Structural Analysis of Prodomain Inhibition of Cysteine Proteases in *Plasmodium* Species.

A mini-thesis submitted in 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 & Biotechnology

Faculty of Science

By

Joyce Njoki Njuguna

June 2012

ABSTRACT

Plasmodium is a genus of parasites causing malaria, a virulent protozoan infection in humans resulting in over a million deaths annually. Treatment of malaria is increasingly limited by parasite resistance to available drugs. Hence, there is a need to identify new drug targets and authenticate antimalarial compounds that act on these targets. A relatively new therapeutic approach targets proteolytic enzymes responsible for parasite‟s invasion, rupture and hemoglobin degradation at the erythrocytic stage of infection. Cysteine proteases (CPs) are essential for these crucial roles in the intraerythrocytic parasite. CPs are a diverse group of enzymes subdivided into clans and further subdivided into families. Our interest is in Clan CA, papain family C1 proteases, whose members play numerous roles in human and parasitic metabolism. These proteases are produced as zymogens having an N-terminal extension known as the prodomain which regulates the protease activity by selectively inhibiting its active site, preventing substrate access. A Clan CA protease Falcipain-2 (FP-2) of *Plasmodium falciparum* is a validated drug target but little is known of its orthologs in other malarial *Plasmodium* species. This study uses various structural bioinformatics approaches to characterise the prodomain‟s regulatory effect in FP-2 and its orthologs in *Plasmodium* species (*P. vivax, P. berghei, P. knowlesi, P. ovale, P. chabaudi and P. yoelii)*. This was in an effort to discover short peptides with essential residues to mimic the prodomain‟s inhibition of these proteases, as potential peptidomimetic therapeutic agents. Residues in the prodomain region that spans over the active site are most likely to interact with the subsite residues inhibiting the protease. Sequence analysis revealed conservation of residues in this region of *Plasmodium* proteases that differed significantly in human proteases. Further prediction of the 3D structure of these proteases by homology modelling allowed visualisation of these interactions revealing differences between parasite and human proteases which will lead to significant contribution in structure based malarial inhibitor design.

DECLARATION

I, **JOYCE NJOKI NJUGUNA**, 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 ……6 JUNE 2012..……...

ACKNOWLEDGMENTS

It is my pleasure to thank those who made writing this thesis possible:

- My supervisor Dr. Özlem Tastan Bishop for your remarkable insight, guidance and encouragement I am truly grateful
- My parents, sisters and Justus for loving, supporting and believing in me throughout this year
- To my colleagues at the Rhodes Bioinformatics Research group Rowan, Alex and Dustin for their friendship and assistance.
- To the Rhodes University for funding and making it possible for me to undertake the course.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

1D	One Dimensional
2D	Two Dimensional
3D	Three Dimensional
BDZ	Benzodiazepine
BLAST	Basic Local Alignment Tool
BLASTP	Basic Local Alignment Tool for Protein sequences
BLOSUM	BLOcks of Amino Acid Substitution Matrix
CP	Cysteine Protease
DNA	Deoxyribonucleic Acid
DOPE	Discrete Optimized Protein Energy
EM	Electron Microscopy
EMBL	European Molecular Biology Laboratory
FFT	Fast Fourier Transform
FP-1	Falcipain-1
FP-2	Falcipain-2
FP-2B	Falcipain-2B
FP-3	Falcipain-3
GDT_TS	Global Distance Test Total Score
HMM	Hidden Markov Models
JTT	Jones Taylor Thornton

MAFFT	Multiple sequence Alignment based on Fast Fourier Transform
MEGA	Molecular Evolutionary Genetic Analysis
MQAP	Model Quality Assessment Program
MSA	Multiple Sequence Alignment
MUSCLE	MUltiple Sequence Comparison by Log – Expectation
NCBI	National Centre for Biotechnology Information
NJ	Neighbour Joining
NMR	Nuclear Magnetic Resonance
PAM	Point Accepted Mutations
PDB	Protein Data Bank
PDB_ID	Protein Data Bank Identification Number
PIC	Protein Interaction Calculator
PROMALS	PROfile Multiple Alignment with Local Structure
PROSA	PROtein Structure Analysis
PSI- BLAST	Position Specific Iterated - Basic Local Alignment Tool for Protein sequences
PSSM	Position Specific Scoring Matrix
RMSD	Root Mean Square Deviation
RNA	Ribonucleic Acid
SH	Sulfhydryl
T-Coffee	Tree-based Consistency Objective Function for alignment Evaluation
UPGMA	Unweighted Pair Group Method with Arithmetic Mean

TYPOGRAPHICAL CONVENTIONS

The Plasmodium species will be referred to by their Latin names

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*

Amino acids will be referred to by their 3-letter abbreviations

SYMBOLS USED

α	Alpha-helix
β	Beta-sheets
Å	Angstrom a measure of distance at atomic level

CHAPTER ONE

1. INTRODUCTION

1.1 MALARIA

Malaria is the most dreadful protozoan infection in humans resulting in over a million deaths each year and is responsible for enormous economic burden in endemic regions (Ettari et al. 2010). The control of malaria is increasingly limited by resistance of the parasite to available drugs. Thus there is a need for identification of new drug targets and validation of potential antimalarial compounds that act on these targets (Rosenthal et al. 1991). Malaria is caused by the *Plasmodium* parasite; there are five *Plasmodium* species that cause malaria in humans namely: *Plasmodium falciparum, Plasmodium vivax, Plasmodium knowlesi, Plasmodium ovale* and *Plasmodium malariae* (White 2008). Other species of *Plasmodium* including *Plasmodium berghei, Plasmodium yoelii* and *Plasmodium chabaudi* are known to cause malaria in rodents.

Among the five species of *Plasmodium* parasites that infect human beings *P. falciparum* is justifiably the most virulent due to high levels of mortality with which it is associated. However the other human malaria causing *Plasmodium* species seem to be less studied and neglected in the light of *P. falciparum. P. vivax* has an estimated global burden of ~80-100 million clinical cases per annum, it is the most widely distributed human malarial parasite causing extensive morbidity. Most cases originate from Southeast Asia and the Western Pacific, a significant number also occur in Africa and South America. Though often considered a benign and self-limiting infection, there is increasing evidence of the overall burden, economic impact and severity of disease from *P. vivax* (Price et al. 2007). The species *P. knowlesi* has a 24 hour life cycle, hence daily schizont rupture rapidly increases parasite load and even a short delay in accurate diagnosis, treatment and management may lead to increase in complications (Cox-Singh et al. 2008). *P. knowlesi* is encountered widespread in Malaysia and has been found in Thailand and China (White 2008). *P. ovale* is a relapsing malaria parasite with a latent liver stage that

often persists for many months and secondary infections can later be generated. Its course of parasitemia is short in comparison to other human infecting malarial parasites. *P. ovale's* natural distribution is in sub-Saharan Africa and the islands of the West Pacific and there many other reports of its presence throughout the world (Collins et al. 2005). Therefore though neglected in comparison to research efforts of *P. falciparum* the other human malarial *Plasmodium* species are of medical importance and hence they were included in this study. The mice infecting *Plasmodium* species *P. berghei, P. yoelii* and *P. chabaudi* are important as they provide experimental models for studying malarial disease processes *in vivo,* thus they were also included in this study.

1.1.1 Infection and lifecycle

Malaria is spread to humans by the infected female anopheles mosquito that inoculates the parasites (sporozoites) into the subcutaneous tissue or directly into the bloodstream. Sporozoites migrate to the liver and invade hepatocytes where they replicate asexually yielding thousands of merozoites. Merozoites are released into blood circulation where they invade red blood cells beginning the erythrocytic cycle. The intraerythrocytic parasite by asexual replication produces about 20 mature merozoites that are then released to invade other erythrocytes. This can lead to thousands of parasite infected red-blood cells in hosts circulation and causes manifestation of disease symptoms that may be fatal if not treated. A small portion of the merozoites are converted into sexual forms of the parasite known as gametocytes that are essential for transmission of the disease from host to the vector thus completing the lifecycle (Miller et al. 2002).

1.1.2 Erythrocytic stage of *Plasmodium* infection

Clinical manifestation of malaria begins at the erythrocytic stage of infection. The asexual parasites (merozoites) invade and multiply within erythrocytes that eventually rupture releasing more merozoites into circulation. The erythrocytic parasites reach high numbers in circulation

causing increased hemolysis leading to clinical symptoms of malaria including fever, malaise, fatigue, nausea, vomiting, diarrhoea, anaemia and jaundice. The intra-erythrocytic parasites take up the host‟s erythrocyte cytosol, into its own food vacuole and breaks down hemoglobin to provide necessary amino acids for parasite protein synthesis (Sijwali et al. 2004). The invasion and rupture of erythrocytes; allowing entry and egress of parasites as well as the hydrolysis of hemoglobin are processes involving the cooperate activity of various classes of malarial proteases (Figure 1.1) (Rosenthal 1998). Among these proteases are cysteine proteases, which have been identified to play a key role in the breakdown of hemoglobin as well as erythrocyte rupture (Rosenthal 2004). Other proteases involved include: aspartic proteases (plasmepsin I, II, IV and HAP) and metallo proteases involved in hemoglobin hydrolysis. Serine proteases (Subtilase I and Subtilase II) are associated with parasite‟s invasion and rupture of erythrocytes. This study will focus on cysteine proteases, due to their crucial role in the *Plasmodium* life cycle (Figure 1.1).

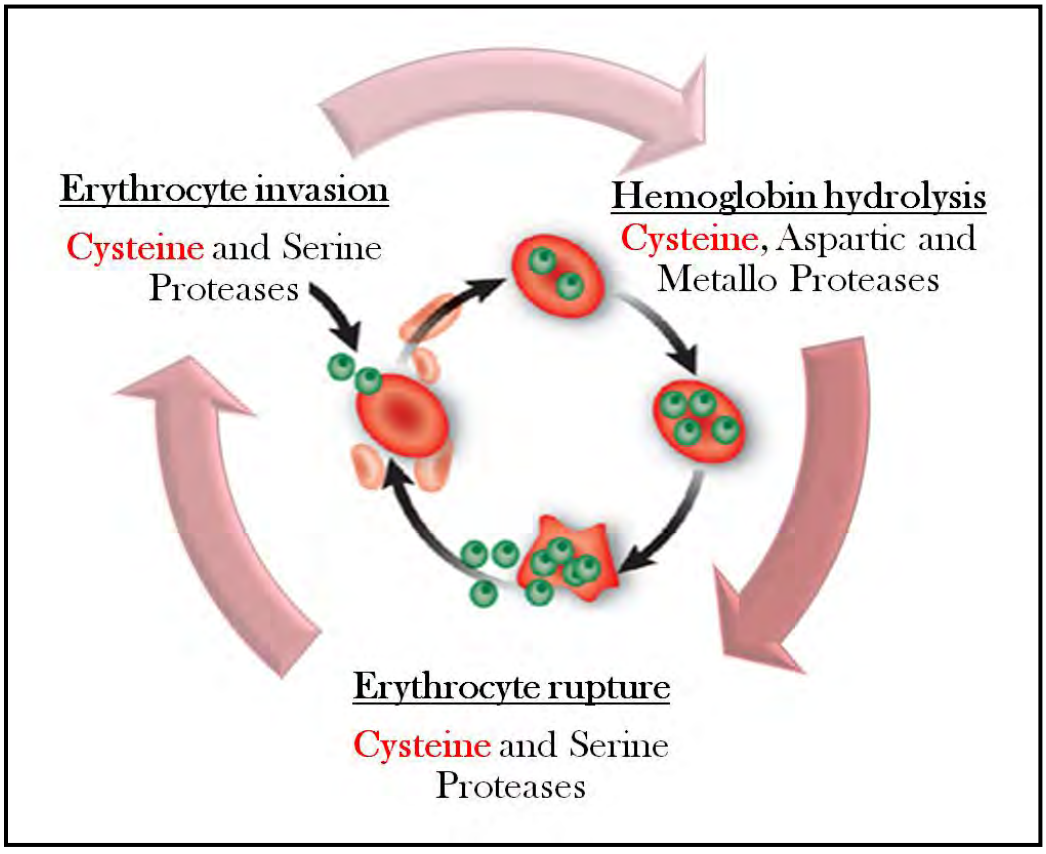

Figure 1.1 Erythrocytic stage of infection, showing the proteases involved at the various stages of erythrocyte invasion, hemoglobin degradation and erythrocyte rapture (Rosenthal 1998). Diagram of intraerythrocytic cycle was adopted from the (http://medic19.blogspot.com/) webpage; the diagram was edited accordingly.

1.2 CYSTEINE PROTEASES

Proteases are a group of enzymes that catalyse the hydrolysis of peptides. Proteases are grouped on the basis of their catalytic mechanisms and substrate specificity. The main catalytic classes include serine, threonine, aspartate, metallo and cysteine protease (Grzonka et al. 2001). Cysteine proteases are of interest to this study as they are essential in the life cycle and pathogenicity of many protozoan parasites and are targets for new anti-parasitic drugs (Rizzi et al. 2011). Cysteine proteases have an essential Cys residue in the active site that mediates hydrolysis of peptide bonds via a nucleophilic Cys thiol-group in a catalytic dyad. The sulfhydryl (-SH) group of Cys side chain and the imidazole of an active site His residue give rise to a thiolate-imidazolium charge relay dyad. This dyad is frequently but not always stabilized by a conserved Asn residue (Rosenthal 2004).

Protease activity is not only determined by the three residues (Cys, His and Asp) forming the catalytic triad, but by other adjacent residues known as subsites residues (Schechter 2005). The active site of a protease has a dual function of binding the substrate and catalyzing the reaction. The efficiency of these two roles dictates the specificity of the protease towards ligand. It is thus possible to determine details on the active site by studying the kinetics of the protease towards different substrates (Schechter & Berger 1967).

Cysteine proteases of parasitic organisms are categorized into clans namely: clan CA, CB and CC (McKerrow 2002). This study focuses on clan CA cysteine proteases also referred to as „papain-like" proteases that utilize Cys, His and Asn residues in their active sites. Clan CA proteases are well characterized in many organisms, the first of these was characterised from the papaya fruit hence the name papain. Clan CA is further subdivided into families based on sequence identities and similarities; family C1 (Papain family) and family C2 (calpain-like) (Lecaille et al. 2002). Majority of parasite proteases belong to the family C1 (Papain family) that is comprised of three subclasses: the cathepsin-L-like, cathepsin-B and cathepsin-F-like proteases (Rosenthal 2004).

Falcipains are the best characterised cysteine proteases in *P. falciparum*, there are four known falcipains: falcipain-1 (FP-1), falcipain-2 (FP-2), falcipain-2B (FP-2B) and falcipain-3 (FP-3) (Rosenthal 2004). Falcipains belong to the family C1 (Papain family) proteases and are within the cathepsin-L-like subclass (Wu et al. 2003).

1.2.1 Functions of cysteine proteases in *Plasmodium* parasite

Cysteine protease specific inhibitors have been used to determine the functions of falcipains in *Plasmodium* as being; invasion and rupture of erythrocytes as well as hemoglobin degradation (Wegscheid et al. 2010). FP-2 is the best characterised falcipains, its specific biologic role in the parasite was determined by the knock-out of FP-2 gene in recombinant parasites. Hemoglobin hydrolysis was altered in the FP-2 knockout parasites demonstrating the crucial role of FP-2 in hemoglobin hydrolysis and validating is as a drug target (Sijwali et al. 2004). Hemoglobin hydrolysis is the best characterised function of falcipains. The intraerythrocytic parasite takes up the host‟s erythrocyte cytosol transporting it via the cytostome to an acidic food vacuole, where hemoglobin is degraded (McKerrow 2002). Hemoglobin is hydrolysed, its heme component is converted to hemozoin pigment and globin is hydrolysed to its constituent amino acids. Hemoglobin hydrolysis is necessary to provide amino acids for parasites own protein synthesis (Francis et al. 1997). The breakdown of hemoglobin also enables maintenance of the erythrocyte‟s osmotic stability preventing premature erythrocyte rupture and provides space for the growing intraerythrocytic parasites. FP-2, FP-2B and FP-3 have also been implicated in hemoglobin degradation (Sijwali et al. 2006).

Erythrocyte invasion by *Plasmodium* parasites involves the action of cysteine proteases on the erythrocyte cell surface. FP-1 specific inhibitors have been shown to block parasites invasion of erythrocytic host in culture (Blackman 2004). The third role of falcipains in the intraerythrocytic parasite is erythrocyte rupture. Falcipains facilitate erythrocyte membrane destabilization hence allowing parasite‟s escape from the erythrocyte confines (Blackman 2008). Studies using E-64 cysteine protease inhibitor have shown that the inhibitor blocks lysis of the parasitophorous

vacuole membrane that encloses the parasite within the erythrocyte. Hence falcipains act against proteins associated with the parasitophorous vacuole and erythrocyte cell membranes to enable rupture and completion of the intraerythrocytic cycle (Rosenthal 2004).

1.2.2 Structural properties of FP-2

Papain-like cysteine proteases are typically synthesised in an inactive zymogen form to prevent potentially harmful premature proteolysis (Lecaille et al. 2002). The zymogen is comprised of an N-terminal extension of the mature enzyme (mature domain) known as the prodomain. The prodomain serves a regulatory role in the protease during secretory transport. It inhibits the protease by spanning over the active site hence preventing substrate access to the catalytic residues (Chowdhury et al. 2002). The zymogen structure is depicted in (Figure 1.2), the prodomain (purple) spanning over the active site (red) in the mature domain (green). The prodomain binds to the active site in the direction opposite to that of regular substrate binding, hence it is resistant to hydrolysis. Activation of the protease occurs by cleaving off the prodomain (Jean et al. 2003).

The prodomain in falcipains has been found to be unusually longer that other Papain-family proteases (Sijwali et al. 2002). Other than its regulatory role, the falcipain prodomain is also known to have a role in intracellular trafficking as well as stabilizing the protease in denaturing alkaline pH environments (Cygler et al. 1997). The prodomain structure of cathepsin-L-like proteases is known to have 2 α-helices at its N-terminal: (α-1 helix and α-2 helix) intersecting at ~60 (Figure 1.2). This is followed by a short α-3 helix that anchors the prodomain to the active site of the mature domain. The C-terminal section of the prodomain overlays the groove between the left and right lobes of the mature domain (Figure 1.2) (Godat et al. 2005).There is currently no available structure for the zymogen form of falcipains. However the crystallographic structure of the mature domain of FP-2 protease has been solved at a resolution of 3.1 Å (Hogg et al. 2006). FP-2 mature domain adopts a fairly typical papain-like structure, having a left and right lobe. The left lobe is predominantly alpha helical, while the right lobe comprises of antiparallel beta sheets and peripheral helices. The active site catalytic residues Cys-285, His-417 and Asn-447 (FP-2 numbering) are located at the junction between the left and right lobes (Ettari et al. 2010) as depicted in (Figure 1.2).

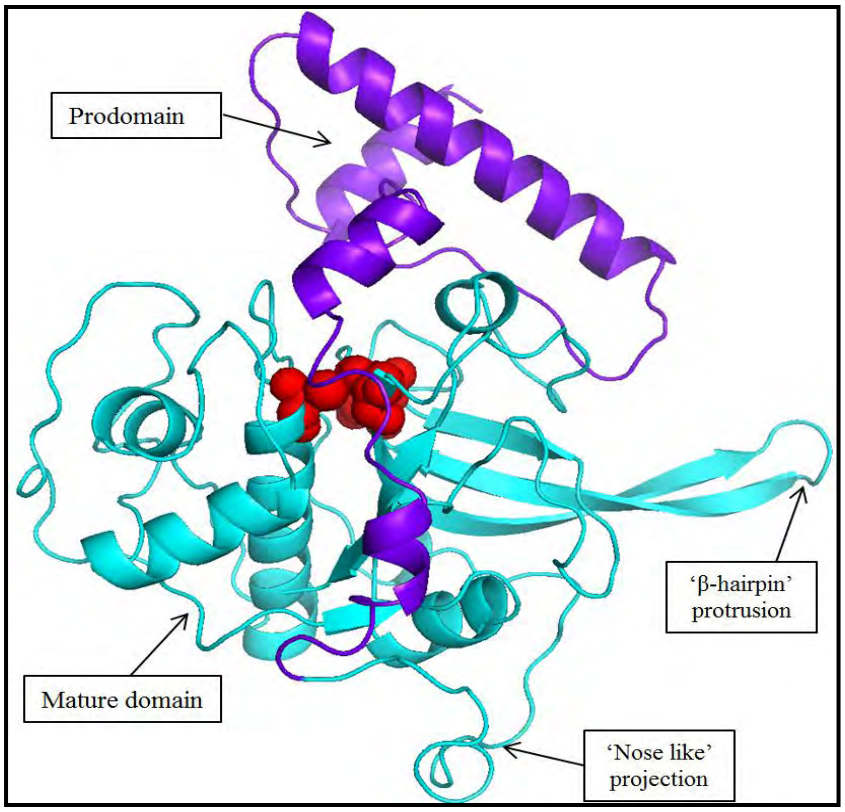

Figure 1.2 Structural features of *Plasmodium* cysteine protease, the structure was attained by homology modelling of FP-2 (see **Chapter 3.3.4**). The mature domain is shown in green, the active site is shown in red located between the left and right lobe of the protease. The N-terminal „nose like" projection and C-terminal insert „β-hairpin" protrusion are indicated. The N-terminal extension (prodomain) that is responsible for protease inhibition is shown in purple lying over the active site and the groove between the left and right lobes.

FP-2 has some unique features that are unusual to the Papain family. There is a 17 amino acid „Nose like" projection at the N-terminal of the mature domain of the protease, it is implicated to be involved in protease folding (Ettari et al. 2010). FP-2 also has a 14-residue „β-hairpin" protrusion known as the „arm", extending away from the protease surface, which is thought to be a hemoglobin binding motif (Hogg et al. 2006). These two features are the cause of the significantly larger dimensions of FP-2 in comparison to other Papain family cysteine proteases. The unique „Nose like" projection at the N-terminus of mature domain have a chaperone like activity allowing proper folding of the mature FP-2 in absence of its prodomain (Sijwali et al. 2002). Studies suggest that these unique features of FP-2 are also found in its orthologs in other *Plasmodium* species and are essential for protease folding, activation and substrate binding (Wang et al. 2006).

1.2.3 Cysteine protease chemotherapy

Malarial proteases are attractive drug targets against malaria due to their critical role in the parasite‟s infection and lifecycle. The treatment of experimental models of parasitic disease with cysteine protease inhibitors such as vinyl sulfones, peptidyl aldehydes and peptidyl fluoromethyl ketones has provided important proof of the drug target concept. However these inhibitors are not selective for the parasite enzyme, hence they may lead to toxicity as they also inhibit the host enzyme (McKerrow 1999). This poses the challenge in drug design as it is difficult to design an inhibitor specific to the parasite enzyme. The complexity of chemotherapy against cysteine proteases is due to abundant presence of these enzymes in viruses, prokaryotes and higher organisms. These enzymes also have broad substrate specificity making it difficult to inhibit individual proteases.

Studies on inhibitor selectivity for parasite proteases suggest that host cells are less sensitive to cysteine protease inhibitors at the concentrations used in cultures or in *in vivo* studies (Selzer et al. 1999). As parasites appear to take up and concentrate inhibitors more efficiently than host cell organelles. Concentrations of proteases within host‟s cell are significantly higher than in the parasites hence even if few proteases are inhibited there may be limited toxicity to the host. In a malaria animal model, a fluoromethyl ketone that inhibits falcipain blocked *P. vinckei* protease activity *in vivo* after a single subcutaneous dose and cured 80% of murine malaria infections when administered for four days (Rosenthal 1998). However due to the challenge of low selectivity of most chemotherapeutic agents, there is a need for development of highly selective inhibitors with minimal toxic effects on the hosts.

1.2.4 Prodomain inhibitory effect on cysteine proteases

Though cysteine proteases are promising therapeutic targets there has been limited ability to attain inhibitors that are selective for the parasite protease. Hence there are efforts in developing alternative biologic inhibitors. One approach is based on the fact that parasitic cysteine proteases are synthesised as zymogens with N-terminal peptide chain extensions (prodomains) (Figure

1.1). Prodomains regulate the catalytic domain activities during secretory transport; preventing possibly damaging premature activation. The prodomains of cysteine proteases in Papain family are specific inhibitors of their cognate proteases. Therefore there is much interest in developing inhibitors that mimic this prodomain selective inhibition of proteases (Scott et al. 2010). One of the approaches is to synthesize small fragments of peptides that span the prodomain and assess their ability to inhibit the mature enzyme and thus identify residues specifically interacting with the active site (Sijwali et al. 2002). In a study small recombinant prodomain peptides of FP-2 were designed and used to evaluate their antimalarial activity (Korde et al. 2008). One of the expressed peptides PP1 (-89aa – 75aa) was shown to significantly inhibit FP-2 in a dose dependent manner, the peptide also showed ability to translocate into the infected erythrocyte and inhibit FP-2 *in-vitro*. As a result of peptide inhibitor treatment, clusters of merozoites enclosed in a delicate membrane covering were observed within the erythrocytes. This was indicative of inhibited erythrocyte rupture hence merozoites did not exit the RBC confine. Prodomain peptide treated parasites also showed accumulation of undigested hemoglobin indicating inhibited hemoglobin hydrolysis (Korde et al. 2008).

However studies on the structure of FP-2 prodomain show that this inhibitory peptide PP1 occurs at the α-1 helix in the N-terminal of the prodomain (see Chapter 2.3.2). Therefore in its native zymogen form peptide PP1 is not in interaction with the active site and is likely not responsible for specific inhibition. Hence a better understanding of the structural features and binding modes of these prodomains to their target cysteine proteases will provide valuable information for the development of effective and highly selective inhibitors (Carmona et al. 1996).

1.2.5 Structural analysis and homology modelling

To enable studying of the prodomain-mature domain interactions in the proteases mentioned it was essential that the 3D structure be available. This would allow for the visualisation of residue interactions and comprehensive interaction analysis. There are millions of protein sequences in the various sequence databases, but only a few tens of thousands of protein structures solved by

X-ray crystallography, NMR and electron microscopy in the Protein Data Bank (Berman et al. 2000). This is because of the difficulty of isolation and solving of 3D protein structures. The goal of protein modelling is to predict a structure from its sequence with an accuracy that is comparable to the structure achieved experimentally. This allows for rapidly generated *in silico* protein models useful in structure-based drug design, analysis of protein function and study of protein interactions. Homology modelling is an alternative way to obtain structural information if experimental techniques are unsuccessful. Many proteins are either too large for NMR analysis or cannot be crystallized for X-ray diffraction. Homology modelling involves identifying appropriate homologs of known protein structure, called templates, alignment of the target (protein to be modelled) sequence with the template structures, model building and refinement, and validation (Tastan et al. 2008). This study will thus use homology modelling to predict the 3D structure of the FP-2 and its orthologs in *P. vivax, P. knowlesi, P. ovale, P. berghei, P. yoelii* and *P. chabaudi*

1.3 PROBLEM STATEMENT AND JUSTIFICATION

Malaria is the world‟s most important parasitic disease infecting 300-500 million people and resulting in over a million deaths annually, mainly among children in sub-Saharan Africa. The control of malaria is increasingly limited by the parasite‟s resistance to available drugs hence there is a need for new strategies for antimalarial therapy. Falcipains are promising drug targets for malaria treatment. Studies on FP-2 have shown its role of hemoglobin hydrolysis as well as erythrocyte invasion and rupture of the intraerythrocytic parasite. Given the crucial role that FP-2 support in the *Plasmodium* parasite it has been validated as drug targets. Cysteine proteases within the Papain superfamily have broad substrate specificity. This poses a challenge in drug design as chemotherapeutic agents developed against them have poor selectivity towards parasitic cysteine protease over the human cysteine protease (Shah et al. 2011). There is a need to design protease inhibitors that are specific to the parasite protease. Prodomain regions of cysteine proteases have been shown to inhibit their cognate enzymes specifically and selectively (Korde et al. 2008). Studying the inhibitory interactions of the prodomain with the mature enzyme, would reveal the essential residues for selective protease inhibition. These can further

be developed as short recombinant peptide inhibitors specific against the protease and potential as peptidomimetic therapeutic agents.

Currently only two studies have been done on prodomain inhibition of FP-2 (Korde et al. 2008 , Pandey et al. 2009). Little is known of prodomain inhibition in the orthologs of FP-2 in other *Plasmodium* species: *P. vivax, P. knowlesi, P. ovale P. berghei, P. yoelii* and *P. chabaudi.* As well as their 3D structures of these orthologs remain unknown. Fewer studies have also been done on the other falcipains FP-1, FP-2B and FP-3. This study seeks to fill this knowledge gap by homology modelling of the 3D structures of the orthologs of FP-2 in the six *Plasmodium* species and in FP-2B and FP-3, showing prodomain to mature domain interactions. The study will use the known crystallographic structures of select papain-family cysteine proteases as templates for homology modelling. Unlike *P. falciparum, in vitro* culture is unavailable for the other human malarial *Plasmodium* species under study, making structural studies more appropriate. This study will analyse the sequences and obtained structures of the malarial proteases against select human papain family proteases (Cathepsin-H, -K, and -L). This is to observe for significant differences between parasite and human proteases. The results obtained will go a long way into broadening knowledge of structure based antimalarial drug design.

1.4 SPECIFIC OBJECTIVES

1. Retrieval of FP-2 and its orthologs from protein sequence databases by reverse BLAST and identification of appropriate templates for model building.
2. Sequence alignment analysis and phylogenetic analysis
3. Model building, refinement and validation of target cysteine proteases
4. Study inhibitory interactions of the prodomain to the mature domain in the obtained cysteine protease models.
5. Identify sequence and structural differences between *Plasmodium* and human cysteine proteases that are useful in development of specific peptide inhibitors

CHAPTER TWO

2. SEQUENCE ALIGNMENT AND ANALYSIS

Falcipain-2 (FP-2) is a papain family cysteine protease from *P. falciparum*. This chapter focuses on the identification of FP-2 orthologs in six *Plasmodium* species namely: *P. vivax, P. knowlesi, P. ovale P. berghei, P. yoelii* and *P. chabaudi.* The protein sequences of FP-2 and its orthologs are analysed via multiple sequence alignment (MSA) and phylogenetic analysis. This is to discover conserved motifs and residues potentially involved in protein function. Papain family cysteine proteases are ubiquitous enzymes occurring both in the *Plasmodium* parasites and its human host. The *Plasmodium* protease sequences are thus compared against their human homologs revealing differences between parasite and human proteases significant in antimalarial drug design.

2.1 INTRODUCTION

Proteins are fundamental in controlling and implementing biological processes in living organisms. This includes enzymatic catalysis, immunological responses, mechanical support, coordinated motion, endocrine function and storage; hence deeper understanding of protein structure and function is of essence. In living organisms the flow of genetic information from the gene to protein level is similar. Genetic information encoded in DNA or RNA sequences is translated to amino acids that build up proteins. The knowledge of genome sequences has been central to understanding the proteins encoded in organisms and their functional implications (Tatusov et al,. 1997). Advances in next generation sequencing technology has allowed for increased speed and efficiency of DNA sequencing, making it easier to obtain the amino acid sequences of proteins (Zhang et al. 2011). Proteins with similarity in sequence of amino acids are divergent from a common ancestral gene and usually have similar structural and functional properties (Altschul 1998).

Evolutionary relatedness of protein families is referred to as homology. Homologs diverge from a common ancestor and homology is inferred by sequence, structural or functional similarity. Homologs are of three types: orthologs, paralogs and xenologs. In most cases orthologs retain similar functional characteristics as they evolve; hence identifying orthologs to a novel protein is essential to its function prediction (Tatusov et al. 1997). Characterising a novel or unknown protein sequence involves comparing it against a sequence database to identify annotated homologs. These annotated homologs can then offer reference for functional and structural analysis of the unknown protein (Pierri et al. 2010).

There are several publicly accessible biological databases to which new nucleotide and amino acid sequences are submitted by the scientific community. These databases store and organise sequence and structural data allowing for its easy retrieval. The available databases range from primary databases that hold raw sequence data, secondary databases having curated and annotated data and specialized databases that focus on information of specific research interest. Among the major biological database resources is the National Center for Biotechnology Information (NCBI) database (http://www.ncbi.nlm.nih.gov/) and the RCSB Protein Data Bank (PDB) (http://www.rcsb.org/pdb/home/home.do) (Rose et al. 2011). To ease retrieval of record in these databases, search tools based on similarity searching are used.

