research articles Journal of proteome research

Mapping the Larval Midgut Lumen Proteome of *Helicoverpa armigera*, a Generalist Herbivorous Insect

Yannick Pauchet,[†] Alexander Muck,[‡] Aleš Svatoš,[‡] David G. Heckel,[†,*] and Susanne Preiss[†]

Department of Entomology and Mass Spectrometry Research Group, Max Planck Institute for Chemical Ecology, Hans-Knöll-Str. 8, D-07745 Jena, Germany

Received September 25, 2007

The gut lumen is the primary site of digestion and detoxification and thus presents conditions hostile to most proteins. We used 2D-gel electrophoresis and MS/MS de novo peptide sequencing to identify the major proteins stable enough to persist in the midgut lumen of caterpillars of the cotton bollworm *Helicoverpa armigera*, a generalist herbivorous insect and a major crop pest worldwide. As expected, we found several enzymes responsible for digestion of carbohydrates, proteins, and lipids. In addition, we identified nondigestive proteins such as a multidomain lipocalin, a protein with pathogen recognition domains, an arginine kinase related to a class of major human allergens, and abundant proteins of unknown function. Identification of the set of proteins that are secreted into the lumen will enable us to further characterize the nutritional and defensive functions of this important intraorganismal space.

Keywords: polyphagous herbivore • digestive enzymes • insect proteomics • tandem mass spectrometry • MS BLAST • EST

Introduction

The lepidopteran larval midgut lumen is a hostile environment for proteins. This follows from the primary function of this compartment as a site of digestion. Here, recently ingested plant material encounters a host of secreted digestive enzymes, which cleave carbohydrates, lipids, and proteins into small molecules that are absorbed into the epithelial cells and passed via the circulatory system to the rest of the body. Yet the insect's own digestive enzymes and defensive molecules must persist and operate in this challenging environment, and studies on the protein level are necessary to understand the function of this complex bioreactor. Although genomic approaches based on cDNA libraries and microarrays can provide information on the identity and abundance of mRNAs produced in the surrounding cells, they provide no information on which proteins are actually secreted into the lumen and offer no insights into their temporal dynamics afterward, which are exclusively determined by post-translational processes.

Most caterpillars (larvae of the insect order Lepidoptera, butterflies, and moths) are herbivorous and must digest large amounts of plant material to achieve high growth rates. The digestive system or gut is a simple tube designed for efficient extraction of nutrients in a constant flow-through system (Figure 1A).[1] After being torn from the leaf or fruit by the larva's sharp mandibles, food particles pass through the relatively small foregut, where they are mixed with enzymes and other digestive secretions from the salivary and mandibular glands. The midgut is the longest section of the gut, and it is here that most of the digestion and absorption of nutrients takes place, generally at a high pH. As the larva eats constantly between molts, a long continuous food bolus passes through the midgut to the hindgut. Here, water is removed, and waste filtered through the Malpighian tubules is added to the bolus which is eventually excreted as frass or fecal pellets.

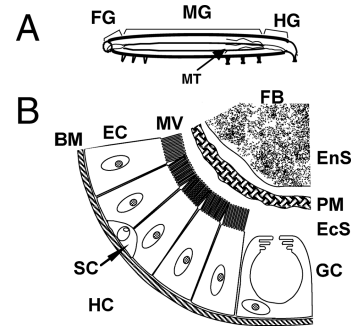

Figure 1. Schematic diagram of the digestive system of *Helicoverpa armigera*. (A) Sagittal section through a fourth-instar larva. Abbreviations: FG, foregut; MG, midgut; HG, hindgut; MT, Malpighian tubules. (B) Portion of a cross-section through the midgut. Abbreviations: FB, food bolus; EnS, endoperitrophic space; PM, peritrophic matrix; EcS, ectoperitrophic space; EC, columnar epithelial cell; MV, microvilli; GC, goblet cell; SC, stem cell; BM, basement membrane; HC, hemocoel. Not to scale.

The midgut itself is a hollow tube of cells encircled by a basement matrix, lying within the hemocoel or body cavity of the larva (Figure 1B). The exterior of the basement matrix is bathed in hemolymph that carries nutrients to the rest of the body. Most of the inner surface of the basement matrix is covered by the basal ends of a single layer of cells, whose apical ends face the interior of the lumen. The two main types are

* Corresponding author. Tel.: +49 3641 57 1500. Fax: +49 3641 57 1502. E-mail: heckel@ice mpg.de.
† Department of Entomology.
‡ Mass Spectrometry Research Group.

10.1021/pr7006208 CCC: $40.75

Published on Web 03/04/2008

columnar epithelial cells, with most of their apical surfaces elaborated into microvilli, and goblet cells, which possess a central cavity partly open to the lumen containing the proton ATPase pump and associated K+/H+ antiporter system responsible for maintaining a high K+ gradient[2] and driving amino acid uptake.[3] Stem cells lie between the basal regions of these two cell types and differentiate to replenish them.

The central cavity or lumen itself is divided into three subcompartments by the peritrophic matrix (PM), which is a hollow meshwork tube of chitinous fibers cross-linked by proteins (Figure 1b).[4] The outer subcompartment (ectoperitrophic space) consists of the spaces between the microvilli and the region between their tips and the outer surface of the PM. Then there is the peritrophic matrix itself, which in most Lepidoptera is secreted by cells at the junction of the foregut and midgut and provides a continuous sheath enclosing the food bolus. PM proteins have been classified according to the degree of their association with the chitinous meshwork.[5] Class 1 PM proteins can be removed by washing with physiological buffers. Class 2 represents the PM proteins extractable by mild detergents like CHAPS. Class 3 includes those only extractable by strong denaturants (e.g., urea), and Class 4 PM proteins are not extractable even by strong denaturants. The innermost subcompartment is the endoperitrophic space containing the mass of food matter being digested.

Proteins secreted into the lumen by the surrounding cells may be blocked or trapped within the peritrophic matrix if they are too large to pass through the meshwork of chitin fibrils or may pass through to contact the bolus of food if sufficiently small. Once inside the endoperitrophic space, proteins are carried posteriorly along with the mass movement of the food bolus and eventually excreted with the frass. However, it has been proposed that some proteins could be recycled by passing through the PM and being carried anteriorly by a countercurrent operating in the ectoperitrophic space.[4,6] To our knowledge, this intriguing suggestion has not yet been confirmed by direct measurements of protein flux.

The main function carried out by lumen proteins is digestion, and we would expect to see a high abundance and activity of enzymes that cleave lipids, carbohydrates, and proteins into smaller molecules that can be efficiently absorbed by the transport systems in the microvillar membranes.[7] In addition, many plants produce toxic compounds to deter herbivory, and the lumen represents a potential early site of enzymatic detoxification or sequestration by binding followed by excretion, to avoid absorbing toxins into the cells along with nutrients. The lumen is also a vulnerable portal of entry by pathogens, including viruses such as NPVs that infect midgut cells before spreading to the rest of the insect,[8] and bacteria such as *Bacillus thuringiensis* that produce pore-forming insecticidal toxins.[9] Proteins to sense the presence of these pathogens, and defensive proteins such as antimicrobial peptides, would also be expected to be present; yet, most studies of insect immunity focus on the hemolymph and fat body and have ignored the midgut.[10] Finally, there are some plant-produced proteins, highly resistant to insect proteases, that can interfere with insect digestive processes;[11,12] thus another expected function of insect-secreted proteins would be to counter these plant antiherbivory defenses.

The purpose of our study was to initiate an exploratory survey of the major proteins secreted into and stably persisting in the lumen cavity of a generalist herbivore, *Helicoverpa armigera* (cotton bollworm). We combined the analytical approaches of 2D-PAGE gel separation and MS/MS de novo peptide sequencing[13] along with information from *H. armigera* midgut cDNA libraries and sequence information from other Lepidoptera. As well as the expected digestive enzymes, novel proteins that could be involved in pathogen recognition were detected, the first such report for the midgut. Proteins similar to sequences from other Lepidoptera of uncharacterized function were found in the lumen for the first time. In addition, a cytosolic protein was found in high abundance, suggesting a means of protein transport into the lumen that does not use the canonical secretory pathway.

