Mapping Proteoforms and Protein Complexes From King Cobra Venom Using Both Denaturing and Native Top-down Proteomics* [S]

Rafael D. Melani‡§, Owen S. Skinner§, Luca Fornelli§, Gilberto B. Domont‡¶, Philip D. Compton§, and Neil L. Kelleher§¶

Characterizing whole proteins by top-down proteomics avoids a step of inference encountered in the dominant bottom-up methodology when peptides are assembled computationally into proteins for identification. The direct interrogation of whole proteins and protein complexes from the venom of Ophiophagus hannah (king cobra) provides a sharply clarified view of toxin sequence variation, transit peptide cleavage sites and post-translational modifications (PTMs) likely critical for venom lethality. A tube-gel format for electrophoresis (called GELFrEE) and solution isoelectric focusing were used for protein fractionation prior to LC-MS/MS analysis resulting in 131 protein identifications (18 more than bottom-up) and a total of 184 proteoforms characterized from 14 protein toxin families. Operating both GELFrEE and mass spectrometry to preserve non-covalent interactions generated detailed information about two of the largest venom glycoprotein complexes: the homodimeric L-amino acid oxidase (~130 kDa) and the multichain toxin cobra venom factor (~147 kDa). The L-amino acid oxidase complex exhibited two clusters of multiproteoform complexes corresponding to the presence of 5 or 6 N-glycans moieties, each consistent with a distribution of N-acetyl hexosamines. Employing top-down proteomics in both native and denaturing modes provides unprecedented characterization of venom proteoforms and their complexes. A precise molecular inventory of venom proteins will propel the study of snake toxin variation and the targeted development of new antivenoms or other biotherapeutics.

The toxic secretions of venomous animals have long been a source of biologically active molecules, yet remain a promising source of new drugs (1, 2). Snake venoms in particular are widely studied and contain a mixture of small molecules, peptides, proteins, and protein complexes which are used to incapacitate prey and to defend against predators (3–5). Snakes produce anywhere from three to several dozen diverse families of toxic proteins. Each family is encoded by multiple multilocus genes generated by gene duplication resulting in a large number of expressed isoforms (6) that can differ greatly even between individuals of the same species (4, 7, 8).

Most current techniques for venom protein analysis and toxin identification are based on bottom-up (BU)1 proteomics approaches where toxins are enzymatically digested and the peptides are analyzed by tandem mass spectrometry (MS/MS) (8–11). Although this technique can generate a comprehensive overview of toxin families present in the venom, such peptide-based toxin studies do not provide isoform-specific identifications and/or distributions and stoichiometries of post-translational modifications (PTMs). As a result, bottom-up experiments provide imprecise information about what specific molecular species/toxins are present and how they interact.

In principle, top-down proteomics (TDPs), a method that measures intact proteins and their fragment ion masses, can

1 The abbreviations used are: BU, bottom-up; 3FTx, three finger toxins; AGC, automatic gain control; BCA, bicinchoninic acid assay; CRISP, cysteine rich secretory protein; CVF, cobra venom factor; DDA, data dependent acquisition; dTD, denaturing top-down; DTT, dithiothreitol; FDR, false discovery rate; GELFrEE, gel-eluted liquid fraction entrapment electrophoresis; HCD, higher-energy collisional dissociation; HexNAc, N-acetyl hexosamine; IGF, insulin-like growth factor; KNT, Kunitz-type serine protease inhibitor; LAOO, L-amino acid oxidase; LC-MS/MS, liquid chromatography online with tandem mass spectrometry; MS/MS, tandem mass spectrometry; NSAF, normalized spectral abundance factor; nTD, native top-down; ODN, ohnin; PLAA2, phospholipase A2; PLRP-S, polymeric reverse phase resin; PTM, post-translational modification; siEF, solution isolectric focusing; SVM, snake venom metalloproteinases; SVNGF, snake venom nerve growth factor; SVSP, snake venom serine protease; SVVEGF, snake venom vascular endothelial growth factor; TDP, top-down proteomics; WAP, waprin.
access new information on protein toxins. TDPs has matured over the past decade and now can be applied to intricate biological samples in a high-throughput manner for proteins up to ~30 kDa (12). One major advantage of TDPs lies in the ability to provide complete molecular specificity of an intact protein to determine its “proteoforms”, the set of molecules that arise from combinations of genetic variation (e.g., sequence variants, isoforms) and PTMs (13). Increased levels of characterization are possible thanks to recent advances in instrumentration (14, 15), intact protein fractionation methods (16–18), and data processing (19, 20). Recently, Calvete and co-workers published a top-down MS analysis of snake venom. Their pipeline coupled LC-MS/MS analysis, intact mass values of reduced and non-reduced proteins, and BU proteomics to assign manually toxin identifications with high attention to locus-specificity. The analysis identified 11 three finger toxins (3FTx), 1 Kunitz-type serine protease inhibitor (KNT), 1 cholin, 1 phospholipase A2 (PLA2), and 1 cysteine rich secretory protein (CRISP) in the venom of the Indonesian king cobra (21).

Although denaturing TDPs provides a powerful technique for characterizing individual proteins <30 kDa, native MS is a growing field where highly purified proteoform complexes (22) larger than 100 kDa in size can be interrogated by top-down MS (12, 22–24). Native MS analysis of purified protein complexes has been reported since the early 1990’s (25, 26). This was followed by gas-phase ejection of subunits (27), with the fragmentation of these by MS/MS and their identification being a more recent advance (23). The development of a separation technique termed native gel-eluted liquid fraction entrapment electrophoresis (native GELFrEE), allowed its linked use with native MS (23) to fully characterize intact complexes from endogenous systems (28). With many proteins performing their functions as members of protein assemblies, their direct observation could provide a highly informative view of the molecular composition and protein-protein interactions in snake venom. Native MS is a relatively new technique to study non-covalent protein-protein interactions in comparison to x-ray diffraction and solution nuclear magnetic resonance (NMR) that are commonly used to characterize the tertiary and quaternary structures of proteins (29). Based on these and other biochemical methods, several toxin complexes were identified with variable stoichiometry (homo- and heteromeric complexes) and from almost all snake venom toxin families (30). Characterization studies have focused on: haditoxin, a 3FTx homodimer (31), crotoxin, a PLA2 heterodimer (32), the homodimer of L-amino acid oxidase (LAAO) from Calloselasma rhodostoma (33), and pseutarin C, a serine protease heterodimer (34).

