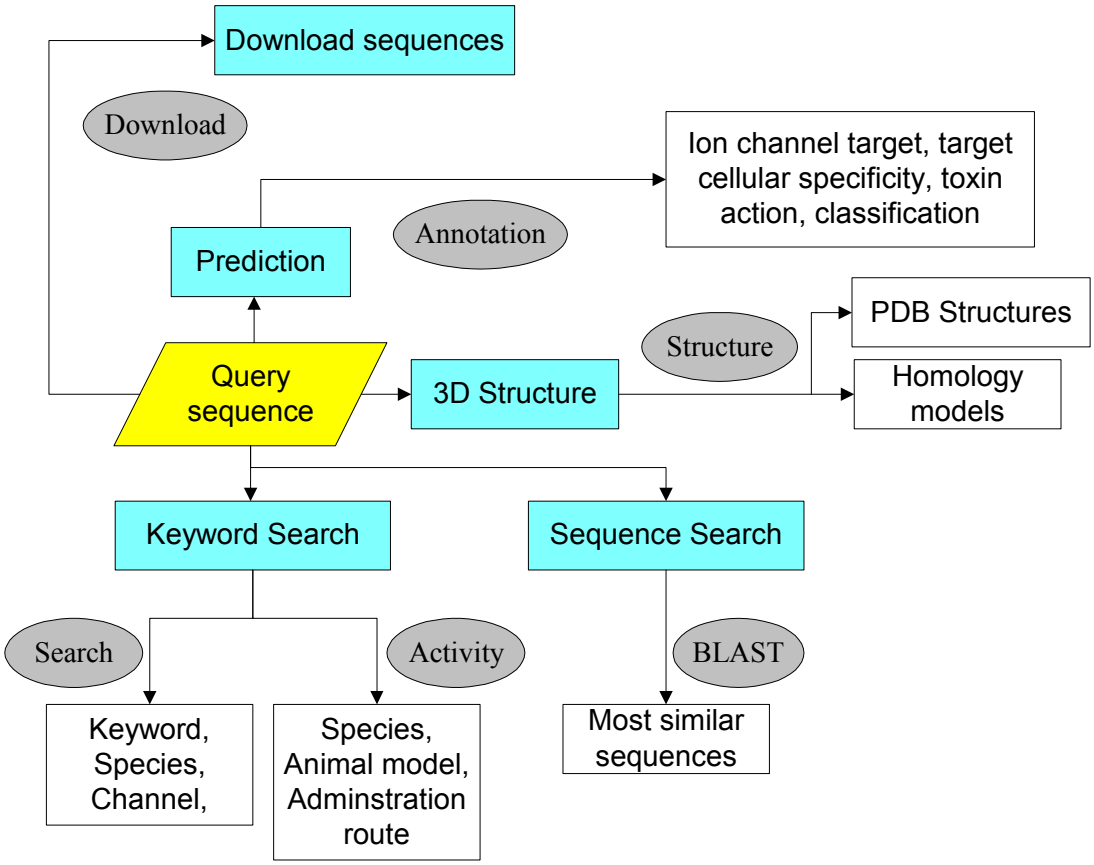

Figure 30 Site map of SCORPION2 database. The six oval shapes represent bioinformatic tools available in the database.

Users can search for scorpion toxin entries using the **Search Scorpion** feature by keywords (e.g. beta toxin, potassium, non-toxic etc.), the specific ion-channel (e.g. sodium, calcium etc.) or scorpion species. The new **Activity Scorpion** feature allows query of scorpion toxin activity through scorpion species, routes of toxin administration or animal models using logical operators (AND, OR). For example, comparison of LD_{50} values by route of toxin administration can be performed by

checking 'intracerebroventricular', 'subcutaneous' and 'AND' options. The LD_{50} values listed corroborates that toxicity in mammalian models are more toxic via intracerebroventricular than subcutaneous route (Gordon and Gurevitz, 2003). This search tool enables identification of the most active toxins in different scorpion venoms where envenomation treatment can be developed (Theakston *et al*., 2003; Gazarian *et al*., 2005). Knowledge of scorpion toxin lethality is important as scorpion stings remains a public health issue (Theakston *et al*., 2003; Gazarian *et al*., 2005). The results of **Search Scorpion** and **Activity Scorpion** features are displayed as lists in a tabular form containing brief descriptions of the records with hyperlinks to individual full data records (**Figure 31**).

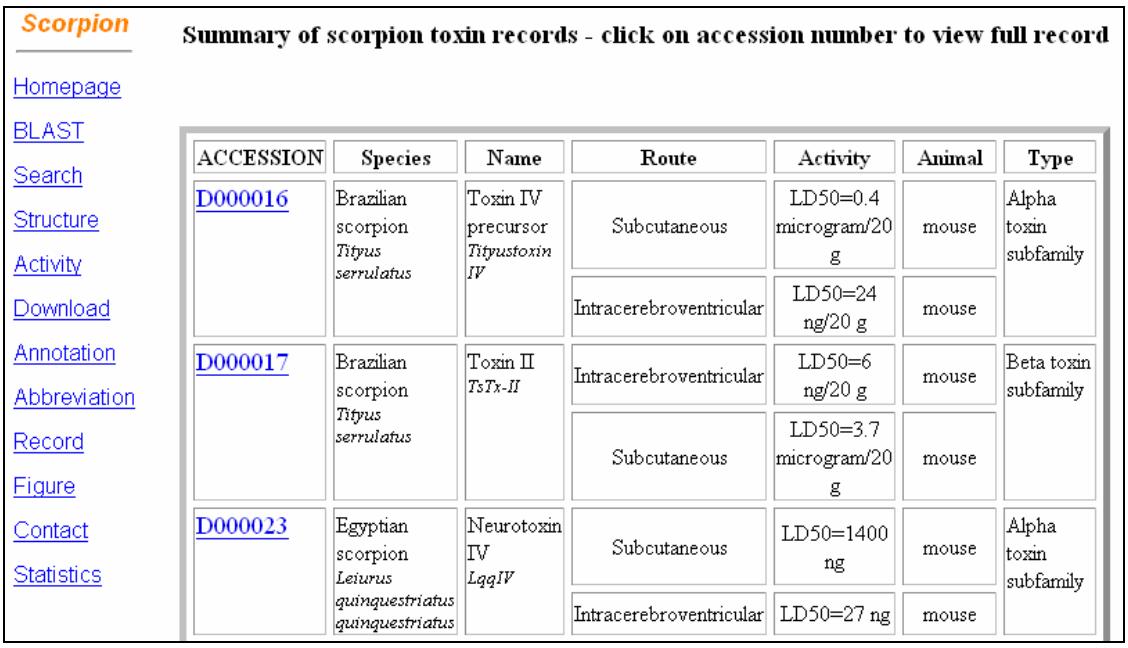

Figure 31 The web interface of the SCORPION2 database. The left frame provides for selection of the various search, extraction and predictive tools. The larger right frame is an example of the output list (partial) of entries that match route keyword search 'intracerebroventricular' and 'subcutaneous'. The accession number fields in the table contain hyperlinks to the full entries in the SCORPION2 database.

The **BLAST Scorpion** feature enables the user to perform sequence comparison using the BLAST (Altschul *et al*., 1997) algorithm. A query sequence, either amino acid or nucleotide, can be compared against all scorpion toxin sequences available in the SCORPION2 database. The users can get the results in either standard BLAST search output or colour-coded multiple sequence alignment generated by the Mview program (Brown *et al*., 1998). The colour-coded alignments indicate the positions of conserved and homologous amino acids in the multiple sequence alignment generated using BLAST searches (**Figure 32**).

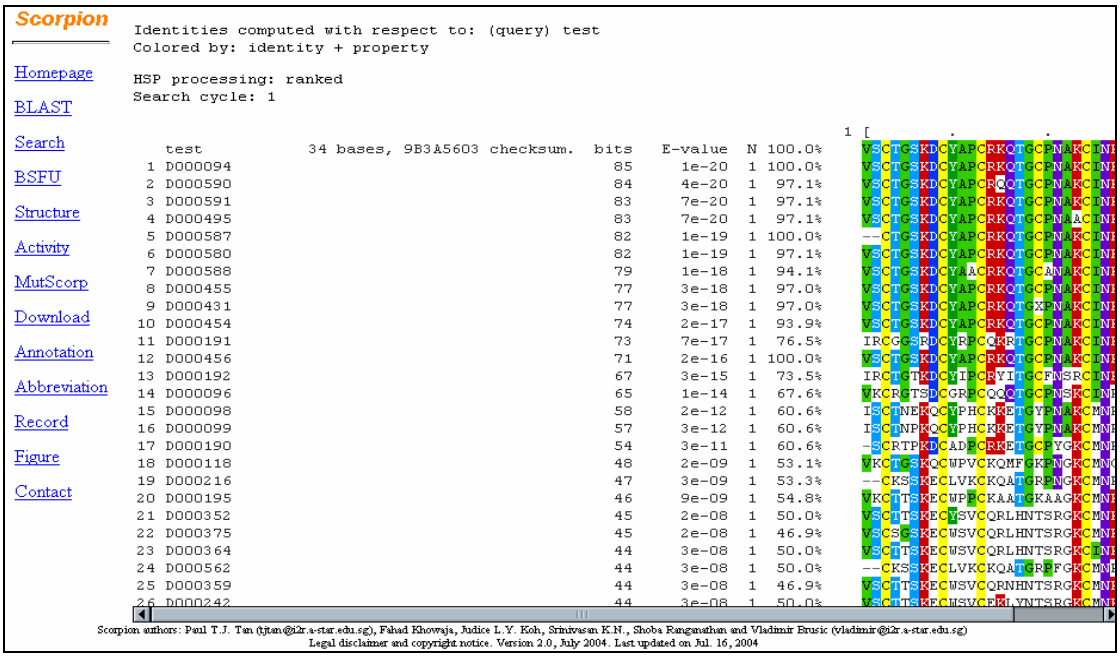

Figure 32 BLAST result upon submission of maurotoxin from *Scorpio maurus palmatus*. Sequences in the SCORPION2 database that are most similar to maurotoxin are ranked in descending order. Positions of amino acid conservation are coloured in the Mview format.

The **Download FASTA** feature enables users to download FASTA formatted files (Pearson, 1990) of amino acid and nucleotide sequences from the SCORPION2 database for sequence analyses of their toxins of interest. The FASTA formatted files are available as zip-compressed files.

The **Scorpion Structure** feature allows users to view and study available 3D structures through either Jmol (http://www.jmol.org/) or Chime (http://www.mdli.com/downloads/) viewer. In addition, users can download the 3D structures in PDB format. These structures include 82 structures extracted from the PDB database and 542 homology models generated for native and mutant toxins as of November 2005. Information on structural template and pairwise sequence identity between template and target (ranging from 32.8 to 100%) are provided in the corresponding SCORPION2 records for each homology model. The quality of each homology model was evaluated by the PROCHECK (Laskowski *et al*., 1993) program and has a Ramachandran plot which can be accessed through a hyperlink. Various display schemes such as spacefill, wireframe and backbone, are available (**Figure 33**).

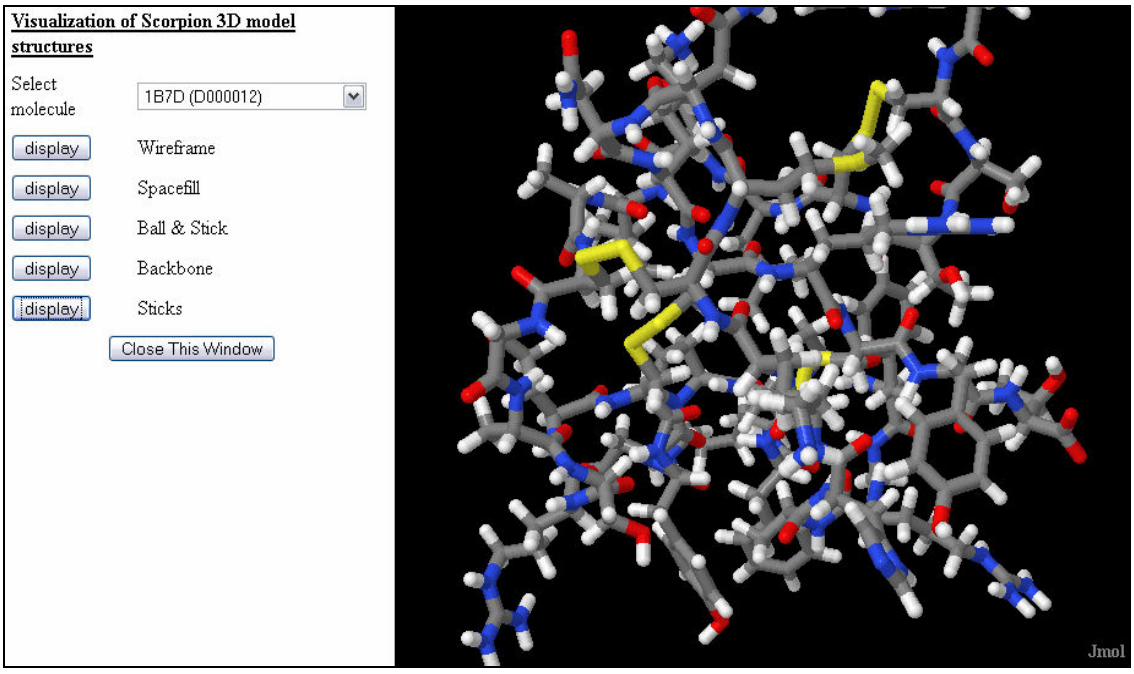

Figure 33 Visualisation of scorpion toxin 3D structures using Jmol. Various schemes of displays such as wireframe, spacefill and sticks are available.

The prediction tool, **Annotate Scorpion**, automatically generates putative functional annotation for the scorpion toxin being queried (discussed in Chapter 5). The annotation result contains the following predicted properties of the query sequence: subfamily of the toxin, ion channel specificity, mechanism of action, nearest neighbour analysis by sequence similarity, and the target cell type. This feature provides multiple sequence alignment, using Clustal W (Thompson *et al*., 1994), of the test sequence along with the nearest neighbour sequences available in the database. The inbuilt intelligence of the **Annotate Scorpion** feature helps the annotation of the fragments of the toxin sequences and identifies novel subfamilies of the query peptides.

6.4.2 Description of the SCORPION2 records

SCORPION2 entries contain fields described in the earlier version of the database (Srinivasan *et al*., 2002a): Date, Accession, References, Species, Name, Mechanism of Action, Toxin_action, Activity, Interaction_site, Target_cell, Toxin Type, Structure, Disulfide, Alpha_helix, Beta_strand, C_end, Sequence, Designation, Translation, AA Alignment, Conflict, and Comment. New fields that contain functional, structural, or other relevant information have been added to SCORPION2. A summary of fields in a SCORPION2 full data record is given in **Table 10**. The description of each field will be divided into database, functional and structural categories.

Table 10 Description of fields in a SCORPION2 record. The field values may be textual, numeric or empty and can be generally divided into database, functional and structural categories. New fieldnames introduced in SCORPION2 are in **boldface**.

Fieldname	**Description**
Database fieldnames	
DBACC	Unique accession number "D" follows by six-digit number
Date	Entry date of record
Last_updated	Last date of updating with relevant literature or information
Accession	Cross-references to corresponding records in GenBank, EMBL, DDBJ, Swiss-Prot and PDB databases if available
Reference	Publication references and hyperlinks to the PubMed database, containing the author names, title and journal reference
Functional fieldnames	
Species	Biological and common name of the scorpion
Name	List of names used in the literature or a specific toxin name
Mechanism of action	Description of the biological mechanism of the toxin activity
Toxin action	Nature of toxin
Activity	Potency of toxin
Target cell	Species specificity of toxin
Toxin type	Subfamily of toxin
Toxin effect	Symptoms observed in animal models upon toxin administration
Binding affinity	Binding affinity of toxin towards a target ion channel
Channel	Target ion channel. Source of ion channel is enclosed in parenthesis if available
Interaction site	Suggest possible residues involved in function of toxin
Comment	Any relevant functional or structural observations
LD_{50}, ED_{50} or PU_{50}	Lethal dosage, effective dosage or paralytic unit at 50% upon administering toxin into animal models, respectively
Injection route	Route of administering the toxin into the animal models
Source	Source of the toxin
Origin	Either native or mutant toxin
Rel_activity	Relative toxicity of mutant toxin
Rel_binding_affinity	Relative binding affinity of mutant toxin
Mutation	Highlights the position and mutation of the sequence such as amino acid substitution, insertion and deletion
Structural fieldnames	
Structure	Access to PDB toxin structures, homology models and other related structural information
Ramachandran	Access to a Ramachandran plot of each homology model generated
CD_spectrum	Annotate circular dichroism spectrum of mutant toxin sequence
Cys_arrangement	Cysteine motif with number of intervening residues (X) in parenthesis
Disulfide	Positions of the residues involved in disulfide bridges
Disulfide_arrangement	One of the eight types of disulfide connectivity patterns currently observed in 393 native scorpion toxin entries
C_end	Nature of C-terminus whether free or post-translationally modified e.g. amidated, sulfoxide
Sequence designation	Can be a precursor, mature or a partial sequence toxin
Translation	Complete amino acid sequence of toxin with the disulfide pattern, together with the signal peptide if present
AA Alignment	Information on the groupings of scorpion toxins
Sequence	Nucleic acid sequence of toxin that describes both the introns and exons if available
Conflict	Observations regarding amino acid discrepancies between records representing the same sequence in different databases

The fields for the database category are as follow. A unique accession number '**DBACC**' has been assigned to each of the records in the SCORPION2 database. The format begins with 'D' follows by a six-digit number. This field is followed by the '**Date**' and '**Last_updated**' which identifies the entry date and date of updating with relevant information, 3D structures or literature. The field '**Accession**' contains hyperlinks to corresponding accession numbers in public databases GenBank, EMBL, DDBJ, Swiss-Prot, and PDB. The field '**References**' provides publication references and hyperlinks to the PubMed database. It contains the author names, title and the journal reference. The **'Comment**' field has been reserved for any relevant comments or observations.

Functional information is provided in the following fields. The field '**Species**' gives the biological and trivial name of the scorpion. The '**Name**' field contains a list of names used in the literature or a specific name, for example "Neurotoxin KTX2 (BmSKTx2) (Bm KK3)" and "makatoxin". Some entries have only one name, while some are known by multiple names. A brief description of the biological mechanism of the toxin activity (such as whether the toxin interacts with Na^+ or K^+ channel) can be found in the field '**Mechanism of Action**'. The '**Toxin_action**' and '**Activity**' fields describe the nature of the toxin and its potency. These fields **'LD_{50}', 'ED_{50}'** or **'PU_{50}'** describe the lethal dosage, effective dosage and paralytic unit at 50% upon administering toxin into animal models, respectively. The animal models are enclosed within parenthesis. '**Injection_route**' is the route of administering toxins into the animal models. Information on binding affinity of a toxin towards a particular ion channel is provided in '**Binding_affinity**' and '**Channel**' fields, respectively. The source of the experimented ion channel, if available, is found in parenthesis in the '**Channel**' field. The possible amino acid residues critical for toxin-channel

interactions are provided in the field '**Interaction_site**'. The species specificity of the toxins is provided in the field '**Target_cell**'. The '**Toxin Type**' field specifies the subfamily of the given toxin. Source (*venom, recombinantly expressed, chemically modified and chemically synthesized*) of the toxin entries are found in '**Source**'. Mutant and wild type toxin entries can be differentiated by either *Mutant* or *Native* in **'Origin'**. The relative effect of mutation on the activity and binding affinity can be obtained from the fields '**Rel_activity**' and '**Rel_binding_affinity**'. The '**Mutation**' field highlights the position and mutation of the sequence such as amino acid substitution, insertion and deletion. Regions of mutation are highlighted in the primary sequence.

The '**Structure**' field gives direct access to the 82 3D structures and other related structural information found in the PDB database. In addition, SCORPION2 contains homology models for 542 toxins that currently lack experimentally determined 3D structural information. The phi (Cα-N bond) versus psi (Cα-C bond) torsion angles for all residues in the homology model can be assessed from the '**Ramachandran**' field. Mutation studies which included circular dichroism spectroscopy to determine if mutation affected structural folding are annotated in the '**CD_spectrum**' field. The positions of the residues involved in disulfide bridges, α-helices or β-strand formations are described in the fields '**Disulfide**', '**Alpha_helix**' and '**Beta_strand**', respectively. The phrase "BY SIMILARITY" in the '**Disulfide**' field stands for putative disulfide bridges determined by similarity to known structures. '**Disulfide_arrangement**' describes one of the eight types of disulfide connectivity patterns currently observed in 393 native scorpion toxin records. Information regarding the nature of the C-terminal, whether free or post-translationally modified can be obtained from the field, '**C_end**'. The '**Sequence Designation**' field describes whether

a particular entry is a precursor sequence, a mature toxin or a partial sequence. The complete amino acid sequences of the toxin with the disulfide pattern, together with the signal peptide if present, are displayed in the field **'Translation'**. The disulfide bridges, the signal peptide and the mature toxin are distinctly colour coded for easy identification and clear understanding. The field '**AA Alignment**' provides the information on the classified groups of scorpion toxins. This feature of the SCORPION2 database is based on multiple sequence alignment of sequences and a unified grouping of toxins, created using Clustal W (discussed in Chapter 3). The Na^+ channel toxins are classified into 18 groups based on sequence similarity and the position of other conserved residues. Similarly, the K^+ channel toxins have been classified into 32 groups. The Cl^- channel toxins, scorpion defensins and other short-chain neurotoxins are grouped separately. All the information on the groupings of scorpion toxins can be obtained from the field '**AA Alignment'**. The nucleic acid sequences of the toxins that describe both the introns and exons are shown, where available, in the '**Sequence**' field. The field **'Conflict'** contains observations regarding amino acid discrepancies between records representing the same sequence in different databases.

6.5 Discussion and conclusion

SCORPION2 is an improved resource of native scorpion toxins and artificial mutant toxins with integrated bioinformatics tools for systematic analysis of their structure-function relationships. Inclusion of mutant scorpion toxins provides information on amino acid identities and positions that may affect toxin function and fold. Designed using data warehousing principles, SCORPION2 is subject-orientated which provides a knowledge base to focus on and analyse scorpion toxin data.

SCORPION2 is a platform where bioinformatic analyses (Chapters 3 and 4), development of extraction and prediction tools for structure-function relationships of scorpion toxins (Chapter 5) are integrated. It is also an example of how to bridge the gap between the fast-paced accumulation of scattered data and their proper management for better utilisation of the current knowledge. It is also intended to allow users to make use of the bioinformatic tools and link them to experimental design in the study of scorpion toxins.

The SCORPION2 is an integrated resource of scorpion toxin data (3D structures, native and mutant toxin sequences) largely compiled from published reports and public databases by keyword and sequence similarity searches. Errors were eliminated by cross-referencing of entries with original papers and other databases. The enlarged and cleaned dataset serves to better represent the current knowledge of their structure-function relationships. The steps of data cleaning, consistency check and data annotation are crucial for the attainment of high quality data which further facilitate detailed analyses. These steps are the most difficult and most time-consuming in the creation of the database. Full automation of annotation helps speed up the process but may introduce false positives and false negatives. Manual annotation by domain expert is more accurate and complete than automated annotation but introduces a degree of random errors not found with automated annotation (Ding *et al*., 2004). Therefore, combining fast automated process with accurate expert inspection could enhance the annotation process. This ideal semi-automated environment provides a human annotator to confirm, edit or reject automatically generated annotations of records.

The unique and novel aspects of the SCORPION2 includes i) 548 newly created homology models which are not available elsewhere, and ii) the functional

prediction module which enable users to infer new knowledge from existing scorpion toxin data. Important structural and functional information can be extracted from mutant toxin records. This information coupled with the available 3D structures provides users with an easy access to information related to potential structure-function relationships for uncharacterised toxins. The prediction tool helps annotation and determination of functional properties of novel peptides. The architecture of the SCORPION2 database allows for easy integration of bioinformatic tools for additional analyses of scorpion toxin data.

SCORPION2 database is a model for management of molecular data of toxins of various venomous animal species. By creating data warehouses of toxins, researchers have the access to more complete data sets for further analysis and interpretation than is available in general-purpose databases. The data warehouses of toxins integrated with analysis tools will enable classification of the data and application of data mining methods for discovery and extraction of new knowledge. The large-scale analysis of structure-function relationships of various toxins could draw a more accurate inference and aid understanding of the correlation among different toxins.

