PROTEINS: Structure, Function, and Bioinformatics 59:131–140 (2005)

A Novel Strategy for the Identification of Toxinlike Structures in Spider Venom

Sergey Kozlov,[1]* Anton Malyavka,[1] Bill McCutchen,[2] Albert Lu,[3] Eric Schepers,[3] Rafi Herrmann,[3] and Eugene Grishin[1]

[1]*Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow, Russia*
[2]*Pioneer Hi-Bred International, Inc., a DuPont Company, Johnston, Iowa*
[3]*DuPont Agriculture and Nutrition, Newark, Delaware*

ABSTRACT **We compared two different approaches to sequence information analysis from the expressed sequence tag (EST) library constructed for the venom glands of the spider *Agelena orientalis*. Some results were more illustrative and reliable by the contig analysis technique, whereas our novel method, with specific structural markers introduced for protein structure detection, allowed us to overcome some limitations of the contig analysis. A novel technique was suggested for the identification in data banks of the spider's ion channel inhibitor toxins using primary structure features common to all spiders. Analysis of about 150 polypeptides made it possible to introduce 3 primary structure motifs for spider toxins: the Principal Structural Motif (PSM), which postulates the existence of 6 amino acid residues between the first and second cysteine residue and the Cys-Cys sequence at a distance of 5–10 amino acid residues from the second cysteine; the Extra Structural Motif (ESM), which postulates the existence of a pair of CXC fragments in the C-region; and the Processing Quadruplet Motif (PQM), which specifies the Arg residue at position −1 and Glu residues at positions −2, −3, and/or −4 in the precursor sequences just before the postprocessing site. In the processed data bank we found 48 toxinlike structures with ion channel inhibitor motifs. These include agelenin earlier isolated from *Agelena opulenta* and 25 more homologous sequences, 15 homologs of μ-agatoxin 2 from the spider *Agelenopsis aperta*, 3 structures with low homology to ω-agatoxin-IIIA, and 4 new structures. Also we showed that toxinlike structures exceed two thirds of the overall database sequences. Proteins 2005;59:131–140.** © 2005 Wiley-Liss, Inc.

Key words: primary structure motif; database processing; proteomic; ion channel inhibitor; structural markers

INTRODUCTION

Spiders produce poisonous secretions that contain biologically active compounds, most of which are toxic for animals. Polypeptide toxins belong to the most extensively studied constituents of spider venoms. The primary structure of many of them is now available. The majority of spider toxins are small, compact molecules with molecular masses ranging from 4 to 7 kDa cross-linked by 3–5 disulfide bonds. They share a common structural motif known as a cystine knot.[1,2] Spider toxins selectively interact with different types of ion channels and membrane receptors. The properties and structure of toxins isolated from spider venoms have been analyzed in detail in several reviews.[3,4] The total number of spider species exceeds 30,000, and most of them have venomous glands. The venoms of some species have been investigated to date, and, usually, only the most active and abundant components have been analyzed. In rare cases, the number of components identified in one species has been as many as 10. Isolation of venom components and cloning experiments have shown that the venom of a single species can contain homologous polypeptides that differ in key point substitutions.[5–9] Characterization of all of the venom polypeptide toxins of a single species presents a challenge to biochemists, as spider venoms can be regarded as a complex natural library of polypeptide components derived from several basic structures with different biological activities and specificities. In the evolution of arachnoids, venom composition has changed for adaptation to varying environments. On the basis of several structural folds, spiders have created a sophisticated toxin array, resulting in venoms of high efficiency for the capture of diverse prey. In this article, we describe a new strategy for analyzing all polypeptides secreted by the spider *Agelena orientalis* based on cDNA library construction and expressed sequence tag (EST) sequencing. To identify the structures of putative toxins involved in interactions with membrane receptors and ion channels, new algorithms for the search of encoded sequences based on primary structure features common to all spider species were formulated. Using these algorithms, 48 toxinlike structures were identified in the database of cDNA from venom glands of the central Asian spider *A. orientalis*.

Grant sponsor: Russian Foundation of Basic Research. Grant sponsor: Ministry of Science and Technology of Russian Federation. Grant sponsor: CRDF. Grant sponsor: PCB RAS Program.

*Correspondence to: Sergey Kozlov, Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, ul.Miklukho-Maklaya, 16/10, 117997 Moscow, Russia. E-mail: serg@ibch.ru

Received 27 July 2004; Accepted 19 October 2004

Published online 1 February 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20390

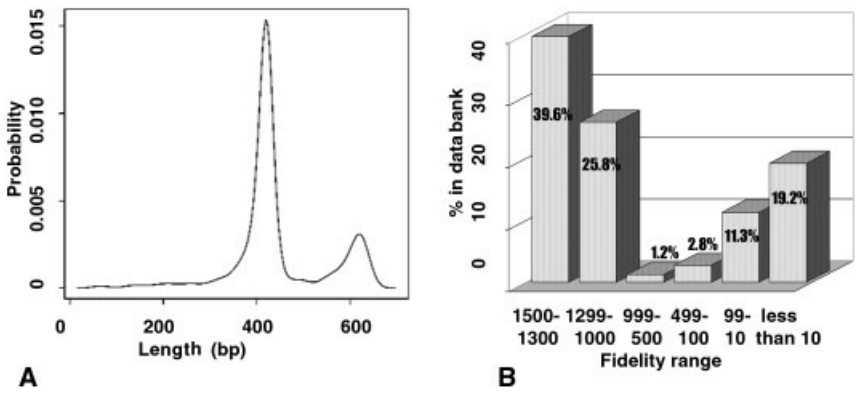

Fig. 1. Properties of the EST database for the venom glands of the central Asian spider *A. orientalis*: (**A**) EST length distribution. (**B**) Distribution of database entries according to fidelity range based on cluster analysis.

MATERIALS AND METHODS

Venom glands of the central Asian spider *A. orientalis* were harvested directly into liquid nitrogen. Total RNA was prepared by tissue homogenization in liquid nitrogen using a mortar and pestle, followed by cell lysis in the presence of TRIzol (Life Technologies).

For cDNA library construction, polyA(+) RNA was purified from the total RNA on an oligo(dT)-cellulose affinity column using the mRNA Purification Kit (Amersham Pharmacia Biotech) according to the manufacturer's protocol. The first-strand cDNA synthesis using Superscript II (Life Technologies) and subsequent second-strand synthesis, linker addition, and directional cloning into the EcoRI and Xhol sites of pBlueScript SK+ (Strategene) were performed in accordance with the instructions provided with the Stratagene cDNA kit (Stratagene). cDNA was purified using a cDNA column (Life Technologies) immediately prior to ligation into the vector.

Sequencing of cDNA library clones was accomplished using the ABI PRISM Big Dye Terminator Cycle Sequencing Ready reaction kit with FS AmpliTaq DNA polymerase (Perkin Elmer) and analyzed on an ABI Model 373 Automated DNA Sequencer.

Sequences from 2166 clones were compared to nucleotide or peptide sequences in GenBank and Peptide Sequence databases using BLASTN or BLASTP programs. The Winstar (DNAstar, Inc.) program modules were used for advanced sequence comparison and alignment.

