A COMPARATIVE GENOMICS APPROACH TOWARDS CLASSIFYING IMMUNITY-RELATED PROTEINS IN THE TSETSE FLY

by

Feziwe Mpondo

A thesis submitted in fulfillment of the requirements for the degree of *Magister Scientiae* in Bioinformatics at the South African National Bioinformatics Institute, University of the Western Cape

Supervisor: Professor Winston Hide
Co-supervisor: Professor Alan Christoffels

September 2009

UNIVERSITY of the
WESTERN CAPE

KEY WORDS

Glossina morsitans

Anopheles gambiae

Drosophila melanogaster

Aedes aegypti

Sleeping sickness

Vector control

Insect immunity

Comparative genomics

Thioester-containing proteins

Phylogenetic analysis

ABSTRACT

Tsetse flies (*Glossina* spp) are vectors of African trypanosome (*Trypanosoma* spp) parasites, causative agents of Human African trypanosomiasis (sleeping sickness) and Nagana in livestock. Research suggests that tsetse fly immunity factors are key determinants in the success and failure of infection and the maturation process of parasites. An analysis of tsetse fly immunity factors is limited by the paucity of genomic data for *Glossina* spp. Nevertheless, completely sequenced and assembled genomes of *Drosophila melanogaster*, *Anopheles gambiae* and *Aedes aegypti* provide an opportunity to characterize protein families in species such as *Glossina* by using a comparative genomics approach. In this study we characterize thioester-containing proteins (TEPs), a sub-family of immunity-related proteins, in *Glossina* by leveraging the EST data for *G. morsitans* and the genomic resources of *D. melanogaster*, *A. gambiae* as well as *A. aegypti*.

A total of 17 TEPs corresponding to *Drosophila* (four TEPs), *Anopheles* (eleven TEPs) and *Aedes aegypti* (two TEPs) were collected from published data supplemented with Genbank searches. In the absence of genome data for *G. morsitans*, 124 000 *G. morsitans* ESTs were clustered and assembled into 18 413 transcripts (contigs and singletons). Five *Glossina* contigs (Gmcn1115, Gmcn1116, Gmcn2398, Gmcn2281 and Gmcn4297) were identified as putative TEPs by BLAST searches. Phylogenetic analyses were conducted to determine the relationship of collected TEP proteins.

Gmcn1115 clustered with DmtepI and DmtepII while Gmcn2398 is placed in a separate branch, suggesting that it is specific to *G. morsitans*.

The TEPs are highly conserved within *D. melanogaster* as reflected in the conservation of the thioester domain, while only two and one TEPs in *A. gambiae* and *A. aegypti* thioester domain show conservation of the thioester domain suggesting that these proteins are subjected to high levels of selection. Despite the absence of a sequenced genome for *G. morsitans*, at least two putative TEPs where identified from EST data.

DECLARATION

I declare "A comparative genomics approach towards classifying immune-related proteins in the tsetse fly" is my own work, that it has not been submitted for any degree in any other university, and that all the sources used or quoted have been indicated and acknowledged by complete references.

Signed: Date: **September 2009**

ACKNOWLEDGEMENTS

My gratitude and thanks to:

My supervisor Winston Hide, for imparting his learned insight and knowledge throughout the programme.

My Co-Supervisor Alan Christoffels, for dedicating his time to the project, guiding and driving it, for also taking time to teach and for being infinitely patient with me.

The Medical Research Council (MRC) for granting me an opportunity to do research at SANBI, for funding and giving me support throughout the MSc programme. The knowledge I gained during my training I hope one day will contribute positively to the development of research in South Africa.

Sarah Mwangi, Kavisha Ramdayal, Magbubah Essack, Mario Jonas, Adele Kruger, Adam Dawe and Sumir Panji, for their constant support and always lending a hand whenever needed.

To the SANBI administration and IT staff Ferial Mullins, Maryam Salie, Dale Gibbs as well as Peter van Huesden, thank you for your support.

To my family for your love and support throughout this journey; especially my mother you have been my rock and you have pushed me to live up to my full potential.

Last but not least, to God (*YHWH*), without whom I would have never made a single day, all honor and glory.

TABLE OF CONTENTS

CONTENTS	PAGE
Title page.....	i
Keywords.....	ii
Abstract.....	iii
Declaration.....	v
Acknowledgements.....	vi
Table of Contents.....	vii
List of Figures.....	viii
List of Tables.....	ix
Abbreviations.....	x
Chapter 1 Introduction and literature review.....	2
Chapter 2 Data and methods.....	32
Chapter 3 Results.....	41
Chapter 4 Discussion.....	59
Chapter 5 Conclusions.....	68
Appendices.....	

LIST OF FIGURES

Chapter 1

Figure 1.1 The anatomical structure of a female tsetse fly..... 4

Figure 1.2 *Trypanosoma* development cycle in the insect and vector and human host..... 6

Figure 1.3 Toll and IMD pathways in invertebrates..... 16

Figure 1.4 A phylogenetic tree depicting thioester-containing proteins..... 19

Figure 1.5 Diagrammatic representation of sequence signatures of insect α-macroglobulins 23

Figure 1.6 An alignment of *D. melanogaster* TEP proteins highlighting key sequence features of..... 24

Chapter 2

Figure 2.1 A schematic diagram of steps followed in the characterization of TEP protein family in *G. morsitans*..... 35

Chapter 3

Figure 3.1 A CLUSTALW alignment of Gmcn1115 and Gmcn1116..... 44

Figure 3.2 A Dot plot matrix of Gmcn1115 and Gmcn1116..... 44

Figure 3.3 Sequence signatures (domains) expressed in TEP proteins of insect species..... 47

Figure 3.4 A phylogenetic tree of TEP proteins using PHYLIP (NJ approach).. 49

Figure 3.5 Gmcn1115 contig mapped to the exon/intron region of DmtepI..... 51

Figure 3.6 The genomic organization of TEP proteins in *Anopheles* chromosome 3 and *Drosophila* chromosome 2..... 53

LIST OF TABLES

Chapter 1

Table 1.1	A summary of invertebrate pattern recognition receptor proteins.....	14

Chapter 3

Table 3.1	TEP homologs identified in literature and protein database searches.....	41
Table 3.2	Putative TEP homologs identified in *G. morsitans* based on sequence similarity (BLASTP) searches	43
Table 3.3	TEP protein families obtained from the Compara_db.....	55
Table 3.4	Compara protein pairs showing significant sequence similarity..	57

ABBREVIATIONS

AMP	Antimicrobial peptide
BAC	Bacterial artificial chromosome
BBB	Blood-Brain-Barrier
Bf	B-factor
BLAST	Basic local alignment search tool
CDD	Conserved Domains Database
CDS	Coding sequences
CLIP	Clip-domain protein
CNS	Central nervous system
DDT	Dichloro-diphenyl-trichloroethane
DNA	Dioxyribosenucleic acid
DRC	The Democratic Republic of Congo
EGF	Epidermal growth factor
EST	Expressed sequenced tags
GNBP	Gram negative bacteria binding protein
GPI	Glycosylphosphatidylinositol
HAT	Human African Trypanosomiasis
HSP	Highest scoring pair
IGGI	International Glossina Genome Initiative
LPS	Lipopolysaccharides
LS	Least squares
MASP	MBL-associated serine proteases
MBL	Mannose-binding lectin
MCL	Markov clustering
MCMC	Markov Chain Monte Carlo
MHC	Major histocompatability complex
ML	Maximum likelihood
NJ	Neighbor joining
ORF	Open reading frame
PAMP	Pathogen-associated molecular pattern
PGN	Peptidoglycan
PGRP	Pattern-recognition receptor protein
pPO	proPhenol oxidases
PRR	Pattern recognition receptor
SCR	Short consensus repeats
SIT	Sterile Insect technique
sp	Species
TEPs	Thioester-containing proteins
VSG	Variant surface glycoprotein
WHO	World Health Organization

Chapter 1

LITERATURE REVIEW AND INTRODUCTION

1.1 Human African Trypanosomiasis in sub-Saharan Africa..... 2

1.2 Insect vector, *Glossina morsitans*..... 3

1.2.1 Vector development cycle and reproduction..... 4

1.3 *Trypanosoma* species and their development cycle..... 5

1.4 Human African Trypanosomiasis, clinical symptoms, drug therapeutics and vector control strategies..... 8

1.4.1 Clinical symptoms of sleeping sickness..... 8

1.4.2 Drug treatment for trypanosomiasis..... 9

1.4.3 Vector control strategies..... 10

1.5 Invertebrate host defense response..... 12

1.5.1 Pathogen recognition..... 13

1.5.2 Signalling..... 15

1.5.3 Pathogen elimination..... 18

1.6 Thioester-containing protein family..... 18

1.7 Comparative genomics and the characterization of immune-related protein families..... 26

1.7.1 The *Glossina morsitans* genome project..... 27

1.8 Study aims and objectives..... 29

CHAPTER 1

LITERATURE REVIEW AND INTRODUCTION

1.1 Human African Trypanosomiasis in sub-Saharan Africa

Human African trypanosomiasis (HAT) or sleeping sickness is a disease caused by *Trypanosoma* parasites and transmitted by tsetse fly vectors (*Glossinidae* spp). HAT has been a major problem for Africa since the beginning of the 20th century, particularly the sub-Saharan region. In 36 sub-Saharan countries, this disease causes approximately 500 000 – 700 000 human infections and approximately 100 000 result in death each year (Mathews, 2005 and WHO, 2006). HAT also causes a wasting trypanosomiasis disease in cattle and game animals, known as nagana. Sleeping sickness has plagued the sub-Saharan region with several sweeps of epidemics, with each episode lasting for several decades (Smith, 1998). As a result, agricultural development and cattle grazing have been constrained in countries such as Uganda and Angola, causing an economic instability directly and indirectly. The Democratic Republic of Congo (DRC), Angola and Southern Sudan are hardest hit by trypanosomiasis causing extensive public health problems, as these countries are impoverished, lacking infrastructure, war-torn and have been afflicted by natural disasters (Aksoy, 2003).

1.2 Insect vector, *Glossina morsitans*

The *Glossinidae* spp are obligate blood feeders (hematophagous), reproducing by a method known as adenotrophic vivaparity. There are 33 species and sub-species in this family, two of which are found in sub-Saharan Africa. *Glossina* spp are yellowish-brown in color, with some having stripes across the abdomen and they have dichoptic eyes distinguishing them from other flies (Jordan, 2003). *Glossina* also have hundreds of labelar teeth that are used to bite into the skin of the host.

The internal structure of *Glossina* comprises of narrow, long salivary glands that extend into the abdominal cavity (Figure 1.1). Salivary glands play a vital role during blood meals as they release anticoagulant enzymes, which help to keep the blood from clotting as the tsetse fly feeds. The pharynx also plays a pivotal role in the blood feeding process as its muscles are used to suction blood from the host. The posterior section of the proventriculus forms part of the fore-and midgut within which blood digestion and absorption occurs (Pollock *et al.*, and Gooding *et al.*, 2005). When the host blood is being transported to other organs it is enclosed in a perithrophic membrane, separating it from the midgut (Lehane, 1996).

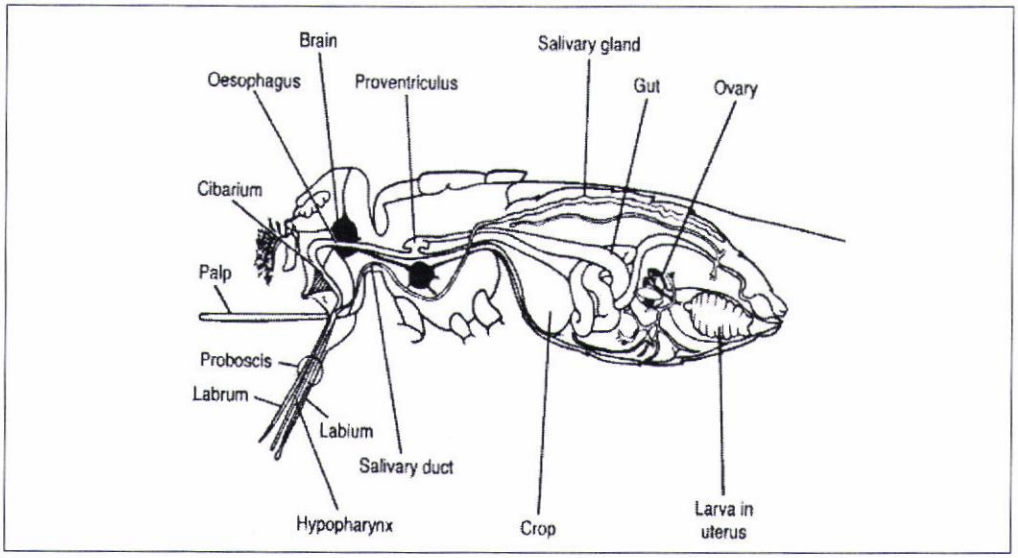


Figure 1.1 The anatomical structure of a female tsetse fly.

The tsetse fly can be divided into two main segments. The first segment is the head containing the labium, eyes and esophagus. The second segment is the body, containing the digestive system (gut, crop and salivary glands), and reproductive organs (From Aksoy, 2005).

1.2.1 Vector development cycle and reproduction

Female tsetse flies mate once per life cycle (90 - 100 days), while the males can mate one or more times during their life cycle. Upon mating the females fertilize their eggs in the uterus, a process that takes approximately four days (Attardo *et al.*, 2006). When fertilization nears completion, the first instar larva starts developing and upon completion, the larva emerges from the egg. The larva remains in the uterus until the development phase is over. During this period the female provides nourishment through the uterine glands (milk glands).

Larval development takes place in three instar stages, with the first stage lasting for 24 hours, the second for 36 hours and the third for 60 hours. At third instar the larva gets deposited onto the soil where it burrows. After several hours the puparium darkens and the larval cuticle hardens and becomes sclerotised (Pollock *et al.*, 1992; Attardo *et al.*, 2006 and Gooding *et al.*, 2006). The puparium takes 4 - 5 weeks to develop, and the young adult emerges from the puparium using its ptilinum (Figure 1.1). At this stage of the life cycle, the tsetse fly can potentially die due to exhaustion caused by the struggle to reach the surface. After emerging, the male and female tsetse flies seek for a blood meal to acquire the energy and nutrients necessary to build flight muscles. In addition, the females also use the meal to rear larvae (Attardo *et al.*, 2006).

1.3 *Trypanosoma* species and their development cycle

Two *Trypanosoma* sub-species are responsible for sleeping sickness in humans; *T. brucei rhodesiense*, which is prevalent in Southern and Eastern Africa and *T. brucei gambiense* mostly predominant in Central, West and some parts of East Africa (Aksoy *et al.*, 2005 and World Health Organization, 2006). *Trypanosoma* spp require a strong ability to adapt to different physiological environments as they are destined to go through rigorous conditions in the tsetse fly and the mammalian host. In the mammalian host, the parasites undergo a complex development cycle (Figure 1.2).

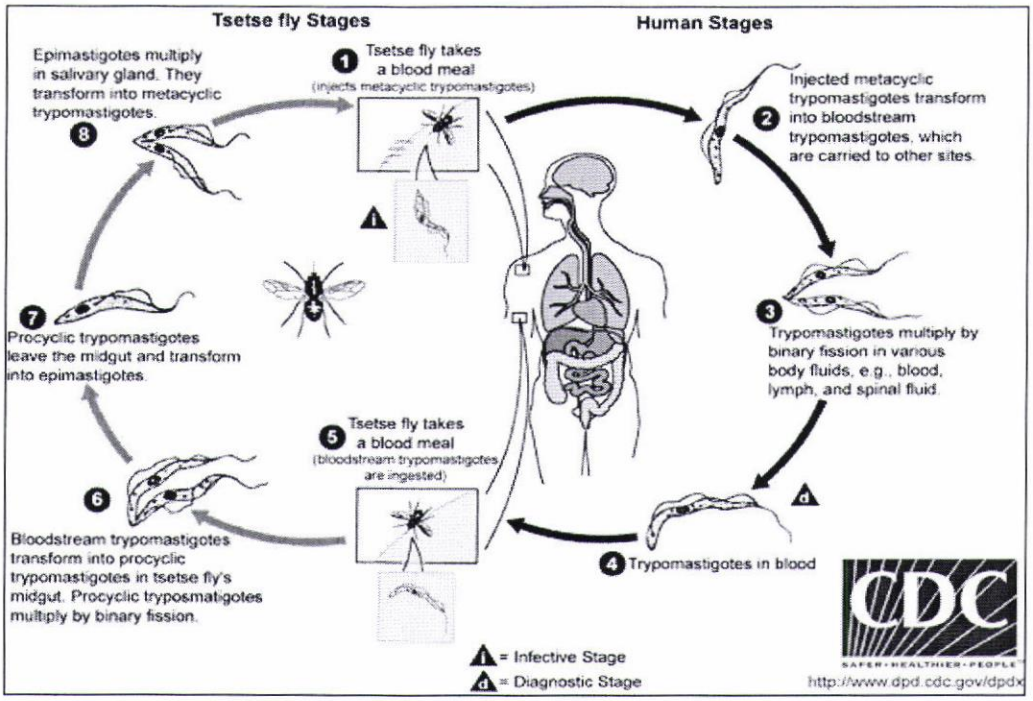


Figure 1.2 *Trypanosoma* development cycle in the insect vector and human host. Part of trypanosome development is carried in the human host (steps 1-4). Another part is carried out in the fly (steps 5-8)
(From http://www.dpd.cdc.gov/dpdc/HTML/TrypanosomiasisAfrican.htm).

The development cycle starts with the parasites establishing themselves in the bloodstream, wherein they express a variable glycoprotein (VSG) coat, which is crucial for the evasion of the mammalian host immune system (Vickerman *et al.*, 1988; Matthews, 2005 and Taylor, 2006). The development cycle proceeds to the next stage and the parasites multiply in the bloodstream with non-proliferative forms replacing

slender forms (G1-phase division arrest). Non-proliferative forms serve as markers to indicate when replication has reached its peak and this step also ensures that the trypanosomes are able to evade the host (Figure 1.2). G1-phase division arrest plays a pivotal role in assuring that the necessary morphological changes required for transmission into the vector take place (Vickerman, 1988; Mathews, 2005 and Aksoy, 2005). The duration of the life cycle within the mammalian host is different for each *Trypanosoma* species. Upon completion of the cycle, the trypanosomes are ready to be transferred to the vector with the next blood meal (Figure 1.2). In the bloodstream of the tsetse fly, the parasites switch the VSG coat to a GPI-anchored procyclins coat in the midgut (Vickerman, 1988 and Roditi, 2002). At this point parasites get extruded from the midgut of the tsetse fly by a process called attrition (Figure 1.2). As a result approximately 25% of the *Trypanosoma* population survives (Aksoy 2005; Mathews, 2005 and Vickerman, 1998). Those that survive are transferred to the salivary glands, forming epimastigotes, which attach themselves to the gland wall using flagellar membranes. Further replication takes place and the parasites undergo another cycle of division arrest and they re-acquire a VSG coat. At the end of division arrest epimastigotes get released in the lumen of the salivary glands (Figure 1.2). The epimastigotes then produce non-proliferative metacyclic forms, which acquire a new coat in preparation for transmission to a new mammalian host (Mathews, 2005 and Taylor, 2006).

