COMPARATIVE STUDY OF CLAN CA CYSTEINE PROTEASES: AN INSIGHT INTO THE PROTOZOAN PARASITES

A mini-thesis submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE OF RHODES UNIVERSITY

by

Coursework / Thesis

in

Bioinformatics and Computational Molecular Biology

in the Department of Biochemistry & Microbiology

Faculty of Science

by

Sipho Dugunye Moyo

February 2015

ABSTRACT

Protozoan infections such as Malaria, Leishmaniasis, Toxoplasmosis, Chaga's disease and African trypanosomiasis caused by the *Plasmodium, Leishmania, Toxoplasma* and *Trypanosoma* genuses respectively; inflict a huge economic, health and social impact in endemic regions particularly tropical and sub-tropical regions. The combined infections are estimated at over a billion annually and approximately 1.1 million deaths annually. The global burden of the protozoan infections is worsened by the increased drug resistance, toxicity and the relatively high cost of treatment and prophylaxis. Therefore there has been a high demand for new drugs and drug targets that play a role in parasite virulence. Cysteine proteases have been validated as viable drug targets due to their role in the infectivity stage of the parasites within the human host. There is a variety of cysteine proteases hence they are subdivided into families and in this study we focus on the clan CA, papain family C1 proteases. The current inhibitors for the protozoan cysteine proteases lack selectivity and specificity which contributes to drug toxicity. Therefore there is a need to identify the differences and similarities between the host, vector and protozoan proteases. This study uses a variety of bioinformatics tools to assess these differences and similarities. The Plasmodium cysteine protease FP-2 is the most characterized protease hence it was used as a reference to all the other proteases and its homologs were retrieved, aligned and the evolutionary relationships established. The homologs were also analysed for common motifs and the physicochemical properties determined which were validated using the Kruskal-Wallis test. These analyses revealed that the host and vector cathepsins share similar properties while the parasite cathepsins differ. At sub-site level sub-site 2 showed greater variations suggesting diverse ligand specificity within the proteases, a revelation that is vital in the design of anti-protozoan inhibitors.

DECLARATION

The research described in this thesis was carried out as part of the one-year MSc coursework and research thesis programme in Bioinformatics and Computational Molecular Biology, from 15 July 2014 to 31 January 2015 under the supervision of Prof Özlem Taştan Bishop.

I, SIPHO DUGUNYE MOYO, declare that this thesis submitted to Rhodes University is my own work and has not previously been submitted for a degree in this or any other university.

Signature ……  ……

Date…6th February 2015……

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to the following that made the completion of this degree and writing of this thesis a success:

- The Lord God almighty for bringing me thus far
- My supervisor Prof. Özlem Tastan Bishop for her unwavering encouragement, advice and guidance
- Ngonidzashe Faya for his constant assistance
- My family and my boyfriend Munya for their support and motivation
- My RUBi colleagues for their assistance and friendship
- NIH Common Fund (H3ABioNet) for the funding

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

2D	Two Dimensional
3D	Three dimensional
ACT	Artemisin-based Combination Therapy
BIC	Bayesian Information Criterion
BLAST	Basic Local Alignment Tool
BLASTP	Basic Local Alignment Tool for Protein sequences
BLOSUM	BLOcks of Amino Acid Substitution Matrix
CNS	Central Nervous System
CT	Computerized Tomography
DNA	Deoxyribonucleic Acid
FP-1	Falcipain-1
FP-2	Falcipain-2
FP-2B	Falcipain-2B
FP-3	Falcipain-3
HIV	Human Immunodeficiency Virus
HMM	Hidden Markov Models
HTML	HyperText Markup Language
JTT	Jones Taylor Thornton
MAFFT	Multiple sequence Alignment based on Fast Fourier Transform
MATLAB	Matrix Laboratory

MEGA	Molecular Evolutionary Genetic Analysis
MEME	Multiple EM for Motif Elicitation
MRI	Magnetic Resonance Imaging
MSA	Multiple Sequence Alignment
NCBI	National Centre for Biotechnology Information
PAM	Point Accepted Mutations
PDB	Protein Data Bank
PDB_ID	Protein Data Bank Identification Number
PROMALS	PROfile Multiple Alignment with Local Structure
PSI- BLAST	Position Specific Iterated - Basic Local Alignment Tool for Protein sequences
PSI-PRED	PSI-blast based secondary structure PREDiction
PWM	Position Weight Matrices
RSCB	Research Collaboratory for Structural Bioinformatics
S1	Sub-site 1
S1’	Sub-site 1’
S2	Sub-site 2
S3	Sub-site 3
T-Coffee	Tree-based Consistency Objective Function for alignment Evaluation
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
WAG	Whelan and Goldman
WHO	World Health Organization

XML EXtensible Markup Language

TYPOGRAPHICAL CONVENTIONS

In this study the species are referred to with their Latin names

E.Histolytica	*Entamoeba Histolytica*
G.gallus	*Gallus gallus*
L.major	*Leishmania major*
L.mexicana	*Leishmania mexicana*
L.tropica	*Leishmania tropica*
L.aethopica	*Leishmania aethopica*
P. falciparum	*Plasmodium falciparum*
P. vivax	*Plasmodium vivax*
P. knowlesi	*Plasmodium knowlesi*
P. ovale	*Plasmodium ovale*
P. Malariae	*Plasmodium Malariae*
P. berghei	*Plasmodium berghei*
P. yoelii	*Plasmodium yoelii*
P. chabaudi	*Plasmodium chabaudi*
T.brucei	*Trypanosoma brucei*
T.cruzi	*Trypanosoma cruzi*
T.gondii	*Toxoplasma gondii*

Amino acid residues are referred to by their 3-letter abbreviations

SYMBOLS USED

α Alpha-helix

β Beta-sheets

Å Angstrom a measure of distance at atomic level

LIST OF WEB SERVERS USED

BLAST -www.ncbi.nlm.nih.gov/BLAST/

EXPRESSO - http://tcoffee.crg.cat/apps/tcoffee/do:expresso

HHpred - http://toolkit.tuebingen.mpg.de/hhpred

MAFFT -http://mafft.cbrc.jp/alignment/server/index.htmL

MEME -http://meme.nbcr.net/meme/cgi-bin/meme.cgi

PROMALS3D -http://prodata.swmed.edu/promals3d/promals3d.phpScan

T-Coffee - http://tcoffee.crg.cat/apps/tcoffee/

CHAPTER ONE

1. INTRODUCTION

1.1 PROTOZOAN PARASITES AND INFECTIONS

Protozoa are defined as microscopic organisms that possess animal-like features and exist as independent single cells (McKerrow et al. 2006). Their infections cause a substantial amount of human morbidity and mortality. The total number of deaths posed by the protozoan infections is estimated at around a million annually (Andrews et al. 2014). This study focuses on five major protozoan infections namely Malaria, Toxoplasmosis, Chaga's disease, Leishmaniasis and African trypanosomiasis. The spread of these infections requires transmission by vectors and in some cases intermediate hosts (Table 1.1).

Table 1.1 A summary of the protozoan infections of interest in this study showing their respective vectors and responsible protozoan parasites.

Protozoa	Protozoa parasite	Vector
Malaria	*Plasmodium spp.*	Female anopheles mosquito
Leishmaniasis	*Leishmania spp.*	Female sand fly
African Trypanosomiasis	*Trypanosoma brucei*	Tsetse fly
Chaga's disease	*Trypanosoma cruzi*	Triatomine bug
Toxoplasmosis	*Toxoplasma gondii*	Contaminated meat and feline contact

The World Health Organization (WHO) states that approximately 50 % of the global population suffers from one parasitic infection or another (World Health Organization, 2010). This chapter will focus on the above infections and summarize their epidemiology, parasite lifecycle, diagnosis and treatment as well as their current treatment and drug targets.

1.1.1 MALARIA

Malaria is the most fatal of all protozoan infections which threatens about 3.3 billion lives resulting in ~0.6 – 1.1 million deaths yearly (Murray et al. 2013). The disease is caused by the *Plasmodium* genus namely; *P. falciparum* being the most severe and fatal type, *P. vivax, P.ovale, P.knowlesi* and lastly *P.Malariae* (Mharakurwa et al. 2012). A majority of Malaria infections and deaths are caused by *P. falciparum* and *P. vivax* infections in both children and adults (World Health Organization 2013). *P.falciparum* is an epidemic in sub-Saharan Africa, while *P. vivax* causes a huge amount of morbidity particularly in the Asia-Pacific and South-American regions (Murray et al. 2013).

The *Plasmodium* parasites are spread by the female anopheles mosquito through an infective bite of the Malaria-infected female anopheles mosquito which inoculates the sporozoites either into the subcutaneous tissue or into the bloodstream during a blood meal (Korde et al. 2008). Almost immediately the sporozoites infect the liver cells. Here they gradually mature into schizonts which eventually disrupt and release merozoites (Lecaille et al. 2002) (Figure 1.1). The *P.vivax* and *P.ovale* species in particular may have a dormant stage (hypnozoites) that lingers in the liver resulting in relapses by invading the bloodstream weeks, or even years later (Collins & Jeffery 2007).

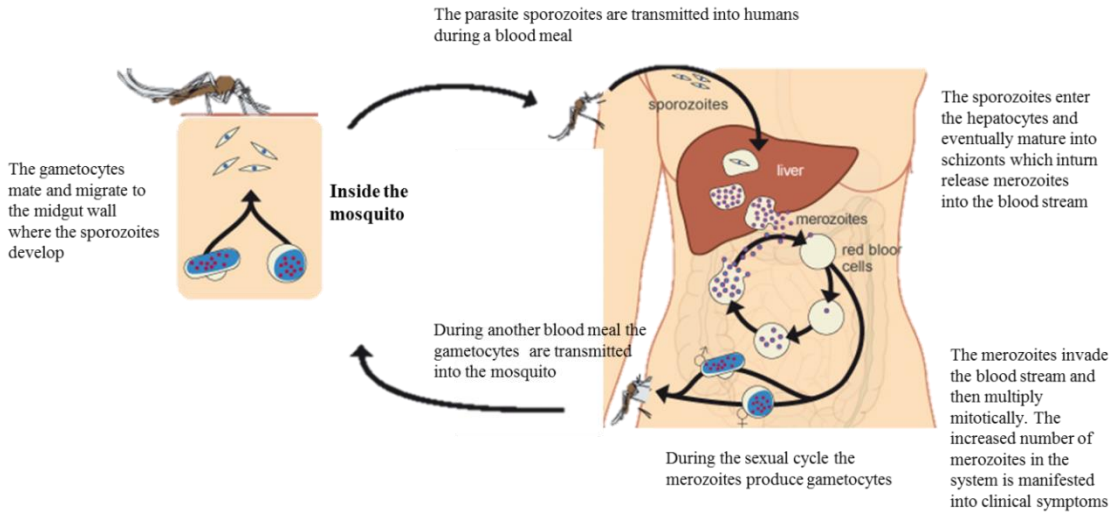


Figure 1.1 An illustration of the *Plasmodium* lifecycle within the human blood stream and the female anopheles mosquito. The drawing was adapted from (http://www.dw.de) webpage and edited accordingly.

As the concentration of the parasite increases in the blood stream the clinical symptoms such as headaches, shaking chills and tiredness are manifested (Autino et al. 2012). The symptoms are used as a basis for diagnosis. However this is a challenging method since the Malaria symptoms are non-specific and there is a substantial overlap with other diseases (Tangpukdee et al. 2009). This results in the over-treatment of Malaria or non-treatment of other diseases in Malaria-endemic regions, and misdiagnosis in non-endemic areas. Therefore to increase the efficiency of the diagnosis procedure the clinical methods are coupled with laboratory methods. The laboratory diagnosis involves the identification of the Malaria parasites in the patients' blood through microscopy (Tangpukdee et al. 2009).

1.1.2 AFRICAN TRYPANOSOMIASIS

This is also commonly known as African sleeping sickness caused by the *Trypanosoma* genus. There are two forms of the disease found in West and East Africa caused by the *Trypanosoma brucei gambiense* and *Trypanosoma brucei rhodesiense*, respectively (Chappuis et al. 2005).

The African sleeping sickness threatens the lives of approximately 60 million people in Sub-Saharan Africa. This threat is spread by the tsetse fly (Barrett et al. 2003). The parasite is transmitted when the infected insect takes a blood meal and secretes the parasite from its salivary gland into the human blood stream (Malvy & Chappuis 2011) (Figure 1.2).

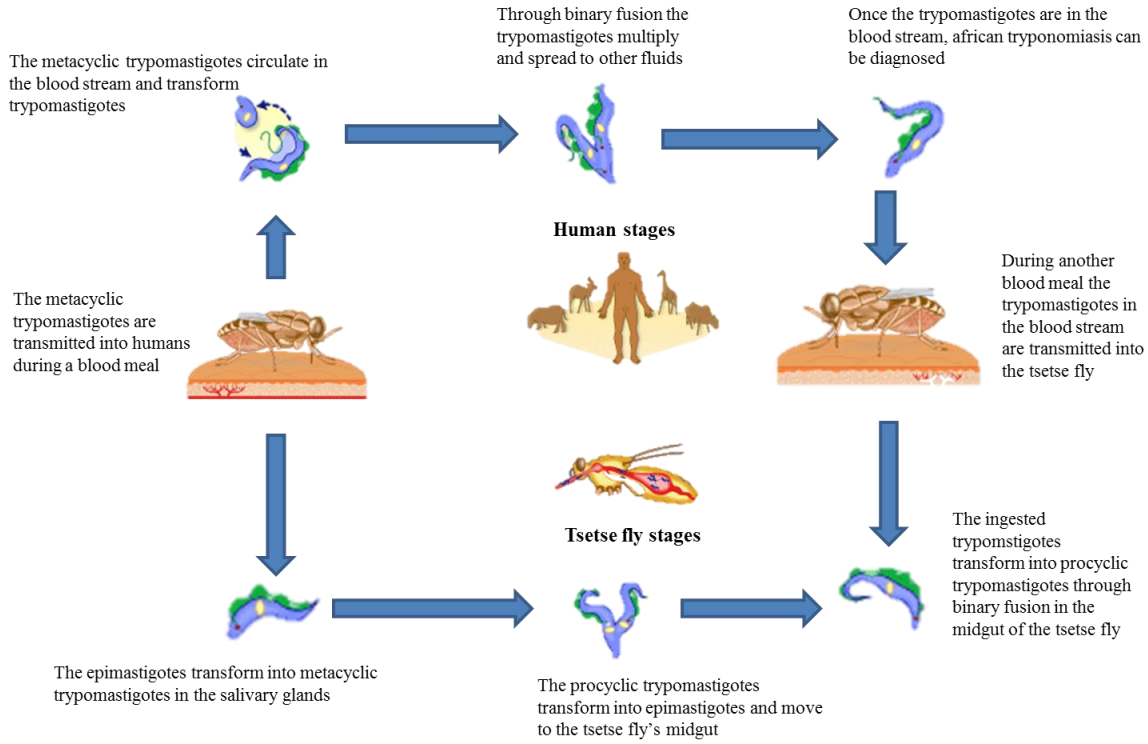


Figure 1.2 A diagrammatic illustration of the lifecycle of *Trypanosoma brucei* within the human host and its tsetse fly vector. The drawing was adapted from (http://www.who.int) and edited accordingly.

The trypomastigotes injected into the blood stream during the fly's blood meal multiply in the blood, spinal fluid and the lymph fluid. The symptoms are manifested as a consequence of the presence of trypomastigotes in these body fluids particularly in the blood (Enanga et al. 2002). The symptoms include fever, arthralgia, headache, and malaise in the hemolymphatic stage and motor and sensory disorders, endocrine disorders and weight loss in the advanced Central Nervous System (CNS) stage (Barrett et al. 2003). The diagnosis of the disease involves straight forward microscopy of the lymph and spinal fluid specimens (Chappuis et al. 2005).

1.1.3 CHAGA'S DISEASE

Chaga's disease is caused by a *Trypanosoma cruzi* infection. It is primarily found in the United States, Latin America and Mexico. Approximately 8-10 million people are reported to be infected by the parasite annually (Rassi Jr et al. 2010). The prevalence of the disease is exacerbated by the kissing bug which transmits *Trypanosoma cruzi*. The insect infects the host by secreting on the host through a blood meal and then defecates on the host. The secretions

cause the host to itch; the scratching forms microabrasions through which the trypomastigotes in the faeces enter the host. The parasite then moves to the blood and lymph systems where it multiplies (Barrett et al. 2003) (Figure 1.3).

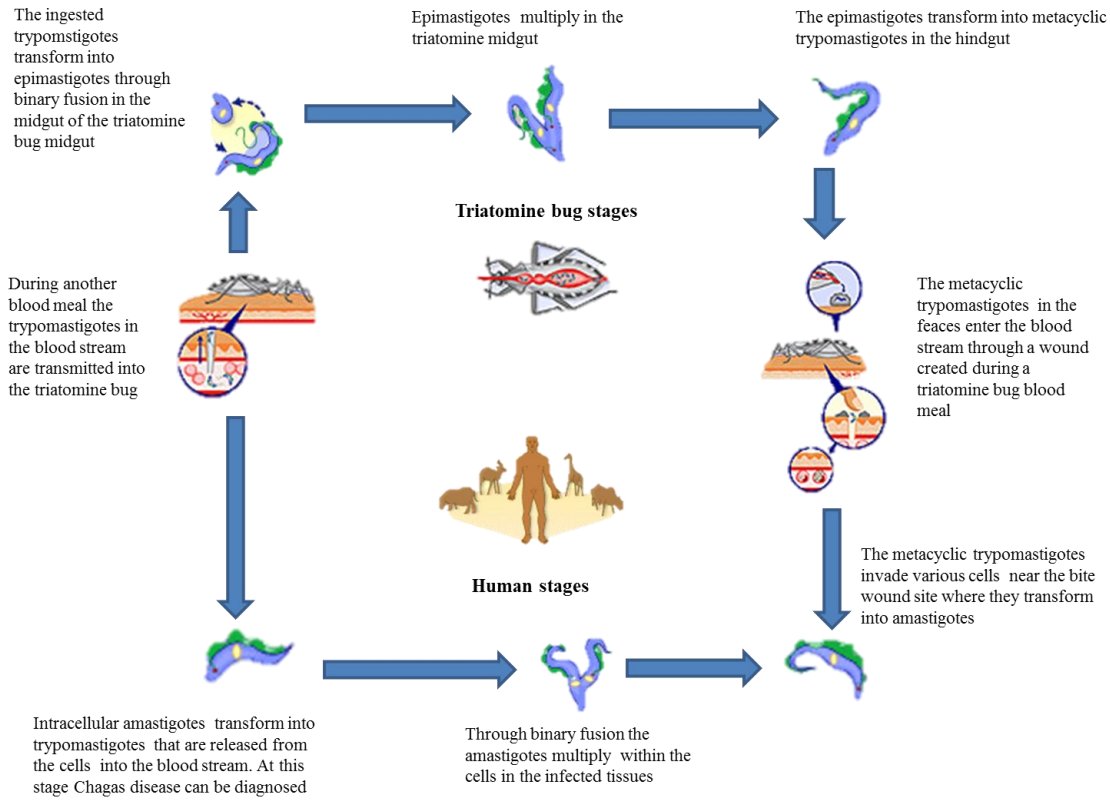


Figure 1.3 An illustration of the *Trypanosoma cruzi* lifecycle showing the human and triatomine bug stages. The diagram was adapted from the (http://www.who.int/disease/Chagas/lifecycle) and edited accordingly.

Chaga's disease primarily affects the spine and the heart. The initial stage is acute and involves symptoms such as malaise, anorexia, fever, vomiting and diarrhoea. The acute initial stage is followed by a chronic infection that usually leads to neurological disorders (Kirchhoff 2011). The acute phase is usually misdiagnosed unlike the chronic phase which is usually diagnosed through direct observation of the trypomastigotes in the patients' blood smear (Antunes et al. 2002).

1.1.4 LEISHMANIASIS

Leishmaniasis is caused by the *Leishmania* species that are intracellular organisms that cause three major disorders in humans namely mucocutaneous Leishmaniasis, visceral Leishmaniasis and cutaneous Leishmaniasis; all have varying immunopathologies hence pose different degrees of mortality and morbidity (Andrews et al. 2014).

There are approximately 12 million estimated infections of Leishmaniasis globally and the associated deaths are estimated at 50 000 annually (Delgado et al. 2008). Approximately 500 000 infections are reported to be cases of visceral Leishmaniasis and 1.5 million of cutaneous Leishmaniasis (Alvar et al. 2012).

The parasite is transmitted by the Phlebotomine sand fly, through an infective bite of the insect. During the blood meal the infected sand fly injects the promastigotes into the uninfected human; the promastigotes are then phagocytosed by macrophages. Within this phagolysome the parasites multiply and destroy the host cell (Wittner & Tanowitz 2000) (Figure 1.4).

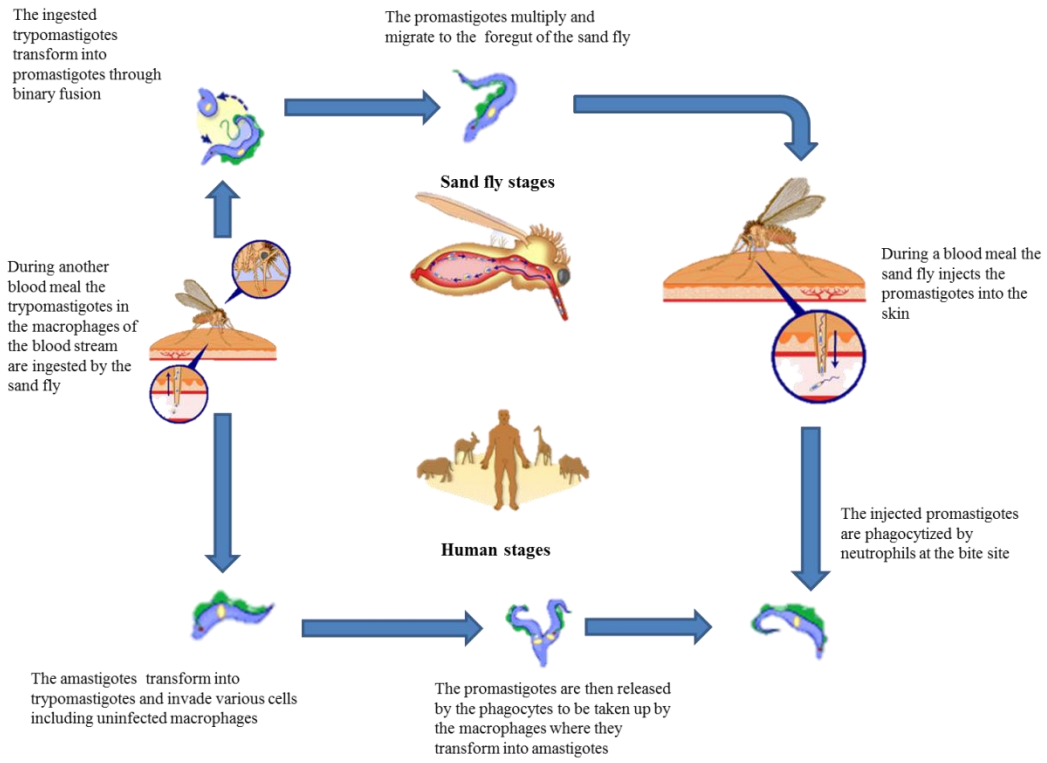


Figure 1.4 A diagrammatic illustration of the *Leishmania* lifecycle in both sand fly and human stages. The diagram was adapted from (http://www.who.int) webpage and edited accordingly.

Currently Leishmaniasis is diagonised using real-time PCR assays, immunoassays, and microscopy and culture systems (Singh & Sivakumar 2003). The microscopy method is preferred in diagonosing Leishmaniasis. The specimens from patients are stained and the *Leishmania* parasite is identified through direct microscopy (Rosenblatt 2009). In some cases microscopy maybe unrealistic hence serological tests such as the dipstick test maybe used. The dipstick tests are rapid, sensitive and inexpensive (Rosenblatt 2009).

1.1.5 TOXOPLASMOSIS

Toxoplasma gondii is the intracellular protozoan that causes Toxoplasmosis. The disease has a global epidemiology and affects about 25–30% of the global population (Montoya & Liesenfeld 2004). It is usually transmitted through the ingestion of meat contaminated with tissue cysts such as pork or ingestion of oocysts shed by household cats. The parasite can also be transmitted from mother to child resulting in foetal death or birth defects (Montoya & Liesenfeld 2004). The parasite is resistant to the immune system (Dubey 2009). The resistant structures allow the parasite to mature within the host resulting in the eventual rupture of the tissue cysts thus releasing the tachyzoite form of the parasite that infects new host cells (Weiss & Dubey 2009) (Figure 1.5).

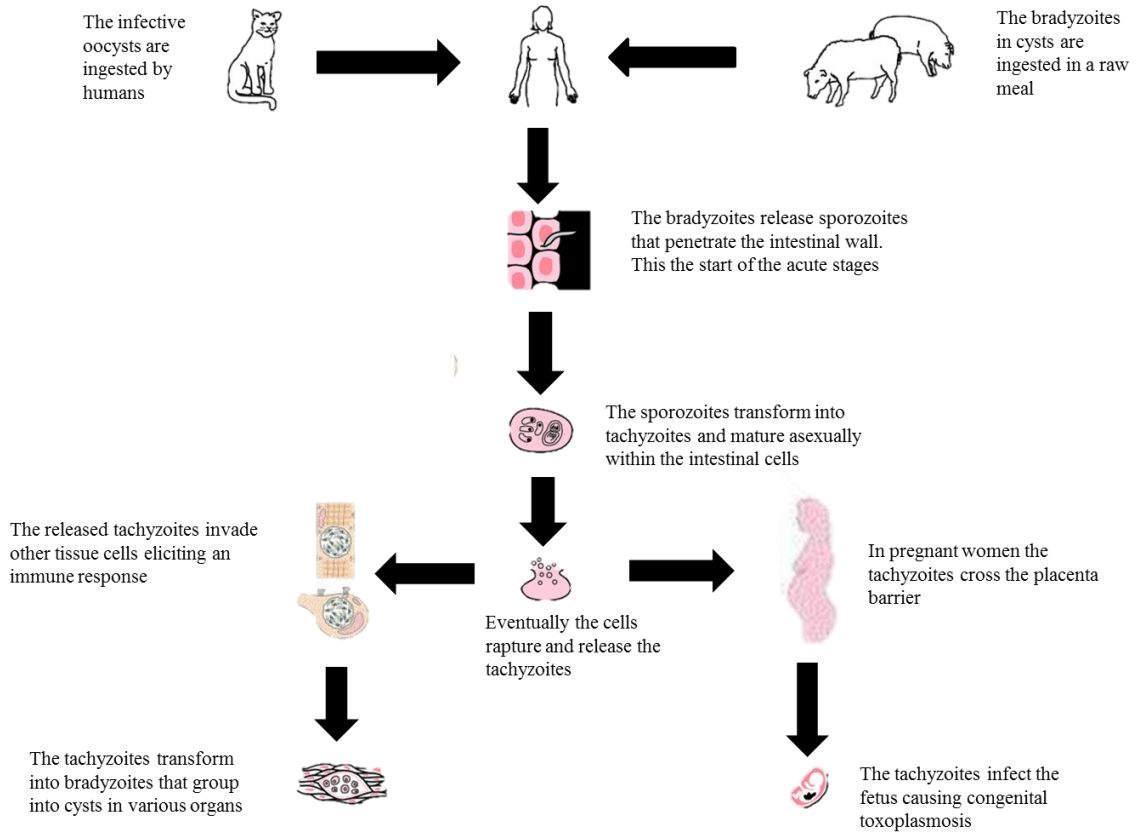


Figure 1.5 An illustration of the transmission and lifecycle of the *Toxoplasma* parasite within the hosts and vectors. The images were adapted from (http://www.scielo.br) and (http://medical-dictionary.thefreedictionary.com) webpages and edited accordingly.

The clinical manifestations of the disease are often moderate if they do exist in healthy individuals; unfortunately in immunocompromised individuals such as human immunodeficiency virus (HIV) patients the symptoms can be severe (Weiss & Dubey 2009). The diagnosis of the disease includes a magnetic resonance image (MRI) or a computerized tomography (CT) scan of the brain. Unfortunately the results from a CT scan or MRI may be inconclusive hence the diagnosis is coupled with microscopy, parasite cultivation or molecular methods (Miller et al. 1998).

1.2 ANTI-PROTOZOAN TREATMENT

The treatment of protozoa infections is largely dependent on chemotherapy. In Leishmaniasis cases antimonial compounds are the first line of treatment in developing countries mainly because they are readily available and have a low cost however they are quite toxic (Sundar &

Chakravarty 2013). In developed countries amphotericin B is preferred due to its reduced toxicity although the treatment is expensive and many patients cannot afford it (Sundar & Chakravarty 2013). This is also the case in treating Chaga's disease. The drugs of choice are usually benznidazole and nifurtimox which have shown great efficacy (Antunes et al. 2002). Nevertheless the treatment requires extended courses resulting in drug toxicity and resistance in some cases (Apt 2010).

African trypanomiasis is treated according to its stages. The disease progresses in two stages namely the hemolymphatic stage and the CNS stage (Brun et al. 2010). During the hemolymphatic stage the patients suffer from headaches, malaise, weakness and fever, while in the CNS stage trypomastigotes invade the CNS and the patients' experience a disruption in their sleep-wake cycle resulting in a strong urge to sleep (Kennedy 2004). This is usually coupled with motor and sensory disorders. In the hemolymphatic stages pentamidine and suramin are the drugs of choice while Eflornithine is preferred in treating the CNS stage (Barrett et al. 2011). These drugs have adverse side effects while having limited efficacy (Delespaux & de Koning 2007).

The same challenges are faced in treating Toxoplasmosis. There are currently limited treatment options for Toxoplasmosis. The acute forms of the disease are treated using a combination of sulfonamide and pyrimethamine, azithromycin is also another alternative (Pereira-Chioccola et al. 2009). Unfortunately adverse side effects of the pyrimethamine combination have been reported and the alternative azithromycin has raised some concerns about its efficacy (Weiss & Dubey 2009).

In Malaria cases the treatment has relied on chloroquine but due to the extensive misuse of this drug there have been cases of drug resistance reported in endemic regions (Wegscheid-Gerlach et al. 2010). Since 2012 Artemisinin-based combination therapies (ACTs) are the best on the market however the therapy is inaccessible in developing countries. Since ACTs are a combination therapy there is misuse of the drug due to the lack of knowledge thus drug resistance is emerging (World Health Organization 2013).

The decreased efficacy, toxicity and inaccessibility associated with anti-protozoan treatments have spurred on the research for new drugs and drug targets. Research groups have focused on

cysteine proteases in an effort to identify novel drug targets due to their role in parasite host cell invasion and rupture (McKerrow et al. 2006).

1.3 CYSTEINE PROTEASES

Proteases are a group of enzymes that facilitate the splitting of proteins into smaller fragments and can be grouped into various classes such as cysteine, serine, aspartate, threonine and metalloproteases, according to their catalytic activity and substrate specificity (Grzonka et al. 2001). Protozoan cysteine proteases are vital for host-cell invasion, host protein hydrolysis and the intracellular maturation of the parasite; hence they are a prerequisite for the survival and the proliferation of the parasite in its vector and host (Hogg et al. 2006). It is these characteristics that arouse interest in cysteine proteases as potential targets for anti-protozoan drugs.

Cysteine proteases exploit the catalytic cysteine residue localized in the active site where it mediates the peptide bond hydrolysis through a nucleophilic attack on the carbonyl carbon. The nucleophilic attack is carried out by the thiol-group in the catalytic dyad formed by the sulfhydryl group of the cysteine residue and the imidazole ring of the histidine residue (McKerrow et al. 2006).The fundamental step for this hydrolysis process is the formation of the S-acyl-enzyme moiety followed by the nucleophilic attack of the thiolate group which releases the C-terminal fragment of the cleaved protein and finally the reaction of the water molecule with the intermediate group releasing the N-terminal fragment of the cleaved product (Pandey et al. 2009).