Chapter summary

- The highlight of this chapter is the implementation of SCORPION2 database. It is the first data warehouse of native and mutant scorpion toxins with integrated bioinformatics tools for users to analyse their structure-function relationships and aid research planning and knowledge discovery. It is an improved resource of scorpion toxin data (nucleotide and primary sequences, 3D structures and relevant references) where users need not query multiple resources.

- The creation of SCORPION2 involves multiple steps: 1) access to toxin data across multiple databases, 2) inspection and subsequent cleaning of errors, 3) annotation of structure-function information from literature, 4) analysis and classification of toxin sequences and their structures, and 5) the design and use of predictive models for simulation of laboratory experiments.
- The success of SCORPION2 serves as an example for the management of molecular toxin data from various venomous animals where the gap between the fast-paced accumulation of scattered toxin data and their proper management can be bridged.

Part III: Chapter 7 Exploring bioinformatic approaches for functional prediction of bioactive scorpion toxins

'The best computer is a man, and it's the only one that can be mass-produced by unskilled labor.'

Wernher Magnus Maximilian von Braun

Scorpion toxins are important pharmacological tools for probing the physiological roles of ion channels which have significant therapeutic potential for an array of medical disorders (Rodriguez de la Vega *et al*., 2003; Blank *et al*., 2004; Bagdany *et al*., 2005). The discovery of new scorpion toxins with different specificities and affinities is needed to further characterise the physiology of ion channels. However, to serve as effective probes, the new toxins identified must not only be specific but also bind with high affinity to targeted ion channel. Binding affinity data of native and artificial mutant scorpion toxins to various ion channel subtypes are available in the literature. These data can be used to develop a tool for predicting toxin strength of binding affinity to specific ion channel. Accurate prediction of toxin binding strength facilitates identification of high-affinity toxin binders which improves efficiency and the economy of experimentation.

This is the first report of a prediction tool based on a bioinformatic-driven approach involving sequence comparison, nearest neighbour analysis, decision rules, scaled binding affinity data, and functional motifs (discussed in Chapter 4) to predict strength of binding affinity to ion channels in the field of venom research.

7.1 Materials and Methods

Native and artificial scorpion toxin data were collected and enriched with binding affinity information from literature (discussed in Section 6.3). A test set of 26 newly identified scorpion toxins with experimentally determined binding affinity data was used for evaluating the prediction accuracy. Binding affinity data were scaled to a common scale (discussed in Section 4.1.1) and classified into four categories, namely 'non-binding', 'low', 'medium' and 'high' binding (**Table 11**). These four categories

were used for sequence comparison of toxin sequences. The functional motifs specific to Na^+ and K^+ channels (discussed in Section 4.2) were incorporated into an algorithm for predicting ion channel specificity and strength of binding affinity. A flowchart of the approach is available in **Figure 34**.

Table 11 Four categories of strength of binding affinity were introduced, namely 'Non-binding', 'Low', 'Medium' and 'High' from the mapping of binding affinity to a common scale discussed in Section 4.1.

Class	1	2	3	4	5	6	7	8
Mapped affinity	Non-binding	Very low	Low	Moderately low	Moderate	Moderately high	High	Very high
Prediction	**Non-binding**	**Low**		**Medium**			**High**	

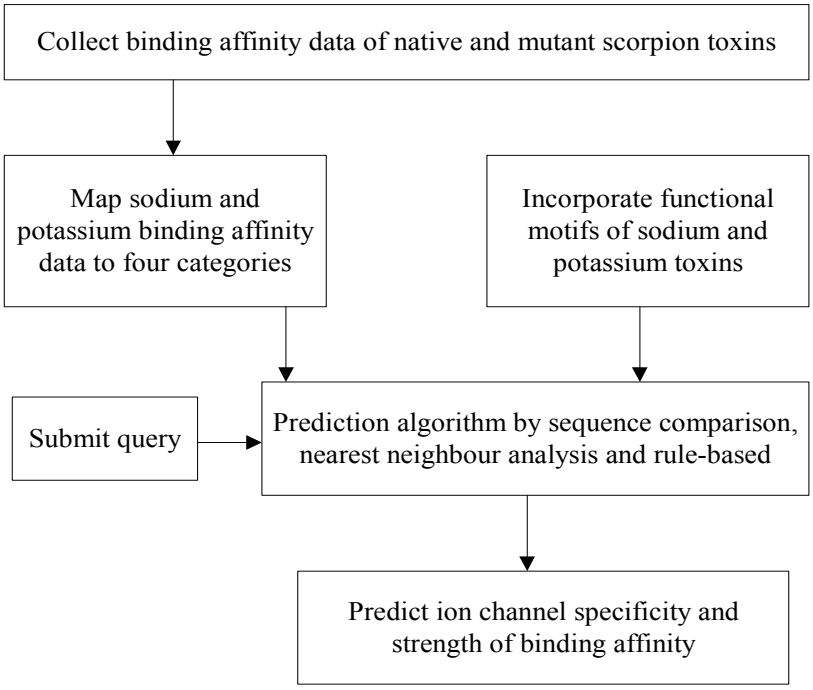

Figure 34 Flowchart of predicting ion channel specificity and strength of binding affinity.

7.1.1 Algorithm for predicting strength of binding affinity of scorpion toxins

Prediction of strength of binding affinity to K^+ and Na^+ ion channels was performed by modifying the algorithm described in Section 5.2.2. This modified module also combines sequence comparison, nearest neighbour analysis and decision rules for assigning a putative classification of a query sequence to a structure-function group. The grouping of the query to similar sequences permits nearest neighbour analysis to assign the scaled binding affinity data in the neighbour sequences to the query. For example, if the query matched neighbour sequences with 'High' and 'Very high' mapped affinities, it was predicted as 'High' affinity and so forth with prediction of 'Medium', 'Low' and 'Non-binding' categories. The query sequence was also compared against scorpion mutant toxin data by BLAST program (Altschul *et al*., 1997) where residue(s) in the query sequence that matched with mutations had the relative binding affinity information extracted. If residues in the query did not match any mutation, it was compared against the functional motifs (discussed in Section 4.2) to highlight the positions and identities of amino acids in the query sequence, which may affect strength of binding affinity.

7.2 Results

The accuracy of predictions was validated by submitting 26 newly identified scorpion toxins with experimental data. All predictions of ion channel specificity were correct (100%) (**Table 12**). For strength of binding affinity, two agreed with experimental data (7.7%) (Meki *et al*., 2000; Alami *et al*., 2003) (**Figure 35, Figure 36**), three were wrong (11.5%) and 21 were not predicted (80.8%). The tool also highlighted positions and residues in the query sequence which may affect the strength of binding affinity.

Validation with experimental data has shown that this approach is not predictive in the strength of binding affinity of scorpion toxins to ion channel. Thus, there is a limit to the use of bioinformatics in prediction of toxin-binding strength but not specificity to ion channels.

Table 12 Predicted ion channel specificity and strength of binding affinity for 26 newly identified scorpion toxins with experimental data. Toxins are arranged alphabetically. Pu represents prediction, Ex represents experimental data, * represents not predicted, NB, L, M and H represent non-binding, low, medium and high binding, respectively. K and Na represent potassium and sodium ion channels, respectively.

Toxin name	**Ion channel**		**Strength**		**Reference**
	Pu	**Ex**	**Pu**	**Ex**	
AmmTX3	K	K	L	M	Vacher *et al.*, 2002
AmmVIII	Na	Na	M	M	Alami *et al.*, 2003
Anuroctoxin	K	K	*	M	Bagdany *et al.*, 2005
BmK37	K	K	*	M	Xu *et al.*, 2004a
BmKIM2	Na	Na	*	L	Peng *et al.*, 2002
BmKSKTx1	K	K	*	L	Xu *et al.*, 2004b
BotIT6	Na	Na	*	M	Mejri *et al.*, 2003
BtK-2	K	K	*	L	Dhawan *et al.,* 2003
Cn12	Na	Na	*	L	del Rio-Portilla *et al.*, 2004
CsEKerg1	K	K	*	M	Nastainczyk *et al.*, 2002
Discrepin	K	K	*	M	D'Suze *et al.*, 2004a
IsTx	K	K	*	L	Yamaji *et al.*, 2004
Kbot1	K	K	*	M	Mahjoubi-Boubaker *et al.*, 2004
KTX3	K	K	M	M	Meki *et al.*, 2000
Lqh6	Na	Na	NB	M	Hamon *et al.*, 2002
Lqh7	Na	Na	NB	M	Hamon *et al.*, 2002
Lqhβ1	Na	Na	*	H	Gordon *et al.*, 2003
Tc30	K	K	*	L	Batista *et al.*, 2002a
Tc32	K	K	*	M	Batista *et al.*, 2002a
Tc48a	Na	Na	*	M	Batista *et al.*, 2004
TsPep1	K	K	*	NB	Pimenta *et al.*, 2003
TsPep2	K	K	*	NB	Pimenta *et al.*, 2003
TsPep3	K	K	*	NB	Pimenta *et al.*, 2003
TtBut-toxin	K	K	*	L	Coronas *et al.*, 2003a
Tz1	Na	Na	*	L	Borges *et al.*, 2004
κ-KTX1.3	K	K	*	NB	Nirthanan *et al.*, 2005

Query name	Query
Query sequence	VGIPVSCKHSGQCIKPCKDAGMRFGKCMNRKCDCTPK
Channel	K
Binding Affinity	Medium
Mutations	The binding affinity may be moderately affected. For SKca sub-channels, position 11 is not R/K; For SKca sub-channels, position 12 is not R/M; For Kv1.3, Shaker and BKca sub-channels, position 33 is not K/R/H;
Range for binding affinity towards potassium channel (nM)	
High	<0.032
Medium	0.032 - <252
Low	252 - <100,000
Non Binding	>=100,000

Figure 35 Prediction of the binding affinity of KTX3 from *Buthus occitanus tunetanus*. It is predicted to target K^+ channel with medium binding affinity which agrees with experimental results (IC_{50} = 0.05 nM) (Meki *et al*., 2000). The positions and residue identity in the query sequence that do not coincide with the functional motifs are highlighted.

Query name	Query
Query sequence	LKDGYIVNDINCTYFCGRNAYCNELCIKLKGESGYCQWASPYGNSCYCY KLPDHVRTKGPGRCND
Channel	Na
Binding Affinity	Medium
Mutations	Mutation(s) in this sequence may affect the following: For Beta type, toxicity to insects and mammals may be affected because position 13 is not K For Beta type, binding affinity may be moderately affected because position 19 is not L For Beta type, binding affinity may be strongly affected because position 22 is not N For Beta type, binding affinity may be strongly affected because position 25 is not E For Beta type, binding affinity may be strongly affected because position 37 is not Y For Beta type, binding affinity may be strongly affected because position 39 is not F For Alpha, Alpha-like type, binding affinity may be moderately affected because position 57 is not I For Alpha, Alpha-like, binding affinity may be strongly affected because position 58 is not R For Beta type, structural integrity may be affected because position 61 is not K For Alpha, Alpha-like, binding affinity may be moderately affected because position 64 is not H/R
Range for binding affinity towards sodium channel (nM)	
High	<0.6
Medium	0.6 - <770
Low	770 - <100,000
Non Binding	>=100,000

Figure 36 Prediction of the binding affinity of AmmVIII from *Androctonus mauretinicus mauretinicus*. It is predicted to target Na^+ channels with medium binding affinity which agrees with experimental results (EC_{50} = 29 nM and 416 nM) (Alami *et al*., 2003). The positions and residue identity in the query sequence that do not coincide with the functional motifs are highlighted.

7.3 Discussion and conclusion

The application of bioinformatic approach to functional prediction of binding affinity of scorpion toxins to ion channels was explored in this chapter. There are several possible reasons for the poor performance in predicting strength of binding affinity where majority were not predicted. First, the number of binding affinity data available is insufficient. The pharmacological data for many scorpion toxins remain to

be determined. With no binding affinity information, nearest neighbour analysis cannot ascribe an unknown function in the nearest neighbour to the query. Second, the sequence similarity of the query is too diverse where no nearest neighbour is found as observed in the new groups proposed for the newly identified scorpion toxins (discussed in Chapter 3). Third, toxin-channel binding processes are extremely complicated, determined by many factors such as shape and size complementary, hydrophobicity and polarity. Also, toxins and channel proteins are flexible molecules, and the energy inventory between the bound and unbound states must be considered in aqueous solution (Gohlke and Klebe, 2002). This information cannot be captured by primary sequence similarity. Lastly, the algorithm may be inadequate to solve the prediction of toxin binding strength to ion channels.

In conclusion, the prediction of strength of binding affinity of scorpion toxins to ion channels cannot be achieved with the nearest neighbour method. The limits of the current approach have been explored in this work and as new data become available, the situation might change.

Chapter summary

- There is a need to discover new scorpion toxins with high specificities and affinities for further physiological characterisation of the ion channels.
- Accurate prediction of toxin binding strength aids identification of high binders where weak or non-binders need not be tested, thus improving the efficacy of experimentation.
- The limits of the current approach to predict toxin binding strength was explored as validated by experimental data.

Part IV: Chapter 8 General discussion

'Life is a continuous exercise in creative problem solving.'

Michael J. Gelb

The introduction of bioinformatics as a discipline in the last quarter of the 20th century, it has revolutionised the approach in which biological research is conducted. Researchers are increasingly conducting searches in public databases for characterised sequences that match their sequences before conducting experiments to determine their function. Bioinformatics narrows down the number of essential experiments needed and thus expedites the discovery process. In this work, the author employed a top-down approach which serves as an emerging alternative strategy for prediction of the structural and functional properties of the large pool of uncharacterised scorpion toxins in contrast to traditional experimental bottom-up approach. Coupling both approaches have the advantage of reducing laborious and time-consuming bench work.

In this thesis, application of bioinformatics to venom research was accomplished by data collection, cleaning and enrichment, classification and analyses, development of prediction tool, and implementation of specialised data warehouse of bioactive toxins in scorpion venom as an example. During the process of data collection and cleaning the public database records, several errors in the data were identified and corrected. Examples include high redundancy due to maintenance of the same sequence in different public databases, discrepancies in primary sequences and conflicting annotation. More biological data artifacts, which affect accuracy of data mining has been studied by Koh *et al*. (2004b). Data checking and correction are thus critical for the improvement of data quality. Interpretation of unclean data is normally inaccurate and errors will be propagated in subsequent analysis where high data quality is important for accurate predictions. Further, the basic information in the records prevented application of data mining. Thus, annotation of functional and structural information from literature to records was manually performed to improve the data

quality. However, manual annotation does not scale up to capacities needed for large-scale analysis of bioactive toxins.

Proper data classification is also needed for the accuracy of predictions. There are two principal approaches taken for data classification, manual and automatic. Manual classifications are based on human expertise, facilitated by bioinformatic analyses, to cluster data into particular groups that share common properties defined by domain experts. Examples include the manually curated Swiss-Prot and PROSITE databases. The other approach is automatic classifications which depend on algorithms or models to generate matrices for similarity or distance that are then processed to determine these groups. Examples of automatic classification algorithms include self-organised maps, artificial neural network, and support vector machines which belong to the fields of artificial intelligence and machine learning (Kapetanovic *et al*., 2004). ProDom and DOMO (Gracy and Argos, 1998) among others, address classification more systematically with automated processes that classify entire protein sequence databases. The advantage of manual classification is the high quality of clustering but the final classification result may be irreproducible because of differences in the experts' knowledge. In contrast, automation is fully reproducible because of fixed rules written in computer programs and scalable to large data set.

A combination of different bioinformatic approaches was performed for the large-scale analyses of toxin data. Multiple alignments of protein sequences are an effective way of identifying conserved amino acids that provide clues to functional relationships among proteins. The patterns of amino acid variability in multiple sequence alignments reveal evolutionary pressure, mutation, recombination and genetic drift that spans millions of years (Valdar, 2002). Conserved residues among related toxin families are usually involved in functional properties of the peptides

(Gilquin *et al*., 2002). Closely interlinked with functional properties of proteins is the correct structural folding of proteins. A linear protein sequence must be folded correctly in spatial organisation for maintaining the optimal positions of functional residues on the protein molecule for interaction. Thus, 3D analysis helps understand spatial properties and molecular structure of toxins, which provide clues for identification of interaction sites and key structure-function features. The key structure-function information from mutation studies of toxins is extracted from the literature for analyses and extraction of functional motifs. Therefore, a combinatorial approach to analyse toxins, as demonstrated in this work, can greatly enhanced understanding of their molecular function.

Prediction of a protein function from primary sequence and tertiary structure is a challenging issue, because homologous proteins often have different function. Though detailed studies on determination of a function in isolated bioactive toxin are performed, researchers cannot be confident that the toxin's full repertoires of biological activities are known. Many methods of function prediction rely on identifying similarity in sequence and structure between a protein of unknown function and one or more characterised proteins (Whisstock and Lesk, 2003). Prediction of function from protein sequence based on detecting sequence similarity or matching motifs have two main weaknesses. First, a similar protein must be available for comparative analysis else prediction is not made. For instance, *Annotate scorpion* tool cannot predict functional properties of κ-hefutoxins from *Heterometrus fulvipes* because no similar sequence was found to adopt a hairpin structure of two short helices cross-linked with two disulfide bridges (Srinivasan *et al*., 2002b; Nirthanan *et al*., 2005) which is different from the CSH fold in majority of scorpion toxins. Second, at least one of the similar proteins must be characterised in order for inference of

function. When this fail, new methods which are not reliant on alignments are appearing such as gene neighbour, inductive logical programming and text analysis (Dobson *et al*., 2004).

The implementation of specialised data warehouse helped to systematise the approach for in-depth analysis and discovery of new knowledge of bioactive toxins in scorpion venom where it captures domain expertise in the analysis and interpretation of the classified toxin sequences. The author envisages this database as a model for management of molecular data of toxins of various venomous animal species. A collection of different specialist toxin databases provides researchers access to the most complete toxin data sets for further analysis, interpretation and formulating hypotheses. The integrated analysis tools in the specialist toxin databases will enable researchers to classify the toxin data and apply data mining methods for discovery and extraction of new knowledge. The large-scale analysis of structure-function relationships of various toxins, supported by bioinformatics, could aid understanding of the correlation among different toxins.

Chapter summary

- Bioinformatics, when applied appropriately by researchers, helps in determining the key experiments to be undertaken, thus improving the efficiency of biology research.
- In this work, bioinformatics was successfully applied to scorpion venom research by data collection, cleaning and enrichment, classification and analyses, development of prediction tool, and implementation of specialised data warehouse of bioactive scorpion toxins. Data cleaning and enrichment is essential for maintaining high data quality needed for accurate prediction.

- A combination of different bioinformatic approaches such as analyses of information from mutation studies, multiple sequence alignments and 3D structures was performed for the large-scale analyses of scorpion toxin data.
- Creation of the scorpion toxin data warehouse aids systematic analysis and discovery of new knowledge of bioactive toxins in scorpion venom.

Part IV: Chapter 9 Conclusion

'Concern for man himself and his fate must always form the chief interest of all technical endeavors, concern for the great unsolved problems of the organisation of labor and the distribution of goods – in order that the creations of our mind shall be a blessing and not a curse to mankind. Never forget this in the midst of your diagrams and equations.'

Albert Einstein

9. Conclusion

This work started with an investigation of the use of bioinformatic-based approach to the large-scale study of structure-function relationships of scorpion toxins. By the systematic application of bioinformatics to scorpion venom research, this work has defined the novel field of venominformatics. Because this thesis is to his knowledge the first research project that deals systematically with bioinformatics in venom research, the author had to take a narrow focus, covering only bioactive toxins from scorpion venom. The author focused on the classification, analyses, development of predictive tool and creation of scorpion toxin data warehouse for large-scale study of scorpion toxins. By completing this project, the author has advanced the knowledge about the bioinformatic-based approach in venom research. The author now summarises his conclusions.

9.1 Large-scale classification

Large-scale classification of all known toxins according to their function is necessary for clarifying the current knowledge, including an overview of the functional repertoire of the toxins. Such knowledge will facilitate functional assignment of newly identified bioactive toxins. Large-scale classification is an important step towards effective information management where relevant biological information can be retrieved from vast amounts of toxin data. Classification of related biological sequences can be used to predict the function of an unknown sequence based on inference of homology between the unknown and the characterised sequences in a class.

Here, the author has described a systematic large-scale classification of 393 currently known scorpion toxin sequences into 62 groups based on ion channel specificity and primary sequence similarity, combined with multiple sequence alignments and phylogenetic analyses. With a large repository of toxin sequences, a broad perspective of the general patterns in their structure and function can be observed. This present large-scale classification of different toxins relates to current knowledge of the field and as more information becomes available in the future, it is expected that revision of classification will be necessary. This classification approach, as a generic tool, can be applied to large-scale classification of other bioactive toxins and families of bioactive peptides.

9.2 Large-scale analysis

With the accumulation of toxin data and increasing information of their complex structure-function properties, there is a need to develop new computational strategies for extraction of information from low-level raw data to support large-scale venom research. The large-scale analysis of structure-function relationships of various animal toxins could draw more accurate inferences and aid our understanding of the correlation among different toxins.

The approach to include structure-function information from mutation studies of scorpion toxins combined with multiple sequence alignment and 3D structure analyses provide biological significance to the extraction of functionally relevant motifs, complementing motifs obtained statistically. This work reports the first binding motifs for four K^+ ion channel subtypes (voltage-dependent K^+ channels, large- and small-conductance Ca^{2+}-activated K^+ channels, and ether-a-go-go K^+ channel), four binding site motifs for Na^+ channels and a conserved motif for Cl^- channels. This

systematic approach of including mutant data of scorpion toxins and 3D structure analyses for extraction of motifs can serve as a model for other bioactive toxins where mutation studies and 3D structures are available.

9.3 Development of functional prediction tool

Laboratory experiments involving large number of sequences are extremely costly and time-consuming. In contrast, the advantages of applying bioinformatics in biological research are lower cost, speed and efficiency. Identification and optimisation of key experiments can be achieved by combining conventional experimental methods for structure-function analysis, such as site-directed mutagenesis and chemical modifications, with bioinformatic-driven structure-function study of toxins for identification of key experiments and help optimisation of experimental design. Bioinformatics bridges the gap between the fast data growth and the slower pace of experimental validation studies by facilitating data analyses and interpretation, and understanding biological processes. The computational algorithms developed from data analyses play an increasingly important role in the formation and testing of hypotheses to determine function of proteins more rapidly.

Database mining for discovery and extraction of new knowledge is new in toxinology, with excellent prospects to facilitate future advances of this field. In addition, the detailed structural and functional grouping of protein sequences can be used for accurate prediction of functional properties of other toxins and other families of bioactive peptides. In this era of large-scale screening using genomics and proteomics, venominformatics will become increasingly important for management and analysis of toxin data and provides a framework for efficient analysis and maximisation of knowledge extraction from large amounts of these pharmacologically

important toxin data.

A systematic approach to the prediction of functions of bioactive toxins in scorpion venom through bioinformatics was accomplished in this work. Detailed classification of scorpion toxin sequences into well-organised groups allows better correlation of structure-function relationships and thereafter, classification of new sequences by the prediction tool. This work demonstrated high accuracy (>90%) of predicting ion channel specificity and toxin family based on sequence comparison, nearest neighbor analysis and decision rules. This generic bioinformatic-driven approach serves as a model for functional prediction of novel toxins from other venomous animals. Functional predictions of bioactive toxins help to minimise the number of experiments performed by narrowing the validation processes with the predicted function. It complements wet-lab experiments as demonstrated in the wide application of bioinformatics in various fields including toxicology (Fielden *et al*., 2002), drug discovery (Gagna *et al*., 2004), genetics (Fishelson and Geiger, 2004) and pharmacology (Ross *et al*., 2004), creating a dry-wet cycle. The application of bioinformatics in the functional prediction of toxin-binding strength to molecular ion channel targets is currently limited as validated by experimental data.

