Bioinformatics-Based Identification of Chemosensory Proteins in African Malaria Mosquito, *Anopheles gambiae*

Zhengxi Li[1]*, Zuorui Shen[1], Jingjiang Zhou[2], and Lin Field[2]

[1] *Department of Entomology, China Agricultural University, Beijing 100094, China;* [2] *Biological Chemistry Division, Rothamsted Research, Harpenden, Herts. AL5 2JQ, UK.*

Chemosensory proteins (CSPs) are identifiable by four spatially conserved Cysteine residues in their primary structure or by two disulfide bridges in their tertiary structure according to the previously identified olfactory specific-D related proteins. A genomics- and bioinformatics-based approach is taken in the present study to identify the putative CSPs in the malaria-carrying mosquito, *Anopheles gambiae*. The results show that five out of the nine annotated candidates are the most possible *Anopheles* CSPs of *A. gambiae*. This study lays the foundation for further functional identification of *Anopheles* CSPs, though all of these candidates need additional experimental verification.

Key words: chemosensory protein, proteomics, bioinformatics, olfaction, African malaria mosquito, *Anopheles gambiae*

Introduction

Anopheles gambiae (*A. gambiae*) is the principal vector of malaria that afflicts more than 500 million people and causes more than 1 million deaths each year. Analysis of the whole genome of *A. gambiae* revealed strong evidence for about 14,000 protein-coding transcripts, which need further annotation and experimental verification (*1*).

Olfaction plays a key role in host selection of agricultural pests and disease vectors. Recent advancement in understanding the molecular mechanism of olfaction is the result of multidisciplinary research efforts by using a variety of model organisms including insects. The reception of pheromones and general odorants is mediated by specific neurons located in specialized cuticular sensilla in insects (*2*). Chemosensory neurons extend their dendrites to a lymphatic cavity, where the soluble and low molecular weight proteins that are supposed to transfer the hydrophobic odorants across the fluid barrier to the receptive dendritic membrane are contained. Molecular cloning and biochemical surveys of insect antennae have identified two abundant but unrelated families of small soluble proteins with proposed odorant transport function, that is, odorant-binding proteins (OBPs) and olfactory specific-D (OS-D) related proteins (*3*, *4*). OS-D related proteins (average 13 kDa) were first identified by subtractive hybridisation experiments using antennae of *Drosophila melanogaster* (*5*, *6*). Many OS-D homologues were subsequently identified based on sequence similarity in different insect orders (Table 1).

OS-D related proteins differ from OBPs in several aspects: they share no sequence similarity with OBPs and contain only four of the six spatially conserved Cysteine residues that are characterised by OBPs (*4*). Although two of the protein families are both represented by multiple genes within a given species, OS-D related proteins are more conserved than OBPs across evolution or between different phyla, with 40%-50% of identical residues even between most distant species. They have been identified in a variety of tissues, while most OBPs appear to be restricted to olfactory tissues. They are common in Orthopteroid (phasmid and grasshopper) and holometabolous (Lepidoptera and Diptera) insects (Table 1) and thus may be present throughout the Neoptera, while OBPs are only known within the holometabolous and hemipteroid lineages (*4*). There is no strong evidence for the physiological role of OS-Ds so far, and they may be involved in chemical communication and perception. To contrast with OBPs that are found in olfactory sensilla, the OS-D related proteins were designated as Chemosensory Proteins (CSPs; ref. *18*, *23*).

*** Corresponding author.**
E-mail: zxli@cau.edu.cn

Table 1 Previously Identified Insect Chemosensory Proteins*

Order	Species	Protein name[a]	Length(a.a)	Accession No.[b]	References
Hymenoptera	*Apis mellifera*	ASP3c	130	AF481963	*7, 8*
Lepidoptera	*Cactoblastis cactorum*	CLP-1	130	U95046	*9*
	Manduca sexta	SAP1	105	AF117574	*10*
		SAP3	126	AF117585	*10*
		SAP2	127	AF117592	*10*
		SAP5	231	AF117594	*10*
		SAP4	127	AF117599	*10*
	Bombyx mori	BmorCSP2	120	AF509238	*11*
		BmorCSP1	127	AF509239	*11*
	Mamestra brassicae	CSP-MbraA1	112	AF211177	*12*
		CSP-MbraA2	112	AF211178	*12*
		CSP-MbraA3	112	AF211179	*12*
		CSP-MbraA4	112	AF211180	*12*
		CSP-MbraA5	112	AF211181	*12*
		CSP-MbraB1	108	AF211182	*12*
		CSP-MbraB2	108	AF211183	*12*
		CSP-MbraA6	128	AF255918	*13*
		CSP-MbraB3	108	AF255919	*13*
		CSP-MbraB4	108	AF255920	*13*
	Heliothis virescens	HvirCSP2	126	AY101511	*14*
		HvirCSP1	114	AY101512	*14*
		HvirCSP3	106	AY101513	*14*
	Mamestra brassicae	SAP	111	AY026760	unpublished
	Helicoverpa armigera	CSP-Harm	127	AF368375	unpublished
	Helicoverpa zea	CSP-Hzea	128	AF448448	unpublished
Diptera	*D. melanogaster*	A10	155	U05244	*15*
		RH70879p	124	BT001865	unpublished
		PEBmeIII	158	U08281	*16*
	Anopheles gambiae	SAP-1	127	AF437891	*17*
Orthoptera	*Schistocerca gregaria*	CSP-sg1	109	AF070961	*18*
		CSP-sg2	109	AF070962	*18*
		CSP-sg3	103	AF070963	*18*
		CSP-sg4	109	AF070964	*18*
		CSP-sg5	109	AF070965	*18*
	Locusta migratoria	OS-D1	103	AJ251075	*19*
		OS-D2	120	AJ251076	*19*
		OS-D3	125	AJ251077	*19*
		OS-D4	125	AJ251078	*19*
		OS-D5	125	AJ251079	*19*
Phasmatodea	*Eurycantha calcarata*	CSP-ec1	107	AF139196	*20*
		CSP-ec2	102	AF139197	*20*
		CSP-ec3	107	AF139198	*20*
Dictyoptera	*Periplaneta americana*	p10	130	AF030340	*21, 22*

*GenBank (04/2003);
[a]Registered names in GenBank;
[b]GenBank accession numbers.