Here, we leverage all recent advances of denaturing and native TDPs to characterize the proteins in Ophiophagus han-nah (king cobra) venom. The results provide the most comprehensive, proteoform-specific view of a venom proteome to date, resolving combinatorial variations arising from proteolytic protein processing, sequence polymorphism, complexation and post-translational modification.

**EXPERIMENTAL PROCEDURES**

**Venom Samples—Ophiophagus hannah** (king cobra) venom milked from two individuals original from Malaysia was purchased from the Kentucky Reptile Zoo, and was pooled by zoo staff. Venom protein concentration was determined using the BCA protein assay kit and BSA as standard following the manufacturer’s instructions (Thermo Pierce, Rockford, IL).

**Denaturing Top-down Proteomics—GELFrEE separation was performed on a GELFREEM$10$ fractionation system (Expedeon, Harstton, Cambridgeshire, UK) using 300 μg of venom sample on a 10% cartridge following the manufacturer’s instructions. Twelve 150 μl fractions were collected manually and precipitated with MeOH/CH$_3$Cl/H$_2$O to remove SDS and salts as described previously (16). Pelleted fractions were resuspended in 30 μl of buffer A (95% H$_2$O, 5% ACN, 0.2% formic acid).

In-solution isoelectric focusing (sIEF) was performed on a 1 mg sample of venom dissolved in 3 ml of 7 M urea, 2 M thiourea, 30 mM DTT, and 60 μl (2%) of Biolyte 3/10 ampholytes, pH range 3.0 to 10.0 (Bio-Rad, Hercules, CA). The separation was performed on a Micro-Rotofor (BioRad) at 1 W constant power for 3 h at 10 °C. Ten fractions were collected and cleaned using homemade C-18 stage tips (35). Fractions were dried and resuspended in 50 μl of buffer A. Nano RPLC-MS/MS analysis was performed using an Ultimate 3000 nano-RPLC (Thermo Scientific Dionex, Sunnyvale, CA) system on an in-house packed polymeric reverse phase resin (PLRP-S, 5 μm, 1000 Å, Phenomenex, Santa Clara, CA) for the trap-column (2 cm x 150 μm i.d.) and analytical column (20 cm x 75 μm i.d.). Fractionated samples were separated with a 90 min. gradient whereas the whole venom was analyzed using a 120 min. gradient. RPLC was used in tandem with an Orbitrap Elite and a Q-Exactive HF (Thermo Scientific, San Jose, CA) for online MS/MS analysis. MS$^1$ (intact protein) spectra were acquired at 120,000 resolving power (at 400 m/z for the Orbitrap Elite and at 200 m/z for the Q-Exactive HF) using four μscans. Data-dependent top two MS$^2$ (protein fragmentation) spectra were acquired using higher-energy collisional dissociation (HCD) at 60,000 resolving power with an isolation window of 15 m/z (Orbitrap Elite) and 4 m/z (Q-Exactive HF).

**Data Analysis—**Intact protein mass and fragment ions were searched against protein sequences and candidate proteoforms from the suborder Serpentes, including sequences, features and modifications for the king cobra (36). This was accomplished by downloading a UniprotKB flatfile on June 10, 2014 and the search database was created using the Database Manager application within ProSightPC v3.0 (Thermo Fisher Scientific). The resulting database was established as a ProSight Warehouse File (.pwf) and contained 331,349 candidate proteoforms. Searches were performed using an iterative, three-tiered, absolute mass search tree (2.5 Da, 200 Da, and 2000 Da intact mass tolerance), and then re-searched using a biomarker search with a 15 ppm tolerance for intact protein mass and a 10 ppm mass tolerance for fragment ions. In order to estimate the false discovery rate (FDR) for each data set, spectra were searched against a scrambled database and their p-scores derived from a Poisson statistical model (37) were fit to a gamma distribution as previously described (16, 38). The E-values corresponding to a 5% FDR cutoff for absolute mass searches are as follows: GELFrEE, 1.1 × 10$^{-6}$; whole venom, 1.7 × 10$^{-6}$; sIEF 2.7 × 10$^{-6}$. Because of the differences in the search space covered by biomarker (“no enzyme”) searches, the E-values corresponding to a 5% instantaneous FDR (a.k.a. q-values) for those were generated separately and were: 1.2 × 10$^{-10}$ for denaturing GELFrEE, 2.6 × 10$^{-11}$ for LC-MS of whole venom and 1.7 × 10$^{-8}$ for sIEF.
Relative protein abundance was quantified by summing ion abundance measurements using in-house protein software built and reported for label-free quantitation (39). All raw data files associated with this study are available at http://massive.ucsd.edu/ with identifier MSV000079351 and all compatible identified proteoforms are assigned unique PFR identifiers, deposited in the Consortium for Top Down Proteomics Proteoform Repository (http://repository.topdownproteomics.org/).