9.4 Data warehouse of scorpion toxins

Biological databases now play a central role in directing medical and biological research with the huge amount of data generated by genomics and proteomics, and high-throughput technologies. The number of newly identified toxin sequences is growing fast, while their functional characterisation is lagging. These toxin data are deposited across different public databases, usually with primary sequence information only. Discrepancies between database records are common. Structural and functional

information, in particular from mutation studies, is available mainly in the literature. Analyses of toxin data are growing increasingly difficult and there is a need for cleaned and well annotated databases of toxins.

Data warehouses of toxins serve as valuable resources for exploration of toxins, allowing users to query complex biological questions that may usually involve searching multiple sources. Researchers have the access to the most complete data sets for analysis and interpretation. This thesis has demonstrated such applications with relatively small data sets of scorpion toxin data (up to 1000 entries), while work with larger data sets will require additional data management and data analysis tools, supported by bioinformatics (Koh *et al*., 2004a). This thesis had systematically collected and combined toxin sequence and 3D structure data, and hypothesis-driven results from literature into highly useful resources for experimental biologists. More than 500 homology models of scorpion toxins have been generated and deposited in the SCORPION2 database. The 3D models of these toxins coupled with multiple sequence alignment of each group in the toxin classification, can be employed to investigate the functional residues of each pharmacological activity. The utilisation of knowledge, particularly of functional information, has been made more efficient by enriched descriptions of toxin entries and integration of bioinformatic tools to facilitate comprehensive analysis of the data and for data mining.

9.5 Evaluation of application of bioinformatics in venom research

The field of venominformatics, a marriage between bioinformatics and venom biology, is in its infancy. But already it has the potential to revolutionise the way that researchers manage venom related data and information. Venominformatics is a systematic approach that facilitates discovery of new knowledge either through direct

discovery, or through support for efficient experimentation by pre-selection of the most interesting toxin candidates. It bears the potential to expedite drug discovery process and accurate prediction of functional properties of novel toxins based on existing data. Because 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 structure-function relationships of diverse types of bioactive toxins.

Conclusion summary

The work described in this thesis is the first and pioneering work in the field of venom research to the author's knowledge. In summary, the following general conclusions can be drawn:

- Bioinformatics is a growing field in the post-genome era and is likely to shape future research in both theoretical and applied venom research. The author has demonstrated the application of bioinformatics in the large-scale study of scorpion toxins.
- The large-scale classification of 393 currently known scorpion toxins provides a global view of their structure-function relationships (described in Chapter 3).
- Inclusion of information from mutation studies of scorpion toxins, combined with multiple sequence alignment and 3D structure analyses supports extraction of functionally relevant binding motifs for scorpion toxins (described in Chapter 4).
- The author has demonstrated that combining bioinformatics and venom research can increase the efficiency of the discovery process where the

functional prediction tool can be used for support and planning of the experiments (e.g. validation of the predicted function of newly identified scorpion toxins, described in Chapter 5). This combination is a viable methodology for acquiring new knowledge in venom research where the growth of new toxin sequences identified is increasing.

- In this work, the author has created the first data warehouse of native and mutant scorpion toxin data that have been cleaned, enriched, classified and analysed, and integrated with bioinformatic tools for efficient and effective discovery of new knowledge.

9.6 Future works

There are two principal directions for the development of bioinformatics in this venom research, namely the development of data warehouses and development of functional prediction methods. The development of data warehouse includes data update on top of data collection, data cleaning and integration of bioinformatic tools. Data update focuses on adding new toxin sequences identified, new 3D structures solved and new information published in literature into the data warehouse. The new data help to verify hypotheses made during analyses of initial dataset while new information can provide insights for further analysis. Important discoveries of sequence relationships have been missed because old or incomplete databases were used (Altschul *et al*., 1994). For instance, original data submitted by sequencers into the major nucleotide sequence database GenBank/EMBL/DDBJ (Benson *et al*., 2005) are not updated unless a revision is submitted by the same group (Wu *et al*., 2003). Thus data remains uninformative though more recent knowledge is available. The author proposes to develop a system that can periodically perform data update and a

platform for domain expert to screen and discard irrelevant data so as to increase efficiency. Complete and up-to-date databases of biological knowledge are critical for information-dependent biological research (Apweiler *et al*., 2004).

The further development of functional prediction methods in venom research depends largely on the identification of biological problems that can be transformed into computational questions. The biological problem of determining the interaction surface of toxin-channel complex can be formulated into computational algorithm where data from thermodynamic mutant cycles of toxin-channel complexes could be analysed. Thermodynamic mutant cycles provide valuable information for studying energetic coupling between amino acids on the interacting surface (Hidalgo and MacKinnon, 1995) and thus the interacting amino acids can be determined not only on the toxins, but also ion channels where potential therapeutics can be developed from this information. Recently, the crystal structure of voltage-dependent K^+ channel was solved (Jiang *et al*., 2003) which provides an opportunity to allow homology modeling of other ion channels and perform computer simulation of toxin docking onto the ion channels for determining toxin-channel interacting residues.

The current sequencing of three venomous animal genomes provides an opportunity for comparative genomic analysis of toxins. Computational methods have been used to compare between genome sequences to predict protein-protein interaction based on the assumption that the genes of functionally interacting proteins tend to be associated with each other on genomes (Huynen *et al*., 2003). Promising results for deduction of specific functions for numerous proteins are obtained from comparative analysis of complete genomes by analysing phylogenetic profiles of protein families, domain fusions, gene adjacency in genomes and expression methods (Galperin and Koonin, 2000; Marcotte, 2000). The functional prediction method used in this thesis is

suitable for functional prediction of toxins that share sequence similarity. Incorporation of comparative genomes can be used to predict function of toxins that lack characterised homologues.

The author anticipates that bioinformatics will increase in importance in venom research field. Given the importance of toxins to scientific, medical and commercial applications, the number of applications stemming from venom research will be huge.

References

Aiyar, J., Withka, J.M., Rizzi, J.P., Singleton, D.H., Andrews, G.C., Lin, W., Boyd, J., Hanson, D.C., Simon, M., Dethlefs, B. *et al*., 1995. Topology of the pore-region of a K^+ channel revealed by the NMR-derived structures of scorpion toxins. Neuron 15, 1169-1181.

Alami, M., Vacher, H., Bosmans, F., Devaux, C., Rosso, J.P., Bougis, P.E., Tytgat, J., Darbon, H. and Martin-Eauclaire, M.F., 2003. Characterisation of Amm VIII from *Androctonus mauretanicus mauretanicus*: a new scorpion toxin that discriminates between neuronal and skeletal sodium channels. Biochem J 375, 551-560.

Ali, S.A., Stoeva, S., Grossmann, J.G., Abbasi, A. and Voelter, W., 2001. Purification, characterisation, and primary structure of four depressant insect-selective neurotoxin analogs from scorpion (*Buthus sindicus*) venom. Arch Biochem Biophys 391, 197-206.

Alonso, D., Khalil, Z., Satkunanthan, N. and Livett, B.G., 2003. Drugs from the sea: conotoxins as drug leads for neuropathic pain and other neurological conditions. Mini Rev Med Chem 3, 785-787.

Altschul, S.F., Boguski, M.S., Gish, W. and Wootton, J.C., 1994. Issues in searching molecular sequence databases. Nat Genet 6, 119-129.

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.

Apweiler, R., Bairoch, A. and Wu, C.H., 2004. Protein sequence databases. Curr Opin Chem Biol 8, 76-80.

Arnon, T., Potikha, T., Sher, D., Elazar, M., Mao, W., Tal, T., Bosmans, F., Tytgat, J., Ben-Arie, N. and Zlotkin, E., 2005. BjalphaIT: a novel scorpion alpha-toxin selective for insects – unique pharmacological tool. Insect Biochem Mol Biol 35, 187-195.

Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton,

G., Nordle, A., Paine, K., Taylor, P., Uddin, A. and Zygouri, C., 2003. PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res 31, 400-402.

Attwood, T.K., 2000. The quest to deduce protein function from sequence: the role of pattern databases. Int J Biochem Cell Biol 32, 139-155.

Bagdany, M., Batista, C.V., Valdez-Cruz, N.A., Somodi, S., Rodriguez de la Vega, R.C., Licea, A.F., Varga, Z., Gaspar, R., Possani, L.D. and Panyi, G., 2005. Anuroctoxin, a new scorpion toxin of the α-KTx 6 subfamily is highly selective for $K_V1.3$ over $IK_{Ca}1$ ion channels of human T lymphocytes. Mol Pharmacol 67, 1034-1044.

Bairoch, A., Boeckmann, B., Ferro, S. and Gasteiger, E., 2004. Swiss-Prot: juggling between evolution and stability. Brief Bioinform 5, 39-55.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C. and Eddy, S.R., 2004. The Pfam protein families database. Nucleic Acids Res 32, D138-D141.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L. and Sonnhammer, E.L., 2000. The Pfam protein families database. Nucleic Acids Res 28, 263-266.

Batista, C.V., del Pozo, L., Zamudio, F.Z., Contreras, S., Becerril, B., Wanke, E. and Possani, L.D., 2004. Proteomics of the venom from the Amazonian scorpion *Tityus cambridgei* and the role of prolines on mass spectrometry analysis of toxins. J Chromatogr B Analyt Technol Biomed Life Sci 803, 55-66.

Batista, C.V., Gomez-Lagunas, F., Rodriguez de la Vega, R.C., Hajdu, P., Panyi, G., Gaspar, R. and Possani, L.D., 2002a. Two novel toxins from the Amazonian scorpion *Tityus cambridgei* that block $K_V1.3$ and Shaker B K^+-channels with distinctly different affinities. Biochim Biophys Acta 1601, 123-131.

Batista, C.V., Zamudio, F.Z., Lucas, S., Fox, J.W., Frau, A., Prestipino, G. and Possani, L.D., 2002b. Scorpion toxins from *Tityus cambridgei* that affect Na^+-channels. Toxicon 40, 557-562.

Becerril, B., Corona, M., Coronas, F.I., Zamudio, F., Calderon-Aranda, E.S., Fletcher, P.L., Jr. Martin, B.M. and Possani, L.D., 1996. Toxic peptides and genes encoding toxin

gamma of the Brazilian scorpions *Tityus bahiensis* and *Tityus stigmurus*. Biochem J 313, 753-760.

Becerril, B., Marangoni, S. and Possani, L.D., 1997. Toxins and genes isolated from scorpions of the genus *Tityus*. Toxicon 35, 821-835.

Benkhadir, K., Kharrat, R., Cestele, S., Mosbah, A., Rochat, H., El Ayeb, M. and Karoui, H., 2004. Molecular cloning and functional expression of the alpha-scorpion toxin BotIII: pivotal role of the C-terminal region for its interaction with voltage-dependent sodium channels. Peptides 25, 151-161.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L., 2005. GenBank. Nucleic Acids Res 33, D34-D38.

Blank, T., Nijholt, I., Kye, M.J. and Spiess, J., 2004. Small conductance Ca^{2+}-activated K^{+} channels as targets of CNS drug development. Curr Drug Targets CNS Neurol Disord 3, 161-167.

Bontems, F., Roumestand, C., Gilquin, B., Menez, A. and Toma, F., 1991. Refined structure of charybdotoxin: common motifs in scorpion toxins and insect defensins. Science 254, 1521-1523.

Borges, A., Alfonzo, M.J., Garcia, C.C., Winand, N.J., Leipold, E. and Heinemann, S.H., 2004. Isolation, molecular cloning and functional characterisation of a novel beta-toxin from the Venezuelan scorpion, *Tityus zulianus*. Toxicon 43, 671-684.

Bradbury, J., 2003. Death, where is thy sting? Drug Discov Today 8, 1099.

Briggs, J.C., 1987. Antitropical distribution and evolution in the Indo-West Pacific Ocean. Syst Zool 36, 237-247.

Brown, N.P., Leroy, C. and Sander, C., 1998. MView: a web-compatible database search or multiple alignment viewer. Bioinformatics 14, 380-381.

Brusic, V., Zeleznikow, J. and Petrovsky, N., 2000. Molecular immunology databases and data repositories. J Immunol Methods 238, 17-28.

Cai, Z., Xu, C., Xu, Y., Lu, W., Chi, C.W., Shi, Y. and Wu, J., 2004. Solution structure of BmBKTx1, a new BK_{Ca}1 channel blocker from the Chinese scorpion *Buthus martensi*

Karsch. Biochem 43, 3764-3771.

Catterall, W.A., 2002. Molecular mechanisms of gating and drug block of sodium channels. Novartis Found Symp 241, 206-218; discussion 218-232.

Ceard, B., Martin-Eauclaire, M. and Bougis, P.E., 2001. Evidence for a position-specific deletion as an evolutionary link between long- and short-chain scorpion toxins. FEBS Lett 494, 246-248.

Chagot, B., Pimentel, C., Dai, L., Pil, J., Tytgat, J., Nakajima, T., Corzo, G., Darbon, H. and Ferrat, G., 2005. An unusual fold for potassium channel blockers: NMR structure of three toxins from the scorpion *Opisthacanthus madagascariensis*. Biochem J 388, 263-271.

Chang, L.S., Chu, Y.P., Cheng, Y.C., Liou, J.C. and Yang, C.C., 2005. Lys-64 of the A chain is involved in the enzymatic activity and neurotoxic effect of beta-bungarotoxin. Toxicon 45, 179-185.

Chang, L.S., Wu, P.F., Liou, J.C., Chiang-Lin, W.H. and Yang, C.C., 2004. Chemical modification of arginine residues of *Notechis scutatus scutatus* notexin. Toxicon 44, 491-497.

Chen, S.W. and Pellequer, J.L., 2004. Identification of functionally important residues in proteins using comparative models. Curr Med Chem 11, 595-605.

Chen, T., Folan, R., Kwok, H., O'Kane, E.J., Bjourson, A.J. and Shaw, C., 2003. Isolation of scorpion (*Androctonus amoreuxi*) putative alpha neurotoxins and parallel cloning of their respective cDNAs from a single sample of venom. Regul Pept 115, 115-121.

Chen, T., Walker, B., Zhou, M. and Shaw, C., 2005. Molecular cloning of a novel putative potassium channel-blocking neurotoxin from the venom of the North African scorpion, *Androctonus amoreuxi*. Peptides 26, 731-736.

Chioato, L. and Ward, R.J., 2003. Mapping structural determinants of biological activities in snake venom phospholipases A2 by sequence analysis and site directed mutagenesis. Toxicon 42, 869-883.

Chuang, R.S., Jaffe, H., Cribbs, L., Perez-Reyes, E. and Swartz, K.J., 1998. Inhibition of T-

type voltage-gated calcium channels by a new scorpion toxin. Nat Neurosci 1, 668-674.

Church, J.E. and Hodgson, W.C., 2002. The pharmacological activity of fish venoms. Toxicon 40, 1083-1093.

Cline, M., Hughey, R. and Karplus, K., 2002. Predicting reliable regions in protein sequence alignments. Bioinformatics 18, 306-314.

Cochrane, G., Aldebert, P., Althorpe, N., Andersson, M., Baker, W., Baldwin, A., Bates, K., Bhattacharyya, S., Browne, P., van den Broek, A., 2006. EMBL Nucleotide Sequence Database: developments in 2005. Nucleic Acids Res 34, D10-D15.

Cohen, L., Karbat, I., Gilles, N., Froy, O., Corzo, G., Angelovici, R., Gordon, D. and Gurevitz, M., 2004. Dissection of the functional surface of an anti-insect excitatory toxin illuminates a putative "hot spot" common to all scorpion beta-toxins affecting Na^+ channels. J Biol Chem 279, 8206-8211.

Cohen, L., Karbat, I., Gilles, N., Ilan, N., Benveniste, M., Gordon, D. and Gurevitz, M., 2005. Common features in the functional surface of scorpion β-toxins and elements that confer specificity for insect and mammalian voltage-gated sodium channels. J Biol Chem 280, 5045-5053.

Collins, F.S., Morgan, M. and Patrinos, A., 2003. The Human Genome Project: lessons from large-scale biology. Science 300, 286-290.

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.

Corona, M., Coronas, F.V., Merino, E., Becerril, B., Gutierrez, R., Rebolledo-Antunez, S., Garcia, D.E. and Possani, L.D., 2003. A novel class of peptide found in scorpion venom with neurodepressant effects in peripheral and central nervous system of the rat. Biochim Biophys Acta 1649, 58-67.

Corona, M., Gurrola, G.B., Merino, E., Cassulini, R.R., Valdez-Cruz, N.A., Garcia, B., Ramirez-Dominguez, M.E., Coronas, F.I., Zamudio, F.Z., Wanke, E. and Possani,

L.D., 2002. A large number of novel Ergtoxin-like genes and ERG K^+-channels blocking peptides from scorpions of the genus *Centruroides*. FEBS Lett 532, 121-126.

Corona, M., Valdez-Cruz, N.A., Merino, E., Zurita, M. and Possani, L.D., 2001. Genes and peptides from the scorpion *Centruroides sculpturatus* Ewing, that recognize Na^+-channels. Toxicon 39, 1893-1898.

Coronas, F.V., de Roodt, A.R., Portugal, T.O., Zamudio, F.Z., Batista, C.V., Gomez-Lagunas, F. and Possani, L.D., 2003a. Disulfide bridges and blockage of Shaker B K^+-channels by another butantoxin peptide purified from the Argentinean scorpion *Tityus trivittatus*. Toxicon 41, 173-179.

Coronas, F.V., Stankiewicz, M., Batista, C.V., Giraud, S., Alam, J.M., Possani, L.D., Mebs, D. and Pelhate, M., 2003b. Primary structure and electrophysiological characterisation of two almost identical isoforms of toxin from *Isometrus vittatus* (family: Buthidae) scorpion venom. Toxicon 41, 989-997.

Corzo, G., Villegas, E. and Nakajima, T., 2001. Isolation and structural characterisation of a peptide from the venom of scorpion with toxicity towards invertebrates and vertebrates. Protein Pept Lett 8, 385-393.

Couraud, F., Jover, E., Dubois, J.M. and Rochat, H., 1982. Two types of scorpion receptor sites, one related to the activation, the other to the inactivation of the action potential sodium channel. Toxicon 20, 9-16.

Cover, T. and Hart, P., 1967. Nearest neighbour pattern classification. IEEE Transactions on Information Theory 13, 21-27.

Curran, M.E., 1998. Potassium ion channels and human disease: phenotypes to drug targets? Curr Opin Biotechnol 9, 565-572.

Dalton, S., Gerzanich, V., Chen, M., Dong, Y., Shuba, Y. and Simard, J.M., 2003. Chlorotoxin-sensitive Ca^{2+}-activated Cl^- channel in type R2 reactive astrocytes from adult rat brain. Glia 42, 325-339.

Dasarathy, B.V., Ed., 1991. Nearest neighbour (NN) Norms: NN pattern classification techniques. IEEE Computer Society Press. Los Alamitos, Calif.

Dauplais, M., Lecoq, A., Song, J., Cotton, J., Jamin, N., Gilquin, B., Roumestand, C., Vita, C., de Medeiros, C.L., Rowan *et al*., 1997. On the convergent evolution of animal toxins. Conservation of a diad of functional residues in potassium channel-blocking toxins with unrelated structures. J Biol Chem 272, 4302-4309.

David, R.M., Krishna, N.R. and Watt, D.D., 1991. Characterisation of cationic binding sites of neurotoxins from venom of the scorpion (*Centruroides sculpturatus* Ewing) using lanthanides as binding probes. Toxicon 29, 645-662.

Davies, N.W., Wiese, M.D. and Brown, S.G., 2004. Characterisation of major peptides in 'jack jumper' ant venom by mass spectrometry. Toxicon 43, 173-183.

del Rio-Portilla, F., Hernandez-Marin, E., Pimienta, G., Coronas, F.V., Zamudio, F.Z., Rodriguez de la Vega, R.C., Wanke, E. and Possani, L.D., 2004. NMR solution structure of Cn12, a novel peptide from the Mexican scorpion *Centruroides noxius* with a typical β-toxin sequence but with α-like physiological activity. Eur J Biochem 271, 2504-2516.

Deshpande, N., Addess, K.J., Bluhm, W.F., Merino-Ott, J.C., Townsend-Merino, W., Zhang, Q., Knezevich, C., Xie, L., Chen, L., Feng, Z. *et al*., 2005. The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. Nucleic Acids Res 33, D233-D237.

Deutsch, C., Price, M., Lee, S., King, V.F. and Garcia, M.L., 1991. Characterisation of high affinity binding sites for charybdotoxin in human T lymphocytes. Evidence for association with the voltage-gated K^+ channel. J Biol Chem 266, 3668-3674.

Devaux, J., Beeton, C., Beraud, E. and Crest, M. 2004. Ion channels and demyelination: basis of a treatment of experimental autoimmune encephalomyelitis (EAE) by potassium channel blockers. Rev Neurol (Paris) 160, S16-S27.

Dhawan, R., Varshney, A., Mathew, M.K. and Lala, A.K., 2003. BTK-2, a new inhibitor of the $K_V1.1$ potassium channel purified from Indian scorpion *Buthus tamulus*. FEBS Lett 539, 7-13.

Dhawan, R.D., Joseph, S., Sethi, A. and Lala, A.K., 2002. Purification and characterisation of

a short insect toxin from the venom of the scorpion *Buthus tamulus*. FEBS Lett 528, 261-266.

Ding, L., Sabo, A., Berkowicz, N., Meyer, R.R., Shotland, Y., Johnson, M.R., Pepin, K.H., Wilson, R.K. and Spieth, J., 2004. EAnnot: a genome annotation tool using experimental evidence. Genome Res 14, 2503-2509.

Dobson, P.D., Cai, Y.D., Stapley, B.J. and Doig, A.J., 2004. Prediction of protein function in the absence of significant sequence similarity. Curr Med Chem 11, 2135-2142.

D'Suze, G., Batista, C.V., Frau, A., Murgia, A.R., Zamudio, F.Z., Sevcik, C., Possani, L.D. and Prestipino, G., 2004a. Discrepin, a new peptide of the sub-family α-KTX15, isolated from the scorpion *Tityus discrepans* irreversibly blocks K^+-channels (IA currents) of cerebellum granular cells. Arch Biochem Biophys 430, 256-263.

D'Suze, G., Sevcik, C., Corona, M., Zamudio, F.Z., Batista, C.V., Coronas, F.I. and Possani, L.D., 2004b. Ardiscretin a novel arthropod-selective toxin from *Tityus discrepans* scorpion venom. Toxicon 43, 263-272.

Dudina, E.E., Korolkova, Y.V., Bocharova, N.E., Koshelev, S.G., Egorov, T.A., Huys, I., Tytgat, J. and Grishin, E.V., 2001. OsK2, a new selective inhibitor of $K_V1.2$ potassium channels purified from the venom of the scorpion *Orthochirus scrobiculosus*. Biochem Biophys Res Commun 286, 841-847.

Dutertre, S., Nicke, A., Tyndall, J.D. and Lewis, R.J., 2004. Determination of alpha-conotoxin binding modes on neuronal nicotinic acetylcholine receptors. J Mol Recognit 17, 339-347.

Escoubas, P., Stankiewicz, M., Takaoka, T., Pelhate, M., Romi-Lebrun, R., Wu, F.Q. and Nakajima, T., 2000. Sequence and electrophysiological characterisation of two insect-selective excitatory toxins from the venom of the Chinese scorpion *Buthus martensi*. FEBS Lett 483, 175-180.

Everhart, D., Cartier, G.E., Malhotra, A., Gomes, A.V., McIntosh, J.M. and Luetje, C.W., 2004. Determinants of potency on alpha-conotoxin MII, a peptide antagonist of neuronal nicotinic receptors. Biochem 43, 2732-2737.

Fajloun, Z., Kharrat, R., Chen, L., Lecomte, C., Di Luccio, E., Bichet, D., El Ayeb, M., Rochat, H., Allen, P.D., Pessah, I.N. *et al*., 2000. Chemical synthesis and characterisation of maurocalcine, a scorpion toxin that activates Ca^{2+} release channel/ryanodine receptors. FEBS Lett 469, 179-185.

Fielden, M.R., Matthews, J.B., Fertuck, K.C., Halgren, R.G. and Zacharewski, T.R., 2002. *In silico* approaches to mechanistic and predictive toxicology: an introduction to bioinformatics for toxicologists. Crit Rev Toxicol 32, 67-112.

Fishelson, M. and Geiger, D., 2004. Optimizing exact genetic linkage computations. J Comput Biol 11, 263-275.

