Pergamon

Insect Biochemistry and Molecular Biology 32 (2002) 1101–1122

Insect Biochemistry and Molecular Biology

www.elsevier.com/locate/ibmb

Toward a description of the sialome of the adult female mosquito *Aedes aegypti*

J.G. Valenzuela [a], V.M. Pham [a], M.K. Garfield [b], I.M.B. Francischetti [a], J.M.C. Ribeiro [a,*]

[a] *Medical Entomology Section, Laboratory of Parasitic Diseases, NIAID, National Institutes of Health, 4 Center Drive, Room 4/126, Bethesda, MD 20892-0425, USA*

[b] *Research Technologies Branch, National Institute of Allergy and Infectious Diseases, Rockville, MD 20852, USA*

Received 6 November 2001; received in revised form 11 February 2002; accepted 15 February 2002

Abstract

To describe the set of mRNA and protein expressed in the salivary glands (sialome) of *Aedes aegypti* mosquitoes, we randomly sequenced a full-length cDNA library of this insect and performed Edman degradation of PVDF-transferred protein bands from salivary homogenates. We found 238 cDNA clusters which contained those coding for 10 of the 11 proteins found by aminoterminal degradation. All six previously described salivary proteins were found in this library. Full-length sequences of 32 novel cDNA sequences are reported, one of which is the product of a transposable element. Among the 31 novel protein sequences are 4 additional members of the D7 protein family; 4 novel members of the antigen 5 family (a protein family not reported in *Aedes*); a novel serpin; a novel member of the 30-kDa allergen of *Ae. Aegypti*; a secreted calreticulin; 2 proteins similar to mammalian angiopoietins; adenosine deaminase; purine hydrolase; lysozyme; a C-type lectin; 3 serine proteases, including one with high similarity to *Bombyx* prophenoloxidase activating enzyme; 2 proteins related to invertebrate immunity; and several sequences that have no significant matches to known proteins. The possible role of these proteins in blood and sugar feeding by the mosquito is discussed. Published by Elsevier Science Ltd.

Keywords: Salivary glands; Proteome; Electrophoresis; Hematophagy

1. Introduction

Saliva of blood-sucking arthropods contains substances that counteract vertebrate host hemostasis. Accordingly, at least one anticlotting, one antiplatelet, and one vasodilator substance usually are found in the salivary glands of such animals. These substances presumably increase arthropod fitness by increasing the speed at which blood is found and imbibed and by decreasing the possibility of being killed by the host during feeding (Ribeiro, 1987; Ribeiro, 1995). These salivary components may also affect parasite transmission by arthropod vectors and thus may serve as vaccine targets against these diseases (Valenzuela et al., 2001). Mosquito salivary glands also contain enzymes associated with sugar feeding (Marinotti et al., 1990; Grossman and James, 1993) as well as lysozyme, which may help to control bacterial growth in the sugar meal while stored in the mosquito crop (Rossignol and Lueders, 1986). Salivary glands of mosquitoes have usually less than 3 μg of protein (Ribeiro et al., 2001). Identification and characterization of such compounds usually is accomplished by tedious accumulations of over 1000 pairs of dissected salivary glands as starting material for fractionation and/or bioassay experiments.

The protein composition of the salivary glands of blood-sucking diptera is not very complex. Usually, less than 20 main protein bands are seen by sodium dodecylsulphate polyacrylamide gel electrophoresis (SDS-PAGE) (Racioppi and Spielman, 1987, Mellink and van Zeben (1976); Poehling (1979). Recent advances in proteomics research indicate that, in general, there is no significant correlation between the amount of cellular mRNA coding for a particular protein and the amount of the protein (Futcher et al., 1999; Gygi et al., 1999; Gygi et al., 2000). On the other hand, a significant correlation exists between amounts of RNA and the coded

* Corresponding author. Tel.: +1-301-496-3066; fax: +1-301-402-4941.

E-mail address: jribeiro@NIH.gov (J.M.C. Ribeiro).

0965-1748/02/$ - see front matter Published by Elsevier Science Ltd.
PII: S0965-1748(02)00047-4

protein only when the 12 most abundant yeast proteins are considered (Gygi et al., 1999). If this correlation holds true for mosquito salivary glands, random sequencing of a non-normalized, non-subtracted organ library should yield information of the most abundant proteins found in the insect salivary proteome.

We report the results following sequencing of 456 clones from a full-length, PCR-constructed, cDNA library from adult female *Aedes aegypti*. The resulting database allowed identification of the cDNA clone corresponding to 10 of 11 salivary gland proteins isolated by SDS-PAGE that yielded aminoterminal sequencing by Edman's degradation. Furthermore, 227 novel partial sequences from *Ae. aegypti* are described, including full-length sequence information for 32 of these. Interestingly, we cannot ascribe functions to the majority of these new sequences, indicating that these salivary proteins may have functions not related directly to hemostasis or sugar feeding, or, alternatively, that they may represent novel molecule classes acting on hemostasis, sugar digestion, or antibacterial activity. Finally, this paper contributes to defining the sialome (the set of RNA messages + proteins expressed in the salivary glands) from females of the mosquito, *Ae. aegypti*.

2. Materials and methods

2.1. Reagents

All water used was of 18 MΩ quality and was produced using a MilliQ apparatus (Millipore, Bedford, MA, USA). Organic compounds were obtained from Sigma Chemical Corporation (St. Louis, MO, USA) or as stated otherwise.

2.2. Mosquitoes

The Liverpool/black eye strain of *Ae. aegypti* was reared under the expert supervision of Mr. André Laughinghouse. Insectary rooms were kept at 26±0.5°C, with a relative humidity of 70% and a 16 h:8 h light:dark photoperiod. Adult female mosquitoes used in the experiments were aged 0–7 days, took no blood meals and were maintained on a diet of 10% Karo syrup solution. Salivary glands from adult female mosquitoes were dissected and transferred to 20 μl Hepes saline (HS; NaCl 0.15 mol l^{-1}, 10 mmol l^{-1} Hepes, pH 7.0) in 1.5 ml polypropylene vials in groups of 20 pairs of glands in 20 μl of HS or as individual glands in 10 μl of HS. Salivary glands were kept at −75°C until needed.

2.3. Salivary gland cDNA library construction

Ae. aegypti salivary gland mRNA was isolated from 80 salivary gland pairs from adult females, at days 1 and 2 after emergence, using the Micro-FastTrack mRNA isolation kit (Invitrogen, San Diego, CA, USA). The PCR-based cDNA library was made following the instructions for the SMART cDNA library construction kit (Clontech, Palo Alto, CA, USA). Two hundred nanograms of *Aedes aegypti* salivary gland mRNA were reverse transcribed to cDNA using Superscript II RNase H- reverse transcriptase (Gibco-BRL, Gaithersburg, MD, USA) and the CDS III/3′ PCR primer (Clontech) for 1 h at 42°C. Second strand synthesis was performed through a PCR-based protocol using the SMART III primer (Clontech) as the sense primer and the CDSIII/3′ primer as antisense primer. These two primers create *Sfi*IA and B sites at the ends of the nascent cDNA. Double-strand cDNA synthesis was carried out on a Perkin Elmer 9700 Thermal cycler (Perkin Elmer Corp., Foster City, CA, USA) and using the Advantage Klen-*Taq* DNA polymerase (Clontech). PCR conditions were the following: 94°C for 2 min; 19 cycles of 94°C for 10 s and 68°C for 6 min. Double-strand cDNA was immediately treated with proteinase K (0.8 μg/μl) for 20 min at 45°C and washed three times with water using Amicon filters with a 100 kDa cut off (Millipore Corp., Bedford MA, USA). The double-strand cDNA was then digested with *Sfi*I for 2 h at 50°C. The cDNA was then fractionated using columns provided by the manufacturer (Clontech). Fractions containing cDNA of more than 400 bp were pooled, concentrated, and washed three times with water using an Amicon filter with a 100 kDa cut off. The cDNA was concentrated to a volume of 7 μl. The concentrated cDNA was then ligated into a Lambda TriplEx2 vector (Clontech), and the resulting ligation reaction was packed using the Gigapack Gold III from Stratagene/Biocrest (Cedar Creek, TN, USA), following manufacturer's specifications. The obtained library was plated by infecting log-phase XL1- Blue cells (Clontech) and the amount of recombinants determined by PCR using vector primers flanking the inserted cDNA and visualized on a 1.1% agarose gel with ethidium bromide (1.5 μg/ml).

2.4. Sequencing of Ae. aegypti cDNA library

The *Ae. aegypti* salivary gland cDNA library was plated to approximately 200 plaques per plate (150 mm Petri dish). The plaques were randomly selected and transferred to a 96-well polypropylene plate containing 100 μl of water per well. The plate was covered and placed on a gyratory shaker for 1 h at room temperature. Five microliters of the phage sample were used as a template for a PCR reaction to amplify random cDNAs. The primers used for this reaction were sequences from the TriplEx2 vector and were named PT2F1 (5′-AAG TAC TCT AGC AAT TGT GAG C-3′), which is positioned upstream of the cDNA of interest (5′ end), and PT2R1 (5′-CTC TTC GCT ATT ACG CCA GCT G-3′) which

is positioned downstream of the cDNA of interest (3′ end). Platinum *Taq* polymerase (Gibco-BRL) was used for these reactions. Amplification conditions were: 1 hold of 75°C for 3 min, 1 hold of 94°C for 2 min and 33 cycles of 94°C for 1 min, 49°C for 1 min, and 72°C for 1 min and 20 s. Amplified products were visualized on a 1.1% agarose gel with ethidium bromide. The concentration of double-strand cDNA was measured by using the Hoechst dye 33258 on a Flurolite 1000 plate fluorometer (Dynatech Laboratories, Chantilly, VA, USA). Three to four microliters of PCR reaction containing between 100 to 200 ng of DNA were then treated with Exonuclease I (0.5 units μl^{-1}) and shrimp alkaline phosphatase (0.1 units μl^{-1}) for 15 min at 37°C and 15 min at 80°C on a 96-well PCR plate. This mixture was used as a template for a cycle sequencing reaction using the DTCS labeling kit from Beckman Coulter Inc. (Fullerton, CA, USA). The primer used for sequencing (PT2F3), is upstream of the inserted cDNA and downstream of the primer PT2F1. Sequencing reaction was performed on a Perkin Elmer 9700 thermacycler. Conditions were 75°C for 2 min, 94°C for 4 min, and 30 cycles of 96°C for 20 s, 50°C for 20 s and 60°C for 4 min. After cycle sequencing the samples, a cleaning step was done using the multiscreen 96-well plate cleaning system from Millipore (Bedford, MA, USA). The 96-well multiscreening plate was prepared by adding a fixed amount (manufacturer's specification) of Sephadex-50 (Amersham Pharmacia Biotech, Piscataway, NJ, USA) and 300 μl of deionized water. After 1 h of incubation at room temperature, the water was removed from the multiscreen plate by centrifugation at 750g for 5 min. After the Sephadex in the multiscreen plate was partially dried, the whole cycle-sequencing reaction was added to the center of each well, centrifuged at 750g for 5 min and the clean sample was collected on a sequencing microtiter plate (Beckman Coulter, Fullerton, CA, USA). The plate was then dried on Speed-Vac SC 110 model with a microtiter plate holder (Savant Instruments Inc, Holbrook, NY, USA). The dried samples were immediately resuspended with 25 μl of deionized ultraPure formamide (J. T. Baker; Phillipsburg, NJ, USA) and one drop of mineral oil was added to the top of each sample. Samples were either sequenced immediately on a CEQ 2000 DNA sequencing instrument (Beckman Coulter Inc.) or stored at −30°C.

2.5. Sequence information cleaning

Raw sequences originating from the DNA sequencer had either of 5 letters in their result: ATCG for the nucleotide bases, and N when the sequencer program could not call a base. Usually the beginning and ends of the sequences have a higher proportion of N calls. Sequences also contain primer and vector sequences used in the library construction. For this reason, the raw sequences were treated by a program written in VisualBasic 6.0 (VB) (Microsoft Corp., Redmond, WA, USA) as follows: (1) The sequences were analyzed in their first 80 base pairs (bp) for groups of four Ns, and, if found, the block of 4 Ns closer to position 80 was used to trim the raw sequence from this 5′ N-rich region. (2) For sequences longer than 110 bp, windows of 10 bp were screened for the occurrence of 4 or more Ns above position 100. The positive window with the smallest position value was used to trim the sequence from the 3′ N-rich region. Sequences thus trimmed and having more than 10% N content were discarded. (3) Good quality and trimmed sequences were then searched for occurrence of the primers used in the library construction (the SMART III primer, as well as the CDS/R primers). A moving window of the size of the primer was searched on the sequence for matches with the primer sequence. If more than 70% match was obtained, or if a contiguous match longer than 50% of the length of the primer was observed, the sequence was trimmed at the beginning or end of the window depending on the expected position of the primer. This simple algorithm avoided errors due to spurious insertions. (4) The trimmed sequence was 'polished' by removing any trailing N. The sequence final N content was assessed, as well as its AT content and length. The final sequence was written to a FASTA format file containing in its definition line the actions taken by the program.