Material and Methods

Insects. The TWB strain of *Helicoverpa armigera* Hübner (Lepidoptera: Noctuidae) was collected from the vicinity of Toowoomba, Queensland, Australia, in January 2003 and maintained in the laboratory in Jena since August 2004, for about 25 generations prior to the start of this study. Neonates destined for proteomic studies were reared on a chemically defined diet containing only casein as the protein source and no plant-derived material, as described by Vanderzant[14] with a few modifications (see Supporting Information 1) at 26 °C with a 16:8 (L:D) photoperiod.

Sample Preparation. Midguts were dissected from 60 actively feeding second-day fifth-instar larvae in ice-cold phosphate-buffered saline (PBS) for each sample. The peritrophic matrix containing the food bolus was pulled out of the midgut with forceps and gently homogenized by 10 strokes in a Potter-Elvehjem homogenizer in PBS pH 7.5 containing a cocktail of protease inhibitors (Complete EDTA-free, Roche Applied Science) in order to release soluble proteins. After centrifugation (30 000*g*, 30 min, 4 °C), the supernatant containing the gut lumen soluble proteins was kept, and protein concentration was determined using the Protein Dye reagent (BioRad) and bovine serum albumin (BSA) as standard.

Separation of Proteins by Two-Dimensional Gel Electrophoresis. Gut lumen proteins were precipitated by 6% trichloroacetic acid using 0.02% sodium deoxycholate as a carrier. After two washes with 100% acetone, the protein pellets were solubilized in IEF lysis buffer (7 M urea, 2 M thiourea, 2% CHAPS, 60 mM DTT, 1% carrier ampholytes and protease inhibitors). Insoluble material was removed by centrifugation (15 000*g*, 20 min, 20 °C). Protein concentration was determined using the 2D-Quant kit (GE Healthcare). Isoelectric focusing was performed on a PROTEAN IEF Cell (BioRad) with 24 cm IPG-strips, pH 3–11 NL (GE Healthcare). IPG strips were loaded with 500 μg of total proteins in 450 μL of IEF lysis buffer and passively rehydrated for 16 h at 20 °C. Isoelectric focusing runs followed a program of 8 h at maximum 50 mA per strip, at 500 V for 1 h; 500−1000 V in 1 h; 1000−8000 V in 3 h; and 8000 V for 3 h. In order to fully reduce and alkylate proteins, the strips were incubated for 30 min in a solution containing 50 mM Tris pH 8.8, 6 M urea, 30% v/v glycerol, 2% (w/v) SDS, and 1% (w/v) DTT, followed by 30 min in the same equilibration solution with DTT substituted by 2.5% (w/v) iodoacetamide. The IPG strips were then sealed with a solution containing 0.3% agarose in SDS-PAGE buffer on top of a 10% or 12% SDS-PAGE gel used for the second dimension. Electrophoresis was performed at a maximum power of 20 W per gel with a voltage limit of 500 V at 20 °C using the Ettan DALT 6 apparatus (GE Healthcare). Gels were fixed for 16 h in 40% ethanol and 10% acetic acid and then stained with PageBlue (Fermentas), which can typically detect down to 5 ng of protein, about 10 times the

minimum detection limit of silver staining. To verify that the observed size range of 25–65 kDa (see Results) was not due to the well-known bias of 2D-gels against larger proteins, an aliquot was separated on a 12% 1D polyacrylamide gel and the same size range of proteins observed (data not shown). Spots were considered for identification only if they were found in three replicate gels with samples from three independent dissections.

Protein Spot Picking and Processing. The protein spots were manually picked from 2D-gels, destained, trypsinized, and extracted as follows.[15] Samples were processed on 96-well microtiter plates with an Ettan TA Digester running the Digester Version 1.10 software (GE Healthcare Bio-Sciences AB). The excised gel plugs were washed for 20 min with 70 μL of acetonitrile/50 mM ammonium bicarbonate a total of four times. Following two additional washes with 70 μL of 70% acetonitrile for 20 min each, the gel plugs were air-dried for 1 h. Trypsin digestion was carried out overnight with 50 ng of porcine trypsin (Promega) in 15 μL of 50 mM ammonium bicarbonate at 37 °C. This was combined with 25 μL of extraction solution (70% acetonitrile and 0.1% trifluoroacetic acid) and incubated for 20 min, and 20 μL was removed to a 96-well plate. The gel plug was then incubated with an additional 40 μL of extraction solution for 20 min, and the solution was transferred to the plate. The 60-μL solution containing the peptide mixture was then vacuum-dried for 1 h.

MALDI-TOF/MS. Dry peptides were dissolved in 10 μL of aqueous 0.1% trifluoroacetic acid. A 1 μL aliquot was mixed with 1 μL of α-cyano-4-hydroxycinnamic acid (alpha matrix, 10 mg/mL in ethanol:acetonitrile, 1:1 [v/v]), and 1 μL of this was spotted onto a metal 96-spot MALDI target plate for cocrystallization. A MALDImicro MX mass spectrometer (Waters) was used in reflectron mode to analyze the tryptic peptides.[15] A strong electrical field was created to accelerate the sample ions into the flight tube toward the detector (5 kV on the sample table, 212 kV on the extraction grid, pulse voltage of 1.95 kV, and 2.35 kV detector voltages). A nitrogen laser (337 nm, 5 Hz) was used for ionization, with energies of ~50 mJ per pulse. MassLynx Version 4.0 software (Waters) was used for data acquisition. Each spectrum was combined from 10 laser pulses. Human Glu-Fibrinopeptide B (1570.6774 D) served as an external lock-mass reference. A tryptic digest of bovine serum albumin (MPrep, Waters) was used to calibrate the mass spectrometer. The MALDI-TOF spectra searches were performed in the Protein Lynx Global Server software, version 2.2 (PLGS 2.2, Waters), against the NCBI_insecta database (downloaded on 20 March 2006 from http://www.ncbi.nlm.nih.gov/database, 286 733 entries). The search parameters were as follows: peptide tolerance of 80 ppm, one missed cleavage, carbamidomethyl modification of cysteines, and possible methionine oxidation. An estimated calibration error of 0.05 D and a minimum of four peptide matches were the criteria for obtaining positive database hits. Results are documented in Supporting Information 2.

NanoLC MS/MS. The MALDI-TOF peptide signal intensities were used to estimate the volume of the remaining sample to be used for the subsequent nanoLC-MS/MS de novo sequence analysis. The appropriate aliquot of tryptic peptides in aqueous 0.1% trifluoroacetic acid (1.5–6 μL) was injected into the CapLC XE nanoLC system (Waters) equipped with a desalting precolumn. A mobile phase flow of 0.1% aqueous formic acid (15 μL/min for 5 min) was used to concentrate and desalt the samples on a 24 × 0.180 mm NanoEase C18 5 μm particle size precolumn (Waters). The samples were eluted on a 150 mm × 75 μm I.D., 3 μm NanoEase Atlantis C18 column (Waters), using an increasing acetonitrile gradient. Phases A (5% MeCN in 0.1% formic acid) and B (95% MeCN in 0.1% formic acid) were linearly mixed using a gradient program set to 5% phase B for 5 min, increased to 40% in 25 min, to 60% in 10 min, and finally to 95% B for 4 min. The eluted peptides were directly transferred through a Teflon capillary union and a metal coated nanoelectrospray tip (Picotip, 50 × 0.36 mm, 10 μm I.D., New Objective) in the NanoElectroSpray source into a Q-TOF Ultima tandem mass spectrometer (Waters). The source temperature was set to 40 °C, cone gas flow to 50 L/h, and the nanoelectrospray voltage was 1.6 kV. The TOF analyzer was used in reflectron mode. The MS/MS spectra were collected at 0.9 s intervals in the range of 50−1700 m/z. A mixture of 100 fmol/mL human Glu-Fibrinopeptide B and 80 fmol/mL reserpine in 0.1% formic acid/acetonitrile (1:1 v/v) was infused through the reference NanoLockSpray source every fifth scan to compensate for mass shifts in the MS survey scan and MS/MS CID (collision induced dissociation) fragmentation modes due to temperature and electronic fluctuations. The data were collected by MassLynx v4.0 software (Waters). PLGS 2.2 (Waters) was used for further data processing (baseline subtraction, smoothing, lock-mass correction for both precursor and fragments, deisotoping and deconvolution using MaxEnt3 module), producing the peak list, de novo peptide sequence identification, and database searches.

