An evolutionary genomics approach towards analysis of genes implicated in transmission of trypanosomes between tsetse fly and mammalian host.

By

Sarah Wambui Mwangi

A thesis presented in fulfillment of the requirements of *Magister Scientae* at the South African National Bioinformatics Institute, University of the Western Cape.

September 2009

Supervisor: Prof. Alan Christoffels

KEYWORDS

Contig

Drosophila melanogaster

Evolution

Glossina morsitans

Orthologue

Polymorphism

Serpin

Singleton

Trypanosome

ABSTRACT

Human African trypanosomiasis is the world’s third most important parasitic disease affecting human health after malaria and schistosomiaisis. The world health organization estimates approximately 60 million people at risk in sub-Saharan Africa and up to 50,000 deaths per year caused by trypanosomiasis. Current management of human African trypanosomiasis relies on active surveillance and chemotherapy of infected patients. Efforts to develop a vaccine to immunize the human host have been hampered by antigenic variation of the parasites cell coat. The advent of the genome era has opened up opportunities for developing novel strategies for interrupting the transmission cycle of trypanosomes, specifically using any of the three players, the human host, the tsetse fly vector and/or the parasite. The human genome has been deciphered and the genomes of several trypanosome species have been sequenced. Sequencing of additional neglected trypanosome species is in progress. The tsetse fly genome is currently being sequenced as part of the genomic activities of the International *Glossina* genome initiative (IGGI). In an attempt to support the tsetse fly sequencing effort, expressed sequence tags (ESTs) from various tissues and developmental stages of *Glossina morsitans* have been generated.

In this study, tsetse fly EST data was analyzed using bioinformatics approaches, focusing on transcripts encoding serpin genes implicated in the immune defenses of tsetse flies. *Glossina morsitans* homologues to *Drosophila melanogaster serpin4, serpin5, and serpin27A* and *Anopheles gambiae serpin10* were identified in the tsetse fly EST contigs. Comparison of the reactive center loop of tsetse fly serpins with human α-1-antitrypsin suggests that these tsetse serpins are inhibitory. Preliminary EST clustering did not succeed in assembling 3564 *Tsal* encoded ESTs into one contig. In this study, these ESTs were assembled together with three published *Tsal* cDNAs. A total of 29 *Tsal*-encoded contigs were generated. An analysis of the sequence variation within the *Tsal* EST assembled contigs identified five single base mismatches namely A-T, T-A, G-T and T-G.

Results from this study form a basis onto which genetic and biochemical experimental studies can be designed, a process that will be successfully carried out once we have a reference genome. Specifically, studies aimed at genetic modification of tsetse flies towards populations that are inhabitable to trypanosomes. Ultimately, this will supplement current vector control strategies towards elimination of human African trypanosomiasis.

17th August 2009

DECLARATION

I declare that *An evolutionary genomics approach towards analysis of genes implicated in transmission of trypanosomes between tsetse fly and mammalian host* is my own work, that it has not been submitted for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Sarah Wambui Mwangi

Sarah Wambui

Signed
17th August 2009.

ACKNOWLEDGEMENTS

All thanks to almighty God, through whom all things possible and impossible come to pass.

This thesis would not have been completed without well-found support from the following: Professor Alan Christoffels, my supervisor and mentor. I am indebted to your enthusiasm towards my progress and for believing in me. It has been a privilege to work under your supervision. To my wonderful family, Oscar and Raquelle, for being my foster parents in South Africa. My mum and dad for their unrelenting sacrifices to see that we pursue our dreams. Thank you all for your moral and financial support. NRF and NBN, for availing the funds to pursue my postgraduate studies. All my colleagues at Christoffels lab, Mario, Mark, Musa, Mushal, Feziwe, Sam and Edwin. You have made my study experience in SANBI very rewarding, your ideas, criticism, and jokes. You are a great crew! Danny, Ed and Tim for all the help with the editing, and for the gift of friendship.

TABLE OF CONTENTS

KEYWORDS ii

ABSTRACT iii

DECLARATION v

ACKNOWLEDGEMENTS vi

TABLE OF CONTENTS vii

LIST OF FIGURES viii

LIST OF TABLES ix

LIST OF ABBREVIATIONS x

CHAPTER ONE: LITERATURE REVIEW AND THESIS RATIONALE 1

CHAPTER TWO: METHODS 27

CHAPTER THREE: RESULTS AND ANALYSIS 36

CHAPTER FOUR: DISCUSSION AND CONCLUSION 56

REFERENCES 65

APPENDICES 77

- APPENDIX I: Outline of PHRED and PHRAP 77
- APPENDIX II: List of Perl scripts 79

LIST OF FIGURES

Figure 1: Life Cycle of Trypanosomes. 5

Figure 2: Lateral view of the internal anatomy of the female *Glossina* species. 9

Figure 3 : Transcriptional activation of antimicrobial peptide production by the toll and IMD pathways. 13

Figure 4: The structure of human α-1-antitrypsin. 14

Figure 5: Alternative serpin conformational states 15

Figure 6: The EST generation protocol. 19

Figure 7: *G.morsitans* serpin domains as predicted by CDD. 41

Figure 8: Bayesian inference tree for insect serpins. 43

Figure 9 : Neighbor-joining tree for insect serpins. 44

Figure 10: Maximum likelihood tree for insect serpins. 45

Figure 11: Intron-exon organization of *G.morsitans* contigs aligned to *D.melanogaster* orthologues. 47

Figure 12: An overlay of cn6243 (magenta) three dimensional model onto α-1-antitrypsin (green) crystal structure (PDB code 2QUG.pdb). 48

Figure 13: An Alignment of amino acids spanning the reactive center loop................................ 49

Figure 14: Graphical representation of the conserved domain for G. morsitans PHRAP generated contig29 putative *Tsal*. 53

Figure 15 :Variation in the contig12 assembly. 54

LIST OF TABLES

Table 1 : Orthologous serpins from selected insects 29

Table 2 : Insect serpins and corresponding *G. morsitans* hits 38

Table 3 : Amino acid overlap between *Drosophila, Anopheles* and *Aedes* (as a query) versus *G. morsitans* ORFs 39

Table 4: Characteristic serpin domain residues as predicted by PFAM and SMART databases . 40

Table 5 : Tissue distribution of putative *Tsal* ESTs.. 51

Table 6 : Conserved domains of *Tsal* ORFs 52

Table 7 : Putative SNPs in GmSg121a03.plk 55

LIST OF ABBREVIATIONS

A	Adenine
Ae	*Aedes aegypti*
Ag	*Anopheles gambiae*
AMP	Antimicrobial peptide
BAC	Bacterial artificial chromosomes
BEAP	Blast extension assembly program
BLAST	Basic local alignment search tool
bp	Base Pair
C	Cytosine
CDD	Conserved domain database
cDNA	Complementary DNA
cn	Contig
cSNP	Coding single nucleotide polymorphism
DDT	Dichloro –Diphenyl-Trichloroethane
DIF	Dorsal-related Immune Factor
Dm	*Drosophila melanogaster*
dMyD88	Myosin differentiation protein
DNA	Deoxyribonucleic acid
EST	Expressed Sequence Tag
FB	Fat body
G	Guanine
GNBP	Gram negative bacteria binding proteins
GTR	General time reversible
GUI	Graphical User Interface
HSP	Heat shock protein
Ig	Immunoglobulin
IGGI	International *Glossina* genome initiative
IMD	Immune deficiency
JTT	Jones Taylor Thorton

kDa	kilo Dalton
LLR	Log-likelihood ratios
MCMCMC	Metropolis-coupled monte carlo markov chain
mRNA	Messenger RNA
NCBI	National center for biotechnology information
NNI	Nearest Neighbor Interchange
ORF	Open reading frame
PAMP	Pathogen associated molecular patterns
PDB	Protein Data Bank
PGRP	Peptidoglycan recognition proteins
PHRAP	Phragment assembly program
PPAE	Prophenoloxidase-activating enzyme
PPO	Prophenoloxidase
PRR	pattern recognition receptors
RCL	Reactive center loop
RNA	Ribonucleic acid
RNA*i*	RNA interference
rTsal	Recombinant tsetse salivary gland protein
SCF	Standard Chromatogram File
Serpin	Serine protease inhibitor
SIT	Sterile insect technique
SNP	Single Nucleotide Polymorphism
srpn	Serpin
STACK	Sequence Alignment and Consensus Knowledge Base
T	Thiamine
Th2	T-helper cells type 2
Tsal	Tsetse salivary gland protein
VSG	Variable surface glycoprotein
WHO	World health organization
α	Alpha
β	Beta

δ	Delta
3D	Three dimensional

CHAPTER ONE: LITERATURE REVIEW AND THESIS RATIONALE

1.1 Overview: tsetse fly (*Glossina*) and trypanosomiasis ………………………………….. 2

1.1.1 Trypanosomiasis ………………………………………………………………….. 2

1.1.2 Life cycle of trypanosomes ……………………………………………………….. 3

1.1.3 Clinical symptoms of human African trypanosomiasis ……………………………… 6

1.1.4 Control of human African trypanosomiasis …………………………………………. 7

1.1.5 Refractoriness to infection by tsetse flies …………………………………………… 10

1.2 Insect immune response ……………………………………………………………... 10

1.2.1 Antimicrobial peptide production …………………………………………………… 11

1.2.2 The toll pathway ……………………………………………………………………... 11

1.3 Serine protease inhibitors (serpins) …………………………………………………….. 12

1.3.1 Role of serpins in antimicrobial peptide production ………………………………… 12

1.3.2 Structure of serpins ………………………………………………………………….... 12

1.3.3 Mechanism of serpin action ………………………………………………………….. 15

1.4 The *Glossina* genome initiative and preliminary results ……………………………….. 17

1.4.1 Expressed sequence tags: A synopsis ………………………………………………... 18

1.4.2 Limitations of EST data ……………………………………………………………… 20

1.4.3 EST clustering and assembly ………………………………………………………… 21

1.4.4 Generation and assembly of *Glossina* morsitans ESTs ……………………………… 21

1.4.5 Genetic polymorphism in ESTs (a case of Tsetse salivary gland protein) …………... 21

1.4.5.1 Genetic polymorphism ………………………………………………………….. 21

1.4.5.2 Single nucleotide polymorphisms ………………………………………………. 22

1.4.5.3 Tsetse salivary gland protein …………………………………………………… 23

Thesis rationale ………………………………………………………………………………. 25

1 LITERATURE REVIEW AND THESIS RATIONALE

1.1 Overview: tsetse fly (*Glossina*) and trypanosomiasis

1.1.1 Trypanosomiasis

Tsetse flies are dipteran insects that belong to the family *Glossinidae* with a single genus *Glossina*. Based on their geographical distribution, behavioral and morphological characteristics, the genus *Glossina* is divided into three species complexes namely; *Morsitans* group (savannah flies), *Palpalis* group (riverine flies) and *Fusca* group (forest flies). Members of this genus are the biological vectors of the flagellate protozoan parasites of the genus *Trypanosoma*. Trypanosomes are single celled organisms that remain in extracellular form in the host and are the causative agents of African trypanosomiasis. The particular ecological environment of parasite and vector is such that the disease is only found in the intertropical regions of Africa (Cecchi *et al.*, 2008).

Human African trypanosomiasis is a severe disease that is lethal if it remains untreated. It is the world's third most important parasitic disease affecting human health after malaria and schistosomiaisis (Kennedy, 2007). According to the world health organization, it is estimated that there are approximately 60 million people at risk in sub-Saharan Africa and up to 50,000 deaths per year are caused by trypanosomiasis (WHO report series 1986; 739).

African trypanosomiasis continues to represent a key factor limiting rural development in vast regions of tropical Africa. With the disease affecting animals and humans, the rural populations who rely primarily on agriculture and animal husbandry for their livelihood are increasingly exposed to the disease. Additionally, weak health systems allied to war and poverty are some of the factors that have led to rapid rate of disease transmission.

Several epidemics struck parts of Africa in the beginning of the 20th century. Constant surveillance markedly reduced the disease such that by the mid 1960's the disease had almost disappeared. However, there has been a resurgence of trypanosomiasis in the last few decades due to reduced surveillance.

1.1.2 Life cycle of trypanosomes

Trypanosomiasis involves a systematic set of interactions between the vector (tsetse fly) and the mammalian host. Six main species of trypanosomes, *T.brucei, T.congolense, T.simiae, T.vivax* and *T.suis* are responsible for African trypanosomiasis. While the *brucei* group is responsible for the human form of trypanosomiasis, *T.congolense, T.simiae, T.vivax* and *T.suis* are responsible for animal trypanosomiasis.

Trypanosomes have become highly evolved and have developed an ordered set of interactions between the parasite and its vector as well as the mammalian host. The life cycle of trypanosomes is characterized by a succession of two stages, (i) a growing stage, that occurs within the fly and is adapted to infection and (ii) a non-growing stage that occurs within the mammalian bloodstream and is adapted to transmission. As the trypanosomes change their environment, they exhibit gross change in morphology, an event that is coupled with adaptive activation and repression of metabolic pathways. These modifications facilitate their adaptation to different environments (reviewed by Vickerman, 1985).

The life cycle begins in the fly and matures in the mammalian host. Developmental stages of trypanosomes are known as trypomastigotes. Upon a blood feed from an infected mammal, the tsetse fly acquires bloodstream trypomastigotes. Trypomastigote establishment (replication) takes place in the fly's midgut (Figure 1) where they divide exponentially into procyclic forms. Procyclic forms, whose morphology is short-stumpy (Figure 1), develop a fully active mitochondrion and utilize proline as the chief source of energy. Alternatively, they may use the Krebs cycle as a source of energy. During this stage, the trypanosomes express an invariant glycoprotein surface coat known as procyclin or procyclic acid repetitive protein (PARP). Procyclic trypomastigotes migrate from the midgut to the proventriculus (Figure 2) where they transform into epimastigotes and continue to replicate. Epimastigotes migrate to the salivary glands where they stop dividing. This non-dividing form is known as the metacyclic (long-slender) form. Metacyclic trypomastigotes begin to express a variable surface glycoprotein (VSG) coat as a pre-adaptation to living in the mammalian bloodstream. The cycle within the fly lasts for approximately three weeks. However, a fly remains infective for life, and the whole infective cycle is probably completed successfully in only one for every ten infected flies (Filipa *et al.*, 2008).

The Metacyclic trypomastigotes are inoculated into the mammalian host through the subcutaneous route (fly bite) where they divide exponentially in the extracellular environment. At this stage mitochondrial functions are repressed and the cells cannot engage in oxidative phosphorylation. As an alternative, they metabolize glucose to pyruvate via a glycolytic pathway that is partially compartmentalized within the glycosomes. Simultaneous to these metabolic activities is the continued expression of the variable surface glycoprotein coat. It is this coat that is recognized as an antigen by the mammalian hosts' immune system. The trypanosomes exhibit continuous variation of this antigen (Hadjuk, 1984; Pays *et al.*, 2000) to allow them to escape the hosts' immune defenses. As the immune system eliminates the trypanosomes covered by previous VSG, individual parasites expressing a new VSG multiply. This creates a continuous wave of parasitemia as bloodstream trypomastigotes are carried to other parts of the body. In the mammalian bloodstream some trypanosomes start transformation from long-slender form to short-stumpy forms in preparation for adaptation to their new environment. Infection exhibits itself at areas where the trypanosomes localize, namely, the bloodstream, lymphatic system or interstitial spaces.

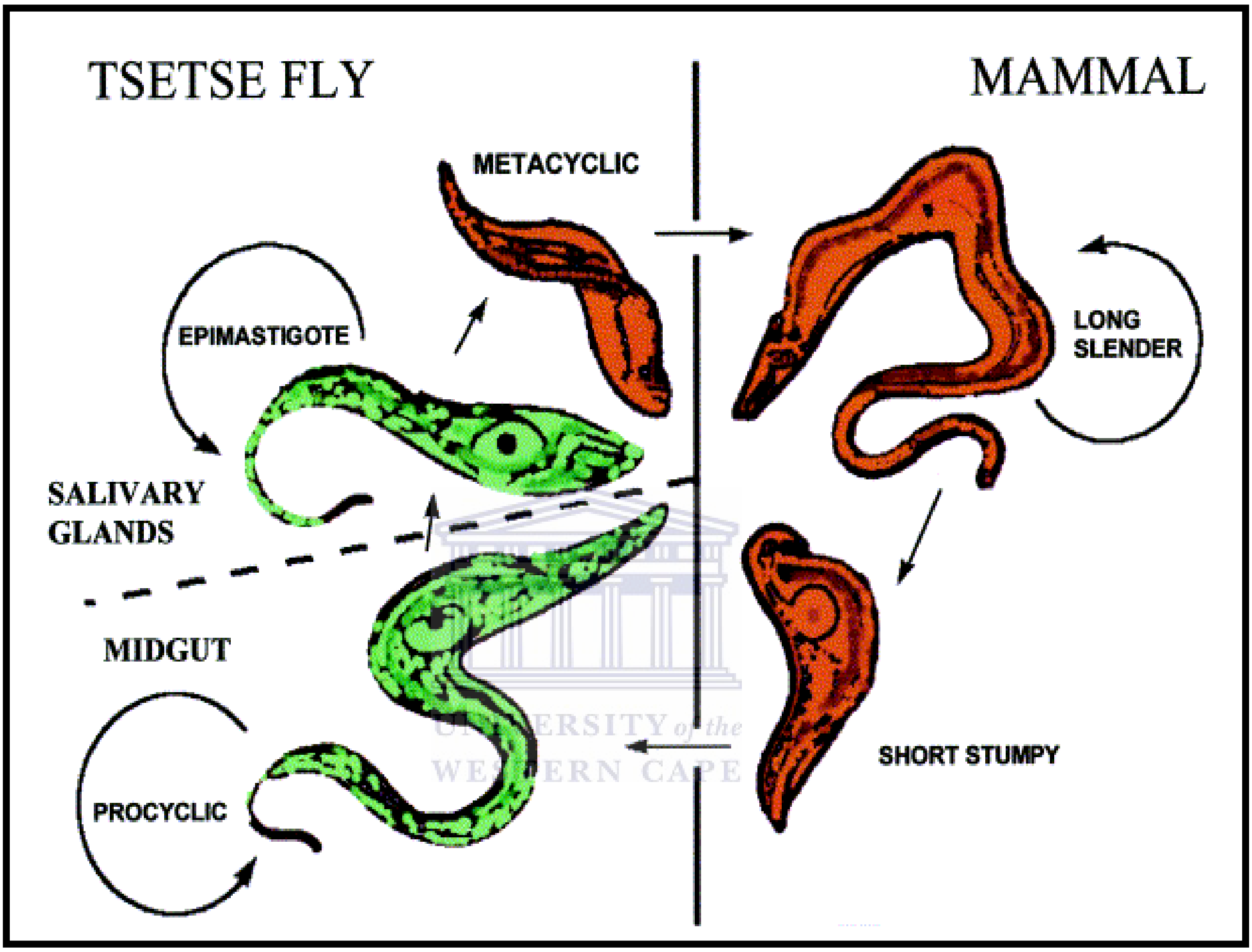


Figure 1: Life Cycle of Trypanosomes.

The life cycle begins in the fly and matures in the mammalian host. Short-stumpy procyclic trypomastigotes are ingested during a blood meal and localize within the fly’s midgut. Here, they express an invariant procyclin coat (green). Procyclic trypomastigotes then migrate to the salivary glands where they transform into metacyclic trypomastigotes (long-slender) that are infective to mammals. During the next blood meal, metacyclic trypomastigotes are inoculated into the host subcutaneously where they express a variable surface glycoprotein (red) that helps them evade the hosts’ immune defenses (El-Sayed *et al.*, 2000).

1.1.3 Clinical symptoms of human African trypanosomiasis

Trypanosome species exist as different strains whose virulence varies within the mammalian and the tsetse fly hosts (Pinder *et al.*, 1987; Rickman 1977; Okolo *et al.*, 1990). Depending on the parasite involved, human African trypanosomiasis may exhibit as either chronic or acute infection. Chronic infection is caused by *Trypanosoma brucei ghambiense,* a strain predominant in Central and West Africa that accounts for 90% of cases of trypanosomiasis. The symptoms do not present for months or even years, rather, they emerge at an advanced stage when the parasites cross the blood brain barrier to invade the central nervous system. Acute infection is caused by *Trypanosoma brucei rhodesiense,* which is found in Eastern and Southern Africa. It accounts for approximately 10% of reported cases where the first clinical symptoms are observed after a few weeks or months. The disease develops rapidly to invade the central nervous system (Dumas and Bouteille 1996).

The parasites are inoculated through the bite of an infected tsetse fly where they multiply in the subcutaneous tissue forming a trypanosomal chancre. The trypanosomal chancre presents the first symptom at the local site of infection at least five days after inoculation. The parasite then migrates as it multiplies in the blood and lymph fluid to form the primary stage of infection that is characterized by symptoms such as fever, lymphadenopathy and hepatosplenomegaly. In the case of acute infection, the symptoms can be severe often due to myocardial involvement. Approximately one tenth of the patients without rapid access to treatment will die. In the case of chronic infection the symptoms at this stage are usually inconspicuous.

In the secondary stage, the parasites cross the blood brain barrier to invade the central nervous system. Invasion of the central nervous system commences after three weeks in the case of acute infection and may take as long as several months in the case of chronic infection. This is when the signs and symptoms of the secondary stage of human African trypanosomiasis present, including chronic encephalopathy associated with mental changes for example, confusion, sensory disturbances and poor coordination. The patient enters a somnolent state due to disturbance of the sleep cycle (thus the name sleeping sickness). This is accompanied by severe body wasting that may advance to coma and ultimately death (Stich *et al.*, 2002).