T. brucei follows the development stages outlined above, while *T. congolense* and *T. vivax* follow a slightly different course. In *T. congolense* trypanosome parasites attach to the hypopharynx instead of the gland wall, the parasites then undergo further development producing mature metacyclic forms. *T.vivax* however, evades the fly by

migrating straight to the foregut instead of the midgut. From there trypanosome parasites take a similar route to that of *T. congolense*, producing mature metacyclins (Vickerman, 1998 and Aksoy, 2005).

1.4 Human African trypanosomiasis, clinical symptoms, drug therapeutics and vector control strategies

1.4.1 Clinical symptoms of sleeping sickness

The bite of a tsetse fly while feeding on mammalian blood can cause the formation of a skin lesion (chancre). Subsequently, parasites multiply in the blood stream, while parasitemia may also be detected in the lymph nodes, spleen and liver. If not diagnosed, as often is the case in many poor sub-Saharan African countries, especially rural areas, the parasites migrate to the central nervous system (CNS) through the Blood-Brain Barrier (BBB). In *T.b. rhodesiense*, parasites cross to the CNS within weeks, while this takes a longer period (months to years) in *T.b. gambiense*. However before the parasites penetrate the BBB, a person will present early phase symptoms (hemolymphatic phase) that includes fever, headache, malaise and lymphadenopathy. As the disease progresses to the second stage (encephalopatic stage) cutaneous lesions, hair loss and reproduction dysfunction can be observed. Crucially, at encephalopatic stage many organs go into distress resulting in heart failure and several endocrine problems (Kennedy, 2005 and Steverding, 2008).

Active screening for individuals presenting sleeping sickness symptoms is vital for preventing many infected people from reaching the encephalopatic stage.

However as disease surveillance has broken down due to civil unrest and other contributing factors, the disease remains undetected for the majority of these poor communities. If not treated, sleeping sickness may lead to death in as many as 10% of infected cases (Aksoy, 2005 and Steverding, 2008). Clinical therapeutics can be used once the disease is diagnosed. Although many of these regimens present numerous undesired side effects, they are the most effective treatment available currently. However, once the disease reaches encephalopatic stage, very few treatment regimens are effective (Aksoy, 2005).

1.4.2 Drug treatment for trypanosomiasis

The treatment course for HAT is physiologically demanding for the person infected with the disease, as many regimens have very intense and toxic side effects. Thus, correct diagnosis is vital, in that it helps establish the progression of infection. There are different compounds available for the treatment of HAT caused by either *T.b. gambiense* or *T.b. rhodesiense,* all of which are the same drugs that have been used for HAT treatment for the past 50 years (Fairlamb, 2003 and Kennedy, 2006). First stage sleeping sickness is treated with Suramin and Pentamidine for *T.b. rhodiense* and *T.b. gambiense* infections respectively. Pentamidine can only be administered through the intramuscular route as intravenous administration causes a severe hypotensive reaction. Second stage infections are treated with Melarsoprol, the only approved drug used to treat both *T.b. rhodiense* and *T.b. gambiense* as it can cross the BBB (Kennedy, 2004). Melarsoprol an intravenously administered drug presents a whole host of side effects including reactive encephalopathy.

Eflornithine treats late-stage HAT caused by *T.b. gambiense*, however it has to be taken by choice, as it is costly and very difficult to administer. Eflornithine has to be infused four times a day at 400 mg kg^{-1}, for 7-14 days. In preliminary combination therapy, Eflornithine shows synergism when used in Melarsoprol-resistant trypanosomiasis suggesting that a combinatorial regimen could be successful in clinical use. However, all possible rationale for this synergism requires further exploration before administration as combination therapy (Fairlamb, 2003 and Chappuis, 2005).

The toxicity, poor efficacy and other reasons that cause current HAT drugs to be ineffective limits the treatment of sleeping sickness, thereby motivating the need for development of new drug targets.

1.4.3 Vector control strategies

It is clear that the treatment of HAT is limited to a few drugs, which are not very effective and the parasite (*Trypanosoma* spp) has a very complex developmental life cycle, thereby making it very difficult to design vaccines for the disease. Another possible way of controlling HAT infections is to look at vector control management strategies. Currently these strategies involve the use of insecticides, fly-reduction, target and traps, as well as aerial spraying. Many of the control strategies have had notable success, especially in farming and agricultural settings (Aksoy, 2003). Nigeria has used Dichloro-Diphenyl-Trichloroethane (DDT) for almost 10 years in ground- and aerial spraying together with other insecticides, resulting in sleeping sickness being eradicated from this country. In South Africa, Botswana and Zimbabwe, HAT has been reduced to very small incidences due to successful use of these methods. In Zanzibar and Burkina-

Faso a systematic approach known as Sterile Insect Technique (SIT) was successfully applied to eradicate a *Glossina* species (*Glossina austeni*). In SIT, genetic tools are applied to sterilize large numbers of male flies, which are released into the environment to mate with females without creating any progeny, thus markedly reducing the population. However, the cycle of rearing and releasing males into the rest of the population has to be repeated approximately every four generations to be successful. SIT works well when used in conjunction with trapping and other vector control methods (Allsopp, 2001 and Aksoy, 2003).

Sustainability of these control strategies has been challenged, as societal issues play a huge role in their implementation. Financial backing is crucial, which is currently non-existent as most of these countries are impoverished and lack infrastructure due to years of civil unrest. Vector resistance at present hampers the use of insecticides, as well as the use of nets and trapping (Aksoy, 2003 and Aksoy *et al.*, 2005). Therefore, there is a pressing need to direct research efforts to molecular research, genomics and comparative genomics, as this knowledge will provide an understanding of vector-parasite interactions. Many studies have focused on studying the biology of the trypanosomes, while there are very few studies that have looked at the biology of the tsetse flies, as there was little or no data available for such work. As part of the efforts to provide control strategies for tsetse flies and the parasite and by extension sleeping sickness, it is important to look at insect immunity and genes responsible for refractoriness (parasite resistance) as these genes can be used as targets for pharmalogical intervention in sleeping sickness.

1.5 Invertebrate host defense responses

Immunity is a mechanism used by organisms for protection against invading microbes (Beck, 1996). Vertebrates use both adaptive and innate immunity as defense mechanisms. Adaptive immunity is further divided into two defense response systems, which are cellular and humoral immunity. Cellular response uses T-lymphocytes to recognize antigens via specific receptors. Humoral immunity mainly uses B-lymphocytes, which upon binding to specific antigens of foreign microbes release antibodies facilitating the elimination of pathogens (Silverman *et al.*, 2001).

In contrast, invertebrate species only have innate immunity as a defense mechanism (Dimipoulos, 2000 and Osta, 2004). Invertebrate innate immunity is divided into two defense systems. The humoral defense system, which includes antimicrobial peptides (AMPs), induction of Lectin synthesis and proPO synthesis. The second defense response includes phagocytosis and encapsulation. The two defense systems overlap in some parts, as many humoral factor molecules induce hemocyte-mediated response. Some of innate immunity defense systems have been studied in invertebrate species such as *Bombyx mori*, *Tenebrio molitor*, *Holotrichia diomphalia* larvae, *Anopheles gambiae* and *Drosophila melanogaster* (Iwanaga, 2005).

Recent studies have looked at tsetse fly immune responses. Like many other insects tsetse flies mount robust immune response against any infection, such that the invading microbes are subjected to harsh conditions, which markedly reduce invading parasites, especially in the midgut (Hao *et al.*, 2003). Studies have shown that there is a link between refractoriness and innate immune response (Hao *et al.*, 2001; Hao *et al.*, 2003

and Hu *et al.*, 2006). Some steps used in humoral and cellular immune response by insects will be reviewed.

1.5.1 Pathogen recognition

Upon encountering a pathogen, invertebrate's pathogen recognition receptors (PRRs) bind to pathogen associated molecular patterns (PAMPs). These PAMPs would be expressed on the surface of the pathogen. They include lipopolysaccharides (LPS) β-1,3glucans and peptidoglycans. There are specific pattern recognition receptors (PRRs) for gram-negative and gram-positive bacteria respectively. Some examples of pattern recognition receptor (PRR) proteins found in different insects are summarized in Table 1.1. Not included in the table are homologs PGRP-LB, PGRP-LC, PGRP-Cx and PGRP-SA identified in G. *morsitans*, where fly ESTs were compared against the *Drosophila* genome (Attardo *et al.*, 2006).

Table1.1 A summary of invertebrate pattern recognition receptor proteins.*

PRRs	Species	Family member	Function / Pathogen recognized
PGRP-S	*D. melanogaster*	PGRP-SA	Gram-positive bacteria and activates Toll pathway
	D. melanogaster	PGRP-SC1B	*Staphylococcus aureus*
	Bombyx. mori	PGRP-S	*Micrococcus. luteus* PGN
	Trichoplusia ni	PGRP-S	Micrococcus luteus
PGRP-L	*D. melanogaster*	PGRP-LB	*Escherichia coli*
		PGRP-LC(a) PGRP-LC(x)	Gram negative bacteria and activates proteases that cleave spatzle
		PGRP-LE	
GNBP/βGRP	*D. melanogaster*	GNBP-1	PGN from *Lactobacillus plantanum*
	Bombyx mori		LPS, β-1,3-glucan Gram-positive bacteria
	Manduca sexta	βGRP1 and 2	β-1,3-glucan
dsR-C	*D. melanogaster*	dsR-CI	β-1,3-glucan
Hemolin	*Manduca sexta*	Hemolin-1	Gram positive bacteria
Immulectin	*Manduca sexta*	Immulectin-2	LPS (Lipid-A and O-specific antigen)
Thioester-containing protein	*D. melanogaster* *A. gambiae* *A. aedes*	TepI-IV Tep1-15 Tep2, 3	LPS

* From Royet, 2004

1.5.2 Signaling

The recognition of foreign invaders by PRRs activates Signaling cascades and induces effector responses (Figure 1.3). Activated PRRs such as PGR-SA and GNBP-1 by gram-positive bacteria bind to cytokine Spatzle thus activating the Toll pathway downstream (Figure 1.3). Upon activation PGRP-LC and PGRP-LE activate the IMD pathway. Activation of either pathway (Toll or IMD) induces production of AMPs such as Attacin, Diptericin and Drosomycin (Figure 1.3)(Attardo *et al.*, 2006 and Wang *et al.*, 2008). Clip-domain Serine proteases (CLIPs) play a key role to signal transduction and modulation as they activate downstream processes, which will lead to AMP synthesis, hemolymph agglutination and melanization and later the killing of invading microbes. Serine proteases are also used to convert inactive pPO to active phenoloxidase (Figure 1.3). These Serine proteases are regulated by Serpins, which bind in an irreversible manner to the active site of the proteases and thereby modulating the signal cascade to ensure that proteases are not activated prematurely (Christophides, 2004; Attardo *et al.*, 2006 and Wang *et al.*, 2008).

Signaling cascades have been studied extensively in *D. melanogaster*. Investigations conducted on tsetse flies and their interaction with parasites upon infection suggests that immune responses may not be induced in the initial stages of infection. Seemingly, immune responses are only triggered later with the increase of parasite infection in the midgut (Lehane *et al.*, 2004).

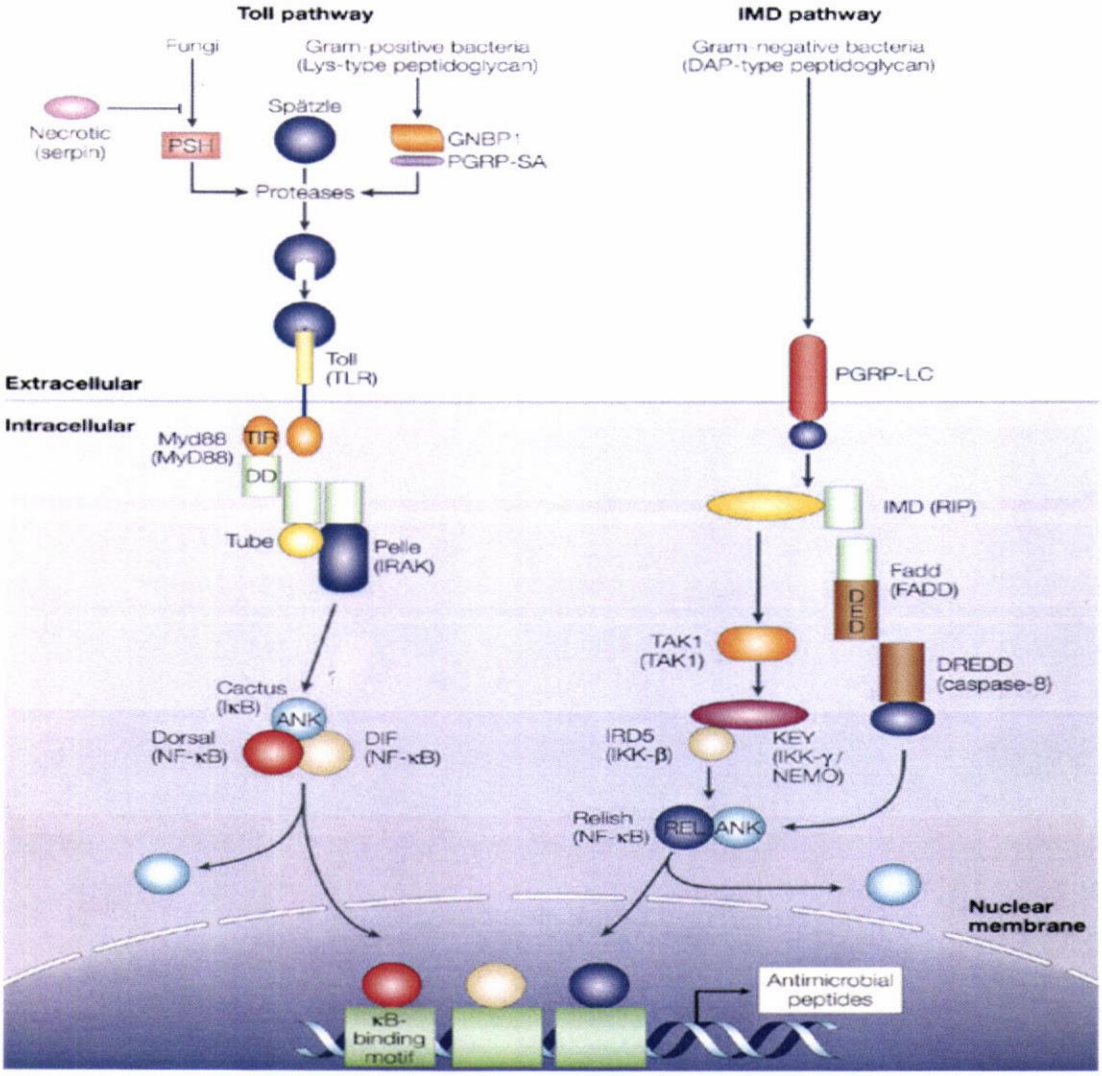


Figure 1.3 Toll and IMD pathways in invertebrates.

Gram (+) bacteria and fungi activate PRRs (PSH, GNBP-1 and PGRP-SA). These interact with a cleaved Spatzle activating the Toll pathway. Downstream interactions of Dorsal and Dif (NF- kβ factors) lead to the expression of genes that encode antimicrobial peptides such as Drosomycin. Gram (-) bacteria and diamino (DAP-like peptidoglycan) activate PGRP-LC, which recruit the immune deficiency (IMD) pathway. The following step involves the interaction of Fadd, Dredd and Relish leading to the expression of antimicrobial peptides (From Lemaitre, 2004).

Binding receptor-proteins in the Toll pathway recruit Myd88/Tube and Pelle death domain proteins (Kurata, 2005). These death-domain proteins assemble to form a complex, which induces the phosphorylation of I_K-B-like inhibitor Cactus using an unknown kinase. The complex causes the dissociation of Rel/ NF-$_K$B transcription factors from Cactus, which is phosphorylated and degraded by the proteasome. Cactus will be translocated to the nucleus to activate various proteins such as antimicrobial peptides (Figure 1.3) (Aggarwal, 2008; Wang, 2004 and Leulier, 2000). Gram-negative bacteria as well as LPS and PGN activate the IMD pathway. When IMD signaling cascades are activated, TAK-1 gets recruited and used to activate the IKK complex (Figure 1.3). Although the molecules involved in the interaction between TAK-1 and IMD interaction are not very well understood, it is known that the DREDD protein plays a role in the signaling process (Figure 1.3). TAK-1 activates NF-$_K$B/Relish transcription factors, prompting pathogen elimination mechanisms such as phagocytosis (Wang, 2004 and Aggarwal, 2008).

Until recently insect immunity was poorly understood and studies conducted using *Drosophila* have helped elucidate mammalian immunity. This knowledge can now be used to conduct comparative research towards elucidating immune cascades in other insects such as *A. gambiae* and *G. morsitans*, this is important for vector and disease control.

1.5.3 Pathogen elimination

Elimination of non-self involves the secretion of antimicrobial peptides (AMPs) into the hemolymph upon infection, carrying out immune reactions (Osta *et al.*, 2004 and Bulet, 2004). Another mechanism of elimination is phagocytosis, whereby small invading microbes are engulfed and degraded by hematocytes. Phagocytic molecules include oenocytoids, adipo-hemocytes, granulocytes and thrombocytoids. In *D. melanogaster* plasmatocytes are used for the disposal of microorganisms and apoptotic cells, while lamellocytes are used for encapsulation and crystal cells execute melanization (Dimipoulos, 2003). Melatonic encapsulation is used to eliminate bigger pathogens, whereby hematocytes fully adhere to attacking microbes forming a capsule.

Thioester-containing proteins (discussed in detail in section 1.6) are a class of proteins used in recognition of foreign microbes (PRRs) by binding directly to surface molecules of PAMPs.

