Journal of Advanced Bioinformatics Applications and Research
ISSN 0976-2604.Online ISSN 2278–6007
Vol 5, Issue3, 2014, pp163-171
http://www.bipublication.com

Genome Wide Analysis of Cecropin Multigene Family In *Anopheles gambiae* Pest Strain

Neelam Sehrawat[1] and Surendra Kumar Gakhar[2]
[1]Department of Genetics, M.D. University, Rohtak-124001 (Haryana)
[2]Faculty of Life Science, M.D. University, Rohtak-124001 (Haryana)
Corresponding author, Email-neelamsehrawat@gmail.com, Mobile- 09034600138

[Received-01/08/2014, Accepted-21/08/2014]

ABSTRACT
Cecropins are most potent immune families found in insect with diversified numbers and features. Cecropins constitute a large family of antibacterial, anti viral, antifungal, anti protozonal, anti malarial, anti cancerous and toxic peptides. Cecropins are isolated from a variety of Lepidoptera, diptera and coleopteran species. In view of the large number of Cecropin genes exists with high sequence variation among them, an attempt has been done in the present study to overview genome wide analysis of Cecropin multigene family in *Anopheles gambiae* PEST Strain. Cecropin genes encodes an inducible 58-67 amino acid peptide and was clusterd into two groups; first viz. B, A and C and second group including 4. The first group cecropin B, A and C was located on chromosomes number X and second group's Cecropin 4 was on chromosome number 2L. The first group Cecropins were organized in positive as well as opposite orientations. Cecropin A and B were in reverse orientation where as Cecropin C and 4 were in forward orientations. All the four Cecropins have two exons and one intron. The genomic map showed that Cecropin A is in between B and C, Cecropin 4 is lying far from first group. The GC content is maximum i.e. 55.04 in Cecropin B. The intergenic region of 2456 bp was observed between Cecropin B and Cecropin A, 1854 bp in Cecropin A and Cecropin C. The intergenic region observed between Cecropin B, A and Cecropin C indicating that these genes were distantly evolved. The phylogenetic tree showed a positive correlation between members of group and physical location on the chromosome, as the length of the intergenic region plays a major role to create newer Cecropin gene families.

Key words: *Anopheles gambiae*, Cecropin, Paralogous gene, Genome, multigene, evolution.

INTRODUCTION
Innate immunity is the first barrier of host defense in animals to kill invading microbes such as bacteria, fungi and viruses [25]. Insect have no specific immunity but only an innate response including cellular and humoral immune responses [8]. The humoral immune response includes rapid synthesis of a number of antimicrobial peptides (AMP), such as Cecropin, attacins, defensin, and diptericins in response to bacterial invasion [1] that were important effectors of the innate immune response in insects. The cellular response consists mainly of phagocytosis and encapsulation. In

addition to a large number of different AMPs, reactive oxygen species (ROS) and nitric oxide (NO) were known to have a role in humoral defense. Most AMPs were synthesized mainly in the fat body, the major immune-responsive tissues, and were secreted into the haemolymph [24]. Among the AMPs, Cecropins were well understood and have been investigated in vertebrates as well as insects. Since the discovery of Cecropin in *Hyalophora cecropia* [18], many Cecropin family genes have been found and isolated in various lepidopteran and dipteran insect [6, 10 & 22]. They are immunogenic peptides exhibiting host defense functions and often act in synergy, providing rapid non-specific defense against invading micro-organisms [20]. Most of these natural broad-spectrum peptide antibiotics have limited cytotoxicity to human cells and are being developed as therapeutics against pathogens resistant to classical antibiotics. The broad antibacterial activity of cecropins is due to the formation of large pores in bacterial cell membranes [3]. The diverse activity spectrum of these peptides may indicate different modes of action [2].

Genes that have originated by gene duplication and retained a certain degree of similarity form multigene families. The different members of a multigene family were often arranged in a compact cluster although due to chromosomal rearrangements subsequent to the gene duplications they might be more or less dispersed in the genome. The members of a multigene family can be functional or nonfunctional which were known as pseudogenes. The functional genes can be very similar as the copies might have retained the same function and be redundant.