1.1.4 Control of human African trypanosomiasis

Prevalence of trypanosomiasis relies on three interacting organisms: the human host, the insect vector and the pathogenic parasite. This reliance on several players is complex, yet, it presents several opportunities since the interruption of any of the players can potentially reduce the rate of disease transmission. Efforts to develop a vaccine to immunize the human hosts have been hampered by antigenic variation of the parasite's cell surface glycoprotein. Current management of Human African Trypanosomiasis relies on active surveillance and treatment of infected patients. Chemotherapy with trypanocidal drugs has for long been used in an attempt to interrupt the parasite's life cycle. Research over the past century has yielded only four clinically approved drugs, Suramin (1916), Pentamidine (1937), Melarsoprol (1946) and Eflornithine (1977) (Fairlamb, 2003). These drugs pose several limitations including, undesirable routes of administration and high levels of toxicity. Moreover, antigenic variation due to change in cell surface coats of trypanosomes increases the parasites' resistance thereby reducing the drugs efficacy levels. In addition to low efficacy of trypanocidal drugs, lack of sensitive diagnostic tools, inaccessibility to endemic rural areas and variable host range threaten chemotherapy as a long term control strategy. The completed genome of the African trypanosome, *Trypanosoma brucei* has provided insights into novel drug targets in the pursuit of developing improved chemotherapeutic agents for example, (i) glycosylphosphatidylinositol anchors that facilitate attachment of components of the variable surface glycoprotein coat and (ii) the isoprenoid metabolism pathway which is susceptible to antifungal agents (Berriman *et al.*, 2005).

Efforts to interrupt the vector have long been an important option for controlling trypanosomiasis. They seek to reduce fly numbers with the ultimate aim of eradicating tsetse flies to create fly free zones (Kabayo *et al.*, 2002). Many techniques have evolved over the years including, slaughter of game animals and clearance of huge areas of woodland. Following discovery of DDT and other persistent insecticides, aerial and ground spraying were used as additional methods for vector control throughout Africa. Over time these methods have raised health and environmental concerns. In the pursuit of environmentally friendly techniques, visual and olfactory traps have been developed together with the sterile insect technique (SIT) that involves irradiating males to sterility. The sterile males are subsequently released into the tsetse-infested zones where they compete with fertile males for fertilizing available females in order to stop the reproduction cycle (Allsop, 2001).

Aksoy and colleagues (2001) have proposed the application of molecular biology techniques to vector genetics in order to develop novel vector based strategies. This will involve characterization of vital insect targets that can be manipulated to obstruct transmission of pathogenic agents. The insect's immune system presents excellent candidature for alteration of the vector competence of the fly with the ultimate aim of disrupting the transmission cycle. This can be achieved by introducing foreign antiparasitic genes that can be passed on to the next progeny (germline transgenesis). However, the viviparous nature of the tsetse is a limiting factor of germline transgenesis because the reproductive ability of the female tsetse fly is approximately five to eight offsprings for an entire lifetime implying that, possibly hundreds of early embryos have to be microinjected. Genetic transformation of maternally inherited tsetse symbionts to express trypanocidal products constitutively (paratransgenesis) has been proposed as a novel technique of vector control (Geiger *et al.,* 2005). When microinjected into the haemolymph of the female parent, recombinant *Sodalis* species are successfully acquired by progeny flies and are passed to multiple generations with high fidelity where they continuously express marker gene products (Cheng and Aksoy 1999). The commensal symbiont *Sodalis glossinidae* is therefore a good candidate for paratransgenesis. In addition, members of *Sodalis* are localized within the midgut (see Figure 2) where trypanosome establishment takes place, thus expression of trypanocidal products by *Sodalis* can potentially block parasite establishment. *Sodalis glossinidae* has co evolved with the fly's immune system to display a high level of resistance to the insect's immune arsenal (Haines *et al.,* 2003).

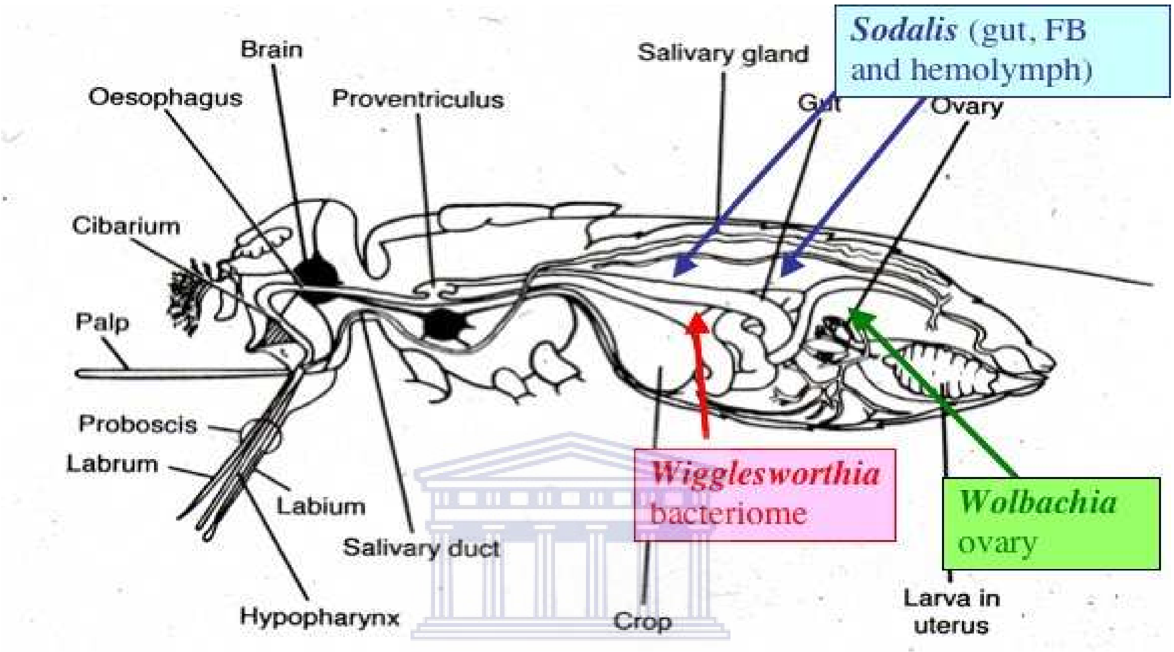


Figure 2: Lateral view of the internal anatomy of the female *Glossina* species.

While inside the fly, the trypanosomes undergo cycles of development and multiplication that involve the gut, proventriculus and salivary glands. Three different species of maternally inherited endosymbiotic microorganisms are highlighted with their respective tissue localization. The endosymbiont *Sodalis* occupies several tissues including the gut, fat body (FB) and haemolymph.

Obtained with permission from Aksoy lab: (http://aksoylab.yale.edu/)

1.1.5 Refractoriness to infection by tsetse flies

The nature of host trypanosome interaction during the trypanosome life cycle dictates that both the vector and the mammalian hosts are vital for production of a fully fledged infection. Despite being efficient vectors for disease transmission, tsetse fly infection prevalence in the field is surprisingly minimal, a phenomena that is referred to as refractoriness. In their review, Welburn and Maudlin (1999) discuss factors that account for refractoriness in tsetse flies including (i) the peritrophic matrix which acts as a first barrier against invasion by trypanosomes, (ii) midgut lectins that are capable of killing trypanosomes by a process resembling apoptosis and (iii) competition with other trypanosome strains, in the case of a mixed infection.

Milligan and colleagues (1995) showed that fly sex also influences the success of any infection with male tsetse flies producing significantly mature trypanosomes and they attribute the observation to the operation of an X-linked gene that prevents maturation of migrating trypanosomes.

Under ideal laboratory conditions, 40% of tsetse flies fed on trypanosomes of the *brucei* group were shown to be refractory while 90% self cure from the third blood meal onwards (Lehane *et al.,* 2007). The inherent mechanisms within the tsetse that facilitate self-curing are yet to be established, however, it has been demonstrated that the insect mounts a strong immune response (Hao *et al.,* 2001; Boulanger *et al.,* 2002; Hu and Aksoy 2006).

1.2 Insect immune response

Studies on the *Drosophila melanogaster* (fruit fly) immune system demonstrate the presence of an effective pathogen surveillance system that could either be cellular or humoral. They include (i) presence of epithelial barriers that form a protective coating around the insect, (ii) protease cascades, which invoke phagocytosis by haemocytes and melanocytes (melanization), (iii) production and release of reactive oxygen intermediates and (iv) production of antimicrobial peptides, (Hoffman *et al.,* 2003).

1.2.1 Antimicrobial peptide production

Antimicrobial peptide (AMP) production, the hallmark of humoral immune response is the primary mechanism by which the insect mounts an immune response. A systemic infection induces transcription of AMPs in the insect’s fat body, a diffuse organ within the insect’s haemocoel that is functionally analogous to the mammalian liver. These peptides are transported into the haemolymph where they accumulate and circulate throughout the haemocoel destroying invading pathogens.

Transcriptional activation of antimicrobial peptide production is activated by the toll and IMD (immune deficiency) pathways (Tanji *et al.*, 2005). While the toll pathway is activated in response to fungi and gram positive bacterial infection, the IMD pathway is activated in response to gram negative bacterial infection. The two pathways involve a series of proteolytic cascades that culminate in transcriptional activation of AMP production. These cascades can be divided into four distinct phases; recognition of infectious agents, signal modulation and amplification, signal transduction and finally effector response - transcription of antimicrobial peptides.

1.2.2 The toll pathway

Recognition of an infectious agent takes place in the haemolymph where pattern recognition receptors (PRRs) such as peptidoglycan recognition proteins (PGRPs) and gram negative bacteria binding proteins (GNBPs) bind to pathogen associated molecular patterns (PAMPs) for example, the lipopolysaccharides and peptidoglycans of invading microorganisms (see Figure 3).In the case of fungal infection, the toll pathway activation starts when a serine protease, Persephone is activated to cleave a cytokine-like polypeptide, Speatzle (the toll ligand) into its active form. On the other hand, gram positive bacterial infection causes direct activation of speatzle through a series of proteolytic cascades (see Figure 3). Active speatzle binds as a dimer to the toll ectodomain (cytoplasmic domain) which subsequently induces activation of toll receptors on the intracytoplasmic membrane of the cell. Activated toll then interacts with three protein kinases, a myosin differentiation protein (dMyD88), pelle and tube. Activation of these protein kinases lead to the end effect of the toll signaling pathway which is a signal dependent phosphorylation of the inhibitor cactus. Phosphorylation of cactus leads to its degradation enabling nuclear translocation of the transcription factor DIF (Dorsal-related immunity factor). DIF up-regulates transcription of the antimicrobial peptides drosomycin (reviewed by Hoffman, 2003).

The widely perceived concept that the toll and IMD pathways act independent of each other has been questioned recently. Tanji and colleagues (2006) demonstrated synergistic action between the two pathways through cross regulation. This is an important finding in the studies of mechanisms of innate immunity as it illustrates how specific ligand binding by distinct upstream pattern recognition receptors can be translated into a broad-spectrum host response.

1.3 Serine protease inhibitors (serpins)

1.3.1 Role of serpins in antimicrobial peptide production

Immune modulation of the toll and IMD pathways is an important process that regulates the immediate result of recognition and ensuing effector functions, particularly transcription of antimicrobial peptides. Signal amplification by the serine protease cascades is under tight regulation by serine protease inhibitors. These serpins process the signal two fold, they amplify strong (true) signals and dampen weak (false) signals. Thus serpins can act as both negative and positive regulators of the toll and IMD pathways. Consequently, serpins pave the way for molecules involved in the signal transduction pathway so that the recognition and amplification of the ‘true’ signal is coupled with transcriptional activation of corresponding antimicrobial peptides (Christophides *et al.*, 2002).

1.3.2 Structure of serpins

Among the families of peptidase inhibitors, serpins are found in all superkingdoms: Eukarya Bacteria and Archaea. Indeed, they are the largest and most diverse class of peptidase inhibitors (Rawlings *et al.*, 2004). A typical serpin is made up of approximately 350-400 amino acids, with elements of secondary structure being composed of three β sheets and nine α helices. Serpin crystal structure appears in five conformational states namely, the latent, native, cleaved, delta (δ), and polymeric states. These states are distinguished by the structure of the reactive center loop (RCL), an exposed flexible stretch of approximately 17 amino acids tethered between β sheets A and C (see Figures 4 and 5).

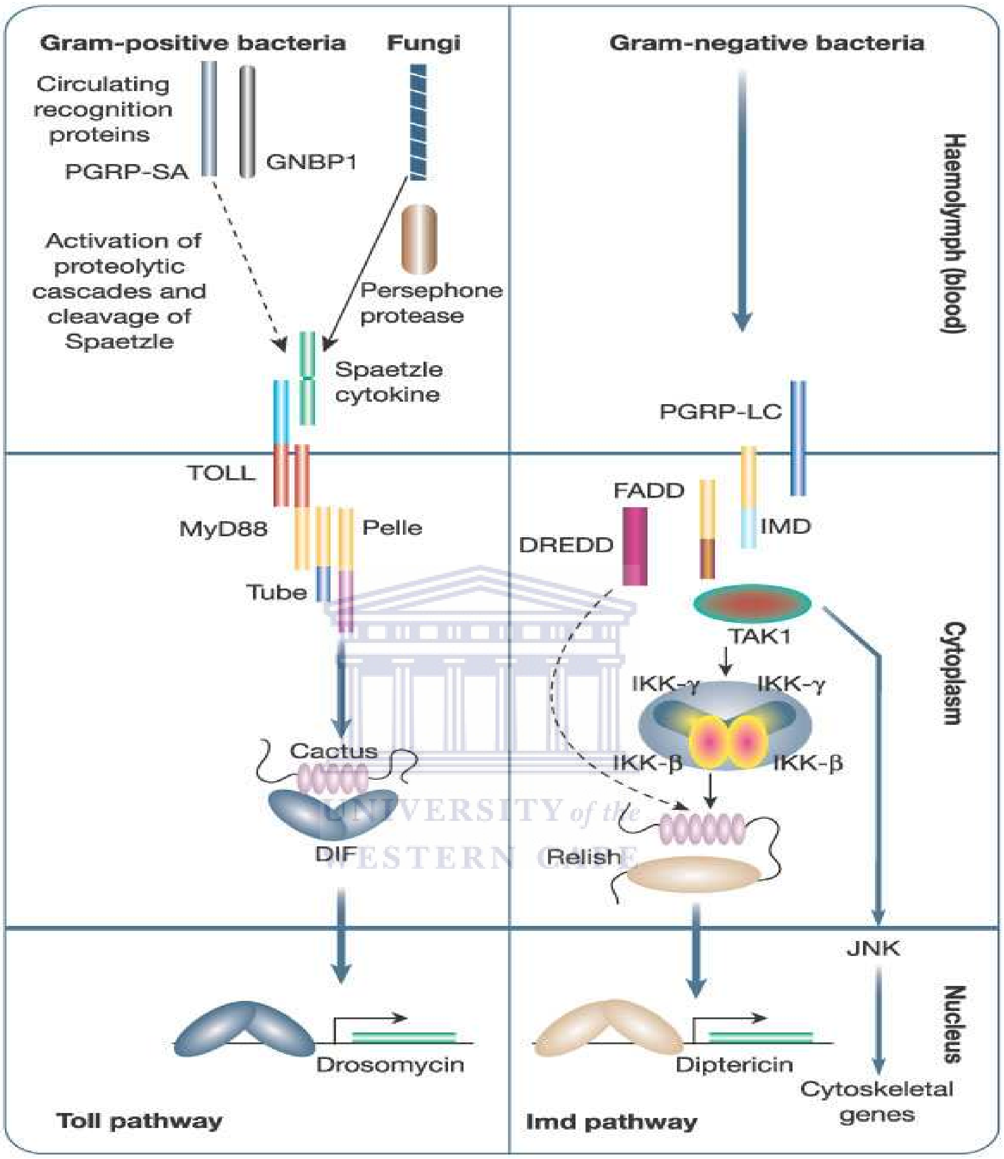


Figure 3: Transcriptional activation of antimicrobial peptide production by the toll and IMD pathways.

While the toll pathway is triggered by gram positive bacteria and fungal infection, the IMD pathway is triggered by gram negative bacteria. Activation of both pathways takes place in the insect haemolymph and molecules involved in signal transduction exert their action in the cytoplasm. Signal transduction culminates in nuclear translocation of transcription factors that up-regulate expression of antimicrobial peptides (Hoffman, 2003).

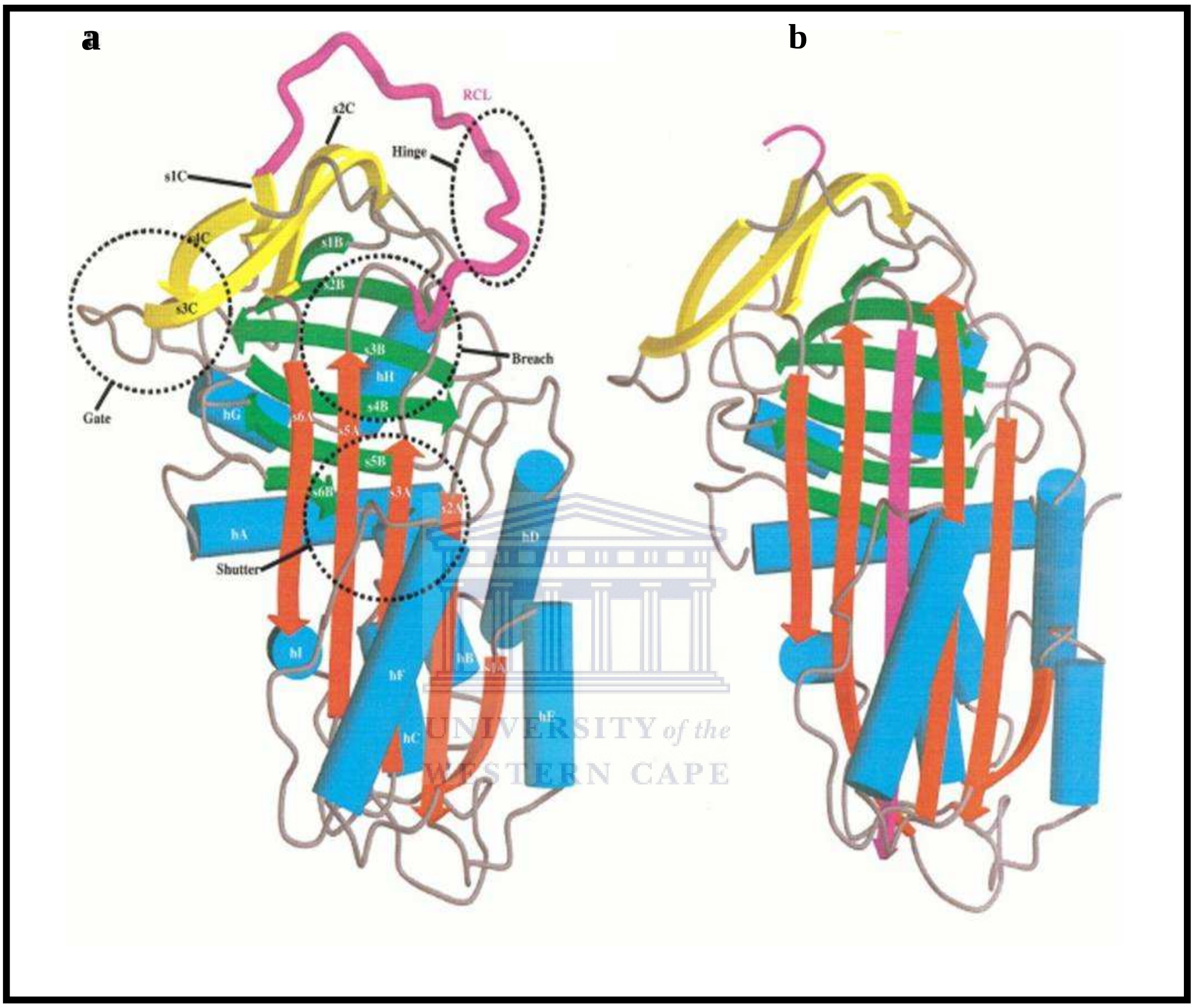


Figure 4: The structure of human α-1-antitrypsin.

Human α-1-antitrypsin was the first inhibitory serpin structure to be solved in its cleaved form.

(a) Serpin native state in which the RCL is tethered between β-sheets A (red) and C (yellow). β-sheet B is in green, and the reactive center loop (RCL) in magenta. α-helices are represented by cylinders colored cyan and denoted by the prefix ‘h’. β-sheets are denoted by the prefix‘s’. The important breach, shutter, gate, and hinge regions are indicated by broken circles. (b) Serpin in its cleaved state in which the RCL is cleaved at the scissile bond and inserts into β-sheet A (Irving *et al.,* 2000).center

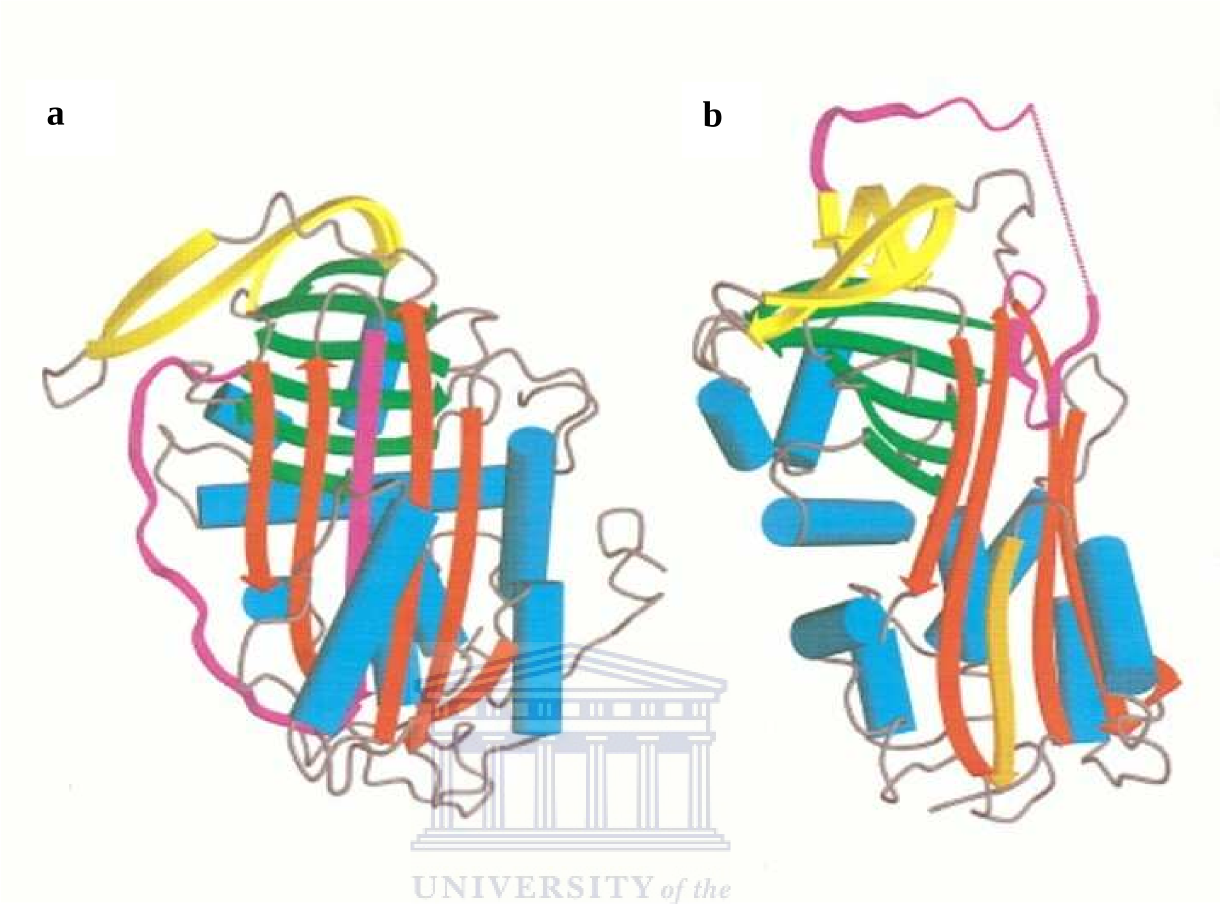


Figure 5: Alternative serpin conformational states.

