research articles Journal of proteome research

Sequence Similarity-Based Proteomics in Insects: Characterization of the Larvae Venom of the Brazilian Moth *Cerodirphia speciosa*

Anna Shevchenko,[†,||] Mirta Mittelstedt Leal de Sousa,[‡,||] Patrice Waridel,[†] Silvia Tolfo Bittencourt,[‡,§] Marcelo Valle de Sousa,[‡] and Andrej Shevchenko*[,†]

Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, 01307 Dresden, Germany, Brazilian Center for Protein Research, Department of Cell Biology, University of Brasilia, 70910-900, Brasilia, DF, Brazil, and FTB College, 72610-300, Recanto das Emas, DF, Brazil

Received January 19, 2005

Using a combination of tandem mass spectrometric sequencing and sequence similarity searches, we characterized the larvae venom of the moth *Cerodirphia speciosa*, which belongs to the *Saturniidae* family of the *Lepidoptera* order. Despite the paucity of available database sequence resources, the approach enabled us to identify 48 out of 58 attempted spots on its two-dimensional gel electrophoresis map, which represented 37 unique proteins, whereas it was only possible to identify 13 proteins by conventional non-error tolerant database searching methods. The majority of cross-species hits were made to proteins from the phylogenetically related *Lepidoptera* organism, the silk worm *Bombyx mori*. The protein composition of the venom suggested that envenoming by *C. speciosa* toxins might proceed through the contact with its hemolymph, similarly to another toxic *Lepidoptera* organism, *Lonomia obliqua.*

Keywords: MS BLAST • homology search • insect proteomics • tandem mass spectrometry • insect venom

The Brazilian moth *Cerodirphia speciosa* belongs to the *Saturniidae* family, the *Hemileucinae* subfamily of *Lepidoptera* order. Lepidopterian larvae having urticating hairs, such as larvae of the *Saturniidae* family, are dangerous to humans. Accidental contact with their hairs and spines results in burning sensations, renal failure, epistaxis, melena, hematuria, severe haemorrhagia, and sometimes, death.[1] However, neither the composition of the venom of caterpillars of the *Cerodirphia* genus, nor specific toxic agent(s) have yet been determined. The proteomic characterization of a toxic fluid, resting in urticating hairs and spines of larvae, could help to elucidate the toxicity mechanisms and develop appropriate medicines.

The *Lepidoptera* order comprises more than 150 000 species of butterflies and moths. Currently, sequences of only 106 highly homologous proteins from multiple species are known for the entire *Saturniidae* family, the *Hemileucinae* subfamily, which contains 142 taxonomically identified species. No proteins from *C. speciosa* are available in a protein database and therefore protein identification could only rely on cross-species matches to sequences from other *Lepidoptera* species, such as the silk worm *Bombyx mori*[2,3] and the tobacco hornworm *Manduca sexta*,[4] or sequences from even more distantly related insects, such as the fruit fly *Drosophila melanogater*[5] or the malaria mosquito *Anopheles gambiae.*[6]

Genomic, EST, and protein sequence databases provide a resource for the identification of proteins by mass spectrometry. In a typical proteomic routine, proteins are digested in-gel,[7] or in-solution,[8] by specific proteases (most often, by trypsin). The unseparated tryptic digest is subjected to peptide mass fingerprinting, in which intact masses of peptides are determined with high accuracy. Alternatively, peptides are sequenced by tandem mass spectrometry, either directly from the unseparated mixture,[9] or after on-line separation by nano-flow reversed-phase chromatography (reviewed in ref 10). Masses of intact tryptic peptides alone (in peptide mass fingerprinting), or together with masses of derived fragment ions (in tandem mass spectrometry), are then correlated with corresponding masses calculated by in silico processing of sequences from databases entries. Despite significant differences in the spectrum-to-sequence correlation and scoring algorithms, and their program realization, conventional protein identification software, such as Mascot[11] and SEQUEST,[12] is primarily capable of exact matching of analyzed peptides to database sequences (reviewed in refs 13,14). Stringent matching dramatically increases the specificity of database searches, helps to overcome the inherent paucity of tandem mass spectra and enables confident identification of proteins using spectra acquired from only a few fragmented peptide precursors. However, any discrepancy between sequences of the analyzed peptides, and sequences of the corresponding database entries, typically results in mismatch and precludes the identification of the protein. This represents a significant bottleneck in the characterization of proteomes of species with unsequenced genomes,[15,16] and especially affects proteomics in insects

* To whom correspondence should be addressed. E-mail: shevchenko@mpi-cbg.de.
[†] Max Planck Institute of Molecular Cell Biology and Genetics.
[‡] University of Brasilia.
[§] FTB College.
[||] Equal contribution of these authors.

10.1021/pr0500051 CCC: $30.25
Published on Web 05/05/2005

because of their remarkable phylogenetic diversity and high population divergence of protein sequences between organisms with undefined genetic backgrounds.[17] We demonstrate here that a combination of mass spectrometry and sequence-similarity database searches helps to circumvent these limitations and paves the way to efficient exploration of proteomes of insects, despite the paucity of available sequence resources.

To characterize the venom of *C. speciosa*, we separated proteins by two-dimensional gel electrophoresis and identified them by mass spectrometry and a combination of Mascot and Mass Spectrometry driven BLAST (MS BLAST) searches.[18] MS BLAST is a sequence similarity searching tool, which utilizes redundant, degenerate and partially inaccurate sequence candidates, obtained by the interpretation of tandem mass spectra of peptides. All candidate sequences, deduced from of all acquired tandem mass spectra are assembled into a single query in an arbitrary order. Importantly, many sequence candidates per each analyzed peptide are accepted, and therefore sequences could be deduced from MS/MS spectra of lower quality that are not amenable for the unequivocal interpretation. Contrary to conventional BLAST searches,[19,20] which employ *E*-values and *p*-values for assessing the statistical confidence of hits, MS BLAST utilizes an alternative scoring scheme, which is based on the pre-computed threshold scores that are set conditionally on the number of retrieved high scoring segment pairs (HSPs) and the total number of fragmented precursors.[21] According to the computational evaluation of the results of searches with more than 20 000 MS BLAST queries assembled from 4500 individual proteins, the rate of false positive identifications did not exceed 3%.[21] Although searches are always performed against a comprehensive sequence database, the rate of true positive identifications typically depends on the phylogenetic distance to the most related reference organism(s) and differs substantially between various phylogenetic lineages. MS BLAST methods were successfully applied to the identification of unknown proteins from the African clawed frog *Xenopus laevis*,[22] Dead Sea alga *Dunaliella salina*,[23] methylotrophic yeast *Pichia pastoris*,[18] holm oak,[24] and other species with unsequenced genomes. However, MS BLAST capabilities for identifying proteins from wild-bred insects that are phylogenetically distant to sequenced reference species have not been evaluated yet.