Sequence comparison is the starting point of structural and functional analysis of proteins. Comparing sequences provides insight on the functional and evolutionary inference of a newly sequenced protein with proteins already present in biological databases. These eases annotation of new proteins as biological knowledge from well characterised homologs can be conferred. Sequence comparison is achieved through alignment, the process by which regions of similarity and residue to residue correspondence is searched between sequences (Xiong 2006). Sequence alignment depicts evolutionary relatedness of the proteins involved, extensive similarity between sequences points out likely common ancestral origin. Analysis of sequence alignment is based on sequence similarity and identity. Sequence identity is when matching residues align, while sequence similarity refers to alignment of residues that are similar in physiochemical properties.

If sequence similarity is high enough it can infer sequence homology and consequently likely structural similarity (Krissinel 2007).

2.1.1 Alignment strategies and algorithms

Sequence comparison can either be by global and/or local alignment. Global alignment is carried out across the complete length of the sequences involved, while local alignment finds and aligns local regions of similarity between the sequences (Altschul 1998). Local alignment is ideal for distantly related sequences that contain only similar domains. Dynamic programing is an effective algorithm for attaining optimal sequence alignment. Needleman-Wunsch algorithm for global alignment and Smith-Waterman algorithm for local alignment employ dynamic programming (Needleman et al. 1970). Dynamic programing seeks to find the maximum matches that can be aligned between two protein sequences while accounting for mismatches. These methods assign scores to matches, define costs for substitutions and gaps and compute the alignments, the higher the total score is the better the alignment is (Morgenstern et al. 1996). Dynamic programming is computationally expensive and time consuming hence unsuitable for aligning multiple sequences.

To assign the various scores dynamic programming uses scoring matrices, to determine the probability of a residue been substituted for another in an alignment based on similarity in physiochemical properties. The most common scoring matrices used are Point Accepted Mutations (PAM) and Blocks of Amino Acid Substitution Matrix (BLOSUM) which are based on real multiple sequence alignments (Henikoff 1992) and have been shown to outperform matrices based purely on the physiochemical properties of amino acids. PAM scoring matrix is based on the fact that amino acids substitution during the evolutionary process tends to conserve physiochemical properties. Thus amino acids of the same size, charge or hydrophobicity are more likely to be substituted for each other. The PAM-1 matrix is based on a real dataset by calculation of the likelihood of one substitution per 100 residues. Higher PAM matrices are derived by multiplying PAM-1 by itself, thus PAM-40 results from multiplying PAM-1 by itself 40 times. The PAM 40 matrix means that in 100 residues there are 40 substitutions, which is

ideal in aligning closely related sequences that have 70-90% similarity. Hence as the evolutionary distance between sequences increases higher numbered PAM matrices are used (Wheeler 2002). BLOSUM scoring matrix is derived from conserved un-gapped blocks of protein sequence alignments. The frequency of all possible amino acid pairs in every column of the block are numerated and used to calculate a matrix. Closely related sequences in the block are most likely to increase the frequency of certain amino acids hence creating bias. This is resolved by clustering similar sequences beyond a certain threshold in a block; hence clustering at 80% similarity threshold gives BLOSUM-80. As the clustering threshold reduces it incorporates more distantly related sequences hence lower BLOSUM matrix numbers detect more distant related sequences (Henikoff 1992). Scoring matrices score matches and substitutions in alignment; additional gaps costs are charged where residues align with nulls the best alignment has the least costs and the highest score.

2.1.2 Database similarity search tools

There are software tools used to search for homologs in biological databases all based on sequences similarity realized through alignments. Basic Local Alignment Search Tool (BLAST) is one of the most frequently used programs for searching for homologs in biological databases. BLAST uses heuristic algorithms to rapidly find optimal alignments in biological databases (Ye et al. 2006). BLAST displays results as a list of the sequence matches ordered by the statistical significance measured by the alignment score and the expectation value (E-value). A higher alignment score depicts a better alignment while the E-value is a measure of the probability that the alignment is a random match from the database. The lower the E-value is the higher the biological significance of the alignment. The BLAST tool is available as a stand-alone program as well as through a variety of web servers, the NCBI-BLAST been one that is commonly accessed (http://blast.ncbi.nlm.nih.gov/Blast.cgi). NCBI-BLAST offers a variety of databases search methods like BLASTP that screens protein databases using a protein query sequence. There are some refinements to the BLAST tool that enhance the program‟s speed and sensitivity. The PSI-BLAST tool is one such enhancement that makes use of position specific scoring matrices (PSSM) to increase sensitivity in detecting remote homologs (Altschul et al. 1998).

HHpred server is another fast database search tool for detecting protein homology across a wide range of protein databases (http://toolkit.tuebingen.mpg.de/hhpred). HHpred server facilitates sensitive and speedy database searches allowing detection of remote homologs and protein structure prediction. The server achieves high sensitivity and accuracy through its HHsearch resource that makes use of Hidden Markov models (HMM-HMM comparison) (Söding et al. 2005). Database searches are performed by first querying the sequence against non-redundant databases using several PSI-BLAST iterations to align with homologs. PSI-BLAST generates sequence profiles that are used for secondary structure prediction by PSIPRED; hence results of the final PSI-BLAST alignment are inclusive of predicted secondary structure (Jones 1999). This final alignment with structure prediction details is used to create a profile HMM that has position specific amino acid, insertion and deletion probabilities (Söding 2005). For each column in the alignment a HMM profile column is added that holds the probabilities of a match, insert and delete state. A match state is the probabilities of the 20 amino acids allowed in that column, while the insert and delete states model the probability of an insert or deletion at the position. The insert and delete probabilities are converted to position specific gap penalties used in HMM-HMM alignments (Eddy 1998). HHsearch is then used to query the consensus HMM profile against a selected database by performing HMM-HMM comparison. HMM profiles are better for homolog detection than traditional position-specific scoring matrices (PSSMs) and achieve better alignments; their increased sensitivity is due to position specific gap penalties (Söding 2005). HHpred results display the alignment and alignment statistics including the probability of homology to the query, percentage identity, similarity and the E-value.

2.1.3 Retrieval of orthologs from biological databases

Retrieving of sequences from biological databases is based on sequence similarity. The BLAST tool is useful in detecting homologs. BLAST hits with high sequence identity and E-values lower than 1e-6 ($1*10^{-6}$) are likely to be homologs to the query sequence (Xiong 2006). Reverse BLAST is an approach used to identify orthologs of a query sequence from a biological database, it is a two-stage process involving a forward and reverse BLAST step. In the forward BLAST, the query sequence is used to perform a BLAST search narrowed to the species from which

orthologs are targeted. The first hit from the forward BLAST results is used as a query in the reverse BLAST step. If the reverse BLAST returns the initial query sequence as its best hit, the two sequences are thus considered to be orthologs (Mazumder et al. 2005). Orthologs commonly have extensive sequence and structural similarity.

There are homologs and orthologs gene databases that can also be searched for putative orthologs (Zhou et al. 2007). The NCBI Clusters of Orthologous Groups of proteins (COG) (http://www.ncbi.nml.nih.gov/COG/) is a database containing orthologs of unicellular organisms currently holding 66 genomes. HomoloGene (http://www.ncbi.nlm.nih.gov/homologene) is a homologs sequence database as well as OrthoMCL DB (http://www.orthomcl.org/cgi-bin/OrthoMclWeb.cgi) which is a database of orthologs holding 150 genomes from both eukaryotic and prokaryotic species (Chen et al. 2006). These databases are specialized to hold orthologs and have search tools like BLASTO making it possible to precisely search and retrieve orthologs (Zhou, et al. 2007).

2.1.4 Multiple sequence alignments

Sequence alignment as discussed earlier is the initial point in functional and structural characterisation of a protein. Homologs of a given protein can be identified through database similarity searches and comparing several homologs gives a better view of the conservation within a family of organisms. Multiple sequence alignment (MSA) helps to reveal ancestral relationships between organisms and conserved residues and motif that may be of functional importance (Taylor 1998). Heuristic approaches have been developed to carry out MSA, this includes progressive alignments, iterative alignment and consistency based alignment (Thompson et al. 1999). Progressive MSA is built up in a stepwise manner by first aligning closely related sequences and subsequently adding more distantly related sequences. There are a number of MSA programs based on progressive alignments, CLUSTAL been the most commonly used. CLUSTAL MSA are built up progressively by initial pairwise alignments between all possible sequences, based on their alignment scores a guide tree is constructed using the Neighbour joining method. The guide tree is used to build the MSA by aligning closely

related sequences and adding sequences progressively according to their location in the guide tree (Chenna et al. 2003). T-Coffee is a consistency based aligner, known to be slow but has increased accuracy in alignment. Consistency based approach is where T-coffee program creates a library comprised of local and global pairwise alignments. These pairwise alignments are generated by other alignment programs like CLUSTAL (Local alignments) and LALIGN (global alignments). The program evaluates the local and global alignments produced, allocating weight to the alignments based on sequence identity. This is followed by library extension of the combined local and global alignments (Niun et al. 2006). T-Coffee then uses the library to come up with a position specific scoring matrix used to construct a guide tree, which is used to direct a MSA using the progressive approach (Di Tommaso et al. 2011). T-Coffee eliminates errors in the early stages of pairwise alignment that can be propagated through the MSA hence improving the progressive method (Pei et al. 2007).

Profile multiple alignment with local structure (PROMALS3D) is a MSA alignment program that performs both sequence and structural alignments (Pei et al. 2008). It is an improved consistency based method that initially performs pairwise alignment of similar sequences scoring alignments using the BLOSUM62 scoring matrix to give groups of pre-aligned sequences. A representative sequence is selected from each group and using PSI-BLAST its homologs are searched from the Uniprot non-redundant reference databases facts (UNIREF90) and Protein structure prediction server (PSIPRED) for secondary structure prediction (Pei et al. 2007). Amino acids sequence profiles are derived from the PSI-BLAST and PSIPRED they are used to come up with a consistency scoring function. The representative sequences are then aligned progressively using the consistency scoring function and all the pre-aligned sequences are incorporated to form a complete MSA (Pei et al. 2008).

Multiple sequence alignment based on fast fourier transform (MAFFT) is a fast consistency based approach of MSA, which builds a preliminary alignment using the progressive method then refines it iteratively to produce an optimal alignment (Pei, Grishin 2007). Distance between all pairs of sequences to be aligned is calculated and used to generate a guide tree using Unweighted pair group method with arithmetic mean (UPGMA) method. MAFFT uses fast

Fourier transform (FFT) analysis to rapidly detect segments of sequence similarity; FFT algorithm aligns the sequences by progressive alignment based on the guide tree (Katoh et al. 2002). The alignment is further refined by generating a better distance matrix based on the pairwise alignments of the initial guide tree. The new distance matrix is used to generate a guide tree for progressive alignment this process is iterated till no better scoring alignment is attained (Katoh et al. 2005). MAFFT has the advantage of speed and producing precise MSA.

MUSCLE is a MSA program that is based on refinement of the progressive method to get optimal alignments. MUSCLE uses *K*mer a distance measure between un-aligned pair of sequences to generate a distance matrix, clustered by UPGMA to produce a guide tree (Tree-1). *K*mer is a measure that does not require alignment giving this first stage the advantage of speed. The guide tree generated (Tree-1) is used to guide a progressive alignment producing a draft MSA. The draft MSA is rather substandard as a result of the *K*mer distance estimation (Edgar 2004). MUSCLE thus generates an improved tree using kimura distance, which is a measure of evolutionary distance between aligned sequences that corrects homoplasy. Homoplasy is the effect of multiple mutations at one position of a sequence that affects accurate evolutionary distance prediction. Kimura distance is a more accurate measure of evolutionary distance than *K*mer and is used to generate a guide tree (Tree-2). By progressive alignment a MSA is generated guided by tree-2, computing of the alignment is only done for branching orders of Tree-2 that differ from Tree-1.

2.1.5 Phylogenetic analysis

To evaluate the evolutionary history of the protein sequences, molecular phylogenetic analysis is done. This allows for studying of evolutionary divergence of sequences using a tree branching pattern, referred to as a phylogenetic tree. The general assumption is that divergence is bifurcating hence a predecessor branches out into two descendants. Construction of a phylogenetic tree involve multiple sequence alignment of the sequences involved, this need be done accurately. Evolutionary models are then used to approximate the evolutionary distances

between sequences and a distance matrix is constructed that is used to generate a phylogenetic tree (Xiong 2006).

MEGA5 is one of the numerous programs available for phylogenetic analysis. MEGA5 is used for the analysis of DNA and protein sequences, the program has tools that align sequences, estimate distance between sequences and generate phylogenetic trees. The program performs alignment of the data using in built MUSLE or CLUSTAL alignment programs or can accept data that is already aligned. Estimating evolutionary distances between the protein sequences is of importance to assessing the sequence divergence; to do this MEGA5 has statistical models otherwise known as evolutionary models. Statistical models like Jones-Thornton-Taylor (JTT) and Dayhoff produce accurate evolutionary distance estimations by accounting for amino acid substitution between sequences and correcting homoplasy. The calculated distances are converted into a distance matrix used by an algorithm to build a phylogenetic tree (Kumar et al. 2004).

Neighbour joining (NJ) is one of the distance based algorithms used in constructing of phylogenetic tree. It involves joining all the taxa into a single node forming a star-like tree, the most closely related pair forms the first node and the next closely related taxa is linked to the first node this is progressive till all taxa have been added creating a complete tree (Saitou et al. 1987). To ensure the most optimal tree topology for a given set of taxa is achieved, several trees are generated each having a different initial pair for the first node and the tree that optimally fits the evolutionary distance is selected. Authenticity of the tree in representing phylogeny of the data is then statistically evaluated using the bootstrap test (Kumar et al. 2004). Bootstrapping involves iteratively altering the data-set and generating trees that are then compared against the original tree to test for robustness. Bootstrap values are indicative of the confidence levels of the topology. MEGA5 allows visualisation of the constructed tree in a Tree Explorer. Phylogenetic trees give a clear view of how various species have evolved within a family, showing the evolutionary divergence or similarity of species involved.

2.2 METHODS

Protein sequences of FP-2 and its orthologs in six *Plasmodium* species were retrieved, aligned and analysed. *Plasmodium* sequences are compared to those of human papain family homologs. The procedures and tools used are discussed here.

2.2.1 Sequence retrieval

The protein sequence of FP-1 (XP_001348727.1), FP-2 (XP_001347836.1), FP-2B (XP_001347832.1) and FP-3 (XP_001347833.1) were retrieved from the NCBI-Entrez database. FP-2 sequence was used as a query for the reverse BLAST, to retrieve its orthologs in six species: *P. vivax* (taxid:5855), *P. berghei str.* ANKA (taxid:5823), *P. knowlesi* (taxid:5850), *P. ovale* (taxid:36330), *P. chabaudi chabaudi* (taxid:31271) and *P. yoelii yoelii* str. 17XNL (taxid:352914). The NCBI BLASTP tool was used for the reverse 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 search for FP-2 orthologs using the reverse BLAST approach was done in two stages; the forward BLAST that points to the orthologs in other species and the reverse BLAST that ideally should return FP-2 as the best hit. However the reverse BLAST returned FP-3 as the best hit instead of FP-2. Therefore it is likely that the retrieved orthologs (Table 2.1) are more closely related to FP-3 than FP-2. To establish this sequence identity of the retrieved orthologs to FP-2 and FP-3 is determined (Tables 2.2). A search was done in the PlasmoDB and HomoloGene database to verify orthologs results obtained from the NCBI. The most probable FP-2 orthologs were selected and used for sequence alignment (Table 2.1). Homologs of FP-2 from the *H. sapiens* species were also retrieved from NCBI. The retrieved human protease sequences were within the papain family, cathepsin-L-Like sub-class; there were essential for comparative analysis of parasite versus human proteases. All retrieved sequences were analysed using the InterPro server (http://www.ebi.ac.uk/Tools/pfa/iprscan/) to scan for functional domains.

2.2.2 Sequence alignments

MSA of the *Plasmodium* sequences were generated to observe for conserved residues and domains within this genus of proteases. Comparative analysis of *Plasmodium* against human homologs cathepsin-L1, cathepsin-K and cathepsin-H was done. MSA was carried out using MAFFT (http://toolkit.tuebingen.mpg.de/mafft/), T-Coffee (http://tcoffee.crg.cat/), CLUSTALX, PROMALS3D (http://prodata.swmed.edu/promals3d/promals3d.php) and MUSCLE (http://www.ebi.ac.uk/Tools/msa/muscle/) alignment programs. The alignments generated by the various programs were compared and the most accurate alignment was selected. None of the used alignment programs is entirely acurate, most generated alignments would have regions that are poorly aligned. The poorly aligned regions have significant effect on the general quality of the alignment. Hence it is necessary to select the best possible alignment.

The alignment programs were accessed via web servers and default alignment parameters were used. A fasta file holding all the sequences to be aligned was uploaded to the program and the MSA was run automatically. An initial alignment was done using the entire protease sequences; to give a view of the local regions of similarity across the full length sequences (Appendix A-I). The sequences were then trimmed to the prodomain and mature domain segment of interest. This allowed for better alignment that focussed on interest regions. Alignments were viewed using BioEdit (Hall 1999) alignment viewer. It was observed that PROMALS-3D alignment program gave the most accurate / best quality alignments. This MSA program is not entirely error free hence it is necessary to manually edit the alignment and improve its quality. This is done by examining the aligned residues and hand adjusting of residues so as to attain a close to ideal alignment. Analysis was done by identifying conservations unique to *Plasmodium* that are likely to influence selective propeptide inhibition of these species.

2.2.3 Phylogenetic analysis

MEGA5 was used to generate a NJ phylogenetic tree that depicted the evolutionary relationships between the *Plasmodium* and human proteases. PROMALS3D alignment was used as the dataset

for tree generation. Contiguous regions of the alignment with minimal gaps were selected as the dataset (Appendix A-II). The Neighbour joining algorithm was used to generate a phylogenetic tree. The parameters used were JTT statistical model and 1000 replicates of bootstrapping analysis were done to determine the confidence level at the inner nodes of the topology.

2.3 RESULTS AND DISCUSSION

Results attained from protease sequence retrieval, sequence and structural alignment and phylogenetic analysis are discussed in this section. All residue numbering used in analysis is in reference of FP-2. The proteases are referred to by their common names indicated in (Table 2.1)

2.3.1 Retrieval of orthologous sequences

FP-2 orthologs in *Plasmodium* species were searched from the NCBI database by reverse BLAST (Table 2.1). The HomoloGene and PlasmoDB database searches reflected similar FP-2 orthologs authenticating our results. Homologs of FP-2 from the *H. sapiens* species that are within the cathepsin-L-like subclass were also retrieved. All the retrieved cysteine proteases will hence forth be referred to by their common names as indicated in (Table 2.1)

Table 2.1 Summary of FP-2 and its homologs retrieved from NCBI. The table shows the accession number, common name and length of the sequence as well as the prodomain and mature domain regions of interest used in sequence alignment.

Species name	Accession Number	Common name	Number of Amino acids		
			Total	Pro-domain	Mature domain
P. falciparum 3D7	XP_001347836.1	Falcipain-2	484	155-243	244-481
P. falciparum 3D7	XP_001347832.1	Falcipain-2B	482	153-241	242-479
P. falciparum 3D7	XP_001347833.1	Falcipain-3	492	161-249	250-489
P. vivax	AAT36263.1	Vivapain-2	487	157-245	246-484
P. knowlesi strain H	XP_002259152.1		495	163-251	252-490
P. berghei strain ANKA	XP_680416.1		470	144-229	230-467
P. yoelii yoelii str. 17XNL	XP_726900.1		472	146-231	232-469
P. chabaudi chabaudi	AAP43630.1	Chabaupain-2	471	144-230	231-468
P. falciparum 3D7	XP_001348727.1	Falcipain-1	569	215-306	307-567
P. ovale	AAC47037.1		317	1-78	79-310
Homo sapiens	AAL23962.1	Cathepsin H	323	18-103	104-307
Homo sapiens	NP_001903	Cathepsin L1	333	25-113	114-331
Homo sapiens	NP_000387	Cathepsin K	329	22-113	114-327

All retrieved sequences were from the clan CA proteases and within the Papain family C1, Cathepsin-L-like subfamily. BLAST as well as InterPro database searching revealed two functional domains associated with this family: peptidase C1A Papain family (mature domain) and proteinase inhibitor-129 cathepsin propeptide (prodomain). The respective prodomain and mature domain regions of the sequence are indicated in (Table 2.1) above. Reverse BLAST results showing the first two hits attained is represented in (Table 2.2). Sequence identity of the of the ortholog sequences to FP-2 and FP-3 was also attained.

Table 2.2 Reverse BLAST results indicating the first and second hit returned by the NCBI-BLAST tool and the percentage identity of the mature domain of each ortholog to FP-2 and FP-3.

Ortholog	**Accession number**	**REVERSE BLAST**		**% Identity of sequences of FP-2 and FP-3**	
		First hit	second hit	FP_2	FP_3
Falcipain-2	XP_001347836.1	FP_2	FP_2B	100	52.8
Falcipain-2B	XP_001347832.1	FP_2B	FP_2	93.1	53.0
Falcipain-3	XP_001347833.1	FP_3	FP_2	52.8	100
Vivapain-2	AAT36263.1	FP_3	FP_2	47.8	54.2
P. knowlesi	XP_002259152.1	FP_3	FP_2	43.8	47.6
P. berghei	XP_680416.1	FP_3	FP_2	40.5	40.7
Chabaupain-2	AAP43630.1	FP_3	FP_2	40.6	41.0
P. yoelii	XP_726900.1	FP_3	FP_2	39.3	40.7
Falcipain-1	XP_001348727.1	FP_2B	FP_2	18.7	20.1
P. ovale	AAC47037.1	FP_1	FP_2	36.1	36.1

FP-3 been the first hit in the reverse BLAST results likely suggests that the sequences were more closely related to FP-3 than to FP-2 based on the E-value and BLAST bit scores. To establish this sequence identity of the retrieved sequences to FP-2 and FP-3 were compared. Most of the sequences have higher sequence identity to FP-3 than to FP-2, with an exception of FP-2B. Further sequence and phylogenetic analysis was done to determine if the higher similarity to FP-3 was significant to the functionality of the proteases structure and function.

2.3.2 Sequence alignment and analysis

Multiple sequence alignments of the retrieved proteases were carried out using MUSCLE, MAFFT, T-Coffee, CLUSTALX and PROMALS3D, the best alignment was selected. PROMALS3D gave the best alignment, given that it combines sequence and structural constraints ensuring alignment of best fit without gaps in non-loop regions. Unlike the other alignment programs PROMALS3D alignment used structural constraints from the homologous structures included in the MSA. Alignments by the other programs are provided in the (Appendix A-III).The sequence identity between aligned sequences was not very high so to ensure best fit of alignment PROMALS3D results were manually inspected and slight hand adjustment was done. Hand adjustment was guided by HHpred structure prediction analysis (see more in the Appendix A-IV). The alignments clearly showed conserved regions in the prodomains of *Plasmodium* proteases (Figure 2.1).

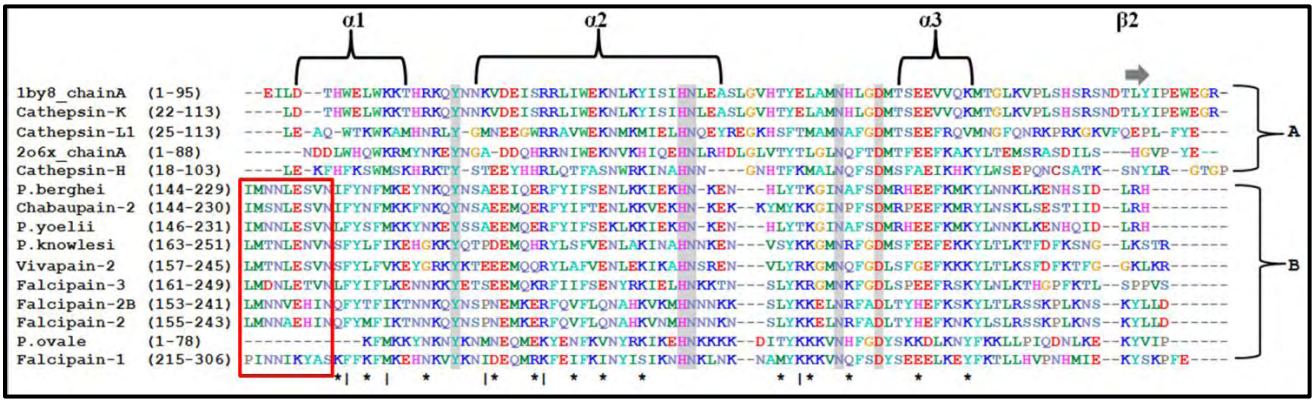

Figure 2.1 Alignment of the prodomain sequences of human and *Plasmodium* proteases in section A and B respectively. Two 3D structures from the papain family are added to the alignment (1BY8 and 2O6X) (in section A), the α-helices and beta sheets positions relative to 206X structure are indicated. The residues highlighted in grey are indicative of strict conservation in both species and the asterisk „*" indicates the aligned regions of high similarity in both species. The vertical line „|" indicate aligned residues whose conservation is unique to *Plasmodium.* A nine residue insert at N-terminal of *Plasmodium* is marked out in red

The alignment showed residue conservation in certain regions indicated by asterisks and grey highlighting in (Figure 2.1). *P. ovale* sequence was notably shorter than the other *Plasmodium* sequences which may be attributed to its incomplete sequencing. It was also observed that the sequences of FP-1 and *P. ovale* were less similar to the other *Plasmodium* sequences due to their distant evolutionary relationship as depicted by the phylogenetic tree in (Figure 2.5). The

prodomain sequences of *Plasmodium* proteases were markedly longer than those of human proteases; having a conserved nine residue N-terminal extension (red outlined in Figure 2.1). This nine residue extension (155-164) LMNNLESVN is unique to the *Plasmodium* species and is likely to have a significant function in the *Plasmodium* proteases. A study suggests that this insert region of the prodomain is responsible for protease inhibition (Korde et al. 2008) (see Chapter 1.2.4). However this is highly unlikely given its location in the prodomain structure.

The general 3D structure of the prodomain in cathepsin-L-like proteases has an N-terminal mini-domain comprising of 2 α-helices intersecting at 60 degrees (Figure 2.2). This is followed by a section of prodomain chain that covers the groove between the two lobes of the mature enzyme blocking the active site and is affixed by a short α-3 helix (Cygler & Mort 1997). These structural regions were mapped onto the sequences and the nine residue extension in *Plasmodium* was found to occur in the α-1 helix region, which as depicted in (Figure 2.2) is not in direct interaction with the active site and is least likely to cause inhibition. The α-1and α-2-helices are known to be key in maintaining structural integrity of the prodomain, while α-3-helix lies adjacent to the active site (red in Figure 2.2) region causing inhibition (LaLonde et al. 1999).

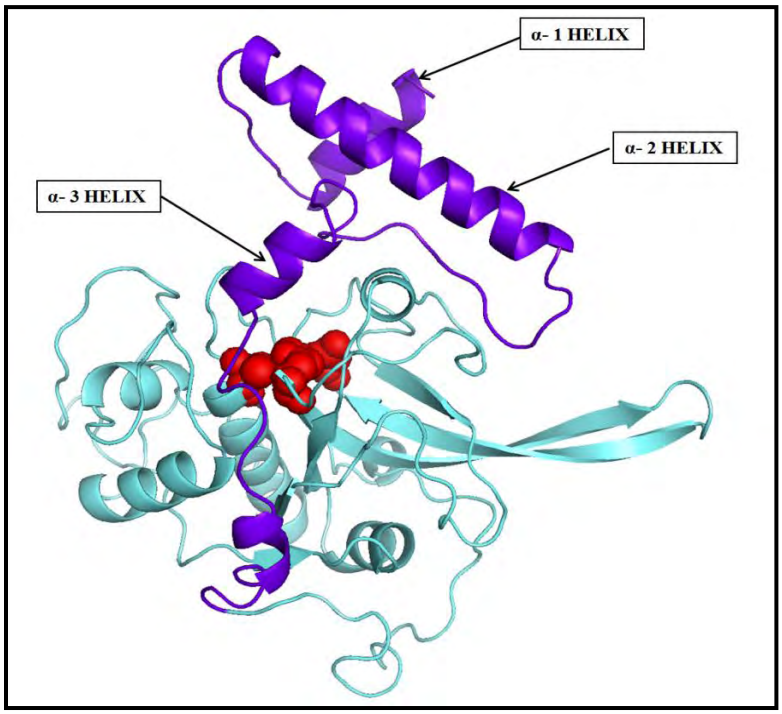

Figure 2.2 Papain family prodomain structure in purple showing the α-1, α-2 and α-3 helix sections. α-3 helix is shown to directly overlay the active site (red) and is most likely to have inhibitory interactions with the residues of the active site. The structure is attained by homology modelling of FP-2 (see **Chapter 3.3.4**).

Conserved residues in the α-1, α-2 and α-3 helix segments of the prodomain were closely examined and significant differences between *Plasmodium* and human proteases in these sections were observed (Table 2.3).

Table 2.3 Conserved residues in the prodomains of *Plasmodium* (grey) and human (green) proteases. The table represents select substitutions observed across the sequences (blank spaces indicate residue conservation). Residues are coloured by physiochemical properties: green-hydrophobic, red-polar acidic, blue-polar basic, black-polar uncharged

Protease	α-1 helix					α-2 helix								α-3 helix and surrounding loop regions													
conservation	165 F	166 Y	168 F	174 K		181 E	185 R	186 F	189 F	192 N	196 I	200 N		209 K	210 G	212 N	214 F	216 D	221 E	222 F	223 K	224 N	225 K	226 Y	227 L	228 S	230 K
Falcipains-2											V				E												R
Falcipains-2B											V				E							S					R
Falcipains-3														R							R	S				N	
Vivapain-2				R					Y													K				T	
P. knowlesi									Y												E	K				T	
P. berghei																						M				N	
Chabaupain-2											V											M	R			N	
P. yoelii																						M				N	
Falcipain-1		F					K								K					L		E	Y	F	K	K	L
P. ovale	-	-					K		Y	K					K				D	L			Y	F	K	T	L
Cathepsin-H	W	E	W					W	L					M	A							H					Q
Cathepsin-K	W	T	W			D		W	R					L	A					V	V	Q		M	T	G	
Cathepsin-L1	W	K	W	R				W	R					M	A		L			I	R	Q	V	M	N	G	E

The α-1 helix region was found to have conserved Phe residues at position 165 and 168 in *Plasmodium* species, both residues are replaced by Trp in the human protease; these residues are proposed to form hydrophobic interactions with Phe 189 that is highly conserved in α-2 helix of the *Plasmodium* prodomain (Table 2.3). These three Phe residues form a hydrophobic core that is vital in stabilizing the prodomain structure (Pandey et al. 2009). Another hydrophobic residue Try 166 is observed to be conserved in the α-1 helix of the *Plasmodium* proteases. However this residue is substituted for Glu and Lys which are polar charged residue in human protease cathepsin-H and cathepsin-L1 respectively. Phe 186 is also relatively conserved in the α-2 helix of *Plasmodium* species and is thought to support the hydrophobic core; it however aligns to Arg a polar basic residue in human cathepsin-K and -L1. The structure of the prodomain is sustained by residue interactions via hydrophobic interactions and hydrogen bonding (Pandey et al. 2009). This facilitates inhibitory association of the prodomain to the mature domain. Hence these

dissimilarities in key hydrophobic residues are likely to have effect on the structures of the *Plasmodium* versus human prodomain affecting how they occlude the mature domain.

There are two well-known conserved motifs in the prodomains of cathepsin-L-like proteases ER(F/W)N(I/V)N (residue 189-200) and GNFD (residue 210-216) (Wiederanders et al.2003), these are observed to be conserved in both *Plasmodium* and human prodomains. These two conserved motifs are essential in maintaining the secondary structure of the prodomain. The ERFNIN motif lies in the α-2 helix fostering interactions with residues in α-1 helix to ensure proper prodomain structure and appropriate inhibitory contact to the catalytic domain. The ERFNIN motif is well conserved in the α-2 helix of both *Plasmodium* and human proteases (Table 2.3). The GNFD motif is found in the loop region right before α-3 helix (Table 2.3). This motif is conserved in both *species* with an exception of the Gly-210 residue which is replaced by a non-polar residue Ala in all the human proteases; it also aligns to Glu in FP-2 and FP-2B and Lys in *P. ovale* and FP-1. The Asp residue of the GNDF motif is highly conserved and is interacts with the Arg-181 residue of the ERFNIN motif to stabilize α-3 helix (Wiederanders et al. 2003).