Native Separations—O. hannah venom was fractionated by Native-GELFrEE as described previously (28). Briefly, 1 mg of venom was suspended in 200 μl of low salt solubilization buffer (50 mM imidazole, 500 mM aminocaproic acid, 1 mM EDTA, pH 7.0) and mixed with 30 μl of 70% glycerol and 0.1% Ponceau S. Fractionation was performed via imidazole-based clear native GELFrEE carried out using 1–12% T gradient tube gel. A voltage range of 200–1000 V was applied for protein separation, and 150 μl fractions were collected manually. Fractions were cleaned with HiPPR detergent-removal columns (Thermo Pierce), following manufacturer instructions, and buffer exchanged into 150 mM ammonium acetate using 30 kDa Amicon Ultra molecular weight cutoff filters (Millipore, Billerica, MA).

Native Mass Spectrometry—Native electrospray ionization was performed using a custom source developed by the Dovichi lab and previously reported (28, 40) with a spray voltage of 1.0–1.2 kV. Mass spectrometry was conducted on a modified Q-Exactive HF (Thermo Fisher Scientific). Briefly, the standard Q-Exactive quadrupole rf drive was modified in order to decrease the resonant rf frequency to 284 kHz and enable higher m/z selection up to m/z 20,000, as described previously (23). Scans were collected manually using 15,000 resolving power at m/z 200 for MS1 spectra and 120,000 resolving power for MS2 and pseudo-MS3 spectra as described previously (23). Average masses and charge states for spectra without isotopic resolution were calculated by optimizing the standard deviation of the most intense peaks using the web tool ESIprot Online (41). Proteins were identified from fragmentation spectra using ProSightPC v3.0 (Thermo Fisher Scientific) and were searched against the Serpente database with 331,349 candidate proteoforms in the same ProSight Warehouse molecular data, downloaded from UniprotKB on June 10, 2014, totaling 49,360 entries. Carboxamidomethylation of cysteines was set as a fixed modification and oxidation of methionine was set as a variable modification. Only full tryptic peptides were considered with a maximum of two missed cleavages. Mass tolerances for MS1 and MS2 were set at 70 and 550 ppm, respectively. XCorr was determined as a primary score, and ZScore was used as a secondary one. Resultant peptides were processed and evaluated by Search Engine Processor (SEPro) v3.2.0.2 using the following filter parameters: 10 ppm deviation from theoretical peptide precursor, peptides longer than six amino acid residues, and a 1% estimated protein-level FDR (46). Label free protein quantitation was performed according to the normalized spectral abundance factor (NSAF) (47). Bottom-up data are available via ProteomeXchange with identifier PXD003403. Annotated spectra and filtered data can be opened and visualized using the package “PatternLab for proteomics” available for free at http://www.patternlabforproteomics.org/.

Experimental Design and Statistical Rationale—Denaturing top-down analyses of the whole venom without prefractionation were performed in five runs of LC-MS/MS (MS technical replicates) to reduce under-sampling associated with the DDA-based method. GELFrEE and sIEF fractionations were performed twice (process replicates) and each obtained fraction was analyzed by LC-MS/MS twice (MS technical replicates) to reduce variance from the separation methods. The venom was fractionated twice by native GELFrEE (process replicates) and the fractions from each were analyzed by direct injection nanospray once. The bottom-up analysis was performed using a single digestion that was analyzed by LC-MS/MS in triplicate (MS technical replicates) to reduce under-sampling associated with the DDA method.

RESULTS

Denaturing Top-down Venomics—Fig. 1 shows the strategy used for proteoform-resolved interrogation of the venom proteome obtained from two individuals of king cobra. The top-down methodology is divided into two major sections: denaturing top-down (dTD) and native top-down (nTD), with the latter employing native GELFrEE and native MS/MS (28). For dTD (Fig. 1A), we began by omitting prior fractionation and directly analyzing the whole venom by LC-MS/MS, yielding an intense set of peaks eluting in the first 30 min. of the total ion chromatogram (Fig. 2A). MS data extracted from these peaks indicated the presence of many proteoforms with masses of 5–13 kDa. A total of 17 proteins (counted as unique UniprotKB accession numbers) were identified at a 5% FDR (Table I and supplemental Table S1). The highly abundant peaks were identified as 3FTx, PLA2, and ohanin. After 30 min. of elution, only the larger (CRISP) and more hydrophobic (3FTx and PLA2) proteins were identified.

To improve experimental dynamic range and obtain better coverage of the venom proteome, we then employed and compared denaturing GELFrEE and solution isoelectric focusing (sIEF) prior to LC-MS/MS analysis (Fig. 1A). The resulting
Proteins identified by absolute mass searching were manually validated to precisely annotate the amino acid substitutions, alternative processing cleavage sites, and presence of extra PTMs, resulting in the identification of 53 additional proteoforms (each different from the 131 mapping to unique accession numbers discussed above). The 53 additional cases map to genes in the 3FTx, PLA2, KNT, and CRISP families; in total, 184 unique proteoforms were identified (Table I and supplemental Table S4). The mass shifts observed in the intact mass of some proteoforms were characterized with complete molecular specificity based on the information available from MS² fragment masses that matched to <10 ppm. Nearly 70% of the 53 distinct proteoforms arose from either amino acid changes in the primary sequence (compared with the canonical sequence present in UniprotKB) and/or PTMs not related to protein proteolytic processing. The next most common modification was proteolytic processing, which was observed on 28 proteoforms at sites other than the annotated cleavage sites reported in the literature or in UniprotKB. In addition, 12 proteoforms harbored variations in both primary sequence and proteolytic processing.

