Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Molecular Graphics and Modelling 24 (2005) 17–24
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Accurate prediction of scorpion toxin functional properties from primary structures

Paul T.J. Tan [a,b], K.N. Srinivasan [c], Seng Hong Seah [a], Judice L.Y. Koh [a], Tin Wee Tan [b], Shoba Ranganathan [a,b,d], Vladimir Brusic [a,e,*]

[a] *Laboratories for Information Technology, Knowledge Discovery Department, Institute for Infocomm Research, I^2R, 21 Heng Mui Keng Terrace, Singapore 119613, Singapore*
[b] *Department of Biochemistry, Faculty of Medicine, National University of Singapore, Singapore*
[c] *Division of Biomedical Sciences, Johns Hopkins in Singapore, Singapore*
[d] *Biotechnology Research Institute, Macquarie University, Sydney, Australia*
[e] *School of Land and Food Sciences and the Institute for Molecular Bioscience, University of Queensland, Brisbane, Qld, Australia*

Received 3 May 2004; accepted 31 January 2005
Available online 9 June 2005

Abstract

Scorpion toxins are common experimental tools for studies of biochemical and pharmacological properties of ion channels. The number of functionally annotated scorpion toxins is steadily growing, but the number of identified toxin sequences is increasing at much faster pace. With an estimated 100,000 different variants, bioinformatic analysis of scorpion toxins is becoming a necessary tool for their systematic functional analysis. Here, we report a bioinformatics-driven system involving scorpion toxin structural classification, functional annotation, database technology, sequence comparison, nearest neighbour analysis, and decision rules which produces highly accurate predictions of scorpion toxin functional properties.

Keywords: Prediction systems; Protein function; Protein classification; Structure–function relationship

1. Introduction

Scorpion toxins are used as research tools for probing physiological function and structure of ion channels [1–5]. Scorpion toxins that exhibit pharmacological activity are used in preparation of antitoxins [6], and studied as potential drug targets [7,8]. Insect-specific toxins are studied as pest control agents [9].

Venom of an individual scorpion contains 50–100 different toxins [10]. Approximately 100,000 different toxins exist across 1500 scorpion species. In contrast, only about 300 distinct toxins from 30 species of scorpion have been characterized and published. Classification of scorpion toxins is important for understanding their structural properties. Criteria used for their classification are ion channel specificity, peptide length, structural properties, and toxin cellular target specificity.

Tytgat et al. [11] classified 49 potassium (K^+) specific toxins into 12 subfamilies by analyzing sequence similarity. Possani et al. [12] analyzed 202 scorpion toxins specific for K^+, sodium (Na^+), calcium (Ca^{2+}), or chloride (Cl^-) channels where K^+ and Cl^- toxins form 14 subfamilies, and Na^+ and Ca^{2+} toxins form 12 subfamilies. Long-chain scorpion toxins containing 60–70 amino acids target Na^+ channels [13]. Short-chain toxins containing 30–40 amino acids recognize K^+ and Cl^- channels [12]. A small number of scorpion toxins of variable peptide lengths affect Ca^{2+} channels. The ability of scorpion chlorotoxin to block Cl^- channels is controversial [14,15], but they form a separate group. Finally, scorpion toxins can be classified according to the species they target (insect, arthropod, crustacean, or mammal). Some toxins exhibit cross-specificity e.g. [16].

Methods for determining function of scorpion toxins include experimental studies of naturally occurring peptides,

* Corresponding author. Tel.: +65 6874 7920; fax: +65 6774 8056.
E-mail address: vladimir@i2r.a-star.edu.sg (V. Brusic).

1093-3263/$ – see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2005.01.003

site-directed mutagenesis, or chemically modified variants. Experimentation is typically supported by computational methods of sequence comparison [11,12], or by the analysis of three-dimensional (3-D) structures of scorpion toxins [17]. Bioinformatic analyses can improve the efficacy of research by indicating critical experiments [18,19]. Bioinformatic approaches involve access to scorpion toxin data across multiple databases, analysis and classification of scorpion toxin sequences and their structures, and the design and use of predictive models for simulation of laboratory experiments.

Recently, we created the SCORPION molecular database [20] that contains fully referenced entries of scorpion toxins. We proposed a new hierarchical classification scheme, underpinned by sequence, structural and functional similarity of sequences. We defined 34 groups of scorpion toxins that have high primary structure (sequence) identity. These groupings were used as a basis for the development of a bioinformatics-based approach for prediction of functional properties of scorpion toxins including toxin type, toxin action, target receptor and target cells (insect-, crustacean-, or mammal-specific). This method termed ''Annotate Scorpion'' combines sequence comparison, nearest neighbour analysis and decision rules. The Annotate Scorpion has been implemented in the SCORPION database (http://research.i2r.a-star.edu.sg/scorpion). The testing of the Annotate Scorpion showed that it can predict functional properties of scorpion toxins with high accuracy. Hereby we describe the method and the algorithm, and present the results of the assessment of prediction accuracy. The methodology reported in this article is of importance because it provides a bioinformatics basis for systematic functional analysis of large sets of toxin sequences.

