RESEARCH ARTICLE

Proteomic analysis of the venom of *Heterometrus longimanus* (Asian black scorpion)

Scott Bringans[1], Soren Eriksen[1], Tulene Kendrick[1], P. Gopalakrishnakone[2], Andreja Livk[1], Robert Lock[1] and Richard Lipscombe[1]

[1] Proteomics International, Perth, Western Australia, Australia
[2] Venom and Toxin Research Programme, Department of Anatomy, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

Received: October 8, 2007
Revised: November 5, 2007
Accepted: November 16, 2007

Venoms have evolved over millions of years into potent cocktails of bioactive peptides and proteins. These compounds can be of great value to the pharmaceutical industry for numerous clinical applications. In this study, a novel proteomic – bioinformatic approach was utilised, where chromatography followed by gel electrophoresis was utilised to separate the venom peptides/proteins of *Heterometrus longimanus (Asian black scorpion)*. Purified peptides were analysed by tandem mass spectrometry, *de novo* sequenced and then homology matched against known peptides in the Swiss-Prot protein database. Numerous potentially biologically active peptide matches were discovered, and a simple scoring system applied to putatively assign functions to the peptides. As a validation of this approach, the functional composition of the experimentally derived proteome is similar to that of other scorpions, and contains a potent mix of toxins, antimicrobials and ionic channel inhibitors.

Keywords:
Bioinformatics / *de novo* sequencing / Peptidomics

1 Introduction

In 2002, more than 2500 bioactive compounds derived from venom were listed in the literature [1]. As a single example, researchers working with cone snails (the *Conidae*) estimate that the venoms of these genera contain more than 50 000 different small peptides, the so-called conotoxins. As of 2003, less than 0.1% of the conotoxins had been characterized pharmacologically [2]. Bioactive peptides can be of great value for the pharmaceutical industry, serving as lead molecules for new drugs or as diagnostic or proof-of-concept tools. The usefulness of toxin peptides lies in their capacity to evade regular control mechanisms through their independence of co-factors and their non-recognition by inhibitors. They are highly specific in their targeting, which is useful in terms of minimizing side effects of any new drug that might be developed.

Venom biodiscovery has traditionally utilised bioassay-guided fractionation, where fractions displaying the desired activity in a specific assay are further investigated. However, venoms are widely recognized to contain many potentially interesting peptides with a broad range of activities. The use of a bioassay-guided approach means that many potential applications of venoms are not investigated, as they do not possess the particular niche of activity being researched. In this study a novel approach was applied. The venom peptides were *de novo* sequenced and characterised based on homology in amino acid composition with known peptides from public databases.

The sensitivity and high-throughput nature of modern mass spectrometers enables the implementation of a proteomic approach to rational drug discovery. The technical situation has changed dramatically over the last decade and

Correspondence: Dr. Scott Bringans, Proteomics International, PO Box 6064, East Perth, Western Australia 6892, Australia
E-mail: scott@proteomics.com.au
Fax: +61-8-9360-2856

Abbreviation: **E-value**, expectation value

using MS coupled with other sensitive methods, it is now possible to identify and characterize peptides of low abundance that would have been overlooked previously.

The objective of this study was to develop a preliminary method to investigate the proteome of a previously uncharacterised venom by utilising the power of MS to effectively map all proteins and peptides present. Peptides derived from the venom were automatically *de novo* sequenced, and matched for homology to sequences in publicly available databases. When mapping an uncharacterised proteome, many of the sequences may be unique. Therefore, while exact identification of some peptides may be possible, it is likely that most will only show partial homology to known sequences. To further characterise the peptides identified, we utilised a simple scoring system to putatively assign functions to the peptides, based on homology to known proteins. This bioinformatic approach describes an effective method to provide an initial characterisation of a previously unstudied proteome.

2 Materials and methods

2.1 Supply of venom

Lyophilized *Heterometrus longimanus (Asian black scorpion)* venom was obtained from a colony of captured scorpions maintained and continuously milked of venom as described by Gopalakrishnakone *et al.* [3].

2.2 Alkylation

Approximately 0.5 mg lyophilized venom was alkylated by the standard method: 500 μL of 10 mM DTT in 25 mM ammonium bicarbonate was added and the solution incubated for 1 h at 56°C. Subsequently, 500 μL of 55 mM iodoacetamide in 25 mM ammonium bicarbonate was added, and the fractions were incubated in the dark at room temperature for 45 min.

2.3 HPLC analysis

All RP-HPLC was performed on a Hewlett Packard Series 1100 HPLC consisting of automated liquid sampler, quaternary pump, degasser, column compartment and a diode array detector (observing at 220 nm), controlled by the ChemStation© software (Agilent Technologies). The column used was a Vydac 5 micron protein-peptide C18 column, with internal dimensions 2 mm × 150 mm. Two solvents were used to elute the loaded material: solvent A: 0.06% TFA in milli-Q water and solvent B: 0.06% TFA in 100% ACN.

The alkylated venom was loaded on to the column and eluted by a 2–40% linear gradient of ACN (0.06% TFA) in 55 min with a flow rate of 0.5 mL/min. Test runs using small amounts of venom (10 μg) indicated that all peptides eluted in this interval. The gradient was then raised from 40 to 100% ACN in 10 min, to wash off any material still binding to the column. Fractions were collected based on peaks identified by test runs and subsequently vacuum-dried.

2.4 1-DE

The alkylated fractions were concentrated by rotary evaporation to approximately 20 μL with 5 μL of concentrated (5 ×) non-reducing SDS-PAGE sample buffer added (20% SDS, 50% glycerol, 2.25 M Tris-HCl pH 6.8, trace of bromophenol blue). The fractions were incubated at room temperature for 30 min and loaded on to 12-lane 10–20% Tris-Tricine Peptide CriterionTM Precast Gels (Bio-Rad, CA, USA) connected to a PowerPac BasicTM power source (Bio-Rad). Precision Plus markers (5 μL; Bio-Rad), with a molecular mass range of 10 to 250 kDa, were loaded on both sides of each gel. Voltage was applied until the tracking dye had migrated out of the gel (120 min). The gels were stained with "Blue Silver" Coomassie stain [4] overnight and the background destained overnight in 10% acetic acid. Additional silver staining of the gels was performed with all visible bands (both Coomassie and silver) excised and stored individually at 4°C pending further analysis.