The toxin family with the largest number of identified proteoforms was the 3FTx (supplemental Fig. S1). Within this family, the long neurotoxin Oh-55 (Q53B58) exhibited the most proteoforms (5 in total; Table S4). The base proteoform of Oh-55 we refer to here as Pf 0 (derived from the canonical sequence deposited in UniprotKB with no additional modifications) was identified with a mass error of −0.53 ppm when compared with the theoretical mass and with extensive fragmentation reflected in the outstanding p-score of 5.2 × 10⁻¹¹⁴ (Fig. 3A). Three additional proteoforms of Oh-55 were identified and characterized with varying sequence lengths which are referred to here as Oh-55 Pf 1, 2, and 3. They are missing C-terminal residues from the mature toxin chain from unexpected proteolytic processing (supplemental Table S4). Furthermore, we identified Oh-55 Pf 4 with seven amino acid mutations from the Oh-55 Pf 0, the largest number characterized in this study (Fig. 3B). When mapped onto the crystal structure, none of these sequence differences were observed in Loop I (residues 1–14), and only one was found in Loop II (residues 19–43). However, five differences were observed in Loop III (residues 44–59) (Fig. 3C) (48, 49). Curiously, all 7 non-synonymous differences (only two are conservative) do not to map to known function-critical residues that directly bind to acetylcholine receptors and impair neuromuscular and neuronal transmission (48, 49).

Proteolytic processing/degradation and non-canonical cleavage sites or poorly annotated proteins were identified and annotated using a Biomarker-type search. In this search the intact proteoform mass is searched against all protein subsequences, similar to “no-enzyme” BU searches but also considering possible disulfide bond formation. Several of the larger proteins present in the venom such as SVMP, LAOO, and CVF were only identified in the denaturing TD workflow as

---

**Fig. 1.** **Strategy used for proteoform-resolved interrogation of king cobra venom.** The main steps used in each of the different top-down and bottom-up strategies are highlighted. Top-down analyses were divided into denaturing (A) and native (B) modes, with bottom-up analysis (C) used for comparative benchmarking. Comparison of protein identifications (D) and toxin families (E) identified by denaturing top-down and bottom-up proteomics.

Fractions from both separation techniques were visualized on SDS-PAGE slab gels (Fig. 2B and 2C, respectively) before being interrogated by LC-MS/MS. GELFrEE generated 12 fractions and separated the most abundant 5–13 kDa proteins into the first four fractions (Fig. 2B). LC-MS/MS analysis of each fraction using the same MS/MS data acquisition and search parameters used for whole venom analysis generated 64 protein identifications (Table I and supplemental Table S2), representing a ~2.8-fold increase when compared with direct venom analysis. On the other hand, siEF partitioned the most abundant proteins into 7 of the 10 collected fractions (Fig. 2C). When these ten fractions were submitted to the same LC-MS/MS analysis, they resulted in 113 protein identifications, nearly twice as many as found via denaturing GELFrEE (Table I and supplemental Table S3). Fig. 2D shows a direct comparison of the resulting identifications, organized by separation method. In total, we identified 131 proteins, with siEF/LC-MS/MS providing 63 unique protein identifications not found using the other two separation strategies. Denaturing GELFrEE/LC-MS/MS and direct venom analysis provided 18 identifications not found after siEF, with 16 of these classified as nontoxins. Of the 131 proteins identified during the automated search process, only 14% arose from absolute mass searches with a MS¹ mass window of less than 2.5 Da. The remaining 86% of identifications resulted from a searches using 2.5 to 2000 Da windows, indicating the large sequence variation present in the venom.
unexpected proteolytic products using the Biomarker search mode (supplemental Table S4). New proteolytic processing events not yet annotated in UniprotKB were observed in ohanin (P83234), and the resultant proteoforms are displayed in Fig. 4. The ohanin proteoforms Pf 1 and 2 differ when compared with Pf 0 by the addition of 1 (Arg129) and 2 (Arg130) amino acids at the C terminus, respectively. Further, we identified three proteoforms originating from the ohanin propeptide (supplemental Fig. S2). They constitute the predicted propeptide with six (Pf 3) and seven (Pf 4) residues cleaved from the N-terminal region, and a third proteoform equivalent to Pf 4 but with the N-terminal Glu137 converted to pyroglutamic acid (Pf 5). The most abundant propeptide proteoform was Pf 3, representing 88.3% (± 0.5%) of the total propeptide abundance followed by ohanin Pf 5, 11.0% (± 0.5%), and Pf 4, 0.8% (± 0.2%).

Native Top-down Venomics—To access the larger proteins present in king cobra venom, we applied a nTD methodology (Fig. 1B). Whole venom was fractionated by native GELFrEE, yielding 14 fractions visualized using a clear native PAGE slab gel (Fig. 5A). The most intense fractions in the gel, N10 and N11, were pooled and analyzed by native mass spectrometry. We observed two distributions of peaks with average masses centered at 130,014.2 ± 0.7 Da and 126,271.0 ± 2.3 Da (Fig. 5B and Table II). The peaks making up each distribution differed in mass by ~203 Da, consistent with the mass of multiple N-acetyl hexosamine (HexNAc) residues (Fig. 5B, inset). Quadrupole-isolation, with a 50 m/z isolation window, of each of the main glycoform distributions, followed by HCD fragmentation using an optimized energy to maximize dissociation, in separate MS/MS experiments produced highly similar fragmentation spectra (supplemental Fig. S3). Neutral
Table I

| Protein family | Whole venom | GELFrEE | sIEF | Non-redundant | Previous work (21) |
|----------------|-------------|---------|------|---------------|-------------------|
| 3FTx           | 12          | 43 (2)  | 50 (27) | 57 (29)       | 11                |
| CRISP          | 1           | 2 (3)   | 2    | 3 (3)         | 1                 |
| CVF            | 1           | 4       | 4    |               |                   |
| Disintegrin    | 2           | 2       |      |               |                   |
| IGF            | 2           | 2       | 2    |               |                   |
| KNT            | 2           | 2 (7)   | 2    | 2 (8)         | 1                 |
| LAAO           | 2           | 1       | 2    |               |                   |
| OHN            | 2           | 3       | 3    |               |                   |
| PLA2           | 1           | 2       | 2    | 2 (9)         | 1                 |
| SVMP           | 14          |         |      |               | 1                 |
| SVNGF          | 1           |         |      |               |                   |
| SVSP           | 1           |         |      |               |                   |
| WAP            | 1           | 1       |      |               |                   |
| Non-toxin      | 1           | 6       | 20   |               | 25                |
| Total          | 17          | 64 (6)  | 113 (47) | 131 (53)     | 15                |