In *Drosophila melanogaster*, the Cecropin multigene family consist of both functional and non functional (pseudogenes) genes and functional genes code for cecropins [10 & 23]. The Cecropin genes occur in a single cluster, measuring about 7 kb in *D. melanogaster*, to 20 kb in *H. cecropia* [1]. In *D. melanogaster*, the 7 kb cluster maps to chromosomal location 99E and contains four expressed genes (CecA1, A2, B and C), two pseudogenes, and the gene encoding the male-specific antibacterial protein, andropin [23]. *D. virilis* lacks the andropin gene, but contains four Cecropin genes and one pseudogene within a 9 kb region [28]. On bacterial infection all functional genes were expressed, mainly in the fat body, although at different times during life cycle. The functional genes Cecropin A1 and Cecropin A2 were essentially expressed in larvae and adults, while, Cecropin B and Cecropin C were mainly expressed during the pupal stage [8]. Four Cecropin genes also appear to be present in the flesh fly, *Sarcophaga peregrina* [9]. The large size of the cluster is in part due to the long distances between the genes and to insertion elements in the introns of CecA and CecD. In *Bombyx mori* cecropin gene family occurred in two groups, first group viz. A and second group including B, D, E and Enbocin [15]. Cecropin A consisted of two sub-groups located on chromosome number 6 of *B. mori* genome. Cecropin B consisted of six sub-groups, cecropin D and E of one each and Enbocin of two. The second sub-group formed in tandem array of multigene family locus over a length of 78.62 kb on chromosome number 26 in *B. mori* genome and was organized in positive as well as opposite orientation. In this connection, an attempt has been made to analyze the organization of cecropin multigene cluster in *Anopheles gambiae* PEST strain.

In *An. gambiae*, Cecropin A has been cloned and characterized [26], the genomic organization and regulation of three cecropin genes, Cec A, B and C has been done [27] but there is no study on molecular evolution of cecropin multigene family in *An. gambiae*. Now whole genomic database of *An. gambiae* is available, utilizing this information, a genome wide screening of cecropin genes to analyze the organized structure of cecropin multigene clusters in the *An. gambiae* PEST strain was attempted.

METHODOLOGY

Retrieval of cecropin sequence

The genomic DNA, cDNA sequences of cecropin gene were retrieved from the NCBI database and those that expressed exclusively in *An. gambiae* were selected. These sequences pertaining to various target tissues were independently deposited by different researchers. The deduced amino acid sequences were obtained through translation of the selected cDNA sequences and were converted in to FASTA format for Clustal X analysis. Then the cDNA sequences were converted into deduced amino acid sequences and the ORF of individual gene was determined using ORF finder (*www.ncbi.nlm.nih.gov/projects/gorf/*). The conserved domain for Cecropin protein was also identified in all translated sequences with the help of SMART 7 software.

Identification of paralogous gene sequences in *An. gambiae*

The *An. gambiae* cDNA sequences were BLAST searched with *An. gambiae* Genome Database (https://www.vectorbase.org) to identify paralogous multigene family. Using the *An. gambiae* database, the functional annotation of genes, paralogous gene sequences, gene products, number of nucleotides and amino acids, GC contents, gene orientation and chromosome mapping were determined. Further, the tools provided in the database were utilized to perform specific genomic BLAST search as well as Map view (a visualization tool that provides a graphical view of selected genes). The organization of paralogous cecropin multigene family on individual scaffold was also analyzed using BLAST search with Gene ID.

Clustal W analysis

Phylogenetic analyses were performed with the multiple sequence alignment using ClustalW through MEGA 6.06 [21]. The Bootstrap consensus NJ tree for cecropin paralogous gene sequences was constructed with the Bootstrap value 1000. A separate Clustal W alignment was performed with 5' upstream regions of all paralogous gene sequences. The sequences were manually edited using BIOEDIT programme and the promoter elements were identified using GENERUNNER programme.

Identification of Signal Peptide

Signal peptide cleavage sites were predicted using Signal IP algorithm (www.cbs.dtu.dk/services/SignalIP), based on the Neural Network and Hidden Markov Model.

ESTs of the *An. gambiae* Cecropin

To determine the expression patterns of the individual paralogous genes, a local BLASTN search was performed against the *An. gambiae* EST database (https://www.vectorbase.org) and Unigene database at NCBI. The majority of EST database sequences originated mainly from the microbial infected fat body, salivary gland, ovary, midgut and hemolymph etc. The specific expressions of the individual paralogous genes were analyzed in EST libraries constructed from different tissues of *An. gambiae*.