Excised Coomassie gel bands were de-stained by three 45-min washes with 25 mM ammonium bicarbonate in 50:50 ACN:water. Silver-stained bands were destained by two 10-min incubations in a 1:1 solution of 30 mM potassium ferricyanide and 100 mM sodium thiosulfate, followed by two 1-h washes with water and one 10-min wash with 25 mM ammonium bicarbonate in 50:50 ACN:water. De-stained and washed gel pieces were vacuum-dried and stored at −20°C pending tryptic digestion.

2.5 Digestion

The following protocol was applied to all excised and dried gel bands: 10 μL trypsin digest solution (12.5 μg/mL trypsin, 25 mM ammonium bicarbonate) was added to each gel piece and incubated overnight at 37°C. The digested peptides were extracted by two 20-min incubations with 10–20 μL ACN containing 1% TFA, depending on the size of the gel piece. Exceptionally, three extractions were applied for large-size gel pieces. The pooled extracts were dried by rotary evaporation and stored at −20°C pending further analysis by MS.

2.6 MS

MALDI TOF/TOF MS/MS was performed on a 4800 MALDI TOF/TOFTM analyzer (Applied Biosystems/MDS SCIEX).

The dry peptide samples were reconstituted in 2 μL standard diluent (30:70 ACN:water). The resulting solution was diluted 1:10 with matrix solution (CHCA, 10 mg/mL) and spotted on a 384-well Opti-TOF stainless steel plate. The spotted samples were analysed using a first run of standard TOF MS. The system was set to perform a second run of MS/

MS focused on the 15 most intensive peaks of the first MS (excluding peaks known to be trypsin). The laser was set to fire 400 times per spot in MS mode and 2000 times per spot in MS/MS mode. Laser intensity was 2800 J (MS) and 3900 J (MS/MS). A mass range of 400–4000 amu with a focus mass of 2100 amu was used.

De novo sequencing of MALDI TOF/TOF derived sequences was performed automatically by the DeNovo Explorer™ version 3.6 software (Applied Biosystems) with the following settings: enzyme: trypsin; fixed modification: carbamidomethyl (C); mass tolerance: 0.2. This software automatically generates candidate sequences and assigns them a score between 0 and 100. The score indicates the degree of matching between the theoretical fragmentation pattern and the fragmentation spectrum that corresponds to the peaks in the peak list. The highest scoring sequence of each peptide was chosen for further analysis, and primary database searching.

All *de novo*-derived sequences from MS/MS spectra were submitted for BLAST search at www.ncbi.nlm.nih.gov/BLAST/, using the following settings for the non-redundant Swiss-Prot database:

Taxonomy = Arachnida (taxid:6854)
Expect threshold = 20000
Word size = 2
Matrix = PAM30
Gap costs = Existence:9 Extension:1
Compositional adjustments = No adjustment
Filter = not selected.

2.7 Assignment of putative functions to gel bands

Following BLAST searching, the *de novo*-derived peptides determined from each gel band were putatively assigned a function using the following scoring system. The most common functions identified from the BLAST results were divided into nine categories: K^+ channel inhibitor, Na^+ channel inhibitor, Ca^{2+} channel inhibitor, Cl^- channel inhibitor, antimicrobial, protease inhibitor, neurotoxin, insect toxin and dermonecrotic toxin. For each peptide, the functions of the homologous BLAST hits [with the three highest expectation values (E-values)] were evaluated from the Swiss-Prot entry. A score of 1 point was added to a functional category each time this function was found in the hits with the top three highest E-values. If the sequence match was to the same protein with an equivalent E-value but from different species, the match was only scored 1 point (for example, band 2 in Table 1 matched the sequence DCLPHL of a Ca^{2+} channel inhibitor in three different species with an E-value = 0.052, this was only given 1 point total for the three matches). Where two or more peptides were generated from a band, points were scored for each peptide as above, and the total points for the band calculated. An additional point was then awarded for each time a protein belonging to the same family was identified in a pair of peptides. The gel band was putatively assigned the function with the highest overall score. Where two or more functions had equivalent scores, the function was labelled as undetermined.

3 Results and discussion

Nineteen fractions were collected during the initial gradient of 2–40% ACN (Fig. 1A). Another two fractions were collected during the final gradient of 40–100% ACN (not displayed). Two large peaks were visible after 4 and 6 min of the run (not displayed) and were due to salts and the large excess of DTT and iodoacetamide added to the venom during the alkylation step. The scorpion peptides appeared to elute evenly along the gradient, from approximately 5 to 40% ACN.

The two gels were stained with "Blue Silver" (Fig. 1B) and 40 visible bands were excised. A further seven bands (41–47) were visualised with additional silver staining but none of these additional silver stained bands produced MS/MS data of sufficient quality for *de novo* sequencing. The fractions 6, 8, 10, 11 and 12 each contained one or two highly abundant low-molecular mass proteins whilst fractions 13 to 21 displayed a wider variety of sizes. From the results of the gels it can be concluded that the majority of the material is of 15 kDa or less. This is consistent with existing literature on the composition of scorpion venom.

All *H. longimanus* derived samples were analysed by MALDI TOF/TOF MS as described in Section 2. MALDI TOF/TOF-derived data were sequenced automatically by the DeNovo Explorer™ software, which generates sequences based on the 'best fit' to the MS/MS data after considering the complexity of the spectra and numerous possible a, b, c, x, y, z ion combinations. The sequence with the top score was submitted for further bioinformatic analysis.