RESULTS

Construction and Analysis of the cDNA Library

Purification of mRNA from *A. orientalis* (Invitrogen) venom glands and subsequent cDNA synthesis (Stratagene) were conducted according to the manufacturer's protocols. The cDNA was directionally cloned into pBlueScript SK+, and 2166 recombinant clones from the resulting cDNA library were sequenced using the M13 Fwd primer. The cDNA length distribution in the spider *A. orientalis* has a pronounced maximum at about 400 nucleotides [Fig. 1(A)].

cDNA fragments less than 800 nucleotides in length were sequenced to generate ESTs. Thus, the maximum derived protein sizes encoded by the cDNA sequences taking into account the 5′-nontranslated regions exceed 170 amino acid residues. This range overlaps the size of virtually all known spider toxins interacting with ion channels.

The initial ESTs were clustered into 37 sequence contigs and 332 singletons. Contigs varied in size and contained from 2 to 1500 homologous sequences. The consensus sequence analysis for some of the contigs revealed homology to known ion channel blocking toxins of spiders (see Table I). As seen in Table I, most frequently sequenced ESTs showed homology to the toxins of spiders belonging to the family Agelenidae including the *Agelena*, *Agelenopsis*, and *Hololena* genera. These toxins act on different types of ion channels. Agelenin, ω-Aga-IIIA, ω-Aga-IVA, ω-Aga-IA and Tx3-6 affect the function of Ca^{2+} channels.[10–15] All curtatoxins, μ-agatoxins, and palutoxin IT1 act on sodium channels,[16–18] and the Tx3-1 toxin selectively inhibits the A-type K^+ current [IA].[19] The specificity of Tx2-1, which is highly neurotoxic for mammals and insects, for a particular type of ion channels is still unknown.[20]

Further database analysis was performed by comparing separate translated polypeptide sequences by FASTA allowed us to avoid assembly errors[21] common during the formation of contigs and to identify all putative sites of point mutations. For convenience, we introduced the term *entry* for one sequenced EST included in the bank and the term *fidelity of entry* for the number of homologous ESTs. If, during grouping of sequences in contigs, the entry of interest falls into the group of 30 homologous nucleotide sequences, its fidelity is 30; if the group consists of 115 clones, the fidelity of all entries will be 115; for all singletons, the fidelity is always 1.

The fidelity value determined by cluster analysis shows if an entry is faithful. If this parameter is high, we can state with assurance that the encoded protein has a conserved structure and/or is well represented in the natural venom. Low fidelity values indicate that either the encoded protein is a rare venom component or that the sequence is wrong. The distribution of entries according to

TABLE I. Homology of Consensus Sequences in Assembled Contigs to Known Polypeptide Toxins of Spider Venoms

Contig	Size	Toxins	Organism	Accession No.
1	1497	Agelenin	Agelena opulenta	P31328
2	30	μ-agatoxins 1, 2, 3, 4, 5, 6	Agelenopsis aperta	P11057, P11058, P60178, P11060, P11061, P11062
		Curtatoxins CT I, CT II, CT III	Hololena curta	P15967, P60177, P15968
		Δ-palulT1	Paracoelotes luctuosus	A59401
3	22	ω-agatoxin IA, IIIA, IIIB	Agelenopsis aperta	A45069, A42335, C54252
		Neurotoxin Tx3-6, Tx2-1	Phoneutria nigriventer	F44336, AAC26166.1
4	12	μ-agatoxins 1, 2, 3, 4, 5, 6	Agelenopsis aperta	P11057, P11058, P60178, P11060, P11061, P11062
		Curtatoxins CT II, CT III	Hololena curta	P60177, P15968
		Δ-palulT1	Paracoelotes luctuosus	A59401
5	4	μ-agatoxins 4	Agelenopsis aperta	P11060
		Curtatoxins CT II, CT III	Hololena curta	P60177, P15968
		Δ-palulT1	Paracoelotes luctuosus	A59401
6	2	ω-agatoxin IIA, IIIA, IIIB	Agelenopsis aperta	C34923, A42335, C54252
		Neurotoxin Tx3-4	Phoneutria nigriventer	D44336

See accession number for reference in protein data banks. Size of contigs measured in a numbers of sequences.

fidelity in the EST bank from the venom glands of *A. orientalis* is shown in Figure 1(B).

About two-thirds of entries belong to structural groups 1 and 2 with very high homology (fidelity above 1000). This means that some structures are true, and their content in the *A. orientalis* venom is rather high. Other putative structures are less represented in the bank and consequently, in the natural venom, since the fidelity of the remaining entries is mostly below 10.

After general analysis of the *A. orientalis* nucleotide sequence bank, we turned to the methods for the identification and classification of genome-encoded proteins. For this purpose, we analyzed all known polypeptide toxins of spiders.

Establishment of Motifs for Structural Characterization

Multiple analysis of the primary structure of any spider toxin using the BLAST algorithm failed to reveal apparent homology with other structural groups of spiders toxins (usually about 15–50%), except with those from closely related species. To date, the structures of about 150 polypeptide toxins acting as ion channel blockers from venoms of different spider species have been published. Attempts to reveal primary structure homology among all known sequences were unsuccessful. However, the majority of toxins share similarity in Cys arrangement and disulfide bridge pattern. They have a common structural motif composed of a triple-stranded antiparallel β-sheet, stabilized by internal disulfide bonds termed cystine knots. Cystine knot are found in growth factors, cyclic plant polypeptides, defensins, protease inhibitors, and polypeptide toxins from snake venoms, arthropods, and mollusks.[22,23]

In other cystine knot polypeptides, the gap between the first and second cysteine residues is variable; however, in spider toxins, which are ion channel blockers, the first 2 Cys residues are always invariable separated by 6 amino acids. In addition, there are no amino acids between the third and fourth cysteine residues. These observations allowed us to introduce the Principal Structural Motif for spider toxins (PSM) that stipulates the existence of 6 amino acid residues between the first and second cysteine residues and the occurrence of a Cys-Cys sequence at a distance of 5–10 amino acid residues from the second cysteine residue. About two-thirds of all structures conform to these criteria; some of them are shown in Fig. 2. There are only 6 exceptions: the TX2-9 toxin (5 amino acids gap between Cys 1 and 2), ω-agatoxin IVA, ω-agatoxin IVB, DTX11 (7 amino acids gap between Cys 1 and 2), and atracotoxin-HVF17 and Magi 6 (4 amino-acid gap between Cys 1 and 2).

PSM CXXXXXXC-----CC or C6CUCC,

where U is a variable number of amino acid residues

We found that among the PSM-containing toxins with more than 6 cysteine residues, Cys 5 and 6, as well as Cys 7 and 8, are separated by a single amino acid residue. This holds true for more than 70% of structures. Therefore, we also described an Extra Structural Motif for spider toxins (ESM), which stipulates the existence of a pair of CXC fragments in the C-region of the polypeptide chain. The PSM and ESM sequences are separated by 2–13 amino acid residues (on average, by 4), while CXC fragments are separated by 3–24 residues.

ESM CXC----CXC or C1CZC1C,

where Z is a variable number of amino acid residues

Analysis of toxin precursors (Fig. 3) allowed us to introduce a third structural motif.