2.6. Searches for known sequence similarities and known protein domains of the cDNA sequences

To obtain information on the possible role of the cDNA sequences, the FASTA file containing all the stripped sequences were blasted against the GenBank non-redundant protein database (NR) from the National Center for Biotechnology Information (NCBI) using the standalone BlastX program found in the executable package at ftp://ftp.ncbi.nlm.nih.gov/blast/executables/ (Altschul et al., 1997). The NR database was regularly downloaded, uncompressed with GUNZIP (found at www.gzip.org/), and formatted for Blast program use with the formatdb program (executables also found at ftp://ftp.ncbi.nlm.nih.gov/blast/executables/) with the help of a program written in PERL code (software found at www.activeperl.com). NCBI sequences are indicated in this manuscript by their acession number as gi|XXXX where XXXX is a unique identifier number. The resulting file was parsed, and the best match was incorporated in the FASTA definition line after the delimiter |. The sequences were next submitted to the standalone program RPSBlast (Altschul et al., 1997) and searched against the Conserved Domains Database (CDD) (found at ftp://ftp.ncbi.nlm.nih.gov/pub/mmdb/cdd/) which includes all Pfam (Bateman et al., 2000) and SMART (Schultz et al., 2000) protein domains. The RPSBlast

result file was parsed as above and the best match incorporated also into the FASTA definition line of the sequence. When all sequences of a particular cluster were blasted against the NR protein database (using the BlastX program), the best protein match was searched for the species from which the NR database sequence originated. If the species was not *Ae. aegypti*, or no matches to the NR were found, the cluster was marked as representing a novel *Ae. aegypti* sequence (indicated by the Y under the Table 1 column marked as 'Novel?'). All cluster sequences that gave a match to an *Aedes aegypti* protein sequence were further individually inspected to verify whether the cDNA sequence represented nearly the same information translated as the protein match, or whether it represented a closely related, but different, protein. In this latter case a Y would also be added to the Table 1 result in the column "New?" for the row of the cluster in question.

2.7. Sequence clustering

The FASTA file containing all sequences were clustered by first comparing (using the BlastN program) each sequence against the formatted database file, using a Blast cutoff score of $1E^{-60}$. The resulting file was used to join in a single cluster all sequences that shared at least one common sequence in the BlastN result. Thus, if sequence A had a $1E^{-60}$ match to B and B had a similar match to C, the 3 sequences would be joined even if A had a less meaningful score in relation to C. The clustering program also made individual FASTA formatted files for each cluster, sorted in descending order of sequence size. When these files contained 2 or more sequences, they were used as input for the sequence aligment program CLUSTALW (Higgins et al., 1996), which was called automatically by the clustering program. CLUSTAL aligment files were thus created for each cluster having 2 or more sequences. This clustering program was also written in VB. Finally, a program was written in VB that combines all the results to create the Table 1 of this paper, except for the "Comments" column. The output of this program is imported into a Microsoft Excel spreadsheet. In the supplemental material, available by e-mail request to JRibeiro@NIH.GOV Table 1 includes hyperlinks to the best NR protein match in the NCBI site, all FASTA files for each individual cluster, CLUSTAL aligment files for each cluster, when available, as well the FASTA file for the whole database. Each cluster was individually analyzed for the probable function of its translation product and assigned a probably secreted, probably housekeeping, and indeterminate function. This decision was based on the best match to the NR protein database and their related sequences as searched online in the NCBI site (www.ncbi.nlm.gov), and on the SMART and/or Pfam matches, including searches of the nature of the domains by online searches of the respective sites.

2.8. Full-length sequencing of selected cDNA clones

An aliquot (4 μl) of the lambda phage containing the cDNA of interest was amplified using the PT2F1 and PT2R1 primers (conditions as described above). The PCR samples were cleaned using the MutiScreen-PCR 96-well filtration system (Millipore). Cleaned samples were sequenced first with PT2F3 primer and subsequently with custom primers. Primer selection for complete sequence of selected full-length cDNA was also assisted by a program (written in VB) that identified unique primer sites within the sequences. To assemble the sequences, the previously known sequence was blasted against the new sequence using the standalone program bl2seq found with the executable package provided at the NCBI ftp site mentioned above. After identifying the regions of overlap, the two sequences were joined. The program attempted to locate a polyA region by using a window of 12 bp, in which 11 As would constitute a polyA. If no polyA was found, a new set of primers would be found to continue extension of the cDNA. The program also generates CLUSTAL alignments of all sequences, produces a consensus output, and the 3 possible translations of this unidirectionally cloned RNA. The final alignment is adjusted by hand. If necessary, the original tracings of the DNA sequencer are reviewed to decide on critical base calls. The translated sequences are submitted as a FASTA file to the SIGNALP server (at http://www.cbs.dtu.dk/services/SignalP/) (Nielsen et al., 1997), which responds by e-mail indicating whether a signal peptide exists and its location. A program written in VB interprets this SIGNALP result file and removes the signal peptide, if it is predicted to exist, to create a mature protein sequence. Molecular weights using average molecular masses for C, H, O, N, P, and S are calculated for all protein sequences-as are pI based on reduced proteins-following the pKa for amino acids within proteins as indicated before (Altland, 1990; Bjellqvist et al., 1994). This program, combined with the program generating Table 1 of this paper, produced an output that can be read by the spreadsheet program Excel to produce the Table 2 shown in this paper. In the supplemental table, available by e-mail request, hyperlinks are given to all proteins.

2.9. SDS-polyacrylamide gel electrophoresis (PAGE)

Nu PAGE 12% Bis-Tris Gel, 1 mm thick, was used (Invitrogen). The gel was run with MOPS buffer according to the manufacturer's instructions. To estimate the molecular weight of the samples, SeeBlue™ markers from Invitrogen (myosin, BSA, glutamic dehydrogenase,

Table 1A
Sequence clusters of salivary gland cDNA's from female *Aedes aegypti* mosquitoes coding for proteins of probably housekeeping function

Line # (1)	Clus # (2)	No. Seq (3)	Best match to NR protein database (4)	E value (5)	Best match to CDD database (6)	E value (5)	Novel (7)	Comments
1	216	1	gi|5733713| chrysoptin precur	6E-31	pfam010095—nucleotidase	4E-35	Y	5'nucleotidase
2	111	1	gi|7290649| Act5C gene product	8E-71	pfam00022 actin	4E-76	Y	Actin
3	175	1	gi|156955| adenine	4E-28			Y	Adenine phosphoriboyl transf.
4	144	1	gi|12862108| putative [Mus musculus]	3E-16	pfam00025 arf	1E-16	Y	ADP rybosylation factor
5	129	1	gi|7300695| AnnIX gene product	1E-78	Smart ANX Annexin repeats	4E-18	Y	Annexin
6	74	1	gi|10726362| CG18745 gene product	3E10			Y	Arrestin?
7	63	2	gi|7302028| CG1746 gene product	9E-34	Pfam00137 ATP-synt—C	5E-22	Y	ATP synthase
8	106	1	gi|905352|H+ ATP synthase [Drosophila	8E-56	pfam00006 ATP-synt—ab	3E-41	Y	ATP synthase
9	122	1	gi|7291477| blw gene product	5E-61	pfam00006 ATP-synt—ab	2E-24	Y	ATP synthase
10	82	1	gi|246435130| putative E1-E2 A…	5E-52			Y	ATPase
11	124	1	gi|246560653| calcyphosine-lik…	3E-38			Y	Calcium binding protein
12	24	4	gi|6690638| N-type calcium c…	0.005	pfam01202 SKI	0.0001	Y	Calcium channel?
13	193	1	gi|2661166| DgZW10 [Drosophila	0.005			Y	Centromere protein
14	138	1	gi|2559010| chaperonin containing	1E-52	pfam00118 cpn60—TCP1	1E-37	T	Chaperonin
15	200	1	gi|7292935| CG10997 gene product	6E-67			Y	Chloride channel
16	123	1	gi|7295410| Bj1 gene product [alt 1]	6E-14			Y	Chromatin binding protein
17	155	1	gi|3649785| chromobox protein (CHCB2)…	2E-30	pfam00385 chromo	2E-14	Y	Chromatin binding protein
18	192	1	gi|7290498| CG3564 gene product	1E-45	pfam01105 EMP24—GP25L	3E-20	Y	Coated vesicle protein
19	40	2	gi|7293206| Cyp1 gene product	4E-72	pfam00160 pro—isomerase	6E-71	Y	Cyclophylin
20	172	1	gi|7291447| CG2852 gene product [alt	2E-33	pfam00160 pro—isomerase	4E-38	Y	Cyclophorin
21	55	2	gi|309068| cytochrome b [Anopheles	3E-83	pfam00033 cytochrome—b—N	2E-59	Y	Cytochrome b
22	118	1	gi|7295470| CG4769 gene product	5E-57	pfam02167 Cytochrome—C1	4E-50	Y	Cytochrome c
23	9	6	gi|5101670| cytochrome oxidase subuni…	5E-57	pfam00510 COX3	3E-65	Y	Cytochrome oxidase
24	26	4	gi|3256672| 279aa long hypothetical p…	0.003	Smart SynN Syntaxin N-terminal domain	0.0003	Y	Cytoskeletal protein?
25	185	1	gi|7290524| CG7010 gene product [alt	6E-55	pfam00676 E1—dehydrog	8E-31	Y	Dehydrogenase
26	60	2	gi|7296723| CRMP geen product [alt l]	2E-44	pfam00744 Dihydrooratase	5E-19	Y	Dihydrooratase
27	96	1	gi|159575| DNA-binding protein [Aedes	0.0000001				DNA binding protein
28	120	1	gi|7296201| CG4164 gene product	3E-36	pfam00226 DnaJ	1E-20	Y	DNAJ protein
29	221	1	gi|7303574| wal gene product	2E-59	pfam00766 ETF—alpha	1E-42	Y	Electron transfer
30	31	4	gi|217274| elongation factor 1 alpha	2E-96	pfam00009 GTP—EFTU	1E-46	Y	Elongation factor
31	151	1	gi|12328436| elongation factor 1 delt…	3E-15			Y	Elongation factor
32	168	1	gi|217274Z elongation factor 1 alpha	1E-102	pfam00009 GTP—EFTU	3E-23	Y	Elongation factor
33	119	1	gi|12862421| hypothetical protein [Ps…	2E-16			Y	Esterase?
34	176	1	gi|559067| ferritin subunit {Aedes	1E-81	pfam00210 ferritin	4E-40		Ferritin
35	186	1	gi|6688879| ferritin subunit (glycosy…	6E-16	pfam00210 ferritin	0.0000002	Y	Ferritin
36	210	1	gi|2565196| non-functional folate	3E-19			Y	Folate binding protein
37	135	1	gi|7297468| CG9520 gene product [alt	4E-88	pfam01762 Galactosyl—T	0.0000003	Y	Galactosyl-transferase
38	79	1	gi|12654561| MpV17 transgene, …	0.0008			Y	Glomerulosclerosis protein
39	48	2	gi|460879| 6-phosphogluconate	1E-48	pfam00393 6PGD	9E-50	Y	Glucose 6P dehydrogenase
40	174	1	gi|7296321| CG5976 gene product	6E-25			Y	Glutaminyl cyclase

(*continued on next page*)

Table 1A (*continued*)