2. Materials and methods

Scorpion toxin data were extracted from public databases and literature, cleaned of errors and enriched with functional information. The primary toxin set was analysed and classified into groups based on primary structure similarity. These groups were used as comparison targets by the ''Annotate Scorpion'' method for prediction of anonymous protein sequences. The secondary data set, containing newly characterized scorpion toxins, was used as a test set for evaluating its prediction accuracy.

2.1. Scorpion toxin data

The 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, while 24 partially sequenced toxins were not analysed, but were later added to the groups. The test set of 52 sequences was obtained from the secondary collection of scorpion toxin sequences where 18 were functionally characterized for their receptor and target cell specificities, and toxin action and toxin type by experimentation, whereas 34 sequences were partially characterized (Fig. 1).

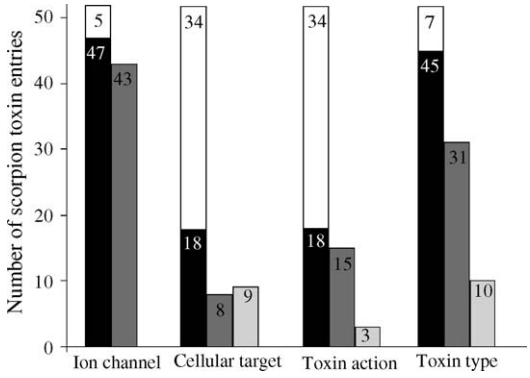

Fig. 1. Accurate prediction of scorpion toxin functions by the Annotate Scorpion module. The four functional properties of 52 scorpion toxin sequences from the test set were extracted from literature and database annotations. (■) Scorpion toxin having experimentally determined functional properties; (□) functionally uncharacterized; (▩) Correct prediction and (▨) partially correct prediction of functional properties.

2.2. Sequence analysis for grouping scorpion toxin data

The majority of the 220 scorpion toxin sequences were categorized into four broad families according to their ion channel specificity, namely Na^+, K^+, Ca^{2+} and Cl^-. One sequence belonged to the scorpion defensin family while the molecular target of one sequence was not reported. Within each family, we further classified sequences into smaller groups using BLAST 2.0 [21] and CLUSTAL W 1.74 [22] programs.

A representative sequence in each family was used to perform a similarity search against the non-redundant (nr) database at NCBI using BLAST search (http://www.ncbi.nlm.nih.gov/BLAST). Next, we analysed the expectation (E) values of the resulting sequences where low E values represent significant similarities. If there was a marked increase in the E value between two consecutive sequences, the larger E value would serve as a cut-off between close homologues and other sequences. Sequences with E-values below this cut-off were clustered into a preliminary group. We confirmed this grouping by searching the sequence closest to the cut-off value against the nr database and confirming the grouping (Fig. 2). The grouped sequences were removed from the list and the process was repeated with subsequent sequences until the list was exhausted.

Within each group, multiple sequence alignment was performed using CLUSTAL W. Sequences with distinct primary structure patterns were further classified into subgroups. We verified these groupings by phylogenetic analysis. Based on the initial alignment of consecutive groups, a resample was performed by the generation of 100

(A)

Sequences producing significant alignments:	Score (bits)	E Value
gi|6094296|sp|P56637|SIXE_BUTJU Excitatory insect toxin BJX...	146	4e-35
gi|3858953|emb|CAA09988.1| (AJ012313) neurotoxin, variant 3...	145	1e-34
gi|4140001|pdb|1BCG| Scorpion Toxin Bjxtr-It	138	1e-32
gi|3649606|gb|AAC61256.1| (AF055672) insect neurotoxin prec...	86	1e-16
gi|3914986|sp|O61668|SIX1_MESMA Excitatory insect selective...	86	1e-16
gi|58309|emb|CAA41265.1| (X58376) Hector insect toxin [Andr...	78	2e-14
gi|3036821|emb|CAA76604.1| (Y17050) neurotoxin (KIT) [Buthu...	76	8e-14
gi|161147|gb|AAA29950.1| (M27705) neurotoxin AaH IT1 [Andro...	74	4e-13
gi|134372|sp|P01497|SIX1_ANDAU Insect toxin 1 precursor (In...	74	5e-13
gi|134375|sp|P15147|SIX2_ANDAU Insect toxin 2 precursor (In...	70	4e-12
gi|69545|pir||XISR1A insect toxin 1 - Sahara scorpion >gi|2...	70	5e-12
gi|134340|sp|P19856|SIX1_LEIQU Insect toxins 1 and 1' (Inse...	65	1e-10
gi|3293263|gb|AAC25688.1| (AF039599) immunogenic venom prot...	43	0.001

(B)