In translated toxin precursors, the initiation Met codon can be easily found, and the position of signal peptides can be determined on the basis of general rules.[56,57] After the signal peptide, there is commonly a propeptide region, with an average size that varies from 10 to 20 amino acids (maximum 44 residues). At position −1, there is an Arg residue in all known toxin precursors, and at positions −2 and/or −4, there is a Glu. The Glu residue may also occur

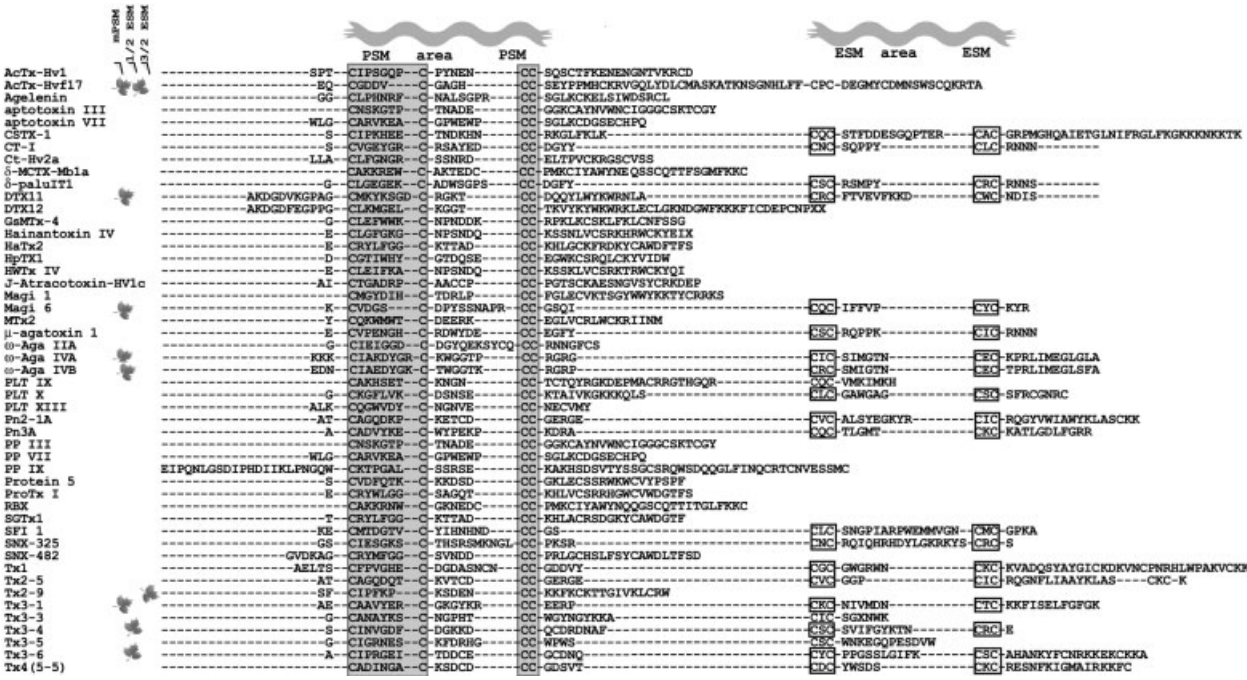

Fig. 2. Sequence alignment in alphabetic order of mature toxins from spiders. Shading shows PSM regions; ESM region sequences are boxed. Modified structural motifs are highlighted by maple leaves. Sequences correspond to the inhibitors of ions channels isolated from different spider species: *Hadronyche versuta*, AcTx-Hv1, AcTx-Hvf17, Ct-Hv2a, J-atracotoxin-Hv1c[24,25]; *Agelena opulenta*, agelenin[26]; *Apomastus schlingeri*, aptotoxins, PP III, PPVII, PP IX[27]; *Cupiennius salei*, CSTX-1[28]; *Hololena curta*, curtatoxins[16]; *Missulena bradleyi*, δ-MSTX-Mb1a[29]; *Paracoelotes luctuosus*, δ-paluIT2[18]; *Diguetia canities*, DTX11, DTX12[30]; *Grammostola spatulata*, GsMTx-4, HaTx2, MTx2[31–33]; *Ornithoctonus hainana*, hainantoxin-IV[34]; *Heteropoda venatoria*, HpTX1[35]; *Ornithoctonus huwena*, HwTx-IV[36]; *Macrothele gigas*, Magi 1 and Magi 6[37]; *Agelenopsis aperta*, μ-agatoxin 1, ω-Aga-IIA, ω-Aga-IVA, ω-Aga-IVB[13,17,38,39]; *Plectreurys tristis*, PLT toxins[5]; *Phoneutria nigriventer*, Tx1, Tx2, Tx3, Tx4 families and Pn2-1A, Pn3A[20,40–44]; *Brachypelma smithii*, Protein 5[45]; *Thrixopelma pruriens*, ProTx-I[46]; *Atrax robustus*, RBX[47]; *Scodra griseipes*, SGTx1[48]; *Segestria florentina*, SFI 1, SNX-325[9,49]; *Hysterocrates gigas*, SNX-482.[50]

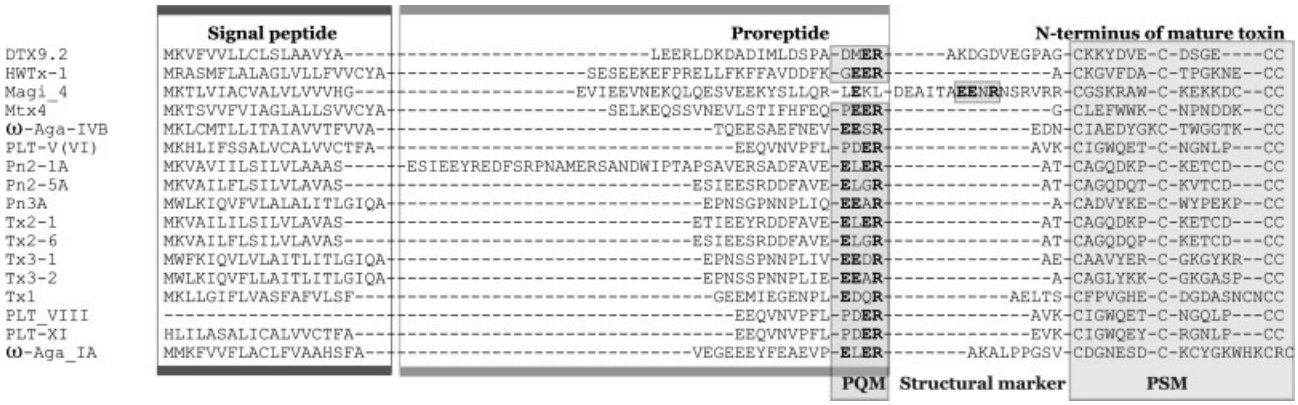

Fig. 3. N-terminal partial sequence of spider toxin precursors. Signal peptides and propeptides are indicated in accordance with SWISS-PROT structural data referring to toxin DTX 9.2,[30] huwentoxin-I,[51] Magi 4,[37] GsMTx4 (Mtx4),[52] ω-Aga toxins IVB and IA,[39,53] PLT toxins,[5] and toxins from *Phoneutria nigriventer*.[40,41,43,54,55] Structural markers: PQM and PSM are shown by shadowing. Amino acid residues in PQM important for motif detection are given in bold.

at position −3 or several times between the −1 and −4 residues in the mature polypeptide chain numbering. Highly conserved positions of Arg and Glu residues, just before the postprocessing site, therefore serve as the third structural motif of the ion channel blocking toxins, named the Processing Quadruplet Motif (PQM).

$$\mathbf{PQM}\ A_{-4}\ A_{-3}\ A_{-2}\ R_{-1}\ X,$$

where at least one amino acid A is Glu and X is the first amino acid residue in the mature polypeptide chain

In the Magi 4 precursor, where propeptide proteolysis involves another site, the primary structure analysis nevertheless shows the presence of the PQM motif in the structure.

All the above-mentioned characteristics are typical for ion channel blocking toxins. Whether the same holds are true for other toxin groups is still to be elucidated.