Using a combination of mass spectrometry and error-tolerant and stringent database search methods, we identified 48 out of 58 attempted protein spots that were detectable on a 2D gel of the caterpillar venom. Sequence-similarity searches enabled the identification of twice as many unique proteins, compared to cross- species identifications by stringent database searches[25] and its confidence was not compromised by natural polymorphism of insect proteins. The abundance of hemolymph proteins in the venom pointed to a similarity between the mechanisms of production and secretion of the venom between caterpillars of *Cerodirphia speciosa* and phylogenetically related toxic organism *Lonomia obliqua*.[26]

Materials and Methods

***C. speciosa* Caterpillars.** *C. speciosa* caterpillars were collected at the Experimental Station of Biology of the University of Brasilia, Brazil, and were fed with plant leaves from the site where they had been found. Cleaning and feeding was performed three times a week, before extraction of the venom. Some caterpillars were allowed to develop to adult moths for subsequent taxonomical classification, which was performed by Amabílio José Aires de Camargo at Centro de Pesquisa Agropecuária dos Cerrados, Brasilia, Brazil.

Venom Extraction. A group of twelve caterpillars, approximately 6 cm long, was kept at −20 °C for 20 min before dissection. Hair fibers were removed at their base using bistouries. Drops of hair secretion leaking from the cuts were collected with a Pasteur pipet and dried in a vacuum centrifuge.

Amino Acid Analysis. In two independent experiments, 1 mg and 150 μg of the hair secretion were dissolved in 100 μL and 75 μL of 0.1 M HCl, respectively. The acid hydrolysis of the hair secretion samples and protein standards was performed in 6 M HCl under vacuum for 24 h at 109 °C. After acid hydrolysis, the samples and standards were solubilized in 75 μL of 0.1 M HCl, and 50 μL were injected into an amino acid analyzer Hitachi L8500 (Tokyo, Japan). The total protein content was calculated by summing up the determined amounts of recovered amino acids.

Separation of Venom Proteins by Two-Dimensional Gel Electrophoresis. This was performed according to the methodology described by Görg et al.[27,28] IPG strips were purchased from Amersham Biosciences (Uppsala, Sweden). Isoelectric focusing was performed on an IPG-phor system (Amersham Biosciences) with 11 cm IPG-strips, pH 3−10. The aliquot of venom (150 μg) was dissolved in 250 μL of a solution containing 7 M urea, 2 M thiourea, 1% DTT, 2% Triton X-100, 0.8% ampholytes (Pharmalyte), 1% of each protease inhibitor (TLCK, PMSF, TPCK, EDTA, pepstatin A, and leupeptin) and traces of bromophenol blue. The sample was centrifuged for 5 min at 10 000 rpm in a bench centrifuge. The rehydration of the IPG strips was performed for 13 h at 20 °C. Isoelectric focusing runs followed a program of 4 h at constant 75 μA per strip, at 500 V/1 h; 1000 V/1 h and 8000 V/2 h. After that, the strips were left incubating for 20 min in a solution (3 mL) containing 50 mM tris pH 8.8, 6 M urea, 30% v/v glycerol, 2% w/v SDS and 125 mM DTT, followed by 20 min in the same equilibration solution with DTT substituted with 125 mM iodoacetoamide. Then IPG strips were rinsed for 5 min in 37.5 mM Tris buffer containing 0.3 M glycine and 0.1 M SDS. The IPG strips were sealed with a solution containing 0.3% agarose in Tris SDS-PAGE buffer for electrophoresis in the second dimension, which was performed in a gradient (7% to 15% acrylamide-bisacrylamide) with 4% stacking gel. Electrophoresis was performed at 25 mA with a voltage limit of 500 V at 15 °C. A solution containing 0.24 M Tris HCl pH 8.8, 12% saccharose, 2% SDS, 1% DTT and traces of bromophenol blue was used as a sample buffer. Gels were stained with Coomasie Brilliant Blue R250. The approximate amount of loaded proteins was estimated by comparing the staining intensity of corresponding spots with the intensity of spots of four protein standards with a concentration determined by amino acid analysis.

Protein Identification by Mass Spectrometry. Individual protein spots were manually excised from 2D gels and in-gel digested with trypsin as described previously.[7] Protein digests were first subjected to peptide mass fingerprinting by matrix-assisted laser desorption/ionization on a Bruker Reflex IV time-of-flight (MALDI-TOF) mass spectrometer equipped with Scout 384 ion source. Probes were prepared by dried-droplet method as described previously.[29] Briefly, 1 μL aliquot of the digest was mixed on the surface of AnchorChip 384/600 targets (Bruker Daltonics, Germany) with a saturated solution of matrix (α-cyano-4-hydroxycinnamic acid) in 1:2 (v/v) solution of 2.5% aqueous TFA: acetonitrile. The mixture was allowed to dry at