Sequences producing significant alignments:	Score (bits)	E Value
gi|134340|sp|P19856|SIX1_LEIQU Insect toxins 1 and 1' (Inse...	111	2e-24
gi|58309|emb|CAA41265.1| (X58376) Hector insect toxin [Andr...	109	7e-24
gi|134375|sp|P15147|SIX2_ANDAU Insect toxin 2 precursor (In...	108	2e-23
gi|134372|sp|P01497|SIX1_ANDAU Insect toxin 1 precursor (In...	105	8e-23
gi|161147|gb|AAA29950.1| (M27705) neurotoxin AaH IT1 [Andro...	105	9e-23
gi|69545|pir||XISR1A insect toxin 1 - Sahara scorpion >gi|2...	99	8e-21
gi|3649606|gb|AAC61256.1| (AF055672) insect neurotoxin prec...	95	2e-19
gi|3914986|sp|O61668|SIX1_MESMA Excitatory insect selective...	90	6e-18
gi|3036821|emb|CAA76604.1| (Y17050) neurotoxin (KIT) [Buthu...	80	7e-15
gi|6094296|sp|P56637|SIXE_BUTJU Excitatory insect toxin BJX...	69	8e-12
gi|3858953|emb|CAA09988.1| (AJ012313) neurotoxin, variant 3...	68	2e-11
gi|4140001|pdb|1BCG| Scorpion Toxin Bjxtr-It	65	1e-10
gi|15825921|pdb|1I6F|A Chain A, Nmr Solution Structure Of T...	38	0.023

Fig. 2. BLAST search results used for classification of scorpion toxin sequences into groups based on high 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 Bjxtr-IT, a Na^+ toxin from *Hottentotta judaica*, against the nr database at NCBI. A marked increase in the *E* values from $1 \times e^{-10}$ to 0.001 would serve as a borderline where 0.001 was designated as the cut-off value. Sequences having *E* values <0.001 were clustered into a preliminary group. (B) Result of submitting Insect toxin 1 from *Leiurus quinquestriatus quinquestriatus*, which was the last sequence before the cut-off value, into nr database. The cut-off value for this second BLAST search was 0.023, where sequences having *E* values smaller than 0.023 were clustered into the group.

bootstrapped data sets using the SEQBOOT program [23] of the phylogeny inference package PHYLIP 3.6a2. For each group, the most parsimonious trees were calculated using the PROTPARS program [23] with the default parameter ordinary parsimony. The strict consensus trees were obtained using the CONSENSE program [23]. The unrooted tree diagrams were generated with the TREEVIEW program [24]. The group and subgroup classification of each scorpion toxin sequence can be found in the SCORPION database entries. The multiple alignments of the grouped toxins and the sequences of the test set of novel toxins from this study are accessible at http://research.i2r.a-star.edu.sg/scorpion/groups0.

2.3. 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 include toxin type, toxin action, ion channel specificity, and cellular target specificity. This module combines sequence comparison, nearest neighbour analysis and heuristics 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 220 scorpion toxin sequences in SCORPION database by BLAST program. A large change in score values separates dissimilar 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 algorithm uses ten 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 (Fig. 2) of the nearest neighbour is <20, than NOT SCORPION TOXIN.
3) If the query sequence is 100% identical to an existing sequence, than IDENTICAL.
4) If the query sequence is 100% identical to a portion of an existing sequence, then PARTIAL SEQUENCE.
5) If the identity to the nearest neighbour is less than 50%, then NEW GROUP.

6) If the difference of bit scores between two consecutive neighbours is $\geq$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.

3. Results

3.1. Classification of scorpion toxin sequences

We classified the 220 scorpion sequences into 34 groups (Table 1). The characteristic of each group was a high sequence similarity, from 34 to 100% pair-wise identity. The 100% identity was obtained for 16 precursor sequences that had identical mature region and variable signal peptide. Na^+ toxins were classified into 13 groups. Na^+ groups 2 and 12 were further classified into two subgroups each. Na^+ groups 1 and 9 were classified into three subgroups each. K^+ toxins were classified into 16 groups, while two groups were defined for Ca^{2+} toxins. Cl^- toxins were classified into two subgroups. Scorpine and an uncharacterized scorpion toxin sequence were assigned to the 'defensin' and 'short chain neurotoxin' groups, respectively.

Our classification of K^+ toxins was similar to the 12 subfamilies classified by Tytgat et al. [11] (Table 1). Our groups 2, 6, 7, and 10–12 were identical, while 3, 4, 5, 8, and 9 corresponded to Tytgat et al. subfamilies, except that our groups had larger number members. We reclassified the toxin PBTx3 of subfamily 1 into group 4 as it shared higher sequence identity with TyKα (55.3%) than BmTx1 or HgTx2 (52.6%). Finally, we introduced four additional groups 13–16 for seven novel scorpion toxins that could not be grouped with other toxins.