Using PLGS 2.2, The CID spectra were first searched against the SwissProt database for the presence of contaminants (mostly trypsin autodigest products and keratins) and then searched against the NCBI_insecta database with the following parameters: fixed precursor ion mass tolerance of 20 ppm for survey peptide, fragment ion mass tolerance of 0.05 Da, estimated calibration error of 0.005 Da, 1 missed cleavage, carbamidomethylation of cysteines and possible oxidation of methionine. The database search results were considered as positive hits when a protein identity likelihood score of 11.9, greater than the 95% confidence threshold, was obtained. The PLGS2.2 program calculates the probability of a correct sequence assignment as the product of the probability of assigning a peptide sequence before evaluating the experimental data (prior) and of the likelihood of the specific peptide fragmentation divided by the normalization factor obtained from experimental spectra. The model also considers the number of possible sequences and is tuned for tight matches. The probability of correct protein identification is then the product of all peptide probabilities (including the probabilities of peptide nonpresence). The reported scores are calculated as the natural logarithm of the protein probability divided by the inverse number of proteins (N) in the database. Thus, the highest possible score is obtained for a protein probability 1 where the score equals ln(N) (=12.566 for NCBI_insecta). Additional information on PLGS 2.2 statistics and scoring has been described by Skilling et al.[16] Results are documented in Supporting Information 3.

MS BLAST. Using PLGS 2.2, CID spectra were interpreted de novo to yield peptide sequences. Sequences with a ladder score (percentage of expected y- and b-ions) exceeding 30 were then used in a homology-based search strategy using the MS BLAST program.[17] MS BLAST was developed to utilize redundant, degenerate and partly inaccurate peptide sequences in similarity searches of protein databases that may be derived from organisms phylogenetically distant from the study species.

All candidate sequences from a given spot exceeding the threshold, even different sequences from the same peptide, are concatenated into a single query separated by dashes in an arbitrary order. The WU-BLAST2 BLASTP search engine (W. Gish 1996, http://blast.wustl.edu) is employed with parameter values that disallow gaps within a peptide and that score only the most significant match in the case of several peptide candidates covering the same region in the target sequence. In addition, the PAM30MS matrix, which accounts for the inability to distinguish I and L residues and allows for unknown residues X, is used in the blastp similarity search.[17] In practice, this enables identification of homologous proteins in other species with many amino acid substitutions, under conditions where spectral searches fail. Scoring of the significance of such matches is not based on *E*- or *p*-values of the individual HSPs (high-scoring segment pairs) but instead on precomputed threshold scores conditional on the number of query peptides and HSP hits. The color-coded output produced by the MS BLAST script identifies in red those target sequences with scores exceeding 99% of queries utilizing randomized peptide sequences by chance. Computational studies[18] have estimated a false positive rate of <3%. Individual searches were performed on the public server at http://dove.embl-heidelberg.de/Blast2/msblast.html. At our request, the MS BLAST server was also installed on the ButterflyBase web page (http://butterflybase.org/) for searching the ButterflyBase EST database from Lepidoptera, exclusive of *Bombyx mori* (34 882 protein sequences).[19] In addition, we installed an in-house MS BLAST server for searching NCBI_insecta and a locally generated EST database from *H. armigera* midgut cDNA libraries (5685 protein sequences). Additional information on MS BLAST statistics and scoring can be found in Shevchenko et al.[17] Results are documented in Supporting Information 4.

Results and Discussion

Characterization of the Gut Lumen Proteome in the Absence of Contaminating Proteins. The identification of insect gut lumen proteins can be complicated by contamination by abundant proteins present in the food source. Some plant proteins, such as ribulose bisphosphate carboxylase, although susceptible to digestive proteolysis, are abundant enough to dominate large regions of the 2D-gel and swamp out any comigrating insect proteins. Others, especially those acting as feeding deterrents, are stable enough such that specific enzymatic activity can be recovered in the frass.[11,12] Although analyzing starved individuals would have reduced this contamination, our aim was to analyze the midgut when it was actively functioning in digestion; therefore, we isolated soluble lumen proteins from larvae reared on our standard artificial diet containing wheat germ and pinto beans. 2D-gels revealed highly abundant spots corresponding to phaseolin (Figure 2B), a common protein present in all kinds of beans, which overwhelmed the larval-specific proteins. In order to further limit such extraneous contamination, for proteomic studies we reared *H. armigera* neonates on a chemically defined diet containing casein as the only protein source[14] (Supporting Information 1). The casein seems to be easily and quickly digested because no casein spots were found in our 2D-gels (Figure 2A). One potential limitation of this diet is that we may not detect enzymes such as chymotrypsins whose expression may be induced mainly by plant-produced compounds. Furthermore, we focused on the last larval instar in order to be able to efficiently dissect and recover the gut lumen. Changes in the gut lumen proteome from earlier larval instars may be expected, especially with respect to serine proteases,[20] and will be investigated in future studies.

The 2D-gel pattern of the gut lumen proteins revealed a fairly low complexity dominated by several highly abundant proteins (Figure 2A). Around 95% of the proteins range between 20 and 65 kDa in molecular weight. This observation can be correlated to the molecular weight cutoff imposed by the peritrophic matrix (PM). One of the major roles of the PM is compartmentalization of digestive events which is mainly achieved by sorting the soluble digestive enzymes according to their molecular weight. Proteins below the cutoff imposed by the PM can pass through it into the endoperitrophic space, and all proteins with a molecular weight above this cutoff remain in the ectoperitrophic space between the PM and the microvilli of the midgut cells.[4] The size range of proteins observed on the 2D-gels suggests that the molecular weight cutoff of the PM in fifth-instar *H. armigera* larvae is around 65 kDa.

Protein Identification. For the identification of major gut lumen proteins, a sequential strategy was employed. First, tryptic digests were performed on spots isolated from the 2D-gels, and MALDI-TOF peptide mass fingerprints were used to query the NCBI_insecta dabase. This identified only four spots corresponding to *H. armigera* protein sequences already present in that database (Supporting Information 2). Next, tryptic peptides were sequenced by nanoLC MS/MS, and the spectra obtained were subjected to searches of NCBI_insecta using the PLGS 2.2 search engine. To broaden our searches, tandem mass spectra were interpreted de novo, and the peptide sequences obtained were used for searches of the same database using the MS BLAST program.[17] As this database is not well represented with Lepidopteran proteins, we did two additional rounds of identification: the first to identify unannotated homologues in ESTs from other Lepidoptera and the second to identify specific *H. armigera* sequences from a midgut cDNA library.

PLGS2.2 searches using the MS/MS spectra obtained from tryptic peptides were able to identify 14 protein spots (Table 1), most commonly from *H. armigera* itself or closely related lepidopteran species such as *S. frugiperda* and *T. ni*. Such matches are not very tolerant of amino acid substitutions, and so the peptides responsible for these hits must have been nearly identical in these species. When amino acid sequences were estimated de novo from the remaining peptides and these sequences were used as queries for MS BLAST searches, additional matches were found, sometimes in the same proteins already identified but more often in distantly related species, increasing the number of identifications to 29. Thus, the imperfect peptide matches detected by MS BLAST from a larger number of different peptides in the same protein increased the support for existing identifications overall and provided new ones. This was still probably limited by the taxonomic bias of the NCBI_insecta database in favor of *Drosophila* and other dipteran insects.

