STRUCTURE PREDICTION OF MATRIX METALLOPROTEINASES IN *Anopheles gambiae* USING BIOINFORMATIC TOOLS

MUTAI JACQUELINE

A Thesis Submitted to the Graduate School in Partial Fulfillment for the Requirements of the Award of a Master of Science Degree in Biochemistry of Egerton University

EGERTON UNIVERSITY

APRIL, 2017

DECLARATION AND RECOMMENDATION

Declaration

This thesis is my original work and has not been submitted or presented for examination in any other institution

Signature …… Date...……..…….……..

Ms. Mutai Jacqueline

SM14/2864/10

Recommendation

This thesis has been submitted with our approval as supervisors according to Egerton University regulations.

Signature …….. Date……..…..……

Dr. Paul Mireji, (PhD)

Yale School of Public Health,

Yale University

Signature …… Date……

Dr. Ramadhan Mwakubambanya, (PhD)

Department of Biochemistry and Molecular Biology,

Egerton University

COPYRIGHT

DEDICATION

I dedicate this thesis to my parents, Dr. Joseph K. Mutai and Mrs. Eddy Mutai, and to my siblings. Their continued support, prayers and encouragement enable me to reach this far and produce this thesis.

God Bless You

ACKNOWLEDGEMENT

I sincerely thank the Almighty God for all his goodness and faithfulness. Without His grace and strength l would not have finished this work.

I thank my supervisors Dr. Paul Mireji, and Dr. Ramadhan Mwakubambanya for their advice and guidance. God bless you richly.

I also want to thank Dr. Rob Skilton, formerly of BECA, ILRI for giving me an opportunity to do the bioinformatics aspect of this thesis at BeCA, Dr Etienne for his guidance and advice. I thank the BeCA team for their continued guidance on this thesis.

I also want to thank Dr. Daniel Masiga for giving me the opportunity to do the wet lab analysis at International Centre for Insect Physiology and Ecology (ICIPE). I also wish to thank Mr. James Kabii for his help and guidance during this duration.

I am also grateful to the great scientist that l have met during this journey, though most of you are pursuing your masters and PhD, in my eyes you are great scientists. Thank you to those great scientists that l have met in ILRI and ICIPE.

ABSTRACT

Human malaria is the most important disease in tropical countries in terms of morbidity and mortality. Malaria transmission involves complex interactions between *Plasmodium falciparum* and *Anopheles gambiae*. For successful establishment of invasion/infection of the *Anopheles gambiae* midgut the parasite must overcome the immune responses of the vector. Matrix metalloproteinase (MMPs) are a family of zinc metalloendopeptidases known to disrupt sub-endothelial membranes, destroy tight junctions and shed active cytokines, chemokines and other MMPs through cleavage from their precursors. The latter function putatively explains the great parasite loss during invasion of the *Anopheles gambiae* midgut by the parasite. The objective of this thesis was to study matrix metalloproteinases in *An. gambiae* as a potential addition to transmission blocking strategies. BLASTp searches of the complete *Anopheles gambiae* genome using *Drosophila melanogaster* MMP resulted in the identification of two Metazoa-like MMP genes. Domains of these proteases were determined through InterProScan. The 3-D structure was determined using MODELLER. The structure was validated using MetaMQAPII, ProSA and PROCHECK. A validation of the presence of MMPs in *Anopheles gambiae* was performed through RNA extraction, cDNA synthesis and Polymerase Chain Reaction amplification. Based on the BLAST output, two MMP genes similar to *Drosophila melanogaster* MMP were found in *An. gambiae* (AGAP006904 and AGAP003929). The proteases were shown to have a prodomain, metalloproteinase domain (catalytic domain) and a hemopexin domain and were classified into superfamily and family through presence of conserved domains and residues in a multiple sequence alignment (MSA). The modeled protein had a similar structural conformation to human pro-collagenase. The results of the amplification showed that AGAP006904 produced a truncated transcript. PCR amplification showed that MMP1 transcript A (AGAP006904**)** is expressed in the larvae, pupae and adult of *An. gambiae*. We can conclude that, *Anopheles* MMP is similar to MMP from humans and Dipterans in structural conformation and domain architecture. The presence of MMP in the 3 stages of *Anopheles gambiae* indicates a possible role in development. Knowledge of the structure and activation of *Anopheles* MMP is vital in understanding how this protein folds, which is vital in coming up with transmission blocking strategies to either inhibit or activate these proteases.

TABLE OF CONTENTS

DECLARATION AND RECOMMENDATION ..... **ii**
COPYRIGHT ..... **iii**
DEDICATION ..... **iv**
ACKNOWLEDGEMENT ..... **v**
ABSTRACT ..... **vi**
TABLE OF CONTENTS ..... **vii**
LIST OF TABLES ..... **x**
LIST OF FIGURES ..... **xi**
LIST OF ABBREVIATIONS ..... **xii**
CHAPTER ONE **1**
INTRODUCTION **1**
1.1. Background Information 1
1.2. Statement of the problem 2
1.3. Objectives 3
1.3.1. General objective 3
1.3.2. Specific objectives 3
1.4. Justification. 3
CHAPTER TWO **4**
LITERATURE REVIEW **4**
2.1. Prevalence of Malaria **4**
2.2. Life cycle of *Plasmodium falciparum* 6
2.2.1. The vertebrate host (human) 6
2.2.2. The vector host (female *Anopheles* mosquito) 7
2.3. Matrix Metalloproteinases 8
2.3.1. Other Metalloproteases 10
2.3.2 *Drosophila melanogaster* MMP 13
2.4. Bioinformatic Programs 13
2.4.1 BLAST 13

2.4.2. Determination of domains. 14
2.4.3 Structural Analysis..... 14
2.4.4 Characterization of Matrix Metalloproteinases into Protein Families..... 17
CHAPTER THREE 18
MATERIALS AND METHODS 18
3.1. Determination of Matrix Metalloproteinase Domain 18
3.2. Predictions of 3-D Structure..... 18
3.3. Characterization of Matrix Metalloproteinases. 19
3.3.1. Gene Validation..... 19
3.3.2. Primer Design..... 19
3.3.3. Extraction of RNA 20
3.3.4. Purity. 21
3.3.5. First Strand cDNA synthesis 21
3.3.6. Polymerase chain reaction 21
3.3.7. PCR product purification 22
3.3.8. Bioinformatic Analysis..... 23
CHAPTER FOUR..... 24
RESULTS AND DISCUSSION 24
4.1. Results..... 24
4.1.1. Determination of Matrix Metalloproteinase's Domain 24
4.1.2. Prediction of 3-D Structure 26
4.2. Discussion 43
CHAPTER FIVE 51
CONCLUSION AND RECOMMENDATION 51
5.1. Conclusion..... 51
5.2. Recommendation..... 51
REFERENCES..... 52
APPENDIX I: Erroneous structure of AGAP003929..... 64
APPENDIX II: Scripts for modeling and assessment of energy..... 65

APPENDIX III: DNA sequences from Sanger sequencing 66

APPENDIX IV: Protein sequences retrieved from NCBI and used in 3-D structure prediction..... 67

LIST OF TABLES

Table 1: MMP Primers: Sequences and their expected product size 20
Table 2: Nanodrop results of RNA 36
Table 2: Nanodrop results of cDNA 36

LIST OF FIGURES

Figure 1: Spatial distribution of *Plasmodium falciparum* Endemicity. 5
Figure 2: Life cycle of *Plasmodium falciparum* 8
Figure 3: Schematic of the domain architecture of human matrix metalloproteinases..…….....10
Figure 4: Generic Structure of matrix metallopeptidase..... 13
Figure 5: InterProScan 25
Figure 6: Predicted 3D structure 26
Figure 7: Prodomain. 27
Figure 8: Interaction of Prodomain..... 28
Figure 9: Metalloproteinase domain A and B. 31
Figure 10: Hemopexin domain A and B. 32
Figure 11: Ramachandran Plot. 33
Figure 12: ProSA……...34
Figure 13: Root Mean Square Deviation (RMSD) 36
Figure 14: This figure show the resluts of PCR amplification.. 37
Figure 15: Multiple Sequence Alignment. 42

LIST OF ABBREVIATIONS

BLAST	Basic Local Alignment Search Tool
CDC	Centre for Disease Control
DNA	Deoxyribonucleic Acid
DOPE	Discrete Optimized protein Energy.
HMM	Hidden Markov Model
MetaMQAPII	Meta-Model Quality Assessment Program
MMP	Matrix Metalloproteinases
mRNA	Messenger Ribonucleic Acid
MSA	Multiple Sequence Alignment
NCBI	National Center for Biotechnology Information
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
Pfam	Protein Families
pfAPI	*Plasmodium falciparum* Annual Parasite Incidence
pfPR	*Plasmodium falciparum* Parasite Rate
ProSA	Protein Structure Analysis
PROVE	PROtein Volume Evaluation
PSSM	Position Specific Scoring Matrix
RMSD	Root Mean Square Deviation
SAVeS	Structural Analysis and Verification Server
SCOP	Structural Classification of Proteins
SG	Salivary Glands
SMART	Simple Modular Architecture Research Tool
TIMP	Tissue Inhibitor for Metalloproteinases

CHAPTER ONE
INTRODUCTION

1.1. Background Information

Malaria, caused by *Plasmodium falciparum*, is one of the most prevalent and lethal diseases affecting humans. Globally, an estimated 3.3 billion people were at risk of malaria in 2011, with people living in sub-Saharan Africa having the highest risk of acquiring malaria: approximately 80% of cases and 90% of deaths estimated occur in the WHO African Region, with children under 5 years of age and pregnant women severely affected (World Malaria Report, 2012). The highest and most lethal incidences of malaria are caused by *Plasmodium falciparum* as the main causative agent, predominantly in Africa (Snow *et al*., 2005). *Plasmodium* is a complex organism completing its life cycle in two different hosts, an invertebrate (female *Anopheles* mosquito) and vertebrate (humans) hosts. *Plasmodium*; *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium ovale* and *Plasmodium malariae* constitute the major species of the parasite. Africa has over 140 recorded *Anopheles* species, of which at least eight are considered to be effective vectors of malaria (Gillies and Coetzee, 1987; Gillies and De Meillon, 1968). In the *An. gambiae* complex, *Anopheles gambiae sensu stricto* (*Anopheles gambiae ss*) and *Anopheles arabiensis* (White, 1974) are the most efficient vectors of human malaria. Other recognized species of the complex are *Anopheles merus*, *Anopheles melas*, *Anopheles quadriannulatus*, *Anopheles quadriannulatus* B *and Anopheles bwambae* (White, 1985). In addition to the *An. gambiae* complex, other species known to be important in malaria transmission in Africa include *Anopheles nili*, *Anopheles moucheti* and *Anopheles funestus* which belong to the Funestus group of which there are two African subgroups (Funestus subgroup includes *Anopheles aruni*, *Anopheles confusus*, *Anopheles funestus sensu stricto*, *Anopheles parensis* and *Anopheles vaneedeni*; Rivulorum subgroup includes *Anopheles brucei*, *Anopheles fuscivenosus*, *Anopheles rivulorum*, and An. rivulorum-like species) (Gillies and Coetzee, 1987; Harbach, 2004). Other species, such as *Anopheles paludis*, *Anopheles mascarensis* and *Anopheles hancocki* play only a limited, secondary and localized role where they are found (Fontenille and Simard, 2004).

Matrix metalloproteases (MMPs) comprise a family of enzymes that cleave protein substrates based on a conserved mechanism involving activation of an active site-bound water molecule by a Zn^{2+}ion. Although the catalytic domain of MMPs is structurally highly similar,

there are many differences with respect to substrate specificity, cellular and tissue localization, membrane binding and regulation that make this a very versatile family of enzymes with a multitude of physiological functions, many of which are still not fully understood (Klein and Bischoff, 2011).

The main physiological function of these proteases was originally ascribed to the modulation and regulation of extracellular matrix (ECM) turnover by direct proteolytic degradation of the ECM proteins (e.g., collagen, proteoglycans and fibronectin) (Woessner 1991). Another important function is the liberation of biologically active proteins such as cytokines, growth factors and chemokines from their membrane-anchored proforms (so-called shedding). MMPs thus contribute to the generation of protein species with vastly differing activities from a single, original gene product.

Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth in the biological information generated by the scientific community. This deluge of genomic information has, in turn, led to an absolute requirement for computerized databases to store, organize, and index the data, and for specialized tools to view and analyze the data.

1.2. Statement of the problem

Malaria is one of the leading causes of child mortality in Africa, Asia, the Middle East, Central and South America and Oceania. The parasite life cycle within the *An. gambiae* vector consists of rapid invasion, followed by rounds of intracellular replication of the parasites. Most malaria intervention research efforts are skewed towards *P. falciparum* in human-parasite interaction and to date, reliable malaria control measures are still elusive. The resurgence of malaria is partly attributed to the absence of an effective vaccine, parasite resistance to antimalarial drugs and *anopheline* mosquito's resistance to insecticides. One of the functions of MMPs in insects is in development and immunity. MMPs can be more sustainable malaria intervention because its inhibition or enhanced expression will curb the development of *Anopheles* and increase immune response, leading to the elimination of *P. falciparum*. Studies can be done to genetically modify *Anopheles* to produce an increase in MMP production leading

to increased liberation of cytokines and chemokines to disrupt parasite development in the mosquito and hence transmission blocking.

1.3. Objectives

1.3.1. General objective

To predict the structure of matrix metalloproteinase in *Anopheles gambiae* using bioinformatics tools

1.3.2. Specific objectives

1. To add to knowledge of matrix metalloproteinase in *An. gambiae*.
2. To determine molecular structure of matrix metalloproteinases in *An. gambiae.*
3. Characterize matrix metalloproteinase in *An. gambiae*

1.4. Justification.

Novel strategies are needed to combat malaria on three fronts: protection (vaccines), prophylaxis/treatment (antimalarial drugs) and transmission blocking. The latter entails either killing mosquitoes using insecticides, prevent mosquito biting through the use of bednets and repellants, blocking parasite development in the vector through genetic manipulation or chemical incapacitation of the mosquitoes. During the past decade, mosquito research has been energized by several breakthroughs including the successful transformation of *Anopheline* vectors, analysis of gene function by RNAi, genome-wide expression profiling using DNA microarrays and most importantly sequencing of the *An. gambiae* genome. The sequencing of *An. gambiae* genome will therefore make it possible to study matrix metalloproteinases in *An. gambiae.*

CHAPTER TWO
LITERATURE REVIEW

2.1. Prevalence of Malaria

Malaria in tropical regions which is caused by the protozoan parasites Plasmodium falciparum and Plasmodium vivax is responsible for 515 million (Snow *et al*., 2005) and 1-3 million deaths annually (Sachs, 2002). Plasmodium vivax, the most widespread parasite causing human malaria, is responsible for an estimated 130-435 million infections annually and is the major cause of malaria in most of Asia and Latin America (Baird, 2007).

In Eastern Mediterranean Region, approximately 55% of the population is at some risk of malaria and about 20% of the population is at high risk. Malaria endemicity varies considerably: 7 countries still have areas of high malaria transmission (Afghanistan, Djibouti, Pakistan, Somalia, South Sudan, Sudan and Yemen); malaria transmission is geographically limited in 2 countries (Iran (Islamic Republic of) and Saudi Arabia) whereas Iraq has not reported locally acquired cases since 2009. *P. falciparum* is the dominant malaria species in Djibouti, Saudi Arabia, Somalia, South Sudan, Sudan and Yemen, while the majority of cases in Afghanistan, Iran (Islamic Republic of) and Pakistan are due to *P. vivax*. Afghanistan, Iran (Islamic Republic of), Iraq, and Saudi Arabia achieved a decrease in malaria cases and case incidence rates of ≥75% between 2000 and 2011 (World Malaria Report, 2012).

In the European Region the total number of reported malaria cases decreased from 33 365 in 9 countries in 2000 to just 226 in 5 countries in 2011. Only 69 of the 226 malaria cases were indigenous; these were reported from Tajikistan and Azerbaijan. No locally-acquired *P.falciparum* cases have been reported since 2008; the last case was reported from Tajikistan. All other *Plasmodium falciparum* malaria cases found in the Region in 2011 were imported (World Malaria Report, 2012).

In South-East Asia Region approximately 70% of the population of 1.8 billion people is at some risk for malaria, with 26% at high risk: 460 million people inhabit areas with a reported incidence of >1 case per 1000 population per year. The majority of confirmed cases in the Region are due to *P. falciparum*, although the proportion varies greatly among countries (World Malaria Report, 2012).

In Western Pacific Region approximately 870 million are at some risk of malaria of which 69 million (8%) people inhabit areas with a reported incidence of ≥1 case per 1000 population per year (World Malaria Report, 2012).

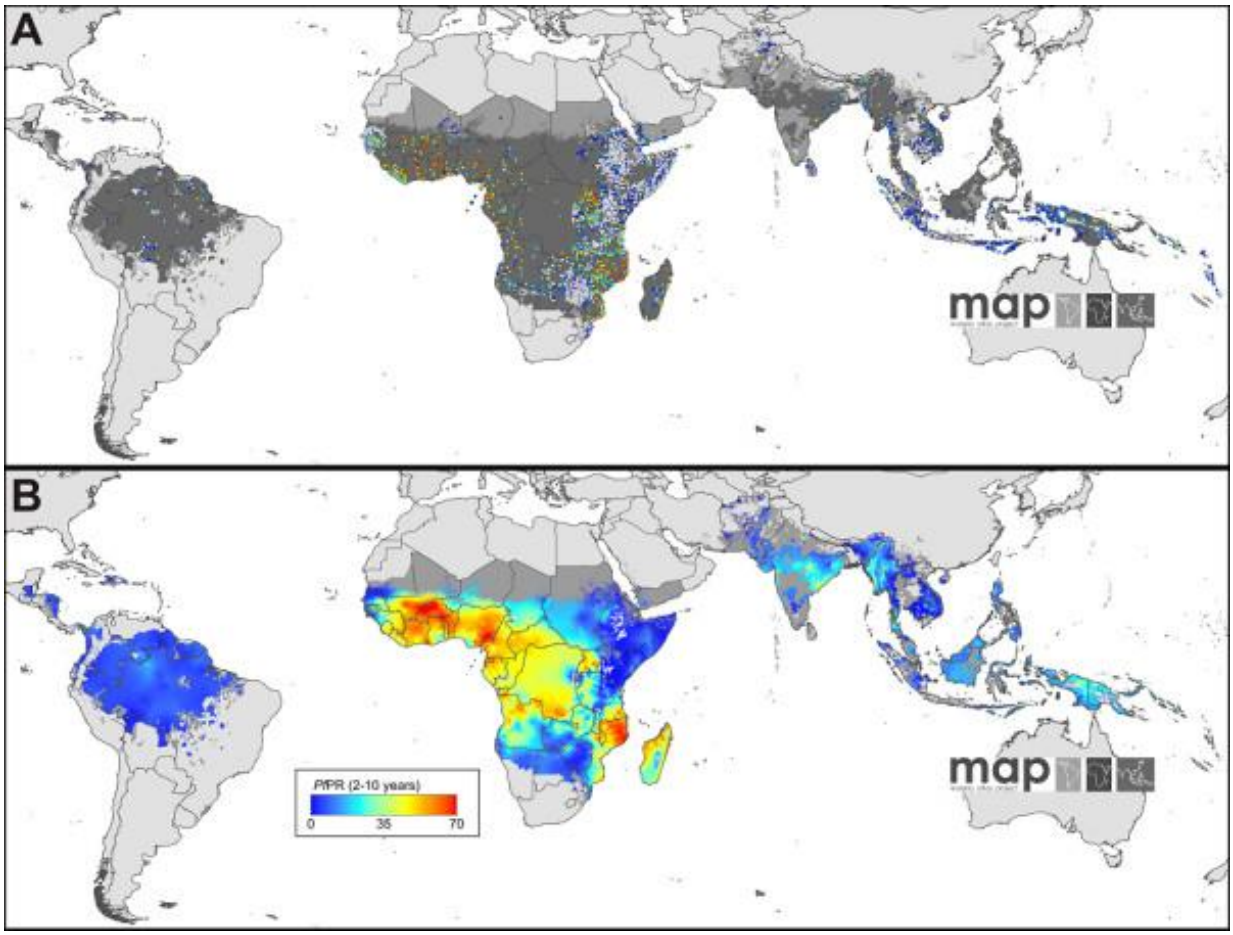


Figure 1: Spatial distribution of Plasmodium falciparum Endemicity (Gething *et al*., 2011)

Panel A shows the 2010 spatial Limits of *P. falciparum* malaria risk defined by PfAPI with further medical intelligence, temperature and aridity Masks. Areas were defined as stable (dark grey areas, wherePfAPI≥0.1 per 1,000 pa), unstable (medium grey areas, where PfAPI< 0.1 per 1,000 pa) or no risk (light grey, where PfAPI = 0 per 1,000 pa). Panel B shows the Model-Based Geostatistical (MBG) point estimates of the annual mean PfPR 2-10 for 2010 within the spatial limits of stable *P. falciparum* malaria transmission, displayed on the same colour scale. Areas of no risk or unstable risk are as in (A) (Gething *et al*., 2011)

2.2. Life cycle of *Plasmodium falciparum*

2.2.1. The vertebrate host (human)

Sporozoites from the saliva of a biting female mosquito are transmitted to either the blood or the lymphatic system of the human host. The parasites block the salivary ducts of the mosquito and as a consequence the insect normally requires multiple attempts to obtain blood. Multiple attempts by the mosquito may contribute to immunological tolerance of the parasite (Guilbride *et al*., 2010). The majority of sporozoites appear to be injected into the subcutaneous tissue from which they migrate into the capillaries. A proportion is ingested by macrophages and still others are taken up by the lymphatic system where they are presumably destroyed. Approximately 10% of the parasites inoculated by the mosquitoes may remain in the skin where they may develop into infective merozoites (Gueirard *et al*., 2010).