Frenal, K., Xu, C.Q., Wolff, N., Wecker, K., Gurrola, G.B., Zhu, S.Y., Chi, C.W., Possani, L.D., Tytgat, J. and Delepierre, M., 2004. Exploring structural features of the interaction between the scorpion toxin CnErg1 and ERG K^+ channels. Proteins 56, 367-375.

Froy, O., Sagiv, T., Poreh, M., Urbach, D., Zilberberg, N. and Gurevitz, M., 1999. Dynamic diversification from a putative common ancestor of scorpion toxins affecting sodium, potassium, and chloride channels. J Mol Evol 48, 187-196.

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.

Fry, B.G. and Wuster, W., 2004. Assembling an arsenal: origin and evolution of the snake venom proteome inferred from phylogenetic analysis of toxin sequences. Mol Biol Evol 21, 870-883.

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.

Fuller, M.D., Zhang, Z.R., Cui, G., Kubanek, J. and McCarty, N.A., 2004. Inhibition of CFTR channels by a peptide toxin of scorpion venom. Am J Physiol Cell Physiol 287, C1328-1341.

Gagna, C.E., Winokur, D. and Clark Lambert, W., 2004. Cell biology, chemogenomics and chemoproteomics. Cell Biol Int 28, 755-764.

Galperin, M.Y., 2005. The Molecular Biology Database Collection: 2005 update. Nucleic Acids Res 33, D5-D24.

Galperin, M.Y. and Koonin, E.V., 2000. Who's your neighbour? New computational approaches for functional genomics. Nat Biotechnol 18, 609-613.

Galvez, A., Gimenez-Gallego, G., Reuben, J.P., Roy-Contancin, L., Feigenbaum, P., Kaczorowski, G.J. and Garcia, M.L., 1990. Purification and characterisation of a unique, potent, peptidyl probe for the high conductance calcium-activated potassium channel from venom of the scorpion *Buthus tamulus*. J Biol Chem 265, 11083-11090.

Gao, Y.D. and Garcia, M.L., 2003 Interaction of agitoxin2, charybdotoxin, and iberiotoxin with potassium channels: selectivity between voltage-gated and Maxi-K channels. Proteins 52, 146-154.

Gattiker, A., Gasteiger, E. and Bairoch, A., 2002. ScanProsite: a reference implementation of a PROSITE scanning tool. Appl Bioinformatics 1, 107-108.

Garcia, M.L., Garcia-Calvo, M., Hidalgo, P., Lee, A. and MacKinnon, R., 1994. Purification and characterisation of three inhibitors of voltage-dependent K^+ channels from *Leiurus quinquestriatus var. hebraeus* venom. Biochem 33, 6834-6839.

Gazarian, K.G., Gazarian, T., Hernandez, R. and Possani, L.D. 2005. Immunology of scorpion toxins and perspectives for generation of anti-venom vaccines. Vaccine 23, 3357-3368.

Giangiacomo, K.M., Ceralde, Y. and Mullmann, T.J., 2004. Molecular basis of alpha-KTx specificity. Toxicon 43, 877-886.

Gilles, N., Blanchet, C., Shichor, I., Zaninetti, M., Lotan, I., Bertrand, D. and Gordon, D., 1999. A scorpion alpha-like toxin that is active on insects and mammals reveals an unexpected specificity and distribution of sodium channel subtypes in rat brain neurons. J Neurosci 19, 8730-8739.

Gilles, N., Gurevitz, M. and Gordon, D., 2003. Allosteric interactions among pyrethroid,

brevetoxin, and scorpion toxin receptors on insect sodium channels raise an alternative approach for insect control. FEBS Lett 540, 81-85.

Gilquin, B., Racape, J., Wrisch, A., Visan, V., Lecoq, A., Grissmer, S., Menez, A. and Gasparini, S., 2002. Structure of the BgK-K_V1.1 complex based on distance restraints identified by double mutant cycles. Molecular basis for convergent evolution of K_V1 channel blockers. J Biol Chem 277, 37406-37413.

Gohlke, H. and Klebe, G., 2002. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. Angew Chem Int Ed Engl 41, 2644-2676.

Goldstein, S.A. and Miller, C., 1993. Mechanism of charybdotoxin block of a voltage-gated K^+ channel. Biophys J 65, 1613-1619.

Gordon, D. and Gurevitz, M., 2003. The selectivity of scorpion alpha-toxins for sodium channel subtypes is determined by subtle variations at the interacting surface. Toxicon 41, 125-128.

Gordon, D., Ilan, N., Zilberberg, N., Gilles, N., Urbach, D., Cohen, L., Karbat, I., Froy, O., Gaathon, A., Kallen, R.G. *et al*., 2003. An 'Old World' scorpion beta-toxin that recognizes both insect and mammalian sodium channels. Eur J Biochem 270, 2663-2670.

Gordon, D., Martin-Eauclaire, M.F., Cestele, S., Kopeyan, C., Carlier, E., Khalifa, R.B., Pelhate, M. and Rochat, H., 1996. Scorpion toxins affecting sodium current inactivation bind to distinct homologous receptor sites on rat brain and insect sodium channels. J Biol Chem 271, 8034-8045.

Gordon, D., Savarin, P., Gurevitz, M. and Zinn-Justin, S., 1998. Functional anatomy of scorpion toxins affecting sodium channels. J Toxicol Toxin Rev 17, 131-159.

Gottlieb, P.A., Suchyna, T.M., Ostrow, L.W. and Sachs, F., 2004. Mechanosensitive ion channels as drug targets. Curr Drug Targets CNS Neurol Disord 3, 287-295.

Goudet, C., Chi, C.W. and Tytgat, J., 2002. An overview of toxins and genes from the venom of the Asian scorpion *Buthus martensi* Karsch. Toxicon 40, 1239-1258.

Gracy, J. and Argos, P., 1998. DOMO: a new database of aligned protein domains. Trends Biochem Sci 23, 495-497.

Grant, M.A., Morelli, X.J. and Rigby, A.C., 2004. Conotoxins and structural biology: a prospective paradigm for drug discovery. Curr Protein Pept Sci 5, 235-248.

Guan, R., Wang, C.G., Wang, M. and Wang, D.C., 2001. A depressant insect toxin with a novel analgesic effect from scorpion *Buthus martensii* Karsch. Biochim Biophys Acta 1549, 9-18.

Guan, R.J., Xiang, Y., He, X.L., Wang, C.G., Wang, M., Zhang, Y., Sundberg, E.J. and Wang, D.C., 2004. Structural mechanism governing *cis* and *trans* isomeric states and an intramolecular switch for *cis*/*trans* isomerization of a non-proline peptide bond observed in crystal structures of scorpion toxins. J Mol Biol 341, 1189-1204.

Hains, P.G., Ramsland, P.A. and Broady, K.W., 1999. Modeling of acanthoxin A1, a PLA2 enzyme from the venom of the common death adder (*Acanthophis antarcticus*). Proteins 35, 80-88.

Hamon, A., Gilles, N., Sautiere, P., Martinage, A., Kopeyan, C., Ulens, C., Tytgat, J., Lancelin, J.M. and Gordon, D., 2002. Characterisation of scorpion alpha-like toxin group using two new toxins from the scorpion *Leiurus quinquestriatus hebraeus*. Eur J Biochem 269, 3920-3933.

Harrison, R.A., 2004. Development of venom toxin-specific antibodies by DNA immunisation: rationale and strategies to improve therapy of viper envenoming. Vaccine 22, 1648-1655.

Hassani, O., Mansuelle, P., Cestele, S., Bourdeaux, M., Rochat, H. and Sampieri, F., 1999. Role of lysine and tryptophan residues in the biological activity of toxin VII (Ts gamma) from the scorpion *Tityus serrulatus*. Eur J Biochem 260, 76-86.

He, X.L., Li, H.M., Zeng, Z.H., Liu, X.Q., Wang, M. and Wang, D.C., 1999. Crystal structures of two alpha-like scorpion toxins: non-proline *cis* peptide bonds and implications for new binding site selectivity on the sodium channel. J Mol Biol 292, 125-135.

He, Y.Y., Lee, W.H. and Zhang, Y., 2004. Cloning and purification of alpha-neurotoxins from

king cobra (*Ophiophagus hannah*). Toxicon 44, 295-303.

Hidalgo, P. and MacKinnon, R., 1995. Revealing the architecture of a K^+ channel pore through mutant cycles with a peptide inhibitor. Science 268, 307-310.

Hulo, N., Sigrist, C.J., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P. and Bairoch, A., 2004. Recent improvements to the PROSITE database. Nucleic Acids Res 32, D134-D137.

Huynen, M.A., Snel, B., von Mering, C. and Bork, P., 2003. Function prediction and protein networks. Curr Opin Cell Biol 15, 191-198.

Huys, I., Olamendi-Portugal, T., Garcia-Gomez, B.I., Vandenberghe, I., Van Beeumen, J., Dyason, K., Clynen, E., Zhu, S., van der Walt, J., Possani, L.D. *et al*., 2004. A subfamily of acidic alpha-K^+ toxins. J Biol Chem 279, 2781-2789.

Huys, I. and Tytgat, J., 2003. Evidence for a function-specific mutation in the neurotoxin, parabutoxin 3. Eur J Neurosci 17, 1786-1792.

Inceoglu, A.B., Hayashida, Y., Lango, J., Ishida, A.T. and Hammock, B.D., 2002. A single charged surface residue modifies the activity of ikitoxin, a beta-type Na^+ channel toxin from *Parabuthus transvaalicus*. Eur J Biochem 269, 5369-5376.

Inceoglu, B., Lango, J., Pessah, I.N. and Hammock, B.D., 2005. Three structurally related, highly potent, peptides from the venom of *Parabuthus transvaalicus* possess divergent biological activity. Toxicon 45, 727-733.

Ivanovski, G., Petan, T., Krizaj, I., Gelb, M.H., Gubensek, F. and Pungercar, J., 2004. Basic amino acid residues in the beta-structure region contribute, but not critically, to presynaptic neurotoxicity of ammodytoxin A. Biochim Biophys Acta 1702, 217-225.

Jiang, Y., Lee, A., Chen, J., Ruta, V., Cadene, M., Chait, B.T. and MacKinnon, R., 2003. X-ray structure of a voltage-dependent K^+ channel. Nature 423, 33-41.

Jiqun, S., Xiuling, X., Zhijian, C., Wanhong, L., Yingliang, W., Shunyi, Z., Xianchun, Z., Dahe, J., Xin, M., Hui, L. *et al*., 2004. Molecular cloning, genomic organization and functional characterisation of a new short-chain potassium channel toxin-like peptide BmTxKS4 from *Buthus martensii* Karsch (BmK). J Biochem Mol Toxicol 18, 187-

195.

Jouirou, B., Mosbah, A., Visan, V., Grissmer, S., M'Barek, S., Fajloun, Z., Van Rietschoten, J., Devaux, C., Rochat, H., Lippens, G. *et al*., 2004. Cobatoxin 1 from *Centruroides noxius* scorpion venom: chemical synthesis, three-dimensional structure in solution, pharmacology and docking on K^+ channels. Biochem J 377, 37-49.

Jover, E., Bablito, J. and Couraud, F., 1984. Binding of beta-scorpion toxin: a physicochemical study. Biochem 23, 1147-1152.

Kapetanovic, I.M., Rosenfeld, S. and Izmirlian, G., 2004. Overview of commonly used bioinformatics methods and their applications. Ann N Y Acad Sci 1020, 10-21.

Karasavvas, K.A., Baldock, R. and Burger, A., 2004. Bioinformatics integration and agent technology. J Biomed Inform 37, 205-219.

Karbat, I., Cohen, L., Gilles, N., Gordon, D. and Gurevitz, M., 2004a. Conversion of a scorpion toxin agonist into an antagonist highlights an acidic residue involved in voltage sensor trapping during activation of neuronal Na^+ channels. Faseb J 18, 683-689.

Karbat, I., Frolow, F., Froy, O., Gilles, N., Cohen, L., Turkov, M., Gordon, D. and Gurevitz, M., 2004b. Molecular basis of the high insecticidal potency of scorpion alpha-toxins. J Biol Chem 279, 31679-31686.

Karp, P.D., Paley, S. and Zhu, J., 2001. Database verification studies of SWISS-PROT and GenBank. Bioinformatics 17, 526-532.

Kini, R.M., 2002. Molecular moulds with multiple missions: functional sites in three-finger toxins. Clin Exp Pharmacol Physiol 29, 815-822.

Kobayashi, Y., Takashima, H., Tamaoki, H., Kyogoku, Y., Lambert, P., Kuroda, H., Chino, N., Watanabe, T.X., Kimura, T., Sakakibara, S. *et al.*, 1991. The cystine-stabilized alpha-helix: a common structural motif of ion-channel blocking neurotoxic peptides. Biopolymers 31, 1213-1220.

Koh, J.L., Krishnan, S.P.T., Hong, S.H., Tan, P.T., Khan, A.M., Lee, M.L. and Brusic, V., 2004a. Bioware: A framework for bioinformatics data retrieval, annotation and

publishing. ACM SIGIR Workshop on Search and Discovery in Bioinformatics, Sheffield, UK.

Koh, J.L.Y., Lee, M.L. and Brusic, V., 2004b. A classification of biological data artifacts. ICDT Workshop on Database Issues in Biological Databases (DBiBD), 8-9th January 2005, Edinburgh, Scotland, UK, 53-57.

Kohling, R., 2002. Voltage-gated sodium channels in epilepsy. Epilepsia 43, 1278-1295.

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-W359.

Kordis, D. and Gubensek, F., 2000. Adaptive evolution of animal toxin multigene families. Gene 261, 43-52.

Korn, S.J. and Trapani, J.G., 2005. Potassium channels. IEEE Trans Nanobioscience 4, 21-33.

Korolkova, Y.V., Bocharov, E.V., Angelo, K., Maslennikov, I.V., Grinenko, O.V., Lipkin, A.V., Nosyreva, E.D., Pluzhnikov, K.A., Olesen, S.P., Arseniev, A.S. *et al*., 2002. New binding site on common molecular scaffold provides HERG channel specificity of scorpion toxin BeKm-1. J Biol Chem 277, 43104-43109.

Korolkova, Y.V., Kozlov, S.A., Lipkin, A.V., Pluzhnikov, K.A., Hadley, J.K., Filippov, A.K., Brown, D.A., Angelo, K., Strobaek, D., Jespersen, T. *et al*., 2001. An ERG channel inhibitor from the scorpion *Buthus eupeus*. J Biol Chem 276, 9868-9876.

Korolkova, Y.V., Tseng, G.N. and Grishin, E.V., 2004. Unique interaction of scorpion toxins with the hERG channel. J Mol Recognit 17, 209-217.

Koschak, A., Bugianesi, R.M., Mitterdorfer, J., Kaczorowski, G.J., Garcia, M.L. and Knaus, H.G., 1998. Subunit composition of brain voltage-gated potassium channels determined by hongotoxin-1, a novel peptide derived from *Centruroides limbatus* venom. J Biol Chem 273, 2639-2644.

Kumar, S., Tamura, K. and Nei, M., 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 5, 150-163.

Kyte, J. and Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of

a protein. J Mol Biol 157, 105-132.

Lacinova, L., 2004. Pharmacology of recombinant low-voltage activated calcium channels. Curr Drug Targets CNS Neurol Disord 3, 105-111.

Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M., 1993. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 26, 283-291.

Lassmann, T. and Sonnhammer, E.L., 2002. Quality assessment of multiple alignment programs. FEBS Lett 529, 126-130.

Lecomte, C., Ferrat, G., Fajloun, Z., Van Rietschoten, J., Rochat, H., Martin-Eauclaire, M.F., Darbon, H. and Sabatier, J.M., 1999. Chemical synthesis and structure-activity relationships of Ts kappa, a novel scorpion toxin acting on apamin-sensitive SK channel. J Pept Res 54, 369-376.

Legros, C., Ceard, B., Vacher, H., Marchot, P., Bougis, P.E. and Martin-Eauclaire, M.F., 2005. Expression of the standard scorpion alpha-toxin AaH II and AaH II mutants leading to the identification of some key bioactive elements. Biochim Biophys Acta 1723, 91-99.

Legros, C., Bougis, P.E. and Martin-Eauclaire, M.F., 2003. Characterisation of the genes encoding Aa1 isoforms from the scorpion *Androctonus australis*. Toxicon 41, 115-119.

Legros, C., Ceard, B., Bougis, P.E. and Martin-Eauclaire, M.F., 1998. Evidence for a new class of scorpion toxins active against K^+ channels. FEBS Lett 431, 375-380.

Lewis, R.J., 2004. Conotoxins as selective inhibitors of neuronal ion channels, receptors and transporters. IUBMB Life 56, 89-93.

Lewis, R.J. and Garcia, M.L., 2003. Therapeutic potential of venom peptides. Nat Rev Drug Discov 2, 790-802.

Li, R.A. and Tomaselli, G.F., 2004. Using the deadly mu-conotoxins as probes of voltage-gated sodium channels. Toxicon 44, 117-122.

Liu, H.L. and Lin, J.C., 2004. Molecular docking of the scorpion toxin Tc1 to the structural model of the voltage-gated potassium channel $K_V1.1$ from human *Homo sapiens*. J

Biomol Struct Dyn 21, 639-650.

Liu, L.H., Bosmans, F., Maertens, C., Zhu, R.H., Wang, D.C. and Tytgat, J., 2005. Molecular basis of the mammalian potency of the scorpion alpha-like toxin, BmK M1. Faseb J 19, 594-596.

Liu, Z., Yang, G., Li, B., Chi, C. and Wu, X., 2003. Cloning, co-expression with an amidating enzyme, and activity of the scorpion toxin BmK ITa1 cDNA in insect cells. Mol Biotechnol 24, 21-26.

Lopez-Gonzalez, I., Olamendi-Portugal, T., De la Vega-Beltran, J.L., Van der Walt, J., Dyason, K., Possani, L.D., Felix, R. and Darszon, A., 2003. Scorpion toxins that block T-type Ca^{2+} channels in spermatogenic cells inhibit the sperm acrosome reaction. Biochem Biophys Res Commun 300, 408-414.

Loret, E.P., Mansuelle, P., Rochat, H. and Granier, C., 1990. Neurotoxins active on insects: amino acid sequences, chemical modifications, and secondary structure estimation by circular dichroism of toxins from the scorpion *Androctonus australis* Hector. Biochem 29, 1492-1501.

Lourenco, W.R., 1994. Diversity and endemism in tropical versus temperate scorpion communities. Biogiographica 70, 155-160.

Lu, Z. and MacKinnon, R., 1997. Purification, characterisation, and synthesis of an inward-rectifier K^+ channel inhibitor from scorpion venom. Biochem 36, 6936-6940.

Maertens, C., Wei, L., Tytgat, J., Droogmans, G. and Nilius, B., 2000. Chlorotoxin does not inhibit volume-regulated, calcium-activated and cyclic AMP-activated chloride channels. Br J Pharmacol 129, 791-801.

Mahjoubi-Boubaker, B., Crest, M., Khalifa, R.B., Ayeb, M.E. and Kharrat, R., 2004. Kbot1, a three disulfide bridges toxin from *Buthus occitanus tunetanus* venom highly active on both SK and K_V channels. Peptides 25, 637-645.

Marcotte, E.M., 2000. Computational genetics: finding protein function by nonhomology methods. Curr Opin Struct Biol 10, 359-365.

Martin-Eauclaire, M.F. and Couraud, F., 1995. Scorpion neurotoxins: effects and mechanisms.

In: Chang, L.W., Dyer, R.S. (Eds.), Handbook of Neurotoxicology, Marcell and Dekker, New York, pp. 683–716.

Maslennikov, I.V., Shenkarev, Z.O., Zhmak, M.N., Ivanov, V.T., Methfessel, C., Tsetlin, V.I. and Arseniev, A.S., 1999. NMR spatial structure of alpha-conotoxin ImI reveals a common scaffold in snail and snake toxins recognizing neuronal nicotinic acetylcholine receptors. FEBS Lett 444, 275-280.

McPherson, A., 1999. Crystallization of biological macromolecules. New York, Cold Spring Harbor Laboratory Press.

Mejri, T., Borchani, L., Srairi-Abid, N., Benkhalifa, R., Cestele, S., Regaya, I., Karoui, H., Pelhate, M., Rochat, H. and El Ayeb, M., 2003. BotIT6: a potent depressant insect toxin from *Buthus occitanus tunetanus* venom. Toxicon 41, 163-171.

Meki, A., Mansuelle, P., Laraba-Djebari, F., Oughideni, R., Rochat, H. and Martin-Eauclaire, M.F., 2000. KTX3, the kaliotoxin from *Buthus occitanus tunetanus* scorpion venom: one of an extensive family of peptidyl ligands of potassium channels. Toxicon 38, 105-111.

Miljanich, G.P., 2004. Ziconotide: neuronal calcium channel blocker for treating severe chronic pain. Curr Med Chem 11, 3029-3040.

Moreno-Murciano, M.P., Monleon, D., Calvete, J.J., Celda, B. and Marcinkiewicz, C., 2003. Amino acid sequence and homology modeling of obtustatin, a novel non-RGD-containing short disintegrin isolated from the venom of *Vipera lebetina obtusa*. Protein Sci 12, 366-371.

Mouhat, S., Jouirou, B., Mosbah, A., De Waard, M. and Sabatier, J.M., 2004a. Diversity of folds in animal toxins acting on ion channels. Biochem J 378, 717-726.

Mouhat, S., Mosbah, A., Visan, V., Wulff, H., Delepierre, M., Darbon, H., Grissmer, S., De Waard, M. and Sabatier, J.M., 2004b. The 'functional' dyad of scorpion toxin Pi1 is not itself a prerequisite for toxin binding to the voltage-gated $K_V1.2$ potassium channels. Biochem J 377, 25-36.

Mouhat, S., Visan, V., Ananthakrishnan, S., Wulff, H., Andreotti, N., Grissmer, S., Darbon,

H., De Waard, M. and Sabatier, J.M., 2005. K^+ channel types targeted by synthetic OSK1, a toxin from *Orthochirus scrobiculosus* scorpion venom. Biochem J 385, 95-104.

Mourier, G., Dutertre, S., Fruchart-Gaillard, C., Menez, A. and Servent, D., 2003. Chemical synthesis of MT1 and MT7 muscarinic toxins: critical role of Arg-34 in their interaction with M1 muscarinic receptor. Mol Pharmacol 63, 26-35.

Murgia, A.R., Batista, C.V., Prestipino, G. and Possani, L.D., 2004. Amino acid sequence and function of a new alpha-toxin from the Amazonian scorpion *Tityus cambridgei*. Toxicon 43, 737-740.

Nastainczyk, W., Meves, H. and Watt, D.D., 2002. A short-chain peptide toxin isolated from *Centruroides sculpturatus* scorpion venom inhibits ether-a-go-go-related gene K^+ channels. Toxicon 40, 1053-1058.

Nirthanan, S., Pil, J., Abdel-Mottaleb, Y., Sugahara, Y., Gopalakrishnakone, P., Joseph, J.S., Sato, K. and Tytgat, J., 2005. Assignment of voltage-gated potassium channel blocking activity to kappa-KTx1.3, a non-toxic homologue of kappa-hefutoxin-1, from *Heterometrus spinifer* venom. Biochem Pharmacol 69, 669-678.

Olamendi-Portugal, T., Gomez-Lagunas, F., Gurrola, G.B. and Possani, L.D., 1998. Two similar peptides from the venom of the scorpion *Pandinus imperator*, one highly effective blocker and the other inactive on K^+ channels. Toxicon 36, 759-770.

Olamendi-Portugal, T., Garcia, B.I., Lopez-Gonzalez, I., Van Der Walt, J., Dyason, K., Ulens, C., Tytgat, J., Felix, R., Darszon, A., and Possani, L.D., 2002. Two new scorpion toxins that target voltage-gated Ca^{2+} and Na^+ channels. Biochem Biophys Res Commun 299, 562-568.

Oren, D.A., Froy, O., Amit, E., Kleinberger-Doron, N., Gurevitz, M. and Shaanan, B., 1998. An excitatory scorpion toxin with a distinctive feature: an additional alpha helix at the C terminus and its implications for interaction with insect sodium channels. Structure 6, 1095-1103.

Padilla, A., Govezensky, T., Possani, L.D. and Larralde, C., 2003. Experimental envenoming

of mice with venom from the scorpion *Centruroides limpidus limpidus*: differences in mortality and symptoms with and without antibody therapy relating to differences in age, sex and strain of mouse. Toxicon 41, 959-965.