For Na^+ toxins, we compared our classification with subfamilies reported by Possani et al. [12] (Table 1). Our Na^+ groups 1, 2, and 7 corresponded to Possani subfamilies 1, 8, and 2. We reclassified sequences spread over Possani subfamilies 9, 10, 11, and 12 into Na^+ groups 3, 4, 5, 6, and 8. The subfamilies 5 and 6 were combined into group 9 and further classified into subgroups a, b, and c. Finally, we introduced groups 10 and 13 containing scorpion toxin sequences that were dissimilar to other groups.

To verify our classifications based on sequence similarity, we performed phylogenetic analysis. The results of phylogenetic analysis were in concordance with our groupings. A representative phylogenetic tree is shown in Fig. 3.

Table 1
Breakdown of 220 scorpion toxin sequences into groups and subgroups

K^+ group	Peptide no.	%Identity	Ref. [11]
1	10	37.8–78.4	[1,4]
2	9	48.7–89.7	Coincide
3	9	68.4–100.0	Similar [1]
4	8	34.2–56.8	Similar [3]
5	3	87.1–90.3	Similar [2]
6	5	42.1–68.4	Coincide
7	2	97.1	Coincide
8	4	79.3–96.6	Similar [1]
9	5	82.1–100.0	Similar [2]
10	2	84.4	Coincide
11	2	97.3	Coincide
12	1	NA	Coincide
13	1	NA	New
14	3	56.2–93.8	New
15	1	NA	New
16	2	92.9	New

Na^+ group	Peptide no.	%Identity	Ref. [12]
1 (a, b, c)	13 (4, 7, 2)	76.4–100.0	[1]
2 (a, b)	8 (3, 5)	57.1–98.4	[8]
3	31	52.9–100.0	[9,10]
4	4	76.6–98.4	[11,12]
5	4	70.1–79.1	[11]
6	12	59.7–100.0	[10]
7	6	50.7–100.0	[2]
8	2	71.2	[12]
9 (a, b, c)	28 (4, 17, 7)	73.5–97.0	[5,6]
10	4	71.6–83.3	–
11	19	57.4–100.0	[7]
12 (a, b)	4 (3, 1)	Partial sequences	[7]
13	1	NA	–

Ca^{2+} group	Peptide no.	%Identity
1	2	81.8
2	1	NA

Cl^- group	Peptide no.	%Identity
1 (a, b)	(7, 5)	55.3–88.6

Defensin	1	NA
Short-chain neurotoxin	1	NA

The first column displays the four ion channel families, defensin and short chain families. Second column lists the number of peptides in each group or subgroup. Third column gives the range of pair-wise identity for mature toxin sequences only. Fourth column represents our K^+ and Na^+ groupings as compared with the subfamilies of references 11 and 12, respectively. Multiple numbers in the same row correspond to scorpion toxin sequences classified into various subfamilies in each reference. Dashes represent no corresponding toxin sequence. For K^+ toxins, groups 2, 6, 7, 10–12 coincide with the respective subfamilies in Tytgat et al. [11]. The number within square brackets represents the number of new toxin sequence(s) not found in Tytgat's subfamily. New groups were introduced for novel scorpion toxin sequences. NA: not applicable.

3.2. Functional assignment of novel scorpion toxins

The functional properties predicted by the Annotate Scorpion module are the toxin subfamily, toxin potency, ion channel specificity and target cell type. The predicted functions of 52 toxin sequences from the test set are shown in (Table 2). Of these, only two could not be annotated. No

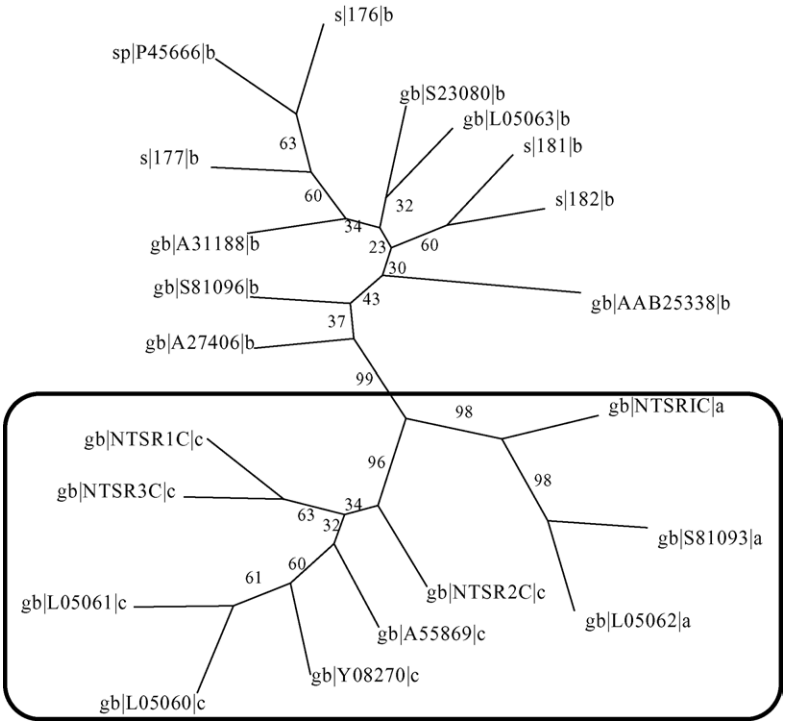

Fig. 3. An unrooted phylogenetic tree of Na^+ toxins from group 9. Abbreviations for data sources: sp: Swiss-Prot; gb: GenBank; s: SCORPION. The numbers correspond to bootstrap values of 100-bootstrapped data set. Sequences inside the box were earlier classified into subfamily 5 [12], whereas we have split them into subgroups 9a and 9c. Our classification of sequences into subgroup 9b corresponds to the subfamily 6 [12].