1.3.1 CYSTEINE PROTEASE NOMENCLATURE

Cysteine proteases are grouped into various clans that have divergent sequence and structure similarity namely; clan CA, CB and CC. This study focuses on the CA clan that utilizes catalytic Cys, His and Asn residues that are in this perpetual order in the primary sequence of the protease (Pandey & Dixit 2012). The CA clan is further subdivided into families according to sequence similarity and identity namely; family C1 and C2. Family C1 (papain‑family) cysteine proteases are the best characterized cysteine proteases of parasite proteases. This family is further categorized into classes namely; cathespsin-L-like, cathepsin-B-like and cathepsin-F-like proteases (McKerrow et al. 2006).

The best characterized *Plasmodium* cysteine proteases are the falcipains, papain family (clan CA) from the cathepsin-L-like subclass enzymes namely; falcipain-1 (FP-1), falcipain-2 (FP-2), falcipain-2B (FP-2B) and falcipain-3 (FP-3) (Selzer et al. 1999). FP-2 is the most studied of all the falcipains hence it is used as a reference in this study.

1.3.2 CYSTEINE PROTEASE STRUCTURE

Cysteine proteases are initially synthesized as zymogens to prevent the potentially premature degradation of the protein. The zymogen displays two major domains, the prodomain and the mature domain (Rosenthal et al. 2002). The prodomain is further differentiated into smaller domains. At the first 35 amino acids of the prodomain N-terminus is the cytosolic prodomain, followed by the transmembrane made up of 20 amino acid residues which is required for entry into the ER and finally the lumenal domain consisting of 188 amino acid residues. The inhibitory domain is located in the C-terminal portion of the prodomain (Rizzi et al. 2011). These domains were elucidated using transfection studies whereby the different chimeras with parts of the N-terminal portion of the prodomain where fused to green fluorescent protein and their localization determined (Wang et al. 2006).

The prodomain spans over the active site thus demonstrating a substrate inhibitory effect. To prevent hydrolysis the prodomain is bound to the active site in the opposite direction to that of substrate binding (Pandey et al. 2009). The prodomain consists of 2 α-helices located at the N-terminal. The prodomain is anchored to the active site of the protease by a short α-helix following the two α-helices. The C-terminal end of the prodomain covers both the left and right lobes of the mature domain (Sijwali et al. 2006). Apart from inhibiting the active site of the inactive zymogen, the prodomain plays a key role in intracellular trafficking. Once the prodomain is cleaved off the cysteine proteases are activated (Wang et al. 2006).

The mature domain consists of two lobes; the right lobe is made up of α-helices while the left lobe is made up of beta sheets and surface α-helices. The catalytic triad (Cysteine, Histidine and Asparagine) is located in the intersection between the two lobes (Wang et al. 2006). The mature domain houses the active site which is subdivided into 4 regions namely, sub-sites 1, 1', 2 and 3 (S1, S1', S2 and S3). Of all the binding pockets S1 is the least defined, it holds the glutamine residue that forms the oxyanion hole that stabilizes the substrate while interacting with the

substrate solvent. S1' is known to hold the highly conserved tryptophan which forms the hydrophobic interactions with the substrate (Sabnis et al. 2003). The S2 is the most conserved of the sub-sites indicating its importance in the catalytic activity of the protease. This region is primarily hydrophobic and it is the region that governs ligand specificity (McKerrow et al. 2006). S3 is responsible for the stabilization of the intermediates during catalysis and interacts with the substrate although it is not involved in substrate specificity (Sabnis et al. 2003) (Figure 1.6).

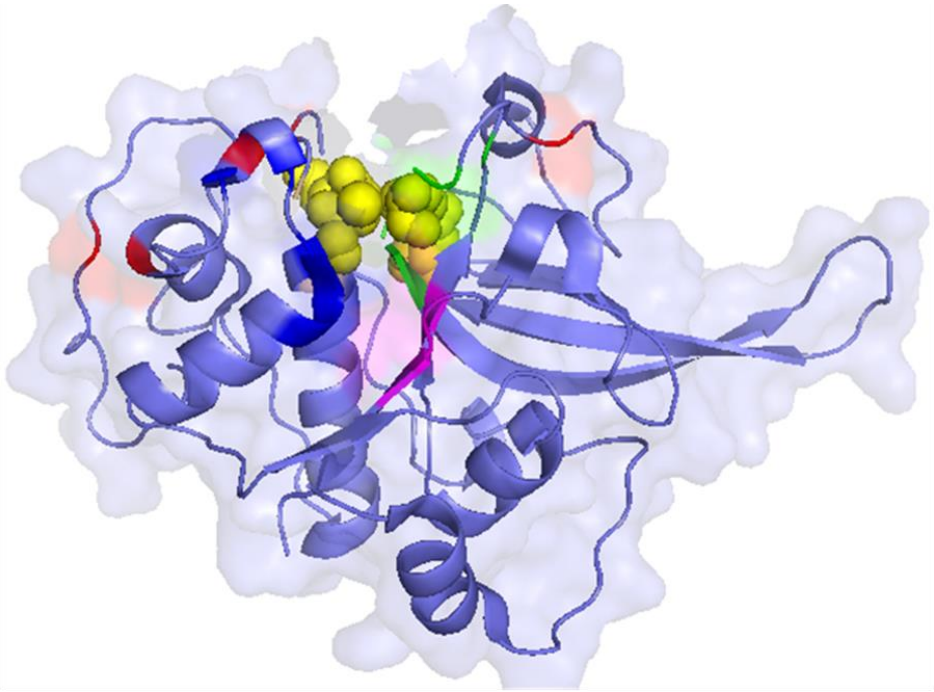


Figure 1.6 The sub-sites are mapped onto the *Plasmodium* structure (2OUL) sub-site 1 is shown in red, sub-site 1' in wheat, sub-site 2 in magenta and sub-site 3 in blue. The active site is shown as yellow spheres (see **Chapter 2.3.2.1**).

1.3.3 FUNCTION OF CYSTEINE PROTEASES IN PROTOZOAN PARASITES

In *T.gondii* cysteine proteases play a vital role in parasite invasion and growth. The protease is present in the vacuolar compartment of the parasite, although its functions in the vacuolar compartment are not known it is believed that the protease is involved in protein degradation to facilitate the parasite maturation and food acquisition within the host (Dou & Carruthers 2011). In *Plasmodium* species protein inhibitor studies showed the role cysteine proteases play in

hemoglobin hydrolysis whereby the free amino acids are used for the sustenance of parasite osmotic stability preventing the premature bursting of the parasite and also provide space for the maturing intraerythrocytic parasite (Rosenthal et al. 2002). Cysteine proteases in *Plasmodium* species have also been implicated in erythrocytic invasion by merozoites and their role in this process was demonstrated using FP-1 inhibitors that blocked the invasion of the erythrocytic host in vitro by *Plasmodium* parasites (Wegscheid-Gerlach et al. 2010). Falcipains have also been implicated in the subsequent host erythrocytic rupture to release merozoites that rapidly invade neighbouring erythrocytes to reinitiate the asexual cycle (Blackman 2004). The cysteine protease role in erythrocytic rupture was illustrated by cysteine protease inhibitors leupeptin and E-64, which showed the aggregation of mature schizonts in cultures treated with the inhibitors (Sijwali et al. 2006).

Inhibitor studies have also shown that cysteine proteases are vital in the virulence of *Leishmania* species. The study was conducted by inhibiting the cathepsin-L of *Leishmania spp*. and the study showed an 80 % decrease in parasite virulence although the parasite continued to grow (Selzer et al. 1999). Meirelle and co-workers in 1992 reported a similar observation in *Trypanosoma* species, when they used irreversible inhibitors for cathepsin-L and discovered that the parasite had reduced ability to occupy cardiomycetes thus reducing its virulence as the newly released amastigotes failed to move away to less inflamed regions to propagate the infection (Abdulla et al. 2008).

1.3.4 CYSTEINE PROTEASES AS DRUG TARGETS

Cysteine proteases play a critical role in the protozoan infection and lifecycle completion making them lucrative drug targets. Studies have been conducted whereby the models where treated with cysteine protease inhibitors and these studies have validated the idea of cysteine proteases as drug targets (McKerrow et al. 2006). Unfortunately these inhibitors lack specificity resulting in toxicity and inhibition of the host cysteine proteases. This is a result of the enzyme being ubiquitous in virtually all organisms thus making it difficult to design specific protease inhibitors (Selzer et al. 1999). Due to this challenge of broad specificity of chemotherapeutic agents among other reasons discussed in section 1.2 there is a growing need for the design of highly selective inhibitors that exhibit minimal toxic effects posed on the host.

1.4 PROBLEM STATEMENT

Protozoan infections account for over a million deaths per year worldwide mainly young children and pregnant women in the sub-Saharan region in Africa (Barratt et al. 2010). A variety of drugs are presently accessible to treat these infections; however, treatment is currently limited by drug resistance, toxicity, and high cost (Bousema & Drakeley 2011). Therefore, there is a need for the development of effective treatment against these infections. Cysteine proteases play a vital role in the infectivity of the protozoan parasite in the human host hence they have been validated as a viable drug target (Sijwali et al. 2006). Cysteine proteases in general have broad substrate specificity which poses a huge challenge in chemotherapy design as the substrate inhibitors display poor selectivity towards the human cysteine proteases and parasite cysteine proteases; which is the main source of drug toxicity (Ettari et al. 2010). It is for this reason that there is a need for the design of protozoan cysteine protease inhibitors that are specific to the parasite.

This study focuses on the differences and similarities between the protozoan cysteine proteases as well as their host and vector cysteine proteases as a step towards designing specific and selective treatment against protozoan infections.

1.4.1 HYPOTHESIS

There are distinct differences between protozoan cysteine proteases and their respective host and vector cysteine proteases primary structure which could be vital in drug design.

1.4.2 AIM AND OBJECTIVES

In previous studies conducted by McKerrow and co-workers, the size and isoelectric points of both the catalytic domain and prodomain of the protozoan and human cysteine proteases were elucidated and compared however the study did not include other properties that influence the activity of the cysteine proteases. Hence to date there is limited knowledge on the differences attributed by the respective protein sequences, between the catalytic domains of the cysteine proteases of protozoa and their respective hosts and vectors. This has resulted in the failure to identify specific inhibitors for the cysteine proteases of protozoan parasites hence the rising mortalities due to the diseases the parasites cause. This study aims to fill this knowledge gap

through establishing the catalytic domain similarities and differences between the protozoans, their host and vector cysteine proteases with regard to their physicochemical properties (molecular weight, aromaticity, isoelectric point, GRAVY and instability index), motifs and evolutionary relationships; through a sequence analysis of the cysteine proteases of the protozoa, host and vector is performed followed by a phylogenetic analysis, motif analysis, physicochemical analysis and finally a statistical analysis to validate the physicochemical analysis. The results obtained from this study will prove vital in expanding the available knowledge regarding specific inhibitor design of anti-protozoans.

1.4.3 SPECIFIC OBJECTIVES

- Retrieve the sequences from the NCBI website and split the sequences into individual domains
- Perform a multiple sequence alignment of the catalytic domain to identify the regions of similarity and divergence among the retrieved sequences
- Evaluate evolutionary relatedness of the sequences through phylogenetic tree calculations using the optimal alignment
- Investigate common and unique patterns through motif analysis of both the full length sequence and the catalytic domain
- Establish the motifs functional or structural relevance by mapping the catalytic domain motifs to the available FP-2 3D structure
- Identify the amino acids responsible for observed patterns through an amino acid composition analysis of the catalytic domain and the sub-site regions
- Investigate the effect the amino acids have on the protein characteristics through physicochemical properties analysis of the catalytic domain and the full length sequences
- Visualize the physicochemical data using box plots
- Statistically validate the box plots data using the Kruskal-Wallis test

CHAPTER TWO

2. SEQUENCE AND PHYLOGENETIC ANALYSIS

The centre of interest in this chapter is to identify the orthologs of FP-2 in the other six *Plasmodium* species which include *P.vivax, P.knowlesi, P.ovale, P.berghei, P.yoelli* and *P.chabaudi* and the homologs from other protozoan parasites namely: *Leishmania, Toxoplasma gondii, Trypanosoma brucei* and *cruzi* and *Entamoeba histolytica*; as well as their respective host and vector organisms. The retrieved sequences were compared using multiple sequence alignment (MSA) and phylogenetic tree calculations to establish the functional residues as well as the evolutionary relationships respectively, which could be vital in antiprotozoan drug design.

2.1 INTRODUCTION

Cysteine proteases are virtually ubiquitous in all organisms and their major role in living organisms is to degrade peptides and proteins. There are also known to be virulence factors in parasites hence they have been implicated in the development and progression of diseases especially protozoan diseases which are the epicentre of this study (Grzonka et al. 2001). Like most proteins, cysteine proteases can be retrieved from open source databases such as NCBI (http://www.ncbi.nlm.nih.gov/). The retrieved sequences can be aligned to provide a platform for the identification and characterization of proteins, phylogeny, sequence similarities and differences inferences (Edgar 2004). The proteins that exhibit similarities are referred to as homologs which can be classified into three groups; orthologs which are evolutionary related proteins but found in different species, paralogs which are a duplication of proteins found within the same organism and xenologs are also related proteins in different organisms which occur through horizontal gene transfer (Tatusov et al. 1997).

2.1.1 SEQUENCE ALIGNMENT APPROACHES AND ALGORITHMS

Sequence alignment involves the determination of insertions and deletions and their locations within the sequences, which would have arisen during sequences divergence from a common genetic ancestor (Xiong 2006). Consequently alignment is the starting point of virtually all comparative study analyses of molecular data such as this one which involves the identification of homologous sites and phylogenetic tree construction (Rose et al. 2011).

A number of sequences can be aligned simultaneously through multiple sequence alignment (MSA) which could be either global and/or local alignment. Global alignment is defined as the alignment across the entire length of the compared sequences that uses the Needleman-Wunsch algorithm (Needleman & Wunsch 1970). On the contrary local alignment aligns the similar local regions between the studied sequences using the Smith-Waterman algorithm. Local alignment method is ideal for evolutionary distant sequences that have similar domains (Smith & Waterman 1981). Both global and local alignments are effectively achieved through an algorithm known as dynamic programing which is a technique that determines optimal alignments by revealing similarities between genes (Xiong 2006).

Dynamic programing achieves optimal alignments through the use of scoring matrices. There are mainly two types of scoring matrices namely Point Accepted Mutations (PAM) and Blocks of Amino Acid Substitution Matrix (BLOSUM) are the most common matrices (Henikoff & Henikoff 1992). In this study the BLOSUM matrices were employed.

BLOSUM is derived from the direct observation of possible residue substitutions in multiple sequence alignments. This scoring matrix was constructed from over 2000 conserved amino acid patterns which are called blocks (un-gapped alignments less than sixty amino acids in length). The frequency of the amino acid substitutions within the un-gapped blocks is calculated into a substitution matrix. The BLOSUM matrix numbers represent the percentage identity values of sequences used to construct the matrix, for example BLOSUM62 shows that the sequences used to construct the matrix share a sequence identity of 62%. For each residue that is aligned to nothing a gap is introduced and is penalized which lowers the alignment score. Optimal alignments would have the highest score and minimal penalties (Henikoff & Henikoff 1992).

2.1.2 DATABASE SIMILARITY SEARCH AND SEQUENCE RETRIEVAL TOOLS

Sequences are retrieved from the databases using search tools such as Basic Local Alignment Search Tool (BLAST) which searches for the query sequence homologs in the database. BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) uses heuristic database searching algorithms to align the query sequence with all sequences within the database (Ye et al. 2006). The sequences within the sequence similarity threshold show that the similarity did not occur by chance and the results are presented as a list sorted according to statistical significance and the E-value. The E-value is the parameter that depicts the probability that the query sequence alignment with the database sequence did not match by random chance hence the higher the value the less significant the alignment (Altschul 1998).

BLAST has a lot of variants that include BLASTN, BLASTP, BLASTX TBLASTN, and TBLASTX. In this study BLASTP is used, it uses protein sequence queries to search against a database of protein sequences. To increase its sensitivity BLAST has a variant PSI-BLAST that uses profiles to find remote sequence homologs by searching against sequence databases in an automated manner (Xiong 2006).

HHpred (http://toolkit.tuebingen.mpg.de/hhpred) is another speedy and sensitive search tool that reveals homologous proteins across a variety of protein databases by identifying remote homologs and predicting protein structures. The server uses Hidden Markov Models (HMM-HMM comparison) (Söding et al. 2005) whereby the initial query sequence is used to identify and align identified homologs using PSI-BLAST. The profiles generated by PSI-BLAST are then used as input into PSIPRED to give results that include comprehensive secondary structure prediction (Jones 1999). This alignment is used to construct HMM profiles that provide the probability of an insert, deletion and match for each HMM profile column in the alignment (Eddy 1998). HHsearch is then employed to query the consensus HMM profile against a database of choice through HMM-HMM comparison (Söding et al. 2005). This search method is more efficient and sensitive compared to the position specific scoring matrices. HHpred output gives alignment statistics together with the alignment. The statistics include the percentage identity, E-value, similarity and the probability of the homology to the query.

2.1.3 MULTIPLE SEQUENCE ALIGNMENTS

Related sequences can be identified through database similarity search as described earlier in this chapter. The retrieved homologs can be compared to get a clearer view of sequence conservation through multiple sequence alignment (MSA) which reveals more conserved motifs, sequence patterns and evolutionary relationships (Taylor 1998). Heuristic approaches have been developed to carry out MSA, such as iterative alignment, progressive alignment and block-based alignment (Thompson et al. 1999). Progressive MSA is engineered in a stepwise cluster which speeds up the alignment. Initially the query sequence is pairwise aligned with each of the sequences using the Needleman–Wunsch global alignment method. The sequence similarity scores or identity scores depending on the substitution matrix used, from the pairwise alignment are recorded. The scores are then used to calculate the evolutionary distances between the sequences to build distance matrix. A phylogenetic tree based on the distance matrix is generated using the neighbour-joining method which indicates the evolutionary closeness of the sequences. The resulting tree is used as a guide for the realignment of the sequences. Using the guide tree, the two closest sequences are re-aligned and converted into a consensus sequence. The most similar sequence is aligned with the consensus sequence using dynamic programing. The more distantly related sequences are added and the resulting consensus sequence is used for the next round of alignment. This process is repeated until all the sequences are added (Xiong 2006).

There are a number of currently available MSA programs whose basis is progressive alignment. CLUSTAL is one of such programs. Unfortunately progressive alignment uses global alignment in some cases and may therefore fail to identify conserved domains and motifs in distantly related sequences with varying lengths. For such sequences the consistency-based method is ideal. T-coffee is a program that uses the consistency based method which uses both the local and global pairwise alignments. T-coffee does not generate its own sequences; it uses CLUSTAL for local alignments and LALIGN for global alignments (Pei & Grishin 2007). The program assesses the alignments and then weights them on the basis of sequence identity. Once this has been done the program proceeds to extend the library of local and global alignments (Nuin et al. 2006). The program then constructs a guide tree based on the position specific scoring matrix built from the library. From this point the program uses progressive alignment to generate an MSA (Di Tommaso et al. 2011).

Another example of a consistency based MSA alignment program is Profile multiple alignment with local structure (PROMALS3D) which provides both sequence and structural alignment (Pei et al. 2008). The program starts off by clustering the sequences followed by performing a pairwise alignment of similar sequences using the scoring matrix BLOSUM62. From each cluster a representative sequence is selected which is referred to as the 'target sequence'. The target sequences are used as input into PSI-BLAST to search for additional homologs from the database Uniprot non-redundant reference databases facts (UNIREF90) and secondary structure prediction from Protein structure prediction server (PSIPRED) (Altschul 1998). A consistency scoring function is generated from the residue profiles obtained from PSI-BLAST and PSIPRED; the scoring function is used to progressively align the representative target sequences. The complete MSA is obtained by adding the pre-aligned clusters to the representative sequence alignment (Pei & Grishin 2007).

Consistency based approaches can be sped up by using the fast Fourier transform approximation; one such program is MAFFT which uses progressive alignment to construct a preliminary alignment (Katoh et al. 2002). MAFFT initially calculates the distance between every pair of input sequences and the distances are used to construct a guide tree using the unweighted pair group method with arithmetic mean (UPGMA) approach. The input sequences are then progressively aligned according to the guide tree branching order. The initial distance matrix is less accurate hence a new guide tree is constructed based on the new distance matrix generated from the initial guide tree. Progressive alignment is re-done using the new guide tree. The new alignment is enhanced by the iterative refinement method whereby the progressive alignment process is iterated until the scoring alignment obtained is worse than the previous one (Katoh et al. 2002).

2.1.4 PHYLOGENETIC ANALYSIS

Multiple sequence alignments provide a platform for revealing sequence identities and similarities as well as reveal conservation among aligned residues; however they are not equipped to provide accurate phylogenetic inferences and protein groupings deductions (Kosiol et al. 2008). To achieve these, phylogenetic trees are used. Phylogenetic trees show the evolutionary relationships among protein sequences however the resulting tree depends on the algorithm and evolutionary assumptions used (Xiong 2006).

Phylogenetic analysis allows protein sequence history to be evaluated. The branching order of the trees illustrates the sequence evolutionary divergence. Phylogenetic tree construction requires a multiple sequence alignment input therefore it is mandatory that the MSA be accurate. The MSA is used to construct the phylogenetic models which are used as evolutionary distance approximates (Xiong 2006).

There are currently a variety of programs that are available for phylogenetic tree calculations. One such program is MEGA5, it takes DNA and protein sequences as input and has an inbuilt alignment tool, the program can also accept already aligned sequences. MEGA5 evaluates the evolutionary distances between the sequences and then generates a phylogenetic tree based on the estimated distances. The evolutionary distances are essential in gauging the sequence divergence and MEGA5 uses statistical models to do this. The models produce distance estimations and convert them to a distance matrix which is then used to construct a phylogenetic tree (Tamura et al. 2011).

The phylogenetic trees can be calculated either by distanced based methods such as the neighbour joining method or character based methods such as maximum likelihood. In this study the maximum likelihood method is used. Maximum likelihood is a character based approach that selects the best tree based on the probabilistic models. The tree with the highest probability of reproducing the observed data is selected as the optimal tree. This method conducts its analysis on the entire alignment unlike the informative sites only (Kosiol et al. 2008). The input is a set of observed sequences as well as a substitution model. The program then calculates the probability of a residue to evolve to a different residue after time, the probability values are determined by the substitution model, these are termed likelihood scores. The likelihood for the topology is the sum of likelihood scores for each branch in the tree. The tree with the highest likelihood scores is considered the optimal tree (Xiong 2006).

The output trees need to be validated for reliability to infer phylogeny using resampling strategies such as bootstrapping and jackknifing. The bootstrap method repeatedly alters the input data and creates tree according to the replicates required (Xiong 2006). The generated trees are compared to the original tree to test for robustness and they are visualized using the MEGA5 tree explorer to clearly show the evolutionary diversity or similarity within a family of species that are being studied (Tamura et al. 2011).

2.2 METHODOLOGY

Sixty-two sequences of clan CA cysteine proteases from protozoan parasites, their hosts and vectors were retrieved, aligned and their phylogenetic divergences or similarities analysed. The FP-2 *Plasmodium* sequence was used as the query sequence. The tools and procedures employed are described below.

2.2.1 DATA RETRIEVAL

The Falcipain-2A (FP-2A) sequence was retrieved from the NCBI databases. Using FP-2A as a query sequence orthologs from the *Leishmania, Toxoplasma, Trypanosoma* and *Entamoeba* species as well as their respective host and vector sequences were retrieved from the NCBI database. The sequences of the homolog structures were identified using HHpred. The FP-2 sequence was used as the template in HHpred to identify the homolog structures and the default parameters were used. The homolog structures with complete structures, a low E-value and 100 % probability, which were sourced from protozoan parasites that are being studied in this project, their respective hosts and vectors were selected and then the sequences were retrieved from the NCBI database A total of 62 FP-2 orthologs were retrieved from the NCBI BLASTP tool through BLAST with default alignment parameters of BLOSUM-62 scoring matrix, a gap existence and extension cost of 11 and 1 respectively and a word-size of 3. The BLAST method occurred in two stages. In the first stage FP-2 was used as the query and it returned the orthologs in other organisms. In the second stage (reverse BLAST) the ortholog sequence was used as the query and ideally it should return FP-2 as the first ortholog hit, however this was not the case in all the orthologs, some returned FP-3 as the first hit. This was probably due to the fact that the orthologs were more closely related to FP-3 than FP-2 to confirm this sequence identities are indicated below (Appendix A-III).

The sequence analysis included 20 parasite species, 16 and 26 host and vector sequences respectively. Only the clan CA proteases were retrieved in this study.

2.2.2 MULTIPLE SEQUENCE ALIGNMENT

The 62 retrieved sequences were then aligned to identify the regions of similarity and divergence among the species and therefore infer homology and evolutionary relationships (Edgar 2004).

The alignment programs MAFFT (Katoh et al. 2002), TCOFFEE (Di Tommaso et al. 2011), EXPRESSO (Armougom et al. 2006) and PROMALS3D (Pei et al. 2008), were used to perform MSA. The homolog structures from FP-2 (2OUL), FP-3 (3BWK), T.gondii (3F75), H.sapiens-L and K (2XU3 and 1BY8 respectively), S.scrofa-H (8PCH) and M.musculus (4BS6) were used as the templates in the PROMALS3D alignment. These programs were accessed from the web server and their default parameters were used. A fasta file containing all the retrieved sequences was uploaded into each of the above mentioned programs and the MSA was run automatically. Since the region of interest in this study was the mature domain only the mature domain was aligned. The alignments were viewed on Jalview alignment viewer and the MAFFT alignment was observed as the best quality alignment.

2.2.3 PHYLOGENETIC TREE CALCULATIONS

The MAFFT alignment which had the most accurate alignment was used as the dataset for tree generation. The calculations were performed on MEGAv5 to generate maximum likelihood phylogenetic trees, which showed the evolutionary relatedness of the cysteine proteases across all the studied organisms. The maximum likelihood approach calculates the probability that the input data in this case the MAFFT alignment will be observed under a particular phylogenetic tree and the selected substitution model. The substitution models provided a set of parameters that describe the evolutionary process. In this study five substitution models were used, WAG+G, WAG+G+I, rtREV+G+I, JTT+G and JTT+G+I since these models had the top BIC (Bayesian Information Criterion) scores. The other parameters included in this study were the site coverage threshold which was set at complete deletion and a bootstrap value of 1000 for the purpose of identifying the confidence levels of the inner nodes of the topology. The other parameters not discussed were set at default.

2.3 RESULTS AND DISCUSSION

In this section the data retrieval, multiple sequence alignment and phylogenetic tree calculation results are reported. The residue numbering is in reference to FP-2. The sequences are referred to by their common names (Appendix A-I).

2.3.1 DATA RETRIEVAL

FP-2 orthologs were retrieved by reverse BLAST from the NCBI database and the sequences of the homolog structures were identified (Appendix A-I).

All the retrieved sequences were from the cathepsin L-like sub-class and they had low E-values and sequences identities were greater than 30% suggesting that the retrieved sequences were significant hits. The sequences had two functional domains namely; the prodomain which is a propeptide that inhibits the activity of the cysteine protease and the catalytic domain. For the purpose of this study only the catalytic domain was analysed. The regions marking the prodomain and the catalytic domain are illustrated in the appendix (A-II). In addition to the identification of the functional domains' regions the first and second hits returned by the NCBI database were attained as well as their respective sequence identities as illustrated in the appendix (A-III).

In some indicated sequences FP-3 was the first hit showing that the sequences were more closely related to FP-3 than FP-2. Further studies of the sequences relatedness were determined by phylogenetic tree calculations.

2.3.2 MULTIPLE SEQUENCE ALIGNMENT

The retrieved sequences were aligned using MAFFT, PROMALS3D, TCOFFEE and EXPRESSO alignment algorithms. The alignments from all the programs are shown in the appendix (A-IV-VII). The alignment generated by the MAFFT algorithm was chosen as the best alignment because it aligned the conserved regions such as catalytic and sub-site residues as well as the insert regions well relative to the other alignment programs that were used and had minimal gaps. The prodomain and the catalytic domain regions were separated and the results reported are for the mature domain.

In analysing the mature domain the catalytic residues were identified. The full alignment can be found in the appendix section (A-IV). The best alignment is shown below (Figure 2.1) but only the regions that are known to be conserved are shown to justify the choice of best alignment.

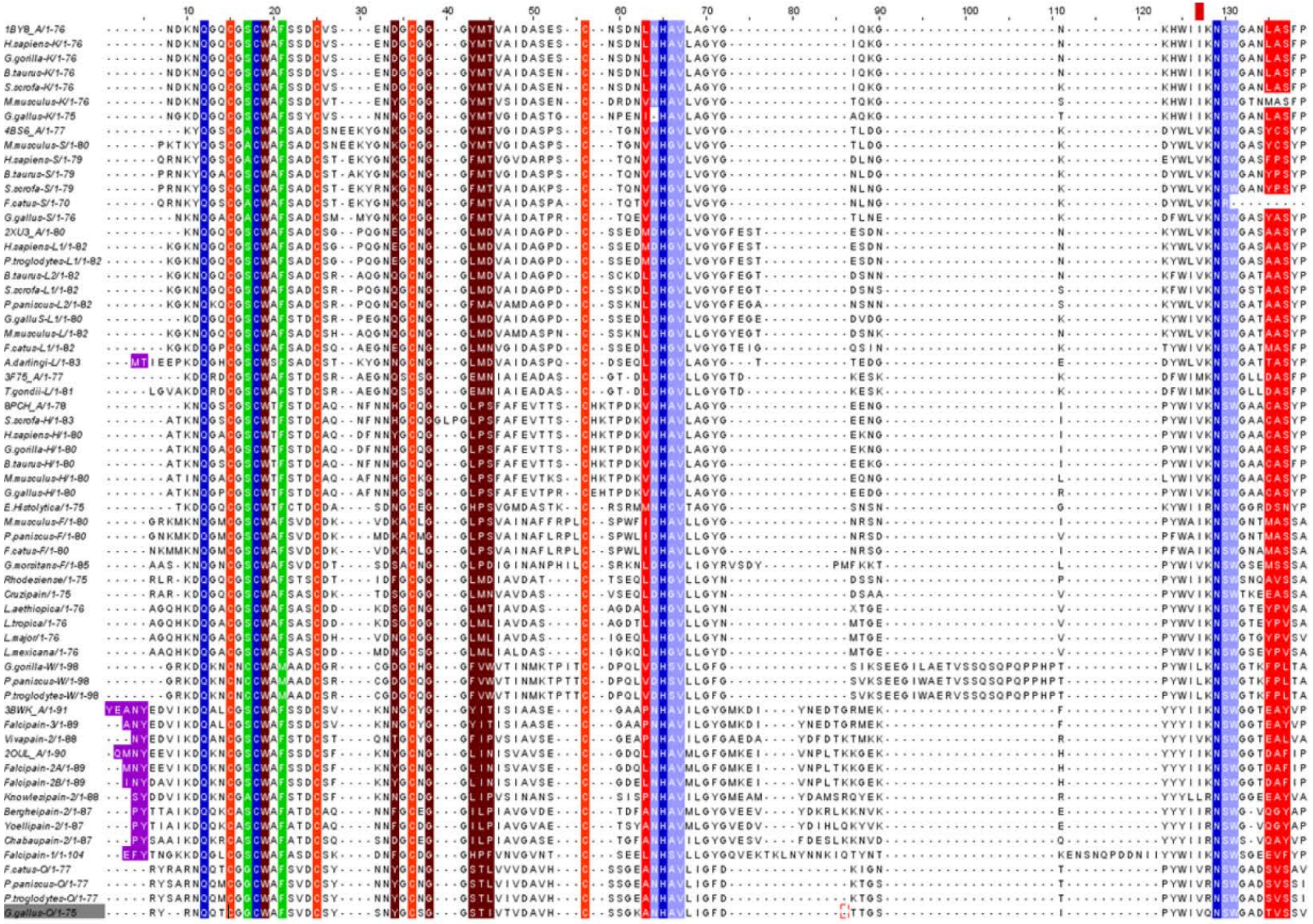


Figure 2.1 MAFFT alignment of the cysteine proteases' catalytic domain as viewed on Jalview. The purple residues indicate the Plasmodium species insert that forms part of the nose-like protrusion, the royal blue residues indicate the catalytic residues while the orange residues indicate the cysteine residues involved in disulphide bridge formation. The green residues illustrate the sub-site 1 residues, the brown residues show the sub-site 2 residues, the light blue residues indicate the sub-site1' residues and the red residues show the sub-site 3 residues.