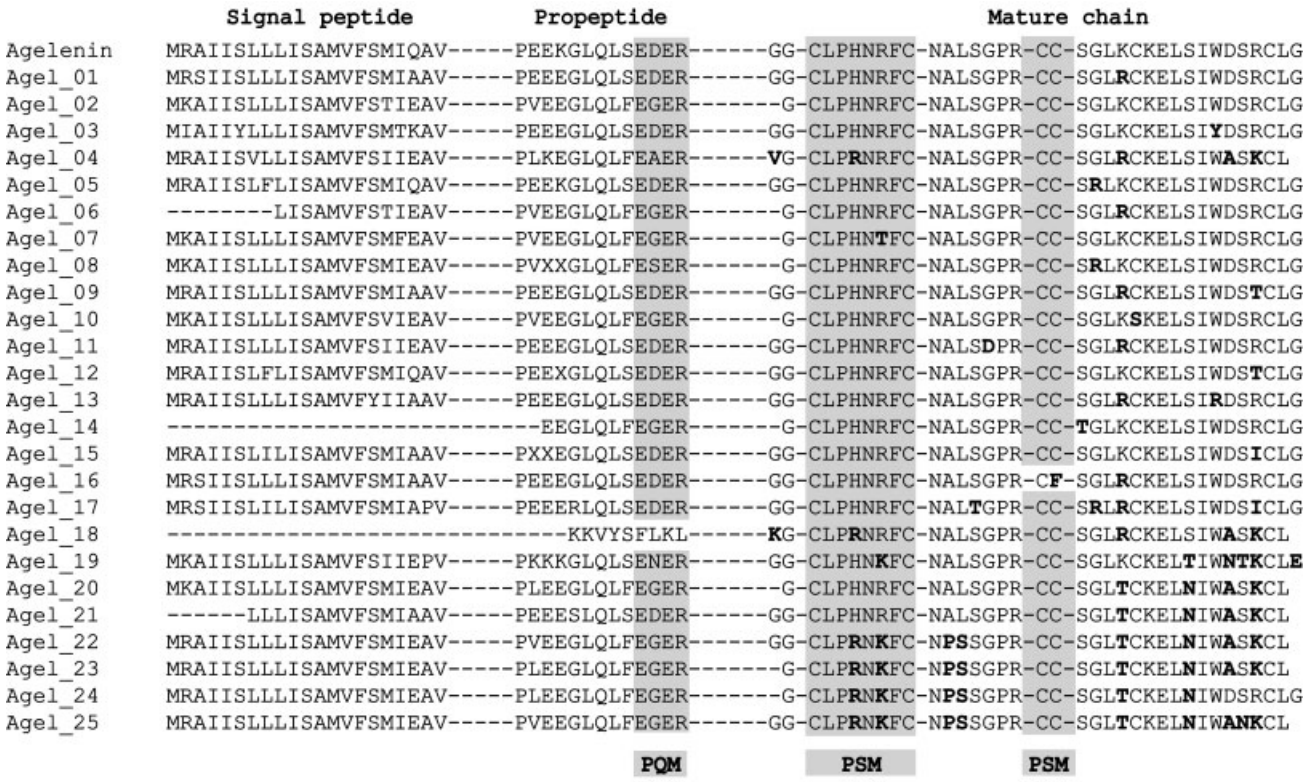

Fig. 4. Sequences of agelenin and ageleninlike peptides deduced from the ESTs data bank by PSM and PQM marker search. Amino acid residues differing from the agelenin sequence are shown in bold. Sequences corresponding to structural markers are indicated by shadowing.

Development of the *A. orientalis* cDNA Database

The *A. orientalis* cDNA database was analyzed using the 3 structural motifs of spider toxins. At first, we identified sequences that contained PSM (structural group A) and PSM+ESM (structural group B) markers. We compared amino acid sequences of mature proteins without taking into account signal peptides and propeptides. In deduced sequences, the mature chains began immediately after the Arg residue in PQM. In rare cases, when the PQM was missing, the amino acid after the signal peptide was taken for the beginning of the mature chain.

The first protein found in the database was agelenin from a closely related species *Agelena opulenta*.[10] In addition, 25 more sequences homologous to agelenin (80–98% similarity) named Agel_01–Agel_25, were identified; their precursors are shown in Figure 4. Agelenin and all ageleninlike structures belong to structural group A and possess only PSM, while in 2 structures, Agel_10 and Agel_16, the distribution of Cys residues is disturbed. The PQM is present in all precursors, except for Agel_18 with an incomplete N-terminal region.

Based on the number of entries encoding mature ageleninlike polypeptides, the proportion of each deduced sequence in the total proteins in this group was calculated. It is of particular interest that Agel_01 is the most abundant component (more than 40.5%), not agelenin (22.5%). The content of 2 other polypeptides, Agel_02 and Agel_22, is also rather high (15% and 5%, respectively).

In addition to 25 agelenin homologs, using the above mentioned motifs for the inhibitors of ion channels of spiders, we succeeded in identifying 15 homologs of another known protein, μ-agatoxin 2 from the spider *Agelenopsis aperta* (Fig. 5). These structures belong to group B and carry all 3 structural motifs—PQM, PSM, and ESM. Only μ-2Aga_16 has a modified PSM and lacks a propeptide (lacks PQM). Seven more structures of putative ion channel blockers were found (Fig. 6); three of them were assigned to structural group A, while the remaining sequences were placed in group B. All these structures are unique, except for AgorTX_B7b and AgorTX_B7c, which differ in a single substitution of Ile for Leu. Three structures, AgorTX-B7a–c, revealed some sequence homology to ω-agatoxin-IIIA[58] from a closely related species, *A. aperta*; the remaining sequences showed no homology to known proteins. Homology level of all deduced structures with known sequences was determined with the BLAST method, and the data obtained are given in Table II.

DISCUSSION

In this work, an *A. orientalis* EST database was analyzed using 2 different systemic approaches. Using the traditional approach, which includes clustering of sequences in contigs, determination of consensus sequences, and its analysis using available databases, 6 groups of putative ion channel blockers were identified. The second approach, involving a search for specific structural mark-

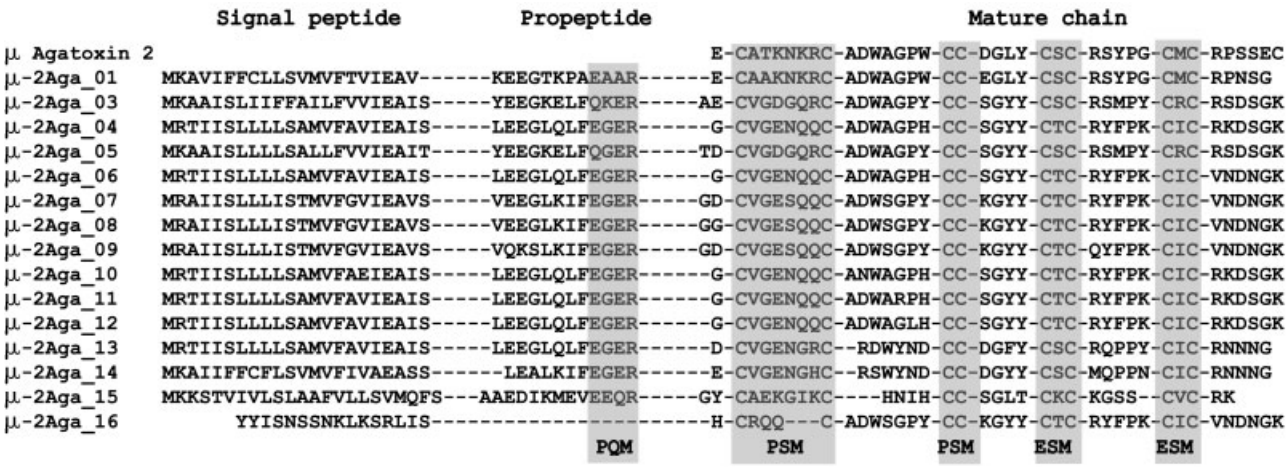

Fig. 5. Structural alignment of μ-agatoxin 2 from *A. aperta*[17] and toxinlike structures deduced from the *A. orientalis* data bank. Sequences corresponding to structural markers are shadowed.

