FALCIPAINS AS MALARIAL DRUG TARGETS

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

Master of Science

by

Coursework / Thesis

in

Bioinformatics and Computational Molecular Biology

Department of Biochemistry, Microbiology & Biotechnology

Faculty of Science

by

AQUILLAH M. KANZI

June 2013

ABSTRACT

Malaria is an infectious disease caused by parasites of the *Plasmodium* genus with mortality rates of more than a million annually, hence a major global public health concern. *Plasmodium falciparum* (*P. falciparum*) accounts for over 90% of malaria incidence. Increased resistance to antimalarial drugs by the *Plasmodium* parasite, coupled with the lack of an effective malaria vaccine necessitates the urgent need for new research avenues to develop novel and more potent antimalarial drugs. This study focused on falcipains, a group of *P. falciparum* cysteine proteases that belong to the clan CA and papain family C1, that have emerged as potential drug targets due to their involvement in a range of crucial functions in the *P. falciparum* life cycle. Recently, falcipain-2 has been validated as a drug target but little is known of its *Plasmodium* orthologs. Currently, there are several falcipain inhibitors that have been identified, most of which are peptide based but none has proceeded to drug development due to associated poor pharmacological profiles and susceptibility to degradation by host cysteine proteases. Non-peptides inhibitors have been shown to be more stable *in vivo* but limited information exists. *In vivo* studies on falcipain-2 and falcipain-3 inhibitors have also been complicated by varying outcomes, thus a good understanding of the structural variations of falcipain *Plasmodium* orthologs at the active site could go a long way to ease *in vivo* results interpretation and effective inhibitor design. In this study, we use bioinformatics approaches to perform comparative sequence and structural analysis and molecular docking to characterize protein-inhibitor interactions of falcipain homologs at the active site. Known FP-2 and FP-3 small molecule non-peptide inhibitors were used to identify residue variations and their effect on inhibitor binding. This was done with the aim of screening a collection of selected non-peptide compounds of South African natural origin to identify possible new inhibitor leads. Natural compounds with high binding affinities across all *Plasmodium* orthologs were identified. These compounds were then used to search the ZINC database for similar compounds which could have better binding affinities across all selected falcipain homologs. Compounds with high binding affinities across all *Plasmodium* orthologs were found.

DECLARATION

I, **AQUILLAH M. KANZI**, hereby declare that this thesis submitted at Rhodes University is my own work and has not been previously submitted for a degree in this or any other university.

Signature:  Date: 03 June, 2013

DEDICATION

To my

Dad and Mom,

Stephen Musumbi and Agnes Kanzi

I will always treasure this.

....Aquillah

.

ACKNOWLEDGEMENT

To begin with, I thank the almighty God for giving me the strength, humility, patience, courage and a sound mind to take up this challenge.

I thank my supervisors, Dr. Özlem Tastan Bishop and Dr. Kevin Lobb who tirelessly and generously made it possible for me to complete this project successfully and on time.

I also thank Rhodes University for granting me the opportunity to study, providing a friendly academic environment and most importantly the financial support through the Henderson Bioinformatics Rhodes Prestigious Scholarship 2012.

I thank my colleagues at Rhodes Bioinformatics Research Group as well for their help and support.

Lastly, special thanks go to my parents, brother and sister (Andrew and Aidah) for their prayers and support.

TABLE OF CONTENTS

ABSTRACT i
DECLARATION ii
DEDICATION iii
ACKNOWLEDGEMENT iv
TABLE OF CONTENTS v
LIST OF FIGURES ix
LIST OF TABLES xii
LIST OF ACRONYMS xiii
SYMBOLS USED xiv
LIST OF AMINO ACIDS xv
CHAPTER ONE 1
1. LITERATURE REVIEW 1
1.1 Introduction 1
1.2 *Plasmodium* life cycle 2
1.3 Proteins involved in *Plasmodium* spp. growth and development 3
1.3.1 Kinases 3
1.3.2 Cysteine proteases 3
1.3.3 Metallopeptidases 3
1.3.4 Other proteases 4
1.4 Cysteine proteases and falcipains 5
1.4.1 Cysteine proteases 5
1.4.2 Evolution and classification 5
1.4.3 Functions of *Plasmodium spp.* cysteine proteases 6
1.5 Falcipains (FPs) 8

1.5.1 Falcipains classification and function 9
1.6 Falcipain structure and function 11
1.6.1 Falcipain prodomain 11
1.6.2 Mature domain 12
1.6.3 Falcipain active site and sub sites 13
1.6.4 Falcipain active site mechanism 14
1.7 Structural approach to falcipain inhibition 15
1.7.1 Inhibition of FPs by small molecular weight inhibitors 16
1.7.2 Non-peptide inhibitors 17
1.7.3 Peptidomimetic inhibitors 18
1.7.4 Inhibition of FPs by small proteins &macromolecular molecules 18
1.8 Falcipain homologs 19
1.9 Research problem statement and justification 20
1.10 Aims and objectives 22
1.10.1 Specific Objectives 22
CHAPTER TWO 23
2. SEQUENCE AND PROTEIN STRUCTURE ANALYSIS 23
2.1 Introduction 24
2.1.1 Sequence alignment and alignment algorithms 25
2.1.2 Database similarity search and sequence retrieval 26
2.1.3 Multiple sequence alignment 27
2.1.4 Phylogenetic analysis 28
2.1.5 Homology Modeling 28
2.2 Methodology 33
2.2.1 Database similarity search and sequence retrieval 34

2.2.2 Multiple sequence alignment 35
2.2.3 Phylogenetic analysis 35
2.2.4 Homology modeling 36
2.3 Results and Discussion 40
2.3.1 Sequence retrieval 40
2.3.2 Multiple sequence alignment and structural analysis 42
2.3.3 Phylogenetic analysis 49
CHAPTER THREE 71
3. MOLECULAR DOCKING 71
3.1 Introduction 72
3.1.1 Rigid docking 73
3.1.2 Flexible docking 74
3.1.3 AutoDock4 as a docking tool 74
3.2 Methodology 78
3.2.1 Data preparation for molecular docking 78
3.2.2 Ligand and protein protonation 80
3.2.3 Grid calculation and docking parameter file preparation 81
3.2.4 Docking simulations 82
3.2.5 Docking analysis 82
3.3 Results and Discussion 83
3.3.1 Docking validation 83
3.3.2 Non-peptide inhibitors 84
3.3.3 Screened natural compounds 99
3.3.4 ZINC database search 102

CHAPTER FOUR 107
4. CONCLUSION AND FUTURE PROSPECTS 107
REFERENCES 109
APPENDICES 125
Appendix 1 125
Appendix 2A 126
Appendix 2B 131

LIST OF FIGURES

Figure 1.1: Life cycle of *Plasmodium spp.*. 4
Figure 1.2: Cysteine protease classification. 6
Figure 1.3: a) 3D structure of cathepsin L (3OF8) and b) FP-2 (3BPF) 12
Figure 1.4: Proteolytic mechanism of papain like cysteine proteases. 15
Figure 2.1: Summary the methodology used for sequence analysis and homology modeling. 33
Figure 2.2: Multiple sequence alignment of FP-2, FP-3 homologs as predicted by MAFFT 46
Figure 2.3: Superimposed FP-2 and FP-3 *Plasmodium* homolog 3D models showing active site residue variations 48
Figure 2.4: Phylogenetic analysis of FP-2 and FP-3 homologs. The tree shows distinct clustering of FP-2 and FP-3 *Plasmodium* homologs from the human homologs. 49
Figure 2.5: MetaMQAP validation results for the templates (a) PDBID: 2OUL and (b) PDBID: 3BWK 53
Figure 2.6: ANOLEA and QMEAN6 model quality evaluation results for the template 2OUL (FP-2). 54
Figure 2.7: ANOLEA and QMEAN6 model quality evaluation results for template 3BWK (FP-3) 54
Figure 2.8: (a) Ramachandran plot for template structure, 2OUL (FP-2) chain A, and (b) 3BWK (FP-3) chain A. 55
Figure 2.9: Superimposed models of all retrieved *Plasmodium* homologs colored according to the MetaMQAPII scores 58
Figure 2.10: BP-2 model quality assessment. 59
Figure 2.11:BPy-2 model quality assessment. 60
Figure 2.12: CP-2 model quality assessment 62
Figure 2.13: KP-3 model quality assessment. 63
Figure 2.14: KP-2 model quality assessment. 65
Figure 2.15:VP-2 model quality assessment. 66
Figure 2.16: VP-3 model quality assessment 67

Figure 2.17: Surface presentations of FP-2, FP-3 and, *Plasmodium* homolog model structures.. 70
Figure 3.1: Summary of molecular docking methodology. 78
Figure 3.2:a) Mu-Leu-homoPhe-VsPH docked to FP-2 and b) Mu-Leu-homoPhe-VsPH docked to FP-3 (3BWK). c) FP-2 with docked 2-cyanopyrimidine inhibitors..... 84
Figure 3.3: Estimated free energy of binding of best docked non-peptide inhibitors against FP-2, FP-3 and their *Plasmodium* homolog 3D models. 85
Figure 3.4: Inhibition constants (Ki) of docked non-peptides as predicted by AutoDock4.2.against selected FP homolog protein structures and 3D models…….. 89
Figure 3.5: FP-2 surface presentation with best docked 2-cyanopyrimidine derivative (2_CPI).. 90
Figure 3.6: FP-3 surface presentation with best docked 2-cyanopyrimidine derivative (2_CPI).. 90
Figure 3.7: BP-2 (*P. Berghei*,) surface presentation with best docked 2-cyanopyrimidine (2_CPI) 91
Figure 3.8: BPy-2 (*P. Yoelii Yoelii*) surface presentation with best docked 2-cyanopyrimidine derivative (2_CPI). 91
Figure 3.9: CP-2 surface presentation with best docked 2-cyanopyrimidinederivative (2_CPI).. 92
Figure 3.10: VP-2 surface presentation with best docked 2-cyanopyrimidine derivative (2_CPI).. 92
Figure 3.11: VP-3 surface presentation with best docked 2-cyanopyrimidine derivative (2_CPI).. 93
Figure 3.12: KP-2 surface presentation with best docked 2-cyanopyrimidine derivative (2_CPI).. 93
Figure 3.13: KP-3 surface presentation with best docked 2–cyanopyrimidinederivative (2_CPI).. 94
Figure 3.14: Cathepsin-L surface presentation with best docked 2–cyanopyrimidine derivative (2_CPI).. 94
Figure 3.15: Estimated free energy of binding map for screened natural compounds.. 99
Figure 3.16: inhibitor constant of 5α-pregna-1, 20-dien-3-one against all FP homologs.. 100

Figure 3.17: a) Surface presentation of FP-2 with best docked natural compound (5α-pregna- 1, 20-dien-3-one).. 101
Figure 3.18: Estimated free energy of binding map for screened compounds (0001-186) from the ZINC database search..... 102
Figure 3.19: Docked ZINC database hits with lowest estimated free energy of binding in the range -9.45 kcal/mol and below as predicted by AutoDock4.2. 104
Figure 3.20: VP-2 with docked compounds from the ZINC database. 105

LIST OF TABLES

Table 2.1: List of retrieved FP *Plasmodium* homologs from NCBI and PlasmoDB. 41

Table 2.2: Sub site 1, 2, 3 and 1' sub site residues of FP *Plasmodium* homologs 47

Table 2.3: FP-2 and FP-3 orthologs with their corresponding templates selected from HHpred.. 52

Table 2.4: Positions of mature domains in the actual protein sequence, selected templates and their coverage against respective targets. 53

Table 2.5: Model scores as predicted by modeler and METAMQAPII protein structure validation program.. 57

Table 3.1: List of ligands used for evaluating docking method accuracy as well as elucidate protein-ligand/inhibitor interactions of calculated FP 3D models. 79

Table 3.2: Active site amino acid residues interacting with the best docked ligand (2-cyanopyrimidine derivative, 2_CPI).. 95

LIST OF ACRONYMS

ADT	- AutoDock Tools
AMBER	- Assisted Model Building with Energy Refinement
ANOLEA	- Atomic Non Local Environment Assessment
BLAST	- Basic Local Alignment Sequence Search Tool
BLOSUM	- Blocks of Amino acid substitution
BP-2	- Berghepain-2 (*P. berghei*)
BPy-2	- Berghepain-2 (*P. yoelii*)
CHARMM	- Chemistry at HARvard Molecular Mechanics
CP-2	- Chabaupain-2
DLG	- Docking Log File
DOPE	- Discrete Optimized Protein Energy
Falcipain	- FP
FP-1	- Falcipain-2
FP-2	- Falcipain-2
FP-2’	- Falcipain-2
FP-3	- Falcipain-3
GA	- Genetic Algorithm
GDT_TS	- Global Distance Tests – Total Score
GPF	- Grid Parameter File
LGA	- Lamarckian Genetic Algorithm
KP-2	-Knowlesipain-2
KP-3	-Knowlesipain-3
MAFFT	- Multiple Alignment using Fast Fourier Transform
MEGA	- Molecular Evolutionary Genetic Analysis
MetaMQAP	- Meta Model Quality Assessment Program
NCBI	- National Center for Biotechnology Information
PAM	- Point Accepted Mutations

PDB	- Protein Data Bank
PIC	-Protein Interaction Calculator
PROCHECK	- PROgram to CHECK the stereochemical quality of protein structures
PROMALS3D	- Profile Multiple Sequence Alignment with Local Structure and 3D constraints
QMEAN6	- Qualitative Model Energy ANalysis
RMS	- Root Mean Square
RMSD	- Root Mean Square Deviation
spp.	- Species
VP-2	- Vivapain-2
VP-3	- Vivapain-3
WHO	- World Health Organization

SYMBOLS USED

Å	- Angstrom a measure of atomic distance
μM	- Micromolar, a measure of concentration equal to 10E-3 mM
nM	- Nanomolar, a measure of concentration equal to 10E-6 mM
β	- Beta, used mainly in reference to beta sheets
α	- Alpha, used mainly in reference to alpha helix
π	- Pie bonds, a type of covalent chemical bonds

LIST OF AMINO ACIDS

Name	3 letter code	1 letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamine	Gln	Q
Glutamic acid	Glu	E
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

CHAPTER ONE

1. LITERATURE REVIEW

1.1 Introduction

Malaria is a major public health concern worldwide being one of the most widely spread diseases in the world, affecting about 500 million people and mortality rates of approximately a million people annually (WHO World Malaria Report, 2011). It is caused by parasitic protozoa of the *Apicomplexan* phylum and the genus *Plasmodium* (Vaughan *et al*. 2009). Malaria is transmitted to humans and other vertebrates through the introduction of sporozoites into the blood stream by a mosquito (female *Anopheles*) bite (Kumar & Mishra, 2012; Steinbuechel & Matuschewski, 2009). Clinical manifestation of malaria is experienced during the erythrocytic stage of the parasite's life cycle. Malaria caused by *Plasmodium falciparum* (*P. falciparum*) is the most dangerous form of malaria, being responsible for the majority of malarial deaths a year worldwide (Perlmann *et al*. 2000; Rich *et al*. 2009; WHO World Malaria Report, 2010). Malaria endemic regions include: Africa, Asia, and South America (Aguiar *et al*. 2012). Malaria caused by *P. falciparum* is more prevalent in sub-Saharan Africa than the other regions of the world where less virulent *Plasmodium* species, which include *Plasmodium vivax* (*P. vivax*), *Plasmodium malariae* (*P. malariae*) and, *Plasmodium ovale* (*P. ovale*), predominate (Aguiar *et al*. 2012; Gollin & Zimmermann, 2007). Conventional methods of malaria treatment and control have been hampered by increasing resistance to antimalarial drugs and mosquito insecticides (Greenwood *et al*. 2008; Wongsrichanalai *et al*. 2002).With the absence of an effective vaccine (Hartjes, 2012), which has proven to be a daunting task far from bearing results, there is need to venture into other potential alternatives that have not been fully exploited such as molecular targets (Liñares & Rodriguez, 2007). Falcipains (FPs), namely FP-1, FP-2, FP-2' and FP-3 are a group of *P. falciparum* cysteine proteases that have emerged as potential molecular targets due to their involvement in a host of crucial functions in the growth and development (Rosenthal, 2004). FPs, together with other proteases, hydrolyze hemoglobin which is the main source of essential amino acids required by the parasite for growth and development during its asexual life cycle (Rosenthal *et al*. 2002).

Inhibition of some of these proteases has been shown to have lethal effects (Sijwali *et al*. 2001; Shenai *et al*, 2000).This has been confirmed by gene disruption studies on FP-2 which resulted in blocking hemoglobin degradation (Rosenthal, 2011). Arresting this cycle presents a prime target for drug design and development.

1.2 *Plasmodium* life cycle

The *Plasmodium* lifecycle is complex (Wirth, 2002) and a highly controlled process involving different stages. It begins by introduction of sporozoites into the vertebrate host e.g. humans (Figure 1.1) through a bite by an infected female *Anopheles* mosquito's salivary glands into the bloodstream during feeding (Kumar & Mishra, 2012). Sporozoites have a nucleus, mitochondrion, apicoplast and a microtubule all linked by long tethering proteins and are highly motile cells (Aly & Matuschewski, 2005). The sporozoites disappear from the bloodstream within minutes and quickly infect liver cells where they differentiate over a number of days (Cowman & Kappe, 2006; Sturm *et al*. 2006). The liver-stage parasites differentiates and undergo asexual multiplication resulting in tens of thousands of schizonts which re-invade the blood stream subsequently infecting erythrocytes while undergoing additional cycles of multiplication producing 12-16 trophozoites in a single schizont (Greenwood *et al*. 2008). The erythrocytic stage is the clinically visible stage of the disease. Not all merozoites differentiate into mature schizonts (Cowman & Kappe, 2006). The rest differentiate into sexual forms, male and female gametocytes which are taken up by the female Anopheles mosquito during feeding. The male gametocytes while in the gut undergo a rapid nuclear division producing eight flagellated microgametes which fertilize the female macrogamete Ghosh & Edwards, 2000). The result is an ookinate which traverses the gut wall and encysts on the exterior of the gut wall as an oocyst. Oocysts later rapture releasing hundreds of sporozoites into the mosquito body cavity, and eventually migrating to the mosquito salivary glands and are transmitted to a human host after a mosquito bite and the cycle goes on (Talman *et al*. 2004). As previously mentioned, the life cycle of *Plasmodium* spp. is complex and is highly controlled and it involves many proteins with diverse and specific functions. These include proteins described below.

1.3 Proteins involved in *Plasmodium* spp. growth and development

There are many proteins that are involved in the growth and development of the *Plasmodium* spp. These include proteases (endopeptidases) such that catalyze hydrolysis of peptide bonds (Rosenthal, 2004; Rzychon *et al*. 2004). For the purpose of this study, only proteins with relevance to the *Plasmodium* spp. will be discussed. These include:

1.3.1 Kinases

Several kinases in *P. falciparum* have been identified. However little is known about them. They are mainly associated with completion of the asexual erythrocytic and replication cycle of *P. falciparum* (Dastidar *et al*. 2012; Dorin-Semblat *et al*. 2011).

1.3.2 Cysteine proteases

It has been established through research that proteases especially cysteine proteases play a critical role in the life cycle as well as pathogenicity of many protozoan parasites including *Plasmodium* spp. (Sajid and Mckerrow, 2002). Cysteine proteases also play major roles in parasite immune evasion, enzyme activation, virulence, tissue and cell invasion and, exystment (Rosenthal, 2011). There are a number of cysteine proteases that have been identified in *P. falciparum* which include FPs, serine repeat antigens, dipeptidyl aminopeptidase 1, dipeptidyl aminopeptidase 3 and a calpain homolog (Wirth, 2002).

1.3.3 Metallopeptidases

Metallopeptidases are important in proteolysis of hemoglobin derived oligopeptides (Sijwali, *et al*. 2001). There are at least two essential metallopeptidases which are encoded in the genome PfA-M1 and Pf-LAP (Dalal & Klemba, 2007). Inhibition of these two proteases is lethal to the parasite as a result of starvation. Pf-LAP is mainly involved in the early stages of the lifecycle before hemoglobin hydrolysis suggesting different functions (Harbut *et al*. 2011).

Falcilysin is zinc metallopeptidase which is involved in degradation of small polypeptides up to 20 amino acids to produce even shorter nucleotides which are usually products of other proteases such as aspartyl and cysteine proteases (Ettari *et al*. 2009).

1.3.4 Other proteases

These include proteases such as aspartyl proteases. Plasmepsins are the most common of aspartyl proteases in *Plasmodium* which are also involved in the degradation of hemoglobin by *P. falciparum*(Gupta *et al*. 2010). Subtilin like protease is also encoded in the genome. At least two of these are known; these are serine proteases expressed in three late asexual stages of *P. falciparum* (Alam *et al*. 2012). Histo-aspartic proteases (HAP) have been crystallized and have unique features which are potential drug targets (Bhaumik *et al*. 2011). Dipeptidyl aminopeptidase 1 cleave dipeptides derived from hemoglobin oligopeptides in the parasite food vacuole (Ettari *et al*. 2009).