Table 1. Proteins Identified in the Venom of *C. speciosa* Caterpillars

spot	protein	accession no.	MW, kDa	organism	acquired MS/MS spectra	peptides matched by MASCOT	de novo sequenced peptides	peptides matched by MS BLAST
1	Arylphorin	A34287	83	*Bombyx mori*	30	5		
2	Prophenoloxidase	O77002	78	*Hyphantria cunea*	22		6	5
3	Transferrin	O97158	73	*Bombyx mori*	17	6		
4	Apolipophorin-II	Q25490	370	*Manduca sexta*	29	2		
5	Aldehyde dehydrogenase	P32872	54	*Saccharomyces cerevisiae*	19	2		
6	CG14639 protein	Q9VMZ5	37	*Drosophila melamogaster*	25		7	3
7	Catalase	P17336	58	*Drosophila melamogaster*	30	(*1*)[a]	8	5
8	Hypothetical protein 13	Q5MGN4		*Lonomia obliqua*	13		4	2
9	Hypothetical protein 13	Q5MGN4		*Lonomia obliqua*	6		4	2
10	IDGF like protein	Q9GV28	48	*Bombyx mori*	23	(*1*)[a]	4	3
11	Hemolin	P25033	46	*Hyalophora cecropia*	19	3		
12	CG16885-PA protein (Articulin), *polym.*[d]	Q8SZM2	29	*Drosophila melamogaster*	19	(*1*)[a]	6	5
13	Follicle-specific yolk polypeptide-4, *polym.*[d]	Q24993	32	*Galleria mellonella*	12		6	2
14	Follicle-specific yolk polypeptide-4, *polym.*[d]	Q24993	32	*Galleria mellonella*	17		5	2
15	Glyceraldehyde-3-phosphate dehydrogenase	P00357	35	*Homarus americanus*	18	4		
16	Ommochrome-binding protein	P31420	31	*Manduca sexta*	17		6	2
17	Larval cuticle protein LCP-30	Q08738	24	*Bombyx mori*	17		2	1
18	24 kDa female-specific fat body protein	O96096	24	*Antheraea yamamai*	19		4	3
19	24 kDa female-specific fat body protein	O96096	24	*Antheraea yamamai*	21		4	3
20	Biliverdin binding protein-I	Q8T118	22	*Samia cynthia ricini*	13		5	2
21	CG7178 protein (Troponin I-like)	Q9VWY2	23	*Drosophila melamogaster*	13		3	3
22	CG32209 protein	Q9VW34	54	*Drosophila melamogaster*	9		3	3
23	CG8756 protein	Q9VW30	62	*Drosophila melamogaster*	23		4	3
24	Calreticulin	Q869E0	46	*Bombyx mori*	16		3	3
25	Arylphorin	A34287	83	*Bombyx mori*	19	6		
26	Actin, alpha and beta tubulins		42	*various species*	PMF[f]			
27	Actin, beta tubulin		42	*various species*	PMF[f]			
28	Antichymotrypsin I precursor, *polim.*	Q03383	40	*Bombyx mori*	36		5	5
29	Tropomyosin	AX399623	32	*Plodia interpunctella*	PMF[f]			
30	Annexin IX-C	Q9NL59	36	*Bombyx mori*	16	4		
31	n.i.,[e] PMF identical to #33, *polim.*				20		7	
32	GASP protein	Q9VNL0	29	*Drosophila melamogaster*	18	3		
33	n.i.,[e] PMF identical to #31, *polim.*				18		3	
34	AgCP4195	Q7QHZ7	26	*Anopheles gambiae*	19		6	4
35	ENSANGP00000024537 ORF	Q7PHG0	27	*Anopheles gambiae*	19		7	2
36	Flexible cuticle protein 12	P45589	11	*Hyalophora cecropia*	9		3	1
37	Flexible cuticle protein 12	P45589	11	*Hyalophora cecropia*	13		4	2
38	n.i.[e]				15		4	
39	n.i.,[e] PMF identical to #40				14		6	
40	n.i.,[e] PMF identical to #39				11		6	
41	n.i.,[e] PMF identical to #46				18		8	
42	Larval cuticle protein 16/17 precursor, *polym*	Q25504	12	*Manduca sexta*	11		7	3
43	Myosin light chain alkali	Q24621	17	*Drosophila pseudoobscura*	11		4	3
44	Larval cuticle protein 16/17 precursor, *polym*	Q25504	12	*Manduca sexta*	13		2	2
45	Larval cuticle protein LCP-17	O02387	15	*Bombyx mori*	11		4	1
46	n.i.,[e] PMF identical to #41				11		4	
47	CG9070 protein (Cuticle protein)	Q9V5W0	11	*Drosophila melamogaster*	15		3	2
48	CG9070 protein (Cuticle protein)	Q9V5W0	11	*Drosophila melamogaster*	15		3	2
49	Larval cuticle protein LCP-17	O02387	15	*Bombyx mori*	15		3	2
50	Larval cuticle protein LCP-17	O02387	15	*Bombyx mori*	11		2	1
51	Larval cuticle protein 16/17 precursor, *polym*	Q25504	12	*Manduca sexta*	12		6	4
52	n.i.[e]				*b*			
53	n.i.[e]				*b*			
54	Muscle-specific protein 20	P14318	20	*Drosophila melamogaster*	*b*	3		
55	n.i.[e]				*b*			
56	AgCP3324 protein	Q7QIH4	35	*Anopheles gambiae*	*b*		30	11[c]
57	Lim protein	Q6SA71	11	*Bombyx mori*	*b*	4		
58	Muscle-specific protein 20	P14318	20	*Drosophila melamogaster*	*b*	3		

[a] Mascot matched only a single peptide with borderline significance. The identification was confirmed by MS BLAST by confident matching of multiple peptides. [b] Spots analyzed by LC−MS/MS and more than 5000 MS/MS spectra were acquired that, however, mostly represented trypsin autolysis products, keratins and other background ions. [c] Protein sequence contains multiple repeating peptide sequence stretch. [d] polym. − indicates samples with multiple polymorphic peptide sequences. [e] n.i. − not identified. [f] PMF − identified by peptide mass fingerprinting and confirmed by LC−MS/MS sequencing and Mascot searches.

room temperature and the entire target was washed with 5% formic acid. If peptide mass fingerprinting did not identify the protein, then peptides were extracted from the gel pieces with 5% formic acid and acetonitrile and the extracts were pooled together and dried down in a vacuum centrifuge. The digests were taken up in 5% formic acid and analyzed by nanoelectrospray tandem mass spectrometry on a modified MDS Sciex QSTAR Pulsar *i* quadrupole time-of-flight (QqTOF) instrument as previously described.[30,31] Several low abundant protein spots (indicated in Table 1) were analyzed by nanoLC−MS/MS on a linear ion trap mass spectrometer LTQ (ThermoElectron Corp., CA) essentially as described in.[32]