Park, C.S. and Miller, C., 1992. Mapping function to structure in a channel-blocking peptide: electrostatic mutants of charybdotoxin. Biochem 31, 7749-7755.

Pearson, W.R., 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 183, 63-98.

Pearson, W.R., 2000. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 132, 185-219.

Pedarzani, P., D'Hoedt, D., Doorty, K.B., Wadsworth, J.D., Joseph, J.S., Jeyaseelan, K., Kini, R.M., Gadre, S.V., Sapatnekar, S.M., Stocker, M. *et al*., 2002. Tamapin, a venom peptide from the Indian red scorpion (*Mesobuthus tamulus*) that targets small conductance Ca^{2+}-activated K^{+} channels and afterhyperpolarization currents in central neurons. J Biol Chem 277, 46101-46109.

Peng, F., Zeng, X.C., He, X.H., Pu, J., Li, W.X., Zhu, Z.H. and Liu, H., 2002. Molecular cloning and functional expression of a gene encoding an antiarrhythmia peptide derived from the scorpion toxin. Eur J Biochem 269, 4468-4475.

Peng, L.S., Zhong, X.F., Huang, Y.S., Zhang, Y., Zheng, S.L., Wei, J.W., Wu, W.Y. and Xu, A.L., 2003. Molecular cloning, expression and characterisation of three short chain alpha-neurotoxins from the venom of sea snake – Hydrophiinae *Hydrophis cyanocinctus* Daudin. Toxicon 42, 753-761.

Pimenta, A.M., Legros, C., Almeida Fde, M., Mansuelle, P., De Lima, M.E., Bougis, P.E. and Martin-Eauclaire, M.F., 2003. Novel structural class of four disulfide-bridged peptides from *Tityus serrulatus* venom. Biochem Biophys Res Commun 301, 1086-1092.

Pimenta, A.M., Martin-Eauclaire, M., Rochat, H., Figueiredo, S.G., Kalapothakis, E., Afonso, L.C. and De Lima, M.E., 2001. Purification, amino-acid sequence and partial characterisation of two toxins with anti-insect activity from the venom of the South American scorpion *Tityus bahiensis* (Buthidae). Toxicon 39, 1009-1019.

Poirot, O., O'Toole, E. and Notredame, C., 2003. Tcoffee@igs: A web server for computing, evaluating and combining multiple sequence alignments. Nucleic Acids Res 31, 3503-3506.

Possani, L.D., Becerril, B., Delepierre, M. and Tytgat, J., 1999. Scorpion toxins specific for Na^+-channels. Eur J Biochem 264, 287-300.

Possani, L.D., Merino, E., Corona, M., Bolivar, F. and Becerril, B., 2000. Peptides and genes coding for scorpion toxins that affect ion-channels. Biochimie 82, 861-868.

Pragl, B., Koschak, A., Trieb, M., Obermair, G., Kaufmann, W.A., Gerster, U., Blanc, E., Hahn, C., Prinz, H., Schutz, G. *et al*., 2002. Synthesis, characterisation, and application of cy-dye- and alexa-dye-labeled hongotoxin$_1$ analogues. The first high affinity fluorescence probes for voltage-gated K^+ channels. Bioconjug Chem 13, 416-425.

Rajendra, W., Armugam, A. and Jeyaseelan, K., 2004. Toxins in anti-nociception and anti-inflammation. Toxicon 44, 1-17.

Ramirez-Dominguez, M.E., Olamendi-Portugal, T., Garcia, U., Garcia, C., Arechiga, H. and Possani, L.D., 2002. Cn11, the first example of a scorpion toxin that is a true blocker of Na^+ currents in crayfish neurons. J Exp Biol 205, 869-876.

Ramos, O.H. and Selistre-de-Araujo, H.S., 2004. Comparative analysis of the catalytic domain of hemorrhagic and non-hemorrhagic snake venom metallopeptidases using bioinformatic tools. Toxicon 44, 529-538.

Ranganathan, R., Lewis, J.H. and MacKinnon, R., 1996. Spatial localization of the K^+ channel selectivity filter by mutant cycle-based structure analysis. Neuron 16, 131-139.

Rodriguez de la Vega, R.C., Merino, E., Becerril, B. and Possani, L.D., 2003. Novel interactions between K^+ channels and scorpion toxins. Trends Pharmacol Sci 24, 222-227.

Rodriguez de la Vega, R.C. and Possani, L.D., 2005. Overview of scorpion toxins specific for Na^+ channels and related peptides: biodiversity, structure-function relationships and evolution. Toxicon 46, 831-844.

Rodriguez de la Vega, R.C. and Possani, L.D., 2004. Current views on scorpion toxins specific for K^+-channels. Toxicon 43, 865-875.

Rogowski, R.S., Collins, J.H., O'Neill, T.J., Gustafson, T.A., Werkman, T.R., Rogawski, M.A., Tenenholz, T.C., Weber, D.J. and Blaustein, M.P., 1996. Three new toxins from the scorpion *Pandinus imperator* selectively block certain voltage-gated K^+ channels. Mol Pharmacol 50, 1167-1177.

Ross, J.S., Schenkein, D.P., Kashala, O., Linette, G.P., Stec, J., Symmans, W.F., Pusztai, L. and Hortobagyi, G.N., 2004. Pharmacogenomics. Adv Anat Pathol 11, 211-220.

Rost, B., 1999. Twilight zone of protein sequence alignments. Protein Eng 12, 85-94.

Sabatier, J.M., Fremont, V., Mabrouk, K., Crest, M., Darbon, H., Rochat, H., Van Rietschoten, J. and Martin-Eauclaire, M.F., 1994. Leiurotoxin I, a scorpion toxin specific for Ca^{2+}-activated K^+ channels. Structure-activity analysis using synthetic analogs. Int J Pept Protein Res 43, 486-495.

Sabatier, J.M., Zerrouk, H., Darbon, H., Mabrouk, K., Benslimane, A., Rochat, H., Martin-Eauclaire, M.F. and Van Rietschoten, J., 1993. P05, a new leiurotoxin I-like scorpion toxin: synthesis and structure-activity relationships of the alpha-amidated analog, a ligand of Ca^{2+}-activated K^+ channels with increased affinity. Biochem 32, 2763-2770.

Saitou, N. and Nei, M., 1987. The neighbour-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4, 406-425.

Santos, A.D., McIntosh, J.M., Hillyard, D.R., Cruz, L.J. and Olivera, B.M., 2004. The A-superfamily of conotoxins: structural and functional divergence. J Biol Chem 279, 17596-17606.

Schonbach, C., Kowalski-Saunders, P. and Brusic, V., 2000. Data warehousing in molecular biology. Brief Bioinform 1, 190-198.

Schroeder, N., Mullmann, T.J., Schmalhofer, W.A., Gao, Y.D., Garcia, M.L. and Giangiacomo, K.M., 2002. Glycine 30 in iberiotoxin is a critical determinant of its specificity for maxi-K versus K_V channels. FEBS Lett 527, 298-302.

Schwede, T., Kopp, J., Guex, N., and Peitsch, M.C., 2003. SWISS-MODEL: an automated

protein homology-modeling server. Nucleic Acids Res 31, 3381-3385.

Selisko, B., Garcia, C., Becerril, B., Delepierre, M. and Possani, L.D., 1996. An insect-specific toxin from *Centruroides noxius* Hoffmann. cDNA, primary structure, three-dimensional model and electrostatic surface potentials in comparison with other toxin variants. Eur J Biochem 242, 235-242.

Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D. and Kahn, D., 2002. ProDom: automated clustering of homologous domains. Brief Bioinform 3, 246-251.

Shakkottai, V.G., Regaya, I., Wulff, H., Fajloun, Z., Tomita, H., Fathallah, M., Cahalan, M.D., Gargus, J.J., Sabatier, J.M. and Chandy, K.G., 2001. Design and characterisation of a highly selective peptide inhibitor of the small conductance calcium-activated K^+ channel, $SK_{Ca}2$. J Biol Chem 276, 43145-43151.

Siew, J.P., Khan, A.M., Tan, P.T., Koh, J.L., Seah, S.H., Koo, C.Y., Chai, S.C., Armugam, A., Brusic, V. and Jeyaseelan, K., 2004. Systematic analysis of snake neurotoxins' functional classification using a data warehousing approach. Bioinformatics 20, 3466-3480.

Smith, C.G. and Vane, J.R., 2003. The discovery of captopril. Faseb J 17, 788-789.

Srinivasan, K.N., Nirthanan, S., Sasaki, T., Sato, K., Cheng, B., Gwee, M.C., Kini, R.M. and Gopalakrishnakone, P., 2001. Functional site of bukatoxin, an alpha-type sodium channel neurotoxin from the Chinese scorpion (*Buthus martensi* Karsch) venom: probable role of the (52)PDKVP(56) loop. FEBS Lett 494, 145-149.

Srinivasan, K.N., Gopalakrishnakone, P., Tan, P.T., Chew, K.C., Cheng, B., Kini, R.M., Koh, J.L., Seah, S.H. and Brusic, V., 2002a. SCORPION, a molecular database of scorpion toxins. Toxicon 40, 23-31.

Srinivasan, K.N., Sivaraja, V., Huys, I., Sasaki, T., Cheng, B., Kumar, T.K., Sato, K., Tytgat, J., Yu, C., San, B.C. *et al*., 2002b. kappa-Hefutoxin1, a novel toxin from the scorpion *Heterometrus fulvipes* with unique structure and function. Importance of the functional diad in potassium channel selectivity. J Biol Chem 277, 30040-30047.

Stampe, P., Kolmakova-Partensky, L. and Miller, C., 1994. Intimations of K^+ channel structure

from a complete functional map of the molecular surface of charybdotoxin. Biochem 33, 443-450.

Strong, P.N., Clark, G.S., Armugam, A., De-Allie, F.A., Joseph, J.S., Yemul, V., Deshpande, J.M., Kamat, R., Gadre, S.V., Gopalakrishnakone, P. *et al*., 2001. Tamulustoxin: a novel potassium channel blocker from the venom of the Indian red scorpion *Mesobuthus tamulus*. Arch Biochem Biophys 385, 138-144.

Sun, X., Chen, X., Zhang, Z., Wang, H., Bianchi, F.J., Peng, H., Vlak, J.M. and Hu, Z., 2002. Bollworm responses to release of genetically modified *Helicoverpa armigera* nucleopolyhedroviruses in cotton. J Invertebr Pathol 81, 63-69.

Sun, Y.M., Bosmans, F., Zhu, R.H., Goudet, C., Xiong, Y.M., Tytgat, J. and Wang, D.C., 2003. Importance of the conserved aromatic residues in the scorpion alpha-like toxin BmK M1: the hydrophobic surface region revisited. J Biol Chem 278, 24125-24131.

Szolajska, E., Poznanski, J., Ferber, M.L., Michalik, J., Gout, E., Fender, P., Bailly, I., Dublet, B. and Chroboczek, J., 2004. Poneratoxin, a neurotoxin from ant venom. Structure and expression in insect cells and construction of a bio-insecticide. Eur J Biochem 271, 2127-2136.

Tan, P.T., Ranganathan, S. and Brusic, V., 2006a. Extraction of functional peptide motifs in scorpion toxins. J Pept Sci 12, 420-427.

Tan, P.T., Veeramani, A., Srinivasan, K.N., Ranganathan, S. and Brusic, V., 2006b. SCORPION2: a database for structure-function analysis of scorpion toxins. Toxicon 47, 356-363.

Tan, P.T., Srinivasan, K.N., Seah, S.H., Koh, J.L., Tan, T.W., Ranganathan, S. and Brusic, V., 2005. Accurate prediction of scorpion toxin functional properties from primary structures. J Mol Graph Model 24, 17-24.

Tan, P.T., Khan, A.M. and Brusic, V., 2003. Bioinformatics for venom and toxin sciences. Brief Bioinform 4, 53-62.

Theakston, R.D., Warrell, D.A. and Griffiths, E., 2003. Report of a WHO workshop on the standardization and control of antivenoms. Toxicon 41, 541-557.

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.

Tsetlin, V.I. and Hucho, F., 2004. Snake and snail toxins acting on nicotinic acetylcholine receptors: fundamental aspects and medical applications. FEBS Lett 557, 9-13.

Tytgat, J., Chandy, K.G., Garcia, M.L., Gutman, G.A., Martin-Eauclaire, M.F., van der Walt, J.J. and Possani, L.D., 1999. A unified nomenclature for short-chain peptides isolated from scorpion venoms: alpha-KTx molecular subfamilies. Trends Pharmacol Sci 20, 444-447.

Vacher, H., Alami, M., Crest, M., Possani, L.D., Bougis, P.E. and Martin-Eauclaire, M.F., 2002. Expanding the scorpion toxin alpha-KTX 15 family with AmmTX3 from *Androctonus mauretanicus*. Eur J Biochem 269, 6037-6041.

Vacher, H., Romi-Lebrun, R., Mourre, C., Lebrun, B., Kourrich, S., Masmejean, F., Nakajima, T., Legros, C., Crest, M., Bougis, P.E. *et al*., 2001. A new class of scorpion toxin binding sites related to an A-type K^+ channel: pharmacological characterisation and localization in rat brain. FEBS Lett 501, 31-36.

Valdar, W.S., 2002. Scoring residue conservation. Proteins 48, 227-241.

Valdez-Cruz, N.A., Batista, C.V., Zamudio, F.Z., Bosmans, F., Tytgat, J. and Possani, L.D., 2004a. Phaiodotoxin, a novel structural class of insect-toxin isolated from the venom of the Mexican scorpion *Anuroctonus phaiodactylus*. Eur J Biochem 271, 4753-4761.

Valdez-Cruz, N.A., Davila, S., Licea, A.F., Corona, M., Zamudio, F., Garcia-Valdes, J., Boyer, L. and Possani, L.D., 2004b. Biochemical, genetic and physiological characterisation of venom components from two species of scorpions: *Centruroides exilicauda* Wood and *Centruroides sculpturatus* Ewing. Biochimie 86, 387-396.

Valentin, E. and Lambeau, G., 2000. What can venom phospholipases A_2 tell us about the functional diversity of mammalian secreted phospholipases A_2? Biochimie 82, 815-831.

Vazquez, J., Feigenbaum, P., Katz, G., King, V.F., Reuben, J.P., Roy-Contancin, L., Slaughter, R.S., Kaczorowski, G.J. and Garcia, M.L., 1989. Characterisation of high affinity binding sites for charybdotoxin in sarcolemmal membranes from bovine aortic smooth muscle. Evidence for a direct association with the high conductance calcium-activated potassium channel. J Biol Chem 264, 20902-20909.

Visan, V., Fajloun, Z., Sabatier, J.M. and Grissmer, S., 2004. Mapping of maurotoxin binding sites on $hK_V1.2$, $hK_V1.3$, and $hIK_{Ca}1$ channels. Mol Pharmacol 66, 1103-1112.

von Segesser, L.K., Mueller, X., Marty, B., Horisberger, J. and Corno, A., 2001. Alternatives to unfractionated heparin for anticoagulation in cardiopulmonary bypass. Perfusion 16, 411-416.

Wagner, S., Castro, M.S., Barbosa, J.A., Fontes, W., Schwartz, E.N., Sebben, A., Rodrigues Pires, O., Jr., Sousa, M.V. and Schwartz, C.A., 2003. Purification and primary structure determination of Tf4, the first bioactive peptide isolated from the venom of the Brazilian scorpion *Tityus fasciolatus*. Toxicon 41, 737-745.

Wang, C.G., Cai, Z., Lu, W., Wu, J., Xu, Y., Shi, Y. and Chi, C.W., 2005. A novel short-chain peptide BmKX from the Chinese scorpion *Buthus martensi* Karsch, sequencing, gene cloning and structure determination. Toxicon 45, 309-319.

Wang, C.G., Gilles, N., Hamon, A., Le Gall, F., Stankiewicz, M., Pelhate, M., Xiong, Y.M., Wang, D.C. and Chi, C.W., 2003. Exploration of the functional site of a scorpion alpha-like toxin by site-directed mutagenesis. Biochem 42, 4699-4708.

Whisstock, J.C. and Lesk, A.M., 2003. Prediction of protein function from protein sequence and structure. Q Rev Biophys 36, 307-340.

White, J., 2000. Bites and stings from venomous animals: a global overview. Ther Drug Monit 22, 65-68.

Wickenden, A.D., 2002a. K^+ channels as therapeutic drug targets. Pharmacol Ther 94, 157-182.

Wickenden, A.D., 2002b. Potassium channels as anti-epileptic drug targets. Neuropharmacology 43, 1055-1060.

Wu, J.J., He, L.L., Zhou, Z. and Chi, C.W., 2002. Gene expression, mutation, and structure-function relationship of scorpion toxin BmP05 active on SK_{Ca} channels. Biochem 41, 2844-2849.

Wu, C.H., Huang, H., Yeh, L.S. and Barker, W.C., 2003. Protein family classification and functional annotation. Comput Biol Chem 27 (1), 37-47.

Wu, Y., Cao, Z., Yi, H., Jiang, D., Mao, X., Liu, H. and Li, W., 2004. Simulation of the interaction between ScyTx and small conductance calcium-activated potassium channel by docking and MM-PBSA. Biophys J 87, 105-112.

Wulff, H., Beeton, C. and Chandy, K.G., 2003. Potassium channels as therapeutic targets for autoimmune disorders. Curr Opin Drug Discov Devel 6, 640-647.

Xu, C.Q., Brone, B., Wicher, D., Bozkurt, O., Lu, W.Y., Huys, I., Han, Y.H., Tytgat, J., Van Kerkhove, E. and Chi, C.W., 2004a. BmBKTx1, a novel Ca^{2+}-activated K^+ channel blocker purified from the Asian scorpion *Buthus martensi* Karsch. J Biol Chem 279, 34562-34569.

Xu, C.Q., He, L.L., Brone, B., Martin-Eauclaire, M.F., Van Kerkhove, E., Zhou, Z. and Chi, C.W., 2004b. A novel scorpion toxin blocking small conductance Ca^{2+} activated K^+ channel. Toxicon 43, 961-971.

Xu, C.Q., Zhu, S.Y., Chi, C.W. and Tytgat, J., 2003. Turret and pore block of K^+ channels: what is the difference? Trends Pharmacol Sci 24, 446-448; author reply 448-449.

Yamaji, N., Dai, L., Sugase, K., Andriantsiferana, M., Nakajima, T. and Iwashita, T., 2004. Solution structure of IsTX. A male scorpion toxin from *Opisthacanthus madagascariensis* (Ischnuridae). Eur J Biochem 271, 3855-3864.

Yao, J., Chen, X., Li, H., Zhou, Y., Yao, L., Wu, G., Zhang, N., Zhou, Z., Xu, T., Wu, H. *et al*., 2005. BmP09, a "long chain" scorpion peptide blocker of BK channels. J Biol Chem 280, 14819-14828.

Yu, K., Fu, W., Liu, H., Luo, X., Chen, K.X., Ding, J., Shen, J. and Jiang, H., 2004. Computational simulations of interactions of scorpion toxins with the voltage-gated potassium ion channel. Biophys J 86, 3542-3555.

Zaki, T.I. and Maruniak, J.E., 2003. Three polymorphic genes encoding a depressant toxin from the Egyptian scorpion *Leiurus quinquestriatus quinquestriatus*. Toxicon 41, 109-113.

Zamudio, F.Z., Conde, R., Arevalo, C., Becerril, B., Martin, B.M., Valdivia, H.H. and Possani, L.D., 1997. The mechanism of inhibition of ryanodine receptor channels by imperatoxin I, a heterodimeric protein from the scorpion *Pandinus imperator*. J Biol Chem 272, 11886-11894.

Zeng, X.C., Peng, F., Luo, F., Zhu, S.Y., Liu, H. and Li, W.X., 2001. Molecular cloning and characterisation of four scorpion K^+-toxin-like peptides: a new subfamily of venom peptides (alpha-KTx14) and genomic analysis of a member. Biochimie 83, 883-889.

Zhang, N., Wu, G., Wu, H., Chalmers, M.J. and Gaskell, S.J., 2004. Purification, characterisation and sequence determination of BmKK4, a novel potassium channel blocker from Chinese scorpion *Buthus martensi* Karsch. Peptides 25, 951-957.

Zhu, S., Bosmans, F. and Tytgat, J., 2004a. Adaptive evolution of scorpion sodium channel toxins. J Mol Evol 58, 145-153.

Zhu, X., Zamudio, F.Z., Olbinski, B.A., Possani, L.D. and Valdivia, H.H., 2004b. Activation of skeletal ryanodine receptors by two novel scorpion toxins from *Buthotus judaicus*. J Biol Chem 279, 26588-26596.

Zhu, S., Huys, I., Dyason, K., Verdonck, F. and Tytgat, J., 2004c. Evolutionary trace analysis of scorpion toxins specific for K-channels. Proteins 54, 361-370.

Zhu, S., Darbon, H., Dyason, K., Verdonck, F. and Tytgat, J., 2003. Evolutionary origin of inhibitor cystine knot peptides. Faseb J 17, 1765-1767.

Zhu, S. and Li, W., 2002. Precursors of three unique cysteine-rich peptides from the scorpion *Buthus martensii* Karsch. Comp Biochem Physiol B Biochem Mol Biol 131, 749-756.

Zhu, S.Y., Li, W.X. and Zeng, X.C., 2001. Precursor nucleotide sequence and genomic organization of BmTXKS1, a new scorpion toxin-like peptide from *Buthus martensii* Karsch. Toxicon 39, 1291-1296.

Zilberberg, N., Froy, O., Loret, E., Cestele, S., Arad, D., Gordon, D. and Gurevitz, M., 1997.

Identification of structural elements of a scorpion alpha-neurotoxin important for receptor site recognition. J Biol Chem 272, 14810-14816.

Zlotkin, E., Kadouri, D., Gordon, D., Pelhate, M., Martin, M.F. and Rochat, H., 1985. An excitatory and a depressant insect toxin from scorpion venom both affect sodium conductance and possess a common binding site. Arch Biochem Biophys 240, 877-887.

Zuo, X.P. and Ji, Y.H., 2004. Molecular mechanism of scorpion neurotoxins acting on sodium channels: insight into their diverse selectivity. Mol Neurobiol 30, 265-278.

Author's Publications

Tan, P.T., Ranganathan, S. and Brusic. V., 2006. Extraction of functional peptide motifs in scorpion toxins. J Peptide Sci 12, 420-427.

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.

Tan, P.T., Srinivasan, K.N., Seah, S.H., Koh, J.L.Y., Tan, T.W., Ranganathan, S. and Brusic, V., 2005. Accurate prediction of scorpion toxin functional properties from primary structures. J Mol Graph Model 24, 17-24.

Tan, P.T., Khan, A.M. and Brusic, V., 2003. Bioinformatics for venom and toxin sciences. Brief Bioinform 4, 53-62

Siew, J.P., Khan, A.M., **Tan, P.T.**, Koh, J.L., Seah, S.H., Koo, C.Y., Chai, S.C., Armugam, A., Brusic, V. and Jeyaseelan, K., 2004. Systematic analysis of snake neurotoxins functional classification using a data warehousing approach. Bioinformatics 20, 3466-3480.

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.

Koh, J.L.Y., Krishnan, S.P.T., Seah, S.H., **Tan, P.T.**, Khan, A.M., Lee, M.L. and Brusic, V., 2004. BioWare: A framework for bioinformatics data retrieval, annotation and publishing. ACM SIGIR Workshop on Search & Discovery in Bioinformatics, Sheffield, UK, July 2004.

Srinivasan, K.N., Gopalakrishnakone, P., **Tan, P.T.**, Chew, K.C., Cheng, B., Kini, R.M., Koh, J.L.Y., Seah, S.H. and Brusic, V., 2002. SCORPION, a molecular database of scorpion toxins. Toxicon 40, 23-31.