Compared with the twenty-nine putative *A. gambiae* OBPs characterized for similarity to OBPs of *Drosophila* and other insects (*24*), no bioinformatics-based annotation has been carried out to identify the CSP candidates of *A. gambiae*. The NMR (Nuclear Magnetic Resonance) solution structure of chemosensory protein Csp2 (1K19_A) from moth *Mamestra brassicae* has been established (*23*), which is the best elucidated insect CSP and would be used as a model for homology modelling. We created an algorithm for identifying the conserved domains present in *Anopheles* putative CSPs through Perl programming.

Results

Conserved domain of insect CSP candidates

Exhaustive queries with all previously identified insect CSP sequences retrieved from GenBank (April 2003) and ClustalX multialignment resulted in an absolutely conserved structure for insect CSPs, that is, Cx(6,8)Cx(18)Cx(2)Cx(3) (Figure 1). We used a program developed by the authors through Perl programming to search the local database that contains the Fasta files of *Anopheles* gDNA (genomic DNA) and cDNA (complementary DNA) sequences downloaded from Ensembl Mosquito Genome Server for the identified pattern in their primary structure. Totally eight sequences were hit, including agCP10968, agCP11079, agCP11481, agCP11484, agCP11532, agCP11545, agCP6514, and agCP12965. The hits were in turn corroborated by BLAST searching in GenBank, and the prediction was made for their biochemical properties and secondary structure.

CSP candidates corroborated by BLAST

The hits obtained through pattern searching were corroborated by BLAST in GenBank (http://ncbi.nlm.nih.gov/). Six sequences (agCP10968, agCP 11079, agCP11481, agCP11484, agCP11532, and agCP11545) were found closely related to the previously identified CSPs, whereas two sequences (agCP6514 and agCP12965) matched no other proteins, so the CSP candidacy of agCP6514 and agCP12965 could not be excluded. The most interesting discovery was that a novel *Anopheles* CSP candidate (agCP11435) had been identified by BLAST searching, though the Expect (E) values are not very high (Table 2).

Biochemical properties, prediction of secondary structure and ORFs of *Anopheles* CSP candidates

The cDNA sequences of the *Anopheles* Genome Project stored in GenBank and in the Ensembl Mosquito Genome Browser (ftp://ftp.ensembl.org/pub/current_mosquito/) had been annotated jointly by the privately funded Celera and EBI (http://www.ensembl.org/Anopheles_gambiae/). All the entries are the results from preliminary prediction. Even the positions of start codon for most of the annotated transcripts were not determined in the database. We have predicted the complete coding sequences (CDSs) and open reading frames (ORFs) for the *Anopheles* CSP candidates identified through pattern searching and BLAST, based on the genomic DNA sequences of the candidates. The redefined ORFs were listed in Table 3, along with the corresponding Celera IDs, GenBank IDs (#EAA), chromosomal locations, and scaffold numbers (#AAAB). Positions of the signal peptides, isoelectric points (pI) and hydrophobicity were also presented. The information contained in the table shows that all of the CSP candidates except agCP12965 and agCP6154 are located on one scaffold (AAAB01008964) of the chromosome 3R. This stimulating discovery may indicate that these olfactory genes were duplicated at some point of *Anopheles* evolution. Meanwhile, all of the CSP candidates have small molecular weights (<15 kDa in most cases) and are hydrophilic (<35% hydrophobic amino acids in most cases). Most of the candidates have a signal peptide, though no signal peptides have been found in two of the sequences, that is, agCP10968 and agCP12965. In fact, we have used several different tools to predict the ORFs of agCP10968, unfortunately no signal peptides have been found, which indicates that a sequencing error may have occurred in the *Anopheles* Genome Project.

Table 4 summarized the secondary structure prediction of *Anopheles* CSP candidates. The predictions showed that most of the candidates could be classified as all-alpha structure, with a high probability to form a globular domain. However, agCP10968 and agCP6514 did not appear to be globular.