In order to increase the number of potentially available sequences from more closely related Lepidoptera, all of the de novo peptide predictions were used in a second round of MS BLAST searches of ButterflyBase. This database combines publicly available Lepidopteran EST data sets downloaded from NCBI_dbEST (including 474 *H. armigera* ESTs coding for 270 putative proteins as of October 2006) and their predicted protein sequences;[19] results are shown in Table 2. The unannotated EST clusters hit in ButterflyBase were used in blastp

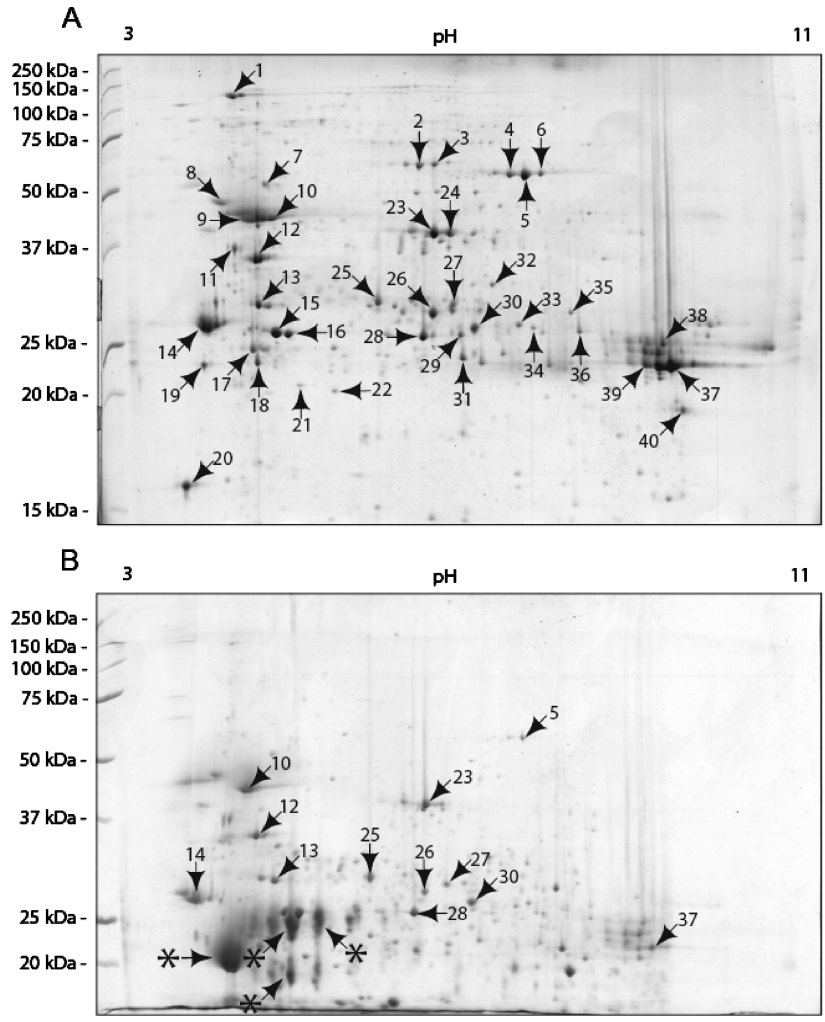

Figure 2. Separation of proteins from *H. armigera* larval gut lumen by 2D-gel electrophoresis followed by staining with PageBlue. (A) Larvae fed on a chemically defined diet containing casein (2nd dimension: 12% SDS-PAGE). (B) Larvae fed on a diet containing pinto beans (2nd dimension: 10% SDS-PAGE). Asterisks indicate protein spots identified as phaseolin. Spots analyzed by mass spectrometry are designated by numbers. Spots annotated with the same number in both gels were identified as the same proteins by tandem MS.

searches of UniRef100 to identify the most similar proteins described from other species. This enabled the assignment of five previously unidentified spots (1, 7, 15, 16, and 35). In four cases, the most similar annotated protein was from a non-Lepidopteran: three of these from a mosquito and one even from a bacterium. Thus, the unannotated and often partial sequences of more closely related Lepidoptera acted as a bridge between the peptide fragments with which they shared high similarity over a low coverage and more distantly related sequences from other insects where a wider coverage was required to detect the lower sequence similarity.

Finally, for a third round of identification, predicted proteins from an in-house *H. armigera* larval midgut EST database were searched using MS BLAST (Table 3). Clones corresponding to hits were recovered from the normalized cDNA library and sequenced. Our final goal was to link each protein spot we analyzed to a full-length *H. armigera* cDNA clone for future experimental studies. We annotated 16 new *H. armigera* proteins according to sequence similarity and released them to GenBank (Table 3). As expected, most are enzymes responsible for the digestion of carbohydrates, proteins, and lipids. However, we also discovered proteins that we did not expect to be found in the gut lumen.

Digestive Enzymes. We identified three major carbohydrate digestive enzymes. Two highly abundant α-amylases were found, each represented by series of spots (4, 5, and 6; 2 and 3) on the 2D-gels indicating proteins of the same molecular weight (54 and 56 kDa, respectively) and slightly different pIs (Figure 2b). Peptides in these spots showed perfect matches to full-length cDNA sequences from the *H. armigera* midgut library (GH13Amy-1 and GH13Amy-2). In addition, GH13Amy-1 also matched the 27 kDa spot #33 which is likely to be a degradation product. We also found, during the second round of identification using ButterflyBase, a less abundant protein (#7) which has similarities to bacterial β-fructosidases. A matching full-length sequence G32FruA-1 was also found in our *H. armigera* larval midgut EST database. This has high similarity to sequences from several bacteria but not the fully sequenced genomes of other insects such as *Drosophila melanogaster* or *Anopheles gambiae*. Thus G32FruA-1 may represent contamination of both the lumen proteome and the cDNA library by bacteria which are part of the midgut microbiota. Alternatively, it could represent a relatively recent gene transfer event, in which case the sequence should also be present in the *H. armigera* genome. Further analysis of genomic DNA is under way to test this hypothesis.

Table 1. Results of PLGS2.2 Searches Using MS/MS Spectra and MS BLAST Searches Using de Novo Peptide Sequences, Against the NCBI_insecta Database

spot	GenBank accession[a]	description[a]	organism[b]	predicted MW[c]	peptide hits PLGS 2.2[d]	PLGS2.2 score[e]	peptide hits MS BLAST[f]	MS BLAST score[g]
1	ni							
2	AAP97393	Alpha-amylase 2	*D. saccharalis*	56.3	2	12.174	5	_
3	AAP97393	Alpha-amylase 2	*D. saccharalis*	56.3	_	_	1	64
4	AAO13754	Alpha-amylase	*Spodoptera frugiperda*	56.4	4	11.9975	_	_
5	AAO13754	Alpha-amylase	*Spodoptera frugiperda*	56.4	6	12.1739	_	_
6	AAP97394	Alpha-amylase 3	*D. saccharalis*	56.2	3	12.1738	_	_
7	ni							
8	ni							
9	AAY46199	Peritrophic membrane chitin binding protein	*Trichoplusia ni*	43.4	_	_	1	83
10	AAY46199	Peritrophic membrane chitin binding protein	*Trichoplusia ni*	43.4	_	_	2	121
11	ni							
12	CAF25189	carboxypeptidase precursor	*Helicoverpa armigera*	47.9	3	12.1738	5	341
13	ni							
14	CAA72965	diverged serine protease	*Helicoverpa armigera*	27.5	4	12.1739	_	_
15	ni							
16	ni							
17	AAT95356	trypsin III precursor	*Sesamia nonagrioides*	nd	_	_	5	313
18	AAT95356	trypsin III precursor	*Sesamia nonagrioides*	nd	_	_	4	197
19	ni							
20	AB062103	P27K precursor	*Bombyx mori*	24.9	_	_	1	60
21	ni							
22	ni							
23	ABC86902	arginine kinase	*Blatella germanica*	39.8	4	12.2391	_	_
	AAZ08496	Gram negative bacteria binding protein 2	*Nasutitermes fumigatus*	41.5	_	_	5	300
24	ABI53568	arginine kinase	*Heterotrigona erythrogaster*	40	_	_	4	175
	AAZ08499	Gram negative bacteria binding protein 2	*Nasutitermes magnus*	41.5	_	_	4	174
25	AB026735	30kP protease A (fragment)	*Bombyx mori*	nd	_	_	3	143
26	AF045138	putative trypsin (fragment)	*Helicoverpa armigera*	nd	_	_	2	151
27	AF045138	putative trypsin (fragment)	*Helicoverpa armigera*	nd	_	_	2	143
28	CAA72953	diverged serine protease	*Helicoverpa armigera*	27.3	8	12.1738	_	_
29	CAA72953	diverged serine protease	*Helicoverpa armigera*	27.3	5	12.1739	_	_
30	AF045138	putative trypsin (fragment)	*Helicoverpa armigera*	nd	_	_	3	196
31	ni							
32	AM085495	carboxypeptidase B precursor	*Helicoverpa zea*	48.4	_	_	4	177
33	AAP92665	Alpha-amylase	*Diatraea saccharalis*	56.2	3	12.1738	_	_
34	ni							
35	ni							
36	AY251276	chymotrypsin precursor	*Spodoptera frugiperda*	30.7	_	_	3	156
37	CAA72962	Trypsin-like protease precursor	*Helicoverpa armigera*	26.9	7	12.1739	_	_
38	CAA72962	Trypsin-like protease precursor	*Helicoverpa armigera*	26.9	5	12.1739	_	_
39	CAA72962	Trypsin-like protease precursor	*Helicoverpa armigera*	26.9	3	11.9978	_	_
40	CAA72962	Trypsin-like protease precursor	*Helicoverpa armigera*	26.9	3	12.1739	_	_

[a] GenBank Accession number and description of best hit protein in NCBI_insecta by PLGS or MS BLAST (ni = not identified). [b] Species of best hit in NCBI_insecta. [c] Predicted molecular weight of best hit (kDa). [d] Number of peptides matching best hit in the PLGS search. [e] PLGS2.2 scoring (please refer to Material and Methods section). [f] Number of peptides matching best hit in MS BLAST search. [g] MS BLAST scoring (please refer to Material and Methods section).