After the infective mosquito bite, the sporozoites rapidly reach the liver (Patricia *et al*., 2004) and traverse several cells by breaching their plasma membrane before they finally invade their target cells (hepatocytes) through the formation of a vacuole (Mota *et al*., 2001). In the hepatocytes, the sporozoite undergoes an initial modeling of the pellicle, with disassembly of the inner membrane complex and the appearance of a bulb that progressively enlarges until the initially elongated sporozoite has transformed into a rounded form. This rounded form then matures within the hepatocyte to a schizont containing thousands of merozoites.

The merozoites are released into the bloodstream upon rapture of the mature schizont where they invade erythrocytes. In the erythrocytes, the merozoites undergo asexual development to form schizonts carrying 16-18 merozoites, a process referred to as the erythrocytic stage. The cycle takes between 48 hours to 72 hours depending on the parasite species: irregular cycle for *P. falciparum*, 48 hours for *P. vivax*, and *P. ovale* and 72 hours for *P. malariae*. The mature schizont raptures releasing new merozoites which invade new erythrocytes. It is this cycle which is responsible for the clinical manifestations of malaria, fever and chills (Cowman and Crabb, 2006).

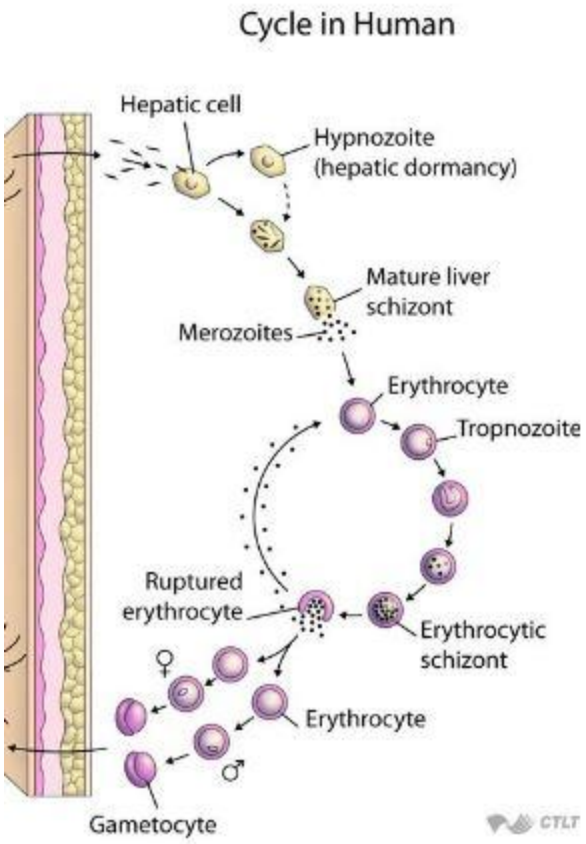


Figure 2-A: Lifecycle and of *Plasmodium* in human host

2.2.2. The vector host (female *Anopheles* mosquito).

Not all invading merozoites divide into schizonts; some differentiate into sexual forms, male and female gametocytes. These gametocytes are taken up by a female *Anopheles* mosquito during a blood meal. Within the mosquito mid-gut, the male gametocyte undergoes a rapid nuclear division, producing 8 flagellating microgametes which fertilize the female macrogamete forming zygote and eventually a motile parasite called ookinete. The ookinete traverses the mosquito mid-gut epithelium and encysts on the exterior of the gut wall as an oocyst. This cycle known as the sporogonic cycle lasts for 7-21 days. The oocyst ruptures, releasing thousands of sporozoites into the mosquito haemoceal. The sporozoites migrate to the mosquito salivary gland (SG) and undergo a developmental cycle, in order to become highly infectious to the mammalian host. Inoculation of the sporozoites into a human host perpetuates the malaria life cycle (Mueller *et al*., 2005).

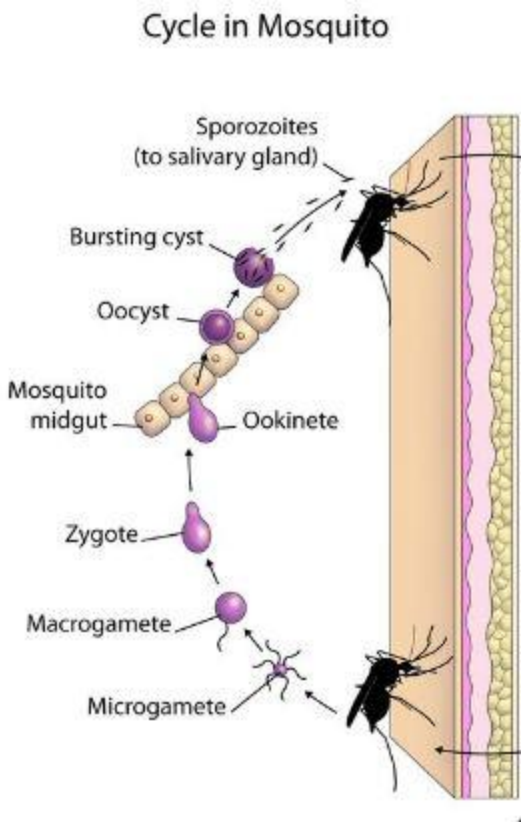


Figure 2-B: Lifecycle of *Plasmodium* in vector host

2.3. Matrix Metalloproteinases

Matrix metalloproteinases (MMPs) comprise a family of enzymes that cleave protein substrates based on a conserved mechanism involving activation of an active site-bound water molecule by a Zn^{2+}ion. Although the catalytic domain of MMPs is structurally highly similar, there are many differences with respect to substrate specificity, cellular and tissue localization, membrane binding and regulation that make this a very versatile family of enzymes with a multitude of physiological functions, many of which are still not fully understood. In recent years the number of identified zinc metalloproteinases/ peptidases has increased dramatically. Members of this superfamily of enzymes are involved in processes as diverse as embryonic development and bone formation, tetanus and botulism toxins, reproduction, arthritis and cancer (Hooper, 1994). Unique signatures within the amino acid sequences of the zinc metalloproteinases have been recently identified which place the enzymes into distinct family groups on the basis of sequence and structural similarities (Jongeneel *et al*., 1989; Vallee and Auld, 1989; Jiang and Bond, 1992; Bode *et al*., 1993; Rawlings and Barrett, 1993). Based on the

sequences around the HEbxH motif zinc metalloproteinases have been classified into five distinct families: thermolysin, astacin, serratia, matrixin, and reprolysin metalloproteinases (Jiang and Bond, 1992). The latter four families have an extended zinc binding site, HEbxHxbGbxHz, where the third histidine acts as the third zinc ligand instead of the more distant glutamic acid in thermolysin. Following the determination of the crystal structures of members of two of these families (astacin from crayfish and adamalysin II (reprolysin family) from snake venom), Bode *et al*., 1993, further classified these latter four families into a superfamily, the 'metzincins', as they all possess a methionine containing turn of similar conformation (the Met-turn). They also suggested that the larger superfamily of zinc metalloproteinases possessing the HEbxH motif be termed the 'zincins'. Although the HEbxH motif has been used extensively to identify zinc binding sites in metalloproteinase when new amino acid sequences are obtained, at least three other zinc binding motifs have been identified in zinc metalloproteinases (Hooper, 1994).

Anopheles metalloproteinase has been shown to have an agonistic role during sporogonic development of *Plasmodium falciparum*. Matrix metalloproteinase have been shown to be involved in vector competences (Goulielmaki *et al*., 2014).

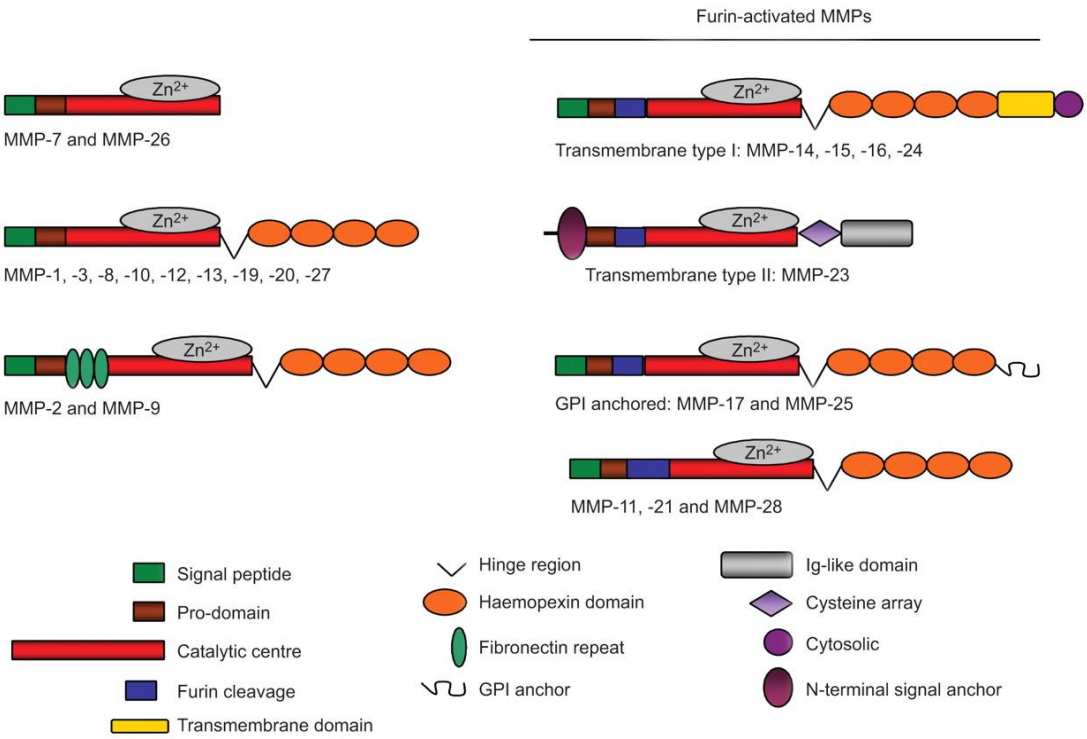


Figure 3: Schematic of the domain architecture of human matrix metalloproteinases. Most MMPs contain a Prodomain (Brown), catalytic domain (Red), a linker (hinge region, V) and a hemopexin domain (Orange). Furin-activated MMPs have a basic (RX{K/R}R sequence before the catalytic domain. Two MMPs (MMP-2 and MMP-9) have 3 fibronectin-like repeats (green) in the catalytic domain prior to the catalytic Zn2+ ion-binding site. Four MMPs (MMP-14, -15, -16, and -24) are anchored to the cell membrane via a c-terminal type-1 transmembrane domain (Yellow) and two MMPs (MMP-17 and -25) are tethered by a glycosylphosphatidylinositol (GPI)-anchor. In contrast MMP-23 is anchored by an N-terminal type-II transmembrane domain (grey). The two minimal domain MMPs (MMP-7, -26) and MMP-23 lack the hemopexin domain and in MMP-23 this domain is replaced by a C-terminal cysteine-rich toxin-like (light purple) domain and an immunoglobulin-like cell adhesion molecule domain (grey)

2.3.1. Other Metalloproteases

The zincins are those zinc metalloproteinases which contain the HEbxH short zinc binding consensus sequence. The zincins are subdivided into gluzincins, metzincins, astacins, serratia and reprolysin.

A number of zinc metalloproteinases have the HEbxH short zinc binding consensus sequence containing the first two zinc ligands and a glutamic acid as the third zinc binding ligand, e.g. thermolysin, endopeptidase-24.11, leukotriene A, hydrolase, etc.

In contrast to the gluzincins, the metzincins have longer zinc binding consensus sequence HEBXHXBGBXH which contains three of the zinc ligands. In addition, this superfamily has a methionine-containing turn of similar conformation (the "Met-turn") (Bode *et al*., 1993). The individual families are distinguished by (i) the residue following the third histidine zinc ligand in the above motif, and (ii) the residues surrounding the methionine in the Met-turn

The astacin family, typified by astacin, a digestive enzyme from the crayfish *Astacus astacus,* consists of several proteins from diverse sources including mammalian metalloendoproteinases, such as meprin (EC 3.4.24.18), and developmentally regulated proteins of man, fruit fly, frog and sea urchin. As with the other families constituting the superfamily of the metzincins three of the zinc ligands are contained within the metzincin consensus sequence which lies within the longer family signature sequence HEBXHXBGFXHEXXRXDRD. One of the distinguishing features of the astacin family is the glutamic acid residue following the third zinc ligating histidine. In addition, in this family, there is a somewhat distant fifth zinc ligand a tyrosine (a bound water molecule being the fourth) in a second highly conserved region which also contains the Met- turn, SBMHY (Bode *et al*., 1993), thus the zinc is penta-coordinated with novel trigonal-bipyramidal geometry.

The Serratia family, which contains several plant pathogen bacterial extracellular proteases including a protease from *Serratia* sp. and protease B and C from *Erwinia chrysanthemi* (Nakahama *et al*., 1986 and Dahler *et al*., 1990), also contains the longer metzincin consensus sequence for the three histidine ligands but in this case the third histidine is followed by a conserved proline. As with the astacin family there is a potential tyrosine fifth zinc ligand in the conserved Met- turn consensus region of SBMSY

The reprolysin family consists of several snake venom proteases, including hemorrhagic toxin and non-hemorrhagic proteins and a number of mammalian reproductive proteins. In this family the third histidine in the consensus sequence containing the three zinc ligands is followed

by a conserved aspartic acid. Unlike the astacin and serratia families the reprolysin family lacks a fifth zinc ligand, leaving the zinc tetrahedrally coordinated (Bode *et al*., 1993 and Gomis-Ruth *et al*., 1993). In place of the tyrosine in the Met- turn is a conserved proline in the consensus sequence CIMXP.

The matrixin family consists of mammalian collagenases, gelatinases, and stromelysins. In this family the metzincin superfamily consensus sequence for the three histidines is followed by a conserved serine. As with the reprolysin family the Met-turn lacks a tyrosine which is replaced by a conserved proline in the consensus sequence ABMYP.

A small group of zinc metalloproteinases are characterized by an inverted zinc binding motif HXXEH for which the name ‘inverzincins’ was proposed. This family, which includes the human, rat and *Drosophila* insulin-degrading enzymes, *Escherichia coli* protease III (pitrilysin) and a yeast processing-enhancing protein, possess an inverted zincin motif lying in the consensus sequence GXXHBXEHBXBXG. Recently the third zinc ligand has been identified as a glutamic acid laying some 82 amino acid residues C-terminal to this motif but not in any consensus sequence (Becker and Roth, 1993).

The carboxypeptidase family, typified by carboxypeptidases A and B but including carboxypeptidases H, M, N, U, mast cell carboxypeptidase A, carboxypeptidase T from *Thermoactinomyces vulgaris* and a carboxypeptidase from *Streptomyces griseus* have a unique short zinc binding motif containing the first two ligands, histidine and glutamic acid, with the third zinc ligand, a histidine, located some distance (108-135 amino acid residues) C-terminal to this motif This family can be further subdivided into three distinct groups on the basis of the sequence around the zinc binding ligands. The first group including carboxypeptidases A and B has the zinc ligands located in the consensus sequences DXGBHXREWBXXA and BHSYSQ. The second group (carboxypeptidase T and the *Streptomyces* carboxypeptidase) are similar in sequence to the above group with the consensus sequences TAXXHAREI-ILTVE and FHTYSE. In contrast the third group (carboxypeptidases H, M and N) has somewhat different consensus sequences BXNMHGXEXBGRE and LHGGXB (Hooper, 1994)

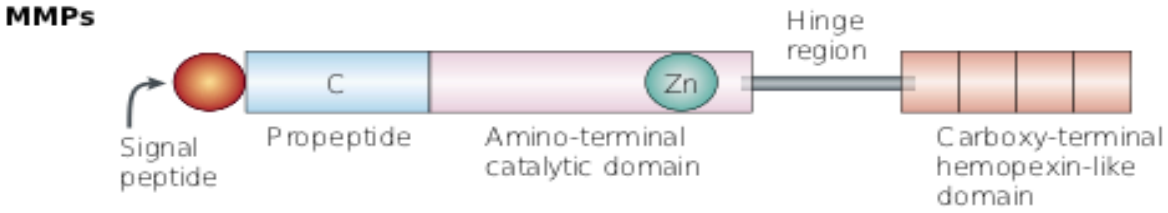


Figure 4: Generic Structure of matrix metallopeptidase. This figure shows the standard structure of MMPs the three main domains, a hinge region and a signal peptide.

Gelatinases (MMP-2 and MMP-9) have a fibronectin type II-like domain inserted into the catalytic site. The hemopexin like domain of MMP contains four repeats. The first and the fourth repeat are connected by a disulphide bridge. The 'C' denotes the cysteine residue that attaches to the catalytic domain to keep the enzyme in the inactive state (Yong *et al*., 2001).

2.3.2 *Drosophila melanogaster* MMP

In the fruit fly *D. Melanogaster*, two MMPs have been identified (Llano *et al*., 2002; Llano *et al*., 2000), which control larval tracheal growth and events of pupal morphogenesis (Page-McCaw *et al*., 2003). This pioneer study generated for the first time mutant organisms which were completely depleted for MMP activity and provided deeper understanding of in vivo roles of individual MMP in the fly development. Additionally, a recent study indicated that both *Drosophila* MMPs modulate the responses of embryonic motor axons of defined neuronal populations to specific guidance cues (Miller *et al*., 2008), indicating suitability of insect model organisms to elucidate novel MMP functions.

2.4. Bioinformatic Programs

2.4.1 BLAST

The BLAST programs are widely used tools for searching protein and DNA databases for sequence similarity. BLAST can be used to search protein database using a DNA query or search a DNA database using a protein query (Altschul *et al*., 1990). These databases are usually hosted

by National Centre for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov/) among other databases.

2.4.2. Determination of domains.

The InterProScan (http://www.ebi.ac.uk/Tools/pfa/iprscan/) (Zdobnov *et al*., 2001) is used for functional and structural analysis of protein sequences. InterProScan is a classification database that provides predictive information about protein sequences. It classifies proteins into families and predicting the presence of domains and important sites. InterPro uses predictive models, known as signatures, provided by several different databases (referred to as member databases) that make up the interpro consortium. Different member databases use different methods to construct their signatures and they have their own particular focus of interest: structural and/or functional domains, protein families, or protein features such as active sites or binding sites. InterProScan is a tool that combines different protein signature recognition from the InterPro member databases, currently- PROSITE, PRINTS, Pfam, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, Gene3D and PANTHER (Mulder *et al*., 2007).

2.4.3 Structural Analysis.

Many reactions within a cell are governed by molecular structure and hence it is very difficult to perform certain types of analysis on sequences without also inferring structure for example we cannot determine how proteins interact with ligands, co-factors from sequence analysis alone. Therefore, the determination of the structure and function of a novel protein is the cornerstone of modern and future biology. Despite the progress of high-throughput structural genomics only 80,000 protein structures have been experimentally determined (Kelley and Sternberg, 2009) and are found in the Protein Data Bank (PDB) (http://www.rcsb.org/pdb). Therefore this calls for the development of powerful techniques that will predict protein structure *ab initio* (computational methods grounded in the simulation of the folding process). Many advances in homology and analogy detection have been achieved in the past decade starting with sequence-structure threading (Sippl, 1990) and structural profiles (Godzik and Skolnick, 1992), including the use of predicted secondary structures (Fischer and Eisenberg, 1996), tertiary structure profiles (Kelley e*t al*., 2000), Hidden Markov Models (Karplus *et al*., 1998 and Eddy,

1998) and most recently profile-profile matching algorithms (Ohlson *et al*., 2004; Panchenko, 2003 and Jaroszewski *et al*., 2005). Some of the software, databases and tools used in structure prediction are discussed below.