The prodomain region spanning the groove between the two domains of the mature protease occludes the active site by interacting with the residues in the substrate binding sites. This region corresponds to the α-3 helix and the loop region after it in the prodomain (Figure 2.2). A conserved motif FKNKYLT is found in this region from residues 222-228 these residues are highly similar in the *Plasmodium* proteases. A pattern of conservation of this motif was observed, in the three murine species *P. berghei, P. chabaudi* and *P. yoelii* the motif was FKMRYLN. In 2 *Plasmodium* species *P. ovale* and FP1 the motif residues were LK(N/E)YFKK. In the proteases FP-2, FP-3, FP-2B, vivapain-2 and *P. knowlesi* the motif residues were FK(S/K)KYLT. This motif was seemingly conserved in accordance to the three major *Plasmodium* clusters depicted by the phylogenetic tree in (Figure 2.5). These prodomain residues are likely to be in interaction with residues of the substrate binding pocket (subsites) in the mature domain, hence obstructing substrate access. Subsites are the residues flanking the catalytic triad (Cys, His and Asn), which interact with ligands. FP-2 has 4 known subsites S1, S1', S2 and S3 (Figure 2.4) (Shah et al. 2011).

The prodomain residues identified to occupy and inhibit the S1' subsite in human cathepsin-L1 and cathepsin-K are Met 92 and Asn/Thr 93 (Chowdhury et al. 2002), these residues align with hydrophobic residues Try-226 and Leu-227 respectively in the FKNK**YL**T motif of the *Plasmodium* species (Table 2.3). This is a significant difference in the S1' inhibition of human and parasitic proteases. In cathepsin-L1 and cathepsin-H the prodomain residues Gln-96 and Glu-86 respectively are known to occupy and inhibit S1 subsite. These align with basic polar residues Lys/Arg-230 in the *Plasmodium* species; hence there are substantial differences in the prodomain residues occupying the various subsites in *Plasmodium* and human proteases. Knowledge of the prodomain residues key in selective inhibition of the various subsites in the proteases is of importance as these are potential in peptidomimetic drug design. The prodomain is known to have selective inhibition of its parent enzymes hence contrasts in residue interaction in parasite and human protease are essential to attaining selective *antimalarial* activity.

Other noted divergences were: Lys-170 a basic polar residue is fully conserved in *Plasmodium* prodomains, however in aligns with Ser and Ala which are uncharged residues in the human cathepsin-H and cathepsin-L1 respectively. Lys-209 is also conserved in *Plasmodium* it aligns with hydrophobic residues Met in cathepsin-H and cathepsin-L1 and Leu in cathepsin-K. These residues within the prodomain mark out differences in human and *Plasmodium* that are likely to affect the structure of their prodomain and its interactions with the mature domain.

Having analysed the prodomain sequences of the proteases, the mature domain was also analysed to identify residues that are around the active site that are likely to interact with the prodomain. The mature domain alignment of human and *Plasmodium* is shown in Figure 2.3. This alignment shows numerous conserved residues across *Plasmodium* and human proteases. The catalytic residues Cys-285, His-417 and Asn-447 (FP-2 numbering) are conserved in both species (Figure 2.3). There is a 17 residue insert observed in the N-terminal of the *Plasmodium* proteases that is absent in the human cathepsin-L1, -K and -H proteases. The insert residues (244-261) are highly similar in *Plasmodium* species with an exception of FP-1 and *P. ovale* that have different residue composition (Figure 2.3).

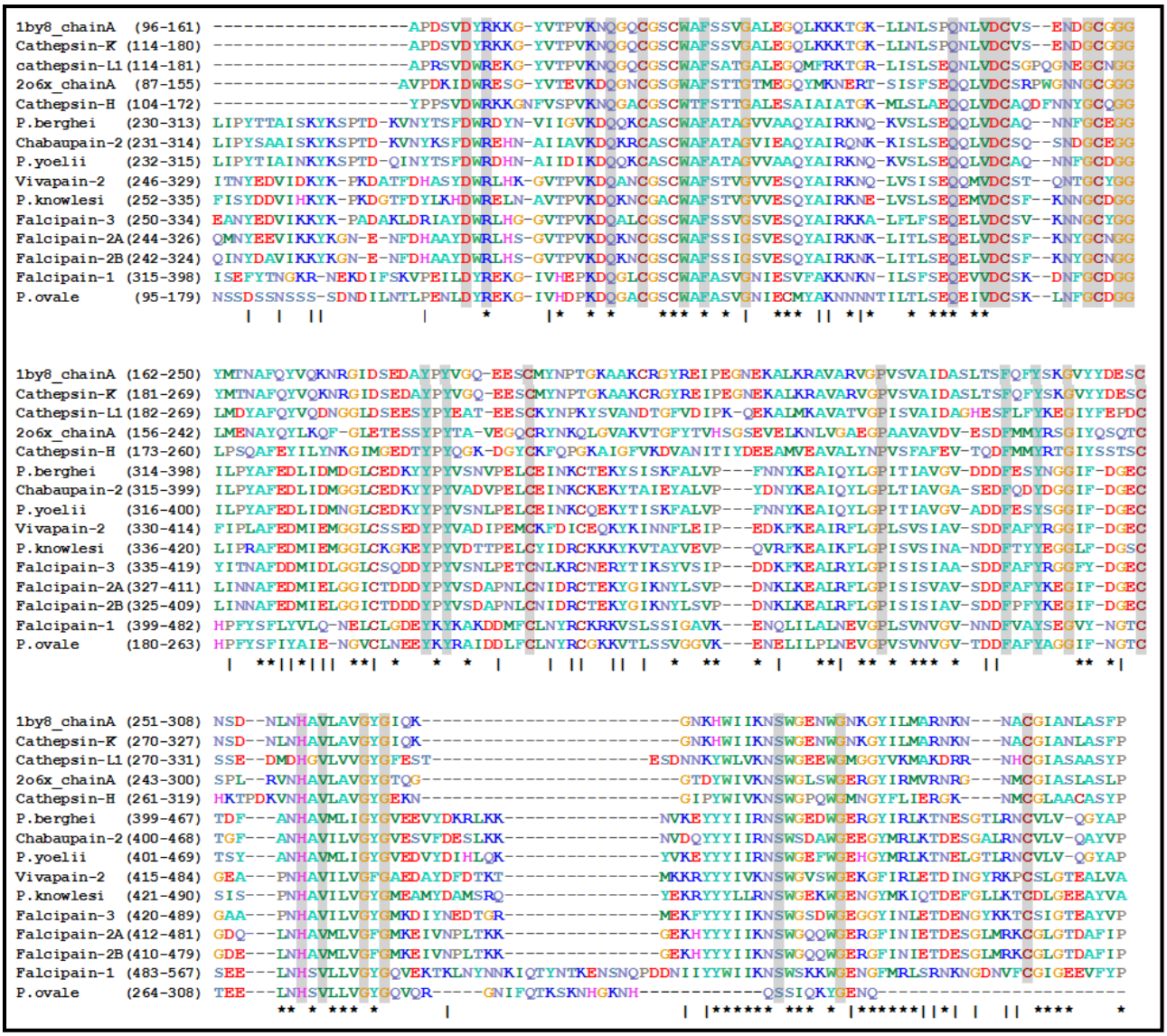

Figure 2.3 Alignment of catalytic domain in *Plasmodium* and human species inclusive of two papain family protease structures 1BY8 and 2O6X. The regions highlighted in grey indicate strictly conserved residues and the asterisk „*" show regions of high similarity and conserved physiochemical properties in both species. The vertical line „|" shows indicates aligned regions with conservation of residues unique to *Plasmodium* species and different in human species.

This N-terminal extension in *Plasmodium* is associated with „chaperone like" function, it enables folding of the mature protease to active conformation in absence of the prodomain (Sijwali et al. 2002). A ten residue insert (429-438) exclusive to the *Plasmodium* species is also observed near the C-terminal (Figure 2.3). This insert is fairly similar in the *Plasmodium* species except in FP-1and *P. ovale* where it is 15 residues longer and fairly different in amino acid residue composition. This insert is thought to be a hemoglobin binding motif (Pandey et al. 2005).

FP-2 is known to have four disulphide bonds that are conserved in all the *Plasmodium* proteases (CYS 282-323), (CYS 316-357), (CYS 342-362), (CYS 411-472) (Hogg et al. 2006). The human proteases also show conservation of these Cys residues with an exception of (CYS 342-362) that is unique to *Plasmodium.* The unique *Plasmodium* disulphide bond (CYS 342-362) is essential in stabilising the structure of the mature domain in active conformation. Disulphide bond (CYS 411-472) in *Plasmodium* forms the upper loops of the S2 and S1' binding sites in FP-2. Residues of the substrate binding pocket (subsites) in the protease sequences were analysed. Conservation of residues in the four subsites (Figure 2.4) were analysed in human and *Plasmodium* proteases (Table 2.4)

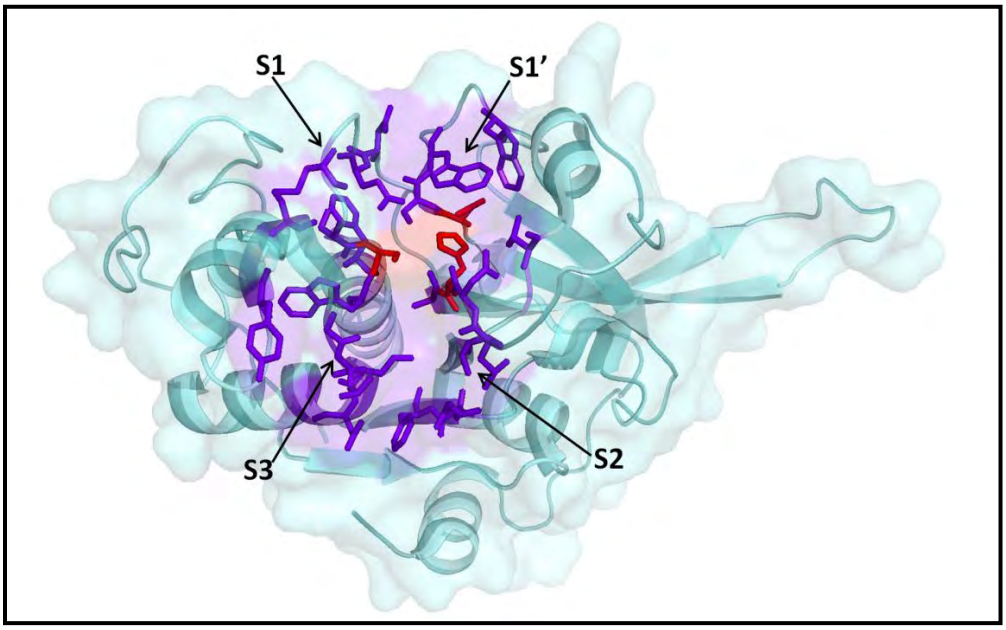

Figure 2.4 Subsites in the *Plasmodium* proteases showing the four subsites (purple) around the catalytic residues (red) in the mature domain (light green) of FP-2.

Table 2.4 Analysis of the subsite residues in the proteases highlighted in blue are the human proteases.

Proteases	S1					S1'						S2							S3			
Residue no.	279	282	284	288	279	395	416	418	419	448	449	286	392	321	325	327	328	329	415	477	478	479
Falcipains-2	Q	C	S	F	C	V	N	A	V	S	W	W	S	Y	G	L	I	N	L	D	A	F
Falcipains-2B	Q	C	S	F	C	V	N	A	V	S	W	W	S	Y	G	L	I	N	L	D	A	F
Falcipains-3	Q	C	S	F	C	V	N	A	V	S	W	W	S	N	G	Y	I	T	P	E	A	Y
Vivapain-2	Q	C	S	F	C	V	N	A	V	S	W	W	S	N	G	L	I	P	P	E	A	Y
P. knowlesi	Q	C	A	F	C	V	N	A	V	S	W	W	S	T	G	F	I	P	P	E	A	L
P. berghei	Q	C	S	F	C	V	N	A	V	S	W	W	A	F	G	I	L	P	A	Q	G	Y
Chabaupain-2	Q	C	S	F	C	A	N	A	V	S	W	W	A	D	G	I	L	P	A	Q	A	Y
P. yoelii	Q	C	S	F	C	V	N	A	V	S	W	W	A	F	G	I	L	P	A	Q	G	Y
Falcipains-1	Q	C	S	F	C	V	N	S	V	S	W	W	N	F	G	H	P	F	L	E	V	F
P. ovale	Q	C	S	F	C	V	N	S	-	-	-	W	N	F	G	H	P	F	L	-	-	-
Cathepsin-H	Q	C	S	F	C	A	N	A	V	S	W	W	A	Y	G	L	P	S	V	C	A	S
Cathepsin-L1	Q	C	S	F	C	A	D	G	V	S	W	W	A	E	G	L	M	D	M	A	A	S
Cathepsin-K	Q	C	S	F	C	A	N	A	V	S	W	W	A	D	G	Y	M	T	L	L	A	S

FP-2 has 4 subsites S1, S1', S2 and S3 (Table 2.4), subsite residues were defined by selecting residues within 14Å from the catalytic residues Cys, His and Asp (Shah et al. 2011). Residues in subsite S1 and S1' were very well conserved in both *Plasmodium* and human proteases with an exception Asn-416 that is substituted for Asp in cathepsin-L1 (Table 2.4). Residues in subsite S2 and S3 were observed to have more substitutions: Tyr/Phe-321 hydrophobic conserved residues in S2 of *Plasmodium* sequences were substituted for polar acidic Asp and Glu residues in cathepsin-L1 and cathepsin-K respectively. Reside Asp-477 is well conserved in subsite S3 of human-malarial *Plasmodium* proteases is substituted with hydrophobic residues Cys, Leu, Ala in cathepsin H, K and L1 respectively. Phe/Try-479 is conserved in subsite S3 of *Plasmodium* species with an exception of vivapain-2; it is substituted with Ser residue in human proteases. These variances in subsite residues between parasite proteases and their human host are useful in designing specific inhibitors.

2.3.3 Phylogenetic analysis

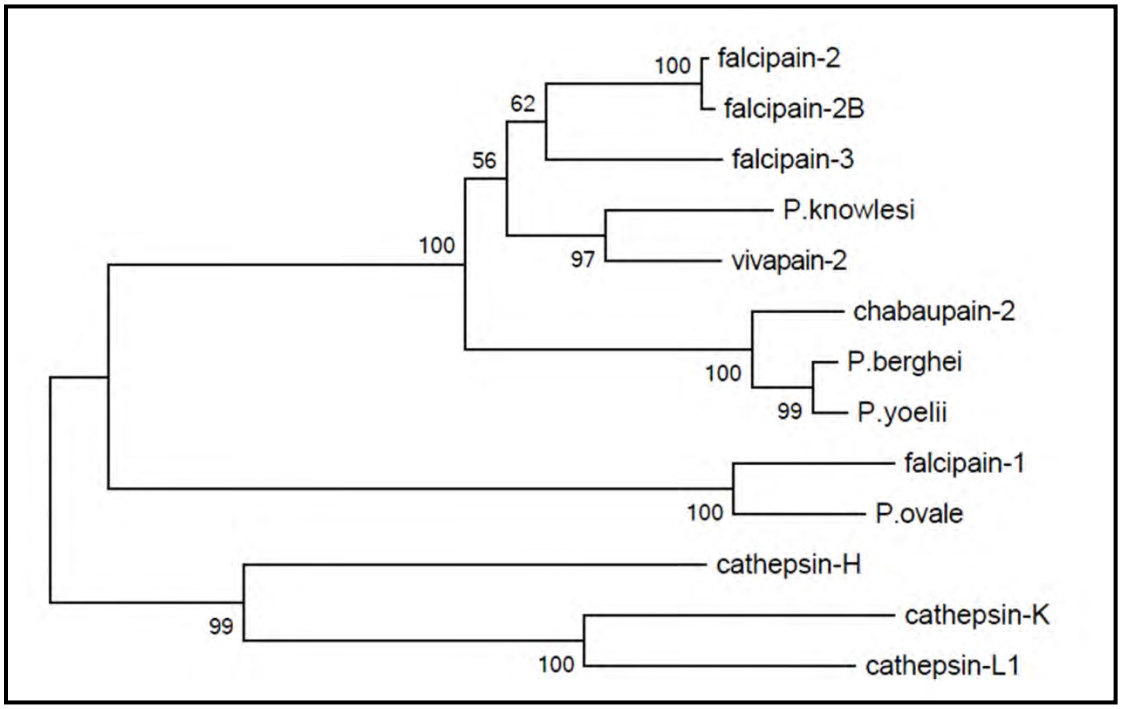

Figure 2.5 Phylogenetic tree of FP-2 and its orthologs in *Plasmodium* and human species based on a PROMALS-3D alignment in the (Appendix A-II). The Neighbour joining tree was generated using MEGA5. Bootstrap statistical analysis values at each inner node specify percentage probability of its occurrence.

The evolutionary relationship of the *Plasmodium* and human proteases was well observed. A region of the sequence alignment with minimal gaps that was likely to depict evolutionary relationship was used to generate the phylogenetic tree as shown in the (Appendix A-II). The Phylogenetic tree shows 4 clusters: *P. ovale* and FP-1 seemingly divergent from the other *Plasmodium* species. The second cluster holds the murine malarial *Plasmodium* species; *P. berghei* and *P. yoelii* branch from a common recent ancestor and are closely related to chabaupain-2. The third cluster is of human malarial proteases; FP-2 and FP-2B which are closely related to FP-3, vivapain-2 and *P. knowlesi*. The fourth cluster having human proteases: cathepsin-L1, cathepsin-K and cathepsin-H. From the analysis of sequence alignments it was noted that residue conservation was grouped in these four clusters. The prodomain region spanning over the active site was also conserved according to these clusters. Hence it is likely that for each of the clusters of *Plasmodium* proteases common peptide inhibitors can be designed.

2.4 CONCLUSION

FP-2 a cysteine protease in *P. falciparum* is a validated drug target for malaria. The prodomain of FP-2 is known to selectively inhibit the protease by obstructing the active site to substrate access. Knowledge of the prodomain residues key in selective inhibition of the mature enzyme in *Plasmodium* is of importance as these are potential in peptidomimetic drug design. The protein sequences of FP-2 and its orthologs in six *Plasmodium* species were analysed. The proteases were found to have a conserved motif in the prodomain section that spans over the active site. Alignment analysis showed that this motif was significantly different in residue composition to human proteases. The subsite residues in *Plasmodium* and human proteases were also analysed and variances were observed in the residue composition of subsite S2 and S3. These dissimilarities in human and *Plasmodium* are crucial to designing specific inhibitors that would target the parasite protease. This analysis was done on protein sequences of the proteases and it would be important to visualise structurally how these residues interact. Therefore there is a need to predict the structure of FP-2 and its orthologs to authenticate the results derived from sequence analysis.

CHAPTER THREE

3. STRUCTURAL ANALYSIS

In this chapter homology models of FP-2 and its orthologs are built to visualise interaction between the prodomain and mature domain. Structural analysis of the regulatory role of the prodomain to mature domain provides insight on the specific inhibition of these proteases. This knowledge will aid in discovery of short peptides that mimic this inhibition as potential antimalarial agents.

3.1 INTRODUCTION

The 3D structure of a protein is crucial to unveiling function of a protein at the atomic level. Protein structure reveals binding sites, domain interactions, ligand binding and spatial relations of subsite residues, these aspects are key to protein function. Advances in high throughput DNA sequencing technology have led to massive increase in generated sequenced data. The protein sequences currently available in biological databases number in the millions whereas the Protein Data Bank holds ~76, 000 protein structures as at October 2011. It is practically impossible to attain the 3D structure of all novel proteins by experimental methods. Computational biology seeks to fill in this sequence to structure gap by predicting the structure and function of uncharacterised proteins (Pierri et al. 2010). The 3D structure of proteins is essential in determining their function and biological significance. Structural analysis of proteins is also known to play a major role in the drug discovery process (Nayeem et al. 2006). Protein structure comparison provides an effective way of finding distant evolutionary related homologs as structures are conserved to a higher degree than sequences (Gherardini et al. 2008). X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and electron microscopy are experimental techniques used to generate the 3D structure of proteins. These techniques are laborious, time consuming and expensive, hence computational techniques like homology modelling are in use to predict the structure of proteins thus subsequently understanding their function.

3.1.1 Protein structure determination

The experimental methods of protein structure determination X-ray crystallography, NMR and electron microscopy are well established and result in optimum and high quality structures (Di Luccio & Koehl 2011). X-ray crystallography is the most accurate technique giving high resolution protein structures. It involves purification and crystallization of the protein that is then exposed to X-rays. The electrons in the protein molecules cause diffraction of the X-ray beam, at different diffraction intensities and directions that are recorded as an electron density map. The electron density maps are interpreted to determine the location of each atom in the molecule and build the protein structure. Crystals are comprised of identical repeating units of the protein molecule referred to as the unit cell, hence amplifying the diffraction of X-rays. The geometry of the unit cell defines the direction of the diffracted beams and the position of atoms in the unit cell affect the intensity of diffraction (Wlodawer et al. 2008). The resulting structure is refined and validated. Quality of crystal structure is assessed by the resolution which is the level of detail at atomic level measured in Å.

Nuclear magnetic resonance (NMR) spectroscopy is a method of determining structure of a protein in solution conditions. This allows studying of proteins in conditions close to their physiological state in living organisms (Wüthrich 2003). NMR capitalises on the magnetic properties of the atomic nuclei by placing the protein molecule in a magnetic field and probing it with radio waves. The resonance frequency is observed and used to characterise individual atomic nuclei. Computational tools are used to interpret the NMR data and build molecular structures that are refined and validated (Di Luccio et al. 2011).

Electron microscopy (EM) is a method based on electron diffraction. This method has the advantage of deriving data on 3D structure of macromolecules that are either too large or heterogeneous for NMR or X-ray crystallography methods (Zhou 2008). The protein sample is prepared either by cryo-freezing (cryoEM) or negative staining with heavy metals for good contrast. EM involves exposing the protein to an electron beam; some energy is absorbed by the

atoms within the protein and the rest is transmitted forming a 2D image. 3D reconstruction is done to the image computationally using methods like single particle reconstruction. 3D images are refinement and verification is done to attain an optimum image. Though experimental techniques in structure determination are reliable, structures of many proteins potential as drug targets are currently unavailable. These structures are essential in assessing „druggability" of the proteins. Therefore homology modelling which is an easier and faster method of attaining protein structure is employed to bridge the gap untill experimentally obtained structures are available (Blundell et al. 2006).

3.1.2 Homology modelling

Homology modelling is based on the principle that the 3D structure of a protein can be predicted given its sequence is similar to proteins of know experimentally determined structure. Homologs diverge from a common ancestor; over time accumulating non-destabilizing mutations hence their protein structures remains relatively similar. Similarity in the backbone structure of proteins increases with higher sequence identity (Bajorath et al. 1993). In the absence of experimentally derived 3D protein structure, homology modelling plays a major role in the processes of drug discovery and this is of interest to this study. The applications of homology modelling in the drug design process include: Identification and analysis of ligand binging sites, validation of these as drug targets and docking ligands to find ideal orientation of compounds in these sites (Hillisch et al. 2004).

Modelling the 3D structure of a protein (target protein) involves a series of steps. To begin with is template identification, which involves searching protein structure databases for a homologous protein of known high-resolution experimentally determined 3D structure (template). The target and template protein sequences are aligned. The main chain of the target protein is modelled based on the coordinates of the template protein and varied regions unrepresented in the template protein are modelled as loop regions. The side chains atoms are built and the resulting model is refined, optimized and verified to assess the quality of the model (Di Luccio & Koehl 2011).

There are several tools available to automate the model building process. Modeller is an efficient homology modelling program that builds 3D models of protein structure from structural alignments and refines obtained models (Sali & Blundell 1993). This program is discussed in greater detail later in this chapter.

3.1.2.1 Template selection

A preliminary and essential stage in model building is selecting a favourable structural template. This involves searching the Protein Data Bank (PDB) for a temple that has high sequence similarity to the target protein. A template is assessed based on its structure resolution and sequence identity to the target protein this determines the quality of the resultant model. Templates with a sequence identity > 30% to the target protein are acceptable as the two proteins are considered to be homologs and are likely to have common 3D structure (Hillisch et al. 2004). Sequence identity between 30-50% between target and template protein results in high quality models that are satisfactory for drug target studies. Sequence identity of 15-30% is considered to be low; hence accurate alignment is required to identify homology between sequences. Models generated at 15-30% identity can be used for protein function studies (Hillisch et al. 2004). Sequence identity below 15% results in models which may not reliably represent the protein structure (Nayeem et al. 2006). In the case that there are many available possible templates, the one with the highest sequence identity, highest resolution and best coverage of the target protein is selected. If no particular template fully structurally represents the target, a combination of multiple templates can be used to model the various domains of the target protein (Moult 2005).

Templates are searched from the protein structure databases using sequence similarity search tools such as PSI-BLAST, HHpred and PDB sequence search. PSI-BLAST is a more sensitive BLAST search that employs position specific scoring matrices (PSSM) profiles. This enables retrieval of distant homologs, with biologically significant sequence similarity hence increasing chances of detecting a structurally annotated homolog (Altschul et al. 1997). However PSI-BLAST is a sequence comparison method and may not effectively detect homologs of known structure, hence protein structure prediction methods like HHpred are used. HHpred

(http://toolkit.tuebingen.mpg.de/hhpred) is a fast and highly sensitive server for structure prediction and detection of structurally annotated homologs (Söding et al. 2005). HHpred server searches several structural databases employing PSIPRED structural predictions and HHsearch (HMM-HMM comparison) increasing sensitivity to detect homologs (for further explanation see Chapter 2.1.4). HHpred results give a list of possible templates ordered by E-value; it also presents accurate structural alignments of the target to the various templates. The best template can then be selected based on resolution, sequence similarity to target and completeness of the structure. PDB sequence search allows the user to use the target protein sequence as a query to search the PDB (Rose et al. 2011). The search gives a list of protein structures that have significant similarity to the target protein sequence; one can then choose a suitable template.

3.1.2.2 Template-target alignment

Having selected a suitable template(s), the sequences of the target and template are aligned. Accuracy of this sequence-structure alignment is a prerequisite for high quality of the resultant model. For optimum alignments structurally equivalent residues need to align, errors made in alignment of residues adversely distort the structure of the model (Venclovas 2003). Target to template alignment takes into consideration both the matching of similar and identical residues and structural correctness of the alignment. It is thus important to use alignment programs that incorporate protein structure prediction such as HHpred and PROMALS-3D (Söding et al. 2005), (Pei et al. 2008). HHpred alignment uses Hidden Markov model (HMM) profiles, which are built from MSA of proteins in the template family. HMM profiles are better at alignment than ordinary PSSMs. This is because HMM profiles incorporate information on position specific probabilities of inserts, deletions and amino acid matches hence ensuring that the alignment algorithm doesn‟t place gaps in structural regions (Dunbrack Jr. 2006). During alignment attention should be given to loop region and regions of low similarity between target and template that are more error prone. Final alignments produced by the alignment programs need be visually reviewed for any mistakes. Manual adjustments can also be done in misaligned regions to attain optimum alignments.

3.1.2.3 Model building

Having attained suitable target-template alignment, the template is used as a guide for the main-chain atom positions this is integrated with loop and side-chain building algorithms (Sali & Blundell 1993). Model building is based on the target-template structural alignment in regions where identical residues align, the templates main-chain and side-chain atom coordinates are copied to the model. In regions where aligned residues differ the main-chain atom coordinates are copied and the side chain is later built. In gap regions of the alignment, that have insertions and deletions, loops are modelled (Xiong 2006).

Loops are mostly found on the surface of the protein as flexible structures connecting between secondary structures (α-helices and β-sheets) (Van & Blundell 1997). Loop regions are challenging to model as they represent regions of the target protein that are structurally different from the template. Loops are often biologically significant as they can be involved in the functional roles of the protein like protein-protein interactions (Di Luccio & Koehl 2011). There are two main approaches used to model loops: database search methods and *ab initio* methods. Database search methods involve; scanning databases of known protein structure (PDB) for protein segments that are of similar length and sequence with the target loop to be modelled. The protein segment need be compatible with the stem residues (end points of main-chain before and after the loop) of the target protein. The best protein segment having suitable orientation and the least steric clashes with the rest of the protein structures is used to model the conformation of the target loop. This loop prediction method is ideal for short-medium sized loops (Van & Blundell 1997). *Ab initio* loop modelling approach involves randomly generating various conformations of the target loop and using an energy scoring function to get the best fit loop. *Ab initio* methods are effective for longer-sized loops and the accuracy of loop prediction depends on the effective sampling of conformational space and evaluation of conformational energy (Xiang et al. 2002). Various methods are used to sample conformational space including; molecular dynamic simulations and Monte Carlo conformational searches (Abagyan & Totrov 1994) ensuring optimum loop prediction.

Once the main-chain has been modelled using the template coordinates and loop building in gap regions, the side-chain conformations are then modelled. Side chain conformation is important especially in the active site region where the interaction of residues is key to the functionality of proteins. The side chains of residues are known to have preferential conformations or rotamers as observed in experimentally derived structures (Bower et al. 1997). The most frequently observed conformations for all 20 amino acids are stored in rotamer libraries. To increase speed and accuracy of deriving residue conformations, backbone dependant rotamer libraries have been developed. These libraries associate main-chain conformations with certain rotamers. To predict the side chain conformation these rotamer libraries are scanned and the most favourable conformations with minimal steric clashes and energy scores are selected (Di Luccio & Koehl 2011).

3.1.2.4 Model refinement

The complete model needs to be refined and optimized in both geometric and energetic aspects so as to attain a structure of stable conformation. Model refinement involves energy minimization and regulating bond lengths and angles of atoms, in order to attain appropriate stereochemistry of the model without distorting its structural conformation (Levitt & Lifson 1969). Local refinement of the structure occurs during loop and side chain building, energy scores are used to access local quality. This allows for detection are correction of errors at the residue level hence improving the model quality. Global refinement is then done to resolve structural irregularities in the entire protein structure. A stable protein conformation (native state) is known to be at local and global minimum of its surface potential energy (Summa & Levitt 2007). Refinement requires methods that effectively sample the conformational space and identify near native-like model conformations. Molecular dynamics simulation (MD) techniques are used to identify near native states enhancing accuracy of conformations selected in the refinement process (Lee et al. 2001). These methods are incorporated into model building programs like Modeller ensuring that optimum models are generated.

3.1.3 Modeller a tool for model building

Modeller is one of the most efficient comparative protein structure modelling programs (Kmiecik et al. 2007). Modeller is a script based program; given the structural alignment file in the pir format, the template atom coordinate files and a script file with the appropriate model building commands the program automatically builds the 3D models of the target (Narayanan Eswer 2006). Modeller uses spatial restraints attained from the template (homology derived restraints) to guide model building of the target protein (Kosmoliaptsis et al. 2011). The forms of spatial restraints used were derived from databases of protein structural alignments, where relationships between equivalent main-chain dihedral angles and Cα-Cα distance of related proteins were obtained. These restraints work together with stereochemistry restraints on the bond length, bond angles and dihedral angles enforced by CHARMM22 to refine the models to proper stereochemistry (MacKerell Jr. et al. 1998). Modelling by spatial restraints is an efficient method of comparative modelling as in addition to homology derived restraints other restraints based on experimentally derived protein structure and general biological knowledge can be incorporated to the process enhancing model quality (Narayanan Eswer 2006). Modeller using its automodel class builds models automatically, going through the backbone, loop and side-chain building steps and finally refining the model. The program‟s output is the atom coordinate files of the models in the PDB format readable by visualisation programs like PyMol (Schrödinger, L.L.C. 2010). Ordinarily several models of the same target are built, hence to select and evaluate the best model Modeller calculates the Discrete Optimized Protein Energy (DOPE) energy. The DOPE score is an atomic-distance-dependent, knowledge-based potential that is derived from a sample of native structures used to assess homology models. DOPE uses an enhanced reference state corresponding to non-interacting atoms in a heterogeneous sphere, whose size reflects that of the sample native structures (Shen & Sali 2006). DOPE score is calculated and normalized by the number of restraints acting on each residue giving a DOPE Z (normalised score). The model with the lowest normalised DOPE Z score is selected; models of medium to high accuracy typically have energy scores below -1 (Eramian et al. 2008).

3.1.4 Model validation

Model validation entails assessing quality of the predicted model. Computational models are typically not as accurate as the true experimentally derived structure hence evaluation of the accuracy of models is import. Validation of proteins involves: checking of stereochemical properties, examining protein folding quality using energy calculations and assessing model compatibility with its amino acid sequence (Kosmoliaptsis et al. 2011). There are a variety of model quality assessment programs (MQAPs) available.