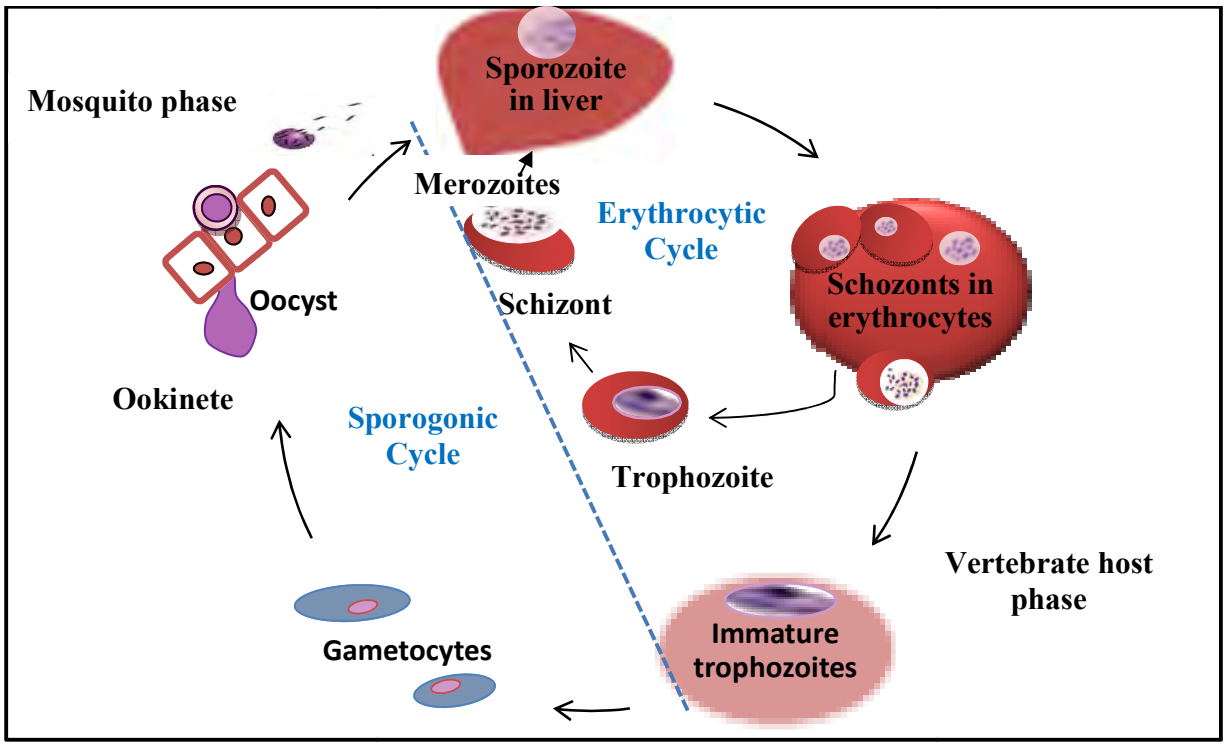


Figure 1.1: Life cycle of *Plasmodium spp.* On the right is the vertebrate host phase while on the left is the mosquito phase of the lifecycle. FPs are found both in schizonts and trophozoites.

1.4 Cysteine proteases and falcipains

1.4.1 Cysteine proteases

This is a group of proteases also referred to as thiol or sulfhydryl proteases which are an important biological group of enzyme catalysts that hydrolyze peptide bonds of polypeptides (Sajid & Mckerrow, 2002). They have a common catalytic mechanism involving nucleophilic cysteine in a catalytic triad of Cys, His and, Asn (Pandey *et al*. 2012). The name cysteine is derived from the catalytic cysteine present in this group of proteases which mediates a nucleophilic attack on a carbonyl carbon of a susceptible peptide (Rosenthal, 2004). FPs belong to this protease classification.

1.4.2 Evolution and classification

The first cysteine protease to be isolated and characterized was the papain form the papaya fruit *Cirica papaya* in 1879. It was also the first cysteine protease to have its protein structure solved (Sajid & Mckerrow, 2002). Since then, homologous groups have been identified and grouped into clans which do not necessarily share high sequence or structural identity, but share a common catalytic mechanism for peptide hydrolysis, which implies probable independent evolution of the proteins (Rosenthal, 2004). Papain group also known as clan CA group is further divided into families based on sequence similarity and identity (Figure 1.2). Among parasitic organisms, cysteine proteases belonging to the C1 family include cathepsin B, cathepsin L-like and C2 or calpain-like (Rosenthal, 2004). FPs belong to the clan CA, family C1 and cathepsin L-like subfamily (Sajid & Mckerrow, 2002). Cysteine proteases are assigned families depending on characteristics such as possession of inserted peptide loops sequence, similarity and, specificity to small peptide substrates. More precise classification takes into account sequence homology spanning the whole sequence, catalytic site Cys, His and Asn (Sajid & Mckerrow, 2002) .

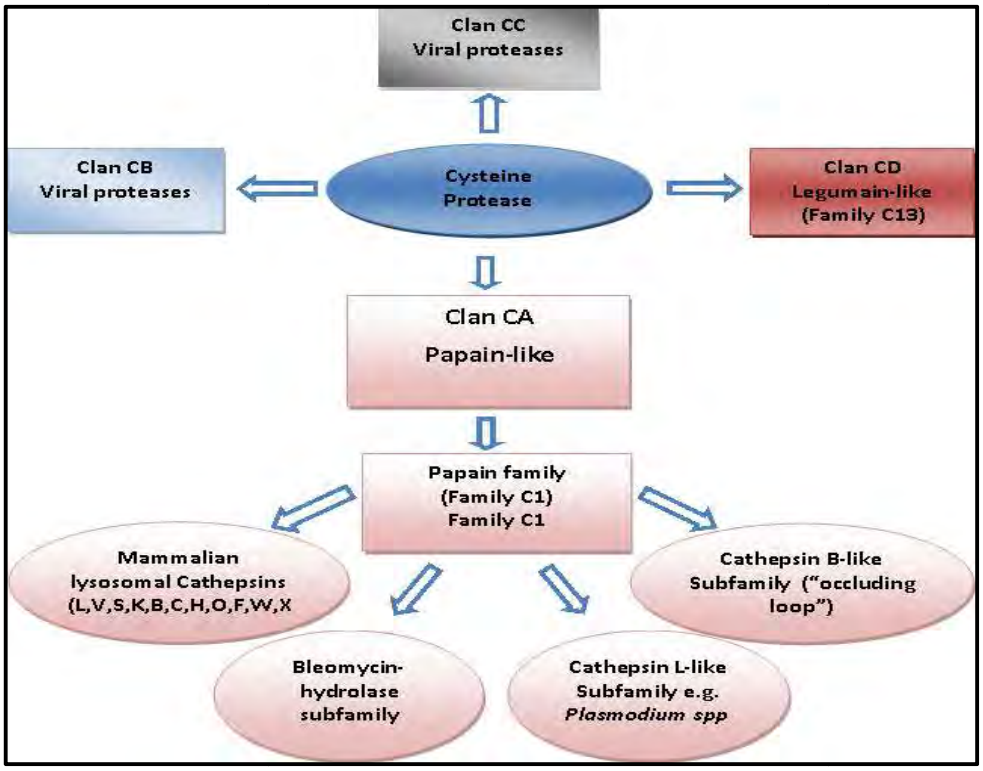


Figure 1.2: Cysteine protease classification. Flow diagram shows different parasite and mammalian cysteine protease groups (Lecaille *et al*. 2002, Sajid & Mckerrow, 2002).

1.4.3 Functions of *Plasmodium spp.* cysteine proteases

Cysteine proteases perform essential functions in the life cycle of *P. falciparum*. These functions make them prime drug targets and hence the attention they have gained for antimalarial chemotherapy development. These functions include: hemoglobin degradation, tissue and cell invasion, protein processing and activation and immune evasion.

1.4.3 Hemoglobin degradation

P. falciparum is dependent on the host for supply of essential amino acids for its growth and development. Merozoites, after release from the hepatocytes, invade erythrocytes which contain amino acid sources important for the growth and development of the parasite. These amino acids are derived from hemoglobin which is degraded by several proteases (Eggleson *et al*. 1999; Gupta *et al*. 2010; Hansen *et al*. 2011; Rosenthal, 2004; Rosenthal, 2011; Salas *et al*. 1995).

Hemoglobin degradation occurs in the parasite's acidic food vacuole enhanced by the cytostome machinery. Several proteases are involved in this process but the exact sequence of events is not well defined (Ettari Roberta *et al*. 2009). There are two proposed pathways. In one of the proposed pathways, aspartyl proteases specifically plasmepsins (Gluzman *et al*. 1994; Goldberg *et al*. 1991) have been postulated to initiate hemoglobin degradation by hydrolyzing peptide bonds of the main chain residues of the native hemoglobin (Banerjee *et al*. 2002), releasing the heme moiety and further degradation by plasmepsins, FP-2 andFP-3 cysteine proteases, falcilysin a metalloproteases and dipeptidyl aminopeptidase 1 (Liu *et al*. 2006). The other postulated pathway maintains that FPs are involved in the initial degradation of hemoglobin after it was deduced that FPs were able to degrade native hemoglobin (Rosenthal *et al*. 1988; Subramanian *et al*. 2009).Both aspartic and cysteine proteases are able to cleave both native and cleaved hemoglobin (Liu *et al*. 2006). FP-2 for instance, is able to cleave native and denatured hemoglobin (Salas *et al*. 1995). Despite the conflicting sequence of events that have been proposed during hemoglobin degradation, the involvement of FPs has been proven (Shenai *et al*. 2002; Sijwali *et al*. 2004; Wang *et al*. 2007; Shenai *et al*. 2000). Plasmepsins knockout studies have shown that they played an important but not essential role in parasite development because they have the ability to compensate for loss of individual plasmepsins (Omara-Opyene *et al*. 2004). FP inhibition is known to be lethal to the parasite as it irreversibly blocks rapture of host cell membrane, thus preventing fresh erythrocyte invasion by the parasites(Sijwali & Rosenthal, 2004). Metalloproteases (falcilysin), breakdown small hemoglobin polypeptides into oligopeptides, which are further degraded by dipeptidyl aminopeptidase into smaller peptides that are pumped out into the parasite cytoplasm (Ettari Roberta *et al*. 2009; Rzychon *et al*. 2004).

1.4.3.2 Tissue and cell invasion

The *Plasmodium* lifecycle involves migrations from one host compartment to another i.e. cell invasion and tissue migration (Ghosh & Edwards, 2000). Several merozoite and erythrocyte proteins are degraded prior or during erythrocyte invasion (Hanspal *et al*. 2002; Sturm *et al*. 2006; Vaughan *et al*. 2009; Rosenthal, 1998). Cysteine protease inhibitors have been shown to have an effect on the invasion process as well as release of late schizonts from erythrocytes due to delayed protein processing (Rosenthal, 2004) .

1.4.3.3 Protein processing and activation.

Merozoites undergo various surface protein processing before or during release from the hepatocytes into the blood stream (Sturm *et al*. 2006). There is research evidence indicating that merozoites surface protein processing is a prerequisite for release. Cysteine inhibitors have been shown to inhibit this process and hence merozoites release (Sajid & Mckerrow, 2002).

1.4.3.4 Immunoevassion

Cysteine proteases have been hypothesized to aid in immune evasion of the host by degrading immune effectors or interfering with cellular immune responses (Sajid & Mckerrow, 2002). There are several examples showing that cysteine proteases cleave immunoglobulins *in vitro*. However, there is still much research being done to verify this hypothesis. Degradation of antibodies by parasite cysteine protease has been documented in some protozoan parasites e.g. *T. cruzi, E. hitolytica* and *G. lamblia* (Sajid & Mckerrow, 2002).

1.5 Falcipains (FPs)

The term FPs is used in reference to a group of *P. falciparum* trophozoite cysteine proteases (Rosenthal, 2004). Research aimed at identifying enzymes responsible for hydrolyzing hemoglobin in the *Plasmodium* trophozoite food vacuole, led to the identification of FPs (Salas *et al*. 1995). *Plasmodium spp.* growth and development is dependent on hydrolysis of hemoglobin as a source of essential amino acids (Shenai & Rosenthal, 2002). The trophozoite food vacuole is slightly acidic but FPs and homologs have an optimum pH of up to 7.0 (Sajid & Mckerrow, 2002; Shenai & Rosenthal, 2002). Hemoglobin is transported to the food vacuole where it is hydrolyzed by multiple enzymes, including cysteine, aspartic and metaloproteases, and cytosolic aminopeptidases, as previously discussed in section 1.4.3, to malarial pigment and globin which is further hydrolyzed to free amino acids (Lew *et al*. 2003).

Gene disruption and biochemical experimental tests have shown that FP-2 and FP-3 are crucial for hemoglobin degradation. Incubation of parasites with cysteine inhibitors blocked globin hydrolysis (Sijwali & Rosenthal, 2004). Hemoglobin hydrolysis inhibition results into a characteristic abnormality in the food vacuole, causing it to fill up with undegraded hemoglobin consequently blocking parasite development (Shenai *et al*. 2000; Sijwali *et al*. 2001).

1.5.1 Falcipains classification and function

There are four known FPs that have been identified and characterized from the *P. falciparum* genome which are: FP-1, FP-2, FP-2' and, FP-3(Rosenthal, 2004). From previous gene disruption analyses of FPs, it has been determined that the four identified FPs do possess independent functions apart from the shared hemoglobinase function between FP-2 and FP-3 (Sijwali *et al*. 2006)

1.5.1.1 Falcipain-1

The function of this particular protease is poorly understood. Falcipain-1 (FP-1) is found in relatively low levels compared to the other FPs in the erythrocytic stage of the parasite. This, coupled with difficulties in production of recombinant protease have limited biochemical studies as well as functional characterization (Kumar *et al*. 2007). However, earlier studies by Salas *et al*. 1995, suggested involvement of FP-1 as a hemoglobinase. Other studies have shown that FP-1 is active during merozoite invasive stage. FP-1 inhibitors successfully blocked parasite invasion of host erythrocytes, but had no effect on parasite hemoglobin hydrolysis (Greenbaum *et al*. 2002). However, later studies have shown that FP-1 alone is neither required for parasite invasion nor intracellular development within erythrocytes, suggesting functional overlap in activity with other proteases. Gene disruption studies have suggested a role in oocyst development in the mosquito midgut by directly activating proteins by proteolytic processing or by degrading midgut endothelium facilitating ookinate migration (Eksi *et al*. 2004).

1.5.1.2 Falcipain-2

Falcipain-2 also (FP-2A), is a *P. falciparum* cysteine protease with hemoglobinase activity that is predominantly expressed in the parasite acidic food vacuole of merozoites, trophozoites and early, schizonts (Shenai *et al*. 2000). Inhibition of FP-2 blocks degradation of hemoglobin and consequently parasite growth and development (Sijwali *et al*. 2006) . This makes it a prime target for drug development. Knockout experiments reveal that hemoglobin hydrolysis is blocked. However parasites recovered from the effect because of expression of FP-3 in the later in the parasite life cycle (Dahl & Rosenthal, 2005; Sijwali & Rosenthal, 2004).

1.5.1.3 Falcipain-2'

Falcipain-2' (FP-2') or falcipain-2B (FP-2B) is almost identical to FP-2. Its functions are also not well characterized biochemically (Singh *et al*. 2006). Experiments in which FP2' protease gene was interfered with did not yield any significant changes in the erythrocytic parasites and may have been as a result of compensatory effects from FP-2. However, due to their high sequence similarity of 99% at the catalytic domain (Rosenthal *et al*. 2004) it can be anticipated to be involved in hemoglobin hydrolysis and that FP-2 inhibitors would have similar activity against FP-2'. This hypothesis seems valid after previous recombinant studies on FP-2' show that it hydrolyses hemoglobin and is inhibited by potent inhibitors in the same way as FP-2 (Singh *et al*. 2006)

1.5.1.4 Falcipain-3

Falcipain-3 (FP-3) is expressed in late trophozoites and early schizonts later in the erythrocytic stage of the parasite's life cycle. It is a more efficient hemoglobinase than FP-2 in more acidic pH conditions where it is quickly processed to an active form and its more active and stable at these conditions (Sijwali *et al*. 2001). The two share 66% (catalytic domain only) sequence identity as per the pair wise sequence alignment in chapter 2.FP-2 and FP-3 contribute more or less equally in digestion of hemoglobin in parasites merozoite/trophozoite food vacuole, but FP-3 is less susceptible against typical cysteine protease inhibitors compared to FP-2.

However, FP-3 is more active against native hemoglobin but is less sensitive to FP inhibitors compared FP-2 (Sijwali *et al*. 2001). Knockout studies have shown that *Plasmodium* parasites could not cope without FP-3 (Sijwali *et al*. 2006). Cysteine protease inhibitors targeting FP-2 and FP-3 block degradation of hemoglobin thus interfering with the development of the parasite, thus potential drug targets (Rosenthal, 2011; Rosenthal *et al*. 2002; Shenai *et al*. 2000).

1.6 Falcipain structure and function

FPs, just like other clan CA parasite cysteine proteases have two main domains, prodomain and mature or active domain (Pandey *et al*. 2009). They are synthesized as inactive zymogens which are converted to active forms following proteolytic cleavage and release of the active domain (Shenai *et al*. 2000).The prodomain serves to inhibit the immature catalytic domain until maturation and subsequent activation (Pandey & Dixit, 2012). The catalytic domain is further divided into the L and R domains (Figure 1.3) with a 'V' shaped active site cleft in between them (Hogg, *et al*. 2006). The L domain is mainly alpha helical and the R domain fold into a β-barrel (Figure 1.3). The α-helical folding of the L domain sterically blocks access to the active site hence the latency of the prodomain (Turk *et al*. 2002).

1.6.1 Falcipain prodomain

The N-terminal of the prodomain is required for trafficking FP-2 and FP-3 as well as transport food into the food vacuole (Subramanian *et al*. 2007). This is made possible by small domains which include 35 cytosolic amino acids, a 20 amino acid trans-membrane domain and a luminal domain of 188 amino acids that make up the N-terminal (Pandey & Dixit, 2012). FPs are synthesized as zymogens and the C-terminal portion of the prodomain is required to inhibit the activity of mature enzyme. The prodomain also has unique motifs that are responsible for inhibition of FPs (Pandey *et al*. 2009). In this portion are two highly conserved motifs ERFNIN and GNFD in the cathepsin-L subfamily of proteases which have been shown to mediate the mature domain inhibition (Pandey & Dixit, 2012). These two conserved motifs are also involved in proper folding and/or maintaining the structure of the prodomain (Pandey *et al*. 2009; Korde *et al*. 2008).

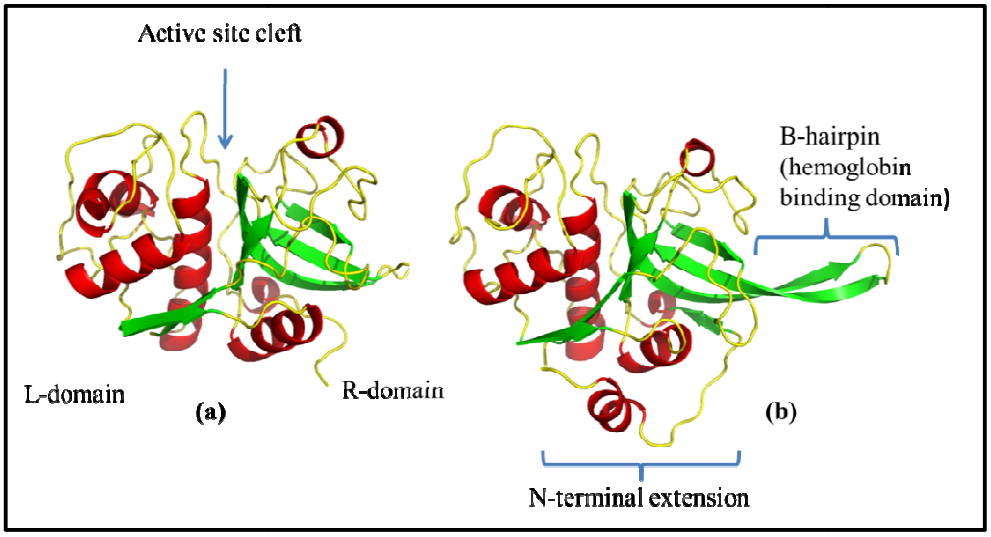


Figure 1.3: a) 3D structure of cathepsin L (3OF8) showing the enzyme active site, L-domain and R-Domain. b) FP-2 (3BPF) protein structures showing additional features not present in cathepsin L.

1.6.2 Mature domain

The mature (catalytic) domain is the active form of FPs as well as other parasite cysteine proteases (Sijwali *et al*. 2002). It dissociates from the prodomain on transportation to the trophozoites food vacuole via the endoplasmic reticulum/Golgi system during which the residues on the N-terminal containing the membrane anchor are proteolytically removed (Dahl & Rosenthal, 2005). In FP-2 enzyme consists of residues beginning from position 244-484 while the F-3 is comprised of residues 243-492 which comprise the mature domain (Kerr, Lee, Pandey, *et al*. 2009). Throughout this study residues will be renumbered according to the mature domain and with the position of the residues in the whole sequence (prodomain and mature domain) indicated in brackets. The N-terminal of the mature domain is required for refolding (Korde *et al*. 2008). FPs have a short N-terminal extension of approximately 20 amino acids which is functionally conserved (Hogg *et al*. 2006; Kerr *et al*. 2009; Pandey *et al*. 2005).

This extension is a unique feature of the mature domain not observed in the family C1A of other cysteine proteases. Previous studies have suggested involvement of the N-terminal extension in the refolding of FP-2 and FP-3 (Pandey *et al*. 2004; Sijwali *et al*. 2002).

Structural and functional analysis of FPs has shown that they have a hemoglobin binding region comprised of 14 residues close to the C-terminal end that forms a structurally distinct extension that differentiates FPs from other clan C1A cysteine proteases (Hogg *et al*. 2006; Pandey *et al*. 2005). This domain is found in between the highly conserved catalytic site of His174 (417) and Asn204 (447) residues in FP-2 and corresponding positions in *Plasmodium* homologs (Chapter 2, Figure 2.2). Studies on this motif suggest that the motif initiates specific interactions with hemoglobin during hydrolysis in the food vacuole of the parasite trophozoite (Pandey *et al*. 2009). Other conserved features in the mature domain are the active site residues which are involved in substrate binding (Chapter 2, Figure 2.2).