(a) Latent antithrombin in which the RCL inserts into the middle of the β sheet A to give a fully antiparallel β sheet, serpins in this state are non inhibitory. (b) δ-Antichymotrypsin. Part of the F-helix is unwound and inserted into the bottom of the β-sheet A (orange) (Irving *et al.*, 2000).

1.3.3 Mechanism of serpin action

The reactive center loop consists of several regions that are important in modulating serpin conformational changes including, the hinge, breach, shutter and gate regions (see Figure 4a). The hinge portion of the reactive center loop that is most proximal to the amino terminus of the serpin holds a consensus pattern that is used to recognize inhibitory serpins (Hopkins *et al.*, 1993);

Consensus pattern: -17[E]-16 [EKR]-15[G]-14[TS]-X-(12, 9) [AGS]-$_{\text{[A/G/S]-}}$

This pattern can be translated as:

17[Glu]-16[Glu or Lys or Arg]-15[Gly]-14[Thr or Ser]-13[any]-(12, 9) [Ala or Gly or Ser] $_{\textbf{4-[Ala or Gly or Ser]-}}$

During the reaction, the hinge region provides mobility essential for the conformational change. The breach and shutter regions facilitate sheet opening and accept the conserved hinge of the RCL as it inserts. The gate facilitates full insertion into β-sheet A without cleavage. As the serpin performs its inhibitory reaction, its reactive center loop is presented as 'bait' to the target protease. The 'bait' residue (also called the P_1 residue) interacts with the S_1 site of the protease causing the cleavage of the peptide bond after the P_1 residue. Amino acids on the carboxyl side of this bond are labeled P'_1- P'_2- P'_3- P'_4- P'_n while those on the amino terminal are labeled in the order P_1- P_2- P_3- P_4- P_n (Danielli *et al.*, 2003). After the scissile bond at residues P_1–P_1' is cleaved, the reactive center loop then begins to insert into the β-sheet A and transports the covalently bound protein with it converting the enzyme to a cleaved (inactive) state (Figure 4b) which is the most stable of all serpin conformations.

Even with the functional diversification, the serpin structure remains conserved (Silverman *et al.,* 2001).Though modifications to the sequence and/or size of the reactive center region of a serpin do not affect other interactions, they strongly influence the type of target enzymes and/or the rate of their inactivation. Particularly the P_1 residue plays an important role in defining the target specificity (Potempa *et al*., 1994). Mutations in the hinge region at P_{12} or P_{14} residues of several inhibitory serpins were shown to translate the amino acids into bulky residues thereby converting them into substrates probably due to a decrease in insertion rate of the crucial reactive center loop site into the A β-sheet of the proteins (Hopkins *et al.*, 1993). These serpins lose their inhibitory function and mainly serve as hormone carriers for example, cortisol-binding globulin (Hammond *et al.*, 1987) or may have roles in the control of cell mobility or as chaperones (the 47-kDa heat shock protein HSP47), (Clarke *et al.*, 1991).

The *Drosophila melanogaster* toll pathway is regulated by *serpin43Ac*. Flies in which *serpin 43Ac* locus is mutated or absent, accumulate speatzle in its active form (Figure 3) leading to constitutive expression of the antifungal peptide drosomycin which exhibits as black necrotic spots on the flies' abdomen (Levashina *et al.*, 1999). This implies that *serpin43Ac* acts as a negative regulator (inhibitor) of the serine protease Persephone.

Serpins are also involved in regulation of yet another pathway of the insect immune response, the melanization pathway. The biosynthetic pathway of melanin begins with oxidation of phenols to quinones by phenoloxidases and polymerization with thiol compounds through a series of redox reactions to form melanin. In *Drosophila melanogaster*, *serpin27A* specifically inhibits the terminal protease prophenoloxidase-activating enzyme (PPAE) thus restricting the ensuing phenoloxidase activity to the infection site to prevent excessive melanization (De gregorio *et al.*, 2002). Recent phylogenetic studies on *Anopheles gambiae* have shown that *serpins 1, 2 and 3* form an orthologous group with *Drosophila melanogaster serpin27A* and two known lepidopteran PPAE inhibiting serpins. Indeed, RNA*i* of *Anopheles gambiae serpin2* caused spontaneous melanization reducing parasite numbers significantly. Sequence similarity analysis of the reactive center loops of serpins in the same cluster as *Anopheles gambiae serpin2* with known prophenoloxidases (PPOs) from the same species depict strong similarities in residues flanking the serpins' reactive center loop as well as adjacent amino acids (Michel *et al.*, 2005).

1.4 The *Glossina* genome initiative and preliminary results

Attempts towards eradication of human African Trypanosomiasis have focused on interrupting the transmission cycle using the three players, namely, the insect vector, the mammalian host and the parasite. Currently, the human genome has already been deciphered (Venter *et al.*, 2001) and the complete genome of *Trypanosoma brucei brucei* has been published (Berriman *et al.*, 2005). Partial shotgun genome sequence of *Trypanosoma brucei ghambiense* at 8-fold coverage is towards completion (www.sanger.ac.uk/Projects/Protozoa).

The tsetse fly genome is currently being sequenced as part of the genomic activities of the International *Glossina* genome initiative (IGGI). Members of the Palpalis group are a good justification for genome sequencing as they are the vectors for *Trypanosoma brucei brucei*, the causative agent of human African trypanosomiasis. Using flow cytometry, the palpalis genome size has been estimated at 7000 Mb while the *morsitans* genome is estimated at 500-600 Mb. It has also been estimated that the Palpalis genome is full of repeat sequences. Thus, *Glossina morsitans* genome has been selected for sequencing and will be used as a model for the more important vector Palpalis (Aksoy *et al.,* 2005).

Parallel to the tsetse genome sequencing effort, the IGGI consortium has generated expressed sequence tags (ESTs) from various tissues and developmental stages of *Glossina morsitans* with the aim of uncovering the entire transcriptome. Additionally, ESTs from a variety of tissues challenged with trypanosomes are being generated in order to identify genes that may be targets for generation of genetically modified tsetse flies that are inhospitable to trypanosomes. In addition, a BAC library for *Glossina morsitans* has been constructed to assist in assembling scaffolds for the sequencing project (ftp://ftp.sanger.ac.uk/pub/pathogens/Glossina/morsitans/). This data is aimed at providing valuable preliminary information on the genome structure, for example, putative open reading frame and repetitive elements encoded by *Glossina morsitans* genome.

1.4.1 Expressed sequence tags: A synopsis

Expressed Sequence Tags (ESTs) are partial cDNA sequences that are produced by single pass random sequencing of a cloned mRNA from either the 5’ or 3’ end (see Figure 6). The resultant sequence is a relatively low quality fragment whose length is limited to approximately 100-800 nucleotides. An EST can therefore be viewed as a small segment of an entire gene that can be used to facilitate identification of unknown genes in a speedy manner. Besides gene discovery, ESTs aid in other processes of genome analysis, for example, complete genome annotation, gene structure identification, establishment of the viability of alternative transcripts, characterization of single nucleotide polymorphisms (SNPs) and proteome analysis (reviewed by Rudd, 2003).

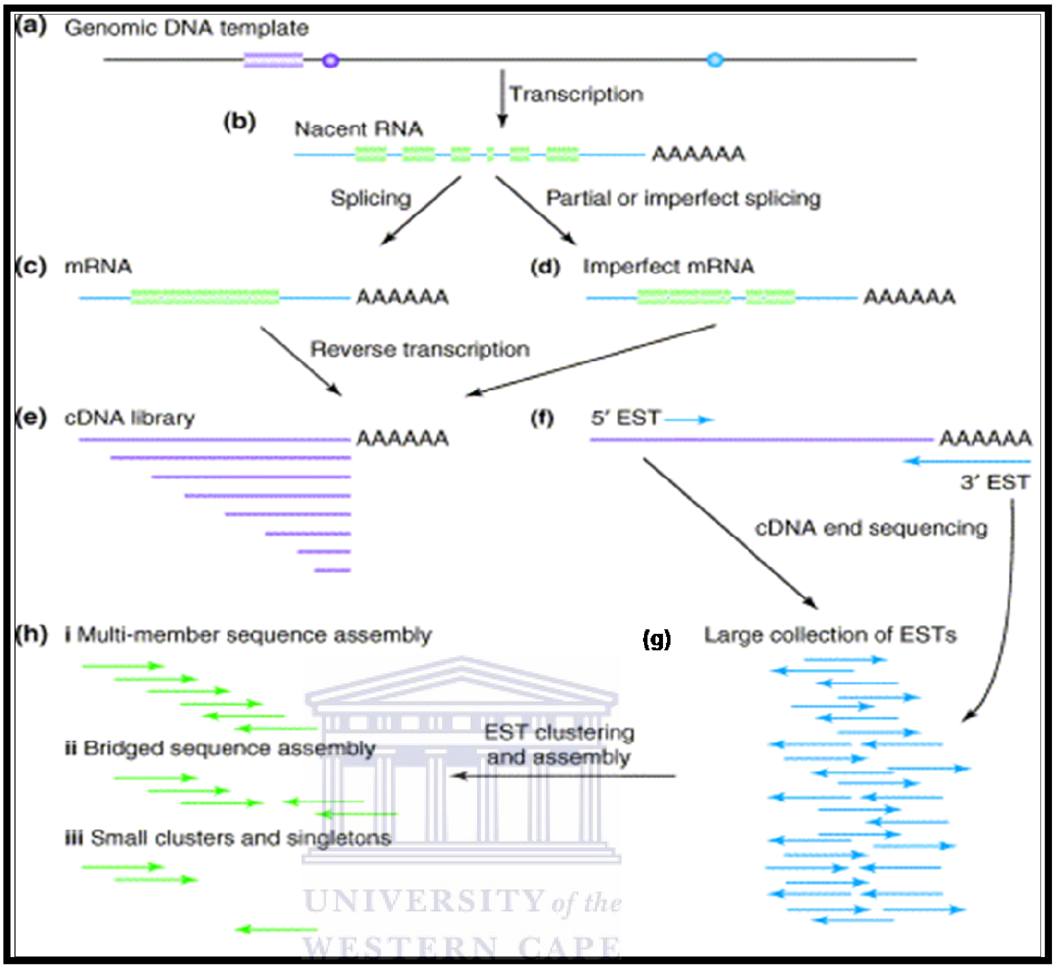


Figure 6: The EST generation protocol.

Genomic DNA template (a) contains regulatory elements and signals that define the location of a gene and recruit the DNA transcription machinery. The genomic fragment is transcribed to yield (b) a nascent RNA, a representative of a gene structure containing exons (green boxes) and introns (blue lines between green boxes) as well as untranslated regions (c) Nascent RNA is spliced perfectly into mRNAs or aberrantly to (d) imperfect mRNA which contains unspliced intron features (e) Both populations of mRNA are reverse transcribed to cDNA with the poly A tail as the selective tag thus the 3’ end of genes are more likely to be represented in cDNA libraries (f) cDNAs undergo single pass sequencing in either direction (g) to generate a pool of ESTs that are (h) clustered and assembled to (hi) clusters that faithfully represent the spliced gene structure of parental cDNA or (hii) pseudo-clusters, when ESTs stem from the same parental cDNA clone or (hiii) several ESTs aggregate into either small clusters or remain as singletons (Rudd, 2003).

1.4.2 Limitations of EST data

Two shortcomings are associated with EST sequences: the overall representation of host genes within a library, and the overall quality of any individual sequence within a collection (Rudd 2003).

Ordinarily, a standard cDNA library will represent the ratio of mRNAs present within a specific tissue under exact conditions at the time of sampling. Thus redundant clones of over expressed genes will be present while there will be poor representation of genes that are moderately expressed. Genes that are not expressed at the time of sampling will be absent. To address this limitation, a normalization procedure in cDNA library construction has been developed (Bonaldo *et al*., 1996). Normalization aims at equalization of overrepresented and under-represented transcripts. Normalization significantly increases gene discovery rate of a given cDNA library. In addition, oligo fingerprinting (Clark *et al.,* 2001; Herwing *et al.*, 2002) has also been used to equalize the relative occurrence of rare transcripts.

With regard to ESTs, sequence quality describes the fidelity with which an EST sequence represents the gene sequence from which it was reverse transcribed and cloned. This relies on the experimental conditions particularly the quality of reverse transcriptase used for conversion of mRNA to cDNA. Due to the tendency to have base calling errors at the beginning and towards the end of the sequencing reaction, a typical EST will have bases along the middle segments being of high quality. Other sources of low sequence quality include contamination from the cloning vector, repeat sequences during base calling, poly A tails that may introduce sequence bias and may cause structural or regulatory RNAs to be cloned. In addition to these are xenocontaminants (sequences from foreign genomes, for example microbial flora).

All these problems coupled with the limited length of ESTs present challenges during downstream analysis of ESTs. Moreover, it is difficult to distinguish natural sequence variations from sequencing artefacts, for example in the case of single nucleotide polymorphisms.

1.4.3 EST clustering and assembly

The length of an EST is limited to a few hundred nucleotides and therefore they are significantly shorter than the length of the cognate gene. Thus incomplete sequence coverage is a hindrance to the process of gene discovery during downstream analysis of ESTs. EST clustering and assembly aims at 'stacking' together overlapping ESTs originating from one transcript (gene). The ultimate aim is to have clones that represent the same gene in one group so that the redundancy of a particular dataset is reduced.

An EST cluster can be defined as "fragmented EST data and (if known) gene sequence data consolidated, placed in correct context and indexed by gene such that all expressed data containing a single gene is in a single index class and each index class contains the information for only one gene" (Burke *et al*., 1999). PHRAP (Appendix I) and CAP3 (Huang and Madan 1999) are the most commonly used programs for sequence clustering and assembly.

1.4.4 Generation and assembly of *Glossina* morsitans ESTs

The EST datasets used in this study were generated by different laboratories as part of the larger International *Glossina* genome initiative. 124,000,000 ESTs were sequenced from 11 tissue and developmental-stage specific EST libraries. Clustering assembly was performed through the IGGI consortium using STACK (Sequence Alignment and Consensus Knowledge Base) (Miller *et al*., 1999; Christoffels *et al*., 2001) at the South African National Bioinformatics Institute. Clustering and assembly resulted in 18,413 contigs/singletons that were computationally annotated and manually curated. The data is available via GeneDB (http://www.genedb.org/genedb/glossina/).

1.4.5 Genetic polymorphism in ESTs (a case of Tsetse salivary gland protein)

1.4.5.1 Genetic polymorphism

Genetic polymorphism implies the existence of multiple alleles at one particular gene locus. For mapping studies, three types of genetic polymorphisms are used: single nucleotide polymorphisms (SNP), microsatellites (variable number tandem repeats) and minisatellites (short tandem repeats). Microsatellites and minisatellites are essentially repeat units of 1-4 and 20-100 nucleotides respectively that occur in both the coding and non-coding regions of DNA.

1.4.5.2 Single nucleotide polymorphisms

Expressed sequence tags are a rich source of identifying polymorphisms in transcribed regions, for example, redundant EST libraries will contain multiple transcripts of the same gene. This redundancy can be useful in detecting polymorphisms in a particular gene especially when the transcripts are from multiple individuals (Guryev *et al.*, 2004).

Single nucleotide polymorphisms (SNPs) denote a difference in a single nucleotide at one particular site. SNPs are the most abundant type of genetic variation. A comparison of any two chromosomes in the human genome estimates occurrence of SNPs every 500 to 1000 bp (Cooper *et al.*, 1985). Genome analysis of other vertebrate species for example, the teleost *Medaka* (Kasahara *et al*., 2007) and the domestic dog (Lindblad-Toh *et al*., 2005) shows that the frequency of SNPs is higher in other organisms. SNPs can occur in the coding (cSNP) as well as the non-coding region of DNA. Coding SNPs often generate polymorphic variants of expressed proteins that may or may not affect their functional properties. If a SNP is associated with a change in protein function, then it is referred to as non-synonymous. Non-synonymous SNPs are mainly results of missense or nonsense mutations which alter the amino acid sequence or form premature stop codons respectively. Single nucleotide polymorphisms that do not alter protein function are referred to as synonymous.

As genetic markers, SNPs present several advantages namely, (i) they can achieve high levels of automation during genotyping (Stickney *et al.*, 2002), (ii) their mode of inheritance is codominant and thus they are suitable for comparative genomic analyses, (Lindblad-Toh *et al*., 2005), (iii) compared to tandem repeat markers, they are more stable because they have a lower mutation rate and (iv) because they are fundamental causes of genetic variation, mapping of SNPs would enable the identification of the causative SNP as well as associated SNP with specific and complex traits. The main limitation of SNPs as a measure of polymorphism is their bi allelic nature, thus they are less informative than multi allelic microsatellite markers. However, this is balanced off by their abundance as well as the use of SNP haplotypes.

A major difficulty when trying to validate SNPs from EST resources is the failure to differentiate a sequencing error and true polymorphism. To circumvent this, Marth and coworkers (1999) suggested that a predicted SNP is deemed valid only if the corresponding quality score on the base is satisfactory. Hayes and colleagues (2007) have proposed another method which involves the creation of two independent datasets, running the SNP detection pipeline and retaining only cross validated SNPs.

1.4.5.3 Tsetse salivary gland protein

In tsetse flies, salivary gland extracts have been found to contain more than twenty proteins whose detection from sera varies with respect to time of exposure and sensitivity of the host (Ellis *et al.*, 1986). Most of these proteins are considered essential in development of procyclic trypomastigotes to metacyclic forms. On the other hand, some of these proteins contain anti-thrombin anticoagulant activity and platelet anti-aggregation activity, both important in prevention of haemostasis (Mant and Parker 1981). Other extracts contain molecules that cause immediate and delayed-type cutaneous hypersensitivity (Ellis *et al.*, 1986).While the anticoagulant molecules reduce clotting, immunoreactive molecules are thought to minimize the inflammatory responses resulting in continuance of an efficient blood flow.

Two proteins that are unique to the tsetse fly sialome (salivary gland transcriptome) have been identified as *Tsal1* and *Tsal2* (Li *et al*., 2001). *Tsal2* comprise two isoforms, *A* and *B* predicted to encode mature proteins of 399 amino acids (45 kDa) and 389 amino acids (43 kDa) respectively. In the same study, Li and colleagues (2001) observed varied expression profiles with respect to tissue distribution and fly sex. While transcripts specific for *Tsal2* can be detected in all developmental stages, *Tsal1* expression is limited to adult and larval stages. Additionally, salivary glands of adult males are found to express higher levels of *Tsal2* in comparison to females while there is no noteworthy sex-based difference for *Tsal1* expression.

To date, sequence and protein domain database searches of both DNA and protein sequences have failed to obtain any significant hits. A study performed on the salivary gland transcriptome of *Culex quinquefasciatus* has shown that *Tsal1* and *Tsal2* exhibit similarities with an abundant cluster of sequences coding for secreted proteins with endonuclease activity (Ribeiro *et al.*, 2004).

In mice and human the *Tsal* protein is immunogenic (Caljon *et al.*, 2006). The saliva-induced immune response is coupled to a T-helper cells type 2 (Th2)-biased cytokine production together with the release of mainly IgG antibodies. Ellis and colleagues (1986) were able to demonstrate that all serum samples contain anti IgGs elicited against the 42-45 kDa (*Tsal*) band of proteins albeit individual variations in immune reactivity against other tsetse salivary components. Recombinant *Tsal1/2* (*rTsal1/2*) based immune screening of Ugandan samples has been shown to efficiently cross detect both *Glossina morsitans* as well as *Glossina fuscipes* exposure (Hide, 1999). In the same study, no *Tsal1/2* cross-reacting antibodies were raised in individuals exposed to the bite of *Anopheline* mosquitoes. Thus, the recognition of anti *rTsal1/2* IgGs have been proposed as an indicator of previous tsetse fly exposure in epidemiological surveys of tsetse fly challenge especially in geographical areas where *Anopheline* mosquitoes occur.

Thesis rationale

Being vectors of trypanosomes, tsetse flies are of major medical, veterinary and economic importance in Africa. All the techniques of vector control present the problem of sustainability because reinvasion is unavoidable unless entire populations are wiped out and cleared regions guarded. With the advent of molecular genetics, understanding host-parasite interactions at the molecular level is imperative for the development of improved novel approaches towards control and ultimately eradication of human African trypanosomiasis. Refractoriness to infection by tsetse flies has been attributed to their robust immune defenses. Thus, examining the components of the tsetse fly immune repertoire would provide insights into the underlying mechanisms of refractoriness.

As of 2008, IGGI had annotated the *Glossina morsitans* transcriptome. This thesis aims at detailed bioinformatic characterization of ESTs coding for some of the immune related genes in tsetse flies using the transcriptome data. Specifically, serpin genes are analyzed by examining evolutionary relationships between putative serpins and serpins from selected insects using phylogenetic tree reconstruction. The largest EST cluster sampled from the *Glossina morsitans* transcriptome contained 3564 ESTs that share similarity to the gene encoding tsetse salivary gland protein (*Tsal).* Despite successful clustering of putative *Tsal* ESTs, repeated automated assembly did not generate one single contig. In this study the *Tsal* cluster was examined at the level of manual EST assembly with the goal of identifying potential sequence variation.