1.6 Thioester-containing protein superfamily

The Thioester-containing protein superfamily is found in various taxa such as mollusks, fish, nematodes, birds and mammals. The TEP superfamily forms part of innate immunity in both vertebrate and invertebrates (Blandin and Levashina, 2004). TEP superfamily proteins function by labeling foreign microbes and activating signaling pathways and cascades, which induce the destruction of invading pathogens. Studies done by Dodds and Law (1998) on the evolution of thioester-containing proteins indicate that this protein family is part of the complement system that predates the

appearance of molecules such as the Major Histocompatability Complex (MHC) and antibodies, which are the main components of the adaptive immune system in vertebrates. However, details of complement system evolution remain unclear and require further investigation. In invertebrates, particularly insects, the thioester-containing superfamily is divided into three sub-families or subgroups namely complement-factors, ∝-2-macroglobulins (∝-2-Ms) and invertebrate TEP proteins. TEP superfamily proteins from various species including some insects are shown in Figure 1.4. There are two features that characterize the TEP superfamily (Dodds and Law, 1998; Blandin and Levashina, 2004)

(i) A canonical β-cysteinyl-γ-glutamyl thioester region with an amino acid signature of [GS] C[GA]E[EQ]

(ii) A high affinity for multiple binding interactions which are conformationally sensitive

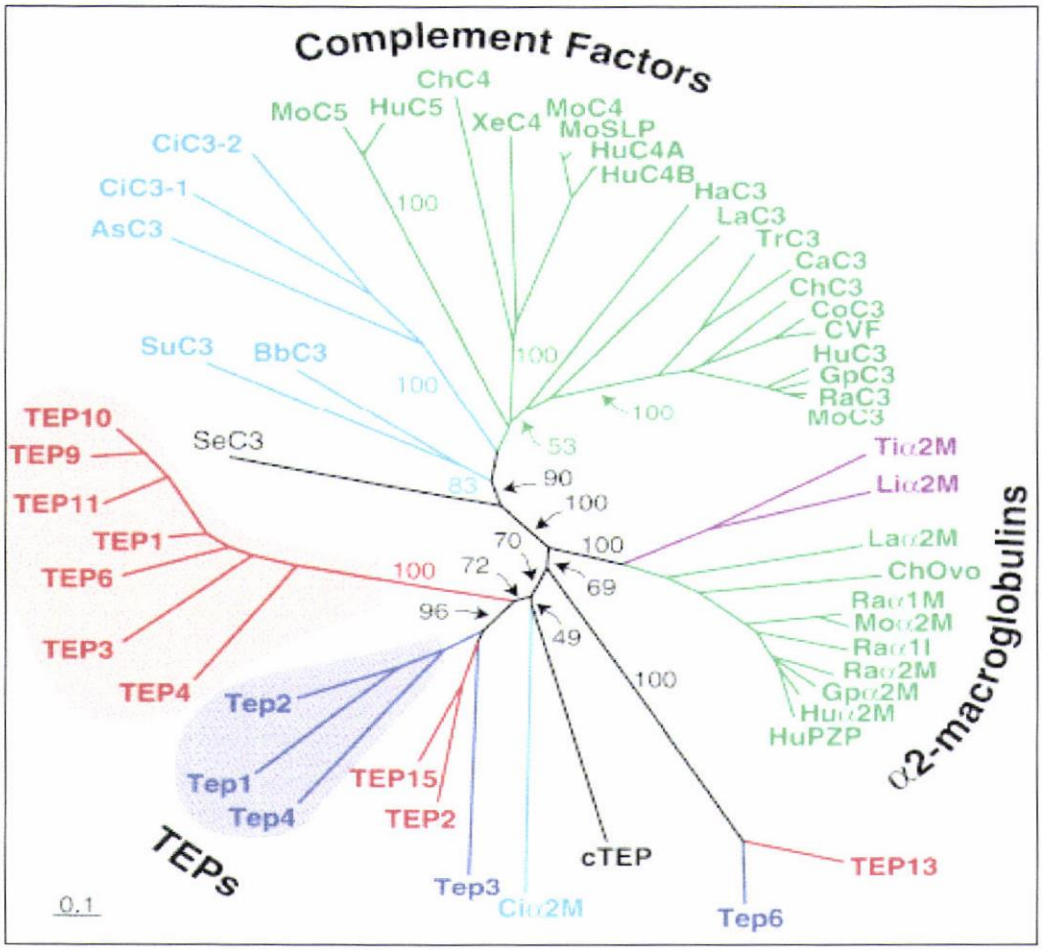


Figure 1.4 A phylogenetic tree depicting thioester-containing proteins.

The superfamily is divided into three subgroups (invertebrate TEPs, α-macroglobulins and complement factors). Representative insect TEPs, *A. gambiae* and *D. melanogaster* TEPS, are shown in red and purple respectively. Complement factors and α-macroglobulins for vertebrate species are shown in green. Primitive deuterostomes are shown in cyan (From Blandin and Levashina, 2004).

The first sub-family, complement factors (C2-C5) in higher vertebrates is comprised of proteins that encode the alternative, classical and lectin pathways that function together to activate the C3 factor, which binds to the surface of microbes. Binding of the C3 factor labels the microbes for termination or the lytic pathway.

In vertebrates the classical pathway is activated by antibody-antigen interactions that bind to Clq/Cls/Clr complexes. The activation of alternative pathway is dependent on formation of the C3 convertase. Mannose-binding lectin (MBL) initiates the lectin pathway by interacting with mannose sugars on bacterial cell surfaces. Additionally, there is an MBL-associated serine protease (MASP), which functions in place of Cls and Clr to activate C2 or C4 in the classic pathway (Smith *et al.*, 1999).

Studies conducted on invertebrate complement factors identified the C3/C4/C5 complex, which is homologous to higher vertebrates (Smith *et al.*, 1999). Work conducted by Nonaka *et al* (1999) also identified a C3 homolog (represented as SpC3 and ASC3) in Sea urchins and Tunates. Analysis of the SpC3 and ASC3 identified features such as the leader region, β-∝, γ-∝ junctions (thioester bond region), a C3 convertase site, cysteines in various conserved sites and a disulfide bridge. Further analysis of SpC3 and ASC3 shows sequence modifications from the vertebrate C3 complement factor, suggesting altered function (Smith *et al.,* 1999 and Nonaka *et al.*, 1999). Invertebrates (Sea urchins) also have a factor B (Bf) known to interact with C3b in vertebrates during the formation of C3 convertase. Comparison of the sequence structure of Bf in Sea urchins, Tunates and humans showed that Sea urchins have three short-consensus repeats (SCR), while human and Tunates have 5 SCRs (Smith *et al.,* 1999). MASP (MBL-associate serine proteases) homologues were also identified in Sea urchins, which are characterized by a CUB domain, an epidermal growth factor (EGF), a second CUB

domain, two SCR domains and a serine protease domain. All homologues identified in Sea urchins and Tunates showed a strong association with alternative and lectin pathway (Smith *et al.,* 1999).

Later studies conducted by Blandin and Levashina (2004) show that there are key sequence features that characterize complement factors such as the anaphylatoxin cleavage site found approximately 70 amino acids downstream of the thioester region. They also have an excision motif with amino acid signature R[RK][RK]R upstream of the thioester region and a catalytic histidine residue located approximately 40-100 amino acids downstream of thioester (Nair, 2005). Attacking pathogens activate the cleavage of complement factors producing a thioester and anaphylatoxin fragments (C2a/C3a/C4a), which are released to site of infection and bind to the pathogen respectively. When anaphylatoxin fragments are released in the site of infection they recruit macrophages to the infected site (Levashina, 2001 and Smith *et al.*, 1999)

The second sub-group, ∝-2-macroglobulins is composed of protease-binding proteins released in the hosts' plasma upon infection by either binding to these attacking proteases blocking access from substrates to their active site (pan-protease inhibitors), or by trapping the protease in "cage" of its macromolecules which also prevents interaction with substrates (α-2-macroglobulins). The latter set of proteins also function as growth factors (Amstrong *et al.*, 1999).

∝-2-macroglobulins have been characterized in species such as Limulus polyphemus and they are characterized three important sequence features (Figure 1.6) (Amstrong *et al.*, 1999).

(i) A domain with different cleavage sites for attacking proteases, known as the bait region (a hypervariable region)

(ii) A thioester region

(iii) Conserved serine residues

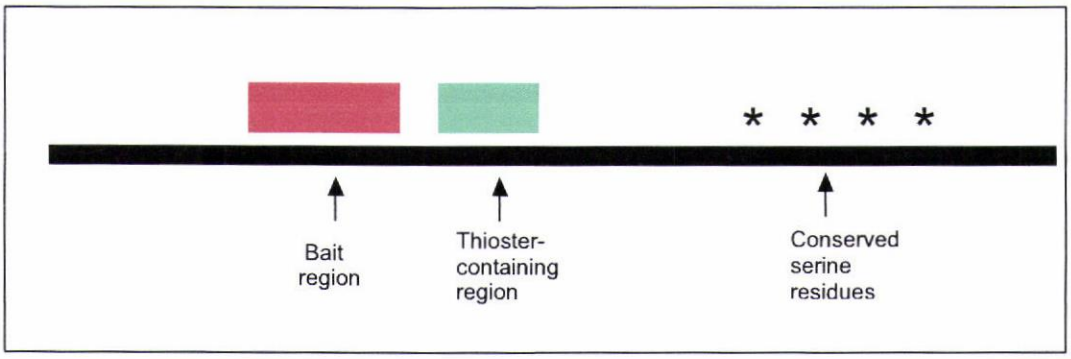


Figure 1.5 Diagrammatic representation of α-2-macroglobulin domains.
Upstream is a bait region (red), followed by a thioester region and conserved serine residues (responsible for thioester binding specificity) located downstream (From Saravanan, 2003).

Up to date invertebrate TEPs (third sub-family) are the best characterized of the three sub-families. Previous studies suggested that invertebrate TEPs were more closely related to complement factors (Smith *et al.*, 1999). In contrast, studies conducted by Blandin and Levashina (2004) show that invertebrate TEPs are more closely related to ∝-2-macroglobulin.

As shown by the phylogenetic tree constructed wherein a long branch separates the complement factor proteins from the invertebrate TEP proteins (Figure 1.4) (Zhang *et al.*, 2008). There is an important need to study the TEP superfamily further to establish sequence and function differences between these three sub-families. Invertebrate TEPs (third sub-family) are discussed further in more detail. TEPs were recently characterized in *D. melanogaster* and *A. gambiae* (Blandin *et al.*, 2004).

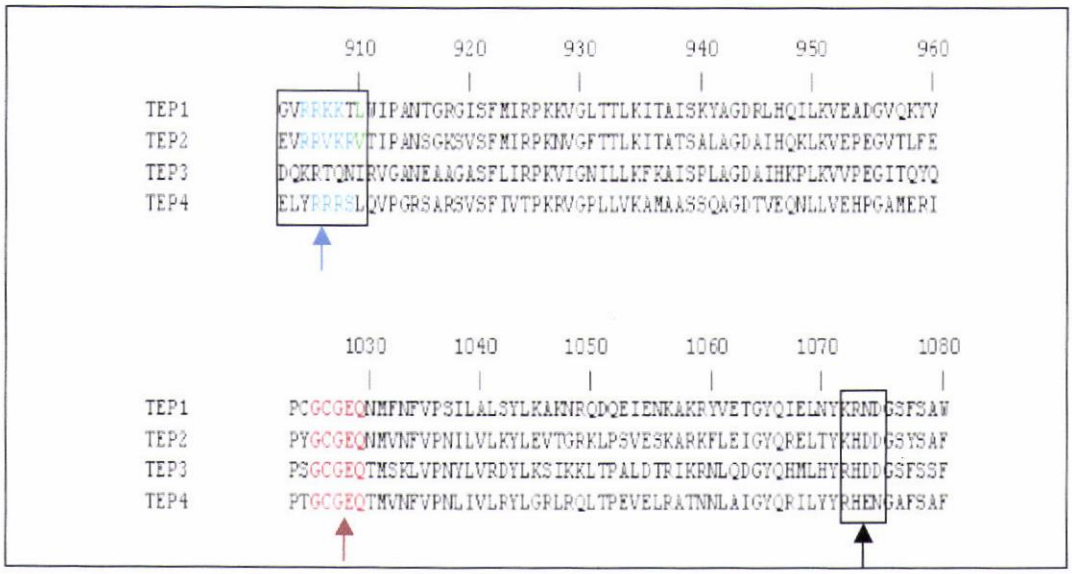


Figure 1.6 Alignment of D. *melanogaster* highlighting key sequence features of TEP proteins.

An excision motif upstream of the thioester region (blue), the canonical thioester region (red) and the catalytic histidine (black box) 40 amino acids downstream of the thioester region (From Nair, 2005)

Most of invertebrate TEPs have a thioester region as observed in thioester-containing superfamily proteins and a structure known as bait-like region similar to the bait region in α-2-macroglobulins (Jiggins, 2006).

There is an excision motif upstream of the thioester region with amino acid arrangements of R[RR]S, R[RK]R and R[RV][KR] (highlighted in Figure 1.5). In *Drosophila* there are four TEP proteins Tep1-4 containing a conserved thioester motif (Figure 1.5). Tep1 has been characterized in the fat body libraries after infection and in immune-challenged larvae. Bacterial infections induce up-regulation of Tep1, Tep2 and Tep4 (Blandin and Levashina, 2004; Obbard *et al.*, 2008). There are 15 TEP (Tep1-15) homologs identified in the *Anopheles gambiae* genome.

Of the fifteen, four pairs are haplotypes, which may be a result of polymorphic variations (Obbard *et al.*, 2008). The thioester motif is only conserved in Tep1 and Tep4 proteins. Nine of the TEP proteins lack this motif and show sequence modification in this region. It is suggested that the TEP proteins without the thioester motif might function as protease inhibitors, making them useful to insect immune responses (Blandin and Levashina, 2004). *Anopheles gambiae* Tep1 is the only TEP protein whose structure has been solved to date (Baxter *et al.*, 2007).

Tep1 was also used as a candidate gene in a study that looked at phagocytosis in mosquito immune response. The phagocytosis immune response studies revealed that Tep1 expression is up-regulated upon encountering bacterial infection. Tep1, a single molecule which is secreted into the hemolymph and gets cleaved at C-terminal upon infection with the C-terminal end fragment binding to gram-negative or gram-positive bacteria. Tep3 and Tep4 show similar results upon infection by bacteria. Tep1 and Tep3 are controlled by a Rel/Cactus cassette and are implicated in the killing of parasites by binding to the surface of ookinetes, which have crossed over to the midgut (Blandin and Levashina, 2007).

Characterization of TEP proteins in *G. morsitans* will provide knowledge that can be potentially used to complement the genome annotation of *Glossina*.

1.7 Comparative genomics and the characterization of immune-related gene families

Comparative genomics is a powerful tool that enables identification of new genes and regulatory elements for use as drug targets in many pathogenic organisms, by aligning unknown sequences onto genes with known function. Comparative genomic methods can be taken a step further and be used to look for chromosomal segments and functional elements across species being compared thereby determining orthologs and paralogs, depending on evolutionary distances of the species being compared (Hardison, 2003 and Sivashankari, 2007). For example a gene of survey comparing three malaria parasite species was conducted, with the aim of identifying homologs, which could be used as putative drug targets. In this study, EST transcripts of *Plasmodium berghei, Plasmodium falcipurum* and *Plasmodium vivax* were compared against existing protein data in public databases and a hundred new homologs were identified (Carlton *et al.*, 2001).

In addition, comparative genomics allows comparison of species that do not have fully sequenced and annotated genomes to those that are annotated. For example *Glossina morsitans* is a medically important vector, however its genome is not yet fully sequenced, therefore available data in the form of expressed sequenced tags (ESTs) will be used to conduct comparative genomics of the tsetse fly against annotated insects.

1.7.1 The *Glossina morsitans* genome project

The Glossina genome project was established by the International Glossina Genome Initiative (IGGI consortium) with the aim of providing data that would be used to aid the development of new control strategies for HAT by providing genomic data (Aksoy *et al.*, 2005). In the first phase of the project, data was published on functional annotation of the midgut and fat body transcriptomes (Lehane *et al., 2003* and Attardo *et al.*, 2006). In the midgut transcriptome 8876 sequence contigs were analyzed and putative function assigned to 4035 of the transcripts, of which 68 were immune-related. The 68 immune-related transcripts were further used in micro-array analysis to determine whether they would be up- or down- regulated in response to bacterial infection (Attardo *et al.*, 2006). In a fat body transcriptome analysis, 3059 consensus sequences were generated and putative function using homology-based methods was assigned. Consensus sequences with assigned putative function were further clustered into functional groups, some of which contained immune-related products (Lehane, 2003). The published data provided a starting point towards generating knowledge that will provide a better understanding of tsetse fly biology. More tissue transcriptome data was made available by the IGGI consortium including ESTs from fat body, head, salivary glands, midgut, reproductive organs, whole body larvae and pupae as well as adult flies (Aksoy, 2007). The ESTs were clustered and assembled, hereafter referred to as the *G. morsitans* transcriptome.

The EST data will be used to aid genome annotation once the tsetse fly genome is sequenced and assembled. Genomic resources are being generated alongside the genome-sequencing project, such as Bacterial Artificial Chromosome (BAC) paired ends for *G. morsitans* constructed with the aim of providing preliminary information to assist the whole genome assembly. The genome assembly of *G. morsitans* is currently at 3x coverage.

1.8 Study aims and objectives

The aim of the MSc project was to characterize putative immune-related genes in *Glossina morsitans* using comparative genomics. Existing resources such as the genome information of *D. melanogaster, A. gambiae and A. aegypti* were used to characterize these putative genes, as they have fully annotated genome sequences. The focus of this research is based on a sub-family of proteins known as thioester-containing proteins (TEPs), which function as pattern recognition receptors (PRRs) that bind directly to foreign invaders such as bacteria and eukaryotic parasites in many organisms including invertebrates. TEPs are also used to initiate phagocytosis and lysis, which eliminate the invader from the immune system. The characterization of this protein family in *G. morsitans* will contribute to the annotation of the genome once the sequencing is done. The putative identification of TEP homologs in *G. morsitans* together with the sequenced genome will provide the necessary information needed to further characterize the regulatory regions of these genes. The knowledge generated here can potentially be used in designing of target molecules in immune-related approaches to compromise or enhance the immune system in the fight of parasite infections.

Therefore this project aims to:

- Apply comparative and phylogenetic methods to identify and confirm TEP homologs in *Glossina.*
- Determine the evolutionary relationships of TEP proteins among *Glossina, Drosophila, Aedes* and *Anopheles.*
- Use the genomic organization of TEP loci in *D. melanogaster* and *A. gambiae* to better understand the evolution of these genes.
- Determine whether there are any family specific expansions in *Glossina.*

Chapter 2

Data and Methods

2.1 Tools and Datasets 31

2.1.1 *Glossina morsitans* dataset..... 31

2.1.2 Thioester-containing proteins datasets..... 32

2.1.3 Compara protein families dataset..... 33

2.1.4 Basic Local Alignment Search Tool (BLAST)..... 36

2.2 Methods..... 36

2.2.1 Similarity searches for TEP homologs in *G. morsitans*..... 36

2.2.2 Functional domain analysis..... 37

2.2.3 Phylogenetic analysis..... 37

2.2.2.3.1 Tree construction approaches..... 37

CHAPTER 2
DATA AND METHODS

2.1 Tools and datasets

2.1.1 *Glossina morsitans* dataset

The available *Glossina morsitans morsitans (Gmm)* EST transcriptome was analyzed. The transcriptome contained 15 615 consensus transcripts, 17 287 singletons and 11 222 predicted proteins. These *G. morsitans* ESTs were clustered using StackPack software (Christoffels *et al.,* 2001) by Mario Jonas generating consensus and singleton sequences. Stackpack is designed to perform rapid clustering of EST sequences. StackPack starts with sub-partitioning the data, then masking is performed using Crossmatch, which removes elements such as vector sequences, genomic repeats and mitochondrial sequences. The d2_clustering algorithm is used to cluster the masked sequences using a 95% identity in any 150bp window as criterion. Consensus sequences are generated using Phrap; sequences that do not group together with the consensus sequences will form singletons. Following the generation of consensus sequences CRAW is used to further refine the sequences and generate sub-clusters, which are linked together using a clone identification label. The open reading frames (ORFs) for the consensus and singleton sequences were also predicted and translated by Mario Jonas at SANBI using ESTScan (http://estscan.sourceforge.net) based on a *Glossina*-specific matrix, which was generated as follows:

1 All blastx-predicted contigs searched against Uniprot and the extrema of each contig match was assumed to be the CDS.

2 The 'artificial' CDS were extracted and used as input for the ESTScan scripts extract_mrna and build a model.

2.1.2 Thioester-containing protein dataset

Protein sequence searches were conducted through the NCBI protein database (http://www.ncbi.nlm.nih.gov/) using keywords such as "thioester-containing proteins AND species name" OR "thiolester-containing protein AND species name " OR "tep proteins AND species name " OR "TEP proteins AND species name".