![Proteoform fragmentation map](image)

**Fig. 3.** Proteoform fragmentation, primary sequence, and 3D structure indicating the locations of sequence variation. Proteoform fragmentation map of the base proteoform (Pf 0) of the long neurotoxin Oh-55 (Q53B58) (A) and of Pf 4 of Oh-55 (B); sequence positions showing amino acid differences from the Pf 0 of Oh-55 are highlighted in gray. J represents the isobaric amino acids Ile/Leu. Mapping of the polymorphic residues (indicated in red) determined by dTD analysis of Pf 4 of Oh-55 onto a 3D structural model, based on long neurotoxin 2 (P01386) (C).

masses extracted from each spectrum searched in absolute mass mode with ProSightPC identified both as L-amino acid oxidase (LAAO, P81383) with p-scores of $3 \times 10^{-53}$ and $2 \times 10^{-51}$, for the heavier and lighter precursors, respectively. Mapping the fragment ions to the primary sequence with a 10 ppm tolerance showed extensive coverage of the C-terminal region (Fig. 5C).

The theoretical mass of mature LAAO without any PTMs is 53,690.9 Da, less than half of the measured mass from the venom sample. Therefore, the observation of the intact mass is consistent with that of a homodimer, which has been previously described in the venom of *Calloselasma rhodostoma* (33). However, the mass difference between the theoretical homodimer (107,382 Da) and the observed average masses of the multiproteoform complexes is 22,632.4 and 18,889.2 Da, equivalent to ~17 and ~15% of the total mass, respectively. These mass differences are consistent with the presence of 6 and 5 glycan structures, respectively, with an average mass of ~3,743 Da for each glycan moiety (Fig. 5D). The glycosylation of LAAO can be associated with the four annotated N-linked glycosylation sites on each monomer (Asn122, Asn238, Asn266, and Asn410) (50). However, our data are consistent with the occupancy of five and six of the eight possible glycosylation sites in the dimer with relative abundances of ~15 and ~85%, respectively. This example illustrates the determination of the identity of a complex, as well as its monomer and PTM stoichiometry in a single mass spectrometry experiment.
The analyses of the subsequent native GELFrEE fractions allowed us to identify the CVF (I2C090), known to be the largest protein present in the venom of *O. hannah*. The observed average mass was 146,565 Da (Table II and supplemental Fig. S4), 5.6 kDa larger than the theoretical predicted mass. The observed mass difference is again consistent with the presence of glycans at either one or both of the two predicted glycosites (Asp181 and Asp209) (51). In the lower-mass native GELFrEE fractions, we identified two dimeric protein complexes: CRISP (Q7ZT98) and PLA2 (Q9DF33) (Table II). Although PLA2 has been previously observed as a dimer, CRISP has not. Although PLA2 and CRISP could be identified using an automated search process, we observed a protein of mass 49,283 Da (Table II) that could not be identified by database searching. Manually processing the fragmentation spectrum allowed for de novo annotation and characterization of the protein by using BLASTp (http://blast.ncbi.nlm.nih.gov) from a 13-residue sequence tag (supple-
Mapping Proteoforms and Protein Complexes From King Cobra Venom

**TABLE II**

Proteins and protein complexes identified using top-down proteomics in native mode (nTD). Protein accession numbers in UniprotKB (Accession #), protein-protein interaction (Complex), average theoretical mass, average observed mass of monomers and complexes masses in parentheses, and post-translational modifications (PTM) are included.

| Protein                                  | Accession # | Complex          | Theoretical mass | Observed mass  | PTM                          |
|------------------------------------------|-------------|------------------|------------------|---------------|------------------------------|
| Cobra venom factor                       | I2C090      | –                | 140,957 Da       | 146,565       | Glycosylation 5,608 Da       |
| L-amino acid oxidase form 1              | P81383      | Homodimer        | 53,691 Da (107,382 Da) | 130,014.2     | Glycosylation 22,637.2 Da    |
| L-amino acid oxidase form 2              | P81383      | Homodimer        | 53,690 Da (107,382 Da) | 126,271.0     | Glycosylation 18,889.2 Da    |
| Cysteine rich secretory protein          | Q72798      | Homodimer        | 24,960 Da (49,920 Da) | 49,918 Da     | –                            |
| Phospholipase A2                         | Q9DF33      | Homodimer        | 13,176 Da (26,352 Da) | 26,375 Da     | –                            |
| Snake venom metalloproteinase            | D3TTC2      | –                | 47,684 Da        | 49,283 Da     | Glycosylation 1,599 Da       |

<sup>2</sup> Centered on the top of the glycoform distribution.

Snake venom toxins undergo an accelerated rate of evolution in the wild; their genes are exposed to higher rates of mutation than those expressing other proteins culminating in elevated rate of amino acid sequence variation (53, 54). These variations are often in the form of single or multiple amino acid substitutions in the toxins’ primary sequence. The resulting “new” toxins have molecular weights that can differ by less than 1 Da for isobaric alterations (e.g. Ile/Leu, Lys/Gln/GlyAla) to >20 Da (Ser/Arg etc.) from the original toxin. Although whole venom LC-MS/MS cannot handle adequately the high dynamic range and GELFrEE cannot easily separate proteoforms with small mass differences, changes in isoelectric point between proteoforms can allow their separation into distinct fractions using sIEF and subsequent fragmentation during LC-MS/MS. With improved separation of proteoforms and the concomitant improvement in experimental dynamic range, sIEF enables more efficient mass spectrometric isolations that allow toxins with similar masses to be identified by automated search algorithms. In this study, venom fractionation by sIEF increased the number of protein identifications from 17 to 113 when compared with whole venom analysis using the same LC-MS/MS pipeline. The observed 664% increase in identifications exemplify the benefit of venom dynamic range reduction by prefractionation, and is analogous to a similar increase observed with Crotalus atrox venom using BU approaches (55).