MetaMQAPII (https://genesilico.pl/toolkit/unimod?method=MetaMQAPII) is a program used to assess quality of computational models of protein structure. Unlike most other MQAPs which are based on global evaluation of the protein structure, MetaMQAPII focusses on local areas of inaccuracy that are common in computational models. MetaMQAPII is a meta-predictor that combines results of 8 other validation programs; PROSA, VERIFY3D, ANOLEA, BALA, PROVE, PROQRES, REFINER and TUNE. This aims to combine strengths of different MQAPs whilst eliminating individual flaws. Most of the above MQAPs showed bias to trivial features including: residue depth in the structure, residue hydrophobicity. MetaMQAPII thus combines the various MQAPs and uses a multivariate regression model that controls bias to trivial parameters hence providing a better quality assessment tool (Pawlowski et al. 2008). MetaMQAPII scores predict the absolute deviation of C-α atoms (backbone atoms) in each residue of the model from its corresponding residue in the native protein structure; this is expressed as the root mean square deviation (RMSD) (Moult et al. 2007). MetaMQAPII also gives a Global Distance Test Total score (GDT_TS), which is a measure of the global structure similarity between proteins (Zemla 2003). GDT-TS score is dependent on the length of the protein; hence the optimum score varies with protein size. MetaMQAPII generates a PDB file, in which it colors the structure by quality; colors are incorporating in the B-factor column enabling visualization of errors in the structure. A spectrum of colors ranging from blue to red is used where blue represents highly scored residues and red poorly scores residues. This eases the prediction of regions of lower quality in the model that can then be further refined (Pawlowski et al. 2008).

Structural Analysis and Verification server (SAVS) is a webserver that runs 5 different MQAPs: PROCHECK, VERIFY3D, WHAT_CHECK, PROVE, and ERRAT. This makes it possible to verify many qualities of a model at once (http://nihserver.mbi.ucla.edu/SAVES/). SAVS is user friendly allowing the user to upload the PDB file, select desired MQAPs and it then displays output of all the selected MQAPs. The programs of interest are discussed further. VERIFY3D (http://nihserver.mbi.ucla.edu/Verify_3D/) is a MQAP that uses a statistical approach to measures the compatibility of the 3D model to its amino acid sequence. This is done by characterizing each of the residues in the 3D model based on environment features like solvent exposure, area of side-chain that is buried by other residues and local secondary structure (Bowie et al. 1991). Based on these features each residue position of the structure is categorized in an environmental class. This results in the translation of the 3D structure to a 1D sequence representation of environmental class. A collection of high resolution structures is used as a reference to obtain scores for each of the 20 amino acids in different environmental classes; these are called 3D-1D scores. The 3D-1D scores are plotted against the residue number in a 21-residues window to assess compatibility of the structure with protein sequence (Luthy et al. 1992).

PROCHECK (http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/) is a MQAP that evaluates the stereochemical quality of a protein structure. This includes assessing the bond length, chirality and ψ and ϕ torsion angles. The output given by PROCHECK is Ramachandran plots analyzing overall and residue by residue geometry. A Ramachandran plot has coloring representing different regions where red represents the most favored region. The ideal is to have 90% of the residues in the most favored regions (Yadav et al. 2010).

PROSA (https://prosa.services.came.sbg.ac.at/prosa.php) this is a MQAP that uses knowledge-based force fields to validate the quality of a model (Sippl 1993). These Knowledge based force fields are energy functions derived from statistical analysis of all experimentally determined protein structures in databases. This gives the statistical average energy function of correctly folded, stable proteins that can be used to assess correctness of modeled structures.(Sippl

1995).The program calculates the surface energy of the protein giving a Z-score (see Chapter 3.1.3) that determines global model quality. The Z-score is interpreted by displaying it in a plot containing Z-scores of all experimentally derived protein structures. The protein Z-score can be observed from the plot to see if it is the desired range of native conformations (Wiederstein & Sippl 2007). PROSA also gives a local model quality energy plot, this plots energy scores against the amino acid positions over a 10 and a 40 residue window. Residues with positive energy values show erroneous parts of the structure.

3.1.5 Prodomain – mature domain interaction analysis

The 3D-structure of cysteine proteases in *Plasmodium* were attained via homology modelling and the inhibitory interactions between the prodomain and mature domain are analyzed. One of the tools used for this analysis is the Protein Interaction Calculator (PIC) server (http://pic.mbu.iisc.ernet.in/). The PIC server is used to identify various interactions between residues in a protein structure or between proteins in a complex. The interactions identified include: hydrophobic interactions, ionic interactions, disulphide bridges, aromatic-aromatic interactions, aromatic-sulphur interactions and cation interactions. The PIC server input is a protein coordinate file in the PDB format and the users are allowed to select which interactions are to be calculated. The program out puts all interactions identified in the protein(s) and these interactions can be visualized using a RasMol and Jmol interface (Tina et al. 2007). The PIC is also able to identify solvent accessibility that is if residues lie on the surface or are buried deep in the protein core this is made possible by calculating a depth parameter. The PIC server utilizes the HBOND program (Mizuguchi et al. 1998) to identify hydrogen bonds between main-chain atoms, main-chain to side-chain atoms, and between side-chain atoms. The user can specify distance cutoffs in angstroms for identifying any of the interactions between residues (Tina et al. 2007).

3.2 METHODS

Homology modeling of FP-2 and its orthologs is done and the interactions of prodomain-mature domain are analyzed. The procedures followed and methods used are discussed here.

3.2.1 Template identification

The target protein sequences of (FP-2, FP-2B, FP-3, *P.knowlesi*, vivapain-2, *P. berghei*, Chabaupain-2 and *P. yoelii*) were used as queries to search for templates using the HHpred server. The default settings of 3 iterations using the HHsearch MSA method were used to search the pdb70 HMM database. The alignment mode used was local alignment. The search gave a list of possible templates (homologous proteins of known experimentally derived 3D structure) ordered by E-value. The appropriate template for each target protein was selected based on its percentage similarity to the target sequence, high structure resolution and completeness of the structure as shown in (Table 3.1). The PDB sequence search was also used; the search takes in the protein sequence as a query and displays a list of all 3D structures of similar proteins this confirmed results attained by the HHpred search.

FP-2 and its orthologs showed highest similarity to the template (PDB Id: 2O6X) that is the cysteine protease of *Fasciola hepatica*. However the crystallographic structure 2O6X was incomplete in the C-terminal of the prodomain and did not fully represent the *Plasmodium* mature domain. A second template was thus used so as to adequately represent the mature domain of *Plasmodium.* The second template used was the crystallographic structure of mature domain of FP-2 (PDB Id: 2OUL) or FP-3 (PDB Id: 3BWK) as shown in (Table 3.1). The selected templates were then retrieved from the PDB and verified; this was done to analyze the quality of the structures been used as this would affect the models attained. Template verification was done using PROSA, MetaMQAPII and PROCHECK MQAPs.

3.2.2 Template-target alignment

Once the appropriate templates for modeling were selected, alignment of target and template sequences was done. Alignments were done using the HHpred server that also created .PIR profiles which are an essential alignment format for modeling. The HHpred attained alignments

were compared to alignments generated using the PROMALS-3D program (See Chapter 2.3.2) and slight hand adjustments were done to achieve optimum alignment of residues which is imperative to obtaining good models. Hand adjustment of the template-target alignments is done by checking for any misaligned residues and adjusting them accordingly.

3.2.3 Modeling of the protease structure

Having attained appropriate target to the template alignment, Modeller9v7 was used to calculate the homology models. All computations were carried out on a Linux server equipped with eight Intel(R) Xeon E5506 processors (2.13 GHZ) with 12 GB of RAM. Modeller automatically generated the models once given an input of the alignment in the PIR format, the protein coordinate files of the templates and a python script with the appropriate commands. The python scripts used in modeling are available in (Appendix B-II). 500 models were generated for each of the proteases; incorporating slow refinement to the model building process gave better results. Modeller outputs PDB files that contain the coordinates of the models generated. To select the best of the 500 built models Modeller is used to calculate the DOPE-Z energy scores of the models. A stable protein conformation (native state) is known to be at local and global minimum of its surface potential energy; hence the model with lowest energy is selected. A python script was run to sort the models by their DOPE-Z energy scores (Appendix B-II), to ease picking out models with lowest energy. The models of each target with the lowest energy scores were selected for further validation by other programs.

3.2.4 Model validation

To evaluate the structures of the best models attained, a number of MQAPs were used including MetaMQAP, PROCHECK and PROSA. The MQAP evaluated various properties of the structure detecting areas of the structure that were inappropriately modeled. Validation made it possible to detect mis-oriented side-chains, errors due to misalignment of residues and erroneous main chain conformation. The MQAPs were accessed via web servers, the PDB file holding model atom coordinates was uploaded and the programs validate models automatically. The evaluated models showed that most errors were found to occur in the C-terminal loop region of the

prodomain extending to the N-terminal of the mature domain. This allowed for loop refinement of specific mis-modeled sections. In some cases however mis-modeled sections was due to misalignment of residues hence adjustment were made to alignments and better models built.

3.2.5 Model refinement

Modeller was used to refine loops by rebuilding selected residues in the model. This was made possible by using a python script that specified the residues to be refined (Appendix B-II). Modeller input is the python script and the model coordinate file. 100 loop-refined models were generated and their DOPE-Z energy scores were computed. Generally an improvement was observed in the energy scores attained after refinement. Three models having the lowest energy scores were selected and validated using MetaMQAP, PROCHECK and PROSA. The most satisfactory model for each of the proteases was selected based on validation results. Having attained the most accurate models protein interactions could now be determined.

3.2.6 Interaction analysis

The PIC webserver was used to calculate the interactions in the generated models. The PDB file was uploaded to the server and the interactions are automatically computed. The interactions that were observed indicated that; the prodomain is anchored to the active site region primarily by hydrophobic interactions within 5 Å of the subsite residues. Numerous hydrogen bonds (having a distance of less than 3.5Å between donor and acceptor atoms) were observed between the main-chains and side chains of prodomain and subsite residues. Ionic/electrostatic interactions within less than 6Å of the prodomain and subsite residues were also observed. The prodomain aromatic residues were observed to form aromatic interactions with subsite residues. Analysis of interactions in each of the models was done and comparison was made to residues of human cathepsin-K and cathepsin-L1.

3.3 RESULTS AND DISCUSSION

Results obtained from template selection, target-template alignment, homology modeling and domain interactions are discussed here. PyMol (Schorödinger, L.L.C. 2010) was used for visualization of all models.

3.3.1 Template selection

Two templates were selected for modeling each of the target proteases. The sequence identity and target coverage of the templates is shown in (Table 3.1)

Table 3.1 Template selection for homology modeling, two templates are selected based on the sequence identify to the target protein, resolution and completeness of the structure

Target sequence	Template	Organism	%similarity	Target sequence coverage by template structure	Structure resolution
Falcipain-2	2O6X_A	*Fasciola hepatica*	35	161-479 (484)	1.40
	2OUL_A	*P.falciparum*	100	258-481 (484)	2.20
Falcipain-2B	2O6X_A	*Fasciola hepatica*	35	159-479 (482)	1.40
	2OUL_A	*P.falciparum*	98	256-479 (482)	2.20
Falcipain-3	2O6X_A	*Fasciola hepatica*	35	168-489 (492)	1.40
	3BWK_A	*P.falciparum*	100	263-489 (492)	2.42
Vivapain-2	2O6X_A	*Fasciola hepatica*	37	164-484 (487)	1.40
	3BWK_A	*P.falciparum*	67	258-484 (487)	2.42
P. knowlesi	2O6X_A	*Fasciola hepatica*	38	170-490 (495)	1.40
	2OUL_A	*P.falciparum*	57	268-490 (495)	2.20
P. berghei	2O6X_A	*Fasciola hepatica*	37	151-467 (470)	1.40
	2OUL_A	*P.falciparum*	52	232-467 (470)	2.20
P. yoelii	2O6X_A	*Fasciola hepatica*	36	153-469 (472)	1.40
	2OUL_A	*P.falciparum*	49	234-469 (472)	2.20
Chabaupain-2	2O6X_A	*Fasciola hepatica*	36	151-468 (471)	1.40
	2OUL_A	*P.falciparum*	54	251-468 (471)	2.20

The template (2O6X) which is the crystallographic structure of a cysteine protease (procathepsin L1) in *Fasciola hepatica* (Liver fluke) showed highest similarity to most of the target proteases. The 2O6X structure had 30% - 38 % similarity to target proteases and a relatively high resolution of 1.40Å (Table 3.1). The template had fairly-good sequence coverage of the target proteases,

however it was incomplete in the C-terminal of the prodomain and it did not adequately represent the target sequences in the mature domain. FP-2 and its orthologs have unique features: a 17 residue insert in the N-terminal of the mature domain and a C-terminal 14-residue insert in the mature domain forming an „arm-like" projection. These features were absent in the template 2O6X; hence a second template was selected for modelling to represent the mature domain. The crystallographic structures of the mature domains of either FP-2 (2OUL chain A) or FP-3 (3BWK chain A) were selected as second templates having 41%-100% to the target proteases as shown in (Table 3.1).

The three selected templates *F. hepatica* (2O6X), Falcipain-2 (2OUL) and Falcipain-3(3BWK) were retrieved from the PDB and validated. Assessment of the quality of these crystallographic structures was essential as it determines the quality of the resultant models. Validation allowed us to realise locally stressed regions of template (that is local regions of the template that had high energy values). PROCHECK was used to assess the stereochemistry of the templates and the results are shown in (Figure 3.1). Template-2O6X was found to have 90.5% residues in most favoured region and the remaining 9.5% residues in additional allowed regions. Thus 2O6X was a good quality template having no residues in disallowed regions. Template-3BWK was found to have 85.6% residues in most favoured regions, 13.4% residues in additional allowed regions, 0.5% residues in generously allowed regions and 0.5% residues in disallowed regions. There were two residues in the disallowed region SER-117 and ASP-119 both occurring in loop regions that are away from the active site and would not have effect on the interest region of models generated using this template. The template-2OUL was found to have 85.5% residues in most favoured regions, 13.4% residues in additional allowed regions and 1.0% residues in generously allowed regions. The PROCHECK validation indicated the templates as having acceptable stereochemistry; further validation was done using the MetaMQAPII server. The MetaMQAPII results are shown in (Figure 3.2), the structures are colored by quality to enhance visualization of errors. Blue represents highly scored residues and red poorly scored residues that are likely to be erroneous.

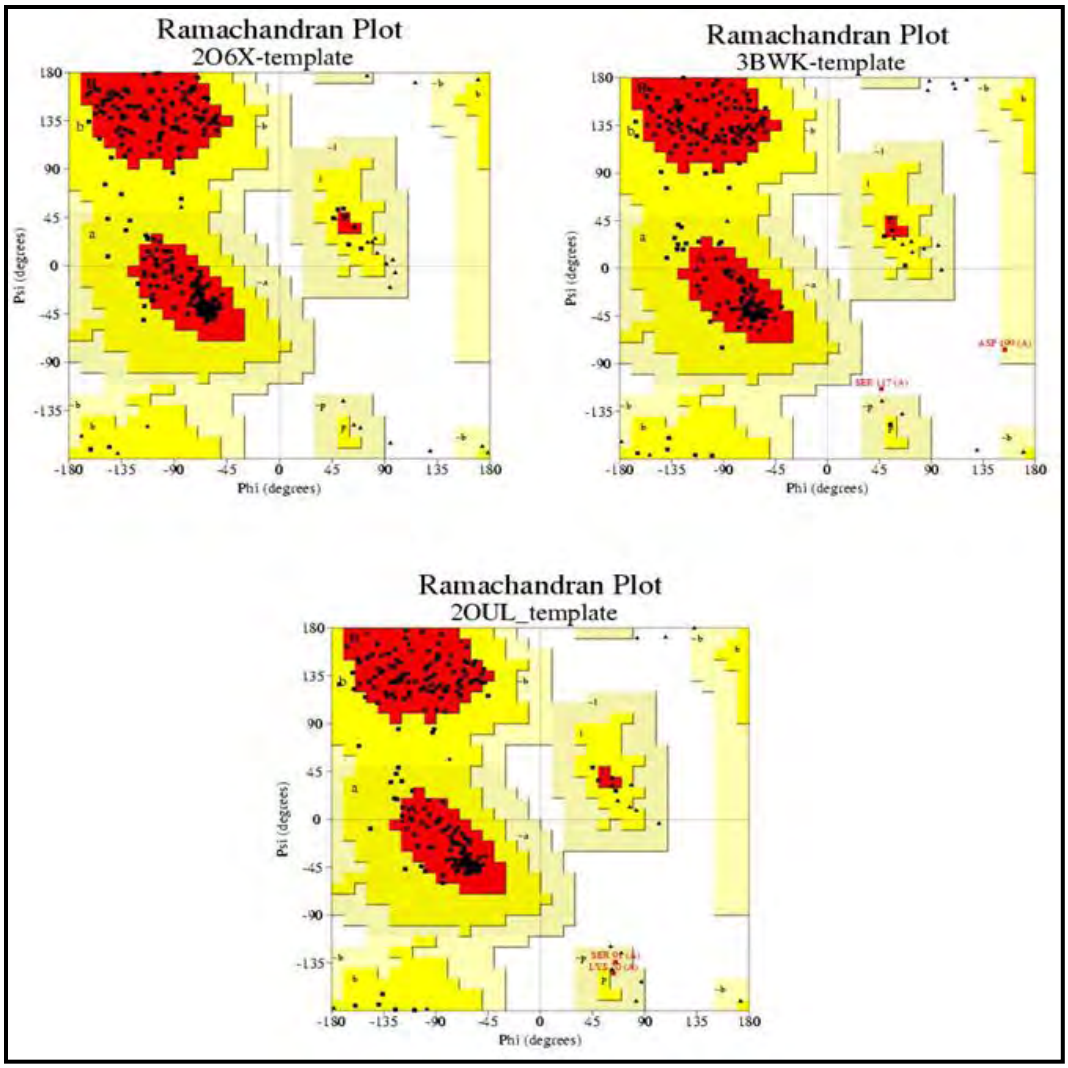

Figure 3.1 Ramachandran plot of templates *F. hepatica*-2O6X (top left), Falcipan-3-3BWK (top right) and Falcipain-2-2OUL (bottom). Red indicates most sterically favoured region, dark-yellow indicates the additional allowed regions, light yellow shows the generously allowed regions and disallowed regions are white.

The MetaMQAPII validation of templates 2OUL and 3BWK gave GDT-TS scores of 66.286 and 64.360 respectively (Figure 3.2). This infers that the structures are of good enough quality to be used for modelling. The subsite region of these structures occurs between the left and right lobe and is mostly blue indicating high scoring residues in this region. The regions that scores poorly (shown in yellow - red) are loop regions that occur in the periphery of the structure away from the subsite residues of interest to this study. These loop regions can be corrected by refinement in the final models that are generated using these templates. The MetaMQAPII validation of templates 2O6X shown at the bottom of (Figure 3.2) gave GDT-TS scores of 69.935. This

template has good structure quality in its mature domain region as depicted by the deep blue colouring. The prodomain structure is also of good quality with an exception of its C-terminal, where there is a 1-turn α-helix that is coloured red. This C-terminal helix region however doesn‟t directly overlay the active site and may not be employed in the study of protease inhibition. Given the GDT-TS scores of 66.286, 64.360 and 69.935 for the templates 2OUL, 3BWK and 2O6X respectively, it is expected that the models generated using these templates will have scores slightly lower than this.

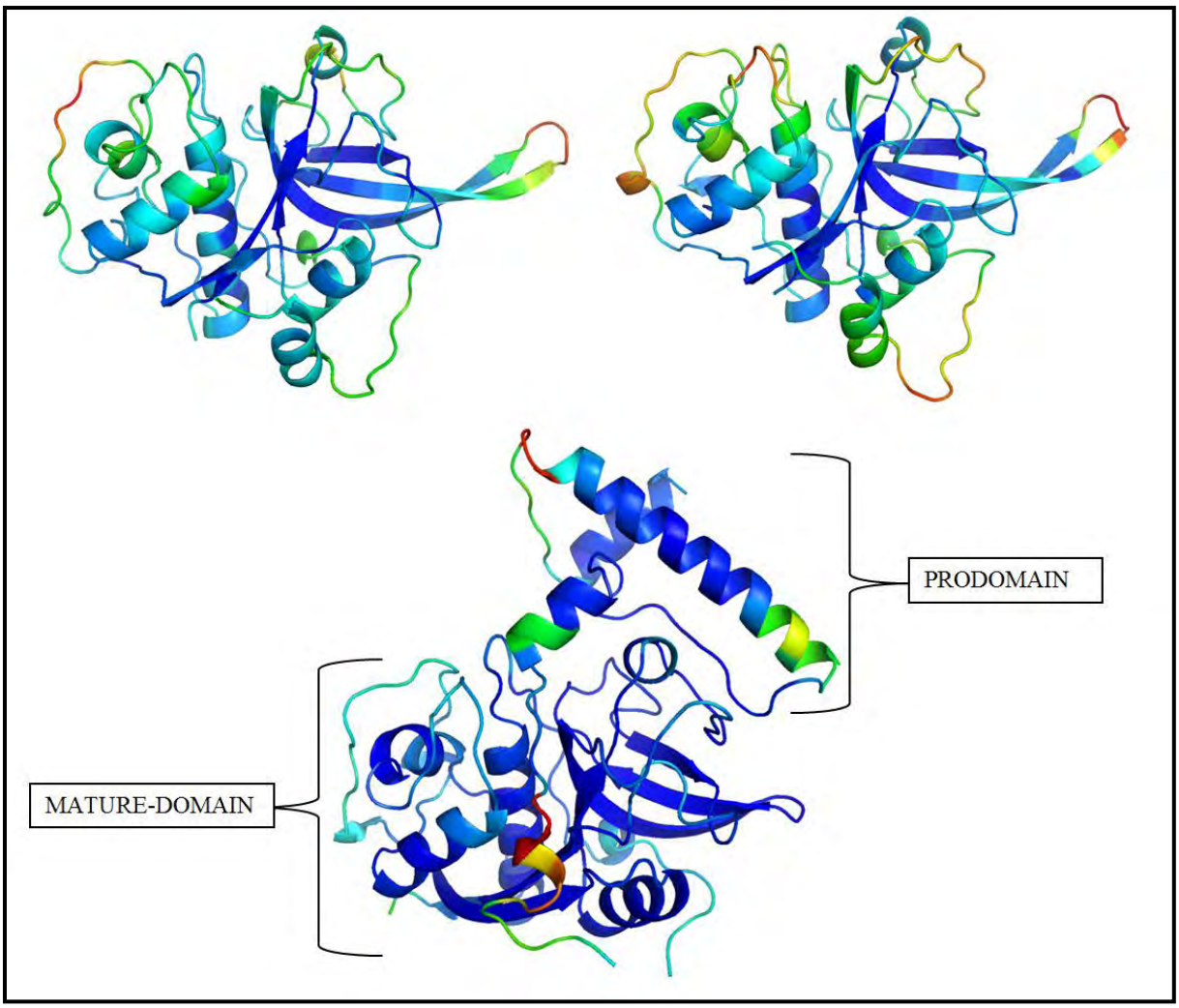

Figure 3.2 MetaMQAPII validation results of templates. Falcipan-3-3BWK (top right), Falcipain-2-2OUL (top left) and *F. hepatica*-2O6X (bottom). The image is coloured by quality in a spectrum of blue to red where blue is highly scored residues and red is poorly scored and erroneous residues. The prodomain and mature domain regions of template 2O6X are indicated. Template 2OUL and 3BWK comprise of the mature domain only and the active site is situated at the groove between the left and right lobes of the structures.

3.3.2 Template-target alignment

Having selected and validated the appropriate templates for modeling, the sequences of the each target and the templates were aligned using the HHpred server. The alignments for each target to the templates are available in the (Appendix B-I).

3.3.3 Model building and refinement

The 3D structures of *Plasmodium* cysteine proteases (FP-2, FP-2B, FP-3, *P.knowlesi*, vivapain-2, *P. berghei*, chabaupain-2 and *P. yoelii*) were built and refined using Modeller. For each of these proteases the best model was selected and validated to check its structure quality. Based on this assessment of the structure loop refinements were performed to improve the model quality. (Table 3.2) gives a detailed summary of the most optimum models attained for each of the proteases under study. The summary indicates the model Dope-Z energy scores calculated by Modeller and some of the scores of model validation that are discussed in greater detail ahead. The 3D structure of the models is shown in figure (Figure 3.3).

Table 3.2 Summary of the homology models attained of the eight proteases, Dope-Z energy scores calculated by Modeller are indicated. GDT-TS validation score by MetaMQAPII server and PROCHECK results are also indicated. The RMSD between the model and template 206X is given (as calculated by superimposing the structures using PyMol).

HOMOLOGY MODEL	Dope Z-score	MetaMQAPII GDT-TS	RMSD	PROCHECK Residues in most favoured region
Falcipain-2A.BL00420001.pdb	-1.012	64.330	0.723	87.8
Falcipain-2B.BL00710001.pdb	-0.896	68.069	0.549	80.1
Falcipain-3.BL00280001.pdb	-1.018	62.655	0.936	86.6
P.knowlesi.BL00240001.pdb	-0.784	62.227	0.640	88.5
Vivapain-2.BL00940001.pdb	-0.823	63.863	0.891	88.4
P.berghei.BL00180001.pdb	-0.771	64.432	0.719	87.5
Chabaupain-2.B99990076.pdb	-0.596	65.645	0.680	86.2
P.yoelii.BL00320001.pdb	-0.374	60.252	0.686	84.8

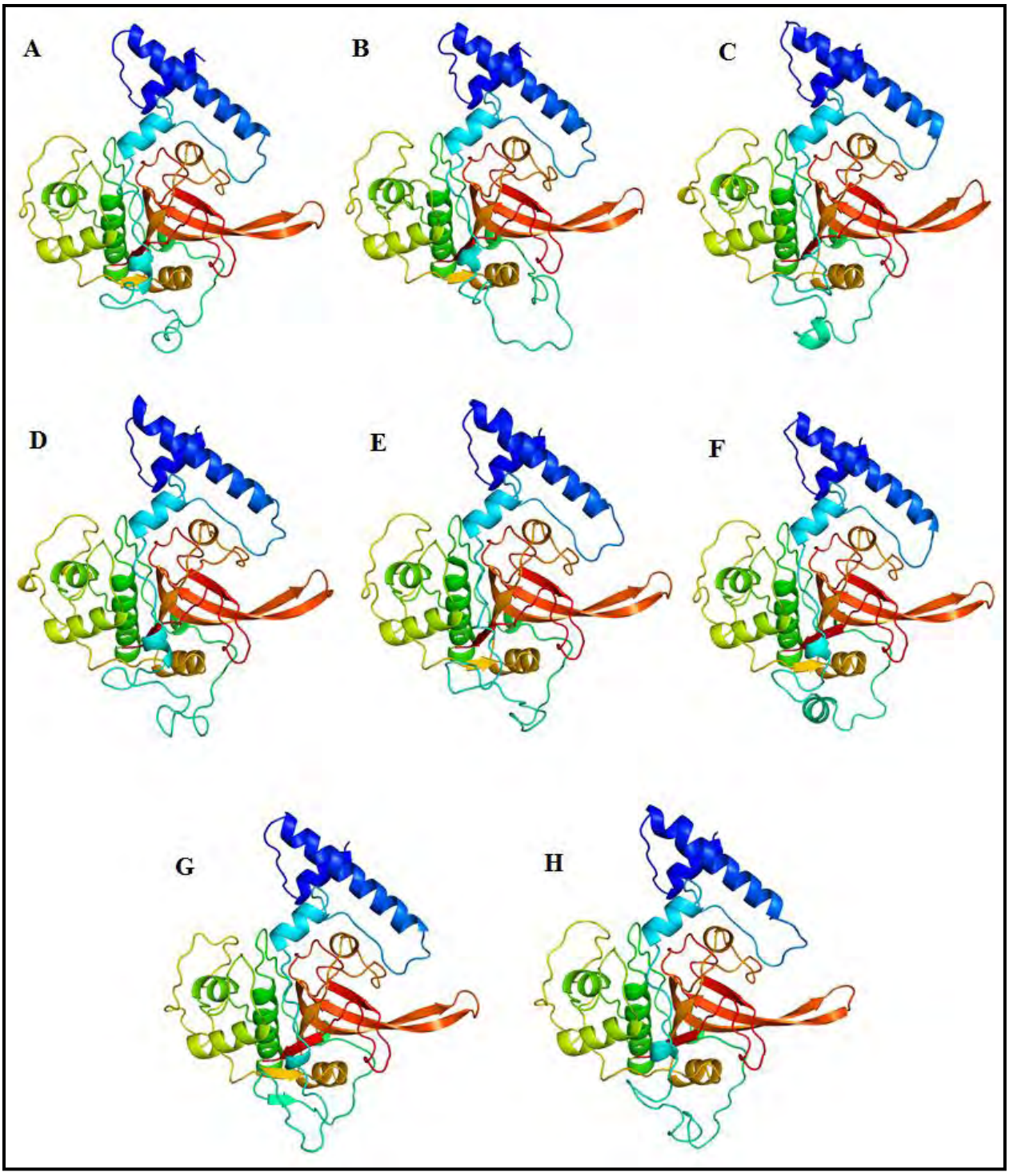

Figure 3.3 3D structure of the best models attained of **(A)** FP-2, **(B)** FP-2B, **(C)** FP-3, **(D)** vivapain-2, **(E)** *P. knowlesi,* **(F)** *P. berghei,* **(G)** chabaupain-2 and **(H)** *P. yoelii.* The models are colored in a spectrum of colors ranging from blue in the N-Terminal to red in the C-terminal. The N-terminal (blue and light blue) regions represent the prodomain, (Green and light green) regions represent the left lobe of the mature domain and the (orange - red) regions show the right lobe of the mature domain.

Figure 3.3 show the best models generated for FP-2 and its 7 orthologs. The general structure of the proteases was analyzed in accordance to the clusters formed in the phylogenetic tree (Figure 2.5). FP2 (Figure 3.3 A) and FP2B (Figure 3.3 B) seem to have a fairly similar topology in the mature domain. However the C-terminal section of the prodomain seems to differ. The α-helix in this C-terminal region is shorter in FP2B that FP2. The structure of FP-3(Figure 3.3C) differs from closely related FP2 and FP2B structures in the mature domain. FP3 is lacking in a β-sheet at the C-terminal of the mature domain that is present in FP2 and FP2B. FP3 also has a helical structure in the left lobe of the mature domain that is not present in FP2 and FP-2B. FP3 structure in the C-terminal of the prodomain is also different from that of FP2 and FP2B this can be attributed to the sequence dissimilarity among these sequences.

The structures of Vivapain-2 (Figure 3.3 D) and P. knowlesi (Figure 3.3 E) were also compared as they seem to be more closely related according to the phylogenetic analysis (Figure 2.5). The structure of *P. knowlesi* has a β-sheet in the C-terminal of the mature domain that is lacking in Vivapain-2. Vivapain-2 also has an additional helical structure in the left lobe of the mature domain that is similar to that in FP-3. The C-terminal of the prodomain of Vivapain-2 has a helical structure that is lacking in *P. knowlesi.* This C-terminal section in *P. knowlesi* has entirely loop structure

The structures of the three murine proteases *P. berghei* (Figure 3.3 F), chaubaupain-2 (Figure 3.3 G) and *P. yoelii* (Figure 3.3 H) were compared. The three protease structures were generally similar in the mature domain. However the structure of *P. yoelii* was lacking in a β-sheet in the C-terminal of the mature domain that was present in chabaupain-2 and *P. berghei*. The C-terminal of the prodomain in *P. berghei* had an additional helical structure that is lacking in chaubapain-2 and *P. yoelii.* In place of this additional helical structure chabaupain-2 had a β-sheet that was absent in *P. yoelii*. The structure of *P. yoelii* in this region is purely comprised of loop structure. The structural differences in the C-terminal of the prodomain of these proteases can be attributed to the varying sequences in this region. This region was also difficult to model due to lack of template coverage hence most of the modeling was done by loop building.

3.3.4 Model validation and interaction analysis

3.3.4.1 FP-2

The best attained 3D model of FP-2 is as shown in part (A) of (Figure 3.3). This model is generated based on the templates 2O6X and 2OUL. The initial model before loop refinement had a DOPE-Z energy score of -0.834 that showed considerable improvement after loop refinement giving an energy score of -1.012. The optimum DOPE-Z score for a structure in its native state is below-1. The model was then validated using PROSA, MetaMQAPII and PROCHECK to assess its structure quality. PROSA validation gave a surface energy Z-score of -7.42 this was plotted along-side Z-scores of all the experimentally derived structures in the PDB (Figure 3.4). The plot showed that the FP-2 model structure is within the desired range of native conformations.