The species names included *Drosophila melanogaster* or *Anopheles gambiae* or *Aedes aegypti*. The same TEP proteins obtained from NCBI were identified in published literature and these TEP proteins are reviewed in section 1.6 (characterized by Blandin, 2004; Jiggins, 2006 and Obbard *et al.*, 2008). These TEPs comprise 15 proteins identified in *A. gambiae,* four and two protein sequences in *D. melanogaster*, and *A. aegypti* respectively. Three of A. *gambiae* proteins (Tep6, Tep7 and Tep14) were excluded from the analysis, as they are isoforms of Tep8, Tep5 and Tep2 respectively, while Tep12 was excluded because the sequence was too divergent from other *A. gambiae* TEP proteins used in the analysis. The 17 TEPs were used to characterize TEP homologs in *G. morsitans*. *Anopheles gambiae* TEPs will be given an "Ag" prefix, *D. melanogaster* TEPs a "Dm" prefix, *A. aegypti* TEPs an "Ae" prefix and *G. morsitans* TEP homologs a "Gm" prefix.

2.1.3 Compara protein family dataset

Compara (Enright *et al.,* 2002) is part of an ensembl database that contains sequence similarity information for all species annotated in ensembl (http://www.ensembl.org). Compara_db is divided into two segments

1) Genomic alignments (DNA-DNA alignment data).

2) Protein families (Protein-protein alignment data), protein trees and homologs defined from protein trees.

A dataset obtained from the second segment (protein families) was analyzed. The protein families were downloaded from

(ftp://ftp.ensembl.org/pub/release-49/mysql/ensembl_compara_49).

Compara_db protein families are constructed by performing an all-against-all similarity search (PSI-BLAST) using a superset of all ensembl protein sequences and sequences obtained from Uniprot (Swissprot or Trembl) (Enright *et al.,* 2002). The second step is to store the results in a square matrix, which is translated into a graph, wherein the nodes represent proteins and edges represent sequence similarity. In the final step the graph is translated into another matrix by employing a mathematical algorithm. The matrix is used as an input to a Markov Chain Clustering based tool (TRIBE_MCL), which clusters the proteins into family classes. The protein classes are stored in Compara MySQL tables. The dataset analyzed was obtained by searching the following MySQL tables

1) Member_table containing values such as member_id, stable_id and description

2) Family_table containing all group homologs

3) Family_member_table containing family_id, member_id and cigar_line.

The tables were searched using MySQL commands to extract a family_id and species_id for each protein (Figure 2.1). The results (12 8041 protein families) were stored in a file (Compara_family.txt), containing a family identifier in the first column and the species identifier in the second column. Protein families containing fly-TEP proteins were extracted using a Unix command "grep <tep_identity> <filename>" from Compara_family.txt, which returned family identifiers and species identifiers for D. *melanogaster*, *A. gambiae* and *A. aegypti* respectively. The output was saved in a new file (Protein_fasta_file1) for each TEP protein as they belonged to different protein families. *A. gambiae*, *D. melanogaster*, *A. aegypti*, *Homo sapiens* and *Mus musculus* sequences (fasta format) were downloaded from a database (Ensembl) and saved in a file (Protein_fasta_file2). Even though the aim was to expand the TEP protein family search for the three insects species, *Homo sapiens* and *Mus musculus* protein sequences were added to determine whether TEP proteins are expanded beyond the invertebrate classes as identified in the literature. *A. gambiae*, *D. melanogaster*, *A. aegypti*, *Homo sapiens* and *Mus musculus* were aligned (Bl2seq pair-wise alignment) using an automated Bioperl script (appendix III). The Bioperl script takes a fasta file as input and creates an individual fasta file for each sequence. The fasta sequence files were aligned using an all-against-all approach with Bl2seq with default parameters and an output (.out) file was created for each alignment. The Bl2seq output files were parsed using a Bioperl parser script (appendix IV), which takes an output file from the Bl2Seq Bioperl script as an input (Figure 2.1). The script parsed the result file for sequence similarity matches with an expectation value of 10^{-2}, percentage identity (greater than 50%) and calculates HSP coverage (greater than 50%). The results of each file are appended to one file for each family (Figure2.1).

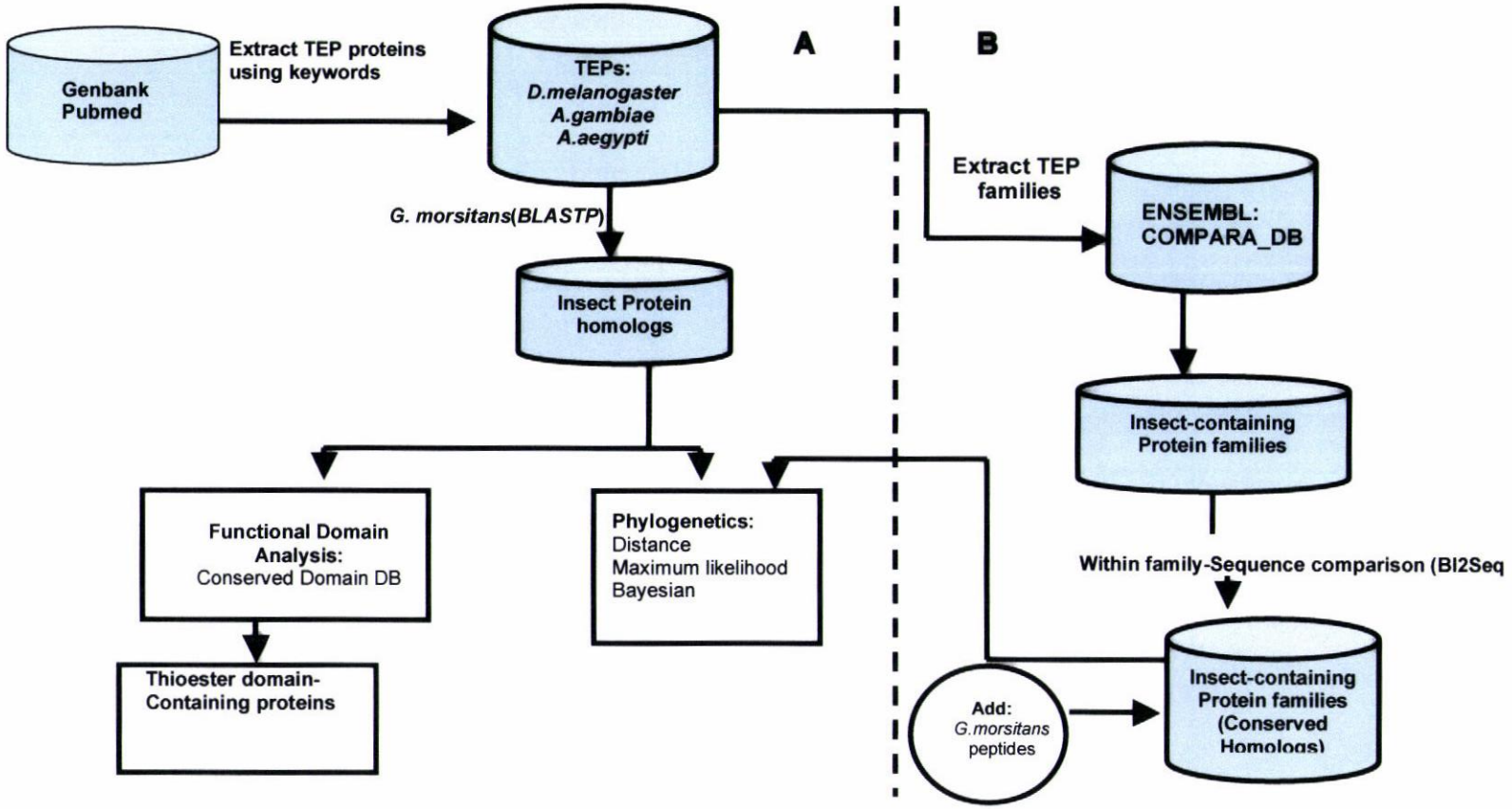


Figure 2.1 A schematic diagram of steps followed in the characterization of the TEP protein family in *G. morsitans*.
A. TEP insect families were collected from databases that are publicly available. Similarity searches were conducted to identity homologs in the *G. morsitans* transcriptome data. Insect protein homologs were used in phylogenetic and functional domain analyses.
B. Proteins were obtained from Compara_db to expand the TEP protein search. Sequence similarity searches were conducted to determine orthology.

2.1.4 Basic local alignment search tool (BLAST)

Basic Local Alignment Search Tool (BLAST) first developed by Altschul *et al.*, in 1990 and later improved in 1997 and 2004 is an algorithm designed to compare DNA and protein sequences. A query sequence is searched against a database of sequences wherein BLAST uses a heuristic approach to scan for word pairs with a score T and a High-scoring Segment Pair (HSP) alignment will be generated. BLAST takes a query file and a database of sequences, both containing fasta sequences (the details of BLAST are outlined in appendix I).

2.2 Methods

2.2.1 Similarity searchers for TEP homologs in *G. morsitans*

Sequence similarity searches were conducted using BLAST, which was pre-installed on a local server. A BLASTP was performed using A. *gambiae* Ensembl version 49 containing 13 621 sequences, *D. melanogaster* version 49 containing 20815 sequences and *A. aegypti* against version 49 containing 16789 sequences the *G. morsitans* transcriptome. For the blast searches a word size of 11 was used, the expectation value cutoff was 10^{-2} and DUST filters were set while other default parameters were unadjusted.

2.2.2 Functional domain analysis

TEP proteins identified in *A. gambiae*, *D. melanogaster*, *A. aegypti* and *G. morsitans* were screened for conserved domains by searching against NCBI Conserved Domains Database (CDD) (www.ncbi.nlm.nih.gov/Structure/cdd/cdd.html). The domain features are reviewed in the literature section 1.6. The program parameters were kept at default.

2.2.3 Phylogenetic analysis

2.2.3.1 Tree construction approaches

Three approaches were employed for the construction of phylogenetic trees for the TEP homologs namely PHYLIP (a neighbor joining approach), PHYML (a maximum likelihood approach) and MrBayes (Bayesian inference).

The neighbor joining approach (NJ) begins with a star tree built under the assumption that there is no clustering data. Phylip (NJ) then refines the topology of a tree by joining neighbors and producing new pairs of neighbors. Once pairs of neighbors are identified, they are combined into composite taxon, and this procedure is repeated until the final tree is produced (Nei and Kumar, 2000). PHYLIP version 3.68 was downloaded from (http://evolution.genetics.washington.edu/getme.html).

The maximum likelihood method was first used for the nucleotide framework (Felsenstein, 1981) and then later applied to amino acid sequences using Dayhoff's transition matrix (Dayhoff *et al.*, 1990 and Kishino *et al.*, 1990).

When compared to other methods, ML programs show the ability to recover the correct tree from simulated data more often than any other methods. Additionally this method uses a statistical framework to compare trees and evolutionary models. A disadvantage of the ML program is its inability to obtain an optimal tree with certainty even from moderate datasets due to computational difficulties. It thus relies on heuristics to obtain a near-optimal tree in reasonable computational time (Guindon, 2003). In ML the problem is more complex because not only does this program depend on the tree topology, it has to put numerical parameters into consideration as well (Chor *et al.*, 2000). PHYML version 2.4.4 was downloaded from (http://atgc.lirmm.fr/phyml)

MrBayes is a tool that is based on Bayes's theorem (Huelsenbeck *et al.*, 2001). MrBayes starts with *a priori* data, which is combined with likelihood to produce posterior probability. The Markov Chain Monte Carlo (MCMC) algorithm is used to formulate or approximate the posterior probability. MCMC perturbs each tree (by integrating all possible combinations) and a new tree is then proposed, which could be accepted or rejected, this step is repeated until a tree with the highest posterior probability is obtained. The challenge for this method is that in complex models, chains can fail to converge due to failure of proposing new states. The convergence problem can be circumvented by running test sets in order to determine the length of time in which the chains can be run to obtain good approximations of posterior probabilities (Huelsenbeck *et al.*, 2001). MrBayes version 3.1.1 was downloaded from (http://mrbayes.csit.fsu.edu/download.php) and installed locally.

The sequences were aligned using CLUSTALW using default parameters. The alignments were edited using Jalview (Clamp *et al.*, 2004) wherein all segments of the

alignment with gaps for all sequences were removed. Edited alignments were saved as fasta files and re-aligned with CLUSTALW and default parameters were unchanged. The alignment output were saved as (.phy) for PHYLIP and PHYML and (.nex) for Mr Bayes. A total of 17 TEP proteins from *D. melanogaster*, *A. gambiae* and *A. aegypti* were identified in the literature and were used to construct trees. Five putative TEP homologs were identified in *G. morsitans,* however Gmcn2281 and Gmcn4297 were more divergent from other TEPs and were excluded in the construction of the trees. Gmcn1115 and Gmcn1116 were generated from the same cluster therefore CLUSTALW alignments were done to determine whether they were isoforms and the results showed that indeed they are isoforms, thus Gmcn1116 was excluded from phylogenetic analysis. PHYLIP (Tuimala, 2006) was used to construct a neighbor-joining tree for the 19 TEP homologs, with 1000 bootstraps. The parameters were kept at default.

PHYML (Tuimala, 2006) was used to construct a tree for the 19 TEP proteins with 1000 bootstraps. The transition/transversion ratio was estimated, the proportion of invariable sites was estimated, the substitution rate was set at 4 and the gamma distribution was kept at default.

MrBayes was also used to generate trees for the TEP homologs using a batch script (see appendix II). The parameters were: evolution model was set at nset6 and rates at gamma, mcmc was set at 10000 in order obtain 1000 replicates from the posterior probability distribution. Once the posterior probability was produced, trees were summarized using the command sumt burnin (set to 250) producing a cladogram and a phylogram of the TEP homologs. All trees were viewed with FigTree software (http://tree.bio.ed.ac.uk/software/figtree/).

Chapter 3

Results

3.1 Thioester-containing proteins dataset..... 41

3.1.1 *Glossina morsitans* thioester-containing protein homologs.... 38

3.1.2 Functional Domain analysis..... 45

3.1.3 Phylogenetic analysis..... 48

3.1.4 Genome organization 51

3.2 Expanding the TEP protein family search..... 54

CHAPTER 3

RESULTS

3.1 Thioester-containing proteins dataset

Invertebrate thioester-containing proteins function by recognizing foreign invaders and can be divided into 3 sub-families: complement factors, α-Macroglobulins and invertebrate TEPs. An analysis of one sub-group (invertebrate TEPs) is described in this chapter.

Table 3.1 TEP homologs identified in literature and protein database searches.

TEP proteins	Protein Identifier (Ensembl)
Anopheles gambiae	
Agtep1	AGAP010815-PA
Agtep2	AGAP008368-PA
Agtep3	AGAP010816-PA
Agtep4	AGAP010812-PA
Agtep5	AGAP010814-PA
Agtep8	AGAP010831-PA
Agtep9	AGAP010830-PA
Agtep10	AGAP010819-PA
Agtep11	AGAP010818-PA
Agtep13	AGAP008407
Agtep15	AGAP008364-PA
Drosophila melanogaster	
DmtepI	FBpp0080369
DmtepII	FBpp00790133
DmtepIII	FBpp0079101
DmtepIV	FBpp0080795
Aedes aegypti	
Aetep2	AAEL008607
Aetep3	AAEL014755

Eleven, four, and two TEP homologs were identified in *A. gambiae*, *D. melanogaster* and A. *aegypti* respectively from literature and protein database searches. Ensembl (version 49) contains 19 *A. gambiae* TEP proteins. In contrast, literature surveys show that there are 15 true TEPs; therefore, Agtep16, Agtep17, Agtep18 and Agtep19 were excluded from the analysis, as they are either TEP isoforms or contained partial sequences.

3.1.1 *Glossina morsitans* thioester-containing protein homologs

Glossina morsitans expressed sequenced tags (ESTs) belonging to anatomical tissues (head, fat body, midgut, salivary glands, reproductive organs, male and female whole body) were clustered and assembled to generate a *G. morsitans* transcriptome (singleton and consensus sequences). Sequence similarity searches (BLASTP) were conducted and five putative TEP homologs were identified in *Glossina morsitans.* Results for contig sequence searches are presented in Table 3.2. Singleton sequence searches yielded no significant results.

Table 3.2 Putative TEP homologs identified in *G. morsitans* based on sequence similarity (BLASTP) searches*.

Putative G.morsitans TEP homologs (Contig ID)	Insect TEP proteins	Percentage identity (%)	Expectation value
Gmcn1115	DmtepII	51	0.0
Gmcn2398	DmtepIV	52	1e-166
Gmcn4297	DmtepIV	50	2e-78
Gmcn2281	DmtepIII	41	2e-169
Gmcn1116	Agtep8	36	6e-18

*Gmcn1115 showed sequence similarity to DmtepII while Gmcn2398 and Gmcn4297 showed sequence similarity to DmtepIV. Gmcn2281 and Gmcn1116 were below the similarity threshold (50%), however, they had significant expectation values, and hence they were considered to be putative TEP proteins.