Processing of Mass Spectrometric Data and Database Searching. Peptide mass fingerprints were used for database searching using Mascot software (Matrix Science Ltd, UK) against the MSDB database downloaded from NCBI (September 2004). Mass tolerance was set to 100 ppm, spectra were calibrated externally and no restrictions were imposed on protein molecular weight or species of origin of analyzed proteins. Uninterpreted tandem mass spectra were first searched by Mascot against the above database to identify proteins comprising tryptic peptides identical to those present in database entries, at a precursor mass tolerance of 0.1 Da and fragment ion mass tolerance of 0.05 Da. Hits were considered significant if their protein score exceeded the threshold score calculated by Mascot software assuming $p < 0.05$ and, depending on the size of the query, was 50−52. Corresponding MS/MS spectra were manually inspected to confirm the match of continuous series of *y*-, *b*-, and *a*- fragment ions. Mascot queries were generated from MS/MS spectra using the processing script Mascot v.1.6b2 as an extension of BioAnalyst QS software from Applied Biosystems (Foster City, CA). Tandem mass spectra acquired by nanoLC−MS/MS method on the LTQ ion trap mass spectrometer were searched by Mascot under the following settings: mass tolerance for precursor and fragment ions was 2 and 0.5 Da, respectfully; fragment ion profile: ESI trap. Cross-species hits were accepted if three criteria were simultaneously met: the protein score exceeded the threshold calculated by Mascot for $p < 0.01$, several peptides exactly matched a database entry and at least one peptide matched with the Expect value lower than 0.1. If no identification by tandem mass spectrometry was achieved, or if the proposed hits were not statistically confident (their scores were below the threshold score suggested by Mascot), tandem mass spectra acquired on a QSTAR Pulsar mass spectrometer were interpreted de novo using BioAnalyst QS software as previously described.[31] Multiple sequence candidates were allowed per each interpreted tandem mass spectrum, and peptide sequences were not necessarily complete. All candidate sequences were merged into a single search string. MS BLAST searches[18,31] were performed against a nonredundant protein database (nrdb95) at http://genetics.bwh.harvard.edu/msblast/ under default settings. Parsing script operating at the MS BLAST web site was applied to identify and color code statistically confident hits in accordance with the MS BLAST scoring scheme.[21] MS BLAST scoring scheme is discussed in detail in the article by Habermann et al,[21] in which the corresponding substitution matrix PAM30MS and table of significance thresholds are provided. MS BLAST searches were also performed against the raw genome sequence of silk worm *Bombyx mori*, which was downloaded from NCBI. Genomic version of MS BLAST will be described in detail elsewhere.

Results and Discussion

Characterization of the Venom by Two-Dimensional Gel Electrophoresis and Mass Spectrometry. The venom sample, containing 50 μg of protein material in total (as estimated by amino acid analysis) was separated by 2D gel electrophoresis and protein spots were visualized by Coomassie staining (Figure 1). Altogether, 58 spots were excised from the gel and in-gel digested with trypsin. For protein identification we combined stringent and sequence-similarity searches in a layered approach (reviewed in[13,15]). First, 1 μL aliquots were taken from protein digests and analyzed by MALDI TOF. Peptide mass fingerprinting relies on the identity of many (typically, 5−10) peptides to sequences from database entries[33] and therefore it was not expected to produce many cross-species hits to homologous proteins from distantly related species. However, the method enabled rapid identification of proteins with a high percentage of sequence identity and consumed only a small amount of tryptic digests, so that the remaining peptide material could be further analyzed by tandem mass spectrometry, if no conclusive identification was achieved by MALDI TOF. Tryptic peptides were further extracted from gel pieces and sequenced by nanoES MS/MS. Uninterpreted MS/MS spectra were subjected to database searches using Mascot software. Although Mascot can only fetch peptides identical to database sequences, its specificity in cross-species identification is higher compared to peptide mass fingerprinting, since tandem mass spectra are rich in sequence-specific fragment ions. Thus, matching of only two to three tandem mass spectra typically yields a statistically confident cross-species hit.[34,35] If no identification was achieved, tandem mass spectra were interpreted de novo and the obtained sequences were used for MS BLAST database searches. De novo interpretation of tandem mass spectra is laborious. However, because of the layered approach to data analysis and database searching, it was only required when analyzed proteins were poorly conserved and both methods of stringent searches (Mascot searches with peptide mass fingerprints and with uninterpreted tandem mass spectra) failed. We found that for low abundant proteins, manual interpretation of tandem mass spectra, acquired on a hybrid quadrupole time-of-flight mass spectrometer, produced data of superior quality, compared to automated de novo sequencing,[36] and therefore, it was used almost exclusively in this work. However, even manual interpretation was often ambiguous and rendered many candidate sequence proposals, rather than a single accurate peptide sequence. Therefore, all sequence proposals obtained from all fragmented peptides were assembled into a MS BLAST query

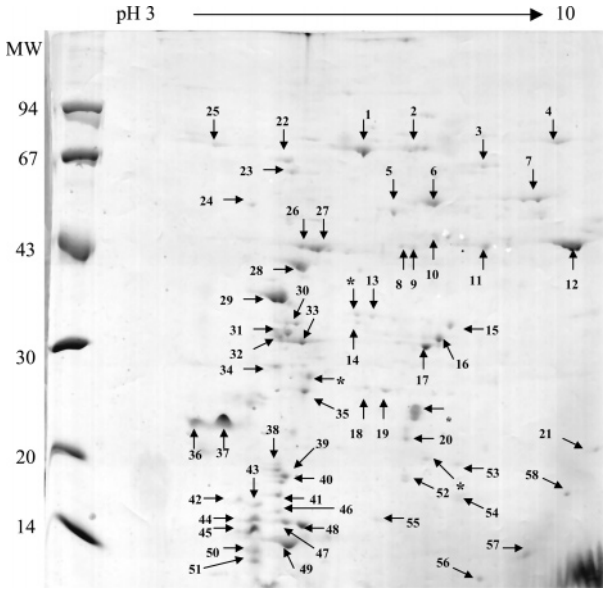


Figure 1. Separation of proteins from *C. speciosa* venom by two-dimensional gel electrophoresis. Protein spots were visualized by Coomassie Brilliant Blue staining. Spots analyzed by mass spectrometry are designated by numbers; their identification is provided in Table 1. Spots designated with asterisks were heavily contaminated with keratins and their analysis revealed only common products of trypsin autolysis.