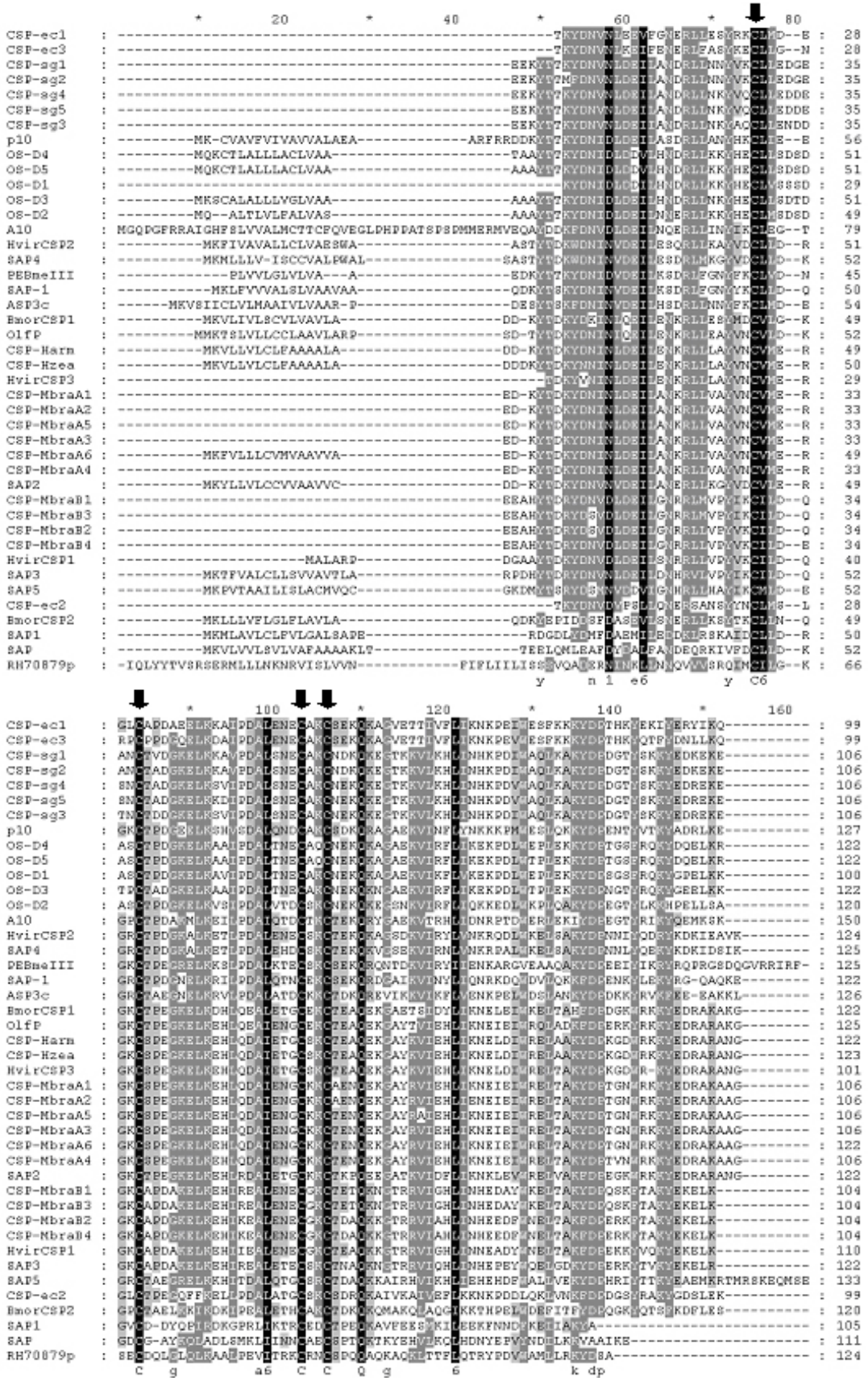

Fig. 1 Multiple alignment showing the absolutely conserved domain in insect CSPs: Cx(6,8)Cx(18)Cx(2)Cx(3). All acronyms in the figure refer to the protein names listed in Table 1.

Table 2 *Anopheles* CSP Candidates Found by CSPMOT and BLAST (E*-values<0.0001)