EST cluster analysis of G. *morsitans* assigns Gmcn1115 to the same cluster (cluster 406) as contig Gmcn1116. The two contigs were aligned and compared using CLUSTALW (Thompson *et al.*, 1994) and GEPARD Dot matrix analysis (Krumsiek *et al.*, 2007) to determine whether Gmcn1116 and Gmcn1115 were isoforms of the same putative transcript (Figures 3.1 and 3.2). In the sequence alignment, residues 1-160 of Gmcn1116 aligned to residues 563-722 of Gmcn1115 (Figure 3.2). The dot plot confirms the results observed in the CLUSTALW alignment, namely that Gmcn1116 maps to the end region of Gmcn1115 (Figure 3.2). Therefore, Gmcn1115 and Gmcn1116 are isoforms of the same transcript.

```
                  ************************************************** *******  *********
Gmcn1115: NSIMSITFKAFDDGKKELSQHRFEVNKDNSLVLQTHVLPKSTRSISLEADGAGSSLIQLS: 563-
622
                  NSIMSITFKAFDDGKKELSQHRFEVNKDNSLVLQTHVLPKSTRS+SLEADG GSSLIQLS
Gmcn1116: NSIMSITFKAFDDGKKELSQHRFEVNKDNSLVLQTHVLPKSTRSLSLEADGVGSSLIQLS: 1-60

               *    *********  *  **  ****************    **   *  *              *   *        *****
Gmcn1115: YQYNLATKDDRPGFKVDIKPKILPSQQLQINICANYQPAVDDEIKESNMAVMEVALPSGY: 623-
682
               Y+YNLATKDD P FK+DIKPKILPSQQLQI +  CA+Y+P  ++I +SNMAV MEV+        LPSGY
Gmcn1116: YRYNLATKDDTPSFKLDIKPKILPSQQLQIEVCASYEPHASEKISQSN MAVMEVSLPSGY: 61-
120

               *******   *********   ******************    ******
Gmcn1115: IADNEKFNDILAVERVQRVDTENSDTKVIVYFDGLVEGEQ: 683-722
               IADNEKF+DILAVERV+RVDTENSDTKVIVYF+GLVEGE+
Gmcn1116: IADN EKFDDILAVERVERVDTENSDTKVIVYFNGLVEGEK: 121 160
```

Figure 3.1 A CLUSTALW alignment of Gmcn1115 and Gmcn1116.

Asterisks denote identical (conserved) amino acid residues, plus (+) signs indicate missing amino acids while red amino acid denote substitutions.

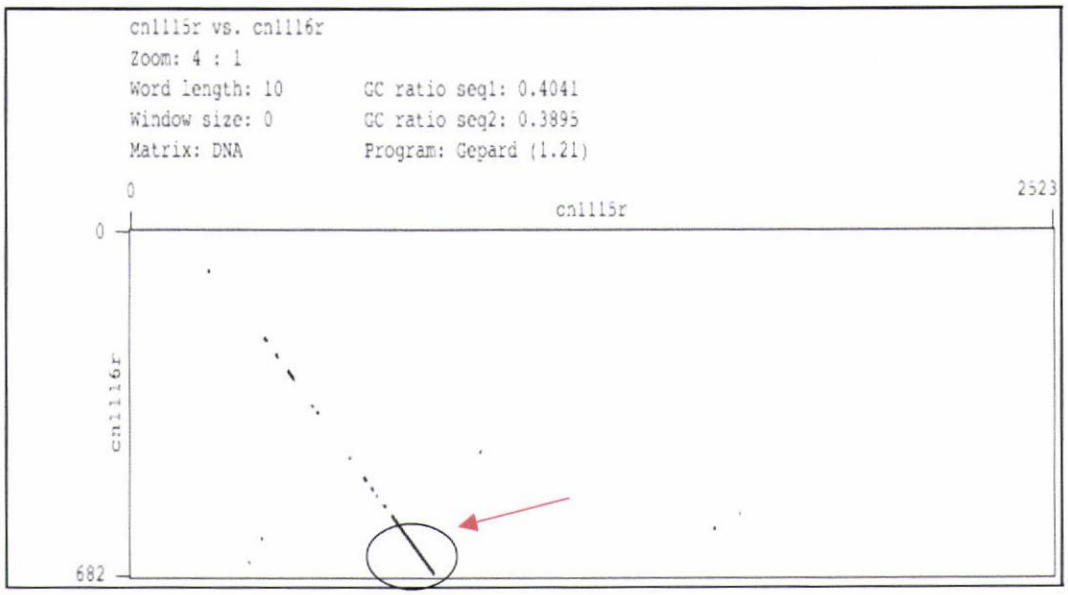


Figure 3.2 A dotplot alignment of Gmcn1115 and Gmcn1116 contigs.

A red arrow marks a region of high similarity between the two sequences; similar to results observed in the CLUSTALW alignment (Figure 3.1) Gmcn1116 (cn1116r) aligns to the end-region of the Gmcn1115 (cn1115r) sequence.

3.1.2 Functional Domain analysis

Conserved sequence signatures (domains) represent functional regions of proteins conserved through evolutionary time. Domains that are not conserved indicate a loss or gain of function due to selection pressures (Fong and Marchler-Bauer, 2008). A total of 19 TEP proteins belonging to *A. gambiae*, *D. melanogaster*, *A. aegypti* and *G. morsitans* were used to conduct functional domain analysis using CDD*. The results of conserved domain patterns identified in TEP proteins identified are presented in Figure 3.3. TEPs have the following domains:

(1) Alpha-2-macroglobulin family (A2M)

(2) Alpha-2-macroglobulin family, N-terminal region (A2M_N_2)

(3) Thioester-containing region (TED domain)

(4) Alpha-macroglobulin complement system (A2M_comp)

(5) Alpha-macroglobulin receptor (A2M_rec)

All *Drosophila* TEP proteins show conservation of four of the TEP domains (A2M, A2M_N_2, TED domain and A2M_rec), suggesting functional significance, especially that of the TED domain (Figure 3.3).

In *A. gambiae* TEPs, Agtep1 and Agtep4 show conservation of all TEP domains (A2M, A2m_N_2, TED domain, A2M_com and A2M_rec) (Figure 3.3). The rest of *Anopheles* TEP proteins show conservation of all domains excluding the TED domain (Figure 3.3).

In *A. aegypti* Aetep3 shows conservation of different TEP domains including the TED domain, while Aetep2 shows conservation of A2M, A2M_comp and A2M_rec. Putative homolog Gmcn1115 shows conservation of A2M, TED domain, A2M-comp and

* (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml).

A2M_rec. Gmcn2398, Gmcn2281 and Gmcn4297 show a conservation of A2M_N_2 and A2M but lack the TED domain (Figure 3.3).

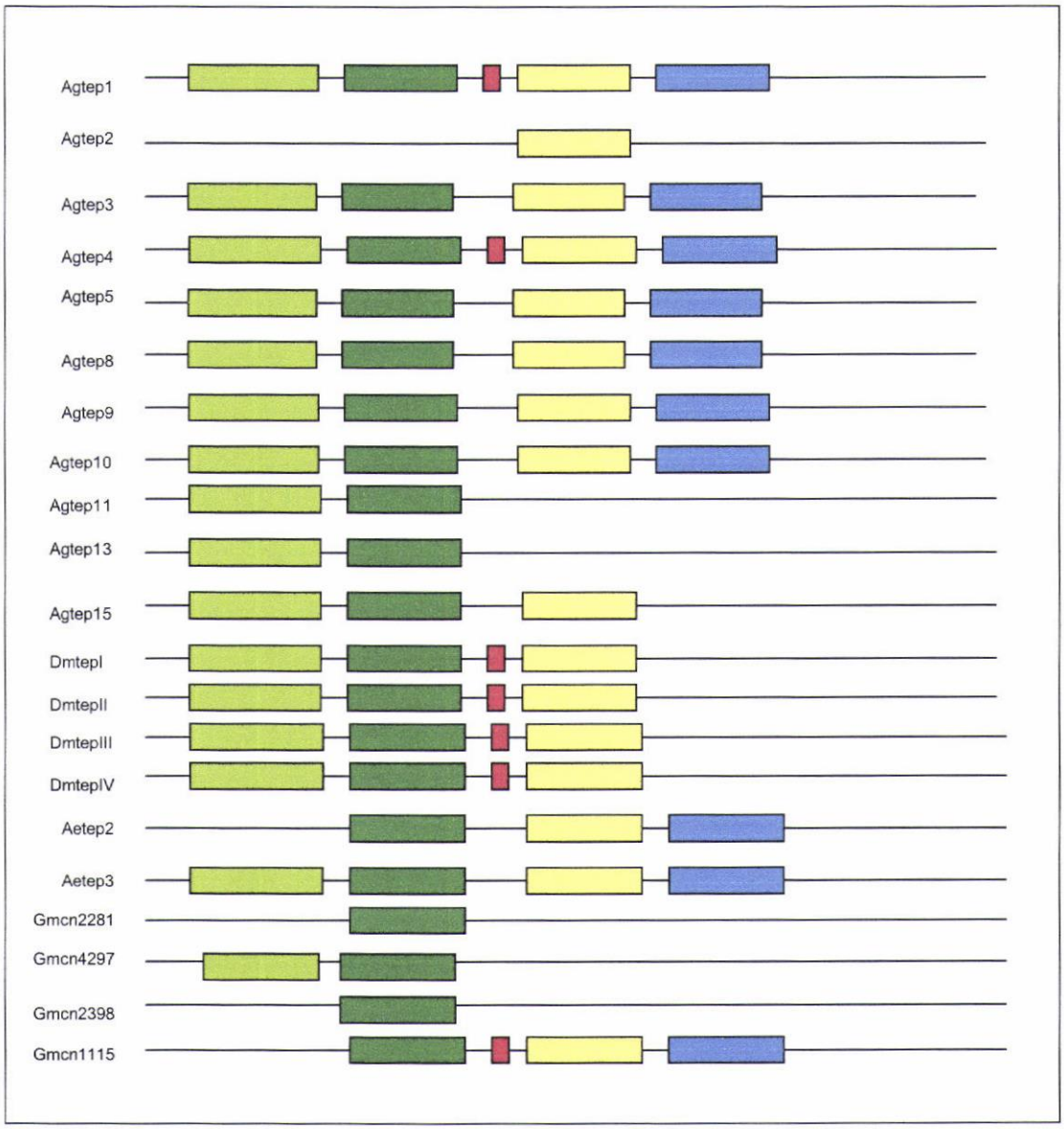


Figure 3.3 Sequence signatures (domains) expressed in TEP proteins of *D. melanogaster*, *A. gambiae*, A. *aegypti* and *G. morsitans.*
Alpha-2-macroglobulin (green), Alpha-2-N-terminal macroglobulin (light green), thioester-containing region (red), Alpha-2-macroglobulin component (yellow) and Alpha-2-macroglobulin receptor (magenta). TEPs that did not match the whole region of A2M_N_2 are denoted with an incomplete oval shape.

3.1.3 Phylogenetic analysis

To characterize *G. morsitans* putative TEPs and determine the evolutionary relationship of *A. gambiae*, *D. melanogaster* and *A. aegypti* TEPs, phylogenetic trees were constructed using the 19 TEP homologs. Agtep6 and Agtep7 protein sequences contained partial sequence information; for this reason, they were excluded from the analysis. Agtep12, Agtep14, Gmcn2281 and Gmcn4297 were also excluded from the analysis as they were shown to be too divergent from the rest of the sequences by the PHYLIP program. Gmcn1116 was omitted from the analysis because it represented the shorter of the two putative isoforms. Results from the distance method implemented in the NJ program of PHYLIP suite (Nei and Kumar, 2000) are presented in Figure 3.4. Cross validation of the phylogenetic trees was performed using a Bayesian based program, MrBayes (Huelsenbeck *et al.*, 2001) and maximum likelihood based program PHYML (Felsenstein, 1981 and Dayhoff *et al.*, 1990). The results for trees constructed using PHYML and MrBayes are shown in appendix V and appendix VI respectively.

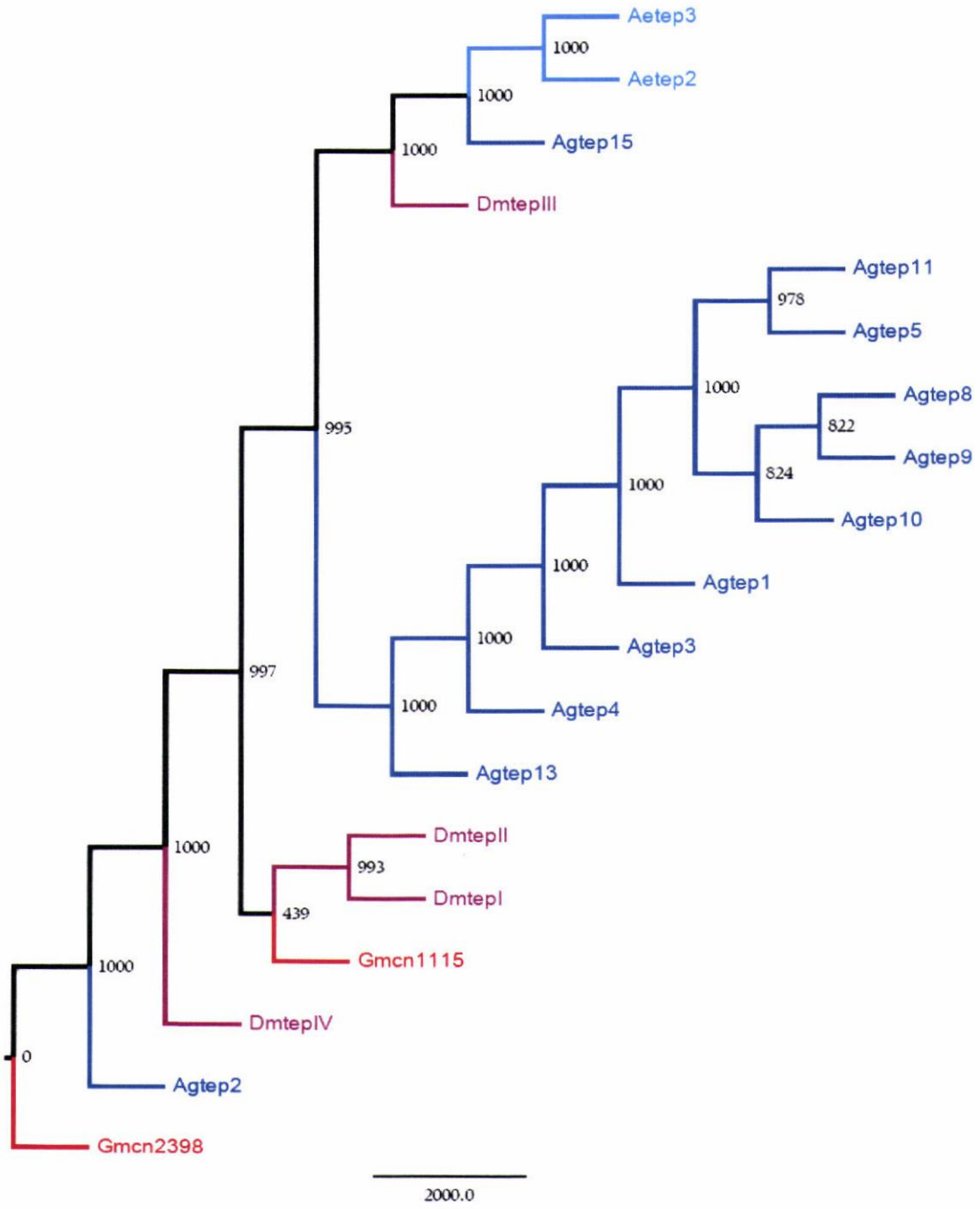


Figure 3.4 A Phylogenetic tree constructed using PHYLIP (NJ approach).

A. gambaie TEP proteins (blue), *D. melanogaster* (purple), *A. aegypti* (cyan) and *G. morsitans* homologs identified using BLASTP (red). All bootstrap values are above 80% indicating strong support for the clades and branches, with the exception of the Gmcn1115 branch, which had a low bootstrap value.

All constructed phylogenetic trees showed similar topologies, however the tree constructed using MrBayes showed slight variation; Agtep2 segregates with the *Anopheles*-specific clade. In contrast, Agtep2 in PHYLIP and PHYML forms a separate branch (see Figure 3.4 and appendix V). All trees are comprised of three clades, an *Anopheles*-specific clade (containing Agtep1, Agtep3, Agtep4, Agtep5, Agtep8, Agtep9, Agtep10 and Agtep11), another clade that contains Aetep2, Aetep3 Agtep15 and DmtepIII as well as a third clade that is comprised of DmtepI, DmtepIII and Gmcn1115. DmtepIV and Gmcn2398 segregate into separate from other TEPs, suggesting that they are more divergent. Gmcn2398 could be a novel TEP protein or a *G. morsitans*. In all trees constructed, there was no evidence of *Drosophila*-specific expansion. In the PHYLIP tree, bootstrap values of 1000 replicates are indicated for most of the branches with the exception of one branch (Figure 3.4).

The phylogenetic trees identified a single clade for Gmcn1115, DmtepI and DmtepII (Figure 3.4). Sequence similarity searches (TBLASTN) were conducted using Gmcn1115 against *Drosophila* chromosome 2L to determine whether Gmcn1115 would map onto the exon or intron region of DmtepI. Two Gmcn1115 fragments mapped to regions of DmtepI. The first fragment maps to DmtepI's exon2 and exon3, covering the intron region, the second fragment maps to exon 4 of DmtepI. The results are presented in Figure 3.5.

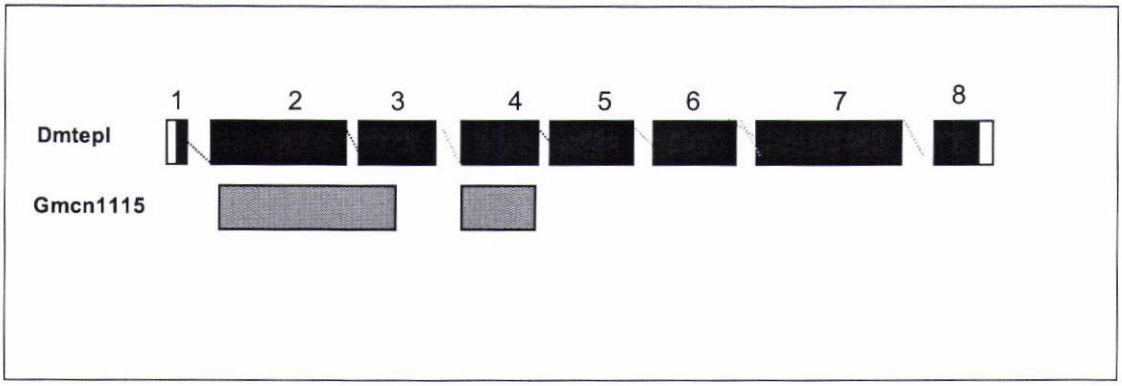


Figure 3.5. Gmcn1115 contig mapped to the exon/intron region of DmtepI.
Black bars denote DmtepI exons, dotted lines in between denote introns. Red bars indicate regions of Gmcn1115 that map to exons2, exon3 and exon4 of DmtepI.

3.1.4 Genome organization

To analyze genomic organization of TEP proteins in *A. gambiae* and *D. melanogaster*, data was obtained from the Ensembl Genome browser *. Genome organization data for *A. aegypti* and *G. morsitans* are not yet available in Ensembl. Genomic organization structures of *Anopheles* and *Drosophila* are presented in Figure 3.6A and Figure 3.6B respectively. *Anopheles gambiae* TEP proteins are located in two arms of chromosome 3 (3L and 3R), 8 TEPs are located on chromosome 3L (Agtep1, Agtep3, Agtep4, Agtep5, Agtep8, Agtep9, Agtep10 and Agtep11) proximal to the centromeric region while 3 are found in chromosome 3R (Agtep2, Agtep13 and Agtep15), proximal to the telomeric region of the right arm of chromosome 3. Agtep2 and Agtep15 are located in close proximity to each other (Figure 3.6A). In *D. melanogaster* all the TEP proteins are

* (http://www.ensembl.org/index.html).

located in Chromosome 2L (Figure 3.6B). DmtepI is located proximal to the centromeric region. DmtepII and DmtepIII are oriented head to head proximal to the centromeric region. DmtepIV is located in close proximity to the centromeric region of the left arm of chromosome 2. The proteins presented in the *Anopheles*-specific expansion are arranged in a cluster along the chromosome (Figure 3.6A). In addition, the orthologous relationship of Agtep8 and Agtep9 observed in phylogenetic trees is also observed in the genomic organization structures, as these two proteins are located in close proximity to each other.