RESULTS AND DISCUSSION

Multigene families often evolve in many ways that violate assumptions necessary for simple and objective gene phylogeny estimation. Some members of the family may evolve at much faster rate and as such are dubbed fast evolving genes. This occurs when one member gene takes on a significantly novel function and thus encounters significantly different selective pressure from other multigene family members. Another usual assumption of molecular tree construction is that each branch of the tree evolves independently from other branches. These families often show coincidental evolution, either indirectly through biased mutational and selective forces or directly by mechanisms such as gene conversions [17].

A total of thirty three *An. gambiae* Cecropin gene sequences were retrieved from the NCBI database (Table 1). Most of the entries submitted for Cec 2 (Cecropin B) gene by [13 & 4]. There were single sequence available for Cecropin A and 4 [14].

[27] has submitted the full genomic sequence of first group of cecropin gene family i.e. Cec B, A and C. Further, the cecropin cDNA as well as protein sequences were BLAST searched with *Anopheles gambiae* PEST genome database (https://www.vectorbase.org). A total of 4 paralogous gene sequences were retrieved from the genomic database of which one sequence belonged to cecropin A, B and C respectively, one to the cecropin 4. The three cecropin A, B and C sequences were located on chromosome number X and Cecropin 4 was on chromosome number 2L. Cecropin A, B and C genes were organized on a single locus while Cecropin 4 on different locus and among them two were organized in reverse orientationi.e. Cecropin B And A, while remaining two in the forward orientation (Cecropin C and 4). Interestingly, the cecropin A was located in between cecropin B and C shown in table-2.

The genomic analysis of *Drosophila* revealed that, the multigene family originated via retro position and DNA based duplication. The presence of repetitive sequences and TEs in the 5' and 3' flanking regions of the multigenes suggested that, gene duplication occurred during the formation of drosomycin multigene family [5]. Earlier reports revealed, only three cecropin families in *Drosophila* viz. A, B and C. Cecropin A had two sub groups viz. A1 and A2, while, B and C had no sub-groups indicating that, evolution of paralogous gene sequences was based on the individual requirement and selection pressure in the individual population.

Bombyx mori genome showed presence of 12 paralogous genes. Presence of the Transposable Elements (TE) in the flanking regions of each paralogous genes confirmed gene duplications giving rise to the cecropin multigene family**.** The presence of six cecropin B sub-groups was a unique feature in the silkworm *B. mori.* Since the cecropin genes B, D, E and Enbocin were located in the same gene locus, it can be inferred that all the four sequences have closely evolved from a single gene. *An. gambiae* genome wide analysis showed that Cecropin A, B and C have same evolution pattern as *B. mori*.

Comparison of the molecular distance of paralogous gene pairs between different and same species indicates the amount of within-species coincidental evolution. In *An. gambiae* three of the four paralogous cecropin genes present formed a tight cluster over about 1.32 kb length of DNA suggesting that, these genes probably originated from a common ancestral by gene duplication and later individuals with multiple genes were selected during evolution. If each protein has a slightly different antibacterial spectrum, the presence of multiple proteins should be

advantageous for the survival of the insects in various pathogenic environments. Genome analysis of model organisms have shown that over one-third of all protein-coding genes belong to multigene family originating from the gene duplications [12].

The multigene cluster of cecropin A, B and C was spread over a region of 1.32 kb in length in *An. gambiae* genomic DNA. Long intergenic regions of 2456 bp was observed between Cecropin B and cecropin A and 1854 bp in Cecropin A and Cecropin C. CecA and CecB are located in opposite orientations separated by a divergent promoter of 1186 bp. CecC is located downstream of CecA. Each of the cecropin genes invariably contained two exons and one intron. However the lengths of the intron regions varied among the cecropin sub-groups. A short intron in the coding region is present in all three Cecropin genes with lengths of 89, 95, 87 and 134 bp for CecA, CecB, CecC and Cec 4 respectively. The complete open reading frames encode a pre protein of 58, 60, 59 and 67 amino acid residues for CecA, CecB, CecC and Cec 4 respectively. The *An. gambiae* cDNA encodes a protein of fifty-eight amino acids showing a high degree of similarity to insect Cecropins, particularly to the recently identified Cecropins of the mosquitoes *Ae. albopictus* [19] and *Ae. aegypti* [17].