The *Plasmodium* species have an N-terminal insert which forms the arm-like protrusion which is believed to be involved in the folding of the prodomain. The *Plasmodium* species also have a C-terminal insert which forms the nose-like protrusion and is thought to be the haemoglobin binding motif. The sequences also showed conservation of six cysteine residues that form disulphide bridges. The *Plasmodium* sequences have an extra pair of cysteine residues that stabilize the structure of the mature domain during its active conformation by forming a loop in the S2 and S1' sub-sites in FP-2.

2.3.2.1 SUB-SITES ANALYSIS

Residues from each of the sub-sites namely; 1, 2, 3 and 1' were analysed (Figure 2.2).

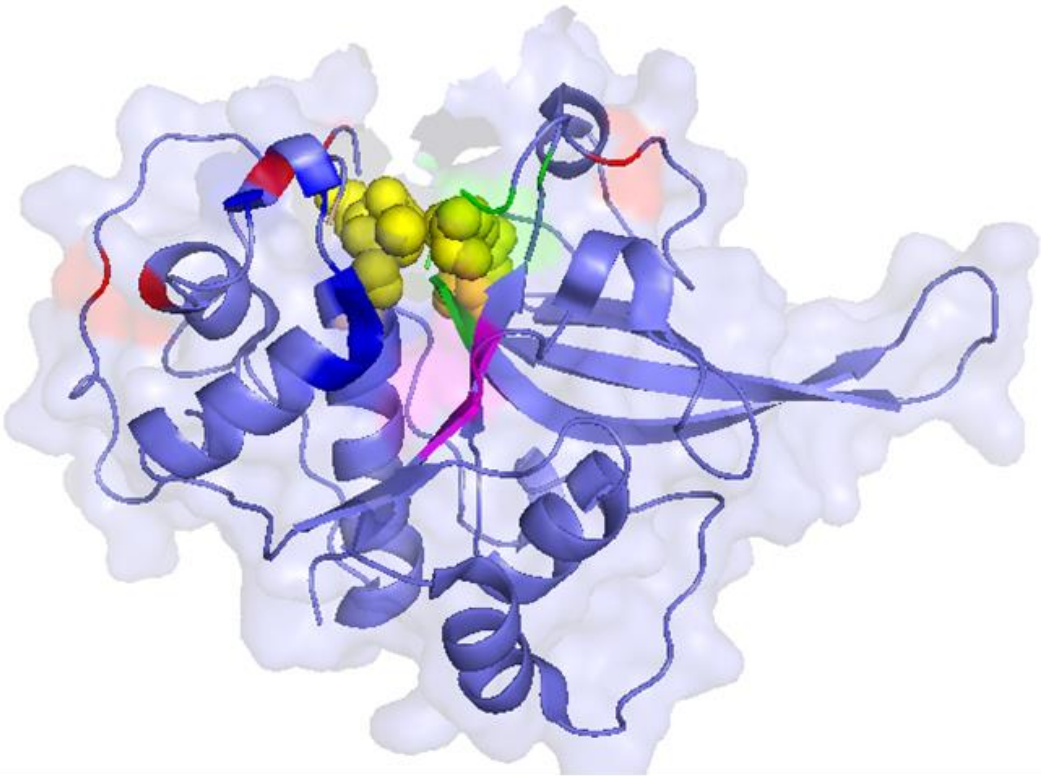


Figure 2.2 Sub-sites map onto the FP-2 structure (2OUL). The catalytic residues are shown in yellow while sub-sites 1, 1', 2 and 3 are shown in wheat, green, magenta and blue respectively. The rest of the protein is slate blue. The conserved cysteine residues are shown in red.

The sub-site regions shown in Figure 2.2 were chosen by selecting the residues known to constitute these regions as indicated in Appendix A-VIII.

Table 2.1 A summary of the sub-site residues of the parasites cysteine proteases' catalytic region. The residue numbering is according *Plasmodium* FP-2.

Species name	Subsite-1			Subsite-1'					Subsite -2							Subsite-3			
	279	282	284	288	395	416	418	419	448	449	286	321	325	327	328	329	392	415	477	478	479
FP-3 (3BWK)	Q	C	S	F	A	N	A	V	S	W	W	N	G	Y	I	T	S	P	N	A	F
FP-2 (2OUL)	Q	C	S	F	V	N	A	V	S	W	W	Y	G	L	I	N	S	L	N	A	F
T.gondii –L (3F75)	Q	C	S	F	A	D	G	V	S	W	W	Q	G	E	M	N	A	L	D	A	F
T.gondii-L	Q	C	S	F	A	D	G	V	S	W	W	Q	G	E	M	A	V	V	D	A	F
L.aethiopica	Q	C	G	F	A	N	G	V	S	W	W	S	G	L	M	A	A	A	N	A	L
L.mexicana	Q	C	S	F	A	N	G	V	S	W	W	N	G	L	M	A	L	L	Q	A	F
L.major	Q	C	S	F	A	D	G	V	S	W	W	N	G	L	M	A	L	L	N	A	F
L.tropica	Q	C	S	F	V	N	A	V	S	W	W	S	G	L	M	A	L	L	R	A	F
T.brucei	Q	C	C	M	V	N	A	V	S	W	W	F	G	L	M	A	M	M	Y	A	F
T.cruzi	Q	C	S	F	A	N	A	V	S	W	W	S	G	L	M	A	L	L	Q	A	F
BP-2	Q	C	S	F	A	N	G	V	S	W	W	F	G	I	L	A	L	A	Y	A	F
YP-2	Q	C	S	F	A	N	G	V	S	W	W	F	G	I	L	A	A	A	Y	A	F
CP-2	Q	C	S	F	A	D	G	V	S	W	W	D	G	I	L	A	P	A	L	A	F
VP-2	Q	C	A	F	V	N	A	V	S	W	W	T	G	F	I	S	A	A	Y	A	F
KP-2	Q	C	S	F	A	N	A	V	S	W	W	N	G	L	I	S	P	P	R	A	F
FP-2	Q	C	S	F	V	N	A	V	S	W	W	Y	G	L	I	S	L	L	N	A	F
FP-2B	Q	C	S	F	V	N	A	V	S	W	W	Y	G	L	I	S	L	L	N	A	F
FP-3	Q	C	S	F	V	N	S	V	S	W	W	N	G	Y	I	S	P	P	N	A	F
FP-1	Q	C	S	F	A	D	G	V	S	W	W	F	G	H	P	N	L	L	N	A	F
E.histolytica	Q	C	S	F	A	N	G	V	S	W	W	N	G	H	P	G	V	V	Q	A	F

From the analysis as indicated in the appendix (A-VIII), the sub-site 1 and 1' residues are conserved while those in sub-site 2 and 3 are variable. The sub-site 1 residues are conserved although the hydrophilic serine residues in all the other sequences are substituted by the

hydrophobic residue alanine in the cathepsin-S proteases and glutamine in cathepsin-O proteases. In the cathepsin-W proteases the serine residue is substituted by the cysteine residue also known as the 'orphan' residue which is a signature of the cathepsin-Ws (Brinkworth et al. 2000). Also within the sub-site the hydrophobic residues tend to alternate between methionine in cathepsin-W proteases and phenylalanine in the rest of the proteases. The glutamine residue is thoroughly conserved due to its catalytic role of stabilizing the oxyanion hole during catalysis.

Although sub-site 2 is highly substituted, the tryptophan and glycine residues are conserved across all species since they are residues responsible for forming hydrophobic interactions with the substrate as well hydrogen bonding respectively. In sub-site 3 alanine and phenylalanine are well conserved across all organisms with the exception of cathepsin-O and cathepsin-F proteases where phenylalanine is substituted by leucine and tyrosine respectively. Valine and tryptophan are well conserved in sub-site 1' to maintain the hydrophobic interactions with substrates. The hydrophilic asparagine residues are substituted by the electrically charged aspartic acid residues in cathepsin-L proteases.

The residue positions of the sub-sites were residue numbers 279, 282, 284, 288 and 279; 395, 416, 418, 419, 448 and 449; 286, 321, 325, 327, 328, 329 and 395 and 415 477 478 479 for sub-sites 1, 1', 2 and 3 respectively.

2.3.3 PHYLOGENETIC ANALYSIS

As previously discussed in the methodology section the cysteine proteases were selected from protozoan species as well as their respective host and vector organisms. Cysteine proteases are virtually ubiquitous and are known to have a conserved mechanism of action therefore substantial differences can be exploited for inhibitor selectivity. To assess the evolutionary relatedness between the proteases a phylogenetic analysis was employed (Figure 2.3).

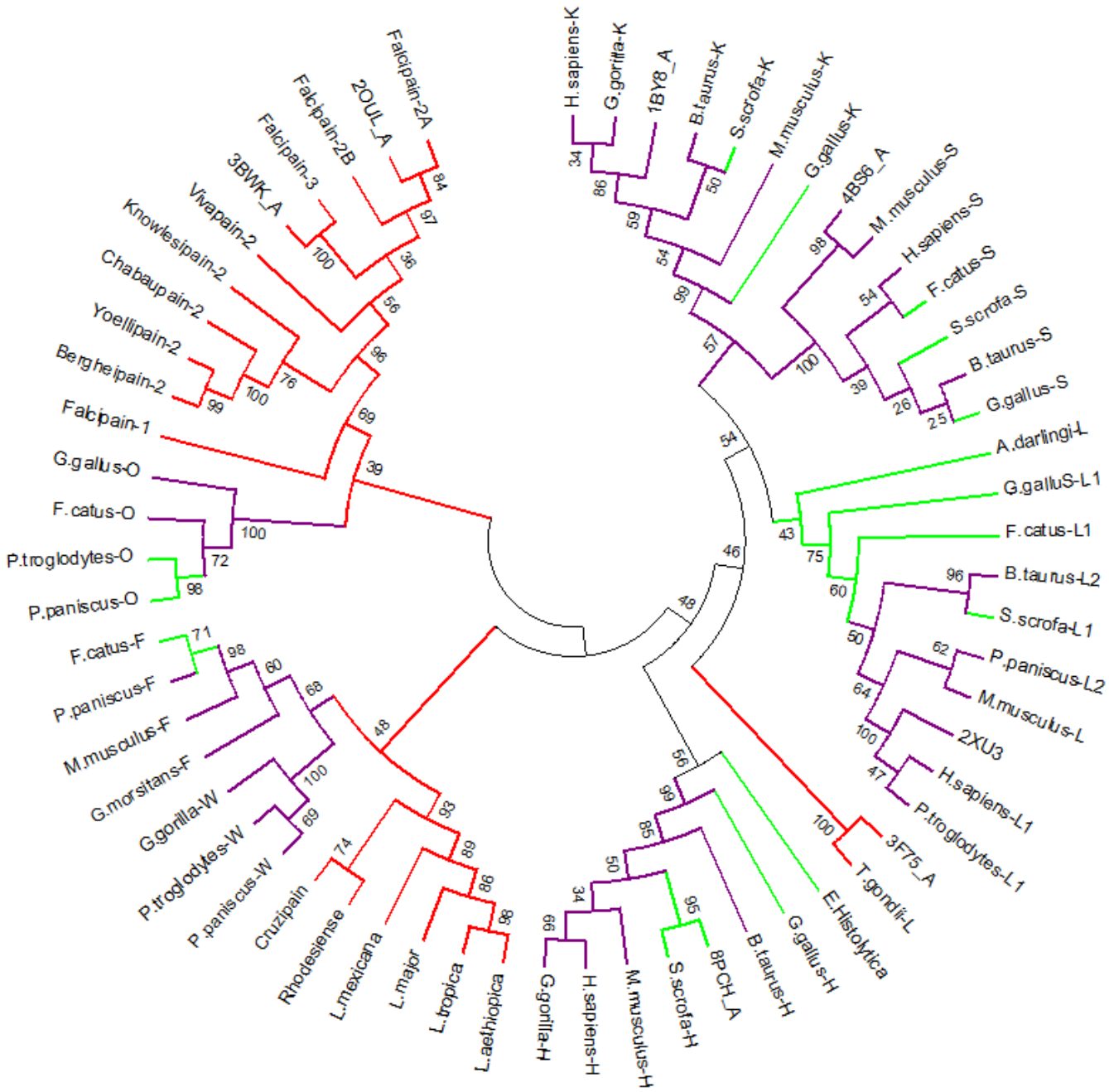


Figure 2.3 Phylogenetic tree of FP-2 and its orthologs based on the alignment generated by the MAFFT algorithm**.** The maximum likelihood tree was generated using MEGA5. The values indicate the bootstrap statistical analysis values at each inner node which specify percentage probability of its occurrence. The tree is based on the WAG+G statistical model. The red branch colours indicate the host and the vector species.

The phylogenetic tree showed that the host and vector species did not cluster according to the organisms they came from but rather the type of cathepsins they were, the K-cathepsins clustered

together and the same was observed for S, L, H, W, F and O-cathepsins. Interestingly the *T.brucei* and *T.cruzi* species clustered with the *Leishmania* proteases showing close relatedness between these species. As expected the *Plasmodium* species clustered together although FP-1 diverged from the rest of the *Plasmodium* species showing that it is distantly related from the other *Plasmodium* species. This is also evident from the multiple sequence alignment as FP-1 has a longer insert relative to other *Plasmodium* species. The murine malarial *Plasmodium* species clustered together, separate from the rest of the *Plasmodium* species, this is interesting because the murine species are known infect primates in Central Africa unlike the other *Plasmodium* species which infect humans. Although KP-2 is known to be a human malarial protease it was observed to be closer related to the murine species instead of the human malarial species. FP-2A and -2B branched out together and were observed to be closely related to FP-3 and VP-2.

The clustering patterns observed in Figure 2.3 may have resulted from specific motifs within related proteases that may lead to them possessing similar physicochemical proteins. In this study these distinct features were further investigated through motif analysis and physicochemical properties analysis as discussed in Chapter 3.

CHAPTER THREE

3. PHYSICOCHEMICAL PROPERTIES AND MOTIF ANALYSIS

This study aims to conduct a comparative analysis of parasitic protozoan cysteine proteases as well as their respective hosts and vectors with special interest in *Plasmodium* cysteine proteases. This chapter will therefore focus on the physicochemical properties and motif analysis of FP-2; a papain family cysteine protease from *P.falciparum*, its orthologs from other protozoan parasites and their respective host and vector homologs. The physicochemical analysis is achieved via BioPython scripting and statistically analysed using box plots and the Kruskal-Wallis test. The motif analysis is performed using MEME. This is done to discover motifs and amino acid residues that are distinct in protozoan parasites and potentially vital in the protein function. Cysteine proteases are virtually ubiquitous in all living organisms therefore a comparative study is imperative in order to unearth distinct differences among the protozoan, host and vector cysteine proteases as a step towards designing selective and specific cysteine protease inhibitors.

3.1 INTRODUCTION

Protein polypeptide chains tend to fold into secondary structural elements. Within the structural elements there are observed patterns of amino acid residues whose abstraction forms a protein motif (Bailey et al. 2006). Protein motifs can be organized into four classes, sequence motifs, sequence-structure motifs, and structure motifs and structure-sequence motifs (Bork & Koonin 1996).

Sequence motifs are amino acid sequence patterns whose topological order can be indirectly inferred. Whereas sequence-structure motifs are primary structure patterns with a secondary structure identifier attached to at least one of the residues. From these motifs the structure of a protein can be interpreted. On the contrary structure motifs are not associated with the sequence motifs, they are the common parts observed in different topologies in known structures. The

structural motifs are used in identifying and classifying proteins. Structure-sequence motifs combine the primary and tertiary structure information. There is no need for a direct implication between the sequences. These motifs can be described in either of two ways, sequence logos or position weight matrices (PWM) (Bailey et al. 2006).

There are available motif discovery algorithms that output their results in each of the two methods. Some motif discovery algorithms display a list of occurrences of the motifs within the sequence instead of the descriptive output. The set of occurrences can be changed to PWMs or regular expressions (Bailey 2008).

Regular expressions describe the pattern of amino acid residues that match the motif. Whereas the position weight matrices define the probability of each residue letter occurring at that particular position (van Helden et al. 1998). Both of these are used by motif discovery algorithms due to their advantages. The regular expressions have an advantage of easy visualization and the statistical significance evaluation can be easily defined on regular expressions. Whereas PWMs have the advantage of a more elaborate description of a motif as it provides the specific position of the motif rather than just the fact that there is a match or not. However unlike regular expressions PWMs are difficult for algorithms to compute since they are computationally expensive (Staden 1989). MEME is an example of a PWM based algorithm. It works by looking for repeated, ungapped sequence patterns within the input sequences (Bailey et al. 2006).

3.1.2 MEME: MOTIF DISCOVERY ALGORITHM

In functionally related proteins, there are common structural characteristics that exist. Within the protein sequences the motifs can represent enzyme active sites among other things (Bork & Koonin 1996). The conserved functional parts can be used to predict the function of unknown proteins and they are known as motifs which can be identified and characterized using various algorithms. However the structures are not always available so to cater for this the local conserved segments can be identified from the primary sequences (Hu *et al*. 2005). In this study the motifs were discovered from the amino acid sequences and identified using MEME.

MEME (Multiple Expectation Maximization Estimation) allows the discovery of motifs in protein sequences and the algorithm is based expectation maximization (EM) technique. The user inputs a set of sequences in fasta format that are believed to possess common motifs. By

default MEME finds up to three motifs, which may be found in some or all of the input sequences. The default motif widths are between 6 and 50 although these default settings maybe changed by the user (Bailey et al. 2006). The width and number of occurrences for each motif is set up in such a way that the probability of finding an equally well conserved motif within random sequences is minimized (E-value) (Bailey 2008).

For each motif with a width w, MEME breaks down the input sequences into w-mers. MEME then uses the model and background components to conclude whether the w-mer is a motif or a background sequence. An increase in number of sequences and sequence length results in an increase in the search space. MEME has an HTML, XML and TXT output (Bailey et al. 2006).

The MEME HTML output shows the discovered motifs as blocks on their relative positions in each of the input sequences. The motifs appear as local multiple alignments of the input sequences. HTML output is generated from the XML output. XML is designed for machine reading as it lacks the visualizations found in the HTML output. The original MEME output is the plain text document output and like the HTML output it is also self-explanatory (Bailey et al. 2006).

Like most motif discovery algorithms MEME follows four basic steps for motif discovery. In the first step the input sequences believed to possess similar motifs are assembled, this is followed by the second step in which the low-complexity regions and known repeating elements in the sequences are masked. This third step involves running the algorithm using the parameters set by the user. In the fourth step the discovered motifs are evaluated, and the motifs that are chance artifacts and do not represent any structurally or functional relevance are removed. During this step the discovered motifs are evaluated on whether they represent a known motif or if they are conserved in orthologous sequences. If they are neither then they are weeded out (Bailey 2008).

3.1.3 PROTEIN PHYSICOCHEMICAL PROPERTIES

Sequence and motif analysis does not take into account the protein physicochemical properties. The physicochemical properties have an effect on the protein function as well as its structure, and their analysis elucidates patterns among the group of proteins being analyzed that may not be obvious in sequence, phylogenetic and motif analysis (Cozzone 2002). In this study six

physicochemical properties are evaluated namely; amino acid composition, aromaticity, GRAVY, instability index, iso-electric point (pI) and molecular weight.

The protein aromaticity influences the stability of the protein through the interactions between the aromatic residues, histidine, tyrosine, tryptophan and phenylalanine (Tartaglia et al. 2004). These residues contain an aromatic ring in which π-electrons are free to cycle around the ring such that aromatic residues with 4.5Å and 7Å apart from each other can form aromatic interactions through the interaction of the circulating electrons (Lanzarotti et al. 2011). Aromaticity is calculated by dividing the total number of aromatic residues with the total number of residues within the sequence (Cozzone 2002).

Aromaticity is not the only property that can be used to determine protein stability; the Grand Average of Hydropathy (GRAVY) can be used to determine the hydrophilic and hydrophobic properties that also influence the protein stability. In GRAVY calculation the hydrophobic and hydrophilic properties of the residues' side chains are taken into account. The GRAVY index is calculated by adding the hydropathy values of the amino acid residues in the protein sequence and then dividing by the total number of residues within the sequence. The positive values show polar proteins while the negative values indicate the non-polar proteins therefore the lower the GRAVY value the more hydrophilic the protein is (Kyte & Doolittle 1982).

Protein stability determines how fast the protein will degrade when exposed to extreme temperatures and pH as well as natural proteolysis, this is called the half-life. The instability index defines the half-life of a protein in a test tube. An instability index above 40 indicates that the protein is unstable whereas a value less than 40 indicates protein stability (Idicula-Thomas & Balaji 2005). The pI can be used to further describe protein solubility. The pI is the pH value where the protein does not carry a net charge. The pI can also be used to define the pH at which the protein is at its minimal solubilty and can therefore be purified in wet laboratory assays such as ion exchange chromatography for further characterization and interaction with potential ligands (Lecaille et al. 2002).

In this study the observed physicochemical properties will be validated using hypothesis testing specifically the Kruskal-Wallis test to test whether the data has the same distribution.

3.1.4 STATISTICAL ANALYSIS: KRUSKAL-WALLIS TEST

The purpose of the Kruskal-Wallis test is to determine whether there is a significant difference between the three or more samples at an alpha level of 0.05 (Kruskal & Wallis 1952). The test basically answers the questions: is there a difference in the samples or the variation that is observed merely occurs by chance and represents the expected differences within a randomly sampled population; these are termed the hypotheses (Delaney & Vargha 2002). The null hypothesis (H_0) states that there are no differences among the samples while the alternative hypothesis (H_1) postulates that there are significant differences among the samples (Chan & Walmsley 1997). To reach the conclusion the final value of the H statistic called the p-value is used. A p-value less than the alpha level 0.05 indicates that the null hypothesis can be rejected and the samples have a significant difference while a p-value greater than the alpha level indicates a lack of sufficient evidence to reject the null hypothesis meaning that the variation observed among the samples occurred by chance therefore they are not different (Kruskal & Wallis 1952).

The Kruskal-Wallis has assumptions regarding its use, firstly it assumes that the samples are independent of each other, secondly the sample scores in one sample group are independent of the scores in the other groups, thirdly the sample scores within the same group are independent of each other and finally the variable under study has a continuous distribution (Chan & Walmsley 1997). Since the sample data in this study satisfied the assumptions the Kruskal-Wallis test was appropriate to use.

3.2 METHODOLOGY

A motif analysis of the protein sequences from FP-2 and its orthologs from the six *Plasmodium* species as well its homologs from the *E.histolytica, T.brucei, T.cruzi, L.major, L.aethiopica, L.mexicana, L.tropica, T.gondii*; their respective hosts and vectors was performed. The sequences were grouped in 3 groups namely; hosts, vectors and parasites and their physicochemical properties calculated. These analyses were performed on both the catalytic domain sequences and the full length sequences. The procedures and tools used are described in this section.

3.2.1 MOTIF ANALYSIS

To analyze the conserved motifs the MEME motif search (http://meme.sdsc.edu/meme/website/meme.html) was used on a set of 62 sequences of both the catalytic domain sequences and the full length sequences. The criterion for selecting the data set is described in the data retrieval section (section 2.2.1). MEME was set to identify up to 100 motifs of 6 and 50 amino acids with zero or one occurrence per sequence, at p-value of <0.0001. The observed motifs were visualized using a heat map which is a 2D graphical representation of motifs represented as colours; generated through Python scripting. The script determined in how many sequences the motifs identified by MEME occurred and then converted it into a percentage. It then used the MAST output to identify the statistically significant motifs using the p-value. The motifs with low p-values were considered significant and were used as input into the heat map while the motifs with high p-values were removed. The significant motifs were then plotted on heat map together with the input sequences whereby the most frequent motifs were shown in red while the least frequent were shown in blue.

3.2.2 PHYSICOCHEMICAL ANALYSIS

The physicochemical properties of the retrieved protein sequences were investigated using BioPython scripting. The properties that were investigated include GRAVY, instability index, aromaticity, isoelectric point, molecular weight and amino acid composition. The physicochemical properties investigation results were visualized using box plots calculated on R

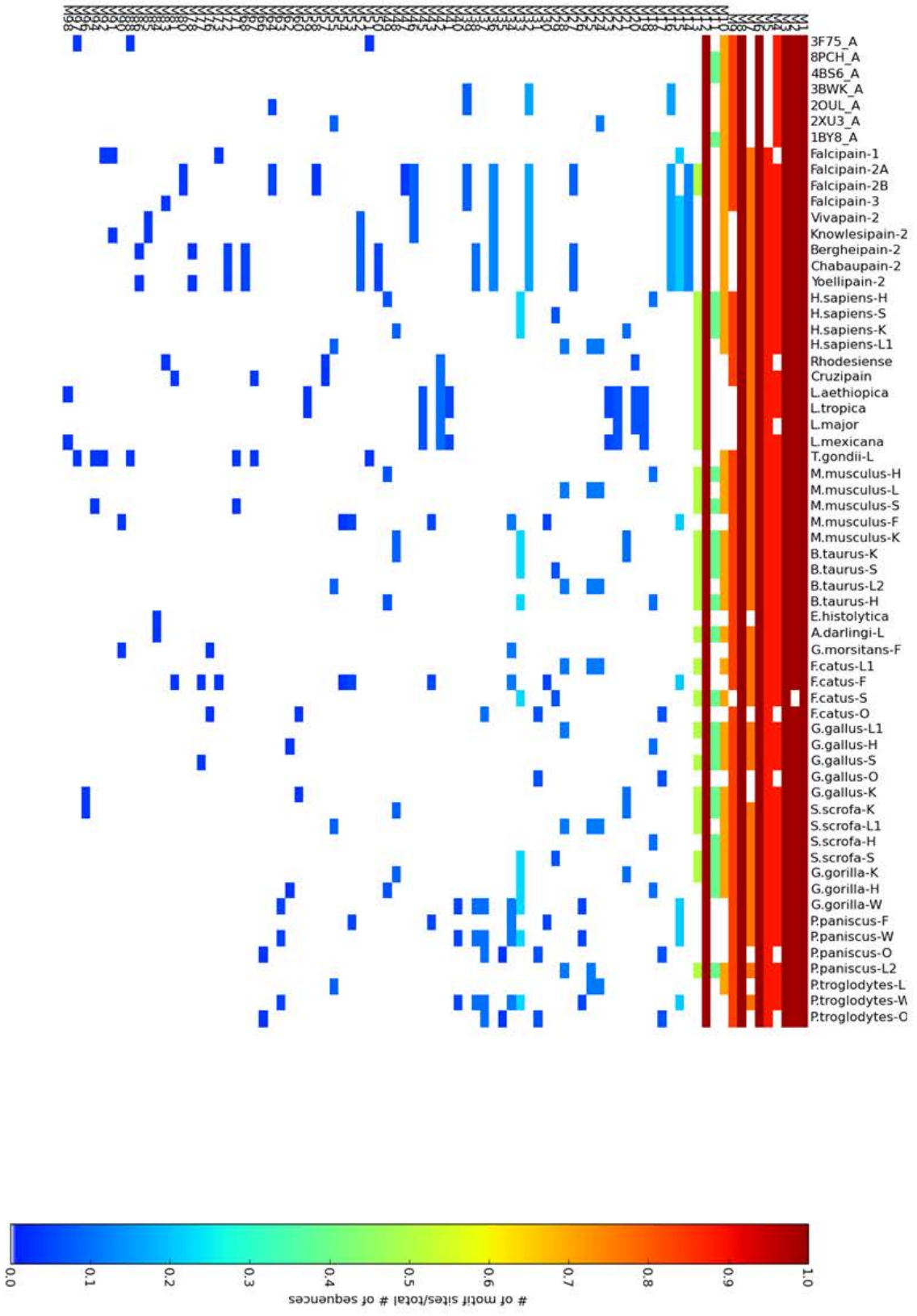
and 3D bar plots in the case of amino acid composition analysis. Both the catalytic domain sequences and full length sequences were evaluated.

3.3 RESULTS AND DISCUSSION

The results obtained from the physicochemical and motif analyses are reported in this section. The sequences are referred to in their common names.

3.3.1 MOTIF ANALYSIS

Motif analysis allows a powerful means of determining protein function (Hodgman, 1989). In this study the motifs from both the catalytic domain and the full length sequences were analyzed using MEME and their conservation was visualized using a heat map (Figure 3.1A and B) generated through Python scripting as the MEME output could not be used for informative visualization.


3F75_A
8PCH_A
4BS6_A
3BWK_A
2OUL_A
2XU3_A
1BY8_A
Falcipain-1
Falcipain-2A
Falcipain-2B
Falcipain-3
Vivapain-2
Knowlesipain-2
Bergheipain-2
Chabaupain-2
Yoellipain-2
H.sapiens-H
H.sapiens-S
H.sapiens-K
H.sapiens-L1
Rhodesiense
Cruzipain
L.aethiopica
L.tropica
L.major
L.mexicana
T.gondii-L
M.musculus-H
M.musculus-L
M.musculus-S
M.musculus-F
M.musculus-K
B.taurus-K
B.taurus-S
B.taurus-L2
B.taurus-H
E.histolytica
A.darlingi-L
G.morsitans-F
F.catus-L1
F.catus-F
F.catus-S
F.catus-O
G.gallus-L1
G.gallus-H
G.gallus-S
G.gallus-O
G.gallus-K
S.scrofa-K
S.scrofa-L1
S.scrofa-H
S.scrofa-S
G.gorilla-K
G.gorilla-H
G.gorilla-W
P.paniscus-F
P.paniscus-W
P.paniscus-O
P.paniscus-L2
P.troglodytes-L
P.troglodytes-W
P.troglodytes-O
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
# of motif sites/total # of sequences

A

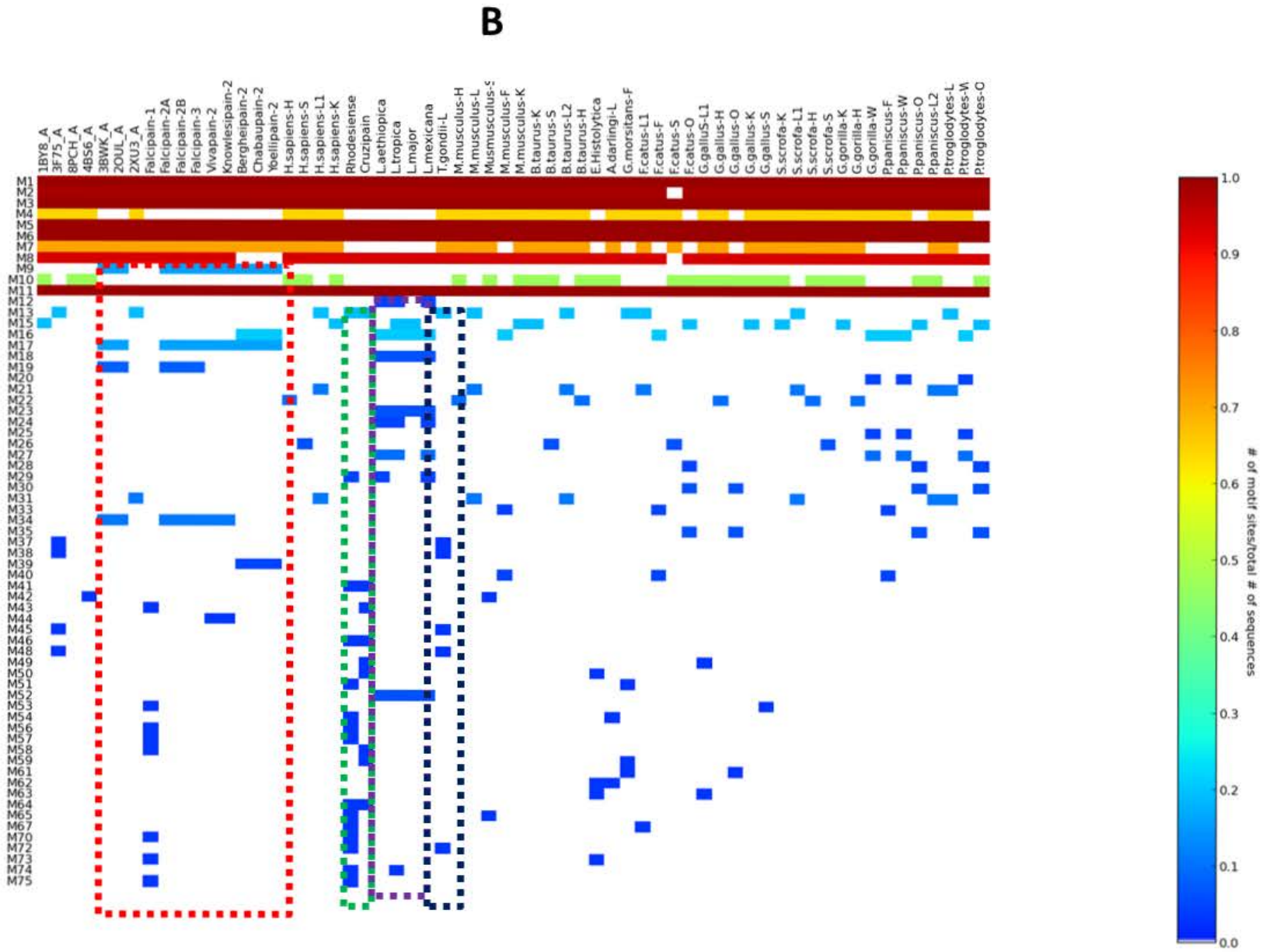


Figure 3.1 Heat map showing motif conservation of the studied cysteine proteases across the full length sequences (A) and the catalytic domain sequences (B) indicated on the x-axis while the motifs are shown on the y-axis. The colour code shows the level of conservation, the red motifs show thorough conservation while blue shows the least conserved motifs and the white background indicates absent motifs.