The large increase in protein identification and proteoform annotation can be attributed to the capability to separate toxins that present similar physicochemical properties. As an example the mass difference between identified Oh-55 Pf 0 and Oh-55 Pf 4 is 15.93 Da, however at the 7+ charge state the two species differ by only 2.28 m/z. In this case, the isotopic envelopes overlap, complicating the efficient isola-
tation of proteoforms using only mass-based separation. These two proteoforms have very similar amino acid composition (90% identity) and hydrophobicity, however the predicted pl difference of one unit (from 8.36 to 7.35) allowed for their complete separation by sIEF and precise analysis by dTD of the two distinct proteoforms, highlighting the importance of pre-fractionation for thorough characterization of venom content.

The increased sensitivity and level of protein characterization obtained by reducing venom complexity by pre-fractionation increased the opportunity to examine the specific molecular states, or proteoforms, for many of the identified proteins. Each proteoform presented here is based not only on a precursor mass (a technique that can often provide false positives because of the presence of isobaric species) but also the masses of the fragment ions produced from gas-phase dissociation of the precursor. The high confidence provided from searching with fragment ions measured at accurate mass (<10 ppm) allows for wide precursor tolerances and “delta M” mode, making it possible to identify proteins with unannotated sequence variations and PTMs. However, proteins that do not have established cleavage sites (proteolytic processing) or that are mis-annotated cannot be identified using this search type unless a large MS^1 tolerance is used in an “open” and error-tolerant search strategy (56). All different branches in the search tree with distinct MS^1 tolerances combined with the deep manual analysis of the obtained identifications allowed us to annotate the 53 unique proteoforms presented here. Of the 53 proteoforms characterized, the largest sequence variation observed was from the 3FTx family, which are the result of multigene families that encode multiple toxin proteoforms and is consistent with the birth-and-death model of gene family evolution (57). This source of toxin variation can now be measured precisely and is related to snakes’ capability to kill different prey (58) and thus to the predator-prey arms race (59).

Toxin variation in populations is well documented in toxicology and enables snakes to adapt, occupy distinct ecological niches, and develop new toxins (36, 59). However, the underlying mechanisms that drive rapid evolutionary variation remain unclear. Access to proteoform-level information can help illuminate major “hot-spots” of snake venom variation. Further, the effect of sequence variation on how PTM profiles of mature toxins in venom can now be determined. The ability to directly readout such combinatorial variation opens the possibility to survey populations and correlate the findings with snakebite pathology and antivenom strategies.

An additional level of information that may be accessed from proteoform-resolved measurements is how toxins are being processed during their maturation from the venom gland to their presence in the milked venom. The final toxin products carry information about how the toxins were proteolytically processed as they reached maturity, and if they contain any PTMs such as pyroglutamic acid, deamidation, and glycosylation. Many toxin families like ohanin, PLA2, snake venom metalloproteinases (SVMP), and cobra venom factor (CVF) are expressed as preprotoxins and the maturation process in many of these cases remains unclear or poorly studied (36).

According to the gene organization of ohanin proposed by Pung and colleagues (60) we identified six proteoforms for this toxin, including 3 proteoforms with different lengths of the mature toxin and 3 proteoforms from the propeptide region. The mature ohanin is a poorly-studied neurotoxin that induces hypolocomotion and hyperalgesia in mice and was first isolated from the venom of king cobra (61). However, there have been no previous studies about the ohanin propeptide. The presence of three propeptide proteoforms can be related to the creation of new peptides with distinct biological activity in a similar way to the prodomains of SVMP (62) or toxins in tandem expression similar to BPP-natriuretic peptides (63). The better characterization of ohanin proteoforms can provide a better understanding of how main-chain toxins and propeptides are processed in the venom gland of snakes. Also, venom degradation by endogenous proteases can generate proteolytic cleavage products and some of them are generated more abundantly during the milking and handling process in the absence of proteinase inhibitors for the metallo- and serine-proteinases present in the venom (64).

For SVMP we identified several cleavage products from the prodomain, consistent with cleavage in the venom gland lumen as observed previously (62). Prodomain degradation in small peptides has been shown to be a mechanism to generate biologically active protease inhibitors that remain inactive in the venom gland and are only activated when the venom is injected into the prey (62). The major decrease in dynamic range from extensive pre-fractionation coupled with improvements in automated MS^2-based search routines and newer instrumentation (i.e. upgrading from a LTQ- XL Orbitrap to a Orbitrap Elite and a Q-Exactive HF) allowed us to identify more than 8 times more proteins than the prior work (Table I) (21) while more precisely mapping proteoforms present in the venom.

Unfortunately, dTD methodologies still have limitations for analyzing intact proteins larger than ~50 kDa, especially those in very low concentrations in the venom, typically lower than 0.5%. The limitations in denaturing mode are related to instrumentation, reduction of signal because of multiple charge states, chemical noise, incomplete desolvation and the presence of multiple PTMs like glycosylation that can frustrate the detection of such signals (65). The use of a native fractionation technique coupled to native MS/MS mitigated many of these problems and proved to be an efficient method to analyze the largest proteins present in the venom. Further, nTD preserves the macromolecular interactions between the toxins and allow the identification and characterization of the complexes, their subunits, and the partial characterization of the glycan structures present in the LAIO complex. Access to
Mapping Proteoforms and Protein Complexes From King Cobra Venom

PTM and toxin interaction information is vital for a complete venom characterization and it is the first time that this information has been obtained using MS in a toxinology study of large proteins.