The phylogenetic analysis of different paralogous cecropin gene sequences revealed that, all cecropin members clustered together shown in figure-1. Among the four cecropin genes, cecropin B and 4 formed separate single clusters, while, cecropin A separate cluster. The two cecropin B and 4 genes formed a single cluster revealing their similar evolution compared to cecropin A genes. Further, Cecropin C was the ancestor among all the Cecropins. Cecropin A, B and C were located on same chmosome X and Cecropin 4 was on chromosome 2 L.

Analysis of different cecropin sequences with respect to the different target tissues based on the EST library information available at http://www.tissue-atlas.org/revealed that, Cecropin A, B and C genes were expressed in different tissues. Maximum expression had been observed in the midgut of bacteria challenged *An. gambiae* (Table- 3). However, the cecropin A gene was expressed in different target tissues like head (5%), salivary gland (4%), malphigian (1%), ovary and testis (1%). Cecropin B gene was expressed in salivary gland (20%), malphigian (15%), testis (2%) and Cecropin C was head (0.2%), salivary gland (5.3%), malphigian (3%) and testis (0.7%). It was interesting to note there Cecropin B and C were not expressed in ovarian tissue There was no expression data available for Cecropin 4.

The promoter sequences of the paralogous genes were compared with the help of Clustal W program. The unique promoter elements of cecropin genes were prominently located in the 5' flanking upstream region such as κβ, GATA, TATAA and CAP site. The promoter elements were also conserved in Cecropin A, B and C. The TATAA boxes of all cecropin A, B and C paralogous gene sequences were located at different position. The location and sequence variations as well as promoter regions were prominent factors in the tissue specificity.

The literature studies showed that mature cecropin protein contained sixty four amino acids of which twenty two located in the N- terminal region functioned as signal peptide [7]. In the present study also, signal peptides were also conserved in the cecropin A and approximately 30% sequence variation was observed in different cecropin genes in *An. gambiae.* However the length of each signal peptide did not vary between the cecropin protein sequences as shown in figure 2.

Analysis of *An. gambiae* genome lead to prediction of 247 putative genes encoding cuticular proteins. A large number of genes were present in gene clusters. Gene clusters of cuticular proteins have also been observed in *Drosophila*, *Tribolium*, *Mosquitoes* and *Silkworm* [16]. Phylogenetic analysis of *An. gambiae* genome showed distinct organization of Cecropin gene family. Cecropin B and Cecropin 4 formed a clade indicating that, these two genes were closely evolved. Cecropin A was in out of the clade (B and 4) showed that, it evolved independently from a common ancestor i.e. Cecropin C. There was no gene duplication observed in *An. gambiae* similar to Drosophila, Tribolium while Cecropin B is exclusively duplicated in Silkworm. In the humoral antimicrobial defence, Cecropin show evidence of existence of multigene family.

CONCLUSION

In the present study the possibility of gene duplication in the Cecropin gene family was predicted through analysis of Cecropin gene sequence in *An. gambiae* genome which revealed presence of four paralogous genes. The Cecropin B, A and C were located on the same gene locus, it can be inferred that all three sequences have closely evolved from a single gene. Cecropin 4 was located on the different genomic location indicated it distinct origin.

In *An. gambiae* three genes out of four paralogus genes present formed a tight cluster of about 1.32 kb length of DNA suggesting that, these original genes probably originated from a common ancestral by gene duplication and later individuals with multiple genes were selected during evolution [15].

Analysis of different cecropin sequences with respect to the different target tissues based on the EST library was carried out indicating that, the cecropin genes of each sub group were selectively expressed in specific target tissue. The location and sequence variations as well as promoter regions were prominent factors in the tissue specificity. The unique promoter elements of cecropin genes were prominently located in the 5' flanking upstream region such as κβ, GATA, TATAA and CAP site. The promoter elements were also conserved depending upon each sub-family. The TATAA boxes of all cecropin A, B and C paralogous gene sequences were located at different position. The location and sequence variations as well as promoter regions were prominent factors in the tissue specificity.