Characterization of these genes from the transcriptome data may well facilitate understanding of *Glossina* biology particularly with respect to its interaction with trypanosomes. This will form a basis onto which further genetic and biochemical experiments can be designed in the pursuit of reducing vector competence by use of genetically modified tsetse flies.

In summary this thesis has two main objectives:

1. To establish the evolutionary relationships among serpin genes from *Glossina morsitans, Drosophila melanogaster, Anopheles gambiae and Aedes aegypti.*

2. To assemble the putative *Tsal* cluster and examine sequence variation within the cluster as a signature for host defense and a potential explanation of the unsuccessful automated assembly.

CHAPTER TWO: METHODS

2.1 Identification of selected immune related insect serpins 28

2.2 Homologue identification and alignment of sequences 28

2.3 Removal of potential alternative splice variants 28

2.4 Domain searches 30

2.5 Alignment editing 30

2.6 Construction of phylogenetic trees 30

2.6.1 Bayesian inference 31

2.6.2 Distance methods 31

2.6.3 Likelihood methods 31

2.7 Prediction of Intron-exon organization 32

2.8 Homology modeling 32

2.9 Identification of the putative reactive center loop 33

2.10 Tsetse salivary gland EST PHRAP assembly 33

2.11 Functional annotation 34

2.12 Assembly viewing 34

2.13 *Tsal* sequence variation and SNP identification 35

2 METHODS

2.1 Identification of selected immune related insect serpins

A list of serpin genes from *A. gambiae* and *D.melanogaster* was compiled based on molecular evidence as regulators of the immune response from published literature. Where there was no experimental evidence, for example the case of *Aedes aegypti,* selection was based on their assignment as orthologues of serpin genes from recent reviews (Christophides *et al.,* 2002; Tanaka *et al.,* 2008) as well as novel ENSEMBL predictions. A total of 27 peptide sequences were downloaded from ENSEMBL, GENBANK AND FLYBASE databases (Table 1) in FASTA format to constitute an initial list that was used for subsequent BLAST searches.

2.2 Homologue identification and alignment of sequences

Protein sequences were demonstrated to be more suitable for homology detection over long periods of evolutionary time (Pearson 1997). Peptide sequences of *G.morsitans* ESTs' contigs and singletons were obtained and merged into one text file. This text file was indexed by FORMATDB program for BLAST searching. The initial dataset of 27 peptide sequences (see Table 1) was used to query the *G.morsitans* peptide database using BLASTP, one of the executables of the BLAST suite (Altschul *et al*., 1990). The expectation cutoff value was set at 1e^{-4} while the rest of the parameters were set to default.

2.3 Removal of potential alternative splice variants

Multiple *G.morsitans* contigs were identified as significant hits against the same *D.melanogaster* serpin (see Table 2). We assume that contigs derived from the same EST cluster represent different isoforms of the same gene. Five contigs originating from cluster 677 matched *D.melanogaster serpin4*. In this example, the longest contig was selected with the assumption that it holds the most biological information. The caveat of this approach is that potential paralogues are deleted together with the splice variants. However, in the absence of a reference *Glossina* genome, paralogues cannot be verified. All serpin putative alternative splice isoforms were filtered using the above-mentioned approach and resulted in the retention of four contigs (see Table 4). These contigs were merged with 27 insect serpin sequences resulting in 31 peptide sequences that were used for phylogenetic reconstruction.

Table 1 : Orthologous serpins from selected insects

ACCESSION NUMBERS			
Anopheles gambiae**[*1]**	***Aedes aegypti*****[*1]**	***Drosophila melanogaster*****[*3]**	**References**
AGAP006909	AAEL005670	FBgn0028990	Michel *et al* .,(2005)
	AAEL0014079		Nappi *et al.*, (2005)
			De Gregorio *et al.*,(2001)
AGAP006911	AAEL005670	FBgn0028990	Michel *et al.*, (2005)
	AAEL0014079		Nappi *et al.*, (2005)
			De Gregorio *et al.*, (2001)
AGAP006910	AAEL005665	FBgn0028990	Michel *et al.*, (2005)
			Nappi *et al.*, (2005)
			De Gregorio *et al.*, (2001)
AGAP009670	AAEL013933	FBgn0031973	Christophides *et al.*, (2001)
AGAP009221	AAEL0014141	FBgn0031973	Christophides *et al.*, (2002)
AGAP009212	AAEL0010769	FBgn0031973	Eappen *et al.*, (2002)
gi|15156470[*2]	AAEL002699	-	Christophides *et al.*, (2002)
AGAP003194	AAEL0011777	FBgn0036968	Christophides *et al.*, (2002)
AGAP003139	AAEL008364	FBgn0028984	Christophides *et al.*, (2002)
		FBgn0038299	Danielli *et al.*, (2003)
gi|187441046[*2]		FBgn0028985	Christophides *et al.*, (2002)
	-		Irving *et al.*, (2001)
			Danielli *et al.*, (2003)
AGAP009213	AAEL014138	FBgn0031973	Michel *et al.*,(2001)

Accessions correspond to the database of sequence origin;

[*1] ENSEMBL; [*2]GENBANK; [*3] FLYBASE;

2.4 Domain searches

Searches for the presence of conserved domains were conducted against the PFAM (http://pfam.sanger.ac.uk/) and SMART (http://smart.embl-heidelberg.de/) profiles with the local alignment setting. Sequence portions spanning the serpin family domain were recorded (see Table 4). In addition a BLASTP of the retained contigs was conducted against NCBI's non-redundant protein database. Graphical representations of the candidate contigs were obtained from NCBI's conserved domain database (CDD) (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) (See Figure 7).

2.5 Alignment editing

Multiple sequence alignment of amino acid sequences was performed by CLUSTALW (Thompson *et al.*, 1994) using default parameters:, BLOSUM62 (for protein) was used as the weight matrix, gap opening penalty set at 10.0 and gap extension penalty at 0.05. The option for hydrophilic gaps was switched on with GPSNDQERK flagged as hydrophilic residues. The option for residue specific gap penalties option was also switched on. The alignment was edited manually using the JALVIEW java alignment editor (Waterhouse *et al.*, 2009), by removal of all columns with gaps. Thereafter, sequences that extended over the length of the longest contig (cn1692, 535 amino acids) were trimmed, resulting in sequences that were 535 characters in length. The edited sequences were aligned once more using CLUSTALW with the default parameters. Two output format options were selected, (i) nexus for further analysis using Bayesian methods and (ii) phylip for further analysis using distance and maximum likelihood methods.

2.6 Construction of phylogenetic trees

Because different methods are based on different principles, each method used for tree reconstruction has its merits and demerits. Likelihood and Bayesian methods provide a statistical framework for selecting the most probable tree compared to distance and parsimony methods that generate a consensus tree. All phylogenetic methods assume an evolutionary process to explain the observed character changes and therefore potentially generate different trees.

In an attempt to analyze the evolutionary relationships between *G. morsitans* putative serpins and serpins from selected insects, the sequence data was subjected to three different methods of phylogenetic reconstruction. The aim of using three methods was to evaluate the consistency of the tree topology.

2.6.1 Bayesian inference

MR BAYES v3.1.2 (Huelsenbeck and Ronquist, 2001) was used to generate phylogenetic inference with the following options. When specifying the model of analysis the GTR model was chosen (Lset nst = 6 rates = gamma). This is a General Time Reversible model with a proportion of invariable sites and a gamma shaped distribution of rates across sites settings. The priors were left at the defaults of the MR BAYES v3.1.2 program. The dataset was analyzed with two independent Markov chains run for 10,000 Metropolis-coupled Monte Carlo Markov Chain (MCMCMC) generations, with tree sampling every 10 generations (samplefreq10). Finally the 100 Bayesian majority rule consensus tree was constructed.

2.6.2 Distance methods

Bootstrap resampling (1000 datasets) was performed on the phylip-formatted alignment using the SEQBOOT program in the PHYLIP v 3.68 suite (Felsenstein, 1989). Distance measurements between proteins in the bootstrapped datasets were computed by PROTDIST within the PHYLIP package. The Jones Taylor Thorton (JTT) model of amino acid substitution (Jones *et al.*, 1992) was employed with no gamma distribution of rates among positions. This was followed by the generation of unrooted tree distance matrices of the datasets using the neighbor-joining algorithm of the program NEIGHBOR within the PHYLIP suite. Finally, a majority rule consensus tree was constructed using the CONSENSE program with *Dm31973* peptide sequence as the outgroup.

2.6.3 Likelihood methods

PHYML command line version 3.0 was implemented for all likelihood measurements (Guindon *et al.*, 2005). The parameters were set as follows: the datasets were analyzed with JTT as the substitution model for amino acid substitution, proportion of invariable sites was set at fixed while the number of substitution rate categories was set at four and the gamma shape parameter was set to be estimated by the program. For tree searching, the starting tree was set at the program's default. BIONJ (Gascuel O, 1997) distance-based tree with NNI (Nearest Neighbor Interchange) was employed as the type of tree improvement. The number of random starting trees was set at five and 100 bootstrap resamplings were performed on our dataset for branch support.

Newick files from all consensus trees were loaded into MEGA (Kumar *et al.*, 2004) for conversion into phylograms.

2.7 Prediction of Intron-exon organization

Nucleotide sequences of *G.morsitans* contigs (see Table 4) were used to perform a TBLASTN against the peptide database of *Drosophila melanogaster* in ENSEMBL. The most significant hit was identified. The "Contig view" option in ENSEMBL allows for a graphical view of synteny conservation among orthologous genes. Graphical representation of the predicted intron-exon organization against putative orthologues was obtained from the "contig view" feature. These intron-exon features were compared to the *G.morsitans* exon boundaries to assess any conservation in intron-exon junctions.

2.8 Homology modeling

We aimed at identifying the structurally conserved regions of the putative *G.morsitans* serpins against a known serpin structure such as α-1-antitrypsin. This serpin (α-1-antitrypsin) was the first inhibitory serpin structure to be solved in its cleaved form (Löbermann *et al.*, 1984) and has been extensively used as a template for active serpins.

The three-dimensional (3D) structure of α-1-antitrypsin was downloaded from Protein Data Bank (PDB http://www.rcsb.org/pdb/explore.do?structureId=1QLP). Homology modeling of putative serpins was carried out using the program MODELLER (Sali and Blundell, 1993), a comparative protein modeling method designed to find the most probable structure for a sequence given its alignment with related structures. The 3D model is obtained by optimally satisfying spatial restraints derived from the alignment. The resultant model was viewed using PYMOL (DeLano, 2002).

2.9 Identification of the putative reactive center loop

The sequence conformation of the reactive center loop largely determines the selectivity of the inhibition. In order to establish whether these putative serpins were inhibitory, the amino acid residues flanking the C terminus where the reactive center loop is located were identified. An alignment of the *G.morsitans* ORFs and their predicted *D.melanogaster* orthologues was performed alongside the sequence of human α-1-antitrypsin using TCOFFEE (Notredam *et al.*, 2000). The consensus pattern of the reactive center loop starts at P_{17} (see Chapter one section 1. 3.3). This pattern was used to facilitate identification of the position of the scissile bond, which is 17 amino acids downstream of the start of the consensus pattern (Whisstock *et al.,* 2000) (see Figure 13).

2.10 Tsetse salivary gland EST PHRAP assembly

The aim of any assembly process is to reduce redundancy of a large set of EST data by constructing consensus sequences (contigs). Contigs are generated from reads that share a given level of similarity (specified during the assembly process). Errors associated with the process of EST generation (discussed in Chapter one section 1.4.2) are taken into account to ensure that the consensus sequences are of high quality.

The 3564 *Tsal* ESTs were assembled together with the three published *Tsal* cDNAs. In the absence of the trace files for the three *Tsal* cDNAs downloaded from GENBANK, a custom Perl script was used to create arbitrary quality scores for them (see Appendix II). Each base in the GENBANK sequences was assigned a quality score of 30. The three GENBANK cDNA files together with their quality scores files were appended to the original dataset of 3564 *Tsal* ESTs and its corresponding quality scores file. The resulting datasets had 3567 *Tsal* EST sequences and 3567 corresponding quality scores in FASTA format. This data was used as the input files to the PHRAP program (see appendix I) using the following UNIX command:

“. /phrap -ace tsal.fasta” where “tsal.fasta contained the 3567 sequences and quality scores and the “-ace” flag ensured the assembly in ace format.

Five files were generated for the PHRAP assembly: (i) The standard error, indicating the status of the PHRAP run, a summary of the results of some of the steps and various warning and error messages, (ii) the ".contigs" file which is a FASTA file that included singletons (contigs consisting of single reads with a match to some other contig but could not be merged consistently with it), (iii) the ".contigs.qual" file which had PHRAP-generated quality scores for the contig bases, (iv) the ".singlets" file which was a FASTA file containing singlet reads and (v) the ".log" file containing various diagnostic information that could be helpful for troubleshooting. In addition to these was the ".ace" output generated for assembly viewing,

2.11 Functional annotation

A major setback with EST sequencing is that it seldom allows determination of a complete cDNA sequence because some genes may be too large such that end sequencing does not cover them. In addition, sequence quality drops towards the end of the sequence reads which can prevent assembly programs from joining overlapping sequences into a single contig. In an effort to confirm that the contigs did not represent untranslated regions of a transcript, conceptual translation using NCBI's Open Reading Frame (ORF) finder (http://www.ncbi.nlm.nih.gov/projects/gorf/) was done. The longest reading frame was selected as the representative novel peptide. These peptides were then searched using BLAST against NCBI's non-redundant protein database to identify putative homologues and conserved domains. Additional searches for domain conservation were done against the PFAM and SMART databases.

2.12 Assembly viewing

The BEAP (Blast Extension and Assembly Program) viewer source code was downloaded from http://www.animalgenome.org/bioinfo/tools/share/BEAP/ and the program installed onto a stand alone computer. Assembly viewing was done using the graphical user interface of the BEAP program (Koltes *et al.*, 2009).

BEAP combines BLAST and CAP3 (Huang and Madan 1999) to retrieve sequences and construct contigs for localized genomic sequences in species with incomplete sequence drafts. Sequence alignments can be viewed graphically with the BEAP viewer, a Java graphical user interface (GUI), that allows users to evaluate contig sequence quality and predict SNPs.

The BEAP viewer facilitates viewing of the sequence alignments from the ".ace" output either in the line overlap representation view or nucleotide alignment view. Only one contig alignment can be viewed at any given time. Within the nucleotide alignment view is a color differentiation option. Bases that do not match with the contig are marked either in red or green for ease of finding areas of low quality and to quickly spot possible single nucleotide polymorphisms. A red base denotes a mismatch in that particular position but with a high quality score while a green base denotes a mismatch with a low score.

2.13 *Tsal* sequence variation and SNP identification

The PHRAP assembly was used as input to screen for *Tsal* variation. The PHRAP assembly output file was converted to GDE format using a customized ace2gde Perl script which was also used to parse the GDE-formatted assembly file to identify sequence variation in each aligned column (see appendix II for command line usage). The quality scores of each variable nucleotide were extracted to confirm their base-calling accuracy: the underlying assumption is that quality scores below 30 have low confidence.

CHAPTER THREE: RESULTS AND ANALYSIS

3.1 Identification of *G.morsitans* homologues of immune related insect serpins ………….. 37

3.2 Serpin domain searches ……………………………………………………………… 37

3.3 Phylogenetic trees analysis ……………………………………………………….... 42

3.4 Intron-exon organization …………………………………………………………….. 46

3.5 Homology modeling ……………………………………………………………………. 46

3.6 Identification of the reactive center loop ………………………………………….. 46

3.7 Assembly of *Tsal* encoded ESTs …………………………………………………….. 50

3.8 Functional annotation of *Tsal* ORFs ………………………………………………….50

3.9 Polymorphism in *Tsal* ESTs ………………….. ……………………………………….. 53

3 RESULTS AND ANALYSIS

3.1 Identification of *G.morsitans* homologues of immune related insect serpins

G.morsitans peptides, predicted from the *Glossina* transcriptome data, were searched against a collection of 27 verified insect serpins. Nine *G.morsitans* putative peptides were identified with percent identity to *Anopheles gambiae, Aedes aegypti and Drosophila melanogaster* ranging from 42% to 66% (Table 2). At least five of the nine *G.morsitans* peptides represented putative homologues to *D.melanogaster serpin4*. Four of them were eliminated based on the assumption that they were putative alternative splice variants (see section 2.3). Another two *G.morsitans* peptides, cn1692 and cn1693, showed significant identity to *D.melanogaster serpin27A* whereas cn1693 was identified as having an additional homologue namely *A. gambiae serpin9*. Contig 1692 was retained for further analysis (see section 2.3).

Calculations of the amino acid coverage of the insect serpin queries and their respective *G. morsitans* hits show that a high percent identity does not correspond to complete overlap. An interesting observation is the low amino acid overlap of cn3298 against its corresponding homologues. Even with the highest percent identity, cn3298 spans less than 30% of its corresponding homologues in all instances (see Tables 2 and 3). Another instance of low amino acid overlap is recorded for cn2215 which notably gives the highest percent identity among all *G.morsitans* peptides that are the most significant hits for *D.melanogaster serpin4* (see Tables 2 and 3).

3.2 Serpin domain searches

The search for conserved domains confirmed that both *G.morsitans* peptides and insect serpins encoded the serpin conserved domain. Both SMART and PFAM databases show that the conserved residues are within the same approximate amino acid range for all *G.morsitans* sequences. *G.morsitans* cn3298 is predicted to encode a serpin domain that is fragmented into two portions (see Table 4 and Figure 7). The SMART protein domain database predicts the smaller portion as a transmembrane domain. It is possible that the poor amino acid overlap observed between cn3298 and its corresponding insect homologues is a result of the fragmentation of the serpin domain. In addition to the domain conservation, NCBI’s CDD also shows the relative position of the reactive center loop (Figure 7). This feature indicated that indeed all the candidate *G.morsitans* contigs possess the amino acid residues that flank the reactive center loop, a signature needed to identify inhibitory serpins.

Table 2 : Insect serpins and corresponding *G. morsitans* hits

SERPIN ID	*G. morsitans* CONTIGS	PERCENT IDENTITY (%)
FBgn0031973	cn3298	66.28
FBgn0028990(*Dm serpin27A*)	cn1692	57.44
FBgn0028990(*Dm serpin27A*)	cn1693	55.67
FBgn0028984 (*Dm serpin5*)	cn6243	50.93
FBgn0028995 (*Dm serpin4*)	cn2217	53.46
FBgn0028995(*Dm serpin4*)	cn2218	50.40
FBgn0028995 (*Dm serpin4*)	cn2216	58.85
FBgn0028995 (*Dm serpin4*)	cn2220	55.64
FBgn0028995 (*Dm serpin4*)	cn2215	58.47
AAEL014141	cn3298	53.85
AAEL014138	cn3298	52.31
*AGAP006911(*Ag serpin2*)	cn1693	43.80
*AGAP003139(*Ag serpin9*)	cn6243	42.25

Ag: Anopheles gambiae

Dm: Drosophila melanogaster

AAE: Aedes Aegypti

* AGAP006911 and AGAP003139 did not meet the 50% identity threshold but they are the most significant hits.

Table 3 : Amino acid overlap between *Drosophila, Anopheles* and *Aedes* (as a query) versus *G. morsitans* ORFs

Insect Serpins	Serpin length (amino acids)	*Query overlap (start-stop)	**Query Coverage (%)	*G. morsitans* contig ID's	*G. morsitans* Length (amino acids)	*G. morsitans* overlap (start-stop)	*G. morsitans* coverage (%)
FBgn0031973	536	447-532	15	cn3298	290	125-209	29
FBgn0028990	447	55-447	87	cn1692	535	129-517	72
FBgn0028990	447	167-447	62	cn1693	326	47-324	85
FBgn0028984	427	1-424	94	cn6243	474	55-471	97
FBgn0028985	392	9-380	94	cn2217	493	95-467	94
FBgn0028985	392	9-383	95	cn2218	474	67-440	78
FBgn0028985	392	266-380	29	cn2215	301	1-115	37
FBgn0028985	392	248-376	32	cn2220	149	1-133	88
FBgn0028985	392	150-357	52	cn2216	208	1-208	99
AAEL014141	465	401-465	13	cn3298	290	147-211	22
AAEL014138	395	329-393	16	cn3298	290	145-209	22
AGAP006911	409	129-409	68	cn1693	326	47-322	67
AGAP003139	447	48-447	89	cn6243	484	89-470	97

*Query corresponds to *D.melanogaster, A.gambiae and A.aegypti* serpin sequences.