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 [25]. Twelve 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.

The accuracy of Annotate Scorpion module was assessed by comparing our predictions with the features of functionally characterized scorpion toxin sequences (Fig. 1). 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).

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 publications 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 characterized sequences, 8 (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 the Annotate Scorpion predicted that it belongs to a new group.

We tested Annotate Scorpion with sequences other than scorpion toxin and in all cases the results were 'no similar

Table 2
Putative functional annotation of 52 new toxin sequences

Toxin name	Group	Ion channel		Toxin type		Toxin action		Species specificity		Reference
		Pu	Ex/Ref.	Pu	Ex/Ref.	Pu	Ex	Pu	Ex	
Aah (putative toxin 1)	New	Na^+	–	α, α′	–	T	–	I, M	–	[29]
Aah (putative toxin 2)	New	Na^+	–	α, α′	–	T	–	I, M	–	[29]
Aah (putative toxin 3)	New	Na^+	–	α, α′	–	T	–	I, M	–	[29]
Aah (putative toxin 4)	3	Na^+	–	α, α′	–	T	–	I, M, C	–	[29]
Aah (putative toxin 5)	6	Na^+	–	α	–	T	–	I, M	–	[29]
*Be*Km-1	New	K^+	K^+	K	K	T	T	M	M	[3]
Bm (ANEPII)	11	Na^+	Na^+	β	β	T	–	I, M	–	Q9BKJ1
Bm (BKTx)	New	Na^+	Na^+	α	α	T	T	I, M	M	[30]
*Bm*32-VI	1a	Na^+	Na^+	β	β	T	T	I	I	[31]
*Bm*33-I	1a	Na^+	Na^+	β	β	T	T	I	I	[31]
*Bm*KdITAP3	11	Na^+	Na^+	β	β	T	T, N	I, M	I,M	[32]
*Bm*KK1	New	K^+	K^+	K	K	T	–	M	–	[33]
*Bm*P01	8	K^+	K^+	K	K	T, N	–	M	–	Q9U522
*Bm*SKTx2	New	K^+	K^+	K	K	T	–	M	–	Q9BJX2
*Bm*Tx3	New	K^+	K^+	K	K	T	T	M	M	[34]
*Bm*TXKS1	New	K^+	K^+	K	K	T, N	–	?	–	[35]
Bom α6a	3	Na^+	Na^+	α, α′	α	T	–	I, M, C	–	[36]
Bom α6b	3	Na^+	Na^+	α, α′	α	T	–	I, M, C	–	[36]
Bom α6c	3	Na^+	Na^+	α, α′	α	T	–	I, M, C	–	[36]
Bom α6d	3	Na^+	Na^+	α, α′	α	T	–	I, M, C	–	[36]
Bom α6e	3	Na^+	Na^+	α, α′	α	T	–	I, M, C	–	[36]
*Bs*IT1	11	Na^+	Na^+	β	β	T	T	I, M	I	[37]
*Bs*IT2	11	Na^+	Na^+	β	β	T	T	I, M	I	[37]
*Bs*IT3	11	Na^+	Na^+	β	β	T	T	I, M	I	[37]
*Bs*IT4	11	Na^+	Na^+	β	β	T	T	I, M	I	[37]
Bt (tamapin)	5	K^+	K^+	K	K	T	T	M	M	[38]
Bt (tamapin 2)	5	K^+	K^+	K	K	T	T	M	M	[38]
Bt (tamulustoxin 1)	New	Na^+	K^+	β	K	T	T	I, M	M	[39]
Bt (tamulustoxin 2)	New	Na^+	K^+	β	K	T	T	I, M	M	[39]
*Cg*2	9c	Na^+	Na^+	β	–	T	–	I, C	–	[12]
*Cll*9	9*	Na^+	Na^+	β	–	T	–	I, C	–	[12]
*Hf*Tx 1	?	?	K^+	?	K	T	–	?	M	[25]
*Hf*Tx 2	?	?	K^+	?	K	T	–	?	M	[25]
Lqh α6a	3	Na^+	Na^+	α, α′	α′	T	–	I, M, C	–	[36]
Lqh α6b	3	Na^+	Na^+	α, α′	α′	T	–	I, M, C	–	[36]
Lqh α6c	3	Na^+	Na^+	α, α′	α′	T	–	I, M, C	–	[36]
*Lqh*ChTx-b	1	K^+	K^+	K	K	T	–	M	–	[36]
*Lqh*ChTx-c	1	K^+	K^+	K	K	T	–	M	–	[36]
*Lqh*ClTx-a	1*	Cl^-	Cl^-	Cl	Cl	T	–	I	–	[36]
*Lqh*ClTx-b	1a	Cl^-	Cl^-	Cl	Cl	T	–	I, M	–	[36]
*Lqh*ClTx-c	1a	Cl^-	Cl^-	Cl	Cl	T	–	I, M	–	[36]
*Lqh*ClTx-d	1a	Cl^-	Cl^-	Cl	Cl	T	–	I, M	–	[36]
*Lqh*IT1-a	1*	Na^+	Na^+	β	β	T	–	I	–	[36]
*Lqh*IT1-b	1*	Na^+	Na^+	β	β	T	–	I	–	[36]
*Lqh*IT1-c	1*	Na^+	Na^+	β	β	T	–	I	–	[36]
*Lqh*IT1-d	1*	Na^+	Na^+	β	β	T	–	I	–	[36]
*Lqh*IT2-13	11	Na^+	Na^+	β	β	T	–	I, M	–	[36]
*Lqh*IT2-53	11	Na^+	Na^+	β	β	T	–	I, M	–	[36]
*Os*K2	New	K^+	K^+	K	K	T	T	M	I	[40]
*Tb*2 II	2b	Na^+	Na^+	α, β	β	T, N	T	M	I, M	[41]
*Tb*IT-I	2*	Na^+	Na^+	α, β	β	T, N	T	I, M	I	[41]
Tst 2	2b	Na^+	Na^+	β	β	T	T	M	M	[42]