The *de novo*-generated peptide sequences (grey shaded bars, Table 1) were submitted for BLAST searching at the NCBI BLAST website using settings based on those suggested in the help section of the website for short peptide sequences (detailed in Section 2). Traditionally, alignments obtained from BLAST searches are assigned an E-value. In general, the lower the E-value, the better the homology, however, the true meaning of this number is more complex. Larger databases will tend towards providing much larger E-values, as the chance that random sequences will match is higher. In addition, searches with short sequences can be virtually identical and still have relatively high E-values. This is because the calculation of the E-value also takes into account the length of the query sequence, and shorter sequences have a high probability of occurring in the database purely by chance. Therefore, for this application, two factors reduce the usefulness of the E-value scoring system. First, the database may be of sufficient size to generate false positive matches, and secondly, as MS-derived sequences are generally short in nature, the E-values will be correspondingly high. In this approach, we have utilised a high E-value

Table 1. Identification of homology matches to peptides derived from a selection of five bands isolated from *Heterometrus longimanus* venom[a)]

Band	Sequence (*de novo* score)	Precursor ion mass	MW From gel	Identity and accession (Swiss-Prot)	Organism	Function	Size (resid.)	Swiss-Prot E-value
1	RDYGPCK (83.4)	**895.27**	**10 kDa**					
	YGPCK			K^+ ch[b)]. toxin alpha-KTx 6.10 (Q6XLL5)	Opistophthalmus carinatus (Scorpion)	K+ ch. inh[c)].	60	0.38 (1st hit)
	DYG CK			Omega-agatoxin-4A (P30288)	Agelepnosis aperta (Spider)	Ca^{2+} ch. inh.	48	1.2 (2nd hit)
	DG PCK			Kunitz-type serine protease inh. BmTI-6 (P83606)	Rhipicephalus micro-plus (Tick)	Serine protease inh.	35	5.4 (3rd hit)
	ECYGPCKEKTGCSSSK (83.2)	**1877.58**	**10 kDa**					
	+CYGPCK++TGC++SK			K^+ ch. toxin alpha-KTx 6.10 (Q6XLL5)	Opistophthalmus carinatus (Scorpion)	K+ ch. inh.	60	9e-7 (1st hit)
	+CY PC+++TGC ++K			K^+ ch. toxin alpha-KTx 6.2 (Mauroto-xin) (P80719)	Scorpio maurus palmatus (Scorpion)	K+ ch. inh.	34	0.003 (2nd hit)
	+CY PC ++TGC ++K			K^+ ch. toxin alpha-KTx 6.13 (Spinoxin) (P84094)	Heterometrus spinifer (Scorpion)	K+ ch. inh.	34	0.005 (3rd hit)
2	RDCLPHLGAR (79.1)	**1194.55**	**10 kDa**					
	DCLPHL			Opicalcin 1 and 2 precursors (P60252)	Opistophthalmus carinatus (Scorpion)	Ca2+ ch. inh.	66	0.052 (=1st hit)
	DCLPHL			Imperatoxin-A (P59868; 1IE6)	Pandinus impera-tor (Scorpion)	Ca2+ ch. inh.	33	0.052 (=1st hit)
	DCLPHL			Maurocalcin (P60254)	Scorpio maurus palma-tus (Scorpion)	Ca2+ ch. inh.	33	0.052 (=1st hit)
	CLPH			Agelenin and Agelenin precursor (Q5Y4Y5)	Agelena orienta-lis & A. Opu-lenta (Spider)	Ca2+ ch. inh.	35/70	4.3 (=2nd hit)
	RDCLPHLG			Stromatoxin-1 (P60991)	Stromatopelma calceatum (Spider)	K+ ch. inh.	34	4.3 (=2nd hit)

Table 1. Continued

Band	Sequence (*de novo* score)	Precursor ion mass	MW From gel	Identity and accession (Swiss-Prot)	Organism	Function	Size (resid.)	Swiss-Prot E-value
	DC PHLG			Ceratotoxin-3 (P84509)	Ceratogyrus cornuatus (Spider)	Na+ ch. inh.	39	5.8 (=3rd hit)
	RDCL			Moubatin precursor (Q04669)	Ornithodoros moubata (Tick)	Anti-platelet aggregation	171	5.8 (=3rd hit)
	DCLP			Female-specific hista-mine-binding protein 1 precursor (O77420)	Rhipicephalus appendi-culatus (Tick)	Suppresses inflamma-tion	190	5.8 (=3rd hit)
	KGSAEKR (64.3)	**775.33**	**10 kDa**					
	AEKR			Imperatoxin-A (P59868)	Pandinus imperator (Scorpion)	Ca2+ ch. inh.	33	13 (=1st hit)
	AEKR			A-latrocrustotoxin (Q9XZCO)	Lactrodectus mactans (Spider)	Presynaptic neurotoxin	1395	13 (=1st hit)
	SAEK			Putative B-neu-rotoxin precursor (Q1T165)	Tityus zulianus (Scorpion)	Na+ ch. inh.	69	24 (2nd hit)
	KG EKR			Psalmopeotoxin-1 precursor(P0C201)	Psalmopoeus cambridgei (Spider)	Antimalarial	86	32 (3rd hit)
4	AGGGCYR (73.1)	**740.21**	**7 kDa**					
	GCYR			K^+ ch. toxin kappa-KTx 1.3 (P83655)	Heterometrus spinifer (Scorpion)	K+ ch. inh.	23	3 (1st hit)
	GGCY			Phaiodotoxin-3 and -2 and precursor (Q5MJP3)	Anuroctonus phaiodac-tylus (Scorpion)	Na+ ch. inh.	72	5.4 (=2nd hit)
	GGCY			Androctonin (P56684)	Androctonus australis (Scorpion)	Anti-microbial	25	5.4 (=2nd hit)