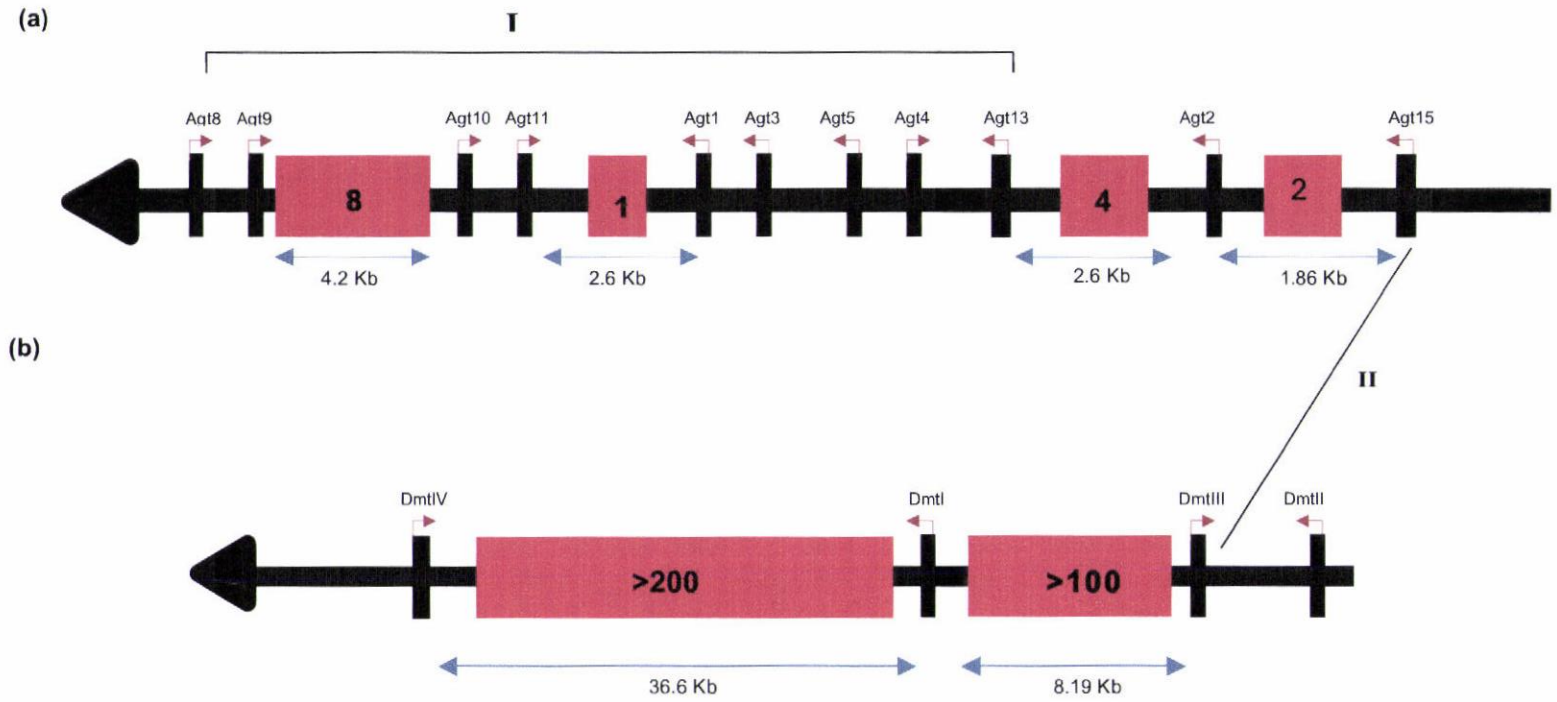


Figure 3.6. The genomic organization of TEP proteins in *Anopheles* Chromosome 3 (Figure 3.6A) and *Drosophila* Chromosome 2 (Figure 3.6B).

Black boxes denote TEP proteins and red boxes denote "non-TEP" proteins. Single headed arrows represent transcription direction while double arrows indicate the distance between TEP proteins. (a) In chromosome 3, bar-I marks a cluster of proteins belong to an *Anopheles*-specific (Figure 3.5). (b) DmtII and DmtIII are arranged in a head to head orientation along the chromosome, while DmtepI and DmtepIV are placed further away from the other two TEPs. Bar-II denotes a 1:1 orthologous relationship between Agt15 and DmtIII.

3.2 Expanding the TEP protein family search

To verify the identification of all TEP orthologs (*A. gambiae*, *D. melanogaster* and *A. aegypti*) that have been identified, a dataset comprised of protein families obtained from Ensembl Compara_db (version 49) was used. Protein families containing *A. gambiae*, *D. melanogaster* and *A. aegypti* were extracted and analyzed. Compara_db protein clusters were constructed based on sequence similarities. The objective was to identify additional TEP homologs that would not have been identified by literature surveys. Additional TEP homologs identified would be added to TEPs identified through literature surveys and phylogenetic trees would be reconstructed. Expanding the TEP protein family in *A. gambiae*, *D. melanogaster* and *A. aegypti* would provide a better understanding of orthologous relationship between the four taxa used in this analysis.

Protein families containing *Anopheles*, *Drosophila* and *Aedes* were extracted; the number of proteins extracted for each family is shown in Table 3.3. Two fasta protein files were created using Compara proteins:

1) Protein_fasta_file1: Insect Compara (*Anopheles*, *Drosophila* and *Aedes*) and all identified TEP homologs
2) Protein_fasta_file2: *Anopheles*, *Drosophila*, *Aedes,* Human and Mouse protein sequences and all identified TEP homologs

Pair-wise sequence similarity analysis was conducted using the BL2SEQ algorithm (Tatusova and Madden, 1999).

Table 3.3: TEP protein families obtained from Ensembl Compara_db.*

TEP proteins identified	Compara Family Identifier	Number of proteins in a cluster
A. gambiae		
Agtep1	Fm_9353	33
Agtep2	Fm_1114	164
Agtep3	Fm_6607	37
Agtep4	Fm_1250	151
Agtep5	Fm_66176	2
Agtep8	Fm_8504	38
Agtep9	Fm_1778	121
Agtep10	Fm_1627	152
Agtep11	Fm_563	556
Agtep13	Fm_8557	49
Agtep15	Fm_524	262
D. melanogaster		
DmtepI	Fm_299	299
DmtepII	Fm_12617	15
DmtepIII	Fm_149	20
DmtepIV	Fm_380	336
A. aegypti		
Aetep2	Fm_595	274
Aetep3	Fm_19508	3

Sequence similarity analysis of Compara insect proteins and TEP proteins yielded no significant results, the criteria of a significant match were: expectation cutoff 10^{-05}, percentage identity (greater than 50%) and coverage (greater than 60%). To determine if there were any other orthologous proteins in the protein families clusters, sequence similarity analysis was conducted using Protein_fasta_file2 created. Results are presented in Table 3.4 the expectation cutoff 10^{-02}, percentage identity (greater than 40%). Results shown in Table 3.4 indicate that many of the proteins that are clustered

* Compara Family Identifier represents family clusters that were extracted.

together in Compara_db do not share significant sequence similarity. A Significant sequence similarity is represented by percentage identity of 50% and above, as well as an expectation value of 10^{-08}. The criteria were made less stringent in order to identify all putative homologs. No additional TEPs were obtained from sequence similarity searches; therefore, this suggests that expansions within the four taxa studied have not occurred.

Table 3.4. Compara protein pairs showing significant sequence similarity.*

Family Identifier	Query	Subject	Percentage Identity (%)	Coverage (%)	E- value
Fm_595					
	ENSP00000312282	ENSP000003760271	54.98	22.30	4.00E-079
	FBpp077622	ENSMUSP00000102799	40.00	1.26	0.013
	ENSMUSP0000010079	ENSP00000360271	64.71	1.59	0.002
	ENST00000358533	ENSP00000360271	45.83	1.59	0.035
	ENSP00000358533	ENSP00000247930	41.18	1.20	0.036
Fm_1114	ENSP00000383516	ENSP00000383519	53.80	30.08	4.00E-059
	ENSP00000311307	FBpp0100635	47.06	1.0	0.022
Fm_299					
	AAEL000580	ENSMUSP00000107070	53.38	9.64	0.018
	AGAP011299	ENSMUSP00000064511	66.67	2.28	0.03
Fm_149					
	AGAP005250	ENSP0000038252	41.18	0.75	0.036
	ENSMUSP00000077433	FBpp0083694	50.00	6.90	0.023
	ENSP00000736085	FBpp0079384	40.70	5.60	0.005
	ENSMUSP00000076751	ENSMUSP000000105681	41.94	6.67	0.001
	AGAP010535	ENSP000000295709	42.42	1.07	8.00E-075
	AAEL002435	ENSP00000375992	41.38	8.48	1.00E-006
	ENSMUSP0000099141	AGAP007390	58.33	1.18	0.028
	ENSMUSP0000059989	ENSMUSP00000092806	47.62	7.80	0.013
	FBpp0079384	ENSMUSP00000333984	44.00	2.36	0.017
	FBpp0079384	ENSP00000361543	40.74	5.26	0.005
	FBpp0085272	ENSMUSP00000059989	47.06	2.33	0.043
	ENSP00000294507	ENSMUSP000000928806	58.33	5.67	0.028
	AGAP007390	ENSMUSP0000099141	58.33	10.0	0.028
Fm_380					
	ENSP00000379111	ENSMUSP00000111055	43.59	1.91	0.049
	ENSP00000351114	ENSP00000355568	40.79	2.82	0.03
	AGAP005160	ENSP00000271452	47.37	2.59	0.031
	ENSMUSP00000096002	ENSP00000221957	40.74	4.15	0.032
	ENSP00000347767	ENSP00000342104	41.67	1.69	0.017
Fm_563					
	FBpp0070148	ENSMUSP0000019779	43.75	7.38	0.026

* The criteria were a percentage identity of 40% and above, as well as an expectation value cutoff of 10^{-02}.

Chapter 4

Discussion

4.1	Thioester –containing protein dataset.....	59
	4.1.1 TEP homologs identified in *Glossina morsitans*	60
4.2	Functional domain analysis.....	62
4.3	The Phylogenetic analysis of TEP proteins.....	63
4.4	Genomic organization.....	65
4.5	Validity of the Compara database	66

CHAPTER 4

DISCUSSION

Insect TEPs and their involvement in immune response were first described in detail by Lageux *et al* (2000), where he looked at Drosophila *melanogaster TEP* proteins. Blandin and Levashina (2004) also conducted phylogenetic analysis on a thioester-containing family using phylogenetic analysis expanding to *A. gambiae*. There is a need to characterize and understand this family more extensively as it is involved in immune response against invading microbes. These proteins can be used as small molecule targets in designing strategies to increase resistance of vectors such as the tsetse fly against parasite infection.

4.1 Insect thioester-containing proteins dataset

Six TEP proteins were identified in *D. melanogaster* (Blandin and Levashina, 2004). Studies conducted by Obbard (2008) showed that there are only four true *D. melanogaster* TEP proteins, hence the exclusion of DmtepV and DmtepVI from this analysis. Work done by Laguex *et al* (2000) on constitutive expression of a complement-like protein in *D. melanogaster* showed that bacterial infection induces the up-regulation of DmtepI, DmtepII and DmtepIV in larvae. Bacterial infection also induces an immune response of DmtepII in adults of *Drosophila* flies (Laguex *et al.*, 2000).

Early studies by Blandin and Levashina (2004) as well as Christophedes *et al* (2002) indicate that there are 19 *A. gambiae* TEP proteins, which is reflected in Ensembl's database (version 49). Subsequent studies by Obbard *et al* (2008) on the evolution of Agtep1, which reviewed all other *Anopheles* TEP proteins showed that there are 15 TEP proteins of which eleven were used in this analysis. Agtep1 plays a role against parasitic infections as observed in knockdown studies of Agtep1, where parasite multiplication was observed upon infection (Obbard *et al.*, 2004).

In literature and protein database (Genbank) searches two *A. aegypti* proteins, Aetep2 and Aetep3 were identified and used in this analysis. There is no published data on *A. aegypti* TEP proteins and knowledge of their involvement in immune responses against foreign invaders is still insubstantial. However, their orthologous relationships to characterized TEP proteins have been established in this study.

4.1.1 TEP homologs identified in *Glossina morsitans*

Sequence similarity searches carried out in this study, identified five putative TEP homologs from the *Glossina morsitans* transcriptome. Gmcn1115, Gmcn2281 are homologs of DmtepII and DmtepIII respectively. Gmcn2398 and Gmcn4297 are both homologs of DmtepIV. Gmcn1115 and Gmcn1116 are isoforms and they are homologs to DmtepII and Agtep8 respectively, suggesting that the two isoforms diverged in order to perform different functions. TEPs produced as a result of alternative splicing were first observed in *Drosophila*, which has DmtepII isoforms (Blandin and Levashina, 2004). *Glossina* EST clusters were generated from ESTs sequenced from different anatomical tissues. Contig Gmcn1115 was generated from a cluster of 22 EST sequences

that are derived from fat body, salivary glands and midgut anatomical tissues. Analysis conducted by Lehane *et al* (2003) to classify immune-related proteins in the midgut identified two clones, which are homologs of DmtepIV. Similarly, Gmcn1115, generated in this thesis, contains a midgut-derived EST sequence, that is homologous to DmtepIV. Given the midgut ESTs identified in this thesis and by Lehane *et al* (2003), it is tempting to speculate that there may be an increased expression of TEP proteins in the insect midgut tissue. Gmcn1116 is represented by one EST, which is derived from the midgut tissue, while Gmcn2398, Gmcn4297 and Gmcn2281 are comprised of two, one and two ESTs respectively. These ESTs for Gmcn2281, Gmcn4297 and Gmcn2398 are derived from fat body tissue. Despite the limited number of EST clones in this study, the observation that the putative TEP homologs are derived from fat body, midgut and salivary gland tissues might be important, as these tissues are vital for the survival of tsetse flies and they are key tissues in the multiplication and maturation process of trypanomosomes (Attardo *et al.*, 2006). The fat body is important in that it caries a function similar to the liver in humans. Additionally, the fat body also releases proteins associated with fecundity and may also contain proteins responsible for refractoriness against trypanosome infection. The midgut tissue is used for blood digestion by tsetse flies. In addition, parasite numbers get reduced in the midgut through a process called attrition. Salivary glands are important in that they are used for the transmission of parasites to metacyclic forms, which will later be transmitted to the mammalian host (discussed in detail in section 1.3) (Attardo *et al.*, 2006). Therefore, identifying immune related proteins in these tissues will help elucidate host-parasite interactions.

4.2 Functional domain analysis

The domain architecture of all *Drosophila* TEPs is identical and suggests conservation of function. The thioester-containing region is conserved in all *Drosophila* TEP proteins, which is contrary to what is observed in the *Anopheles* and *Aedes* TEP functional domains analyzed, as some of the TEPs in the two organisms lack a TED domain in their sequence signatures. *Glossina morsitans* TEP homologs might be representing incomplete sequences and therefore it cannot be concluded that they do not have the TED domain as part of their sequence signature. *Drosophila* TEPs lack an A2M-receptor domain that is seen in *Anopheles*, *Drosophila* and *Aedes*. The A2M-receptor domain is responsible for binding receptors of attacking molecules, thereby facilitating endocytosis (Xiao *et al.*, 2000). As the A2M-receptor domain is lacking in *Drosophila,* endocytosis is possibly assigned to other immune-related proteins.

The thioester-region domain is conserved in two of the *Anopheles* TEP proteins (Agtep1 and Agtep4) and absent in the other nine TEP proteins analyzed. Absence of the thioester-region domain from the rest of *Anopheles* TEP proteins suggests these TEPs may have a modified function. The nine TEP proteins possibly perform a function that is similar to Alpha-2-macroglobulin, in that they still retained the bait-like region, which in this instance is possibly used to trap and clear the proteases or other invading microbes from the immune system. The bait region is found in Alpha-2-macroglobulin family N-terminal region and Alpha-2-macroglobulin family domains.

Gmcn2398 shows conservation of two domains Alpha-2-macroglobulin family N-terminal region and Alpha-2-macroglobulin family domains. The putative homolog (Gmcn2398) may have a modified function from Gmcn1115 TEP homolog, as it lacks the thioester-region domain. Gmcn2398 possibly functions by using the bait- like region to recognize and eliminate invading microbes as well.

It is most likely that Gmcn1115 has a function that is similar to Agtep1 and Agtep4 as it contains similar domain architecture as the two TEP proteins.

Results in this study suggest that gain or loss of TEP functional domain varies from species to species due to selection pressures. In addition, *Drosophila melanogaster* has fewer copies of TEP proteins than *Anopheles gambiae*. Hematophagy added to the fact that *A. gambiae* is a vector of trypanosomes may play a role in increased selection pressure, as more protein copies may be needed to fight immune challenges. The genome of *G. morsitans* is not yet fully sequenced, therefore, there may be more copies of TEP proteins, which have not yet been characterized.

4.3 Phylogenetic analysis of TEP proteins

Trees generated using Neighbor joining, Maximum likelihood and Bayesian (Figure 3.4, appendix V and appendix VI) show similar topologies, with the exception of one branch (Agtep2) that is placed into an *Anopheles*-specific clade in a tree drawn with MrBayes (appendix VI). Of the 19 TEP homologs from *A. gambiae*, *D. melanogaster*, *A. aegypti* and *G. morsitans,* one species-specific expansion was observed, which is an *Anopheles*-specific cluster. In previous studies conducted Agtep15 and Agtep2 cluster together (Blandin and Levashina, 2004). However, in this study these two TEP proteins

were placed in different clusters; Agtep15 clusters with Aetep2, Aetep3 and DmtepIII, whereas Agtep2 either forms a separate branch or is placed into the *Anopheles*-specific clade, suggesting that these two TEP proteins are divergent. There is no *Drosophila*-specific expansion observed in the constructed trees, instead DmtepI, DmtepII and Gmcn1115 are placed in one cluster. The topology of the phylogenetic trees suggests that Gmcn1115 is closely related to DmtepIV as the latter is placed in a single branch close to Gmcn1115 (Figure 3.4).