Line # (1)	Clus # (2)	No. Seq (3)	Best match to NR protein database (4)	E value (5)	Best match to CDD database (6)	E value (5)	Novel (7)	Comments
41	37	3	gi|11385473| glutathione S-t…	0.015	pfam00769ERM	0.0001	Y	Glutatione transferase
42	160	1	gi|7296006| GlyP gene product	2E-64	pfam00343 phosphorylase	2E-44	Y	Glycogen phosphorylase
43	62	2	gi|1066808| heat shock protein 82	6E-63	pfam00183 HSP90	1E-67	Y	Heat shock protein
44	128	1	gi|3426021| heat shock 70 kD protein…	1E-92	pfam00012 HSP70	3E-67	Y	Heat shock protein
45	201	1	gi|7299978| Hsc 70-4gene product	3E-70	pfam00012 HSP70	3E-76	Y	Heat shock protein
46	202	1	gi|1066808| heat shock protein 82	4E-89			Y	Heat shock protein
47	211	1	gi|7290308| EG:25E8.1 gene product	3E-32	pfam00012 HSP70	90.00000002	Y	Heat shock protein
48	159	1	gi|10727768| CG18812 gene product	4E-54	LOAD:hismacro hismacro	9E-22	Y	Histone
49	161	1	gi|7296159| CG3662 gene product	0.00000006			Y	Integral membrane protein
50	219	1	gi|7292705| CG1967 gene product	1E-35	pfam01105 EMP24—GP25L	4E-16	Y	Integral membrane protein
51	212	1	gi|7300004| Surf4 gene product	3E-38	pfam02077 SURF4	1E-37	Y	Integral proteins of ER
52	66	1	gi|7296765| CG11999 gene product	2E-37	pfam02366 pMT	0.00000004	Y	Mannosyl transferase
53	224	1	gi|7292004| CG13887 gene product	1E-30			Y	Membrane protein
54	51	2	gi|8573405| ATPase 6 [Drosophila	9E-42	pfam00119 ATP-synt A	0.00000003	Y	Mitochondrial ATP-synthase
55	53	2	gi|309058| cytochrome c oxidase subunit	1E-59	pfam00115 COX1	6E-58	Y	Mitochondrial enzyme
56	33	3	gi|14438862| ADP/ATP carrier protein	2E-78	pfam00153 mito—carr	4E-22	Y	Mitochondrial protein
57	165	1	gi|7299538| CG18347 gene product	2E-35	pfam00153 mito—carr	3E-15	Y	Mitochondrial protein
58	228	1	gi|7295020| CG5012 gene product	4E-28			Y	Mitochondrial ribosomal protein
59	194	1	gi|7298975| CG10435 gene product	0.00000005			Y	Myosin
60	171	1	gi|7290598| Mlc-c gene product	4E-68			Y	Myosin non-muscle
61	203	1	gi|7292278| CG12093 gene product	2E-19			Y	NAD metabolism
62	198	1	gi|7292912| CG9940 gene product	7E-42			Y	NAD synthase?
63	68	1	gi|7290517| Cbp80 gene product	3E-22			Y	Nuclear cap binding protein
64	232	1			pfam02168 Occludin	0.00000006	Y	Occludin
65	206	1	gi|2708713| ornithine decarboxylase	4E-23	pfam02100 ODC—AZ	2E-14	Y	ODC antizyme
66	92	1	gi|2665654| polyadenylate binding	9E-17			Y	Poly(A) binding protein
67	215	1	gi|64970| polyA binding protein	2E-40	pfam00076 rrm	9E-10	Y	PolyA binding protein
68	231	1	gi|7296685| CG2100 gene product	2E-42			Y	PolyA polymerase
69	109	1	gi|7298546|1(2)37Cc gene product	2E-67	pfam01145 Band—7	8E-42	Y	Prohibitin
70	158	1	gi|7292848| REG gene product	4E-35	pfam02251 PA27—alpha	0.00000001	Y	Proteasome protein
71	167	1	gi|7304058| CG2038 gene product	9E-45	Smart PINT motif in proteasome subunits	0.0000002	Y	Proteasome protein
72	107	1	gi|7299214Z CG8286 gene product	3E-36	pfam00515 TPR	0.0008	Y	Protein kinase?
73	233	1	gi|12856947| putative [Mus musculus]	2E-45			Y	Protein transporter
74	177	1	gi|729676| CG5364 gene product	2E-18			Y	Ribophorin
75	12	6	gi|7297568| sop gene product	2E-66	pfam00333 Ribosomal—S5	1E-25	Y	Ribosomal protein
76	27	4	gi|6911249| 60S acidic ribos…	9E-85	pfam00466 Ribosomal—L10	2E-35	Y	Ribosomal protein
77	60	2	gi|7291436| CG4046 gene product	4E-62	pfam00380 Ribosomal—S9	7E-59	Y	Ribosomal protein
78	65	1	gi|7300661| RpS20 gene product	3E-50	pfam00338 Ribosomal—S10	2E-33	Y	Ribosomal protein
79	71	1	gi|7290065| RpL36 gene product	3E-43	pfam01158 Ribosomal—L36e	2E-31	Y	Ribosomal protein
80	75	1	gi|5690418| ribosomal protei…	2E-63	pfam00833 Ribosomal—S17E	3E-55	Y	Ribosomal protein
81	76	1	gi|10728547| CG2998 gene product	2E-18	pfam01200 Ribosomal—S28E	7E-24	Y	Ribosomal protein

Table 1A (*continued*)

Line # (1)	Clus # (2)	No. Seq (3)	Best match to NR protein database (4)	E value (5)	Best match to CDD database (6)	E value (5)	Novel (7)	Comments
82	80	1	gi|7291277| RpL29 gene product	9E-18	pfam01779 Ribosomal—L29e	4E-14	Y	Ribosomal protein
83	94	1	gi|7294504| RpS4 gene product [alt 2]	2E-54	pfam00900 Ribosomal—S4e	8E-22	Y	Ribosomal protein
84	99	1	gi|4239713| acidic ribosomal protein	2E-30	pfam00427 60s—ribosomal	9E-33	Y	Ribosomal protein
85	113	1	gi|7302162| CG12775 gene product	6E-62	pfam01157 Ribosomal—L21e	5E-40	Y	Ribosomal protein
86	115	1	gi|7289746| Qm gene product [alt 1]	3E-97	pfam00826 Ribosomal—L10e	3E-90	Y	Ribosomal protein
87	121	1	gi|7293234| Rbp2 gene product	5E-38	pfam000776 rrm	0.0000003	Y	Ribosomal protein
88	142	1	gi|7292864| CG2033 gene product [alt	5E-64	pfam00410 Ribosomal—S8	9E-41	Y	Ribosomal protein
89	164	1	gi|4585827| ribosome associated membr…	3E-18			Y	Ribosomal protein
90	184	1	gi|1359478| put. S3a ribosomal protein	4E-39	pfam01015 Ribosomal—S3Ae	2E-49	Y	Ribosomal protein
91	204	1	gi|293233| ribosomal protein L8 [Aedes	2E-86	pfam00181 Ribosomal—L2	7E-52	Y	Ribosomal protein
92	170	1	gi|7304089| Pabp2 gene product	1E-45			Y	RNA binding protein
93	136	1	gi|7296264| M(2)21AB gene product	1E-20	pfam00438 S-AdoMet—synt	0.000000002	Y	S-Adenosylmethionine synthase
94	127	1	gi|10727678| 14-3-3zeta gene product	3E-65	pfam00244 14-3-3	E3-56	Y	Signal transduction
95	134	1			pfam00769 ERM	0.0006	Y	Signal transduction
96	159	1	gi|7289106| CG17454 gene product	1E-39	Smart TUDOR Tudor domain	0.00000004	Y	Splicing factor
97	234	1	gi|7292391| Scsalpha gene product	4E-51			Y	Succinyl-CoA synthetase
98	183	1	gi|290221| Cu/Zn-superoxide dismutase	2E-50	pfam00080 sodcu	1E-71	Y	Superoxide dismutase
99	81	1	gi|12654441| similar to S. pom…	2E-67	pfam00085 thiored	0.0006	Y	Thioredoxin
100	112	1	gi|7294310| Pdi gene product	2E-32	pfam00085 thiored	8E-36	Y	Thioredoxin
101	222	1	gi|7716428| thioredoxin 1 [A…	4E-43	pfam00085 thiored	6E-32	Y	Thioredoxin
102	237	1	gi|9937480| thyroid hormone…	2E-47			Y	Thyroid hormone receptor prn.
103	188	1	gi|7290336| CG8636 gene product	9E-49			Y	Translation initiation factor
104	126	1	gi|12025457| activating tran…	4E-21	Smart BRLZ basic region leucin zipper	5E-11	Y	Transcription factor
105	180	1	gi|7293771Z Taf60 gene product	8E-59			Y	Transcription initiation factor
106	100	1	gi|6469517| translation init…	7E-49	pfam01287 eIF-5A	2E-46	Y	Translation initiation factor
107	90	1	gi|7295855| CG17259 gene product	7E-22	pfam02403 Seryl—tRNA—N	0.0000005	Y	tRNA synthase
108	95	1	gi|7303353| CG4062 gene product	0.00000001	pfam01041 DegT—DnrJ—EryC1	0.0004	Y	tRNA synthase
109	103	1	gi|7301089| CG5977 gene product	4E-25			Y	tRNA synthase
110	163	1	gi|7295855| CG17259 gene product	3E-62	pfam00587 tRNA-synt—2b	3E-13	Y	tRNA synthase
111	28	1	gi|12653955| Similar to tropom…	0.056	pfam00261 Tropomyosin	0.000009	Y	Tropomyosin
112	110	1	gi|7295194| Hn gene product	4E-27	LOAD:ACT ACT	0.00002	Y	Tryptophan hydroxylase
113	131	1	gi|9837286| polyubiquitin C…	5E-72	pfam00240 ubiquitin	6E-42	Y	Ubiquitin
114	195	1	gi|7295143| CG7375 gene product	3E-21			Y	Ubiquitin conjugating enzyme
115	89	1	gi|10442628| V-ATPase 16 kD…	5E-13	pfam00137 ATP-synt—C	0.00000004	Y	V-ATPase subunit
116	98	1	gi|7290447| CG2934 gene product	1E-58	pfam01992vATP-synt—AC39	3E-40	Y	V-ATPase subunit
117	145	1	gi|10728827| VhaSFD gene product [alt	2E-41			Y	V-ATPase subunit
118	213	1	gi|7302940| Vha44 gene product [alt	2E-51			Y	V-ATPase subunit

(1) Refers to table line number; (2) Cluster order in database sorted by decreasing order of cDNA abundance representation; (3) Number of sequences in cluster; (4) Best match to NR protein database of NCBI provided by BlastX program. Cutoff E value=0.1; (5) E value indicating significance of the match to the protein or domain databases; (6) Best domain match to the Conserved Domain Database of the NCBI. Cutoff E value=0.001; (7) Indicates sequence novelty by a Y when the best NR protein match is not from *Aedes aegypti*.

Table 1B
Sequence clusters of salivary gland cDNA's from female *Aedes aegypti* mosquitoes coding for probably secreted proteins

Line # (1)	Clus # (2)	No. Seq (3)	Best match to NR protein database (4)	E value (5)	Best match to CDD database (6)	E value (5)	Novel (7)	Comments
1	10	6	gi|2114497| 30 kDa salivary gland	4E-30	pfam01454 MAGE	5E-14		30 kD antigen
2	5	8	gi|2114497| 30 kDa salivary gland	5E-19	pfam02386 TrkH	0.0006	Y	30 kD antigen (new member)
3	54	2	gi|8745398| putative adenosi...	3E-10			Y	Adenosine deaminase
4	29	4	gi|2190949| amylase [Aedes aegypti]	2E-57				Amylase
5	25	4	gi|4204688| fibrinogen-like pr...	4E-17	Smart FBG Fibrinogen-related domains	1E-25	Y	Angiopoietin - Fibrinogen binding domain
6	52	2	gi|3413516| Hakata antigen [Homo	2E-13	Smart FBG Fibrinogen-related domains	8E-21	Y	Angiopoietin - Fibrinogen binding domain
7	11	6	gi|8927462| antigen 5 precur...	2E-19	pfam00188 SCP	3E-17	Y	Antigen 5 protein
8	15	5	gi|7292977| CG9400 gene product	2E-31	pfam00188 SCP	5E-20	Y	Antigen 5 protein
9	34	3	gi|7292977| CG9400 gene product	1E-28	pfam00188 SCP	2E-20	Y	Antigen 5 protein
10	42	2	gi|7292977| CG9400 gene product	4E-26	pfam00188 SCP	3E-12	Y	Antigen 5 protein
11	19	5	gi|763502| apyrase [Aedes aegypti]	7E-97	pfam01009 5—nucleotidase	2E-42		Apyrase
12	59	2	gi|556272| apyrase [Aedes aegypti]	4E-97	pfam01009 5—nucleotidase	7E-50		Apyrase
13	4	11	gi|3929312| fimbriae-associated	0.007			Y	Bacterial adhesion molecule
14	39	3	gi|7299219| Crc gene product	3E-45	pfam00262 calreticulin	4E-35	Y	Calreticulin
15	150	1	gi|7296094| CG5397 gene product	4E-26	pfam00135 COesterase	1E-16	Y	Carboxylesterase
16	58	2	gi|4335872| chymotrypsin 1 [Anopheles	2E-23	Smart Tryp—SPc Trypsin-like serine	1E-25	Y	Chymotrypsin - Immunity?
17	101	1	gi|159559| D7 protein [Aedes aegypti]	4E-22			Y	D7 family
18	18	5	gi|159557| D7 protein [Aedes aegypti]	1E-107				D7 protein
19	1	22	gi|159557| D7 protein [Aedes aegypti]	5E-22			Y	D7 protein family
20	23	4	gi|159559| D7 protein [Aedes aegypti]	0.051			Y	D7 protein family
21	96	1	gi|159575| DNA-binding protein [Aedes	1E-07				DNA binding protein
22	3	19	gi|159566| alpha-1,4-glucosidase [Aedes	1E-111	pfam00128 alpha-amylase	2E-21		Glucosidase
23	20	5	gi|4808550| putative gram ne...	7E-42			Y	Gram-negative binding protein
24	35	3	gi|7362608| putative infection respon...	2E-25			Y	Immunity?
25	16	5	gi|567099| lysozyme [Bombyx mori]	2E-35	Smart LYZ1 Alpha-lactalbumin /	8E-36	Y	Lysozyme
26	43	2	gi|7292051| CG9134 gene product	3E-16	pfam00059 lectin—c	8E-21	Y	Macrophage mannose receptor
27	44	2	gi|7292051| CG9134 gene product	1E-14	Smart CLECT C-type lectin (CTL) or	1E-15	Y	Macrophage mannose receptor
28	72	1	gi|7304088| CG2297 gene product	4E-23	pfam01395 PBP—GOBP	3E-08	Y	Odorant binding protein
29	205	1	gi|7302980| Lis1 gene product	3E-36			Y	PAF acetyl hydrolase
30	85	1	gi|7301922| CG9737 gene product	1E-07			Y	Prophenoloxidase activating
31	13	6	gi|7297954| CG5418 gene product	3E-15	pfam01156 IU—nuc—hydro	1E-18	Y	Purine hydrolase
32	32	4	gi|4868337| trypsinogen RdoT...	3E-08	Smart Tryp—SPc Trypsin-like serine	5E-10	Y	Serine protease - Immunity?
33	36	3	gi|7291783| CG15873 gene product	4E-10	Smart Tryp—SPc Trypsin-like serine	5E-10	Y	Serine protease - Immunity?
34	6	8	gi|3411116| FXa-directed	4E-81				Serpin - Anticlotting
35	49	2	gi|3411116| FXa-directed	3E-18			Y	Serpin (new protein)
36	227	1	gi|7291768| CG3376 gene product	1E-93			Y	Sphingomyelin phosphodiesterase
37	93	1	gi|7295121| CG7188 gene product	0.0003			Y	Testes protein in mammals
38	21	4	gi|4336720| salivary vasodilatory	1E-32				Vasodilator

(1) Refers to table line number; (2) Cluster order in database sorted by decreasing order of cDNA abundance representation; (3) Number of sequences in cluster; (4) Best match to NR protein database of NCBI provided by BlastX program. Cutoff E value=0.1; (5) E value indicating significance of the match to the protein or domain databases; (6) Best domain match to the Conserved Domain Database of the NCBI. Cutoff E value =0.001; (7) Indicates sequence novelty by an Y when the best NR protein match is not from *Aedes aegypti*.