1.6.3 Falcipain active site and sub sites

The active site of FPs is typical to the cysteine protease active site, located in a grove between the structurally distinct domains of the papain-like fold (Figure1.3 a). At the heart of it, is the highly conserved catalytic amino acid residues of Cys, His and Asn (Rosenthal, 2004). FP-2 and FP-3 have an additional conserved amino acid residue at the active site which is Gln36 (279)/45 (287) which has been shown to be involved during the initial binding of the substrate (Kerr *et al*. 2009). Papain-like cysteine proteases and related proteases which includes FPs, have a large active site which is divided into pockets at each side of the active site comprised of S1-S3 and S1' (Chapter 2, Figure 2.3 and Table 2.2). Each pockets accommodate an amino acid residue at positions corresponding to P1-P4 and P1'-P3' of the peptide substrate (Kerr *et al*. 2009; Sajid & Mckerrow, 2002). The S1 sub site is the least characterized of the four sub sites. It has a conserved Gln that forms the "oxyanion hole" important for substrate stabilization of the hemithioketal transient state formed between the substrate and protease during hydrolysis (Sabnis *et al*. 2003; Kerr, Lee, Pandey, *et al*. 2009). The S2 substrate pocket of FP-2 and FP-3 is predominantly hydrophobic with preference for substrates with a hydrophobic residue particularly Leu at the P2 position but can accommodate Phe and Val (Sijwali et el. 2001; Shenai *et al*. 2000). The S2 substrate pocket is a major determinant of specificity that can be exploited

as a drug target (Kerr *et al*. 2009; Kerr *et al*. 2009), Pandey & Dixit, 2012). The S3 is a Gly rich region (Chapter 2, Table 2.2) which is highly conserved across all papain-like cysteine proteases.

It is has been shown to form extra hydrogen bonds with the substrate/inhibitor which strengthens the binding (Kerr *et al*. 2009; Sabnis *et al*. 2003). The S1' region contributes to further hydrophobic interactions between substrate and protease. At this site is a highly conserved Trp residue across all cysteine proteases that has been implicated in these hydrophobic interactions (Sabnis *et al*. 2003).

1.6.4 Falcipain active site mechanism

The FP proteolytic mechanism is similar to that of other cysteine proteases. Figure 1.4 summarizes the flow of reactions during substrate hydrolysis process. It starts with deprotonation (Figure 1.4 a) of the thiol in the enzymes active site Cys by an adjacent basic side chain with His residue (Rzychon *et al*. 2004). The deprotonated Cys anionic sulfur initiates a nucleophilic attack on the substrate's carbonyl carbon and in the process a new carboxy-terminus of the substrate to the Cys thiol is formed hence the name thiol proteases (Figure 1.4 b). Finally to regenerate the free enzyme, the thioester bond is subsequently hydrolyzed generating a carboxylic acid moiety on the remaining substrate fragment (Figure 1.4 c and d) (Rzychon *et al*. 2004). The enzymes are active at slightly acidic environments of pH of 4.5-6.5 (Turk *et al*. 2002). In Cathepsins the catalytic site residues of His and Cys form an ion pair which is stabilized by Asn through a hydrogen bond (Lecaille *et al*.2002). The nucleophilic Cys residue is usually deprotonated (Figure 1.4 a) prior to substrate hydrolysis as previously described hence *a priori* activated. Hydrolysis (Figure 1.4 b) of the substrate starts with the nucleophilic thiolate Cys attacking the scissile bond of the substrate to form a tetrahedral intermediate stabilized by an oxyanion hole (Lecaille *et al*. 2002; Kerr *et al*. 2009; Sabnis *et al*. 2003). Acylation (Figure 1.4b) then follows where the C-terminal portion of the substrate is released after a change in the tetrahedral intermediate into an enzyme-thiol ester (acyl enzyme). This step is followed by hydrolysis of the acyl enzyme with water forming a second tetrahedral intermediate which undergoes deacylation eventually splitting into a free enzyme consequently releasing the N-terminal portion (Figure 1.4 c and d) of the substrate (Lecaille *et al*. 2002). Lastly His is restored to its deprotonated form by a released fragment of the substrate with an amine terminus (Rzychon *et al*. 2004).

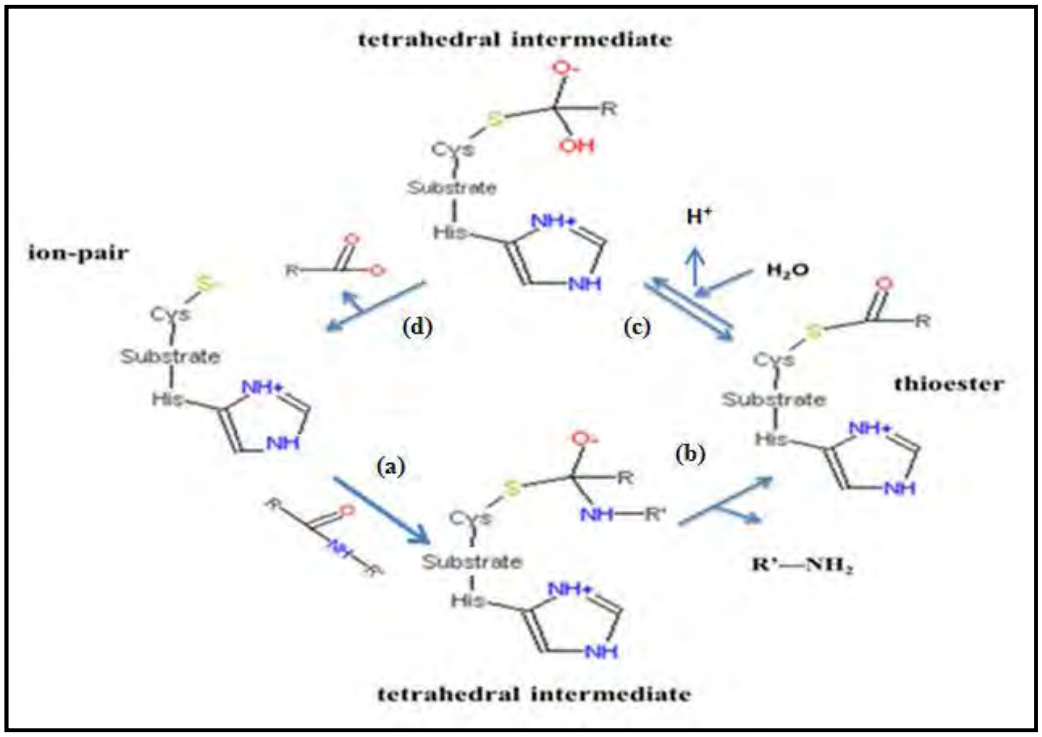


Figure 1.4: Proteolytic mechanism of papain like cysteine proteases. The reaction starts with a) deprotonation, b) hydrolysis, c) acylation and, d) deacylation (Lecaille *et al*., 2002).

1.7 Structural approach to falcipain inhibition

As previously mentioned in section 1.4 of this chapter, cysteine proteases including FPs possess unique functional properties thus potential molecular targets for malaria drug development (Rosenthal, 2004). FP-2 has been validated as an antimalarial drug target, while FP-3 is a potential drug target. FP-3 has been shown to be more active against hemoglobin compared to FP-2. Knockout studies on FP-3 have shown lethal effects on the parasite (Sijwali & Rosenthal, 2004). There are several well-known FP inhibitors, some of which have been co crystalised with FP-2 and FP-3 (Kerr *et al*. 2009; Kerr *et al*. 2009; Pandey *et al*. 2006; Rzychon *et al*. 2004; Verissimo *et al*. 2008; Ettari *et al*. 2009). These protein-inhibitor complexes provide structure activity relationship information that could be used as a basis for inhibitor design.

Structure guided approach to drug design has proven to be valuable in the design of potent and highly selective small molecule inhibitors of various proteases e.g. aspartyl protease in HIV (McKerrow *et al*. 1999). Currently the structure of FP-2 and FP-3 in complex with hemoglobin has not yet been solved. However, from the available FP-2/3-inhibitor complex structures (Hogg *et al*. 2006b; Kerr *et al*. 2009; Kerr *et al*. 2009; Wang *et al*. 2007), it is possible to obtain critical information on the events that underline molecular recognition events (Redzynia *et al*. 2009). Experimental data obtained is often based on information from structure based inhibitor design whose theoretical principles provide putative or proposed ligands which are synthesized and tested (Pandey & Dixit, 2012). A lot has been done to characterize known cysteine protease inhibitors to identify important inhibitor-protein interactions and selectively factors of FPs and related *Plasmodium* homologs (Brady & Cameron, 2004; Kerr *et al*. 2009; Na *et al*. 2004; Shenai *et al*. 2003). In practice, the structure of a cysteine protease inhibitor should have an electrophilic moiety to interact with the cysteine residue of the enzyme and one or two series of substituents, P1-P4 and/or P1'-P3', to interact with the different pockets of the protease (Coterón *et al*. 2010). This, however, depends on the nature and method of binding and subsequent inhibition intended which could be substrate like binding, partial substrate like binding, blockage of active center and backward binding which could result in covalent or non-covalent interactions (Lecaille *et al*. 2002; Rzychon *et al*. 2004). These inhibitors can be grouped into small molecular weight molecules (peptide based), nonpeptides inhibitors, peptidomimetic inhibitors and macromolecular inhibitors which are discussed below.

1.7.1 Inhibition of FPs by small molecular weight inhibitors

From the crystal structures of FP-2 and FP-3 co crystalised with E-64, Leupeptin and vinyl sulfone inhibitors, these small molecular weight inhibitors have been shown to cause irreversible inhibition of FP-2 and FP-3 at low molecular ranges (Kerr *et al*. 2009; Kerr *et al*. 2009b).

Information gathered from structure activity relationship studies on peptidyl vinyl sulfones have revealed important interaction properties of FP-2 and FP-3 with small molecule inhibitors (Liñares & Rodriguez, 2007; Shenai *et al*. 2003). From the crystal structures, active site residues important for inhibitor interactions have been identified (Kerr *et al*. 2009). The enzymatic inhibitory activity of these inhibitors have been studied in FP-2, FP-3 and related *Plasmodium* homologs which has resulted to important information regarding sub site specificities and preferences (Chan *et al*. 2005; Na *et al*. 2004; Shenai *et al*. 2000; Shenai *et al*. 2003; Sijwali *et al*. 2001; Singh and Rosenthal, 2001). Some of these have also been tested *in vivo* producing good results (Olson *et al*. 1999). Despite excellent inhibitory abilities of these inhibitors, limitations such as low absorption rates through cell membranes, susceptibility to other host protease degradation, poor pharmacological profiles, and reduced bioavailability have slowed their development into drugs (Ettari *et al*. 2009). This information could be used to guide design of more potent inhibitors.

1.7.2 Non-peptide inhibitors

Derivatives of existing antimalarials such as isoquinolones have been modified to contain the isoquinolone ring as the basic structure (Batra *et al*., 2003). Design and synthesis has been enhanced by availability of structure activity relationship information obtained from complexed crystal structures of FP-2 and FP-3 with known small molecular weight inhibitors such as E-64 and leupeptin respectively (Kerr *et al*. 2009). Other non-peptide inhibitors include chalcone, the natural precursors of flavonoids with known antimalarial activity. Chalcone derivatives e.g. alkoxylated and hydoxylated chalcones and thiosemocarbazones have experimentally been shown to exhibit *Plasmodium* cysteine protease inhibitor activity (Chipeleme *et al*. 2007; Chiyanzu *et al*. 2003; Ettari *et al*. 2009; Greenbaum *et al*. 2004).

1.7.3 Peptidomimetic inhibitors

Peptidomimetic inhibitors are modified peptide inhibitors in which a non-peptidic scaffold is incorporated into the amino acid backbone. This bears several advantages including increased selectivity as a result of the bioactive compound being stabilized, thus reduced sensitivity to the human host proteases (Verissimo *et al*. 2008). Examples of non-peptidic scaffolds that have been largely studied include peptidimimetics based on 1, 4-Benzodiazepine scaffold and those based on a pyridone ring scaffold (Verissimo *et al*. 2008).

Other examples include peptidomimetic nitriles and vinyl containing peptidomimetics (Ehmke *et al*. 2011; Ettari *et al*. 2011; Verissimo *et al*. 2008). Incorporation of a non-peptide scaffold locks the amino acid backbone to a defined conformation improving the pharmacokinetics and pharmacodynamics of the inhibitor (Ettari *et al*. 2009).

1.7.4 Inhibition of FPs by small proteins &macromolecular molecules

These inhibitors are polypeptide in nature and are generally present within the organisms as regulatory proteins. Cysteine protease endogenous inhibitors include cystatin, chagasin and falstatin (Florent *et al*. 2005; Pandey *et al*. 2006; Redzynia *et al*. 2009). Cystatin inhibits most papain family cysteine proteases with high affinity, hence ideal for co crystalization with FPs (Mukherjee *et al*. 2007; Pandey *et al*. 2006; Rzychon *et al*. 2004). Co crystalization of chagasin or cystatin with FPs shows interactions at prime sites where information can be obtained to clarify FPs binding modes and identify structural requirements for broad spectrum reactivity as well as identify specificity determinants (Pandey *et al*. 2006; Redzynia *et al*. 2009; Santos *et al*. 2006; Wang *et al*. 2006; Wang *et al*. 2007). These complexes can be used to guide structure based design of potent malaria parasite inhibitors (Pandey & Dixit, 2012).

1.8 Falcipain homologs

A number of FP orthologs have been identified since the identification of FPs (Caldeira *et al*. 2009; Chan *et al*. 2005; Desai & Avery, 2004; Na *et al*. 2004). These orthologs can be identified via sequence comparison to FPs. FP-1 orthologs have been shown to have at least 55% sequence identity. FP-2 and FP-3 also have homologs from other human malaria parasites, primate parasites and murine parasites which have been shown to have approximately 50% sequence identity (Rosenthal *et al*. 2002). FP homologs from human malaria parasites have been identified e.g. Vivapain-2 (VP-2) and Vivapain-3 (VP-3) form *P. vivax* (Desai & Avery, 2004) and have been experimentally proven to express hemoglobinase activity (Na *et al*. 2004). Other FP homologs of interest include those from *P. knowlesi* a major primate malaria parasite which has been implicated in symptomatic malaria in humans acquired experimentally, accidentally or naturally (Jongwutiwes *et al*. 2004).

FP homologs from rodent parasites *P. yoelii, P. chabaudi* and *P. berghei* are also of major interest as they are commonly used as animal models for malarial drug study projects (Carlton *et al*. 2002; Jambou *et al*. 2011; Rosenthal *et al*. 2002; Stephens *et al*. 2012). Human homologs include lysosomal cathepsins which perform essential cellular functions and have also been indicated in disease such as cancer, metastasis, rheumatoid arthritis, osteoarthritis, immune diseases, metabolic syndromes and atherosclerosis. These homologs have been successfully targeted for drug development (Reiser *et al*. 2010). It is therefore, important to assess the interaction characteristics of current and future FP inhibitors on these homologs which could prove useful in the design of new and potent FP-2 and FP-3 inhibitors with broad spectrum activity against *Plasmodium* orthologs but selective against host cysteine proteases.

1.9 Research problem statement and justification

Currently, there is no effective vaccine against malaria and the most effective protection method close to a vaccine is not practical for widespread use (Schuldt & Amalfitano, 2012). Conventional malaria control methods have been hampered by resistance to mosquito insecticides and antimalarials, related high cost and long term toxicity, and poor pharmacological profiles (Aguiar *et al*. 2012; Liñares & Rodriguez, 2007; Wongsrichanalai & Meshnick, 2008; Wongsrichanalai *et al*. 2002).

Common strategies in research for development of new antimalarial drugs include innovative exploitation of old drugs, optimization of existing drugs and, validation of molecular targets based on existing knowledge of parasite physiology and biology (Liñares & Rodriguez, 2007; Rosenthal, 2003). The first two are constantly deterred by resistance problems (Chavain *et al*. 2009). *Plasmodium* molecular drug targets, on the other hand, have received much attention especially FP-2 and FP-3 which have emerged as potential targets in recent years (Rosenthal, 2011). However, no drug has been developed yet hence the need for more studies on potential inhibitors.

Structure guided molecular drug design relies on structure activity relationships information obtained from protein (molecular target) crystal structures (Batra *et al*. 2003; Pandey & Dixit, 2012). Unfortunately, there are only a few FP-inhibitor complexes present probably due to associated difficulties in crystallization (Coterón *et al*. 2010).

Comparative protein modeling could provide a fast and efficient method for obtaining protein structural information (Hillisch *et al*. 2004). Molecular docking has been applied in high through put screening of compounds producing viable pharmaceutical leads (Alvarez, 2004; Desai *et al*. 2004), hence, its proposed application in this study.

FP-2 and FP-3 has been the center of focus for drug development which is logical considering *P. falciparum* causes majority of malaria infections. However, *P. knowlesi* and *P. vivax* are equally important as far as malaria incidence is concerned (Barnwell *et al*. 2007; Mendis *et al*. 2001; Jongwutiwes *et al*. 2004).

FP homologs have been identified in *P. chabaudi*, *P. berghei, P. yoelii*, (Caldeira *et al*. 2009; Chan *et al*. 2005) but limited information about the their 3D structures exist.

The importance of the rodent *Plasmodium* cannot be overlooked as they are used as models for *in vivo* antimalarial drug tests (Carlton, 2002; Jambou *et al*. 2011; Stephens *et al*. 2012), hence the need for more structure activity relationship studies to aid interpretation of *in vivo* results which has been a problem (Rosenthal *et al*. 2002). None of these *Plasmodium* FP homologs have been crystallized and *in silico* studies (both homology modeling and docking) are limited (Ettari *et al*. 2009), thus more information is needed.

Most FP inhibitors are peptide based, which have been shown to be susceptible to host proteases, have poor pharmacological profiles and low absorption rates through cell membranes. Consequently, only a few have proceeded to clinical trials (Ettari Roberta *et al*. 2009). Peptidomimetic and non-peptide inhibitors have been shown to increase backbone stability and reduced susceptibility to other proteases (Ehmke *et al*. 2011; Ettari *et al*. 2011; Pérez *et al*. 2012; Verissimo *et al*. 2008) hence, there is need for increased screening of non-peptide inhibitors which are less susceptible to host proteases hence capable of more *in vivo* activity (Desai *et al*. 2006; Teixeira etal. 2011). Natural products are rich in non-peptide compounds thus could a source of potential inhibitors for FPs and related *Plasmodium* homologs (Rosenthal, 2003).

Natural products have been a source of drugs and drug leads for a long time. Natural compounds have unique structure and stereochemistry that allow them to selectively interact with biological target molecules (Ginsburg & Deharo, 2011). An earlier study revealed 62% of small molecule drugs in use from 1981 to 2006 were natural products, derived from natural products or were developed based on natural products (Newman & Cragg, 2007). Natural products have also been shown to have a higher hit rate for potential drugs compared to high throughput screening of synthetic compounds (Weissman & Leadlay, 2005). The South African biodiversity is rich with natural resources which could be a source of potential FP inhibitors. As at present there is no documentation of natural products of South African origin against FPs, hence the need for this study.

1.10 Aims and objectives

The main aim of this study was to perform comparative sequence and protein structure analysis of FP-2 and FP-3 *Plasmodium* homologs *in silico*, using bioinformatics tools and molecular docking experiments to screen non-peptide compounds obtained from South African natural sources against FP-2, FP-3, and related *Plasmodium* homologs.

1.10.1 Specific Objectives

i. Retrieve FP-2 and FP-3 *Plasmodium* and human homolog sequences from NCBI and PlasmoDB biological databases.

ii. Perform comparative multiple sequence analysis of FP-2 and FP-3 homologs to deduce residue variations in the sub site regions whose effect on substrate binding will be analysed through docking studies.

iii. Build homology models for FP-2 and FP-3 orthologs of *P. vivax, P. ovale*, *P. malariae, P. knowlesi, P. berghei, P. yoelii yoelii* and *P. chaubadi*

iv. Perform docking studies between the selected compounds using both FP-2 and FP-3 and the created protein model structures of FP-2 and FP-3 to identify significant inhibitor-enzyme binding, inhibition and interaction characteristics.

v. To deduce important differences of the inhibitor protein interactions between FP-2 and FP-3 homologs at the binding sub sites associated with sequence variations.

vi. Build a small compound library combining some known FP inhibitors and a number of South African natural compounds that might have significant FP binding and inhibition characteristics.

CHAPTER TWO

2. SEQUENCE AND PROTEIN STRUCTURE ANALYSIS

This chapter was aimed at identifying FP-2 and FP-3*Plasmodium* homologs from various databases using various bioinformatics search tools and methods. *Plasmodium* spp. especially *P. falciparum* and *P. vivax* cause most malaria incidences all over the world (WHO World Malaria Report, 2011). *P. knowlesi* is a primate *Plasmodium* but has been documented to cause malaria in humans (Jongwutiwes *et al*. 2004). The treatment of malaria is generalized although species variations are known to cause treatment difficulties with *P. falciparum* being resistant to most antimalarials but *P. vivax* has few cases reported (Wongsrichanalai *et al*. 2002). *P. berghei*, *P. yoelii* and, *P. chaubadii* are rodent infecting *Plasmodium* used as animal models in malaria drug development experiments (Rosenthal *et al*. 2002). As previously described in Chapter 1, section 1.3, cysteine proteases are ubiquitous in nature, also being found in humans as well. It is, therefore, important to evaluate the similarities and differences among these homologs to deduce variations that could be exploited for inhibitor selectivity. A clear understanding of the sequence variations among these homologs could have important implications to the current understanding of the structure and function of these proteins.

A few 3D structures of FP-2, FP-3 are available. However, structures of FP-2 and FP-3 homologs from other *Plasmodium* species have not been solved. The availability of structural bioinformatics methods such as homology modeling can be used to calculate 3D protein structures (Hillisch *et al*. 2004; Jacobson & Sali, 2004). Through comparative structural analysis, it is possible to identify important residues of these structures that influence protein-ligand interactions at the active site. The *Plasmodium* FP-2 and FP-3 homologs may share relatively high sequence similarity(Chan *et al*. 2005; Desai & Avery, 2004; Na *et al*. 2004; Rosenthal *et al*. 2002), but differences in sub site residues may have an effect on the overall binding characteristics which could complicate development of a broad spectrum inhibitor and interpretation of *in vivo* results (Chapter 1, section 1.9). However, variations observed between the *Plasmodium* homologs and the human homologs could be exploited to enhance inhibitor selectivity.