** Coverage was calculated as the number of overlapping amino acids divided by the total amino acid length. Notably, cn3298 spans less than 30% of its whole length as well as the corresponding insect serpin queries

Table 4: Characteristic serpin domain residues as predicted by PFAM and SMART databases

G. morsitans	**LENGTH**	**PFAM**	**E VALUE**	**SMART**	**E**
CONTIG ID	**(AMINO ACIDS)**	**RESIDUES**		**RESIDUES**	**VALUE**
cn1692	535	155 – 512	2.2 $^{e\ -65}$	151 – 512	5.89$^{e\ -72}$
cn2217	493	101 – 466	9.1 $^{e\ -121}$	104 – 466	2.91 $^{e\ -126}$
cn3298	290	144-209	5 $^{e\ -13}$	3-209	1.78e^{+01}
cn3298 Transmembrane fragment		5-42	0.00042	238-260	-
cn6234	474	31- 415	1.1 $^{e\ -83}$	47 – 415	1.54 $^{e\ -84}$

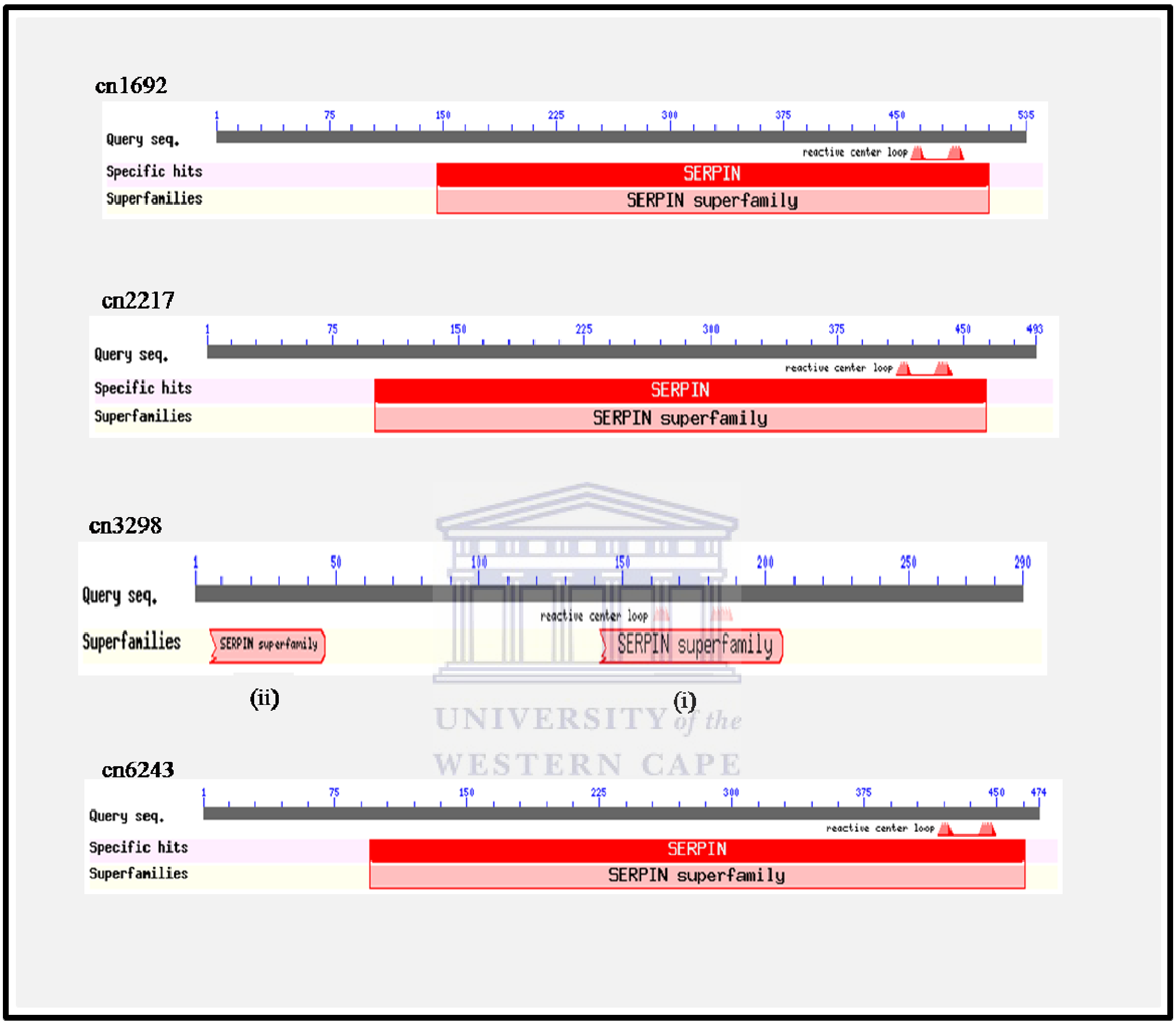


Figure 7: *G.morsitans* serpin domains as predicted by CDD.

Grey bars correspond *to G.morsitans* peptides. The relative positions of the reactive center loops are indicated as peaks above the serpin domain regions (red bars).

3.3 Phylogenetic trees analysis

All the methods employed in phylogenetic reconstruction, that is, Bayesian inference, maximum likelihood and neighbor-joining maintain a consistent tree topology. However, there is slight variation in bootstrap values. Three out of four contigs (cn1692, cn2217, cn6243) are composed of ESTs that originate from a variety of tissues. These contigs form orthologous groups with *Drosophila melanogaster* serpins. On the other hand, cn3298 was sampled from the fat body (chief immune response organ). This contig (cn3298) has no clear 1:1 orthologue in all trees (see Figures 8-10). In all cases the terminal nodes are supported by high bootstrap values. Notably, the neighbor-joining algorithm does not have strong branch support for four internal nodes (Figure 9). Clustering among *D.melanogaster* and *A. gambiae* serpins is concordant with data from previous phylogenetic analysis on insect immunity genes (Christophides *et al.*, 2002; Zou *et al.*, 2009; Tanaka *et al.*, 2008).

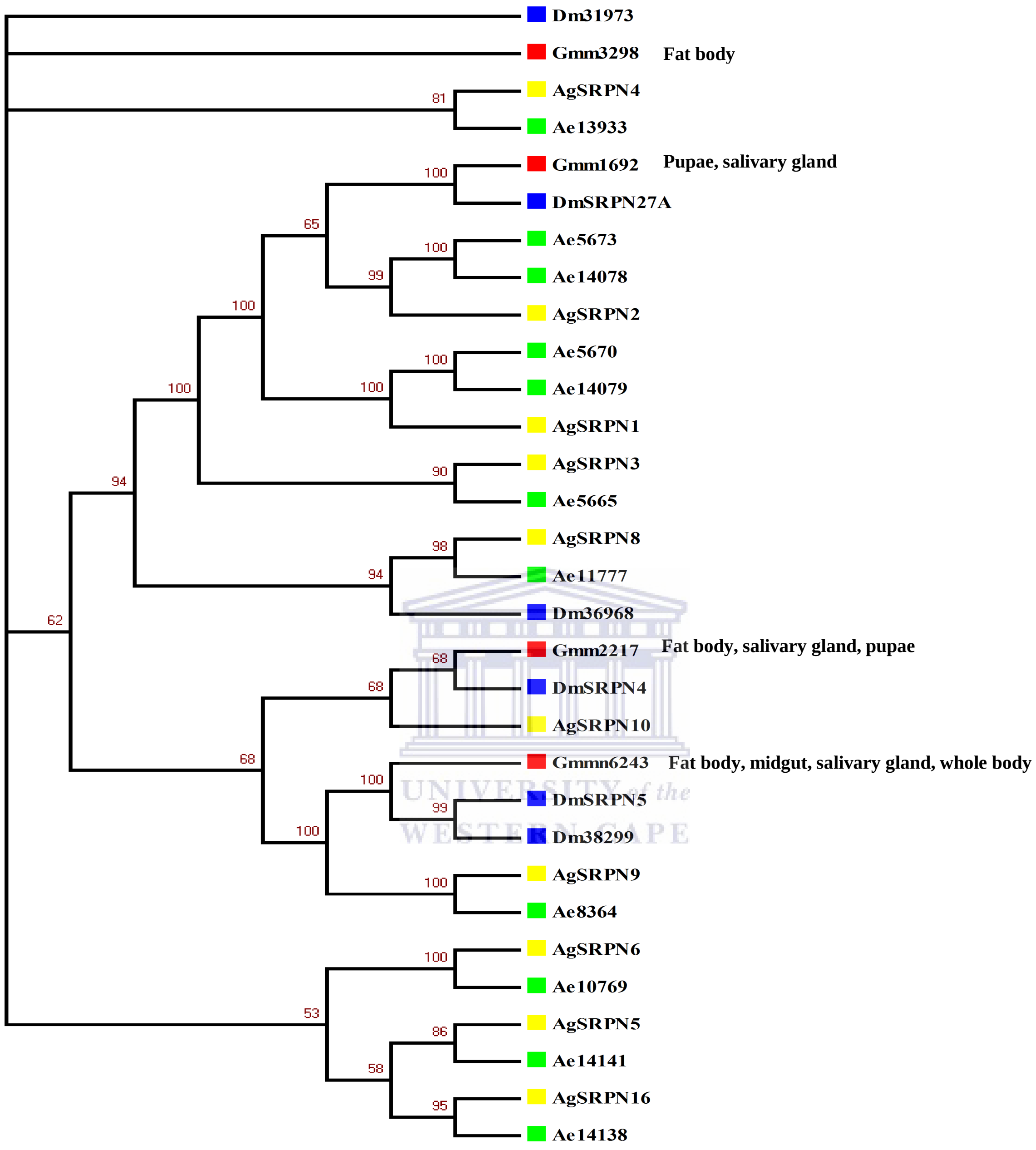


Figure 8: Bayesian inference tree for insect serpins.

The tree was inferred from 27 insect serpin peptide sequences and four *G.morsitans* putative serpins. The tree was generated using MR BAYES v3.1.2 (Huelsenbeck and Ronquist, 2001). Bootstrap values calculated from 1000 resamplings are indicated at the nodes as a percentage. *G.morsitans* contigs (red) are indicated alongside their tissue of origin. Other serpin sequences are indicated as green: *Aedes aegypti*, cyan: *Drosophila melanogaster*, yellow: *Anopheles gambiae*

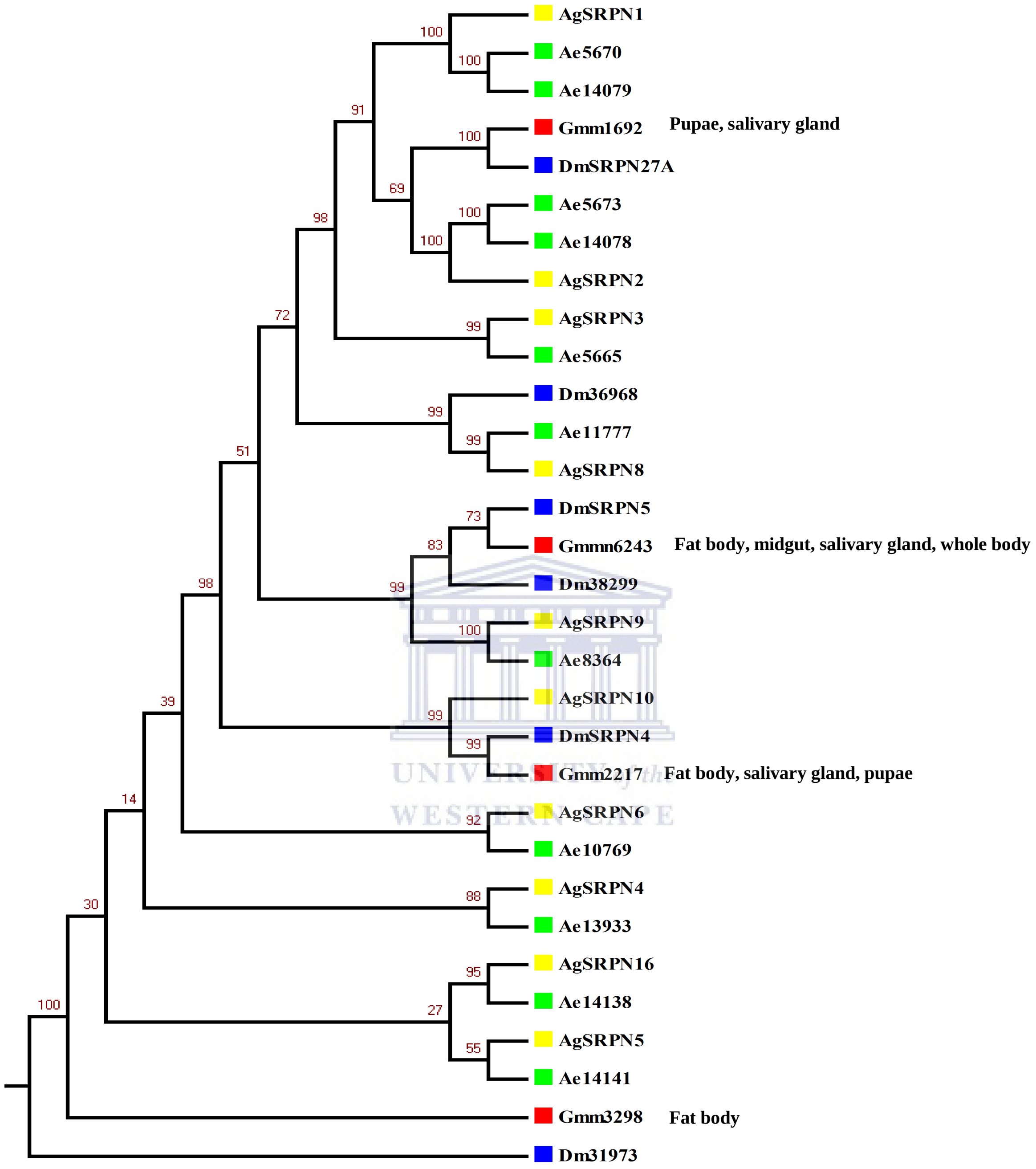


Figure 9: Neighbor -joining tree for insect serpins.

The tree was inferred from 27 insect serpin peptide sequences and four *G.morsitans* putative serpins. The tree was generated using the neighbor-joining algorithm of PHYLIP v 3.68 suite (Felsenstein, 1993). Bootstrap values calculated from 1000 resamplings are indicated at the nodes as a percentage. *G.morsitans* contigs (red) are indicated alongside their tissue of origin. Other serpin sequences are indicated as green: *Aedes aegypti*, cyan: *Drosophila melanogaster*, yellow: *Anopheles gambiae*.

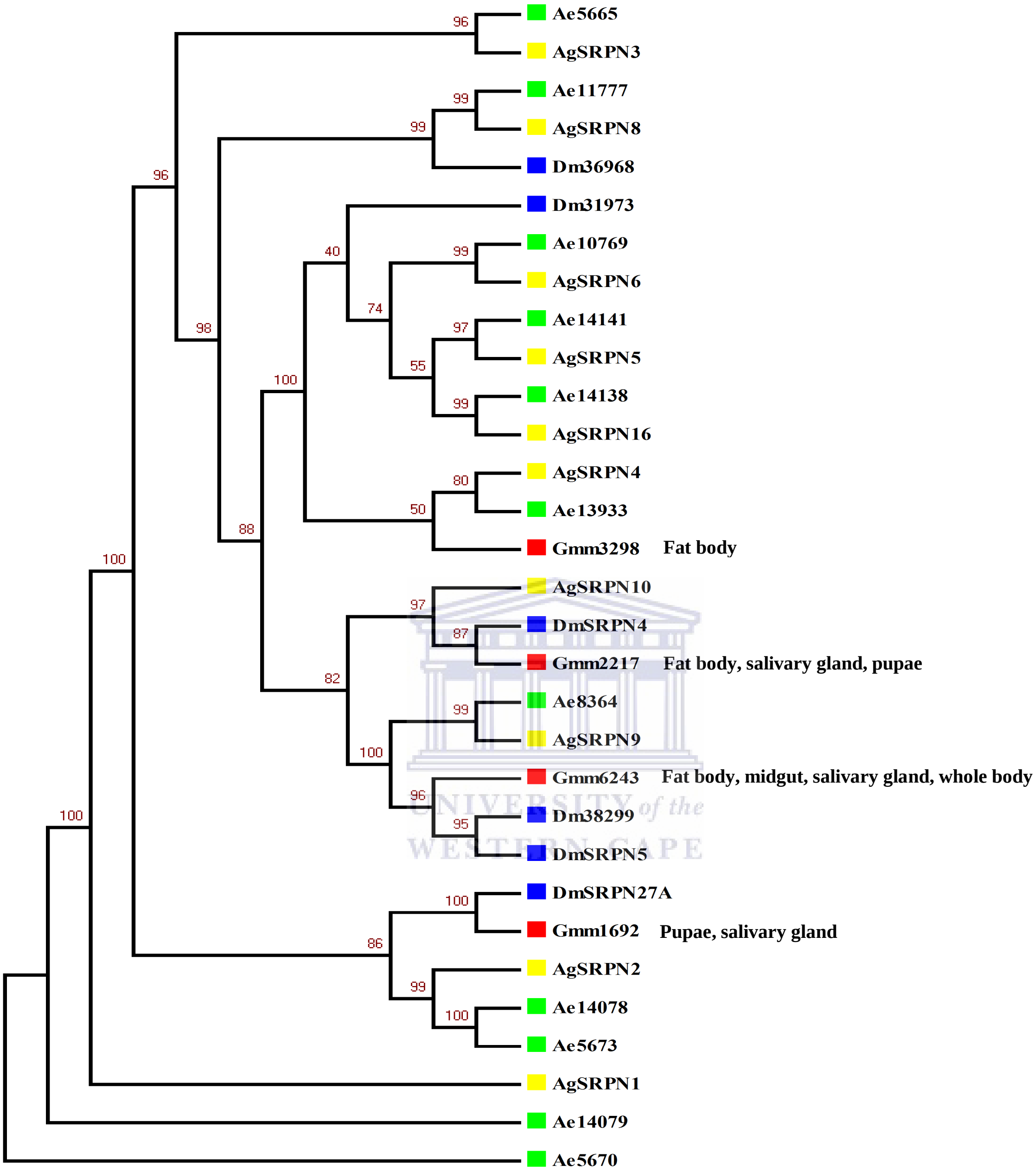


Figure 10: Maximum likelihood tree for insect serpins.

The tree was inferred from 27 insect serpin peptide sequences and four *G. morsitans* putative serpins. The tree was generated using PHYML version 3.0 (Guindon *et al.*, 2005). Bootstrap values calculated from 100 bootstrap resamplings are indicated at the nodes as a percentage. *G.morsitans* contigs (red) are indicated alongside their tissue of origin. Other serpin sequences are indicated as green: *Aedes aegypti*, cyan: *Drosophila melanogaster,* yellow*: Anopheles gambiae*.

3.4 Intron-exon organization

The results obtained from the phylogenetic studies indicated that most of the *G.morsitans* serpins had orthologues with *D.melanogaster* serpins, therefore a TBLASTN was performed against the *D. melanogaster* ENSEMBL database. Serpins from *D.melanogaster* and *G. morsitans* were examined for intron-exon junctions to verify their putative orthologous relationships. In general, conservation of splice sites is observed for cn2217 and cn6243. However, this does not hold entirely true for cn2217 whose orthologue has an additional exon within a large intronic portion of the contig (Figure 11). Contigs 1692 and 3298 fragments span one exon of respective *D.melanogaster* orthologues (Figure 11).

3.5 Homology modeling

We aimed at identifying the structurally conserved regions of the putative *G.morsitans* serpins against a known serpin structure. In particular, the structure of α-1-antitrypsin was employed as the template serpin structure. Elements of secondary structure are conserved for cn6243 indicating the peptide sequence of cn6243 may be a complete representative of its cognate gene (Figure 12). However, homology modeling of cn2217, cn3298 and cn1692 against α-1-antitrypsin resulted in 3D alignments that do not depict a high degree of conservation of secondary structure elements (data not shown). The assumption is that there are critical amino acids missing in the peptide sequences of these contigs thus, optimal spatial restraints cannot be derived from the alignment.

3.6 Identification of the reactive center loop

Residues within the RCL are shown with the assumption that these are true inhibitory serpins and there are no insertions or deletions in this region. The position of the scissile residues (P'_1 P_1) is indicated (see Figure 13). There is a high degree of conservation of residues along the RCL, in particular, the environment flanking both the C and N terminus. The results suggest that cn1692 is indeed an inhibitory serpin since it shares identical P'_1 P_1 residues with *D.melanogaster serpin27A*. The P'_1 P_1 residues are critical since they determine the target specificity of the proteins. *D.melanogaster serpin27A* has been shown to be an inhibitor of the melanization pathway by facilitating site-specific melanization at the infection site. However, this cannot be established for cn6243 whose P_1 residue is a gap. In addition, cn2217 and cn3298 have different

P'_1 P_1 residues (Figure 13), thus, they may not share same target specificity with cn1692.

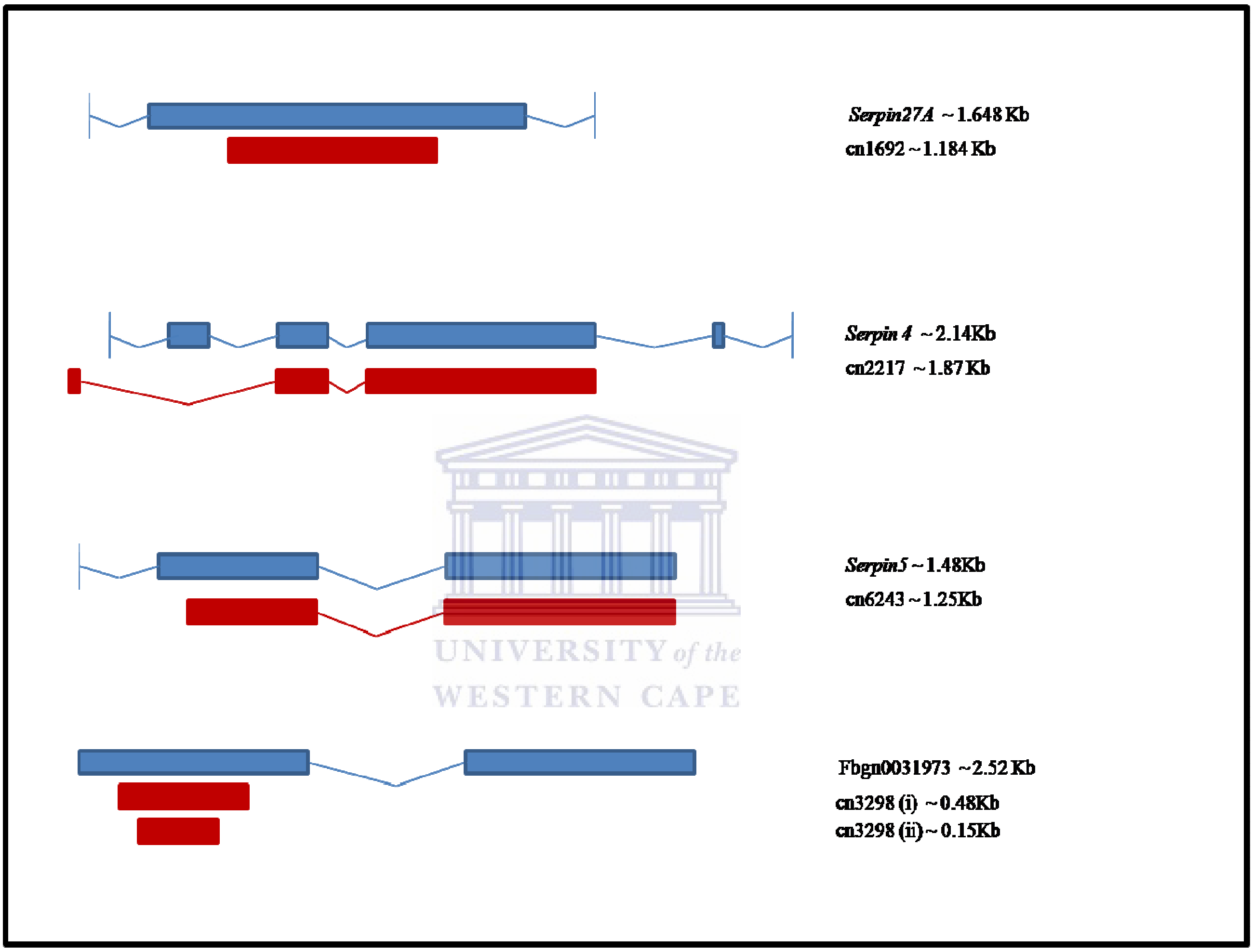


Figure 11: Intron-exon organization of *G.morsitans* contigs aligned to *D.melanogaster* orthologues.