and searched against a protein database, against a nonredundant EST database and against the genome sequence of another *Lepidoptera* organism, the silkworm *Bombyx mori*.[3] An example of MS BLAST protein identification is presented at the Figure 2. Altogether, from 58 attempted spots 48 spots (>80%) were identified (Table 1). As was anticipated, all these identifications were produced by cross-species matching to known relatively conserved proteins from other insects. A majority of cross-species hits was made to proteins from the phylogenetically related silk worm *Bombyx mori*, and only a fraction of hits corresponded to proteins from other insect species in agreement with previously reported computational estimates.[21] The identified spots represented 38 individual proteins. MALDI peptide mass fingerprinting identified only three strongly conserved cytoskeleton proteins—actin, tubulin, and tropomyosin. Tandem mass spectrometry and stringent (Mascot) database searches identified further 11 proteins. In three cases a single peptide was matched to a database entry and the statistical significance of hits, represented by their MOWSE score,[11] was borderline. These hits were subsequently validated by MS BLAST searches as described below in detail. Taken together, conventional methods of database searching (peptide mass fingerprinting and Mascot) enabled the identification of 14 (38%) of all proteins.

Tandem mass spectra acquired from the digests of yet unidentified protein spots were interpreted de novo and sequence candidates were submitted to MS BLAST searches. Searches against a protein database additionally identified 25 proteins, thus increasing the success rate by 2-fold, compared to conventional database mining methods (Table 1). While Mascot searches in three protein spots only matched a single protein sequence with marginal significance, MS BLAST confidently matched three to five peptides. However, it would be prudent to note that even statistically confident identifications do not necessarily support direct assignment of the biological function.[37] For sequence-similarity based identifications it was particularly important to determine if the identification indicated a full length sequence similarity between the sequenced and reference proteins, or a local similarity of a conserved sequence domain in otherwise functionally different proteins. To this end, we carefully inspected the list of MS BLAST hits to check if the same peptide sequences match other nonhomologous proteins with lower confidence.

In attempt to improve the identification success rate, we also tried MS BLAST searches against EST sequences from *Bombyx mori*. Although in many instances, MS BLAST hit EST sequences with one to five fully matching peptides, subsequent *blastx* searches with full-length sequences of correspondent EST clones produced no new identifications, in addition to those already made in a protein database.

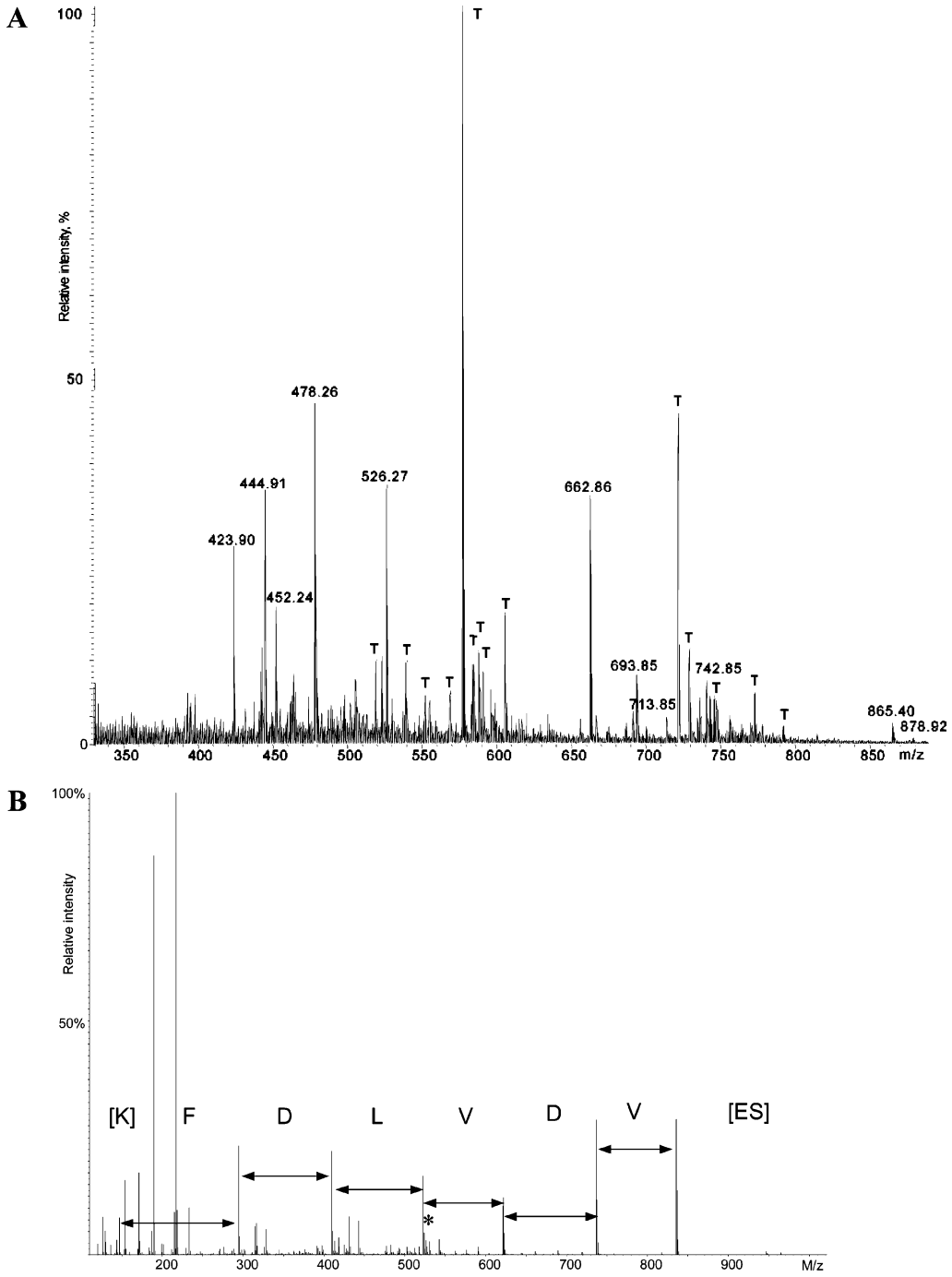
Strong polymorphism of insect protein sequences often hampers their identification. Since the intact masses, and masses of fragments ions of variant peptides, are different, they are not recognized by conventional database searching software. However, MS BLAST searches tolerate multiple mismatches and amino acid substitutions between queried sequences (i.e., in peptide sequences determined by mass spectrometry) and database sequences, and therefore tolerate high degeneracy of protein sequences in individual wild-bred insects. In the analysis of five protein spots we observed from two to four variants of the same peptide sequence, and they were successfully matched to the corresponding database sequences by MS BLAST (Table 3). The genome of another organism of the *Lepidoptera* family—silk worm *Bombyx mori*—has recently become available. All MS BLAST queries were subsequently searched against a raw genome sequence consisting of several contigs by a modified version of MS BLAST, which is based on the *tblastn* (rather than *blastp*) search algorithm. Although these searches also did not produce any new identifications, compared to MS BLAST searches against a protein database, in many cases genomic MS BLAST helped to narrow down the list of homologous protein hits in a few insect species to the correspondent *Bombyx mori* gene and thus improved their functional annotation.