Table 1. Continued

Band	Sequence (*de novo* score)	Precursor ion mass	MW From gel	Identity and accession (Swiss-Prot)	Organism	Function	Size (resid.)	Swiss-Prot E-value
	KLCVCYE (83.9)	**971.35**	**7 kDa**					
	K CVCY+			Venom protein PN16C3 (P84032)	Phoneutria nigriventer (Spider)	Cysteine protease inh.	128	0.21 (1st hit)
	CVCY			K^+ ch. toxin a-KTX 15.7, 15.1, 15.3, 15.8, 15.5, 15.4, 15.2 (Q5KOE0)	Androctonus amoreuxi & australis, Mesobuthus martensii (Scorpion)	K+ ch. inh.	37-60	1.2 (=2nd hit)
	VCYE			Kunitz serine protease inh. (P84556)	Rhipicephalus sanguineus (Tick)	Serine protease inh.	25	2.2 (3rd hit)
	ASTCYR (61.8)	**757.23**	**7 kDa**					
	TCYR			Neurotoxin 8-related gene product 1/2/3 precursor (Q2YHM1)	Androctonus mauretanicus mauretanicus (Scorpion)	Na+ ch. inh.	86	1.9 (=1st hit)
	TCYR			4 kDa defensin (P41965)	Leiurus quinquestriatus hebraeus (Scorpion)	Antimicrobial	38	1.9 (=1st hit)
	TCYR			Defensin precursor (Q86QI5)	Dermacentor variabilis (Tick)	Antimicrobial	74	1.9 (=1st hit)
	TCYR			4 kDa defensin (P56686)	Androctonus australis (Scorpion)	Antimicrobial	37	1.9 (=1st hit)
	STCY			Kunitz-type serine protease inh. BmTI-4 (P83605)	Rhipicephalus microplus (Tick)	Serine protease inh.	25	3.5 (2nd hit)
	ASTC			K^+ ch. toxin alpha-KTx 18.2 (P0C1X5)	Tityus discrepans (Scorpion)	K+ ch. inh.	34	11 (3rd hit)

Table 1. Continued

Band	Sequence (*de novo* score)	Precursor ion mass	MW From gel	Identity and accession (Swiss-Prot)	Organism	Function	Size (resid.)	Swiss-Prot E-value
	NLKCR (61.8)	**690.26**	**7 kDa**					
	NLKC			Venom protein PN10C5 (P84015)	Phoneutria nigriventer (Spider)	Unknown function	33	3.9 (1st hit)
	NL CR			Insecticidal toxin DTX11 (P55816)	Diguetia canities (Spider)	Na+ ch. inh.	58	23 (2nd hit)
	KAGGGGVR (74.4)	**701.28**	**7 kDa**					
	A GGVR			Spidroin-1 (Dragline silk fibroin 1) (P19837)	Nephila clavipes (Spider)	Dragline silk subunit	747	15 (1st hit)
	K GGGG			Acanthoscurrin-1 and -2 precursor (Q8I948; Q8I6R7)	Acanthoscurria gomesiana (Spider)	Anti-microbial	156	27 (2nd hit)
	KA GG			Insect toxin 5 (Lqh IT5) (P81240)	Leiurus quinquestriatus hebraeus (Scorpion)	Na+ ch. inh.	61	87 (=3rd hit)
	KA GG			Neurotoxin 5 (CsEv5) (P58779)	Centruroides sculpturatus (Scorpion)	Na+ ch. inh.	59	87 (=3rd hit)
12	KNDITGYFSSK (71.2)	**1259.44**	**6 kDa**					
	G+FSSK			Insecticidal toxin DTX9.2 precursor (P49126)	Diguetia canities (Spider)	Na+ ch. inh.	94	2.4 (1st hit)
	DI YF SK			Alpha-latroinsectotoxin precursor (Q02989)	Latrodectus mactans (Spider)	Neurotoxin	1141	14 (2nd hit)
	DI YF			Sphingomyelin phosphodiesterase D precursor (Q202J4)	Ixodes scapularis (Tick)	Dermonecrotic Toxin	364	18 (=3rd hit)
	KN TGY			Neurotoxin 6 (P56743)	Androctonus australis (Scorpion)	Insect toxin	66	18 (=3rd hit)

Table 1. Continued

Band	Sequence (*de novo* score)	Precursor ion mass	MW From gel	Identity and accession (Swiss-Prot)	Organism	Function	Size (resid.)	Swiss-Prot E-value
	DITG			Mite group 2 allergen Lep d 2 precursor (P80384)	Lepidoglyphus destructor (Mite)	Allergen	141	18 (=3rd hit)
	NSCTRLDCRFVVPWE (37.0)	**1938.72**	**6 kDa**					
	TRL+ FV+			40S ribosomal protein S4 (Q4PMB3)	Ixodes scapularis (Tick)	Ribosomal protein	262	5.4 (=1st hit)
	LDCR			Insecticidal toxin DTX9.2 precursor (P49126)	Diguetia canities (Spider)	Na+ ch. inh.	94	5.4 (=1st hit)
	SCTCR			Sphingomyelin phosphodiesterase D and D3 precursor (Q8I912)	Loxosceles laeta and Loxosceles intermedia (Spider)	Dermonecrotic Toxin	304	9.8 (=2nd hit)
	SCTCR			Neurotoxin BmKX-A1-S31 and X-29S precursor (Q7Z0H5)	Mesobuthus martensii (Scorpion)	Unknown function	55	9.8 (=2nd hit)
	PWE			Sphingomyelin phosphodiesterase D,D1,D2,D5 precursors (Q1KY79)	Loxosceles laeta, L. Arizonica and L. Intermedia (Spider)	Dermonecrotic Toxin	285-311	18 (=3rd hit)
	SCT DC			Non-toxic venom protein PNTx16C1 (P8399)	Phoneutria nigriventer (Spider)	Unknown function	68	18 (=3rd hit)
	PWE			Toxin SFI 1, 2, 3, 5, 6, 7, 8 (P61095)	Segestria florentina (Spider)	Insect toxin	46	18 (=3rd hit)
	CVDLTGYFSSK (92.6)	**1276.45**	**6 kDa**					
	G+FSSK			Insecticidal toxin DTX9.2 precursor (P49126)	Diguetia canities (Spider)	Na+ ch. inh.	94	2.4 (1st hit)
	+LTGY			Tropomyosin	5 sp. tick/mite inc. Dermatophagoides farinae	Muscle contraction	284	7.6 (=2nd hit)