2.1 Introduction

Currently biological data such as proteins, DNA and RNA sequences are stored in databases. To retrieve these sequences, sequence alignment analysis is applied in various existing biological databases by use of sophisticated search tools such as BLAST (Altschul *et al*. 1990). Sequence alignment provides a platform for identifying and characterizing protein structure and function infer phylogeny, deduce sequence similarities and differences as well as other manipulations (Edgar *et al*. 2006). This data, combined with structural information obtained from protein 3D structure could yield valuable information regarding the target sequence.

It has been shown that protein structure and function are more conserved than the sequence. As such, it is possible to have different proteins sharing low sequence similarity and identity but performing similar functions due to similar protein folds e.g. *Plasmodium* and human cysteine proteases (Sajid & Mckerrow, 2002; Lecaille *et al*. 2002). Protein structures are best determined by experimental procedures such as X-ray crystallography or nuclear magnetic resonance (NMR). X-ray crystallography is however expensive and requires a lot of time because crystallizing proteins can be slow and at times difficult to achieve as observed in the case of large proteins (Krieger *et al*. 2003). NMR is good for solving 3D structures of biomolecules in solution considering many proteins perform their functions in body or cellular fluids (Tamm & Liang, 2006). When these methods are not feasible, computational methods such as homology modeling can be used to obtain the structure of a protein from its amino acid sequence as the structure is uniquely determined by the amino acid sequence (Krieger *et al*. 2003; Hillisch *et al*. 2004; Venclovas, 2012).

This chapter was aimed at gathering information from FP-2, FP-3 and related *Plasmodium* and human homologs using sequence alignment as applied in sequence retrieval, multiple sequence alignment analysis and homology modeling. This was done for comparative functional and structural analyses. Sequence alignment and specific programs used for the above mentioned purposes are outlined below.

2.1.1 Sequence alignment and alignment algorithms

Sequence alignment is a way of comparing nucleic acid (DNA/RNA) and amino acid sequences by revealing regions of similarities that have structural, functional or evolutionary significance which is achieved by a character-wise comparison of a pair or multiple sequences. There are two major sequence alignment algorithms i.e. local and global alignment. Local alignments find similarities between sequences based on relatively conserved regions in the sequence (Smith and Waterman, 1981), while global alignments compare sequences entirely regardless of long stretches of low sequence similarity (Needleman *et al*. 1970). Local and global alignments use scoring matrices to select optimally aligned sequences by assigning scores to matches between the aligned sequences as well as account for mismatches and inserted gaps. Scores also account for the amino acid substitution frequencies and the frequency of each amino acid in the homologs protein sequences (depending on the substitution matrix used) that have occurred over time during evolution (Chuong and Katoh, 2008). There are two popular scoring matrices; BLOSUM (blocks amino acid substitution matrix) and PAM (point accepted mutations) used in multiple sequence alignments.

BLOSUM scoring matrices were developed to account for sequences which are more divergent and it is based on a direct observation of all possible amino acid substitutions in a multiple sequence alignment (Henikoff & Henikoff, 1992). BLOSUM uses the actual sequence identity percentage from the sequences to construct scoring matrices hence BLOSUM62 implies that sequences used for the matrix construction have an average sequence identity of 62% (Henikoff & Henikoff, 1992). Sequences with lower percentage identities would require a lower BLOSUM matrix and the reverse for sequences with a higher percentage of sequence identities.

PAM on the other hand, was developed based on closely related protein sequences. It was based on the assumption that homologous sequences perform similar function hence residue variations could alter function of the sequence was termed significant (Dayhoff *et al*. 1978). PAM matrices are based on evolutionary divergence between closely related sequences. PAM1 substitution matrix was derived from calculating the probability of one substitution per 100 residues. To derive higher PAM matrices, PAM1 is multiplied by itself. For instance, PAM80 is a product of PAM1 multiplied by itself eighty times which means there are a number of transformations in

between before the current amino acid state (Dayhoff *et al*. 1978). PAM80 corresponds to 50% sequence similarity.

This means highly divergent protein sequences will require a higher PAM to reflect the numerous substitutions experienced over the evolutionary time. It is therefore, important to choose a scoring method that is relevant to the target sequences when performing sequence alignments and database similarity search to obtain optimal results.

2.1.2 Database similarity search and sequence retrieval

Database similarity search is a means of searching for sequences that may share sequence similarity/identity. It is usually a primary step towards assigning function to a protein sequence since; sufficient homology implies a shared evolutionary ancestor and probably, structure and biological function. It is based on pair wise sequence alignment where the sequence in question (query) is aligned through sequences in the database (Altschul *et al*. 1990). Database similarity search is mostly performed using Basic Local Alignment Tool (BLAST). BLAST is a heuristic word based method for database similarity search which is less accurate than dynamic programming but is faster and provides hits with acceptable statistical scores e.g. expect values (E-value) which measures the probability of sequences aligning by chance, sequence similarities and sequence coverage. Many databases incorporate the BLAST tool for database similarity searches for instance BLAST (http://blast.ncbi.nlm.nih.gov) is used to access the National Centre for Biotechnology Research (NCBI) as well as other databases (Sijwali & Rosenthal, 2004). Position specific iterated-basic local alignment tool (PSI-BLAST), a BLAST variant is another common database similarity search tool modified to use a position specific score matrix (profile) instead of the query sequence and the associated substitution matrix that is able to detect distant homologs (Altschul *et al*. 1997; Forres *et al*. 2006). It is also possible to search for sequences based on their structural similarity especially when searching for templates for homology modeling using methods that incorporate secondary structure prediction such as HHpred (http://protevo.eb.tuebingen.mpg.de/hhpred) which utilizes PSI-BLAST previously described and HHblits which uses profile Hidden Markov Models (HMMs) which is able to detect remote homologs (Söding *et al*. 2005). In comparative studies such as this, it is important to accurately identify homologs to avoid incorrect inferences. Due to this, several important factors are

considered when retrieving sequence homologs that include; percentage identity, E-values and sequence coverage sequence (Krieger *et al*. 2003).

2.1.3 Multiple sequence alignment

Multiple sequence alignment (MSA) is a method that compares a set of sequences to identify similarities and identities within the sequences to infer homology among protein/DNA sequences by assigning percentage scores and can be done by comparing the whole sequence (global alignment) or certain parts of the sequences (local alignment) as mentioned in Chapter 2, section 2.1.1. It is a common practice in comparative and functional bioinformatics to align homologous sequences to locate functional and structural domains or important motifs that may be present in the aligned sequences. It is also a prerequisite to most phylogenetic analysis programs (Edgar & Batzoglou, 2006; Golubchik *et al*. 2007). There are a number of multiple sequence analysis methods which all use different algorithms. In this study MAFFT and PROMALS3D were used.

MAFFT is both web based (http://www.ebi.ac.uk/Tools/msa/mafft) and standalone multiple sequence alignment program based on two main techniques. These are: Fast Fourier transform (FFT) that rapidly identifies homologous regions using sequence volume and residue polarity (Katoh *et al*. 2002). The second technique is based on a simple scoring system (FFT-NS-2 and FFT-NS-i) that have reduced computational time, increased accuracy even in long gapped sequences as well as distantly related sequences. MAFFT implements progressive (FFT-NS-2) as well as iterative (FFT-NS-2) heuristic methods for refining the sequence alignments (Katoh *et al*. 2002). MAFFT has been shown to be much more effective than other alignment methods with comparable accuracy.

PROMALS3D or profile multiple alignment with local structure and 3D constrains (Promals3D) on the other hand, is a web based (http://prodata.swmed.edu/promals3d/promals3d.php) multiple sequence alignment program (Pei *et al*. 2008). Similar sequences are aligned and scored using a weighted function based on sum of pairs of BLOSUM62. A representative of each cluster is then used to search for additional homologs using PSI-BLAST and PSIPRED (Altschul *et al*. 1997; Mcguffin *et al*. 2000).

A Hidden Markov Model (HMM) of the profile-profile and predicted secondary structures is then applied to representative pairs to obtain sequence based restrains from calculated posterior probabilities (Pei *et al*. 2008). The final multiple sequence alignment is a result of the progressive alignment of the pre aligned pairs which are eventually merged. The result is based on a consistency scoring function derived from the posterior probabilities. PROMALS3D adds structural constraints derived from sequences with known structures to the sequence based constraints. The result is an alignment that contains both sequence and structural information of the input protein sequences. An advantage of this program is the ability to provide an alignment which is consistent both at the sequence and structural level (Pei *et al*. 2008).

2.1.4 Phylogenetic analysis

Multiple sequence analysis provides a means for identifying sequence identities, similarities and show conservation among aligned protein sequences (Chapter 2, section 2.1.4). However, in order to infer evolutionary relationships and deduce protein groupings such as orthology, paralogy as well as draw other conclusions e.g. evolutionary distances from sequence data, molecular phylogenetics is used (Kosiol, Bofkin, & Whelan, 2006). Phylogenetic trees illustrate evolutionary relationship among a group of organisms, a family of nucleic acids or protein sequences. The outcome of the phylogenetic tree is dependent on the tree building algorithm and evolutionary assumptions considered (McCormack *et al*. 2009). Results from phylogenetic analyses could for instance be used to assess the viability of targeting proteins that are shared among species, performing similar functions, but having underlying evolutionary dissimilarities.

2.1.5 Homology Modeling

The 3D structure of a protein can reveal biologically important properties and features of a protein. Homology modeling or comparative modeling is a method that predicts protein structures based on sequence homology with known structures (Hillisch *et al*. 2004; Sali, 2011; Venclovas, 2012).

If two sequences share a high sequence similarity over a comparable length, then they are likely to fold into similar protein structures (Krieger *et al*. 2003; Hillisch *et al*. 2004; Jacobson & Šali, 2004). Using the sequence with known structure (template), the 3D structure of the sequence in question (target) can be predicted based on the alignment between the target and template (s) sequence (s) (Eswar *et al*. 2007). It is a fast method to provide information on protein function, interactions and antigenic properties. When combined with other methods such as molecular docking and molecular dynamics, it can be used for rational structure based inhibitor or drug design (Alvarez, 2004; Desai *et al*. 2006; Gschwend, Good, & Kuntz, 1996; Hillisch *et al*. 2004; Kitchen *et al*. 2004). The Homology modeling procedure is summarized below.

2.1.5.1 Template (s) selection

This is the first and perhaps the most important step in protein structure modeling because it defines the predicted structure of the target sequence and the rest of the modeling process. The sequence similarity between the target sequence and possible template should be high enough i.e. in the safe modeling zone of 40% (Rost, 1999; Venclovas, 2012) and above. However, protein modeling can be achieved with sequence similarities below 40% (Krieger *et al*. 2003; Venclovas, 2012). The methods used to identify template hits include those applied in sequence retrieval (Chapter 2, section 2.1.2 and 2.1.3). The importance of selecting the best template cannot be overlooked therefore accurate selection of the modeling template is very important as the template sequence backbone atom coordinates are simply copied to the aligned target sequence backbone atoms during modeling (Baker & Šali, 2001; Šali, 2011). The quality of protein structure prediction using homology modeling can be improved by using multiple templates and an optimal sequence alignment between the target and templates (Fernandez-Fuentes *et al*. 2007).

2.1.5.2 Template - target sequence alignment

After identifying the template (s) sequence (s), MSA is used to align the sequences to obtain an optimal alignment (global or local) for accurate homology models to be calculated (Cavasotto & Phatak, 2009; Edgar & Batzoglou, 2006). In homology modeling it is done for a number of reasons which include increasing confidence on the quality of multiple sequence alignment obtained during template selection as well as to achieve an alignment of the template and target sequence if present in the alignment results (Venclovas, 2012). If many template sequence hits are identified, MSA can be used to identify which template best aligns with the target sequence (Cheng, 2008). In special cases, where template and target sequences share low homology, manual adjustments may be required to achieve the optimal alignment between the template and target sequences (Tastan Bishop *et al*. 2008). By including other homologous sequences in the alignment, regions of high and low conservation are highlighted thus placing gaps and inserts correctly in the sequence alignment (Krieger *et al*. 2003). With an optimal MSA obtained, modeling can commence.

2.1.5.3 Model building

Model building is the actual calculation of the protein 3D model from the selected template (s) involving backbone generation, side chain and loop generation. In most cases, where the target and template share high sequence identity, the aligned template backbone coordinates are copied to the target structure which also applies to the side chains (Šali, 2011). The loop region however is different due to the high rate of variation at this site which requires optimization or post modeling refinements (Rieger *et al*. 2003; Tastan Bishop *et al*, 2008). There are a variety of programs that are used for homology modeling. This can be either web based e.g. HHpred (Söding *et al*. 2005) and SWISS-MODEL (Arnold *et al*. 2006) or script based and run as a standalone computer program, e.g. MODELLER (Šali, 2011). As for the web based programs, the user only needs to input the target protein and select the templates from the alignment hits and the server outputs the protein structure and a couple of model validation results.

MODELLER on the other hand is a standalone computer program used to build 3D protein structures of proteins and their assemblies that satisfy certain spatial restraints of the amino acid sequence (Šali, 2011). The 3D protein structures are calculated by optimization of a probability density function with a variable target function in Cartesian space using molecular dynamics, simulated annealing and conjugate gradient. It uses user input alignment of sequences to be modeled with known homologous structures to calculate a model with all non-hydrogen atoms. Restraints used are derived from the homologous structures and their alignment with the target sequence (Šali *et al*. 2007).

The web based programs have their limitations in that user manipulations are limited unlike the standalone MODELLER program in which the user has total control of the whole modeling procedure from template selection, sequence alignment, model building including loop optimizations and refinement and validation. Other methods include; MODBASE, MODWEB, MOULDER (Eswar, 2003).

2.1.5.4 Model refinement

Accurate prediction of protein structures that can be used to infer biological functions or be applied for structure based drug design has been a challenge in structural bioinformatics since there are few effective sampling methods and energy functions to search the entire protein conformational space accurately (Raval *et al*. 2012; Xiang, 2006). Regions with poor structural alignments for instance, loop regions where insertions and gaps occur, are modeled to the best conformation possible, maintaining both geometric and energy constraints to acceptable limits without distorting the protein stereo chemical architecture of the protein (Xiang, 2006). Model refinement is done both in the global and local environment of the protein. Local refinement takes care of the loop regions, side chains and backbone atoms that make up the helix and sheets is done simultaneously using methods similar to those applied for loops and side chain building process (Li& Friesner, 2004). Global refinement makes sure that the overall structure of the protein has minimal irregularities. Both global and local energy refinements incorporate energy scores which describe the stability of a protein. Homology modeling is based on statistical constraints thus results need to be evaluated to ensure their quality.

2.1.5.5 Model evaluation and validation

Computational methods are prone to errors which mean generated protein structures may exhibit inaccuracies and deviations from the native structure therefore necessitating structure validation before being used for any further analysis (Pawlowski *et al*. 2008; Ginalski, 2006). The physiochemical rules include stereo chemical correctness of the structures which are implicitly assessed by comparison to profiles obtained from known experimentally determined structures of high quality (Laskowski *et al*. 1993). Deviations from these profiles imply problematic regions in the structure which require further refinement. Most of these problematic regions are found in the flexible loop regions which are modeled incorrectly, from misaligned portions in the sequence alignment, poor template selection, misplaced side chains and alignment errors (Baker & Sali, 2001) and are mostly carried forward from the template hence the need for careful template selection (Krieger *et al*. 2003; Tastan Bishop *et al*. 2008; Venclovas, 2012).

These regions can be improved by performing sequence realignment and modeling or loop optimizations or refinement if only the loops are the only problematic regions (Venclovas, 2012). There are several programs used for Model evaluation and validation for instance; PROCHECK, ANOLEA, QMEAN and MetaMQAPII scores and DOPE Z scores from MODELLER which all use different approaches in their validation (Chapter 2, section 2.2.4). It is always advisable to use different programs because there is none gives accurate results. This is partly because there are a few validation programs that can discriminate between globally incorrect structures and approximately correct ones (Pawlowski *et al*. 2008).

2.2 Methodology

Five FP-2 and two FP-3 homologs were retrieved from PlasmoDB using FP-2 and FP-3 as query sequences respectively. Four FP homologs of human cysteine proteases; cathepsin K, H, L and, S were obtained from NCBI using FP-2 as the query sequence using methods detailed in section 2.2.1 below.

Sequence retrieval, multiple sequence alignment, homology modeling for the identified FP-2 and FP-3 *Plasmodium* homologs, model quality assessed and structure analyses were performed using respective methods. Figure 2.1 is an overview of the methodology.

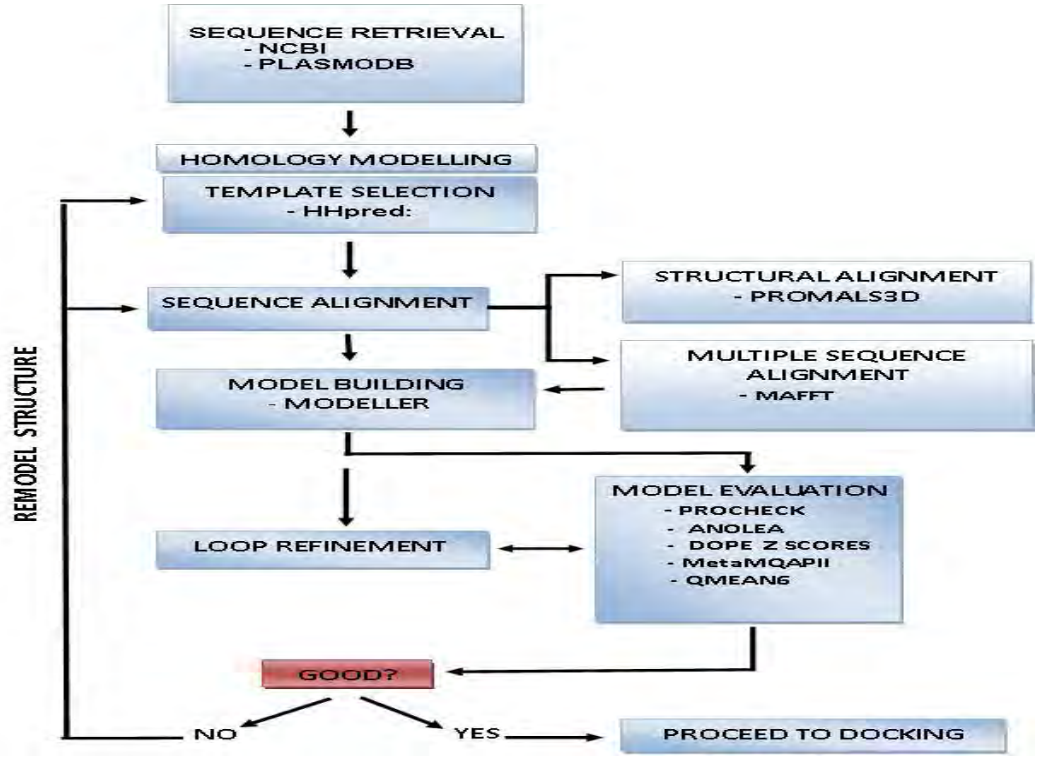


Figure 2.1: Summary the methodology used for sequence analysis and homology modeling. Sequence analysis was done first followed by homology modeling but both were used concurrently for comparative analysis.

Of importance to note was that the sequence identities depicted are not reflective of FP-2 and FP-2 whole sequence but the catalytic domain only. The prodomain is not involved in substrate/inhibitor binding, thus was ignored. For the purpose of this study, sequences were renumbered. The first number is the amino acid residue position in the mature domain, followed in brackets by the actual residue position in the whole sequence (prodomain and mature domain).

2.2.1 Database similarity search and sequence retrieval

FP-2 and FP-3 sequence orthologs were retrieved from PlasmoDB (http://plasmodb.org) using reverse BLAST with PF3D7_1115700 (FP-2) and PF3D7_1115400 (FP-3) as query sequences respectively in the following *Plasmodium* taxids; *P. berghei* strain ANKA, *P. chaubadi* strain *chaubadi chaubadi, P. knowlesi* strain H, *P. vivax* strain *Sal-1* and *P. yoelii* strain 17XNL. Sequences with percentage sequence identities greater than 40% were selected (Rost, 1999). Statistical measures mainly E-values were critical to sequence selection with E-values lower than 1.0e-5 considered as significant. Sequence coverage was also considered in the sequence selection criterion.

Human homologs within the papain and cathepsin L like subclass were obtained from NCBI as well as other human cysteine protease including cathepsin K, H, and S. One of the main aims of the study was to analyse inhibitor and receptor interactions thus comparative analysis of the human cysteine protease and FP-2/FP-3 homologs was done to highlight differences that could be exploited to enhance inhibitor selectivity. Their respective crystal structures of cathepsin S, K and L were retrieved from RCSB PDB (http://www.rcsb.org/pdb/home).

BLAST parameters for the NCBI search were set to have an expect value of 10, gap penalty of 11 and an extension penalty of 1, BLOSUM62 scoring matrix and a word size of 3. For the PlasmoDB BLAST, the expect value was set at 10 and BLOSUM62 which are the default. For PlasmoDB BLAST search, default settings were used. These were expect value of 10, Maximum descriptions/alignments (V=B) of 50 and the lower complexity turned off.

2.2.2 Multiple sequence alignment

Multiple sequence alignment was done using all the retrieved FP-2 and FP-3 homologs for comparative analysis. The web based MAFFT and PROMALS3D programs were used. As for MAFFT program, the following sequence alignment parameters were used with BLOSUM62 set as the substitution matrix, gap penalty and gap extension penalty of 1.53 and 0.123 respectively. The number of tree building steps was adjusted to 2 with maximum iteration also set at 2. The fast Fourier transform (FFTS) was applied for the local pairs and the alignment output saved in the ClustalW format. PROMALS3D default parameters were used. The PSI-BLAST expect value parameter was adjusted to 0.0001 for both cut-off derived from sequence profiles from PSI-BLAST search results and those against homologs with 3D structures.