Table 1C
Sequence clusters of salivary gland cDNA's from female *Aedes aegypti* mosquitoes coding for proteins of uncertain function

Line # (1)	Clus # (2)	No. Seq (3)	Best match to NR protein database (4)	E value (5)	Best match to CDD database (6)	E value (5)	Novel (7)	Comments
1	147	1	gi|7294922| CG3424 gene product [alt	4E-28			Y	Aminoacid transporter?
2	105	1			pfam00968 Dwarfin	2E-09	Y	Dwarfin domain
3	207	1	gi|7296922| CG8895 gene product [alt	3E-38	pfam02453 Reticulon	7E-41	Y	Reticulon protein
4	2	19					Y	
5	7	7					Y	
6	8	6					Y	
7	14	5					Y	
8	17	5					Y	
9	22	4					Y	
10	30	4					Y	
11	38	3					Y	
12	41	2					Y	
13	45	2					Y	
14	46	2					Y	
15	47	2					Y	
16	56	2					Y	
17	57	2					Y	
18	61	2					Y	
19	64	1					Y	
20	67	1					Y	
21	69	1					Y	
22	70	1					Y	
23	73	1					Y	
24	77	1					Y	
25	78	1					Y	
26	83	1	gi|7304069| CG2280 gene product	3E-22	Smart CUB Domain first found in C1r	7E-07	Y	
27	84	1	gi|7299592| CG10038 gene product [alt	3E-13			Y	
28	86	1					Y	
29	87	1	gi|12849932| putative [Mus musculus]	3E-08			Y	
30	88	1					Y	
31	91	1					Y	
32	97	1	gi|7291118| CG2972 gene product	3E-19	pfam02000 DUF133	7E-08	Y	
33	102	1					Y	
34	104	1					Y	
35	108	1	gi|7293475| CG15043 gene product	2E-17			Y	
36	114	1					Y	
37	116	1					Y	
38	117	1					Y	
39	125	1					Y	
40	130	1					Y	
41	132	1					Y	
42	133	1					Y	
43	137	1					Y	

(*continued on next page*)

Table 1C (*continued*)

Line # (1)	Clus # (2)	No. Seq (3)	Best match to NR protein database (4)	E value (5)	Best match to CDD database (6)	E value (5)	Novel (7)	Comments
44	139	1	gi|7303552| CG13189 gene product	8E-22			Y	
45	140	1	gi|7291185| CG1534 gene product [alt	2E-11			Y	
46	141	1					Y	
47	143	1	gi|12855203| putative [Mus musculus]	4E-34			Y	
48	146	1					Y	
49	148	1					Y	
50	149	1					Y	
51	152	1					Y	
52	153	1					Y	
53	154	1	gi|7297376| CG7830 gene product	3E-56			Y	
54	156	1					Y	
55	157	1					Y	
56	162	1					Y	
57	166	1					Y	
58	173	1					Y	
59	178	1					Y	
60	179	1			pfam00531 death	0.00002	Y	
61	181	1	gi|7300465| CG6005 gene product	5E-36			Y	
62	182	1					Y	
63	187	1					Y	
64	189	1	gi|7299620| CG5196 gene product	0.00008			Y	
65	190	1	gi|7292344| CG12017 gene product	0.0004			Y	
66	191	1	gi|7294854| CG6707 gene product [alt	0.000004			Y	
67	196	1	gi|7297585| CG4658 gene product	7E-24			Y	
68	197	1	gi|7291803| CG13585 gene product	2E-13			Y	
69	199	1					Y	
70	208	1	gi|7293775| CG9231 gene product	5E-28			Y	
71	209	1					Y	
72	214	1					Y	
73	217	1					Y	
74	218	1					Y	
75	220	1					Y	
76	223	1					Y	
77	225	1	gi|7293364| CG8408 gene product	2E-07			Y	
78	226	1	gi|7298721| CG14404 gene product	5E-10			Y	
79	229	1	gi|7303621| CG13211 gene product [alt	0.000001			Y	
80	230	1					Y	
81	235	1					Y	
82	236	1					Y	

(1) Refers to table line number; (2) Cluster order in database sorted by decreasing order of cDNA abundance representation; (3) Number of sequences in cluster; (4) Best match to NR protein database of NCBI provided by BlastX program. Cutoff E value=0.1; (5) E value indicating significance of the match to the protein or domain databases; (6) Best domain match to the Conserved Domain Database of the NCBI. Cutoff E value =0.001; (7) Indicates sequence novelty by an Y when the best NR protein match is not from *Aedes aegypti*.

Table 2
Thirty two novel full-length cDNA sequences from *Aedes aegypti* salivary glands

Sequence name (1)	NCBI accession number	Best NR protein match (click for all matches) (2)	E value (3)	Best CDD match (click for all matches) (4)	E value (3)	Predicted mass of polypeptide (kDa)	Predicted cleavage site by SignalP server (5)	Predicted mass of mature protein (kDa)	pI of Mature peptide	Comments
AEA—CLU1	16225992	gi|118216|D7—AEDAE D7 PROTEIN PRECU	3.00E-52			38.6	21	36.3	9.45	New D7 protein member
AEA—CLU23	16225995	gi|159559| (M33157) D7 protein [Aed	0.013			17.7	19	15.5	9.66	Short D7 protein
AEA—CLU61	18568330					16.9	21	14.5	5.52	Short D7 protein
AEA—CLU8	18568332					18.2	18	16.2	5.11	Short D7 protein
AEA—CLU5	18568322	gi|14423642|ALL3—AEDAE 30 KDA SALIV	1.00E-28			23.6	19	21.4	4.58	Similar to 30 kDa allergen
AEA—CLU11	18568278	gi|8927462|AF259957—1 (AF259957) an	9.00E-34	Smart smart00198 SCP	1.00E-22	29.5	20	27.3	9.44	Antigen 5 family
AEA—CLU15	18568284	gi|7292977| (AE003495) CG9400 gene	2.00E-44	Smart smart00198 SCP	3.00E-26	28.9	21	26.7	9.62	Antigen 5 family
AEA—CLU34	18568308	gi|7292998| (AE003496) Ag5r2 gene p	4.00E-42	Smart smart00198 SCP	1.00E-24	28.4	20	26.3	9.84	Antigen 5 family
AEA—CLU42	18568316	gi|4887102|AF132511—1 (AF132511) an	1.00E-45	Smart smart00198 SCP	1.00E-19	29.7	27	26.5	10	Antigen 5 family
AEA—CLU13	18568280	gi|7292911| (AE003493) CG12177 gene	2.00E-43	pfam01156 IU—nuc—hydro	3.00E-33	37.9	15	36.3	4.82	Purine hydrolase
AEA—CLU54	18568326	gi|8745398|AF234182—1 (AF234182) pu	1.00E-142	pfam00962 A—deaminase	1.00E-47	60.6	35	56.9	6.05	Adenosine deaminase
AEA—CLU150	18568286	gi|7296094| (AE003586) CG5397 gene	1.00E-150	pfam00135 COesterase	3.00E-67	68.9	25	66	8.21	Carboxylesterase
AEA—CLU30	18568304	gi|7302317| (AE003790) CG9455 gene	2.00E-06	pfam00079 serpin	3.00E-08	47.1	18	45.1	5.26	Serpin
AEA—CLU39	18568312	gi|416844|CRTC—DROME CALRETICULIN P	1.00E-143	pfam00262 calreticulin	1.00E-105	46.8	19	44.8	4.35	Calreticulin
AEA—CLU25	18568298	gi|120092|FIBA—PARPA FIBRINOGEN-LIK	2.00E-34	Smart smart00186 FBG	4.00E-43	33.7	19	31.5	4.98	Angiopoietin related protein
AEA—CLU52	18568324	gi|7291373| (AE003456) CG6676 gene	2.00E-30	Smart smart00186 FBG	7.00E-41	34.1	18	31.9	5.6	Angiopoietin related protein
AEA—CLU44	18568318	gi|7292051| (AE003471) CG9134 gene	5.00E-15	pfam00059 lectin—c	2.00E-12	17.7	20	15.6	5.03	C- type lectin
AEA—CLU16	18568288	gi|1346474|LYC—BOMMO LYSOZYME PRECU	5.00E-36	Smart smart00263 LYZ1	1.00E-29	15.9	22	13.6	9.06	Lysozyme
AEA—CLU20	18568294	gi|2467083| (AJ001042) putative gra	2.00E-92	pfam00722 Glyco—hydro—16	3.00E-10	41.9	24	39.4	9.67	Immunity recognition protein

(*continued on next page*)

Table 2 (*continued*)

Sequence name (1)	NCBI accession number	Best NR protein match (click for all matches) (2)	E value (3)	Best CDD match (click for all matches) (4)	E value (3)	Predicted mass of polypeptide (kDa)	Predicted cleavage site by SignalP server (5)	Predicted mass of mature protein (kDa)	pI of Mature peptide	Comments
AEA—CLU85	18568334	gi|7301922| (AE003773) CG9737 gene	7.00E-68	Smart smart00020 Tryp—SPc	8.00E-51	45.1	24	42.5	5.41	Prophenol oxidase activator
AEA—CLU58	18568328	gi|4335866| (AF051778) chymotrypsin	1.00E-35	Smart smart00020 Tryp—SPc	2.00E-41	32.2	21	30	10.04	Chymotrypsin
AEA—CLU32	18568306	gi|9857909|AF271695—1 (AF271695) tr	1.00E-22	Smart smart00020 Tryp—SPc	1.00E-44	32.2	32	28.7	6.5	Serine protease
AEA—CLU35	18568310	gi|7362608| (AJ237664) putative inf	8.00E-26			9.1	22	6.8	9.09	Immunity related protein
AEA—CLU26	18568300	gi|13195256|AF323442—1 (AF323442) 2	9.00E-06			66.2	24	63.6	5.59	Similar to *Plasmodium* adhesion protein
AEA—CLU28	18568302	gi|14775649| similar to RESTIN (CYT	2.00E-04			64.4	19	62.2	5.84	Similar to sperm protein STAT domain
AEA—CLU2	18568292	gi|9631598| Asn/Thr/Ser/Val rich pr	0.001			58.3	17	56.5	5.39	Similar to *Ricketsia* antigen
AEA—CLU24	18568296					36.2	21	34	5.66	Unknown protein
AEA—CLU45	18568320					30.5	20	28.4	4.82	Unknown protein
AEA—CLU4	18568314	gi|7521941|T17451 fimbriae-associat	0.002			18.6	19	16.7	4.93	Unknown protein
AEA—CLU14	18568282					10.7	25	7.8	8.61	Unknown peptide
AEA—CLU17	18568290					10.1	18	8.3	6.38	Unknown peptide
AEA—CLU7	18643261					N/A				MITE transposable element

(1) Sequence name includes the cluster from which sequence originated. Clusters were sorted out by order of decreasing cDNA representation abundance; (2) Best match to the non-redundant protein database from GeneBank when the program Blastx was used to search the database; (3) E value indicating significance of the protein sequence match; (4) Best matches and E values to the Conserved Domains Database from the National Center for Biotechnology Information; (5) MW of the putative protein; (6) Cleavage site of signal peptide as predicted by the SignalP server (http://www.cbs.dtu.dk/services/SignalP/) (Nielsen et al., 1997); (7) MW of the predicted mature protein; (8) Isoelectric pH of the mature protein.

alcohol dehydrogenase, carbonic anhydrase, myoglobin, lysozyme, aprotinin, and insulin, chain B) were used. Salivary gland homogenates were treated with NuPAGE LDS sample buffer (Invitrogen). Twenty pairs of homogenized salivary glands per lane (approximately 20 μg protein) were applied when visualization of the proteins bands stained with Coomassie Blue was required. For aminoterminal sequencing of the salivary proteins, 20 homogenized pairs of glands were electrophoresed and transferred to polyvinylidene difluoride (PVDF) membrane using 10 mM CAPS, pH 11.0, 10% methanol as the transfer buffer on a Blot-Module for the XCell II Mini-Cell (Invitrogen). The membrane was stained with Coomassie Blue in the absence of acetic acid. Stained bands were cut from the PVDF membrane and subjected to Edman degradation using a Procise sequencer (Perkin-Elmer Corp.). To find the cDNA sequences corresponding to the amino acid sequence—obtained by Edman degradation of the proteins transferred to PVDF membranes from PAGE gels—we wrote a search program (in VB) that checked these amino acid sequences against the three possible protein translations of each cDNA sequence obtained in the mass sequencing project. This program was written using the same approach utilized in the BLOCKS (Henikoff and Henikoff, 1994) or Prosite (Sibbald et al., 1991) databases. The program is very useful when mixed sequence information occurs, as for example, amino terminal sequences deriving from a mix of equal peptides. In this case, two different cDNA sequences may be unambiguously found as matches.