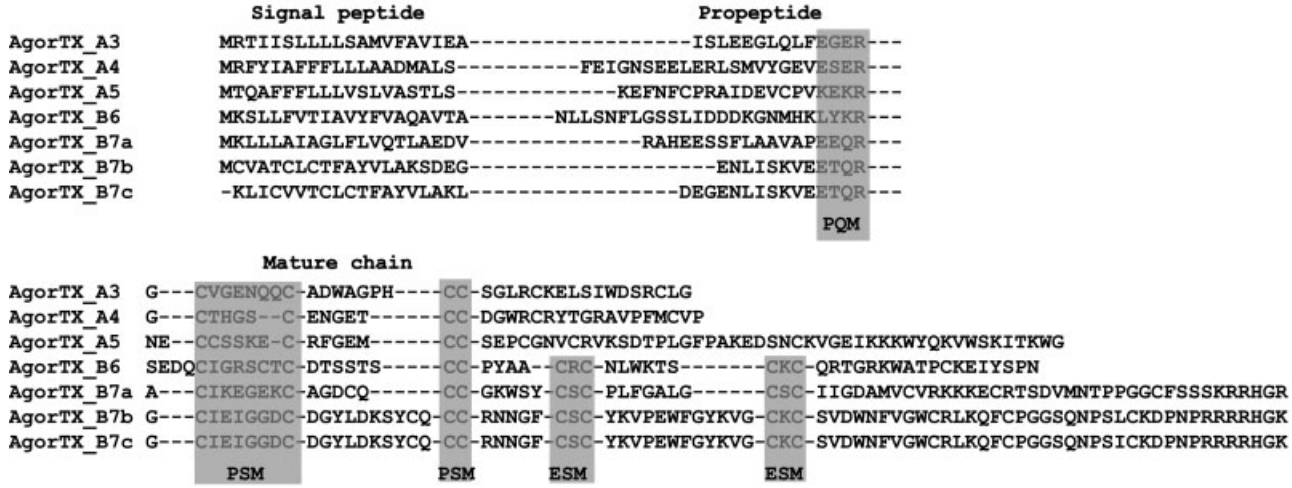

Fig. 6. Deduced amino acid sequences of dissimilar toxin precursors from the *A. orientalis* EST data bank divided onto mature chain, propeptide, and signal peptide in accordance with PQM marker position. Structural markers (PQM, PSM, and ESM) are shadowed.

ers (PSM, ESM, PQM) in translated sequences, led to identification of 3 structurally related groups and 5 unique structures. The results obtained by both approaches will be similar to those in the case of searches for structures homologous to already described sequences, for example, groups of agelenin, μ-agatoxin 2, and ω-agatoxin IIIA homologs. Polypeptide structures with low homology to known structures (see Table II) cannot be identified by the traditional procedure. Such an approach would lead to loss of at least 4 of the 48 structures of putative toxins in the data bank studied.

It should be noted that some ESTs contained gaps, substitutions, and unreadable nucleotides, resulting in stop or senseless codons. This partial EST considerably decreased the number of correct sequences, and together with other mistakes constituted more than 21% of all banks (Fig. 7). The traditional approach employing contigs is therefore preferable for quantitative analysis. It enables an estimate of the average distribution of the total group of homologous structures rather than of an individual sequence in the species examined.

In summary, it should be noted that both approaches to EST database analysis have advantages and limitations. Some results are more illustrative and reliable using the contig analysis, while other data are better interpreted using structural markers. Currently, the use of markers for data analysis is more laborious and less automated. Further development of genome analysis using structural motifs and the development of new algorithms for automated searches will become a good supplement to the widely used BLAST search. New algorithms using only direct comparison of amino acids in 2 sequences by searching for structural formalized characters will facilitate the identification of biologically active natural compounds in different species. Common structural motifs have been already discovered in several groups of proteins and are currently exploited for database analysis including ESTs.[59–62] A well-known thioredoxinlike motif, as well as

TABLE II. Analysis of an Agelena orientalis Data Bank for the Presence of the PQM, PSM, and ESM Structural Motifs Deduced for Ion Channel Inhibitors

Protein	Entry	Avg. F	%	Homolog	PQM	PSM	ESM
Agel 00	304	1358	100	Agelenin	+	+	−
Agel 01	547	1313	98	Agelenin	+	+	−
Agel 02	202	1346	98	Agelenin	+	+	−
Agel 03	27	1213	98	Agelenin	+	+	−
Agel 04	27	1126	89	Agelenin	+	+	−
Agel 05	4	1226	98	Agelenin	+	+	−
Agel 06	3	1073	96	Agelenin	+	+	−
Agel 07	2	1318	96	Agelenin	+	+	−
Agel 08	2	1308	96	Agelenin	+	+	−
Agel 09	2	1266	96	Agelenin	+	+	−
Agel 10	2	973	96	Agelenin	+	+	−
Agel 11	1	1433	96	Agelenin	+	+	−
Agel 12	1	1421	98	Agelenin	+	+	−
Agel 13	1	1396	96	Agelenin	+	+	−
Agel 14	1	1376	96	Agelenin	+	+	−
Agel 15	1	1359	98	Agelenin	+	+	−
Agel 16	1	1359	96	Agelenin	+	−	−
Agel 17	1	1218	92	Agelenin	+	+	−
Agel 18	1	1154	89	Agelenin	−	+	−
Agel 19	1	1152	89	Agelenin	+	+	−
Agel 20	1	471	89	Agelenin	+	+	−
Agel 21	1	333	92	Agelenin	+	+	−
Agel 22	63	1154	83	Agelenin	+	+	−
Agel 23	19	1049	81	Agelenin	+	+	−
Agel 24	2	1028	85	Agelenin	+	+	−
Agel 25	1	1198	81	Agelenin	+	+	−
AgorTX_A3	1	1131	—	Not found	+	+	−
AgorTX_A4	6	6	—	Not found	+	+/−	−
AgorTX_A5	2	1	—	Not found	+	+/−	−
μ-2Aga 01	10	64	95	μ-AgaTx2	+	+	+
μ-2Aga 03	1	44	78	μ-AgaTx2	+	+	+
μ-2Aga 04	71	897	76	μ-AgaTx2	+	+	+
μ-2Aga 05	3	184	78	μ-AgaTx2	+	+	+
μ-2Aga 06	1	1032	74	μ-AgaTx2	+	+	+
μ-2Aga 07	45	842	70	μ-AgaTx2	+	+	+
μ-2Aga 08	9	1019	70	μ-AgaTx2	+	+	+
μ-2Aga 09	1	1022	68	μ-AgaTx2	+	+	+
μ-2Aga 10	1	366	75	μ-AgaTx2	+	+	+
μ-2Aga 11	1	1023	75	μ-AgaTx2	+	+	+
μ-2Aga 12	1	1030	75	μ-AgaTx2	+	+	+
μ-2Aga 13	10	42	75	μ-AgaTx2	+	+	+
μ-2Aga 14	18	43	72	μ-AgaTx2	+	+	+
μ-2Aga 15	12	12	68	μ-AgaTx2	+	+	+
μ-2Aga 16	1	57	68	μ-AgaTx2	−	+/−	+
AgorTX_B6	2	1	—	Not found	+/−	+	+
AgorTX_B7a	21	22	49	ω-AgaTx3A	+	+	+
AgorTX_B7b	1	1	57	ω-AgaTx3A	+	+	+
AgorTX_B7c	1	1	58	ω-AgaTx3A	+	+	+

Value + designates the presence of a marker, − denotes the absence of a marker, +/− shows structural modifications in markers. The entry shows the total number of ESTs coding for the identical mature polypeptide chain. Average fidelity (Avg. F) was calculated for each protein as a mean for all entries. The percentage of homology to the nearest related structure was calculated by FASTA.