Table 1. Continued

Band	Sequence (*de novo* score)	Precursor ion mass	MW From gel	Identity and accession (Swiss-Prot)	Organism	Function	Size (resid.)	Swiss-Prot E-value
	CVDL			Alpha-latrotoxin-associated LMWP precursor (P49125)	Latrodectus mactans (Spider)	Unknown function	88	7.6 (=2nd hit)
	DL GY			Peptide BmKb1 precursor (Q718F4)	Mesobuthus martensii (Scorpion)	Antimicrobial	74	18 (=3rd hit)
	VDLT			Tubulin alpha chain (Q8WQ47)	Lepidoglyphus destructor (Mite)	Microtubule	450	18 (=3rd hit)
14	ECLMGRGK (76.0)	**1113.4**	**11 kDa**					
	ECLM			Toxin DW13.3 (P60979)	Kukulcania hibernalis (Spider)	Ca2+ ch. inh.	74	1.9 (1st hit)
	+CLM			Putative beta-neurotoxin Td3 and Td11 precursor (Q1I177)	Tityus discrepans (Scorpion)	Na+ ch. inh.	73	11 (=2nd hit)
	+GRGK			Hemocyanin C chain (Q9NFL6)	Aphonopelma sp (Spider)	Oxygen transport	629	20 (3rd hit)
	YLLKTSTCAK (89.7)	**1184.49**	**13 kDa**					
	LK STC			Kunitz-type serine protease inh. BmTI-4 (P83605)	Rhipicephalus microplus (Tick)	Protease inh.	25	10 (=1st hit)
	YLL TS			NADH-ubiquinone oxidoreductase chain 5 (Q9ZYM7)	Rhipicephalus sanguineus (Tick)	Mitochondrial enzyme	552	10 (=1st hit)
	TCAK			Huwentoxin-11 (P68425)	Ornithoctonus huwena (Spider)	Serine protease inh.	55	14 (=2nd hit)
	TCAK			Neurotoxin II.22.5 (P18927)	Centruroides limpidus tecomanus (Scorpion)	Na+ ch. inh.	30	14 (=2nd hit)
	LL +TC			Neurotoxin Cex1 and Cex12 precursors (Q68PH4)	Centruroides exilicauda (Scorpion)	Na+ ch. inh.	87	34 (=3rd hit)

Table 1. Continued

Band	Sequence (*de novo* score)	Precursor ion mass	MW From gel	Identity and accession (Swiss-Prot)	Organism	Function	Size (resid.)	Swiss-Prot E-value
	LL +TC			Neurotoxin 9 precursor (Q95WC9)	Centruroides sculpturatus (Scorpion)	Na+ ch. inh.	84	34 (=3rd hit)
	VNSFASKPGR (64.6)	**1062.45**	**13 kDa**					
	KPGR			Alpha-neurotoxin TX11 precursor (Q9NJC7)	Mesobuthus martensii (Scorpion)	Na+ ch. inh.	85	19 (1st hit)
	VN FA			Kunitz-type serine protease inh. BmTI-A (P83609_1)	Rhipicephalus microplus (Tick)	Protease inh.	68	25 (=2nd hit)
	FASK			Alpha-latroin-sectotoxin precursor (Q02989)	Latrodectus mactans (Spider)	Neurotoxin	1411	25 (=2nd hit)
	ASKP			Raventoxin-1 (P61233)	Macrothele raveni (Spider)	Neurotoxin	43	34 (3rd hit)
	SARTDGYECINDDGQ (67.2)	**1700.65**	**13 kDa**					
	YEC+ DD			Neurotoxin V (P46066)	Centruroides sculpturatus (Scorpion)	Na+ ch. inh.	63	0.39 (1st hit)
	DGY CINDD			Insect toxin 2 (P59863)	Buthus occitanus tunetanus (Scorpion)	Na+ ch. inh.	60	4.1 (2nd hit)
	RT+GY			Omega-agatoxin-3D (P81746)	Agelenopsis aperta (Spider)	Ca2+ ch. inh.	37	5.4 (3rd hit)
	KGTFNICCK (68.0)	**1127.41**	**13 kDa**					
	ICCK			K^+ ch. toxin gamma-KTx 3.2 (Q86QV5)	Centruroides elegans (Scorpion)	K+ ch. inh.	43	2.9 (1st hit)
	KGTF			Putative beta-neurotoxin Td3 precursor (Q1I177)	Tityus discrepans (Scorpion)	Na+ ch. inh.	73	17 (=2nd hit)

Table 1. Continued

Band	Sequence (*de novo* score)	Precursor ion mass	MW From gel	Identity and accession (Swiss-Prot)	Organism	Function	Size (resid.)	Swiss-Prot E-value
	+CCK			K^+ ch. toxin gamma-KTx 3.1, 3.3 and 5.1 (P59939)	Centruroides noxius and C. Sculpturatus (Scorpion)	K+ ch. inh.	43-47	17 (=2nd hit)
	F+IC			ATP-dependent DNA helicase 2 subunit 1 (Q26228)	Rhipicephalus appendiculatus (Tick)	DNA binding protein	600	17 (=2nd hit)
	+CCK			Oxytoxin-1 (P83288)	Oxyopes kitabensis (Spider)	Na+ ch. inh.	69	17 (=2nd hit)
	N+CCK			Probable neurotoxin pcD-1008 precursor (Q9BLM2)	Androctonus australis (Scorpion)	Neurotoxin/ Ionic ch. inh.	72	30 (3rd hit)
	YGETCKAGK (73.5)	**1013.37**	**13 kDa**					
	GETCK			Neurotoxin Pn3-6A precursor (P0C2S7)	Phoneutria nigriventer (Spider)	Ca2+ ch. inh.	93	1.2 (=1st hit)
	GETCK			Probable neurotoxin PKTx23C3 (P83902)	Phoneutria keyserlingi (Spider)	Ca2+ ch. inh.	58	1.2 (=1st hit)
	GETCK			Venom protein PNTx22C3 (P84011)	Phoneutria nigriventer (Spider)	Unknown function	58	1.2 (=1st hit)
	ETCK G			Plectoxin-9 (P36989)	Plectreurys tristis (Spider)	Insect toxin	46	3.9 (2nd hit)
	ETCK			K^+ ch. toxin kappa-KTx 1.1, 1.2 and 1.3 (P83655)	Heterometrus spinifer and H. fulvipes (Scorpion)	K+ ch. inh.	22-23	7.0 (=3rd hit)
	KVFGRYECLNYDGGA (69.4)	**1748.62**	**13 kDa**					
	LNYDG			Alpha-latrotoxin precursor (P23631)	Lactrodectus mactans (Spider)	Presynaptic neurotoxin	1401	0.93 (1st hit)