Protein Data Bank (PDB)

The Protein Data Bank (PDB; http://www.rcsb.org/pdb/) is the single worldwide archive of structural data of biological macromolecules. The Protein Data Bank (PDB) was established in the 1970's at the Brookhaven National laboratories (BNL) as an archive for biological macromolecular crystal structure. By the 1980's the number of biological structures deposited increased dramatically (Berman *et al*., 2000) and this was due to the improvement in technology for all aspects of the crystallographic process. Depositors in the PDB have varying expertise ranging from X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy, cryoelectron microscopy and theoretical modeling (Berman *et al*., 2006).

MODELLER

MODELLER is a computer program for comparative protein structure modeling (http://salilab.org/modeller) (Sali and Blundell, 1993 and Fiser *et al*., 2000). In MODELLER the input is an alignment file of the sequence to be modeled and the template structure and a script file. MODELLER then generates a model containing all non-hydrogen atoms. MODELLER performs comparative modeling by satisfaction of spatial restraints (Sali and Blundell, 1993). The spatial restraints can be derived from a number of different sources and these include related protein structures (comparative modeling), NMR experiments (NMR refinement), rules of secondary structure packing (Combinatorial modeling), cross-linking experiments, fluorescence spectroscopy, image reconstruction in electron microscopy, site-directed mutagenesis, intuition, residue-residue and atom-atom potentials of mean force. These restraints can operate on distances, angles, dihedral angles, pair of dihedral angles and some other spatial features defined by atoms or pseudo atoms (Sali and Blundell, 1993; Mackerell *et al*., 1998 and Sali and Overington, 1994). MODELLER can also perform additional auxiliary tasks such as alignment of two proteins sequences or their profiles, multiple alignment of protein sequences and/or structures, calculation of phylogenetic trees and de novo modeling of loops in protein structures (Fiser *et al*., 2000).

MetaMQAPII

Evaluation of model accuracy is an essential step in protein structure prediction. The existing methods for quality assessment of protein models (MQAPs) are usually based either on a physical effective energy which can be obtained from fundamental analysis of particle forces or on an empirical pseudo energy derived from known protein structures (Lazaeidis and Karplus, 2000). The evaluation of protein model quality data can be done on MetaMQAPII (Pawlowski *et al*., 2008) which is made up of 8 Model Quality Assessment Programs (MQAP) methods: VERIFY3D (Luthy *et al*., 1992), PROSA2003 (Sippl,1993), PROVE (Pontius *et al*., 1996), ANOLEA (Melo and Feytmans, 1998), BALA-SNAPP (Krishnamoorthy and Tropsha, 2003), TUNE (Lin *et al*., 2002), REFINER (Boniecki *et al*., 2003) and PROQRES (Elofsson, 2006).
MetaMAQP measures the quality of a modeled 3D structure by use of a GDT_TS score. The GDT ("Global Distance Test') algorithm searches for the largest set of residues that deviates no more than a specified distance cutoff. Results are reported as the percentage of residues under a given distance cutoff. A popular measure is the "GDT Total Score",

$$GDT\_TS = (P_1 + P_2 + P_4 + P_8)/4,$$

Where, P_d is the fraction of residues that can be superimposed under a distance cutoff of d Å, which reduces the dependence on the choice of the cutoff by averaging over four different distance cutoff values.

Ramachandran Plots

Methods have been developed to assess the stereochemical quality of any protein structure both globally and locally using various criteria. Several parameters can be derived from the coordinates of a given structure. Global parameters include the distribution of phi, psi and chi 1 torsion angles, and hydrogen bond energies. There are clear correlations between these parameters and resolution; as the resolution improves, the distribution of the parameters becomes more clustered. Additional indicators of local irregularity include proline phi angles, peptide bond planarities, disulfide bond lengths, and their chi 3 torsion angles. These stereochemical parameters have been used to generate measures of stereochemical quality which provide a simple guide as to the reliability of a structure, in addition to the most important measures, resolution and R-factor.

Ramachandran Plots are determined from the PROCHECK software found at the Structural Analysis and Verification Server (http://nihserver.mbi.ucla.edu/SAVES/).The

PROCHECK suite of programs provides a detailed check on the stereochemistry of a protein structure. Its outputs comprise a number of plots in PostScript format and a comprehensive residue-by-residue listing. These give an assessment of the overall quality of the structure as compared with well refined structures of the same resolution and also highlight regions that may need further investigation. The PROCHECK programs are useful for assessing the quality not only of protein structures in the process of being solved but also of existing structures and of those being modeled on known structures (Laskowski *et al*., 1993).

2.4.4 Characterization of Matrix Metalloproteinases into Protein Families

Multiple sequence alignment (MSA) is a method most widely used in molecular biology to align a set of amino acids or nucleotide sequences (Feng and Doolittle, 1987). The most closely related sequences are aligned first and then these groups are gradually aligned together, keeping the early alignments fixed. This only works when the sequences are only closely related. In more difficult cases where the sequences are less than 30% identical then this automatic method doesn't becomes unreliable (Thompson *et al*., 1997). The automatic alignments generated have to be manually or automatically refined. Several sequences editors and viewers have been developed to allow the user to manually view the alignments and manually modify or edit an alignment (De Rijik and De Wachter, 1993; Galtier *et al*., 1996). MSA viewers/editors include SeaView (http://pbil.univ-lyon1.fr/software/seaview.html) and Jalview (http://www.jalview.org). CLUSTAL O webserver (http://www.ebi.ac.uk/Tools/msa/clustalo/) was used to generate multiple sequence alignments. Clustal Omega (Goujon *et al*., 2010) is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between three or more sequences.

CHAPTER THREE
MATERIALS AND METHODS

This study was based on protein sequence data of *An. gambiae* PEST strain and *D. melanogaster*. The study employed *in silico* approaches using various web-based, stand-alone bioinformatics tools and browsers to perform database searches, sequence and functional analyses. Other computational methods included in-house program scripts. *Drosophila melanogaster* was used as the reference insect as it has been studied widely using bioinformatics tools and also in the wet laboratories. *Anopheles gambiae* has 2 genes for MMP; AGAP006904, AGAP003929. AGAP006904 has 3 transcripts (transcript-A 544bp, transcript-B 1710bp, and transcript-C 1681bp). In this study we will try to amplify the 2 genes and all the transcript.

3.1. Determination of Matrix Metalloproteinase Domain

The sequences for *Drosophila melanogaster* MMP were searched and retrieved from the National Center for Biotechnology Information (NCBI) database. These sequences were then inserted into a BLASTp program to search for similar proteins in the *Anopheles gambiae* genome and the default BLASTp options were used. Three genes with relative homologies to *Drosophila* MMP were found. They had the following accession numbers AGAP006904, AGAP011870, and AGAP003929. Protein sequences (AGAP006904 and AGAP003929) which belong to *An. gambiae* MMP were downloaded from NCBI. The protein sequences were then submitted to the InterProScan server found at (www.ebi.ac.uk/Tools/pfa/iprscan) to determine inherent domains.

3.2. Predictions of 3-D Structure

The three-dimensional structure of human proMMP1 (PDB ID: 1SU3 chain A and B, at 2.20 Å resolution) and the three-dimensional structure of Hemopexin-like domain of MMP14 (PDB ID: 3C7X chain A, at 1.7 Å resolution) were used as template for homology modeling. The comparative modeling of AGAP006904 and AGAP003929 was performed using a restraint-based approach using MODELLER9v10 (Sali and Blundell, 1993).A set of 50 models for each target protein was constructed. The resulting 3-D structure models were sorted according to scores calculated from discrete optimized protein energy (DOPE) scoring function (Shen and Sali, 2006). Models with a DOPE score of -3 to -5 were selected for further downstream

analysis. Model Quality assessment programs, MetaMQAPII, ProSA and SAVeS were used to validate the models. The best model was selected based on a GDT_TS score of more than 65%, for Ramachandran plot more than 85% of residues found in allowed region and a z-score of -7.35 which is relative to that of native proteins. A Root Mean Square Deviation (RMSD) was then calculated. The models were then viewed and generated using PyMOL (http://www.pymol.org/).

3.3. Characterization of Matrix Metalloproteinases.

The protein sequences of *Anopheles gambiae* strain PEST, AGAP006904 and AGAP003929 were characterized into superfamily and family. These were compared to protein sequences that have been grouped into the various superfamilies and families. A fasta file of similar sequences was obtained from a non-redundant protein database running BLASTP, which is protein to protein blast. A multiple sequence alignment was generated using ClustalO.

3.3.1. Gene Validation

To demonstrate that the gene used for the homology model is present in *Anopheles gambiae*, polymerase chain reaction amplification of the MMP gene from genomic DNA of *Anopheles gambiae* was performed. *Anopheles gambiae* was obtained from the Animal Rearing and containment Unit (ARCU) at the International Centre for Insect Physiology and Ecology (ICIPE). PCR products were purified and sequenced and a multiple sequence alignment was generated to compare these MMP genes to other known MMP genes.

3.3.2. Primer Design

Two putative MMP; AGAP006904 and AGAP003929 sequences were retrieved from NCBI (http://www.ncbi.nlm.nih.gov/). The two sequences were translated to protein sequences using ExPASY translate tool (http://web.expasy.org/translate/) to know where the longest open reading frames (ORF) for MMP lie. Based on DNA sequence of the target identified frames, primers were designed manually to target the longest ORF of the two sequences. The best primer parameters were selected using Sequence Manipulation Suite (http://www.bioinformatics.org/sms/).

Table 1: MMP Primers: Sequences and their expected product sizes.

Gene	Transcript	Primer Sequence (5'-3')	Product Size
(AGAP006904)	A	Forward: ggaacccgacgagcgggaacctgct Reverse: cgaaacacgggatcatatccacggt	544bp
	B	Forward: actctaccaccactacccgcggcaacgga Reverse: tactgacactggtgccggtgatcgtggaa	1710bp
	C	Forward: tactgacactggtgccggtgatcgtggaa Reverse: ccagcaagaaacccgcgagggctgttgta	1681bp
(AGAP003929)		Forward: ttgccaccgcgttctgtgaagag Reverse: ccacttggcgagaaaatgcacta	2049bp

3.3.3. Extraction of RNA

The three developmental stages of female *An. gambiae* (larvae, pupae and adult) were obtained from the Animal Rearing and Containment Unit at the International Centre for Insect Physiology and Ecology (ICIPE). The mosquitoes were anaesthetized by chilling on ice for 5 minutes, then placed in a 1.5 ml microcentrifuge tubes. RNA was extracted using the Direct-zol RNA MiniPrep (Fermentas). Five hundred microliters of TRI-reagent was added to the Eppendorf tubes and the samples were homogenized using a squisher homogenizer. Particulate matter was removed by centrifuging at 12,000rpm for one minute. Five hundred microliters of ethanol was added directly to the sample homogenate in the TRI-reagent. The mixtures were then loaded into a Zymo-spin IIC column in a collection tube and centrifuged at 12,000rpm for 1 minute. The columns were then transferred into a collection tubes and the flow through were discarded. Four hundred microliters of RNA wash buffer was added to the columns and centrifuged at 12000rpm for 1 minute and the flow through were discarded. DNase 1 was added to the columns and

incubated at 37^{o}C for 15 minutes and centrifuged at 12,000rpm for 30 seconds. The columns were washed twice and centrifuged at 12,000rpm for 1 minute using 400 µl Direct-zol™ RNA PreWash5 and the flow-through was discarded. Seven hundred microliters of RNA Wash Buffer5 was added to the columns and centrifuged at 12,000rpm for 1 minute to ensure complete removal of the wash buffer. The column was carefully transferred into 3 RNase-free tubes and RNA was eluted by adding 30 µl of DNase/RNase-Free water directly to the column matrix and centrifuged at 20,000rpm for 1 minute.

3.3.4. Purity.

The purity of the isolated DNA was determined by reading absorbance at 260nm and 280 nm. The ratio of $Å_{260}/Å_{280}$ indicated the purity of the sample. Pure RNA samples exhibit $Å_{260}/Å_{280}$ ratios of 2.0. A lower ratio than 1.7 confirms contamination of the sample.

3.3.5. First Strand cDNA synthesis

cDNA was synthesized using the RevertAid H Minus First Strand cDNA Synthesis Kit (Thermo Scientific). A twenty microliter (20µl) first strand cDNA synthesis reaction was set up comprising of 1 µl of total RNA, 1 µl of oligo $(dT)_{18}$ primer, 10 µl of nuclease-free water, to make a total volume of 12 µl. Four microliter of 5X Reaction Buffer, 1 µl of RiboLock RNase Inhibitor (20 u/µl), 2 µl of 10 mM dNTP Mix, 1 µl of RevertAid H Minus M-MuLV Reverse Transcriptase (200 u/µl). The mixture was mixed gently. For oligo $(dT)_{18}$ or gene-specific primed cDNA synthesis, incubation was done for 60 min at 42°C. The reaction was terminated by heating at 70°C for 5 min. The reverse transcription reaction product was directly used in PCR application.

3.3.6. Polymerase chain reaction

Polymerase chain reaction (PCR) was done using. DreamTaq Green PCR master mix (Thermo Scientific, Waltham, MA) which has a 3'-5' exonuclease activity that increases the fidelity of the amplification. A 10 µl reaction mixture were set up as follows: 3 µl of nuclease-free water, 5 µl of Dream Taq master mix, 0.5 µM of each primer and 1 µl of 1st strand cDNA. Amplification was done using ProFlex thermal cycler (applied biosystems by life technologies, California) using the following cycling conditions, AGAP006904-PA initial denaturation at 95^{o}C for 3 minutes, then 35 cycles of denaturation at 95^{o}C for 1 minute, annealing at 61^{o}C for 1 minute and

Extension at 72°C for 1 minute. This was followed by a final extension at 72°C for 10 minutes. The PCR products were electrophoresed on 1.2% ethidium bromide stained Agarose gel for 1 hour 30 minutes at 70 V (BIO-RAD model 200/2.0 POWER SUPPLY). The gel was photographed under UV light. Amplification was done using ProFlex thermal cycler (applied biosystems by life technologies, California) using the following cycling conditions, AGAP006904-PB initial denaturation at 95°C for 3 minutes, denaturation at 95°C for 20 seconds, annealing at 67°C for 1 minute, Extension at 72°C for 1 minute and a final extension at 72°C for 15 minutes for 35 cycles. The PCR products were electrophoresed on 1.2% ethidium bromide stained Agarose gel for 1 hour 30 minutes at 70 V (BIO-RAD model 200/2.0 POWER SUPPLY). The gel was photographed under UV light

Amplification was done using ProFlex thermal cycler (applied biosystems by life technologies, California) using the following cycling conditions, AGAP006904-PC initial denaturation at 95°C for 3 minutes, denaturation at 95°C for 1 minute, annealing at 55°C for 1 minute, Extension at 72°C for 1 minute and a final extension at 72°C for 15 minutes for 35 cycles. The PCR products were electrophoresed on 1.2% ethidium bromide stained Agarose gel for 1 hour 30 minutes at 70 V (BIO-RAD model 200/2.0 POWER SUPPLY). The gel was photographed under UV light.

Amplification was done using ProFlex thermal cycler (applied biosystems by life technologies, California) using the following cycling conditions, AGAP003929 initial denaturation at 95°C for 3 minutes, denaturation at 95°C for 20 seconds, annealing at 50°C for 1 minute, Extension at 72°C for 1 minute and a final extension at 72°C for 10 minutes for 35 cycles. The PCR products were electrophoresed on 1.2% ethidium bromide stained Agarose gel for 1 hour 30 minutes at 70 V (BIO-RAD model 200/2.0 POWER SUPPLY). The gel was photographed under UV light. Unfortunately, AGAP003929 was not amplified at the desired length. Therefore, the results for PCR amplification of AGAP003929 will not be presented.

3.3.7. PCR product purification

The PCR products were excised from the gel using a sterile blade. The PCR products (total volume of 25 µl) were purified using the ISOLATE II PCR and Gel Kit (BIOLINE). One volume of sample was mixed with 2 volume of Binding buffer CB. An ISOLATE II PCR and Gel column were placed in a 2ml collection tube and the samples were loaded and centrifuged at 11000rpm for 30s and the flow through was discarded. Seven hundred microliters (700) µl of wash buffer CW was added to ISOLATE II PCR and Gel column and centrifuged at 11000rpm

for 30 seconds. The flow through was discarded and the column placed back in the collection tube. The ISOLATE II PCR and Gel column were centrifuged at 11,000rpm for 1 minute to remove residual ethanol. The ISOLATE II PCR and Gel column were placed in a 1.5ml microcentrifuge tube. 30 µl of elution buffer C was added directly onto the silica membrane, incubated at room temperature for 1 minute and centrifuged at 11,000xg for 1 minute. Five microliters of the clean products were loaded onto 1.2% ethidium-bromide stained agarose gel to confirm recovery of the cDNA before sequencing. The products were visualized under a UV illumination to confirm the size of the bands using DNA molecular weight marker (DNA marker, BioLabs) electrophoresed alongside the products.

3.3.8. Bioinformatic Analysis

AGAP006904-PA sequences were edited to remove ambiguous base calls and primer sequences using the BioEdit software program. A search to identify similar protein sequences to MMP1 transcript A was performed using tBLASTx algorithm of NCBI. The retrieved sequences were aligned using Clustal Omega.

CHAPTER FOUR
RESULTS AND DISCUSSION

4.1. Results

The results presented for the gene validation are from transcript A of AGAP006904 which was amplified in the lab. Transcript A is truncated and for us to understand the structure of MMP we will use transcript B sequences from NCBI, since transcript B has already been amplified.

4.1.1. Determination of Matrix Metalloproteinase's Domain

The domains of AGAP006904 transcript A were determined using the InterProScan server as summarized in Fig 4.

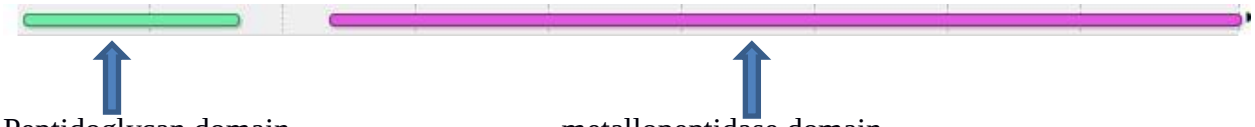


Figure 5 A: This shows the domains of transcript A of AGAP006904. This transcript has 2 domains (Peptidoglycan domain and metallopeptidase domain) and is the truncated transcript of the gene AGAP006904.

Figure 5 B: This shows the domains of transcript B of AGAP006904. This transcript has 3 domains (Peptidoglycan domain, metallopeptidase domain and the hemopexin domain) and is the longest transcript of the AGAP006904.

Peptidoglycan domain metallopeptidase domain hemopexin domain

Figure 5 C: This shows the domains of AGAP003929. This transcript has 3 domains (Peptidoglycan domain, metallopeptidase domain and the hemopexin domain).

The results show the domains for transcript A of AGAP006904, transcript B of AGAP006904 and AGAP003929. The peptidoglycan domain is a binding domain that is found on the N or C terminal of proteins, it found in the pfam domain. The metallopeptidase domain is the functional domain of metalloprotease and here is where catalysis takes place, it's found in the CATH family. The hemopexin domain is found in transcript B of AGAP006904 and AGAP003929, it consists of 4 repeats, and it binds to substrates and tissue inhibitors of MMP. Its classification is found in the SMART database.

4.1.2. Prediction of 3-D Structure

The 3D structure of both AGAP006904 and AGAP003929 were predicted using MODELLER and figures generated using PyMOL. However, after structural analysis and validation using DOPE score, MQAPs and RMSD, the structure of AGAP003929 was found to be erroneous and these errors included high energy Discrete Optimized Protein Scores (DOPE) and Root Mean Square Deviation (RMSD) scores. Therefore the results presented here are for the correctly predicted structure of AGAP006904.