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 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 neighbor found, the query is assigned 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 references. Dashes (–) represent no experimental characterization or no functional annotation in references. Swiss-Prot accession numbers refer to direct submission of toxin sequences. Question marks (?) represent sequences not functionally annotated.

record found'. We have shown that Annotate Scorpion is robust and highly accurate in assigning putative annotation to previously unseen scorpion toxins. The toxin sequences from the test set have been incorporated into the prediction module as templates for functional annotation of new toxin sequences.

4. Discussion and conclusions

We have developed a generic bioinformatic functional prediction tool for functional annotation of scorpion toxins that can help reduce the number of experiments conducted for characterizing novel scorpion toxin sequences. The bioinformatics-based approach of collecting, cleaning, annotating and classifying scorpion sequences into 34 groups allowed us to predict 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 [26,27]. If these annotations were not corrected, our predictions would be less accurate.

We have classified scorpion toxins using BLAST and CLUSTAL W results, which are in agreement with the phylogenetic analysis. Our grouping of K^+ scorpion toxins was in good agreement with Tytgat et al. [11]. However, our grouping of Na^+ toxins differed from Possani et al. [12]. The phylogenetic analysis and highly accurate predictions of scorpion toxin groups and functional properties indicate that current groupings are suitable for scorpion toxin characterization.

Detailed classification of scorpion toxin sequences into well-organized groups allows better correlation of structure–function relationship. Sequences that are similar (in primary, secondary and tertiary structures) often perform similar function. Most scorpion toxins share the cysteine-stabilized α-helix fold [28]. 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. Our 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 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.

Improvements can still be made in the annotation of the subfamily for Na^+ toxins and the cellular target type. We are currently working on fine classification of scorpion toxins into basic structure–function units where a unit contains scorpion toxin sequences with high sequence identity and share similar functional properties such as high binding affinity towards the same ion channel.

With a rapidly increasing number of scorpion toxin sequences in SCORPION database, general patterns in structure and function can be inferred. Similar approaches can be applied to many other fields of research such as the functional annotation of protein families. We foresee that the detailed structural and functional grouping of protein sequences can be used for accurate prediction of functional properties of other toxins and families of bioactive peptides.

References

[1] S. Cestele, M. Stankiewicz, P. Mansuelle, M. De Waard, B. Dargent, N. Gilles, M. Pelhate, H. Rochat, M.F. Martin-Eauclaire, D. Gordon, Scorpion alpha-like toxins, toxic to both mammals and insects, differentially interact with receptor site 3 on voltage-gated sodium channels in mammals and insects, Eur. J. Neurosci. 11 (1999) 975–985.

[2] G.B. Gurrola, B. Rosati, M. Rocchetti, G. Pimienta, A. Zaza, A. Arcangeli, M. Olivotto, L.D. Possani, E. Wanke, A toxin to nervous, cardiac, and endocrine ERG K^+ channels isolated from *Centruroides noxius* scorpion venom, FASEB J. 13 (1999) 953–962.

[3] Y.V. Korolkova, S.A. Kozlov, A.V. Lipkin, K.A. Pluzhnikov, J.K. Hadley, A.K. Filippov, D.A. Brown, K. Angelo, et al. An ERG channel inhibitor from the scorpion *Buthus eupeus*, J. Biol. Chem. 276 (2001) 9868–9876.