In phylogenetic analyses Gmcn2398 contig was placed in a separate branch and was therefore used as an outgroup. The phylogenetic tree constructed using the Neighbor joining method shows strong bootstrap support, except for clade containing Gmcn1115, DmtepI and DmtepII, which had a low bootstrap value (Figure 3.4). As Gmcn1115 shows similarity to DmtepI in phylogenetic trees constructed (Figure 3.4). TBLASTN was performed via the Ensembl blast server version 49. The aim was to determine whether Gmcn1115 would cover any exon/intron boundaries thereby showing that Gmcn1115 is a product of alternative splicing. Two fragments of Gmcn115 map to DmtepI, the first fragment maps to exon1, exon2 covering the exon/intron boundaries. The second short fragment maps to exon3. The results therefore confirm that Gmcn1115 is a putative product of alternative splicing.

4.4 Genome organization of TEP proteins

The phylogenetic analysis shows that there are species-specific expansions within *A. gambiae* TEP protein. Genome organization structure analysis was conducted to determine how the TEP proteins were organized in two of the insect species (*A. gambiae* and *D. melanogaster*) chromosomes. The fact that all *A. gambiae* and *D. melanogaster* TEP proteins are located on the chromosome 3 and chromosome 2 (respectively) suggests that they are paralogs. Phylogenetic trees constructed (Figure 3.4, appendix IV and appendix V) show a species-specific expansion for 9 of 11 *A. gambiae* proteins. These genes are located in a cluster (tandem array) along chromosome 3, suggesting that they were produced as a result of tandem duplication. Similar chromosomal arrangement of paralogs has been observed for other gene families such as homeobox genes (Popovici *et al.*, 2001). Agtep5 and Agtep11 appear to have arisen from a more recent duplication event they are also located next to each other, so is Agtep8, Agetp9 and Agtep10.

There are clusters shown phylogenetic trees and were observed in genome organization structures as well such as Agtep1 and Agtep3 as well Agtep4 and Agtep13 that are placed in branches that are close to each other (Figure 3.4 and Figure 3.6A). Additionally, another cluster that is observed in the phylogenetic tree and the genome organization structure is Agtep8 and Agtep9 (Figure 3.6A). The Genome organization structure of *A. gambiae* also shows that there is a cluster of 'non-TEP' proteins located in between the TEP proteins and they are all novel proteins (Figure 3.6A).

In *D. melanogaster* only DmtepII and DmtepIII appear to form a cluster. DmtepI is placed approximately 100 proteins (8.19 Kb) away from DmtepIII. DmtepIV is placed approximately 200 proteins (36.6 Kb) from DmtepI. The dispersal of these paralogs may

be linked to the fact that they have conserved the thioester-containing region, which is key to the function of TEP proteins. In contrast the *A. gambiae* TEP proteins that are organized in clusters only have two proteins that have conserved the thioester-containing region.

4.5 Validity of the Compara database

To identify more TEP protein homologs, insect proteins (*A. gambiae*, *D. melanogaster* and *A. aegypti*) were extracted from Compara_db and sequence similarity searches were performed using an all-against-all approach for each protein clusters.

Sequence similarity searches conducted yielded no significant results suggesting that none of the insect proteins put in the clusters are homologs. To ascertain whether the families downloaded from Compara_db that were clustered with the TRIBE-MCL algorithm (Enright *et al.*, 2002) shared homology, human and mouse proteins were added and sequence similarity searches were conducted. Eighteen proteins (5% of 426) yielded significant results indicating shared homology. Protein family clusters in the Compara database based on sequence homology, therefore, it is assumed that the proteins clustered together share evolutionary history. In test datasets used by Enright *et al* (2002) to asses the accuracy of TRIBE-MCL 87% of the families clustered shared homology. However, the algorithm is subject to the following drawbacks (Enright *et al.*, 2002):

1) Proteins can share one or two domain without sharing significant sequence homology causing the algorithm to cluster unrelated proteins

2) Some protein clusters may contain unrelated family members due to repeated sequence patterns produced a multiple times as opposed to having promiscuous domains still causing erroneous clustering.

In results obtained in this analysis, none of the selected insect protein families shared sequence similarity. The results of the second dataset analysis shows that 6 out of 17 families had at least one protein pair that shared significant sequence similarity, suggesting that the majority of proteins extracted (Table 3.3 and Table 3.4) do not share sequence similarity. It is possible that the criteria set to extract significant matches in this analysis was different from that used in TRIBE-MCL when they were selecting matches that would be stored in the square matrix for clustering into protein families. It is also possible that the chosen dataset is subject to the first drawback mentioned above; the protein clusters shared a few domains without having significant similarity overall.

Chapter 5
CONCLUSIONS

The availability of fully sequenced genomes of *Drosophila melanogaster* and *Anopheles gambiae* provide key resources for studying genes from organisms whose genome have not been fully sequenced (Holt *et al.*, 2002 and Adams *et al.*, 2000). In addition, the draft of *A. aegypti* is available also adding to available genome resources (Nene *et al.*, 2007). The *Drosophila* insect has been used extensively as a model organism in studies such as comparative genomics as it well annotated. *Anopheles* and *Aedes* represent valuable resources for vector comparative genomics, as they are vectors and hematophagous insects.

In this study, a family of immune-related proteins has been characterized using comparative genomics. Protein sequences from *A. gambiae*, *D. melanogaster* and *A. aegypti* were used in sequence similarity searches against *G. morsitans* EST transcriptome data. Studying immune-related proteins will help elucidate interactions at molecular level between tsetse flies and *Trypanosoma* spp, thus adding knowledge to host and parasite control strategies.

Five *G. morsitans* TEP homologs (Gmcn2281, Gmcn1115, Gmcn1116, Gmcn2398 and Gmcn4297) were identified through sequence similarity searches. Of these, only three (Gmcn2398, Gmcn1115 and Gmcn1116) were shown to be more similar to *Drosophila* and *Anopheles* TEP proteins.

Results of phylogenetic and functional domain analyses of the identified TEP homologues indicate an *Anopheles* expansion.

Gmcn1115 appears to be closely related to DmtepI and DmtepII, while Gmcn2398 segregates to a separate branch from the rest of the TEP homologs, suggesting that Gmcn2398 *Glossina*-specific TEP protein. No *Glossina*-specific expansions were observed. Gmcn2398 and Gmcn4297 appear to be more closely related to α-macroglobulins as opposed to invertebrate TEP proteins another sub-family of thioester-containing superfamily.

Efforts to identify additional TEP homologs in *A. gambiae*, *D. melanogaster* and *A. aegypti* were unsuccessful. The insect protein family clusters downloaded from the Compara-db do not seem to have significant sequence similarity.

The genome organization structures of *A. gambiae* TEP proteins indicate the presence of novel 'non-TEP' proteins located between TEPs. It would be interesting to characterize these proteins and determine whether they have diverged from the same ancestral protein as TEPs. Sequencing of the *G. morsitans* genome is underway and will allow the identification and characterization of more immune-related proteins, possibly including putative TEP homologs, thereby increasing the possibility of successful vector and parasite control strategies.

REFERENCES

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., et al. (2000). The genome sequence of Drosophila melanogaster. Science, 287(5461), 2185-2195.

Aggarwal, K. & Silverman, N. (2008). Positive and negative regulation of the Drosophila immune response. BMB Rep, 41(4), 267-277.

Aksoy, S. (2003). Control of tsetse flies and trypanosomes using molecular genetics. Vet Parasitol, 115(2), 125-145.

Aksoy, S., Berriman, M., Hall, N., Hattori, M., Hide, W., & Lehane, M. J. (2005). A case for a Glossina genome project. Trends Parasitol, 21(3), 107-111.

Aksoy, S. & Rio, R. V. (2005). Interactions among multiple genomes: tsetse, its symbionts and trypanosomes. Insect Biochem Mol Biol, 35(7), 691-698.

Allsopp, R. (2001). Options for vector control against trypanosomiasis in Africa. Trends Parasitol, 17(1), 15-19.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. J Mol Biol, 215(3), 403-410.

Armstrong, P. B. & Quigley, J. P. (1999). Alpha2-macroglobulin: an evolutionarily conserved arm of the innate immune system. Dev Comp Immunol, 23(4-5), 375-390.

Attardo, G. M., Guz, N., Strickler-Dinglasan, P., & Aksoy, S. (2006). Molecular aspects of viviparous reproductive biology of the tsetse fly (Glossina morsitans morsitans): regulation of yolk and milk gland protein synthesis. J Insect Physiol, 52(11-12), 1128-1136.

Baxter, R. H., Chang, C. I., Chelliah, Y., Blandin, S., Levashina, E. A., & Deisenhofer, J. (2007). Structural basis for conserved complement factor-like function in the antimalarial protein TEP1. Proc Natl Acad Sci U S A, 104(28), 11615-11620.

Beck, G. & Habicht, G. S. (1996). Immunity and the invertebrates. Sci Am, 275(5), 60-3, 66.

Blandin, S. & Levashina, E. A. (2004). Thioester-containing proteins and insect immunity. Mol Immunol, 40(12), 903-908.

Blandin, S., Shiao, S. H., Moita, L. F., Janse, C. J., Waters, A. P., Kafatos, F. C., et al. (2004). Complement-like protein TEP1 is a determinant of vectorial capacity in the malaria vector Anopheles gambiae. Cell, 116(5), 661-670.

Blandin, S. A. & Levashina, E. A. (2007). Phagocytosis in mosquito immune responses. Immunol Rev, 219, 8-16.

Bulet, P., Stocklin, R., & Menin, L. (2004). Anti-microbial peptides: from invertebrates to vertebrates. Immunol Rev, 198, 169-184.

Carlton, J. M., Muller, R., Yowell, C. A., Fluegge, M. R., Sturrock, K. A., Pritt, J. R., et al. (2001). Profiling the malaria genome: a gene survey of three species of malaria parasite with comparison to other apicomplexan species. Mol Biochem Parasitol, 118(2), 201-210.

Center for Disease Control (CDC). (2008) http://www.dpd.cdc.gov/dpdc HTML/TrypanosomiasisAfrican.htm)

Conserved Domains Database (2008) www.ncbi.nlm.nih.gov/Structure/cdd/cdd.html.

Chappuis, F., Udayraj, N., Stietenroth, K., Meussen, A., & Bovier, P. A. (2005). Eflornithine is safer than melarsoprol for the treatment of second-stage Trypanosoma brucei gambiense human African trypanosomiasis. Clin Infect Dis, 41(5), 748-751.

Chor, B., Hendy, M. D., Holland, B. R., & Penny, D. (2000). Multiple maxima of likelihood in phylogenetic trees: an analytic approach. Mol Biol Evol, 17(10), 1529-1541.

Christoffels, A., van Gelder, A., Greyling, G., Miller, R., Hide, T., & Hide, W. (2001). STACK: Sequence Tag Alignment and Consensus Knowledgebase. Nucleic Acids Res, 29(1), 234-238.

Christophides, G. K., Vlachou, D., & Kafatos, F. C. (2004). Comparative and functional genomics of the innate immune system in the malaria vector Anopheles gambiae. Immunol Rev, 198, 127-148.

Clamp, M., Cuff, J., Searle, S. M., & Barton, G. J. (2004). The Jalview Java alignment editor. Bioinformatics, 20(3), 426-427.

Dayhoff, M. O., Barker, W. C., & McLaughlin, P. J. (1974). Inferences from protein and nucleic acid sequences: early molecular evolution, divergence of kingdoms and rates of change. Orig Life, 5(3), 311-330.

de Koning, H. P. (2008). Ever-increasing complexities of diamidine and arsenical crossresistance in African trypanosomes. Trends Parasitol, 24(8), 345-349.

Dimopoulos, G., Casavant, T. L., Chang, S., Scheetz, T., Roberts, C., Donohue, M., et al. (2000). Anopheles gambiae pilot gene discovery project: identification of mosquito innate immunity genes from expressed sequence tags generated from immune-competent cell lines. Proc Natl Acad Sci U S A, 97(12), 6619-6624.

Dodds, A. W. & Law, S. K. (1998). The phylogeny and evolution of the thioester bond-containing proteins C3, C4 and alpha 2-macroglobulin. Immunol Rev, 166, 15-26.

Dottorini, T., Nicolaides, L., Ranson, H., Rogers, D. W., Crisanti, A., & Catteruccia, F. (2007). A genome-wide analysis in Anopheles gambiae mosquitoes reveals 46 male accessory gland genes, possible modulators of female behavior. Proc Natl Acad Sci U S A, 104(41), 16215-16220.

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res, 30(7), 1575-1584.

Enserink, M. (2008). Epidemiology. Lower malaria numbers reflect better estimates and a glimmer of hope. Science, 321(5896), 1620.

Ensembl database (2008) http://www.ensembl.org

Fairlamb, A. H. (2003). Chemotherapy of human African trypanosomiasis: current and future prospects. Trends Parasitol, 19(11), 488-494.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol, 17(6), 368-376.

Felsenstein, J. (1989) PHYLIP - Phylogeny Inference Package (Version 3.2).Cladistics. 5: 164-166.

FigTree software: (2006) http://tree.bio.ed.ac.uk/software/figtree/.

Fong, J. H. & Marchler-Bauer, A. (2008). Protein subfamily assignment using the Conserved Domain Database. BMC Res Notes, 1, 114.

Gooding, R. H. & Krafsur, E. S. (2005). Tsetse genetics: contributions to biology, systematics, and control of tsetse flies. Annu Rev Entomol, 50, 101-123.

Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol, 52(5), 696-704.

Hao, Z., Kasumba, I., & Aksoy, S. (2003). Proventriculus (cardia) plays a crucial role in immunity in tsetse fly (Diptera: Glossinidiae). Insect Biochem Mol Biol, 33(11), 1155-1164.

Hardison, R. C. (2003). Comparative genomics. PLoS Biol, 1(2), E58.

Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., et al. (2002). The genome sequence of the malaria mosquito Anopheles gambiae. Science, 298(5591), 129-149.

Hu, C. & Aksoy, S. (2006). Innate immune responses regulate trypanosome parasite infection of the tsetse fly Glossina morsitans morsitans. Mol Microbiol, 60(5), 1194-1204.

Huang, W., Dolmer, K., Liao, X., & Gettins, P. G. (2000). NMR solution structure of the receptor binding domain of human alpha(2)-macroglobulin. J Biol Chem, 275(2), 1089-1094.

Huelsenbeck, J. P. & Bollback, J. P. (2001). Empirical and hierarchical Bayesian estimation of ancestral states. Syst Biol, 50(3), 351-366.

Huelsenbeck, J. P. & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics, 17(8), 754-755.

Iwanaga, S. & Lee, B. L. (2005). Recent advances in the innate immunity of invertebrate animals. J Biochem Mol Biol, 38(2), 128-150.

Jiang, Z., Wu, X. L., Michal, J. J., & McNamara, J. P. (2005). Pattern profiling and mapping of the fat body transcriptome in Drosophila melanogaster. Obes Res, 13(11), 1898-1904.

Jiggins, F. M. & Kim, K. W. (2006). Contrasting evolutionary patterns in Drosophila immune receptors. J Mol Evol, 63(6), 769-780.

Kennedy, P. G. (2004). Human African trypanosomiasis of the CNS: current issues and challenges. J Clin Invest, 113(4), 496-504.

Kennedy, P. G. (2006). Diagnostic and neuropathogenesis issues in human African trypanosomiasis. Int J Parasitol, 36(5), 505-512.

Kishino, H. & Hasegawa, M. (1990). Converting distance to time: application to human evolution. Methods Enzymol, 183, 550-570.

Kopacek, P., Weise, C., Saravanan, T., Vitova, K., & Grubhoffer, L. (2000). Characterization of an alpha-macroglobulin-like glycoprotein isolated from the plasma of the soft tick Ornithodoros moubata. Eur J Biochem, 267(2), 465-475.

Krumsiek, J., Arnold, R., & Rattei, T. (2007). Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics, 23(8), 1026-1028.

Kurata, S. (2004). Recognition of infectious non-self and activation of immune responses by peptidoglycan recognition protein (PGRP)-family members in Drosophila. Dev Comp Immunol, 28(2), 89-95.

Kurata, S. (2006). Recognition and elimination of diversified pathogens in insect defense systems. Mol Divers, 10(4), 599-605.

Kurata, S., Ariki, S., & Kawabata, S. (2006). Recognition of pathogens and activation of immune responses in Drosophila and horseshoe crab innate immunity. Immunobiology, 211(4), 237-249.

Lagueux, M., Perrodou, E., Levashina, E. A., Capovilla, M., & Hoffmann, J. A. (2000). Constitutive expression of a complement-like protein in toll and JAK gain-of-function mutants of Drosophila. Proc Natl Acad Sci U S A, 97(21), 11427-11432.

Lehane, M. & Billingsley, P. (1996). The Biology of the Insect Midgut. Springer.

Lehane, M. J. (2005). The Biology of Blood-Sucking in Insects. Cambridge University Press.

Lehane, M. J., Aksoy, S., Gibson, W., Kerhornou, A., Berriman, M., Hamilton, J., et al. (2003). Adult midgut expressed sequence tags from the tsetse fly Glossina morsitans morsitans and expression analysis of putative immune response genes. Genome Biol, 4(10), R63.

Lehane, M. J., Aksoy, S., & Levashina, E. (2004). Immune responses and parasite transmission in blood-feeding insects. Trends Parasitol, 20(9), 433-439.

Leulier, F., Rodriguez, A., Khush, R. S., Abrams, J. M., & Lemaitre, B. (2000). The Drosophila caspase Dredd is required to resist gram-negative bacterial infection. EMBO Rep, 1(4), 353-358.

Lemaitre, B. (2004). The road to Toll. Nature Reviews immunology, 1(4), 521-527.

Levashina, E. A., Moita, L. F., Blandin, S., Vriend, G., Lagueux, M., & Kafatos, F. C. (2001). Conserved role of a complement-like protein in phagocytosis revealed by dsRNA knockout in cultured cells of the mosquito, Anopheles gambiae. Cell, 104(5), 709-718.

Little, T. J. & Cobbe, N. (2005). The evolution of immune-related genes from disease carrying mosquitoes: diversity in a peptidoglycan- and a thioester-recognizing protein. Insect Mol Biol, 14(6), 599-605.

Loker, E. S., Adema, C. M., Zhang, S. M., & Kepler, T. B. (2004). Invertebrate immune systems--not homogeneous, not simple, not well understood. Immunol Rev, 198, 10-24.

Mathew, A. & Rothman, A. L. (2008). Understanding the contribution of cellular immunity to dengue disease pathogenesis. Immunol Rev, 225, 300-313.

Matthews, K. R. (2005). The developmental cell biology of Trypanosoma brucei. J Cell Sci, 118(Pt 2), 283-290.

Moore, A., Richer, M., Enrile, M., Losio, E., Roberts, J., & Levy, D. (1999). Resurgence of sleeping sickness in Tambura County, Sudan. Am J Trop Med Hyg, 61(2), 315-318.

MrBayes package: http://mrbayes.csit.fsu.edu/download.php.

Nair, S. V., Ramsden, A., Raftos, D. A. (2005). Ancient origins: complement in invertebrates. Invertebrate Survival Journal, 1(2), 91-104.