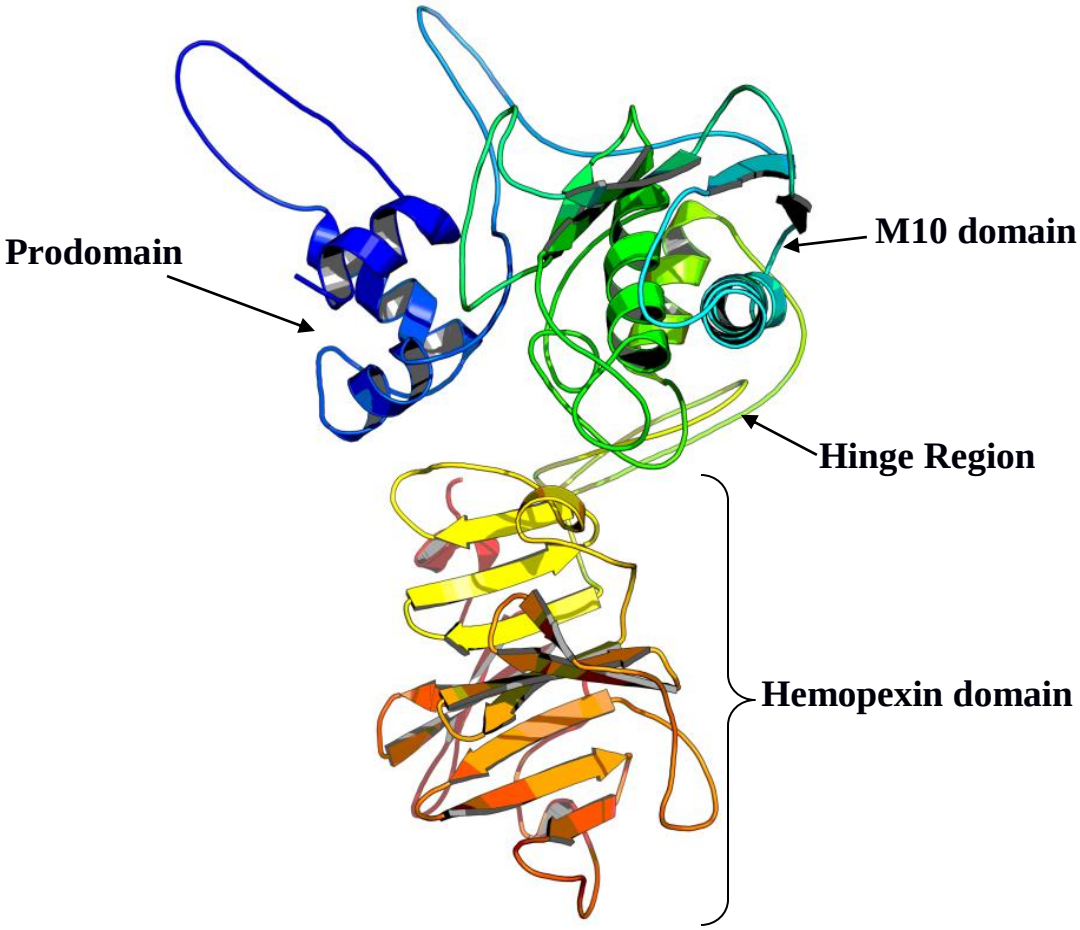


Figure 6: Predicted 3D structure. This model was generated by MODELLER and viewed using PyMOL. The figure shows 3 main domains and a linker region: Prodomain (blue), Metalloproteinase (cyan and green), Linker region (green) and Hemopexin (yellow-orange)

The structure shows four segments, a prodomain, a catalytic domain and a hinge region and a hemopexin domain (Hpx) (Figure 6). The ellipsoid shaped prodomain docks to the active site in the metalloproteinase (M10) domain. In addition, it interacts with the Hpx domain, contributing to a compact structural arrangement of matrix metalloproteinase. The overall fold of the catalytic domain and Hpx domain is characteristic of other MMPs.

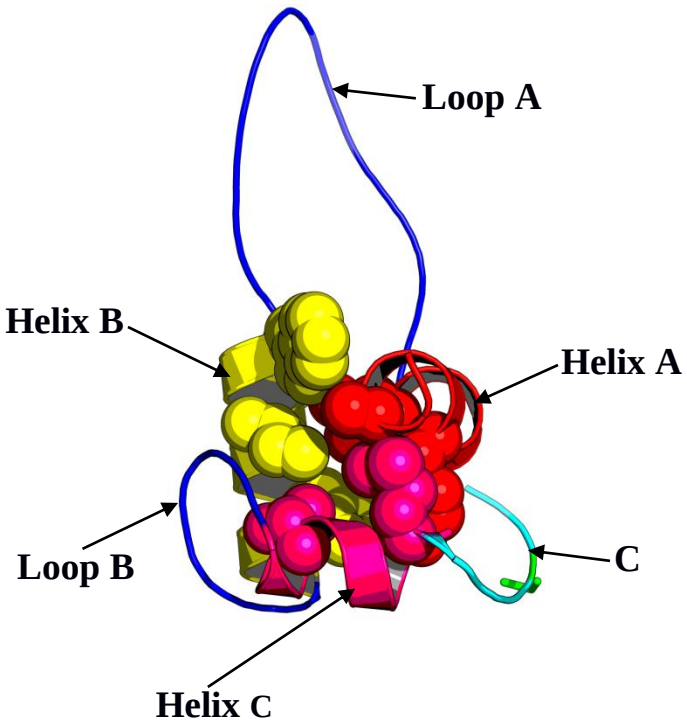


Figure 7: Prodomain. A shows the prodomain of modeled 3D structure of AGAP006904, it shows the side chains of amino acids that are used to stabilize the three helices through hydrophobic interactions. It also shows the three helices, two loops and the cysteine switch of the prodomain.

The prodomain. The main body of the prodomain is formed by three helix bundles with a left-handed twist and helices more or less perpendicular to each other stabilized by a hydrophobic domain. The orientation of Helix A (Gln^{2}-Gly^{11}) is almost parallel with edge strand in the proteinase domain. Helix A is followed by loop A (Tyr^{12}-Asp^{29}) which points away from the catalytic domain. Helix B (Thr^{30}-Ala^{42}) is directed toward the catalytic domain in an almost perpendicular orientation to the active site helix and edge strand. Helix B interacts with other residues in the prodomain provided by Helix A and C and loop B. At the end of Helix B residues Phe^{41} and Ala^{42} interact with residues His^{205}-Ser^{206}-Asp^{207} and Ala^{214}-Pro^{215} of the proteinase domain.

Loop B (Gly^{43}-Asp^{51}) turns up along helix B and then kinks at Gly^{48} towards helix C. Within the prodomain, loop B mainly interacts with helices B and C. It is stabilized by hydrophobic

interactions and hydrogen bonding; the side chain interaction of residue Asp^{51} with Gln^{39} from helix B especially contributes to this stabilization. Residues in this loop, in addition interact with the Hpx domain. This interaction is mainly hydrophobic involving Phe^{289}-Lys^{290}-Gly^{291} of the loop between β-sheets 2 and 3 on blade 1 of the Hpx domain and Gly^{43}-Leu^{44}-Asn^{45} in loop B and Asp^{51}-Glu^{53} of helix B in the prodomain. In addition the side chain of Glu^{53} forms a hydrogen bond to the side chain of Tyr^{295} and a weak hydrogen bond to Asn^{300}. Furthermore a main chain hydrogen bond is formed between Asn^{45} N and Phe^{289} CO. Helix C (Gly^{52}-Met^{58}) points towards the active side cleft of the catalytic domain. It mainly interacts with other helices in the prodomain and loop B.

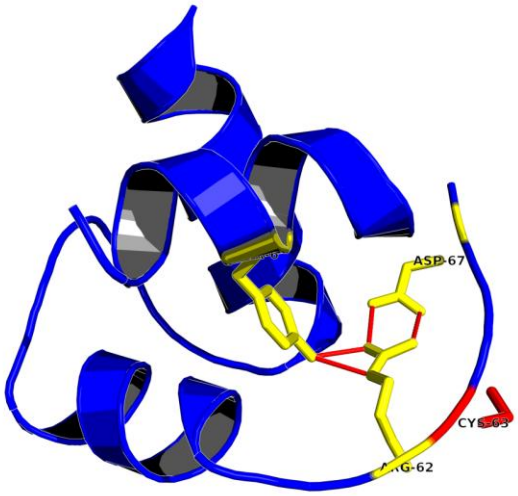


Figure 8: Interaction of Prodomain. This figure shows the interaction of Arg^{62} of C, which is highly conserved in MMPs with Tyr^{6} of HA and Asp^{67} of C. This interaction stabilizes the cystein switch loop against the back of the prodomain.

The cysteine switch sequence PRCGxPD enters the active side cleft on the right hand side in a standard orientation: In Figure 7 it enters on the bottom left side. The cysteine switch region makes β-sheet backbone contacts with the edge strand of the catalytic domain in a parallel orientation (residues Gly^{157} CO, Leu^{159}, and Ala^{160} CO) and with residues Pro^{165} CO and Tyr^{168} N in an antiparallel orientation. The cysteine switch loop also is bent through the Arg-Asp salt

bridge of the PRCGxPD motif and is located on top of the third histidine (His^{205}). The intact prodomain shield this salt bridge via one Tyr and two Phe side chains from bulk water. Pro^{61} is positioned in a hydrophobic pocket formed by catalytic domain residues Leu^{159}, Tyr^{217} and Tyr^{220}. Arg^{62} which is strictly conserved in MMPs, is oriented towards the interior of the prodomain and plays a role in stabilizing the cysteine switch loop region against a bulk of the prodomain with the help of Asp^{67}, with which it forms a salt bridge (Figure 8). Asp^{67} also interacts with His^{205} which is positioned closer to the catalytic zinc in proMMP. Because of the orientation of Cys^{63}, which coordinates with its Sγ to the active site Zn^{2+}, toward the catalytic zinc and Arg^{62} toward the prodomain, the S1' and S2' substrate specificity pockets are empty in contrast to when the proteinase domain interacts with the substrate. It is this interaction between Cys^{63} and the active site zinc that keeps the enzyme in its zymogen state.

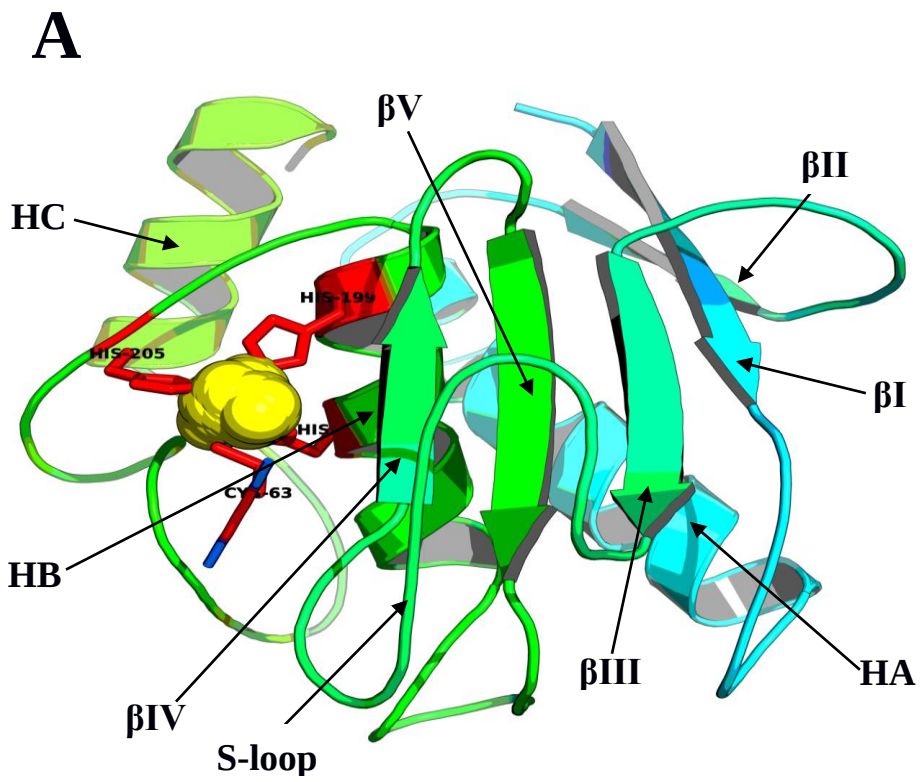


Figure 9 A: Metalloproteinase domain. This shows the active site of Anopheles MMP. The 3 histidine residues: His196, His^{199} and His^{205} and Cys^{63} coordinate the Zn^{2+} ion (Yellow). Structural perturbation in this region activates the proteinase.

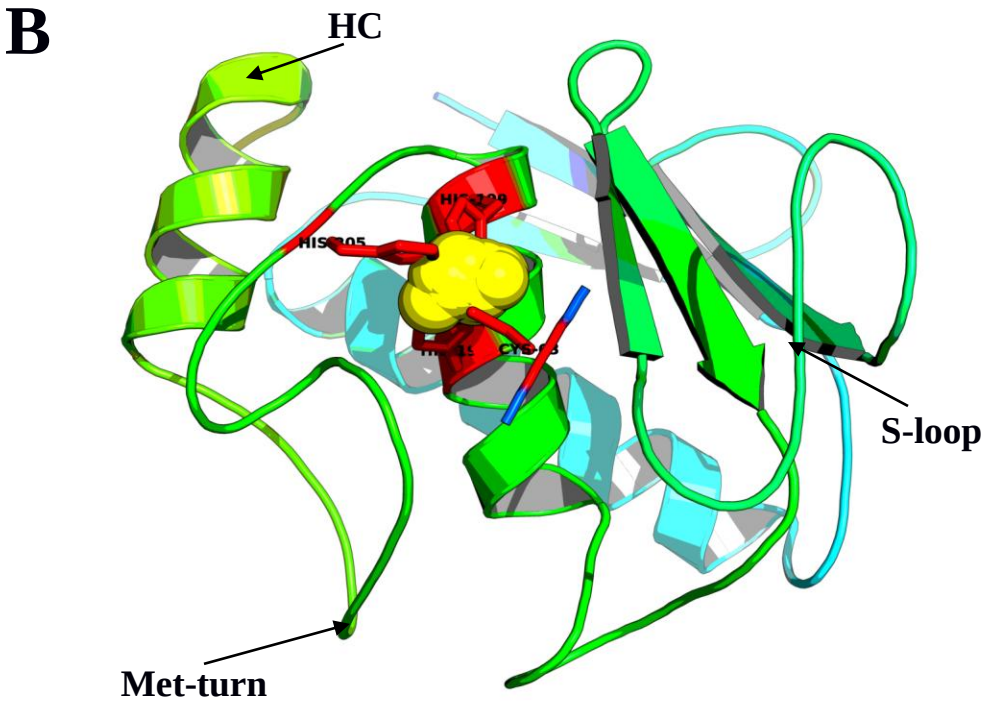


Figure 9 B: Metalloproteinase domain. This shows the met-turn that is conserved in all MMPs.

The structure of the metalloproteinase domain of *Anopheles* AGAP006904 is similar to those of other MMPs in showing an active site cleft notched into the front surface and separating the smaller "lower subdomain" from the larger "upper subdomain". This cleft extends horizontally and across the molecule and would bind a peptide substrate from left to right. The upper subdomain encompasses a characteristic five-stranded, highly twisted β-sheet flanked by three surface loops on its convex side and by two regular α-helices on its concave side. Four of the five β-strands are aligned in a parallel fashion, only the cleft sided edge strand (βIV) runs in the opposite direction. The chain passes β-sheet βI, α-helices HA, β-sheet βII, βIII, before entering the so-called "S-loop", which is fixed through the structural zinc ion (Figure 9 A). The structural zinc is liganded in a tetrahedral coordination sphere made up by His^{195}, Glu^{196}, His^{199} and His^{205}. After the βV strand, the chain passes through the large open βV-HB loop, a segment of high variability and a source of substrate specificity within the MMP family (Maskos and Bode, 2003).

It is then followed by the long horizontally extending active site helix HB. This helix includes the first and second Histidine of the zinc binding motif His^{195}, and His^{199} and Glu^{196}, the residue essential for catalysis. The chain then bends down, exhibits the third zinc ligand (His^{205}), followed by the 1-4 tight "Met-turn" (Figure 9 B) (Bode *et al*., 1993), and then forms the lower primed site surface with the Pro^{215}-Ser^{216}-Tyr^{217} wall-forming segment of the S1' specificity pocket. The S 1' apparently plays a significant role in determining the substrate specificity in the active enzymes. The domain is completed by the following 'specificity loop' and the C-terminal helix HC. The catalytic domain also contains two tetrahedrally –coordinated Zn^{2+}ions: a "structural" zinc ion and a "catalytic" zinc ion whose ligands include the side chain of the three histidyl residues in the signature HEbxHxbGbxHz sequence.

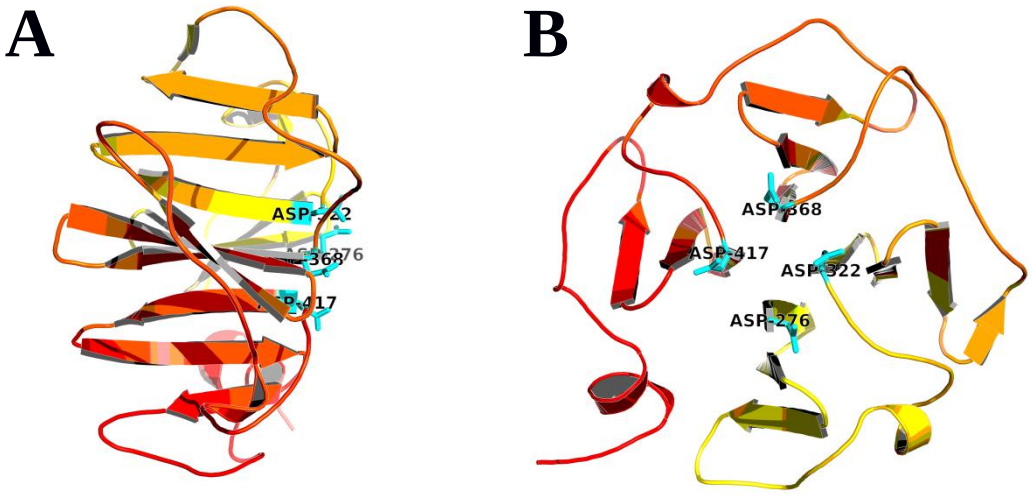


Figure 10: Hemopexin domain A and B: **Figure A** shows the hemopexin domain β-sheets. **Figure B** shows the four charged Asp residues arranged around the tunnel entrance.

The result shows that the C-terminal Hpx domain is connected to the catalytic metalloproteinase domain through a 39-residue linker. The Hpx shows approximately 4-fold symmetry with a structure like a four-bladed propeller (Li *et al*., 1995; Gohlke *et al*., 1996). Each blade is formed by three antiparallel β-strands and short peptide loops with short helices

connecting these blades. The first strands of each β-sheet contribute to the formation of a central solvent accessible channel. The strands of each twisted blade are connected in a W-like topology with the first strand forming the central pore and defining the direction of this channel. This arrangement is stabilized by a disulphide bridge formed Cys^{270} in the beginning of blade 1 and Cys^{463} at the end of blade IV. In all hemopexin-like domains (up to four) charged-uncompensated Asp residues are arranged around the tunnel entrance (Figure 10). These side chains form salt bridges to neighboring β-strand.