Appendix 1

The 62 groups of 393 scorpion native toxin sequences were classified based on ion channel specificity and primary sequence similarity using BLAST, multiple sequence alignment and phylogenetic analysis. The scorpion toxin sequences were broadly classified into four broad groups, namely Na^+, K^+, Ca^{2+} and Cl^-. Each broad group was then classified into groups and subgroups by BLAST and multiple sequence alignment, and verified by phylogenetic analysis. 135 K^+ toxins were classified into 32 groups, 222 Na^+ toxins were classified into 18 groups, while 8 Ca^{2+} toxins were classified into four groups. K^+ group 6 was further classified into three subgroups each. Na^+ groups 2 and 9 were classified into four subgroups each, Na^+ group 1 into three subgroups and Na^+ group 12 into two subgroups. 19 Cl^- toxins were classified into two subgroups. Scorpine and nine scorpion toxin sequence with no annotated molecular target were assigned to the 'defensin' and 'orphan' groups, respectively.

Legend: First column shows the accession numbers in SCORPION2. Second column shows aligned scorpion toxin sequences. Dashes are introduced for aligning sequences. Matured toxin is represented by green single amnio acid notations, signal peptide by brown and fragment region by red. Conserved residues within groups/subgroups are highlighted in grey. Pairs of disulfide bridges are shown by different highlights.

Sodium channel toxins
Group 01 Subgroup 01a

DBACC

D000003 ----------KKNGYAVDSSGKVSECLLNNYCNNICTKVYYATSGYCCLLSCYCFGLDDDKAVLKIKDATKSYCDVQIN---

D000184 ----------KKNGYAVDSSGKVSECLLNNYCNINCTKVYYATSGYCCLLSCYCFGLDDDKAVLKIKDATKSYCDVQIIN--

D000136 MKFFLIFLVIFPIMGVLGKKNGYAVDSSGKVSECLLNNYCNNICTKVYYATSGYCCLLSCYCFGLDDDKAVLKIKDATKSYCDVQIIG--

D000165 MKFFLIFLVIFPIMGVLGKKNGYAVDSSGKVAECLFNNYCNNECTKVYYADKGYCCLLKCYCFGLADDKPVLDIWDSTKNYCDVQIIDLS

Subgroup 01b

DBACC

D000233 MKFFLLFLVVLPIMGVLGKKNGYAVDSKGKAPECFLSNYCNNECTKVHYADKGYCCLLSCYCFGLNDDKKVLEISGTTKKYCDFTIIN

D000234 MKFFLLFLVVLPIMGVLGKKNGYAVDSKGKAPECFLSNYCNNECTKVHYADKGYCCLLSCYCFGLNDDKKVLEISDTTKKYCDFTIIN

D000006 ----------KKNGYAVDSSGKAPECLLSNYCYNECTKVHYADKGYCCLLSCYCVGLSDDKKVLEISDARKKYCDFVTIN

D000007 ----------KKNGYAVDSSGKAPECLLSNYCYNECTKVHYAEKGYCCLLSCYCVGLSDDKKVLEISDARKKYCDFVTIN

D000824 ----------KKNGYAVDSSGKAPECLLSNYCNNECTKVHYADKGYCCLLSCYCFGLSDDKKVLEISDTRKKYCDYTIIN

D000825 ----------KKNGYAVDSSGKAPECLLSNYCNNECTKVHYADKGYCCLLSCYCFGLSDDKKVLDISDTRKKYCDYTIIN

D000005 MKFLLLFLVVLPIMGVLGKKNGYAVDSSGKAPECLLSNYCYNECTKVHYADKGYCCLLSCYCFGLNDDKKVLEISDTRKSYCDTPIIN

D000139 ----------KKDGYAVDSSGKAPECLLSNYCYNECTKVHYADKGYCCLLSCYCFGLNDDKKVLEISDTRKSYCDTPIIN

D000004 MKFLLLFLVVLPIMGVFGKKNGYAVDSSGKAPECLLSNYCNNECTKVHYADKGYCCLLSCYCFGLNDDKKVLEISDTRKSYCDTTIIN

D000125 ----------KKNGYAVDSSGKAPECLLSNYCNNQCTKVHYADKGYCCLLSCYCFGLNDDKKVLEISDTRKSYCDTTIIN

D000137 MKFLLLFLVVLPIMGVLGKKNGYAVDSSGKAPECLLSNYCNNECTKVHYADKGYCCLLSCYCFGLNDDKKVLEISDTRKSYCDTTIIN

D000235 MKFFLLFLVVLPIMGVLGKKNGYAVDSKGKAPECFFSNYCNNECTKVHYAEKGYCCLLSCYCVGLNDDKKVMEISDTRKKICDTTIIN

D000236 MKFFLLFLVVLPIMGVLGKKNGFAVDSNGKAPECFFDHYCNSECTKVYYAEKGYCCTLSCYCVGLDDDKKVLDISDTRKKLCDFTLFN

Subgroup 01c

DBACC

D000001 MKFFLMCLIIFPIMGVLGKKNGYPLDRNGKTTECSGVNAIAPHYCNSECTKVYYAESGYCCWGACYCFGLEDDKPIGPMKDITKKYCDVQIIPS

D000002 MKFFLMCLIIFPIMGVLGKKNGYPLDRNGKTTECSGVNAIAPHYCNSECTKVYYAKSGYCCWGACYCFGLEDDKPIGPMKDITKKYCDVQIIPS

Group 02 Subgroup 02a

DBACC

D000009 MKGMILFISCLLLIGIVVECKEGYLMDHEGCKLSCFIRPSGYCGSECKIKKGSSGYCAWPACYCYGLPNWVKVWDRATNKCGKK

D000012 MKGMILFISCLLLIGIVVECKEGYLMDHEGCKLSCFIRPSGYCGRECGIKKGSSGYCAWPACYCYGLPNWVKVWDRATNKCGKK

D000008 MKGMILFISCLLLIDIVVGGKEGYLMDHEGCKLSCFIRPSGYCGRECTLKKGSSGYCAWPACYCYGLPNWVKVWDRATNKCGKK

Subgroup 02b

DBACC

D000021 KEGYAMDHEGCKFSCFPRPAGFCDGYCKTHLKASSGYCAWPACYCYGVPSNIKVWDYATNKC

D000246 KEGYAMDHEGCKFSCFIRPSGFCDGYCKTHLKASSGYCAWPACYCYGVPSNIKVWDYATNKC

D000017 KEGYAMDHEGCKFSCFIRPAGFCDGYCKTHLKASSGYCAWPACYCYGVPDHIKVWDYATNKC

D000222 KEGYAMDHEGCKFSCFIRPAGFCDGYCKTHLKASSGYCAWPACYCYGVPDHIKVWDYATNKC

Subgroup 02c

DBACC

D000010 ------------------GREGYPADSKGCKIT-CFLTAAGYCNTECTLKKGSSGYCAWPACYCYGLPESVKIWTSETNKC---

D000162 MKRMILFISCLLLIDIVVGGREGYPADSKGCKIT-CFLTAAGYCNTECTLKKGSSGYCAWPACYCYGLPDSVKIWTSETNKCGKK

D000172 ------------------GKEGYPTDKRGCKLT-CFFT----------

D000245 ------------------GKEGYPVDSRGCKVT-CFFTGAGYCDKECKLKKASSGYCAWPACYCYGLPDSVPVYDNASNKCB--

D000631 ------------------GKEGYPADSKGCKVT-CFFTGVGYCDTECKLKKASSGYCAWPACYCYGLPDSASVWDSATNKC---

D000632 -------------------KDGYPVDSKGCKLS-CVAN--NYCDNQCKMKKASGGHCYAMSCYCEGLPENAKVSDSATNIC---

D000626 ------------------LKDGYPTNSKGCKISGCLPGENKFCLNECQ----------

Subgroup 02d

DBACC

D000694 MTRFVLFICCFFLIGMVVECKDGYLVGNDGCKYSCFTR-PGTYCANECSRVKGKDGYCYAWMACYCYSMPNWVKTWDRATNRCGRGK

D000696 --------------------KDGYLVGNDGCKYNCLTR-PGHYCANECSRVKGKDGYCYAWMACYCYSMPDWVKTWSRSTNRCGR--

D000695 ------------------KKEGYLVGNDGCKYGCITR-PHQYCVHECELKKGTDGYCAYWLACYCYNMPDWVKTWSSATNKCK---

D000822 ------------------NKDGYLMEGDGCKMGCLTRKKASYCVDQCKEVGGKDGYCYAWLSCYCYNMPDSVEIWDSKNNKCGK--

D000697 MKGMIMLISCLMLIDVVVESKNGYIIEPKGCKYSCFWG-SSTWCNRECKFKKGSSGYC-AWPACWCYGLPDNVKIFDYYNNKCGK--

Group 03

DBACC

D000031 ------------------GRDAYIADSENCTYTC--ALNPYCNDLCTKNGAKS----GYCQWAGRYGNACWCIDLPDKVPIRISG-SCR--

D000209 MNYLIVISFALLLMTGVESGRDAYIAKKENCTYFC--ALNPYCNDLCTKNGAKS----GYCQWAGRYGNACWCIDLPDKVPIRIPG-PCIGR

D000047 ------------------GRDAYIADSENCTYFC--GSNPYCNDVCTENGAKS----GYCQWAGRYGNACYCIDLPASERIKEPG-KCG--

D000210 MNYLIVISFALLLMTSVESGRDAYIADSENCTYFC--GSNPYCNDLCTENGAKS----GYCQWAGRYGNACWCIDLPDKVPIRIPG-PCRGR

D000027 ------------------ERDGYIVQLHNCVYHC--GLNPYCNGLCTKNGATSG---SYCQWMTKWGNACYCYALPDKVPIKWLDPKCY--

D000076 ---------------AEIKVRDGYIVYPNNCVYHC--GLDPYCNDLCT--GA----------

D000025 MNHLVMISLALLLLLGVESVRDAYIAKNYNCVYEC--FRDAYCNELCTKNGASS----GYCQWAGKYGNACWCYALPDNVPIRVPG-KCHRK

D000029 ------------------VRDAYIAKNYNCVYEC--FRDSYCNDLCTKNGASS----GYCQWAGKYGNACWCYALPDNVPIRVPG-KCH--

D000030 ------------------VRDAYIAQNYNCVYFC--MKDDYCNDLCTKNGASS----GYCQWAGKYGNACWCYALPDNVPIRIPG-KCHS-

D000228 ------------------VRDAYIAQNYNCVYTC--FKDAHCNDLCTKNGASS----GYCQWAGKYGNACWCYALPDNVPIRIPG-KCHRK

D000164 ----------LLMTGVESGRDAYIAKNYNCVYHC--FRDDYCNGLCTENGADS----GYCYLAGKYGNACWCINLPDDKPIRIPG-KCHRR

D000223 MNHLVMISLALLLMTGVESGRDAYIAQNYNCVYHC--ALNPYCNDLCTKNGAKS----GYCQWFGSSGNACWCIDLPDNVPIKVPG-KCHRK

D000224 MNHLVMISLALLLMTGVESGRDAYIAQNYNCVYHC--ALNPYCNDLCTKNGAKS----GYCQWFGSNGNACWCIDLPDSVPIKVPG-KCHRK

D000225 MNHLVMISLALLLMTGVESGRDAYIAQNYNCVYHC--FVNPYCNDLCTKNGAES----GYCQWFTSSGNACWCINLPDSVPIKIPG-KCHRK

D000075 ----------VRDAYIAQNYNCVYTC--FKNDYCNDICTKNGAXX----GYC----------

D000226 ----------VRDAYIAQNYNCVYDC--ARDAYCNDLCTKNGAKS----GYCEWFGPHGDACWCIDLPNNVPIKVEG-KCHRK

D000227 ----------VRDAYIAQNYNCVYDC--ARDAYCNELCTKNGAKS----GHCEWFGPHGDACWCIDLPNNVPIKVEG-KCHRK

D000229 ----------VRDAYIAQNYNCVYHC--GRDAYCNELCSKNGAKSRTRGGYCHWFGPHGDACWCIDLPNNVPIKVEG-KCHRK

D000230 ----------VRDAYIAQNYNCVYAC--ARDAYCNDLCTKNGARS----GLFATFGPHGDACWCIALPNNVPLKVQG-KCHRK

D000211 MNYLVMISFAFLLMTGVESVRDAYIAQNYNCVYHC--ARDAYCNELCTKNGAKS----GSCPYLGEHKFACYCKDLPDNVPIRVPG-KCHRR

D000267 MNYLVMISLA-LLIAGVDSARDAYIAKNDNCVYEC--FQDSYCNDLCTKNGAKS----GTCDWIGTYGDACLCYALPDNVPIKLSG-ECHR-

D000028 ----------GRDAYIAQPENCVYEC--AKNSYCNDLCTKNGAKS----GYCQWLGRWGNACYCIDLPDKVPIRIEG-KCHF-

D000144 ----------GRDAYIAQPENCVYEC--AKSSYCNDLCTKNGAKS----GYCQWLGRWGNACYCIDLPDKVPIRIEG-KCHFA

D000068 ----------GRDAYIAQPENCVYEC--AQNSYCNDLCTKNGATS----GYCQWLGKYGNACWCKDLPDNVPIRIPG-KCHF-

D000143 ----------GRDAYIAQPENCVYEC--AKNSYCNDLCTKNGAKS----GYCQWLGKYGNACWCEDLPDNVPIRIPG-KCHF-

D000045 ----------ARDAYIAKPHNCVYECYNPKGSYCNDLCTENGAES----GYCQILGKYGNACWCIQLPDNVPIRIPG-KCH--

D000040 MNYLVMISFALLLMTGVESVRDAYIAKPENCVYHC--ATNEGCNKLCTDNGAES----GYCQWGGRYGNACWCIKLPDRVPIRVPG-KCHR-

D000156 MNYLVMISFALLLMTGVESVRDAYIAKPENCVYHC--ATNEGCNKLCTDNGAES----GYCQWGGKYGNACWCIKLPDDVPIRVPG-KCHR-

D000126 MNYLVMISFALLLMKGVESVRDAYIAKPENCVYHC--AGNEGCNKLCTDNGAES----GYCQWGGRYGNACWCIKLPDDVPIRVPG-KCHR-

D000042 ----------VRDAYIAKPENCVYEC--ATNEYCNKLCTDNGAES----GYCQWVGRYGNACXCIKLPDRVPIRVWG-KCHG-

D000037 MNYLVMISFALLLMKGVESVRDAYIAKPENCVYEC--GITQDCNKLCTENGAES----GYCQWGGKYGNACWCIKLPDSVPIRVPG-KCQR-

D000038 MNYLVMISFALLLMTGVESVRDAYIAKPHNCVYEC--ARNEYCNDLCTKNGAKS----GYCQWVGKYGNGCWCIELPDNVPIRVPG-KCHR-

D000213 MNYLVMISFALLLMTGVESVRDAYIAKPHNCVYEC--ARNEYCNDLCTKNGAKS----GYCQWVGKYGNGCWCKELPDNVPIRVPG-KCHR-

D000142 ----------VRDAYIAKPHNCVYEC--ARNEYCNNLCTKNGAKS----GYCQWSGKYGNGCWCIELPDNVPIRVPG-KCH--

D000141 ----------VRDAYIAKPHNCVYEC--ARNEYCNDLCTKDGAKS----GYCQWVGKYGNGCWCIELPDNVPIRIPG-NCH--

D000155 MNYLVMVSFALLLMTGVESVRDGYIALPHNCAYGC--LLNEFCNDLCTKNGAKI----GYCNIQGKYGNACWCIELPDNVPIRVPG-RCHPS

D000634 ----------VRDGYIALPHNCAYGC--LNNEYCNNLCTKDGAKI----GYCNIVGKYGNACWCIQLPDNVPIRVPG-RCHPA

D000611 ----------GEDGYIADGDNCTYIC--TFNNYCHALCTDKKGDS----GACDWWVPYGVVCWCEDLPTPVPIRGSG-KCR--

Group 04

DBACC

D000069 MNYLVMISLALLLMTGVESVRDGYIVDSKNCVYHCVPPCDGLCKKNGAKSGSCGFLIPSGLACWCVALPDNVPIKDPSYKCHSR

D000070 MNYLIMFSLALLLVIGVESGRDGYIVDSKNCVYHCYPPCDGLCKKNGAKSGSCGFLVPSGLACWCNDLPENVPIKDPSDDCHKR

D000071 MNYLVMISLALLLMIGVESKRDGYIVYPNNCVYHCVPPCDGLCKKNGGSSGSCSFLVPSGLACWCKDLPDNVPIKDTSRKCTR-

```
D000072  MNYLVMISLALLLMIGVESKRDGYIVYPNNCVYHCIPPCDGLCKKNGGSSGSCSFLVPSGLACWCKDLPDNVPIKDTSRKCTR-
D000625  MNYLVMISLALLLMIGVESVRDGYIVYPHNCVYHCIPSCDGLCKENGATSGSCGYIIKVGIACWCKDLPENVPIYDRSYKCYR-
```

Group 05

DBACC

```
D000074  -MSSLMISTAMKGKAPYRQVRDGYIAQPHNCAYHCLKISSGCDTLCKENGATSGHCGHKSGHGSACWCKDLPDKVGIIVHGEKCHR
D000078  ------------------GVRDGYIAQPHNCVYHCFPGSGGCDTLCKENGATQGSSCFILGRGTACWCKDLPDRVGVIVDGEKCH-
D000621  ------------------VRDGYIAQPENCVYHCIP---DCDTLCKDNGGTGGHCGFKLGHGIACWCNALPDNVGIIVDGVKCHK
D000622  ------------------VRDGYIAKPENCAHHCFPGSSGCDTLCKENGGTGGHCGFKVGHGTACWCNALPDKVGIIVDGVKCH-
D000073  ------------------VRDGYIAQPENCVYHCFPGSSGCDTLCKEKGGTSGHCGFKVGHGLACWCNALPDNVGIIVEGEKCHS
D000077  ------------------GRDGYIAQPENCVYHCFPGSSGCDTLCKEKGATSGHCGFLPGSGVACWCDNLPNKVPIVVGGEKCH-
D000628  MNYLVMISLALLFMIGVESARDGYIAQPNNCVYHCIPLSPGCDKLCRENGATSGKCSFLAGSGLACWCVALPDNVPIKIIGQKCTR
```

Group 06

DBACC

```
D000023  ------------------GVRDAYIADDKNCVYTCGSNSYCNTECTKNGAESGYCQWLGKYGNACWCIKLPDKVPIR--IPGK-CR--
D000665  ------------------GVRDAYIADDKNCVYTCGANSYCNTECTKNGAESGYCQWFGKYGNACWCIKLPDKVPIR--IPGK-CR--
D000817  MNYLVVICFALLLMT-VVESGRDAYIADNLNCAYTCGSNSYCNTECTKNGAVSGYCQWLGKYGNACWCINLPDKVPIR--IPGA-CRGR
D000627  MNYLITISLALLLMTGVASGVRDGYIADAGNCGYTCVANDYCNTECTKNGAESGYCQWFGRYGNACWCIKLPDKVPIK--VPGK-CNGR
D000014  MNYLVMISLALLFVT-GVESVKDGYIVDDVNCTYFCGRNAYCNEECTKLKGESGYCQWASPYGNACYCYKLPDHVRTK--GPGR-CHGR
D000179  --------------------IKDGYIVDDVNCTYFCGRNAYCNEECTKLKGESGYCQWASPYGNACYCYKLPDHVRTK--GPGR-CR--
D000019  --------------------VKDGYIVDDRNCTYFCGRNAYCNEECTKLKGESGYCQWASPYGNACYCYKVPDHVRTK--GPGR-CN--
D000688  MNYLVMISLALLFMT-GVESLKDGYIVNDINCTYFCGRNAYCNELCIKLKGESGYCQWASPYGNSCYCYKLPDHVRTK--GPGR-CNDR
D000022  --------------------LKDGYIVDDRNCTYFCGTNAYCNEECVKLKGESGYCQWVGRYGNACWCYKLPDHVRTV--QAGR-CRS-
D000026  --------------------LKDGYIVDDKNCTFFCGRNAYCNDECKKKGGESGYCQWASPYGNACWCYKLPDRVSIK--EKGR-CN--
D000032  --------------------LKDGYIIDDLNCTFFCGRNAYCDDECKKKGGESGYCQWASPYGNACWCYKLPDRVSIK--EKGR-CN--
D000268  MNYLVMISLALLFMT-GVESKKDGYIVDDKNCTFFCGRNAYCNDECKKKGAESGYCQWASPYGNACYCYKLPDRVSTK--KKGG-CNGR
D000151  MNYMVIISLALLVMT-GVESVKDGYIADDRNCPYFCGRNAYCDGECKKNRAESGYCQWASKYGNACWCYKLPDDARIM--KPGR-CNGG
D000163  MNYLVFFSLALLLMT-GVESVKDGYIADDRNCPYFCGRNAYCDGECKKNRAESGYCQWASKYGNACWCYKLPDDARIM--KPGR-CNGG
D000203  MNYLVFFSLALLLMT-GVESVRDGYIADDKNCAYFCGRNAYCDDECKKNGAESGYCQWAGVYGNACWCYKLPDKVPIR--VPGK-CNGG
D000614  MNYLVFFSLALLLMT-GVESVRDGYIADDKNCAYFCGRNAYCDDECKKKGAESGYCQWAGVYGNACWCYKLPDKVPIR--VPGK-CNGG
D000212  MNYLVFFSLALLVMT-GVESVRDGYIADDKNCAYFCGRNAYCDDECKKNGAESGYCQWAGVYGNACWCYKLPDKVPIR--VPGK-CNGG
```