2.10. Statistical tests

Statistical tests were performed with SigmaStat version 2.0 (Jandel Software, San Rafael, CA). Kruskal–Wallis ANOVA on ranks was performed and multiple comparisons were done by the Dunn method. Dual comparisons were made with the Mann–Whitney rank sum test.

3. Results

We sequenced 456 random clones from a PCR-based cDNA library from the salivary glands of adult female *Aedes aegypti*. After grouping this database in clusters, as described in the methods section, 238 individual clusters were found. The distribution of the clusters as a function of their decreasing number of sequences is shown in Fig. 1. Table 1 indicates the similarities of the putative translated products of such clusters when compared with the nonredundant (NR) protein database of the NCBI and the protein domains found on the translated products, as indicated by their best SMART or Pfam domain match. Table 1A reports on the 118 clusters that are associated with probable housekeeping (H) function, Table 1B on 38 clusters probably associated with a secreted (S) product, and Table 1C on 82 clusters referring to products of unknown (U) function. The electronic table, available by e-mail request, contains the complete table and its links as described in methods. Although the determination of a cluster as being of the H, S, or U class is somewhat arbitrary, it is interesting to observe that the clusters in the S group have an average of 4.6±0.782 sequences (mean±S.E.M.), while the H and U clusters average 1.33±0.087 and 1.63±0.236 sequences per cluster, respectively. Accordingly, while only 14.8% of the clusters are arbitrarily indicated as related to secretory products, the totality of the sequences within these clusters represent 35.3% of the set comprising all 456 sequences. When tested by the nonparametric Kruskal–Wallis one-way analysis of variance assay (ANOVA), these differences were highly significant ($P<0.01$). Multicomparisons performed by the Dunn's test indicated that the S group was significantly different from the other two groups ($P<0.05$). It is also interesting that when comparing the sequences of the 238 clusters with known proteins in the NR protein database, 225 (95%) were not described before as belonging to *Ae. aegypti*, although it is possible that related nucleotide sequences have been deposited as expressed sequence tags (EST) in other databases.

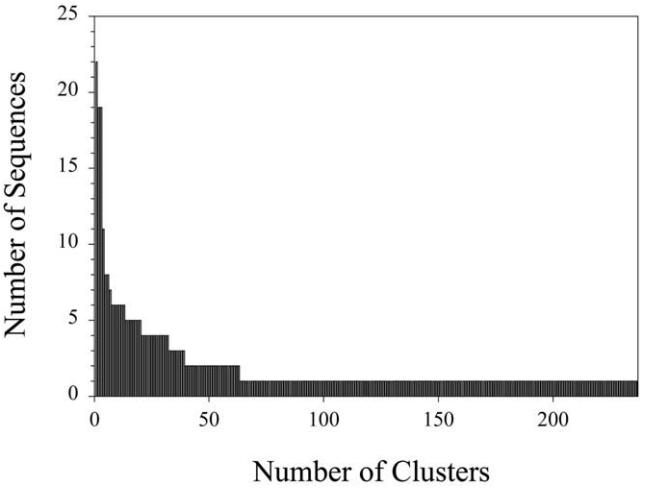

Fig. 1. Number of similar cDNA sequences within a cluster as a function of cluster number, sorted by abundance of sequences per cluster. Clusters were grouped by the output of the BlastN program with a cut off of 1E-60. For more details, see Methods.

Among the probable housekeeping cDNA sequences found in our database, many different ribosomal proteins, t-RNA synthases, elongation factors, heat shock and chaperone proteins such as cyclophilin and chaperonin were encountered, as well as many enzymes linked to respiratory metabolism and three cDNA sequences probably associated with subunits of the V-ATPase. A cDNA coding for a protein with high similarity to tryptophan hydroxylase was also found. Indi-

cation of many other enzymes or proteins are indicated in the column labelled Comments in Table 1A. The electronic version of the table has additional columns with hyperlinks to the best match of the NR database, links to all NR matches found for the cluster, as well all matches to the CDD. It also contains FASTA formatted files for each cluster, and CLUSTAL aligments of each cluster having two or more sequences.

Inspection of the clusters arbitrarily assigned to the group of cDNA sequences probably associated to secreted proteins (Table 1B) reveals the presence of 4 cDNA clusters related to proteins of the antigen-5 family (Schreiber et al., 1997) (all not previously described in *Aedes aegypti*), four clusters related to the D7 protein (James et al., 1991) (3 of which are novel members), two clusters associated with the salivary factor Xa-directed anticlotting of *Ae. aegypti* (Stark and James, 1998) (one of which is a new member), 2 clusters related to the *Ae. aegypti* 30-kDa salivary allergen (Simons and Peng, 2001) (of which one cluster is a new member), two clusters having the C-type lectin signature and related to the macrophage mannose receptor (none described before in *Aedes*), and two clusters having a fibrinogen domain and related to vertebrate angiopoietins (Hackett et al., 2000) (none previously described in insects). In addition to two clusters associated with the previously described salivary apyrase of *Ae. aegypti* (Champagne et al., 1995) (one of which is a novel member), six other clusters were found, each associated with a different enzyme as follows: 1) PAF-acetyl hydrolase, an enzyme destroying the platelet-aggregating compound PAF (Derewenda and Ho, 1999); 2) sphingomyelin phosphodiesterase; 3) carboxylesterase, an enzyme activity previously described in saliva of *Aedes* (Argentine and James, 1995); 4) amylase; 5) glucosidase, two previously described enzymes related to mosquito sugar feeding (James et al., 1989; Grossman and James, 1993); and 6) purine hydrolase, an enzyme thought to occur only in plants, prokaryotes and protozoa (Versees et al., 2001). Additionally, we found four clusters associated with proteins or enzymes related to invertebrate immunity (none of which has been previously described in *Aedes*) as follows: 1) lyzozyme; 2) gram-negative binding protein, a protein involved in initial steps of prophenoloxidase cascade activation (phenoloxidase is the enzyme responsible by invertebrate immunity reactions such as melanotic encapsulation) (Kim et al., 2000); and 3) Two serine proteases possibly associated with the amplification steps of the phenoloxidase cascade, including one having very high similarity to an insect prophenoloxidase activating enzyme. Finally, we found four clusters coding for proteins related to 1) calreticulin, a protein described before in the salivary glands of blood-sucking ticks and insects (Jaworski et al., 1996; Jaworski et al., 1995); 2) bacterial adhesion proteins; 3) a mammalian testes protein; and 4) the known *Ae. aegypti* salivary vasodilator, sialokinin (Champagne and Ribeiro, 1994; Beerntsen et al., 1999). From these 35 clusters, 27 have a coding region translating for proteins not previously described in *Ae. aegypti*.

To determine the main proteins expressed in adult female *Ae. aegypti* salivary glands, we submitted salivary gland homogenates to SDS-PAGE, transferred the proteins to PVDF membranes, and attempted to obtain Edman degradation information on the Coomassie Blue-stained bands that were cut out of the membrane. The gel shows relatively low complexity, with ~ 15 clearly visible stained bands. Aminoterminal information was successfully obtained for eleven of such protein bands, 10 of which were found in the translation products of clusters containing two or more sequences (these 10 sequences are shown in Fig. 2). The exception was for the sequence X-D-I-K-Y-A-D-K-D-F-L-M-K-Q-K-F for which no coding cDNA was found in the mosquito salivary gland cDNA library. This sequence has 93% identity with the aminoterminal region of insect hexamerins (such as *Drosophila* CAA27066 sequence), a reserve hemolymph protein normally made in the fat body and that could have contaminated the salivary glands, or that could have been uptaken by the salivary glands (Burmester and Scheller, 1999; Capurro et al., 2000).

The 10 clusters corresponding to the amino terminal protein sequences found in SDS-PAGE gel bands have an average sequence number of 6.4±1.7 (mean±S.E.M.; n=10), a number significantly larger than the average sequence numbers of the H and U groups ($P<0.05$, Dunn's multicomparison test), but not significantly different from the arbitrarily selected S group of cDNA clusters. Of the ten clusters related to the determined aminoterminal sequences shown in Fig. 1, eight were from clusters preassigned to probably related to secreted products, and two were from the undetermined group of clusters.

Three of the ten amino terminal sequences obtained refer to previously described proteins, apyrase (NCBI accession number gi|556272), the 30-kDa salivary gland allergen (gi|2114497), and the D7 protein (gi|159559). Of the seven novel sequences, two D7-related proteins were found and labeled D7B and D7C in Fig. 2. Two low molecular weight peptides were found, matching clusters that gave no significant protein matches or domain matches on the NR or CDD databases, respectively. These are named AE01 and AE02 in Fig. 2. An additional protein related to the described 30 kDa salivary allergen of *Aedes* was found. This protein has no known function. The aminoterminal sequence DENHSILI… matched cDNA sequences coding for proteins with similarity to angiopoietin. These proteins are involved in the growth regulation of vascular tissues in mammals and were never described before in blood-sucking insects or ticks. Finally, the sequence NYXDQ… matched cDNA sequences that coded for proteins related to the antigen-5 family of proteins.

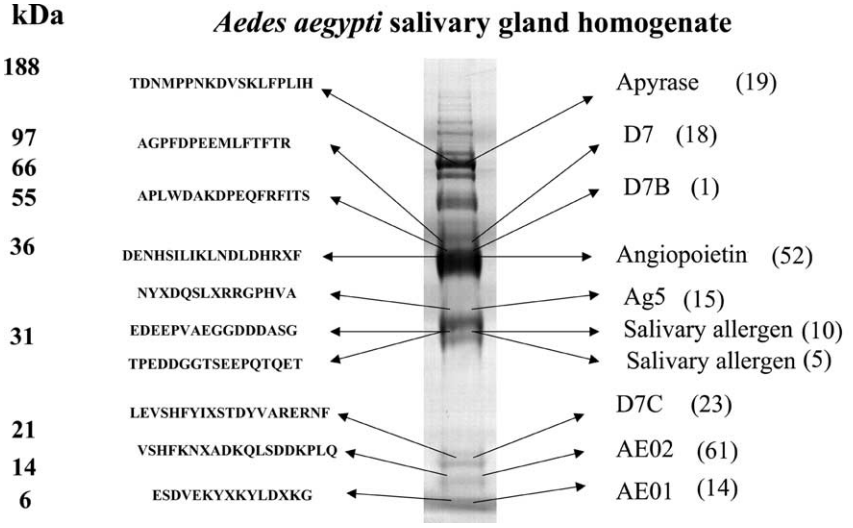

Fig. 2. Coomassie Blue-stained PVDF membrane resulting from transfer of SDS-PAGE of 80 pairs of homogenized glands of adult female *Aedes aegypti*. Molecular weight markers are shown in the left side of the figure, as well as the amino acid sequences obtained by Edman degradation for each band. The right side of the figure indicates the cluster of cDNA sequences where a match was found, and their number of sequences in parenthesis.

cDNA sequences coding for this family of proteins have been cloned from the salivary glands of blood-sucking sand flies and tsetse *Glossina morsitans*. Their function in blood feeding is unknown.

To investigate further the nature of the cDNA sequences possibly coding for secreted proteins, we selected 32 clones to obtain their full-length sequence. These clones were selected based on whether they matched N terminal sequences described in Fig. 2, whether they were arbitrarily selected as possibly secreted and having a role in preventing host hemostasis or invertebrate immunity, and whether they belonged to either the uncertain or housekeeping group of clusters but having a large number of sequences in the cluster. Primary protein sequence deducted from translation of the full-length cDNA sequences were compared with the NR protein database using the BlastP program; their domain structure was further analyzed with RPSBlast using the CDD database. The molecular weight (MW) of the putative protein was calculated, as well as the presence or not of a signal peptide (Nielsen et al., 1997). In the event a signal peptide was predicted to exist, the MW of the mature protein was also calculated. Finally, the pI of the reduced mature protein was calculated. Table 2 reports the summary information for 32 novel cDNA clones thus fully sequenced.