2.2.3 Phylogenetic analysis

Phylogenetic analysis was performed using MEGA (Molecular Evolutionary Genetic Analysis) version 5.05 (Tamura *et al* . 2011). The maximum likelihood method of phylogenetic inference was selected. Maximum likelihood is used to calculate the probability of observing the data in this case the protein multiple sequence alignment under a certain phylogenetic tree and the chosen model of substitution. It finds the optimal set of parameters in the tree and the best model of substitution describing the given data set. The model of substitution is a set of parameters that clearly describes the evolutionary process, e.g. amino acid substitutions (Kosiol *et al*. 2006). For this study the substitution type used was amino acid, WAG+G substitution model as selected by the BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion).

Gamma (G) evolutionary distance correction was set as specified in the best selected evolutionary model (WAG). Other parameters included site coverage cut off which was set at 95%; the tree inference method chosen was the Nearest-Neighbor Interchange (NNI) and a bootstrap value of 1000. The parameters not mentioned were left as is in the default setting. The bootstrap consensus tree was chosen with the cut-off for significant nodes set at 70%.

2.2.4 Homology modeling

Homology modeling was done for each protein homolog retrieved to build 3D protein structures for structural analysis and docking purposes. The procedure used in building these models is explained in detail in the sub-sections below.

2.2.4.1 Template selection and multiple sequence alignment

The HHpred server (http://protevo.eb.tuebingen.mpg.de/hhpred) was used for the template search. The template search was done for FP-2, FP-3 and related homologs retrieved (Table 2.1). Two templates with the highest sequence identity scores, best sequence coverage, resolution and E- values chosen from the list of template hits produced by HHpred search.

The MSA obtained by MAFFT and PROMALS3D were compared to the structural alignment obtained from HHpred. These two were used to perform any necessary adjustments to the sequence alignment if there were any. Templates selected were validated using MetaMQAPII ANOLEA, QMEAN and PROCHECK to ascertain their quality and suitability for use (Benkert *et al*. 2011; Melo *et al*. 1997; Pawlowski *et al*. 2008; Laskowski *et al*. 1993). Manual adjustments were made where applicable.

2.2.4.2 Model building

Modeling was done using MODELLER version 9.10 (Šali, 2011). MODELLER scripts were obtained from the MODELLER manual and customised to include user specifications such as the ability to build models using multiple templates (*Supplementary data/Chapter2/MODELING_SCRIPTS/h_modeller.py*). The model assessment method chosen was DOPE (Discrete Optimized Protein Energy).

The script was also modified to produce models in the superimposed position with the template(s) used. Model refinement was done slowly using molecular dynamics. For each protein, 200 models were produced and based on the DOPE score (John & Šali, 2003). DOPE Z scores were calculated using a separate MODELLER based script and the top ten models with the lowest energy scores selected. 3D model structures were visualized and manipulated using PyMOL v 3.5 and Accelrys Discovery Studio v 3.1.

2.2.4.3 Model evaluation and validation

The model evaluation programs used employ different criteria to perform model quality assessment. Methods used were global and local model evaluation programs and programs to check the stereo correctness of the calculated models. Below is a detailed description of each evaluation program and how it was applied for model validation.

2.2.4.3.1 Global and local model evaluation programs

i. DOPE scores and DOPE Z scores

DOPE was used to assess the model in the refinement process during homology (Chapter 2, section 2.2.4, sub-section 2.2.4.2). DOPE Z scores were calculated using a separate customised python based MODELLER script that could select the top ten best based on DOPE Z scores (*Supplementary data/Chapter2/MODELING_SCRIPTS/modeller_z_score.py*) (Table 2.5). DOPE score are derived from a statistical potential homology model assessment method is based on an atomic dependent statistical potential from a sample of native structures that does not depend on any adjustable parameters DOPE (John & Šali, 2003). Statistical potentials are widely used methods for protein structure analysis, in modeling and quality assessment as well. The method is implemented in MODELLER where it is used during modeling iterations.

It is based on improved reference state corresponding to non-interacting atoms in a homogenous sphere. The radius of the sphere is derived from the input native structure hence considering the finite and spherical shape of the native structures. In nutshell it looks at the overall quality of the model as a whole (Shen & Šali, 2006). DOPE Z score on the other hand looks at the local model quality by looking at the atomic distance of the heavy atoms in the model. It is used to facilitate the comparison of different protein sequences. A profile is usually generated with information pertaining problematic regions in the model (Pieper *et al*. 2011). Top models were selected and considered for further quality evaluation.

ii. MetaMQAPII

Meta Model Quality Assessment Program (MetaMQAPII), is a web based (https://genesilico.pl/toolkit/unimod?method=MetaMQAPII) model quality assessment method. It uses a combination of eight programs; VERIFY3D, PROSA2003, PROVE, ANOLEA, BALA-SNAPP, TUNE, REFINER, and PROQRES. The output is composite energy score incorporated in to the models B-factor column of the coordinate file of models (Pawlowski *et al*. 2008).

The energy distribution in the 3D model can be visualized using a protein structure visualizer such as PyMOL. The energy profiles were colored from blue (stable) to red (unstable). Generally values greater than zero are high energy (unstable regions) while scores below zero are low energy regions (stable regions)

Outputs are the modified coordinate file, and a log file with evaluation results from the eight individual programs as well as RMSD and GDT-TS scores based on similar proteins in the database (Pawlowski *et al*. 2008). The active site of the proteins was of main interest, local model inaccuracies flagged in these areas by MetaMQAPII were considered for refinement. Models used were those selected in sub-section 2.2.4.3.1.

iii. ANOLEA (Atomic Non Local Environment Assessment)

ANOLEA is a web based server for homology model quality evaluation. The ANOLEA program used in this study was anchored in the SWISS-MODEL work space for model structure assessment (http://swissmodel.expasy.org/workspace).

ANOLEA computes the energy involving non-local interactions between all heavy atoms of the 20 standard amino acids giving an energy distribution profile of the protein sequence. High energy regions in the profile correspond to erroneous possible interacting regions of the protein (Melo & Feytmans, 1998). Regions flagged as erroneous (high energy regions) were compared with the other validation programs and the protein sequence. This was done to ensure none of these erroneous regions were in the active site. If erroneous regions in the active site were found further model refinement was done. Top models on the basis of DOPE Z scores were analysed.

iv. QMEAN6

QMEAN6 performs structural quality assessment of the protein structures based on six statistical potentials derived from a set of non-redundant high resolution crystal structures from PDB and outputs energy profiles of the protein structure, Z scores and QMEAN6 scores (Benkert *et al*. 2008). These statistical potentials are solvation potential, torsion angle potential, distance dependent pair wise potential, structural similarity between model and target protein, secondary structure and solvent accessibility agreement. The output is given as individual scores which are made composite using a Z score, and an energy profile is generated. Scores close to zero indicate stable regions while scores towards ten are unstable regions (Benkert *et al*. 2008). QMEAN6 is therefore able to perform global and stereochemical model evaluation.

2.2.4.3.2 Programs for evaluating model stereochemistry

i. PROCHECK

Protein stereochemistry is important for the stability and proper functioning of proteins. PROCHECK looks at the stereo correctness of the proteins compared to related protein structures of the same resolution and highlights regions that may require further optimization in terms of quality. PROCHECK is mainly used for X-ray crystal structures (Laskowski *et al*. 1993) but could also be used for homology models just to have an idea about the stereochemistry of the models which should be backed by other evaluation and validation programs.

PROCHECK evaluation was done using the web based SWISS-MODEL workspace for structure assessment (sub-section 2.2.4.3.4). Best models selected in the basis of DOPE Z scores were used as input. The Ramachandran output was considered.

ii. Protein Interaction Calculator (PIC)

PIC (Tina *et al*. 2007), is not a model quality evaluation program. In this study, it was used to assess the correctness of disulfide bonds found in cysteine proteases. FP-2, FP-3 and related *Plasmodium* homologs have four sets of disulfide bridges compared to three among related papain-like cysteine proteases (Hogg *et al*. 2006). MODELLER has a provision for including

disulfide bonds during modeling (Sali, 2011). If these bonds are present in the template protein, the program automatically detects them and tries to replicate the same for the model. Normal cysteine protease disulfide bridges fall in the range of up to 2.2Å. These bond lengths were then compared to the templates disulfide bridges. Models in which the disulfide bonds were not in these range were re-modeled.

2.2.4.4 Loop refinement

MODELLER has an inbuilt loop refinement step which models the loops to the best possible conformation with the lowest energy and best conformation. However, sometimes further loop refinement may be required. Loop refinement was done for models that had loop regions near or at the active site of the protein structure flagged erroneous by MetaMQAPII, ANOLEA and QMEAN6 model validation programs. A customised MODELLER script for loop refinement was used (*Supplementary_data/Chapter2/MODELING_SCRIPTS/loop_modeller.py*) and 100 models produced. Models were refined slowly using molecular dynamics minimization method and were assessed using DOPE scores. The DOPEZ score was calculated as described in Chapter 2, section 2.2.4.3.1. Model evaluation and validation was repeated with methods described in Chapter 2, section 2.2.4.3.

2.3 Results and Discussion

2.3.1 Sequence retrieval

Five FP-2 and two FP-3 homologs were retrieved from the PlasmoDB database using FP-2 and FP-3 as the query sequence respectively. Four human homologs of FP-2 were retrieved from NCBI namely cathepsin K, H, L and, S (Table 2.1). VP-2 had the highest percentage sequence identity score of 62% against FP-2 while FP-3 homolog of *P. knowlesi* (Str. H) had 60%. The remaining FP-2 and FP-3 orthologs had sequence identities above 45%.

The human homologs on the other hand had sequence identities below 40% with cathepsin K having the highest percentage identity score of 39%. It is important to note that these statistics in Table 2.1 are based on the catalytic domain only and not the whole protein (prodomain and

catalytic domain). From the percentage sequence identities, it can be deduced that the sequences obtained using FP-2 and FP-3 as query sequences are homologous.

Apart from the human homologs the remaining sequences showed high degree of similarities and since they share the same genus, then they can be referred to as orthologous. The human homologs seemed to be divergent though the sequence lengths were varied.

Accession number	**FP homolog**	**Abbrev.**	**Organism**	**% Sequence identity**		**% coverage**	**E - value**
				FP-2	FP-3		
PVX_091415	Vivapain-2	VP-2	*P. vivax*	62	67	99	3.1e-88
PCHAS_091190	Chaubapain-2	CP-2	*P. chabaudi (Str. chabaudi)*	50	48	100	2.9e-69
PKH_091250	Knowlesipain-2	KP-2	*P. knowlesi (Str. H)*	56	57	98	1.1e-83
PBANKA_093240	Berghepain-2	BP-2	*P. berghei (str. ANKA)*	51	47	100	1.1e-71
PY00783	Berghepain-2	BPy-2	*P. yoelii-yoelii*	48	47	100	2.6e-68
PVX_091410	Vivapain-3	VP-3	*P. vivax*	57	57	98	2.4e-81
PKH_091260	Knowlesipain-3	KP-3	*P.knowlesi (Str. H)*	60	60	98	8.8e-84
gi|157830076	Cathepsin-K		*Homo sapiens*	39	41	90	1.0e-41
CAA30428.1	Cathepsin-H		*Homo sapiens*	37	34	90	1.0e-41
gi|313754420	Cathepsin-L		*Homo sapiens*	37	38	90	1.0e-45
gi|130749675	Cathepsin-S		*Homo sapiens*	37	37	90	1.0e-41

Abbrev = abbreviation

Table 2.1: Listed are retrieved FP homologs. *Plasmodium* homologs were obtained from PlasmoDB while the human homologs were obtained from GeneBank. Percentage coverage and E-values are based on FP-2 as query. FP-3 orthologs (VP-3, KP-3), were retrieved using FP-3 as the query sequence in a BLAST search of PlasmoDB. *P. knowlesi* FP-2 and FP-3 direct homologs (PKH_091250, PKH_091260) were renamed Knowlesipain-2 and Knowlesipain-3 respectively in of this study. E-values apply to the query sequence used.

2.3.2 Multiple sequence alignment and structural analysis

To analyse protein sequences, the best multiple sequence alignment was obtained from MAFFT although the PROMALS3D (Appendix 1) was used during homology modeling for manual adjustments of the modeling sequence alignment. Analysis was done with more emphasis directed at the binding/active site of the homologs where the protein and inhibitor interact.

2.3.2.1 Inserts

From the alignment (Figure 2.2 and Table 2.2), it was observed that there was conservation among the homologs but with some residues clearly showing non-conservation. Clearly visible were two inserts in the N-terminal and C-terminal that are absent in the human cysteine proteases. The first insert was a stretch of about 17 residues which in FP-2 has been shown to be important for the proper folding of the mature domain before assuming its active conformation (Pandey *et al*. 2004). The insert conserved varied across all the *Plasmodium* homologs (Figure 2.2).

The second insert is located at the extreme end of the C-terminal and it is comprised of 14 residues that structurally form a β-hairpin. In FP-2 this portion has been associated with hemoglobin binding (Pandey *et al*. 2005; Wang *et al*. 2007). This region in FP-2 and FP-3 is well conserved but variations were observed among the rest of the *Plasmodium* homologs which puts into doubt as to whether this insert performs the same function as in FP-2 but this is yet to be deduced. These two inserts are explicitly visible in the structures of the *Plasmodium* cysteine proteases when compared to the human homologs.

2.3.2.2 Active site residues

Cysteine proteases have sub site pocket comprised of S1, S2, S3 and S1' which correspond to P1, P2, P3 and P1' positions of ligands situated in a cleft between the structurally conserved R and L domains (Figure 2.5) of the papain-like fold (Kerr *et al*. 2009; Sajid & Mckerrow, 2002). Some of the sub site residues were conserved while others showed variations.

At the sequence level, it was not possible to deduce the effects of these variations on the structure and function of the proteins. This can only be shown via protein/ligand interactions

which are covered in Chapter 3. The highly conserved residues could however be proof that they probably have important functions structurally and functionally. The sub site residues were selected based on a 6Å radius from the catalytic Cys of all *Plasmodium*FP homologs and compared to those of the FP-2 and FP-3 and it was evident that the sub site residues were varied across the homologs (Table 2.2).

i. Catalytic residues

The catalytic site residues of Cys, His and Asn, that form a catalytic triad in FP-2, FP-3 as is in all papain-like cysteine proteases (Rosenthal, 2004). This typical clan C1A cysteine protease feature was observed at the corresponding positions in all retrieved homologs (Figure 2.1). From the alignment it was clear that the catalytic residues were conserved across all homologs from both *Plasmodium* and human sources.

ii. S1 sub site

In the S1 sub site, Gln36 (279 actual residue numbering) which forms the oxyanion hole in FP-2 was conserved in the corresponding positions in all homologs as is in cysteine proteases. This residue is important in the formation of additional hydrogen bonds that strengthen peptide based inhibitor binding together with the catalytic Cys, His and Asn (Kerr *et al*. 2009; Shenai *et al*. 2003). The nonpolar Gly40 (281) in FP-2 is fairly conserved in all but CP-2, BPy-2and BP-2 which had an Ala instead. The Cys in this particular sub site was conserved across all homologs and its one of the disulfide bridge forming pairs (Hogg *et al*. 2006). Asn81 (324) in FP-2 was replaced with Tyr90 (335) in the corresponding position of FP-3. The residues at corresponding positions in the other homologs were varied but belonged to the polar group of amino acids suggesting a conserved function. From these analyses, the S1 sub site was greatly conserved with a few differences observed in CP-2, BPy-2and BP-2. These variations were considered during docking analysis to find out if these substitutions had significant effects on inhibitor binding.

Little is known about the S1 because it is the least characterized sub site in cysteine proteases (Sabnis *et al*. 2003) sub site although it has a significant role in the initial binding of peptide based small inhibitors such as E-64, leupeptin, and vinyl sulfones (Kerr *et al*. 2009; Shenai *et al*. 2003).

iii. S2 sub site

The S2 sub site has been defined as the major pocket influencing ligand specificity in cysteine proteases characterized by a high abundance of hydrophobic residues (Pandey & Dixit, 2012) However, some polar residues were found in this sub site for instance FP-2, VP-2, FP-3, VP-3 and KP-2 had a Ser residue present at this sub site (Table 2.2). One major notable difference was that the *Plasmodium* homologs had a polar residue at the S2 hollow end of the pocket. The human homologs on the other hand had hydrophobic residues at the same position. It was not clear at this point whether this observation could have a significant impact on protein-ligand interactions. However, if any, it would be observed during molecular modeling (Chapter 3). FP-2 has Leu84, I85 and Asp234 (477) lining the wall at the S2 sub site entrance.Asp234 (477) has been associated with the enlarged S2 pocket compared to FP-3 which has the larger Glu243 (485) instead. The other residues in FP-3 lining the wall at the S2 entrance in FP-3 are Tyr93 and Ile94 with the bulkier Tyrosine associated with narrowing the S2 entrance in combination with Pro181(423) results to a smaller S2 sub site compared to FP-2 (Kerr *et al*. 2009). Lining the S2 entrance in CP-2, BPy-2and, BP-2 were Ile, Leu and Gln at the corresponding the positions. Ile and Leu were interchanged when compared to FP-2 hence may not yield any significant difference however, the inclusion of Glu which is has more or less similar size and flexibility to Glu243(485) in FP-3 may influence the size of the S2 pocket entry point. VP-2 had Phe, Ile and Glu while VP-3 had Asn, Ile and Gln lining the S2 entrance (Table 2.2). VP-2 and Vivapain-3 preferredGlu235 (480), Gln234 (486) respectively at the S2 entrance but VP-2 had Leu84 (327) in FP-2 substituted with a more rigid Phe85(418). KP-2 differed from FP-2 in that Leu172 (415) and D234 (477) were replaced by a Pro173 (424) and Glu234 (486) respectively. KP-3 was the most varied of all the *Plasmodium* homologsas was the human homologs (Figure 2.2).

The S2 sub site is the major determinant of specificity hence the observed variations could have significant implications to substrate specificity mediated by restricted accessibility to the S2 sub site as well as substrate recognition and catalytic efficiency.

Cathepsin L and S seemed to have similar residues at this sub site. From previous analysis, when Cathepsin L is compared to cathepsin B the S2 appears narrow and clear. This is because Asp162, Met161, Asp160 and Ala214 form a wall lining the right hand of the cathepsin L S2 pocket (Fujishima *et al*. 1997).

FP-2, FP-3, VP-2 and 3 have been shown to have preference for peptide inhibitors with a Leu at the P2 position such as E-64, Leupeptin, and peptidyl vinyl sulfones as their active sites are lined with residues able to make nonpolar contacts with their respective inhibitors(Na, Kim, *et al*. 2004; Pandey & Dixit, 2012). It would be important to find out if the other FP-2, FP-3 *Plasmodium* homologs will have similar preferences and if the human homologs differ. In the human FP-2,FP-3 homologs cathepsin L which is the closer to the *Plasmodium* homologs is also characterized by hydrophobic residues in the S2 sub site and prefers bulky and hydrophobic residues at the P2 position of the substrate preferably Phe to Leu. Cathepsin K prefers Ile, Val and Leu at the P2 position but can also accommodate Pro (Lecaille, Brömme, & Lalmanach, 2008).There are amino acid substitutions and conformational variations at this sub site that may be significant for inhibitor selectivity.

iv. S3 sub site

The S3 sub site is a Gly rich region which was conserved among all homologs typical to papain like cysteine proteases (Figure 2.2 and Table 2.2). Another conserved residue was Asn at position 77 (280) in FP-2 and corresponding positions across all homologs. The remaining positions in the S3 sub site were comprised of a mixture of polar and hydrophobic residues. This site is important as it is involved during initial substrate binding (Sabnis *et al*. 2003). In most clan CA cysteine proteases, the sub site has been identified to be crucial for tethering small peptidyl based inhibitors to the main chain of the S3 sub site (Bhaskar R Shenai *et al*. 2003). The Glycine and Tyrosine in this sub site also contributes to the hydrophobic interactions between the ligand and the protease however the Tyr78 (321) in FP-2 was not conserved across all the homologs.

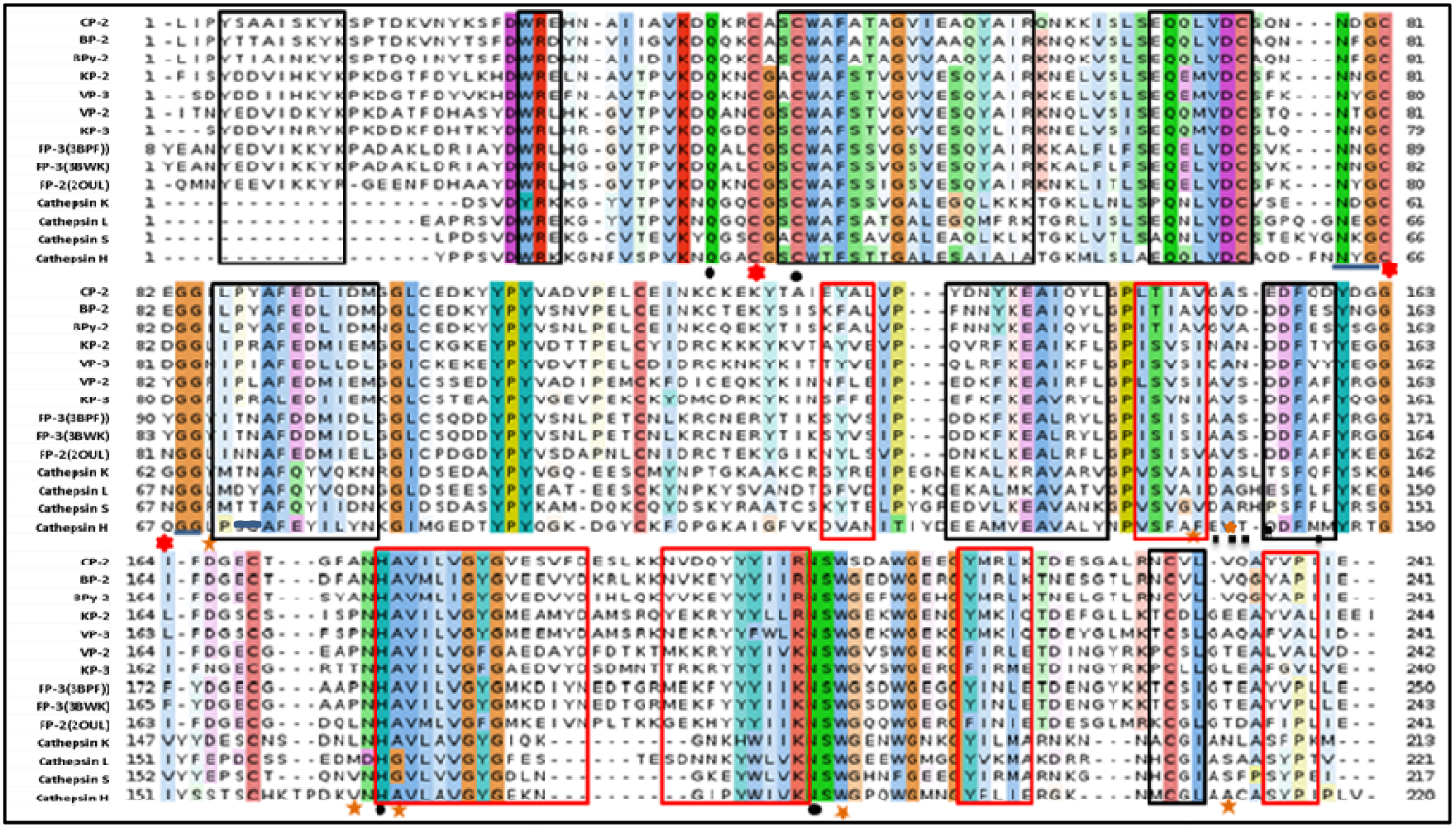


Figure 2.2: Multiple sequence alignment of FP-2, FP-3 homologs as predicted by MAFFT. Residues coloured in dark shades are the most conserved while residues lighter shades are less conserved. Highlighted in black boxes are the helices and in red are the sheets. The non-highlighted residues form the loop regions.