Table 1. Continued

Band	Sequence (*de novo* score)	Precursor ion mass	MW From gel	Identity and accession (Swiss-Prot)	Organism	Function	Size (resid.)	Swiss-Prot E-value
	YECL+Y			Neurotoxin V (P46066)	Centruroides sculptu-ratus (Scor-pion)	Na+ ch. inh.	63	2.2 (=2nd hit)
	KVF L+YDG			Allergen Mag (P39673)	Dermato-phagoides farinae (Mite)	Allergen	341	2.2 (=2nd hit)
	CLNY			40S ribosomal protein S23 (Q86FP7)	Dermacentor variabilis (Tick)	Ribosomal protein	143	3.0 (=3rd hit)
	YECL			Toxin CsE M1, Toxin Css II and Toxin 2 precursor (P56646)	Centruroides sculpturatus, Centruroides noxius, Cen-truroides suf-fusus suffusus (Scorpion)	Na+ ch. inh.	65-84	3.0 (=3rd hit)

a) For the complete table see the Supporting Information.
b) ch. = channel
c) inh. = inhibitor

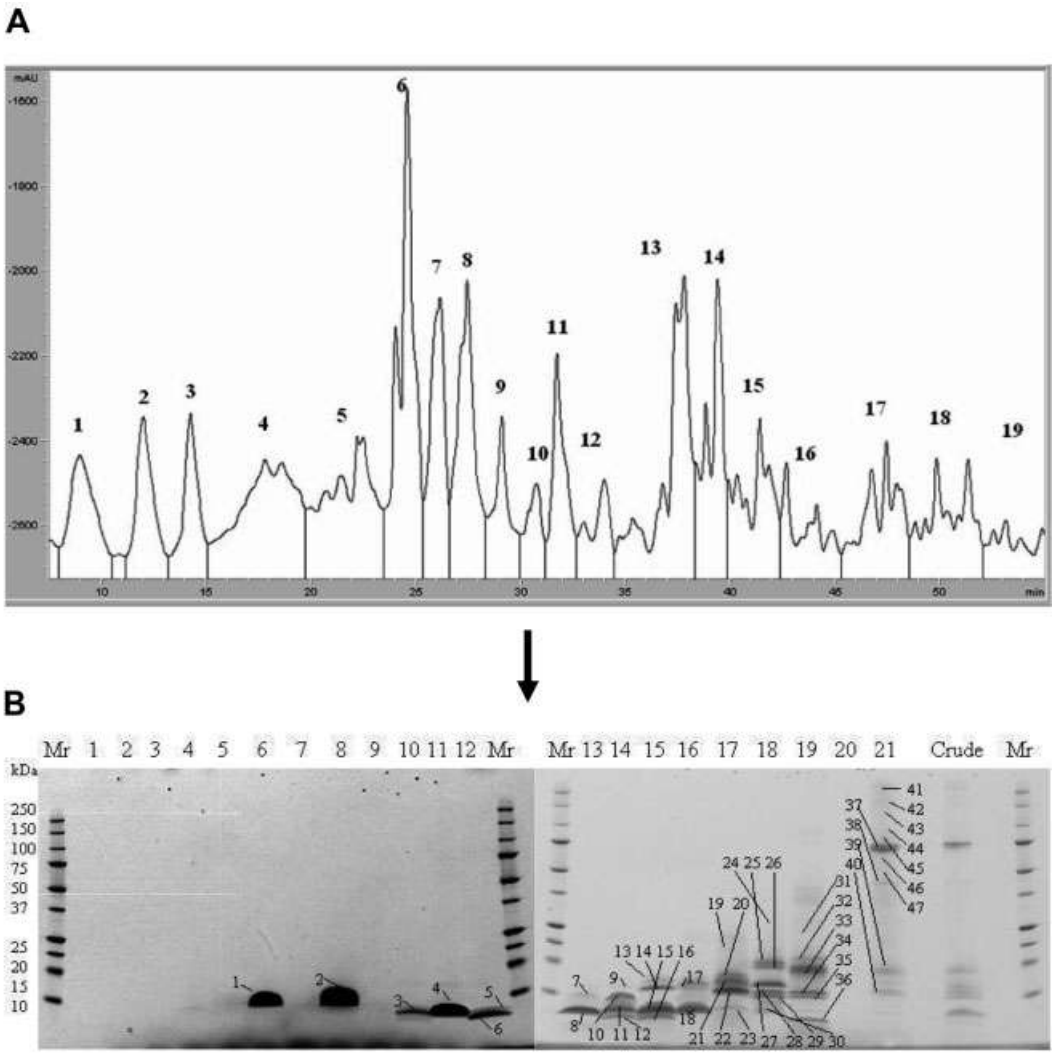

Figure 1. Two-dimensional separation of alkylated *H. longimanus (Asian black scorpion)* venom. The first dimension (A) is an RP-HPLC with a gradient of 2–40% ACN, 0.06% TFA, in 55 min. The gradient was then raised from 40–100% in 10 min (not shown). Fractions were manually collected as numbered and subsequently run on 12-lane 10–20% Tris-Tricine Peptide Criterion™ Precast gels (B). Lanes 20 and 21 contain the fractions collected from 40–70% and 70–100% ACN, respectively (not displayed in the HPLC chromatogram). Alkylated crude venom was loaded in lane 23. Markers were Precision Plus markers (Bio-Rad). The stain displayed is "Blue Silver" Coomassie. Bands were excised as numbered and analysed by MALDI TOF/TOF MS.

threshold, which allows reporting of lower scoring matches, but also decreases the confidence in the protein identification.