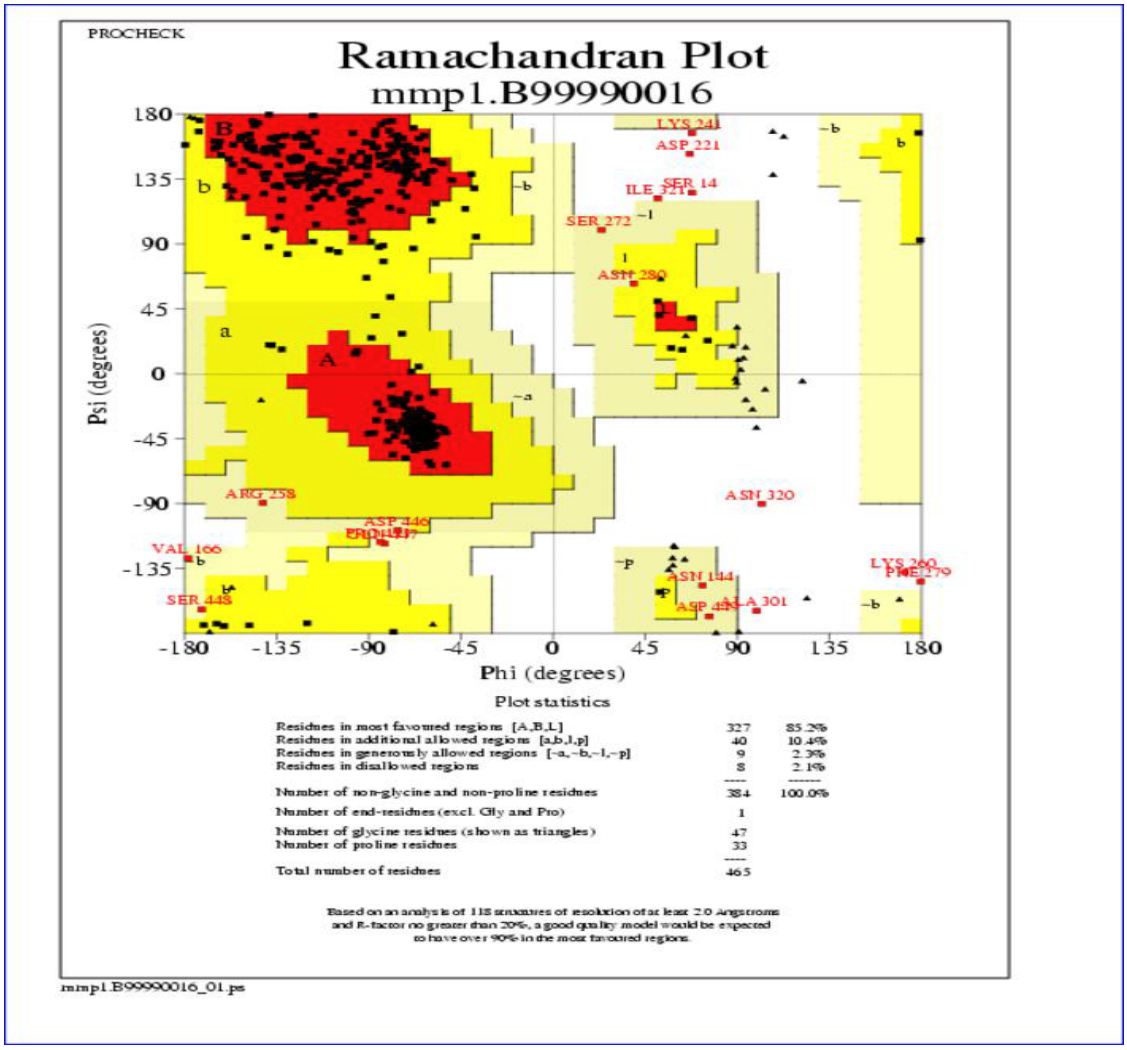
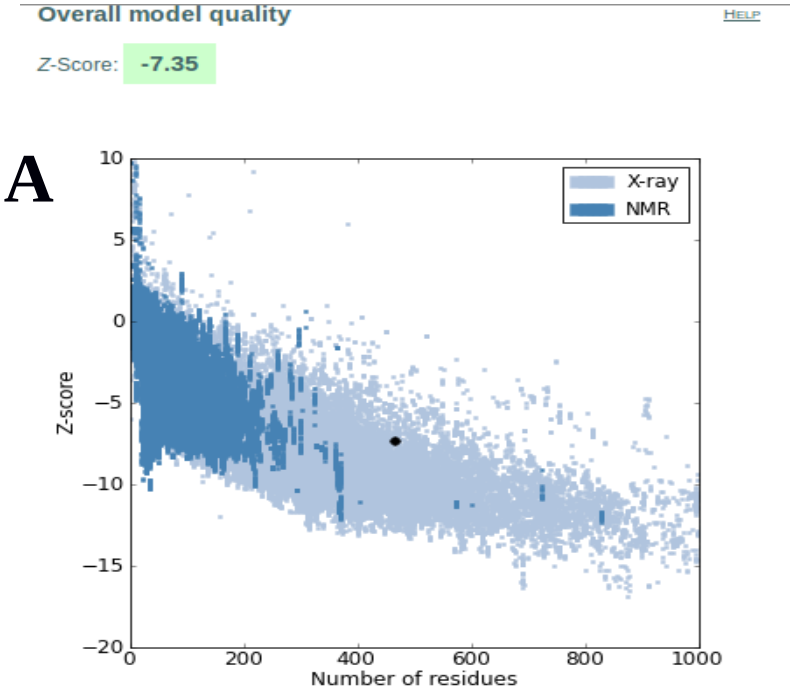


Figure 11: Ramachandran Plot. This figure shows a Ramachandran Plot for the modeled 3D structure of AGAP006904. The plot shows that most of the residues are found in the most favoured region (85.2%) and very few in the disallowed region (2.1%).

Figure 11 shows a ramachandran plot generated from the PROCHECK program, which is a model quality assessment program used for structure verification. The ramachandran plot is divided into four quadrants. The quadrants have three colors red, yellow and beige. The red sections of the quadrants are labeled using upper case letters A, B and L. The yellow sections of the quadrants are labeled using lower case letters a, b, l and p. The beige sections of the quadrants are labeled using lower case letters preceded by a hyphen (-) –a, -b, -l and –p. The black spots on the quadrants are amino acid residues found in the different colored sections of the four quadrants. Below the plot is a display of plot statistics.



.

Figure 12 A: ProSA. This shows that the modeled 3D structure of AGAP006904 is found within the range for proteins in the X-ray crystallography group and has a Z-score of -7.35, which is within the range of native proteins.

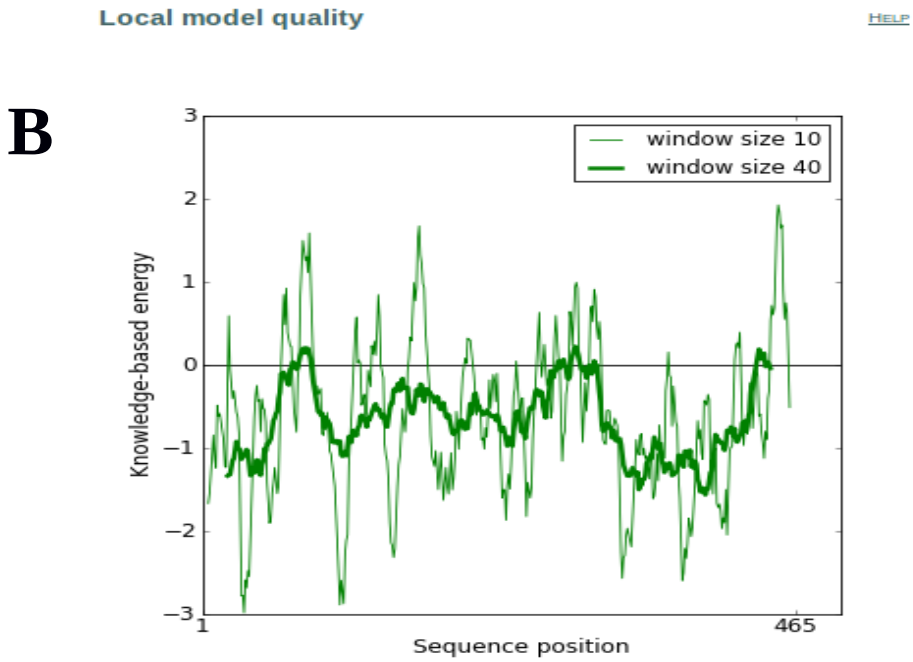


Figure 12 B: ProSA. This shows that most of the residues along the sequence of the protein are negative.

In a given structure the interaction energy e_{ij} between amino acid residues at positions *i* and *j* along the chain is the sum of the interaction energies between the atoms of the respective residues. The z-score obtained can be interpreted as an overall quality index of a particular fold. A more detailed view of the energy distribution in protein folds is obtained from the residue interaction energies e_{ij}. The interaction energies e_{ij}, *i*, *j*=1....,l form the energy matrix E of a conformation where the sequence length l corresponds to the dimension of the matrix. From E the interaction energy $e_i = \sum_j e_{ij}$ of a particular residue I with respect to all other residues is derived. When e*i* is plotted as a function of *i*, we obtain an energy graph displaying the energy distribution of a sequence structure pair in terms of sequence position. In energy graphs positive values point to strained sections of the chain whereas negative value corresponds to stable parts of the molecule. Figure 12 shows a ProSA output. Figure 12A shows a plot of number of residues against Z-score. This plot shows the distribution of structures that have been solved using X-ray crystallography (light blue) and Nuclear Magnetic Resonance (navy blue) based on their Z-scores and number of residues. The overall model quality is measured using z-score that is displayed on the top left hand side of the diagram. Figure 12 B shows a plot of sequence

position against knowledge-based energy. This plot also shows two lines; one corresponding to a window size of 40 (~40 residues) and the other a window size of 10 (~10 residues).

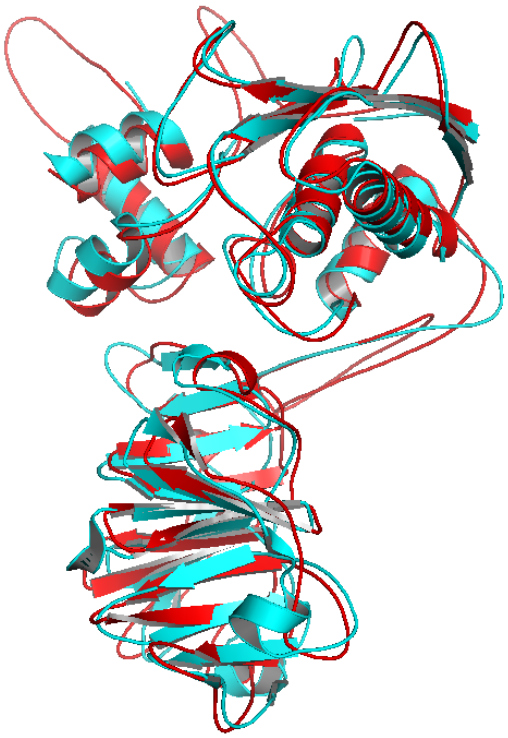


Figure 13: Root Mean Square Deviation (RMSD). The calculated RMSD is 1.695 Å which indicates a good model. The modeled structure (Red) and template file 1SU3.pdb (Cyan). The RMSD error range for proteins is 2 Å to 3 Å

Figure 13 shows the superimposition of two structures 1su3.pdb (cyan) and Anopheles MMP (red). This figure shows the almost perfect superimposition of this two structures. In this figure the loop between helix A and helix B in the prodomain of 1su3.pdb(cyan) are not visible in the electron density and is though to be flexible in soution (Jozic et al., 2004). The prodomain and the metalloproteinase domain show a good structural alignment between modeled strucutre (red) and 1su3.pdb (cyan).The linker region of the modeled structure (red) has a loop and does not align well with that of 1su3.pdb because 1su3.pdb has a 16-residue linker (Jozic et al., 2004) and the modeled structure (red) has a 39-residue linker.

Table 2: Nanodrop results of RNA

Sample ID	**Nucleic Acid conc**	**Unit**	**Å260**	**Å280**	**260/280**
Larvae	2533.1	ng/µl	63.327	30.793	2.06
Pupae	1751.6	ng/µl	43.789	21.251	2.06
Adult	129.4	ng/µl	3.234	1.676	1.93

Table 2 shows the nanodrop results of RNA extracted from the three developmental stages of *Anopheles gambiae*. The columns shows nucleic acid concentration which have to be high in order to rule out contamination during nucliec acid extraction procedures. Absorbance ratios 260/280 are indicative of the purity of DNA or RNA samples. In the 260/280 column we can conclude the extracted RNA is pure.

Table 3: Nanodrop results of cDNA.

Sample ID	**Nulceic Acid Conc.**	**Unit**	**Å260**	**Å280**	**260/280**
Larvae	2699.9	ng/µl	53.997	31.104	1.74
Pupae	2804.4	ng/µl	56.088	32.214	1.74
Adult	2582.9	ng/µl	51.658	30.175	1.71

Table 3 shows the nanodrop results of cDNA extracted from the three developmental stages of *Anopheles gambiae*. The columns shows nucleic acid concentration which have to be high in order to rule out contamination during cDNA synthesis. Absorbance ratios 260/280 are

indicative of the purity of cDNA samples. In the 260/280 column we can conclude that the cDNA synthesized are relatively pure.

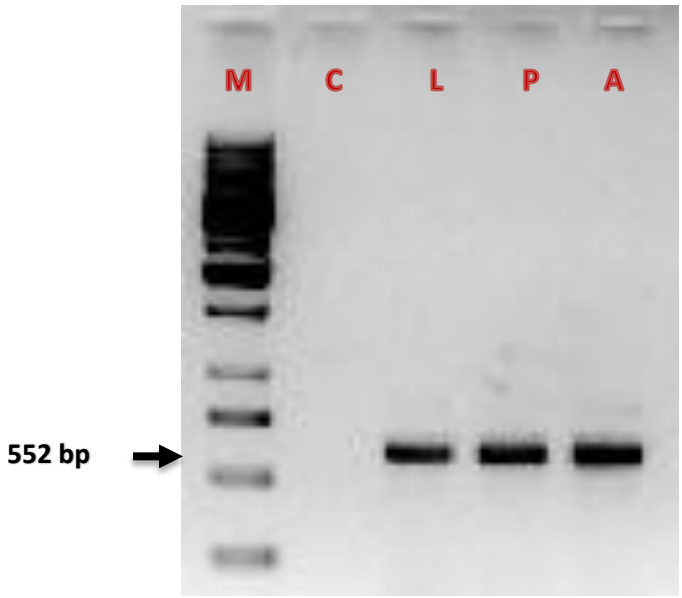


Figure 14: PCR products of AGAP006904-PA. **M**: 1kb DNA ladder (Thermo Scientific), **C**: negative control, **L**: Larvae, **P**: Pupae, **A**: Adult.