Peptide ID	Previously identified insect CSPs						
agCP10968	SAP-1	ASP3c	SAP2	CSP-Harm	HvirCSP2	OS-D3	CSP-sg1
	(6e-18)	(7e-18)	(6e-17)	(1e-16)	(2e-16)	(2e-16)	(2e-16)
	OS-D1	CSP-sg4	CSP-sg2	CSP-MbraA6	OS-D4	CSP-Hzea	OS-D5
	(2e-16)	(3e-16)	(3e-16)	(3e-16)	(4e-16)	(5e-16)	(3e-16)
	SAP4	PEBmeIII	CSP-sg5	OS-D2	A10	CSP-ec3	CSP-MbraA3
	(7e-16)	(9e-16)	(9e-16)	(2e-15)	(2e-15)	(2e-15)	(2e-15)
	CSP-MbraA1	CSP-MbraA2	CSP-sg3	p10	CSP-MbraA5	SAP3	
	(2e-15)	(3e-15)	(3e-15)	(3e-15)	(7e-15)	(8e-15)	
agCP11079	SAP-1	PEBmeIII	ASP3c	p10	HvirCSP2	SAP4	A10
	(1e-45)	(1e-34)	(2e-34)	(4e-28)	(3e-27)	(1e-26)	(2e-26)
	OS-D2	CSP-MbraA6	OS-D3	SAP3	BmorCSP1	SAP5	CSP-sg1
	(6e-26)	(2e-23)	(4e-23)	(7e-23)	(8e-23)	(9e-23)	(1e-22)
	CLP-1	CSP-MbraA3	CSP-sg4	HvirCSP1	CSP-sg2	OS-D4	OS-D5
	(2e-22)	(3e-22)	(4e-22)	(5e-22)	(6e-22)	(7e-22)	(7e-22)
	CSP-sg4	OS-D1	CSP-MbraA2	CSP-sg3	SAP2	CSP-MbraA3	agCP11435
	(1e-21)	(1e-21)	(2e-21)	(3e-21)	(4e-21)	(4e-21)	(5e-08)
agCP11481	ASP3c	PEBmeIII	A10	SAP4	HvirCSP2	SAP5	SAP3
	(4e-28)	(3e-27)	(3e-26)	(3e-25)	(4e-25)	(4e-24)	(2e-23)
	CSP-sg2	SAP-1	CSP-sg1	OS-D2	CSP-MbraA6	CLP-1	CSP-sg5
	(1e-22)	(1e-22)	(2e-22)	(3e-22)	(7e-22)	(1e-21)	(2e-21)
	CSP-sg4	CSP-MbraA3	CSP-sg3	CSP-MbraA2	OS-D3	HvirCSP1	CSP-MbraA1
	(2e-21)	(9e-21)	(1e-20)	(2e-20)	(3e-20)	(4e-20)	(6e-20)
	CSP-MbraA3	CSP-MbraB1	SAP2	CSP-MbraA4	CSP-MbraA5	BmorCSP1	agCP11435
	(7e-20)	(8e-20)	(1e-19)	(2e-19)	(2e-19)	(3e-19)	(1e-06)
agCP11484	SAP-1	ASP3c	PEBmeIII	HvirCSP2	p10	CSP-MbraA6	OS-D2
	(3e-55)	(1e-31)	(5e-32)	(6e-29)	(4e-28)	(3e-27)	(9e-27)
	SAP4	BmorCSP1	CLP-1	SAP2	A10	SAP3	SAP5
	(1e-26)	(2e-26)	(2e-26)	(1e-25)	(6e-25)	(4e-24)	(6e-24)
	CSP-Hzea	CSP-MbraA2	HvirCSP1	CSP-MbraA1	CSP-MbraA3	CSP-Harm	CSP-MbraA4
	(3e-23)	(5e-23)	(5e-23)	(7e-23)	(7e-23)	(9e-23)	(2e-22)
	CSP-MbraA5	CSP-sg1	CSP-sg4	CSP-sg2	OS-D3	OS-D1	CSP-ec1
	(2e-22)	(4e-22)	(8e-22)	(1e-21)	(2e-21)	(3e-21)	(3e-21)
	CSP-sg5	agCP11435					
	(3e-21)	(2e-08)					
agCP11532	RH70879	ASP3c	SAP5	CSP-sg4	CSP-MbraA6	CSP-sg5	CSP-sg2
	(5e-30)	(4e-10)	(6e-08)	(6e-08)	(1e-07)	(1e-07)	(2e-07)
	HvirCSP2	CSP-sg1	SAP2	CSP-ec3	BmorCSP2	CSP-sg3	OS-D3
	(2e-07)	(3e-07)	(4e-07)	(7e-07)	(7e-07)	(9e-07)	(1e-06)
	CLP-1	CSP-MbraA3	CSP-MbraA1	CSP-MbraA2	CSP-MbraA4	SAP4	CSP-MbraA5
	(1e-06)	(1e-06)	(1e-06)	(1e-06)	(1e-06)	(2e-06)	(2e-06)
	OS-D2	OS-D1	A10	CSP-Hzea	HvirCSP3	CSP-ec1	
	(3e-06)	(5e-06)	(1e-05)	(1e-05)	(1e-05)	(3e-05)	
agCP11545	SAP-1	PEBmeIII	ASP3c	p10	HvirCSP2	A10	OS-D2
	(3e-40)	(2e-33)	(2e-32)	(1e-30)	(6e-30)	(1e-28)	(1e-27)
	SAP4	CSP-MbraA6	CSP-sg1	BmorCSP1	SAP3	CSP-sg4	CSP-sg2
	(3e-26)	(1e-25)	(3e-24)	(3e-24)	(3e-24)	(6e-24)	(6e-24)

Table 2 ***Continued***

Peptide ID	Previously identified insect CSPs						
agCP11545	OS-D3	SAP5	CSP-sg5	CSP-sg3	HvirCSP1	OS-D4	OS-D1
	(1e-23)	(2e-23)	(3e-23)	(3e-23)	(4e-23)	(5e-23)	(7e-23)
	CSP-MbraA3	CSP-MbraA5	CLP-1	CSP-ec1	CSP-MbraA2	CSP-Hzea	agCP11435
	(2e-22)	(2e-22)	(2e-22)	(5e-22)	(7e-22)	(1e-21)	(7e-09)
agCP6514	no matches[a]						
agCP12965	no matches[a]						
agCP11435[b] A10	OS-D1	OS-D3	OS-D5	OS-D4	SAP-1	SAP4	
	(2e-12)	(5e-10)	(7e-10)	(1e-09)	(2e-09)	(2e-08)	(3e-08)
	PEBmeIII	OS-D2	HvirCSP2	ASP3c	CSP-sg5	CSP-ec2	CSP-sg4
	(3e-08)	(4e-08)	(8e-08)	(1e-07)	(3e-07)	(5e-07)	(5e-07)
	CSP-sg3	CSP-sg2	CSP-sg1	CSP-ec1	CSP-ec3	p10	CSP-MbraA6
	(6e-07)	(9e-07)	(9e-07)	(2e-06)	(2e-06)	(5e-05)	(5e-05)
	CSP-Hzea	CSP-Harm	CSP-MbraA3	CSP-MbraA1			
	(1e-04)	(1e-04)	(1e-04)	(1e-04)			

[a] No significant hits obtained by BLAST;

[b] BLAST identified;

* The E value is a parameter that describes the number of hits one can "expect" to see just by chance when searching a database of a particular size. All acronyms of insect CSPs refer to protein names listed in Table 1.