The Swiss-Prot database was searched with a taxonomy filter of Arachnids. The taxonomy was deliberately narrowed to Arachnids to reduce random matches of no intrinsic value, which were evident in the larger unfiltered Swiss-Prot database. All the hits with the three highest E-values for the Swiss-Prot matches are shown in Table 1 for a subset of 19 different peptides. This subset was chosen to demonstrate the breadth of results, ranging from high homology, to those that show a greater degree of variability (for the complete Table of 78 peptides, refer to the Supporting Information).

The results from Table 1 (and the Supporting Information Table) are promising in that many of the hits comprise proteins which would be expected to be found in scorpion venoms, such as ionic channel inhibitors and antimicrobial proteins [5, 6]. The presence of some non-toxin, random matches, such as the spider dragline silk subunit, Spindroin-1 (peptide 11) is likely due to the lower stringency search parameters required for searching short peptide sequences. These parameters were chosen in an attempt to give the optimal balance between finding the maximum number of true venom protein matches, while excluding random matches. The total number of annotations for known venoms, toxins or secreted proteins in the Swiss-Prot database was

calculated as 1255 of 7796 entries (16%). When the identity of the top hit (by E-value) for our *de novo*-generated peptides was determined, 66 of 78 (84%) matched to known venom proteins. Therefore, this enrichment of hits to venom proteins indicates that the homology matching of the short peptide sequences is likely producing true venom protein matches, above that expected by chance.

In order to further characterise the experimentally derived proteome we sought to obtain some indication of the function of the *de novo*-generated peptides. Thus, we designed a scoring system to allow us to putatively assign functions to each gel band, based on homology to known venom proteins contained within the Swiss-Prot database (refer to Section 2). Figure 2 represents graphically the

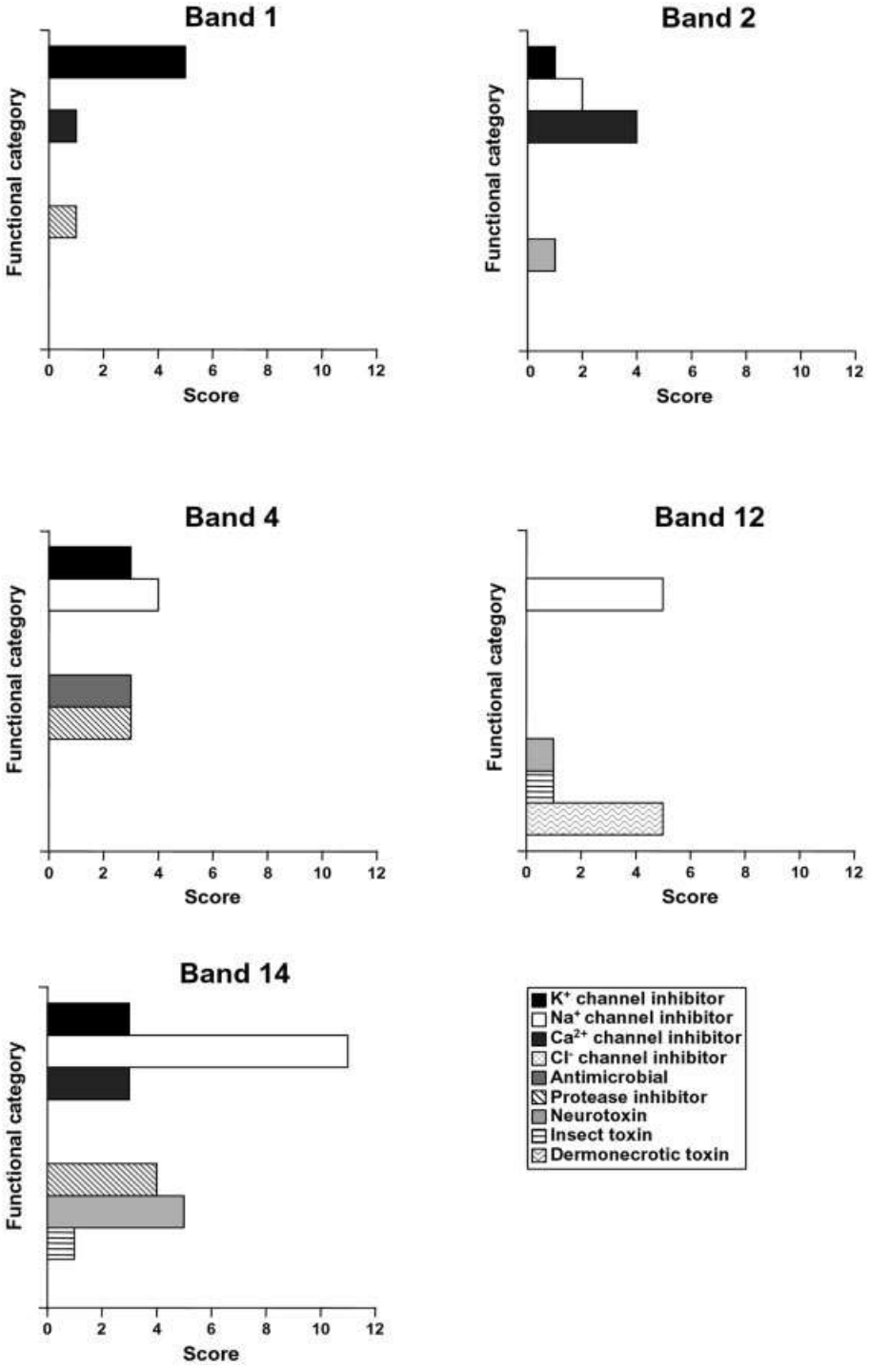

Figure 2. Assignment of putative functions to gel bands using pBLAST and our scoring system. The peptides generated by *de novo* sequencing of gel bands were BLASTed against the Swiss-Prot database, and hits with the three highest E-values annotated. The hits were divided into nine functional categories, and assigned a point for every function within the top hits. A bonus point was scored if the same family of protein was identified in more than one peptide derived from the same gel band.