TRASH and THAP motifs, describe Cys residue distribution, whereas KDEL and FFAT motifs operate with other conserved amino acid sequences.

A question arises as to whether mutations discovered in spider genomes at the mRNA level occur naturally or represent experimental errors in gene structure determination. To answer this, it is necessary to consider the fidelity parameter of each entry (equal to the number of ESTs in the original sequence contig by cluster analysis). Obvious errors in structure determination caused by losses of gene fragments will lead to unique sequences and low fidelity values. However, the omission or misreading of one nucleotide, although changing the primary structure of the encoded protein, is not significant in a gene homology search, and the fidelity values for such entries will be rather high. Due to numerous noncoding regions, small

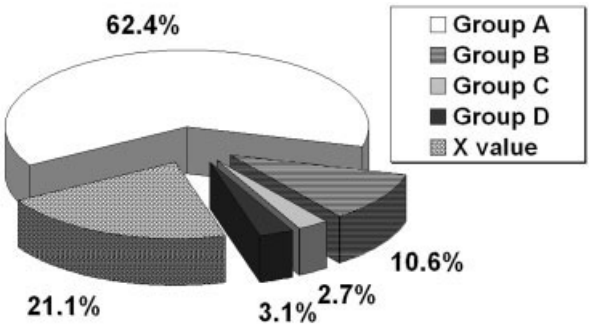

Fig. 7. Clustering of sequences in *A. orientalis* EST database. Group A corresponds to structures with PSM; group B comprises sequences with PSM and ESM; group C includes structures without markers; group D consists of incomplete sequences, and X denotes incorrect sequences.

gene fragments acquire high fidelity levels during computation. It follows from these considerations that both very low and high fidelity values for a particular entry cast doubt on the faithfulness of the results. The most reliable information may be obtained for those entries whose fidelity values are close to the number of deduced sequences in the group.

The distribution of deduced sequences in the *A. orientalis* data bank, based on the presence or absence of 2 main structural motifs, PSM and ESM, is shown in Figure 7. As is seen, the group of toxin-like structures without ESM (group A) prevails in the venom of *A. orientalis* (more than 62%). In this group, agelenin homologs comprise 90%, and 9% of ageleninlike structures are found in group X. The total number of ageleninlike structures estimated by the method used amounts to 1300. As mentioned above for Fig. 1(B), clones with the fidelity values above 1000 constitute 40%, and 26% of all entries. It follows that during analysis of EST contigs and deduced protein sequences, the fidelity is about 1000 and, consequently, the faithfulness of structures is high. Conversely, if the deduced mature structure is homologous to ageleninlike sequences, but its EST is not accurate, the fidelity value will be substantially below 1000, and the faithfulness of this structure will be problematic. For example, it is probable that Agel_20, Agel_21, and AgorTX-A3 are erroneous structures (Table II). For group B, accounting for 10% of the sequences, the picture is different. Eighty-eight percent of deduced polypeptides (204 entries) belong to μ-agatoxin 2 homologs. The average fidelity values for 15 polypeptides vary from 2 to 1000 depending on the protein. This can be explained by the fact that some toxinlike structures in the μ-agatoxin 2 group are similar to agelenin in the regions that differ from the mature polypeptide chain. This indicates that the genes for the corresponding polypeptides are related. Therefore we believe that virtually all deduced sequences are correct and hope to identify most of them in the spider's venom in the near future.

Even if some incomplete sequences have errors, the analysis of the *A. orientalis* cDNA bank and the data obtained for other spider toxins suggest ongoing point mutagenesis of proteins in arthropods, resulting in the appearance of slightly modified structures in addition to the evolutionarily conserved ones.

CONCLUSION

In this work, we established the principles of structural organization for spider polypeptide toxins. These principles for the ion channel blockers can be successfully applied for the analysis of sequences from conus snails, scorpions, and other species. We show that natural mutagenesis, which is difficult to detect by direct analysis of venom proteins, is an ongoing process in the evolution of the arthropod genome. Our result is good correlated with data obtained for α-KTx family of scorpion's toxins by the evolutionary trace (ET) analysis technique.[63]

Three formalized structural motifs for polypeptide ion channel blockers have been developed. The presence of these motifs in amino acid sequences suggests that they belong to the spider toxin class. Using these formalized parameters, we performed a detailed analysis of a cDNA bank from the venom glands of *A. orientalis*. We determined the structure of the agelenin precursor isolated earlier from the closely related species *A. opulenta* and discovered 25 new related polypeptides. Furthermore, we identified 15 toxin precursors with high homology to μ-agatoxin 2 and 7 toxinlike sequences, which have low structural homology to all known spider toxins.

This investigation forms the basis for further analysis of the maturation and functional expression of polypeptide toxins. [Nucleotide sequences and deduced amino acids structures have been deposited in the GenBank database under GenBank Accession Numbers AY681297–AY681344.]

ACKNOWLEDGMENTS

Our thanks to Maureen Dolan and Will Krespan for cDNA library construction and sequencing.

REFERENCES

1. Pallaghy PK, Nielsen KJ, Craik DJ, Norton RS. A common structural motif incorporating a cystine knot and a triple-stranded beta-sheet in toxic and inhibitory polypeptides. Protein Sci 1994;3: 1833–1839.
2. Narasimhan L, Singh J, Humblet C, Guruprasad K, Blundell T. Snail and spider toxins share a similar tertiary structure and "cystine motif." Nat Struct Biol 1994;1:850–852.
3. Escoubas P, Diochot S, Corzo G. Structure and pharmacology of spider venom neurotoxins. Biochimie 2000;82:893–907.
4. Grishin E. Polypeptide neurotoxins from spider venoms. Eur J Biochem 1999;264:276–280.
5. Quistad GB, Skinner WS. Isolation and sequencing of insecticidal peptides from the primitive hunting spider, *Plectreurys tristis* (Simon). J Biol Chem 1994;269:11098–11101.
6. Skinner WS, Adams ME, Quistad GB, Kataoka H, Cesarin BJ, Enderlin FE, Schooley DA. Purification and characterization of two classes of neurotoxins from the funnelweb spider, *Agelenopsis aperta*. J Biol Chem 1989;264:2150–2155.
7. Corzo G, Villegas E, Gomez-Lagunas F, Possani LD, Belokoneva OS, Nakajima T. Oxyopinins, large amphipathic peptides isolated from the venom of the wolf spider *Oxyopes kitabensis* with cytolytic properties and positive insecticidal cooperativity with spider neurotoxins. J Biol Chem 2002;277:23627–23637.
8. Johnson JH, Bloomquist JR, Krapcho KJ, Kral RMJ, Trovato R, Eppler KG, Morgan TK, DelMar EG. Novel insecticidal peptides from *Tegenaria agrestis* spider venom may have a direct effect on