Phylogenetic analysis of *An. gambiae* genome showed distinct organization of Cecropin gene family. Cecropin B and Cecropin 4 formed a clade indicating that, these two genes were closely evolved. Cecropin A was in out of the clade (B and 4) showed that, it evolved independently from a common ancestor i.e. Cecropin C. There was no gene duplication observed in *An. gambiae* similar to Drosophila, Tribolium while Cecropin B is exclusively duplicated in Silkworm. In the humoral antimicrobial defence, Cecropin show evidence of existence of multigene family.

The information retrieved from the Vectorbase, *Anopheles gambiae* PEST genome database allowed us to draw comprehensive conclusions regarding adaptive evolution as well as functional significance of cecropin multigene family in *An. gambiae.* This forms a very vital basis to understand evolution of the immune system genes of *An. gambiae* with respect to interaction with the natural diversified pathogens in the ever changing environment.

ACKNOWLEDGEMENT

We are thankful to Department of Biotechnology (DBT), New Delhi for providing IPLS grant.

REFERENCES

1. Boman, H.G, (1994), Cecropins: antibacterial peptides from insects and pigs, Phylogenetic Perspectives in Immunity. pp 3-17.
2. Bulet, P, et al., (2005), Insect antimicrobial peptides: structures, properties and gene regulation, Protein and Peptide Letters. Vol-12 issue 1, pg 3-11.
3. Christensen, B, et al., (1988), Channel-forming properties of cecropins and related model compounds incorporated into planar lipid membranes, Proceedings of the National Academy of Sciences (USA). Vol- 85, issue 14, pg 5072-5076.
4. Cohuet, A, et al., (2008), SNP discovery and molecular evolution in *Anopheles gambiae*, with special emphasis on innate immune system, BMC Genomics. Vol- 9, pg 227.
5. Deng, X. J, et al., (2009), Gene expression divergence and evolutionary analysis of the drosomycin gene family in Drosophila melanogaster, Journal of Biomedicine and Biotechnology. doi: 10.1155/2009/315423.[PMID: 19888430].
6. Dickinson, L, et al., (1988), A family of bacteria-regulated, cecropin D-like peptides from Manduca sexta, Journal of Biological Chemistry. Vol - 263 issue 36, pg 19424-9. [PMID: 3143727]
7. Futahashi, R, et al., (2008), Genome-wide identification of cuticular protein genes in the silkworm, Bombyx mori, Insect Biochemistry and Molecular Biology. Vol-38 issue 12, pg 1138-46. [PMID: 19280704].
8. Hultmark, D. (1993), Immune reactions in Drosophila and other insects: a model for innate immunity, Trends in Genetics.Vol- 9 issue 5, pg 178-83. [PMID: 8337755].
9. Kanai, A. and Natori, S, (1989), Cloning of gene cluster for sarcotoxin 1, antibacterial proteins of Sarcophaga peregrina, FEBS Lett. Vol- 258, pg 199-202.