● = Catalytic residues — = S3 sub site residues ★ = S2 sub site residues

★ = S1 sub site residues ■ = S1' sub site residues

Protein ID.	Sub site 1(S1)	Sub site 2 (S2)	Sub site 3 (S3)	Sub site 1' (S1')
FP-2	Q279-G283-C323-N324	L327-I328-S392-L415-D477	K319-N320-Y321-G325-G326	A394-V395-S396-A406-W449
VP-2	Q282-G286-C326-Y327	F330-I331-S395-P418-E480	Q322-N323-T324-G326-G327	A397-V398-S399-A158-W403
CP-2	Q267-A271-C311-E312	I315-L316-A380-A403-Q464	N307-N308-D309-G313-G314	G382-V383-D384-Q388-W437
KP-2	Q288-G292-C332-D333	L336-I337-S401-P424-E486	K328-N329-N330-G334-G335	N403-A404-N405-T409-W458
BP-2 (*P. berghei*)	Q264-A268-C308-E309	I312-L313-A377-A400-Q461	N304-N305-F306-G310-G311	G379-V380-D381-E385-W434
BPy-2 (*P. yoelii*)	Q268-A272-C312-D313	I316-L317-A381-A404-Q465	N308-N309-F310-G314-G315	G383-V384-A385-E389-W438
FP-3	Q287-G291-C331-Y332	Y335-I336-S400-P423-E485	K327-N328-N329-G333-G334	A402-A403-S404-A408-W457
VP-3	Q288-G292-C332-D333	N336-I337-S401-P424-Q486	K328-N329-Y330-G334-G335	C403-A404-N405-V409-W458
KP-3	Q274-G278-C318-D319	F322-I323-N387-T410-E472	Q314-N315-N316-G320-G321	A389-V390-S391-A395-W444
Cathepsin-K	Q19-G23-C61-G62	Y67-M68-A134-L160-L209	E59-N60-D61-G65-G66	D136-A137-S138-Q143-W184
Cathepsin-L	Q20-G24-C66-N67	L70-M71-A136-M162-A215	P60-N63-E64-G68-G69	D138-A139-G140-L145-W190
Cathepsin-S	Q19-G23-C66-N67	F70-M71-G137-V162-F211	E60-N63-K64-G68-G69	D139-V140-R141-F146-W186 W186
Cathepsin-H	Q20-G24-C66-Q67	L70-P71-A137-V164-C212	D60-N63-Y64-G68-G69	E139-V140-T141-M145-W188

Table 2.2: Sub site 1, 2, 3 and 1' sub site residues. *Plasmodium* homologs were renumbered according to the actual protein sequence. Human homologs were numbered according to papain numbering.

2.3.2.3 Special features

FP-2 and FP-3 have a total of eight cysteine residues in the catalytic domain. These cysteine residues form a set of four disulfide bonds. This feature was confirmed in all other *Plasmodium* homologs (*Supplementary_data/Chapter2/Homology_modeling.docx/Table 1*). However, the human homologs had three corresponding sets of these disulfide bonds. This extra set was found between Cys99 (342)-Cys119 (362) of FP-2 and corresponding positions in FP-3 and all other *Plasmodium* homologs. At the corresponding positions in cathepsin K was Asp80-Thr99, Asp 85-Lys104 in cathepsin L, Asp85-Lys104 in cathepsin S and Met85-Gly104 in cathepsin H. The Asp corresponding to Cys99 (342) seemed conserved apart from Cathepsin H which was substituted for Met. The residues corresponding to Cys119 (362) were varied with only Lys shared between cathepsin L and S. This extra disulfide bridge has been suggested to provide support for the long loop spanning the lateral surface of the L domain (Hogg *et al*. 2006).

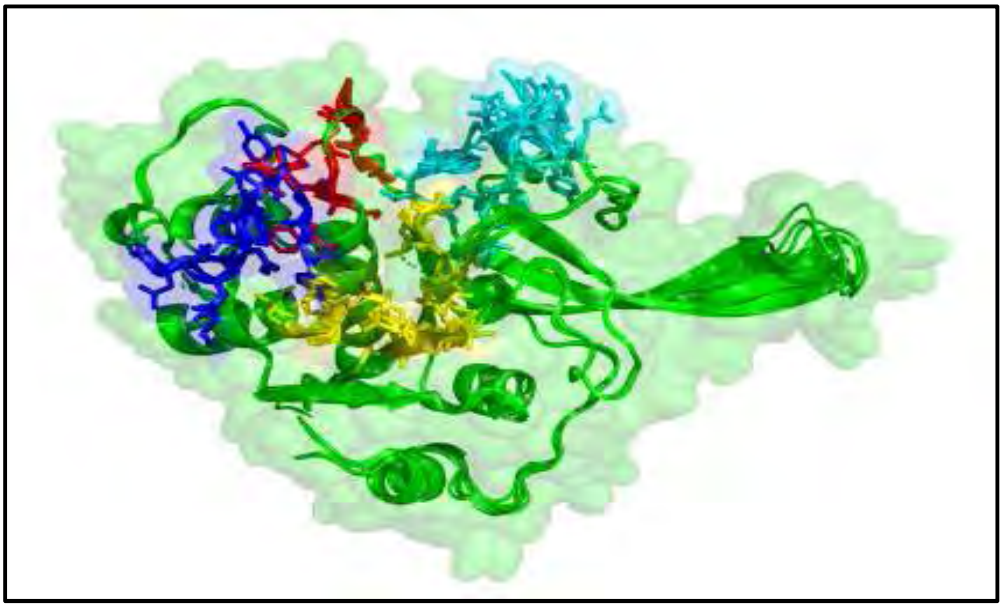


Figure 2.3: Superimposed FP-2 and FP-3 *Plasmodium* homolog 3D models showing active site residue variations. Colored in red are the S1 sub site residues, S2 in yellow, S3 in blue and S1' sub site in cyan.

2.3.3 Phylogenetic analysis

As previously discussed in Chapter 1, section 1.3, FP-2 and FP-3 homologs from both *Plasmodium* and human retrieved for comparative analysis in this study (Table 2.1) all belong to the clan CA, Papain family (family C1). The *Plasmodium* cysteine protease belong to the cathepsin L-like subfamily (Sajid & Mckerrow, 2002). Human lysosomal cathepsins are abundant in nature performing crucial functions such as intracellular protein degradation. They have been implicated in disease as well (Jenko *et al*. 2003). Because of the conserved mechanism of substrate degradation among these proteins (Chapter 1, section 1.3), only substantial differences in the protein sequences can be exploited for inhibitor selectivity. Phylogenetic analysis was used to assess the degree of evolutionary difference between the *Plasmodium* and human FP-2, FP-3 homologs (Figure 2.4).The most notable difference was the explicit clustering of the human and *Plasmodium* homologs into two distinct groups.

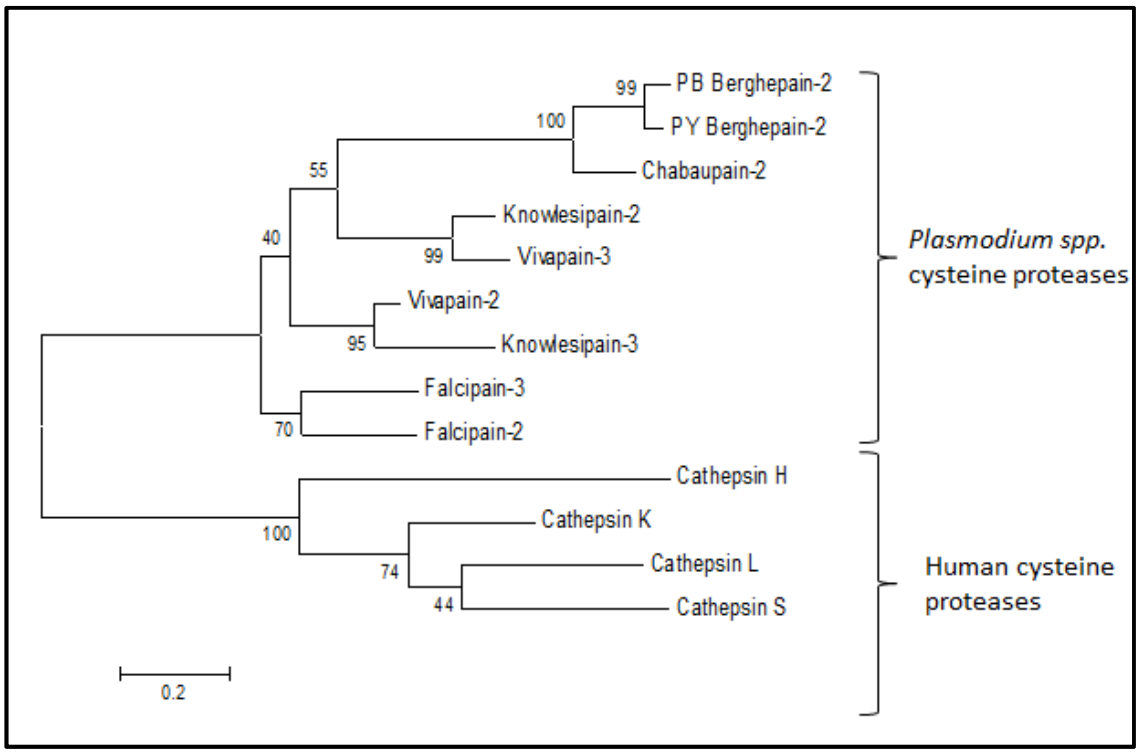


Figure 2.4: Phylogenetic analysis of FP-2 and FP-3 homologs. The tree shows distinct clustering of FP-2 and FP-3 *Plasmodium* homologs from the human homologs.

This observation was in agreement with low sequence identity between FP-2 and the human cathepsin cysteine proteases (Table 2.1). FP-2 and FP-3 clustered together consistent with the high degree of pair wise sequence identity of 68% calculated from the multiple sequence alignment (Figure 2.2). VP-2 and KP-3 retrieved using FP-3 as the query sequence clustered together and were the closest group to FP-2 and FP-3. These two shared a pair wise sequence identity of 76%. The other closest group to FP-2 and FP-3 was VP-3 and Knowlepain-2 which share 83% sequence identity. Chaubapain-2 did not cluster with any of the proteins although it shared an older hypothetical taxonomic unit with BP-2 and BPy-2. Differences observed from the multiple sequence alignment of the rodent (*P. yoelii yoelii*, *P. chabaudi* and *P. Berghei*) *Plasmodium* homologs could explain their distinct clustering from the *Plasmodium* homologs and primate (*P. knowlesi*) infecting homolog (Table 2.2) which also accounts for their lower percentage sequence identities against FP-2 and FP-3 (Table 2.1) when compared with the other *Plasmodium* homologs. Rodent models, though not the best compared to the scarcely available primate models, are widely used as primary models in malaria drug and vaccine development research as well as other *in vivo* experiments (Carlton *et al*. 2002; Jambou *et al*. 2011; Stephens *et al*. 2012). Therefore, an understanding of their differences with FP-2 and FP-3 at protein sequence and functional level is required they to educate the interpretation of experimental results when transitioning to more detailed *in vivo* studies (Chan *et al*. 2005).

2.3.4 Homology Modeling

Seven protein models structures of *Plasmodium* homologs VP-2, VP-3, BP-2, BPy-2, CP-2, KP-2 and KP-3. Models of good quality were obtained. The following sections details results from the whole homology modeling procedure.

2.3.4.1 Template selection

The first step in homology modeling is template selection where protein 3D structure(s) with a similar sequence to the target are identified. Templates were searched for each target sequence using the web based program HHpred (http://toolkit.tuebingen.mpg.de/hhpred). The template search results are summarized in Table 2.3. Two templates were consistently identified during the search among all homologs.

These are 2OUL (FP-2) and 3BWK (FP-3) from PDB. Interestingly, 3BWK the highest scoring in most homologs with an exception of CP-2, BP-2, BPy-2 which are all FP-2 orthologs when the expect value was considered. An expect value of 1.0 E-4 was set as the cut-off point for similar structures identified through the sequence identities. Both VP-2 and VP-3 had 3BWK as the highest scoring template. The templates were further scrutinized for their suitability by assessing the sequence coverage (Table 2.4) and structure resolution. 2OUL had the best resolution of 2.20Å while 3BWK had 2.42Å. Sequence identity, template coverage and the structure resolution are important in comparative modeling because coordinates of identical residues in the template structure and target sequence are simply copied to the target's model structure (Krieger *et al*. 2003).This implies that errors originating from the template structure could be transferred to the model structure if a poor template was selected. A high resolution structure implies more accurate structure thus was selected. The template quality was assed using MetaMQAPII and PROCHECK (Figure 2.4 and 2.5).The two templates had more or less the same resolution; sequence identities differing by at most 4% and more or less equal template coverage (Table 2.3, 2.4).Templates were chosen depending on their sequence identity percentage score with the highest scoring template selected except for VP-2, 3 and KP-3 in which both templates were used. KP-2 was modeled using 3BWK as the template. The rest of the retrieved protein sequences were modeled with 2OUL as the template.

2.3.4.2 Sequence alignment correction

From the multiple sequence alignment in Figure 2.1, 2.2 above, FP-2 and FP-3 was included as well. The MAFFT multiple sequence alignment is only able to give information on the residue similarities and differences among the sequences but does not consider structural information. This alignment alone was not sufficient to carry out modeling hence manual adjustments were made using the structural alignments which were obtained from HHpred target-template sequence alignment and PROMALS3D structural sequence alignment (Appendix 1).

FP homolog	Template ID.	Protein	Template organism	% Sequence Identity	E-value	Resolution Å
VP-2	2OUL	FP -2	*P. falciparum*	63	3.3e-66	2.20
	3BWK	FP -3	*P. falciparum*	67	2.7e-70	2.42
CP-2	2OUL	FP -2	*P. falciparum*	50	1.5e-64	2.20
	3BWK	FP- 3	*P. falciparum*	49	5.8e-65	2.42
KP-2	2OUL	FP -2	*P. falciparum*	56	3.9e-67	2.20
	3BWK	FP -3	*P. falciparum*	58	1.7e-68	2.42
BP-2	2OUL	FP- 2	*P. falciparum*	52	4.7e-64	2.20
	3BWK	FP -3	*P. falciparum*	48	1.3e-65	2.42
BPy-2	2OUL	FP -2	*P. falciparum*	49	1.7e-63	2.20
	3BWK	FP -3	*P. falciparum*	48	4.3e-66	2.42
VP-3	2OUL	FP -2	*P. falciparum*	57	4.3e-67	2.20
	3BWK	FP -3	*P. falciparum*	58	5.6e-69	2.42
KP-3	2OUL	FP -2	*P. falciparum*	57	5.5e-66	2.20
	3BWK	FP -3	*P. falciparum*	60	1.1e-67	2.42

Table 2.3: FP-2 and FP-3 orthologs with their corresponding templates selected from HHpred. Templates were selected based on their percentage sequence identity, resolution, E-values and sequence coverage.

2.3.4.3 Template evaluation and validation

Before proceeding with the homology modeling, the chosen templates were first evaluated to validate their suitability. Homology modeling as previously discussed (Chapter 2, section 2.1.5) is based on sequence homology Krieger *et al*. 2003; Venclovas, 2012), and hence identical residues have their backbone coordinates copied to the model. It is important therefore to ensure that the template protein structure chosen is of high quality with the least number of problematic regions. The methods used for template validation were the same as those used on the model protein structures.

Homolog (Target)	Position In whole sequence	Mature domain numbering	Template(s)	Template coverage	
				Target	Template
VP-2	246-487	1-242	2OUL	2-242	2-241(241)
			3BWK	1-242	2-243(243)
CP-2	231-471	1-241	2OUL	1-241	1-241(241)
			3BWK	2-241	3-243(243)
KP-2	252-495	1-244	2OUL	1-242	1-241(241)
			3BWK	2-244	3-243(243)
BP-2	228-468	1-241	2OUL	1-241	1-241(241)
			3BWK	2-241	2-242(243)
BPy-2	232-472	1-241	2OUL	1-242	1-242(241)
			3BWK	2-241	3-243(243)
VP-3	253-493	1-241	2OUL	2-241	3-241(241)
			3BWK	2-241	4-243(243)
KP-3	240-479	1-240	2OUL	2-240	4-241(241)
			3BWK	1-240	4-243(243)

Table 2.4: Positions of FP *Plasmodium* homolog mature domains in the actual protein sequence, selected templates and their coverage against respective targets.

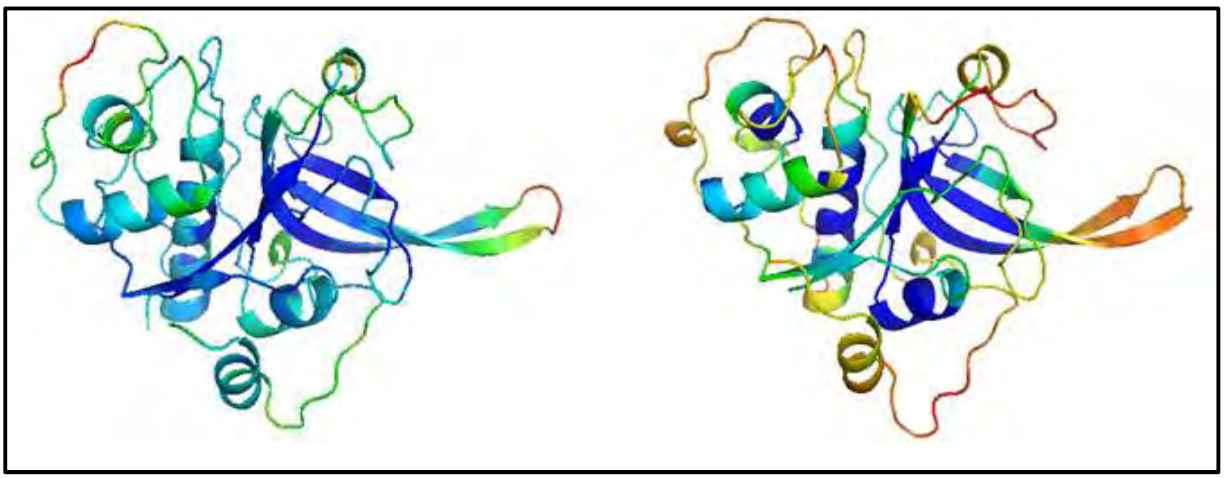


Figure 2.5: MetaMQAP validation results for the templates (a) PDBID: 2OUL and (b) PDBID: 3BWK.The MetaMQAPII score is color coded from blue (stable) to red (unstable) regions.

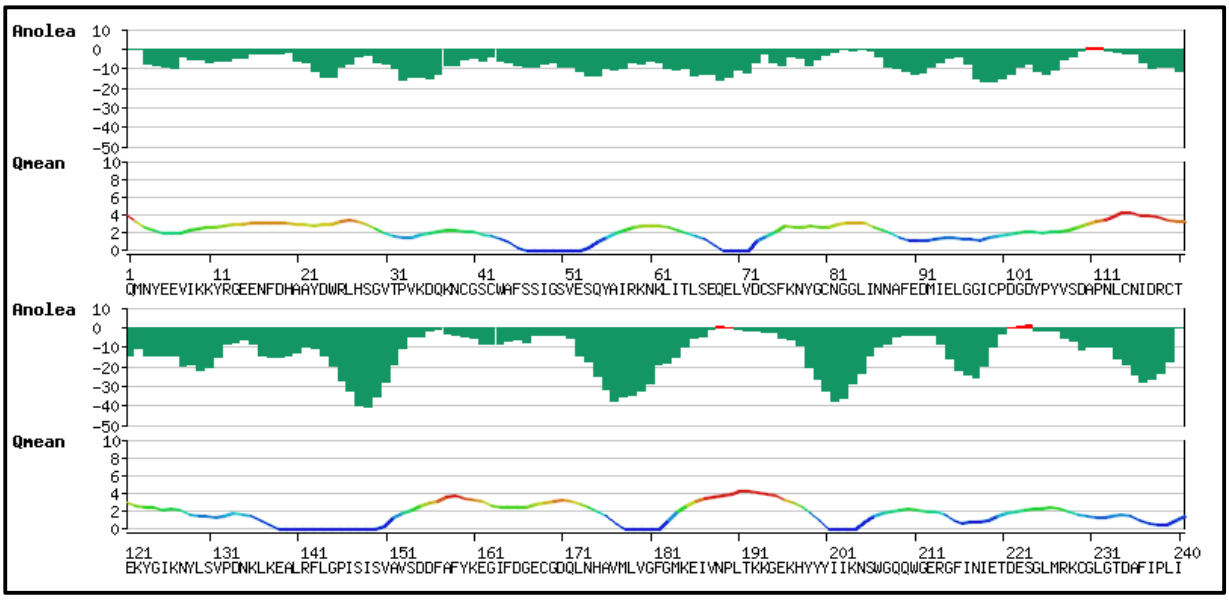


Figure 2.6: ANOLEA and QMEAN6model quality evaluationresults for the template 2OUL (FP-2). The QMEAN energy profile is color coded with blue (stable) to red (unstable).Score values greater than zero in ANOLEA correspond to high energy regions while in QMEAN, scores above 2 represent high energy (unstable) regions.