the insect central nervous system. Arch Insect Biochem Physiol 1998;38:19–31.
9. Lipkin A, Kozlov S, Nosyreva E, Blake A, Windass JD, Grishin E. Novel insecticidal toxins from the venom of the spider *Segestria florentina*. Toxicon 2002;40:125–130.
10. Inui T, Hagiwara K, Nakajima K, Kimura T, Nakajima T, Sakakibara S. Synthesis and disulfide structure determination of agelenin: identification of the carboxy-terminus as an amide form. Peptide Res 1992;5:140–144.
11. Yan L, Adams ME. The spider toxin omega-Aga IIIA defines a high affinity site on neuronal high voltage-activated calcium channels. J Biol Chem 2000;275:21309–21316.
12. Mintz IM, Venema VJ, Adams ME, Bean BP. Inhibition of N- and L-type Ca2+ channels by the spider venom toxin omega-Aga-IIIA. Proc Natl Acad Sci USA 1991;88:6628–6631.
13. Mintz IM, Venema VJ, Swiderek KM, Lee TD, Bean BP, Adams ME. P-type calcium channels blocked by the spider toxin omega-Aga-IVA. Nature 1992;355:827–829.
14. Scott RH, Dolphin AC, Bindokas VP, Adams ME. Inhibition of neuronal Ca2+ channel currents by the funnel web spider toxin omega-Aga-IA. Mol Pharmacol 1990;38:711–718.
15. Vieira LB, Kushmerick C, Reis HJ, Diniz CR, Cordeiro MN, Prado MA, Kalapothakis E, Romano-Silva MA, Gomez MV. PnTx3-6 a spider neurotoxin inhibits K+-evoked increase in [Ca2+](i) and Ca2+-dependent glutamate release in synaptosomes. Neurochem Int 2003;42:277–282.
16. Stapleton A, Blankenship DT, Ackermann BL, Chen TM, Gorder GW, Manley GD, Palfreyman MG, Coutant JE, Cardin AD. Curtatoxins: neurotoxic insecticidal polypeptides isolated from the funnel-web spider *Hololena curta*. J Biol Chem 1990;265:2054–2059.
17. Skinner WS, Adams ME, Quistad GB, Kataoka H, Cesarin BJ, Enderlin FE, Schooley DA. Purification and characterization of two classes of neurotoxins from the funnel web spider, *Agelenopsis aperta*. J Biol Chem 1989;264:2150–2155.
18. Corzo G, Escoubas P, Stankiewicz M, Pelhate M, Kristensen CP, Nakajima T. Isolation, synthesis and pharmacological characterization of delta-palutoxins IT, novel insecticidal toxins from the spider *Paracoelotes luctuosus* (Amaurobiidae). Eur J Biochem 2000;267:5783–5795.
19. Kushmerick C, Kalapothakis E, Beirao PS, Penaforte CL, Prado VF, Cruz JS, Diniz CR, Cordeiro MN, Gomez MV, Romano-Silva MA, Prado MA. *Phoneutria nigriventer* toxin Tx3-1 blocks A-type K+ currents controlling Ca2+ oscillation frequency in GH3 cells. J Neurochem 1999;72:1472–1481.
20. Cordeiro MN, Diniz CR, Valentim AC, von Eickstedt VR, Gilroy J, Richardson M. The purification and amino acid sequences of four Tx2 neurotoxins from the venom of the Brazilian "armed" spider *Phoneutria nigriventer*. FEBS Lett 1992;310:153–156.
21. Huang X, Madan A. CAP3: a DNA sequence assembly program. Genome Res 1999;9:868–877.
22. Craik DJ, Daly NL, Waine C. The cystine knot motif in toxins and implications for drug design. Toxicon 2001;39:43–60.
23. Norton RS, Pallaghy PK. The cystine knot structure of ion channel toxins and related polypeptides. Toxicon 1998;36:1573–1583.
24. Fletcher JI, Smith R, O'Donoghue SI, Nilges M, Connor M, Howden ME, Christie MJ, King GF. The structure of a novel insecticidal neurotoxin, omega-atracotoxin-HV1, from the venom of an Australian funnel web spider. Nat Struct Biol 1997;4:559–566.
25. Szeto TH, Wang XH, Smith R, Connor M, Christie MJ, Nicholson GM, King GF. Isolation of a funnel-web spider polypeptide with homology to mamba intestinal toxin 1 and the embryonic head inducer Dickkopf-1. Toxicon 2000;38:429–442.
26. Hagiwara K, Sakai T, Miwa A, Kawai N, Nakajima T. Complete amino acid sequence of a new type of neurotoxin from the venom of the spider, *Agelena opulenta*. Biomed Res 1990;11:181–186.
27. Skinner WS, Dennis PA, Li JP, Quistad GB. Identification of insecticidal peptides from venom of the trap-door spider, *Aptostichus schlingeri* (Ctenizidae). Toxicon 1992;30:1043–1050.
28. Kuhn-Nentwig L, Schaller J, Nentwig W. Purification of toxic peptides and the amino acid sequence of CSTX-1 from the multicomponent venom of *Cupiennius salei* (Araneae:Ctenidae). Toxicon 1994;32;287–302.
29. Gunning SJ, Chong Y, Khalife AA, Hains PG, Broady KW, Nicholson GM. Isolation of delta-missulenatoxin-Mb1a, the major vertebrate-active spider delta-toxin from the venom of *Missulena bradleyi* (Actinopodidae). FEBS Lett 2003;554:211–218.
30. Krapcho KJ, Kral RMJ, Vanwagenen BC, Eppler KG, Morgan TK. Characterization and cloning of insecticidal peptides from the primitive weaving spider *Diguetia canities*. Insect Biochem Mol Biol 1995;25:991–1000.
31. Suchyna TM, Johnson JH, Hamer K, Leykam JF, Gage DA, Clemo HF, Baumgarten CM, Sachs F. Identification of a peptide toxin from *Grammostola spatulata* spider venom that blocks cation-selective stretch-activated channels. J Gen Physiol 2000;115:583–598.
32. Swartz KJ, MacKinnon R. An inhibitor of the Kv2.1 potassium channel isolated from the venom of a Chilean tarantula. Neuron 1995;15:941–949.
33. Oswald RE, Suchyna TM, McFeeters R, Gottlieb P, Sachs F. Solution structure of peptide toxins that block mechanosensitive ion channels. J Biol Chem 2002;277:34443–34450.
34. Liu Z, Dai J, Chen Z, Hu W, Xiao Y, Liang S. Isolation and characterization of hainantoxin-IV, a novel antagonist of tetrodotoxin-sensitive sodium channels from the Chinese bird spider *Selenocosmia hainana*. Cell Mol Life Sci 2003;60:972–978.
35. Sanguinetti MC, Johnson JH, Hammerland LG, Kelbaugh PR, Volkmann RA, Saccomano NA, Mueller AL. Heteropodatoxins: peptides isolated from spider venom that block Kv4.2 potassium channels. Mol Pharmacol 1997;51:491–498.
36. Peng K, Shu Q, Liu Z, Liang S. Function and solution structure of huwentoxin-IV, a potent neuronal tetrodotoxin (TTX)-sensitive sodium channel antagonist from Chinese bird spider *Selenocosmia huwena*. J Biol Chem 2002;277:47564–47571.
37. Corzo G, Gilles N, Satake H, Villegas E, Dai L, Nakajima T, Haupt J. Distinct primary structures of the major peptide toxins from the venom of the spider *Macrothele gigas* that bind to sites 3 and 4 in the sodium channel. FEBS Lett 2003;547:43–50.
38. Adams ME, Bindokas VP, Hasegawa L, Venema VJ. Omega-agatoxins: novel calcium channel antagonists of two subtypes from funnel web spider (*Agelenopsis aperta*) venom. J Biol Chem 1990;265:861–867.
39. Adams ME, Mintz IM, Reily MD, Thanabal V, Bean BP. Structure and properties of omega-agatoxin IVB, a new antagonist of P-type calcium channels. Mol Pharmacol 1993;44:681–688.
40. Kalapothakis E, Penaforte CL, Leao RM, Cruz JS, Prado VF, Cordeiro MN, Diniz CR, Romano-Silva MA, Prado MAM, Gomez MV, Beirao PSL. Cloning, cDNA sequence analysis and patch clamp studies of a toxin from the venom of the armed spider (*Phoneutria nigriventer*). Toxicon 1998;36:1971–1980.
41. Kalapothakis E, Penaforte CL, Beirao PS, Romano-Silva MA, Cruz JS, Prado MA, Guimaraes PE, Gomez MV, Prado VF. Cloning of cDNAS encoding neurotoxic peptides from the spider *Phoneutria nigriventer*. Toxicon 1998;36:1843–1850.
42. Diniz CR, Cordeiro MN, Junor LR, Kelly P, Fischer S, Reimann F, Oliveira EB, Richardson M. The purification and amino acid sequence of the lethal neurotoxin Tx1 from the venom of the Brazilian "armed" spider *Phoneutria nigriventer*. FEBS Lett 1990; 263:251–253.
43. Cordeiro MN, de Figueiredo SG, Valentim AC, Diniz CR, von Eickstedt VR, Gilroy J, Richardson M. Purification and amino acid sequences of six Tx3 type neurotoxins from the venom of the Brazilian "armed" spider *Phoneutria nigriventer* (Keys). Toxicon 1993;31:35–42.
44. de Figueiredo SG, de Lima ME, Nascimento Cordeiro M, Diniz CR, Patten D, Halliwell RF, Gilroy J, Richardson M. Purification and amino acid sequence of a highly insecticidal toxin from the venom of the Brazilian spider *Phoneutria nigriventer* which inhibits NMDA-evoked currents in rat hippocampal neurones. Toxicon 2001;39:309–317.
45. Kaiser II, Griffin PR, Aird SD, Hudiburg S, Shabanowitz J, Francis B, John TR, Hunt DF, Odell GV. Primary structures of two proteins from the venom of the Mexican red knee tarantula (*Brachypelma smithii*). Toxicon 1994;32:1083–1093.
46. Middleton RE, Warren VA, Kraus RL, Hwang JC, Liu CJ, Dai G, Brochu RM, Kohler MG, Gao YD, Garsky VM, Bogusky MJ, Mehl JT, Cohen CJ, Smith MM. Two tarantula peptides inhibit activation of multiple sodium channels. Biochemistry 2002;41:14734–14747.
47. Sheumack DD, Claassens R, Whiteley NM, Howden ME. Complete amino acid sequence of a new type of lethal neurotoxin from