10. Kylsten, P. et al., (1990), The cecropin locus in Drosophila; a compact gene cluster involved in the response to infection, EMBO Journal. Vol- 9, pg 217. [PMID: 2104802]
11. Lowenberge, C. et al., (1999), Antimicrobial activity spectrum, cDNA cloning, and mRNA expression of a newly isolated member of the Cecropin family from the mosquito vector *Aedes aegypt*,. Journal of Biological Chemistry. Vol- 274 issue 29, pg 20092-20097
12. Meisel, P. R, (2009), Repeat mediated gene duplication in the Drosophila pseudoobscura genome, Gene. Vol 438 issue 1, pg 1-7. [PMID: 19272434].
13. Mendes, A.M, et al., (2008), Conserved mosquito/parasite interactions affect development of *Plasmodium falciparum* in Africa, PLoS Pathogens. Vol - 4: e1000069
14. Mongin, E, et al., (2004), The Anopheles gambiae genome: an update, Trends Parasitol. Vol- 20 issue 2, pg 49-52.
15. Ponnuvel, P. M. et al., (2010), Molecular evolution of the cecropin multigene family in silkworm Bombyx mori, Bioinformation. Vol-5 issue 3, pg 97-103.
16. Xia, Q, *et al.* (2008) The genome of a lepidopteran model insect, the silkworm Bombyx mori, Insect Biochemistry and Molecular Biology. Vol-38 issue 12, pg 1036-45. [PMID: 19121390].
17. Roach, J.C, et al., (2005), The evolution of vertebrate Toll-like receptors, Proceedings of National Academy of Sciences. Vol- 102 issue 27, pg 9577-82. [PMID: 15976025].
18. Steiner, H.D et al., (1981), Sequence and specificity of two antibacterial proteins involved in insect immunity. Nature. Vol- 292 issue 5820, pg 246-248.
19. Sun, D, et al., (1998), Peptide sequence of an antibiotic cecropin from the vector mosquito, *Aedes albopictus*, Biochemical Biophysical Research Communications. Vol-249 issue 2, pg 410-415.
20. Tamang, D.G, et al., (2006), The cecropin superfamily of toxic peptides, Journal of Molecular Microbiology and Biotechnology. Vol-11 issue 1-2, pg 94-103.
21. Tamura, K et al., (2013), MEGA6: Molecular Evolutionary Genetics Analysis version 6.0, Molecular Biology and Evolution. Vol- 30, pg 2725-2729.
22. Taniai, K. et al., (1992), Isolation and nucleotide sequence of cecropin B cDNA clones from the silkworm, Bombyx mori, Biochemistry Biophysics Acta. Vol- 1132 issue 2, pg 203-6. [PMID: 1390892]
23. Tryselius, Y, et al., (1992), *CecC*, a cecropin gene expressed during metamorphosis in *Drosophila* pupae, Eur. J. Biochem. Vol- 204, pg 395-399.
24. Uvell, H. and Engstrom, Y, (2007), A multilayered defense against infection: combinatorial control of insect immune genes, Trends in Genetics. Vol- 23 issue 7, pg 342-9. [PMID: 17532525].
25. Vilmos, P. and Kurucz, E, (1998), Insect immunity: evolutionary roots of the mammalian innate immune system, Immunology Letters. Vol- 62 issue 2, pg 59-66. [PMID: 9698099].
26. Vizioli, J, et al., (2000), Cloning and analysis of a cecropin gene from the malaria vector mosquito, *Anopheles gambiae*, Insect Mol. Biol. Vol- 9, pg 75-84.
27. Zheng, X. L. and Zheng, A.L, (2002) Genomic organization and regulation of three cecropin genes in *Anopheles gambiae*, Insect Molecular Biology. Vol- 11 issue 6, pg 517-525.
28. Zhou, X, et al., (1997), Identification and characterization of the Cecropin antibacterial protein gene locus in Drosophila virilis, J Mol Evol. Vol-44, 272-281.

Table 1: Details of cecropin gene sequences of *An. gambiae* retrieved from the NCBI database

Sr. No.	Cecropin sub family	Accession No.	Gene/Protein	Reference	Length
1	Cecropin B_ANOGA (1)	XM311224	Cec B	Mongin *et al.*,2004	718 bp
2	Cecropin A (1)	AY353563	Cec A	Mongin *et al.*,2004	511 bp
3	CecA,cecB,cecC	AF525673	cecA,cecB,cecC	Zheng *et al.*, 2002	13,259bp
4	Cecropin B (29)	AM900862	cec2	Mendes et al.,2008	280 bp
5		AM900858	cec2	Mendes et al.,2008	280 bp
6		AM900856	cec2	Mendes et al.,2008	280 bp
7		AM900854	cec2	Mendes et al.,2008	280 bp
8		AM900852	cec2	Mendes et al.,2008	280 bp
9		AM900850	cec2	Mendes et al.,2008	280 bp
10		AM900861	cec2	Mendes et al.,2008	280 bp
11		AM900859	cec2	Mendes et al.,2008	280 bp
12		AM900857	cec2	Mendes et al.,2008	280 bp
13		AM900855	cec2	Mendes et al.,2008	280 bp
14		AM900853	cec2	Mendes et al.,2008	280 bp
15		AM900851	cec2	Mendes et al.,2008	280 bp
16		AM900849	cec2	Mendes et al.,2008	280 bp
17		AM774784	cec2	Cohuet et al.,2008	280 bp
18		AM774782	cec2	Cohuet et al.,2008	280 bp
19		AM774780	cec2	Cohuet et al.,2008	280 bp
20		AM774778	cec2	Cohuet et al.,2008	280 bp
21		AM774776	cec2	Cohuet et al.,2008	280 bp
22		AM774774	cec2	Cohuet et al.,2008	280 bp
23		AM774772	cec2	Cohuet et al.,2008	280 bp
24		AM774770	cec2	Cohuet et al.,2008	280 bp
25		AM774785	cec2	Cohuet et al.,2008	280 bp
26		AM774783	cec2	Cohuet et al.,2008	280 bp
27		AM774781	cec2	Cohuet et al.,2008	280 bp
28		AM774779	cec2	Cohuet et al.,2008	280 bp
29		AM774777	cec2	Cohuet et al.,2008	280 bp
30		AM774775	cec2	Cohuet et al.,2008	280 bp
31		AM774773	cec2	Cohuet et al.,2008	280 bp
32		AM774771	cec2	Cohuet et al.,2008	280 bp
33	Cecropin 4(1)	XM565481	Cec 4	Mongin et al., 2004	204bp