The gel shows the amplification of AGAP006904-PA**,** which is a 552 bp product. This gene has been amplified across all the developmental stages of *Anopheles gambiae* indicating that it is found in all the developmental stages. A tBLASTx analysis was run after the sequencing of the purified PCR product. The results indicated that it was similar to AGAP006904-PA that is found in the NCBI database.

```
Hydra_MMP                 -----------------MFISLGFNLFF-----IFYVAQSLPVEHIQVPRKTLDYLANLG
Tribolium_castaneum_1     -------------------MRHT-VISFTLFATLLQLSK-----SAPSGSSALLYLSQYG
Tribolium_castaneum       -------------------MRHT-VISFTLFATLLQLSK-----SAPSGSSALLYLSQYG
Drosophila_melanogaster_1 ----------------MTNCQSS---VFIVVGTLFSIMAAAQSAPVSTTTQAEIYLSQFG
Drosophila_melanogaster   ----------------MTNCQSS---VFIVVGTLFSIMAAAQSAPVSTTTQAEIYLSQFG
Manduca_sexta             MRIN---NNVRRTGVMAVNMRNSLRILWTTVAAFVFLTRSSAAPTFGGPDKATMYLAQYG
Bombyx_mori_2             MRIIKNTKGISRGIMAMMTMRGGLRILWTVAAAGVLLTRSSAAPTFGTTDKATMYLAQYG
Bombyx_mori_1             MRIIKNTKGISRGIMAMMTMRGGLRILWTVAAAGVLLTRSSAAPTFGTTDKATMYLAQYG
Nasonia_vitripennis       ---------MSLNIHTQLQHYAGLKYFIFFVFIVLYFQKESVVLASFPATGAMRYLSQYG
AGAP006904-L              ------------------------------------------------------------
AGAP006904-A              ------------------------------------------------------------
AGAP006904-P              ------------------------------------------------------------

Hydra_MMP                 YYTLS---TEVGSINNEKEIRNSIENLQRFAGIPVSGILDAPTQELIETPRCGLPDFKKP
Tribolium_castaneum_1     YLGGNLRSLNSSALTDERVLRKAVEDFQSFAGLDVTGELDDRTLKEMQLPRCGVKDKVGT
Tribolium_castaneum       YLGGNLRSLNSSALTDERVLRKAVEDFQSFAGLDVTGELDDRTLKEMQLPRCGVKDKVGT
Drosophila_melanogaster_1 YLPASARNPASSGLHDQRTWVSAIEEFQSFAGLNITGELDAETMKLMSLPRCGVRDRVGT
Drosophila_melanogaster   YLPASARNPASSGLHDQRTWVSAIEEFQSFAGLNITGELDAETMKLMSLPRCGVRDRVGT
Manduca_sexta             YLSPQVRNPSSGHIMDESSWRRAIAEFQSFAGLNATGELDEETSKVMSLPRCGVKDKVGF
Bombyx_mori_2             YLSPSVRNPSSGHIMDESSWRRAIAEFQSFAGLNATGELDDQTNEMMSLPRCGVRDKVGF
Bombyx_mori_1             YLSPSVRNPSSGHIMDESSWRRAIAEFQSFAGLNATGELDDQTNEMMSLPRCGVRDKVGF
Nasonia_vitripennis       YLPPL--NPTNGGLVSEQTMQRAIAEFQSFAGLNITGVLDTDTAALMSMPRCGVKDRVGA
AGAP006904-L              ------------------------MEFQSFAGLNVTGELDGETMQLMSLPRCGVKDKVGF
AGAP006904-A              ------------------------MEFQSFAGLNVTGELDGETMQLMSLPRCGVKDKVGF
AGAP006904-P              ------------------------MEFQSFAGLNVTGELDGETMQLMSLPRCGVKDKVGF
                                                  ::* ***:  :* **  *   :. ****: *

Hydra_MMP                 NE-SRNRRYTLQGTTWKKNELTWKLLNNNNDGLTRGEIETTLHKAFSMWEAVTNLKFRQL
Tribolium_castaneum_1     GD-NRAKRYALQGSRWKVKNLNYKISK-YPKNLNTKEVDKEIHRAFSVWSQYTDLTFTPS
Tribolium_castaneum       GD-NRAKRYALQGSRWKVKNLNYKISK-YPKNLNTKEVDKEIHRAFSVWSQYTDLTFTPS
Drosophila_melanogaster_1 GD-SRSKRYALQGSRWRVKNLTYKISK-YPKRLKRVDVDAEIGRAFAVWSEDTDLTFTRK
Drosophila_melanogaster   GD-SRSKRYALQGSRWRVKNLTYKISK-YPKRLKRVDVDAEIGRAFAVWSEDTDLTFTRK
Manduca_sexta             GE-SRAKRYALQGSRWRVKNLTYKISK-YPSKLNHAEVDAELAKAFSVWTDYTDLTFTQK
Bombyx_mori_2             GE-SRAKRYALQGSRWRVKNLTYKISK-YPSRLNRAEVDAELAKAFSVWSDYTDLTFTQK
Bombyx_mori_1             GE-SRAKRYALQGSRWRVKNLTYKISK-YPSRLNRAEVDAELAKAFSVWSDYTDLTFTQK
Nasonia_vitripennis       SSDGRSKRYPLQSSGWRVKKLTYKISK-YPRVLEKGAVDKEIAKAFSVRSDYTNLQFTPK
AGAP006904-L              GSDTRSKRYALQGSRWKVKDLTYRISK-YPRRLERTAVDKEIAKAFGVWSEYTDLRFTPK
AGAP006904-A              GSDTRSKRYALQGSRWKVKDLTYRISK-YPRRLERTAVDKEIAKAFGVWSEYTDLRFTPK
AGAP006904-P              GSDTRSKRYALQGSRWKVKDLTYRISK-YPRRLERTAVDKEIAKAFGVWSEYTDLRFTPK
                          ..  * :** **.: *: :.*.::: :     *    ::  : :**.:    *:* *
```

```
Hydra_MMP                   QINENKKADIEIKFAQGYHDDPYSFDGFGGTLAHAFYPHTNEGLSGDVHFDDAEKFTIES
Tribolium_castaneum_1       K----GSAHIEIRFESGEHGDGDPFDGPGGTLAHAYFPV----FGGDAHFDASEKWTINS
Tribolium_castaneum         K----GSAHIEIRFESGEHGDGDPFDGPGGTLAHAYFPV----FGGDAHFDASEKWTINS
Drosophila_melanogaster_1   T---SGPVHIEIKFVESEHGDGDAFDGQGGTLAHAFFPV----FGGDAHFDDAELWTIGS
Drosophila_melanogaster     T---SGPVHIEIKFVESEHGDGDAFDGQGGTLAHAFFPV----FGGDAHFDDAELWTIGS
Manduca_sexta               R---SGQVHIEIRFEKGEHGDGDPFDGPGGTLAHAYFPV----YGGDAHFDDAEMWSINS
Bombyx_mori_2               R---SGQVHIEIRFEKGEHGDGDPFDGPGGTLAHAYFPV----YGGDAHFDDAEMWSINS
Bombyx_mori_1               R---SGQVHIEIRFEKGEHGDGDPFDGPGGTLAHAYFPV----YGGDAHFDDAEMWSINS
Nasonia_vitripennis         K---SGQVHIEIRFERGEHGDGDPFDGPGGTLAHAYFPV----YGGDAHFDDTEQWTINS
AGAP006904-L                K---TGAVHIDIRFEENEHGDGDPFDGPGGTLAHAYFPV----YGGDAHFDDAEQWTIDK
AGAP006904-A                K---TGAVHIDIRFEENEHGDGDPFDGPGGTLAHAYFPV----YGGDAHFDDAEQWTIDK
AGAP006904-P                K---TGAVHIDIRFEENEHGDGDPFDGPGGTLAHAYFPV----YGGDAHFDDAEQWTIDK
                                   ..*:*:*  . *.*   *** *******::*      .**.*** :* ::* .

Hydra_MMP                   PEGRSLLWVAVHEIGHSIGLEHSNVKEALMFPWYRVQDVRDIQLSDDDVLGIQSIYGSKK
Tribolium_castaneum_1       YRGTNLFQVAAHEFGHSLGLSHSDVREALMAPFYRGYDP-LFELHEDDIQGIQALYGKKT
Tribolium_castaneum         YRGTNLFQVAAHEFGHSLGLSHSDVREALMAPFYRGYDP-LFELHEDDIQGIQALYGKKT
Drosophila_melanogaster_1   PRGTNLFQVAAHEFGHSLGLSHSDQSSALMAPFYRGFEP-VFKLDEDDKAAIQSLYGRKT
Drosophila_melanogaster     PRGTNLFQVAAHEFGHSLGLSHSDQSSALMAPFYRGFEP-VFKLDEDDKAAIQSLYGRKT
Manduca_sexta               RRGTNLFQVAAHEFGHSLGLSHSDVRSALMAPFYRGFDP-AFQLDQDDIQGIQALYGHKT
Bombyx_mori_2               RRGTNLFQVAAHEFGHSLGLSHSDVRSALMAPFYRGYDP-AFQLDQDDVQGIQSLYGHKT
Bombyx_mori_1               RRGTNLFQVAAHEFGHSLGLSHSDVRSALMAPFYRGYDP-AFQLDQDDVQGIQSLYGHKT
Nasonia_vitripennis         FRGTNLFQVAAHEFGHSLGLSHSDVKSALMAPFYRGYDP-DFLLESDDIQGIQALYGSKN
AGAP006904-L                PRGTNLFQVAAHEFGHSLGLSHSEIPFALVARFY--------------------------
AGAP006904-A                PRGTNLFQVAAHEFGHSLGLSHSDIPFSLVSRFYRA------------------------
AGAP006904-P                PRGTNLFQVAAHEFGHSLGLSHSDVR-----------------------------------
                             .* .*: **.**:***:**.**:

Hydra_MMP                   SIMLPST----------V---------TPTKKMK----NSFQMKAVILDKSTGVTYAFND
Tribolium_castaneum_1       RKPGGGGGGYDDSDFQGSNPGGHR-VPAPAPTPVDSNLCKNPKIDTIFNSAEGYTYIFKG
Tribolium_castaneum         RKPGGGGGGYDDSDFQGSNPGGHR-VPAPAPTPVDSNLCKNPKIDTIFNSAEGYTYIFKG
Drosophila_melanogaster_1   NQLRPTN----------VYPATTQRPYSPPKVPLDDSICKDSKVDTLFNSAQGETYAFKG
Drosophila_melanogaster     NQLRPTN----------VYPATTQRPYSPPKVPLDDSICKDSKVDTLFNSAQGETYAFKG
Manduca_sexta               QTDIGGGSVGGG----GLVPSVPRAT-TQQPSAEDPALCADPRIDTIFNGADGSTFVFKG
Bombyx_mori_2               QTDIGGG---GG----GLIPSVPRAT-TQQPSAEDPALCADPRVDTIFNSADGSTFVFKG
Bombyx_mori_1               QTDIGGG---GG----GLIPSVPRAT-TQQPSAEDPALCADPRVDTIFNSADGSTFVFKG
Nasonia_vitripennis         EDGG-------------ESNIHQRTTHLPPSSEEDSQLCSNPKIDTIFNSAEGDTFVFRG
AGAP006904-L                ------------------------------------------------------------
AGAP006904-A                ------------------------------------------------------------
AGAP006904-P                ------------------------------------------------------------


Hydra_MMP                   DEFYKINNDLKKTEGPFTVSSLFPEVN-SVNSGYMDSDGKLIFFKGTRYYTYKNFSDRKL
Tribolium_castaneum_1       DKYWKLTEESVAPGYPKAISSGWPGLPGDIDAAFTYKNGKTYFFKGSKYWRYKGRKVDGD
```

```
Tribolium_castaneum           DKYWKLTEESVAPGYPKAISSGWPGLPGDIDAAFTYKNGKTYFFKGSKYWRYKGRKVDGD
Drosophila_melanogaster_1     DKYYKLTTDSVEEGYPQLISKGWPGLPGNIDAAFTYKNGKTYFFKGTQYWRYQGRQMDGV
Drosophila_melanogaster       DKYYKLTTDSVEEGYPQLISKGWPGLPGNIDAAFTYKNGKTYFFKGTQYWRYQGRQMDGV
Manduca_sexta                 EHYWRLTEDGVAAGYPRLISRAWPNLPGNIDAAFTYKNGKTYFFKGSKYWRYNGQKMDGD
Bombyx_mori_2                 DHYWRLTEDGVAAGYPRLISRAWPGLPGNIDAAFTYKNGKTYFFKGSKYWRYNGQKMDGD
Bombyx_mori_1                 DHYWRLTEDGVAAGYPRLISRAWPGLPGNIDAAFTYKNGKTYFFKGSKYWRYNGQKMDGD
Nasonia_vitripennis           DLYWKLTTDGVESGYPRLISTSWKNLPGNIDAAFTYKNGKTYFFKGSKYWRYIGRKMDGD
AGAP006904-L                  ------------------------------------------------------------
AGAP006904-A                  ------------------------------------------------------------
AGAP006904-P                  ------------------------------------------------------------

Hydra_MMP                     LESGSIFDKYKGIKSGVKTIDAAFVWN-NGRTYVFIEDEYYRFGGTKNVLDAGFPRKIVD
Tribolium_castaneum_1         Y-PKEIS---EGFTGIPDDLDAAMVWSGNGKIYFFKGAKFWRFDPSQRPPVKSTYPKPIS
Tribolium_castaneum           Y-PKEIS---EGFTGIPDDLDAAMVWSGNGKIYFFKGAKFWRFDPSQRPPVKSTYPKPIS
Drosophila_melanogaster_1     Y-PKEIS---EGFTGIPDHLDAAMVWGGNGKIYFFKGSKFWRFDPAKRPPVKASYPKPIS
Drosophila_melanogaster       Y-PKEIS---EGFTGIPDHLDAAMVWGGNGKIYFFKGSKFWRFDPAKRPPVKASYPKPIS
Manduca_sexta                 Y-PKEIS---EGFTGIPDNIDAALVWSGNGKIYFYKGSKFWRFDPAQRPPVKATYPKPLS
Bombyx_mori_2                 Y-PKDIS---EGFTGIPDNLDAALVWSGNGKIYFYKGSKFWRFDPAQRPPVKATYPKPLS
Bombyx_mori_1                 Y-PKDIS---EGFTGIPDNLDAALVWSGNGKIYFYKGSKFWRFDPAQRPPVKATYPKPLS
Nasonia_vitripennis           Y-PKDIS---EGFTGIPDNLDAVTVWTGNGKIYFYKGTKFWRFDPLQKPPVKSTYPKLIS
AGAP006904-L                  ------------------------------------------------------------
AGAP006904-A                  ------------------------------------------------------------
AGAP006904-P                  ------------------------------------------------------------

Hydra_MMP                     NWTGVPKNIDSVFVWRNGVTYFFKGSLFYRVNEKGQV----LLNYPKSISGAWLNFPNK-
Tribolium_castaneum_1         NWEGVPNNLDAAFKWTNGYTYFYKGDAYYRFNDRAFAVDKASPAFPRAIAYWWLGCSNAP
Tribolium_castaneum           NWEGVPNNLDAAFKWTNGYTYFYKGDAYYRFNDRAFAVDKASPAFPRAIAYWWLGCSNAP
Drosophila_melanogaster_1     NWEGVPNNLDAALKYTNGYTYFFKGDKYYRFHDARFAVDSATPPFPRPTAHWWFGCKNTP
Drosophila_melanogaster       NWEGVPNNLDAALKYTNGYTYFFKGDKYYRFHDARFAVDSATPPFPRPTAHWWFGCKNTP
Manduca_sexta                 NWDGIPDNIDAALQYTNGYTYFFKGGSYWRFNDRTFSVDADNPQFPRSTAFWWLGCSNAP
Bombyx_mori_2                 NWDGIPDNIDAALQYTNGYTYFFKGGSYWRFNDRLFSVDTDNPQFPRSTAFWWLGCSSAP
Bombyx_mori_1                 NWDGIPDNIDAALQYTNGYTYFFKGGSYWRFNDRLFSVDTDNPQFPRSTAFWWLGCSSAP
Nasonia_vitripennis           NWEGIPDNVDASITY-HGYTYFFKDNAYFRFNDRSFSVDVASPSFPRSIAYWWFGCRSTS
AGAP006904-L                  ------------------------------------------------------------
AGAP006904-A                  ------------------------------------------------------------
AGAP006904-P                  ------------------------------------------------------------

Hydra_MMP                     ------------------------------------------------------------
Tribolium_castaneum_1         QGTIGTKN------------------------YRRP----ASH-----------------
Tribolium_castaneum           QGTIGTSESRGWLLDEESDQDYAGSDTLDTENYRRP----ASH-----------------
Drosophila_melanogaster_1     SSTAVGD---HQSNDEPIVPEVAERTGNGAMSQSKLTSSSAVSTVITTILMCLVSKLIVS
Drosophila_melanogaster       SSTGFKR---RGYKNKNN------------------------------------------
```

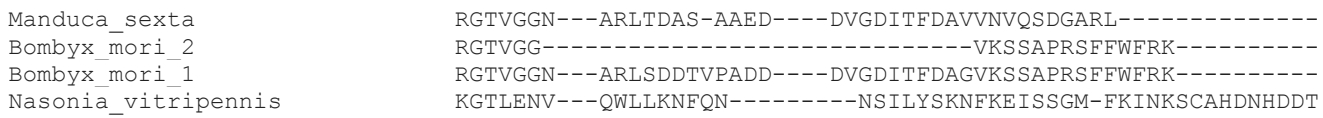


Figure 15: Multiple Sequence Alignment. This figure shows a multiple sequence alignment of MMPs from *Drososphila melanogaster*, *Tribolium, castaneum*, *Manduca sexta*, *Bombyx mori*, *Nasonia vitripennis*, AGAP006904-Larvae, AGAP006904-Pupae, AGAP006904-Adult and the outgroup *Hydra vulgaris*. In the MSA you can see that the prodomain contains the conserved cysteine in PRCGxD, the metalloproteinase domain contains the consensus sequence HebxHxbGbxHz.

Figure 15 indicates that AGAP006904-Larvae, AGAP006904-Pupae and AGAP006904-Adult are MMPs as they have conserved sequences with other known MMPs. MMPs are known to have 3 main domains: a prodomain with the conserved consensus sequence **PRCG**XX**D**, a metalloprotease domain with a consensus conserved sequence **HE**X**GH**XXXXX**H**S, and a hemopexin domain. All MMPs also have a conserved methionine just after the consensus sequence found in the metalloproteinases domain. The AGAP006904-PA lacks a methionine as it appears to be truncated. The conserved residues are also important in the structural integrity of the 3D structures of MMPs, for example Arg (**R**) and Asp (**D**) are used to pack the cysteine switch against the main body of the prodomain by interacting mainly with tyrosine residues in helix A.

4.2. Discussion

The protein sequences AGAP006904 and AGAP003929 were predicted to have two domains (metalloproteinase and hemopexin repeats) and one binding site (peptidoglycan-like binding site). These domains and the nature of their interactions determines the functions of the MMP proteins. Annotations is based on the detailed examinations of the protein structure which is essential for understanding the precise molecular functions of the domains and its contribution to the function of the whole protein.

Anopheles AGAP006904 and AGAP003929 belong to the MEROPS peptidase family M10 (clan MA (M)), subfamily M10A. The protein fold of the peptidase domain for members of this family resembles that of thermolysin, the type example for clan MA. These two protein sequences are extracellular metalloproteases, such as collagenase and stromelysin, which degrade the extracellular matrix, are known as matrixins. They are zinc-dependent, calcium-activated proteases synthesised as inactive precursors (zymogens), which are proteolytically cleaved to yield the active enzyme (Wilhelm *et al*., 1989, Lepage and Gache, 1990). All matrixins and related proteins possess 2 domains: an N-terminal domain, and a zinc-binding active site domain. The N-terminal domain peptide, cleaved during the activation step, includes a conserved PRCGVPDV octapeptide, known as the cysteine switch, whose Cys residue chelates the active site zinc atom, rendering the enzyme inactive (Sanchez-lopez *et al*., 1988, Park *et al*., 1991). The active enzyme degrades components of the extracellular matrix, playing a role in the initial steps of tissue remodeling during morphogenesis, wound healing, angiogenesis and tumour invasion (Wilhelm *et al*., 1989, Lepage and Gache, 1990).

Both AGAP006904 and AGAP003929 had a peptidoglycan binding like domain (PGBD). The PGBD may have a general peptidoglycan binding function. It has a core structure consisting of a closed, three-helical bundle with a left-handed twist. It is found at the N or C terminus of a variety of enzymes involved in bacterial cell wall degradation (Krogh *et al*., 1998; Dideberg *et al*., 1982). Many of the proteins having this domain are as yet uncharacterized. However, some are known to belong to MEROPS peptidase family M15 (clan MD), subfamily M15A metallopeptidases. A number of the proteins belonging to subfamily M15A are non-

peptidase homologues as they either have been found experimentally to be without peptidase activity, or lack amino acid residues that are believed to be essential for the catalytic activity. Eukaryotic enzymes can contain structurally similar PGBD-like domains. Matrix metalloproteinases (MMP), which catalyze extracellular matrix degradation, have N-terminal domains that resemble PGBD. Examples are two human MMPs, gelatinase A (MMP-2), which degrades type IV collagen (Seiki, 1999, stromelysin-1 (MMP-3), which plays a role in arthritis and tumour invasion (Smeets *et al*., 2003; Hornebeck and Maquart, 2003), and gelatinase B (MMP-9) secreted by neutrophils as part of the innate immune defense mechanism (Van den Steen *et al*., 2003). Several MMPs are implicated in cancer progression, since degradation of the extracellular matrix is an essential step in the cascade of metastasis (Yoshizaki *et al*., 2002).

The AGAP006904 and AGAP003929 also had a hemopexin-like domain. Hemopexin-like domains have been found in two other types of proteins, vitronectin (Yoneda *et al*., 1998), a cell adhesion and spreading factor found in plasma and tissues, and matrixins MMP-1, MMP-2, MMP-3, MMP-9, MMP-10, MMP-11, MMP-12, MMP-14, MMP-15 and MMP-16, members of the matrix metalloproteinase family that cleave extracellular matrix constituents (Das *et al*., 2003). These zinc endopeptidases, which belong to MEROPS peptidase subfamily M10A, have a single hemopexin-like domain in their C-terminal section. It is suggested that the hemopexin domain facilitates binding to a variety of molecules and proteins, for example the Hpx repeats of some matrixins bind tissue inhibitor of metallopeptidases (TIMPs).

Based on the classification of domains and analysis, we can say that the putative function of transcript A of AGAP006904 and transcript A of AGAP006904 are degradation of proteoglycans,gelatin, and other constituents of the extracellular matrix(stromelys in) and transcript B of AGAP006904 are degradation of collagen fibers (collagenase).

The structure generated by MODELLER shows four segments, a prodomain, a catalytic domain, a linker region and a hemopexin (Hpx) domain (Figure 6). The ellipsoid/egg-like shaped prodomain, docks to the active site in the metalloproteinase domain. In addition, it interacts with the Hpx domain contributing to a compact structural arrangement of proMMP. The overall fold

of the catalytic and Hpx domain is in agreement with the characteristic of other MMPs (Jozic et al., 2004).

Numerous structures of human MMPs have been solved in the last decade. Most published structures are catalytic domains of human MMP-1, MMP-3, MMP-7, MMP-8, MMP-9, MMP-11, MMP-12, MMP-13, MMP-14 and MMP-16. ProMMP-1 is a complete enzyme that has been solved and shows its collegenolysis activity (Jozic *et al*., 2004). ProMMP-2 is also a complete proenzyme solved either alone (Morgunova *et al*., 1999) or as the proMMP-2-TIMP-2 complex (Morgunova *et al*., 2002) and one complete active two-domain structure has been determined for pig MMP-1 (Li *et al*., 1995). This is the first structure of a complete MMP in *Anopheles gambiae*. MMP shows three distinct domains namely a prodomain, a catalytic domain and an Hpx domain. This structure also shows an interaction between the prodomain and the Hpx domain. Comparison of the structure resolved in this work is characteristically similar to that of other MMPs found in humans.

The prodomain shows the characteristic three-helix bundle with a left-handed twist. In human proMMP1 the residues of a 'bait' region in loop 1 dictates which proteinases can initiate the activation of this proMMP1 (Jozic *et al*., 2004). Residues of this bait region are not known in *Anopheles gambiae*. Once the bait region is cleaved, autocatalytic cleavage occurs between Thr^{64}-Leu^{65} and the peptide bond in the junction of the prodomain and the catalytic domain becomes exposed and susceptible to proteolysis. In the case of human proMMP-1, this latter cleavage is done by MMP-3 (Suzuki *et al*., 1990), MMP-2 (Crabbe *et al*., 1994) and MMP-7 (Imai *et al*., 1995). In proMMP-2, MT-MMPs cleave the loop (Tyr^{58}-Asn^{66}) that connects helices A and helices B. In *Anopheles gambiae* the site for MMP autocatalytic cleavage is between Thr^{54}-Leu^{57}. However the molecules that are involved in this autocatalytic cleavage are yet to be elucidated. These residues are part of Helix C and it is clear from the prodomain structure that the peptide bond between these two residues is oriented towards the interior of the prodomain. Therefore this peptide bond is inaccessible for cleavage in the intact prodomain. Both in *Anopheles gambiae* and human pro-MMP1 removal of Helix A, as a result of 'bait' region cleavage exposes the Thr^{64}-Leu^{65} peptide bond, but additional conformational changes are required before autocatalytic cleavage can take place within this α-helical structure. In the presence of MMP-3, Gly^{70}-Phe^{71} bond is cleaved and fully activates MMP (Suzuki *et al*., 1990).

This action also requires the initial cleavage of the bait region and a considerable change in conformation of the remaining propeptide to allow Phe^{71} to fit into the S1' pocket of MMP-3.

The three-helix bundle of the prodomain allows for a compact, stable structure. The main stability comes from the hydrophobic core formed within the bundle and the fact that each helix interacts with the two others, giving the domain a natural rigidity. Removal of one of the helices would have a severe impact on its stability because the two remaining helices only interact with each other. This suggests that the three helix bundle design of the prodomain is in fact ideally suited to act as a trigger mechanism for the sequential activation mechanism that is observed for most MMPs.

The actual cysteine switch region is packed against the main body of the prodomain with Arg^{62} and Asp^{67} interacting mainly with Tyr residue form helix A (Tyr^{6}) (Figure 8). Arg^{62} and Asp^{67} are absolutely conserved in the cysteine switch and Tyr^{6} that form H-bonds in this arrangement are also conserved, together with other two tyrosine residues in human MMPs for example MMP-1, MMP-2, MMP-3, MMP-7, MMP-8, MPP-9, MMP-10, MMP-12 and MMP-13. In humans, proMMP-9 has a glycine insertion between the two tyrosines at the end of helix A, but the second tyrosine still makes the same hydrogen bond as Tyr^{6} in proMMP-1. This indicates that the interactions of the cysteine switch with helix A residues are highly conserved. Thus, it maybe postulated that the cleavage in the bait region and the loss of helix A directly affect the cysteine switch region by the loss of a few interactions between them. This therefore leads to the destabilization of the Cys-Zn^{2+} interaction, resulting in a partially active enzyme. Because of the trigger mechanism of destabilization of the propeptide, this is probably associated with structural perturbation around the junction of pro- and the catalytic domain as demonstrated by biochemical observations of human proMMP-1 and other proMMPs (Nagase, 1997).

The region Gly^{70}-Ser^{77} is visible in *Anopheles gambiae* proMMP, however in human proMMP-1 this region has residues Gln^{80}-Asn^{87} and is not visible. Although the overall structure of this loop follows a similar conformation in all proMMPs solved thus far, there is a considerable amount of variation in the local details of this region of proMMP-2, proMMP-3 and proMMP-9. In both proMMP-3 and proMMP-9, the phenylalanine that is to become the N-terminus is buried in a hydrophobic pocket contributed to by Val^{65} (number in *Anopheles* MMP) of the cysteine switch. Cleavage of the peptide bond at this position to generate a fully active

enzyme will require a considerable conformational change around this residue to fit the phenylalanine in the S1' pocket of the activating MMP enzyme. In *Anopheles* MMP it is likely to expect Phe^{71} and Phe^{81} in human proMMP-1 to be buried in the pocket formed by Val^{65} and Val^{75}respectively.