In the BU comparison the only toxin family below 50 kDa that was not identified by TDPs was the natriuretic peptide family, represented in the BU identifications by the VBNJ2J, a similar toxin to “Natriuretic peptide Oh-NP” (D9I9X89). There is no prior protein evidence using classical biochemical methods for these gene products and the basis for protein annotation was information from the genome and transcriptome (36, 66). We suggest that sequence variance from different venom samples in addition to non-annotated cleavage sites and low concentrations may have contributed to the non-identification of this protein family by TDPs. Comparing our BU identifications to recently published papers that used BU for protein identification in O. hannah venom, we were able to identify 80 and 40% more protein families than the identifications presented by Petras et al. (21) and Vonk and co-workers (36), respectively. Notable differences may be related to the different methodologies applied, instrumentation used, and/or variation in snake venom sample (4, 7, 8). Our BU quantitation results showed large amounts of 3FTx in the venom that is consistent with the previous studies of the king cobra venom proteome (21, 36, 67) and with the rapid neurotoxicity because of muscular transmission blockade (i.e. paralysis) observed in human envenomation (68, 69). The comparison with the proteomics gold standard BU showed that TDPs is now a feasible technique for detection, identification, and characterization of even low-abundance proteoforms in snake venom. Even at <1% abundance, highly potent toxins can play import roles in human envenomation and in prey submission by themselves or through synergistic activity with the main toxins (70).

The combination of denaturing and native TDPs to explore the venom proteome of O. hannah has provided a superlative catalogue of venom protein content. Using our pipeline, the dTD approach led to the identification of 184 proteoforms from 15 protein families in a high-confidence process. These results not only represent a ~1,100% increase in protein identification when compared with the pipeline presented by Calvete and co-workers, but also provided the first proteoform-specific look at the rapidly evolving sequences of the venom proteome. The use of nTD provided a new alternative to study toxins larger than 50 kDa including glycoproteins and their complexes. Using intact mass and fragmentation information, toxins can be identified and characterized even if not present in the protein database using de novo sequencing. The use of both nTD and dTD in “venomics” stands ready to address many of the open challenges in venom toxicology. It allows for a precise molecular catalogue of toxin proteoforms and protein complexes. Thus, TDPs is poised to provide significant contributions to snake natural history and ecology. Furthermore, better understanding of venom will generate new knowledge about snakebites and antivenom production with important impacts in human and animal health.

* The W. M. Keck Foundation generously provided funding for this work. This material is based upon work supported by FAPERJ research grant 100.039/2014 from the government of Rio de Janeiro State - Brazil for R.D.M., by the Swiss National Science Foundation under Early Postdoc Mobility fellowship for L.F., and by the National Science Foundation Graduate Research Fellowship under fellowship number 2014171659 for O.S.S. The work was also supported by NIH GM 067193, and the Office for Research at Northwestern University. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

This article contains supplemental material.

† To whom correspondence may be addressed. Departments of Chemistry and Molecular Biosciences, the Chemistry of Life Processes Institute, and the Proteomics Center of Excellence, Northwestern University, 2–100 Hogan Hall, 2205 Tech Drive, Evanston, IL, 60208. Tel.: +1 847 467-4362; E-mail: n-kelleher@northwestern.edu; Proteomics Unit, Rio de Janeiro Proteomics Network, Depar-tements of Bioquimica and Bioquimica e Bioenergetica, Universidade Federal do Rio de Janeiro, Av. Athos da Silveira Ramos, 149, CT A-542, Cidade Universitaria, 21941-908 Rio de Janeiro, RJ, Brazil. Tel.: +55 21 3938-7353; E-mail: gilberto@iq.ufrj.br.

REFERENCES

1. Escobas, P., and King, G. F. (2009) Venomics as a drug discovery platform. Expert Rev. Proteomics 6, 221–224
2. McCleary, R. J. R., and Kini, R. M. (2013) Non-enzymatic proteins from snake venoms: A gold mine of pharmacological tools and drug leads. Toxicon 62, 56–74
3. Fox, J. W., and Serrano, S. M. T. (2005) Snake toxins and hemostasis. Toxicon 45, 949
4. Georgieva, D., Arni, R. K., and Betzel, C. (2008) Proteome analysis of snake venoms: pharmacological insights. Expert Rev. Proteomics 5, 787–797
5. Vetter, I., Davis, J. L., Rash, L. D., Anangi, R., Mobli, M., Aiewood, P. F., Lewis, R. J., and King, G. F. (2011) Venomics: a new paradigm for natural products-based drug discovery. Amino Acids 40, 15–28
6. Hargreaves, A. D., Swain, M. T., Hegarty, M. J., Logan, D. W., and Mulley, J. F. (2014) Restriction and Recruitment—Gene Duplication and the Origin and Evolution of Snake Venom Toxins. Genome Biol. Evol. 6, 2088–2095
7. Calvete, J. J. (2013) Snake venomics: From the inventory of toxins to biology. Toxicon 75, 44–62
8. Fox, J. W., and Serrano, S. M. T. (2008) Exploring snake venom proteomes: multifaceted analyses for complex toxin mixtures. Proteomics 8, 909–920
9. Calvete, J. J., Juarez, P., and Sanz, L. (2007) Snake venomics. Strategy and applications. J. Mass Spectrom. 42, 1405–1414
10. Juarez, P., Sanz, L., and Calvete, J. J. (2004) Snake venomics: Characterization of protein families in Sistrurus barbouri venom by cytoe mapping, N-terminal sequencing, and tandem mass spectrometry analysis. Proteomics 4, 327–338
11. Melani, R. D., Goto-Silva, L., Nogueira, F. C. S. N., Junqueira, M., and Domont, G. B. (2014) In Venom Genomics and Proteomics, eds Gopalakrishnakone P, Calvete JJ (Springer Netherlands, Netherlands), pp 1–12. 1st Ed.
12. Compton, P. D., and Kelleher, N. L. (2012) Spinning up mass spectrometry for whole protein complexes. Nat. Methods 9, 1065–1066
13. Smith, L. M., and Kelleher, N. L. (2013) Proteof orm: a single term describing protein complexity. Nat. Methods 10, 186–187
14. Bogdanov, B., and Smith, R. D. (2005) Proteomics by FTICR mass spectrometry: Top down and bottom up. Mass Spectrom. Rev. 24, 168–200
15. Perry, R. H., Cooks, R. G., and Noll, R. J. (2008) Orbitrap mass spectrometry: Instrumentation, ion motion and applications. Mass Spectrom. Rev. 27, 661–699
surface of venom phospholipase A2 enzymes. J. Mol. Evol. 48, 125–132