Table 2: Genomic organization, orientation and chromosomal location, GC content number of nucleotide and amino acids and exon - intron of paralogous cecropin gene sequences of *An. gambiae.*

Paralogous gene sequence ID	Name of the protein	GC content	Number of nucleotides	Number of amino acids	Position	Orientation	No. of exons & introns		Chromosomal location
							Exon	Intron	
AGAP000693	Cecropin A	48.83	511	58	12,440,533-12,441,132	← (reverse)	2	1	Chromosome X
AGAP000692	Cecropin B	55.04	518	59	12,437,473-12,438,077	← (reverse)	2	1	Chromosome X

AGAP000694	Cecropin C	47.11	718	60	12,442,174-12,442,986	→ (Forward)	2	1	Chromosome X
AGAP006722	Cecropin 4	49.26	204	67	37,233,354-37,233,690	→ (Forward)	2	1	Chromosome 2L

Table 3- Expression pattern of *An. gambiae* cecropin in different tissues based on data obtained from http://www.tissue-atlas.org/.

Expressed in tissue/organ	Cecropin A	Cecropin B	Cecropin C
Carcase	17%	31%	14%
Head	5%	-	0.2%
Salivary gland	4%	20%	5.3%
Midgut	71%	32%	77%
Malphigian	1%	15%	3%
Ovary	1%	-	-
Testis	1%	2%	0.7%
unknown	-	-	0.9%

Figure 1- Neighbor-joining tree of 4 paralogous Cecropin gene percentage of bootstrap values (based on the 1000 replication) for the main branching nodes shown on the tree. The paralogous gene sequences of *An. gambiae* genome database are indicated by vector database gene locations and chromosomal numbers

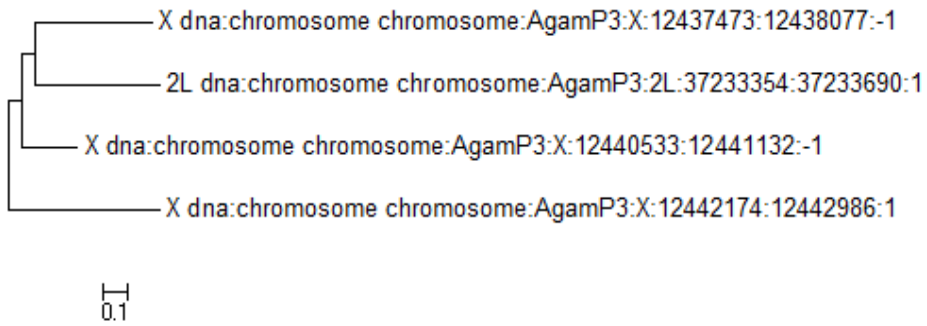

Figure -2: Alignment of the amino acid sequences from *An. gambiae* genomic database. The N-terminal signal peptides of are indicated by underlining.

```
>AGAP000693 MNFSKIFIFVVLAVLLLCS-QTEAGRLK-KLGKKIEGAGKRVFKAAEKALPVVAGVKAL--------G  58
>AGAP000692 MNF-KLIFLVALVLMAAFLGQTEGRRFK-KFLKKVEGAGRRVANAAQKGLPLAAGVKGLV-------G  59
>AGAP000694 MNFTKLFILVAIAVLVVVGVQPVDGAPRwKFGKRLEKLGRNVFRAAKKALPVIAGYKAL--------G  60
>AGAP006722 MNVSKLFVIVLLATLLLFGGQAEAGHLK-KFGKKLEKVGKNVFHAVEKVVPVLQGIQDLRdkkngqrG  67
```