MEME generated a total of 98 and 75 motifs from the full sequences and the catalytic domain respectively (Fig. 3.1A and B). The motifs are represented according to conservation, 6 common conserved motifs were observed in both the full length sequences and the catalytic domain.

3.3.1.1 FULL LENGTH SEQUENCE MOTIF ANALYSIS

MEME generated a total of 98 motifs from the full length sequences (Figure 3.1A). Six thoroughly conserved motifs were discovered namely; motifs 1, 2, 3, 6, 8 and 12. There were other motifs that were absent from the sequences of the homolog structures suggesting that the motifs were conserved in the prodomain these were motifs 5 and 7. Motif 5 represents GNFD motif while motif 7 represents the ERFNIN motif; both are required in the folding and stabilization of the prodomain with salt bridges which are essential in the inhibition of the mature domain by the prodomain (Pandey et al. 2009). Motif 7 was observed to be absent from the O-cathepsins, they displayed the pattern S(A/G)R/EFSHN instead of the common ERFNIN motif however the motif serves the same purpose as the ERFNIN motif. This suggests that the O-cathepsins prodomain might fold in a unique way and the mature domain inhibition by the prodomain could occur differently although further studies would have to be conducted to confirm this. The motif analysis also revealed 2 unique motifs only found in the *Plasmodium* species with the exception of FP-1, the LMNNLESVN and the motifs suggesting that it could be vital in the function of the *Plasmodium* proteases. The motif is thought to form the prodomain insert region that facilitates the prodomain inhibition (Korde et al. 2008). As discussed earlier (section 2.3.3) FP-1 diverged from the other *Plasmodium* species suggesting that it is distantly related to them. The motif analysis confirmed this observation as it showed that FP-1 lacked motifs common in other *Plasmodium* proteases that were vital in their function.

3.3.1.2 CATALYTIC DOMAIN MOTIF ANALYSIS

The same 6 motifs that were conserved across all the sequences including the sequences of the homolog structures observed in the full length sequence (Figure 3.1A) were also observed in the catalytic motif analysis (Figure 3.1B). Since MEME identifies motifs regardless of their biological significance the conserved motifs were mapped onto the FP-2 structure and their position compared to that of the sub-sites essential for catalytic activity to determine their functional importance (Figure 3.2).

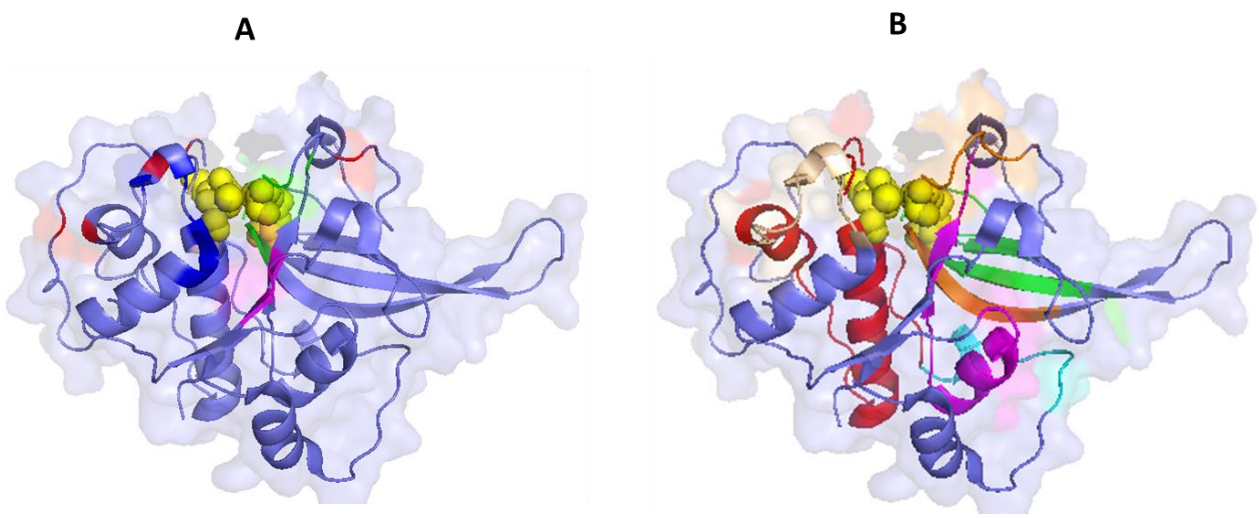
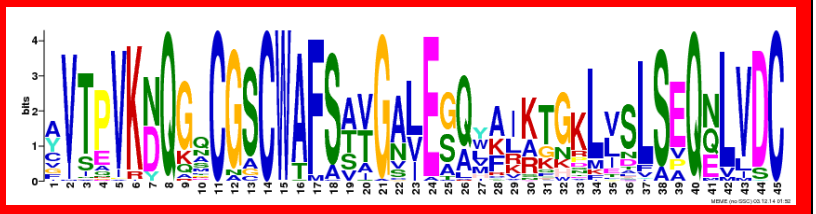
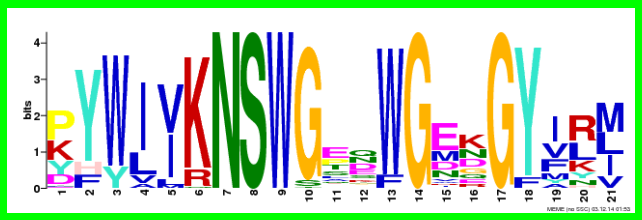


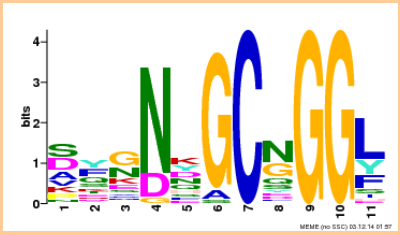
Figure 3.2 Sub-site residues and motif mapping onto the FP-2 structure (2OUL). A) The catalytic residues are shown in yellow spheres while sub-site 1 is shown in wheat, sub-site 1' is in green, sub-site 2 is shown in magenta and sub-site 3 is shown in blue. The rest of the protein is shown in slate blue. B) The conserved residues are shown in green, the *Plasmodium* unique motifs are shown in red and the motif that differs from the *Leishmania* species is shown in purple

When compared to the sub-site residues map, 5 of the 6 conserved motifs were found to be located in the sub-site regions suggesting that the motifs are conserved in order to maintain the activity of the proteases. Interestingly there was another conserved motif found in the region next to the FP-2 nose-like protrusion (shown in cyan in Figure 3.2B and motif 6 in Table 3.1). The motif is known to play a vital role in the folding of the protease by stabilizing the initial interactions of the folding process (Wang et al. 2006). However, despite the conservation of this motif, in *Plasmodium* species this motif is not sufficient to facilitate the folding of the protease hence the presence of the nose-like protrusion only found in *Plasmodium* species (Pandey et al. 2009).

The positions and sequence logos of the conserved motifs in both the catalytic domain and the full length sequences were summarized and illustrated (Table 3.1).

Table 3.1 A summary of the conserved motifs in both the catalytic domain and the full length sequences of the cysteine proteases according to the MEME HTML output. The background shading indicates the colour of the motif in the mapping (Figure 3.2B). The motif positions on the sequence indicate the range of positions across all the sequences.

Motif number	Sequence position	Sequence Logos
1	12 -70	
2	180 - 230	
3	118 – 155	
5 (6 in full length sequence analysis)	154 - 182	

6 (8 in full length sequence analysis)	1 – 25	
11 (12 in full length sequence analysis)	60 - 83	

3.3.1.3 UNIQUE CATALYTIC DOMAIN MOTIFS

The motif analysis of the catalytic domain also revealed motifs that were unique to certain species. The *Plasmodium* species unique motifs were observed as shown in a red box in Figure 3.2B, however within the *Plasmodium* species it was noted that FP-1 did not share the common motifs as other *Plasmodium* proteases and it had its own motifs and this observation is explained by the clustering of the *Plasmodium* proteases in the phylogenetic tree (Chapter 2, Figure 2.5) in which FP-1 diverges from the rest of the *Plasmodium* species. The conserved motif 14 in the *Plasmodium* species forms the nose-like protrusion which is necessary for correct folding of the protease. This motif was absent in FP-1 which is a distantly related protease to the other *Plasmodium* proteases and studies have shown that adding the sequence that forms the nose-like structure to FP-1 results in the lack of enzyme activity of FP-1 due to the fact that the residue Glu-20 that mediates the folding is absent in FP-1 as well as other related proteases. Also within the *Plasmodium* species the falcipains were observed their unique motif (motif 19) which interestingly forms the arm-like protrusion in the structure. Motif 19 is located on the distal end of the right side of the protease. Since the motif is located at a distance greater than 25 Å from the active site it is thought to participate in hemoglobin binding (Pandey & Dixit 2012). It is not clear how hemoglobin binding occurs in the other *Plasmodium* proteases but it can be inferred

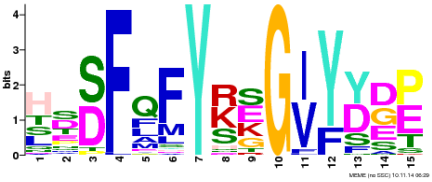
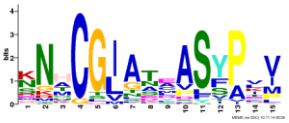
that the unique motifs observed in the murine and other human malarial proteases could participate in hemoglobin binding.

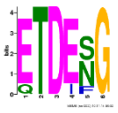
The *Leishmania* species were also shown to exhibit unique motifs shown in a purple box in Figure 3.2B. The *Leishmania* motifs could not be mapped onto a structure to infer the possible role of the motifs as there is currently no available crystal structure. However in future studies the *Leishmania* structures can be modelled and the unique motifs mapped to elucidate their role and the results confirmed with wet laboratory assays.

The *Trypanosoma* and *T.gondii* species each had unique motifs shown in green and black boxes respectively (Figure 3.2B) and these were successful mapped onto the available structures. The *T.gondii* unique motifs 38 and 45 are located in the distal end of the right domain of the protease and their role during the infective cycle is not clear, however motif 48 is known to participate in the folding of the protein (Kim, 2004). When the *Trypanosoma* unique motifs were mapped onto the cruzipain structure (1ME4) and they were not located in functionally relevant regions suggesting that the species shares functional motifs with other cysteine proteases such as motif 13 which makes up sub-site 2, the region that governs substrate specificity. This could also infer that the *Trypanosoma* species in this study each have their own motifs that define their structure and function.

The heat map also supported the results obtained from phylogenetic analysis discussed in Chapter 2 as each type of cathepsins had their own unique motifs which could be explained by the observation that they clustered together in the phylogenetic tree regardless of the organism they came from (Chapter 2, Figure 2.5). The unique motifs mainly found in the parasites were reported and summarized (Table 3.2).

Table 3.2 A summary of the species cysteine proteases' unique motifs according to the MEME HTML output. The motif positions on the sequence indicate the range of positions across all the sequences.

Motif number	Species name	Sequence position	Sequence Logos
Motif 4	All except the protozoan parasites and cathepsin O	68 - 111	
Motif 7	All except Trypanosoma, Leishmania, Cathepsin F and O	140 -169	
Motif 8	All except murine Plasmodium	214 - 248	
Motif 10	All except Protozoan parasites, Cathepsin L, F and W	94 - 134	
Motif 16 and 39	Murine *Plasmodium*	219 -239 182 - 196	 

Motif 19	Falcipains except Falcipain 1	182 -196	
Motif 34	Human Malarial *Plasmodium*	218 - 226	
Motif 39 and 64	*Trypanosoma*	182 – 196, 236 - 337	See appendix
Motif 43,53,56-58	Falcipain-1	83 – 93, 191 - 198, 110 - 119, 1 - 6, 96 - 101	See appendix
Motif 16, 18, 23 and 52	*Leishmania*	115 – 135, 212 – 232, 3 – 13, 107 - 114	See appendix
Motif 37, 38, 45 and 48	*T,gondii*	110-115, 212 – 217, 184 – 189, 9 -14	See appendix

Motif analysis on its own is insufficient in providing a conclusion about the distinct patterns among the studied proteases. Therefore to further elucidate and understand the differences it is ideal to zoom into the sequences and evaluate the physicochemical properties and well as the amino acid composition since protein characteristics are influenced by the residues within the sequence.

3.3.2 PHYSICOCHEMICAL PROPERTIES ANALYSIS

To understand protein function diversity, it is vital to understand physicochemical properties of the constituent amino acids, even though physicochemical properties go beyond the mere sum of properties of the various component amino acids (Cozzone 2002). It is possible to infer protein diversity within the same family of proteins based on their different physicochemical properties and in this study these properties were calculated for all the retrieved sequences (Appendix A-I). The results were averaged for each group and the summary tabulated (Table 3.3).

Table 3.3 A summary of all the calculated physiochemical properties of the catalytic domain of the parasites, vectors and hosts cysteine proteases using BioPython. The blue background indicates the properties of parasites group while the red background indicates the vectors group and the green background highlights the hosts group.

Property	**Parasites**		**Vectors**		**Hosts**	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
Aromaticity	0.112	±0.025	0.114	±0.008	0.115	±0.009
GRAVY	-0.269	±0.187	-0.390	±0.159	-0.404	±0.154
PI	5.168	±0.863	6.486	±1.330	7.210	±1.437
Instability index	37.290	±9.459	27.039	±6.584	29.017	±6.584
Mr (kDa)	28.601	±3.804	25.033	±0.983	25.664	±1.429

All the groups seem to have similar aromaticity indices of about 0.11 and very low standard deviations which show that they have the same content of the aromatic residues. However the GRAVY summary showed negative averages indicating that the proteins are hydrophilic although the standard deviations were quite high meaning that in each of the groups there is variation in the GRAVY values therefore there is a need to observe the data distribution more closely. When observing the pI of each of the groups there was a distinct difference in the results. The parasite group showed a pI of 5.168 while the vectors and hosts showed a pI of 6.486 and 7.210 respectively indicating that the different groups have minimal solubility at varying pH and the low respective standard deviations show that most of the data in each of the groups did not vary significantly. The pH values indicate the organisms in each group had overlapping pH where they can be purified for further analysis in a wet laboratory. The

instability indices for each of the groups was below 40 showing that the proteins across all the groups have long half-lives therefore they are stable, however there are large standard deviations that were observed which illustrate that in as much as most of the proteins were stable with an instability index below 40 there is variation in the data distribution. Finally the molecular weight calculations showed that the vector and host groups have similar weights of about 25 kDa while the parasite group is slightly larger with a weight of 28.601 kDa. The standard deviations are low meaning that there is little variation in the molecular weight values.

3.3.2.1 CATALYTIC DOMAIN AND FULL LENGTH SEQUENCE GRAVY ANALYSIS

Based on the amino acid sequence alone it is possible to determine whether a segment or the entire protein is hydrophobic or hydrophilic in nature through GRAVY calculations. The hydropathicity of a protein provides insight in the nature of interactions and ligands that can interact with the protein (Kyte & Doolittle 1982).Hydropathicity evaluates the hydrophobicity or hydrophilicity of a protein. The greater the hydropathic index the more hydrophobic the protein and the more negative the index the more hydrophilic the protein is.

The aim of this study is to compare the physicochemical properties of different groups hence box plots are the most appropriate method of showing the data. Boxplots show and compare the distribution of the properties' values in quarters. The lower quartile shows the value below which 25% of the data is contained while the upper quartile shows the value above which 25% of the data is contained. The box plot also shows the median value within the range of data as well as the outliers which are defined as any data point that is calculated 1.5 the interquartile range whether below the lower quartile or above the upper quartile.

In this study the GRAVY calculations were visualized using boxplots and the levels of hydropathicity in the proteins evaluated (Figure 3.3).

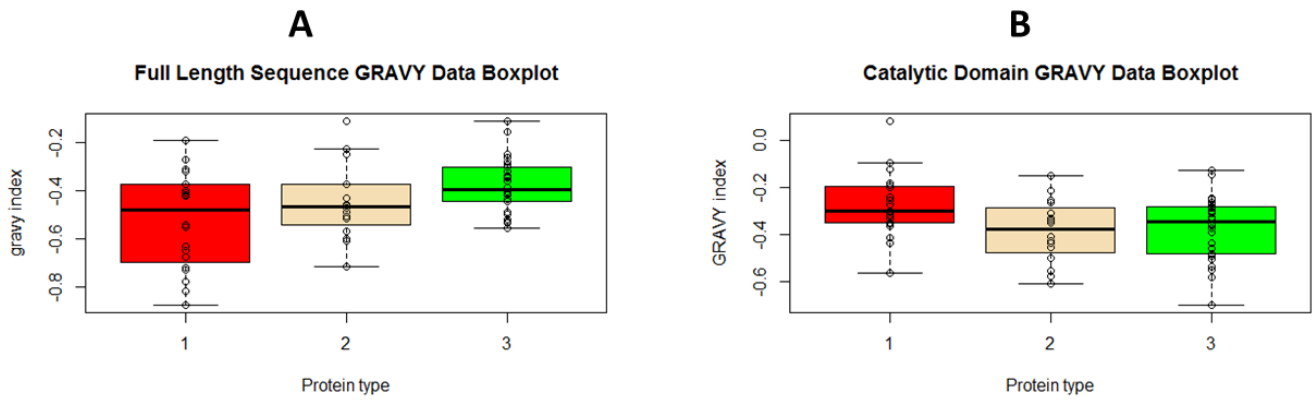


Figure 3.3 Box plots showing the GRAVY data of the full length sequences (A) and the catalytic domain (B) of the parasites, vectors and host cysteine proteases shown in red, wheat and green respectively. GRAVY is calculated by adding the hydropathic value for each residue and dividing by the length of the sequence.

A direct observation of the boxplots showed that most of the data was in the negative range indicating that the proteases interact well with water. The boxplot medians in the full length sequences were at the same point while those of the catalytic domain of the host and vectors differed from that of the parasites. This suggested that in full length sequences all the data had the same distribution and the variations noted were expected in samples, whereas in the catalytic domain the host and vector samples were not so different from each other but differed from the parasites group. When the Kruskal-Wallis test was performed and it returned a p-value of 0.84 and 0.02254 for the full length sequences and the catalytic domain respectively. Since the p-value of the full length sequences was higher than the alpha level 0.05, there was insufficient evidence to reject the null hypothesis therefore concluding that all the full length sequences are the same while the catalytic domain sequences have significant variation since their p-value was lower than the alpha level. To confirm the differences between the catalytic domain and the full length sequences the Kruskal-Wallis test was used and it returned a p-value of 0.06003, a value above 0.05 suggesting that since there is no significant difference between the full length sequences and the catalytic domain hence the prodomain in the full sequences does not affect the hydropathicity of the proteases.

3.3.2.2 CATALYTIC DOMAIN AND FULL LENGTH SEQUENCE AROMATICITY ANALYSIS

In a number of studies aromatic residues have been implicated in forming aromatic interactions within the protease for stabilization or with the substrate (Lanzarotti et al. 2011). Aromatic interactions can only occur in the presence of aromatic residues hence in this study the relative frequencies of the aromatic residues in the retrieved sequences were investigated and illustrated using a box plot (Figure 3.4).

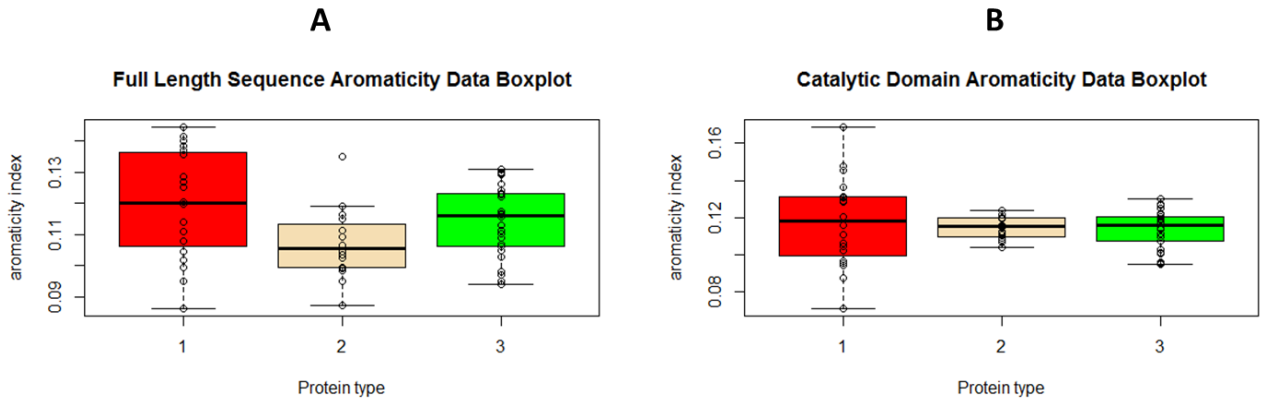


Figure 3.4 Box plots showing the aromaticity data of the full length sequences (A) and the catalytic domain (B) of the parasites, vectors and host cysteine proteases shown in red, wheat and green respectively. The aromaticity index is determined by the relative frequencies of the aromatic amino acids, phenylalanine (Phe), tyrosine (Tyr) and tryptophan (Trp) within the sequences.

In the full length sequence boxplot (Figure 3.4A) the medians were not at the same point implying that the samples might be different, however when the Kruskal-Wallis test was performed it returned a p-value of 0.1965 which is above the alpha level 0.05 suggesting that the variability in the positions of the medians showed mere differences that are expected within populations.

The medians observed from each of the groups in the catalytic domain boxplots are virtually at the same point suggesting that the data follows the same distribution (Figure 3.4B). To confirm this observation the Kruskal-Wallis test was used and it returned a p-value of 0.9359. Since the returned p-value was greater than 0.05 as the hypothesis testing is done 95% level of significance, there was insufficient evidence to reject the null hypothesis that states that the data

follows the same distribution. Therefore it could be concluded that the proteins in each of groups have the same content of aromatic residues.

When the full length sequences aromaticity was compared to that of the catalytic domain the medians were virtually at the same point although the distribution of the catalytic domain of the host and vectors was not as wide as that of the full length sequences. When the test statistic was applied a p-value of 0.6209 was obtained hence it could be concluded that there is no difference in aromaticity of the full length sequences and the catalytic domain hence the despite the presence of the prodomain in the full length sequences it does not affect the protease aromaticity suggesting that the prodomain is not rich in aromatic residues. This is probably due to the fact aromatic residues stabilize proteins and since the prodomain is eventually degraded to activate the protease it does not need to be highly stable to avoid an energy expensive proteolysis of the prodomain.

3.3.2.3 CATALYTIC DOMAIN AND FULL LENGTH SEQUENCE INSTABILITY INDEX ANALYSIS

The ability of a protein to retain its activity and structural conformation was evaluated by calculating the instability index of both the full length sequences and the catalytic domain. The distribution of the instability indices was visualized using box plots (Figure 3.5).

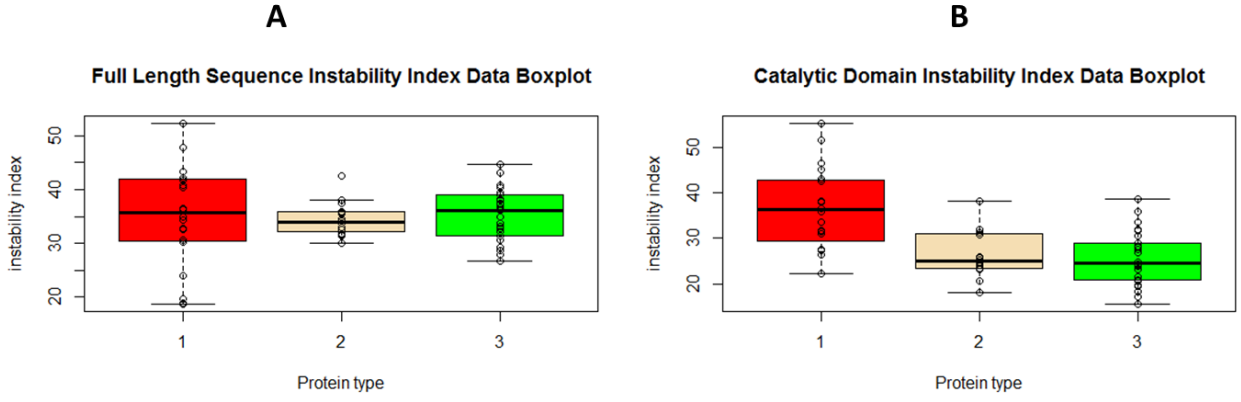


Figure 3.5 Box plots showing the instability index data of the full length sequences (A) and the catalytic domain sequences (B) of the parasites, vectors and host cysteine proteases shown in red, wheat and green respectively. The instability index is calculated by evaluating the N-terminal residue.

In the full length sequences box plot the groups' data had varying distribution patterns, the parasites group instability indices varied from 20 to above 50 while vectors groups' indices were between 30 and 40 suggesting that most of the proteases in the group had the same stability with the exception of the *G.gallus* cathepsin K which was above 40. The hosts' group data ranged from 30 to 45 almost like that of the vectors. When the test statistic was applied it returned a p-value of 0.2166 suggesting that even though the data was distributed differently in each group, the proteases had the same stability.

The parasite group median was positioned above that of the vectors and hosts meaning that the host and vector proteins have similar instability indices (Figure 3.5B). The parasites group showed some unstable proteins that had an instability index above 40 namely the *Leishmania, T.gondii* and *P.chabaudi* species. The Kruskal-Wallistest returned a p-value of 0.002742 which was less than 0.05 therefore the null hypothesis stipulating that the data had the same distributions was rejected and there was variation in the groups' instability index. The Kruskal-Wallis test was also performed on the host and vector groups to confirm that indeed the proteins had similar distributions. The test returned a p-value of 0.84 therefore failing to reject the null hypothesis hence the host and vector groups have similar distributions. When comparing the full length sequences to the catalytic domain the host and vector groups had much lower medians suggesting that the catalytic domain is more stable than the entire full sequence. Interestingly the parasite medians in both plots remained the same meaning that the stability of the catalytic domain and the full length sequence could be the same. When the Kruskal-Wallis test was applied a p-value of 4.674e-05 was obtained therefore it could be concluded that the stabilities of the catalytic domain and the full length sequences were different and by observation of the box plots (Figure 3.5) the catalytic domain is more stable than the full length sequences. It could be inferred that the prodomain affects the stability of the proteases and without it the proteases are more stable.

3.3.2.4 CATALYTIC DOMAIN AND FULL LENGTH SEQUENCE ISOELECTRIC POINT ANALYSIS

To further establish the differences between the proteases the environments at which they have minimal solubility were established through pI calculations and these were visualized using a box plot (Figure 3.6).

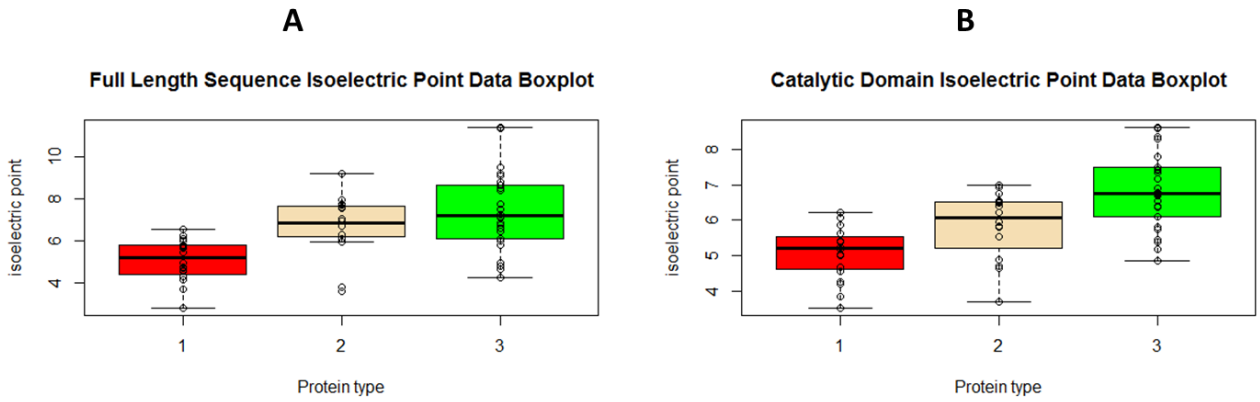


Figure 3.6 Box plots showing the isoelectric points of the full length sequences (A) and the catalytic domain (B) of the parasites, vectors and host cysteine proteases shown in red, wheat and green respectively.

In Figure 3.6A the host and vector medians were at the same point suggesting they had similar pI of about 7. However they had varying distributions the hosts group had a much wider distribution suggesting that the proteases within the group had varying pI, ranging from about 5 to 11. The parasites group had a much lower median and a distribution range below 7. When the test statistic was performed it gave a p-value of 0.008671 meaning that the protease groups have varying pI.

In Figure 3.6B the groups showed a difference in the median positions also illustrating that each of the groups had different pI. This was confirmed by performing a Kruskal-Wallis test which returned a p-value of 2.679e-05 which was interpreted as indicating that there is a significant difference between the pI of the groups. From direct observation of the plot it can be seen that the host and vectors groups are have minimal solubility in the acidic range although the parasitic proteases have much lower pI. The vectors group showed that they have minimal solubility pH range from 5 to 9.

When the full length sequences plot was compared to that of the catalytic domain the pI of the hosts and vectors in catalytic domain plot were much lower (Figure 3.6). The vectors group was basic in the full length sequences plot seemed to be acidic in the catalytic domain. Although the hosts' median in both plots was at the same point the distribution narrowed in catalytic domain from a range of pH 5 to 11 to a range of pH 5 to 9. The differences were assessed using the Kruskal-Wallis test that returned a p-value of 2.804e-07 indicating that there is a significant difference between the two groups which could be interpreted as that the prodomain plays a role in the solubility environment of the proteases.