The architecture of the catalytic domain known as the matrixin fold (Stocker and Bode, 1995) has five-stranded β-sheets and three α-helices. This structure is highly conserved in MMPs and is unaffected by the insertion of fibronectin domains in MMP-2 and MMP-9. The catalytic domain of *Anopheles* proMMP is similar to proMMP-1, proMMP-2 and proMMP3 in humans. The same residues form the substrate binding pocket and the co-ordination of the Zn^{2+} is similar. Also, the binding site for the structural $Zn^{2+}$is identical to a well conserved motif found in all known MMP structures.

The hinge/linker region in MMPs is a segment of 15-65 amino acids. In *Anopheles* proMMP, the hinge region has 39 amino acids. The structures of human and porcine MMP-1 show an elongated peptide segment which is in close contact with the catalytic and the hemopexin-like domain (Jozic *et al*., 2004, Li *et al*., 1995). Its importance is in the stability of the enzyme and also for the degradation of complex substrates such as collagen. Recently, it has been shown to contribute to the binding and unwinding of collagen. The hinge region contains 6 residues of proline which are thought to be highly conserved in collagenases. These proline residues are in direct contact with the metalloproteinase domain and Hpx domain thus stabilizing the domain arrangement in MMP-1. In accordance with these structures, mutagenesis experiments such as the alanine scanning analysis in the hinge region of MMP-8 showed an 98.5% drop in activity when four prolines in the hinge are replaced by alanines (Knauper *et al*., 1997).

The Hemopexin (Hpx) domain contains a C-terminal hemopexin-like domain, which was named that due to its sequence similarity to hemopexin, a plasmaheme-binding and heme-transport protein (Murphy *et al*., 1992). Some hemopexin-like domains have been shown to be involved in substrate recognition and specificity. This has especially been demonstrated for the subfamily of collageneases where the Hpx-like domain has been shown to be important in the cleavage of a triple-helical collagen (Faber *et al*., 1995). The Hpx-like domain has the shape of an oblate ellipsoidal disc.

The evaluation of homology models was based on the Global distance test_Total Score (GDT_TS), superimposition of modeled structure and template structure, stereochemical analysis using PROCHECK. The GDT_TS score was 68.065. GDT_TS varies between 0 and 100, with values approaching 100 as models become better. The GDT_TS cut-off for homology-based models is 65. This therefore shows that the model generated which has a GDT_TS score of 68 scan be used in the drug discovery.

The structure modeled by MODELLER 9V10 shows that the amino acid percentage in the favorable region is 85.2% and amino acid percentage in the disallowed region is 2.1%. Altogether, 100% of the residues are in favored and allowed regions. In the Ramachandran plot analysis, the residues were classified according to its regions in the quadrangle. The red area in the graph indicates the most allowed regions where there are few steric clashes, whereas the yellow area represent allowed regions. Glycine is represented by triangles, and other residues are represented by squares (Figure 11). Analysis of PROCHECK reveals that all residues are within the limits of the Ramachandran plot. Therefore, it can be considered as a good model.

Root mean square deviation (RMSD) is the measure of the average distance between the atoms (usually the backbone atoms) of superimposed proteins. Small rmsd values correspond to higher quality predictions than larger values. Figure 13 shows the superimposition of the modeled 3D structure (Red) and template file 1SU3.pdb (Cyan). The calculated RMSD was 1.695 Å which is within the range of allowed structural errors.

The z-score for the modeled structure was determined using ProSA. This program scores structures according to how well each residue fits into its structural environment based on criteria derived from statistical analysis of high resolution structures in the Protein Data Bank (PDB) (Penhoat *et al*., 2005). In Figure 12B, the graphs obtained are typical for native sequence structure pairs. Graphs obtained from a small window size (~10 residues) show few positive peaks. Graphs obtained from a large window size (~40 residues) stay almost below zero (Sippl, 1993). Figure 12 A shows the overall model quality by obtaining a z-score of -7.35, which is in the range of native structures.

The amplified gene showed that this protein has a 552 base pair (Figure 14). It also shows that it is expressed in larvae, pupae and adult. This suggests that these proteases maybe involved in processes of embryogenesis, metamorphosis and development. The gene AGAP003929 could

not be amplified, requiring further wet laboratory analysis. In *Drosophila*, MMPs have been co-opted for maggot specific metamorphosis processes including head evasion and notum connation (Page-McCaw, 2008). Both processes have recently been demonstrated to depend on imaginal disc eversions which require basement membrane remodeling by particularly MMPs (Srivastava et al., 2007). We postulate that silencing of MMP1 in *Anopheles* may result in altered basement membrane remodeling thereby leading to abnormal pupal tissue development and differentiation. Moreover, beside basement membrane remodeling other molecular processes such as specific cell migration and autophagic cell death processes, which are known to be regulated by MMPs in mammals, may be also involved in *Anopheles* and will be investigated in future studies. Since insect MMP-1s are most closely related to mammalian MMP-19 and MMP-28, there is a possibility that some functions may have been conserved between insects and mammals. Indeed, it has been shown that MMP-28 expression is induced during dermal wound healing (Lohi *et al*., 2001) and that MMP-19 functions in cutaneous homeostasis by modulating epidermal proliferation (Sadowski *et al*., 2003; Sadowski *et al*., 2003), cutaneous immune responses (Beck *et al*., 2008) as well as skin tumorigenesis (Pendas *et al*., 2004; Jost *et al*., 2006). Furthermore, human MMP-19 associates with the surface of human monocytes and macrophages (Mauch *et al*., 2002) and impacts the maturation and response of mammalian T-cells that are involved in both innate and adaptive immunity (Beck *et al*., 2008). Thus, insect model organisms such as *Anopheles gambiae*, lacking adaptive immune system may help to elucidate direct role of MMPs in innate specific immune responses.

The protein sequences of AGAP006904 (Larvae, Pupae and Adult) together with other known matrix metalloproteinases from other organism in the class insect were aligned with Clustal Omega. Figure 15 shows that the propeptide domain and catalytic domain of MMP are evolutionarily conserved across species. There is also a conserved methionine across all the species. This strictly conserved methionine contains a tight 1, 4-beta turn forming a hydrophobic cleft for the catalytic zinc ion, this turn is also known as the ‘Met-turn’. This conserved methionine in AGAP006904 (Larvae, Pupae and Adult) shows that they belong to the metzincins superfamily under zinc metalloendopeptidases. The individual families of this superfamily can be distinguished by (i) the residue that immediately follows the third histidine zinc ligand in the consensus sequence HEXXHXXGXXH found in the catalytic domain, (ii) the residue surrounding the methionine in the ‘Met-turn’. In figure 15, the multiple sequence alignment

shows that these 2 proteins have a conserved serine (s) following the third histidine residues in the conserved consensus sequence in the metalloproteinase (M10) domain. Since the predicted protein has similar domain and structural conformation to the human pro-collagenase, we can say that the hypothetical function of *Anopheles gambiae* MMP is in the breaking down of collagen.

CHAPTER FIVE
CONCLUSION AND RECOMMENDATION

5.1. Conclusion

The AGAP006904 has similar structural conformation to human proMMP-1 (1su3.pdb). It putatively has similar activation mechanism and conserved regions such as the cysteine switch PRCGV, catalytic domain HEFGHSLGLSHS and Met-turn sequences MxP to human MMP and other MMPs in Diptera. Based on the structural arrangement of transcript B of AGAP006904 we can conclude that the putative function is degradation of collagen. Transcript A of AGAP006904 has been found to be expressed in the 3 developmental stages of An. gambiae (Larvae, Pupae and Adult), implying possible roles of MMP in development.

5.2. Recommendation

I recommend that:

1. Gene silencing of MMP should be done to further understand its function, since we have shown that MMP-1 is expressed in both the larvae, pupae and adult.
2. qPCR should be performed to determine the levels of expression of MMP-1 across the three developmental stages.
3. Further studies on the structural similarity of MMP-1 in *Anopheles* and the mammals in their various habitat should be carried out.

REFERENCES

Altschul S. F., Gish W., Miller W., Myers E. W., and Lipman D. J., (1990). "The Basic Local Alignment Search tool." *Journal of Molecular Biology* **215**: 403-410

Baird J. K. (2007). "Neglect of *Plasmodium vivax* malaria." *Trends in Parasitology* **23**: 533-539.

Beck I. M, Ruckert R., Brandt K., Mueller M. S., Sadowski T., Brauer R., Schirmacher., Mentlein R., Sedlacek R. (2008). "MMP19 is essential for T cell development and T cell-mediated cutaneous immune responses." *PLoS ONE* **3**: 2343

Becker A. B. and Roth R. A. (1993). "Identification of Glutamate-169 as the Third Zinc-binding Residue in Proteinase III, a Member of the Family of Insulin-degrading Enzymes." *The Biochemical Journal* **292**: 137–142.

Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N., and Bourne P. E. (2000). "The Protein Data Bank." *Nucleic Acids Research* **28**: 235–242.

Berman H. M., Burley S. K., Chiu W., Sali A., Adzhubei A., Bourne P. E., Bryant S. H., Dunbrack R. L., Fidelis K., Frank J., Godzic A., Henrick K., Joachimiak A., Heyman B., Jones D., Markley J. L., Moult J., Montelione G. T., Orengo C., Rossmann M. G., Rost B., Saibil H., Schwede T., Standley D. M., and Westbrook J. D. (2006). "Outcome of a Workshop on Archiving Structural Models of Biological Macromolecules." *Structure* **14**: 1211–1217.

Bode W., Gomis-Rüth F. X., and Stöckler W. (1993). "Astacins, Serralysins, Snake Venom and Matrix Metalloproteinases Exhibit Identical Zinc-binding Environments (HEXXHXXGXXH and Met-turn) and Topologies and Should Be Grouped into a Common Family, the 'Metzincins'." *FEBS Letters* **331**: 134–140.

Boniecki M., Rotkiewicz P., Skolnick J., and Kolinski A. (2003). "Protein Fragment Reconstruction Using Various Modeling Techniques." *Journal of Computer-aided Molecular Design* **17**: 725–738.

Bru C., Courcelle E., Carrère S., Beausse Y., Dalmar S., and Kahn D. (2005). "The ProDom Database of Protein Domain Families: More Emphasis on 3D." *Nucleic Acids Research* **33**: 212–215.

Crabbe T., O'Connell J. P., Smith B. J., and Docherty A. J. (1994). "Reciprocated matrix metalloproteinase activation: a process performed by interstitial collagenase and progelatinase A." *Biochemistry* **33**: 14419-14425

Cowman A. F., and Crabb B. S. (2006). "Invasion of Red Blood Cells by Malaria Parasites." *Cell* **124**: 755–766.

Dahler G. S., Barras F., and Keen N. T. (1990). "Cloning of Genes Encoding Extracellular Metalloproteases from Erwinia Chrysanthemi EC16." *Journal of Bacteriology* **172**: 5803–5815.

Das S., Mandal M., Chakraborti T., Mandal A., and Chakraborti S. (2003). "Structure and Evolutionary Aspects of Matrix Metalloproteinases: a Brief Overview." *Molecular and Cellular Biochemistry* **253**: 31–40.

Dideberg O., Charlier P., Dive G., Joris B., Frère J. M., and Ghuysen J. M. (1982). "Structure of a Zn^{2+}-containing D-alanyl-D-alanine-cleaving carboxypeptidase at 2.5Å resolution." *Nature* **299**:469-70.

Eddy S. R. (1998). "Profile Hidden Markov Models." *Bioinformatics review* **14**: 755–763.

Elofsson A. (2006). "Identification of Correct Regions in Protein Models Using Structural, Alignment, and Consensus Information." *Protein Science* **15**: 900–913.

Faber H. R., Groom C. R., Baker H. M., Morgan W. T., Smith A., and Baker E. N. (1995). "1.8 A Crystal Structure of the C-terminal Domain of Rabbit Serum Haemopexin." *Structure* **3**: 551–559.

Feng D. F., and Doolittle R. F. (1987). "Progressive Sequence Alignment as a Prerequisite to Correct Phylogenetic Trees." *Journal of Molecular Evolution* **25**: 351-360

Fischer D. and Eisenberg D. (1996). "Protein Fold Recognition Using Sequence-derived Predictions." *Protein Science : A Publication of the Protein Society* **5**: 947–955.

Fiser A., Do R. K., and Sali A. (2000). "Modeling of Loops in Protein Structures." *Protein Science : A Publication of the Protein Society* **9**: 1753–1773.

Fontenille D. and Simard F. (2004). "Unravelling complexities in human malaria transmission dynamics in Africa through a comprehensive knowledge of vector populations." *Comparative Immunology, Microbiology and Infectious Diseases* **27**: 357-375.

Galtier N., Gouy M., and Gautier C. (1996). "SEA VIEW and PHYLO_ WIN: Two Graphic Tools for Sequence Alignment and Molecular Phylogeny." *Oxford University Press* **12**: 543–548.

Gething P. W., Anand P. P., Smith D. L., Guerra C. A., Elyazar R. F. I., Johnston G. L., Tatem A. J., and Hay S. I. (2011). "A New World Malaria Map : Plasmodium Falciparum Endemicity in 2010." *Malaria Journal* **10**: 378.

Gillies M.T. and Coetzee M. (1987). "A Supplement to the Anophelinae of Africa South of the Sahara Johannesburg." *The South African Institute for Medical Research* **55**: 143.

Gillies M.T. and De Meillon B. (1968). "The Anophelinae of Africa South of the Sahara (Ethiopian zoogeographical region) Johannesburg." *Second edition publications of the South African Institute for Medical Research* **54**: 343

Godzik A., and Skolnick J. (1992). "Sequence-structure Matching in Globular Proteins: Application to Supersecondary and Tertiary Structure Determination." *Proceedings of the National Academy of Sciences of the United States of America* **89**: 12098–102.

Gohlke U., Gomis-Rüth F. X., Crabbe T., Murphy G., Docherty A. J., and Bode W. (1996). "The C-terminal (haemopexin-like) Domain Structure of Human Gelatinase A (MMP2): Structural Implications for Its Function." *FEBS Letters* **378**: 126–130.

Gomis-Rüth F. X., Kress L. F., and Bode W. (1993). "First Structure of a Snake Venom Metalloproteinase: a Prototype for Matrix Metalloproteinases/collagenases." *The EMBO Journal* **12:** 4151–4157.

Goujon M., McWilliam H., Li W., Valentin F., Squizzato S., Paern J., and Lopez R. (2010). "A new bioinformatics analysis tools framework at EMBL-EBI." *Nucleic acids research* **38**: 695-699

Goulielmaki E., Siden-Kiamos I., and Loukeris T. G. (2014). "Functional characterization of Anopheles matrix metalloprotease 1 reveals its agonistic role during sporogonic development of malaria parasit." *American Society for microbiology* **82**: 4865-4877.

Gueirard P., Tavares J., Thiberge S., Bernex F., Ishino T., Milon G., Franke-Fayard B., Janse C. J., Menard R., and Amino R. (2010). "Development of the malaria parasite in the skin of the mammalian host." *Proceeding of National Academy of Sciences of U.S.A* **107**: 18640–18645.

Guilbride D. L., Gawlinski P., Guilbride P. D., and Rodrigues M. M. (2010). "Why functional pre-erythrocytic and blood stage malaria vaccines fail: a meta-analysis of fully protective immunizations and novel immunological model." *PLOS One* **5**:10685.

Harbach R. E. (2004). "The classification of genus Anopheles (Diptera: Culicidae): a working hypothesis of phylogenetic relationships." *Bulletin of Entomological Research*, **94**:537-553.

Hooper N. M. (1994). "Families of zinc metalloproteases." *FEBS Letters* **354**: 1–6.

Hornebeck W. and Maquart F. X. (2003). "Proteolyzed transcript as a template for the regulation of tumor pogression." *Biomedicine and Pharmacotherapy* **57**: 223-230.

Imai K., Yokohama Y., Nakanishi I., Obuchi E., Fujii Y., Nakai N., and Okada Y. (1995). "Matrix Metalloproteinase 7 (Matrilysin) from Human Rectal Carcinoma Cell:Activation of the precursor, interaction with other matrix metalloproteinases and enzymic properties." *Journal of Biological Chemistry* **270**: 6691-6697

Jaroszewski L., Rychlewski L., Li Z., Li W., and Godzik A. (2005). "FFAS03: a Server for Profile--profile Sequence Alignments." *Nucleic Acids Research* **33**: 284–288.

Jiang W., and Bond J. S. (1992). "Families of Metalloendopeptidases and Their Relationships." *FEBS Letters* **312**: 110–114.

Jongeneel C. V., Bouvier J., and Bairoch A. (1989). "A Unique Signature Identifies a Family of Zinc-dependent Metallopeptidases." *FEBS Letters* **242**: 211–214.

Jost M., Folgueras A. R., Frerart F., Pendas A. M., Blacher S., Houard X., Berndt S., Munaut C., Cataldo D., Alvarez J., Melen-Lamalle L., Foidart J. M., López-Otín C., and Noël A. (2006) "Earlier onset of tumoral angiogenesis in matrix metalloproteinase-19-deficient mice."*Cancer Research* **66**: 5234–5241.

Jozic D., Bourenkov G., Lim N-G., Visse R., Nagase H., Bode W., and Maskos K. (2005). "X-ray structure of human proMMP-1: new insights into procollagenase activation and collagen binding." *Journal of Biological Chemistry* **280**: 9578-9585.

Karplus K., Barrett C., and Hughey R. (1998). "Hidden Markov Models for Detecting Remote Protein Homologies." *Bioinformatics (Oxford, England)* **14**: 846–856.

Kelley L. A., MacCallum R. M., Sternberg M. J. E. (2000). "Enhanced genome annotation using structural profiles in the program 3D-PSSM." *Journal of Molecular Biology* **299**: 499-520.

Kelley L. A., and Sternberg M. J. (2009). "Protein Structure Prediction on the Web: a Case Study Using the Phyre Server." *Nature Protocols* **4**: 363–371.

Klein T. and Bischoff R. (2011). "Phyiology and Pathophysiology of matrix metalloproteinases." *Amino Acids* **41**: 271-290

Knauper V., Docherty A. J., Smith B., Tschesche H., and Murphy G. (1997). "Analysis of the contribution of the hinge region of human neutrophil collagenase (HNC, MMP-8) to stability and collagenolytic activity by alanine scanning mutagenesis." *FEBS Letter* **405**: 60–64

Krishnamoorthy B. and Tropsha A. (2003). "Development of a Four-body Statistical Pseudo-potential to Discriminate Native from Non-native Protein Conformations." *Bioinformatics* **19**: 1540–1548.

Krogh S., Jørgensen S. T., and Devine K. M. (1998). "Lysis Genes of the Bacillus Subtilis Defective Prophage PBSX." *Journal of Bacteriology* **180**: 2110–2117.

Laskowski R. A., MacArthur M. W., Moss D. S., and Thornton J. M. (1993). "PROCHECK: a program to check the stereochemical quality of protein structure." *Journal of Applied Crystallography* **26**: 283-291

Lepage T., and Gache C. (1990). "Early Expression of a Collagenase-like Hatching Enzyme Gene in the Sea Urchin Embryo." *The EMBO Journal* **9**: 3003–3012.

Li J., Brick P., O'Hare M. C., Skarzynski T., Lloyd L. F., Curry V. A., Clark I. M., Bigg H. F., Hazleman B. L., and Cawston T. E. (1995). "Structure of Full-length Porcine Synovial Collagenase Reveals a C-terminal Domain Containing a Calcium-linked, Four-bladed Beta-propeller." *Structure* **3**: 541–549.

Lin K., May A. C., and Taylor W. R. (2002). "Threading Using Neural nEtwork (TUNE) : a measure of Protein Sequence–Structure Compatibility." *Bioinformatics* **18**: 1350–1357.

Llano E., Adam G., Pendas A. M., Quesada V., Sanchez L. M., Santamaria I., Noselli S., and Lopes-Otin C. (2002). "Structural and enzymatic characterization of *Drosophila* Dm2-MMP, a membrane-bound matrix metalloproteinase with tissue-specific expression." *Journal of Biological Chemistry* **277**: 23321-23329.

Llano E., Pendas A. M., Aza-Blanc P., Kornbag T. B., and Lopez-Otin C. (2000). "Dm1-MMP, a matrix metalloproteinase from *Drosophila* with a potential role in extracellular matrix remodeling during neural development." *Journal Of Biological Chemistry* **275**: 35978-35985.

Lohi J., Wilson C. L., Roby J. D., and Parks W. C. (2001). "Epilysin, a novel human matrix metalloproteinase (MMP-28) expressed in testis and keratinocytes and in response to injury."*Journal of Biological Chemistry* **276**: 10134–10144.