Nene, V., Wortman, J. R., Lawson, D., Haas, B., Kodira, C., Tu, Z. J., et al. (2007). Genome sequence of Aedes aegypti, a major arbovirus vector. Science, 316(5832), 1718-1723.

Nei, M & Kumar, S. (2000). Molecular Evolution and Phylogenetics. Oxfords University Press. New York.

Nonaka, M. & Yoshizaki, F. (2004). Evolution of the complement system. Mol Immunol, 40(12), 897-902.

Obbard, D. J., Callister, D. M., Jiggins, F. M., Soares, D. C., Yan, G., & Little, T. J. (2008). The evolution of TEP1, an exceptionally polymorphic immunity gene in Anopheles gambiae. BMC Evol Biol, 8, 274.

Oduol, F., Xu, J., Niare, O., Natarajan, R., & Vernick, K. D. (2000). Genes identified by an expression screen of the vector mosquito Anopheles gambiae display differential molecular immune response to malaria parasites and bacteria. Proc Natl Acad Sci U S A, 97(21), 11397-11402.

Osta, M. A., Christophides, G. K., Vlachou, D., & Kafatos, F. C. (2004). Innate immunity in the malaria vector Anopheles gambiae: comparative and functional genomics. J Exp Biol, 207(Pt 15), 2551-2563.

PHYLIP suite: http://evolution.genetics.washington.edu/getme.html.

PHYML suite: http://atgc.lirmm.fr/phyml.

Pollock, J.N., (1992). Tsetse and trypanosomiasis information quarterly. Food and Agriculture of the United Nations. 1.

Popovici, C., Leveugle, M., Birnbaum, D., & Coulier F. (2001). Homeobox gene clusters and the human paralogy map. FEBS letter, 491, 237-242.

Roditi, I. & Liniger, M. (2002). Dressed for success: the surface coats of insect-borne protozoan parasites. Trends Microbiol, 10(3), 128-134.

Royet, J. (2004). Infectious non-self recognition in invertebrates: lessons from Drosophila and other insect models. Mol Immunol, 41(11), 1063-1075.

Saravanan, T., Weise, C., Sojka, D., & Kopacek, P. (2003). Molecular cloning, structure and bait region splice variants of alpha2-macroglobulin from the soft tick Ornithodoros moubata. Insect Biochem Mol Biol, 33(8), 841-851.

Silverman, N. & Maniatis, T. (2001). NF-kappaB signaling pathways in mammalian and insect innate immunity. Genes Dev, 15(18), 2321-2342.

Sivashankari, S. & Shanmughavel, P. (2007). Comparative genomics - a perspective. Bioinformation, 1(9), 376-378.

Smith, L. C., Azumi, K., & Nonaka, M. (1999). Complement systems in invertebrates. The ancient alternative and lectin pathways. Immunopharmacology, 42(1-3), 107-120.

Smith, D. H., Pepin, J., & Stich, A. H. (1998). Human African trypanosomiasis: an emerging public health crisis. Br Med Bull, 54(2), 341-355.

Steverding, D. (2008). The history of African trypanosomiasis. Parasit Vectors, 1(1), 3.

Taylor, J. E. & Rudenko, G. (2006). Switching trypanosome coats: what's in the wardrobe? Trends Genet, 22(11), 614-620.

Tatusova, T. A & Madden, T. L. (1999). Blast 2 sequences - a new tool for comparing protein and nucleotide sequences. FEMS Microbiology Letters, 174, (2), 247-250.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res, 22(22), 4673-4680.

Vickerman, K., Tetley, L., Hendry, K. A., & Turner, C. M. (1988). Biology of African trypanosomes in the tsetse fly. Biol Cell, 64(2), 109-119.

Wang, J., Hu, C., Wu, Y., Stuart, A., Amemiya, C., Berriman, M., et al. (2008). Characterization of the antimicrobial peptide attacin loci from Glossina morsitans. Insect Mol Biol, 17(3), 293-302.

Wang, X., Fuchs, J. F., Infanger, L. C., Rocheleau, T. A., Hillyer, J. F., Chen, C. C., et al. (2005). Mosquito innate immunity: involvement of beta 1,3-glucan recognition protein in melanotic encapsulation immune responses in Armigeres subalbatus. Mol Biochem Parasitol, 139(1), 65-73.

Waterhouse, R. M., Wyder, S., & Zdobnov, E. M. (2008). The Aedes aegypti genome: a comparative perspective. Insect Mol Biol, 17(1), 1-8.

World Health Organization fact sheet. (2006). Human African trypanosomiasis. http://www.who.int/mediacentre/factsheets/fs259/en/

Xiao, T., DeCamp, D. L., & Spran, S. R. (2000). Structure of a rat alpha 1-macroglobulin receptor-binding domain dimer. Protein Sci, 9(10), 1889-1897.

Zhang, H., Song, L., Li, C., Zhao, J., Wang, H., Gao, Q., et al. (2007). Molecular cloning and characterization of a thioester-containing protein from Zhikong scallop Chlamys farreri. Mol Immunol, 44(14), 3492-3500.

Appendix I

BLAST takes a query file and database of sequences, both containing fasta sequences. The algorithm follows 11 steps:

(a) Filters for regions of low-complexity or regions with sequence repeats (they are masked with X or N), because they will obscure the scoring system
(b) A word list is generated between two sequences aligned
(c) BLAST uses a substitution matrix to find high scoring alignments (HSPs)
(d) A tree is generated, which is used to compare HSPs to the database sequences
(e) The algorithm then iterates steps (i-iv)
(f) The tool subsequently looks for matches for the remainder of the HSPs, which would possibly be used as a seed for ungapped alignments
(g) Then an extension step is conducted for the HSP, with the aim of increasing the alignments
(h) All high scoring HSPs are listed
(i) Gumbel's extreme value distribution (EVD) is then used to evaluate the HSPs to extract those that statistically significant
(j) A check for the HSPs is conducted whereby alignments are evaluated to determine if any could be merged
(k) Results are produced

Appendix II

Script: mrb.txt

The script is used to run MrBayes phylogenetic programme. The script specifies the commands required for the program to run.

```
begin mrbayes;
 set autoclose = yes nowarn = yes;
 execute Tep_hmlgs1_ed.nex;
 Lset nst = 6 rates = gamma;
 mcmc nruns = 1 ngen = 10000
  samplefreq = 10 file = Tep_hmlgs1_ed.nex1;
end;
```

Appendix III

BioPerl script1: BlastMatrix2.pl

Bioperl module script written by Alan Christoffels

```
#! /usr/bin/perl -w
use strict;
#The script was used to run an automated Bl2seq alignment
#Point to perl packages

use lib '/cip0/research/feziwe/bioperl/bioperl-run';
use lib '/cip0/research/feziwe/bioperl/bioperl-live';
use Bio::SeqIO;
use Bio::Seq;

##Point program to working directory, open and read fasta files using the Seq object,
then split to separate files for input to the Bl2Seq programme and print into new files
with an extension .fa

my $seqfile =
"/cip0/research/feziwe/New_tep_analysis/Bl2_seq/Aedes_tep_homologs/Aetep3/Aetep3
.fa";
my $workdir =
"/cip0/research/feziwe/New_tep_analysis/Bl2_seq/Aedes_tep_homologs/Aetep3";
my $n = 0;
my $in = Bio::SeqIO->new(-file=>$seqfile, -format=>"Fasta");
while (my $s = $in->next_seq()) {
        $n++;
        my $out = Bio::SeqIO->new(-file=>">$workdir/$n.fa", -format=>"Fasta");
        $out->write_seq($s);
}
## Block runs Bl2Seq (i.e seq1 vs seq2 vs seq3, an all against all alignment in the
working directory and then prints and output of each alignment)
foreach my $i(1..$n) {
        foreach my $j(1..$n) {
                if ($i==$j) {
                        #print "... "
                } else {
                        my $out = "$workdir/$i"."_"."$j.out";
                        system("bl2seq -p blastp -i $workdir/$i.fa -j $workdir/$j.fa -o
$out");
                        #parse the bl2seq output and print the identities.
                }
        }

   #print "\n";
   #open ( OT,">>Agtep1_blseq.out1");
```

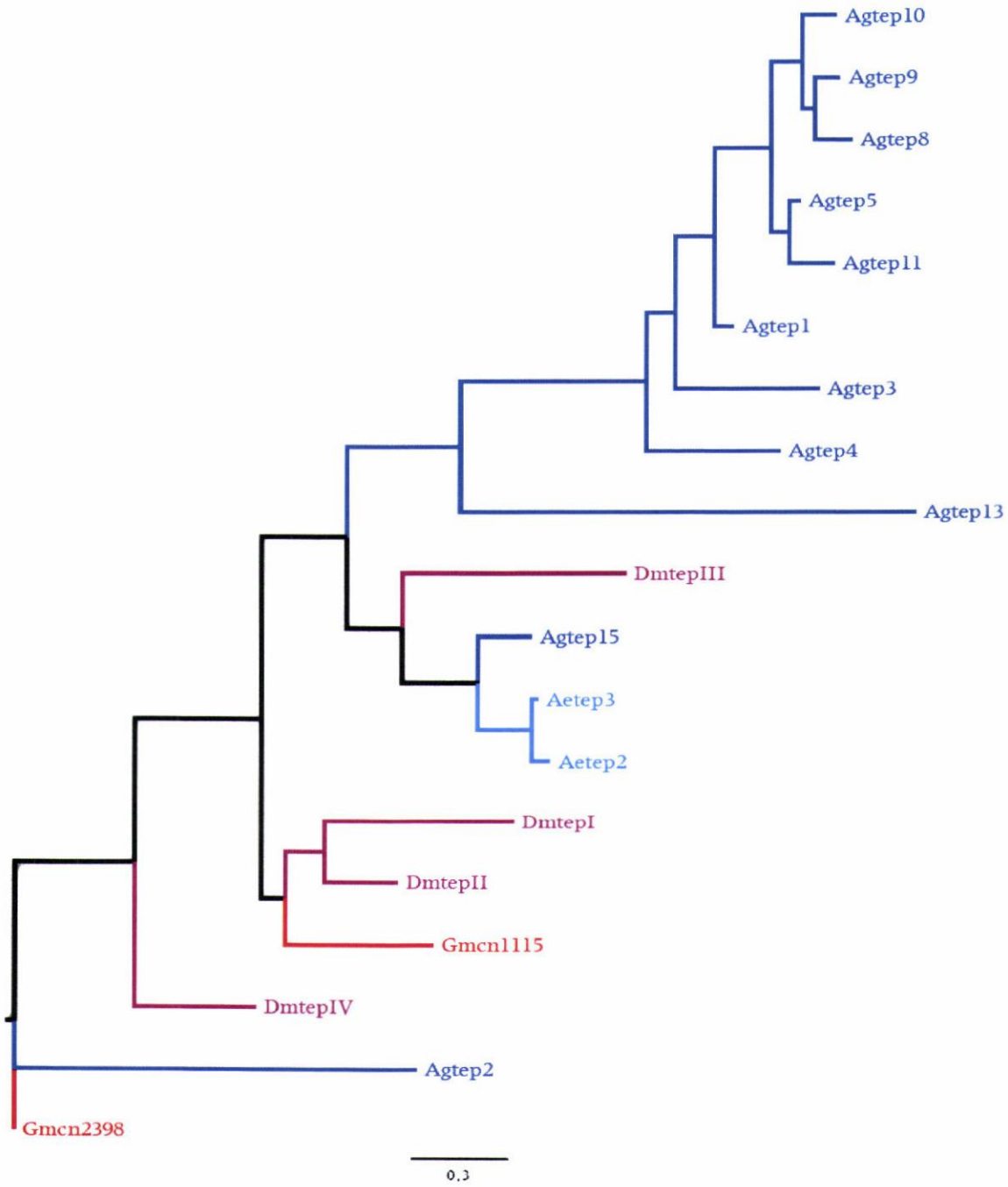
```
#print "Agtep1_blseq.out1\n";
#close (OT);
```

Appendix IV

```
BioPerl script2: BlastMatrix2.pl
# Bioperl module script written by Alan Christoffels
# /usr/bin/perl -w

### The script was used to parse BL2Seq pair-wise alignment output files and print out
percentage identity, E-value as well as the Coverage
##Point to directory containing bio-perl packages and use search module

use lib '/cip0/research/feziwe/bioperl/bioperl-run';
use lib '/cip0/research/feziwe/bioperl/bioperl-live';
use strict;
use Bio::SearchIO;
use Bio::SearchIO::blast;

### Takes path of the directory as an argument and saves reads in files within that
directory and saves them into an array. The array is used to execute the sub-routine
which parses out for features from the blast report using bioperl objects

open (OUTFILE,"> Agtep1_pars.out");

my $dir = shift @ARGV;

opendir(D, $dir);
my @files = readdir(D);
foreach my $file (@files) {
        next unless (-f "$dir/$file");
        next unless ($file =~/out/);
        my $outstring = &parser("$dir/$file");
        print "$outstring\n";
}

sub parser {
        my $f = shift;

        my $searchio = new Bio::SearchIO(-format=>'blast', -file =>$f);
        while (my $result = $searchio->next_result()) {
                if ($result->hits) {
                        my $qname  = $result->query_name;
                my $qlen  = $result->query_length;
                  #print OUTFILE "qname=$qname $qlen\n";
                while (my $hit = $result->next_hit) {
                my $hit_name = $hit->name;
```

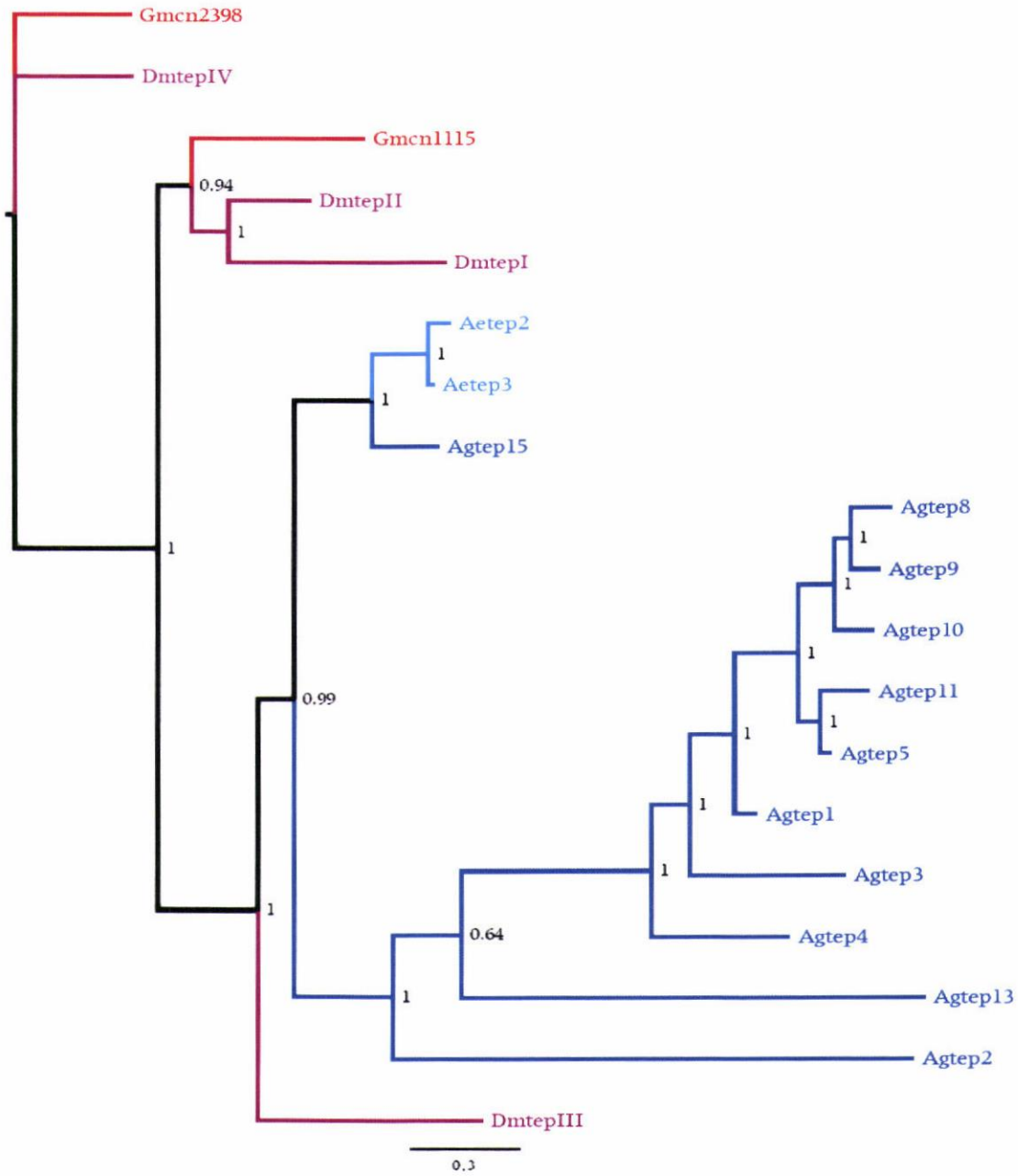
```
                my $desc = $hit->description;
                my $hlen = $hit->length;
                my $aln_len_subject = $hit->length_aln('sbjct');
                my $aln_len_query = $hit->length_aln('query');
                                #print OUTFILE "$hlen $qlen\t";

                FF:
                while (my $hsp = $hit->next_hsp) {
                        my $qstrand = $hsp->strand('Query');
                        my $hstrand = $hsp->strand('Hit');
                        my $len = $hsp->length('total');

        my $len = $hsp->length('query');
        my $len = $hsp->length('hit');
                        my $cons = $hsp->num_conserved;
                        my $eval = $hsp->evalue;
                                        my ($h_start,$h_end) = $hsp->range('hit');
                                        my ($q_start,$q_end) = $hsp->range('query');

        my $hitcov = $cons/$hlen*100;
        my $var =$hitcov;
        my $var1= $hsp->percent_identity;
        my $var1 = sprintf "%.2f", $var1;
        my $var = sprintf "%.2f", $var;
                                        my $string = $qname." ".$hit->name." ".$cons."
".$hlen." ".$qlen." ".$var1." ".$var." ".$eval;
                                        return($string);


                                        #print OUTFILE "$qname\t";
        #print OUTFILE $hit->name, "\t";
        #print OUTFILE "$cons\t";
        #3print OUTFILE "$hlen $qlen\t";
        #print OUTFILE "$var1\t";
        #print OUTFILE "$var\t";
        #print OUTFILE "$eval\n";
        #print "$desc\t";
        #print $hit->name, "\t";
        #print $hsp->percent_identity,"\t";
        #print "$eval\t";
        #print "$cons\t";
        #print $hsp->hsp_length,"\t";
        #print $aln_len_subject,"\t";
        #print $aln_len_query,"\n";
                                        #print "identity = $cons\n";
                                }
                        }
                }
```

Appendix V



A phylogenetic tree constructed using PHYML (maximum likelihood) suite

Appendix VI



A phylogenetic tree constructed using MrBayes (Bayesian inference) package