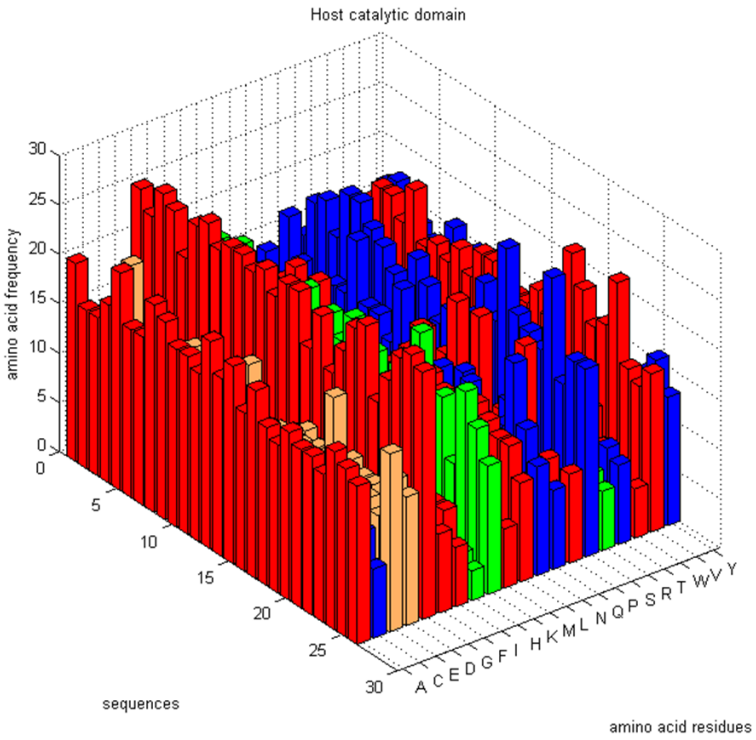
After determining the various properties of the proteases it was crucial to understand the amino acids underlying the varying physicochemical properties. Consequently an amino acid composition analysis was performed and reported in the next section.

3.3.3 AMINO ACID COMPOSITION ANALYSIS

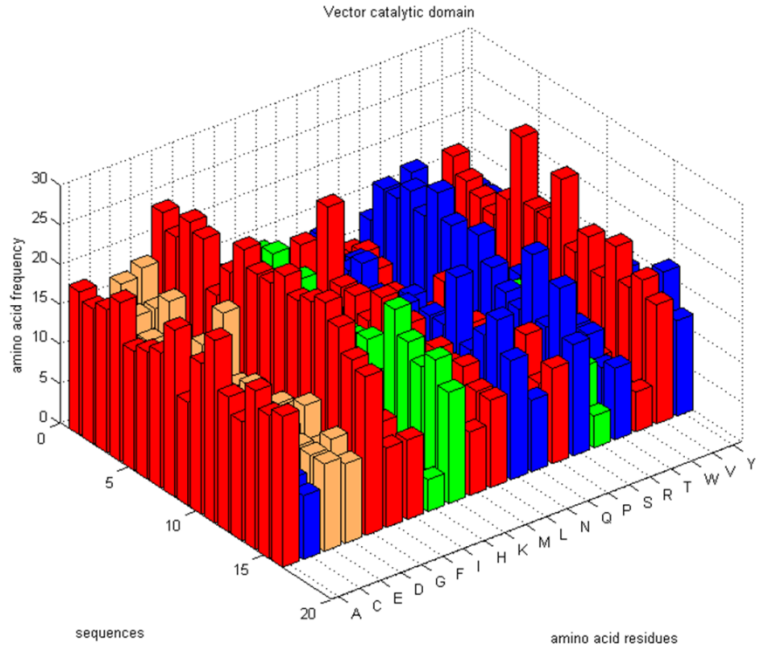
3.3.3.1 CATALYTIC DOMAIN ANALYSIS

The amino acid composition of proteins gives them their distinct features and properties. The physiochemical properties discussed in section 3.3.2.2 are a result of the amino acid residues present in the proteases. In this study the amino acid compositions for each group were investigated and visualized using bar plots generated from MATLAB (Figure 3.7).

A


Host catalytic domain
amino acid frequency
0
5
10
15
20
25
30
sequences
A C E D G F I H K M L N Q P S R T W V Y
amino acid residues

B


Vector catalytic domain
amino acid frequency
0
5
10
15
20
25
30
sequences
A C E D G F I H K M L N Q P S R T W V Y
amino acid residues

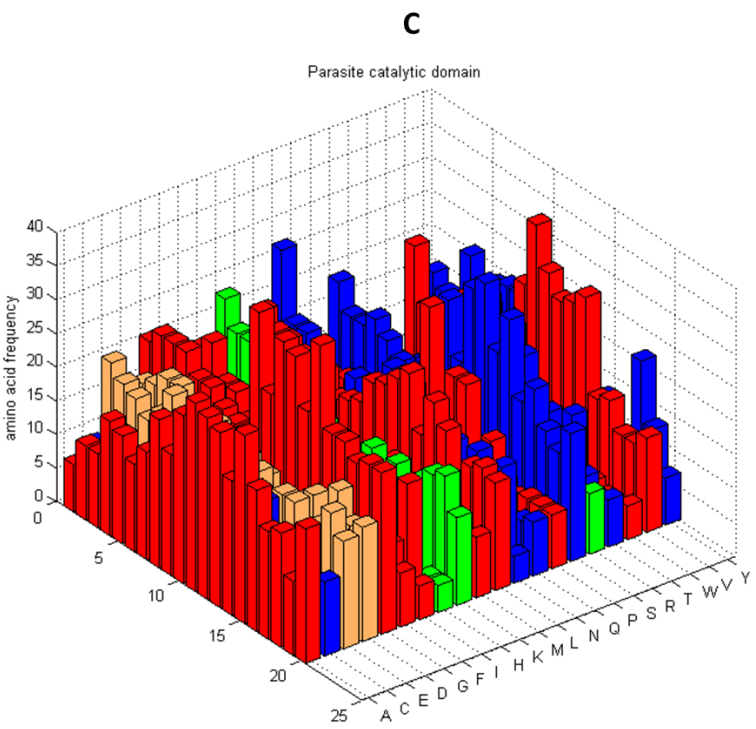


Figure 3.7 3D bar plots showing the catalytic domain amino acid composition in hosts (A), vectors (B) and parasites (C) cysteine proteases generated in MATLAB. The individual residues, sequences and the amino acid frequencies are illustrated on the x, y and z axis respectively. The hydrophobic residues are shown in red while the hydrophilic residues are shown in blue. The negatively charged residues are indicated orange and the positively charged residues

The host and vector sequences were observed to have similar patterns of amino acid compositions. There were no distinctions between the host and vector groups. This could have been due to the fact that the host and vector sequences do not cluster according to the organisms they come from but rather the type of cathepsins they are. The patterns that were observed between the host and vector groups followed the clustering pattern in the phylogenetic analysis (section 2.3.3). The cathepsin H proteases seemed to favour the alanine residues while the cathepsins L and F favoured the threonine residues (Figure 3.7A and B).

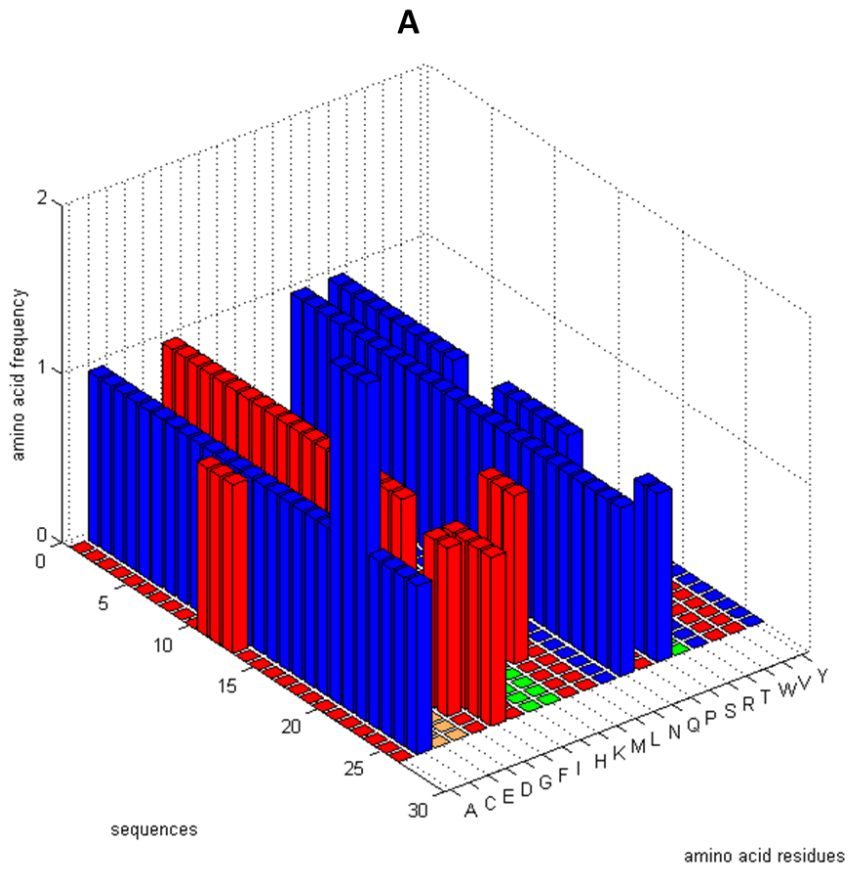
The parasites group was observed to favour less alanine, phenylalanine and asparagine but preferred threonine. The *Trypanosoma* and *Leishmania* species were observed to favour threonine relative to other parasite sequences. The *Plasmodium* species preferred the hydrophilic glutamine residues and less hydrophobic alanine residues. The *Leishmania* and *Trypanosoma* residues were observed to favour hydrophobic residues unlike the *Plasmodium* species which favoured the hydrophilic residues and the negatively charged glutamic acid and aspartic acid. In

as much there were distinct patterns that were observed among and within the parasite group their catalytic significance could not be inferred therefore it was important to do a sub-site analysis.

3.3.3.2 SUB-SITES ANALYSIS

3.3.3.2.1 SUB-SITE 1 ANALYSIS

Sub-site 1 is the least defined of all the sub-sites and it does not contribute much to substrate specificity, however it is the region that initiates catalysis therefore there is a need to understand the amino acids present in each group that participate in initiating catalysis (Sabnis et al. 2003). The sub-site 1 amino acid compositions are also illustrated using a bar plot (Figure 3.7).



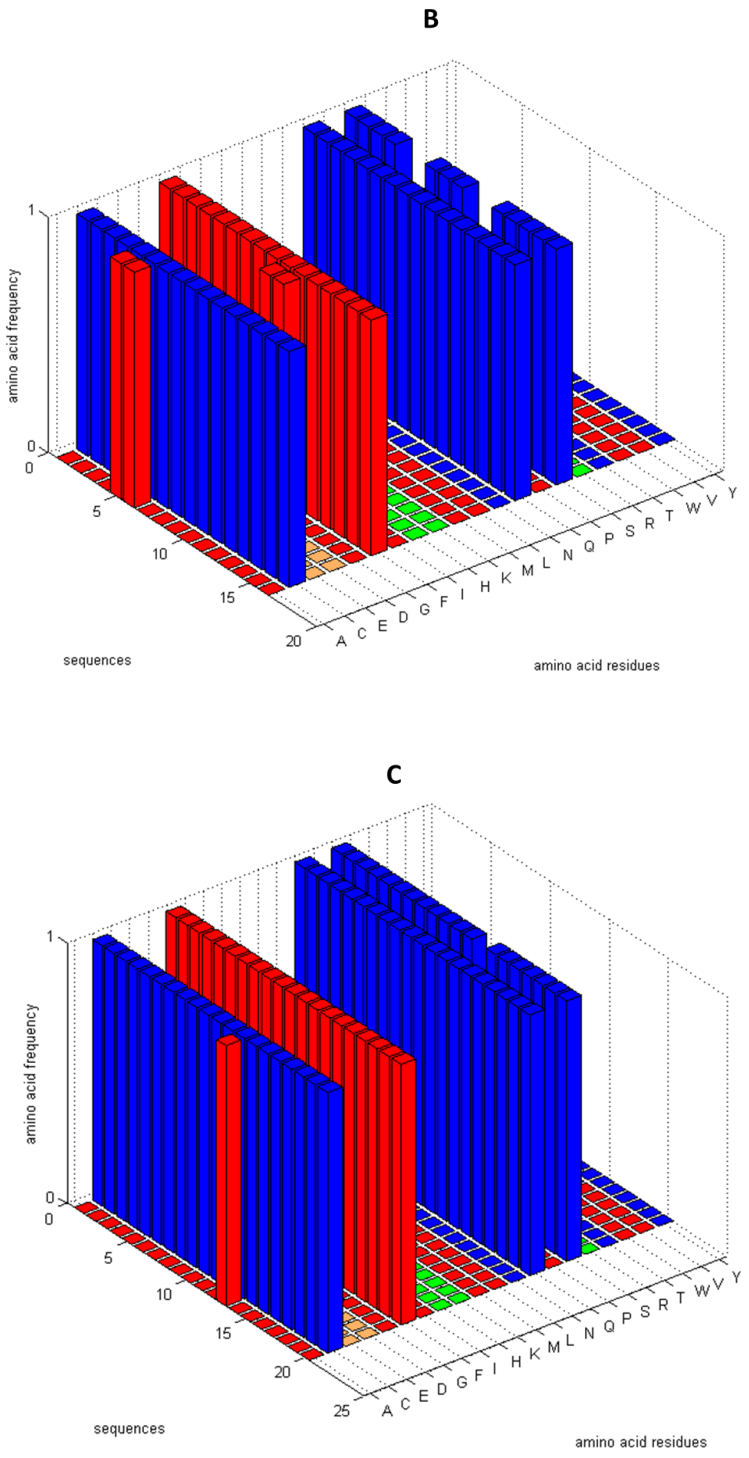
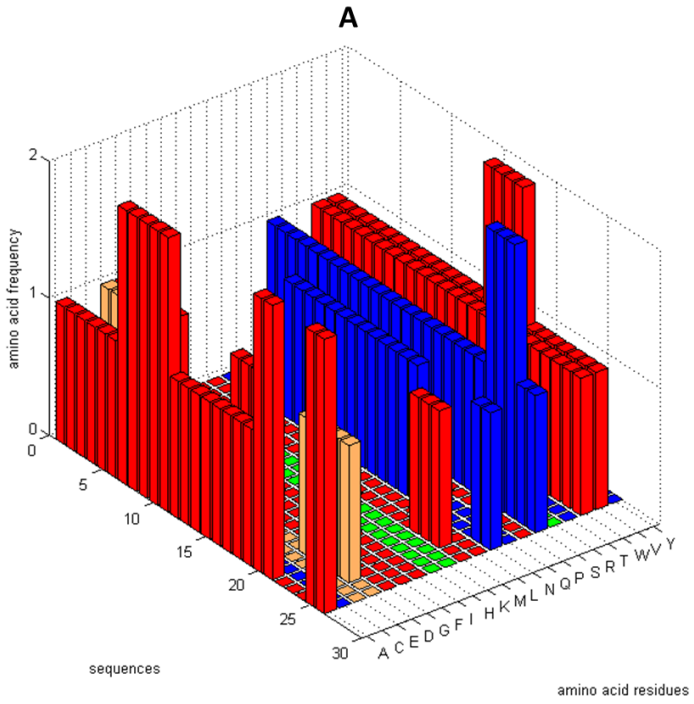


Figure 3.8 3D bar plots showing the sub-site 1 amino acid composition in hosts (A), vectors (B) and parasites (C) cysteine proteases generated in MATLAB. The individual residues, sequences and the amino acid frequencies are illustrated on the x, y and z axis respectively. The hydrophobic residues are shown in red while the hydrophilic residues are shown in blue. The negatively charged residues are indicated orange and the positively charged residues are shown in green.

Sub-site 1 is very well conserved throughout all the groups; glutamine is conserved as it plays a role in stabilizing the oxyanion intermediate during substrate catalysis (Sabnis et al. 2003). This region is primarily hydrophilic as it is located in the exterior of the protein where it interacts with the substrate solvent.

3.3.3.2.2 SUB-SITE 1' ANALYSIS

Sub-site 1' is known not to be the primary site of substrate specificity however the variations in amino acid compositions in the parasite, host and vector groups could generate notable activity differences. In this study the variations in this region were investigated and the results displayed in a bar plot (Figure 3.9).



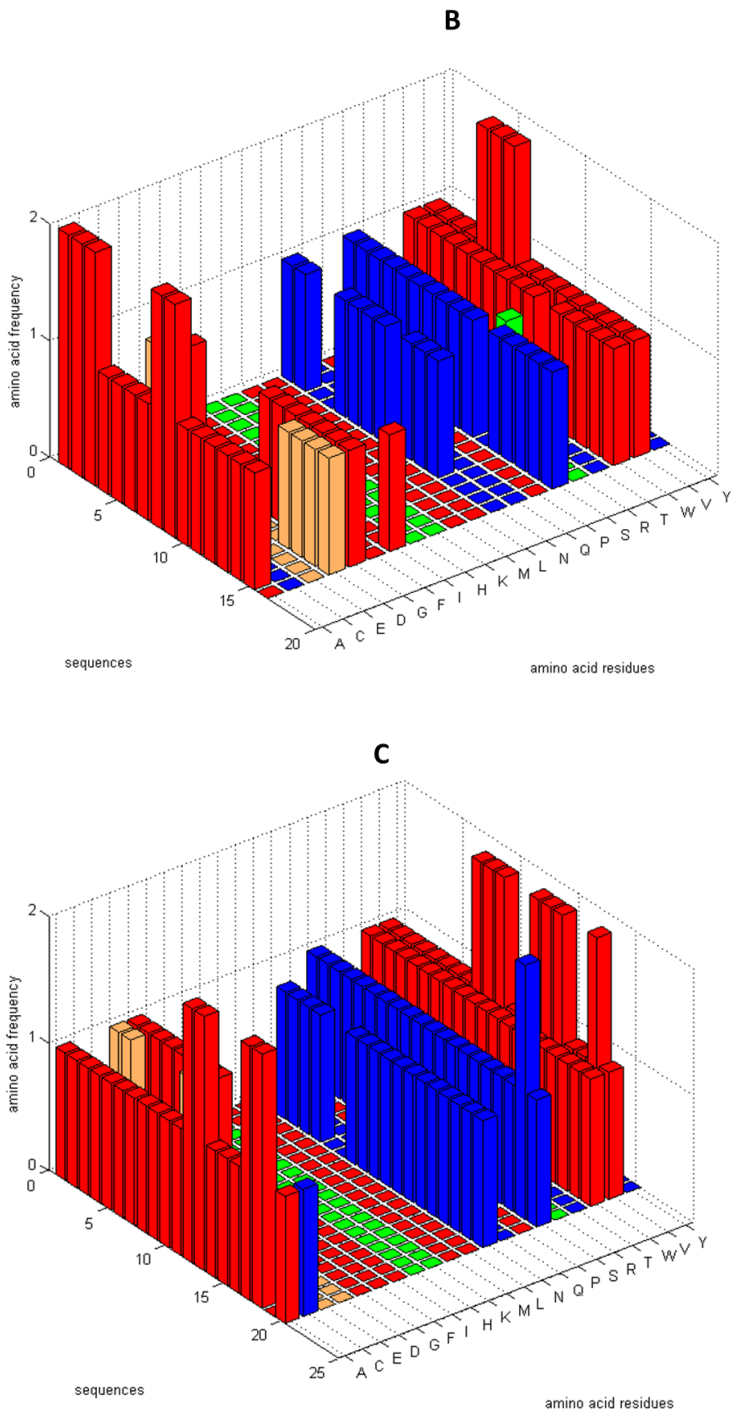


Figure 3.9 3D bar plots showing the sub-site 1' amino acid composition in hosts (A), vectors (B) and parasites (C) cysteine proteases generated in MATLAB. The individual residues, sequences and the amino acid frequencies are illustrated on the x, y and z axis respectively. The hydrophobic residues are shown in red while the hydrophilic residues are shown in blue. The negatively charged residues are indicated orange and the positively charged residues are shown in green.

From the direct observation of the plots it can be observed that they have similar hydrophobic and hydrophilic distributions. The host and the vector groups (Figure 3.8A and B) are more negatively charged relative to the parasites group.

In the hosts group the same pattern observed in the phylogenetic analysis and motif analysis (section 3.3.1.2) was observed (Figure 3.8B). Cathepsins of the same types had similar amino acid compositions. F, K and O cathepsins share the same preferences; they all favour the alanine residue. Whereas the L-cathepsins prefer the negative charged aspartic acid. The W and H cathepsins prefer methionine and tyrosine respectively instead of alanine. The S-cathepsins favour the glycine residue. It can therefore be inferred that these patterns imply activity differences within the cathepsins.

In the vectors group a similar pattern was observed as the O, F and K cathepsins favour alanine with the exception of *G.gallus*-O (Figure 3.8B). On the contrary the H-cathepsins favour tyrosine instead of alanine and the L-cathepsins favour aspartic acid, serine and alanine with the exception of *A.darlingi* which favour isoleucine. From this observation it can be noted that the H-cathepsins are different from the rest of the other cathepsins as they do not favour alanine.

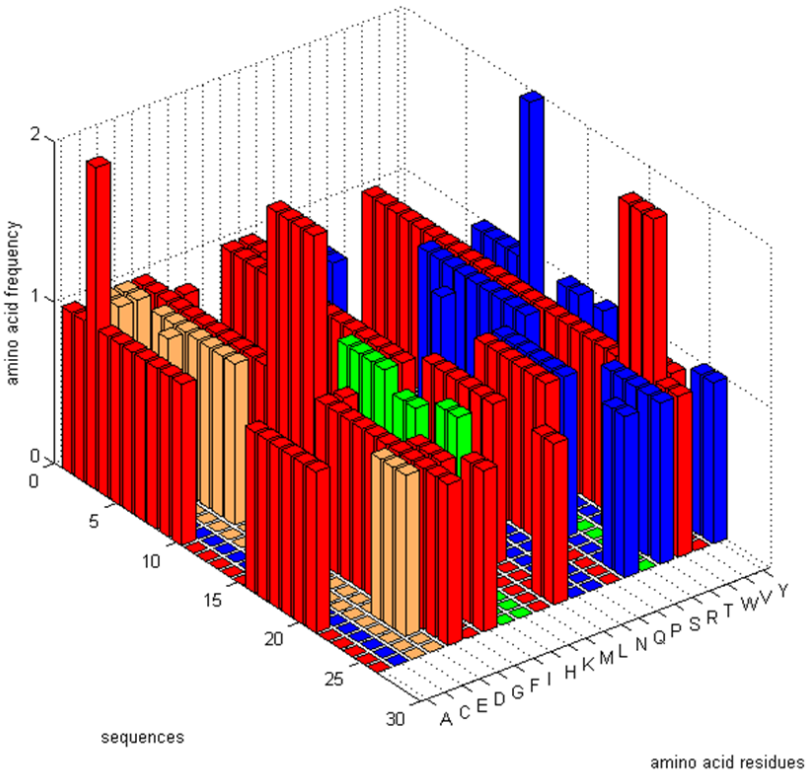
In the parasites group (Figure 3.8C) FP-2A and FP-2B like FP-1 and the murine malarial proteases favour the tyrosine residues although FP-1 also favours serine better than any other of the proteases within the same group, whereas the FP-3, KP-2 and VP-2 favour alanine.

In sub-site 1' there are noted differences in size and hydrophobicity considering that residues such as alanine and glycine have a much smaller mass relative to other residues hence the proteases that favour them are likely to have smaller sub-site 1'. Tryptophan is conserved in all groups as it is the residue that stabilizes the substrate through hydrophobic interactions (Sabnis *et al*. 2003).

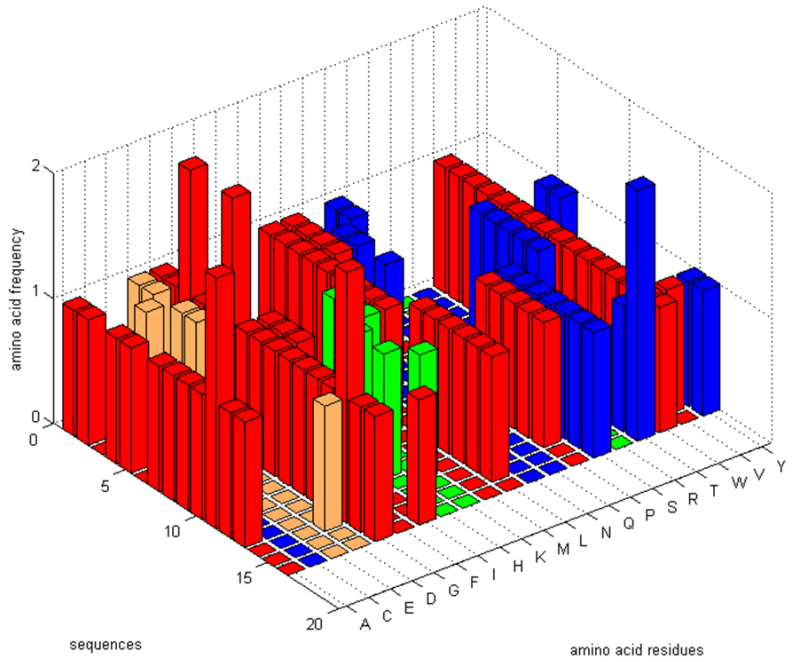
3.3.3.2.3 SUB-SITE 2 ANALYSIS

Sub-site 2 is the primary site of substrate preference therefore it is the region that arouses a lot of interest in designing highly selective inhibitors (Alves *et al*. 2003). It is for this reason that the amino acid composition analysis was performed at sub-site 2. The 3D plots generated from MATLAB are used to illustrate the analysis (Figure 3.6).

A


amino acid frequency
0
1
2
0
5
10
15
20
25
30
sequences
A C E D G F I H K M L N Q P S R T W V Y
amino acid residues

B


amino acid frequency
0
1
2
0
5
10
15
20
sequences
A C E D G F I H K M L N Q P S R T W V Y
amino acid residues

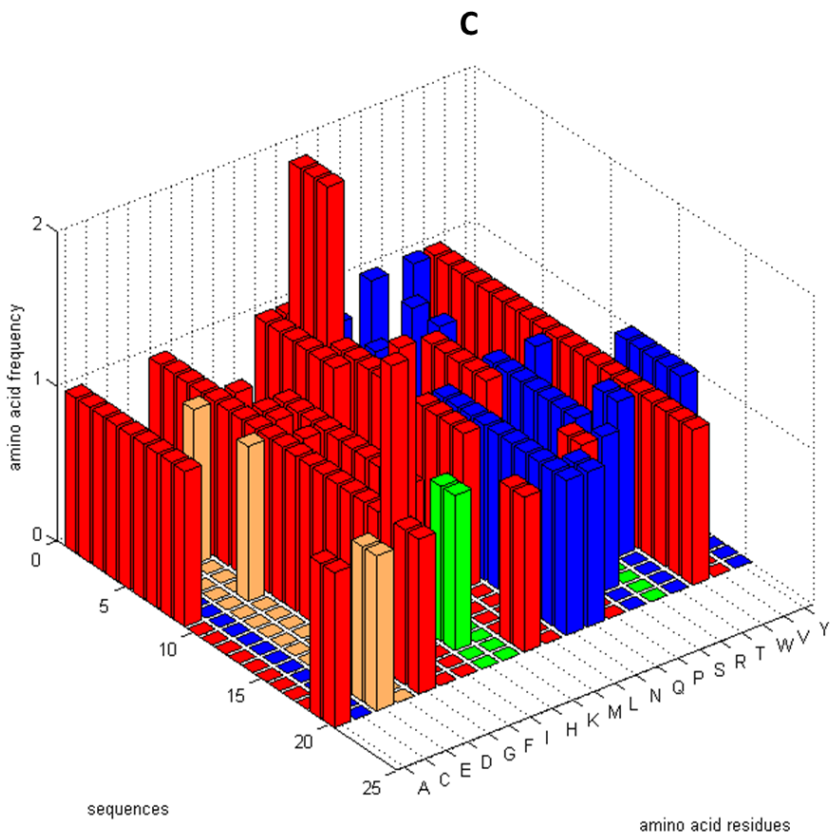


Figure 3.10 3D bar plots showing the sub-site 2 amino acid composition in hosts (A), vectors (B) and parasites (C) cysteine proteases generated in MATLAB. The hydrophobic residues are shown in red while the hydrophilic residues are shown in blue. The negatively charged residues are indicated orange and the positively charged residues are shown in green.

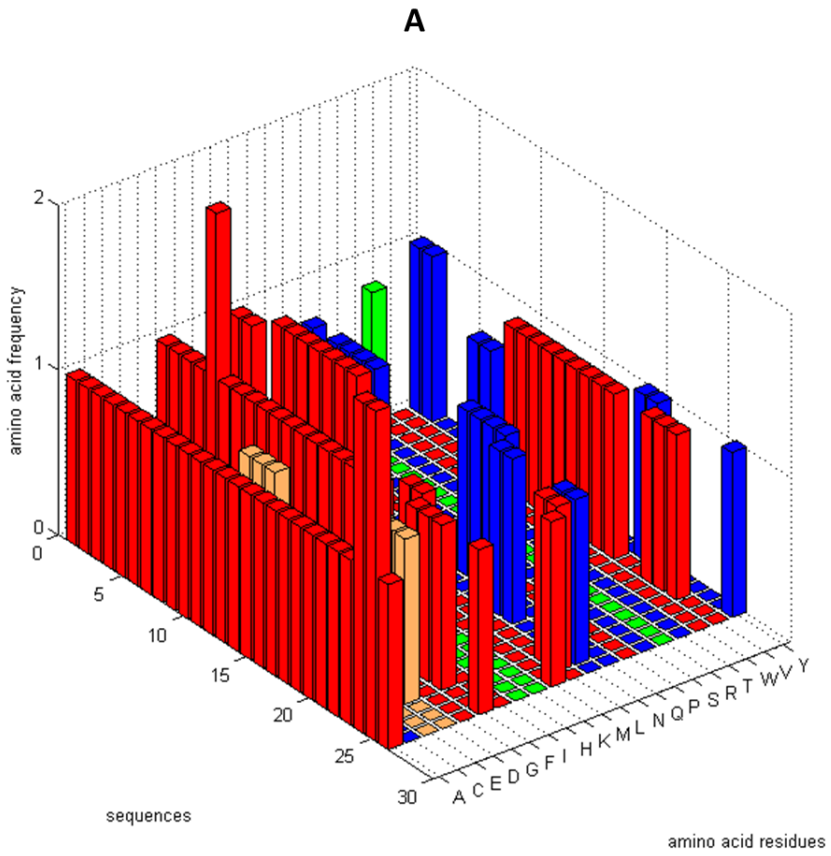
At sub-site 2 it was observed that the site is primarily hydrophobic with conserved tryptophan and glycine residues that are known to form hydrophobic interactions with the ligand. In the parasites group the FP-2A, 2B and 3 were found to favour more lysine residues, this was interesting as the lysine residues are known to occupy the bottom of the falcipain binding pocket influencing the size of the pocket and consequently the type of ligand that can bind in the pocket. The *Leishmania* species were noted to favor the hydrophilic tyrosine residues unlike the other protozoan parasites that favour the hydrophobic alanine residues which are known to play a role in substrate preferences. Thus the *Leishmania* residues are relatively more hydrophilic. The *Entamoeba* species seemed to favour the positive charged histidine residue while the human malarial *Plasmodium* species favour the methionine residues. The *Leishmania, Trypanosoma* and

the murine *Plasmodium* species were observed to be more hydrophilic relative to the human *Plasmodium* species.

In the hosts group the human proteases were observed to be the most hydrophobic when compared to the other host organisms. Both the host and vector groups were more positively and negatively charged while the parasites group was relatively more hydrophobic.

3.3.3.2.4 SUB-SITE 3 ANALYSIS

This is the glycine-rich region that is known to aid in the orientation of the substrate through hydrophobic interactions. Sub-site 3 also participates in substrate recognition and has little substrate discrimination (Sabnis et al. 2003). The variations in amino acid compositions in all the groups was investigated and shown using a 3D bar plot (Figure 3.10).