The most abundant salivary cDNA (cluster 1) coded for protein sequence having high similarity to the D7 protein family (Table 2). This new D7 protein is named D7Bclu1. Another cDNA coding for a truncated member of the D7 protein family was D7Cclu23. Both of these D7-related proteins, as well as the previously known D7 protein (NCBI accession number gi|159559), had their aminoterminal sequence found by Edman degradation as shown in Fig. 1. Further, the gel mobilities of the proteins are consistent with the MW deduced from the protein sequence of the mature peptides (Table 2). Together, these data indicate that the translation products of the two novel D7 cDNA sequences are expressed abundantly in *Ae. aegypti* salivary glands. We additionally found 2 new short D7 proteins that are similar to D7Cclu23, but have no similarities to other proteins in the NR database. Fig. 3 shows the alignment of the 5 mature D7 proteins. Note the conservation of the cysteine residues and that D7Cclu23, AEA—clu61 and AEA—clu8 are shorter, truncated versions of the other two proteins. It is interesting that cDNA sequences coding for short D7-like proteins have been observed from salivary gland cDNA libraries of anopheline mosquitoes (Arca et al., 1999). The D7 protein family has no known function. It belongs to the superfamily of odorant binding proteins (Rothemund et al., 1999; Hekmat-Scafe et al., 2000; Graham et al., 2001) and may function by binding mediators of vertebrate hemostasis to help blood feeding, as is seen in several members of the lipocalin family in the saliva of hematophagous hemiptera or ticks that bind histamine (Ribeiro and Walker, 1994; Paesen et al., 1999) or adenine nucleotides (Francischetti et al., 2000). Unique to the short D7 protein (D7Cclu23) is the KGD triad (shown in reverse colors in Fig. 2) flanked by nearby cysteines, a motif found in disintegrins (more commonly found as RGD rather than KGD), which are proteins that interfere with fibrinogen binding to platelets (Scarborough et al., 1993). Accordingly, D7Cclu23 could have a platelet inhibitory activity or otherwise disrupt assembly of fibrinogen domain-containing proteins. We have recently analyzed the different forms of the D7 family of mosquitoes and sand flies, which included D7Cclu23 and D7Bclu1 (Valenzuela et al., 2002).

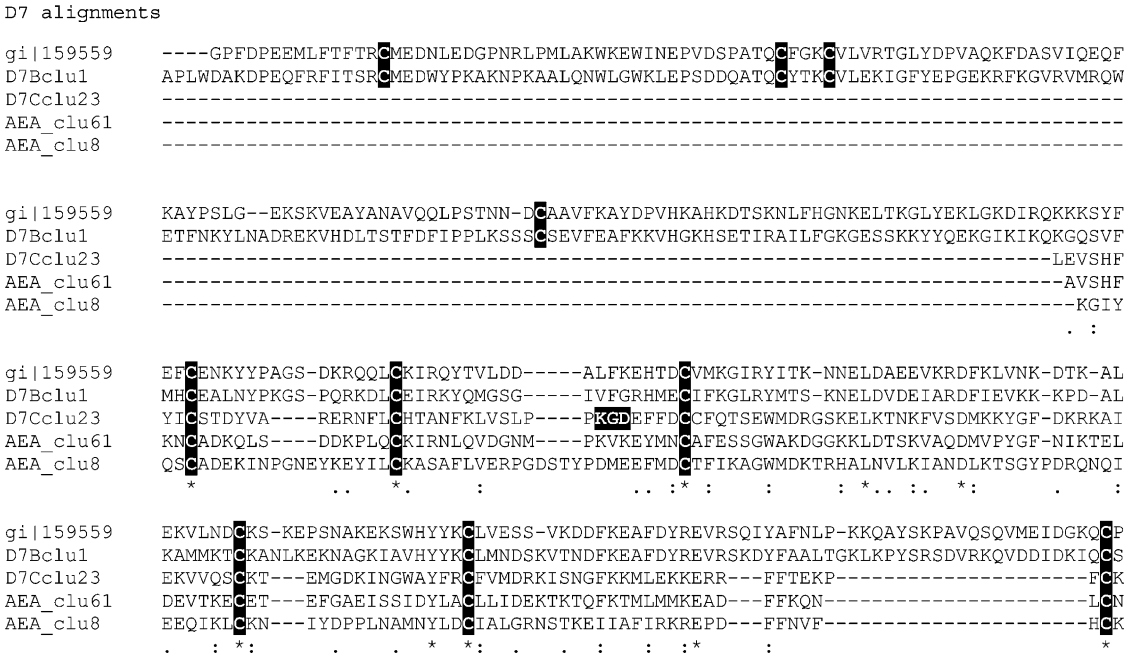

Fig. 3. Clustal alignment of the mature (signal peptide removed) novel D7 protein sequences (D7Bclu1, D7Cclu23, AEA—clu61 and AEA—clu8) with the known D7 protein (AE—D7|159559|, NCBI acession number gi|159559). The symbol ∗ indicates identity between the sequences, whereas the symbol : indicates conserved amino acid substitution. The conserved cysteines are marked in gray background.

The fifth most abundant cDNA cluster coded for a protein sequence having high similarity to the previously described 30-kDa salivary gland antigen of *Ae. aegypti* (gi|2114497) (Simons and Peng, 2001) (Table 2). AEA—Clu5 is a shorter relative of gi|2114497, having a predicted MW of 21.4 kDa, while gi|2114497 is predicted to have a mature mass of 25 kDa.The alignment of AEA—Clu5 with gi|2114497 is shown in Fig. 4. Note the conservation of the 4 cysteines and other residues. The function of this family of proteins is unknown.

AEA—Clu11 codes for a protein with high similarity to the antigen-5 precursor of the tsetse *Glossina morsitans* (gi|8927462) (Li et al., 2001), a member of the single-cell family of proteins (SCP), containing extracellular proteins of unknown function (Schreiber et al., 1997). A signal peptide indicative of secretion is found, producing a predicted mature MW of 27.3 kDa (Table 2). The aminoterminal sequence NYCD… was detected from a gel band corresponding to a MW of 32 kDa. Three additional full-length clones were sequenced with similarity to antigen-5 proteins: AEA—Clu15, AEA—Clu34, and AEA—Clu42 (Fig. 5). These clones have nucleotide sequence coding for proteins with a signal peptide, have mature molecular mass of 26.5–27.3 kDa, and are all basic in nature, with pI above 9.2. Their alignment is shown in Fig. 4. Note the conservation of the eight cysteines and a centrally located tryptophan. The possible function of this widespread protein family in feeding is unknown.

AEA—Clu13 codes for protein with high similarity to purine hydrolases (Table 2). It contains a putative signal peptide, and the mature protein has a predicted MW of 36.3 kDa and a pI of 4.82. This enzyme was previously found only in plants, bacteria, and protozoa but not in

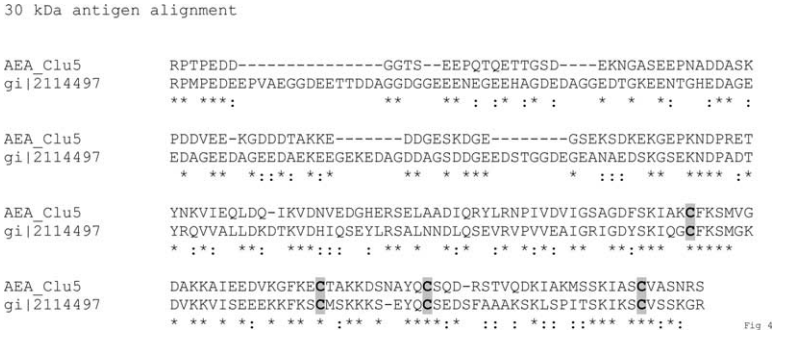

Fig. 4. Clustal alignment of the mature novel protein sequence (AEA—Clu5) related to the previously described 30-kDa salivary allergen of Aedes aegypti (NCBI accession number gi|2114497). Symbols as in Fig. 3.

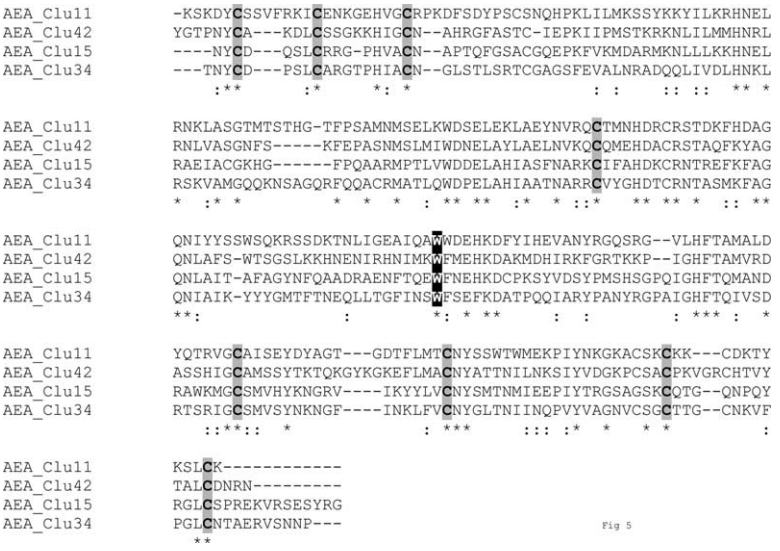

Fig. 5. Clustal aligments of 4 novel sequences containing similarity to proteins of the antigen-5 family of proteins. Symbols as in Fig. 3. A conserved tryptophan residue is also marked.

animals (Versees et al., 2001). We report elsewhere (Ribeiro and Valenzuela, submitted) on the salivary purine hydrolase activity of *Ae. aegypti*, which is one of the richest known sources of this enzyme. The possible role of this enzyme during blood feeding-combined with the recently described salivary adenosine deaminase of the same mosquito (Ribeiro et al., 2001), probably coded by full-length clone sequence here reported in AEA—Clu54 (Table 2)—is to complete the catabolism of the vertebrate host adenosine to hypoxanthine, thus destroying a mediator of mast cell degranulation (Tilley et al., 2000). AEA—Clu150 codes for a secreted carboxylesterase, consistent with the previously described abundant esterase activity of adult female *Aedes aegypti* (Argentine and James, 1995).

Two cDNA clusters, AEA—Clu30 and AEA—Clu39, are possibly related to antihemostatic activity (Table 2). Both have predicted signal peptides indicative of secretion. AEA—Clu30 has a domain signature of serpin (serine protease inhibitor), similarly to the known *Ae. aegypti* FXa-directed salivary anticlotting protein (Stark and James, 1998). The alignment of these two proteins is shown in Fig. 6. The specificity of this novel serpin remains to be investigated. AEA—Clu39 is a member of the calreticulin family (Table 2), secreted forms of which have been shown to have antithrombotic and angiogenic activities (Johnson et al., 2001).

Two cDNA sequences were found (AEA—Clu25 and AEA—Clu52) coding for different proteins having both a fibrinogen domain and high similarity to proteins of the angiopoietin family (Table 2). The aminoterminal DENH… shown in Fig. 1 matches the aminoterminal of the mature translation product of AEA—Clu52, indicating that this protein is expressed and probably secreted. Fig. 7 shows the alignments of two mosquito angiopoietins with two mammalian angiopoietin-related proteins. Note the remarkable conservation of tryptophan and cysteine residues. Also, these two mosquito angiopoietins represent truncated versions of the mammalian angiopoietins, containing only their carboxyterminal moieties. Nearly 200 residues of the aminoterminal portion of the mammalian angiopoietins are lacking in the mosquito forms. Angiopoietins are proteins associated with angiogenesis, some stimulating (such as angiopoietin-1) and some inhibiting (such as angiopoietin-2) this process (Hackett et al., 2000). Their aminoterminals have a coiled-coil region, and their carboxyterminal regions have a fibrinogen domain, the region represented in these novel mosquito translation products. A role for these truncated mosquito molecules in angiogenesis and tissue repair remains to be elucidated.

AEA—Clu44 codes for a protein having the C-type lectin signature, and has high similarity to proteins annotated as lectins in the NR database (Table 2). Hemaglutinins have been described in the salivary glands of anopheline mosquitoes (Johnson et al., 2001), but are not known to exist in *Ae. aegypti* mosquitoes. C-type lectin molecules have been recently described in the salivary glands of phlebotomine sand flies, one such molecule having been purified with the salivary anticlotting activity of *Lutzomyia longipalpis* (Charlab et al., 1999). Some snake venom anticlotting proteins are also of this class of proteins (Arocas et al., 1997). Lectins are also known to participate in invertebrate immunity reactions (Lackie, 1980; Ratcliffe, 1985). Within this context, it is interesting that several cDNA clones were found in this work coding for proteins associated with mosquito immunity, as follows: 1) AEA—Clu16, coding for a lysozyme enzyme; 2) AEA—Clu20, coding for a protein with very high similarity to the putative gram-negative

Fig. 6. Clustal aligment of the novel salivary serpin (AEA—Clu30) with the Factor Xa-directed anticlotting of *Aedes aegypti*, gi|3411116.