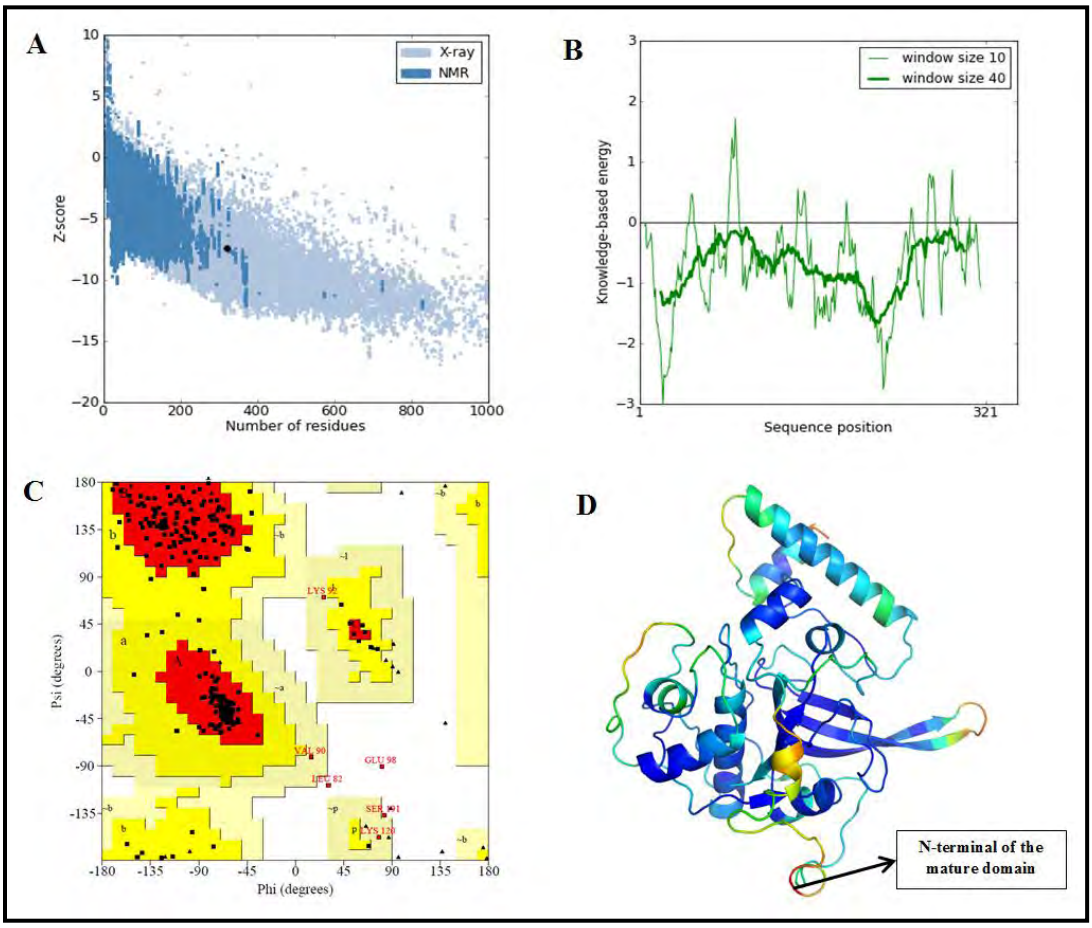

Figure 3.4 Validation of FP-2 structure part (A) is the PROSA global energy plot FP-2 structure is plotted as the black dot other PDB structures are in blue and light blue dots. Part (B) is the PROSA local plot in a 10 and 40 residue window. Part (C) the PROCHECK validation ramachandran plot and part (D) is the MetaMQAPII structure validation colored by quality in a spectrum of blue (good) to red (bad).

PROSA local model quality plot shows the energy per residue. Most residues in the model seemed to be appropriately modeled having low energy scores (below the zero line), with an exception of one major peak observed (Figure 3.4 B). The peak coincided with the N-terminal of the mature domain where a 17 residue insert unique to *Plasmodium* occurs. This region was difficult to model as it was absent in one of the templates (2O6X) used for modelling; hence the predicted structure in this region may not be fully accurate. However the N-terminal of the mature domain does not lie in our area of interest in the structure and will not affect observation of the desired interactions in the structure (Figure 3.5). MetaMQAPII validation gave a GDT-TS score of 64.330 which is a fairly good structure. MetaMQAPII validation reflected the problematic region of this structure (colored red) as the N-terminal of the mature domain this coincided with the PROSA results. Loop refinement seemed to improve this region of the structure though energy scores of residues in this region remained high (Figure 3.4D). PROCHECK verification gave a ramachandran plot with 87.8% of the residue in the most favored region. Judging from the structure in the active site region and the prodomain section overlaying it the structure was considered fit for use in structural analysis.

The PIC enabled calculation of the interactions between the prodomain and mature domain residues. α-3 helix affixes the prodomain to the active site, residues 213-236 in this prodomain region were found to be in interaction with the subsites (Figure 3.5). The subsites S1 and S1' (located in the anterior section of the substrate binding pocket) were occupied by prodomain residues; Arg-213 (S1), Phe-214 (S1'), Tyr-219 (S1'), Phe-222 (S1'), Lys-223 (S1'), Lys-225 (S1) and Tyr-226 (S1). Several hydrogen bonds, hydrophobic and aromatic contacts occur between these prodomain residues and residues in subsite S1 and S1'. Each of these residue interactions were analyzed as depicted in part A of (Figure 3.5). Arg-213 a polar basic prodomain residue forms an ionic interaction with Asp-278 (S1). The side-chain NH2 group of Arg-213 also forms a hydrogen bond to the main-chain carbonyl oxygen in Trp-449 (S1'). Three aromatic prodomain residues Phe-214, Phe-222 and Tyr-226 form aromatic interactions with Trp-449 (S1'). This aromatic stacking anchors the prodomain to the subsite S1'. Phe-214 also forms hydrophobic interactions with non-polar residues Ala-400 and Trp-453 in subsite S1'. Tyr-219 and Phe-222 non-polar prodomain residues form hydrophobic interactions to Val-395 (S1'). Phe-222 is also in

hydrophobic interaction with and Ala-400 (S1'). The Lys-223 and Lys-225 polar basic prodomain residues form an ionic bonds with Asp-413 (S1') and Asp-352(S1) respectively.

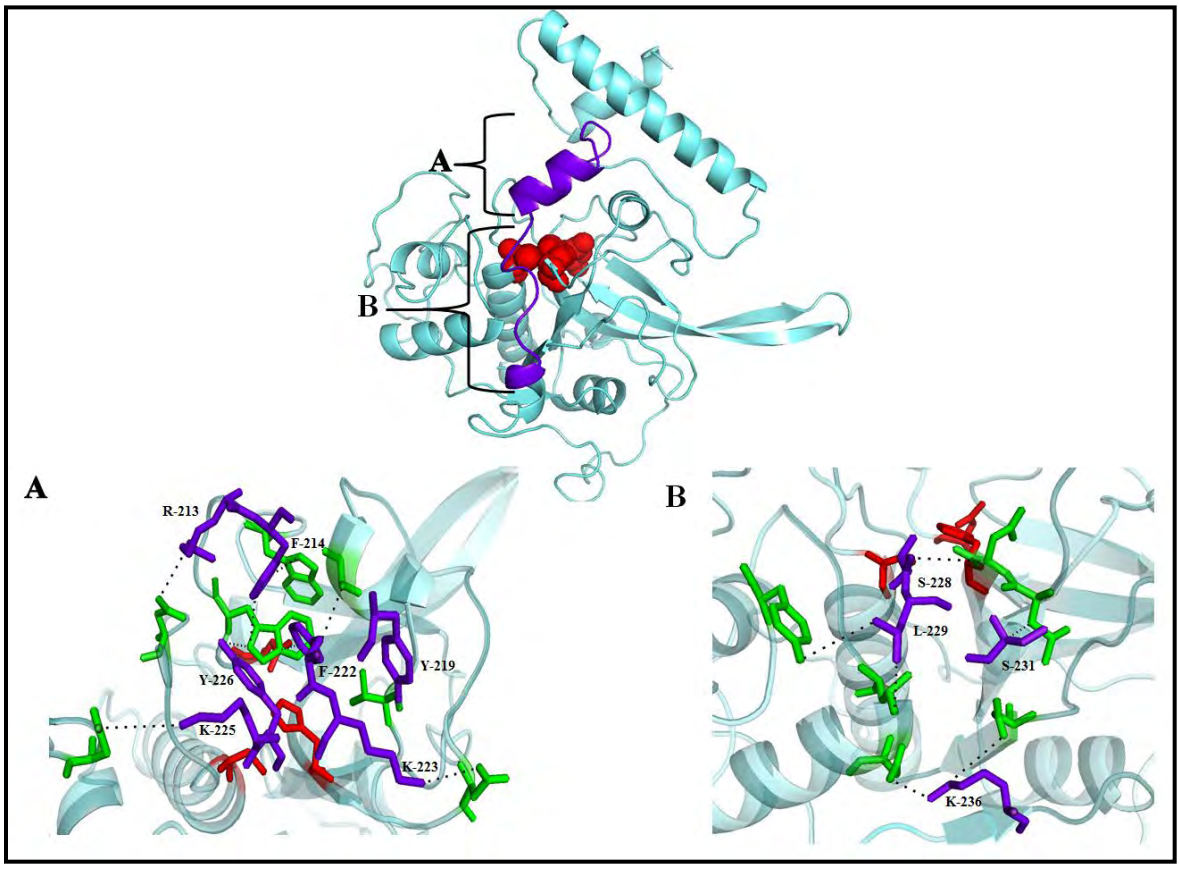

Figure 3.5 Prodomain interactions to FP-2 subsites; the top diagram shows in purple the prodomain region (213-236) interacting with subsite residues (red indicates the active site). A- represents the prodomain residues (purple sticks) interacting with residues in subsite S1 and S1' (green sticks) and B represents interactions in subsite S2 and S3. Prodomain residues are labeled.

The subsite S2 and S3 are occupied by prodomain residues: Ser-228 (S2), Leu-229 (S3), Ser-231 (S2) and Lys-236 (S3) (Part B Figure 3.5). The main-chain amide nitrogen atoms of residues Ser-228 and Ser-231 form hydrogen bonds to the carbonyl oxygen of residues Leu-415 (S2) and Asn-416 (S2) respectively. This anchors the prodomain to subsite S2. The side-chain of Ser-228 was also found at a distance of 3.96Å to the catalytic sulphur of active site Cys-285, there is likely hydrogen bonding between the two residues. Leu-229 a polar prodomain residue forms hydrophobic contact to Tyr-321(S3) and Leu-327 (S3). A salt bridge occurs between Lys-236 a polar basic prodomain residue and Asp-477 (S2). The side-chain of Lys-236 also forms a hydrogen bond to the side chain of Asn-329 (S3).

3.3.4.2 FP-2B

The model of FP-2B is as shown in part B of (Figure 3.3); the model was attained based on templates 2O6X and 2OUL with a DOPE Z score of -0.896. The FP-2B model was validated using PROSA, VERIFY3D and MetaMQAPII (Figure 3.6). PROSA gave a surface energy score of -6.42, that in a plot with other structures from the PDB was shown to lie in a conformationaly acceptable region. PROSA local model quality plot showed a prominent peak region indicative of an inacuracy in the structure (Figure 3.6 B). The region corresponds to the C-terminal of the prodomain extending to the adjacent N-terminal of the mature domain in the structure. This region as discussed earlier was problematic to model due to lack of template coverage hence the structure was predicted by loop building and may not have been entirely accurate. This region is however not of interest to this study. MetaMQAPII validation gave a GDT-TS score of 68.069.

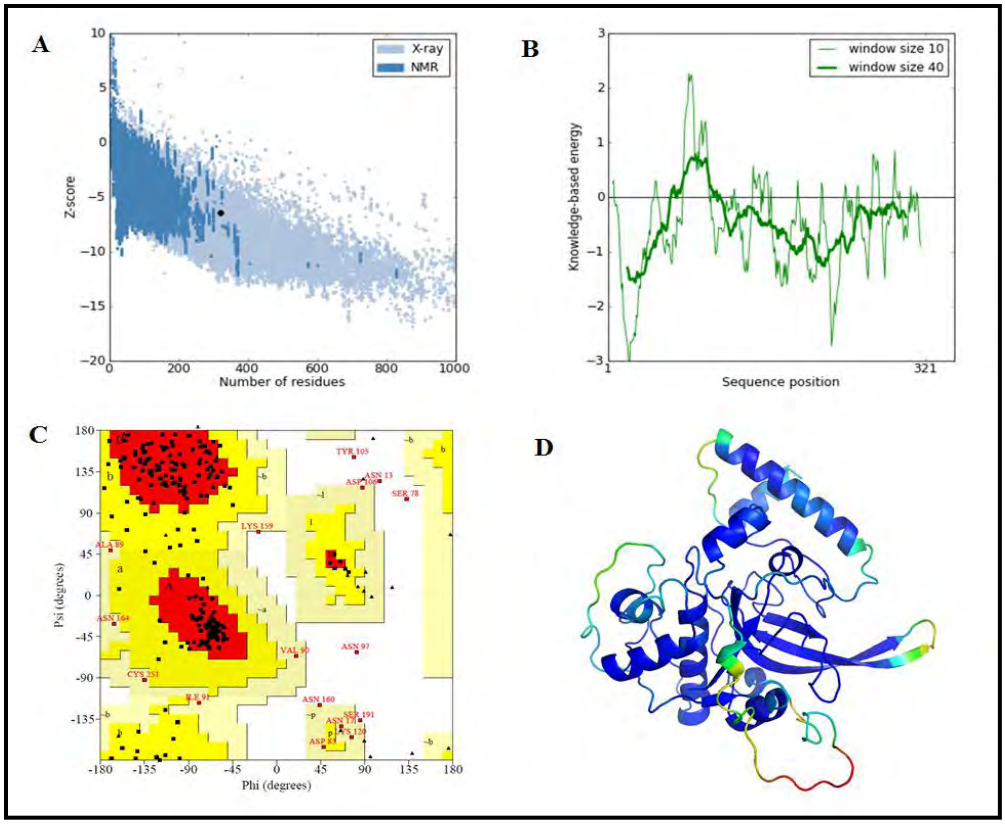

Figure 3.6 Validation of FP-2B structure: part (A) is the PROSA global energy plot FP-2 structure is plotted as the black dot other PDB structures are in blue and light blue dots. Part (B) is the PROSA local plot in a 10 and 40 residue window. Part (C) the PROCHECK validation ramachandran plot and part (D) is the MetaMQAPII structure validation colored by quality in a spectrum of blue (good) to red (bad).

PROCHECK analysis gave a plot showing 80.1% of the residues were in allowed regions. However 2.4% of the model residues were found in dissallowed regions these residues occur in the C-termainal of the prodomain and the N-terminal of the mature domain in the structure. Validation inferred that the structure quality in the subsites and the overlaying prodomain was good enough for use in interaction analysis.

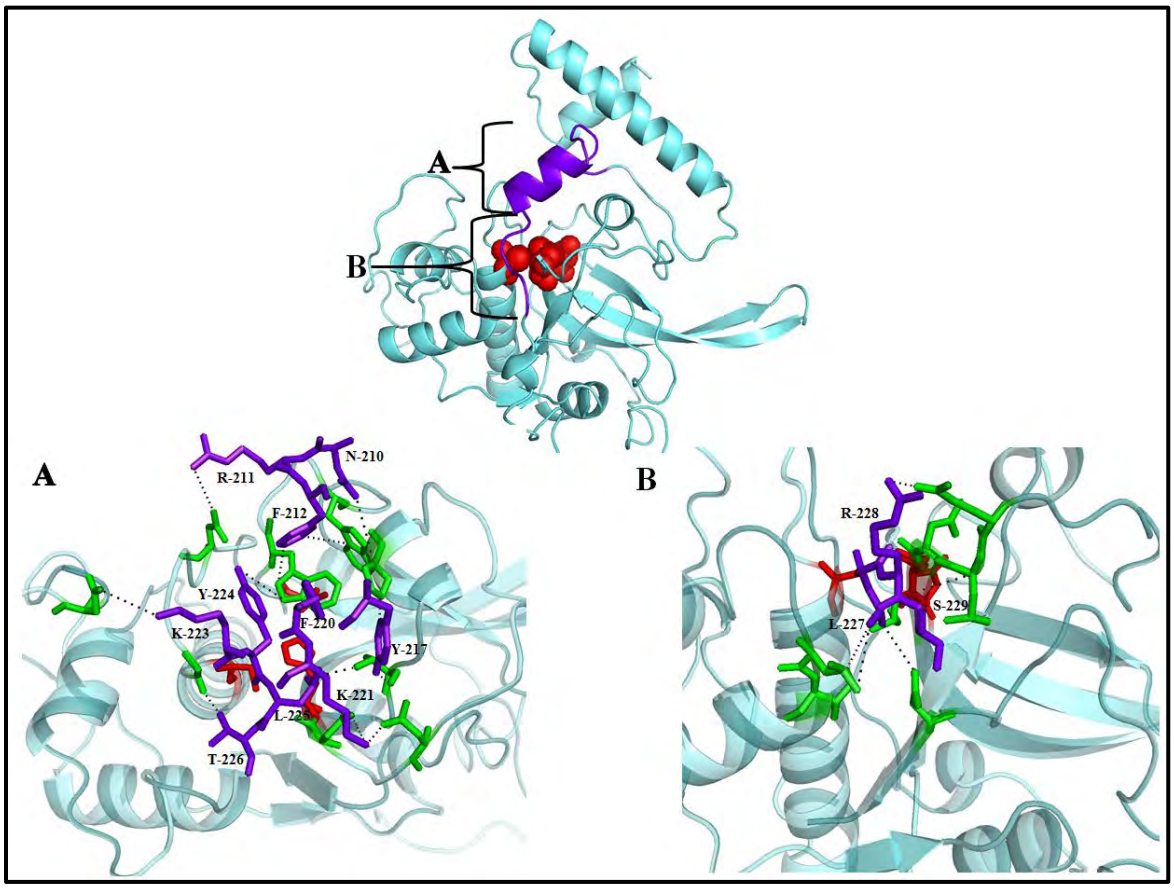

Figure 3.7 Prodomain interactions to FP-2B subsites; the top diagram shows in purple the prodomain region (210-229) interacting with subsite residues (red indicates the active site). A- represents the prodomain residues (purple sticks) interacting with residues in subsite S1 and S1' (green sticks) and B represents interactions in subsite S2 and S3. Prodomain residues are labeled.

Prodomain residues between positions 210-229 were found to be interaction with the subsites (Figure 3.7). Subsite S1 and S1' were occupied by prodomain residues: Asn-210 (S1'), Arg-211 (S1), Phe-212 (S1'), Try-217 (S1'), Phe-220 (S1'), Lys-221 (S1'), Lys-223 (S1), Tyr-224 (S1), Leu-225 (S1') and Thr-226 (S1). Each of these residue contacts to S1 and S1' were analyzed (Figure 3.7). The side-chain of prodomain residue Asn-210 forms a hydrogen-bond to the carbonyl oxygen of Pro-398 (S1'). Arg-211 a basic polar prodomain residue forms an ionic bond

with Asp-276 (S1). Three aromatic residues of the prodomain Phe-212, Phe-220 and Tyr-224 are in aromatic interaction to Trp-447(S1'). These aromatic side-chains residues adequately obstruct subsite S1' and are supported by additional hydrophobic interactions. The aromatic residues Phe-212 and Phe-220 form hydrophobic contacts with Trp-451 (S1') and Val-393 (S1') respectively. A non-polar prodomain residue Tyr-217 forms hydrophobic interactions with Val-393 (S1') and Pro-398 (S1'). The side-chain of Try-217 also forms a hydrogen bond to Asp-411 (S1').

The prodomain interacts via salt bridges to the S1 subsite causing inhibition. Basic polar prodomain residue Lys-221 is forms an ionic bond with Asp-411 (S1) and Glu-412 (S2). Another basic polar prodomain residue Lys-223 forms a salt bridge with residue Glu-350 (S1). The side-chain of Tyr-224 is in hydrogen bonding with the carbonyl oxygen of Lys-278 (S1). The non-polar prodomain residues Leu-225 and Thr-226 are in hydrophobic contact with Gly-281 (S1) and Val-393 (S1') respectively. Hence the Subsite S1 and S1' are effectively obstructed.

Three prodomain residues Leu-227, Arg-228 and Ser-229 occupy the subsite S2 and S3 subsites (Part B Figure 3.7). Leu-227 forms hydrophobic interactions with Leu-413(S2), Ala-416(S2), Leu-325 (S3) and Ile (S3). A salt bridge is formed between prodomain polar basic residue Arg-228 and Glu-412 (S2). The side-chain of prodomain residue Ser-229 is in hydrogen bonding with Asp-475(S2). These residues occupy the groove which substrates follow to access the catalytic residues, by forming contacts to the subsites S2 and S3 hence inhibiting the protease.

3.3.4.3 FP-3

The model of FP-3 is attained using template 2O6X and 3BWK (part C of Figure 3.3). PROSA validation of the model gave a surface energy score -7.11 that was plotted with other structures in the PDB and found to be conformational acceptable (Figure 3.8 B). PROSA local model quality plot showed that the structure was acceptable with most residues having low energy scores. MetaMQAPII validation gave a GDT-TS score of 62.655 and depicted some high energy residues in the N-terminal of the mature domain. This inaccuracy in the model was as a result of

the 3BWK template used in modelling that also showed high residues in the N-terminal of the mature domain (Figure 3.2 A). PROCHECK analysis gave a ramachandran plot placing 86.6% residues in the most favored regions. FP-3 model structure quality was generally acceptable with emphasis been placed on the active site region and the prodomain region overlapping it.

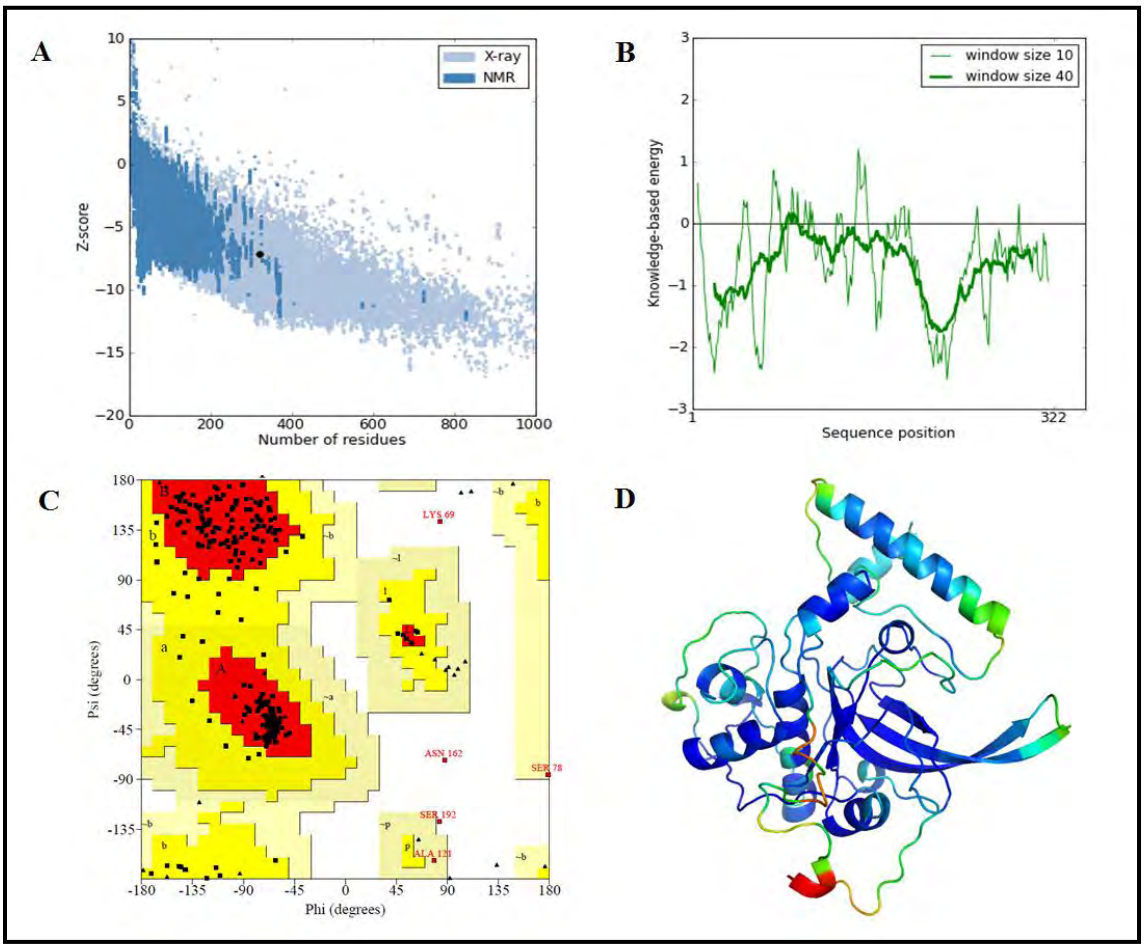

Figure 3.8 Validation of FP-3 structure: part (A) is the PROSA global energy plot FP-2 structure is plotted as the black dot other PDB structures are in blue and light blue dots. Part (B) is the PROSA local plot in a 10 and 40 residue window. Part (C) the PROCHECK validation ramachandran plot and part (D) is the MetaMQAPII structure validation colored by quality in a spectrum of blue (good) to red (bad).

Interactions between the prodomain and the subsite residues were analysed, residues 218-235 of the prodomain were in contact with the active site region (Figure 3.9). Subsite S1 and S1' were found to be occupied by prodomain residues; Asn-218 (S1'), Lys-219 (S1), Phe-220 (S1'), Phe-228 (S1'), Lys-231 (S1), Tyr-232 (S1) and Leu-233 (S1'). The individual interactions of these prodomain residues with the S1 and S1' subsites were analysed. The side-chain of prodomain residue Asn-218 forms a hydrogen bond with Asp-460 (S1'). A salt bridge is formed between

prodomain residue Lys-219 and Asp-286 (S1). The main-chain carbonyl oxygen of Lys-219 also forms a hydrogen bonding with Asp-460 (S1'). Three aromatic prodomain residues Phe-220, Phe-228 and Tyr-232 are in aromatic contact with Trp-457 occupying subsite S1'. Phe-220 and Phe-228 also form hydrophobic interactions with Trp-461 (S1') and Ala-408 (S1') respectively. The side-chain of prodomain residue Lys-231 is in hydrogen bonding with Tyr-332 (S1). Non-polar prodomain residueTyr-232 forms hydrophobic interactions to Ala-288 and Leu-289 in subsite S1. Leu-233 forms hydrophobic interactions with Trp-457 (S1') and Ala-403 (S1').

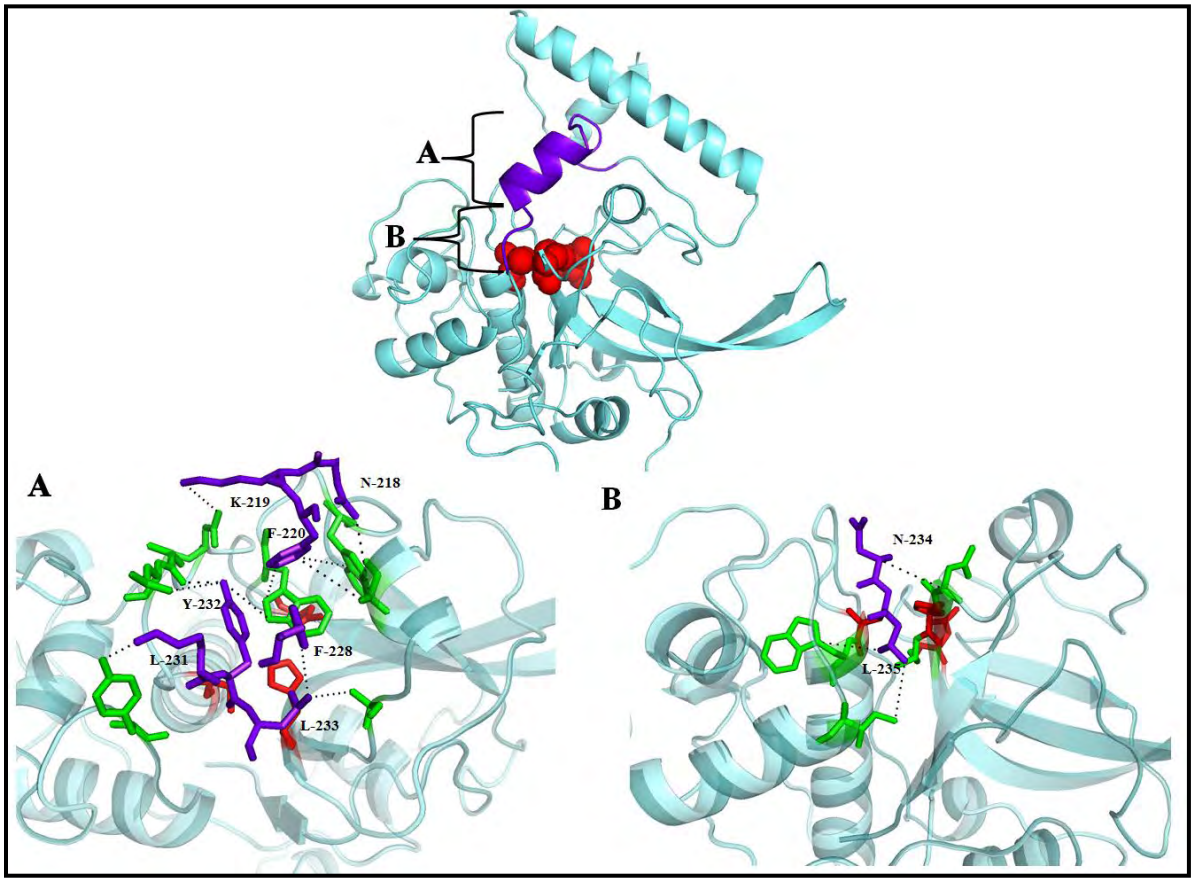

Figure 3.9 Prodomain interactions to FP-3 subsites; the top diagram shows in purple the prodomain region (218-235) interacting with subsite residues (red indicates the active site). A- represents the prodomain residues (purple sticks) interacting with residues in subsite S1 and S1' (green sticks) and B represents interactions in subsite S2 and S3. Prodomain residues are labeled.