Luthy R., Bowie J. U., and Eisenberg D. (1992). "Assessment of protein models with three-dimensional profiles." *Nature* **356**: 83-85.

MacKerell Jr A. D., Bashford D., Bellot M., Dunbrack Jr R. L., Evanseck J. D., Field M. J., Fischer S., Gao J., Guo H., Ha S., Joseph-McCarthy D., Kuchnir L., Kuczera K., Lau F. T. K., Mattos C., Michnick S., Ngo T., Nguyen D. T., Prodhom B., Reiher W. E., Roux B., Schlenkrich M., Smith J. C., Stote R., Straub J., Watanabe M., Wiorkiewicz-Kuczera J., Yin D., and Karplus M. (1998). "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins." *The Journal of Physical Chemistry* **102**: 3586–3616.

Maskos K. and Bode W. (2003). "Structural basis of matrix metalloproteinases and tissue inhibitors of metalloproteinases." *Molecular Biotechnology* **25**: 241-266

Mauch S., Kolb C., Kolb B., Sadowski T., and Sedlacek R. (2002). "Matrix metalloproteinase-19 is expressed in myeloid cells in an adhesion-dependent manner and associates with the cell surface."*Journal of Immunology* **168**: 1244–1251.

Melo F. and Feytmans E. (1998). "Assessing protein structures with a non-local atomic interaction energy." *Journal of Molecular Biology*. **277:** 1141–1152

Miller C. M., Page-McCaw A., and Broihier H. T. (2008). "Matrix metalloproteinases promote motor axon fasciculaton in the *Drosophila* embryo." *Development* **135**: 95-109

Morgunova E., Tuuttila A., Bergmann U., Isupov M., Lindqvist Y., Schneider G., Tryggvason K. (1999). "Structure of human pro-matrix metalloproteinase-2: activation mechanism revealed." *Science* **284**:1667–1670.

Morgunova E., Tuuttila A., Bergmann U., Isupov M., Lindqvist Y., Schneider G., and Tryggvason K. (2002). "Structural insight into the complex formation of latent matrix

metalloproteinase 2 with tissue inhibitor of metalloproteinase 2." *Proceedings of the National Academy of Sciences U.S.A* **99**: 7414-7419.

Mota M. M., Pradel G., Vanderberg J. P., Hafalla J. C., Frevert U., Nussenzweig R. S., Nussenzweig V., and Rodríguez A. (2001). "Migration of Plasmodium Sporozoites Through Cells Before Infection." *Science* **291**: 141–144.

Mulder N. J., Apweiler R., Attwood T. K., Bairoch A., Bateman A., Binns D., Bork P., Buillard V., Cerutti L., Copley R., Courcelle E., Das U., Daugherty L., Dibley M., Finn R., Fleischmann W., Gough J., Haft D., Hulo N., Hunter S., Kahn D., Kanapin A., Kejariwal A., Labarga A., Langendijk-Genevaux P. S., Lonsdale D., Lopez R., Letunic I., Madera M., Maslen J., McAnulla C., McDowall J., Mistry J., Mitchell A., Nikolskaya A. N., Orchard S., Orengo C., Petryszak R., Selengut J. D., Sigrist C. J., Thomas P. D., Valentin F., Wilson D., Wu C. H., and Yeats C. (2007). "New Developments in the InterPro Database." *Nucleic Acids Research* **35**: 224–228.

Mueller A. K., Labaied M., Kappe S. H., Matuschewski K. (2005). "Genetcally modified Plasmodium parasites as a protective experimental malaria vaccine." *Nature* **433:** 164-167.

Murphy G., Allan J. A., Willenbrock F., Cockett M. I., O'Connell J. P., and Docherty A. J. (1992). "The Role of the C-terminal Domain in Collagenase and Stromelysin Specificity." *The Journal of Biological Chemistry* **267**: 9612–9618.

Nagase H. (1997). "Activation mechanisms of matrix metalloproteinases". *Biological Chemistry* **378**: 151–160.

Nakahama K., Yoshimura K., Marumoto R., Kikuchi M., Lee I. S., Hase T., and Matsubara H. (1986). "Cloning and Sequencing of *Serratia* protease gene." *Nucleic Acids Research Acids* **14**: 5843–5855.

Ohlson T., Wallner B., and Elofsson A. (2004). "Profile-profile Methods Provide Improved Fold-recognition: a Study of Different Profile-profile Alignment Methods." *Proteins* **57**: 188–197.

Page-McCaw A., Serano J., Sante J. M., Rubin G. M. (2003). "*Drosophila* matrix metalloproteinases are required for tissue remodelling, but not embryonic development." *Developmental Cell* **4**: 95-106.

Page-McCaw A. (2008). "Remodeling the model organism: matrix metalloproteinase functions in invertebrates." *Seminars in Cell and Developmental Biology* **19**: 14–23.

Panchenko A. R. (2003). "Finding Weak Similarities Between Proteins by Sequence Profile Comparison." *Nucleic Acids Research* **31**: 683–689.

Park A. J., Matrisian L. M., Kells A. F., Pearson R., Yuan Z. Y., and Navre M. (1991). "Mutational Analysis of the Transin (rat Stromelysin) Autoinhibitor Region Demonstrates a Role for Residues Surrounding the 'Cysteine Switch'." *The Journal of Biological Chemistry* **266**: 1584–1590.

Pawlowski M., Gajda M. J., Matlak R., and Bujnicki J. M. (2008). "MetaMQAP : A Meta-server for the Quality Assessment of Protein Models." *BMC Bioinformatics* **20**: 1–20.

Pendas A. M., Folgueras A. R., Llano E., Caterina J., Frerard F., Rodríguez F., Astudillo A., Noel A., Birkedal-Hansen H., and Lopez-Otin C. (2004) "Diet induced obesity and reduced skin cancer susceptibility in matrix metalloproteinase 19-deficient mice." *Molecular and Cellular Biology* **24**: 5304–5313.

Penhoat C. H., Li Z., Atreya H. S., Kim S., Yee A., Xiao R., Murray D., Arrowsmith C. H., and Szyperski T. (2005). "NMR solution structure of Thermotogamaritima protein TM1509 reveals a Zn-metalloprotease-like tertiary structure." *Journal of Structural and Functional Genomics* **6**:51-62.

Pontius J., Richelle J., and Wodak S. J. (1996). "Deviations from Standard Atomic Volumes as a Quality Measure for Protein Crystal Structures." *Journal of Molecular Biology* **264**: 121–136.

Rawlings N. D., and Barrett A. J. (1993). "Evolutionary Families of Peptidases." *The Biochemical Journal* **290**: 205–218.

Sachs J. D. (2002). "A new global effort to control malaria." *Science* **298**: 122-124.

Sadowski T., Dietrich S., Koschinsky F., and Sedlacek R. (2003). "Matrix metalloproteinase 19 regulates insulin-like growth factor-mediated proliferation, migration, and adhesion in human keratinocytes through proteolysis of insulin-like growth factor binding protein-3." *Molecular Biology of the Cell* **14**: 4569–4580.

Sadowski T., Dietrich S., Muller M., Havlickova B., Schunck M., Proksch E., Muller M. S., and Sedlacek R. (2003). "Matrix metalloproteinase-19 expression in normal and diseased skin: dysregulation by epidermal proliferation." *Journal of Investigative Dermatology* **121**: 989–996.

Sali A., and Overington J. P. (1994). "Derivation of Rules for Comparative Protein Modeling from a Database of Protein Structure Alignments." *Protein Science : A Publication of the Protein Society* **3**: 1582–1596.

Sali A., and Blundell T. L. (1993). "Comparative Protein Modelling by Satisfaction of Spatial Restraints." *Journal of Molecular Biology* **234**: 779-815

Sanchez-Lopez R., Nicholson R., Gesnel M. C., Matrisian L. M., and Breathnach R. (1988). "Structure-function Relationships in the Collagenase Family Member Transin." *The Journal of Biological Chemistry* **263**: 11892–11899.

Seiki M. (1999). "Membrane-type matrix metalloproteinases." *APMIS* **107**:137-143.

Shen M., and Sali A. (2006). "Statistical Potential for Assessment and Prediction of Protein Structures." *Protein Science* **15**: 2507–2524.

Sippl M.J. (1990). "Calculation of Conformational Ensembles from Potentials of Mean Force: An Approach to the Knowledge-based Prediction of Local Structures in Globular Proteins." *Journal of Molecular Biology* **213**: 859-883

Sippl M. J. (1993). "Recogniton of errors in three-dimensional structures of proteins." *Proteins* **17**: 355-362.

Smeets T. J. M., Barg E. C., Kraan M. C., Smith M. D., Breedveld F. C., and Tak P. P. (2003). "Analysis of the cell infiltrate and expression of proinflammatory cytokines and matrix metalloproteinases in arthroscopic synovial biopsies: comparison with synovial samples from patients with end stage, destructive rheumatoid arthritis." *Annals of the Rheumatic Diseases* **62**: 635–638

Snow R. W., Guerra C. A., Noor A. M., Myint H. Y., and Hay S. I. (2005). "The global distribution of clinical episodes of Plasmodium Falciparum malaria." *Nature* **434**: 214-217

Srivastava A., Pastor-Pareja J. C., Igaki T., Pagliarini R., and Xu T. (2007) "Basement membrane remodeling is essential for Drosophila disc eversion and tumor invasion". *Proceedings of the National Academy of Sciences USA* **104**: 2721–2726.

Stocker W. and Bode W. (1995). "Structural features of a superfamily of zinc-endopeptidases: the metzincins." *Current Opinion in Structural Biology* **5**: 383-390.

Suzuki K., Enghild J. J., Morodomi T., Salvesen G., and Nagase H. (1990). "Mechanisms of activation of tissue procollagenase by matrix metalloprotease 3 (stromelysin)." *Biochemistry* **29**: 10261-10270

Thompson J. D., Gibson T. J., Plewniak F., Jeanmougin F., and Higgins D. G. (1997). "The CLUSTAL_X Windows Interface: Flexible Strategies for Multiple Sequence Alignment Aided by Quality Analysis Tools." *Nucleic Acids Research* **25**: 4876–4882.

Van den Steen P. E., Wuyts A., Husson S. J., Proost P., Van Damme J., and Opdenakker G. (2003). "Gelatinase B/MMP-9 and Neutrophil collagenase/MMP-8 Process the Chemokines Human GCP-2/CXCL6, ENA-78/CXCL5 and Mouse GCP-2/LIX and Modulate Their Physiological Activities." *European Journal of Biochemistry* **270**: 3739–3749.

White G. B. (1974). "Anopheles gambiae complex and disease transmission in Africa". *Transactions of the Royal Society of Tropical Medicine and Hygiene* **68**:278-302.

Wilhelm S. M., Collier I. E., Marmer B. L., Eisen A. Z., Grant G. A., and Goldberg G. I. (1989). "SV40-transformed Human Lung Fibroblasts Secrete a 92-kDa Type IV Collagenase Which

Is Identical to That Secreted by Normal Human Macrophages." *The Journal of Biological Chemistry*.**264**: 17213–17221.

Woessner J. F. Jr. (1991) "Matrix metalloproteinases and their inhibitors in connective tissue remodeling." *FASEB Journal*. **5**: 2145–2154

WHO (2012) The World Malaria Report. WHO, Geneva.

White G. B. (1985) "Anoheles bwambae sp.n., a malaria vector in the Semliki Valley , Uganda and its relationships with other sibling species of the An gambiae complex (Diptera: culicidae)." *Systematic Entomology* **10**:501-522

Yoneda A., Ogawa H., Kojima K., and Matsumoto I. (1998). "Characterization of the ligand binding activities of vitronectin: interaction of vitronectin with lipids and identification of the binding domains for various ligands using recombinant domains." *Biochemistry* **37**:6351-6360.

Yong V. W., Power C., Forsyth P., and Edwards D. R. (2001). "Metalloproteinases in Biology and Pathology of the Nervous System." *Nature Reviews: Neuroscience* **2**: 502–511.

Yoshizaki T., Sato H., and Furukawa M. (2002). "Recent advances in the regulation of matrix metalloproteinase 2 activation: from basic research to clinical implication (Review)." *Oncology Reports* **9**: 607-611.

APPENDICES

APPENDIX I: Erroneous structure of AGAP003929

APPENDIX II: Scripts for modeling and assessment of energy

1. Script for homology modeling of 3D structure using a template file

```
from modeller import *
from modeller.automodel import * # loads the automodel class
log.verbose() # request verbose output
env = environ() # Create a modellereny to build in model
env.io.atom_files_directory=''

a = automodel(env,
        alnfile='mmp1.pir', # alignment filename
        knowns= ('1su3_A','3c7x_A'),    # Template file name
        sequence='mmp1',    #Sequence_name
        assess_methods=(assess.DOPE, assess.GA341))
a.starting_model = 1 # index of first model
a.ending_model = 50 # index of last model
a.md_level = refine.very_slow # slow refinement
a.make()    # do actual modelling
```

2. Script for determining the discrete optimized protein energy for every model generated

```
# Example for: model.assess_normalized_dope()

from modeller import *
from modeller.scripts import complete_pdb

env = environ()
env.libs.topology.read(file='$(LIB)/top_heav.lib')
env.libs.parameters.read(file='$(LIB)/par.lib')

# Read a model previously generated by Modeller's automodel class
mdl = complete_pdb(env, 'mmp1.B99990050.pdb')

zscore = mdl.assess_normalized_dope()
```

APPENDIX III: DNA sequences from Sanger sequencing

>MM1P1 Adult

CGGCGGGAACCTGCTGGACCAGGACACCTGGGAGAAAGCCATCATGGAGTTCCAGA
GCTTTGCCGGGCTGAATGTTACCGGCGAGCTGGACGGCGAAACGATGCAGCTCATG
TCGCTGCCCCGGTGTGGCGTGAAGGATAAGGTTGGCTTTGGGTCCGACACCCGCTCG
AAGCGCTACGCCCTGCAGGGCAGCCGCTGGAAGGTGAAGGATCTTACCTACCGGAT
ATCCAAGTACCCGAGGCGGCTGGAACGGACGGCGGTGGATAAGGAGATCGCGAAA
GCGTTCGGCGTGTGGAGCGAGTACACGGATTTGCGCTTTACGCCGAAGAAAACGGG
CGCAGTTCATATCGACATTAGGTTCGAGGAGAACGAACACGGTGATGGTGATCCGTT
TGACGGACCGGGCGGCACTCTGGCCCACGCGTACTTCCCCGTGTACGGTGGTGATGC
ACACTTTGACGACGCCGAACAGTGGACGATTGATAAGCCACGCGGGACGAATCTGT
TCCAGGTGGCAGCGCACGAGTTTGGCCACTCGCTGGGTCTGAGTCACTCCGACATAC
CATTCTCACTGGTGTCACGGTTCTACCGGGCA

>MMPI Larvae

GGAACCCCGGCGAGCGGGAACCTGCTGGACCAGGACACCTGGGAGAAAGCCATCAT
GGAGTTCCAGAGCTTTGCCGGGCTGAATGTTACCGGCGAGCTGGATGGCGAAACGA
TGCAACTCATGTCGCTGCCCCGGTGTGGCGTGAAGGATAAGGTTGGCTTTGGGTCCG
ACACCCGCTCGAAGCGTTACGCCCTGCAGGGCAGCCGCTGGAAGGTGAAGGATCTT
ACCTACCGGATATCCAAGTACCCGAGGCGGCTGGAACGGACGGCGGTGGATAAGGA
GATCGCGAAAGCGTTTGGCGTGTGGAGCGAGTACACGGATTTGCGCTTTACGCCGA
AGAAAACGGGCGCCGTTCATATCGACATTAGGTTCGAGGAGAACGAACACGGTGAT
GGTGATCCGTTTGACGGACCGGGCGGCACCCTGGCCCACGCGTACTTCCCCGTGTAC
GGTGGTGATGCACACTTTGACGACGCCGAACAGTGGACGATTGATAAGCCACGCGG
GACGAATCTGTTCCAGGTGGCAGCGCACGAGTTTGGCCACTCGCTGGGTCTGAGTCA
CTCCGAGATACCATTCGCACTGGTGGCACGGTTCTAC

>MMPI Pupae

CGGCGGGAACCTGCTGGACCAGGACACCTGGGAGAAAGCCATCATGGAGTTCCAGA
GCTTTGCCGGGCTGAATGTTACCGGCGAGCTGGACGGCGAAACGATGCAGCTCATG
TCGCTGCCCCGGTGTGGCGTGAAGGATAAGGTTGGCTTTGGGTCCGACACCCGCTCG

AAGCGCTACGCCCTGCAGGGCAGCCGCTGGAAGGTGAAGGATCTTACCTACCGGAT
ATCCAAGTACCCGAGGCGGCTGGAACGGACGGCGGTGGATAAGGAGATCGCGAAA
GCGTTCGGCGTGTGGAGCGAGTACACGGATTTGCGCTTTACGCCGAAGAAAACGGG
CGCAGTTCATATCGACATTAGGTTCGAGGAGAACGAACACGGTGATGGTGATCCGTT
TGACGGACCGGGCGGCACTCTGGCCCACGCGTACTTCCCCGTGTACGGTGGTGATGC
ACACTTTGACGACGCCGAACAGTGGACGATTGATAAGCCACGCGGGACGAATCTGT
TCCAGGTGGCAGCGCACGAGTTTGGCCACTCGCTGGGTCTGAGTCACTCCGACATAC
CATTCTCACTGGTGTCACGGTTCTACCGGGCA

APPENDIX IV: Protein sequences retrieved from NCBI and used in 3-D structure prediction

>gi|157020567|gb|AGAP006904| AGAP006904-PA [Anopheles gambiae str. PEST]
MLRNHAHWIRALAIVLVGACAAGTASPVSTTPQAELYLSQFGYLSPKYRNPTSGNLLDQ
DTWEKAIMEFQSFAGLNVTGELDGETMQLMSLPRCGVKDKVGFGSDTRSKRYALQGS
RWKVKDLTYRISKYPRRLERTAVDKEIAKAFGVWSEYTDLRFTPKKTGAVHIDIRFEEN
EHGDGDPFDGPGGTLAHAYFPVYGGDAHFDDAEQWTIDKPRGTNLFQVAAHEFGHSL
GLSHSDVRSALMAPFYRGYDPVFRLDSDDIQGIQTLYGTKTRNPGGGAGATPTRTPRPK
TPTEMDSELCTSPKIDAIFNTADGSTYAFKGDKYYKLTENAVAEGYPKKISDGWPGLPG
NIDAAFTYKNGKTYFFQGTKYWRYQGRTIDGDYPKEISEGFTGVPDHLDAAMVWGGN
GKIYFYKGSKFWRFDPLKRPPVKSTYPKPISNWEGVPNSVDAALQYTNGYTYFFKDDK
YYRFNDRTFTVDQSDPPFPRPTAHWWYGCKNTPSTFNTLGNVRLQKSDEHPYDIGDLA
QDAADTDDQPDPDGYSGNGASTITGTSVSITTALAGFLLGAYLVSRC

>gi|347970929|ref|AGAP003929| AGAP003929-PA [Anopheles gambiae str. PEST]
MARYPLHLFGLCFLCLISLRLMHGAPAVPTKEMIDFMRRFGYLEKGPTQAEALYSGEAII
DAIKHVQKFGALPQTGVLDRRTIELMSAPRCGVVDVMQHDQSLRHRRYVIGSESWRKR

RITYFIANWSSKVGEDAVAKFMAKAFGEWSKYSKLRFVRVYDPSADIIVGFGSGHHGD
NYPFDGPGNVLAHAFYPYEMNAYGGDVHFDEDENWKENSTHLSEGVDFYSVAIHELG
HSLGLAHSPVYSSLMFPYYKGIAQGTLDYDDILAMYQLYIQNPHITDEPDWMYTTEAST
TVDEYGTVTPAVPRLPDLDSEPHYPEPDPLPTSTPALSSTTEVYDIPITFVGDYETVDDHIS
RHHAQSPPPTSVTTPPEDRPPAAIPSYVPVPDICSGSFDAIGLLRGEIFIFKGAYLWRLTEK
YRIKTGYPVRIWQVFRGFPKTVSHIDAVYERLDDNAIVLFSGRFYWVFDALNFLHPEVR
PLTDFGLPEELRRIDAALVWPKNNKTYLFAGDRFWRYNDTAGEMDEGYPSSMDRWFGI
PNNIDAATAVASGKFEACQEAETEETHTTRHGSREWRNCVREEAGNEARPRWARRTSS
KGTTTGCTIMSGSARNVDTLGGRVIYGSVADKMRMLNASRAPNAVGLSECISTATRTFN
IGRKVRRCVTAYVYCIFKPKLLNSAFSRQVAK