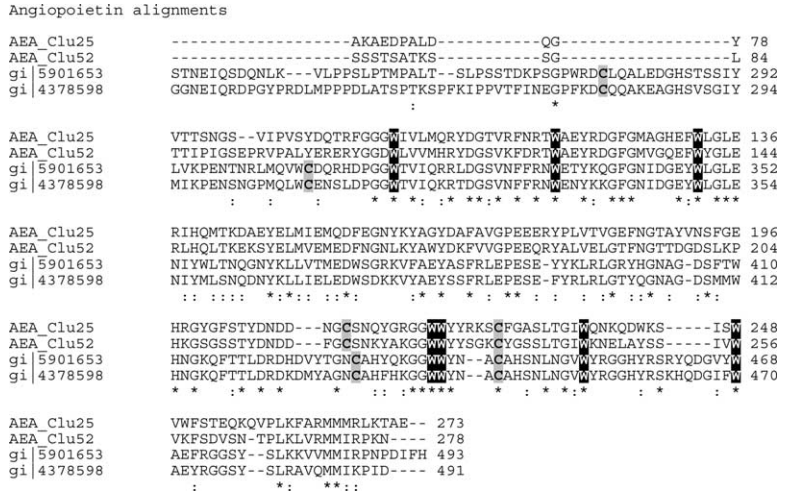

Fig. 7. Clustal alignments of two novel salivary proteins (AEA—Clu25 and AEA—Clu52-predicted mature proteins) containing similarity to the carboxyterminal region of mammalian angiopoietins (gi|5901653 and gi|4378598). Symbols as in Fig. 3. Conserved cysteines and tryptophan residues are shown in gray and black background, respectively.

binding protein of the mosquito *Anopheles gambiae* (Dimopoulos et al., 1997; Richman et al., 1997; Muller et al., 1999), 3) AEA—Clu85, a serine protease with high similarity to prophenoloxidase-activating enzymes (Jiang et al., 1998); 4) AEA—Clu32 and AEA—Clu58 code also for secreted serine proteases; and 5) AEA—Clu35, coding for a 6.8-kDa peptide with high similarity to a peptide sequence deposited in GenBank and annotated as *An. gambiae* infection-responsive short peptide.

The second most abundant cDNA in our database, AEA—Clu2, codes for a protein containing a signal peptide (mature protein molecular weight=56.5), indicating its secretory role. It has weak sequence similarity to *Rickettsia* antigen. An additionally possibly secreted protein is coded by AE—Clu26, with similarity to the the malarial parasite *Plasmodium* adhesion protein. The function of these proteins are unknown. Similarly, AEA—Clu4, AEA—Clu24, AEA—Clu26, and AEA—Clu45 all have putative signal peptides indicative of secretion but have no known indication of their function by similarity searches of the databases AEA—Clu28 codes for a mature protein of over 60 kDa with some similarity to a human testes antigen, the function of which is also unknown. AEA—Clu14 and AEA—Clu17, both having predicted signal peptides indicative of secretion, code for peptides under 10 kDa in size. Amino terminal sequences for both peptides were found by Edman degradation as shown in Fig. 1. Their gel location is also consistent with their predicted molecular weight, as is the predicted aminoterminal sequence fol-

lowing removal of the signal peptide. These peptides have a unique sequence and their possible functions are also unknown. AEA—Clu7 was found to give no matches to the NR protein database, and its putative translation product has an excessive number of stop codons in all frames. Its nucleotide sequence matched MITE transposable elements (Tu, 1997, 2001) (gi|9754772 *Ae. aegypti* clone 101, Pony-Aa-A1 MITE repeat region) when the sequence was compared to the nonredundant EST nucleotide database (NCBI). It is very likely that this abundant transcript (the 7th most abundant cluster) represents a non-coding RNA (ncRNA), which have been implicated, among other things, in gene regulation (Eddy, 2001).

We initiated full-length sequencing of the single sequence found in cluster 225, which yielded similarity to PAF-acetyl hydrolase (Table 1); however, we found that sequence to yield a translation polypeptide sequence not having a secretion signal peptide. Furthermore, the sequence had very high similarity to intracellular PAF-acetyl hydrolases, indicating that PAF may be a signal transduction molecule in mosquito salivary glands. For this reason, full-length sequencing of this clone was discontinued.

4. Discussion

Presently, there are 6 different sequences deposited in GenBank and known to be expressed in the salivary glands of *Ae. aegypti*. These are for amylase (Grossman and James, 1993), alpha-1,4 glucosidase (James et al., 1989), apyrase (Champagne et al., 1995), D7 protein (James et al., 1991), 30-kDa allergen (Simons and Peng, 2001), and the vasodilator sialokinin (Beerntsen et al., 1999; Champagne and Ribeiro, 1994). To extend the knowledge on the compositional diversity and complexity of the salivary glands of this mosquito, a two-pronged approach was used in this paper: A full-length, unidirectionally cloned, PCR-based library of the salivary glands was made from 80 pairs of salivary glands, and ~ 500 clones were randomly sequenced. From this library, 32 novel clones were fully sequenced, 31 of which have a putative signal peptide indicative of secretion, and one of which is a transposable element product. In another approach, 20 pairs of glands were submitted to SDS-PAGE. Ten proteins yielded aminoterminal sequence with matching cDNA sequences from the library clusters, seven of which were not described before, including two new members of the D7 protein family, one related to the 30 kDa *Aedes* allergen, one related to the antigen-5 protein family, two novel peptides, and one angiopoietin-related proteins. The 456 randomly sequenced clones reported in this paper contained all six previously known salivary proteins and identified ten of the 11 aminoterminal sequences obtained by Edman degradation of SDS-PAGE-separated salivary proteins. The only amino terminal sequence that was not present in the cDNA library, as pointed out in the results section, is most likely a hexamerin, an abundant hemolymph storage protein that is normally synthesized in the insect fat body and that may have contaminated or have been uptaken by the salivary glands. Taken together, the library construction approach, without normalization or subtraction, can yield useful information about the predominant proteins from insect salivary glands. The success of this approach may result from the relatively extreme low complexity of the salivary gland proteins, and the correlation displayed by mRNA and their coded protein in the case of abundantly translated proteins (Gygy et al., 1999).

Of the 31 novel protein sequences that appear to be related to secreted proteins in the salivary gland of *Ae. aegypti* (Table 2), inclusive of seven whose protein expression was confirmed by Edman degradation, we can ascribe a function to less than one half. The enzymes adenosine deaminase, purine hydrolase, carboxylesterase, and lysozyme are probably the sources of the enzymatic activities previously described (Rossignol and Lueders, 1986; Argentine and James, 1995; Ribeiro et al., 2001), (Ribeiro and Valenzuela, submitted). Adenosine deaminase hydrolyses the mast cell-degranulating adenosine to inosine, and purine hydrolase hydrolyses inosine to hypoxanthine and ribose. The cDNA coding for a lysozyme is consistent with the previously reported lysozyme activity found in the anterior sugar feeding regions of the salivary glands (Rossignol and Lueders, 1986) postulated to have a role in controlling bacterial growth in the sugar meal ingested with saliva and stored in the mosquito crop (Clements, 1992, pp. 536). Lysozyme is a poor bactericidal compound, and its presence is usually associated with other more potent compounds. Of interest, several proteins, such as the three secreted salivary serine proteases, are probably associated with the activation of the polyphenol oxidase activation cascade, responsible to the melanotic immune reactions of arthropods. Indeed, AEA—Clu85 is very similar to *Manduca* prophenoloxidase-activating enzyme (Jiang et al., 1998). A protein associated with recognition of gram-negative bacteria was also found, as were a C-type lectin and a peptide induced in mosquito infection (for more details see Results). Together, these data indicate that mosquito saliva may have an active polyphenol oxidase-activating system. It cannot be excluded though, that these serine proteases may be playing other roles, such as activation of anti-hemostatic pathways of which the C protein or plasmin activation could be their targets. The novel serpin described in this paper may have an anticlotting effect in addition to that already reported for the FXa-directed anticlotting of *Aedes* (Stark and James, 1998) or it may act in the regulation of the salivary polyphenol-activating cascade. The identification of calretic-

ulin, as well as the two angiopoietins, may indicate that saliva affects endothelial cells and/or plays a role in the cellular matrix of the skin (Pike et al., 1998; Yancopoulos et al., 2000), a role not yet determined in the salivary glands of blood-sucking arthropods. We cannot ascribe a function for the remaining 16 proteins described in this paper, representing 55% of the 29 sequences here reported, although a KGD domain found in the short D7 protein is indicative of an interaction with a fibrinogen receptor. These proteins and peptides with unknown function represent some of the most abundantly expressed in the salivary glands, as revealed by the Edman degradation experiment of SDS-PAGE-separated proteins (Fig. 2). Although the emphasis in mosquito salivary gland research has been on proteins associated with sugar digestion or vertebrate hemostasis, it is possible that these novel molecules may play a role elsewhere, as in preventing mast cell degranulation, a notion reinforced by the presence of the enzymes adenosine deaminase and purine hydrolase as well as by a previously reported inhibitory activity of *Aedes* salivary homogenates on TNF production by mast cells (Bissonnette et al., 1993). A role in angiogenesis, as well as antibacterial or antifungal activity are other possible functions of salivary proteins, as indicated above.

It should be remarked that the functional discussion above assumes that the cDNA sequences described in this paper are of a salivary, secretory nature, an assumption that we have no proof for the majority of the putative proteins. Development of antibodies, perhaps using DNA vaccines (Valenzuela et al., 2001), may help to localize such proteins in the secretory vacuoles of the mosquito salivary glands. Transcriptome research is descriptive in nature and lacks an 'a priori' hypothesis. It uses modern computational tools to analyze all transcripts from a given tissue leading to a wealth of data from which new hypothesis about salivary glands can be drawn.

The search for a function within this majority group of salivary proteins will dominate post-sialome research. The availability of a full-length sequence database for the majority of the salivary gland proteins allows a novel approach to this subject, rather than the former tedious methodology based on HPLC purification and sequencing starting with thousands of pairs of salivary glands. Because the complexity of the salivary glands is low, it is probable that a full protein mass estimate will be useful to produce only one or a few candidate proteins. Indeed, the majority of the proteins of unknown function are abundantly expressed in the salivary glands. High-throughput protein expression or DNA vaccine construction can help in expressing the candidate proteins or to create immunoreagents that can block or adsorb the activity to confirm the identity of the candidate protein. Combining these approaches will permit very fast identification of novel biologically active compounds in the salivary glands of blood-sucking insects and ticks and a better understanding of the evolutionary path these organisms took toward adaptation to this unique diet.

This work may also help to advance the understanding of mosquito salivary gland protein evolution. It has been proposed that salivary proteins of blood-sucking diptera may have an unusual degree of polymorphism because the insect may benefit from antigenic variation of its salivary proteins (Lanzaro et al., 1999). In this regard, the multiple clustal alignments of the cDNA sequences within library clusters having four or more sequences will be useful for those researchers interested in the population genetics of such proteins. These aligments are available in the electronic version of Table 1. For example, inspection of the abundant cluster 11 (file AEn2e60-4aln.txt) shows several base changes and one sequence with a single insertion of three bp. cDNA sequences of housekeeping genes (nuclear and/or mitochondrial), also indicated in the electronic Table 1, can serve as control genes not under the pressure of antigenic variation. Such alignments may become the starting point for further population genetics studies to verify the excess of polymorphism in the secreted salivary proteins of *Aedes aegypti*.

Acknowledgements

We are grateful to Drs Robert Gwadz, Thomas Kindt and Louis Miller for encouragement and support, and to Nancy Shulman for editorial assistance. We are also thankful for helpful comments by two anonymous reviewers.

References

Altland, K., 1990. IPGMAKER: a program for IBM-compatible personal computers to create and test recipes for immobilized pH gradients. Electrophoresis 11, 140–147.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Arca, B., Lombardo, F., de Lara Capurro, M., della Torre, A., Dimopoulos, G., James, A.A., Coluzzi, M., 1999. Trapping cDNAs encoding secreted proteins from the salivary glands of the malaria vector *Anopheles gambiae*. Proc. Natl. Acad. Sci. USA 96, 1516–1521.

Argentine, J.A., James, A.A., 1995. Characterization of a salivary gland-specific esterase in the vector mosquito, *Aedes aegypti*. Insect Biochem. Mol. Biol. 25, 621–630.

Arocas, V., Castro, H.C., Zingali, R.B., Guillin, M.C., Jandrot-Perrus, M., Bon, C., Wisner, A., 1997. Molecular cloning and expression of bothrojaracin, a potent thrombin inhibitor from snake venom. Eur. J. Biochem. 248, 550–557.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., Sonnhammer, E.L., 2000. The Pfam protein families database. Nucleic Acids Res. 28, 263–266.

Beerntsen, B.T., Champagne, D.E., Coleman, J.L., Campos, Y.A.,

James, A.A., 1999. Characterization of the Sialokinin I gene encoding the salivary vasodilator of the yellow fever mosquito, *Aedes aegypti*. Insect Mol. Biol. 8, 459–467.