Identified Proteins and Possible Sources of Venom Toxicity. *C. speciosa* caterpillar venom is a complex mixture of proteins of intracellular (larval cuticle proteins, calreticulin, aldehyde dehydrogenase 2, and catalase) or hemolymph (storage proteins, apolipophorin II, pro-phenol oxidase subunit 1, hemolin, and transferrin) origin. It also contains abundant housekeeping proteins, involved in energy metabolism (alcohol dehydrogenase, prophenoloxidase, and others), iron and calcium storage (transferrin and calreticulin), lipid storage and transport (apolipophorin) as well as ubiquitous components of cytoskeleton (actin, myosin, troponin), and a Ca-dependent lipid binding protein (annexin). Insect-specific proteins were presented by several cuticle and chitin-binding proteins, which are the abundant components of insect integument and yolk, and hemolin, which is involved in immune response. Interestingly, we found that spots 8 and 9 contained a homologue of the hypothetical protein 13 (Q5MGN4) from the highly hemorrhagic caterpillar *Lonomia obliqua.*

Although all major spots on the 2D gel were identified, we did not find common toxic components of many animal venoms, such as phospholipases, proteases and neurotoxins.[38] We previously identified a number of phospholipases and neurotoxic molecules from the venom of the snake, *Bothrops atrox,* by cross-species matching to a variety of species, apart from reptiles (Shevchenko et al., unpublished). Phospholipases from *D. melanogaster* (CG14507-PB) and *A. gambiae* (AgCP13927) share 69% of the sequence identity and their homologues were found by full-length protein BLAST searching in the *B. mori* genome and therefore, it is likely that *C. speciosa* phospholipases would be identifiable by sequence similarity searches.[21] Since we also did not observe phospholipase activity in the crude venom, we concluded that phospholipases might only be present in a very minute amount and could not be responsible for the venom toxicity.

Abundance of hemolymph proteins prompted us to speculate that *C. speciosa* might be using similar mechanisms of production and secretion of the venom, as the hemorrhagic caterpillar *Lonomia obliqua*.[26] Veiga et al described epithelium that underlines spines and comprises specialized cells producing the venom, which is then deposited in the extracellular space between epithelium and spine cuticles and is predominantly concentrated at the tips of the spines.[39] If venom-secreting cells are located similarly in *C. speciosa*, then the bulk of collected secretion would be represented by hemolymph, residing in the interior of bristles, along with intracellular proteins that might originate from injured epithelium. Arocha-Piñango and Layrisse demonstrated[40] that envenoming is carried out through the injection of hemolymph and other body fluids of the caterpillars. The hemolymph toxicity of *Lepidoptera* species has been previously observed.[40−42]

Although we could not identify proteins whose toxicity is apparent, we cannot exclude that they might be present in



C -BTALGDALDKTSLK- BATLGDALDKTSLK - BXXLGDALDKTSLK -BSEVDVLDFK- BESVDVLDFK - BXXVDVLDFK –BXXSEFNPDFLTVANK- BFTSEFNPDFLTVANK- BTFSECNPDFLTVANK - BXXDVLNQWANEK-BLVQPLFYTESK-BDVNLELPR -FEESAPHLVK-BXXVELPYNEP-BMVELPYNETD-BLYVGNQYVNNDK-

Figure 2. MS BLAST identification of spot #28. Panel A: nanoES mass spectrum of its the unseparated tryptic digest. Peaks of trypsin autolysis products are designated with "T". Peptide precursor ions whose tandem mass spectra were acquired and interpreted are designated with corresponding *m*/*z*. Panel B: MS/MS spectrum of the precursor ion with *m*/*z* 526.27 (designated with an asterisk). De novo sequencing was performed by considering mass differences between the adjacent peaks in the series of fragments that belong to the ions containing C-terminal amino acid residue (y-ions). In high *m*/*z* region the spectrum contained only a few intense ions and corresponding stretch of peptide sequence could be deduced unambiguously. In lower *m*/*z* region a few combinations of fragment ions whose mass differences match the masses of amino acid residues or their combinations, were possible. The sequence shown in the panel was deduced by considering the most abundant fragment ions and was not necessarily correct. A few, equally probable, sequence proposals were used for database searches by MS BLAST. Panel C: MS BLAST query that comprises all sequence proposals obtained by the interpretation of all tandem mass spectra assembled in an arbitrary order and spaced by a minus symbol. B stands for a generic trypsin cleavage site (R or K) preceding peptide sequences. The results of MS BLAST search are presented in Table 2.