The prodomain residues Asn-234 and Leu-235 were found to occupy subsite S2 and S3 (Figure 3.9). Asn-234 forms main-chain hydrogen bonding with the carbonyl oxygen in Asn-424 (S3). The non-polar prodomain residue Leu-235 forms hydrophobic interactions with Trp-294 (S2), Ile-336 (S2) and Ala-426 (S3)

3.3.4.4 P. knowlesi

This structure was modelled using template 2O6X and 2OUL. PROSA structure validation gave a global surface energy score of -6.59, which in a global plot with energy scores of other structures in the PDB showed that the model is in the desired range of structure conformation. PROSA local model quality assessment reveals two peaks at the C-terminus of both the mature domain and the prodomain. A closer look at the residues in this region showed there was low similarity in aligned residues between target and template hence making it error prone and it is likely that the C-backbone and the side-chains of these residues were wrongly predicted. MetaMQAPII validation presented a faulty α-2 helix region in the prodomain this is attributed to low similarity of target to template sequence in this region. MetaMQAPII gave a GDT-TS score of 62.227. PROCHECK analysis showed 88.5%.of the residues were in the most favored regions. The rest of the model was of good quality and acceptable for use in interaction analysis

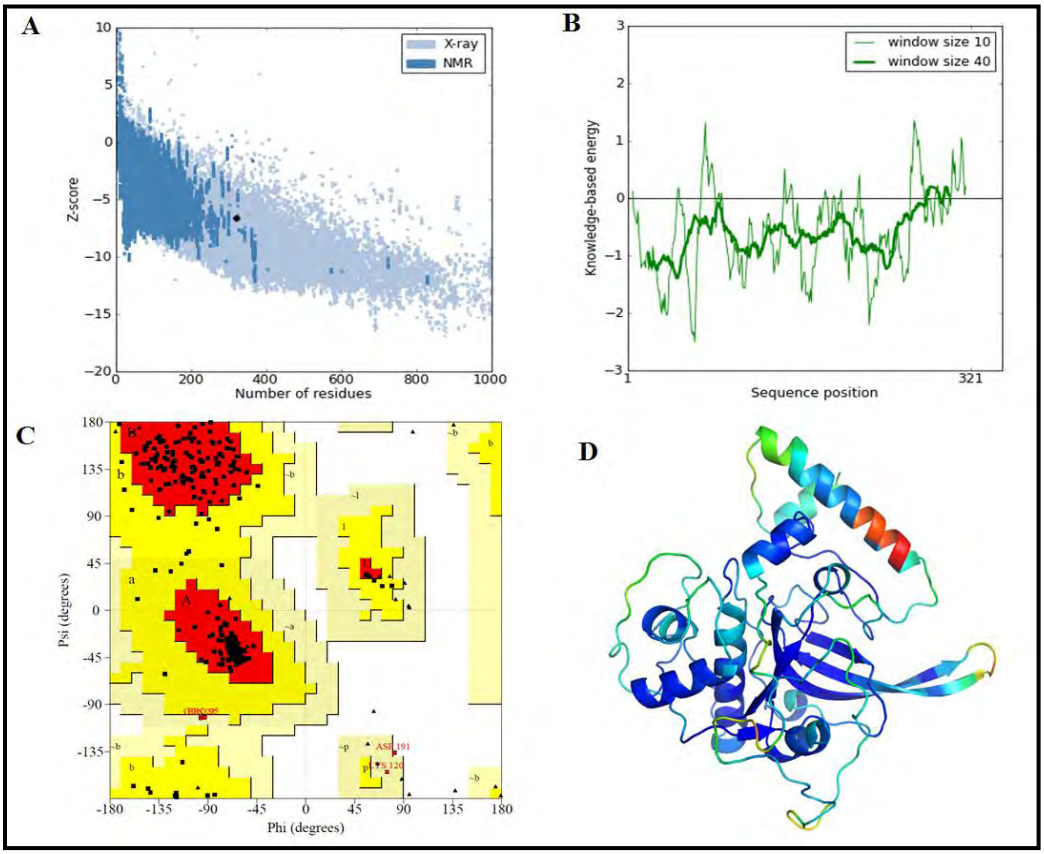

Figure 3.10 Validation of *P. knowlesi* structure: part (A) is the PROSA global energy plot FP-2 structure is plotted as the black dot other PDB structures are in blue and light blue dots. Part (B) is the PROSA local plot in a 10 and 40 residue window. Part (C) the PROCHECK validation ramachandran plot and part (D) is the MetaMQAPII structure validation colored by quality in a spectrum of blue (good) to red (bad).

Interaction analysis of *P. Knowlesi* showed that prodomain residues 220-239 lie in the active site region (Figure 3.11). The subsites S1 and S1' were occupied by prodomain residues; Asn-220 (S1'), Arg-221 (S1), Phe-222 (S1'), Phe-227 (S1'), Phe-230 (S1'), Lys-233 (S1) and Tyr-334 (S1). Each of these residues was analyzed to view interactions with the subsites. The side-chain of prodomain residue Asn-220 is in hydrogen bonding with the main-chain carboxylic oxygen of Tyr-410 (S1'). Three aromatic residues Phe-222, Phe-230 and Tyr-234 form aromatic interactions with Trp-458 (S1). Phe-222 and Phe-227 are also in hydrophobic interaction with Trp-462 and Ala-404 respectively. These aromatic and hydrophobic interactions occupy subsite S1'. Subsite S1 is occupied by prodomain residue Lys-233 that is in hydrogen bonding with Asn-290 (S1). A basic polar prodomain residue Arg-221 forms an ionic bond with Asp-289 (S1).

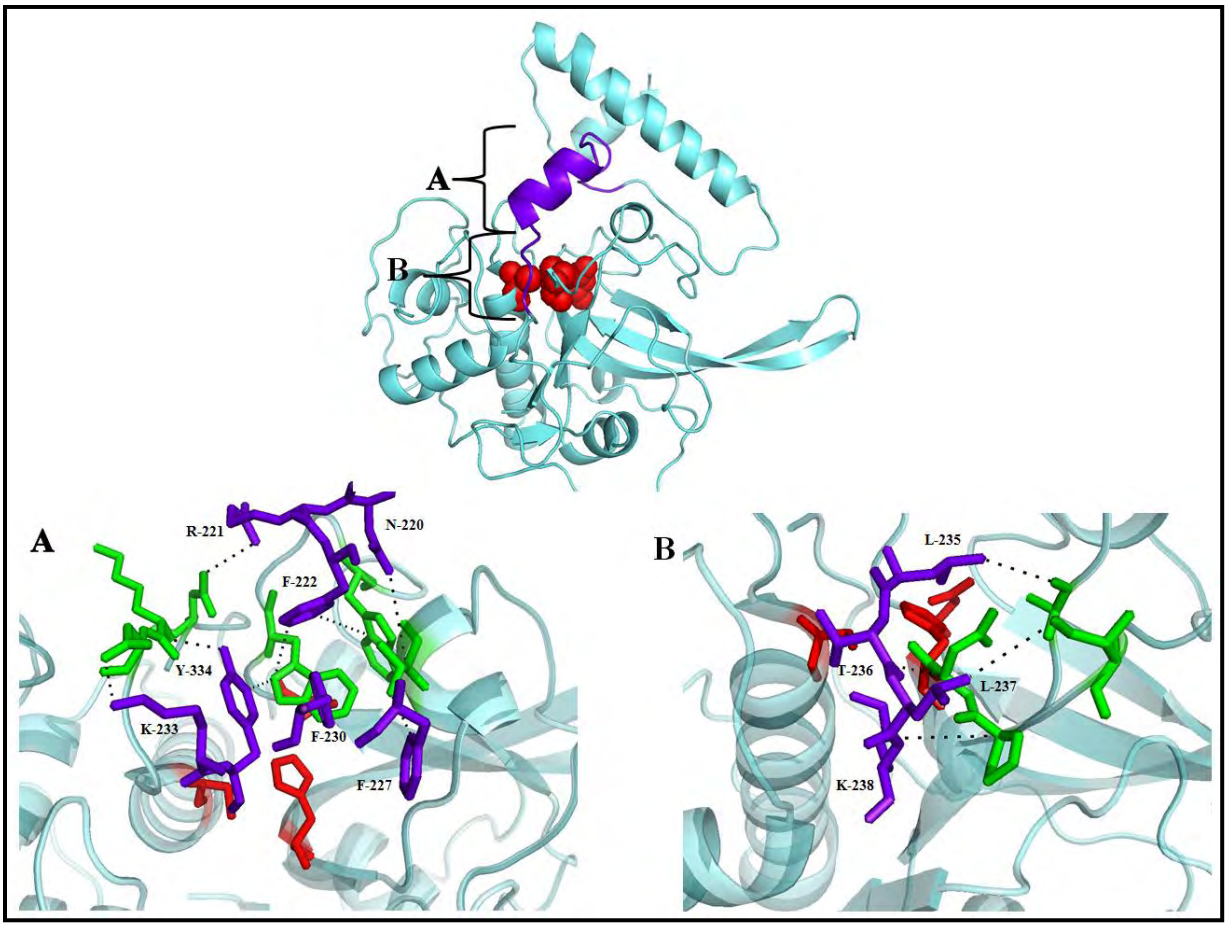

Figure 3.11 Prodomain interactions to *P. knowlesi* subsites; the top diagram shows in purple the prodomain region (220-239) interacting with subsite residues (red indicates the active site). A- represents the prodomain residues (purple sticks) interacting with residues in subsite S1 and S1' (green sticks) and B represents interactions in subsite S2 and S3. Prodomain residues are labeled.

The prodomain residues Leu-235, Thr-236, Leu-237 and Lys-238 were found to occupy the subsite S2 and S3 (Figure 3.11). Non-polar prodomain residues Leu-235 and Leu-237 form hydrophobic interactions with Ala-404 (S2) and Ile-422 (S2) respectively. Thr-236 forms main-chain hydrogen bonding with Asn-425 (S2) .The side-chain of Lys-238 is at a distance of 3.64Å from the catalytic sulphur of Cys-294 there is likely hydrogen bonding. The main-chain of Lys-238 is also in hydrogen bonding with the main-chain carboxylic oxygen of Pro-424 (S2)

3.3.4.5 Vivapain-2

The model of vivapain-2 was attained based on template 2O6X and 3BWK. PROSA structure validation gave a global surface energy score of -6.78 that in a global plot showed that the model lied in the desired range of structure conformation (Figure 3.12 B).

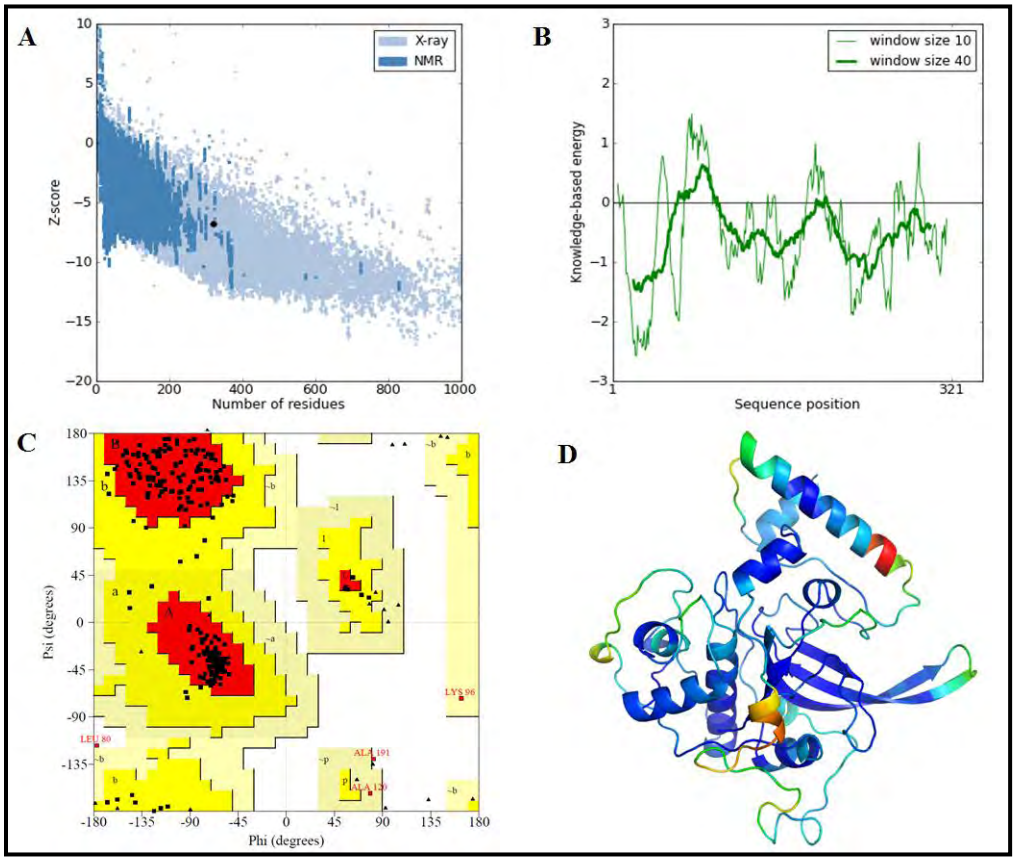

Figure 3.12 Validation of vivapain-2 structure: part (A) is the PROSA global energy plot FP-2 structure is plotted as the black dot other PDB structures are in blue and light blue dots. Part (B) is the PROSA local plot in a 10 and 40 residue window. Part (C) the PROCHECK validation ramachandran plot and part (D) is the MetaMQAPII structure validation colored by quality in a spectrum of blue (good) to red (bad).

PROSA local model quality assessment revealed that most residues in the structure had minimal energy. However a prominent peak was observed at the C-terminus of the prodomain. This coincided with MetaMQAPII validation that showed high energy scores at the C-terminal and α-2 helix of the prodomain. The target residues in this region had low similarity in to those of the template hence it is likely that the C-backbone and the side-chains of these residues were not accurately modelled. MetaMQAPII validation gave a GDT-TS score of 63.869. PROCHECK analysis showed 88.4% of the residues were in the most favored regions. The rest of the model was of good quality and acceptable for use in interaction analysis.

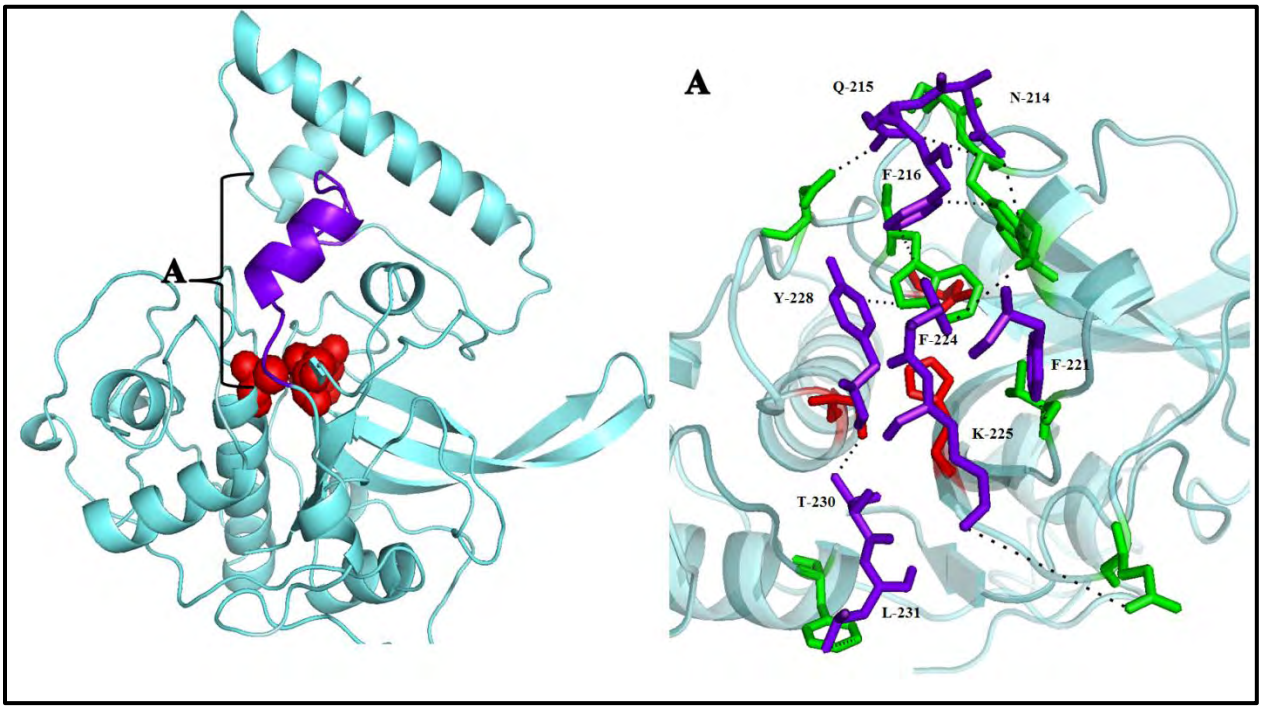

Figure 3.13 Prodomain interactions to vivapain-2 subsites; the top diagram shows in purple the prodomain region (214-231) interacting with subsite residues (red indicates the active site). A-Represents the prodomain residues (purple sticks) interacting with residues in subsites (green sticks). Prodomain residues are labeled.

Interaction analysis showed that prodomain residues (214-231) were in interaction with the active site (Figure 3.13). The prodomain residues in contact with subsite S1 and S1' include: Asn-214 (S1'), Glu-215 (S1), Phe-216 (S1'), Phe-221(S1'), Phe-224 (S1'), and Tyr-228 (S1'). The individual interactions of these residues were observed. The side-chains of prodomain residues Asn-214 and Glu-215 forms hydrogen bonds to the main-chain carboxyl oxygen atoms of Ala-401 (S1') and Trp-452 (S1') respectively. Glu-215 is also in hydrogen bonding with the side-

chain of Asp-281 (S1). Three aromatic prodomain residues Phe-216, Phe-224 and Try-288 form aromatic interactions with Trp-452 anchoring the prodomain to the subsite S1'. Phe-216 was also found to be in hydrophobic interactions with Trp-456 (S1'). Non-polar prodomain residues Phe-221 and Phe-224 both form hydrophobic interactions with Val-398 and Ala-403 in subsite (S1). The prodomain residues Lys-225 (S2), Thr-230 and Leu-231 were found to occupy the subsite S2 and S3. A salt bridge is formed between Lys-225 and Glu-418(S2). Thr-230 is in side-chain hydrogen bond with the catalytic sulphur Cys-288. Leu-231 is in hydrophobic interaction with Phe-330 (S3) (Figure 3.13)

3.3.4.6 P. berghei

The model of *P. berghei* is attained using template 2O6X and 2OUL. PROSA validation gave a surface energy score -7.82 that was plotted with other structures in the PDB and found to be conformational acceptable (Figure 3.14 B). PROSA local model quality plot showed that the structure was good most residues having low energy scores. MetaMQAPII validation gave a GDT-TS score of 64.432 and depicted some high energy scoring residues in the of α-2 helix of the prodomain. A closer look at the sequence alignment in this region showed low similarity between template and target. PROCHECK analysis gave a ramachandran plot placing 87.5% residues in the most favored regions. This model structure was very well modelled and the validation showed good structure quality acceptable for interaction analysis (Figure 3.14 D)

The prodomain residues (200-219) were observed to interact with the subsites (Figure 3.15). Subsites S1 and S1' were occupied by prodomain residues; Asn-200 (S1'), Phe-202 (S1'), His-207 (S1'), Phe-210 (S1'), Lys-211 (S1'), and Try-214 (S1). The individual interactions of these prodomain residues with the S1 and S1' subsites were analysed. The side-chain of Asn-200 interacts via hydrogen bonds to carboxylic oxygen in the main-chain of Glu-287 (S1'). The aromatic prodomain residues Phe-202, Phe-210 and Tyr-214 form aromatic interactions to Typ-436 (S1'). The three prodomain residues Phe 202, Phe-210 and Tyr-214 were also in hydrophobic interaction with Trp-440, Val-382 and Trp-440 in subsite S1' respectively. Two salt

bridges occur between prodomain residues His-207 and Lys-211 with subsite residues Asp-384(S1') and Asp-400 (S2).

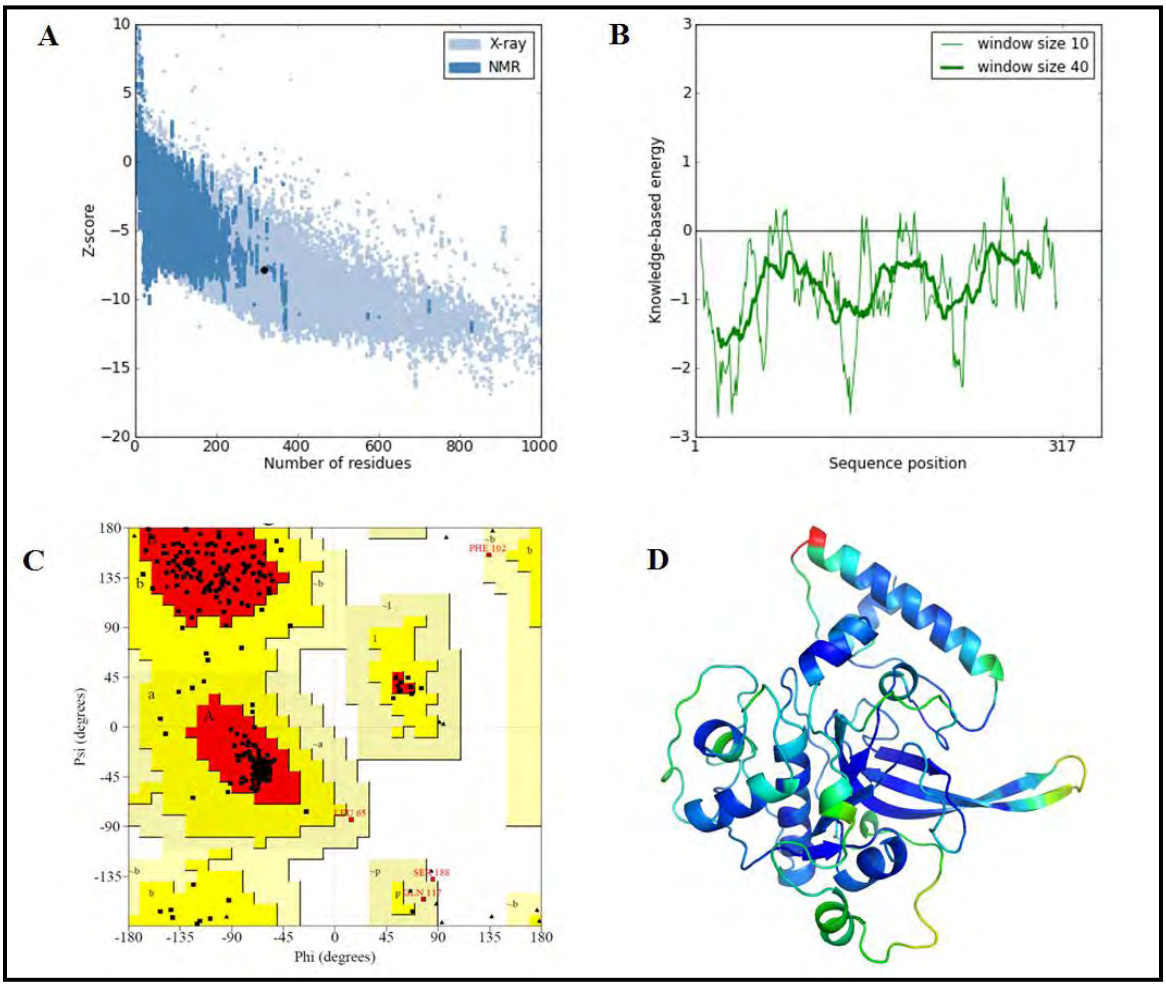

Figure 3.14 Validation of *P. berghei* model: part (A) is the PROSA global energy plot FP-2 structure is plotted as the black dot other PDB structures are in blue and light blue dots. Part (B) is the PROSA local plot in a 10 and 40 residue window. Part (C) the PROCHECK validation ramachandran plot and part (D) is the MetaMQAPII structure validation colored by quality in a spectrum of blue (good) to red (bad).

S2 and S3 were occupied by prodomain residues Leu-215, Asn-217 and Leu-219. Leu-215 forms hydrophobic interactions with Ala-250 (S3) (Figure 3.15). Asn-217 is in main-chain hydrogen bonding with Gly-313 (S3). The side-chain of Asn-217 is also at a distance of 3.77Å to the catalytic sulphur of Cys-272 and is likely in hydrogen bonding. Leu-219 is in hydrophobic interaction with Phe-401 and Ala-402 (S2)

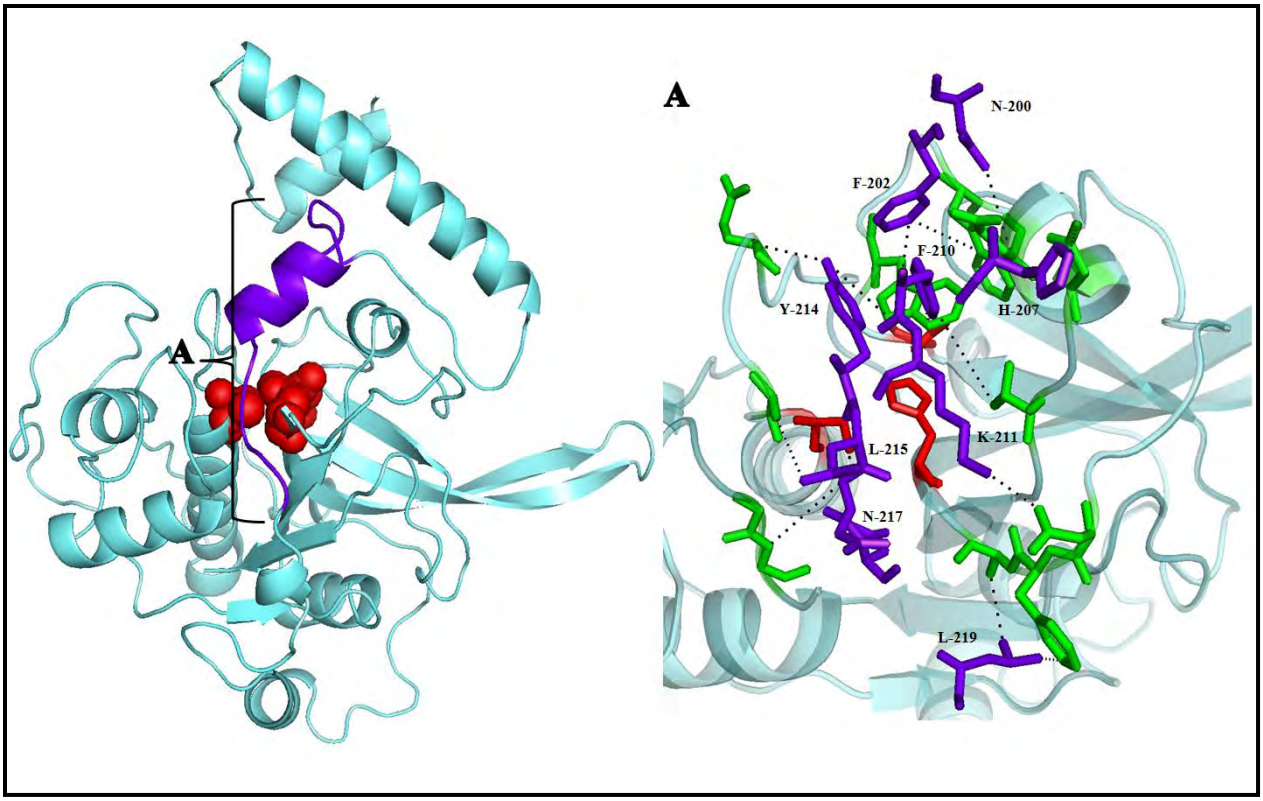

Figure 3.15 Prodomain interactions to P. berghei subsites; the top diagram shows in purple the prodomain region (200-219) interacting with subsite residues (red indicates the active site). A- Represents the prodomain residues (purple sticks) interacting with residues in subsites (green sticks). Prodomain residues are labeled.

3.3.4.7 Chabaupain-2

The model of chabaupain-2 was modelled based on templates 2O6X and 2OUL. PROSA validation gave a surface energy score -7.19 that was plotted with other structures in the PDB and found to be conformational acceptable (Figure 3.16 B). PROSA local model quality plot showed that the structure was good most residues having low energy scores with an exception of the C-terminal of the prodomain. MetaMQAPII validation gave a GDT-TS score of 65.645 and depicted some high energy residues in the C-terminal of the prodomain and in the α-2 helix. PROCHECK analysis gave a ramachandran plot placing 86.2% residues in the most favored regions. This model structure quality was very well modelled the validation showed good structure quality and was acceptable for interaction analysis

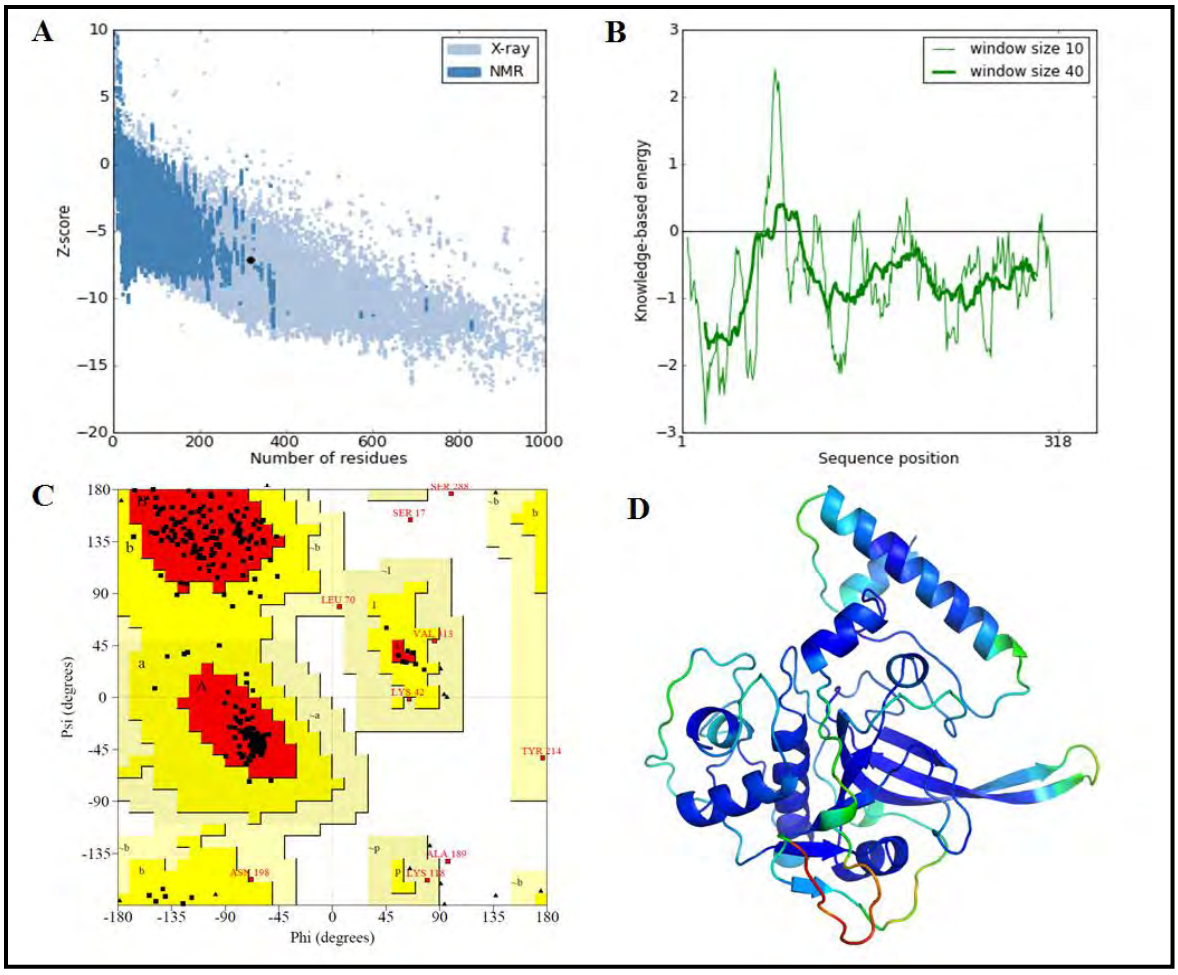

Figure 3.16 Validation of chabaupain-2 model: part (A) is the PROSA global energy plot FP-2 structure is plotted as the black dot other PDB structures are in blue and light blue dots. Part (B) is the PROSA local plot in a 10 and 40 residue window. Part (C) the PROCHECK validation ramachandran plot and part (D) is the MetaMQAPII structure validation colored by quality in a spectrum of blue (good) to red (bad).

Interactions were evaluated residue (201-220) of the prodomain were found to be in interaction with the active site region of the protease. The S1 and S1' were occupied by the prodomain residues: Asn-201 (S1'), Pro-202 (S1'), Phe-203 (S1'), Phe-211 (S1') and Tyr-215 (S1). The interactions of each residue were defined. The side-chain of Asn-201 formed a hydrogen bond to the main-chain carboxylic oxygen of Glu-238 S1'). Pro-202 forms a hydrophobic interaction with Ala-440 in (S1'). Three aromatic prodomain residues Phe-203, Phe-211 and Try-215 form aromatic interactions to Trp-437. Phe-203 is also found in hydrophobic interaction with Trp-441. The side-chain of Try-215 forms a hydrogen bond to the main-chain of Lys-268 (S1).