Bissonnette, E.Y., Rossignol, P.A., Befus, A.D., 1993. Extracts of mosquito salivary gland inhibit tumour necrosis factor alpha release from mast cells. Parasite Immunol. 15, 27–33.

Bjellqvist, B., Basse, B., Olsen, E., Celis, J.E., 1994. Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. Electrophoresis 15, 529–539.

Burmester, T., Scheller, K., 1999. Ligands and receptors: common theme in insect storage protein transport. Naturwissenschaften 86, 468–474.

Capurro, M.L., Moreira-Ferro, C.K., Marinotti, O., James, A.A., de Bianchi, A.G., 2000. Expression patterns of the larval and adult hexamerin genes of *Musca domestica*. Insect Mol. Biol. 9, 169–177.

Champagne, D., Ribeiro, J.M.C., 1994. Sialokinins I and II: Two salivary tachykinins from the Yellow Fever mosquito, *Aedes aegypti*. Proc. Natl. Acad. Sci. USA 91, 138–142.

Champagne, D.E., Smartt, C.T., Ribeiro, J.M., James, A.A., 1995. The salivary gland-specific apyrase of the mosquito *Aedes aegypti* is a member of the 5′-nucleotidase family. Proc. Natl. Acad. Sci. USA 92, 694–698.

Charlab, R., Valenzuela, J.G., Rowton, E.D., Ribeiro, J.M., 1999. Toward an understanding of the biochemical and pharmacological complexity of the saliva of a hematophagous sand fly *Lutzomyia longipalpis* [In Process Citation]. Proc. Natl. Acad. Sci. USA. 96, 15155–15160.

Clements, A.N., 1992. Biology of Mosquitoes: Development, Nutrition and Reproduction. Chapman & Hall, London.

Derewenda, Z.S., Ho, Y.S., 1999. PAF-acetylhydrolases. Biochem. Biophys. Acta. 1441, 229–236.

Dimopoulos, G., Richman, A., Muller, H.M., Kafatos, F.C., 1997. Molecular immune responses of the mosquito *Anopheles gambiae* to bacteria and malaria parasites. Proc. Natl. Acad. Sci. USA. 94, 11508–11513.

Eddy, S.R., 2001. Non-coding RNA genes and the modern RNA world. Nat. Rev. Genet. 2, 919–929.

Francischetti, I.M., Ribeiro, J.M., Champagne, D., Andersen, J., 2000. Purification, cloning, expression, and mechanism of action of a novel platelet aggregation inhibitor from the salivary gland of the blood-sucking bug, *Rhodnius prolixus*. J. Biol. Chem. 275, 12639–12650.

Futcher, B., Latter, G. I, Monardo, P., McLaughlin, C.S., Garrels, J. I, 1999. A sampling of the yeast proteome. Mol. Cell Biol. 19, 7357–7368.

Graham, L.A., Tang, W., Baust, J.G., Liou, Y., Reid, T.S., Davies, P.L., 2001. Characterization and cloning of a *Tenebrio molitor* hemolymph protein with sequence similarity to insect odorant-binding proteins. Insect Biochem. Mol. Biol. 31, 691–702.

Grossman, G.L., James, A.A., 1993. The salivary glands of the vector mosquito, *Aedes aegypti*, express a novel member of the amylase gene family. Insect Mol. Biol. 1, 223–232.

Gygi, S.P., Corthals, G.L., Zhang, Y., Rochon, Y., Aebersold, R., 2000. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. Proc. Natl. Acad. Sci. USA. 97, 9390–9395.

Gygi, S.P., Rochon, Y., Franza, B.R., Aebersold, R., 1999. Correlation between protein and mRNA abundance in yeast. Mol. Cell Biol. 19, 1720–1730.

Hackett, S.F., Ozaki, H., Strauss, R.W., Wahlin, K., Suri, C., Maisonpierre, P., Yancopoulos, G., Campochiaro, P.A., 2000. Angiopoietin 2 expression in the retina: upregulation during physiologic and pathologic neovascularization. J. Cell Physiol. 184, 275–284.

Hekmat-Scafe, D.S., Dorit, R.L., Carlson, J.R., 2000. Molecular evolution of odorant-binding protein genes OS-E and OS-F in Drosophila. Genetics. 155, 117–127.

Henikoff, S., Henikoff, J.G., 1994. Protein family classification based on searching a database of blocks. Genomics 19, 97–107.

Higgins, D.G., Thompson, J.D., Gibson, T.J., 1996. Using CLUSTAL for multiple sequence alignments. Methods Enzymol. 266, 383–402.

James, A.A., Blackmer, K., Marinotti, O., Ghosn, C.R., Racioppi, J. V, 1991. Isolation and characterization of the gene expressing the major salivary gland protein of the female mosquito, *Aedes aegypti*. Mol. Biochem. Parasitol. 44, 245–253.

James, A.A., Blackmer, K., Racioppi, J. V, 1989. A salivary gland-specific, maltase-like gene of the vector mosquito, *Aedes aegypti*. Gene 75, 73–83.

Jaworski, D.C., Higgins, J.A., Radulovic, S., Vaughan, J.A., Azad, A.F., 1996. Presence of calreticulin in vector fleas (Siphonaptera). J. Med. Entomol. 33, 482–489.

Jaworski, D.C.S.F.A., Lamoreaux, W., Coons, L.B., Muller, M.T., Needham, G.R., 1995. A secreted calreticulin protein in Ixodid tick (*Amblyomma americanum*) saliva. J. Insect Physiol. 41, 369–375.

Jiang, H., Wang, Y., Kanost, M.R., 1998. Pro-phenol oxidase activating proteinase from an insect, Manduca sexta: a bacteria-inducible protein similar to Drosophila easter. Proc. Natl. Acad. Sci. USA 95, 12220–12225.

Johnson, S., Michalak, M., Opas, M., Eggleton, P., 2001. The ins and outs of calreticulin: from the ER lumen to the extracellular space. Trends Cell Biol. 11, 122–129.

Kim, Y.S., Ryu, J.H., Han, S.J., Choi, K.H., Nam, K.B., Jang, I.H., Lemaitre, B., Brey, P.T., Lee, W.J., 2000. Gram-negative bacteria-binding protein, a pattern recognition receptor for lipopolysaccharide and beta-1,3-glucan that mediates the signaling for the induction of innate immune genes in Drosophila melanogaster cells. J. Biol. Chem. 275, 32721–32727.

Lackie, A.M., 1980. Invertebrate immunity. Parasitology 80, 393–412.

Lanzaro, G.C., Lopes, A.H., Ribeiro, J.M., Shoemaker, C.B., Warburg, A., Soares, M., Titus, R.G., 1999. Variation in the salivary peptide, maxadilan, from species in the *Lutzomyia longipalpis* complex. Insect Mol. Biol. 8, 267–275.

Li, S., Kwon, J., Aksoy, S., 2001. Characterization of genes expressed in the salivary glands of the tsetse fly, *Glossina morsitans*. Insect Mol. Biol. 10, 69–76.

Marinotti, O., James, A., Ribeiro, J.M.C., 1990. Diet and salivation in female *Aedes aegypti* mosquitoes. J. Insect Physiol. 36, 545–548.

Mellink, J.J., Van Zeben, M.S., 1976. Age related differences of saliva composition in *Aedes aegypti*. Mosquito News 36, 247.

Muller, H.M., Dimopoulos, G., Blass, C., Kafatos, F.C., 1999. A hemocyte-like cell line established from the malaria vector *Anopheles gambiae* expresses six prophenoloxidase genes. J. Biol. Chem. 274, 11727–11735.

Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G., 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. 10, 1–6.

Paesen, G.C., Adams, P.L., Harlos, K., Nuttall, P.A., Stuart, D. I, 1999. Tick histamine-binding proteins: isolation, cloning, and three-dimensional structure. Mol. Cell. 3, 661–671.

Pike, S.E., Yao, L., Jones, K.D., Cherney, B., Appella, E., Sakaguchi, K., Nakhasi, H., Teruya-Feldstein, J., Wirth, P., Gupta, G., Tosato, G., 1998. Vasostatin, a calreticulin fragment, inhibits angiogenesis and suppresses tumor growth. J. Exp. Med. 188, 2349–2356.

Poehling, M.M., 1979. Distribution of specific proteins in the salivary gland lobes of Culicidae and their relation to age and blood-sucking. J. Insect Physiol. 25, 3–8.

Racioppi, J. V, Spielman, A., 1987. Secretory proteins from the salivary glands of adult *Aedes aegypti* mosquitoes. Insect Biochem. 17, 503.

Ratcliffe, N.A., 1985. Invertebrate immunity-a primer for the non-specialist. Immunol. Lett. 10, 253–270.

Ribeiro, J.M., Charlab, R., Valenzuela, J.G., 2001. The salivary adenosine deaminase activity of the mosquitoes *Culex quinquefasciatus* and *Aedes aegypti*. J. Exp. Biol. 204, 2001–2010.

Ribeiro, J.M.C., 1987. Role of arthropod saliva in blood feeding. Ann. Rev. Entomol. 32, 463–478.

Ribeiro, J.M.C., 1995. Blood-feeding arthropods: Live syringes or invertebrate pharmacologists? Infect. Agents Dis. 4, 143–152.

Ribeiro, J.M.C., Walker, F.A., 1994. High affinity histamine-binding and anti-histaminic activity of the salivary NO-carrying heme protein (nitrophorin) of *Rhodnius prolixus*. J. Exp. Med. 180, 2251–2257.

Richman, A.M., Dimopoulos, G., Seeley, D., Kafatos, F.C., 1997. Plasmodium activates the innate immune response of *Anopheles gambiae* mosquitoes. Embo J. 16, 6114–6119.

Rossignol, P.A., Lueders, A.M., 1986. Bacteriolytic factor in the salivary glands of *Aedes aegypti*. Comp. Biochem. Physiol. 83B, 819–822.

Rothemund, S., Liou, Y.C., Davies, P.L., Krause, E., Sonnichsen, F.D., 1999. A new class of hexahelical insect proteins revealed as putative carriers of small hydrophobic ligands. Structure Fold Des. 7, 1325–1332.

Scarborough, R.M., Naughton, M.A., Teng, W., Rose, J.W., Phillips, D.R., Nannizzi, L., Arfsten, A., Campbell, A.M., Charo, I.F., 1993. Design of potent and specific integrin antagonists. Peptide antagonists with high specificity for glycoprotein IIb-IIIa. J. Biol. Chem. 268, 1066–1073.

Schreiber, M.C., Karlo, J.C., Kovalick, G.E., 1997. A novel cDNA from *Drosophila* encoding a protein with similarity to mammalian cysteine-rich secretory proteins, wasp venom antigen 5, and plant group 1 pathogenesis-related proteins. Gene 191, 135–141.

Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P., Bork, P., 2000. SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res. 28, 231–234.

Sibbald, P.R., Sommerfeldt, H., Argos, P., 1991. Automated protein sequence pattern handling and PROSITE searching. Comput. Appl. Biosci. 7, 535–536.

Simons, F.E., Peng, Z., 2001. Mosquito allergy: recombinant mosquito salivary antigens for new diagnostic tests. Int. Arch. Allergy Immunol. 124, 403–405.

Stark, K.R., James, A.A., 1998. Isolation and characterization of the gene encoding a novel factor Xa-directed anticoagulant from the yellow fever mosquito, Aedes aegypti. J. Biol. Chem. 273, 20802–20809.

Tilley, S.L., Wagoner, V.A., Salvatore, C.A., Jacobson, M.A., Koller, B.H., 2000. Adenosine and inosine increase cutaneous vasopermeability by activating A(3) receptors on mast cells. J. Clin. Invest. 105, 361–367.

Tu, Z., 1997. Three novel families of miniature inverted-repeat transposable elements are associated with genes of the yellow fever mosquito, *Aedes aegypti*. Proc. Natl. Acad. Sci. USA 94, 7475–7480.

Tu, Z., 2001. Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, *Anopheles gambiae*. Proc. Natl. Acad. Sci. USA 98, 1699–1704.

Valenzuela, J.G., Charlab, R., Gonzalez, E.C., Miranda-Santos, I.K.F., Marinotti, O., Francischetti, I.M. and Ribeiro, J.M.C. (2002) The D7 family of salivary proteins in blood sucking Diptera. *Insect Molec.* Biol. In press.

Valenzuela, J.G., Belkaid, Y., Garfield, M.K., Mendez, S., Kamhawi, S., Rowton, E.D., Sacks, D.L., Ribeiro, J.M., 2001. Toward a defined anti-leishmania vaccine targeting vector antigens. Characterization of a protective salivary protein. J. Exp. Med. 194, 331–342.

Versees, W., Decanniere, K., Pelle, R., Depoorter, J., Brosens, E., Parkin, D.W., Steyaert, J., 2001. Structure and function of a novel purine specific nucleoside hydrolase from Trypanosoma vivax. J. Mol. Biol. 307, 1363–1379.

Yancopoulos, G.D., Davis, S., Gale, N.W., Rudge, J.S., Wiegand, S.J., Holash, J., 2000. Vascular-specific growth factors and blood vessel formation. Nature 407, 242–248.