Two different lipase-like sequences were identified, one from protein spots #15 and #16 and the other one from spot #35. They were first identified using ButterflyBase, and we subsequently found matching sequences (HaLix-3 and HaLix-5) in the *H. armigera* EST database. The peptides obtained from the two spots #15 and #16 (differing in pI but not molecular weight), when used to search the ButterflyBase and our EST database using MS BLAST, gave the same hits in both cases. The sequence coverage is not sufficient for concluding whether these two spots represent different post-translationally modified forms of the same protein or two different proteins, differing by only a few residues due to their origin from a recent gene duplication.

Two prominent spots represented metallo-carboxypeptidases. One of them (#12) has already been characterized from *H. armigera* (HaCA42, GenBank CAF25189) as showing specificity for C-terminal glutamate residues.[21] The full protein sequence has a predicted molecular weight of 46 kDa, but the active form after cleavage of the signal peptide and prodomain has a molecular weight of 35 kDa[21] as observed on our gels. The second (#32) matched a full-length sequence from the *H. armigera* midgut library Cpep-2 with the highest sequence

Table 2. Results of MS BLAST Searches Using de Novo Peptide Sequences Against the ButterflyBase Database and BLASTP Searches Using ButterflyBase Predicted Protein Sequences Against UniRef100

spot	ButterflyBase identification[a]	species[b]	length of protein (aa)[c]	peptide hits[d]	blastp vs UniRef100[e]	UniRef100 identification[f]	species[g]	*E* value[h]
1	HAP00010_1	*H. armigera*	253	4	Chlorophyllide A binding protein precursor	Q2WBZ0	*B. mori*	3.00E−46
2	HAP00587_1	*H. armigera*	174	2	Alpha-amylase 2	Q7YXJ4	*D. saccharalis*	3.00E−72
3	HAP00587_1	*H. armigera*	174	1	Alpha-amylase 2	Q7YXJ4	*D. saccharalis*	3.00E−72
4	HAP00151_1	*H. armigera*	220	4	Alpha-amylase	Q8IA46	*S. frugiperda*	4.00E−81
5	HAP00151_1	*H. armigera*	220	4	Alpha-amylase	Q8IA46	*S. frugiperda*	4.00E−81
6	HAP00151_1	*H. armigera*	220	4	Alpha-amylase	Q8IA46	*S. frugiperda*	4.00E−81
7	PIP00419_1	*P. interpunctella*	162	1	Beta-fructosidase FruA	Q8GM36	*B. megaterium*	5.00E−45
8	ni							
9	HAP00067_1	*H. armigera*	254	4	Peritrophic membrane chitin binding protein	Q3B9L9	*T. ni*	5.00E−81
10	HAP00067_1	*H. armigera*	254	2	Peritrophic membrane chitin binding protein	Q3B9L9	*T. ni*	5.00E−81
11	ni							
13	ni							
15	HAP00113_2	*H. armigera*	182	5	Lipase	Q173Q6	*A. aegypti*	4.00E−29
16	HAP00113_2	*H. armigera*	182	4	Lipase	Q173Q6	*A. aegypti*	4.00E−29
17	HAP00411_1	*H. armigera*	263	6	Trypsin AiT6	Q9NB92	*A. ipsilon*	1.00E−129
18	HAP00411_1	*H. armigera*	263	3	Trypsin AiT6	Q9NB92	*A. ipsilon*	1.00E−129
19	ni							
20	SFP03734_1	*S. frugiperda*	233	1	P27K precursor	Q8T113	*B. mori*	3.00E−68
21	ni							
22	ni							
23	HAP00343_1	*H. armigera*	356	1	Arginine kinase	Q95PM9	*P. interpunctella*	0
	HAP00503_1	*H. armigera*	304	4	Gram negative bacteria binding protein 2	Q2N3Y1	*N. exitiosus*	1.00E−147
24	SFP00297_3	*S. frugiperda*	375	5	Gram negative bacteria binding protein 2	Q2N3Y1	*N. exitiosus*	1.00E−147
25	HAP00073_1	*H. armigera*	282	4	30kP protease A	Q9XY10	*B. mori*	4.00E−58
31	ni							
32	ni							
33	HAP00151_1	*H. armigera*	220	4	Alpha-amylase	Q8IA46	*S. frugiperda*	4.00E−81
34	ni							
35	HAP00417_1	*S. frugiperda*	331	3	Lipase	Q173Q0	*A. aegypti*	1.00E−36

[a] Cluster ID of best hit in ButterflyBase by MS BLAST (ni = not identified). [b] Species of best hit in ButterflyBase. [c] Length (number of amino acids) of predicted ButterflyBase protein. [d] Number of peptides matching best hit in ButterflyBase in MS BLAST search. [e] Result of blastp search using ButterflyBase predicted protein against UniRef100. [f] UniRef100 Accession number. [g] Species of best hit in UniRef100. [h] *E*-value of best hit in blastp search against UniRef100.

similarity to a carboxypeptidase B from *Helicoverpa zea* resistant to the potato carboxypeptidase inhibitor that has been crystallized.[22] The apparent amounts of those two proteins on the 2D-gels confirmed previous findings that carboxypeptidase A activity predominates in the gut lumen of *H. armigera* larvae, with only a low level of carboxypeptidase B activity present.[23]

Fifteen protein spots (13, 14, 17, 18, 25–30, 36–40) corresponded to at least nine different serine proteases, many previously described from *H. armigera* (Tables 1 and 3). Spot #36 was very similar to a chymotrypsin from *Spodoptera frugiperda* (Table 2), with no corresponding sequence from *H. armigera* in GenBank or our libraries. We did not detect any of the 14 previously described chymotrypsins[24] from *H. armigera*, indicating that they are not among the most abundant proteins in larvae fed the chemically defined diet. Three identified sequences (SerProx-2, -4, and -6) are serine proteases that cannot readily be classified as chymotrypsins or trypsins. The majority of spots matched sequences similar to insect trypsins. In many cases, two or three spots (usually close together on the gel but sometimes widely separated) hit the same sequence and are grouped together in Table 3. Because of incomplete peptide coverage, we could not determine whether amino acid differences or post-translational modifications were responsible for the differing mobilities on the gel. Even the cleavage of the prodomain to produce the catalytically active enzyme[25] may occur at different positions, depending on the cleavage specificities of different activating proteases, resulting in a series of spots corresponding to different cleavage products of the same translated polypeptide. Moreover, many of the previously described sequences are so similar that they cannot be distinguished on the basis of the limited peptide sequence available, especially with the large cluster corresponding to spots 37–39. This reflects the well-established fact that serine proteases are clustered in large gene families in Lepidoptera.[24,26,27]

Peritrophic Matrix Binding Protein. Peptides from protein spots #9 and #10 were similar to TnPM-P42, a chitin binding protein originally purified from the PM of *Trichoplusia ni* larvae.[28] Unlike most PM proteins, this one lacks the typical peritrophin chitin-binding domain but instead possesses a chitin deacetylase domain. This was hypothesized to be involved in the strong chitin binding of recombinant TnPM-P42 observed by the authors, leading them to classify it as a type 3 PM protein.[28] However, an antibody to TnPM-P42 detected it not only in the PM but also abundantly in regurgitated midgut fluid of *T. ni*, showing the existence of a