All introns are indicated as lines while the exons are colored blue for *D.melanogaster* and red for *G.morsitans*. The approximate size of the exons is indicated alongside each structure as approximate kilobases (~kb).Contig 3298 has two exons (i) and (ii) which span one exon of its corresponding *D.melanogaster* orthologue.

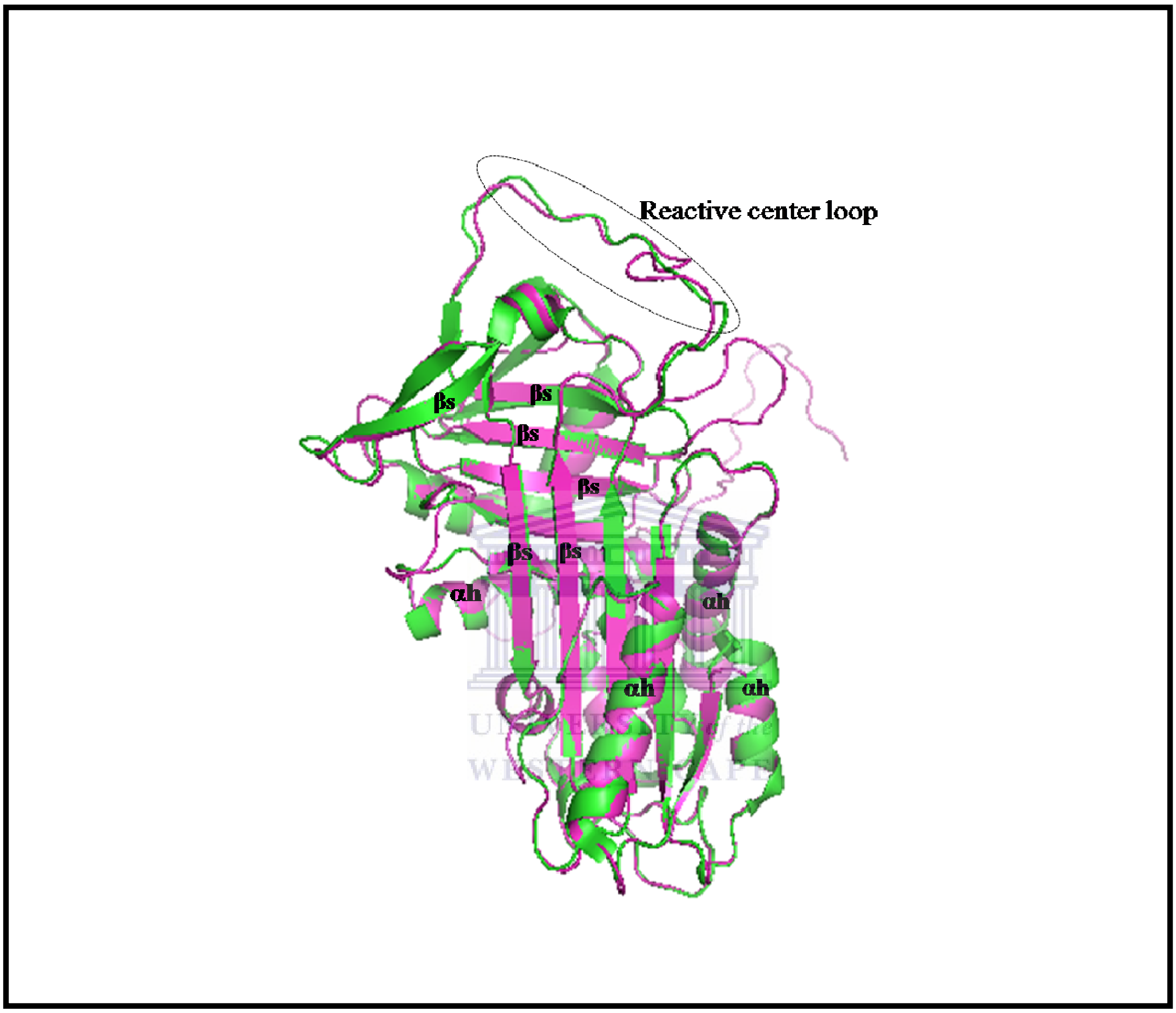
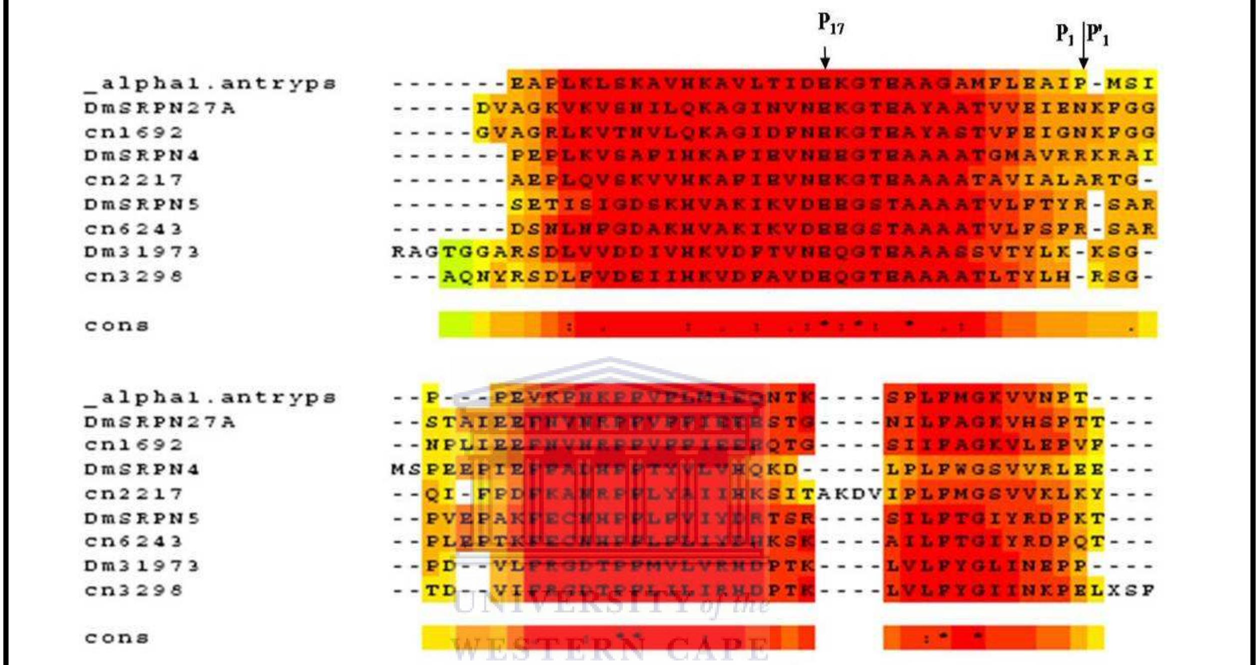


Figure 12: An overlay of cn6243 (magenta) three dimensional model onto α-1-antitrypsin (green) crystal structure (PDB code 2QUG.pdb).

This alignment depicts conservation of most of the secondary structure elements. The α-helices are labeled as (αh) while the β-sheets are labeled as (βs). The region spanning the reactive center loop is also indicated by a broken circle.



```
_alpha1.antryps    -------EAPLKLSKAVHKAVLTIDEKGTEAAGAMFLEAIP-MSI
DmSRPN27A          -----DVAGKVKVSNILQKAGINVNEKGTEAYAATVVEIENKFGG
cn1692             -----GVAGRLKVTNVLQKAGIDFNEKGTEAYASTVFEIGNKFGG
DmSRPN4            -------PEPLKVSAPIHKAPIEVNEEGTEAAAATGMAVRRKRAI
cn2217             -------AEPLQVSKVVHKAPIEVNEKGTEAAAATAVIALARTG-
DmSRPN5            -------SETISIGDSKHVAKIKVDEEGSTAAAATVLFTYR-SAR
cn6243             -------DSNLNFGDAKHVAKIKVDEEGSTAAAATVLFSFR-SAR
Dm31973            RAGTGGARSDLVVDDIVHKVDFTVNEQGTEAAASSVTYLK-KSG-
cn3298             ---AQNYRSDLFVDEIIHKVDFAVDEQGTEAAAATLTYLH-RSG-

cons

_alpha1.antryps    --P---PEVKFNKPFVFLMIEQNTK----SPLFMGKVVNPT----
DmSRPN27A          --STAIEEFNVNKPFVFFIREESTG----NILFAGKVHSPTT---
cn1692             --NPLIEEFNVNKPFVFFIEEKQTG----SIIFAGKVLEPVF---
DmSRPN4            MSPEEPIEFFADHPFTYVLVHQKD-----LPLFWGSVVRLEE---
cn2217             --QI-FPDFKANRPFLYAIIHKSITAKDVIPLFMGSVVKLKY---
DmSRPN5            --PVEPAKFECNHPFLFVIYDRTSR----SILFTGIYRDPKT---
cn6243             --PLEPTKFECNHPFLFLIYDHKSK----AILFTGIYRDPQT---
Dm31973            --PD--VLFRGDTPFMVLVRHDPTK----LVLFYGLINEPP----
cn3298             --TD--VIFRGDTPFLLLIRHDPTK----LVLFYGIINKPELXSF

cons
```

Figure 13: An Alignment of amino acids spanning the reactive center loop.

Putative *G.morsitans* serpins were aligned against corresponding *D.melanogaster* orthologues and human α-1-antitrypsin. This alignment was generated by TCOFFEE (Notredame *et al.*, 2000). The start of the reactive center loop consensus pattern is indicated at P_{17}. Degree of conservation is represented by the color schemes as:

BAD AVG GOOD

3.7 Assembly of *Tsal* encoded ESTs

Tsal1, *Tsal2a* and *2b* were assembled with 3564 *Tsal* ESTs as described in the methods (see chapter 2 section.2.10). Twenty-nine contigs and two singlets were generated by the PHRAP assembly (Table 5). The two singlets do not seem to represent novel proteins because, (i) quality scores of the first singlet read (GMsg139h02.plkw) are relatively low with almost half of the sequence being converted to N's, a possible explanation as to why it does not align to any other EST and (ii) After conceptual translation, the second singlet (GMsg14c05.qlkSCF) gives a short peptide (64 amino acids) which does not give a significant hit to NCBI's non-redundant protein database. Additionally, no domains or motifs could be predicted with confidence using SMART and PFAM databases. Six contigs contained one read while the largest contig (contig29) was made up of 1228 reads including the *Tsal1* isoform (Table 5). *Tsal2* isoforms do not assemble with any other reads. They are the constituents of contig7. Contig29 has the widest tissue distribution. Most reverse complements assemble within the same contig while there is a high rate of overlaps observed for those contigs with greater than 100 reads. Quality scores indeed increase the stringency of the assembly. For example, an assembly of all ESTs without provision of the quality scores file generates 615 contigs while inclusion of quality scores in the assembly process reduces the number of contigs to 29. Notably, the same singletons are retained with and without the provision of quality scores.

3.8 Functional annotation of *Tsal* ORFs

Open reading frames that are less than 99 amino acids are rarely annotated as proteins unless they show significant homology to known proteins from other species or members of gene families (reviewed by Reich, 2000). Four *Tsal* contigs were predicted to be less than 100 amino acids (see Table 6). Twenty out of twenty-nine contigs gave the nuclease domain as their conserved domain (see Figure 14 for a representative). Homology searches of the conceptually translated ORFs against NCBI's non-redundant protein database identified *Tsal1* and *Tsal2* as the most significant hits.

Table 5 : Tissue distribution of putative *Tsal* ESTs

PHRAP GENERATED CONTIG IDs	LENGTH (BP)	NUMBER OF READS								
		SG	MG	FB	PUM	PUF	TUF	**Tsal1*	**Tsal2*	TOTAL
1	1091	1								1
2	1161	1								1
3	1396	1								1
4	892	1								1
5	1198	1								1
6	699		1							1
7	1263								2	2
8	2197	2								2
9	886	2								3
10	2016	4								4
11	1202	4								4
12	1633	4	1							5
13	1463		3		2					5
14	1443	11								11
15	1436	13								13
16	1537	16								16
17	1396	17								17
18	1477	20								20
19	1516	22								22
20	1542	27								27
21	1576	32								32
22	1521	35								35
23	1584	55								55
24	1680	105	1							106
25	1755	115	1							116
26	2821	366	2							368
27	2715	452	1	1						454
28	2807	1011	4	1						1016
29	2495	1216	5	2	2	1	1	1		1228

The average contig length is 1599 base pairs. While the bulk of reads originate from the salivary glands, contig 29 has the widest tissue distribution.

SG – Salivary gland; PUM – Male pupae; MG – Midgut;

PUF – Female pupae; TUF – Combined tissue data; FB – Fat body;

**Tsal1* and *Tsal2* cDNA sequences were downloaded from GENBANK.

Table 6 : Conserved domains of *Tsal* ORFs

Contig 1D	Length (amino acids)	CDD	PFAM	SMART
1	117	Lac Z	Dihydropiridine Sensitive L type calcium channel	x
2	123	x	Shikimate quinate dehydrogenase	x
3	126	x	X	x
4	215	x	X	x
5	108	x	X	x
6	54	x	X	x
7	388	Nuclease	Endonuclease	Endonuclease
8	14	x	X	x
9	54	x	X	x
10	35	x	X	x
11	164	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
12	164	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
13	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
14	418	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
15	359	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
16	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
17	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
18	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
19	164	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
20	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
21	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
22	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
23	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
24	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
25	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
26	164	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
27	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
28	356	Nuclease	DNA/RNA non specific endonuclease	Endonuclease
29	399	Nuclease	DNA/RNA non specific endonuclease	Endonuclease

Columns indicated by ‘x’ denote that no domains or motifs could be predicted with confidence.

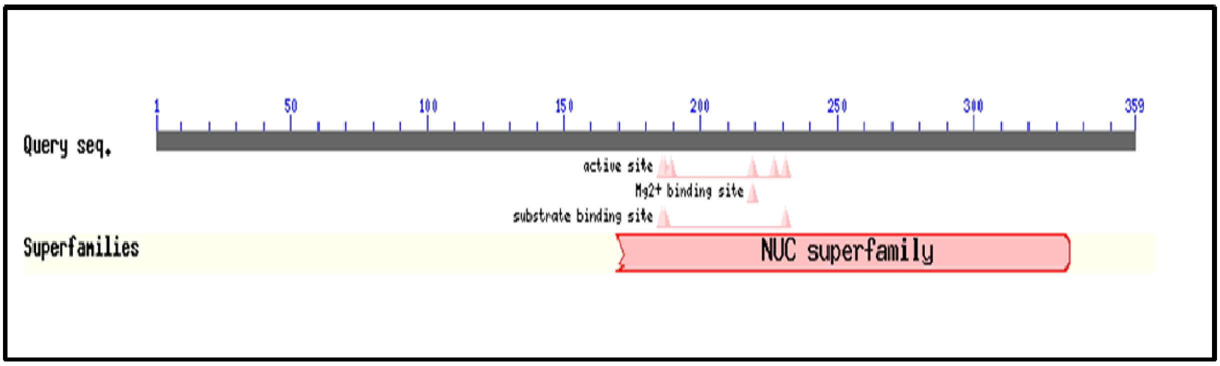


Figure 14: Graphical representation of the conserved domain for *G. morsitans* PHRAP generated contig29 putative *Tsal*.

Contig29 was generated as a result of PHRAP assembly of *Tsal* ESTs. A BLASTP search was carried out with contig29 ORF against NCBI non-redundant protein database. The grey bar corresponds to *G.morsitans* contig29 ORF. The nuclease superfamily conserved domain is depicted as the red bar. The relative positions of the substrate binding sites, Mg^{2+} binding site and the active site are indicated as peaks above the nuclease domain region.

3.9 Polymorphism in *Tsal* ESTs

There was a high degree of variation observed within the reads in all PHRAP generated contigs. A predicted SNP is deemed valid only if the corresponding quality score on the base is satisfactory (Marth *et al.*, 1999). Additionally, a SNP is confirmed if its neighboring nucleotides are conserved over an interval. For example, Hayes and colleagues (2007) have proposed an interval of at least 50 bases. Single base mismatches were identified on contig12 using both a customized Perl script and the BEAP viewer (Figure 15). In addition, the corresponding quality scores of the putative SNPs were verified. Most of the variation observed was in the read GmSg121a03.plk with two instances of substitution of A to T at consensus positions 362 and 363. In addition, two instances of substitution from T to A occur at positions 368 and 445. Substitution for position 391 is T to G, while G is substituted with T in position 411(Table 7).

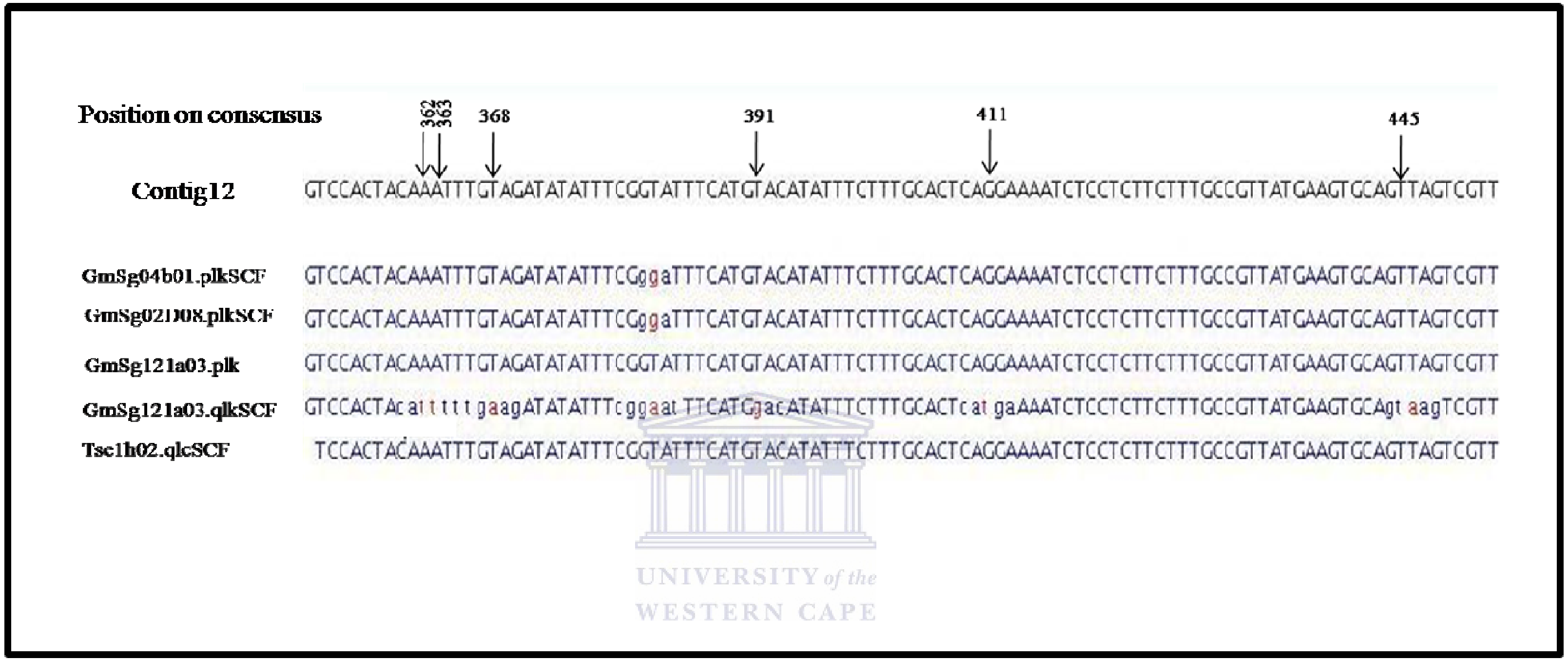

Position on conscnsus
362
363
368
391
411
445
Contig12
GTCCACTACAAATTTGTAGATATATTTCGGTATTTCATGTACATATTTCTTTGCACTCAGGAAAATCTCCTCTTCTTTGCCGTTATGAAGTGCAGTTAGTCGTT
GmSg04b01.plkSCF
GTCCACTACAAATTTGTAGATATATTTCGggaTTTCATGTACATATTTCTTTGCACTCAGGAAAATCTCCTCTTCTTTGCCGTTATGAAGTGCAGTTAGTCGTT
GmSg02D08.plkSCF
GTCCACTACAAATTTGTAGATATATTTCGggaTTTCATGTACATATTTCTTTGCACTCAGGAAAATCTCCTCTTCTTTGCCGTTATGAAGTGCAGTTAGTCGTT
GmSg121a03.plk
GTCCACTACAAATTTGTAGATATATTTCGGTATTTCATGTACATATTTCTTTGCACTCAGGAAAATCTCCTCTTCTTTGCCGTTATGAAGTGCAGTTAGTCGTT
GmSg121a03.qlkSCF
GTCCACTAcattttt gaagATATATTTcggaat TTCATGgacATATTTCTTTGCACTcat gaAAATCTCCTCTTCTTTGCCGTTATGAAGTGCAgt aagTCGTT
Tsc1h02.qlcSCF
TCCACTACAAATTTGTAGATATATTTCGGTATTTCATGTACATATTTCTTTGCACTCAGGAAAATCTCCTCTTCTTTGCCGTTATGAAGTGCAGTTAGTCGTT

Figure 15: Variation in the contig12 assembly.