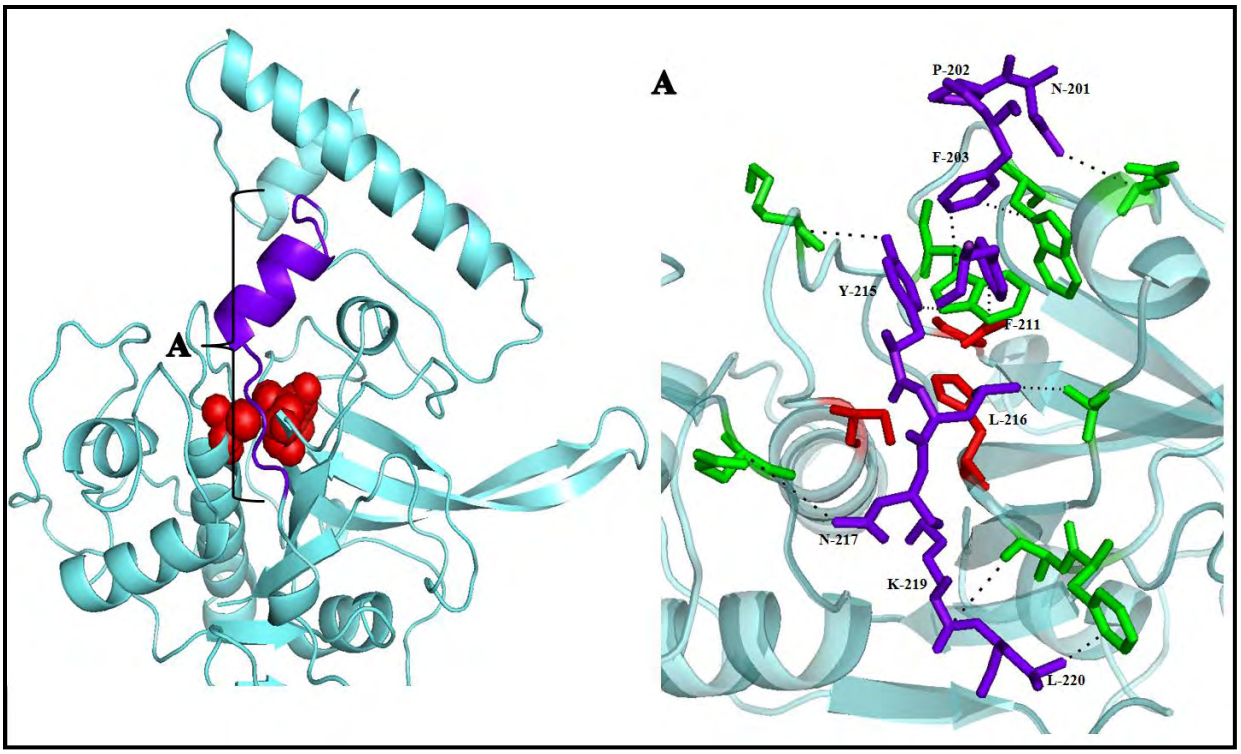

Figure 3.17 Prodomain interactions to chabaupain-2 subsites; the top diagram shows in purple the prodomain region (201-220) interacting with subsite residues (red indicates the active site). A- Represents the prodomain residues (purple sticks) interacting with residues in subsites (green sticks). Prodomain residues are labeled.

The residues of the prodomain in contact with S2 and S3 include Leu-216, Asn-217, Lys-219 and Leu-220 (Figure 3.17). Non-polar prodomain residues Leu-216 and Leu-220 form hydrophobic interactions with Ala-383 and Phe-404 respectively in subsite S2. The side-chain of Asn-217 is in hydrogen bonding with the carboxylic oxygen of Glu-312 (S3). Lys-219 is in main-chain hydrogen bonding with Ala-403 (S2)

3.3.4.8 P. yoelii

The model of *P. yoelii* was modelled based on templates 2O6X and 2OUL. PROSA validation gave a surface energy score -7.09 that was plotted with other structures in the PDB and found to be conformational acceptable (Figure 3.18 B). PROSA local model quality plot showed that the structure was good most residues having low energy scores with an exception of the C-terminal

of the prodomain. MetaMQAPII validation gave a GDT-TS score of 60.252 and depicted some high energy residues in the C-terminal of the prodomain. This C-terminus of the prodomain was missing in the template used hence it is likely that the predictions in this region was erroneous. PROCHECK analysis gave a ramachandran plot placing 84.8% in the most favored regions. However 2.4% residues were found in disallowed regions a close look at this residues shows they occurs from residue 77-89 which is the C-terminal of the prodomain. This model structure quality was very well modelled the validation showed good structure quality and was acceptable for interaction analysis

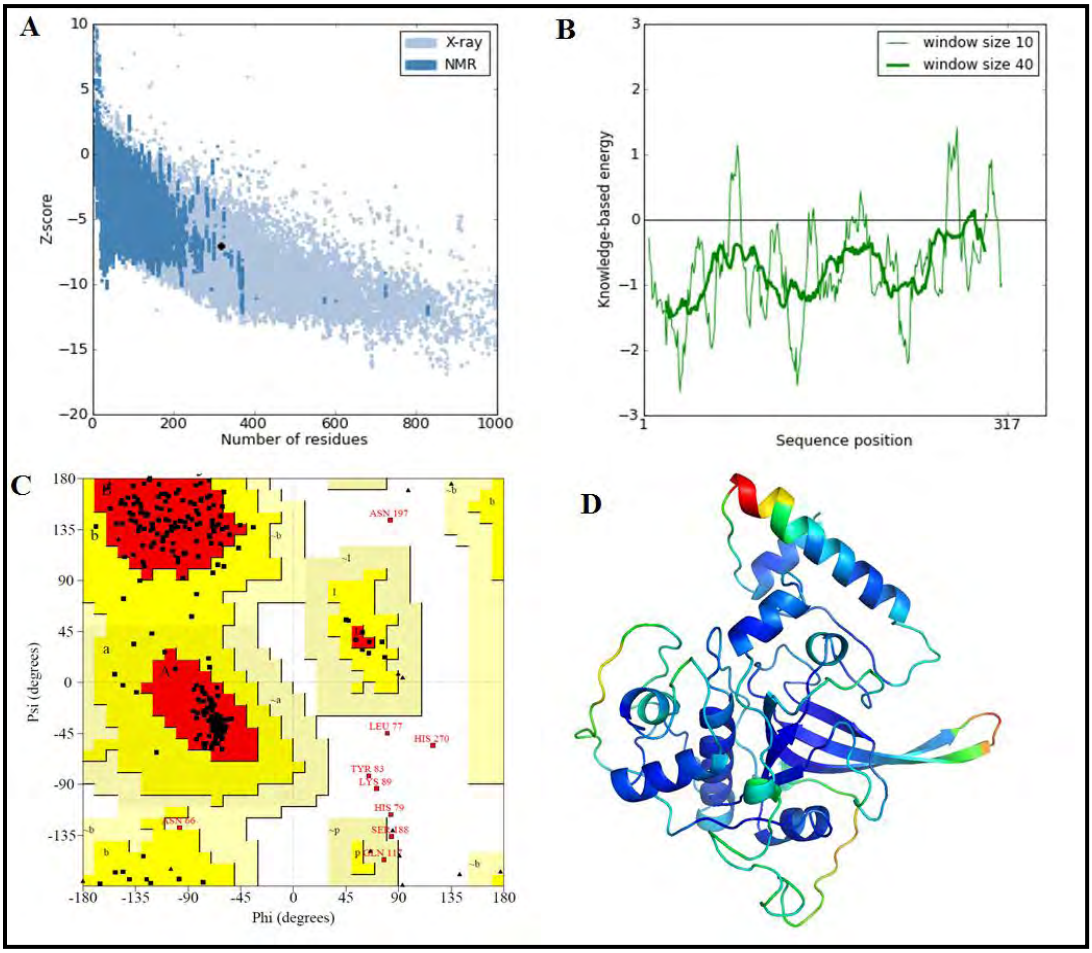

Figure 3.18 Validation of *P. Yoelii* model: part (A) is the PROSA global energy plot FP-2 structure is plotted as the black dot other PDB structures are in blue and light blue dots. Part (B) is the PROSA local plot in a 10 and 40 residue window. Part (C) the PROCHECK validation ramachandran plot and part (D) is the MetaMQAPII structure validation colored by quality in a spectrum of blue (good) to red (bad).

Interaction analysis prodomain region (202-221) interacts with the subsites (Figure 3.19). S1 and S1' were occupied by prodomain residues Asn-202 (S1'), Phe-204 (S1'), His-209 (S1'), Phe-212 (S1'), Tyr-216 (S1'). The side-chain of Asn-202 forms a hydrogen bond to the main-chain of Glu-389 (S1'). Three prodomain residues; Phe-204, Phe-212 and Tyr-216 form aromatic interactions to Trp-438 in subsite S1'. Phe-204 and Phe-212 are also found in hydrophobic contact to Trp-448 and Val-384 in subsite S1'.

The prodomain residues Asn-218, Asn-219, Leu-221 occupy S2 and S3 (Figure 3.19). The side-chains of prodomain residues Asn-218 and Asn-219 form hydrogen bonds to the carboxylic oxygen of Gly-315 (S3). Asn-218 also forms main-chain hydrogen bonding to Asn-405 (S2). The side chain of Asn-218 is at a distance of 3.66Å to the side-chain of catalytic Cys-274. Hydrophobic prodomain residue Leu-221 forms hydrophobic interactions to Ala-404, Val-462 and Val-464 in the S2 subsite.

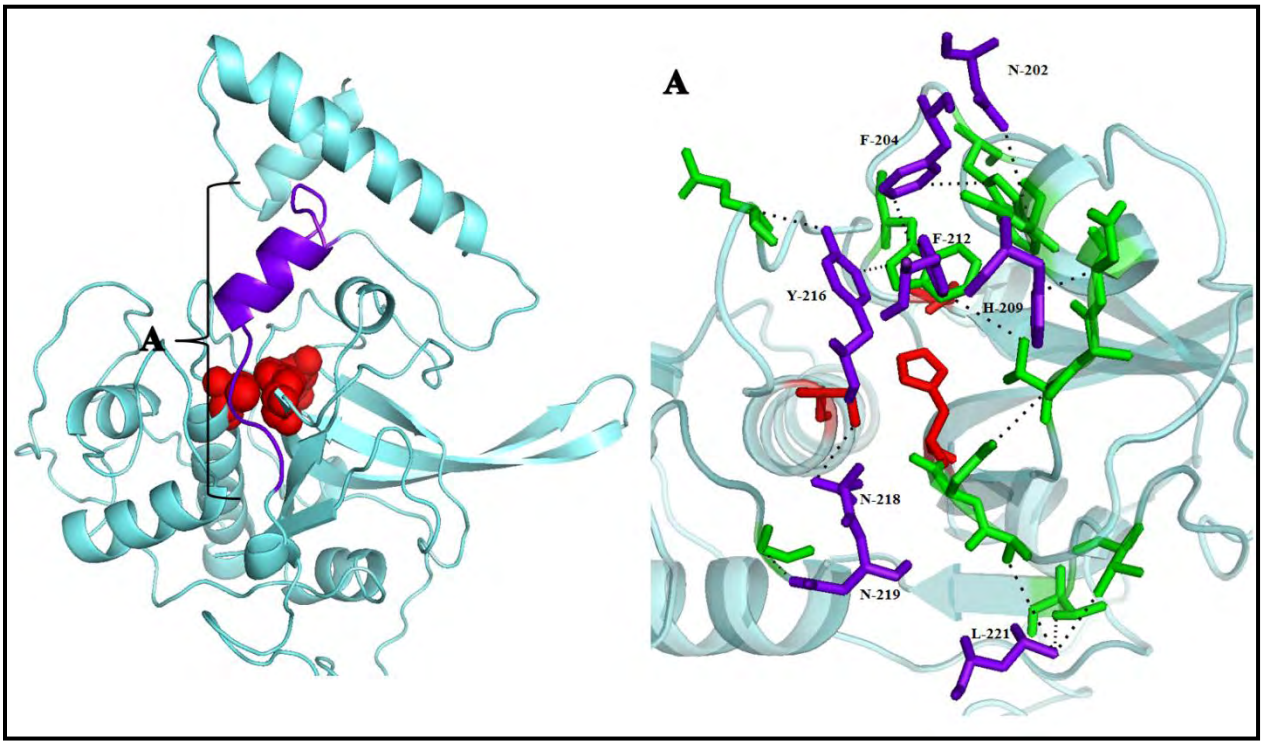

Figure 3.19 Prodomain interactions to *P. yoelii* subsites; the top diagram shows in purple the prodomain region (202-221) interacting with subsite residues (red indicates the active site). A- Represents the prodomain residues (purple sticks) interacting with residues in subsites (green sticks). Prodomain residues are labeled.

3.3.5 Summary of interactions

The prodomain residues responsible for inhibition are summarized in (Table 3.3). Showing the residues inhibiting subsites S1, S1', S2 and S3. The corresponding residues in human proteases: cathepsin-L1 and cathepsin-H are indicated.

Table 3.3 Summary of prodomain residues inhibiting the various subsites. Corresponding human protease residues are grey highlighted. Darker shade of grey points out major substitutions observed in the human proteases

Protease	Subsite S1 and S1'									Subsite S2 and S3			
FP-2		R-213	F-214	Y-219	F-222	K-223	K-225	Y-226		S-228	L-229		S-231
FP-2B	N-210	R-211	F-212	Y-217	F-220	K-221	K-223	Y-224	L-225	T-226	L-227	R-228	S-229
FP-3	N-218	K-219	F-220		F-228		K-231	Y-232	L-233	N-234	L-235		
Vivapain-2	N-214	Q-215	F-216	F-221	F-224	K-225		Y-228		T-230	L-231		
P. Knowlesi	N-220	R-221	F-222	F-227	F-230		K-233	Y-334	L-235	T-236	L-237	K-238	
P. berghei	N-200		F-202	H-207	F-210	K-211		Y-214	L-215		N-217		L-219
Chabaupain-2	N-201	P-202	F-203		F-211			Y-215	L-216	N-217		K-219	
P. Yoelii	N-202		F-204	H-209	F-212			Y-216		N-218	N-219		L-221
Cathepsin –L1	N-78	A-79	F-80	S-85	F-88	R-89	V-91	M-92	N-93	G-94	F-95	Q-96	N-97
Cathepsin-K	N-76	H-77	L-78	S-83	V-86	V-87	K-89	M-90	T-91	G-92	L-93	K-94	V-95

The prodomain segment in interaction with the subsites spans from residue 212 to 231 (FP-2 numbering). At the beginning of this segment, there are three conserved residues (212-214) Asn, Arg and Phe (NRF) observed to occupy and hence inhibit the S1' in the proteases (Table 3.3). After which an essential aromatic Try/Phe/His residue (219) is observed in most of the proteases to inhibit subsite S1' via hydrophobic interactions. This aromatic residue is replaced by serine in human proteases. Next is a sequence of interacting residues (222-227) FKKYL that are essential in binding to the S1 and S1' subsite via aromatic, hydrophobic, ionic and hydrogen bonds. The subsequent four residues (228-231) NLKS, obstruct the S2 and S3 subsites via hydrophobic and hydrogen bonding.

Having identified these residues and their fostering interactions, it was necessary to pick out a short contiguous section that contains most of the interacting residues. Hence we propose the most suitable prodomain peptide responsible for inhibition of FP-2, FP-2B, FP-3, vivapain-2 and *P.knowlesi* is residues 222-232 (FP-2 numbering). The sequence of the peptide in FP-2 is FKNKYLSLRS; it is a ten residue continuous section of the prodomain inhibiting all four subsites. The peptide sequence varies slightly in each of the proteases;

- FP-2 (FKNKYLSLRS)
- FP-2B (FKSKYLTLRS)
- FP-3 (FRSKYLNLKT)
- Vivapain-2- (FKKKYLTLKS)
- *P. ovale* (FEKKYLTLKT).

We also propose a prodomain peptide (residues 222-233) responsible for inhibition of the 3 murine *Plasmodium* species proteases. The peptide sequence varies slightly in each species:

- *P. berghei* (FKMKYLNNKLK)
- Chabaupain-2 (FKMRYLNSKLS)
- *P. yoelii* (FKMKYLNNKLK)

3.4 CONCLUSION AND FUTURE WORK

Prodomains have been shown to have selective inhibition towards their parent enzymes (Cygler et al. 1997) and this project aimed to design short peptide that could mimic this inhibition. Having attained peptide sequences that are likely to have selective inhibition of the target proteases. The next step is to test the binding affinity and specificity of these peptides. This would be made possible by synthesis of these peptides in the lab. Assays can then be carried out to test the inhibition effect of prodomain peptides on intraerythrocytic *Plasmodium* parasites in culture. Hemoglobin degradation of the peptide treated *Plasmodium* parasites can be monitored to verify inhibition of cysteine protease activity. An alternative approach would be to amplify the genes of the cysteine proteases and express recombinant proteases in *E. coli*. *In vitro* assays can then be conducted to test the ability of the peptides to inhibit substrate hydrolysis of the proteases.

A number of peptide based inhibitors against FP-2 have been identified (Ettari et al. 2007). However the practicality of using peptide inhibitors is limited by many factors including peptide ability to transverse cell membranes, peptide susceptibility to proteolysis and orientation of peptide binding in the active site. This factors need to be put into consideration in peptidomimetic drug design. This is achieved by replacement of natural residues with unnatural amino acids or incorporating a non-peptidic scaffold to the peptide sequence to increase potency and stability against proteolysis as well as enhance oral bioavailability. Peptidomimetic inhibitors of FP-2 have been designed based on a 1,4-benzodiazepine (BDZ) scaffold, which mimics the fragment D-Ser-Gly, that binds reversibly to the active site inhibiting the enzyme (Micale et al. 2006). BDZs are in a class of drugs known to have good oral bioavailability and be well tolerated. The peptidomimetic was potent and selective against FP-2 and FP-2B having significant antimalarial activity. Hence it is promising that the peptides identified in this study can be developed into antimalarial peptidomimetic.

REFERENCES

Abagyan, R. & Totrov, M. 1994, "Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins", *Journal of Molecular Biology,* vol. 235, no. 3, pp. 983-1002.

Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. 1998, "Gapped blast and psi-blast: a new generation of protein database search programs", *FASEB Journal,* vol. 12, no. 8. Pp 3389-3

Altschul, S.F. 1998, "Fundamentals of database searching", *Trends in biotechnology,* vol. 16, no. Supplement 1, pp. 7-9.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. 1990, "Basic local alignment search tool", *Journal of Molecular Biology,* vol. 215, no. 3, pp. 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. 1997, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs", *Nucleic acids research,* vol. 25, no. 17, pp. 3389-3402.

Aurrecoechea, C., Brestelli, J., Brunk, B.P., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O.S., Heiges, M., Innamorato, F., Iodice, J., Kissinger, J.C., Kraemer, E., Li, W., Miller, J.A., Nayak, V., Pennington, C., Pinney, D.F., Roos, D.S., Ross, C., Stoeckert Jr., C.J., Treatman, C. & Wang, H. 2009, "PlasmoDB: A functional genomic database for malaria parasites", *Nucleic acids research,* vol. 37, no. SUPPL. 1, pp. D539-D543.

Bajorath, J., Stenkamp, R. & Aruffo, A. 1993, "Knowledge-based model building of proteins: Concepts and examples", *Protein Science,* vol. 2, no. 11, pp. 1798-1810.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. 2000, "The Protein Data Bank", *Nucleic acids research,* vol. 28, no. 1, pp. 235-242.

Blackman, M.J. 2004, "Proteases in host cell invasion by the malaria parasite", *Cellular microbiology,* vol. 6, no. 10, pp. 893-903.

Blackman, M.J. 2008, "Malarial proteases and host cell egress: An 'emerging' cascade", *Cellular microbiology,* vol. 10, no. 10, pp. 1925-1934.

Blundell, T.L., Sibanda, B.L., Montalvão, R.W., Brewerton, S., Chelliah, V., Worth, C.L., Harmer, N.J., Davies, O. & Burke, D. 2006, "Structural biology and bioinformatics in drug design: Opportunities and challenges for target identification and lead discovery", *Philosophical Transactions of the Royal Society B: Biological Sciences,* vol. 361, no. 1467, pp. 413-423.

Bower, M.J., Cohen, F.E. & Dunbrack Jr., R.L. 1997, "Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: A new homology modeling tool", *Journal of Molecular Biology,* vol. 267, no. 5, pp. 1268-1282.

Bowie, J.U., Luthy, R. & Eisenberg, D. 1991, "A method to identify protein sequences that fold into a known three-dimensional structure", *Science,* vol. 253, no. 5016, pp. 164-170.

Carmona, E., Dufour, É., Plouffe, C., Takebe, S., Mason, P., Mort, J.S. & Ménard, R. 1996, "Potency and selectivity of the cathepsin L propeptide as an inhibitor of cysteine proteases", *Biochemistry,* vol. 35, no. 25, pp. 8149-8157.

Chen, F., Mackey, A.J., Stoeckert Jr., C.J. & Roos, D.S. 2006, "OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups", *Nucleic acids research,* vol. 34, no. Database issue, pp. D363-368.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G. & Thompson, J.D. 2003, "Multiple sequence alignment with the Clustal series of programs", *Nucleic acids research,* vol. 31, no. 13, pp. 3497-3500.

Chowdhury, S.F., Sivaraman, J., Wang, J., Devanathan, G., Lachance, P., Qi, H., Ménard, R., Lefebvre, J., Konishi, Y., Cygler, M., Sulea, T. & Purisima, E.O. 2002, "Design of noncovalent inhibitors of human cathepsin L. from the 96-residue proregion to optimized tripeptides", *Journal of Medicinal Chemistry,* vol. 45, no. 24, pp. 5321-5329.

Collins, W.E. & Jeffery, G.M. 2005, "Plasmodium ovale: Parasite and disease", *Clinical microbiology reviews,* vol. 18, no. 3, pp. 570-581.

Cox-Singh, J., Davis, T.M.E., Lee, K.-., Shamsul, S.S.G., Matusop, A., Ratnam, S., Rahman, H.A., Conway, D.J. & Singh, B. 2008, "Plasmodium knowlesi malaria in humans is widely distributed and potentially life threatening", *Clinical Infectious Diseases,* vol. 46, no. 2, pp. 165-171.

Cygler, M. & Mort, J.S. 1997, "Proregion structure of members of the papain superfamily. Mode of inhibition of enzymatic activity", *Biochimie,* vol. 79, no. 11, pp. 645-652.

Di Luccio, E. & Koehl, P. 2011, "A quality metric for homology modeling: The H-factor", *BMC Bioinformatics,* vol. 12, pp. 12-48

Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J.-., Taly, J.-. & Notredame, C. 2011, "T-Coffee: A web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension", *Nucleic acids research,* vol. 39, no. SUPPL. 2, pp. W13-W17.

Dunbrack Jr., R.L. 2006, "Sequence comparison and protein structure prediction", *Current opinion in structural biology,* vol. 16, no. 3, pp. 374-384.

Eddy, S.R. 1998, "Profile hidden Markov models", *Bioinformatics,* vol. 14, no. 9, pp. 755-763.

Edgar, R.C. 2004, "MUSCLE: Multiple sequence alignment with high accuracy and high throughput", *Nucleic acids research,* vol. 32, no. 5, pp. 1792-1797.

Eramian, D., Eswar, N., Shen, M.-. & Sali, A. 2008, "How well can the accuracy of comparative protein structure models be predicted?", Protein Science, vol. 17, no. 11, pp. 1881-1893

Ettari, R., Bova, F., Zappalà, M., Grasso, S. & Micale, N. 2010, "Falcipain-2 inhibitors", *Medicinal research reviews,* vol. 30, no. 1, pp. 136-167.

Ettari, R., Nizi, E., Di Francesco, M.E., Dude, M.-., Pradel, G., Vičík, R., Schirmeister, T., Micale, N., Grasso, S. & Zappalà, M. 2008, "Development of peptidomimetics with a vinyl sulfone warhead as irreversible falcipain-2 inhibitors", Journal of medicinal chemistry, vol. 51, no. 4, pp. 988-996.

Francis, S.E., Sullivan Jr., D.J. & Goldberg, D.E. 1997, “Hemoglobin metabolism in the malaria parasite *Plasmodium falciparium*”, *microbiology,*vol. 51, pp 97-123.

Gherardini, P.F. & Helmer-Citterich, M. 2008, "Structure-based function prediction: Approaches and applications", *Briefings in Functional Genomics and Proteomics,* vol. 7, no. 4, pp. 291-302.

Godat, E., Chowdhury, S., Lecaille, F., Belghazi, M., Purisima, E.O. & Lalmanach, G. 2005, "Inhibition of a cathepsin L-like cysteine protease by a chimeric propeptide-derived inhibitor", *Biochemistry,* vol. 44, no. 31, pp. 10486-10493.

Grant, M.A. 2011, "Integrating computational protein function prediction into drug discovery initiatives", *Drug Development Research,* vol. 72, no. 1, pp. 4-16.

Grzonka, Z., Jankowska, E., Kasprzykowski, F., Kasprzykowska, R., Łankiewicz, L., Wiczk, W., Wieczerzak, E., Ciarkowski, J., Drabik, P., Janowski, R., Kozak, M., Jaskólski, M. & Grubb, A. 2001, "Structural studies of cysteine proteases and their inhibitors", *Acta Biochimica Polonica,* vol. 48, no. 1, pp. 1-20.

Hall, T.A. 1999. “BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT”.

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,* vol. 89, no. 22, pp. 10915-10919.

Hillisch, A., Pineda, L.F. & Hilgenfeld, R. 2004, "Utility of homology models in the drug discovery process", *Drug discovery today,* vol. 9, no. 15, pp. 659-669.

Hogg, T., Nagarajan, K., Herzberg, S., Chen, L., Shen, X., Jiang, H., Wecke, M., Blohmke, C., Hilgenfeld, R. & Schmidt, C.L. 2006, "Structural and functional characterization of falcipain-2, a hemoglobinase from the malarial parasite Plasmodium falciparum", *Journal of Biological Chemistry,* vol. 281, no. 35, pp. 25425-25437.

Jean, L., Hackett, F., Martin, S.R. & Blackman, M.J. 2003, "Functional characterization of the propeptide of Plasmodium falciparum subtilisin-like protease-1", *Journal of Biological Chemistry,* vol. 278, no. 31, pp. 28572-28579.

Jones, D.T. 1999, "Protein secondary structure prediction based on position-specific scoring matrices", *Journal of Molecular Biology,* vol. 292, no. 2, pp. 195-202.

Katoh, K., Kuma, K.-., Toh, H. & Miyata, T. 2005, "MAFFT version 5: Improvement in accuracy of multiple sequence alignment", *Nucleic acids research,* vol. 33, no. 2, pp. 511-518.

Katoh, K., Misawa, K., Kuma, K.-. & Miyata, T. 2002, "MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform", *Nucleic acids research,* vol. 30, no. 14, pp. 3059-3066.

Kmiecik, S., Gront, D. & Kolinski, A. 2007, "Towards the high-resolution protein structure prediction. Fast refinement of reduced models with all-atom force field", *BMC Structural Biology,* vol. 7.pp. 7-18

Korde, R., Bhardwaj, A., Singh, R., Srivastava, A., Chauhan, V.S., Bhatnagar, R.K. & Malhotra, P. 2008, "A prodomain peptide of Plasmodium falciparum cysteine protease (falcipain-2) inhibits malaria parasite development", *Journal of medicinal chemistry,* vol. 51, no. 11, pp. 3116-3123.

Kosmoliaptsis, V., Dafforn, T.R., Chaudhry, A.N., Halsall, D.J., Bradley, J.A. & Taylor, C.J. 2011, "High-resolution, three-dimensional modeling of human leukocyte antigen class I structure and surface electrostatic potential reveals the molecular basis for alloantibody binding epitopes", *Human immunology,* . vol. 10, pp 1016-1027

Krissinel, E. 2007, "On the relationship between sequence and structure similarities in proteomics", *Bioinformatics,* vol. 23, no. 6, pp. 717-723.

Kumar, S., Tamura, K. & Nei, M. 2004, "MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment", *Briefings in bioinformatics,* vol. 5, no. 2, pp. 150-163.

LaLonde, J.M., Zhao, B., Janson, C.A., D'Alessio, K.J., McQueney, M.S., Orsini, M.J., Debouck, C.M. & Smith, W.W. 1999, "The crystal structure of human procathepsin K", *Biochemistry,* vol. 38, no. 3, pp. 862-869.

Lecaille, F., Kaleta, J. & Br m me, D. 2002a, Human and parasitic Papain-like cysteine proteases: Their role in physiology and pathology and recent developments in inhibitor design", *Chemical reviews,* vol. 102, no. 12, pp. 4459-4488.

Lee, M.R., Tsai, J., Baker, D. & Kollman, P.A. 2001, "Molecular Dynamics in the Endgame of Protein Structure Prediction", *Journal of Molecular Biology,* vol. 313, no. 2, pp. 417-430.

Levitt, M. & Lifson, S. 1969, "Refinement of protein conformations using a macromolecular energy minimization procedure", *Journal of Molecular Biology,* vol. 46, no. 2, pp. 269-279.

Luthy, R., Bowie, J.U. & Eisenberg, D. 1992, "Assesment of protein models with three-dimensional profiles", *Nature,* vol. 356, no. 6364, pp. 83-85.

MacKerell Jr., A.D., Bashford, D., Bellott, M., Dunbrack Jr., R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F.T.K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D.T., Prodhom, B., Reiher III, W.E., Roux, B., Schlenkrich, M., Smith, J.C., Stote, R., Straub, J., Watanabe, M., Wiórkiewicz-Kuczera, J., Yin, D. & Karplus, M. 1998, "All-atom empirical potential for molecular modeling and dynamics studies of proteins", *Journal of Physical Chemistry B,* vol. 102, no. 18, pp. 3586-3616.

Mazumder, R., Natale, D.A., Murthy, S., Thiagarajan, R. & Wu, C.H. 2005, "Computational identification of strain-, species- and genus-specific proteins", *BMC Bioinformatics,* vol. 6.pp 543-550

McKerrow, J.H. & Sajid, M. 2002, "Cysteine proteases of parasitic organisms", *Molecular and biochemical parasitology,* vol. 120, no. 1, pp. 1-21.

McKerrow, J.H. 1999, "Development of cysteine protease inhibitors as chemotherapy for parasitic diseases: Insights on safety, target validation, and mechanism of action", *International journal for parasitology,* vol. 29, no. 6, pp. 833-837.

Micale, N., Kozikowski, A.P., Ettari, R., Grasso, S., Zappalà, M., Jeong, J.-., Kumar, A., Hanspal, M. & Chishti, A.H. 2006, "Novel peptidomimetic cysteine protease inhibitors as potential antimalarial agents", Journal of medicinal chemistry, vol. 49, no. 11, pp. 3064-3067.

Miller, L.H., Baruch, D.I., Marsh, K. & Doumbo, O.K. 2002, "The pathogenic basis of malaria", *Nature,* vol. 415, no. 6872, pp. 673-679.

Mizuguchi, K., Deane, C.M., Blundell, T.L., Johnson, M.S. & Overington, J.P. 1998, "JOY: Protein sequence-structure representation and analysis", *Bioinformatics,* vol. 14, no. 7, pp. 617-623.

Morgenstern, B., Dress, A. & Werner, T. 1996, "Multiple DNA and protein sequence alignment based on segment-to-segment comparison", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 93, no. 22, pp. 12098-12103.

Moult, J. 2005, "A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction", *Current opinion in structural biology,* vol. 15, no. 3 SPEC. ISS., pp. 285-289.

Moult, J., Fidelis, K., Kryshtafovych, A., Rost, B., Hubbard, T. & Tramontano, A. 2007, "Critical assessment of methods of protein structure prediction - Round VII", *Proteins: Structure, Function and Genetics,* vol. 69, no. SUPPL. 8, pp. 3-9.

Narayanan Eswer, B.W, 2006, "Comparative Protein Structure Modeling Using Modeller", *Current Protocols in Bioinformatics,* vol. 5, no. 6. pp. 1-30.

Nayeem, A., Sitkoff, D. & Krystek Jr., S. 2006, "A comparative study of available software for high-accuracy homology modeling: From sequence alignments to structural models", *Protein Science,* vol. 15, no. 4, pp. 808-824.

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,* vol. 48, no. 3, pp. 443-453.

Nuin, P, Wang, Z, & Tillier, E 2006, "The accuracy of several multiple sequence alignment programs for proteins", *BMC Bioinformatics*, 7, p. 471.

Pandey, K.C., Barkan, D.T., Sali, A. & Rosenthal, P.J. 2009, "Regulatory elements within the prodomain of falcipain-2, a cysteine protease of the malaria parasite Plasmodium falciparum", *PLoS ONE,* vol. 4, no. 5.pp. 5694-5703.

Pandey, K.C., Wang, S.X., Sijwali, P.S., Lau, A.L., McKerrow, J.H. & Rosenthal, P.J. 2005, "The Plasmodium falciparum cysteine protease falcipain-2 captures its substrate, hemoglobin, via a unique motif", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 102, no. 26, pp. 9138-9143.

Pawlowski, M., Gajda, M.J., Matlak, R. & Bujnicki, J.M. 2008, "MetaMQAP: A meta-server for the quality assessment of protein models", *BMC Bioinformatics,* vol. 9.pp.403-423

Pei, J. & Grishin, N.V. 2007, "PROMALS: Towards accurate multiple sequence alignments of distantly related proteins", *Bioinformatics,* vol. 23, no. 7, pp. 802-808.

Pei, J., Kim, B.-. & Grishin, N.V. 2008, "PROMALS3D: A tool for multiple protein sequence and structure alignments", *Nucleic acids research,* vol. 36, no. 7, pp. 2295-2300.