Table 3. Results of MS BLAST Searches Using de Novo Peptide Sequences Against *H. armigera* ESTs and cDNA Sequences and BLASTP Searches Using *H. armigera* Protein Sequences Predicted from cDNA Against UniRef100

spot	*H. armigera* assigned name	GenBank[a]	hits[b]	aa[c]	MW[d]	pI[e]	blastp vs UniRef100[f]	UniRef100[g]	species[h]	*E* value[i]
1	multidomain lipocalin pentacalin-1	EF600047*	9	927	101.5	4.5	Chlorophyllide A binding protein precursor	Q2WBZ0	*B. mori*	0
2,3	alpha-amylase GH13Amy-2	EF600048*	6	435#	49.1	5.8	α-amylase 2	Q7YXJ4	*D. saccharalis*	0
4,5,6,33	alpha-amylase GH13Amy-1	EF600049*	5	500	56	6.7	α-amylase	Q8IA46	*S. frugiperda*	0
7	fructosidase GH32FruA-1	EF600050*	5	479	54	4.8	Beta-fructosidase FruA	Q8GM36	*B. megaterium*	5.00E−83
9,10	Chitin binding PM protein ChitDeac-1	EF600051*	5	390	43.5	4.6	Peritrophic membrane chitin binding protein	Q3B9L9	*T. ni*	1.00E−156
11	protein of unknown function HaPUF-1	EF600052*	2	259#	29	4.2	no significant hit found			
12	carboxypeptidase Cpep-1	CAF25189	7	424	47.9	5.2	carboxypeptidase precursor	Q6H962	*H. armigera*	0
13	protease SerProx-1	EF600053*	3	261	27.4	4.9	Trypsin-like proteinase T23	Q6R560	*O. nubilalis*	4.00E−77
14	protease SerProx-2	CAA72965	4	256	27.5	4.35	diverged serine protease	O18449	*H. armigera*	0
15,16	Lipase HaLix-3	EF600061*	7	292#	30.5	5.1	Lipase	Q173Q0	*A. aegypti*	1.00E−40
17,18	protease SerProx-3	EF600054*	6	263	28	5.4	Trypsin AiT6	Q9NB92	*A. ipsilon*	1.00E−129
20	protein of unknown function HaPUF-2	EF600055*	1	230	25.5	5.1	p27K precursor	Q8T113	*B. mori*	3.00E−71
23,24	arginine kinase ArgK-1	EF600057*	5	355	40	5.9	arginine kinase	Q95PM9	*P. interpunctella*	0
23,24	beta-1,3-glucan recognition protein GH16betaGRP-1	EF600056*	6	375	41.8	6.3	Gram negative bacteria binding protein 2	Q2N3Y1	*N. exitiosus*	5.00E−163
25	protease SerProx-4	EF600058*	5	353#	37.5	3.8	30kP protease A	Q9XY10	*B. mori*	5.00E−87
26,27,30	protease SerProx-5	EF600059*	4	260	27.6	6.2	Trypsin Hz8	Q9NB81	*H. zea*	1.00E−122
28,29	protease SerProx-6	CAA72953	5	256	27.3	5.65	diverged serine protease	O18439	*H. armigera*	0
32	carboxypeptidase Cpep-2	EF600060*	4	428	48.4	5.9	Carboxypeptidase B precursor	Q3T905	*H. zea*	1.00E−167
35	Lipase HaLix-5	EF600062*	3	332#	35.58	8.2	Lipase	Q173Q1	*A. aegypti*	1.00E−49
37,38 39,40	protease SerProx-7	CAA72962	7	254	26.9	10.8	Trypsin-like protease	O18447	*H. armigera*	0

[a] GenBank Accession Number (* from the present study). [b] Number of peptides matching the target in the MS BLAST search. [c] Length (amino acids) of predicted *H. armigera* protein (# 5′ truncated). [d] Predicted molecular weight of *H. armigera* protein (kDa). [e] Predicted pI of *H. armigera* protein. [f] Result of blastp search using *H. armigera* predicted protein against UniRef100. [g] UniRef100 Accession Number. [h] Species of best hit in UniRef100. [i] *E*-value of best hit in blastp search against UniRef100.

soluble form in that species. Peptides from spots #9 and #10 showed a perfect match to a predicted protein sequence Ha-ChitDeac-1 from our *H. armigera* midgut cDNA library, which in turn shows an overall amino acid identity of 65% to TnPM-P42 as well as a predicted signal peptide. Our preparative methods would not have released a type 3 PM protein from the peritrophic matrix (and in fact none of the previously characterized type 3 peritrophin-type PM proteins were found in this study), but there may be a soluble isoform of Ha-ChitDeac-1 that would have been recovered from the lumen or by washing the PM with PBS. It should be noted that the antibody of Guo et al.[28] did detect the PM-bound form in the closely related species *Heliothis virescens* and *H. zea* but not a soluble form in the regurgitated midgut fluid.

Lipocalin-Like Protein. Peptides from protein spot #1 were very similar to the *Bombyx mori* multidomain-lipocalin chlorophyllide A-binding protein.[29] Spot #1 has a molecular weight between 100 and 150 kDa in our 2D-gels which is consistent with the molecular weight (around 105 kDa) of the soluble form purified from the *B. mori* midgut soluble protein fraction.[29] The molecular weight of the full-length protein in *B. mori*, however, was estimated to be 302 kDa and contains 15 predicted lipocalin domains. The authors suggested that the 105 kDa polypeptide they isolated was a degraded form of the full-length protein.[29] When we sequenced the full-length cDNA corresponding to the protein in *H. armigera*, we obtained a predicted protein of 927 amino acids (Table 3) with a predicted molecular weight of 101.5 kDa. We named this protein pentacalin on the basis of its five lipocalin-like domains. We found an amino-terminal signal peptide consisting of 20 amino acids using SignalP 3.0[30] and a predicted carboxy-terminal GPI-modification site using the big-PI predictor[31] with a cleavage site after Ser905. We also found a predicted GPI-modification site on the *B. mori* protein. More study in both species is required in order to determine whether they are both modified and thus anchored to the plasma membrane of midgut cells via a GPI moiety and by what process they could be found as soluble forms in the gut lumen. It is also possible that this protein, with a molecular weight above 100 kDa, is not in the endoperitrophic space but instead has become trapped in the peritrophic matrix which appears to exclude most other highly abundant proteins larger than 65 kDa.

Human-Allergenic Enzyme. Two neighboring spots on our 2D-gels (spots #23 and 24) both showed evidence of peptides from two distinct proteins: an arginine kinase and a glucan recognition protein. The first protein was identified as a homologue of a catalytically active arginine kinase from the Indianmeal moth *Plodia interpunctella*.[32] A full-length clone from the *H. armigera* library yielded a protein sequence (HaArgK-1) with 94% amino acid identity with the Indianmeal moth enzyme. The abundance of this protein in the gut lumen was surprising because arginine kinases affect the intracellular balance between ADP and ATP and are typically active within the cell, for example, in mitochondria in the tobacco hornworm *Manduca sexta*[33] or in the cytosol fraction of flagellated trypanosomatids.[34] Moreover, we found no predicted N-terminal signal peptide on any of the available insect arginine kinase sequences including HaArgK-1, suggesting that this protein would not be secreted by the classical pathway. The enzyme from the Indianmeal moth was discovered because some people are extremely allergic to it. The enzyme reacts with serum IgE from about 12% of patients suffering from allergic symptoms indoors.[32] Since the Indianmeal moth is a common pest of stored grain and flour, human contact with arginine kinase excreted in the frass in contaminated food is a likely source of exposure, especially if high concentrations of