This Figure represents consensus sequence of *Tsal* contig12 together with its constituent reads. The color differentiation option was selected to highlight putative SNPs. In this particular view putative SNPs are highlighted in red. Four reads denoted by the prefix "GmSg" originate from salivary gland are while one read (Tse1h02.qlcSCF) originates from the midgut. Most of the variation is observed in GmSg121a03.qlkSCF (see Table 7).

Table 7 : Putative SNPs in GmSg121a03.plk

Base position on consensus	Type of substitution	Quality score
362	A/T	12
363	A/T	15
368	T/A	10
391	T/G	10
411	G/T	15
445	T/A	12

CHAPTER FOUR: DISCUSSION AND CONCLUSION

4.1 Discussion …… 57

4.1.1 Evolutionary relationships between insect serpins and *G.morsitans* putative serpins.. 57

4.1.2 Reactive center loops …… 60

4.1.3 Sequence variation in *Tsal* EST data …… 61

4.2 Conclusion and direction for future work …… 63

4 DISCUSSION AND CONCLUSION

4.1 Discussion

4.1.1 Evolutionary relationships between insect serpins and *G.morsitans* putative serpins

Orthologous relationships established in this study are mainly with *Drosophila melanogaster* sequences. This is an observation that is shared with studies that have carried out functional annotation and/or phylogenetic analysis of sequences from insects such as *Anopheles gambiae, Manduca sexta, Bombyx mori* and *Glossina morsitans* (Christophides *et al.,* 2002; Zou *et al.*, 2009; Tanaka *et al.*, 2008; Lehane *et al.*, 2007). Both the genomes of *Anopheles gambiae* and *Drosophila melanogaster* have been sequenced and annotated (Holt et al., 2002; Celinker and Rubin 2003). It is possible that these genes have been deleted from the *Anopheles gambiae* genome during the evolutionary process.

A notable observation is the absence of a clear *G.morsitans* orthologue for *D.melanogaster serpin43Ac*, an extensively studied serpin. Despite extensive BLAST searches, a significant hit could not be obtained (data not shown). In their studies of the immune genes of *Anopheles gambiae* and *Bombyx mori*, Christophides *et al*., (2002) and Tanaka *et al.*, (2008) did not identify an orthologue of *D.melanogaster serpin43Ac* for serpins from *Anopheles gambiae* and *Bombyx mori* respectively. *D.melanogaster serpin43Ac* (*nec*) has been implicated in the control of the toll mediated antifungal IMD pathway (Levashina *et al.,* 1999) by acting as an inhibitor of the clip domain serine protease, Persephone. The inference that can be drawn from this observation is that perhaps unlike *Drosophila melanogaster*, members of *Bombyx, Anopheline and Glossina* genera may not be overly challenged by fungal infections. Thus in their evolutionary processes, the functions of Persephone and *nec* may have diverged to serve other roles or cope with other regular challenges. Since genome sequencing is in progress, it is possible that the sequence of this gene is not yet available. Interestingly, *D.melanogaster serpin43Ac* shows weak sequence similarity with cn2215 (one of the sequences that was filtered with the presumption that it was a paralogue of cn2217). The putative orthologues of cn2217 are *A.gambiae serpin10* and *D.melanogaster serpin4*. In the event that it is established that cn2217 and cn2215 are indeed true paralogues and their functions are elucidated, the conclusion would be that the role of

D.melanogaster serpin43Ac diverged to perform the function of cn2215.

Different phylogenetic approaches confirm the clustering of cn2217 clusters with *A.gambiae serpin10* and *D.melanogaster serpin4*. Danielli and colleagues (2003) have established that alternative splicing of *A.gambiae serpin10* gives rise to four isoforms. In the same study, biochemical characterization of the inhibitory potential of three recombinant *serpin10* genes show that the isoforms are expressed in tissues involved in insect defense specifically, haemocytes and midgut epithelium. At least two isoforms are transcriptionally up-regulated during parasite establishment within the midgut potentially implicating them in antiparasitic action and/or parasite tolerance. *D.melanogaster serpin4* has been reported among the extensive arrays of genes that are upregulated in response to microbial challenge (Irving *et al.,* 2001). *D.melanogaster serpin4* gene encodes at least two different serpin proteins generated by alternative splicing of the last coding exon. This serpin was also shown to play a role in regulation of peptide maturation in *Drosophila* (Osterwalda *et al*., 2004). Interestingly, the spliced transcripts *Ag serpin10* and *Dm serpin4* segregate with contig 2217. It is likely that cn2217 also represents an alternative splice variant because of the other *Glossina* contigs (cn2215, cn2216, cn2218 and cn2220) that shared the same parent cluster with cn2217 (see section 2.3). Prediction and confirmation of alternative splice variants awaits the sequenced *Glossina* genome. It would be essential to establish the exact role that these homologues play in facilitating immune defenses in order to explain whether there has been pressure for diversification in order to deal with new challenges thus evolution by duplication followed by divergence, to produce a diverse set of paralogues.

Contig 6243 forms an orthologous group with two *Drosophila melanogaster* serpins deemed to be paralogues according to novel ENSEMBL predictions, *D.melanogaster serpin5* and *FBgn0038299(Dm38299)*. The topology of the clade containing cn6243 suggests that a second copy exists for *Glossina* or alternatively, that the second copy has been deleted over time. These paralogues would have arisen before tsetse flies and *Drosophila* species diverged. Besides facilitating wing expansion during early *Drosophila* development (Charron *et al*., 2008), *D.melanogaster serpin5* has been shown to be upregulated in response to immune stimulation by oligonucleotide DNA microarrays (Irving *et al.,* 2001). Within the same cluster is *A.gambiae* *serpin9*. Currently there is no data on the exact physiological roles of genes in this cluster namely, *D.melanogaster serpin5*, *FBgn0038299 (Dm38299)* and *A.gambiae serpin9*. Homology modeling of the 3D alignment of α-1- antitrypsin and cn6243 depicts conservation of elements of secondary structure including residues at the reactive center loop. Comparative modeling is important in establishing evolutionary relationships as it enables the detection of sequence similarity between the target and the template. The 3D structure of proteins from the same family is more conserved than the amino acid sequences. Presence of most of the elements of secondary structure including the reactive center loop, show that indeed cn6243 may be a complete representative of the cognate gene. Thus, this serpin presents an excellent candidate for biochemical and genetic experiments in an effort to understand its exact role.

G.morsitans cn1692 forms an orthologous group with *D.melanogaster serpin27A* and *A.gambiae serpin2*. *D.melanogaster serpin27A* and *A.gambiae serpin2* have been implicated in regulation of the melanization cascade (Michel *et al.*, 2005) by facilitating the site-specific localization of the ensuing melanotic response during the insect's immune response. The target specificity of cn1692 is probably the same as that of *D.melanogaster serpin27A* as they both possess identical scissile bond residues (see Figure 13).

A search of the serpin domain in cn3298 revealed that this domain is represented by two fragments (see Figure 7). Domain-(ii) exhibits extensive coiling thus it is predicted to be a transmembrane portion. A BLAST search against *Drosophila melanogaster* proteins on ENSEMBL identified FBgn0031973 as the best hit. Interestingly, cn3298 gives the highest percent identity against its corresponding insect serpin queries whereas the percentage coverage is poor (see Table 3). Thus, a high percent identity is not always an indicator of sufficient sequence coverage due to divergence across a gene with localized conserved regions. The fragmentation of cn3298 domain into two portions may account for poor amino acids overlap in the alignment. The amino acid sequences flanking the reactive center loop of cn3298 and selected serpins (see Figure 13) are conserved, indicating that it may indeed be an inhibitory serpin. The congruence of the phylogenetic tree reconstruction is perturbed by cn3298 to some degree, a fact that could be attributed to poor amino acid overlap in the alignments.

4.1.2 Reactive center loops

The sequence and conformation of the RCL largely determines the selectivity of inhibition, thus sequence alignments of the serpin C terminus comprising the RCL are particularly revealing (see Figure 13). The RCL has received much attention during the study of serpin mechanism of action for the reason that it holds the critical amino acids considered to control much but not all of the inhibitory specificity of the serpins. In this study, we were able to establish a high degree of conservation in the amino acid environment flanking the reactive center loop, indicating that these serpins may indeed be inhibitory. In addition, we were able to determine the position of the scissile bond (P'_1 P_1). The scissile bond residues are identical for cn1692 and its orthologue *D.melanogaster serpin27A.* Therefore, cn1692 may have a regulatory role in *G. morsitans* defense responses, similar to *D.melanogaster serpin27A* because of common target specificity. We could not establish the P_1 residue of the scissile bond for the cn6243. Albeit the high degree of conservation in the region spanning the RCL the scissile bonds of cn2217 and cn3298 are not identical, a strong indication that they may not have the same target specificity. However, this cannot be used to rule out the possibility that these contigs are indeed inhibitory serpins. In addition, this observation can be attributed to the evolutionary process of these genes that may have been accompanied by mutations that led to changes in the amino acid composition. The high degree of conservation of the RCL is an indicator of evolutionary conservation of serpin genes that participate in the signal transduction pathways. Genes in the signal transduction pathways have remained highly conserved throughout the different insect orders (Tanaka *et al.*, 2008). Conserved gene evolution of insect signal transduction repertoires reflects the essential requirement of these genes for common immune mechanisms against infectious microorganisms.

4.1.3 Sequence variation in *Tsal* EST data

Tissue distribution of ESTs shows that *Tsal* transcripts are not only expressed in the salivary gland but in other tissues, for example, the fat body and midgut (Table 5). While Li and colleagues (2001) had established that *Tsal1* expression is limited to adult and larval stages, the results of the assembly show that *Tsal1* assembles with three transcripts that originated from the pupae. The assembly of *Tsal1* with the highest number of transcripts with a varied tissue distribution indicates the possibility of multiple isoforms for *Tsal* that may not have been characterized.

Homology searches against NCBI non-redundant protein database identified *Anopheles gambiae* and *Culex quinquefasciatus* endonucleases as the next most significant hits albeit low percentage identity. This is consistent with the study performed by Calvo and Ribeiro (2006) on the sialotranscriptome of *Culex quinquefasciatus* whereby *Tsal1* and *Tsal2* exhibit similarities with an abundant cluster of sequences coding for secreted proteins with endonuclease activity. Sixty nine percent of the contigs generated by PHRAP assembly of putative *Tsal* ESTs are predicted to be nucleases more specifically endonucleases, by SMART. PFAM denotes them to be either DNA or RNA endonucleases. However, contigs in which less than five ESTs assembled (excluding contig 7) are predicted to encode other conserved domains, for example, Dihydropiridine and Shikimate quinate dehydrogenase (see table 6). The underlying sequences in these contigs do not share a high level of similarity with the rest of the ESTs in the cluster thus their assembly as singletons and possible possession of variant conserved domains.

We observe a high degree of variation in the contigs. The high degree of variation coupled with low quality scores for contig twelve’s single base pair mismatches hampered the identification of specific SNPs. Accurate identification of *Tsal* SNPs requires resequencing of salivary gland ESTs to obtain accurate base calls. In addition, *Tsa1l, Tsal2a* and *Tsal2b* do not assemble as one contig indicating that their underlying sequences of potential alternative splice variants do not have a high degree of similarity. Variation in the *Tsal* cluster is concordant with Lanzaro and colleagues (1999) who have proposed that salivary gland proteins of blood-sucking dipterans may have unusual degree of polymorphism because the insect may benefit from antigenic variation of its salivary proteins. Our findings could be used in advancing the understanding of tsetse salivary gland protein evolution.

In their study, Ghosh and Mukopadhyay (1998) demonstrate that natural or host anti-sandfly salivary gland antigen immunization reduces sand fly’s blood feeding efficiency and increases mortality in the post bloodmeal vector population. They attribute this to the fact that anti-sandfly saliva immune sera probably bind with the respective antigen-presenting sites of the sandfly salivary gland components causing the sandfly death. A similar study on tsetse flies by Caljon *et al.,* (2006) demonstrates that, immunization with recombinant *Tsal1* (*rTsal1*) does not lead to increased tsetse mortality. The implication of this is that perhaps the *Tsal* protein does not possess antigen binding sites.

The majority (69%) of all conceptually translated putative *Tsal* ORFs are predicted to encode the nuclease domain thus it is our interest to discover the role of nucleases in organisms specifically in dipterans. In *Drosophila melanogaster*, DNase II deficiency has been shown to impair innate immunity function by reducing total haemocyte numbers (Seong *et al.*, 2006). In humans, DNase II was shown to have a role in immunity protection where genetic information may be transferred from apoptotic cells to living cells after being phagocytosed, a potential threat to the genome (Bergsmedh *et al.*, 2006). We propose that *Tsal* may have an immune role (indeed there are transcripts from the fat body which is the major immune response organ) in the tsetse fly and that probably its deposition causes degradation of the apoptotic cells of the human host cell to prevent integration of host DNA into the tsetse fly. This may be an acquired function in the evolutionary process of the tsetse fly to protect itself and it may be linked with the vector competence of the tsetse flies. Apoptosis is usually accompanied by chromosomal DNA degradation and is important for homeostasis of metazoans (Liu *et al.*, 2008). The existence of multiple isoforms may be a mechanism to deal with a wide array of hosts for the tsetse fly.

4.2 Conclusion and direction for future work

With members of the serpin superfamily present over all major taxonomic groups, structural diversification of serpins has occurred to fulfill their roles in myriad of biological processes. Thus, several studies have focused on serpin evolution in all the taxonomic groups. In this study, we focused on a comparative genomics approach to identify putative serpins orthologues from unpublished data of the *Glossina morsitans* transcriptome. By use of data mining and phylogenetic reconstruction, we have gained insight into possible biological functions of putative serpins. However, all these data remains to be confirmed by genetic and biochemical experiments, a process that will be successfully carried out once we have a reference genome.

Ideally, one would want to look at how the immune repertoires of the parasites and *Glossina morsitans* have coevolved to establish whether the evolutionary process has facilitated vector refractoriness. For example, mapping of the contigs onto chromosomes and subsequent identification of gene architectures to provide an insight into the intron-exon organization as well as relative positions of the gene regulatory machinery. This data provides an impetus into structural and functional studies of this family of proteins with a specific focus on the putative paralogues. Paralogues are very important in studying the evolutionary process because they derive from gene family expansions and gene losses, or cases of exceptionally high sequence divergence where they acquire other functions. In the event that these paralogues have evolved to perform different functions then multiple regulatory elements present in the promoter region will need to be dissected.

Studies on the evolutionary patterns of the immune repertoire of *Drosophila melanogaster* and disease vector mosquitoes has established that signal modulation uses a vast reservoir of serine proteases and their inhibitors. Most of these genes have evolved species-specific functions with gene duplication and expansions among certain genes. Complete *Glossina morsitans* serpin sequences would provide insights in the context of a wider array of members of insect serpins in terms of their evolutionary patterns. In addition complete gene architectures will help identify whether these serpins harbor signal peptides either at the N or C terminus to give insight into cellular localization of the proteins.

Identification of the exact function of *Tsal*, for example, by RNA*i* will be particularly informative because it is unique to *Glossina* thus it may offer some insight into the exclusive pathway of *Glossina* evolution. Probably, *Tsal* genes may have diversified to gain other functions that may not be present in other haematophagous insects. This may be an evolutionary process that helps *Glossina morsitans* carve its niche.

REFERENCES

Aksoy S., Berriman M., Hall N., Hattori M., Hide W., Lehane M.J., (2005). A case for a *Glossina* genome project. *Trends Parasitology*. **21(3):**107-11.

Aksoy S., Maudlin I, Dale C., Robinson A.S., O'Neill S.L., (2001). Prospects for control of African trypanosomiasis by tsetse vector manipulation. *Trends in Parasitology.* **17(1):**29-35.

Allsopp R. (2001). Options for vector control against trypanosomiasis in Africa. *Trends in Parasitology*. **17:**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.

Bergsmedh A., Szeles A., Henriksson M., Bratt A., Folkman M.J., Spetz A.L., Holmgren L . (2001). Horizontal transfer of oncogenes by uptake of apoptotic bodies. *Proc Natl Acad Sci U S A*. **98(11):**6407-11.

Berriman M., Ghedin E., Hertz-Fowler C., Blandin G., Renauld H., Bartholomeu D.C., Lennard N.J., Caler E., Hamlin N.E., Haas B *et al.* (2005). The genome of the African trypanosome *Trypanosoma brucei*. *Science*. **309 (5733):** 416-22.

Bonaldo M.F., Lennon G., Soares M.B. (1996). Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res***. 6(9):**791-806.

Boulanger N., Brun R., Ehret-Sabatier L., Kunz C., Bulet P. (2002). Immunopeptides in the defense reactions of *Glossina morsitans* to bacterial and *Trypanosoma brucei brucei* infections. *Insect Biochem Mol Biol*. **32(4):**369-75.

Burke J., Davison D., Hide W. (1999). d2_cluster: a validated method for clustering EST and full-length cDNA sequences. *Genome Res.* **9(11):**1135-42.

Caljon G., Van Den Abbeele J., Sternberg J.M., Coosemans M., De Baetselier P., Magez S. (2006). Tsetse fly saliva biases the immune response to Th2 and induces anti-vector antibodies that are a useful tool for exposure assessment. *Int J Parasitology.* **36(9):**1025-35.

Calvo E., Ribeiro J.M. (2006). A novel secreted endonuclease from Culex quinquefasciatus salivary glands. *J Exp Biol.* **209:**2651-9.

Cecchi G., Mattioli R.C., Slingenbergh J., de la Rocque S. (2008). Land cover and tsetse fly distributions in sub-Saharan Africa. *Med Vet Entomol*. **22(4):**364-73.

Celniker S.E., Rubin G.M. (2003). The *Drosophila melanogaster* genome. *Annu Rev Genomics Hum Genet.* **4:**89-117.

Charron Y., Madani R., Combepine C., Gajdosik V., Hwu Y., Margaritondo G'., Vassalli J.D.(2008). The serpin Spn5 is essential for wing expansion in *Drosophila melanogaster*. *Int J Dev Biol* **7:** 933- 942.

Cheng Q., Aksoy S., (1999). Tissue tropism, transmission and expression of foreign genes in vivo in midgut symbionts of tsetse flies. *Insect Molecular Biology.* **8(1):**125-132.

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):**234-8.

Christophides G.K., Zdobnov E., Barillas-Mury C., Birney E, Blandin S., Blass C., Brey P.T., Collins F.H., Danielli A., Dimopoulos G.*et al.* (2002). Immunity-related genes and gene families in *Anopheles gambiae*. *Science.* **298:**159-165.

Clamp M., Cuff J., Searle S.M., Barton G.J., (2004). The Jalview Java alignment editor. *Bioinformatics*. **20(3):**426-7.

Clark M.D., Hennig S., Herwig R., Clifton S.W., Marra M.A., Lehrach H., Johnson S.L. (2001). An Oligonucleotide Fingerprint Normalized and Expressed Sequence Tag Characterized Zebrafish cDNA Library *Genome Res*. **11:**1594-1602.

Clarke E.P., Cates G.A., Ball E.H., Sanwal B.D., (1991). A collagen-binding protein in the endoplasmic reticulum of myoblasts exhibits relationship with serine protease inhibitors. *J Biol Chem*. **266(26):**17230-5.

Cooper D.N., Smith B.A., Cooke H.J., Niemann S., Schmidt J. (1985). An estimate of unique DNA sequence heterozygosity in the human genome. *Hum Genet.* **69(3):**201-5.

Danielli A., Kafatos F.C., Loukeris T.G. (2003). Cloning and characterization of four *Anopheles* gambiae serpin isoforms, differentially induced in the midgut by Plasmodium berghei invasion. *J Biol Chem*. **278(6):**4184-93.

De Gregorio E., Han S.J., Lee W.J., Baek M.J., Osaki T., Kawabata S., Lee B.L., Iwanaga S., Lemaitre B., Brey P.T. (2002). An immune-responsive Serpin regulates the melanization cascade in *Drosophila*. *Dev Cell*. **4:**581-92.

DeLano W.L. (2002). The PyMOL Molecular Graphics System on the World Wide Web (http://www.pymol.org).

Dumas M and Bouteille B. (1996). Human African trypanosomiasis. *C R Seances Soc Biol Fil*. **190 (4):**395-408.

El-Sayed N.M., Hegde P., Quackenbush J., Melville S.E., Donelson J.E. (2000). The African trypanosome genome. *Int J Parasitol*. **30(4):**329-45.

Ellis J.A., Shapiro S.Z., ole Moi-Yoi O., Moloo S.K., (1986). Lesions and saliva-specific antibody responses in rabbits with immediate and delayed hypersensitivity reactions to the bites of *Glossina morsitans centralis*. *Vet Pathol.* **23(6):**661-7.

Epidemiology and control of African trypanosomiasis. Report of a WHO committee. Geneva Switzerland: World Health Organization Technical report series, 1986:739.

Ewing B., Green P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome.* **8(3):**186-94.

Ewing B., Hillier L., Wendl M.C., Green P. (1998). Base-calling of automated sequencer traces using PHRED. I. Accuracy assessment. *Genome Res*. **8(3):**175-85.

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

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

Filipa F., Jorge C., Andreia F., Nicolas N., Pedro N., Agustin B., João P., Maria O., Jorge S., Jorge A., Sónia C. (2008). An alternative approach to detect Trypanosoma in *Glossina* (Diptera, *Glossinidae)* without dissection. *J Infect Developing Countries.* **2(1):**63-67.

Gascuel O. (1997). BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol.* **14:**685–695.

Geiger A., Ravel S., Frutos R., Cuny G. (2005). Sodalis glossinidius (Enterobacteriaceae) and vectorial competence of *Glossina palpalis gambiensis* and *Glossina morsitans morsitans* for *Trypanosoma congolense* savannah type. *Curr Microbiol*. **51(1):**35-40.

Ghosh K.N, Mukhopadhyay J. (1998). The effect of anti-sandfly saliva antibodies on *Phlebotomus argentipes* and *Leishmania donovani*. *Int J Parasitol.* **28(2):**275-81.

Green P. (1999). PHRAP Documentation http://www.phrap.org/

Guindon S., Gascuel O. (2003). "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood." *Systematic Biology*. **52(5):**696-704.