the venom of the funnel-web spider *Atrax robustus*. FEBS Lett 1985;181:154–156.
48. Marvin L, De E, Cosette P, Gagnon J, Molle G, Lange C. Isolation, amino acid sequence and functional assays of SGTx1. The first toxin purified from the venom of the spider *Scodra griseipes*. Eur J Biochem 1999;265:572–579.
49. Newcomb R, Palma A, Fox J, Gaur S, Lau K, Chung D, Cong R, Bell JR, Horne B, Nadasdi L et al. SNX-325, a novel calcium antagonist from the spider *Segestria florentina*. Biochemistry 1995;34:8341–8347.
50. Newcomb R, Szoke B, Palma A, Wang G, Chen XH, Hopkins W, Cong R, Miller J, Urge L, Tarczy-Hornoch K, Loo JA, Dooley DJ, Nadasdi L, Tsien RW, Lemos J, Miljanich G. Selective peptide antagonist of the class E calcium channel from the venom of the tarantula *Hysterocrates gigas*. Biochemistry 1998;37:15353–15362.
51. Liang SP, Zhang DY, Pan X, Chen Q, Zhou PA. Properties and amino acid sequence of huwentoxin-I, a neurotoxin purified from the venom of the Chinese bird spider *Selenocosmia huwena*. Toxicon 1993;31:969–978.
52. Ostrow KL, Mammoser A, Suchyna T, Sachs F, Oswald R, Kubo S, Chino N, Gottlieb PA. cDNA sequence and in vitro folding of GsMTx4, a specific peptide inhibitor of mechanosensitive channels. Toxicon 2003;42:263–274.
53. Santos AD, Imperial JS, Chaudhary T, Beavis RC, Chait BT, Hunsperger JP, Olivera BM, Adams ME, Hillyard DR. Heterodimeric structure of the spider toxin omega-agatoxin IA revealed by precursor analysis and mass spectrometry. J Biol Chem 1992;267:20701–20705.
54. Diniz MR, Paine MJ, Diniz CR, Theakston RD, Crampton JM. Sequence of the cDNA coding for the lethal neurotoxin Tx1 from the Brazilian "armed" spider *Phoneutria nigriventer* predicts the synthesis and processing of a preprotoxin. J Biol Chem 1993;268:15340–15342.
55. Cordeiro MN, Diniz CR, Valentim AC, von Eickstedt VR, Gilroy J, Richardson M. The purification and amino acid sequences of four Tx2 neurotoxins from the venom of the Brazilian "armed" spider *Phoneutria nigriventer*. FEBS Lett 1992;310:153–156.
56. von Heijne G. Patterns of amino acids near signal-sequence cleavage sites. Eur J Biochem 1983;133:17–21.
57. Perlman D, Halvorson HO. A putative signal peptidase recognition site and sequence in eukaryotic and prokaryotic signal peptides. J Mol Biol 1983;167:391–409.
58. Venema VJ, Swiderek KM, Lee TD, Hathaway GM, Adams ME. Antagonism of synaptosomal calcium channels by subtypes of omega-agatoxins. J Biol Chem 1992;267:2610–2615.
59. Hosoda A, Kimata Y, Tsuru A, Kohno K. JPDI, a novel endoplasmic reticulum-resident protein containing both a BiP-interacting J-domain and thioredoxin-like motifs. J Biol Chem 2003;278:2669–2676.
60. Loewen CJ, Roy A, Levine TP. A conserved ER targeting motif in three families of lipid binding proteins and in Opi1p binds VAP. EMBO J 2003;22:2025–2035.
61. Ettema TJ, Huynen MA, de Vos WM, van der Oost J. TRASH: a novel metal-binding domain predicted to be involved in heavy-metal sensing, trafficking and resistance. Trends Biochem Sci 2003;28:170–173.
62. Roussigne M, Kossida S, Lavigne AC, Clouaire T, Ecochard V, Glories A, Amalric F, Girard JP. The THAP domain: a novel protein motif with similarity to the DNA-binding domain of P element transposase. Trends Biochem Sci 2003;28:66–69.
63. Zhu S, Huys I, Dyason K, Verdonck F, Tytgat J. Evolutionary trace analysis of scorpion toxins specific for K-channels. Proteins 2004;54:361–370.