Table 3 Chromosomal Location, New ORFs, Signal Peptides, and Biochemical Properties of *Anopheles* CSP Candidates

Celera_ID	GB_ID	Chrom	Scaffold No.	Original length (a.a.)	New ORF length (a.a.)	Signal peptide	MW (kDa)	pI	Hydrophobic a.a. (%)
agCP10968	EAA12703	3R	AAAB01008964	127	109	none	12.3	9.5	25.7
agCP11079	EAA12353	3R	AAAB01008964	143	127	1-17	14.8	5.4	33.1
agCP11481	EAA12591	3R	AAAB01008964	137	123	1-19	14.3	9.4	29.3
agCP11484	EAA12322	3R	AAAB01008964	149	127	1-17	14.7	8.6	33.1
agCP11532	EAA12601	3R	AAAB01008964	150	117	1-33	12.9	9.8	41.0
agCP11545	EAA12338	3R	AAAB01008964	141	126	1-17	14.6	8.6	31.0
agCP11435	EAA12702	3R	AAAB01008964	102	137	1-16	15.7	5.0	35.0
agCP12965	EAA05664	3L	AAAB01008834	173	137	none	14.1	3.5	23.4
agCP6514	EAA10937	2L	AAAB01008960	132	117	1-19	13.0	8.6	25.6

Table 4 Secondary Structure of *Anopheles* CSP Candidates*

Peptide ID	Predicted secondory structure (%)				Globularity
	Helix	Sheet	Loop	Class	
agCP10968	25.70	15.50	58.80	mixed	appears not to be globular
agCP11079	72.40	0.00	27.60	all-alpha	may be globular, but it is not as compact as a domain
agCP11481	71.50	0.00	28.50	all-alpha	may be globular, but it is not as compact as a domain
agCP11484	71.70	1.60	26.80	all-alpha	appears as compact, as a globular domain
agCP11532	72.70	0.00	27.40	all-alpha	appears as compact, as a globular domain
agCP11545	68.20	2.40	29.40	all-alpha	may be globular, but it is not as compact as a domain
agCP11435	75.90	0.00	24.10	all-alpha	may be globular, but it is not as compact as a domain
agCP12965	13.10	16.80	70.10	mixed	appears as compact, as a globular domain
agCP6514	10.30	0.00	89.70	mixed	appears not to be globular

* Prediction server—http://www.sbg.bio.ic.ac.uk/3dpssm.

Homology modelling

The NMR solution structure of chemosensory protein Csp2 of *Mamestra brassicae* (1K19_A) was retrieved from Protein Data Bank (PDB ID: 1K19). The 3-D structure of this model molecule is shown in Figure 2A, which is characterized by two disulfide bonds (CysI-CysII, CysIII-CysIV), instead of three disulfide bonds (CysI-CysIII, CysII-CysV and CysIV-CysVI) characterized by OBPs. The model has a typical hydrophobic core that is supposed to act as a pocket for ligand binding. The pocket is formed by hydrophobic amino acids, surrounded by hydrophilic amino acids.

Homology modelling of the *Anopheles* CSP candidates was made in Swiss-PdbViewer (Figure 2). The figures showed that most candidates folded similarly as 1K19_A. However, a structurally weak linkage occurred between the two disulfide bridges of agCP11435 (Figure 2D), though a hydrophobic pocket was formed in the core of its structure. Further sequence analysis was carried out, which showed that the structural abnormality might be caused by a surplus sequence insertion (GRLACLALVL; Figure 3).

Discussion

Sequence similarity in CSP primary structure is significantly higher than that in OBPs. OBP sequences of the same species could be less similar than those of different species, whereas CSP sequences are more conserved at species level as well as between phylogenetically distant groups. In the present study, the sequence similarity between agCP11484 and other CSP candidates is: 77%, 76%, 47%, 39%, 21%, 15%, 13%, and 8%, respectively. Moreover, the positions of two disulfide bridges were highly conserved. The structural conservation should be extremely important in forming a strong hydrophobic core that function as an odorant-binding site.

Thirty-eight OBP candidates have been annotated in *D. melanogaster*, whereas only 29 OBP candidates have been conceptually identified in *A. gambiae*, based on sequencing and genome analysis (*10, 24–26*). In the present study, we located nine CSP candidates in *Anopheles* genome, among which five candidates, *i.e.* agCP11079, agCP11481, agCP11484, agCP11532, and agCP11545, are the most possible *Anopheles* CSPs, considering the sequence similarity in their primary structure, the biochemical properties such as hydrophobicity and molecular weight (<15 kDa), in particular the secondary structure and 3-D