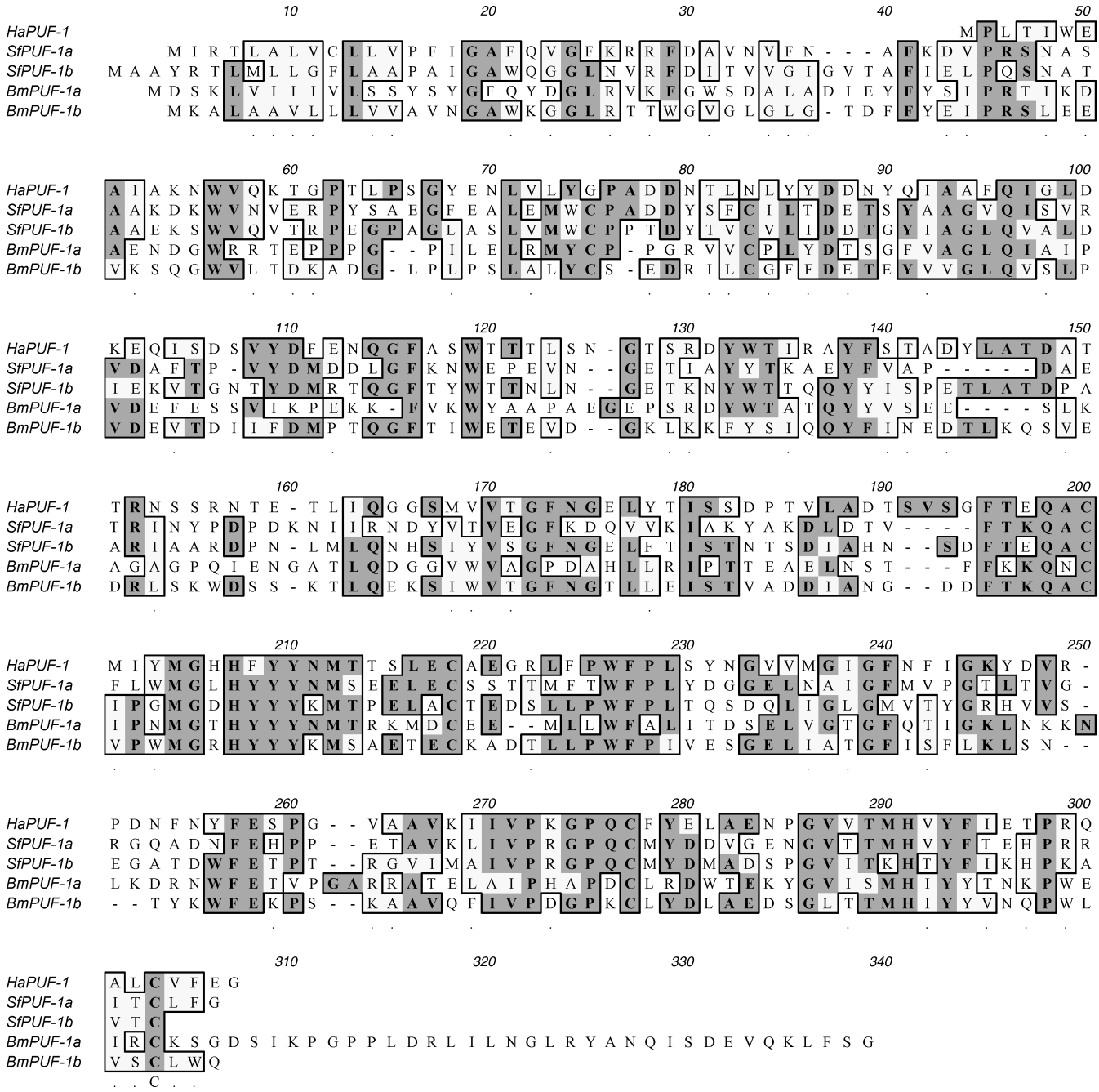

Figure 3. Alignment of HaPUF-1 (protein of unknown function) found in the lumen with sequences predicted from ESTs from *Bombyx mori* (BmPUF-1a and -1b) and *Spodoptera frugiperda* (SfPUF-1a and -1b). Identical residues are boxed with dark shading, and chemically similar residues are boxed with light shading.

the enzyme are present in the insect midgut lumen. A high degree of stability, both in the lumen and in the environment, would be required for such potent antigenicity. The reasons for its presence in the lumen in various Lepidopteran species as well as the means of its transport there remain a mystery.

Protein Implicated in Insect Immune Response. The second distinct protein detected using de novo sequenced peptides from spots #23 and #24 is similar to proteins implicated in the insect immune system's recognition of bacterial pathogens. These peptides matched a full-length clone from the *H. armigera* library encoding a protein we have named GH16betaGRP-1. This protein has 60% amino acid identity with proteins from several termite and mosquito species that have been annotated as Gram negative bacteria binding proteins (GNBPs). Those proteins are members of the so-called pattern recognition receptors (PRRs) well studied in insect hemolymph.[35] GNBPs, together with βGRPs (β-1,3-glucan recognition proteins), have a strong affinity for β-1,3-glucans which are major components of fungal and bacterial cell walls.[35] They contain a conserved carboxyl-terminal glucanase-like domain and a less conserved amino-terminal domain[35] and are members of Family 16 of the glycoside hydrolases. Several previously isolated proteins from Lepidoptera belong to this family (Carbohydrate-Active EnZymes "CAZy" database, http://afmb.cnrs-mrs.fr/CAZY/).[36] These include two *B. mori* proteins, p50,[37] and a β-1,3-glucan recognition protein,[38] each with about 30% amino acid identity with GH16betaGRP-1. Both of them are involved in the recognition of Gram negative and positive bacteria and yeast

and activation of the prophenoloxidase cascade. Two proteins from *M. sexta*, the β-1,3-glucan recognition proteins-1 and -2, play a similar role.[39,40] These proteins had all been purified and characterized from larval or pupal hemolymph. To date, we have found no example in the literature of such proteins present in the gut lumen. Although a β-1,3-glucanase could potentially play a role in digestion of plant and fungal cell walls, this enzyme activity has not been reported from lepidopteran midguts. Further biochemical characterization will more directly address the possibility of an immune-related function of this protein in the gut lumen.

Two Lepidopteran-Specific Proteins of Unknown Function. Two peptides from spot #11 produced perfect matches to a full-length sequence from an *H. armigera* library from integument with predicted protein sequence HaPUF-1 (*H. armigera* protein of unknown function 1). This has no recognizable similarity to any previously described protein but shows 21−35% amino acid identity to protein sequences predicted from ESTs from two other Lepidoptera, *B. mori* and *S. frugiperda* (Figure 3). ESTs corresponding to BmPUF-1a come from the pupal or adult brain, antenna, or embryo; those corresponding to BmPUF-1b come from the middle silkgland or larval epidermis; and all *Spodoptera* ESTs are from a larval fat body library. Thus, HaPUF-1 appears to be a member of a small, rapidly evolving family of proteins expressed in a variety of lepidopteran tissues, and this is the first evidence of occurrence in the lumen.

A peptide from spot #20 produced a perfect match to a portion of a full-length predicted protein sequence from a larval integument *H. armigera* library HaPUF-2, another protein of unknown function. This in turn has 52% amino acid identity to a protein purified from the hemolymph of *M. sexta* of unknown function[41] and 54% identity to a protein originally purified from the subesophageal gland of *B. mori* and subsequently found in fat body, hemocytes, midgut, and Malpighian tubules, but not the silk gland.[42] A protein from *Galleria melonella* hemolymph with 57% identity resulted in increased phenoloxidase activity when added to hemolymph plasma in vitro.[43] Most ESTs from *S. frugiperda* coding for a protein with 65% sequence identity come from the fat body, with a few from hemocytes and the midgut. Homologous sequences in the Diptera could also be identified by BLAST searches, including three predicted proteins of unknown function from *D. melanogaster*, CG14629-PA (38% sequence identity), CG9917-PA (33%), and CG11378-PA (27%), and one protein from *A. gambiae* (HG1 at 39%) discovered during a proteomic survey of hemolymph.[44] All these proteins are classified under InterPro domain IPR009832 (Protein of unknown function DUF1397). Spot #20 represents a very abundant protein of about 16 kDa, while the predicted size from the full-length cDNA is 25.5 kDa, raising the possibility that the form found in the lumen is a degradation product of a protein normally most abundant in the hemolymph, fat body, and other nondigestive tissues.

Conclusion

This study illustrates the utility of the proteomic approach in probing the multiple functions of a large and important intraorganismal extracellular space of insects, the midgut lumen. The complement of proteins persisting and functioning in this harsh environment represents an equilibrium between secretion versus elimination by proteolysis and bulk flow. Our initial study has been limited to a snapshot at a single point in time under somewhat artificial conditions. Further work will be required to quantify secretion rates and half-lives of proteins, to understand the dynamic processes responsible for the observed steady state and how it is shifted by the presence of plant-derived nutrients and secondary compounds.

The most useful search tool we employed in this study was MS BLAST.[17] In searches of the same target database, NCBI_insecta, we consistently recovered more hits with MS BLAST than with the proprietary software of a leading manufacturer with more stringent matching conditions. Guidance in discovering homologues in other Lepidoptera was also greatly facilitated by the availability of an MS BLAST server installed on ButterflyBase, and we encourage the same practice be adopted by those responsible for Web sites devoted to specific groups of organisms. Although a quantitative comparison of different search engines is not in the scope of our inquiry here, we feel that this search tool offers distinct advantages in identifying distantly related or rapidly evolving homologues of proteins from nonmodel organisms, identified by de novo sequencing of peptides. In this study, we were able to validate these identifications with full-length cDNA sequences from the same species and tissues, confirming the utility of the approach.