Guryev V., Berezikov E., Malik R., Plasterk R.H., Cuppen E. (2004) Single nucleotide polymorphisms associated with rat expressed sequences. *Genome Res*. **14(7):**1438-43.

Haines L.R, Hancock R.E., Pearson T.W. (2003). Cationic antimicrobial peptide killing of African trypanosomes and Sodalis glossinidius, a bacterial symbiont of the insect vector of sleeping sickness. *Vector Borne Zoonotic Dis*. **3(4):**175-86.

Hajduk S.L. (1984). Antigenic variation during the developmental cycle of *Trypanosoma brucei*. *J Protozool.* **31(1):**41-7.

Hammond G.L., Smith C.L., Goping I.S., Underhill D.A., Harley M.J., Reventos J., Musto N.A., Gunsalus G.L., Bardin C.W. (1987) Primary structure of human corticosteroid binding globulin, deduced from hepatic and pulmonary cDNAs, exhibits homology with serine protease inhibitors. *Proc Natl Acad Sci U S A.* **84(15):**5153-7.

Hao Z., Kasumba I., Lehane M.J., Gibson W.C., Kwon J., Aksoy S. (2001). Tsetse immune responses and trypanosome transmission: implications for the development of tsetse-based strategies to reduce trypanosomiasis. *Proceedings of the National Academy of Sciences of the United States of America.* **98(22):**12648–12653.

Hayes B.J., Nilsen K., Berg P.R., Grindflek E., Lien S. (2007). SNP detection exploiting multiple sources of redundancy in large EST collections improves validation rates. *Bioinformatics.* **23(13):**1692-3.

Herwig R., Schulz B., Weisshaar B., Hennig S., Steinfath M., Drungowski M., Stahl D., Wruck W., Menze A., O'Brien J., Lehrach H., Radelof U. (2002). Construction of a ‘unigene’ cDNA clone set by oligonucleotide fingerprinting allows access to 25 000 potential sugar beet genes. *Plant J.* **32:**845–857.

Hide G. (1999). History of sleeping sickness in East Africa. *Clin Microbiol Rev.* **12(1):**112-25.

Hoffman J.A. (2003). The immune response of *drosophila*. *Nature* **426:** 33-38.

Holt R.., Subramanian G.M., Halpern A., Sutton G.G., Charlab R., Nusskern D.R., Wincker P., Clark A.G., Ribeiro J.M., Wides R *et al.*(2002). The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science.* **298(5591):**129-49.

Hopkins P.C., Carrel R.W., Stone S.R. (1993). Effects of mutations in the hinge region of serpins. *Biochemistry.* **32:**7650-7657.

Hu C. and Aksoy S. (2006). Innate immune responses regulate trypanosome parasite infection of the tsetse fly *Glossina morsitans morsitans*. *Mol Microbiol*. **60(5):**1194-204.

Huang X and Madan A. (1999). CAP3:A DNA sequence assembly program. *Genome Res*. **9(9):**868-77.

Huelsenbeck, J. P. and Ronquist F. (2001). MRBAYES: Bayesian inference of phylogeny. *Bioinformatics*. **17:**754-755.

Irving J.A., Pike R.N., Lesk A.M., Whisstock J.C. (2000). Phylogeny of the serpin superfamily: implications of patterns of amino acid conservation for structure and function. *Genome Res*. **12:**1845-1864.

Irving P., Troxler L., Heuer T.S., Belvin M., Kopczynski C., Reichhart J.M., Hoffmann J.A., Hetru C. (2001). A genome-wide analysis of immune responses in *Drosophila. Proc Natl Acad Sci.* **98(26):**15119–15124.

Jones D.T., Taylor W.R., Thornton J.M. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences (CABIOS).* **8(3):**275-282.

Kabayo J.P. (2002). Aiming to eliminate tsetse from Africa. *Trends in Parasitology.* **18:**473-475.

Kasahara M., Naruse K., Sasaki S., Nakatani Y., Qu W., Ahsan B., Yamada T., Nagayasu Y., Doi K., Kasai Y., Jindo T, *et al.* (2007). The Medaka draft genome and insights into vertebrate genome evolution. *Nature.* **447(7145):** 714-9.

Kennedy P.G.E. (2007). The fatal sleep. Endiburgh, United Kingdom: Luath Press limited.

Koltes J.E., Hu Z.L., Fritz E., Reecy J.M., (2009). BEAP: The BLAST Extension and Alignment Program- a tool for contig construction and analysis of preliminary genome sequence. *BMC Res Notes*. **22(2):**11.

Kumar S., Tamura K., Nei M. (2004). MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform.* **5(2):**150-63.

Lanzaro G.C., Lopes A.H., Ribeiro J.M., Shoemaker C.B., Warburg A., Soares M., Titus R.G. (1999). Variation in the salivary peptide, maxadilan, from species in the Lutzomyia longipalpis complex. *Insect Mol Biol*. **8(2):**267-75.

Lehane M.J., Gibson W., Lehane S.M. (2007). Differential expression of fat body genes in *Glossina morsitans morsitans* following infection with *Trypanosoma brucei brucei.* *Int J Parasitol.* **38(1):**93-101.

Levashina E.A., Langley E., Green C., Gubb D., Ashburner M., Hoffmann J.A., Reichhart J.M. (1999). Constitutive activation of toll-mediated antifungal defense in serpin-deficient *Drosophila*. *Science.* **285:**1917-1919.

Li S., Kwon J., Aksoy S. (2001). Characterization of genes expressed in the salivary glands of the tsetse fly *Glossina morsitans morsitans*. *Insect Molecular biology* **10(1):**69-76.

Lindblad-Toh K., Wade C.M., Mikkelsen T.S., Karlsson E.K., Jaffe D.B., Kamal M., Clamp M., Chang J.L., Kulbokas E.J 3rd, Zody M.C., *et al.* (2005). Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature.* **438(7069):**803-19.

Liu M.F., Wu X.P., Wang X.L., Yu Y.L., Wang W.F., Chen Q.J., Boireau P., Liu M.Y. (2008). The functions of Deoxyribonuclease II in immunity and development. *DNA Cell.* **27(5):**223-8.

Löbermann H., Tokuoka R., Deisenhofer J., Huber R. (1984). "Human a1-proteinase inhibitor. Crystal structure analysis of two crystal modifications, molecular model and preliminary analysis of the implications for function." *J. Mol. Biol.* **177:** 531-556.

Mant M.J and Parker K.R. (1981) Two platelet aggregation inhibitors in tsetse (*Glossina*) saliva with studies of roles of thrombin and citrate in in-vitro platelet aggregation. *Br J Haematol.* **48(4):**601-8.

Marth G.T., Korf I., Yandell M.D., Yeh R.T., Gu Z., Zakeri H., Stitziel N.O., Hillier L., Kwok P.Y., Gish W.R. (1999). A general approach to single-nucleotide polymorphism discovery. *Nat Genet.* **23(4):**452-6.

Michel K., Budd A., Pinto S., Gibson T.J., Kafatos F.C. (2005). *Anopheles gambiae* SRPN2 facilitates midgut invasion by the malaria parasite Plasmodium berghei. *EMBO Rep.* **6(9):**891-7.

Miller R.T., Christoffels A.G., Gopalakrishnan C., Burke J., Ptitsyn A.A., Broveak T.R., Hide W.A. (1999). A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res.* **9(11):**1143-55.

Milligan P.J., Maudlin I., Welburn S.C. (1995). Trypanozoon: infectivity to humans is linked to reduced transmissibility in tsetse II. Genetic mechanisms. *Exp. Parasitol.* **81:**409–415.

Nagaraj S.H., Gasser R.B., Ranganathan S. (2007). A hitchhiker's guide to expressed sequence tag (EST) analysis. *Briefings in Bioinformatics.* **8(1):**6-21.

Notredame C., Higgins D., Heringa J (2000). T-Coffee: A novel method for multiple sequence alignments. *Journal of Molecular Biology*. **302:** 205-217.

Okolo C.J., Jenni L., Molyneux D.H., Wallbanks K.R. (1990). Surface carbohydrate differences of *Glossina* salivary glands and infectivity of *Trypanosoma brucei gambiense* to *Glossina*. *Ann Soc Belg Med Trop*.**70(1):**39-47.

Osterwalder T., Kuhnen A., Leiserson W.M., Kim Y.S., Keshishian H. (2004). *Drosophila* serpin 4 functions as a neuroserpin-like inhibitor of subtilisin-like proprotein convertases. *J Neurosci.* **24(24):**5482-91.

Pays E., Lips S., Nolan D., Vanhamme L., Pérez-Morga D. (2001). The VSG expression sites of *Trypanosoma brucei*: multipurpose tools for the adaptation of the parasite to mammalian hosts. *Mol Biochem Parasitol*.**114(1):**1-16.

Pearson W.R. (1997). Identifying distantly related protein sequences. *Comput Appl Biosci.* **13:** 325-332.

Pinder M., Fumoux F., van Melick A., Roelants G.E (1987). The role of antibody in natural resistance to African trypanosomiasis. *Vet Immunol Immunopathol*. **17(1-4):**325-32.

Potempa J., Korzus E., Travis J. (1994). The Serpin Superfamily of Proteinase Inhibitors: Structure, Function, and Regulation. *J. Biol. Chem*. **269:** 15957-15960.

Rawlings N.D., Tolle D.P., Barrett A.J. (2004). Evolutionary families of peptidase inhibitors. *Biochem J*. **378:**705–716.

Ribeiro J.M., Charlab R., Pham V.M., Garfield M., Valenzuela J.G. (2004). An insight into the salivary transcriptome and proteome of the adult female mosquito *Culex pipiens quinquefasciatus. Insect Biochem Mol Biol*. **34(6):**543-63.

Rickman L.R. (1977). Variation in the test responses of clone derived *Trypanosoma* (Trypanozoon) *brucei brucei* and T.(T).b. *rhodesiense* relapse antigenic variants, examined by a modified blood incubation infectivity test and its possible significance in Rhodesian sleeping sickness transmission. *Med J Zambia*.**11(2):**31-7.

Rudd S. (2003). Expressed sequence tags: alternative or complement to whole genome sequences? *Trends Plant Sci*. **8(7):**321-9.

Saitou N. and Nei M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4(4):406-25.

Sali A and Blundell T.L. (1993). Comparative protein modeling by satisfaction of spatial restraints. J Mol Biol .234(3):779-815.

Seong C.S., Varela-Ramirez A., Aguilera R.J. (2006). DNase II deficiency impairs innate immune function in *Drosophila*. *Cell Immunol*. **240(1):**5-13.

Silverman G.A., Bird P.I., Carrell R.W., Church F.C., Coughlin P.B., Gettins P.G., Irving J.A., Lomas D.A., Luke CJ., Moyer R.W., Pemberton P.A., Remold-O'Donnell E., Salvesen G.S., Travis J., Whisstock J.C. (2001). The Serpins Are an Expanding Superfamily of Structurally Similar but Functionally Diverse Proteins. *J Biol. Chem.* **276:** 33293-33296.

Staden R. (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.* **6(7):**2601-10.

Stich A., Abel P.M., Krishna S. (2002). Human African trypanosomiasis. *BMJ*. **27; 325(7357):**203-6.

Stickney H.L., Schmutz J., Woods I.G., Holtzer C.C., Dickson M.C., Kelly P.D., Myers R.M., Talbot W.S.(2002). Rapid mapping of zebrafish mutations with SNPs and oligonucleotide microarrays. *Genome Res*.**12(12):**1929-34.

Tanaka H., Ishibashi J., Fujita K., Nakajima Y., Sagisaka A., Tomimoto K., Suzuki N., Yoshiyama M., Kaneko Y., Iwasaki T., Sunagawa T., Yamaji K., Asaoka A., Mita K., Yamakawa M. (2008). A genome-wide analysis of genes and gene families involved in innate immunity of *Bombyx mori*. *Insect Biochem Mol Biol*. **38(12):**1087-110.

Tanji T., Hu X., Weber AN., Ip YT. (2007). Toll and IMD Pathways Synergistically Activate an Innate Immune Response in *Drosophila melanogaster*. *Molecular and Cellular Biology*. **27(12):**4578-4588.

Tanji T and Ip Y.T. (2005). Regulators of the Toll and Imd pathways in the *Drosophila* innate immune response. *Trends in immunology*. **26(4):** 193-198.

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.* **22bghj(22):** 4673-80.

Venter J.C., Adams M.D., Myers E.W., Li P.W., Mural R.J., Sutton G.G., Smith H.O., Yandell M., Evans C.A., Holt R.A *et al.* (2001). The sequence of the human genome. *Science.* **29(5507):** 1304-1351.

Vickerman K. (1985). Developmental cycles and biology of pathogenic trypanosomes. *Br Med Bull*. **41(2):** 105-14.

Waterhouse A.M., Procter J.B., Martin D.M., Clamp M., Barton G.J. (2009). Jalview Version 2 - a multiple sequence alignment editor and analysis workbench. Bioinformatics. [Epub ahead of print].

Welburn S.C., Maudlin I., Ellis D.S. (1998) Rate of trypanosome killing by lectins in midguts of different species and strains of *Glossina. Med Vet Entomol*. **3(1):**77-82.

Welburn S.C., Maudlin I. (1999). Tsetse-Trypanosome interactions: Rites of passage. *Parasitology today*. **15(10):**399-403.

Whisstock J.C., Pike R.N., Jin L., Skinner R., Pei X.Y., Carrell R.W., Lesk A.M (2000) Conformational changes in serpins: II. the mechanism of activation of antithrombin by heparin. *Journal of Molecular Biology*. **301(5):**1287-1305.

Zou Z., Picheng Z., Weng H., Mita K., Jiang H. (2009). A comparative analysis of serpin genes in the silkworm genome. *Genomics.* **93(4):**367-75.

APPENDICES

APPENDIX I: Outline of PHRED and PHRAP

Conversion of chromatograms to computer readable form by PHRED

Conversion of DNA sequencer trace files to computer readable form is the first step of any assembly process; this was done by the program PHRED (Ewing and green, 1998). PHRED reads the DNA sequencer trace data and examines the peaks around each base to assign a quality score. The scores provide a measure of sequence quality and are calculated using the logarithmic function obtained from Ewing and Green (1998).

$$q=10 * \log 10 (P)$$

Where;

q = quality score

p = probability that a particular base was an error

Each base call is therefore logarithmically linked to error probability. Quality scores range from 4 to 60 with higher values corresponding to better sequence quality. For example, a quality score of 30 represents 1/1000 chances of being incorrect while a score of 20 represents 1/100 chances of being incorrect. PHRED scores can therefore be used to extract portions of high quality in any given sequence or the entire sequence depending with the biological question being answered during downstream EST analysis.

Due to the inherent nature of errors that are associated with any EST sequencing project bases within the middle segment of a given EST sequence typically contain the high quality bases while the extreme ends normally contain vector contaminants and repeats. PHRED then writes the base calls in FASTA, PHD (text files that consist of base call and quality information, used during contig editing) or SCF (Standard Chromatogram File) formats. The quality scores are written to FASTA or PHD files (Ewing and Green, 1998; Ewing *et al.*, 1998).

Summary of the steps in a PHRAP assembly

1. PHRAP reads the sequence from the input file as well as the corresponding quality file.

2. PHRAP identifies and trims homopolymeric regions (those that consist almost entirely a single base). These regions are most likely due to poor data quality and may give rise to false matches.

3. PHRAP identifies all potentially overlapping pairs of sequences. Two sequences must have at least one "word" (whose default is set at 14 bases) and a minimum alignment score (default is set at 30).

4. PHRAP computes adjusted quality scores for each base in each overlapping read taking into consideration read orientation and sequencing chemistry. In addition PHRAP calculates log-likelihood ratio (LLR) scores. LLR scores compare the hypothesis that the reads truly overlap to the hypothesis that they are from 95% similar repeats. A pairwise match tends to have a positive LLR score if the two reads overlap, whereas it tends to have a negative LLR score if the two reads are from different repeats. Possible problem clones like chimeric and deletion reads are also identified and withheld from the assembly.

5. PHRAP merges reads into contigs (consensus) starting with pairwise overlaps with the highest LLR scores. The consensus sequence is "pieced" together from individual sequence reads with the highest adjusted quality score at any base.

APPENDIX II: List of Perl scripts

Perl script for creating quality scores for three *Tsal* cDNAs downloaded from GENBANK.

```
#!/usr/bin/perl
use warnings  ;
use strict ;

# Takes a fasta file, determines the length of the sequence
# and creates bogus phred quality scores for each sequence
# The default phred score is 15
# Usage: perl <scriptname> fastafile1 fastafile2....

foreach my $file(@ARGV){
  my $qual_out = $file.".qual" ;
  open(FASTA_IN, $file) ;
  open(QUAL_OUT, ">>$qual_out") ;
  $/ = "\n>" ;          # This line will make the 'while' loop below
deal
                # with a complete FASTA entry at a time
  while(my $sequence = <FASTA_IN>){
    $sequence =~ s/\>//g ;        # remove all ">" characters
    my @array = split(/\n/, $sequence) ;
    my $head = shift(@array) ;
    my $sequence_string = join("", @array) ;
    print              QUAL_OUT              phred_score_producer($head,
length($sequence_string)), "\n" ;
  }
  close(FASTA_IN) ;
  close(QUAL_OUT) ;
}

# SUBROUTINES
sub phred_score_producer{
  my ($header, $sequence_length) = @_ ;
  my $fasta_entry = ">$header" ;
  for( my $index = 0 ; $index < $sequence_length ; $index++ ){
    if(($index/20) == int($index/20)){          # this makes sure my
lines
                                        # are 20 characters long
      $fasta_entry .= "\n30" ;
    }else{ $fasta_entry .= " 30" ;
    }
  }
  return($fasta_entry) ;
}
```

Perl script for extracting read names from a “.ace” file output

```
#! /usr/bin/perl

$/="\nCO";# stop at a line that ends with CO
my $file = shift @ARGV;# initialises our file as an array
open(FH $file);

while (<>) {
	next if ($_=~/^AS/);#match stdin with anything that starts with
AS
	my @lines = split("\n", $_);#create an array that will contain
the lines
	my $contigname = shift @lines;#read the first element of the
lines array and store it in $contigname
	$contigname =~/^\s+(\S+)/;#in the store for contig name pick up
the second item (S+) after the first space s+
	my $contig = $1; #take the pattern in the first set of brackets
which is also equal to $1 and put it into my new variable called
$contig.
	foreach my $line(@lines) {
		if ($line =~/^AF\s+(\S+)/) {
			my $readname = $1;
			print "$contig\t$readname\n";
		}
	}
}
```

Perl script for extracting single base pair mismatches from a “.GDE” file.

```
#! /usr/bin/perl
#
# background: test.pl reads in an ace file and generates a list of
contigIDs vs ESTnames. eg., Contig1 Est1 est2 est3
# --output file= Ace_ESTnames
#
# acefile converted to GDE format using AMOS. However the estnames
have a number and brackets at the end (ESTname(447))
# --output file =tsal.fasta.gde
#
# stripped off the number at the end of the name and saved the
sequences into a new file using ConvertGdeName.pl
# /ConvertGdeName.pl tsal.fasta.gde3 ----output was
tsal.fasta.gde3.new
#
# Use bioperl index script to index the output of ConvertGdeName.pl -
index_file called GDESEQ
# usage: ~/cvs_src/bioperl-live/scripts/index/bp_index.PLS -dir .
GDESEQ ./tsal.fasta.gde3.new
#
# SNP_count.pl
# reads in output from test.pl and store the information in a hash
table.
#
# found bug - est can go beyond contiglength
use strict;
use Bio::SeqIO;
use Bio::Seq;
use Bio::Index::Fasta;

my $estnames = shift;
my %hash;
my             $in             =             Bio::Index::Fasta->new(-
filename=>"/home/sarah/tsal/SNP_results/TSALSEQ");
#my $out = Bio::SeqIO->new(-file=>">$f.test", -format=>"Fasta");

open(F, $estnames);
while (<F>) {
      chomp;
      my @i = split(/\s+/);
      my $contig = shift @i;
      @{$hash{$contig}} = @i;
}
close(F);

foreach my $contig(keys %hash) {
      #print "$contig\n";
      my $contig_seq = $in->fetch($contig);
      #print $contig_seq->id."\n";

      my $contiglen = length($contig_seq->seq);
      my @est_comp;
      foreach my $est(@{$hash{$contig}}) {
            my $est_seq = $in->fetch($est);
            my ($start_string, $end_string);
            if ($est_seq->desc =~/\<(.*)\>$/) {
                  my $str = $1;
                  #print "$str\n";
                  # start and stop are pos where EST aligns with other
ESTs
                  # these are positions relative to the consensus
                  my ($start, $stop) = split(/\s+/, $str);
                  #print "start = $start stop=$stop\n";
                  if ($start > 1) {
                        my $start_string_len = ($start - 1);
                        #print "$start_string_len\n";
                        $start_string = "";
                        foreach my $j(1..$start_string_len) {
                            $start_string .="N";
                        }
                  } else {
                        $start_string = "";
                  }

                  #deal with the end_string
                  if ($stop < $contiglen) {
                        my $end_string_len = ($contiglen - $stop);
                        $end_string = "";
                        foreach my $j(1..$end_string_len) {
                            $end_string .="N";
                        }
                  } else {
                        $end_string="";
                  }
            }
            #print "$est .$start_string.\n";
            #print "$est .$end_string\n";
            my $new_est = $start_string.$est_seq->seq().$end_string;
            #print "$est     .$new_est.\n";

            my @new_estseq = split("", "\U$new_est");
            push @est_comp, [@new_estseq];
      }
      my ($j, $i, %temphash);
      for ($j=0; $j< $contiglen; $j++) {
      %temphash=();
      for ($i=0; $i<= $#est_comp; $i++) {
      #print $dna[$i][$j]."\n";
      #$temphash{$est_comp[$i][$j]}++ if ($est_comp[$i][$j] ne "N");
      $temphash{$est_comp[$i][$j]}++;
```

```
                #print "$i:$j .$est_comp[$i][$j]. - ";
      }
      if (scalar(keys %temphash) ==1) {
           #conserved pos
           print "pos-idx-$j";
                my ($n) = keys %temphash;
                print " $n ***\n";
      } else {
           print "pos-idx-$j";
         foreach my $base(keys %temphash) {
           print " $base\[$temphash{$base}\]";
         }
         print "\n";
      }
     }
}
```