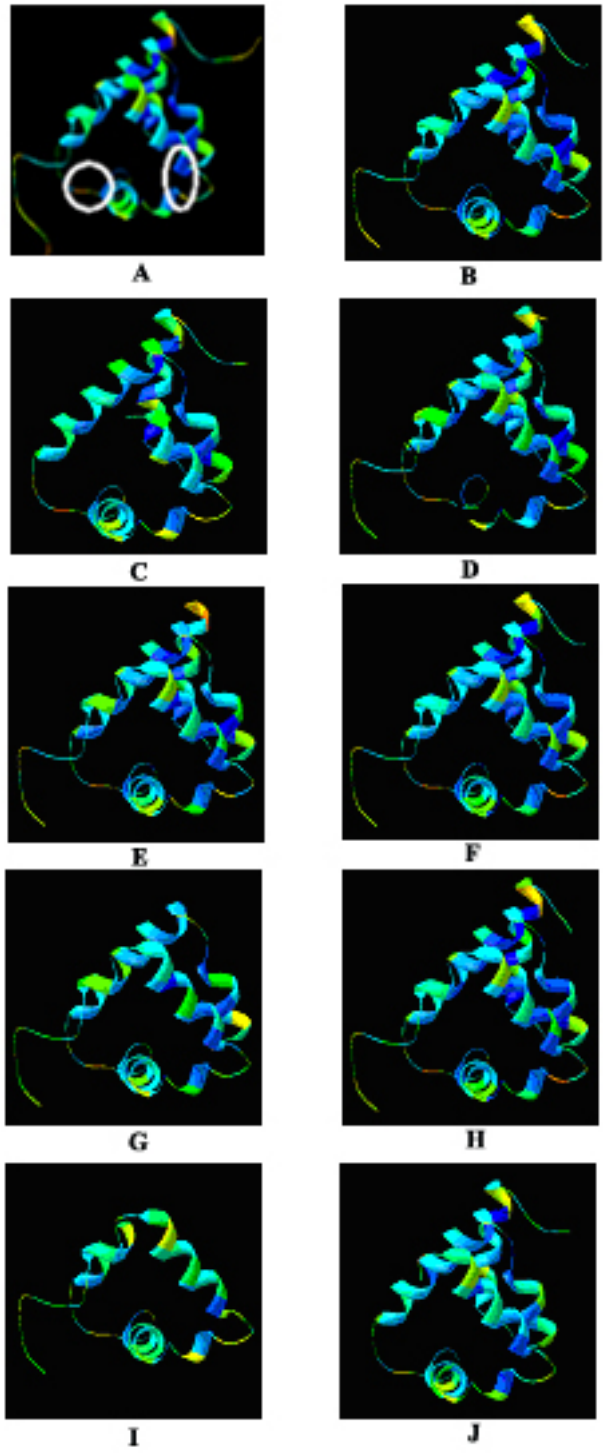

Fig. 2 Homology modelling of *Anopheles* CSP candidates with 1K19_A_MbraCSP (PDB ID: 1K19) as the model. A. 3D-structure of 1K19_A_MbraCSP showing the position of two disulfide bridges; B-J: predicted 3D-structure of *Anopheles* CSPs, including agCP11079, agCP10968, agCP11435, agCP11481, agCP11484, agCP11532, agCP 11545, agCP12695, and agCP6514. The figures are generated in Swiss-PdbViewer.

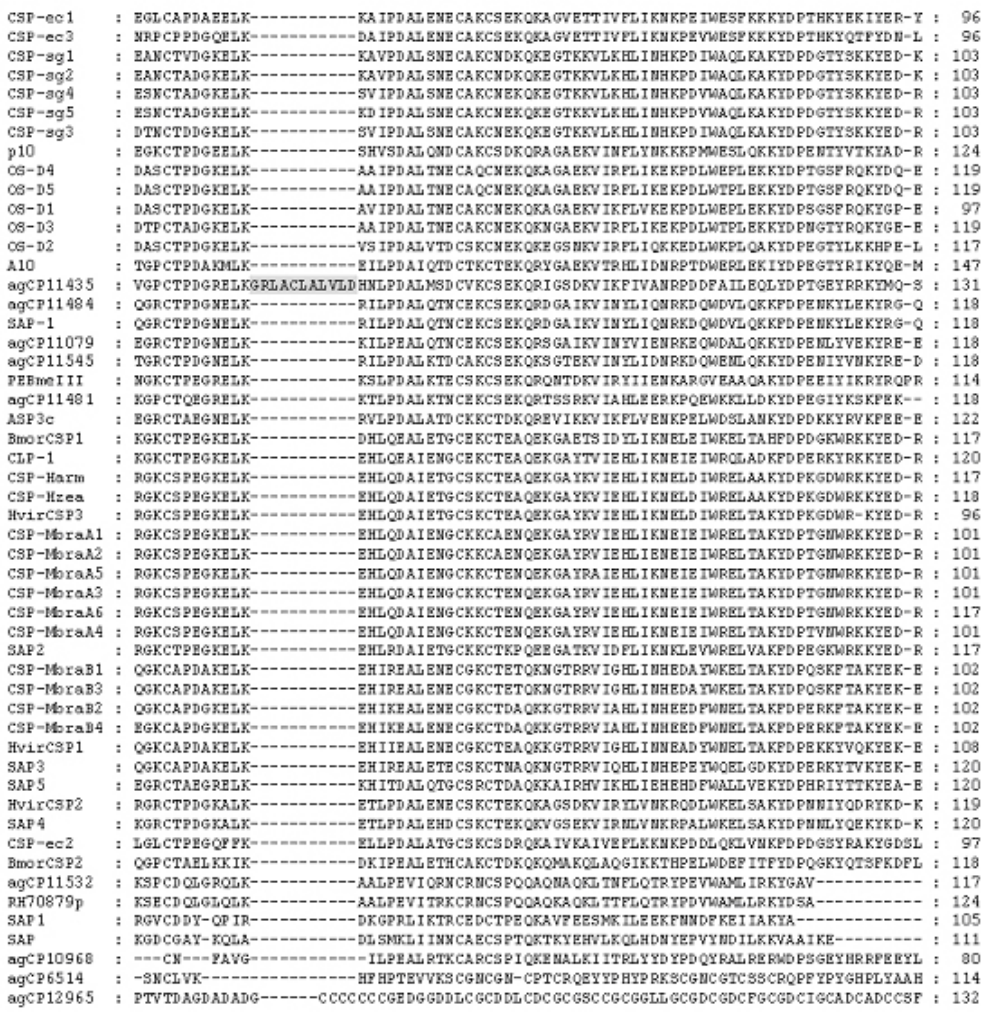

Fig. 3 Partial alignment of candidate *Anopheles* CSPs with the previously identified CSPs shows the surplus sequence of agCP11435 that causes its 3D-structural abnormality. All acronyms of the gene names refer to the protein names listed in Table 1.

structure. It would be very interesting to discuss the number of insect OBPs and CSPs (29 OBPs to 9 CSPs for *Anopheles*), because the numbers may be closely related to the functions. It has been shown that the previously identified OBPs are spatially distributed in insect olfactory sensilla, while CSPs are distributed in different tissues, including non-olfactory tissues. Pheromone-binding proteins (PBPs) are typical insect OBPs, which are highly specific. However, CSPs are less specific, which indicate that one CSP may be able to bind more than one odorant. We postulated that more OBPs might be needed to bind different odorants specifically, although any further hypothesis on the physiological function of CSPs can only be made when reliable experimental evidence has been presented, such as identification of the specific ligands that show CSPs function as chemosensory proteins.