D000150 MNYLVFFSLALLLMT-GVGSVRDGYIADDKNCPYFCGRNAYCDDECKKNGAESGYCQWAGVYGNACWCYKLPDKVPIR--VPGK-CNGG

D000063 ----------VRDGYIADDKDCAYFCGRNAYCDEECKK-GAESGKCWYAGQYGNACWCYKLPDWVPIKQKVSGK-CN--

D000244 ----------VRDGYIADDKNCAYFCGRNAYCDEECIINGAESGYCQQAGVYGNACWCYKLPDKVPIR--VSGE-CQQ-

D000039 ----------ARDAYIADDRNCVYTCALNPYCDSECKKNGADGSYCQWLGRFGNACWCKNLPDDVPIR-KIPGEECR--

Group 07

DBACC

D000015 MVVVCLLTAGTEGKKDGYPVEYDNCAYICWNYDNAYCDKLCKDKKADSGYCYWVHILCYCYGLPD---SEPTKTNGKCKSGKK

D000016 LVVVCLLTAGTEGKKDGYPVEYDNCAYICWNYDNAYCDKLCKDKKADSGYCYWVHILCYCYGLPD---SEPTKTNGKCKSGKK

D000020 ----------KKDGYPVEADNCAFVCFGYDNAYCDKLCGDKKADSGYCYWVHILCYCYGLPD---NEPTKTNGKC-----

D000024 ----------KKDGYPVEGDNCAFACFGYDNAYCDKLCKDKKADDGYCVWS-PDCYCYGLPEHILKEPTKTSGRC-----

D000011 ----------KKDGYPVDSGNCKYECLKDD--YCNDLCLERKADKGYCYWGKVSCYCYGLPD---NSPTKTSGKCNPA--

D000180 ----------KIDGYPVDYWNCKRICW-YNNKYCNDLCKGLKADSGYCWGWTLSCYCQGLPD---NARIKRSGRCRA---

Group 08

DBACC

D000050 GRDGYVVKNGTNCKYSCEIGSEYEYCGPLCKRKNAKTGYCYAFACWCIDVPDDVKLYGDDGTYCSS

D000055 ARDGYIVHDGTNCKYSCEFGSEYKYCGPLCEKKKAKTGYCYLFACWCIEVPDEVRVWGEDGFMCWS

Group 09 Subgroup 09a

DBACC

D000013 MNSLLMITACFVLIGTVWAKDGYLVDAKGCKKNCYKLGKNDYCNRECRMKHRGGSYGYCYGFGCYCEGLSDSTPTWPLPNKTCSGK

D000121 ----------KDGYLVDAKGCKKNCYKLGKNDYCNRECRMKHRGGSYGYCYGFGCYCEGLSDSTPTWPLTNKTC---

D000018 MNSLLIITACLVLIGTVWAKDGYLVDVKGCKKNCYKLGENDYCNRECKMKHRGGSYGYCYGFGCYCEGLSDSTPTWPLPNKRCGGK

D000035 ----------KDGYLVEKTGCKKTCYKLGENDFCNRECKWKHIGGSYGYCYGFGCYCEGLPDSTQTWPLPNKTC---

D000603 MNSLLMITACLVLIGTVWAKDGYLVEKTGCKKTCYKLGENDFCNRECKWKHIGGSYGYFYGFGCYCEGLPDSTQTWPLPNKTCGKK

D000638 MNSLLMITACLVVIGTVWAKEGYLVDVKGCKKNCWKLGDNDYCNRECKWKHIGGSYGYCYGFGCYCEGLPDSTQTWPLPNKTCGKK

Subgroup 09b

DBACC

D000059 ----------KEGYIVNYHTGCKYTCAKLGDNDYCLRECK----------

D000616 MNSLLIIAACLALIGTVWAKEGYIVNYHTGCKYECFKLGDNDYCLRECKLRHGKGSGGYCYAFGCWCTHLYEQAVVWPLPKKKCNGK

D000134 ----------KEGYIVNYHDGCKYECYKLGDNDYCLRECKLRVGKGAGGYCYAFACWCTHLYEQA----------

D000673 MNSLLMITACLALIGTVWAKEGYIVNYHDGCKYECYKLGDNDYCLRECKLRYGKGAGGYCYAFGCWCTHLYEQAVVWPLPKKRCNGK

D000133 ----------KEGYIVNYYTGCKFACAKLGDNDYCLRECKARYGKGAGGYCYAFGCWCTHLYEQAVVWPLPK------

D000674 MNSLLMITACLAVIGTVWAKEGYIVNYYDGCKYACLKLGENDYCLRECKARYYKSAGGYCYAFACWCTHLYEQAVVWPLPNKTCYGK

D000181 ----------KEGYLVNSYTGCKFECFKLGDNDYCLRECRQQYGKGSGGYCYAFGCWCTHLYEQAVVWPLPNKTCN--

D000182 ----------KEGYLVNSYTGCKFECFKLGDNDYCKRECKQQYGKSSGGYCYAFGCWCTHLYEQAVVWPLPNKTCN--

D000044 ----------KEGYLVNSYTGCKYECLKLGDNDYCLRECRQQYGK-SGGYCYAFACWCTHLYEQAVVWPLPNKTCN--

D000051 ----------KEGYLVSKSTGCKYECLKLGDNDYCLRECKQQYGKSSGGYCYAFACWCTHLYEQAVVWPLPNKTCN--

D000041 ---LLIITACLALIGTVWAKEGYLVDKNTGCKYECLKLGDNDYCLRECKQQYGKGAGGYCYAFACWCTHLYEQAIVWPLPNKRCSGK

D000123 ----------KEGYLVNHSTGCKYECYKLGDNDYCLRECK----------

D000176 ----------KEGYLVNHSTGCKYECYKLGDNDYCLRECKQQYGKGAGGYCYAFGCWCTHLYEQAVVWPLPKKTCN--

D000612 MNSLLMITACLALVGTVWAKEGYLVNHSTGCKYECYKLGDNDYCLRECKQQYGKGAGGYCYAFGCWCTHLYEQAVVWPLPKKTCNGK

D000177 ----------KEGYLVNHSTGCKYECFKLGDNDYCLRECKQQYGKGAGGYCYAFGCWCNHLYEQAVVWPLPKKTCN--

D000057 ----------KEGYIVNLSTGCKYECYKLGDNDYCLRECKQQYGKGAGGYCYAFGCWCTHLYEQAVVWPLPKKTCT--

D000056 ----------KEGYLVNHSTGCKYECFKLGDNDYCLRECRQQYGKGAGGYCYAFGCWCTHLYEQAVVWPLPNKTCS--

D000043 MNSLLMITACLALVGTVWAKEGYLVNSYTGCKYECFKLGDNDYCLRECKQQYGKGAGGYCYAFGCWCTHLYEQAVVWPLKNKTCNGK

D000054 ----------KEGYLVELGTGCKYECFKLGDNDYCLRECKARYGKGAGGYCYAFGCWCTQLYEQAVVWPLKNKTCR--

D000122 ----------KKDGYLVNKYTGCKVNCYKLGENKFCNRE----------

D000633 MNSLLMITACLVLFGTVWAKEGYLVNTYTGCKYICWKLGENKYCIDECKE--IGAGYGYCYGFGCYCEGFPENKPTWPLPNKTCGRK

Subgroup 09c

DBACC

D000053 ----------KEGYLVNKSTGCKYGCFWLGKNENCDKECKAKNQGGSYGYCYSFACWCEGLPESTPTYPLPNKSCS--

D000672 MNSLLMITACLAEIGTVWAKEGYLVNKSTGCKYGCFWLGKNENCDKECKAKNQGGSYGYCYSFACWCEGLPDSTPTYPLPNKSCSKK

D000666 MNSLLMITACLVLFGTVWAKEGYLVNKSTGCKYGCFWLGKNENCDMECKAKNQGGSYGYCYSFACWCEGLPDSTPTYPLPNKSCSKK

D000671 MNSLLIITACLVLF--VWAKEGYLVNKSTGCKYGCFWLGKNENCDMECKAKNQGGSYGYCYSFACWCEGLPDSTPTYPLPNKSCSKK

D000046 MNSLLMITACLVLFGTVWAKEGYLVNKSTGCKYGCFWLGKNEGCDKECKAKNQGGSYGYCYAFGCWCEGLPESTPTYPLPNKTCSKK

D000048 ----------KEGYLVKKSDGCKYDCFWLGKNEHCNTECKAKNQGGSYGYCYAFACWCEGLPESTPTYPLPNKSC---

D000608 MNSLLIITACFALVGTVWAKEGYLVKKSDGCKYDCFWLGKNEHCDTECKAKNQGGSYGYCYAFACWCEGLPESTPTYPLPNKSCGKK

D000600 MNSLLMITACFALVGTVWAKEGYLVKKSDGCKYDCFWLGKNEHCDLECKAKNQGGSYGYCYAFACWCEGLPESTPTYPLPNKSCGKK

D000609 MNSLLMITACFALVGTVWAKEGYLVKKSDGCKYDCFWLGENEGCDKECKAKNQGGSYGYCYAFACWCEGLPESTPTYPLPNKSCGKK

D000669 ----------TVSAKEGYLVKKSNGCKYECFKLGENEHCDTECKAPNQGGSYGYCDTFECWCEGLPESTPTWPLPNKSCGKK

D000599 MNSLLIITVCLFLIGTVWAKEGYLVNKSTGCKYDCFWLGKNEHCDLECKAKNQGGSYGYCYAFACWCEGLPESTPTYPLPNKSCGKK

D000058 MNSLLIITACLFLIGTVWAKEGYLVNKSTGCKYGCLKLGENEGCDKECKAKNQGGSYGYCYAFACWCEGLPESTPTYPLPNKSCSRK

D000604 MNSLLIITACLFLIGTVWAKEGYLVNKSTGCKYGCLKLGENEGCDKECKAENQGGSYGYCYAFACWCEGLPESTPTYPLPNKSCSRK

D000605 MNSLLIITACFALVGTVWAKEGYLVNKSTGCKYGCLKLGENEGCDKECKAKNQGGSYGYCYAFACWCEGLPESTPTYPLPNKSCSRK

D000606 MNSLLMITACLFLIGTVWAREGYLVNKSTGCKYGCLKLGENEGCDKECKAKNQGGSYGYCYAFACWCEGLPESTPTYPLPNKSCSRK

D000607 MNSLLMITACLFLIGTVWAKEGYLVNKSTACKYGCLKLGENEGCDKECKAKNQGGSYGYCYAFACWCEGLPESTPTYPLPNKSCSRK

D000052 ------------------KEGYLVKKSDGCKYGCLKLGENEGCDTECKAKNQGGSYGYCYAFACWCEGLPESTPTYPLPNKSC---

D000602 MNS-LLITACLFLIGTVWAKEGYLVNKSTGCKYGYLKLGENEGCDKECKAKNQGGSYGYCYAFACWCEGLPESTPTYPLPNKSCGKK

D000601 MNSLLMITACLFLIGTVWAKEGYLVNKSTGCKYGCLKLGENEGCDKECKAKNQGGSYGYCYAFACWCEGLPESTPTYPLPNKSCGKK

D000049 MNSLLMITACLFLIGTVWAKEGYLVNKSTGCKYGCLLLGKNEGCDKECKAKNQGGSYGYCYAFGCWCEGLPESTPTYPLPNKSCSKK

D000138 ------ITACLVLIGTVCAKEGYLVNKSTGCKYNCLILGENKNCDMECKAKNQGGSYGYCYKLACWCEGLPESTPTYPIPGKTCRTK

D000630 ------------------KEGYMVNKSTGCSYSCPKTGESVYCDKECKAKNQGGSYGFCQYSNCWCEGLPESTPTWPLDDKPCD--

D000635 MNSLLMITACLVLFGTVWSEKGYLVHEDTGCRYKCTFSGENSYCDKECKS--QGGDSGICQSKACYCQGLPEDTKTWPLIGKLCGRK

D000670 MNSLLMIIGCLVLIGTVWTKEGYLVNMKTGCKYGCYELGDNGYCDRKCKA--ESGNYGYCYTVGCWCEGLPNSKPTWPLPGKSCSGK

D000252 ------------------KDGYLVNKSTGCKYSCIENINDSHCNEECISSIRKGSYGYCYKFYCYCIGMPDSTQVYPIPGKTCSTE

Subgroup 09d

DBACC

D000253 MNSLLMITACLILIGTVWAEDGYLFDKRKRCTLACIDKTGDKNCDRNCKKEGGSFGHCSYSACWCKGLPGSTPISRTPGKTCKK

D000636 MNSLLMITTCLILIGTVLAEDGYLFDKRKRCTLECIDKTGDKNCDRNCKNEGGSFGKCSYFACWCKGLPGITPISRTPGKTCKI

Group 10

DBACC

D000159 MKIIIFLIVCSFVLIGVKADNGYLLNKYTGCKIWCVINNESCNSECKLRRG-NYGYCYFWKLACYCEGAPKSELWAYETNKCNGKM

D000685 MKIIIFLIVSSLMLIGVKTDNGYLLNKATGCKVWCVINNASCNSECKLRRG-NYGYCYFWKLACYCEGAPKSELWAYATNKCNGKL

D000160 MKTVIFLIVSSLLLIGVKTDNGYLLDKYTGCKVWCVINNESCNSECKIRGG-YYGYCYFWKLACFCQGARKSELWNYNTNKCNGKL

D000062 ------------------EHGYLLNKYTGCKVWCVINNEECGYLCNKRRGGYYGYCYFWKLACYCQGARKSELWNYKTNKCD--L

D000090 ------------------EDGYLLNRDTGCKVSCG----TCRY-CND------------------------------------

Group 11

DBACC

D000036 MKLLLLLIVSASMLIESLVNADGYIRKRDGCKLSCLFGNEGCNKECKSYGGSYGYCWTWGLACWCEGLP-DEKTWKSE-TNTCG---

```
D000598 MKLLLLLIVSASMLIESLVNADGYIRKRDGCKVSCLFGNEGCDKECKSYGGSYGYCWTWGLACWCEGLP-DEKTWKSE-TNTCG---
D000597 MKLLLLLIVSASMLIESLVNADGYIRKRDGCKVSCLFGNEGCDKECKAYGGSYGYCWTWGLACWCEGLP-DDKTWKSE-TNTCG---
D000033 ----------------------DGYIRKRDGCKVSCLFGNEGCDKECKAYGGSYGYCWTWGLACWCEGLP-DDKTWKSE-TNTCG---
D000034 ----------------------DGYIRRRDGCKVSCLFGNEGCDKECKAYGGSYGYCWTWGLACWCEGLP-DDKTWKSE-TNTCG---
D000081 ----------------------DGYIRRRDGCKVSCLFGNEGCDKECKAYGGSYGYCWTWGLACWCEGLP-DDKTWKSE-TNTCG---
D000124 MKLLLLLIVSASMLIESLVNADGYIKRRDGCKVACLIGNEGCDKECKAYGGSYGYCWTWGLACWCEGLP-DDKTWKSE-TNTCGGKK
D000231 MKLLLLLIVSASMLIESLVNADGYIKRRDGCKVACLVGNEGCDKECKAYGGSYGYCWTWGLACWCEGLP-DDKTWKSE-TNTCGGKK
D000060 MKLLLLLVISASMLLECLVNADGYIRKKDGCKVSCIIGNEGCRKECVAHGGSFGYCWTWGLACWCENLP-DAVTWKSS-TNTCGRKK
D000146 ----------------------MDGYIRGSNGCKVSCLWGNEGCNKECRAYGASYGYCWTWGLACWCEGLP-DDKTWKSE-SNTCG---
D000272 ----------------------DGYIRGSNGCKVSCLWGNEGCNKECRAYGASYGYCWTWGLACWCEGLP-DDKTWKSE-SNTCG---
D000683 MKLFLLLLISASMLIDGLVNADGYIRGSNGCKVSCLWGNEGCNKECRAYGASYGYCWTWGLACWCQGLP-DDKTWKSE-SNTCGGKK
D000148 MKLFLLLVISASMLIDGLVNADGYIRGSNGCKVSCLWGNEGCNKECKAFGAYYGYCWTWGLACWCQGLP-DDKTWKSE-SNTCGGKK
D000675 MKLFLLLVISASMLIDGLVNADGYIRGSNGCKVSCLWGNEGCNKECKAFGAYYGYCWTWGLACWCEGLP-DDKTWKSE-SNTCGGKK
D000149 ----------------------MDGYIRGSNGCKISCLWGNEGCNKECKGFGAYYGYCWTWGLACWCEGLP-DDKTWKSE-SNTCG---
D000676 MKLFLLLVFFASMLIDGLVNADGYIRGSNGCKISCLWGNEGCNKECKGFGAYYGYCWTWGLACWCEGLP-DDKTWKSE-SNTCGGKK
D000623 MKLSLLLVISASMLIDGLVNADGYIRGSNGCKISCLWGNEGCNKECKGFGAYYGYCWTWGLACWCEGLP-DDKTWKSE-SNTCGGKK
D000678 MKLFLLLVISASMLIDGLVNADGYIRGSNGCKVSCLWGNEGCNKECGAYGASYGYCWTWGLACWCEGLP-DDKTWKSE-SNTCGGKK
D000271 MKLSLLLVISASMLIDGLVNADGYIRGSNGCKVSCLWGNDGCNKECRAYGASYGYCWTWGLACWCEGLP-DDKTWKSE-SNTCGGKK
D000080 ----------------------DGYIRGSDNCKVSCLLGNEGCNKECRAYGASYGYCWTVKLAQDCEGLP-DT---------------
D000158 MKLFLLLVISASMLIDGLVNADGYIRGSNGCKVSCLLGNEGCNKECRAYGASYGYCWTWKLACWCEGLP-DDKTWKSE-SNTCGGKK
D000204 ----------------------DGYIKGKSGCRVACLIGNQGCLKDCRAYGASYGYCWTWGLACWCEGLP-DNKTWKSE-SNTCG---
D000145 MKLFLLLVIFASMLNDGLVNADGYIRGSDGCKVSCLWGNDFCDKVCKKSGGSYGYCWTWGLACWCEGLP-DNEKWKYE-SNTCGSKK
D000256 ----------------------DGYILMRNGCKIPCLFGNDGCNKECKAYGGSYGYCWTYGLACACEGQPEDKKHLNYH-KKTC----
D000061 ----------------------DGYIRGGDGCKVSCVIDHVFCDNECKAAGGSYGYCWGWGLACWCEGLP-ADREWKYE-TNTCG---
D000220 ----------------------DGYIKRHDGCKVTCLINDNYCDTECKREGGSYGYCYSVGFACWCEGLP-DDKAWKSE-TNTCD---
D000232 MKLLLLLIITASMLIEGLVNADVYIRRHDGCKISCTVNDKYCDNECKSEGGSYGYCY--AFGCWCEGLP-NDKAWKSE-TNTCGGKK
D000221 ----------------------DGYIKGYKGCKITCVINDDYCDTECKAEGGTYGYCWKWGLACWCEDLP-EDKRWKPE-TNTC----
D000157 ----------------------DGYPKQKDGCKYSCTINHKFCNSVCKSNGGDYGYCWFWGLACWCEGLPDN-KMWKYE-TNTCG---
D000663 ----------------------DGYPKQKNGCKYDCIINNKWCNGICKMHGGYYGYCWGWGLACWCEGLPED-KKWWYE-TNKCGR--
D000637 --------------------ARDGYPVDEKGCKLSCLINDKWCNSACHSRGGKYGYCYTGGLACYCEAVPDNVKVWTYE-TNTC----
D000257 ----------------------DGYIKKSKGCKVSCVINNVYCNSMCKSLGGSYGYCWTYGLACWCEGLPNA-KRWKYE-TKTCK---
D000258 ----------------------DGYILNSKGCKVSCVVSIVYCNSMCKSSGGSYGYCWTWGLACWCEGLPNS-KRWTSS-KNKCN---
```