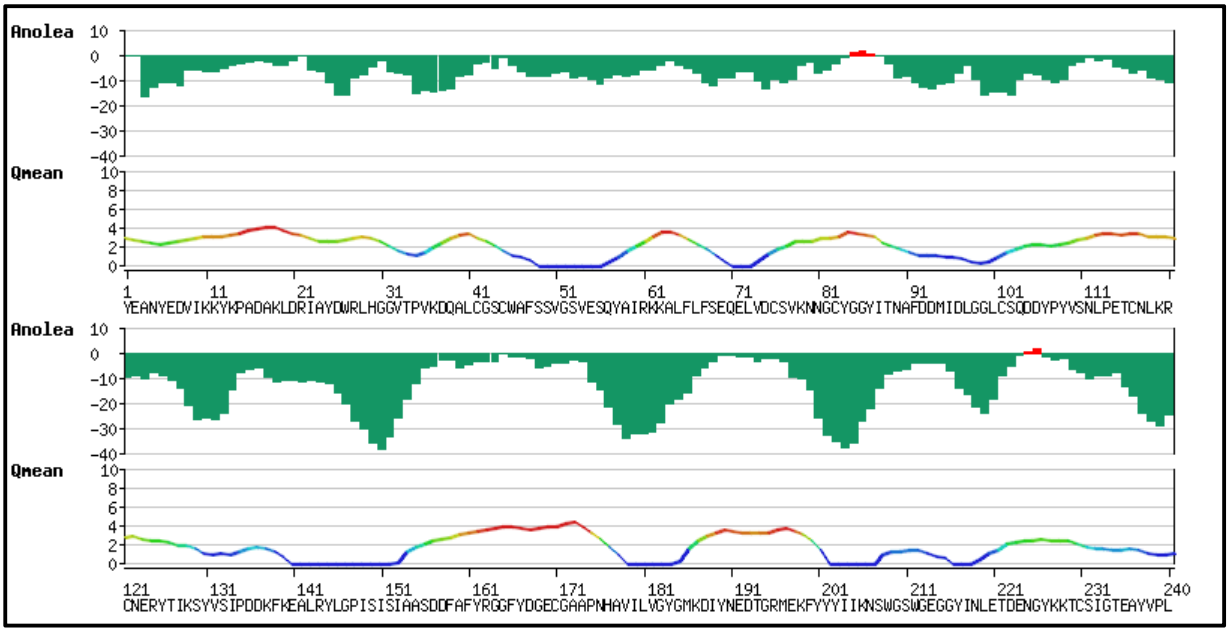


Figure 2.7: ANOLEA and QMEAN6 model quality evaluation results for template 3BWK (FP-3). The QMEAN energy profile is color coded from blue (stable) to red (unstable). Score values greater than zero in ANOLEA correspond to high energy regions while in QMEAN, scores above 2 represent high energy (unstable) regions.

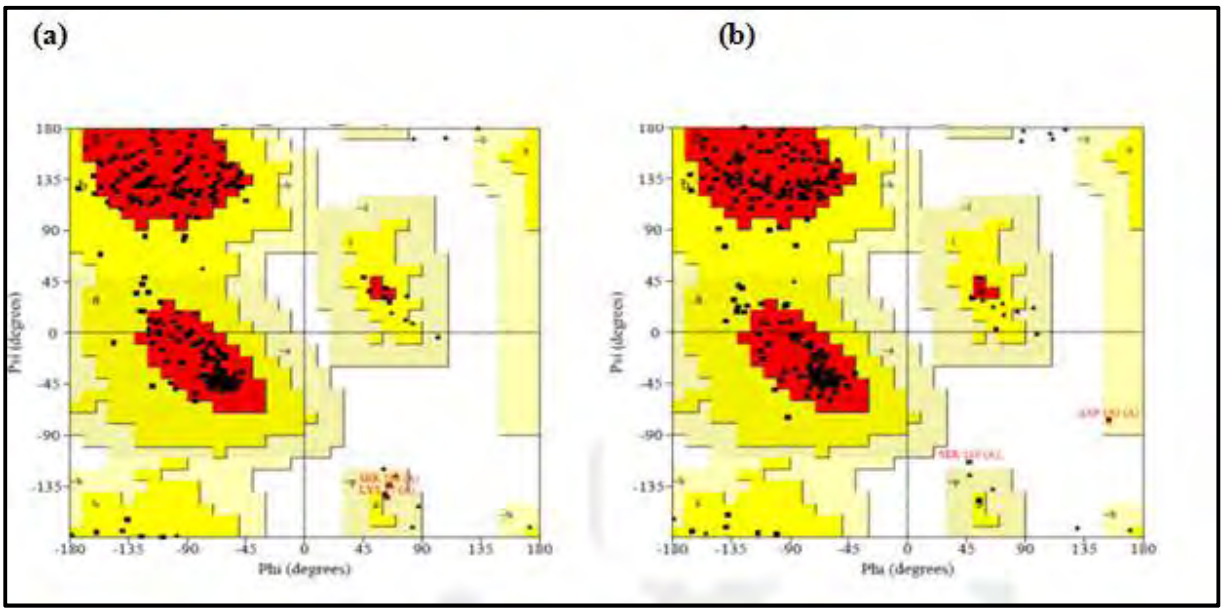


Figure 2.8: a) Ramachandran plot for template structure, 2OUL (FP-2) chain A, and (b) 3BWK (FP-3) chain A.

From the ANOLEA results, the templates were of high quality with only a few problematic regions. The QMEAN6 results for the two templates however seemed to differ with the ANOLEA prediction. However the residues around the active site region were not flagged by the validation programs. The Ramachandran plot for 2OUL (Figure 2.8 a) had two residues Ser108 and Lys37 in sterically non allowed regions. 3BWK also had two residues Ser110 (352) and Asp192 (434) in the same region none of which are in the active site of these proteins. Overall the templates were of acceptable quality as per the methods used for validation.

2.3.4.4 Homology modeling results

MODELLER was used to generate 3D models for the homolog sequences. MODELLER has scripts which can include user specific requirements for special features in the protein structure. In this case, cysteine proteases have disulfide bonds which have to be in the correct stereochemistry to maintain the structural integrity. MODELLER is however able to detect these bonds if present in the template and tries to replicate the same to the models (Šali *et al*. 2011).

The templates used had these disulfide bonds hence no user specifications were used. Loop optimization was not included in the initial model building but rather after validation. For each homolog, 200 models were generated and ranked by their normalized DOPE z scores.

2.3.4.5 Model validation

Among all the models, KP-3 had the highest normalized Z score closer to the templates'. It also had the best GDT-TS score and RMSD as calculated by MetaMQAPII (Table 2.5). The GDT-TS score calculated by MetaMQAPII was calculated based on the number of backbone atoms that confer similar conformation to backbone atoms of protein structures in a database. The RMSD calculates the deviations of the target backbone atom conformations to similar protein structures in the database as well (Pawlowski *et al*. 2008). Positive normalized DOPE scores are likely to be poor models while those with scores from -0.5 and below tend to be near native (Shen & Šali, 2006). This means that all the models in Table 2.5 had good quality based on this parameter. The idealized minimum GDT-TS score as per MetaMQAPII platform is between 56-90%. Again the models had values in the 60-69% range which was relatively good. The RMSD for good models according to MetaMQAPII evaluation is values below 2Å which is indicative of closeness of the model to the native protein structure (Pawlowski *et al*. 2008). However, RMSD evaluation can be biased and inaccurate when the protein being modeled is distantly related to the template and has most regions correctly predicted but incorrectly predicted region are very remote to the template hence a poor RMSD score. Sequence length also affects the outcome of the RMSD score (Li *et al*. 2011). The GDT-TS score is on the other hand one of the many accepted and widely used model evaluation method giving more reliable results compared to RMSD.

GDT-TS or Global Distance Test – Total Score, uses a set of distance thresholds to perform model and template quality evaluation. Scores are calculated based on matched pair wise alignment of the structures within 1Å, 2Å, 4Å and 8Å distance thresholds and an average score given as the final score (Li *et al*. 2011). From the Table 2.5 the RMSD as per MetaMQAPII were quite good considering the length of the sequences (approximately 250 residues). The GDT-TS scores were also on the acceptable range 56-90% (Pawlowski *et al*. 2008).

On performing template to target MSA it was observed that both BP-2 and BPy-2 and VP-3 were closer to FP-2 than FP-3 which was in agreement with observations made from the pair wise sequence alignment (Table 2.1 and 2.3) and the phylogenetic analysis (Figure 2.4).

Protein Homolog	DOPE Z-score	GDT-TS	RMSD	Template RMSD 2OUL (FP-2)	Template RMSD 3BWK (FP-3)
VP-2	-1.2714	65.393	2.370	0.348	0.286
CP-2	-0.8962	63.797	2.499	0.178	0.488
KP-3	-0.9650	63.422	2.558	0.498	0.179
BP-2	-0.8768	60.477	2.700	0.183	0.476
BPy-2	-0.8227	63.589	2.478	0.283	0.357
VP-3	-1.0265	64.627	2.476	0.154	0.467
KP-3	-1.3292	69.062	2.082	0.301	0.308

Table 2.5: Model scores as predicted by modeler and METAMQAPII protein structure validation program. The template structures (2OUL and 3BWK) had DOPE Z scores of -1.4468 and -1.4071 respectively.

The models were of acceptable quality according to these methods as is shown in Figure 2.9 below. Further analysis was however performed to ascertain the quality of the calculated models because different model quality validation methods employ different parameters.

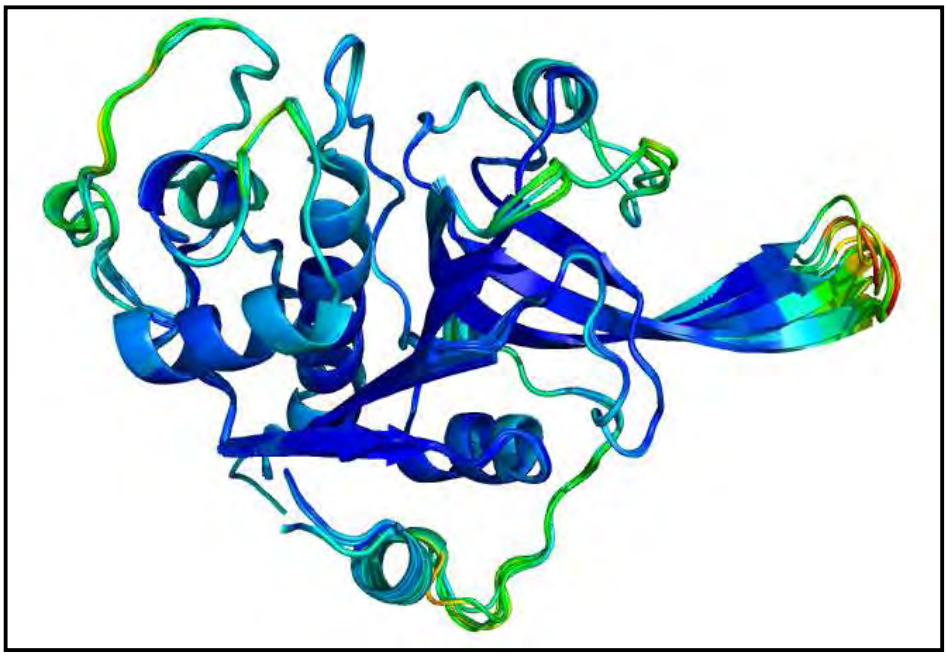


Figure 2.9: Superimposed models of all retrieved *Plasmodium* homologs colored according to the MetaMQAPII scores. Color code ranges from blue (favorable) to red (unfavorable) regions. The models aligned well with the template structures 2OUL (FP-2) and 3BWK (FP-3).

2.3.4.5.1 BP -2

Figure 2.10 shows the 3D structure of BP-2 from *P. berghei*. From the MetaMQAPII energy scores the active site region of the structure which is the main focus of interest was well modeled. The PROCHECK results showed that 87% of the residues are in allowed regions 12% in additionally allowed regions, 0.5% in generously allowed areas and the remaining 0.5% in the disallowed regions comprising of two residues Ser109 and Gln38. These residues are at corresponding positions to those of 2OUL template (Figure 2.8 a) hence the error originated from the template. From ANOLEA and MetaMQAPII analysis, Gln38 was not flagged as erroneous.

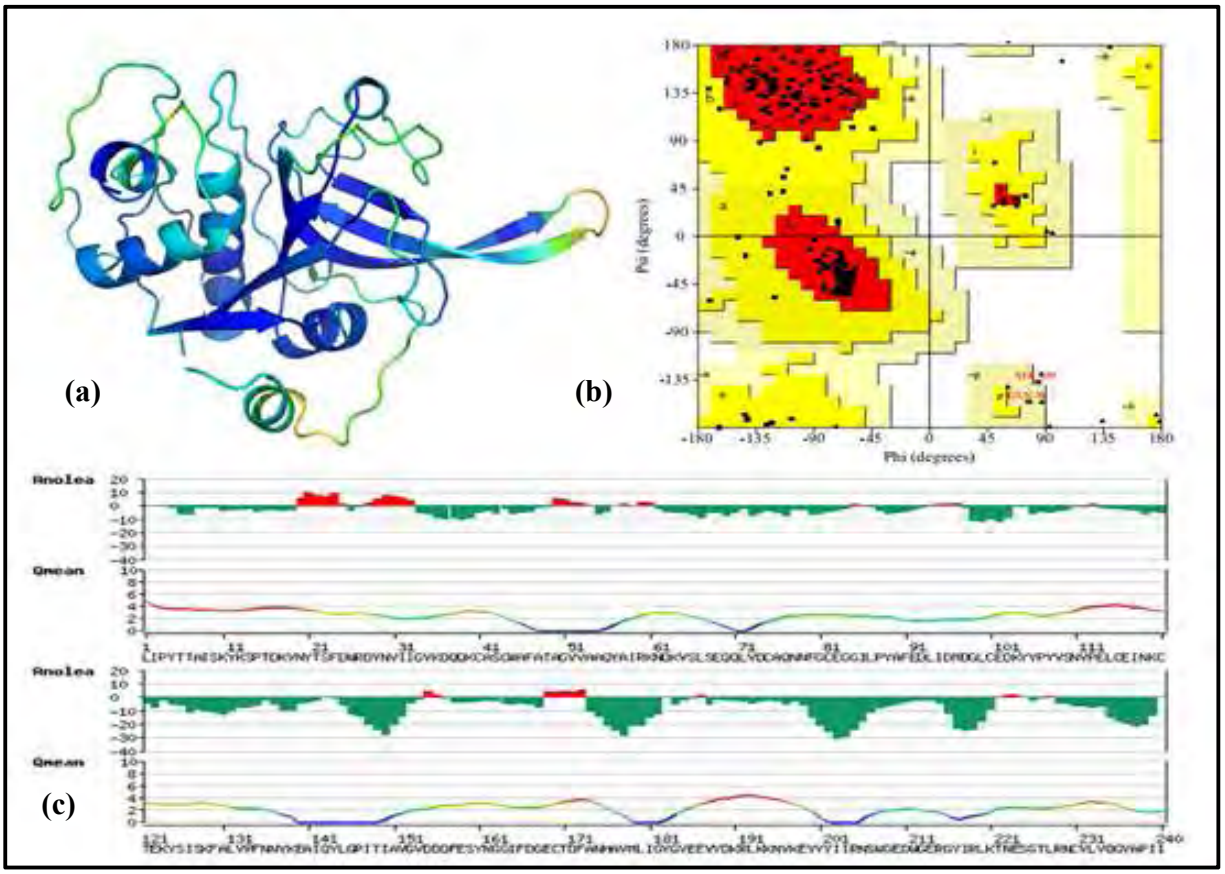


Figure 2.10: a) MetaMQAPII energy profile for BP-2. b) PROCHECK analysis and, c) ANOLEA evaluation and QMEAN6 energy profile.The MetaMQAPII and QMEAN scoresare color coded from blue (stable) to red (unstable) regions.

The regions highlighted by ANOLEA as problematic were in the loop regions of the structure and not in the active site. The QMEAN graph seemed to flag regions not flagged by ANOLEA for instance the first 20 residues at the C-terminal were flagged unfavorable by QMEAN while ANOLEA did not. The ANOLEA results were however consistent with MetaMQAPII results. Overall BP-2 (*P. berghei*) structure was good enough to proceed to docking studies. The model also had a good normalized Z score of -0.8768, GDT-TS of 60.477 and 2.700Å RMSD according to MetaMQAPII. BP-2 was modeled using 2OUL as the template. An analysis of the template to model RMSD revealed a deviation of 0.183Å which is quite low and near nativity.

2.3.4.5.2 BPy-2

The BPy-2 (*P. yoelii yoelii*) had a few problematic regions mostly at the loop regions. These are the regions flagged by ANOLEA. QMEAN6 and ANOLEA (Figure 2.11 c) conflicted in the scoring of residues between positions 45 to 55 (papain numbering).

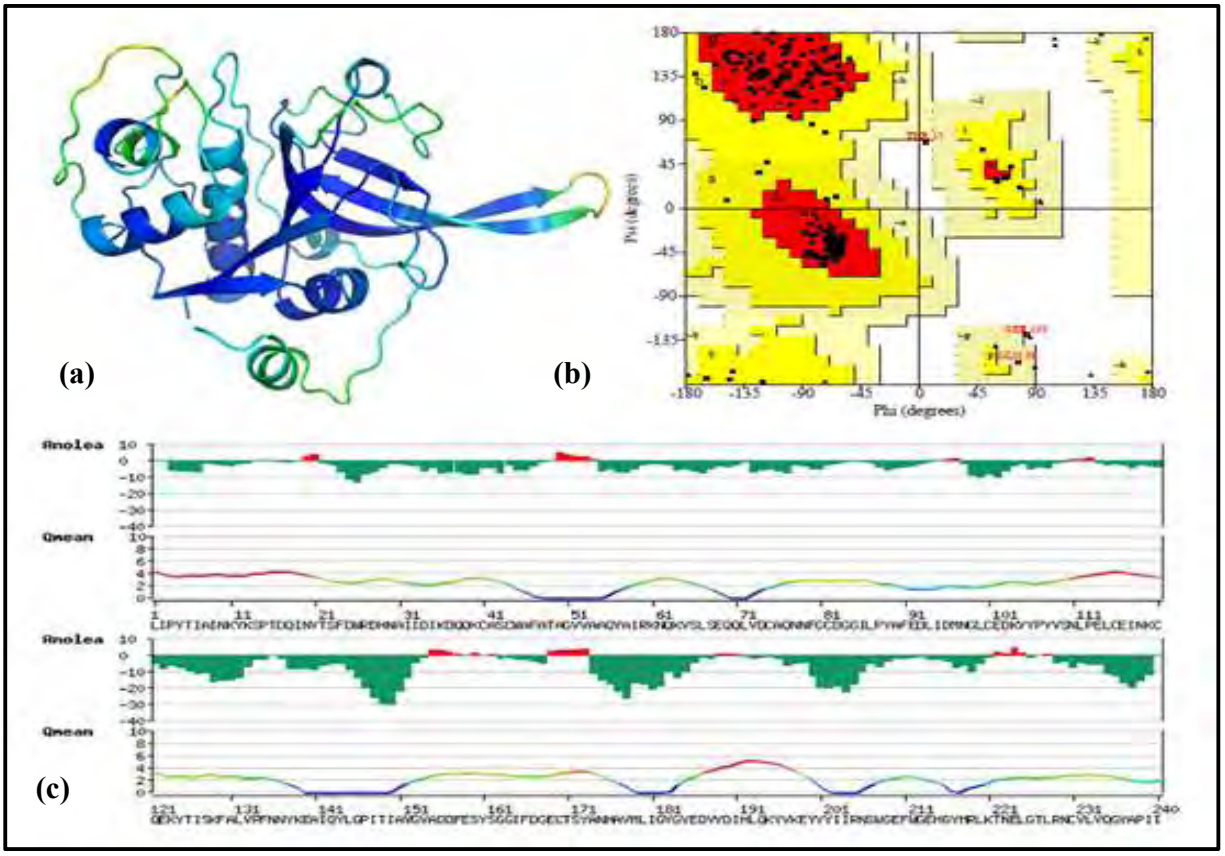


Figure 2.11: a) MetaMQAPII energy profile for BPy-2. b) PROCHECK analysis and, c) ANOLEA evaluation and QMEAN6 energy profile.The MetaMQAPII and QMEAN scoresare color coded from blue (stable) to red (unstable) regions.

QMEAN6 also differed with MetaMQAPII (Figure 2.11 a) on the N-terminal residues flagging them as unfavorable, whereas ANOLEA and MetaMQAPII did not. The PROCHECK (Figure 2.11b) results showed 89% of the residues were in most favored regions, 9 % in additionally allowed regions, 0.5% in generously allowed regions and 0.9% in disallowed regions. The active site of the protein was well modeled as none of the residues were flagged. However the two residues flagged at the disallowed region in the template 2OUL affected the corresponding residues for this model.