Phillips, A.J. 2006, "Homology assessment and molecular sequence alignment", *Journal of Biomedical Informatics,* vol. 39, no. 1, pp. 18-33.

Pierri, C.L., Parisi, G. & Porcelli, V. 2010, "Computational approaches for protein function prediction: A combined strategy from multiple sequence alignment to molecular docking-based virtual screening", *Biochimica et Biophysica Acta - Proteins and Proteomics,* vol. 1804, no. 9, pp. 1695-1712.

Price, R.N., Tjitra, E., Guerra, C.A., Yeung, S., White, N.J. & Anstey, N.M. 2007, "Vivax malaria: neglected and not benign", *The American Journal of Tropical Medicine and Hygiene,* vol. 77, no. 6 Suppl, pp. 79-87.

Rizzi, L., Sundararaman, S., Cendic, K., Vaiana, N., Korde, R., Sinha, D., Mohmmed, A., Malhotra, P. & Romeo, S. 2011, "Design and synthesis of protein-protein interaction mimics as Plasmodium falciparum cysteine protease, falcipain-2 inhibitors", *European journal of medicinal chemistry,* vol. 46, no. 6, pp. 2083-2090.

Rose, P.W., Beran, B., Bi, C., Bluhm, W.F., Dimitropoulos, D., Goodsell, D.S., Prlić, A., Quesada, M., Quinn, G.B., Westbrook, J.D., Young, J., Yukich, B., Zardecki, C., Berman, H.M. & Bourne, P.E. 2011, "The RCSB Protein Data Bank: Redesigned web site and web services", *Nucleic acids research,* vol. 39, no. SUPPL. 1, pp. D392-D401.

Rosenthal, P.J. 1998, "Proteases of malaria parasites: New targets for chemotherapy", *Emerging Infectious Diseases,* vol. 4, no. 1, pp. 49-57.

Rosenthal, P.J. 2004, "Cysteine protease of malarial parasite", *International journal for parasitology,* vol. 34, pp. 1489.

Rosenthal, P.J., Wollish, W.S., Palmer, J.T. & Rasnick, D. 1991, "Antimalarial effects of peptide inhibitors of a Plasmodium falciparum cysteine proteinase", *Journal of Clinical Investigation,* vol. 88, no. 5, pp. 1467-1472.

Saitou, N. & Nei, M. 1987, "The neighbor-joining method: a new method for reconstructing phylogenetic trees", *Molecular biology and evolution,* vol. 4, no. 4, pp. 406-425.

Sali, A. & Blundell, T.L. 1993, "Comparative protein modelling by satisfaction of spatial restraints", *Journal of Molecular Biology,* vol. 234, no. 3, pp. 779-815.

Schechter, I. (2005). “Mapping of the active site of proteases in the 1960s and rational design of inhibitors/drugs in the 1990s”, *Current Protein and Peptide Science*, vol.6 pp. 501-512.

Schechter, I. and Berger, A., 1967,“On the size of the active site in proteases. I. Papain” *Biochemical and biophysical research communications*, 27(2), pp. 157-162

Scott, C.J. & Taggart, C.C. 2010, "Biologic protease inhibitors as novel therapeutic agents", *Biochimie,* vol. 92, no. 11, pp. 1681-1688.

Selzer, P.M., Pingel, S., Hsieh, I., Ugele, B., Chan, V.J., Engel, J.C., Bogyo, M., Russell, D.G., Sakanari, J.A. & Mckerrow, J.H. 1999, "Cysteine protease inhibitors as chemotherapy: Lessons from a parasite target", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 96, no. 20, pp. 11015-11022.

Shah, F., Mukherjee, P., Gut, J., Legac, J., Rosenthal, P.J., Tekwani, B.L. & Avery, M.A. 2011, "Identification of novel malarial cysteine protease inhibitors using structure-based virtual screening of a focused cysteine protease inhibitor library", *Journal of Chemical Information and Modeling,* vol. 51, no. 4, pp. 852-864.

Shen, M.-. & Sali, A. 2006, "Statistical potential for assessment and prediction of protein structures", Protein Science, vol. 15, no. 11, pp. 2507-2524.

Sijwali, P.S. & Rosenthal, P.J. 2004, "Gene disruption confirms a critical role for the cysteine protease falcipain-2 in hemoglobin hydrolysis by Plasmodium falciparum", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 101, no. 13, pp. 4384-4389.

Sijwali, P.S., Koo, J., Singh, N. & Rosenthal, P.J. 2006, "Gene disruptions demonstrate independent roles for the four falcipain cysteine proteases of Plasmodium falciparum", *Molecular and biochemical parasitology,* vol. 150, no. 1, pp. 96-106.

Sijwali, P.S., Shenai, B.R. & Rosenthal, P.J. 2002, "Folding of the Plasmodium falciparum cysteine protease falcipain-2 is mediated by a chaperone-like peptide and not the prodomain", *Journal of Biological Chemistry,* vol. 277, no. 17, pp. 14910-14915.

Sippl, M.J. 1993, "Recognition of errors in three-dimensional structures of proteins", *Proteins: Structure, Function and Genetics,* vol. 17, no. 4, pp. 355-362.

Sippl, M.J. 1995, "Knowledge-based potentials for proteins", *Current opinion in structural biology,* vol. 5, no. 2, pp. 229-235.

S ding , J. 2005, Protein homology detection by HMM-HMM comparison", *Bioinformatics,* vol. 21, no. 7, pp. 951-960.

Söding, J., Biegert, A. & Lupas, A.N. 2005, "The HHpred interactive server for protein homology detection and structure prediction", *Nucleic acids research,* vol. 33, no. SUPPL. 2, pp. W244-W248.

Summa, C.M. & Levitt, M. 2007, "Near-native structure refinement using in vacuo energy minimization", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 104, no. 9, pp. 3177-3182.

Tastan, A.Ö.B., De Beer, T.A.P. & Joubert, F. 2008, "Protein homology modelling and its use in South Africa", *South African Journal of Science,* vol. 104, no. 1-2, pp. 2-6.

Tatusov, R.L., Koonin, E.V. & Lipman, D.J. 1997, "A genomic perspective on protein families", *Science,* vol. 278, no. 5338, pp. 631-637.

Taylor, W.R. 1998, "Dynamic sequence databank searching with templates and multiple alignment", *Journal of Molecular Biology,* vol. 280, no. 3, pp. 375-406.

Thompson, J.D., Plewniak, F. & Poch, O. 1999, "A comprehensive comparison of multiple sequence alignment programs", *Nucleic acids research,* vol. 27, no. 13, pp. 2682-2690.

Tina, K.G., Bhadra, R. & Srinivasan, N. 2007, "PIC: Protein Interactions Calculator.", *Nucleic acids research,* vol. 35, no. Web Server issue, pp. W473-476.

Van Vlijmen, H.W.T. & Karplus, M. 1997, "PDB-based protein loop prediction: Parameters for selection and methods for optimization", *Journal of Molecular Biology,* vol. 267, no. 4, pp. 975-1001.

Venclovas, C. 2003, "Comparative Modeling in CASP5: Progress Is Evident, but Alignment Errors Remain a Significant Hindrance", *Proteins: Structure, Function and Genetics,* vol. 53, no. SUPPL. 6, pp. 380-388.

Wang, S.X., Pandey, K.C., Somoza, J.R., Sijwali, P.S., Kortemme, T., Brinen, L.S., Fletterick, R.J., Rosenthal, P.J. & McKerrow, J.H. 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,* vol. 103, no. 31, pp. 11503-11508.

Wegscheid-Gerlach, C., Gerber, H.-. & Diederich, W.E. 2010, "Proteases of plasmodium falciparum as potential drug targets and inhibitors thereof", *Current Topics in Medicinal Chemistry,* vol. 10, no. 3, pp. 346-367.

Wheeler, D. 2002, "Selecting the right protein-scoring matrix", *Current protocols in bioinformatics / editoral board, Andreas D.Baxevanis ...[et al.],* vol. Chapter 3.pp. 1-6

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L.Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Miller, V., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R.L., Tatusova, T.A., Wagner, L. & Yaschenko, E. 2007, "Database resources of the National Center for Biotechnology Information", *Nucleic acids research,* vol. 35, no. SUPPL. 1, pp. D5-D12.

White, N.J. 2008, "Plasmodium knowlesi: The fifth human malaria parasite", *Clinical Infectious Diseases,* vol. 46, no. 2, pp. 172-173.

Wiederanders, B., Kaulmann, G. & Schilling, K. 2003, "Functions of propeptide parts in cysteine proteases", *Current Protein and Peptide Science,* vol. 4, no. 5, pp. 309-326.

Wiederstein, M. & Sippl, M.J. 2007, "ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins", *Nucleic acids research,* vol. 35, no. Web Server issue, pp. W407-410.

Wlodawer, A., Minor, W., Dauter, Z. & Jaskolski, M. 2008, "Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures", *FEBS Journal,* vol. 275, no. 1, pp. 1-21.

Wu, Y., Wang, X., Liu, X. & Wang, Y. 2003, "Data-mining approaches reveal hidden families of proteases in the genome of malaria parasite", *Genome research,* vol. 13, no. 4, pp. 601-616.

Wüthrich, K. 2003, "NMR studies of structure and function of biological macromolecules (Nobel Lecture)", *Journal of Biomolecular NMR,* vol. 27, no. 1, pp. 13-39.

Xiang, Z., Soto, C.S. & Honig, B. 2002, "Evaluating conformational free energies: The colony energy and its application to the problem of loop prediction", *Proceedings of the National Academy of Sciences of the United States of America,* vol. 99, no. 11, pp. 7432-7437.

Xiong, J. 2006, Essential Bioinformatics*, Cabridge University Press, New York, USA*.

Yadav, M., Singh, A., Rathaur, S. and Liebau, E., 2010. “Structural modeling and simulation studies of Brugia malayi glutathione-S-transferase with compounds exhibiting antifilarial activity: Implications in drug targeting and designing”, *Journal of Molecular Graphics and Modelling*, 28(5), pp. 435-445

Ye, J., McGinnis, S. & Madden, T.L. 2006, "BLAST: Improvements for better sequence analysis", *Nucleic acids research,* vol. 34, no. WEB. SERV. ISS., pp. W6-W9.

Zemla, A. 2003, "LGA: A method for finding 3D similarities in protein structures", Nucleic acids research, vol. 31, no. 13, pp. 3370-3374.

Zhang, J., Chiodini, R., Badr, A. & Zhang, G. 2011, "The impact of next-generation sequencing on genomics", *Journal of Genetics and Genomics,* vol. 38, no. 3, pp. 95-109.

Zhou, Y. & Landweber, L.F. 2007, "BLASTO: a tool for searching orthologous groups", *Nucleic acids research,* vol. 35, no. Web Server issue, pp. W678-682.

Zhou, Z.H. 2008, "Towards atomic resolution structural determination by single-particle cryo-electron microscopy", *Current opinion in structural biology,* vol. 18, no. 2, pp. 218-228.

APPENDIX

APPENDIX A: SEQUENCE RETRIVAL AND ALIGNMENT

I. Full length alignments of retrieved sequences (PROMALS 3D)

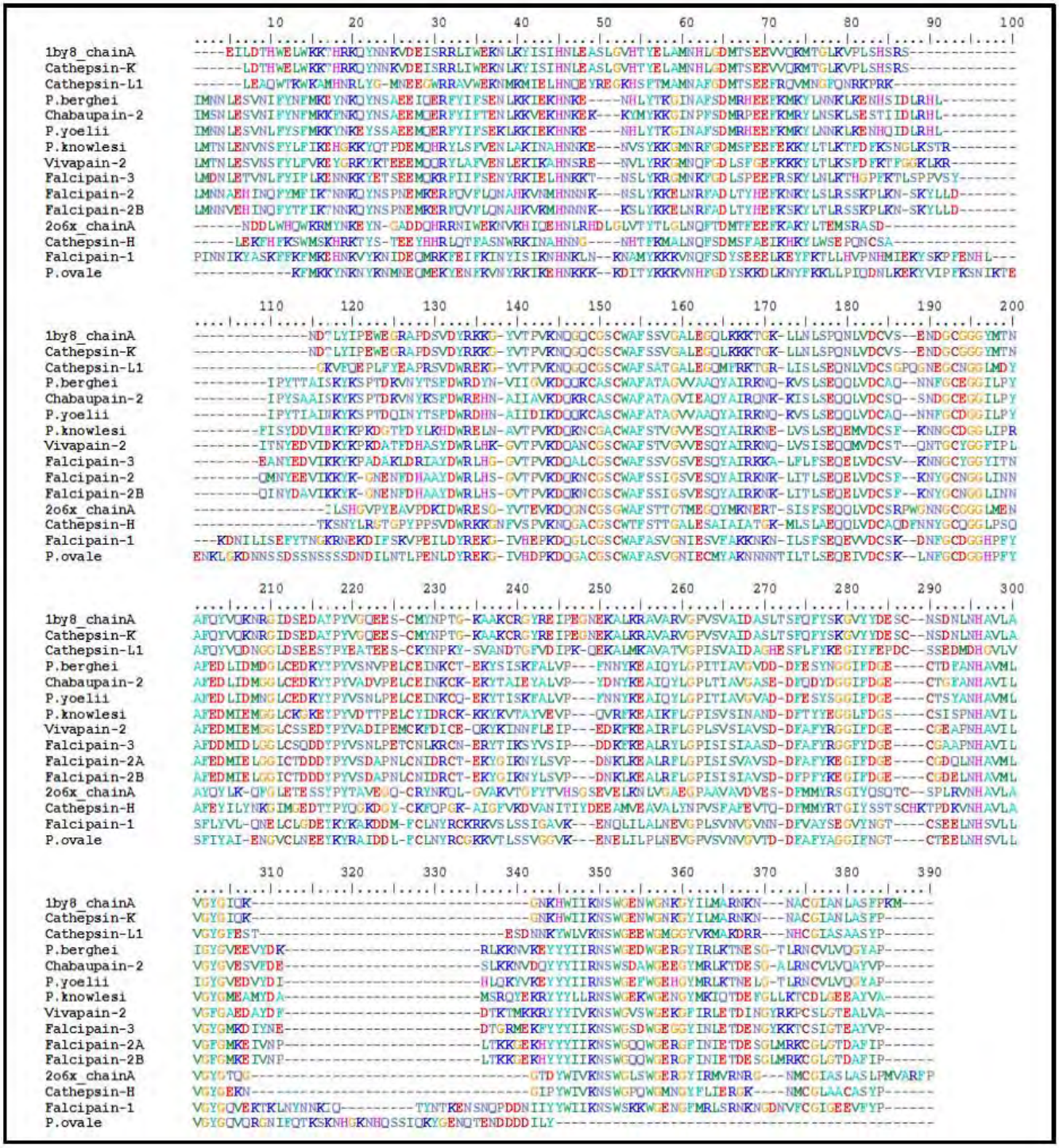

Figure A-1: Showing full length alignment of all sequences it depicts the prodomain and catalytic domains.

II. PROMALS-3D alignment used as dataset for phylogenetic analysis

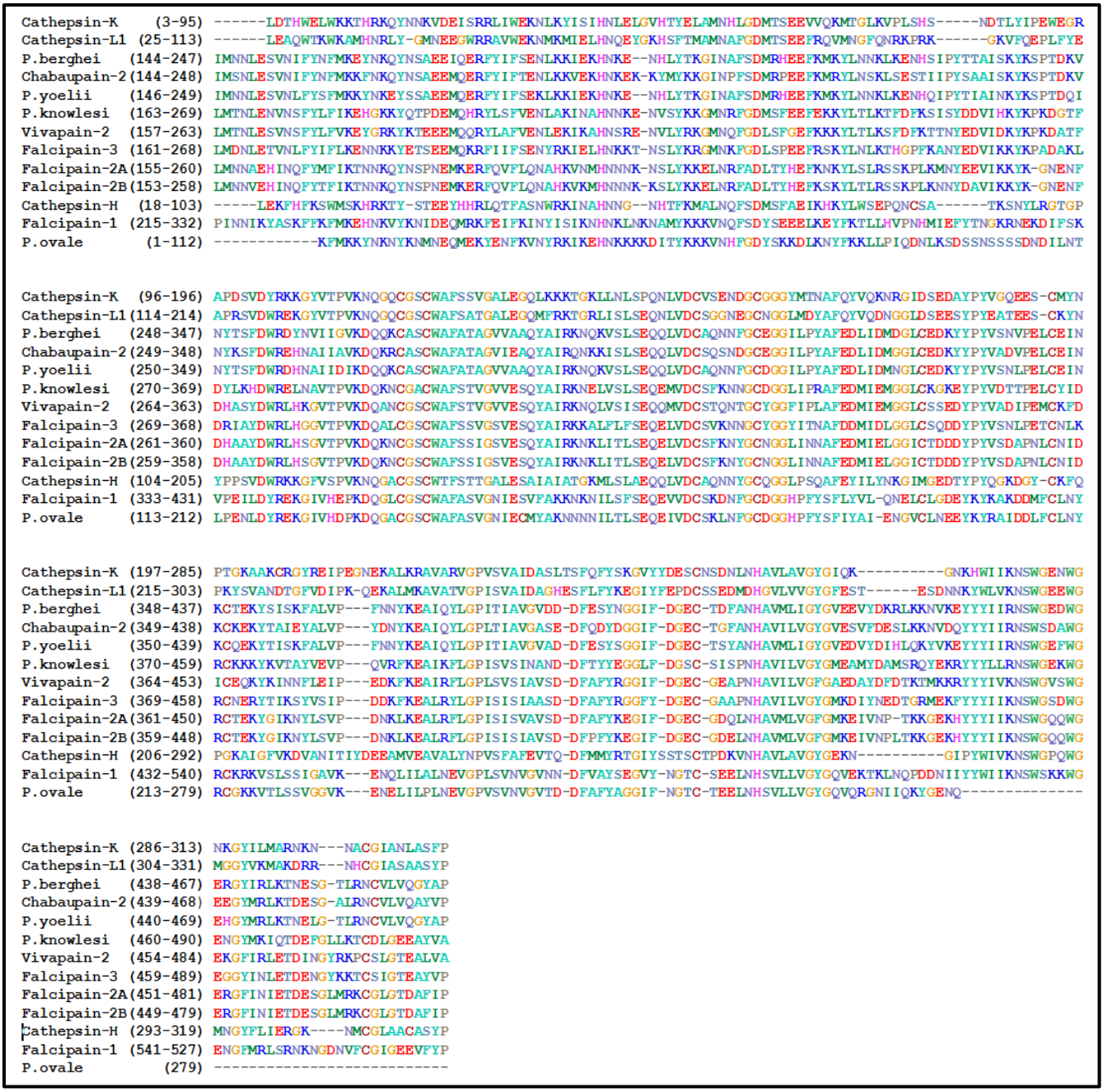

Figure A-2. Alignment of *Plasmodium* and human species used in generating the neighbour joining tree. Alignment was done using PROMALS3D program. Contiguous un-gapped sections of the full length alignment in (Appendix A-I) were used as dataset for phylogenetics

III. Comparison of prodomain alignments generated using the various multiple sequence alignment programs

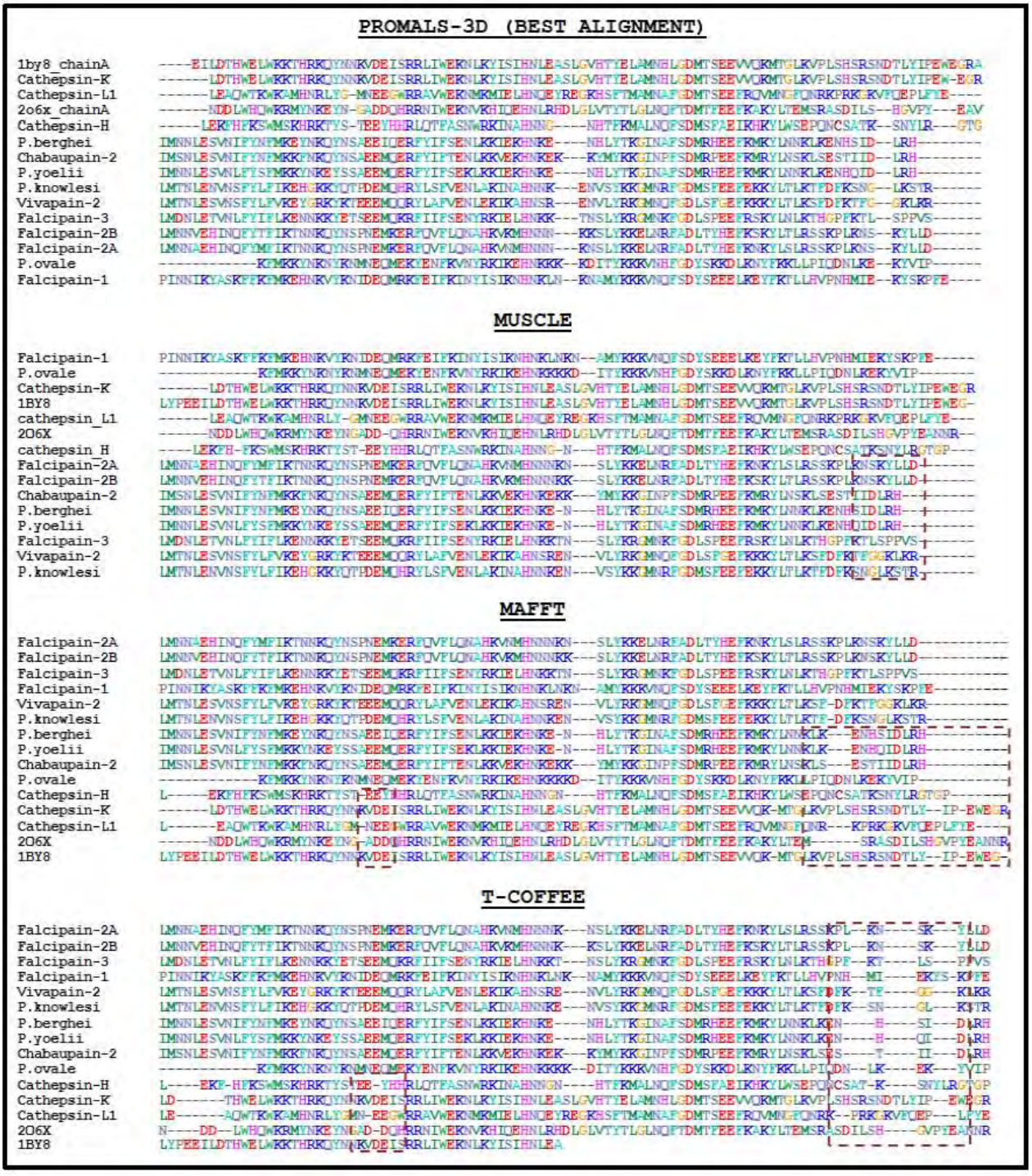

Figure A-3. Comparison of the prodomain alignment by the four alignment programs PROMALS 3D, MUSCLE, MAFFT and T-COFFEE. PROMALS3D gives the best alignment the other alignment programs seem to have inaccuracy (highlighted in red dashes). MUSCLE and MAFFT programs misalign the C-terminal region. T-COFFEE seems to put in unnecessary gaps in the C-terminal region of the alignment.

IV. HHpred structural alignment of FP-2 and the crystallographic structure of cysteine protease in Fasciola hepatica (2O6X).

The HHpred alignment has columns (Q ss_pred and T ss_pred) for the structure prediction of FP-2 and 2O6X respectively, and a column for the actual structure (T ss_dssp). where the three columns are at a consensus it indicates accuracy of alignment. This guide was used to hand adjust alignments.

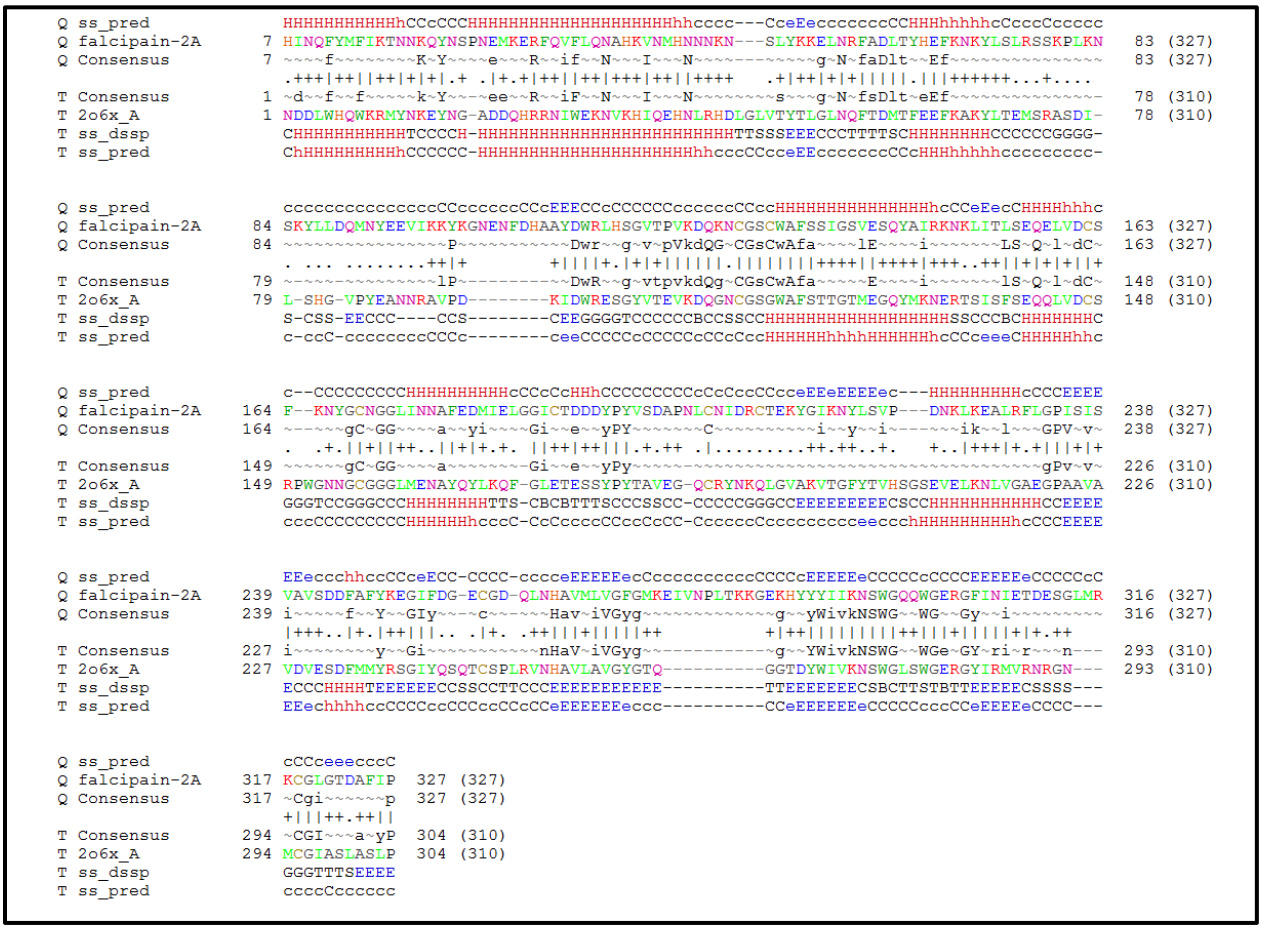

Figure A-4 Structural alignment of FP-2 and 2O6X; Q ss_pred and T ss_pred are predicted secondary structure for the FP-2 and 2O6X, T ss_dssp is the actual secondary structure of 2O6X. Upper and lowercase amino acids show high-60% and low-40% conservation. Symbols rate the alignment as: '|' very good, '+' good, '.' neutral, '-' bad and '=' very bad.

APPENDIX B: HOMOLOGY MODELLING

I. Template-target alignments for homology modelling.

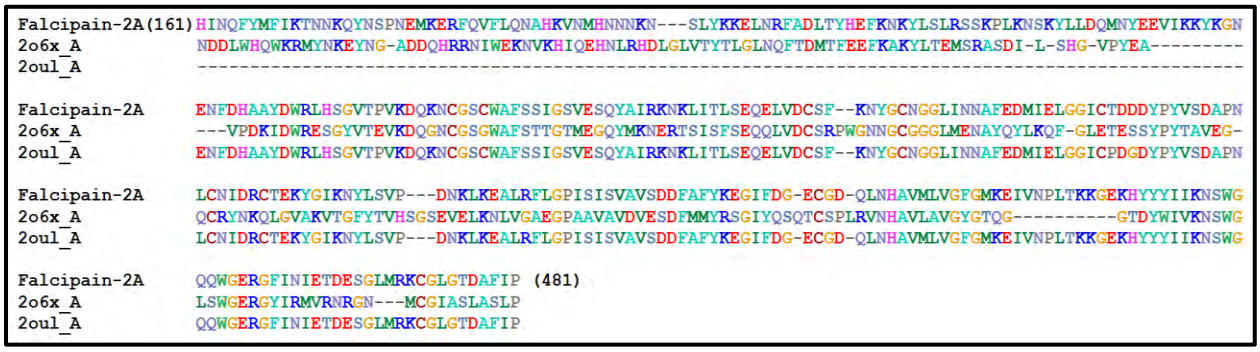

Figure B-1 The alignment for FP-2 with templates 2O6X and 2OUL

Figure B-2 The alignment for FP-2B with templates 2O6X and 2OUL

Figure B-3 The alignment for FP-3 with templates 2O6X and 3BWK

Figure B-4 The alignment for vivapain-2 with templates 2O6X and 3BWK

Figure B-5 The alignment for *P. knowlesi* with templates 2O6X and 2OUL

Figure B-6 The alignment for *P. berghei* with templates 2O6X and 2OUL

Figure B-7 The alignment for *chabaupain-2* with templates 2O6X and 2OUL

Figure B-8 The alignment for *chabaupain-2* with templates 2O6X and 2OUL

II. Scripts for modelling

- Script used to run the automodel class that generates models.

```
# Homology modelling by the automodel class

from modeller.automodel import *    # Load the automodel class

log.verbose()    # request verbose output
env = environ()  # create a new MODELLER environment to build this model in

# directories for input atom files
env.io.atom_files_directory = '/home/joyce/backup/Modelling/FP_2A'

a = automodel(env,
              alnfile  = 'FP2_2templates.pir',     # alignment filename
              knowns   = ('2o6x_A','2oul_A'),               # codes of the templates
              sequence = 'falcipain-2A')                    # code of the target
a.starting_model= 1                 # index of the first model
a.ending_model  = 500                 # index of the last model
                                      # (determines how many models to calculate)
a.final_malign3d = True     # generate superimposed templatesand model (*_fit.pdb files)
a.md_level = refine.very_slow
a.make()                            # do the actual homology modelling
```

- Scripts used to calculate Dope-Z scores. Script written by Matthys Kroon

```
import subprocess

ofile = open("zdope_scores.txt","w")
ofile.write("z-DOPE-score filename\n")
ofile.close()
models = []

for model in open("modellist").readlines():
    models.append(model.strip())

for model in models:
    subprocess.call("mod9v7 zdope_single.py "+model,shell=True)
    subprocess.call("mv zdope_single.log zdope."+model[:-4],shell=True)
exitk
#print models
```

- Script used to sort models in descending order by their Dope-Z scores. Script written by Matthys Kroon

```
ifile=open("zdope_scores.txt")
lines = ifile.readlines()
ifile.close()
scores = []
for line in lines:
    scores.append(line.rsplit())

scores.sort()
ofile=open("sorted_zdope_scores.csv","w")
for line in scores:
    ofile.write(str(line[0])+","+line[1]+",\n")
```

- Script used for loop refinement, script written by Rowan Hatherley

```
# Loop refinement of an existing model
from modeller import *
from modeller.automodel import *

log.verbose()
env = environ()

# directories for input atom files
env.io.atom_files_directory = ['.']

# Create a new class based on 'loopmodel' so that we can redefine
# select_loop_atoms (necessary)

class MyLoop(dope_loopmodel):
    # This routine picks the residues to be refined by loop modeling
    def select_loop_atoms(self):
        # selection of residues to be refined (inclusive)
        return selection(self.residue_range('80:', '84:')
                         (self.residue_range('90:', '94:'))



m = MyLoop(env,
           inimodel='falcipain-2A.B99990337.pdb', # initial model of the target
           sequence='falcipain-2A',
           loop_assess_methods=assess.normalized_dope             # code of the target
       )

m.loop.starting_model= 1            # index of the first loop model
m.loop.ending_model = 100            # index of the last loop model
m.loop.md_level = refine.very_slow   # loop refinement method
m.make()
```