D000259 ----------DGYIKGNKGCKVSCVINNVFCNSMCKSSGGSYGYCWSWGLACWCEGLPAA-KKWLYAATNTCG---

Group 12 Subgroup 12a

DBACC

D000130 KDGYLMEPNGCKLGCLTRPAKYCWXEE--

D000131 KDGYLVGTDGCKYGCFTRPGHFCANEECL

D000132 KDGYLMGADGCKLCVLTAPYDYCACE---

Subgroup 12b

DBACC

D000064 ADGYVKGKSGCKISCFLDNDLCNADCKYYGGKLNSWCIPDKSGYCWCPNKGWNSIKSETNTC

Group 13

DBACC

D000153 MKAALLLVIFSLMLIGVLTKKSGYPTDHEGCKNWCVLNHSCGILCEGYGGSGYCYFWKLACWCDDIHNWVPTWSRATNKCRAK

Group 14

DBACC

D000264 MNYLVMISFALLLVIGVESVRDGYFVEPDNCVVHCMPSSEMCDRGCKHNGATSGSCKAFSKGGNACWCKGLR

D000266 MNYLVMISFALLLVIGVESVRDGYFVEPDNCVIYCMPSSEVCDRGCKHNGATSGTCKEFSKGGNVCWCKGLR

D000265 MNYLVMISFALLLVIGVESVRDGYFVEPDNCLVYCMPSPEICDRGCKRYGATSGFCKEFSKGENFCWCKGLR

Group 15

DBACC

D000833 ADVPGNYPLDKDGNTYTCLELGENKDCQKVCKLHGVQYGYCYAFSCWCKEYLDDKD-SV

D000834 ADVPGNYPLDKDGNTYTCLELGENKDCQKVCKLHGVQYGYCYAFFCWCKE-LDDKDVSV

D000701 ADVPGNYPLDKDGNTYKCLKLGENKDCQKVCKLHGVQYGYCYAFECWCKEYLDDKD-SV

D000277 ADVPGNYPLDKDGNTYKCFLLGGNEECLNVCKLHGVQYGYCYASKCWCEYLEDDKD-SV

D000684 ADVPGNYPLDKDGNTYKCFLLGENEECLNVCKLHGVQYGYCYASKCWCEYLEDDKD-SV

Group 16

DBACC

D000819 MKTIPLLFLLFIYFECDGKFIRHKDESFYECGQLIGYQQYCVDACQAHGSKEKGYCKGMAPFGLPGGCYCPKLPSNRVKMCFGALESKCA

D000821 ------------------KFIRHKDESFYECGQLIGYQQYCVNACQAHGSKEKGYCKGMAPFGLPGGCYCPKLPSNRVKMCFGALESKCA

D000820 ------------------KFIRHKDESFYECGQSIGYQQYCVDACQAHGSKEKGYCKAMAPFGLPGGCYCPKLPSNRVKMCFGALESKCA

Group 17

DBACC

D000681 MVKMQVIFIAFIAVIACSMVYGDSLSPWNEGDTYYGCQRQTDEFCNKICKLHLASGGSCQQPAPFVKLCTCQGIDYDNSFFFGALEKQCPKLRE

Group 18

DBACC

D000702 RDGYPLASNGCKFGCSGLGENNPTCNHVCEKKAGSDYGYCYAWTCYCEHVAEGTVLWGDSGTGPCRS

Potassium channel toxins: Alpha, Beta, Gamma and Delta Ktx

Alpha KTx 01

DBACC

D000100 MKILSVLLLALIICSIVGWSEAQFTNVSCTTSKECWSVCQRLHNTSRGKCMNKKCRCYS

D000241 MKILSVLLLALIICSIVGWSEAQFTDVSCTTSKECWSVCQRLHNTSIGKCMNKKCRCYS

D000242 ----------------------QFTNVSCTTSKECWSVCEKLYNTSRGKCMNKKCRCYS

D000102 ----------------------QFTQESCTASNQCWSICKRLHNTNRGKCMNKKCRCYS

D000117 MKISFLLLLAIVICSIG-WTEAQFTNVSCSASSQCWPVCKKLFGTYRGKCMNSKCRCYS

D000097 ----------------------QFTDVDCSVSKECWSVCKDLFGVDRGKCMGKKCRCYQ

D000118 MKISF-LLLALVICSIGWSEAQFTDVKCTGSKQCWPVCKQMFGKPNGKCMNGKCRCYS

D000195 ----------------------KFIDVKCTTSKECWPPCKAATGKAAGKCMNKKCKCQ-

Alpha KTx 02

DBACC

D000197 KVIDVKCTSPKQCLPPCKAQFGD----------------

D000198 TVIDVKCTSPKQCLPPCAKQ--------------------

D000194 TVIDVKCTSPKQCLPPCKAQFGIRAGAKCMNGKCKCYPH

D000067 TVIDVKCTSPKQCLPPCKEIYGRHAGAKCMNGKCKC---

D000066 TIINVKCTSPKQCLPPCKAQFGQSAGAKCMNGKCKCYPH

D000196 TFINVKCTSPKQCLPACKEKFGX-AAGKCMNGKCK----

D000065 ITINVKCTSPQQCLRPCKDRFGQHAGGKCINGKCKCYP-

D000089 TIINVKCTSPKQCSKPCKELYGSSAGAKCMNGKCKCYNN

D000175 TIINEKCFATSQCWTPCKKAIGS-LQSKCMNGKCKCYNG

Alpha KTx 03

DBACC

D000083 ----------GVPINVSCTGSPQCIKPCKDAGMRFGKCMNRKCHCTPK-

D000084 ----------GVPINVPCTGSPQCIKPCKDAGMRFGKCMNRKCHCTPK-

D000200 ----------GVPINVKCRGSPQCIQPCRDAGMRFGKCMNGKCHCTPQ-

D000082 ----------GVPINVKCTGSPQCLKPCKDAGMRFGKCINGKCHCTPK-

D000085 ----------GVEINVKCSGSPQCLKPCKDAGMRFGKCMNRKCHCTPK-

D000086 ----------GVIINVKCKISRQCLEPCKKAGMRFGKCMNGKCHCTPK-

D000087 MKVFSAVLIILFVCSMIIGINAVRIPVSCKHSGQCLKPCKDAGMRFGKCMNGKCDCTPK-

D000617 ----------VGIPVSCKHSGQCIKPCKDAGMRFGKCMNRKCDCTPK-

D000119 MKVFFAVLITLFICSMIIGIHGVGINVKCKHSGQCLKPCKDAGMRFGKCINGKCDCTPKG

D000173 MKVFFAVLITLFISSMIIGIHGVGINVKCKHSGQCLKPCKDAGMRFGKCINGKCDCTPKG

Alpha KTx 04

DBACC

D000091 ----------VFINVKCRGSPECLPKCKEAIGKAAGKCVN--------

D000092 ----------VFINVKCRGSPECLPKCKEAFGKAAGKCVN--------

D000093 ----------VFINVKCRGSPECLPKCKEAIGKAAGKCMN--------

D000088 ----------VFINAKCRGSPECLPKCKEAIGKAAGKCMNGKCKCYP-

D000147 ----------VFINVKCTGSKQCLPACKAAVGKAAGKCMNGKCKCYT-

D000699 ----------VFINVKCRGSKECLPACKAAVGKAAGKCMNGKCKCYP-

D000216 ----------EVDMRCKSSKECLVKCKQATGRPNGKCMNRKCKCYPR

D000095 MKVLYGILIIFILCSMFYLSQEVVIGQRCYRSPDCYSACKKLVGKATGKCTNGRCDC---

Alpha KTx 05

DBACC

D000107 ----------TVCNLRRCQLSCRSLGLLGKCIGVKCECVKH--

D000168 MHNYYKIVLIMVAFFAVIITFSNIQVEGAVCNLKRCQLSCRSLGLLGKCIGDKCECVKHGK

D000103 ----------AFCNLRMCQLSCRSLGLLGKCIGDKCECVKH--

D000248 ----------AFCNLRRCELSCRSLGLLGKCIGEECKCVPY--

D000249 ----------AFCNLRRCELSCRSLGLLGKCIGEECKCVPH--

Alpha KTx 06a

DBACC

D000190 -ASCRTPKDCADPCRKETGCPYGKCMNRKCKCNRC-

D000818 QKECTGPQHCTNFCRKNK-CTHGKCMNRKCKCFNCK

D000094 -VSCTGSKDCYAPCRKQTGCPNAKCINKSCKCYGC-

D000096 LVKCRGTSDCGRPCQQQTGCPNSKCINRMCKCYGC-

Alpha KTx 06b

DBACC

D000191 ----------IEAIRCGGSRDCYRPCQKRTGCPNAKCINKTCKCYGCS

D000192 ----------DEAIRCTGTKDCYIPCRYITGCFNSRCINKSCKCYGCT

D000592 MNAKFILLL-VLTTMMLLPDTKGAEVIRCSGSKQCYGPCKQQTGCTNSKCMNKVCKCYGCG

D000593 MNAKFILLLLVVTTTTLLPDAKGAEIIRCSGTRECYAPCQKLTGCLNAKCMNKACKCYGCV

D000595 MNAKFILLLLVVTTTMLLPDTQGAEVIKCRTPKDCADPCRKQTGCPHGKCMNRTCRCNRCG

D000596 MNAKFILLLLVVATTMLLPDTQGAEVIKCRTPKDCAGPCRKQTGCPHGKCMNRTCRCNRCG

D000594 MNAKFILLLLVVTTTILLPDTQGAEVIKCRTPKDCADPCRKQTGCPHAKCMNKTCRCHRCG

Alpha KTx 06c

DBACC

D000829 MKVAYLLVLFTIMMLANDASLVHTNIPCRGTSDCYEPCEKKYNCARAKCMNRHCNCYNNCPWR

Alpha KTx 07

DBACC

D000098 ----------TISCTNEKQCYPHCKKETGYPNAKCMNRKCKCFGR

D000099 RGSVDYKDDDDKTISCTNPKQCYPHCKKETGYPNAKCMNRKCKCFGR

Alpha KTx 08

DBACC

D000113 ----------VSCEDCPDHCSTQKARAKCDNDKCVCEPI

D000114 ----------VSCEDCPDHCSTQKARAKCDNDKCVCEPK

D000112 ----------VSCEDCPEHCSTQKAQAKCDNDKCVCEPI

D000169 MSRLYAIILIALVFNVVMTITPDMKVEAATCEDCPEHCATQNARAKCDNDKCVCEPK

D000263 MSRLYAIILIALVFNVIMTIIPDMKVEAATCEDCPEHCATQNARAKCDNDKCVCEPK

Alpha KTx 09

DBACC

D000174 MIVLFTLVLIVLAMNVTMAIISDPVVEAVGCEECPMHCKGKNANPTCDDGVCNCNV----

D000693 ----------VGCEECPMHCKGKHAVPTCDDGVCNCNV----

D000167 MSRLFTLVLIVLAMNVMMAIISDPVVEAVGCEECPMHCKGKNANPTCDDGVCNCNV----

D000170 MSRLFTLVLIVLAMNVMMAIISDPVVEAVGCEECPMHCKGKNAKPTCDDGVCNCNV----

D000115 ----------VGCEECPMHCKGKNAKPTCDNGVCNCNV----

D000116 ----------VGCEEDPMHCKGKQAKPTCCNGVCNCNV----

D000687 ----------VGCAECPMHCKGKMAKPTCENEVCKCNIGKKD

Alpha KTx 10

DBACC

D000129 MEGIAKITLILLFLFVTMHTFANWNTEAAVCVYRTCDKDCKRRGYRSGKCINNACKCYPYGK

D000208 ----------VACVYRTCDKDCTSRKYRSGKCINNACKCYPY--

Alpha KTx 11

DBACC

D000214 DEEPKESCSDEMCVIYCKGEEYSTGVCDGPQKCKCSD

D000215 DEEPKETCSDEMCVIYCKGEEYSTGVCDGPQKCKCSD

D000686 DEEPKETCSDDMCVIYCKGEEFSTGACDGPQKCKCS-

Alpha KTx 12

DBACC

D000193 WCSTCLDLACGASRECYDPCFKAFGRAHGKCMNNKCRCYT

D000700 WCSTCLDLACGASRECYDPCFKAFGRAHGKCMNNKCRCYT

Alpha KTx 13

DBACC

D000202 ACGSCRKKCKGSGKCINGRCKCY

Alpha KTx 14

DBACC

D000243 MKIFFAILLILAVCSMAIWTVNGTPFAIKCATDADCSRKCPGNPSCRNGFCACT

D000682 MKIFFAILLILAVCSMAIWTVNGTPFAIKCATNADCSRKCPGNPPCRNGFCACT

D000274 MKIFFAILLILAVCSMAIWTVNGTPFAIKCATDADCSRKCPGNPPCRNGFCACT

D000262 MKIFFAILLILAVCSMAIWTVNGTPFEVRCATDADCSRKCPGNPPCRNGFCACT

Alpha KTx 15

DBACC

D000183 MKFSSIILLTLLICSMSIFGNCQIETNKKCQGGS-CASVCR-RVIGVAAGKCINGRCVCYP-

D000689 ----------QIETNKKCQGGS-CASVCR-KVIGVAAGKCINGRCVCYP-

D000624 MKFSSIILLTLLICSMSIFGNCQVETNKKCQGGS-CASVCR-RVIGVAAGKCINGRCVCYP-

D000618 MKFSSIILLTLLICSMSIFGNCQVQTNVKCQGGS-CASVCR-REIGVAAGKCINGKCVCYRN

D000835 MKFSSIILLTLLICSMSIFGNGQVQTNKKCKGGS-CASVCA-KEIGVAAGKCINGRCVCYP-

D000269 MKFSSIILLTLLICSMSKFGNCQVETNVKCQGGS-CASVCR-KAIGVAAGKCINGRCVCYP-

D000826 ----------QIDTNVKCSGSSKCVKICIDRYYNTRGAKCINGRCTCYP-

Alpha KTx16

DBACC

D000152 MKIFSILLVALIICSISICTEAFGLIDVKCFASSECWTACKKVTGSGQGKCQNNQCRCY

D000620 MKIFSILLVALIICSISICTEAFGLIDVKCFASSECWIACKKVTGSVQGKCQNNQCRCY

D000201 ----------DLIDVKCISSQECWIACKKVTGRFEGKCQNRQCRCY

D000101 ----------GLIDVRCYDSRQCWIACKKVTGSTQGKCQNKQCRCY

D000108 ----------GLIDVRCYDSSQCE----------

Alpha KTx 17

DBACC

D000276 MKFIIVLILISVLIATIVPVNEAQTQCQSVRDCQQYCLTPDRCSYGTCYCKTTGK

Alpha KTx 18

DBACC

D000698 TGPQTTCQAAMCEAGCKGLGKSMESCQGDTCKCKA

Alpha KTx 19

DBACC

D000613 AACYSSDCRVKCVAMGFSSGKCINSKCKCYK

Alpha KTx 20

DBACC

D000273 ACGPGCSGSCRQKGDRIKCINGSCHCYP

Alpha KTx 21

DBACC

D000270 MNRLTTIILMLIVINVIMDDISESKVAAGIVCKVCKIICGMQGKKVNICKAPIKCKCKKG

Alpha KTx 22

DBACC

D000250 RCHFVVCTTDCRRNSPGTYGECVKKEKGKECVCKS

D000251 RCHFVICTTDCRRNSPGTYGECVKKEKGKECVCKS

Alpha KTx 23

DBACC

D000830 -DPCYEVCLQQHGNVKECEEACKHPVE-

D000831 -DPCYEVCLQQHGNVKECEEACKHPVEY

D000832 NDPCEEVCIQHTGDVKACEEACQ-----

Alpha KTx 24

DBACC

D000166 CQNECCGISSLRERNYCANLVCINCFCQGRTYKICRCFFSIHAIR

Alpha KTx 25

DBACC

D000677 MQKLFIVFVLFCILRLDAEVDGKTATFCTQSICQESCKRQNKNGRCVIEAEGSLIYHLCKCY

Beta-KTx

Beta KTx 01

DBACC

D000187 MQRNLVVLLFLGMVALSSCGLREKHFQKLVKYAVPEGTLRTIIQTAVHKLGKTQFGCPAYQGYCDDHCQDIKKEEGFCHGFKCKCGIPMGF

D000189 MQRNLVVLLFLGMVALSSCGLREKHVQKLVKYAVPVGTLRTILQTVVHKVGKTQFGCPAYQGYCDDHCQDIKKEEGFCHGFKCKCGIPMGF

D000188 --RKLALLLILGMVTLASCGLREKHVQKLVA-LIPNDQLRSILKAVVHKVAKTQFGCPAYEGYCNDHCNDIERKDGECHGFKCKCAKD---

Beta KTx 02

DBACC

D000186 MMKQQFFLFLAVIVMISSVIEAGRGKEIMKNIKEKLTEVKDKMKHSWNKLTSMSEYACPVIEKWCEDHCAAKKAIGKCEDTECKCLKLRK

Beta KTx 03

DBACC

D000823 DNGYLLNKYTGCKIWCVINNESCNSECKLRRGNYGYCYFWKLACYCEGAPKSELWAYETNKCNGKM

Gamma-KTx

Gamma KTx01

DBACC

D000247 MKISFVLLLTLFICSIGWSEARPTDIKCSESYQCFPVCKSRFGKTNGRCVNGFCDCF-

D000619 MKISFVLLLTLFICSIGWSEARPTDIKCSASYQCFPVCKSRFGKTNGRCVNGLCDCF-

Gamma KTx02

DBACC

D000647 DRDSCVDKSKCSKYGYYGQCDECCKKAGDRAGNCVYFKCKCNP

D000653 DRDSCVDKSKCSKYGYYGQCDKCCKKAGDRAGNCVYFKCKCNQ

D000657 DRDSCVDKSKCAKYGYYGQCDECCKKAGDRAGNCVYLKCKCNQ

D000655 DRDSCVDKSRCGKYGYYGQCDDCCKKAGDRAGTCVYYKCKCNP

D000659 DRDSCVDKSRCGKYGYYGQCDECCKKAGDRAGTCVYYKCKCNP

D000658 DRDSCVDKSQCGKYGYYGQCDECCKKAGERVGTCVYYKCKCNP

D000654 DRDSCVDKSKCGKYGYYGQCDECCKKAGDRAGTCVYYKCKCNP

D000662 ERDSCVEKSKCGKYGYYGQCDECCKKAGDRAGTCVYYKCKCNP

D000650 DRDSCVDKSKCGKYGYYHQCDECCKKAGDRAGNCVYYKCKCNP

D000645 DRDSCVDKSKCGKYGYYGQCDECCKKAGDRAGICEYYKCKCNP

D000651 DRDSCVDKSKCAKYGYYYQCDECCKKAGDRAGTCEYFKCKCNP

D000656 DRDSCVDKSQCAKYGYYYQCDECCKKAGDRAGTCEYFKCKCNP

Gamma KTx03

DBACC

D000648 --------------------GRDSCVNKSRCAKYGYYSQCEVCCKKAGHKGGTCDFFKCKCKV----

D000649 --------------------DRDSCVDKSRCAKYGYYGQCEVCCKKAGHRGGTCDFFKCKCKV----

D000644 --------------------DRDSCVDKSRCAKYGYYQQCEICCKKAGHRGGTCEFFKCKCKV----

D000661 --------------------DRDSCVDKSRCAKYGYYGQCEVCCKKAGHNGGTCMFFKCMCVNSKMN

D000639 --------------------DRDSCVDKSRCAKYGYYQECTDCCKKYGHNGGTCMFFKCKCA-----

D000641 --------------------DRDSCVDKSRCAKYGHYQECTDCCKKYGHNGGTCMFFKCKCA-----

D000218 MKVLILIMIIASLMIMGVEMDRDSCVDKSRCAKYGYYQECQDCCKNAGHNGGTCMFFKCKCA-----

D000219 --------------------DRDSCVDKSRCAKYGYYQECQDCCKNAGHNGGTQMFFKCKAP-----

D000660 --------------------DRDSCVDKSKCGKYGYYQECQDCCKNAGHNGGTCVYYKCKCNP----

D000642 --------------------DRDSCVDKSRCSKYGYYQECQDCCKKAGHNGGTCMFFKCKCA-----

D000640 --------------------DRDSCVDKSRCAKYGYYQECQDCCKKAGHSGGTCMFFKCKCA-----

D000643 --------------------DRDSCVDKSRCAKYGYYQECQDCCKKAGHNGGTCMFFKCKCA-----

D000646 --------------------DRDSCVDKSRCQKYGNYAQCTACCKKAGHNKGTCDFFKCKCT-----

D000652 --------------------DRDSCVDKSRCQKYGPYGQCTDCCKKAGHTGGTCIYFKCKCGAESGR

Delta-KTx

Delta KTx01

DBACC

D000260 GHACYRNCWREGNDEETCKERC-

D000261 GHACYRNCWREGNDEETCKERCG

D000816 GFGCYRSCWKAGHDEETCKRECS

Calcium channel toxins

Group 01

DBACC

D000610 MKPSLIIVTFIVVFMAISCVAADDEQETWIEKRGDCLPHLKRCKENNDCCSKKCKRRGTNPEKRCR

D000703 MKPSLIIVTFIVVFMTISCVAADDEQETWIEKRGDCLPHLKRCKENNDCCSKKCKRRGANPEKRCR

D000178 ----------GDCLPHLKLCKENKDCCSKKCKRRGTNIEKRCR

D000185 ----------GDCLPHLKRCKADNDCCGKKCKRRGTNAEKRCR

Group 02

DBACC

D000135 MHTPKHAIQRISKEEMEFFEGRCERMGEADETMWGTKWCGSGNEATDISELGYWSNLDSCCRTHDHCDNIPSGQTKYGLTNEGK

YTMMNCKCETAFEQCLRNVTGGMEGPAAGFVRKTYFDLYGNGCYNVQCPSQRRLARSEECPDGVATYTGEAGYGAWAINKLNG

Group 03

DBACC

D000180 KIDGYPVDYWNCKRICWYNNKYCNDLCKGLKADSGYCWGWTLSCYCQGLPDNARIKRSGRCRA

Group 04

DBACC

D000827 VGCEECPAHCKGKNAKPTCDDGVCNCNV

D000828 VGCEECPAHCKGKNAIPTCDDGVCNCNV

Chloride channel toxins

Group 01 Subgroup 01a

DBACC

D000111 ----------CGPCFTTDPYTESKCATCCGGRGKCVGPQCLCNRI-

D000171 ----------CGPCFTKDPETEKKCATCCGGIGRCFGPQCLCNRGY

D000238 MKFLYGIAFIAVFLTVMIVTDIEACGPCFTTDHQTEQKCAECCGGIGKCYGPQCLC-RG-

D000240 MKFLYGIAFIAVFLTVMIVTDIEACGPCFTTDHQTEQKCAECCGGIGKCYGPQCLCNRG-

D000239 MKFLYGIAFIAVFLTVMIATHIEACGPCFTTDRQMEQKCAECCGGIGKCYGPQCLC-RG-

D000104 ----------MCMPCFTTDPNMAKKCRDCCGGNGKCFGPQCLCNR--

D000206 ----------MCMPCFTTDHNMAKKCRDCCGGNGKCFGPQCLCNR--

D000207 ----------MCMPCFTTDPNMANKCRDCCGGGKKCFGPQCLCNR--

D000110 ----------MCMPCFTTRPDMAQQCRACCKGRGKCFGPQCLCGYD-

D000154 MKFLYGIVFIALFLTVMFATQTDGCGPCFTTDANMARKCRECCGGIGKCFGPQCLCNRI-

D000615 MKFLYGIVFIALFLTVMFATQTDGCGPCFTTDANMARKCRECCGGNGKCFGPQCLCNRE-

D000199 ----------RCKPCFTTDPQMSKKCADXCGGX-KX----------

Subgroup 01b

DBACC GenBank

D000105 ----------RCSPCFTTDQQMTKKCYDCCGGKGKGKCYGPQCICAPY

D000109 S06667 ----------RCKPCFTTDPQMSKKCADCCGGKGKGKCYGPQCLC---

D000106 A48850 ----------MCMPCFTTDHQMARKCDDCCGGKGRGKCYGPQCLCR--

D000237 MKFLYGIVFIALFLTVMIATHTEAMCMPCFTTDHQMARKCDDCCGGKGRGKCYGPQCLCRG-

D000205 P60268 ----------MCMPCFTTDHQTARRCRDCCGGRGR-KCFG-QCLCGYD

D000140 ----------RCGPCFTTDPQTQAKCSECCGRKGG-VCKGPQCICGIQ

D000629 AF481880 MKFLYGTILIAFFLTVMIATHSEARCPPCFTTNPNMEADCRKCCGGRGY--CASYQCICPGG

Defensin scorpion toxins

Group 01

DBACC GenBank

D000217 AJ292361 MNSKLTALIFLGLIAIAYCGWINEEKIQKKIDERMGNTVLGGMAKAIVHKMAKNEFQCMANMDMLGNCEK

HCQTSGEKGYCHGTKCKCGTPLSY

Short Chain Neurotoxins
Group 01

DBACC GenBank

D000128 VTMGYIKDGDGKKIAKKKNKNGRKHVEIDLNKVG

Group 02

DBACC GenBank

D000664 VSIGIKCDPSIDLCEGQCRIRYFTGYCSGDTCHCS

Group 03

DBACC GenBank

D000667 AY147405 MKIFFAVLVILVLFSMLIWTAYGTPYPVNCKTDRDCVMCGLGISCKNGYCQSCTR

D000668 AY125328 MKIFFAVLVILVLFSMLIWTAYGTPYPVNCKTDRDCVMCGLGISCKNGYCQGCTR

Group 04

DBACC GenBank

D000679 AF159979 MEIKYLLTVFLVLLIVSDHCQAFLFSLIPSAISGLISAFKGRRKRDLNG

Group 05

DBACC GenBank

D000680 AF159977 ENLGEDCENLCKQQKATDGFCRQPHCFCTDMPDNYATRPDTVDPIM

Group 06

DBACC

D000691 TVKCGGCNRKCCAGGCRSGKCINGKCQCY

D000692 TVKCGGCNRKCCPGGCRSGKCINGKCQCY

D000690 KPKCGLCRYRCCSGGCSSGKCVNGACDCS

Appendix 2

The 20 natural amino acids have different physicochemical properties which include both physical (e.g. length of R-group side chain, branch or single chain etc.) and chemical (e.g. hydrophobicity, charged or uncharged groups etc.) characteristics (**Figure 37**). A single amino acid residue can be described by various overlapping physical (**Table 13**) and chemical properties. Studying their physicochemical properties can help in analysis of structure-function relationships, particularly in mutation studies. For example, neutralisation of K27 in Kaliotoxin resulted in low binding affinity towards K^+ channels which demonstrated that positive charged is important for toxin-channel interaction (Aiyar *et al.*, 1995).

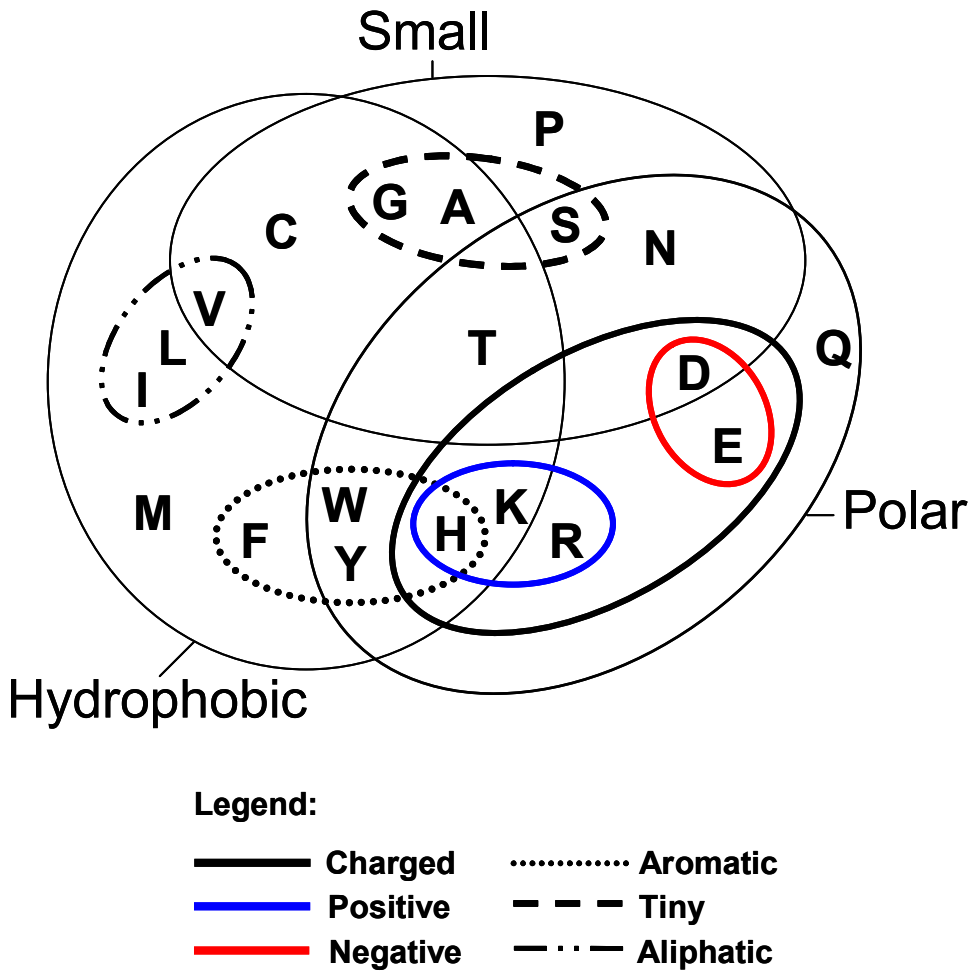

Figure 37 Venn diagram displaying the interrelationships of the 20 naturally occurring amino acids based on their physicochemical properties. Glycine whose side-chain group is a hydrogen atom can fit into either hydrophobic or hydrophilic environment. Proline is the only cyclic amino acid which usually cause kinks in polypeptide chains. Adapted from Taylor, 1986.

Table 13 Physical properties of the 20 L-α-amino acids. * also called aspartatic acid, # also known as glutamic acid, ‡ molecular weights given are those of the neutral, free amino acids; residue weights are obtainable by subtraction of one equivalent of water molecule (18 g/mol). § measures the relative hydrophobicity among amino acids where positive value indicates hydrophobicity while negative value represents hydrophilicity based on (Kyte and Doolittle, 1982).

Name of α-amino acid	Symbol		Mol. Wt.‡	pI value	Hydropathy Index§
	3-Letter	1-Letter			
Alanine	Ala	A	89.09	6.00	1.8
Arginine	Arg	R	174.20	11.15	-4.5
Asparagine	Asn	N	132.12	5.41	-3.5
Aspartate*	Asp	D	133.10	2.77	-3.5
Cysteine	Cys	C	121.15	5.02	2.5
Glutamine	Gln	Q	146.15	5.65	-3.5
Glutamate#	Glu	E	147.13	3.22	-3.5
Glycine	Gly	G	75.07	5.97	-0.4
Histidine	His	H	155.16	7.47	-3.2
Isoleucine	Ile	I	131.18	5.94	4.5
Leucine	Leu	L	131.18	5.98	3.8
Lysine	Lys	K	146.19	9.59	-3.9
Methionine	Met	M	149.21	5.74	1.9
Phenylalanine	Phe	F	165.19	5.48	2.8
Proline	Pro	P	115.13	6.30	-1.6
Serine	Ser	S	105.09	5.68	-0.8
Threonine	Thr	T	119.12	5.64	-0.7
Tryptophan	Trp	W	204.23	5.89	-0.9
Tyrosine	Tyr	Y	181.19	5.66	-1.3
Valine	Val	V	117.15	5.96	4.2