assignment of a functional category to the five gel bands displayed in Table 1. This shows that for some bands, such as band 1, the highest scoring and therefore most likely putative function can be clearly determined. However, in cases such as band 4, several functional categories score very closely (some bands score identically), and therefore the putative function is more difficult to assign confidently. In these instances, it may be useful to consider the E-value and quality of the individual matches in an attempt to putatively assign a function. Although each gel band was subjected to two steps of purification, the possibility remains that a band may be composed of two or more proteins with similar physicochemical properties. This may help to explain a result such as that observed for band 12, where two functionally different categories show equally high scores.

Using this scoring system, we were able to assign putative functions to 70% of the gel bands. The composition of the experimentally derived proteome is shown in Fig. 3. The most predominant functional group of proteins identified was ionic channel inhibitors, which represent more than half the proteins. Of these, Na^+ channels were the most prevalent. This is consistent with previous reports, where Na^+ channels comprise a major component of scorpion venom [7, 8]. Other functional categories comprised smaller proportions of the proteome, including antimicrobial proteins and toxins with specific activity. This is also consistent with previous proteomic analyses of the composition of venom [6, 9–11]. These findings indicate that the assignment of putative function based on the homology to sequenced proteomes can be effective.

Future directions for this approach include investigating the use of alternative *de novo* algorithms to DeNovo Explorer™. The development of a BLAST scoring algorithm to allow identification of high homology matches without relying solely on E-value would be useful for better quantitating the homology matches for short peptides. Furthermore, the development of a protein identification algorithm that could score multiple peptide matches to a protein, in the same way that MASCOT or SEQUEST perform protein identification using sequence information, would be of great benefit in determining homology and strengthening the validity of the match.

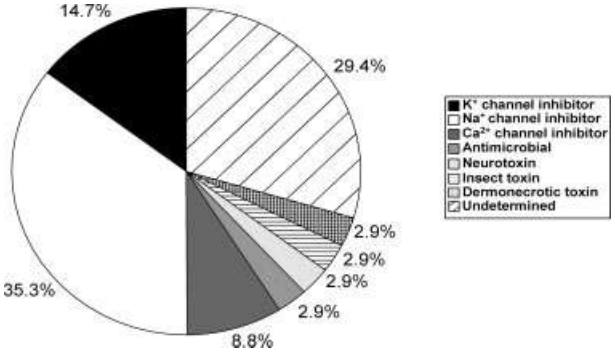

Figure 3. Assignment of functions to *Heterometrus longimanus* venom proteome. For each of the 34 gel bands purified from *H. longimanus* venom, a score was calculated for the seven functional categories as shown in Fig. 2. Where possible, the putative function was determined. The percentage each functional category contributes to the venom proteome is indicated.

4 Concluding remarks

Venoms are some of the more interesting collections of peptides and proteins that nature produces. Evolutionary processes have honed their effectiveness and potency and provided homological similarities between peptide sequences. Although much work has been carried out to analyse their contents, there still remains many organisms whose proteomes are yet to be examined. Here, we developed a novel method that more effectively maps the proteome of venoms while providing an innovative way to derive the contents of the venom and catalogue the putative activity profile of each component. The *de novo*-sequenced peptides of the venom were homology matched against a public database and the resulting matches examined for their putative function. The proteome for *H. longimanus* venom derived in this way correlates well with the venom proteomes published for other species of scorpions, thereby supporting the validity of this approach.

These analyses were performed in facilities provided by the Lotterywest State Biomedical Facility-Proteomics node, Western Australian Institute for Medical Research, and the WA State Agricultural Biotechnology Centre, Murdoch University.

The authors have declared no conflict of interest.

5 References

[1] Theakston, R. D., Kamiguti, A. S., A list of animal toxins and some other natural products with biological activity. *Toxicon* 2002, *40*, 579–651.

[2] Lewis, R. J., Garcia, M. L., Therapeutic potential of venom peptides. *Nat. Rev.* 2003, *2*, 790–802.

[3] Gopalakrishnakone, P., Cheah, J., Gwee, M. C., Black scorpion (*Heterometrus longimanus*) as a laboratory animal: maintenance of a colony of scorpion for milking of venom for research, using a restraining device. *Laboratory animals* 1995, *29*, 456–458.

[4] Casey, T. M., Arthur, P. G., Bogoyevitch, M. A., Proteomic analysis reveals different protein changes during endothelin-1- or leukemic inhibitory factor-induced hypertrophy of cardiomyocytes *in vitro*. *Mol. Cell. Proteomics* 2005, *4*, 651–661.

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

[6] Zhijian, C., Feng, L., Yingliang, W., Xin, M., Wenxin, L., Genetic mechanisms of scorpion venom peptide diversification. *Toxicon* 2006, *47*, 348–355.

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

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

[9] Batista, C. V., D'Suze, G., Gomez-Lagunas, F., Zamudio, F. Z. *et al.*, Proteomic analysis of *Tityus discrepans* scorpion venom and amino acid sequence of novel toxins. *Proteomics* 2006, *6*, 3718–3727.

[10] Gurevitz, M., Karbat, I., Cohen, L., Ilan, N. *et al.*, The insecticidal potential of scorpion beta-toxins. *Toxicon* 2007, *49*, 473–489.

[11] Zeng, X. C., Corzo, G., Hahin, R., Scorpion venom peptides without disulfide bridges. *IUBMB life* 2005, *57*, 13–21.