It must be noted that agCP11435 may be a good model for discussion on its structural relations to its functions. We hope agCP11435 is a CSP candidate, but its structural peculiarity may provide some space for exploring its binding behaviour. It is expected that a minor insertion into the structural core may lead to destruction of the fundamental functions of a protein.

No strong evidence has been provided for the physiological role of CSPs so far, though researchers believed they might be involved in chemical communication and perception. We checked the spatial expression pattern of the putative *Anopheles* CSP candidates (data unpublished) and found they were distributed not only in mosquito antennae (olfactory tissues), but also in non-olfactory tissues such as heads stripped off antennae and maxillary palps, legs and

bodies. The preliminary results indicated that CSPs might have other functions than olfaction. Further studies will be focused on functional research, such as ligand identification and crystallization of CSP recombinants.

Materials and Methods

Defining conserved domains in insect CSPs

The amino acid sequences of previously identified CSPs were downloaded from GenBank. All the sequences were then aligned in ClustalX 8.1 (*27*) using Multiple Alignment Mode with the default gap-penalty parameters. The multiple alignment was manually checked, and the absolutely conserved Cysteine residues in the alignment were defined as CSP motifs.

Pattern search, selection of *Anopheles* CSP candidates in the whole genome sequences of *A. gambiae*

The Fasta files of gDNA, cDNA and the according translated peptide sequences of *A. gambiae* were downloaded from the Ensembl Mosquito Genome Browser (ftp://ftp.ensembl.org/pub/current_mosquito/). A pattern-searching program was created through standard PERL programming and named as CSPMOT by the authors. This program is run in Windows Commander, and can scan a local database for a pattern defined by the users. In this case, the CSP motif was used as a pattern to match every sequence stored in a local database that contains the Fasta files downloaded from the public databases. When a sequence matches the pattern, the program will display the sequence name, the position of the first residue where the pattern matches, and the pattern match alignment.

BLAST

BLAST was performed at NCBI (http://ncbi.nlm.nih.gov/BLAST/; April 2003). The *Anopheles* CSP candidates identified by the program described above were used as queries to Blast GenBank. The E values smaller than 0.0001 were accepted. The E value is a parameter that describes the number of hits one can "expect" to see just by chance when searching a database of a particular size.

Annotations

GENSCAN (http://genes.mit.edu/GENSCAN.html; ref. *28*, *29*) and ORF Finder (http://www. ncbi.nlm.nih.gov/gorf/gorf.html) were used to predict the full-length genes, based on gDNA and cDNA sequences of *Anopheles* CSP candidates. The annotated CDSs were then used for calculation of hydrophobicity, pI and molecular weight of the CSP candidates, by using the comprehensive biosoftware Vector NTI (InforMax Inc., Bethesda, USA). The positions of signal peptides were determined online (http://www.cbs.dtu.dk/services/SignalP2.0/; ref. *30*, *31*), and finally the secondary structure, in particular the globularity, was predicted (http://maple.bioc.columbia.edu/predictprotein/; ref. *32–34*).

Homology modelling

Csp2 of *Mamestra brassicae* (PDB ID: 1K19; ref. *35*) is an elucidated insect chemosensory protein, which was used as a model to predict the 3-D structure of *Anopheles* CSP candidates in Swiss-Pdb Viewer (http://us.expasy.org/spdbv/; ref. *36*). Swiss-Pdb Viewer is a program that allows analysis of several proteins at the same time. It could be used to deduce 3-D structure of proteins and compare the active sites of different molecules.

Acknowledgements

The authors thank Rothamsted International for its support of Dr. Zhengxi Li's post-doctoral fellowship in Biological Chemistry Division, Rothamsted Research, UK. Special thanks should be given to Division of Science and Technology, China Agricultural University for its financial support to the completion of the manuscript.

References

1. Holt, R.A., *et al.* 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* 298: 129-149.
2. Singh, R.N. and Nayak, S.V. 1985. Fine structure and primary sensory projections of sensilla on the maxillary palp of *Drosophila melanogaster* Meigen (Diptera:

Drosophilidae). *Int. J. Insect Morphol. Embryol.* 14: 291-306.
3. Pelosi, P. and Maida, R. 1995. Odorant-binding proteins in insects. *Comp. Biochem. Physiol. B. Biochem. Mol. Biol.* 111: 503-514.
4. Vogt, R.G., *et al.* 1999. Odorant-binding proteins diversity and distribution among the insect orders, as indicated by LAP, an OBP-related protein of the true bug *Lygus lineolaris* (Hemiptera, Heteroptera). *Chem. Senses* 24: 481-495.
5. McKenna, M.P., *et al.* 1994. Putative *Drosophila* pheromone-binding proteins expressed in a subregion of the olfactory system. *J. Biol. Chem.* 269: 16340-16347.
6. Pikielny, C.W., *et al.* 1994. Members of a family of *Drosophila* putative odorant-binding proteins are expressed in different subsets of olfactory hairs. *Neuron* 12: 35-49.
7. Danty, E., *et al.* 1998. Separation, characterization and sexual heterogeneity of multiple putative odorant-binding proteins in the honeybee *Apis mellifera* L. (Hymenoptera: Apidea). *Chem. Senses* 23: 83-91.
8. Briand, L., *et al.* 2002. Characterization of a chemosensory protein (ASP3c) from honeybee (*Apis mellifera* L.) as a brood pheromone carrier. *Eur. J. Biochem.* 269: 4586-4596.
9. Maleszka, R. and Stange, G. 1997. Molecular cloning by a novel approach, of a cDNA encoding a putative olfactory protein in the labial palps of the moth *Cactoblastis cactorum*. *Gene* 202: 39-43.
10. Robertson, H.M., *et al.* 1999. Diversity of odorant-binding proteins revealed by an expressed sequence tag project on male *Manduca sexta* moth antennae. *Insect Mol. Biol.* 8: 501-518.
11. Picimbon, J.F., *et al.* 2000. Purification and molecular cloning of chemosensory proteins from *Bombyx mori*. *Arch. Insect Biochem. Physiol.* 44: 120-129.
12. Nagnan-Le Meillour, P., *et al.* 2000. Chemosensory proteins from the proboscis of *Mamestra brassicae*. *Chem. Senses* 25: 541-553.
13. Jacquin-Joly, E., *et al.* 2001. Functional and expression pattern analysis of chemosensory proteins expressed in antennae and pheromonal gland of *Mamestra brassicae*. *Chem. Senses* 26: 833-844.
14. Picimbon, J.F., *et al.* 2001. Identity and expression pattern of chemosensory proteins in *Heliothis virescens* (Lepidoptera, Noctuidae). *Insect Biochem. Mol. Biol.* 31: 1173-1181.
15. Pikielny, C.W., *et al.* 1994. Members of a family of *Drosophila* putative odorant-binding proteins are expressed in different subsets of olfactory hairs. *Neuron* 12: 35-49.
16. Dyanov, H.M. and Dzitoeva, S.G. 1995. Method for attachment of microscopic preparations on glass for in situ hybridization, PRINS and in situ PCR studies. *BioTechniques* 18: 822-824.
17. Biessmann, H., *et al.* 2002. Isolation of cDNA clones encoding putative odourant binding proteins from the antennae of the malaria-transmitting mosquito, *Anopheles gambiae*. *Insect Mol. Biol.* 11:123-132.
18. Angeli, S., *et al.* 1999. Purification, structural characterization, cloning and immunocytochemical localization of chemoreception proteins from *Schistocerca gregaria*. *Eur. J. Biochem.* 262: 745-754.
19. Picimbon, J.F., *et al.* 2000. Chemosensory proteins of *Locusta migratoria* (Orthoptera, Acrididae). *Insect Biochem. Mol. Biol.* 30: 233-241.
20. Marchese, S., *et al.* 2000. Soluble proteins from chemosensory organs of *Eurycantha calcarata* (Insecta, Phasmatodea). *Insect Biochem. Mol. Biol.* 30: 1091-1098.
21. Normura, A., *et al.* 1992. Purification and localization of p10, a novel protein that increases in nymphal regenerating legs of *Periplaneta americana* (American cockroach). *Int. J. Dev. Biol.* 36: 391-398.
22. Kitabayashi, A.N., *et al.* 1998. Molecular cloning of cDNA for p10, a novel protein that increases in nymphal regenerating legs of *Periplaneta americana* (American cockroach). *Insect Biochem. Mol. Biol.* 28: 785-790.
23. Campanacci, V., *et al.* 2001. Chemosensory protein from the moth *Mamestra brassicae*. Expression and secondary structure from 1H and 15N NMR. *Eur. J. Biochem.* 268: 4731-4739.
24. Vogt, R.G., *et al.* 2002. A comparative study of odorant-binding protein genes: differential expression of the PBP1-GOBP2 gene cluster in *Manduca sexta* (Lepidoptera) and the organization of OBP genes in *Drosophila melanogaster* (Diptera). *J. Exp. Biol.* 205: 719-744.
25. Galindo, K. and Smith, D.P. 2001. A large family of divergent *Drosophila* odorant-binding proteins expressed in gustatory and olfactory sensilla. *Genetics* 159: 1059-1072.
26. Graham, L.A. and Davies, P.L. 2002. The odorant-binding proteins of *Drosophila melanogaster*: annotation and characterization of a divergent gene family. *Gene* 292: 43-55.
27. Thompson, J.D., *et al.* 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25: 4876-4882.
28. Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268: 78-94.
29. Burge, C.B. and Karlin, S. 1998. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* 8: 346-354.
30. Nielsen, H., *et al.* 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering* 10: 1-6.

31. Nielsen, H. and Krogh, A. 1998. Prediction of signal peptides and signal anchors by a hidden Markov model. In *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB 6)*, pp.122-130. AAAI Press, California, USA.
32. Rost, B. 1996. Predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* 266: 525-539.
33. Rost, B. and Sander, C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232: 584-599.
34. Rost, B. and Sander, C. 1994. Conservation and prediction of solvent accessibility in protein families. *Proteins* 20: 216-226.
35. Campanacci, V., *et al.* 2001. Recombinant chemosensory protein (CSP2) from the moth *Mamestra brassicae*: crystallization and preliminary crystallographic study. *Acta Crystallogr. D. Biol. Crystallogr.* 57: 137-139.
36. Guex, N. and Peitsch, M.C. 1997. SWISS-MODEL and the Swiss-Pdb Viewer: an environment for comparative protein modelling. *Electrophoresis* 18: 2714-2723.