Acknowledgment. We thank Dave Murray (Queensland Department of Primary Industries) for the initial collection of insects, Chris D. Jiggins (University of Cambridge) and Alexie Papanicolaou for installing an MS BLAST interface on ButterflyBase, Heiko Vogel for searching his *H. armigera* larval midgut EST database and for sequencing the positive clones, Zhudong Liu for providing *H. armigera* neonates, and Bianca Ulitzsch for excellent technical assistance. Financial support was provided by the Max-Planck-Gesellschaft.

Supporting Information Available: (1) Ingredients for artificial diet. (2) MALDI peptide fingerprint query results. (3) MS/MS spectrum search results. (4) MS BLAST query results. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Chapman, R. F. In *Comprehensive Insect Physiology, Biochemistry and Pharmacology*; Kerkut, G. A., Gilbert, L. I., Eds.; Pergamon Press: Oxford, 1985; Vol. 4, pp 87–164..
(2) Wieczorek, H.; Gruber, G.; Harvey, W. R.; Huss, M.; Merzendorfer, H.; Zeiske, W. *J. Exp. Biol.* **2000**, *203*, 127–135.
(3) Giordana, B.; Sacchi, V. F.; Parenti, P.; Hanozet, G. M. *Am. J. Physiol.* **1989**, *257*, R494–R500.
(4) Terra, W. R. *Arch. Insect Biochem. Physiol.* **2001**, *47*, 47–61.
(5) Tellam, R. L.; Wijffels, G.; Willadsen, P. *Insect Biochem. Mol. Biol.* **1999**, *29*, 87–101.
(6) Terra, W. R. *Ann. Rev. Entomol.* **1990**, *35*, 181–200.
(7) Terra, W. R.; Ferreira, C. *Comp. Biochem. Physiol., B: Biochem. Mol. Biol.* **1994**, *109*, 1–62.
(8) Andreadis, T. G. In *Epizootiology of insect diseases*; Fuxa, J. R., Tanada, Y., Eds.; Wiley: New York, 1987, pp 159–176..
(9) Schnepf, E.; Crickmore, N.; Van Rie, J.; Lereclus, D.; Baum, J.; Feitelson, J.; Zeigler, D. R.; Dean, D. H. *Microbiol. Mol. Biol. Rev.* **1998**, *62*, 775–806.
(10) Siva-Jothy, M. T.; Moret, Y.; Rolff, J. *Adv. Insect Physiol.* **2005**, *32*, p 1–48.
(11) Chen, H.; Wilkerson, C. G.; Kuchar, J. A.; Phinney, B. S.; Howe, G. A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 19237–42.
(12) Felton, G. W. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 18771–2.
(13) Shevchenko, A.; de Sousa, M. M.; Waridel, P.; Bittencourt, S. T.; de Sousa, M. V.; Shevchenko, A. *J. Proteome Res.* **2005**, *4*, 862–9.
(14) Vanderzant, E. S. *Ann. Entomol. Soc. Am.* **1968**, *61*, 120–125.
(15) Giri, A. P.; Wunsche, H.; Mitra, S.; Zavala, J. A.; Muck, A.; Svatos, A.; Baldwin, I. T. *Plant Physiol.* **2006**, *142*, 1621–41.
(16) Skilling, J.; Denny, R.; Richardson, K.; Young, P.; McKenna, T.; Campuzano, I.; Ritchie, M. *Comp. Funct. Genomics* **2004**, *5*, 61–68.
(17) Shevchenko, A.; Sunyaev, S.; Loboda, A.; Bork, P.; Ens, W.; Standing, K. G. *Anal. Chem.* **2001**, *73*, 1917–26.

(18) Habermann, B.; Oegema, J.; Sunyaev, S.; Shevchenko, A. *Mol. Cell. Proteomics* **2004**, *3*, 238–249.
(19) Papanicolaou, A.; Joron, M.; McMillan, W. O.; Blaxter, M. L.; Jiggins, C. D. *Mol. Ecol.* **2005**, *14*, 2883–97.
(20) Chougule, N. P.; Giri, A. P.; Sainani, M. N.; Gupta, V. S. *Insect Biochem. Mol. Biol.* **2005**, *35*, 355–67.
(21) Bown, D. P.; Gatehouse, J. A. *Eur. J. Biochem.* **2004**, *271*, 2000–2011.
(22) Bayés, A.; Comellas-Bigler, M.; Rodríguez de la Vega, M.; Maskos, K.; Bode, W.; Aviles, F. X.; Jongsma, M. A.; Beekwilder, J.; Vendrell, J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 16602–16607.
(23) Bown, D. P.; Wilkinson, H. S.; Gatehouse, J. A. *Insect Biochem. Mol. Biol.* **1998**, *28*, 739–49.
(24) Bown, D. P.; Wilkinson, H. S.; Gatehouse, J. A. *Insect Biochem. Mol. Biol.* **1997**, *27*, 625–38.
(25) Bayes, A.; Sonnenschein, A.; Daura, X.; Vendrell, J.; Aviles, F. X. *Eur. J. Biochem.* **2003**, *270*, 3026–35.
(26) Mazumdar-Leighton, S.; Broadway, R. M. *Insect Biochem. Mol. Biol.* **2001**, *31*, 645–57.
(27) Mazumdar-Leighton, S.; Broadway, R. M. *Insect Biochem. Mol. Biol.* **2001**, *31*, 633–44.
(28) Guo, W.; Li, G.; Pang, Y.; Wang, P. *Insect Biochem. Mol. Biol.* **2005**, *35*, 1224–34.
(29) Mauchamp, B.; Royer, C.; Garel, A.; Jalabert, A.; Da Rocha, M.; Grenier, A. M.; Labas, V.; Vinh, J.; Mita, K.; Kadono, K.; Chavancy, G. *Insect Biochem. Mol. Biol.* **2006**, *36*, 623–33.
(30) Bendtsen, J. D.; Nielsen, H.; von Heijne, G.; Brunak, S. *J. Mol. Biol.* **2004**, *340*, 783–95.
(31) Eisenhaber, B.; Bork, P.; Eisenhaber, F. *J. Mol. Biol.* **1999**, *292*, 741–58.
(32) Binder, M.; Mahler, V.; Hayek, B.; Sperr, W. R.; Scholler, M.; Prozell, S.; Wiedermann, G.; Valent, P.; Valenta, R.; Duchene, M. *J. Immunol.* **2001**, *167*, 5470–7.
(33) Chamberlin, M. *J. Exp. Biol.* **1997**, *200*, 2789–96.
(34) Pereira, C. A.; Alonso, G. D.; Torres, H. N.; Flawia, M. M. *J. Eukaryotic Microbiol.* **2002**, *49*, 82–5.
(35) Kanost, M. R.; Jiang, H.; Yu, X. Q. *Immunol. Rev.* **2004**, *198*, 97–105.
(36) Henrissat, B.; Davies, G. *Curr. Opin. Struct. Biol.* **1997**, *7*, 637–44.
(37) Lee, W. J.; Lee, J. D.; Kravchenko, V. V.; Ulevitch, R. J.; Brey, P. T. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 7888–93.
(38) Ochiai, M.; Ashida, M. *J. Biol. Chem.* **1988**, *263*, 12056–62.
(39) Jiang, H.; Ma, C.; Lu, Z. Q.; Kanost, M. R. *Insect Biochem. Mol. Biol.* **2004**, *34*, 89–100.
(40) Ma, C.; Kanost, M. R. *J. Biol. Chem.* **2000**, *275*, 7505–14.
(41) Samaraweera, P.; Law, J. H. *Insect Mol. Biol.* **1995**, *4*, 7–13.
(42) Tan, A.; Tanaka, H.; Sato, N.; Yaguchi, M.; Nagata, M.; Suzuki, K. *J. Insect Biotechnol. Sericol.* **2003**, *72*, 41–50.
(43) Park, S. Y.; Kim, C. H.; Jeong, W. H.; Lee, J. H.; Seo, S. J.; Han, Y. S.; Lee, I. H. *Dev. Comp. Immunol.* **2005**, *29*, 43–51.
(44) Paskewitz, S. M.; Shi, L. *Insect Biochem. Mol. Biol.* **2005**, *35*, 815–824.

PR7006208