[4] B. Lebrun, R. Romi-Lebrun, M.F. Martin-Eauclaire, A. Yasuda, M. Ishiguro, Y. Oyama, O. Pongs, T. Nakajima, A four-disulphide-bridged toxin, with high affinity towards voltage-gated K^+ channels, isolated from *Heterometrus spinnifer* (Scorpionidae) venom, Biochem. J. 328 (1997) 321–327.

[5] Y.J. Li, Y. Liu, Y.H. Ji, BmK AS: new scorpion neurotoxin binds to distinct receptor sites of mammal and insect voltage-gated sodium channels, J. Neurosci. Res. 61 (2000) 541–548.

[6] I. Zenouaki, R. Kharrat, J.M. Sabatier, C. Devaux, H. Karoui, J. Van Rietschoten, M. el Ayeb, H. Rochat, In vivo protection against *Androctonus australis hector* scorpion toxin and venom by immunization with a synthetic analog of toxin II, Vaccine 15 (1997) 187–194.

[7] C.Y. Wang, Z.Y. Tan, B. Chen, Z.Q. Zhao, Y.H. Ji, Antihyperalgesia effect of BmK IT2, a depressant insect-selective scorpion toxin in rat by peripheral administration., Brain Res. Bull. 53 (2000) 335–338.

[8] D.F. Rogers, Scorpion venoms: taking the sting out of lung disease, Thorax 51 (1996) 546–548.

[9] O. Froy, N. Zilberberg, N. Chejanovsky, J. Anglister, E. Loret, B. Shaanan, D. Gordon, M. Gurevitz, Scorpion neurotoxins: structure/function relationships and application in agriculture, Pest Manag. Sci. 56 (2000) 472–474.

[10] W.R. Lourenco, Diversity and endemism in tropical versus temperate scorpion, Biogiographica 70 (1994) 155–160.

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

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

[13] L.D. Possani, B. Becerril, M. Delepierre, J. Tytgat, Scorpion toxins specific for Na^+-channels, Eur. J. Biochem. 264 (1999) 287–300.