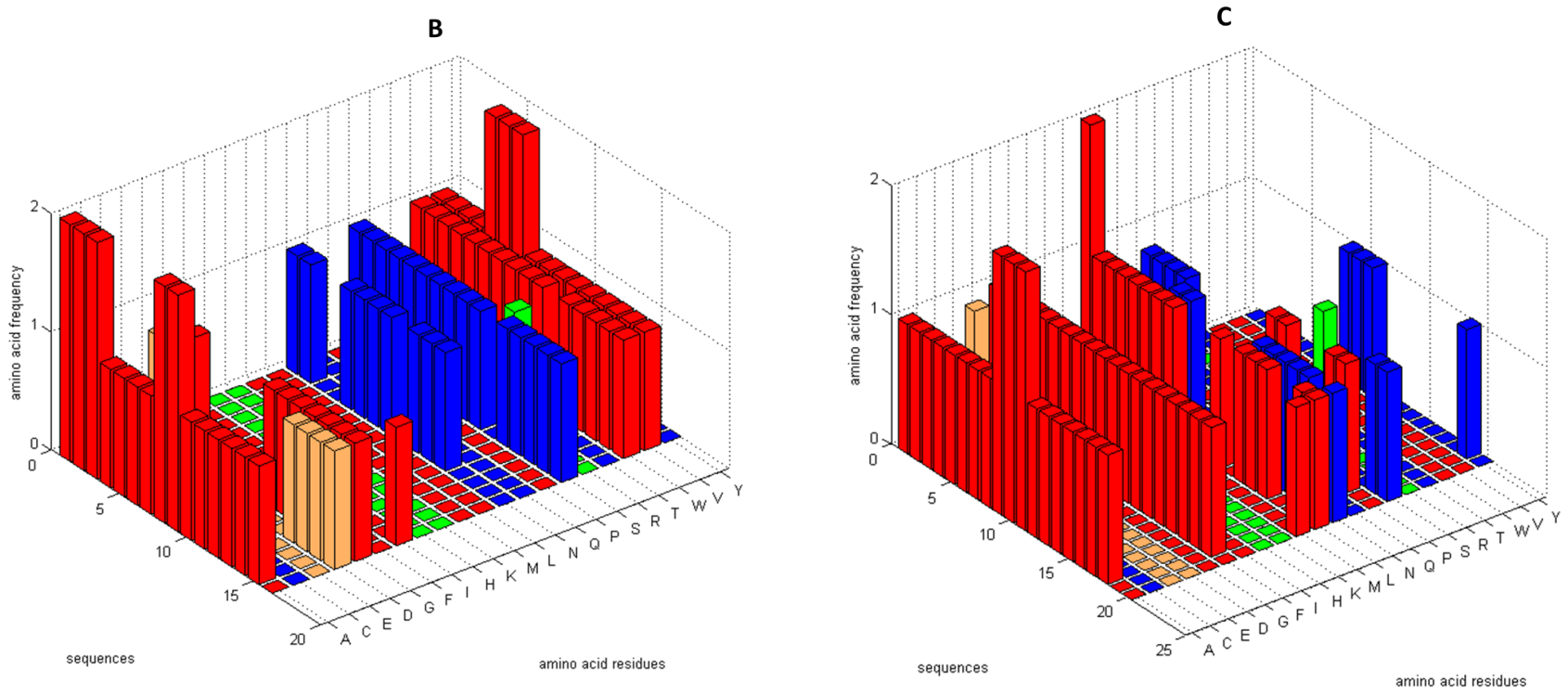


Figure 3.11 3D bar plots showing the sub-site 3 amino acid composition in hosts (A), vectors (B) and parasites (C) cysteine proteases generated in MATLAB. The hydrophobic residues are shown in red while the hydrophilic residues are shown in blue. The negatively charged residues are indicated orange and the positively charged residues are shown in green.

The host sub-site 3 is primarily hydrophobic with alanine being the most conserved residue. The O-cathepsins were observed to favour alanine more than any of the proteases as well as leucine and methionine. The W and S cathepsins favour aspartic acid while the F-cathepsins favour tyrosine.

In the vectors group sub-site 3 was also observed to be also hydrophobic with valine and tyrosine residues being conserved. This region was also observed to favour alanine and serine. The W-cathepsins were observed to favour negatively charged aspartic acid whereas the F, W and O-cathepsins were observed to be the glycine rich proteases.

The parasites group was observed to be hydrophobic as the cathepsins appeared to favour alanine and phenylalanine. The murine cathepsins were observed to favour the alanine residues while *T.gondii*-L favours phenyalanine. FP-1, 2A and 2B, the *Trypanosoma* species favour the asparagine unlike the *Leishmania* species which favour glutamine. Of all the proteases in this group FP-1 preferred leucine and lysine instead of alanine and phenyalanine preferred by the rest of the proteases in this group. The *Trypanosoma* and *Leishmania* cathepsins are the only hydrophilic ones in the parasite group suggesting that their sub-site 3 is most likely to recognize polar substrates while the rest of the proteases recognize non-polar substrates.

Of all the groups in this analysis the vectors group is the only group that favours tyrosine and valine. The studied cathespsins are observed to be hydrophobic with the exception *Trypanosoma* and *Leishmania* cathepsins which are hydrophilic. There are notable differences in amino acid preferences thus demonstrating sub-site 3 binding pockets vary in size from one cathepsin type to another.

CHAPTER FOUR

4. CONCLUSIONS AND FUTURE WORK

This study is a step towards elucidating the differences as well similarities in the primary structure and physicochemical properties of the protozoan parasites, their host and vector cysteine protease catalytic domains; which are essential in understanding the activity of the proteases. Considering cysteine proteases are validated viable drug targets in protozoan infections, this study is vital in drug design since limited knowledge of these differences or lack thereof has resulted in drug toxicity. This study sought to answer the question: Are there any primary structure and physicochemical differences or similarities between the catalytic domains of cysteine proteases of protozoan parasites and their respective hosts and vectors, which play a role in the protease activity and/or substrate preferences?

In the initial stage of the study 62 L-like cathepsin sequences were retrieved and classified in three groups, the hosts, vectors and parasites. The phylogenetic analysis showed that the cathepsins in the host and vector groups clustered according to their cathepsins instead of the organisms they came from while in the parasites group they clustered according to their organisms. The motif analysis confirmed these observations as the same cathepsin types in the host and vector groups shared unique motifs whereas each parasite organisms had their own unique motifs. These unique motifs were observed to occur in the regions outside the catalytic zone. Motif analysis also showed 6 conserved motifs across all organisms, 5 of which were in the catalytic region while the sixth one is involved in protein folding away from the catalytic region. These observations did not contribute to the final conclusions.

The physicochemical analysis showed that the cathepsins studied had similar aromaticity and pI. The parasite cathepsins are less stable and larger in molecular weight relative to the host and vector cathepsins. The GRAVY scores also revealed that the parasites group was more hydrophobic. This revelation was confirmed by the amino acid composition analysis which showed a higher content of hydrophobic residues relative to the host and vector groups.

The amino acid acid analysis was particularly useful and informative. It showed that sub-site 1 is well conserved unlike sub-site 2 which is highly variable although the residues involved in hydrophobic interactions remained conserved. The sub-site 2 variability aroused interest as this is the region that is responsible for ligand specificity and preference. This observation sheds light on the ligand preferences of the various groups. There variations mainly occurred in the residues that line the bottom of the sub-site 2 binding pocket thus influencing the size of the binding pocket and the ligand that can be accommodated during catalysis. It can be inferred that the cathepsins use size as means to select their ligands. For instance within the parasites group there were also distinct patterns, the falcipains with the exception of FP-1 favoured the lysine which could suggest that the binding pocket of these species is much wider hence it can accommodate ligands of a much bigger size relative to the other protozoa parasites. The *Leishmania* species was noted to favour the hydrophilic tyrosine residue instead of alanine which plays a role in ligand preference hence we could conclude that the *Leishmania* species would probably prefer polar-based ligands while the other parasites prefer hydrophobic ligands. The host and vector groups were also noted to be negatively and positively charged while the parasite group is more hydrophobic, these observations ould possibly play role in ligand specificity as the parasite cathepsins would prefer non-polar ligands while the host and vector cathepsins would prefer charged ligands.

As stated earlier this study aimed to identify the primary structure and physicochemical properties differences and similarities within the host, vector and parasite groups' cysteine proteases. Having identified the differences and similarities the next step is to undertake a physicochemical analysis of each of the sub-site regions and validate them with an appropriate test statistic that takes into account the short sequence lengths of the sub-site regions. The ligand preferences concluded from this study can be further studied through structure-based virtual screening of the suggested ligand types from available inhibitor libraries. The protozoan cysteine protease structures that are not currently available can be modelled through homology modelling using available structures from other protozoans as templates to facilitate the structure-based virtual screening process.

This study has offered a perspective on the ligand preferences of the protozoan parasites' cysteine proteases compared to their hosts and vectors by evaluating the protease physicochemical properties and statistically validating them using the Kruskal-Wallis test. As a direct result of the methodology the differences in this study are among the three groups thus a statistical analysis such as Kolmogorov-Smirnov test can be used to identify the group that is different from the rest by using combinations of two for the three groups.

Current inhibitors for cysteine proteases in protozoan parasites lack selectivity and specificity resulting in drug toxicity. Therefore this study shows promise in the design of specific anti-protozoan inhibitors that can be designed by exploiting the parasite preferences elucidated in this study particularly in the sub-site 2 region.

REFERENCES

Abdulla, M.-H. et al., 2008. RNA interference of Trypanosoma brucei cathepsin B and L affects disease progression in a mouse model. *PLoS neglected tropical diseases*, 2, p.e298.

ALTSCHUL, S., 1998. Fundamentals of database searching. *Trends in Biotechnology*, 16, pp.7–9.

Alvar, J. et al., 2012. Leishmaniasis worldwide and global estimates of its incidence. *PLoS ONE*, 7.

Andrews, K.T., Fisher, G. & Skinner-Adams, T.S., 2014. Drug repurposing and human parasitic protozoan diseases. *International Journal for Parasitology: Drugs and Drug Resistance*, 4(2), pp.95–111.

Antunes, A.C.M. et al., 2002. Cerebral trypanosomiasis and AIDS. *Arquivos de Neuro-Psiquiatria*, 60, pp.730–733.

Apt, W., 2010. Current and developing therapeutic agents in the treatment of Chaga's disease. *Drug Design, Development and Therapy*, 4, pp.243–253.

Armougom, F. et al., 2006. Expresso: Automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. *Nucleic Acids Research*, 34.

Autino, B. et al., 2012. Pathogenesis of Malaria in tissues and blood. *Mediterranean Journal of Hematology and Infectious Diseases*, 4.

Bailey, T.L., 2008. Discovering sequence motifs. *Methods in Molecular Biology*, 452, pp.231–251.

Bailey, T.L. et al., 2006. MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, 34.

Barratt, J.L.N. et al., 2010. Importance of nonenteric protozoan infections in immunocompromised people. *Clinical Microbiology Reviews*, 23, pp.795–836.

Barrett, M.P. et al., 2011. Drug resistance in human African trypanosomiasis. *Future microbiology*, 6, pp.1037–47.Barrett, M.P. et al., 2003. The trypanosomiases. In *Lancet*. pp. 1469–1480.

Blackman, M.J., 2004. Proteases in host cell invasion by the Malaria parasite. *Cellular Microbiology*, 6, pp.893–903.

Bork, P. & Koonin, E. V, 1996. Protein sequence motifs. *Curr Opin Struct Biol*, 6, pp.366–376.
Bousema, T. & Drakeley, C., 2011. Epidemiology and infectivity of *Plasmodium falciparum* and *Plasmodium vivax* gametocytes in relation to Malaria control and elimination. *Clinical Microbiology Reviews*, 24, pp.377–410.

Brinkworth, R.I. et al., 2000. Host specificity in blood feeding parasites: A defining contribution by haemoglobin-degrading enzymes? *International Journal for Parasitology*, 30, pp.785–790.

Brun, R. et al., 2010. Human African trypanosomiasis. *Lancet*, 375, pp.148–159.

Chan, Y. & Walmsley, R.P., 1997. Learning and understanding the Kruskal-Wallis one-way analysis-of-variance-by-ranks test for differences among three or more independent groups. *Physical therapy*, 77, pp.1755–1762.

Chappuis, F. et al., 2005. Options for field diagnosis of human African trypanosomiasis. *Clinical Microbiology Reviews*, 18, pp.133–146.

Collins, W.E. & Jeffery, G.M., 2007. Plasmodium Malariae: Parasite and disease. *Clinical Microbiology Reviews*, 20, pp.579–592.

Cozzone, A.J., 2002. Proteins : Fundamental Chemical Properties. *Life Sciences*, pp.1–10.
Delaney, H.D. & Vargha, A., 2002. Comparing several robust tests of stochastic equality with ordinally scaled variables and small to moderate sized samples. *Psychological methods*, 7, pp.485–503.

Delespaux, V. & de Koning, H.P., 2007. Drugs and drug resistance in African trypanosomiasis. *Drug resistance updates : reviews and commentaries in antimicrobial and anticancer chemotherapy*, 10, pp.30–50.

Delgado, O. et al., 2008. Cutaneous Leishmaniasis imported from Colombia to Northcentral Venezuela: Implications for travel advice. *Travel Medicine and Infectious Disease*, 6, pp.376–379.

Dou, Z. & Carruthers, V.B., 2011. Cathepsin proteases in toxoplasma gondii. *Advances in Experimental Medicine and Biology*, 712, pp.49–61.

Dubey, J.P., 2009. History of the discovery of the life cycle of Toxoplasma gondii. *Int J Parasitol*, 39, pp.877–882.Eddy, S., 1998. Profile hidden Markov models. *Bioinformatics*, 14, pp.755–763.

Edgar, R.C., 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32, pp.1792–1797.

Enanga, B. et al., 2002. Sleeping sickness and the brain. *Cellular and Molecular Life Sciences*, 59, pp.845–858.

Ettari, R. et al., 2010. Falcipain-2 inhibitors. *Medicinal Research Reviews*, 30, pp.136–167.

Grzonka, Z. et al., 2001. Structural studies of cysteine proteases and their inhibitors. *Acta Biochimica Polonica*, 48, pp.1–20.

Van Helden, J., André, B. & Collado-Vides, J., 1998. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of molecular biology*, 281, pp.827–842.

Henikoff, S. & Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, 89, pp.10915–10919.

Hogg, T. et al., 2006. Structural and functional characterization of Falcipain-2, a hemoglobinase from the Malarial parasite Plasmodium falciparum. *The Journal of biological chemistry*, 281, pp.25425–25437.

Idicula-Thomas, S. & Balaji, P. V, 2005. Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in Escherichia coli. *Protein Sci*, 14, pp.582–592.

Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology*, 292, pp.195–202.

Katoh, K. et al., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic acids research*, 30, pp.3059–3066.

Kennedy, P.G.E., 2004. Human African trypanosomiasis of the CNS: current issues and challenges. *The Journal of clinical investigation*, 113, pp.496–504.

Kirchhoff, L. V., 2011. Epidemiology of American Trypanosomiasis (Chaga's Disease). *Advances in Parasitology*, 75, pp.1–18.

Korde, R. et al., 2008. A prodomain peptide of Plasmodium falciparum cysteine protease (falcipain-2) inhibits Malaria parasite development. *Journal of Medicinal Chemistry*, 51, pp.3116–3123.

Kosiol, C. et al., 2008. Patterns of positive selection in six mammalian genomes. *PLoS Genetics*, 4.

Kruskal, W.H. & Wallis, W.A., 1952. Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47, pp.583–621.

Kyte, J. & Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. *Journal of molecular biology*, 157, pp.105–132.

Lanzarotti, E. et al., 2011. Aromatic-aromatic interactions in proteins: Beyond the dimer. *Journal of Chemical Information and Modeling*, 51, pp.1623–1633.

Lecaille, F., Kaleta, J. & Brömme, D., 2002. Human and parasitic Papain-like cysteine proteases: Their role in physiology and pathology and recent developments in inhibitor design. *Chemical Reviews*, 102, pp.4459–4488.

Malvy, D. & Chappuis, F., 2011. Sleeping sickness. *Clinical Microbiology and Infection*, 17, pp.986–995.

McKerrow, J.H. et al., 2006. Proteases in parasitic diseases. *Annual review of pathology*, 1, pp.497–536.

Mharakurwa, S. et al., 2012. Malaria epidemiology and control in Southern Africa. *Acta Tropica*, 121, pp.202–206.

Miller, R.F. et al., 1998. Magnetic resonance imaging, thallium-201 SPET scanning, and laboratory analyses for discrimination of cerebral lymphoma and Toxoplasmosis in AIDS. *Sexually transmitted infections*, 74, pp.258–264.

Montoya, J.G. & Liesenfeld, O., 2004. Toxoplasmosis. In *Lancet*. pp. 1965–1976.

Murray, C.J. et al., 2013. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*, 380, pp.2197–2223.

Needleman, S.B. & Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48, pp.443–453.

Nuin, P.A.S., Wang, Z. & Tillier, E.R.M., 2006. The accuracy of several multiple sequence alignment programs for proteins. *BMC bioinformatics*, 7, p.471.

Pandey, K.C. et al., 2009. Regulatory elements within the prodomain of falcipain-2, a cysteine protease of the Malaria parasite Plasmodium falciparum. *PLoS ONE*, 4.

Pandey, K.C. & Dixit, R., 2012. Structure-function of falcipains: Malarial cysteine proteases. *Journal of Tropical Medicine*.

Pei, J. & Grishin, N. V, 2007. PROMALS: towards accurate multiple sequence alignments of distantly related proteins. *Bioinformatics (Oxford, England)*, 23, pp.802–808.

Pei, J., Kim, B.H. & Grishin, N. V., 2008. PROMALS3D: A tool for multiple protein sequence and structure alignments. *Nucleic Acids Research*, 36, pp.2295–2300.

Pereira-Chioccola, V.L., Vidal, J.E. & Su, C., 2009. Toxoplasma gondii infection and cerebral Toxoplasmosis in HIV-infected patients. *Future microbiology*, 4, pp.1363–1379.

Rassi Jr, A., Rassi, A. & Marin-Neto, J.A., 2010. Chaga's disease. *Lancet*, 375, pp.1388–402. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20399979.

Rizzi, L. et al., 2011. Design and synthesis of protein-protein interaction mimics as Plasmodium falciparum cysteine protease, falcipain-2 inhibitors. *European Journal of Medicinal Chemistry*, 46, pp.2083–2090.

Rose, P.W. et al., 2011. The RCSB Protein Data Bank: Redesigned web site and web services. *Nucleic Acids Research*, 39.

Rosenblatt, J.E., 2009. Laboratory diagnosis of infections due to blood and tissue parasites. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, 49, pp.1103–1108.

Rosenthal, P.J. et al., 2002. Cysteine proteases of Malaria parasites: targets for chemotherapy. *Current pharmaceutical design*, 8, pp.1659–1672.

Sabnis, Y. a et al., 2003. Probing the structure of falcipain-3, a cysteine protease from Plasmodium falciparum: comparative protein modeling and docking studies. *Protein science : a publication of the Protein Society*, 12, pp.501–509.

Selzer, P.M. et al., 1999. Cysteine protease inhibitors as chemotherapy: lessons from a parasite target. *Proceedings of the National Academy of Sciences of the United States of America*, 96, pp.11015–11022.

Sijwali, P.S. et al., 2006. Gene disruptions demonstrate independent roles for the four falcipain cysteine proteases of Plasmodium falciparum. *Molecular and Biochemical Parasitology*, 150, pp.96–106.

Singh, S. & Sivakumar, R., 2003. Recent advances in the diagnosis of Leishmaniasis. *Journal of Postgraduate Medicine*, 49, p.55.

Smith, T.F. & Waterman, M.S., 1981. Identification of common molecular subsequences. *Journal of molecular biology*, 147, pp.195–197.

Söding, J., Biegert, A. & Lupas, A.N., 2005. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, 33.

Staden, R., 1989. Methods for calculating the probabilities of finding patterns in sequences. *Computer applications in the biosciences : CABIOS*, 5, pp.89–96.

Sundar, S. & Chakravarty, J., 2013. Leishmaniasis: an update of current pharmacotherapy. *Expert opinion on pharmacotherapy*, 14, pp.53–63.

Tamura, K. et al., 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, 28, pp.2731–2739.

Tangpukdee, N. et al., 2009. Malaria diagnosis: a brief review. *The Korean journal of parasitology*, 47, pp.93–102.

Tartaglia, G.G. et al., 2004. The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. *Protein science : a publication of the Protein Society*, 13, pp.1939–1941.

Tatusov, R.L., Koonin, E. V & Lipman, D.J., 1997. A genomic perspective on protein families. *Science (New York, N.Y.)*, 278, pp.631–637.

Taylor, W.R., 1998. Dynamic sequence databank searching with templates and multiple alignment. *Journal of molecular biology*, 280, pp.375–406.

Thompson, J.D., Plewniak, F. & Poch, O., 1999. A comprehensive comparison of multiple sequence alignment programs. *Nucleic acids research*, 27, pp.2682–2690.

Di Tommaso, P. et al., 2011. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res*, 39, pp.W13–7 ST – T–Coffee: a web server for the multiple.

Wang, S.X. et al., 2006. Structural basis for unique mechanisms of folding and hemoglobin binding by a Malarial protease. *Proceedings of the National Academy of Sciences of the United States of America*, 103, pp.11503–11508.

Wegscheid-Gerlach, C., Gerber, H.-D. & Diederich, W.E., 2010. Proteases of Plasmodium falciparum as potential drug targets and inhibitors thereof. *Current topics in medicinal chemistry*, 10, pp.346–367.

Weiss, L.M. & Dubey, J.P., 2009. Toxoplasmosis: A history of clinical observations. *International Journal for Parasitology*, 39, pp.895–901.

Wittner, M. & Tanowitz, H.B., 2000. Leishmaniasis in infants and children. *Seminars in Pediatric Infectious Diseases*, 11, pp.196–201.

World Health Organization, 2013. *World Malaria Report 2013*, Available at: www.who.int.

Xia, X., 2007. *Bioinformatics and the cell: Modern computational approaches in genomics, proteomics and transcriptomics*,

Xiong, J., 2006. Essential Bioinformatics. *Vasa*, 1, p.362.

Ye, J., McGinnis, S. & Madden, T.L., 2006. BLAST: Improvements for better sequence analysis. *Nucleic Acids Research*, 34.

APPENDIX

APPENDIX A: DATA RETRIEVAL, SEQUENCE ALIGNMENT AND PHYLOGENETIC ANALYSIS

I .The summary of the retrieved sequences showing the accession number, E-value, sequence identity and query cover to the query sequence FP-2.

Species name	Accession number	Common name	E-value	% Identity	%query cover	Positives	Gaps	Score
P.falciparum	3BWK_A	Falcipain-3	2.60E-59	68	-	98	-	394
P.falciparum 3D7	2OUL_A	cysteine protease	1.60E-68	98	-	99	-	450
S.scrofa	8PCH_A	Cathepsin H	9.20E-53	40	-	95	-	348
T.gondii RH	3F75_A	Cathepsin L	8.20E-53	40	-	96	-	349
M.musculus (house mouse)	4BS6_A	Cathepsin S	1.40E-54	38	-	94	-	361
Homo sapiens	1BY8_A	Procathepsin K	8.40E-55	40	-	68	-	379
Homo sapiens	2XU3_A	Cathepsin L	6.10E-54	37	-	96	-	356
Homosapiens	AAH02642.1	Cathepsin S	8.00E-41	32	92	49	11	153
Homosapiens	AAL23961.1	Cathepsin H	4.00E-49	35	89	51	12	176
Homo sapiens	NP_000387.1	Cathepsin K	4.00E-47	34	92	53	10	171
Homo sapiens	AAH12612.1	Cathepsin L	3.00E-46	32	66	54	12	168
B.taurus	NP_001029607.1	Cathepsin K	8.00E-49	35	90	52	10	174
B.taurus	NP_001028787.1	Cathepsin S	1.00E-40	32	91	49	11	151

B.taurus	NP_001029557.1	Cathepsin H	8.00E-47	34	89	51	12	169
B.taurus	NP_776457.1	Cathepsin L2	7.00E-47	32	90	53	11	169
M.musculus	AAD32136.1	Cathepsin L	5.00E-45	32	91	52	12	164
M.musculus	NP_031828.2	Cathepsin K	3.00E-50	34	92	52	10	179
M.musculus	AAC05781.1	Cathepsin S	5.00E-41	32	89	50	11	154
M.musculus	AAG28508.1	Cathepsin F	2.00E-33	34	60	48	10	134
M.musculus	AAH06878.1	Cathepsin H	3.00E-50	34	93	52	11	179
G.gorilla	XP_004026652.1	Cathepsin K	1.00E-47	34	65	53	10	170
G.gorilla	XP_004056662.1	Cathepsin H	9.00E-49	34	65	51	12	173
G.gorilla	XP_004051582.1	Cathepsin W	1.00E-35	28	61	49	9	137
P.paniscus	XP_008952375.1	Cathepsin F	4.00E-38	33	66	48	10	139
P.paniscus	XP_003828695.1	Cathepsin W	1.00E-37	28	61	49	9	138
P.paniscus	XP_003820654.1	Cathepsin L2	5.00E-47	33	65	51	10	163
P.paniscus	XP_003822607.1	Cathepsin O	6.00E-34	34	53	50	14	126
P.troglodytes	XP_001170363.1	Cathepsin W	2.00E-35	28	61	49	9	137
P.troglodytes	XP_517502.2	Cathepsin O	6.00E-32	34	53	50	14	126
P.troglodytes	XP_003312197.1	Cathepsin L1	1.00E-46	32	65	54	12	168
T.gondii	ABY58967.1	TgCathepsinL	2.00E-59	35	75	52	9	205

L.aethiopica	AEE42617.1	cysteine protease	1.00E-50	35	64	53	8	181
L.mexicana	CAA78443.1	cysteine protease	6.00E-43	32	61	51	8	160
L.major	AFN27092.1	cysteine protease	5.00E-49	34	81	52	8	175
L.tropica	AFN27126.1	cysteine protease	3.00E-49	35	64	52	8	178
T.brucei	CAC67416.1	Rhodesiense	4.00E-43	34	61	50	9	161
T.cruzi	AAB41118.1	Cruzipain	2.00E-44	34	61	50	10	164
P.berghei ANKA	XP_680416.1	Berghepain-2	2.00E-135	45	100	64	3	402
P.yoelii yoelii 17XNL	XP_726900.1	Yoellipain-2	1.00E-126	44	100	63	4	385
P.chabaudi chabaudi	AAP43630.1	Chabaupain-2	1.00E-131	45	100	63	3	403
P.vivax	AAT36263.1	Vivapain-2	1.00E-171	50	100	69	1	502
P.knowlesi strain H	XP_002259152.1	Knowlesipain-2	2.00E-157	45	99	66	5	468
P.falciparum 3D7	XP_001347836.1	Falcipain-2	0	100	100	100	0	990
P.falciparum 3D7	XP_001347832.1	Falcipain-2B	0	92	100	95	0	905
P.falciparum 3D7	XP_001347833.1	Falcipain-3	0	55	100	70	1	538
P.falciparum 3D7	XP_001348727.1	Falcipain-1	4.00E-60	35	62	57	8	211
E.histolytica HM-1:IMSS	XP_653254.1	cysteine protease	2.00E-36	30	65	45	12	137
S.scrofa	NP_999057.1	Cathepsin L1	6.00E-47	32	65	53	10	168
S.scrofa	NP_999467.1	Cathepsin K	3.00E-48	34	65	53	10	172

S.scrofa	XP_005663551.1	Cathepsin S	4.00E-39	32	65	49	11	146
S.scrofa	ACB59246.1	Cathepsin H	1.00E-45	35	62	52	13	163
G.gallus	NP_001161481.1	Cathepsin L1	3.00E-49	34	66	56	13	175
G.gallus	NP_990302.1	Cathepsin K	1.00E-43	35	56	51	12	159
G.gallus	NP_001026516.1	Cathepsin S	1.00E-42	30	66	52	12	156
G.gallus	AEC13302.1	Cathepsin H	8.00E-44	33	67	51	11	159
G.gallus	NP_001026300.1	Cathespsin O	4.00E-30	30	57	48	13	120
F.catus	XP_003995514.1	Cathepsin L1	2.00E-44	33	66	50	10	161
F.catus	XP_00399069.2	Cathepsin S	9.00E-33	32	53	49	9	127
F.catus	XP_006937669.1	Cathepsin F	2.00E-33	35	60	50	11	131
F.catus	XP_003984969.1	Cathespsin O	2.00E-31	33	53	50	13	125
A.darlingi	ETN64675.1	Cathepsin I	7.00E-56	34	66	56	7	191
G.morsitans	ABC48936.1	Cathepsin F	6.00E-45	32	61	53	8	159

II. A summary of FP-2 and its retrieved orthologs from the NCBI database. The accession numbers, sequence length and the prodomain and mature domain regions are also indicated.

Species name	**Accession number**	**Common name**	**Number of amino acids**		
			Total	**Prodomain**	**Mature domain**
P.falciparum	3BWK_A	Falcipain-3	243		1-243
P.falciparum	2OUL_A	cysteine	241		1-241

3D7		protease			
S.scrofa	8PCH_A	Cathepsin H	220		1-220
T.gondii RH	3F75_A	Cathepsin L	224		99-330
M.musculus (house mouse)	4BS6_A	Cathepsin S	225		1-225
Homo sapiens	1BY8_A	Procathepsin K	314		88-314
Homo sapiens	2XU3_A	Cathepsin L	220		1-241
Homo sapiens	AAH02642.1	Cathepsin S	331	1-101	102-331
Homo sapiens	AAL23961.1	Cathepsin H	335	1-103	104-335
Homo sapiens	NP_000387.1	Cathepsin K	329	1-102	103-329
Homo sapiens	AAH12612.1	Cathepsin L	333	1-101	102-333
B.taurus	NP_001029607.1	Cathepsin K	334	1-107	108-334
B.taurus	NP_001028787.1	Cathepsin S	331	1-101	102-331
B.taurus	NP_001029557.1	Cathepsin H	335	1-104	105-336
B.taurus	NP_776457.1	Cathepsin L2	334	1-101	102-334
M.musculus	AAD32136.1	Cathepsin L	334	1-101	102-334
M.musculus	NP_031828.2	Cathepsin K	329	1-102	103-329
M.musculus	AAC05781.1	Cathepsin S	340	1-109	110-340
M.musculus	AAG28508.1	Cathepsin F	462	1-235	236-462
M.musculus	AAH06878.1	Cathepsin H	333	1-101	102-333
G.gorilla	XP_004026652.1	Cathepsin K	329	1-102	103-329
G.gorilla	XP_004056662.1	Cathepsin H	335	1-103	104-33
G.gorilla	XP_004051582.1	Cathepsin W	376	1-115	116-376
P.paniscus	XP_008952375.1	Cathepsin F	368	1-141	142-368
P.paniscus	XP_003828695.1	Cathepsin W	376	1-115	116-376
P.paniscus	XP_003820654.1	Cathepsin L2	334	1-101	102-334
P.paniscus	XP_003822607.1	Cathepsin O	321	1-93	94-321
P.troglodytes	XP_001170363.1	Cathepsin W	376	1-115	116-376
P.troglodytes	XP_517502.2	Cathepsin O	318	1-90	91-318
P.troglodytes	XP_003312197.1	Cathepsin L1	278	47-278	47-278
T.gondii	ABY58967.1	TgCathepsinL	421	1-189	190-421
L.aethiopica	AEE42617.1	cysteine protease	443	1-111	112-443
L.mexicana	CAA78443.1	cysteine protease	443	1-111	112-443

L.major	AFN27092.1	cysteine protease	348	1-111	112-348
L.tropica	AFN27126.1	cysteine protease	443	1-111	112-443
T.brucei	CAC67416.1	Rhodesiense	450	1-113	114-450
T.cruzi	AAB41118.1	Cruzipain	383	1-110	111-383
P.berghei ANKA	XP_680416.1	Berghepain-2	470	1-229	230-470
P.yoelii yoelii 17XNL	XP_726900.1	Yoellipain-2	472	1-231	232-472
P.chabaudi chabaudi	AAP43630.1	Chabaupain-2	471	1-230	231-471
P.vivax	AAT36263.1	Vivapain-2	487	1-244	245-487
P.knowlesi strain H	XP_002259152.1	Knowlesipain-2	495	1-250	251-495
P.falciparum 3D7	XP_001347836.1	Falcipain-2	487	1-241	242-484
P.falciparum 3D7	XP_001347832.1	Falcipain-2B	482	1-239	240-482
P.falciparum 3D7	XP_001347833.1	Falcipain-3	492	1-248	249-492
P.falciparum 3D7	XP_001348727.1	Falcipain-1	569	1-316	317-569
E.histolytica HM-1:IMSS	XP_653254.1	cysteine protease	308	1-81	82-308
S.scrofa	NP_999057.1	Cathepsin L1	334	102-334	102-334
S.scrofa	NP_999467.1	Cathepsin K	330	104-330	104-330
S.scrofa	XP_005663551.1	Cathepsin S	331	102-331	102-331
S.scrofa	ACB59246.1	Cathepsin H	297	63-297	63-297
G.gallus	NP_001161481.1	Cathepsin L1	353	121-353	121-353
G.gallus	NP_990302.1	Cathepsin K	334	108-334	108-334
G.gallus	NP_001026516.1	Cathepsin S	353	1-125	126-353
G.gallus	AEC13302.1	Cathepsin H	329	98-329	98-329
G.gallus	NP_001026300.1	Cathespsin O	321	96-321	96-321
F.catus	XP_003995514.1	Cathepsin L1	333	102-333	102-333
F.catus	XP_00399069.2	Cathepsin S	312	115-312	115-312

F.catus	XP_006937669.1	Cathepsin F	461	235-461	235-461
F.catus	XP_003984969.1	Cathespsin O	390	163-390	163-390
A.darlingi	ETN64675.1	Cathepsin I	344	111-344	111-344
G.morsitans	ABC48936.1	Cathepsin F	471	237-471	237-471

III. The retrieved sequences through reverse BLAST showing the first and second hits as returned by NCBI. The table indicates the protein sequence accession numbers, as well as the sequence identities of FP-2 and FP-3 are also indicated.