Table 2. Identification of the Protein from Spot 28 by MS BLAST[a]

m/*z*	sequence proposals	MS BLAST alignments
444.91	BTALGDALDKTSLK	Query: 4 LGDALDKTSLK 14
666.86	BATLGDALDKTSLK	LGD +DKTSLK
452.24	BXXLGDALDKTSLK	Subject: 29 LGDTIDKTSLK 39
526.27	BSEVDVLDFK	Query: 16 BSEVDVLDF 24
	BESVDVLDFK	+SEVD ++F
	BXXVDVLDFK	Subject: 136 RSEVDNINF 144
865.40	BFTSE***F***NPDFLTVANK	Query: 31 EFNPDFLTVANK 42
893.92	BXXSEFNPDFLTVANK	EF+P FLTVA K
878.92	BTFSE***C***NPDFLTVANK	Subject: 104 EFDPKFLTVA IK 115
693.85	BXXDVLNQWANEK	Query: 5 DVLNQWANE 13
		D++N WA+E
		Subject: 152 D I INRWADE 160
662.86	BLVQPLFYTESK	Query: 16 BLVQPLFYTE 25
		+L Q LFYTE
		Subject: 221 RLLQSLFYTE 230
478.26	BDVNLELPR	Query: 72 BDVNLELP 79
		+D+ LE+P
		Subject: 285 RD IE IE IP 292
423.90	FEESAPHLVK	Query: 93 FE ESAPHLVK 102
		F+E AP +VK
		Subject: 315 FQEPAPAI VK 324
742.85	BMVELPYNE***T***D	Query: 48 BMVELPYNE 56
		+M+ELPY E
		Subject: 237 KM I ELPYK E 245
713.85	BLYVGNQYVNNDK	Query: 2 B L YVGNQY 9
		++YV + QY
		Subject: 115 K I YVSD QY 122

[a] All sequenced peptides were matched to Q03383 Antichymotrypsin I precursor from distantly related *Lepitoptera* organism silk worm *Bombyx mori*. Not a single sequenced peptide was identical to the corresponding database sequence.

Table 3. Matching Polymorphic Peptide Sequences by MS BLAST

spot	*m*/*z*	candidate sequence	MS BLAST alignments[a]
42	890.7	NF***D***FETEDGL***N***R	Query: 3 FDFETEDGLNR 13
			F FETEDG+ R
			Subject: 39 FGFETEDGISR 49
44	989.4	NF***D***FETEDGL***A***R	Query: 3 FDFETEDGLAR 13
			F FETEDG+ R
			Subject: 39 FGFETEDGISR 49
51	981.9	NF***Q***FETEDGL***N***R	Query: 3 FQFETEDGLNR 13
			F FETEDG+ R
			Subject: 39 FGFETEDGISR 49

[a] MS BLAST search matched three similar peptides derived from spot nos. 42, 44, and 51 to the same region of Larval cuticle protein 16/17 (ac.no Q25504) from tobacco horn worm *Manduca sexta*. Natural polymorphism of peptide sequences did not compromise the statistical confidence of MS BLAST alignments.

seven spots, which, according to their peptide mass fingerprints, represent five proteins that remained unidentified. Their further characterization would require accurate de novo sequencing,[30,43] synthesis of degenerate oligonucleotide primers, creating a cDNA library and cloning by a PCR-based method.[44] As a provisional solution, we provide here corresponding MS BLAST queries in the Supporting Information, so that database searches can be repeated and proteins identified in the future once the related sequences from other insect species appear in a database.

Conclusion

We demonstrated here that sequence similarity searches substantially expand the boundaries of proteomics in insects, whose genomes are not known, and pave the way for numerous biological applications that require accurate charting of the protein composition of insect body fluids. The ongoing sequencing of insect genomes and ESTs both contribute to increased representation of protein sequences in databases and will enable the characterization of proteins from more phylogenetically distant species. Importantly, sequence similarity searches tolerated remarkable polymorphism of protein sequences and enabled efficient identification of proteins collected from a population of wild-bred organisms with completely undefined genetic backgrounds. Although in the present work de novo sequencing of peptides heavily relied on manual interpretation of spectra, remarkable progress has been achieved in developing software-assisted interpretation of spectra acquired in LC−MS/MS runs[45] and combined in a fully automated sequence-similarity identification pipeline.

Acknowledgment. The authors are grateful for Carlos André Ornelas Ricart (Brazilian Center for Protein Research, Brasilia) for expert assistance in protein separation and useful discussion. We are indebted to Dr. Henrik Thomas (MPI CBG, Dresden) for his valuable assistance in MALDI peptide mass fingerprinting, Dr. Youri Kravatskiy for his assistance with genomic MS BLAST searches and to Ms. Judith Nicholls for critical reading of the manuscript.

Supporting Information Available: Corresponding MS BLAST queries. This material is available free of charge via the Internet at http://pubs.acs.org.

Note Added after ASAP Publication: This manuscript was originally published on the Web 05/05/2005 with a misprinted letter and imperfect alignment in Table 2. The version published on the Web (05/18/2005) and in print is correct.