[14] C. Maertens, L. Wei, J. Tytgat, G. Droogmans, B. Nilius, Chlorotoxin does not inhibit volume-regulated, calcium-activated and cyclic AMP-activated chloride channels, Br. J. Pharmacol. 129 (2000) 791–801.
[15] S. Dalton, V. Gerzanich, M. Chen, Y. Dong, Y. Shuba, J.M. Simard, Chlorotoxin-sensitive Ca^{2+}-activated Cl^- channel in type R2 reactive astrocytes from adult rat brain, Glia 42 (2003) 325–339.
[16] N. Gilles, C. Blanchet, I. Shichor, M. Zaninetti, I. Lotan, D. Bertrand, D. Gordon, 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 (1999) 8730–8739.
[17] G. Ferrat, C. Bernard, V. Fremont, T.J. Mullmann, K.M. Giangiacomo, H. Darbon, Structural basis for alpha-K toxin specificity for K^+ channels revealed through the solution 1H NMR structures of two noxiustoxin-iberiotoxin chimeras, Biochemistry 40 (2001) 10998–11006.
[18] P.T. Tan, A.M. Khan, V. Brusic, Bioinformatics for venom and toxin sciences, Brief. Bioinform. 4 (2003) 53–62.
[19] R.C. Rodriguez de la Vega, E. Merino, B. Becerril, L.D. Possani, Novel interactions between K^+ channels and scorpion toxins, Trends Pharmacol. Sci. 24 (2003) 222–227.
[20] K.N. Srinivasan, P. Gopalakrishnakone, P.T. Tan, K.C. Chew, B. Cheng, R.M. Kini, J.L. Koh, S.H. Seah, V. Brusic, A molecular database of scorpion toxins, Toxicon 40 (2002) 23–32.
[21] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic Acids Res. 25 (1997) 3389–3402.
[22] J.D. Thompson, D.G. Higgins, T.J. Gibson, 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 (1994) 4673–4680.
[23] PHYLIP (Phylogeny Inference Package) version 3.6a2, Department of Genetics, University of Washington, Seattle, 2001.
[24] R.D.M. Page, TreeView: an application to display phylogenetic trees on personal computers, Comput. Appl. Biosci. 12 (1996) 357–358.
[25] K.N. Srinivasan, V. Sivaraja, I. Huys, T. Sasaki, B. Cheng, T.K. Kumar, K. Sato, J. Tytgat, C. Yu, B.C. San, et al. 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 (2002) 30040–30047.
[26] D.A. Oren, O. Froy, E. Amit, N. Kleinberger-Doron, M. Gurevitz, B. Shaanan, 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 (1998) 1095–1103.
[27] D. Gordon, P. Savarin, M. Gurevitz, S. Zinn-Justin, Functional anatomy of scorpion toxins affecting sodium channels, J. Toxicol. Toxin Rev. 17 (1998) 131–159.
[28] F. Bontems, C. Roumestand, P. Boyot, B. Gilquin, Y. Doljansky, A. Menez, F. Toma, Three-dimensional structure of natural charybdotoxin in aqueous solution by 1H-NMR. Charybdotoxin possesses a structural motif found in other scorpion toxins, Eur. J. Biochem. 196 (1991) 19–28.
[29] B. Ceard, M. Martin-Eauclaire, P.E. Bougis, Evidence for a position-specific deletion as an evolutionary link between long- and short-chain scorpion toxins, FEBS Lett. 494 (2001) 246–248.
[30] K.N. Srinivasan, S. Nirthanan, T. Sasaki, K. Sato, B. Cheng, M.C. Gwee, R.M. Kini, P. Gopalakrishnakone, Functional site of bukatoxin, an alpha-type sodium channel neurotoxin from the Chinese scorpion (*Buthus martensi* Karsch) venom: probable role of the 52PDKVP(56) loop, FEBS Lett. 494 (2001) 145–149.
[31] P. Escoubas, M. Stankiewicz, T. Takaoka, M. Pelhate, R. Romi-Lebrun, F.Q. Wu, T. Nakajima, Sequence and electrophysiological characterization of two insect-selective excitatory toxins from the venom of the Chinese scorpion *Buthus martensi*, FEBS Lett. 483 (2000) 175–180.
[32] R. Guan, C. Wang, M. Wang, D. Wang, A depressant insect toxin with a novel analgesic effect from scorpion *Buthus martensii* Karsch, Biochim. Biophys. Acta 1549 (2001) 9–18.
[33] X.C. Zeng, F. Peng, F. Luo, S.Y. Zhu, H. Liu, W.X. Li, Molecular cloning and characterization of four scorpion K(+)-toxin-like peptides: a new subfamily of venom peptides (alpha-KTx14) and genomic analysis of a member, Biochimie 83 (2001) 883–889.
[34] H. Vacher, R. Romi-Lebrun, C. Mourre, B. Lebrun, S. Kourrich, F. Masmejean, T. Nakajima, C. Legros, M. Crest, P.E. Bougis, M.F. Martin-Eauclaire, A new class of scorpion toxin binding sites related to an A-type K^+ channel: pharmacological characterization and localization in rat brain, FEBS Lett. 501 (2001) 31–36.
[35] S.Y. Zhu, W.X. Li, X.C. Zeng, Precursor nucleotide sequence and genomic organization of BmTXKS1, a new scorpion toxin-like peptide from *Buthus martensii Karsch*, Toxicon 39 (2001) 1291–1296.
[36] O. Froy, T. Sagiv, M. Poreh, D. Urbach, N. Zilberberg, M. Gurevitz, Dynamic diversification from a putative common ancestor of scorpion toxins affecting sodium, potassium, and chloride channels, J. Mol. Evol. 48 (1999) 187–196.
[37] S.A. Ali, S. Stoeva, J.G. Grossmann, A. Abbasi, W. Voelter, Purification, characterization, and primary structure of four depressant insect-selective neurotoxin analogs from scorpion (*Buthus sindicus*) venom, Arch. Biochem. Biophys. 391 (2001) 197–206.
[38] P. Pedarzani, D. D'hoedt, K.B. Doorty, J.D. Wadsworth, J.S. Joseph, K. Jeyaseelan, R.M. Kini, S.V. Gadre, S.M. Sapatnekar, et al. 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 (2002) 46101–46109.
[39] P.N. Strong, G.S. Clark, A. Armugam, F.A. De-Allie, J.S. Joseph, V. Yemul, J.M. Deshpande, R. Kamat, S.V. Gadre, P. Gopalakrishnakone, et al. Tamulustoxin: a novel potassium channel blocker from the venom of the Indian red scorpion *Mesobuthus tamulus*, Arch. Biochem. Biophys. 385 (2001) 138–144.
[40] E.E. Dudina, Y.V. Korolkova, N.E. Bocharova, S.G. Koshelev, T.A. Egorov, I. Huys, J. Tytgat, E.V. Grishin, *Os*K2, a new selective inhibitor of Kv1.2 potassium channels purified from the venom of the scorpion *Orthochirus scrobiculosus*, Biochem. Biophys. Res. Commun. 286 (2001) 841–847.
[41] A.M. Pimenta, M. Martin-Eauclaire, H. Rochat, S.G. Figueiredo, E. Kalapothakis, L.C. Afonso, M.E. De Lima, Purification, amino-acid sequence and partial characterization of two toxins with anti-insect activity from the venom of the South American scorpion *Tityus bahiensis* (Buthidae), Toxicon 39 (2001) 1009–1019.
[42] B. Becerril, M. Corona, F.I. Coronas, F. Zamudio, E.S. Calderon-Aranda, P.L. Fletcher Jr., B.M. Martin, L.D. Possani, Toxic peptides and genes encoding toxin gamma of the Brazilian scorpions *Tityus bahiensis* and *Tityus stigmurus*, Biochem. J. 313 (1996) 753–760.