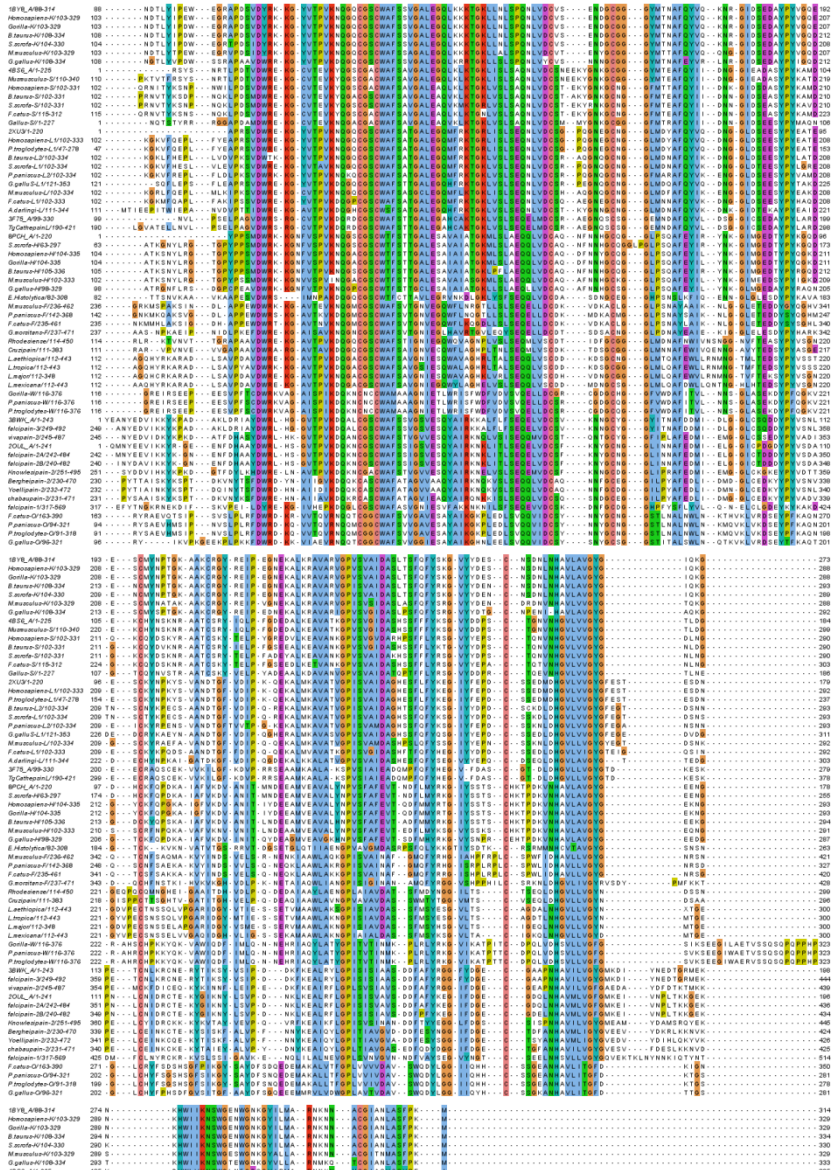
Species name	**Accession number**	**Common name**	**REVERSE BLAST**		**% Identity of sequences**	
			1st hit	**2nd hit**	**FP-2**	**FP-3**
Homosapiens	AAH02642.1	Cathepsin S	FP-3	FP-2	32	33
Homosapiens	AAL23961.1	Cathepsin H	FP-2	FP-2B	35	32
Homo sapiens	NP_000387.1	Cathepsin K	FP-3	FP-2	34	35
Homo sapiens	AAH12612.1	Cathepsin L	FP-3	FP-2B	32	34
B.taurus	NP_001029607.1	Cathepsin K	FP-2	FP-2B	35	35
B.taurus	NP_001028787.1	Cathepsin S	FP-3	FP-2	32	34
B.taurus	NP_001029557.1	Cathepsin H	FP-2	FP-2B	34	34
B.taurus	NP_776457.1	Cathepsin L2	FP-3	FP-2B	32	34
M.musculus	AAD32136.1	Cathepsin L	FP-2B	FP-2	32	32
M.musculus	NP_031828.2	Cathepsin K	FP-2	FP-2B	34	36
M.musculus	AAC05781.1	Cathepsin S	FP-3	FP-2B	32	34
M.musculus	AAG28508.1	Cathepsin F	FP-3	FP-2B	34	36

M.musculus	AAH06878.1	Cathepsin H	FP-2	FP-2B	34	33
G.gorilla	XP_004026652.1	Cathepsin K	FP-3	FP-2	34	35
G.gorilla	XP_004056662.1	Cathepsin H	FP-2	FP-2B	34	32
G.gorilla	XP_004051582.1	Cathepsin W	FP-2	FP-3	28	28
P.paniscus	XP_008952375.1	Cathepsin F	FP-3	FP-2B	33	34
P.paniscus	XP_003828695.1	Cathepsin W	FP-2	FP-3	29	28
P.paniscus	XP_003820654.1	Cathepsin L2	FP-2B	FP-3	33	33
P.paniscus	XP_003822607.1	Cathepsin O	FP-2	FP-2B	34	32
P.troglodytes	XP_001170363.1	Cathepsin W	FP-3	FP-2	29	28
P.troglodytes	XP_517502.2	Cathepsin O	FP-2	FP-2B	34	32
P.troglodytes	XP_003312197.1	Cathepsin L1	FP-2B	FP-3	34	35
T.gondii	ABY58967.1	TgCathepsinL	FP-2B	FP-2	35	38
L.aethiopica	AEE42617.1	cysteine protease	FP-3	FP-2	35	38
L.mexicana	CAA78443.1	cysteine protease	FP-3	FP-2	32	38
L.major	AFN27092.1	cysteine protease	FP-3	FP-2	34	37
L.tropica	AFN27126.1	cysteine protease	FP-3	FP-2	35	38
T.brucei	CAC67416.1	Rhodesiense	FP-2	FP-2B	34	35
T.cruzi	AAB41118.1	Cruzipain	FP-2	FP-2B	34	34
P.berghei ANKA	XP_680416.1	Berghepain-2	FP-3	FP-2	45	43
P.yoelii yoelii 17XNL	XP_726900.1	Yoellipain-2	FP-3	FP-2	44	44
P.chabaudi chabaudi	AAP43630.1	Chabaupain-2	FP-3	FP-2	45	43

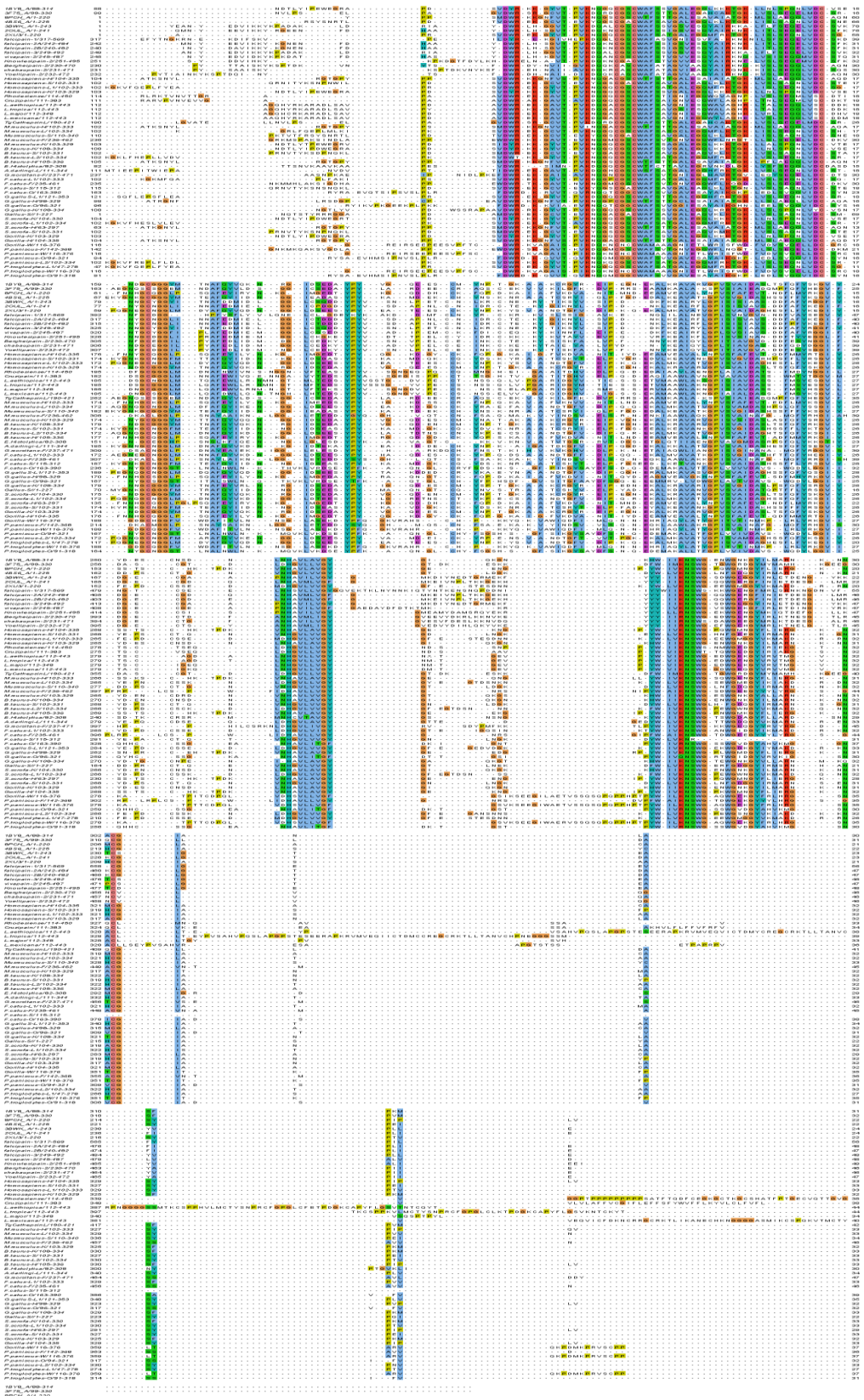
P.vivax	AAT36263.1	Vivapain-2	FP-3	FP-2	50	55
P.knowlesi strain H	XP_002259152.1	Knowlesipain-2	FP-3	FP-2	45	50
P.falciparum 3D7	XP_001347836.1	Falcipain-2	FP-2	FP-2B	100	55
P.falciparum 3D7	XP_001347832.1	Falcipain-2B	FP-2B	FP-2	93	54
P.falciparum 3D7	XP_001347833.1	Falcipain-3	FP-3	FP-2	55	100
P.falciparum 3D7	XP_001348727.1	Falcipain-1	FP-2B	FP-2	35	36
E.histolytica HM-1:IMSS	XP_653254.1	cysteine protease	FP-2B	FP-2	30	30
S.scrofa	NP_999057.1	Cathepsin L1	FP-2B	FP-3	32	34
S.scrofa	NP_999467.1	Cathepsin K	FP-3	FP-2	35	36
S.scrofa	XP_005663551.1	Cathepsin S	FP-3	FP-2B	32	34
S.scrofa	ACB59246.1	Cathepsin H	FP-2	FP-2B	35	33
G.gallus	NP_001161481.1	Cathepsin L1	FP-3	FP-2B	34	35
G.gallus	NP_990302.1	Cathepsin K	FP-3	FP-2B	35	38
G.gallus	NP_001026516.1	Cathepsin S	FP-3	FP-2B	30	34
G.gallus	AEC13302.1	Cathepsin H	FP-2	FP-2B	33	33
G.gallus	NP_001026300.1	Cathespsin O	FP-2B	FP-2	30	29
F.catus	XP_003995514.1	Cathepsin L1	FP-2B	FP-3	33	35
F.catus	XP_00399069.2	Cathepsin S	FP-3	FP-2B	31	35
F.catus	XP_006937669.1	Cathepsin F	FP-3	FP-2B	35	35
F.catus	XP_003984969.1	Cathespsin O	FP-2B	FP-2	33	30

A.darlingi	ETN64675.1	Cathepsin I	FP-2	FP-2B	34	35
G.morsitans	ABC48936.1	Cathepsin F	FP-3	FP-2B	32	33

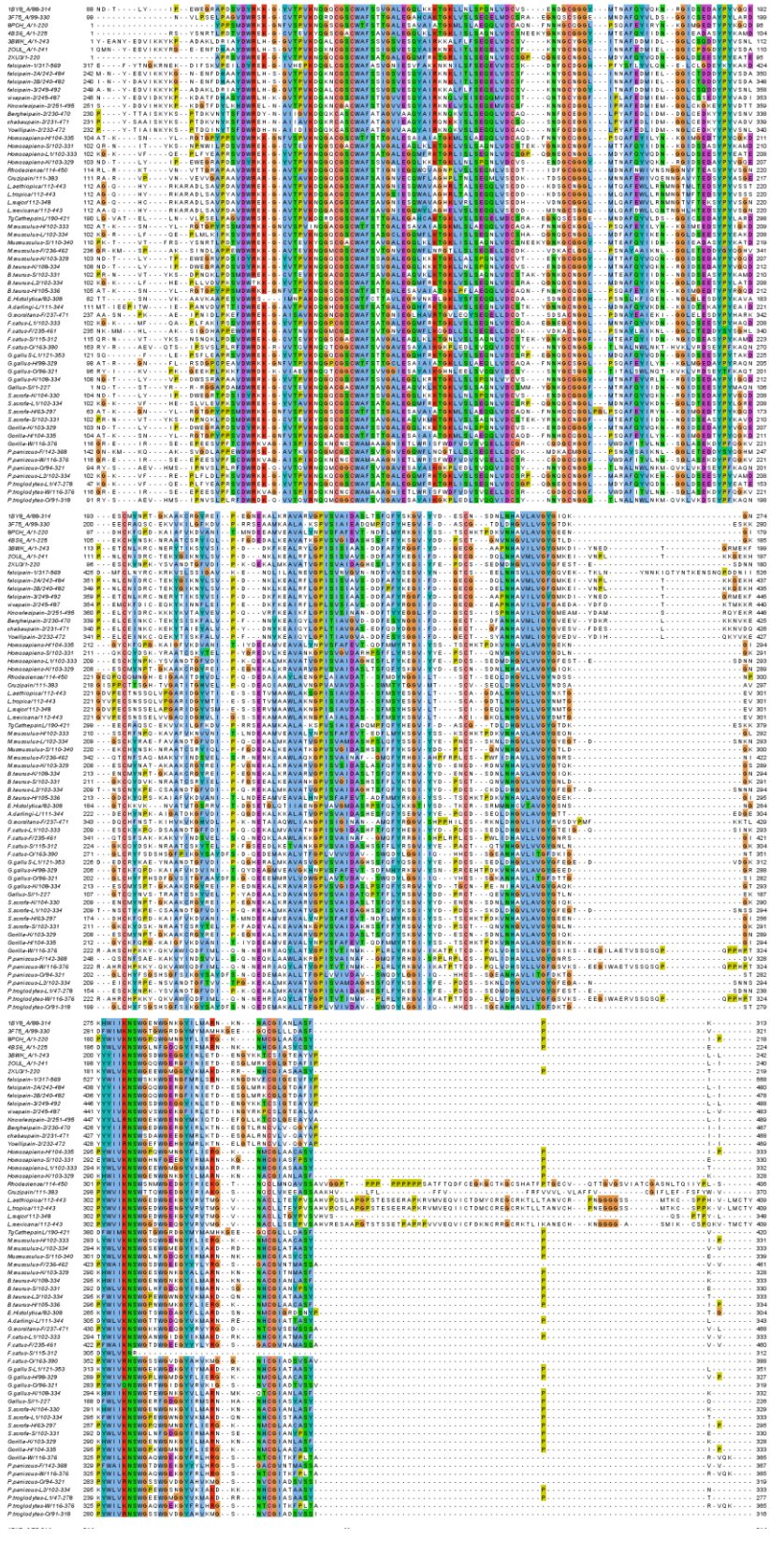
IV. MAFFT multiple sequence alignment which was selected as the best alignment



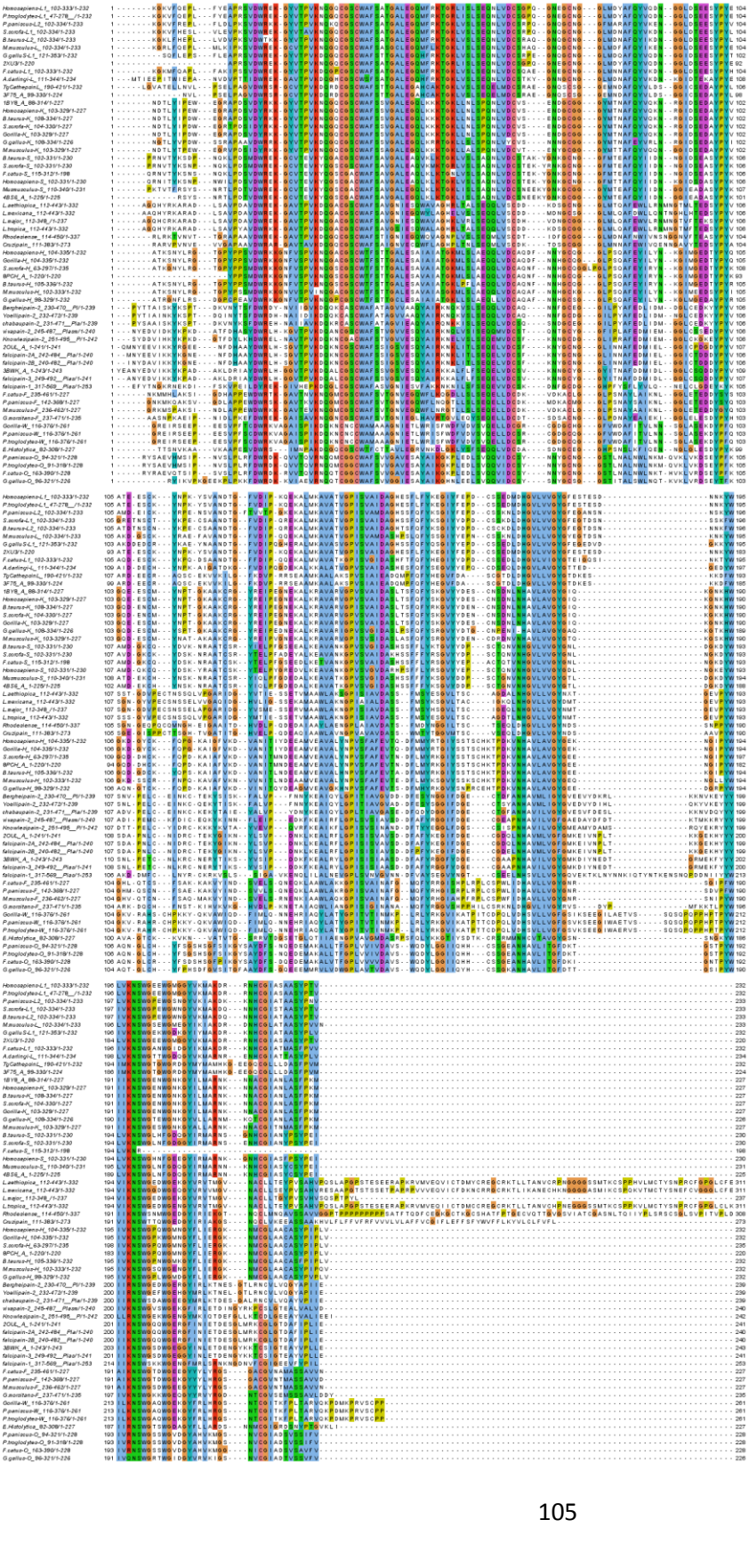
V. EXPRESSO multiple sequence alignment of the catalytic domain



VI. T-COFFEE multiple sequence alignment of the catalytic domain



VII. PROMALS3D multiple sequence alignment of the catalytic domain



VIII. The residues corresponding to each of the sub-sites.

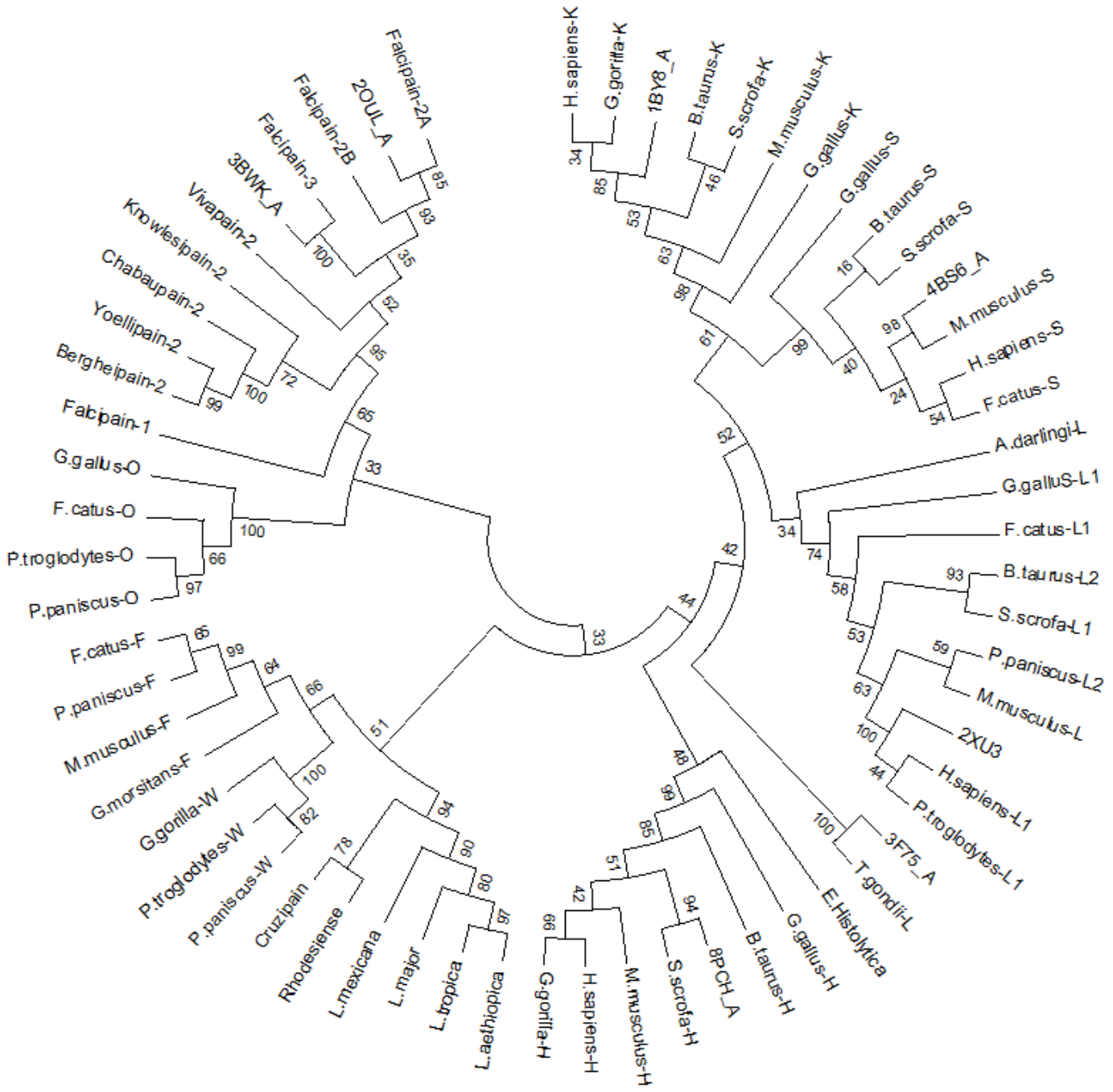
Species name	Accession number	Common name	Sub-site 1				Sub-site 1'					
			279	282	284	288	279	395	416	418	419	448
P.falciparum	3BWK_A	Falcipain-3	Q	C	S	F	A	N	A	V	S	W
P.falciparum 3D7	2OUL_A	cysteine protease	Q	C	S	F	V	N	A	V	S	W
S.scrofa	8PCH_A	Cathepsin H	Q	C	S	F	V	N	A	V	S	W
T.gondii RH	3F75_A	Cathepsin L	Q	C	S	F	A	D	G	V	S	W
M.musculus (house mouse)	4BS6_A	Cathepsin S	Q	C	A	F	A	N	G	V	S	W
Homo sapiens	1BY8_A	Procathepsin K	Q	C	S	F	A	N	A	V	S	W
Homo sapiens	2XU3_A	Cathepsin L	Q	C	S	F	A	N	G	V	S	W
Homosapiens	AAH02642.1	Cathepsin S	Q	C	S	F	A	N	A	V	S	W
Homosapiens	AAL23961.1	Cathepsin H	Q	C	S	F	A	N	A	V	S	W
Homo sapiens	NP_000387.1	Cathepsin K	Q	C	A	F	V	N	A	V	S	W
Homo sapiens	AAH12612.1	Cathepsin L	Q	C	S	F	A	D	G	V	S	W
B.taurus	NP_001029607.1	Cathepsin K	Q	C	S	F	A	D	G	V	S	W
B.taurus	NP_001028787.1	Cathepsin S	Q	C	S	F	V	N	A	V	S	W
B.taurus	NP_001029557.1	Cathepsin H	Q	C	S	F	A	D	G	V	S	W
B.taurus	NP_776457.1	Cathepsin L2	Q	C	S	F	A	N	G	V	S	W
M.musculus	AAD32136.1	Cathepsin L	Q	C	S	F	A	D	G	V	S	W
M.musculus	NP_031828.2	Cathepsin K	Q	C	S	F	A	N	A	V	S	W
M.musculus	AAC05781.1	Cathepsin S	Q	C	A	F	A	N	G	V	S	W
M.musculus	AAG28508.1	Cathepsin F	Q	C	S	F	A	D	A	V	S	W
M.musculus	AAH06878.1	Cathepsin H	Q	C	S	F	A	N	A	V	S	W
G.gorilla	XP_004026652.1	Cathepsin K	Q	C	S	F	M	D	S	V	S	W
G.gorilla	XP_004056662.1	Cathepsin H	Q	C	S	F	M	D	S	V	S	W
G.gorilla	XP_004051582.1	Cathepsin W	Q	C	C	M	A	N	A	V	S	W
P.paniscus	XP_008952375.1	Cathepsin F	Q	C	S	F	A	N	A	V	S	W
P.paniscus	XP_003828695.1	Cathepsin W	Q	C	C	M	V	N	A	V	S	W
P.paniscus	XP_003820654.1	Cathepsin L2	Q	C	S	F	V	N	A	V	S	W
P.paniscus	XP_003822607.1	Cathepsin O	Q	C	S	F	A	D	A	V	S	W
P.troglodytes	XP_001170363.1	Cathepsin W	Q	C	S	F	A	D	G	V	S	W
P.troglodytes	XP_517502.2	Cathepsin O	Q	C	S	F	M	D	S	V	S	W
P.troglodytes	XP_003312197.1	Cathepsin L1	Q	C	S	F	A	D	G	V	S	W
T.gondii	ABY58967.1	TgCathepsinL	Q	C	S	F	A	D	G	V	S	W
L.aethiopica	AEE42617.1	cysteine protease	Q	C	G	F	A	N	G	V	S	W
L.mexicana	CAA78443.1	cysteine protease	Q	C	S	F	A	N	G	V	S	W
L.major	AFN27092.1	cysteine protease	Q	C	S	F	A	D	G	V	S	W
L.tropica	AFN27126.1	cysteine protease	Q	C	S	F	V	N	A	V	S	W
T.brucei	CAC67416.1	Rhodesiense	Q	C	C	M	V	N	A	V	S	W
T.cruzi	AAB41118.1	Cruzipain	Q	C	S	F	A	N	A	V	S	W
P.berghei ANKA	XP_680416.1	Berghepain-2	Q	C	S	F	A	N	G	V	S	W
P.yoelii yoelii 17XNL	XP_726900.1	Yoellipain-2	Q	C	S	F	A	N	G	V	S	W
P.chabaudi chabaudi	AAP43630.1	Chabaupain-2	Q	C	S	F	A	D	G	V	S	W
P.vivax	AAT36263.1	Vivapain-2	Q	C	A	F	V	N	A	V	S	W
P.knowlesi strain H	XP_002259152.1	Knowlesipain-2	Q	C	S	F	A	N	A	V	S	W
P.falciparum 3D7	XP_001347836.1	Falcipain-2	Q	C	S	F	V	N	A	V	S	W

P.falciparum 3D7	XP_001347832.1	Falcipain-2B	Q	C	S	F	V	N	A	V	S	W
P.falciparum 3D7	XP_001347833.1	Falcipain-3	Q	C	S	F	V	N	S	V	S	W
P.falciparum 3D7	XP_001348727.1	Falcipain-1	Q	C	S	F	A	D	G	V	S	W
E.histolytica HM-1:IMSS	XP_653254.1	cysteine protease	Q	C	S	F	A	N	G	V	S	W
S.scrofa	NP_999057.1	Cathepsin L1	Q	C	S	F	A	D	G	V	S	W
S.scrofa	NP_999467.1	Cathepsin K	Q	C	S	F	A	N	G	V	S	W
S.scrofa	XP_005663551.1	Cathepsin S	Q	C	S	F	A	N	A	V	S	W
S.scrofa	ACB59246.1	Cathepsin H	Q	C	S	F	A	N	C	V	S	W
G.gallus	NP_001161481.1	Cathepsin L1	Q	C	S	F	A	N	A	V	S	W
G.gallus	NP_990302.1	Cathepsin K	Q	C	G	F	V	N	A	V	S	W
G.gallus	NP_001026516.1	Cathepsin S	Q	C	G	F						
G.gallus	AEC13302.1	Cathepsin H	Q	C	S	F	V	N	A	V	S	W
G.gallus	NP_001026300.1	Cathespsin O	Q	C	S	F	A	N	A	V	S	W
F.catus	XP_003995514.1	Cathepsin L1	Q	C	S	F	A	D	G	V	S	W
F.catus	XP_00399069.2	Cathepsin S	Q	C	S	F	A	N	G	V	S	W
F.catus	XP_006937669.1	Cathepsin F	Q	C	S	F	A	D	A	V	S	W
F.catus	XP_003984969.1	Cathespsin O	Q	C	A	F	A	N	A	V	S	W
A.darlingi	ETN64675.1	Cathepsin I	Q	C	A	F	I	D	G	V	S	W
G.morsitans	ABC48936.1	Cathepsin F	Q	C	S	F	A	D	G	V	S	W