References

(1) Arocha-Pinango, C. L.; de Bosch, N. B.; Torres, A.; Goldstein, C.; Nouel, A.; Arguello, A.; Carvajal, Z.; Guerrero, B.; et al. *Thromb. Haemost.* **1992**, *67*, 402−407.
(2) Mita, K.; Morimyo, M.; Okano, K.; Koike, Y.; Nohata, J.; Kawasaki, H.; Kadono-Okuda, K.; Yamamoto, K.; et al. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100*, 14121−14126.
(3) Mita, K.; Kasahara, M.; Sasaki, S.; Nagayasu, Y.; Yamada, T.; Kanamori, H.; Namiki, N.; Kitagawa, M.; et al. *DNA Res.* **2004**, *11*, 27−35.
(4) Robertson, H. M.; Martos, R.; Sears, C. R.; Todres, E. Z.; Walden, K. K.; Nardi, J. B. *Insect Mol. Biol.* **1999**, *8*, 501−518.
(5) Adams, M. D.; Celniker, S. E.; Holt, R. A.; Evans, C. A.; Gocayne, J. D.; Amanatides, P. G.; Scherer, S. E.; Li, P. W.; et al. *Science* **2000**, *287*, 2185−2195.
(6) Holt, R. A.; Subramanian, G. M.; Halpern, A.; Sutton, G. G.; Charlab, R.; Nusskern, D. R.; Wincker, P.; Clark, A. G.; et al. *Science* **2002**, *298*, 129−149.
(7) Shevchenko, A.; Wilm, M.; Vorm, O.; Mann, M. *Anal. Chem.* **1996**, *68*, 850−858.
(8) Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd *Nat. Biotechnol.* **2001**, *19*, 242−247.
(9) Wilm, M.; Shevchenko, A.; Houthaeve, T.; Breit, S.; Schweigerer, L.; Fotsis, T.; Mann, M. *Nature* **1996**, *379*, 466−469.
(10) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198−207.
(11) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551−3567.
(12) Eng, J. K.; McCormack, A. L.; Yates, J. R. *J. Am. Soc. Mass Spectrom.* **1994**, *5*, 976−989.
(13) Liska, A. J.; Shevchenko, A. *Trends Anal. Chem.* **2003**, *22*, 291−298.

(14) Fenyo, D. *Curr. Opin. Biotechnol.* **2000**, *11*, 391−395.
(15) Liska, A. J.; Shevchenko, A. *Proteomics* **2003**, *3*, 19−28.
(16) van Wijk, K. J. *Plant Physiol.* **2001**, *126*, 501−508.
(17) Zdobnov, E. M.; von Mering, C.; Letunic, I.; Torrents, D.; Suyama, M.; Copley, R. R.; Christophides, G. K.; Thomasova, D. et al. *Science* **2002**, *298*, 149−159.
(18) Shevchenko, A.; Sunyaev, S.; Loboda, A.; Bork, P.; Ens, W.; Standing, K. G. *Anal. Chem.* **2001**, *73*, 1917−1926.
(19) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. *Nucleic Acids Res.* **1997**, *25*, 3389−3402.
(20) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. *J. Mol. Biol.* **1990**, *215*, 403−410.
(21) Habermann, B.; Oegema, J.; Sunyaev, S.; Shevchenko, A. *Mol. Cell. Proteomics* **2004**, *3*, 238−249.
(22) Liska, A. J.; Popov, A. V.; Sunyaev, S.; Coughlin, P.; Habermann, B.; Shevchenko, A.; Bork, P.; Karsenti, E. et al. *Proteomics* **2004**, *4*, 2707−2721.
(23) Liska, A. J.; Shevchenko, A.; Pick, U.; Katz, A. *Plant Physiol.* **2004**, *136*, 2806−2817.
(24) Jorge, I.; Navarro, R. M.; Lenz, C.; Ariza, D.; Porras, C.; Jorrin, J. *Proteomics* **2004**.
(25) Candas, M.; Loseva, O.; Oppert, B.; Kosaraju, P.; Bulla, L. A., Jr. *Mol. Cell. Proteomics* **2003**, *2*, 19−28.
(26) Seibert, C. S.; Shinohara, E. M.; Sano-Martins, I. S. *Toxicon* **2003**, *41*, 831−839.
(27) Görg, A.; Postel, W.; Günther, S.; Weser, J. *Electrophoresis* **1985**, *6*, 599−604.
(28) Görg, A.; Postel, W.; Gunther, S. *Electrophoresis* **1988**, *9*, 531−546.
(29) Thomas, H.; Havlis, J.; Peychl, J.; Shevchenko, A. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 923−930.
(30) Shevchenko, A.; Chernushevich, I.; Wilm, M.; Mann, M. *Mol. Biotechnol.* **2002**, *20*, 107−118.
(31) Shevchenko, A.; Sunyaev, S.; Liska, A.; Bork, P.; Shevchenko, A. *Meth. Mol. Biol.* **2002**, *211*, 221−234.
(32) Mayya, V.; Rezaul, K.; Cong, Y. S.; Han, D. *Mol. Cell. Proteomics* **2005**, *4*, 214−223.
(33) Lester, P. J.; Hubbard, S. J. *Proteomics* **2002**, *2*, 1392−1405.
(34) Koller, A.; Washburn, M. P.; Lange, B. M.; Andon, N. L.; Deciu, C.; Haynes, P. A.; Hays, L.; Schieltz, D. et al. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 11969−11974.
(35) Cooper, B.; Eckert, D.; Andon, N. L.; Yates, J. R.; Haynes, P. A. *J. Am. Soc. Mass Spectrom.* **2003**, *14*, 736−741.
(36) Arif, S. A.; Hamilton, R. G.; Yusof, F.; Chew, N. P.; Loke, Y. H.; Nimkar, S.; Beintema, J. J.; Yeang, H. Y. *J. Biol. Chem.* **2004**, *279*, 23933−23941.
(37) Rappsilber, J.; Mann, M. *Trends Biochem. Sci.* **2002**, *27*, 74−78.
(38) Bon, C. *Biochimie* **2000**, *82*, 791−792.
(39) Veiga, A. B.; Blochtein, B.; Guimaraes, J. A. *Toxicon* **2001**, *39*, 1343−1351.
(40) Arocha-Pinango, C. L.; Layrisse, M. *Lancet* **1969**, *1*, 810−812.
(41) Amarant, T.; Burkhart, W.; LeVine, H., 3rd; Arocha-Pinango, C. L.; Parikh, I. *Biochim. Biophys. Acta* **1991**, *1079*, 214−221.
(42) Arocha-Pinango, C. L.; Marval, E.; Guerrero, B. *Biochimie* **2000**, *82*, 937−942.
(43) Uttenweiler-Joseph, S.; Neubauer, G.; Christoforidis, S.; Zerial, M.; Wilm, M. *Proteomics* **2001**, *1*, 668−682.
(44) Shevchenko, A.; Wilm, M.; Mann, M. *J. Protein Chem.* **1997**, *16*, 481−490.
(45) Zhang, Z. *Anal. Chem.* **2004**, *76*, 6374−6383.

PR0500051