54. Nakashima, K., Nobuhisa, I., Deshimaru, M., Nakai, M., Ogawa, T., Shimohigashi, Y., Fukumaki, Y., Hattori, M., Sakaki, Y., and Hattori, S. (1995) Accelerated evolution in the protein-coding regions is universal in crotalinae snake venom gland phospholipase A2 isozyme genes. Proc. Natl. Acad. Sci. U.S.A. 92, 5605–5609

55. Fox, J. W., Ma, L., Nelson, K., Sherman, N. E., and Serrano, S. M. T. (2006) Comparison of indirect and direct approaches using ion-trap and Fourier transform ion cyclotron resonance mass spectrometry for exploring viperid venom proteomes. Toxicon 47, 700–714

56. Chick, J. M., Kolippakkam, D., Nusinow, D. P., Zhai, B., Rad, R., Huttlin, E. L., and Gygi, S. P. (2015) A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat. Biotechnol. 33, 743–749

57. Fry, B. G., Wüster, W., Kini, R. M., Brusic, V., Khan, A., Venkataraman, D., and Rooney, A. P. (2003) Molecular evolution and phylogeny of elapid snake venom three-finger toxins. J. Mol. Evol. 57, 110–129

58. Daltry, J. C., Wüster, W., and Thorpe, R. S. (1996) Diet and snake venom evolution. Nature 379, 537–540

59. Heatwole, H., and Poran, N. S. (1996) Resistances of sympatric and allopatric eels to Sea Snake venoms. Copeia 1996, 136–147

60. Pung, Y. F., Kumar, S. V., Rajagopalan, N., Fry, B. G., Kumar, P. P., and Kini, R. M. (2006) Ohanin, a novel protein from king cobra venom: Its cDNA and genomic organization. Gene 371, 246–256

61. Pung, Y. F., Wong, P. T. H., Kumar, P. P., Hodgson, W. C., and Kini, R. M. (2005) Ohanin, a novel protein from King Cobra venom, induces hypocomotion and hyperalgesia in mice. J. Biol. Chem. 280, 13137–13147

62. Portes-Junior, J. A., Yamanouye, N., Carneiro, S. M., Knittel, P. S., Sant’Anna, S. S., Nogueira, F. C. S., Junqueira, M., Magalhães, G. S., Domont, G. B., and Moura-da-Silva, A. M. (2014) Unraveling the processing and activation of snake venom metalloproteinases. J. Proteome Res. 13, 3338–3348

63. James Graham, R. L., Graham, C., McClean, S., Chen, T., O’Rourke, M., Hirst, D., Theakston, D., and Shaw, C. (2005) Identification and functional analysis of a novel bradykinin inhibitory peptide in the venoms of New World Crotalinae pit vipers. Biochem. Biophys. Res. Commun. 338, 1587–1592

64. Tashima, A. K., Zelanis, A., Kitano, E. S., Lanzer, D., Melo, R. L., Rioli, V., Sant’Anna, S. S., Schenberg, A. C. G., Camargo, A. C. M., and Serrano, S. M. T. (2012) Peptidomics of three Bothrops snake venoms: Insights into the molecular diversification of proteomes and peptidomes. Mol. Cell. Proteomics MCP 11, 1245–1262

65. Compton, P. D., Zambdborg, L., Thomas, P. M., and Kelleher, N. L. (2011) On the scalability and requirements of whole protein mass spectrometry. Anal. Chem. 83, 6868–6874

66. Lei, W., Zhang, Y., Yu, G., Jiang, P., He, Y., Lee, W., and Zhang, Y. (2011) Cloning and sequence analysis of an Ophiophagus hannah cDNA encoding a precursor of two natriuretic peptide domains. Toxicon 57, 811–816

67. Chang, H.-C., Tsai, T.-S., and Tsai, I.-H. (2013) Functional proteomic approach to discover geographic variations of king cobra venoms from Southeast Asia and China. J. Proteomics 89, 141–153

68. Tin-Myint, Rai-Mra, Maung-Chit, Tun-Pe, and Warrell, D. (1991) Bites by the King Cobra (Ophiophagus hannah) in Myanmar: Successful treatment of severe neurotoxic envenoming. QJM 80, 751–762

69. Veto, T., Price, R., Slusby, J. F., and Carter, J. A. (2007) Treatment of the first known case of king cobra envenomation in the United Kingdom, complicated by severe anaphylaxis. Anaesthesia 62, 75–78

70. Bordon, K. C. F., Perino, M. G., Giglio, J. R., and Arantes, E. C. (2012) Isolation, enzymatic characterization and antiedematogenic activity of the first reported rattlesnake hyaluronidase from Crotalus durissus terrificus venom. Biochimie 94, 2740–2748

71. Carvalho, P. C., Junqueira, M., Valente, R. H., and Domont, G. B. (2008) Caitlu: A tool to graphically represent peptide sequence coverage and domain distribution. J. Proteomics 71, 486–489