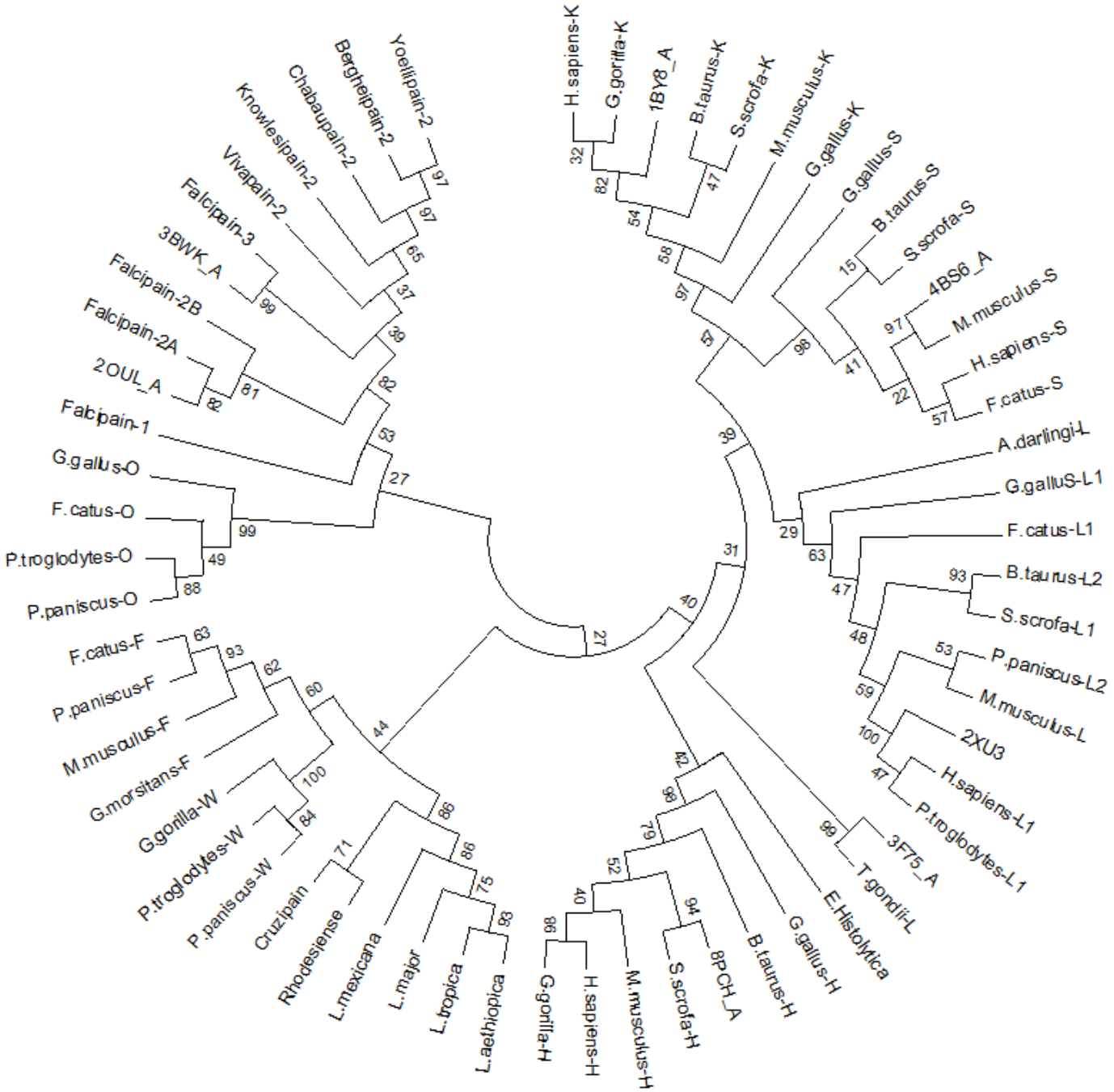
Species name	**Accession number**	**Common name**	**Sub-site2**							**Sub-site3**			
			286	321	325	327	328	329	392	415	477	478	479
P.falciparum	3BWK_A	Falcipain-3	W	N	G	Y	I	T	S	P	N	A	F
P.falciparum 3D7	2OUL_A	cysteine protease	W	Y	G	L	I	N	S	L	N	A	F
S.scrofa	8PCH_A	Cathepsin H	W	A	G	L	P	S	A	V	Q	A	F
T.gondii RH	3F75_A	Cathepsin L	W	Q	G	E	M	N	A	L	D	A	F
M.musculus (house mouse)	4BS6_A	Cathepsin S	W	K	G	Y	M	T	G	V	E	A	F
Homo sapiens	1BY8_A	Procathepsin K	W	D	G	Y	M	T	A	L	N	A	F
Homo sapiens	2XU3_A	Cathepsin L	W	E	G	L	M	D	A	M	Y	A	F
Homosapiens	AAH02642.1	Cathepsin S	W	K	G	F	M	T	G	V	T	A	F
Homosapiens	AAL23961.1	Cathepsin H	W		G	L	P	S	A	V	Q	A	F
Homo sapiens	NP_000387.1	Cathepsin K	W	D	G	Y	M	T	A	L	N	A	F
Homo sapiens	AAH12612.1	Cathepsin L	W	E	G	L	M	D	A	M	Y	A	F
B.taurus	NP_001029607.1	Cathepsin K	W	D	G	Y	M	T	A	V	Q	A	F
B.taurus	NP_001028787.1	Cathepsin S	W	K	G	F	M	T	G	V	T	A	F
B.taurus	NP_001029557.1	Cathepsin H	W	H	G	L	P	S	A	L	N	A	F
B.taurus	NP_776457.1	Cathepsin L2	W	Q	G	L	M	D	A	L	N	A	F
M.musculus	AAD32136.1	Cathepsin L	W	Q	G	L	M	D	A	V	E	A	F
M.musculus	NP_031828.2	Cathepsin K	W	Y	G	Y	M	T	S	V	Q	A	F
M.musculus	AAC05781.1	Cathepsin S	W	K	G	Y	M	T	G	V	E	A	F
M.musculus	AAG28508.1	Cathepsin F	W	K	G	L	P	S	A	L	F	A	F
M.musculus	AAH06878.1	Cathepsin H	W	H	G	L	P	S	A	I	N	A	Y
G.gorilla	XP_004026652.1	Cathepsin K	W	D	G	Y	M	T	A	L	N	A	F
G.gorilla	XP_004056662.1	Cathepsin H	W	Y	G	L	P	S	A	V	Q	A	F
G.gorilla	XP_004051582.1	Cathepsin W	W	D	G	F	V	W	I	V	D	A	F

P.paniscus	XP_008952375.1	Cathepsin F	W	K	G	L	P	S	A	A	N	A	L
P.paniscus	XP_003828695.1	Cathepsin W	W	D	G	F	V	W	T	L	A	F	L
P.paniscus	XP_003820654.1	Cathepsin L2	W	Q	G	F	M	A	A	L	Q	A	F
P.paniscus	XP_003822607.1	Cathepsin O	W	Y	G	S	T	L	I	I	N	A	Y
P.troglodytes	XP_001170363.1	Cathepsin W	W	D	G	F	V	W	I	V	D	A	F
P.troglodytes	XP_517502.2	Cathepsin O	W	Y	G	S	T	L	I	L	Q	A	F
P.troglodytes	XP_003312197.1	Cathepsin L1	W	E	G	L	M	D	A	L	N	A	F
T.gondii	ABY58967.1	TgCathepsinL	W	Q	G	E	M	N	A	V	D	A	F
L.aethiopica	AEE42617.1	cysteine protease	W	S	G	L	M	T	A	A	N	A	L
L.mexicana	CAA78443.1	cysteine protease	W	N	G	L	M	L	A	L	Q	A	F
L.major	AFN27092.1	cysteine protease	W	N	G	L	M	L	A	L	N	A	F
L.tropica	AFN27126.1	cysteine protease	W	S	G	L	M	L	A	L	R	A	F
T.brucei	CAC67416.1	Rhodesiense	W	F	G	L	M	D	A	M	Y	A	F
T.cruzi	AAB41118.1	Cruzipain	W	S	G	L	M	N	A	L	Q	A	F
P.berghei ANKA	XP_680416.1	Berghepain-2	W	F	G	I	L	P	A	A	Y	A	F
P.yoelii yoelii 17XNL	XP_726900.1	Yoellipain-2	W	F	G	I	L	P	A	A	Y	A	F
P.chabaudi chabaudi	AAP43630.1	Chabaupain-2	W	D	G	I	L	P	A	P	L	A	F
P.vivax	AAT36263.1	Vivapain-2	W	T	G	F	I	P	S	A	Y	A	F
P.knowlesi strain H	XP_002259152.1	Knowlesipain-2	W	N	G	L	I	P	S	P	R	A	F
P.falciparum 3D7	XP_001347836.1	Falcipain-2	W	Y	G	L	I	N	S	L	N	A	F
P.falciparum 3D7	XP_001347832.1	Falcipain-2B	W	Y	G	L	I	N	S	L	N	A	F
P.falciparum 3D7	XP_001347833.1	Falcipain-3	W	N	G	Y	I	T	S	P	N	A	F
P.falciparum 3D7	XP_001348727.1	Falcipain-1	W	F	G	H	P	F	N	L	N	A	F
E.histolytica HM-1:IMSS	XP_653254.1	cysteine protease	W	N	G	H	P	S	G	V	Q	A	F
S.scrofa	NP_999057.1	Cathepsin L1	W	Q	G	L	M	D	A	L	Y	S	F
S.scrofa	NP_999467.1	Cathepsin K	W	D	G	Y	M	T	A	L	N	A	F
S.scrofa	XP_005663551.1	Cathepsin S	W	K	G	F	M	T	A	L	Q	A	F
S.scrofa	ACB59246.1	Cathepsin H	W	H	G	L	P	S	A	V	R	A	F
G.gallus	NP_001161481.1	Cathepsin L1	W	Q	G	L	M	D	A	M	N	S	L
G.gallus	NP_990302.1	Cathepsin K	W	N	G	Y	M	T	G	V	E	A	F
G.gallus	NP_001026516.1	Cathepsin S	W	K	G	F	M	T	A	I	N	A	F
G.gallus	AEC13302.1	Cathepsin H	W	H	G	L	P	S	A	V	Q	A	F
G.gallus	NP_001026300.1	Cathespsin O	W	Y	G	S	T	I	T	A	T	A	L
F.catus	XP_003995514.1	Cathepsin L1	W	E	G	L	M	N	G	L	N	A	F
F.catus	XP_00399069.2	Cathepsin S	W	K	G	F	M	T	A	V	E	A	F
F.catus	XP_006937669.1	Cathepsin F	W	K	G	L	P	S	A	I	N	A	Y
F.catus	XP_003984969.1	Cathespsin O	W	Y	G	S	T	L	V	A	N	A	L
A.darlingi	ETN64675.1	Cathepsin I	W	N	G	L	M	D	A	L	N	A	F
G.morsitans	ABC48936.1	Cathepsin F	W	S	G	L	P	D	G	L	N	A	Y

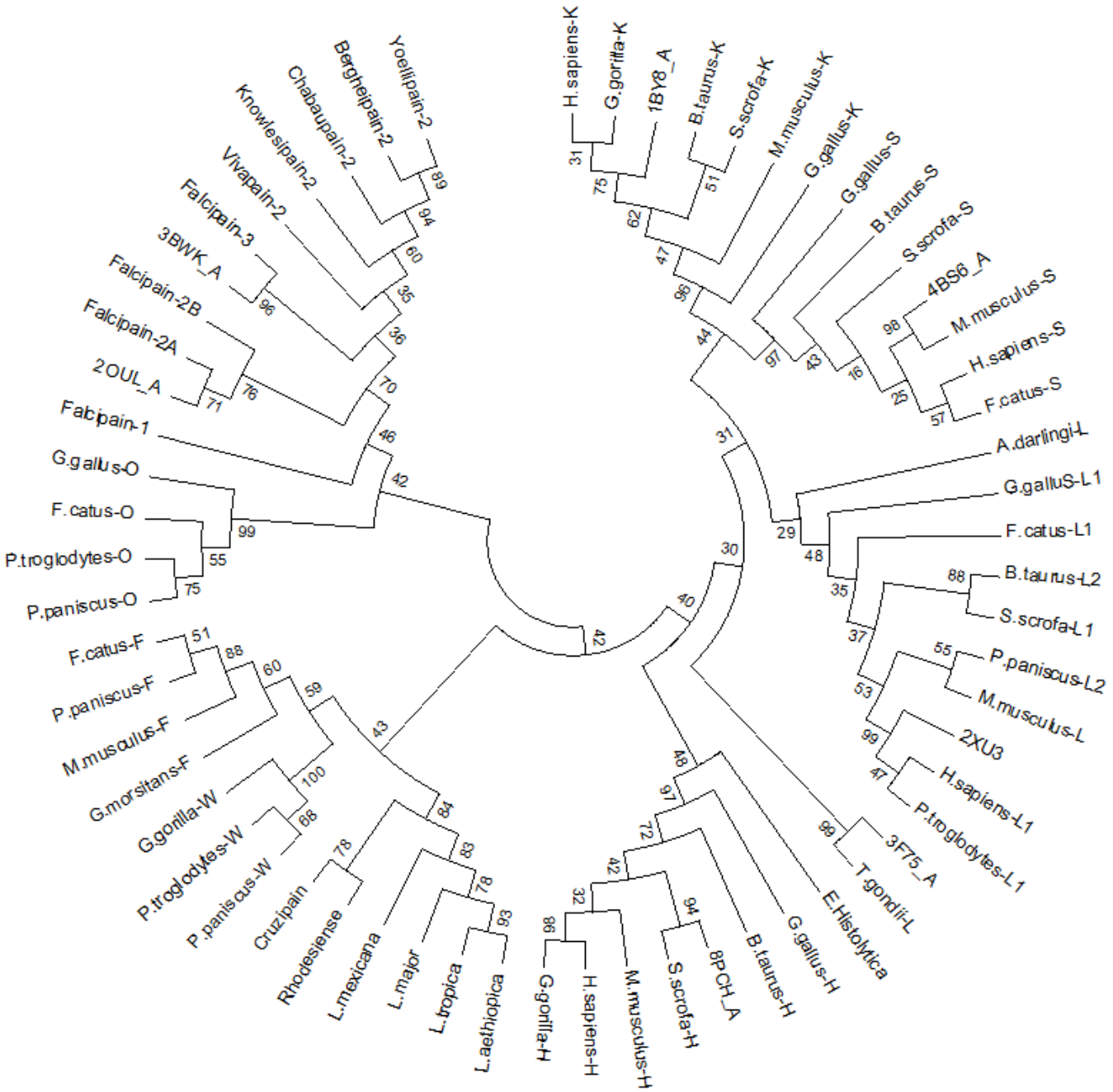
IX. Phylogenetic tree generated from the MAFFT alignment using the JTT+G substitution model.



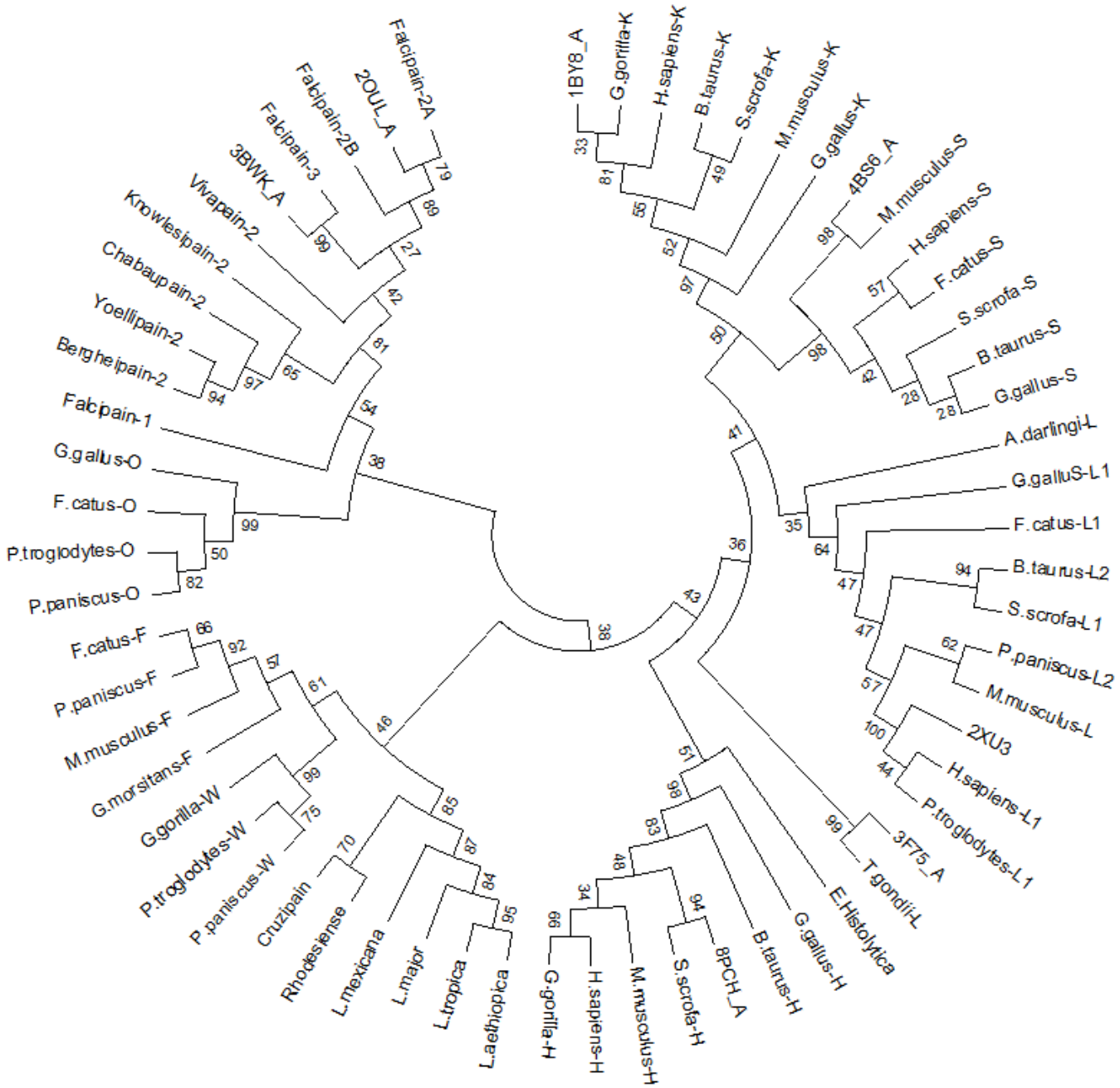
X. Phylogenetic tree generated from the MAFFT alignment using the JTT+G+I substitution model.



XI. Phylogenetic tree generated from the MAFFT alignment using the rtREV+G+I substitution model.

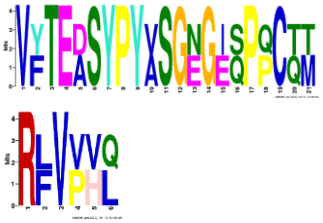
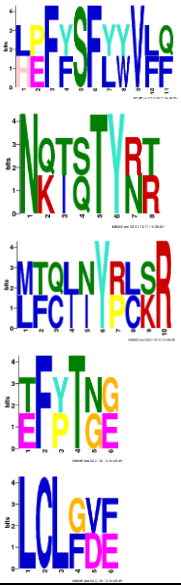
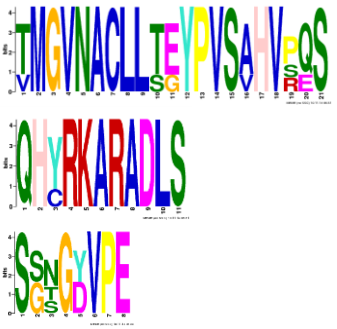


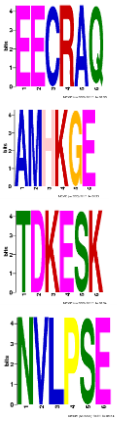
XI. Phylogenetic tree generated from the MAFFT alignment using the WAG+G+I substitution model.



APPENDIX B: MOTIF AND PHYSICOCHEMICAL ANALYSIS

I. A summary of the species unique motifs according to the MEME HTML output.

Motif number	**Species name**	**Sequence position**	**Regular expression**
Motif 41 and 64	*Trypanosoma*	95 – 115 , 236 - 337	
Motif 43,53,56-58	Falcipain-1	83 – 93, 191 - 198, 110 - 119, 1 - 6, 96 -101	
Motif 18, 23 and 52	*Leishmania*	212 – 232, 3 – 13, 107 - 114	

Motif 37, 38, 45 and 48	*T,gondii*	110-115, 212 – 217, 184 – 189, 9 -14	

II. The script used to calculate physicochemical properties

```
#Aromaticity counter for protein sequences within a fasta file
#Sipho Moyo (g10M3876)
#29 September 2014

from Bio.Seq import Seq
from Bio.SeqUtils.ProtParam import ProteinAnalysis
from Bio.SeqUtils import ProtParamData
from Bio import SeqIO

# Entering input files
prot_seqs = raw_input("Enter protein sequences file name: ")
sequences=open(prot_seqs, "r")

#writing into a file
aromaticity_data=open("aromaticity_"+ prot_seqs,"w")
first_line = "Name" + '\t'+ '\t'+ '\t' + " \t aromaticity" + '\n' + '\n'
aromaticity_data.write(first_line)

for files in list(SeqIO.parse(sequences, "fasta")):
    myprot = ProteinAnalysis(str(files.seq))
    aromaticity = myprot.aromaticity()
    seq_ID = files. id #gives the sequence ID
    line = seq_ID + '\t' + '\t' + (str(aromaticity)[1:-1])+ '\n'
    aromaticity_data.write(line) #writes the amino acid composition and corresponding sequence ID
```

```
#Gravy counter for protein sequences within a fasta file
#Sipho Moyo (g10M3876)
#29 September 2014

from Bio.Seq import Seq
from Bio.SeqUtils.ProtParam import ProteinAnalysis
from Bio.SeqUtils import ProtParamData
from Bio import SeqIO

# Entering input files
prot_seqs = raw_input("Enter protein sequences file name: ")
sequences=open(prot_seqs, "r")

#writing into a file
gravy_data=open("gravy_"+ prot_seqs,"w")
first_line = "Name" + '\t'+ "\t" + "\t" + " \t gravy_index" + '\n' + "\n"
gravy_data.write(first_line)

#molecular weight counter
for files in list(SeqIO.parse(sequences, "fasta")):
    myprot = ProteinAnalysis(str(files.seq))
    gravy = myprot.gravy()
    seq_ID = files. id
    line = seq_ID + '\t' + '\t' + (str(gravy)[1:-1])+ '\n'
    gravy_data.write(line)
```

```
#This program calculates the instability index for protein sequences within a fasta file
#Sipho Moyo (g10M3876)
#29 September 2014

from Bio.Seq import Seq
from Bio.SeqUtils.ProtParam import ProteinAnalysis
from Bio.SeqUtils import ProtParamData
from Bio import SeqIO

# Entering input files
prot_seqs = raw_input("Enter protein sequences file name: ")
sequences=open(prot_seqs, "r")

#writing into a file
instability_index_data=open("instability_index_"+ prot_seqs,"w")
first_line = "Name" + '\t' + '\t'+ '\t' + " \t instability_index" + '\n' + '\n'
instability_index_data.write(first_line)

for files in list(SeqIO.parse(sequences, "fasta")):
    myprot = ProteinAnalysis(str(files.seq))
    instability_index = myprot.instability_index()
    seq_ID = files. id #collects the sequence identities
    line = seq_ID + '\t' + '\t' + (str(instability_index)[1:-1])+ '\n'
```

```
#Isoelectric point counter for protein sequences within a fasta file
#Sipho Moyo (g10M3876)
#29 September 2014

from Bio.Seq import Seq
from Bio.SeqUtils.ProtParam import ProteinAnalysis
from Bio.SeqUtils import ProtParamData
from Bio import SeqIO

# Entering input files
prot_seqs = raw_input("Enter protein sequences file name: ")
sequences=open(prot_seqs, "r")

#writing into a file
isoelectric_point_data=open("isoelectric_point_"+ prot_seqs,"w")
first_line = "Name" + '\t' + "\t" + "\t" + " \t isoelectric_point" + '\n' + "\n"
isoelectric_point_data.write(first_line)

for files in list(SeqIO.parse(sequences, "fasta")):
    myprot = ProteinAnalysis(str(files.seq))
    isoelectric_point = myprot.isoelectric_point()
    seq_ID = files. id
    line = seq_ID + '\t' + '\t' + (str(isoelectric_point)[1:-1])+ '\n'
    isoelectric_point_data.write(line)
```

```
#Molecular weight counter for protein sequences within a fasta file
#Sipho Moyo (g10M3876)
#29 September 2014


from Bio.Seq import Seq
from Bio.SeqUtils.ProtParam import ProteinAnalysis
from Bio.SeqUtils import ProtParamData
from Bio import SeqIO

# Entering input files
prot_seqs = raw_input("Enter protein sequences file name: ")
sequences=open(prot_seqs, "r")

#writing into a file
molweight_data=open("mol_weight_"+ prot_seqs,"w")
new_line = "Name" + '\t'+ " \t mol_weight" + '\n'
molweight_data.write(new_line)

#molecular weight counter
for files in list(SeqIO.parse(sequences, "fasta")):
    prot_data = ProteinAnalysis(str(files.seq))
    mol_weight = [int(prot_data.molecular_weight())]
    seq_ID = files. id
    line = seq_ID + '\t' + '\t' + (str(mol_weight)[1:-1])+ '\n'
    molweight_data.write(line)
```

III. Script used calculate the amino acid compositions

```
#Amino acid composition counter for protein sequences within a fasta file
#Sipho Moyo (g10M3876)
#29 September 2014


from Bio.Seq import Seq
from Bio.SeqUtils.ProtParam import ProteinAnalysis
from Bio.SeqUtils import ProtParamData
from Bio import SeqIO

# Entering input files
prot_seqs = raw_input("Enter protein sequences file name: ")
sequences=open(prot_seqs, "r")
seq_ID=[]

#writing into files
aa_data=open("amino_acids_"+ prot_seqs,"w")
first_line = "Name" + '\t'+ "\t" + " \t aa_composition" + '\t' + '\n' + "\n"
aa_data.write(first_line)

#amino acid counter
for files in SeqIO.parse(sequences, "fasta"):
    myprot = ProteinAnalysis(str(files.seq))
    amino_acid_count = myprot.count_amino_acids()
    seq_ID = files. id #gives the sequence ID
    line = seq_ID + '\t' + '\t' + (str(amino_acid_count)[1:-1])+ '\t' + '\n'
    aa_data.write(line) #writes the amino acid composition and corresponding sequence ID
```

IV. A summary of the physicochemical properties of the host, parasite and vector groups shown in blue, green and red backgrounds respectively.

Species name	Aromaticity	Gravy index	Instability index	Isoelectric point	Molecular weight
H.sapiens-H/104-335	0.13362	-0.19698	28.47	6.91	25432
H.sapiens-S/102-331	0.12609	-0.51087	26.56	8.29	25593
H.sapiens-L1/102-333	0.12500	-0.53707	35.06	4.73	25635
H.sapiens-K/103-329	0.10573	-0.58238	29.66	8.68	24969
M.musculus-H/102-333	0.11638	-0.15000	38.00	6.99	25268
M.musculus-L/102-334	0.10730	-0.40858	18.90	5.61	25521
Musmusculus-S/110-340	0.12121	-0.44589	28.12	6.51	25278
M.musculus-F/236-462	0.11894	-0.28767	34.35	6.47	25007
M.musculus-K/103-329	0.11013	-0.51938	28.56	8.68	24882
B.taurus-K/108-334	0.10573	-0.62115	21.37	8.80	25078
B.taurus-S/102-331	0.11739	-0.43870	23.02	7.20	25206
B.taurus-L2/102-334	0.11159	-0.44850	16.92	5.56	25389
B.taurus-H/105-336	0.12069	-0.28017	26.02	8.11	25456
G.gorilla-K/103-329	0.10573	-0.58238	29.22	8.68	24955
G.gorilla-H/104-335	0.12931	-0.20690	27.56	7.68	25406
G.gorilla-W/116-376	0.09962	-0.35594	37.74	8.60	29339
P.paniscus-F/142-368	0.10573	-0.31498	38.99	7.04	24919
P.paniscus-W/116-376	0.10345	-0.41226	36.24	8.73	29462
P.paniscus-O/94-321	0.11842	-0.10439	25.36	6.15	25030
P.paniscus-L2/102-334	0.11588	-0.46094	14.87	8.85	25504

P.troglodytes-L1/47-278	0.12500	-0.53707	35.06	4.73	25635
P.troglodytes-W/116-376	0.10345	-0.42682	36.24	8.84	29517
P.troglodytes-O/91-318	0.11842	-0.10439	25.36	6.15	25030
1BY8_A/88-314	0.10573	-0.58238	29.66	8.68	24969
4BS6_A/1-225	0.12000	-0.44667	28.43	6.11	24634
2XU3	0.11818	-0.53000	34.69	4.68	24168
A.darlingi-L/111-344	0.10256	-0.54060	28.26	4.65	25668
G.morsitans-F/237-471	0.09787	-0.46979	28.34	6.07	26057
F.catus-L1/102-333	0.11638	-0.41767	30.76	5.09	25419
F.catus-F/235-461	0.11013	-0.23656	35.16	7.09	24909
F.catus-S/115-312	0.11111	-0.50152	24.94	6.88	21748
F.catus-O/163-390	0.11842	-0.14825	22.46	6.31	25091
G.gallus-L1/121-353	0.12069	-0.59698	22.97	4.70	25599
G.gallus-H/98-329	0.11638	-0.22888	25.35	6.23	25411
G.gallus-O/96-321	0.12389	-0.10929	20.83	6.22	24962
G.gallus-K/108-334	0.11062	-0.43584	39.11	9.14	24806
Gallus-S/	0.11013	-0.34493	28.15	6.35	24778
S.scrofa-K/104-330	0.10573	-0.61762	21.74	8.79	25053
S.scrofa-L1/102-334	0.11159	-0.47511	20.24	5.60	25409
S.scrofa-H/63-297	0.11915	-0.30000	26.72	6.48	25789
S.scrofa-S/102-331	0.12609	-0.54304	29.55	8.37	25524
8PCH_A/1-220	0.12273	-0.27500	28.04	5.81	24305
Falcipain-1/317-569	0.12648	-0.49842	33.73	5.56	28821
Falcipain-2A/242-484	0.11667	-0.38167	36.11	4.97	27047
Falcipain-2B/240-482	0.11667	-0.36167	37.66	4.96	26998
Falcipain-3/249-492	0.14108	-0.38216	24.95	4.80	27055
Vivapain-2/245-487	0.13333	-0.27667	26.35	4.70	27017
Knowlesipain-2/251-495	0.13223	-0.42686	35.84	4.93	27686
Bergheipain-2/230-470	0.13808	-0.33640	34.08	4.81	27140
Chabaupain-2/231-471	0.12971	-0.30879	49.25	4.56	26886
Yoellipain-2/232-472	0.14226	-0.24644	37.08	4.81	27226
Rhodesiense/114-450	0.07122	-0.14243	33.49	4.68	35574
Cruzipain/111-383	0.13553	0.36264	26.29	4.76	29679
L.aethiopica/112-443	0.08133	-0.18705	44.02	5.02	35531
L.tropica/112-443	0.08434	-0.16747	48.13	5.80	35714
L.major/112-348	0.08861	-0.09198	38.21	5.48	25454
L.mexicana/112-443	0.08735	-0.13163	45.02	5.67	35609
T.gondii-L/190-421	0.08621	-0.25733	54.31	4.77	25198
E.Histolytica/82-308	0.08370	-0.44934	25.90	8.52	24486
3BWK_A/1-243	0.14403	-0.39877	24.51	4.75	27347
2OUL_A/1-241	0.11618	-0.38797	35.83	4.98	27156
3F75_A/99-330	0.08929	-0.32366	55.03	4.82	24401

V. A summary of the p-values obtained from the Kruskal-Wallis test for the physicochemical properties among all the three groups.

Property	P-value		
	Catalytic domain	Full length sequences	Catalytic domain and full length sequences
Aromaticity	0.9359	0.1965	0.6209
Gravy index	0.0225	0.8400	0.0600
Instability index	0.0027	0.2166	4.67E-05
Isoelectric point	2.68E-05	0.0087	2.80E-07
Molecular weight	0.0002	0.0019	3.17E-14