The Gln38 residue (268 in whole sequence) which is part of the S1 sub site was however not flagged erroneous by MetaMQAPII and AOLEA and the error was associated with the corresponding residue in the template structure. The rest of the active site residues were not flagged erroneous by any of the model quality evaluation programs used. The model had a normalized Z score of -0.8227, GDT-TS of 63.589 and RMSD of 2.478Å as calculated by MetaMQAPII (Table: 2.5) which was quite good. The template (2OUL) model RMSD was 0.283Å. Overall the model had acceptable quality hence was used for molecular further structural analysis and molecular docking.

2.3.4.5.3 CP-2

CP-2 was of quite good quality according to MetaMQAPII (Figure 2.12 a) apart from a few residues in the N-terminal. QMEAN and ANOLEA (Figure 2.12 c) flagged some residues as unfavorable although conflicting but were in agreement on residues between positions 165 to 175 (mature domain numbering). PROCHECK (Figure 2.12 b) showed that 86% were in most allowed regions, 12% in additionally allowed regions 0.9% in generously allowed regions and 0.5% in disallowed regions. This was a Serine residue at position 208 (438 in whole sequence) and the error was not carried forward from the template structure 2OUL. The other two residues in the disallowed regions were Ala109 (339) and Lys38 (268) corresponding to the residues in the disallowed regions of the template protein structure (2OUL) in the sequence alignment. FROM the ANOLEA analysis the Lys38 (268) was not flagged to be erroneous but QMEAN6 energy profile had a relatively low score. Lys109 (339) was not flagged erroneous by ANOLEA, MetaMQAPII and QMEAN6. The Ser208 flagged by PROCHECK to be in disallowed regions was not flagged by MetaMQAPII and ANOLEA. However QMEAN6 gave it a low score. These residues were not found in the active site region which was the main site of interest. Considering the RMSD of 0.178Å for the model against its template, a normalized Z score of -0.8962 (Table 2.5) and a GDT-TS score of 63.797Å from MetaMQAPII analysis, the overall assessment of the Chaubapain-2 the model quality was good for use in further analysis.

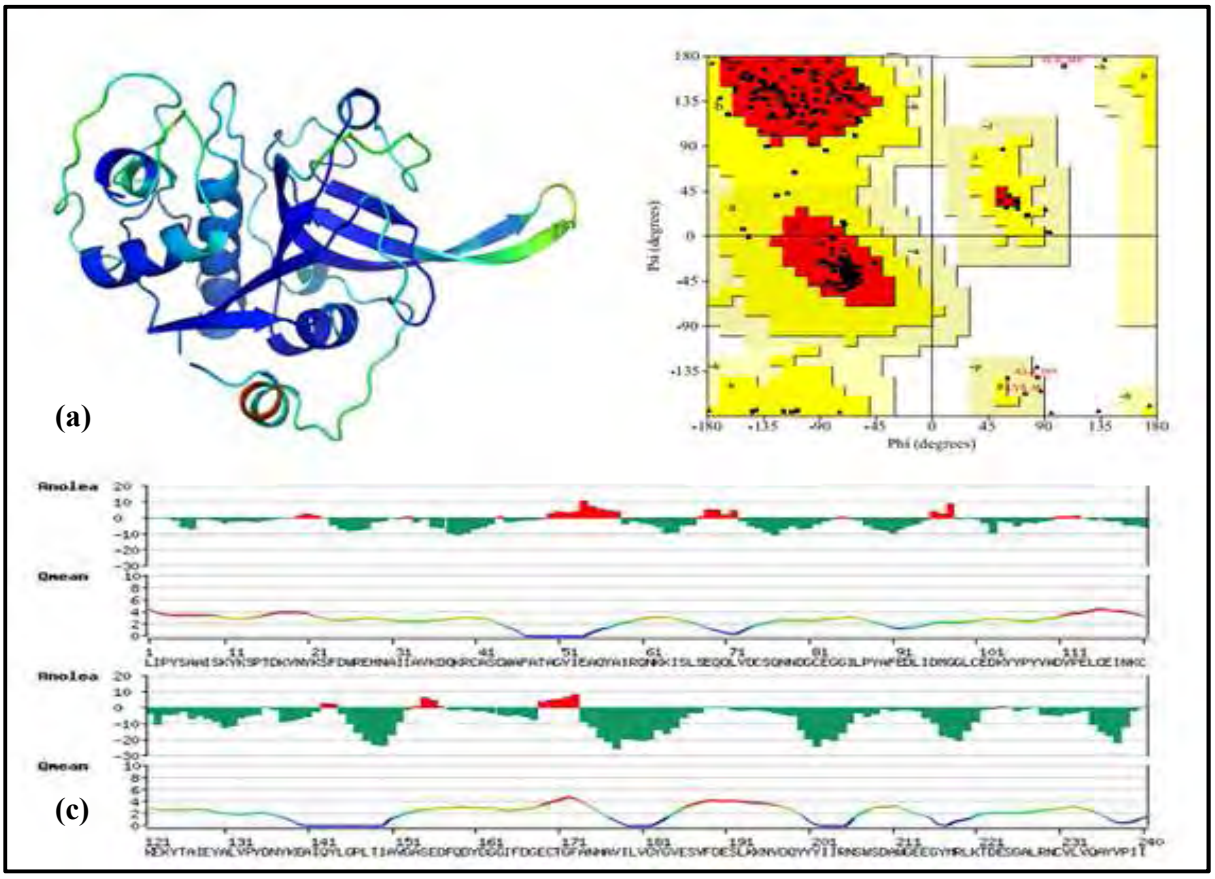


Figure 2.12: a) MetaMQAPII energy profile for CP-2 (*P. chabaudi*). b) PROCHECK analysis and, c) ANOLEA evaluation and QMEAN6 energy profile. The MetaMQAPII and QMEAN scores are color coded from blue (stable) to red (unstable) regions.

2.3.4.5.4 KP-3

KP-3, an ortholog of FP-3, (Figure 2.13) had the loop regions flagged as unstable by MetaMQAPII, however ANOLEA (Figure.13 c) only flagged a few residues in the loop region near the N-terminal and the 14 residue insert forming the β-hairpin. In the case stereochemistry, this protein had the best with 92% of residues in the most favored regions, 8% in additionally allowed regions 0.5% in generously allowed regions and none in the disallowed regions. This finding concurs with its DOPE Z score from MODELLER of -1.3292 which was the highest among all protein homologs. The model's RMSD against the template proteins was 0.301Å for 2OUL and 0.308Å for 3BWK.

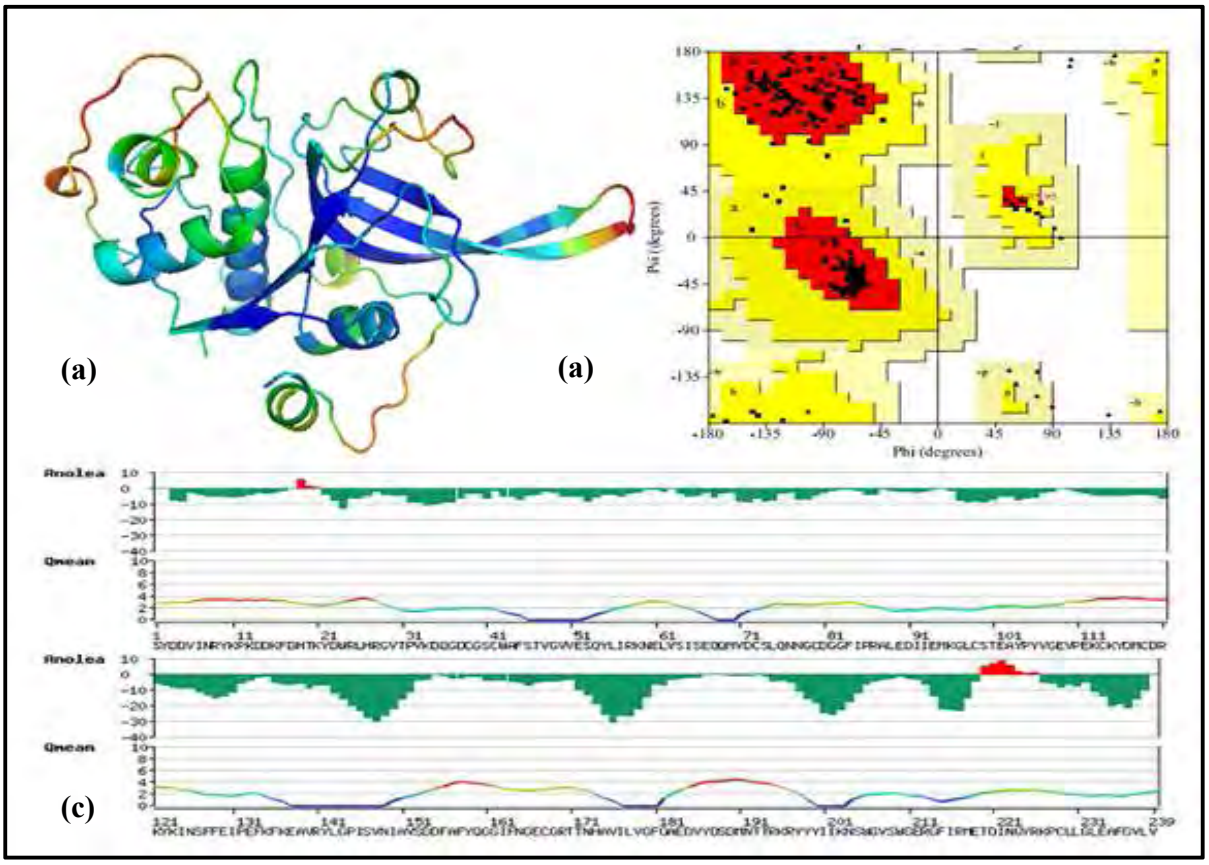


Figure 2.13: a) MetaMQAPII energy profile for KP-3. b) PROCHECK analysis and, c) ANOLEA evaluation and QMEAN6 energy profile. The MetaMQAPII and QMEAN scores are color coded from blue (stable) to red (unstable) regions.

The GDT-TS score according to MetaMQAPII was 69.062 which were the highest among all models. The low energy scores in the loop regions were associated with the 3BWK template structure (Figure 2.4 b). The active site region was of a quality which could be explained by the high level of conservation at this site. The model was selected for further analysis although. Further loop optimizations would however improve the overall model quality.

2.3.4.5.5 KP-2

KP-2 was retrieved using FP-2 as the query sequence. It had few residues flagged erroneous by ANOLEA (Figure 2.14). The model had good energy scores according to MetaMQAPII with a few unstable regions mostly in the loops. PROCHECK (Figure 2.14 b) showed that 91%of the residues were in the allowed regions, 7% in additionally in the allowed, 0.9% in the generously allowed and 0.9% in the disallowed regions (Asp109 and Lys38). Asp109 (360) and Lys38 (290) are at corresponding positions in the sequence alignment with the two residues of 3BWK template flagged at the disallowed regions hence the error was carried forward from the template. These two residues were however not flagged by ANOLEA and QMEAN6 as erroneous. From the MetaMQAPII energy profile which is incorporated in the model structure, this residue is found at the loop region connecting the two sheets that form the nose region or β-hairpin of the protein structure. This region from the sequence alignment (Figure 2.2) is highly varied. From the template structure as well, the region was depicted to have poor quality hence the poor quality at the corresponding position in resulting model for the *P. knowlesi* FP-2 ortholog. The model had a normalized Z score of -0.9650 and RMSD of 0.179Å to its template structure (3BWK) which was quite low hence the model structure was closer to the template. Combining these results with the MetaMQAPII results, RMSD of 2.558Å and GDT-TS of 63.422, the model quality was good enough to be considered for further analysis.

2.3.4.5.6 VP-2

VP-2 had a good energy distribution according to MetaMQAPII model quality evaluation (Figure 2.15 a) with most residues flagged as stable apart from a few in the loop connecting the beta sheets that form the β-hairpin. Cys81 (326 in the whole sequence) is located loop region which forms part of the S1 sub site of VP-2 was previously flagged erroneous by MetaMQAPII hence loop refinement was done and the cysteine disulfide bridge bond formed by the Cys81 (326) and Cys40 (285) verified using PIC (*Supplementary data/Chapter 2: Supplementary data/Figure 8*) and is the model shown in Figure 2.14 a. The binding site did not have any residues flagged unfavorable. From PROCHECK (Figure 2.15 b), 87% of the residues occupied most allowed regions, 9% in the additionally allowed, 0.5% in the generously allowed and 0.9% in disallowed regions. ANOLEA (Figure 2.15 c) flagged residues at the extreme ends of both the

N and C terminals and a few residues within the sequence. QMEAN (Figure 2.14 c) flagged most of the loop regions erroneous which could be as a result of the variations in these areas and the poor quality of the corresponding regions in the template corresponding with MetaMQAPII evaluation results. VP-2 was modeled using both templates. The GDT-TS score reported by MetaMQAPII was 65.393 and an RMSD of 2.370Å which were at acceptable ranges. The template to model RMSD was 0.348Å and 0.286Å for 2OULand 3BWK respectively. RMSD values towards zero or nativity are normally the best. In the case of VP-2 model it was closer to 3BWK than it was to 2OUL. In general the model was good enough to be used for further analysis.

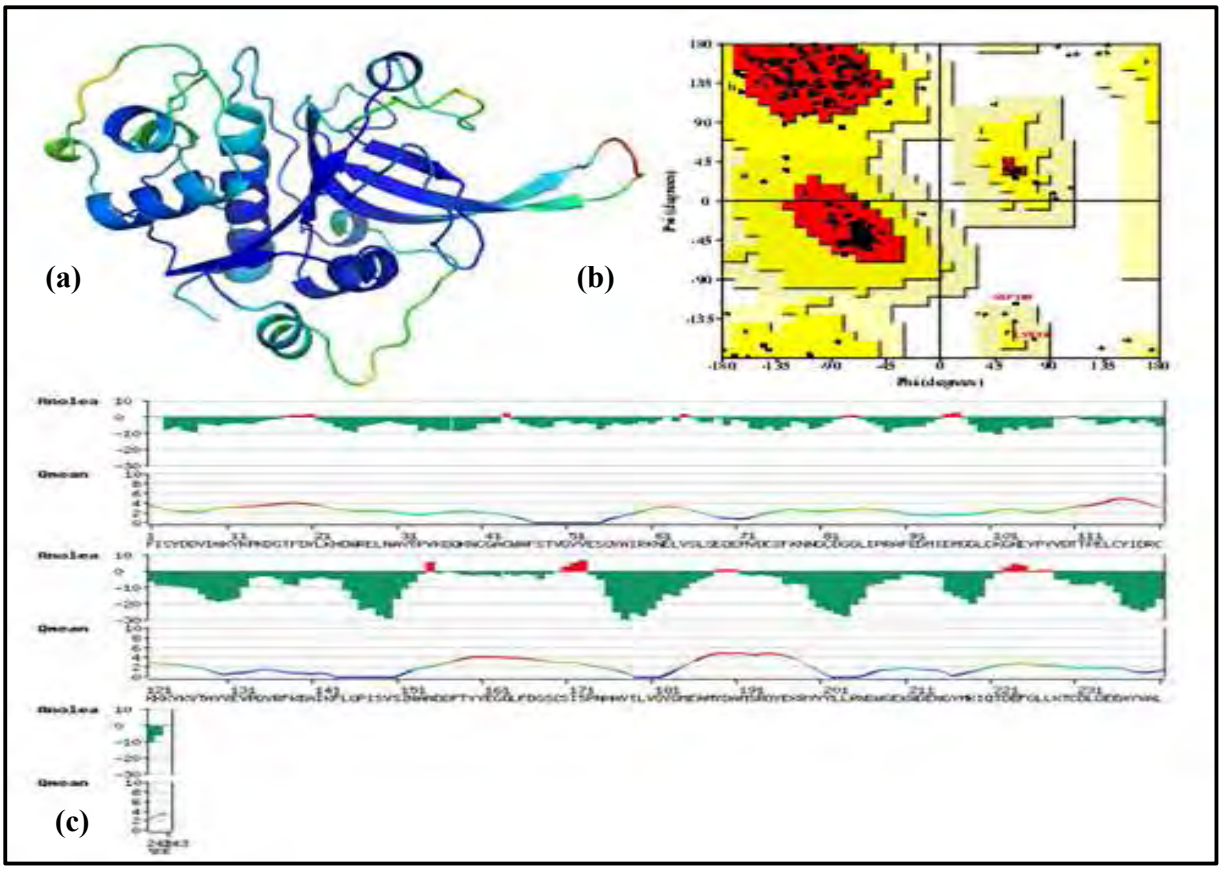


Figure 2.14: a) MetaMQAPII energy profile for KP-2. b) PROCHECK analysis and, c) ANOLEA evaluation and QMEAN6 energy profile. The MetaMQAPII and QMEAN scores are color coded from blue (stable) to red (unstable) regions.

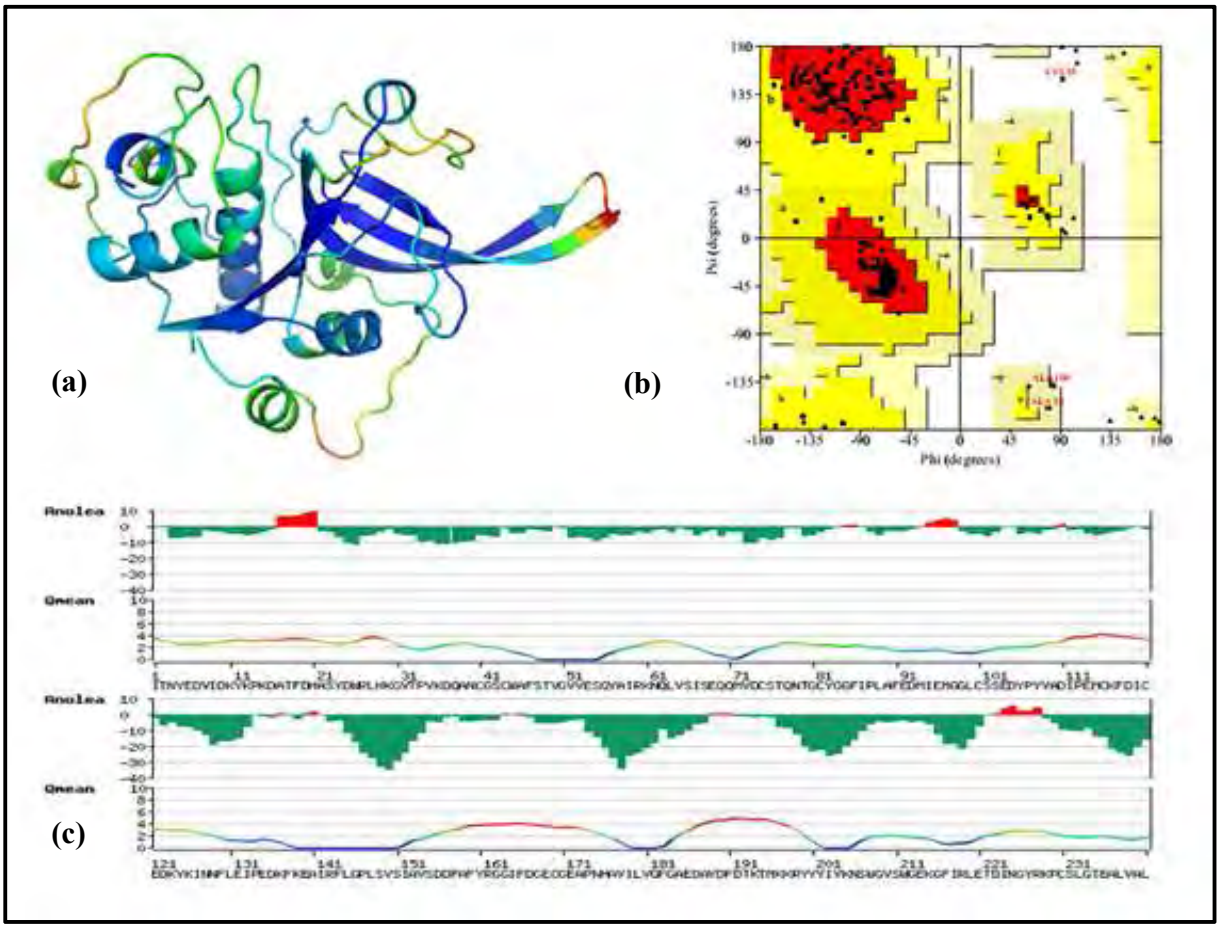


Figure 2.15: a) MetaMQAPII energy profile for VP-2. b) PROCHECK analysis and, c) ANOLEA evaluation and QMEAN6 energy profile. The MetaMQAPII and QMEAN scores are color coded from blue (stable) to red (unstable) regions.

2.3.4.5.7 VP-3

VP-3 as per MetaMQAPII (Figure 2.16 a) model quality evaluation was very stable except for the loop region connecting the sheets forming the β-hairpin which was also flagged unfavorable by ANOLEA although QMEAN showed it as relatively stable. This error was most likely carried forward from the template structures in Figure 2.5 a, and b. The active site residues were not flagged unfavorable by any of the validation methods. PROCHECK (Figure 2.16 b) predicted 87% of the residues to be in the most allowed regions, 10.0% in the additionally allowed regions, 0.9% in the generously allowed regions and 0.5% consisting of a Lys13 (265) residue at position which was also captured by ANOLEA and QMEAN (Figure 2.16 c).This residue is positioned in the highly variable N-terminal insert which could explain why it was predicted to be in

disallowed regions. Asp106 (358) and Lys37 (289) flagged by PROCHECK to be at generously allowed regions do not fall in the active site. This feature was probably carried forward from the 2OUL template (Figure 2.8 a) which had two residues at the same position as per PROCHECK evaluation. Careful analysis of the multiple sequence alignment confirmed that the two residues were corresponding to those of the template 2OUL. Looking at the normalized Z score, Vivapain-3 had -1.0265 which is in the acceptable range, a GDT-TS of 64.627 and RMSD of 2.476Å as per MetaMQAPII evaluation. The RMSD of the model to the template was 0.154Å for 2OUL and 0.467Å for 3BWK in which case the model was closer to the 2OUL template. Overall the model had sufficient quality to be used for further analysis.

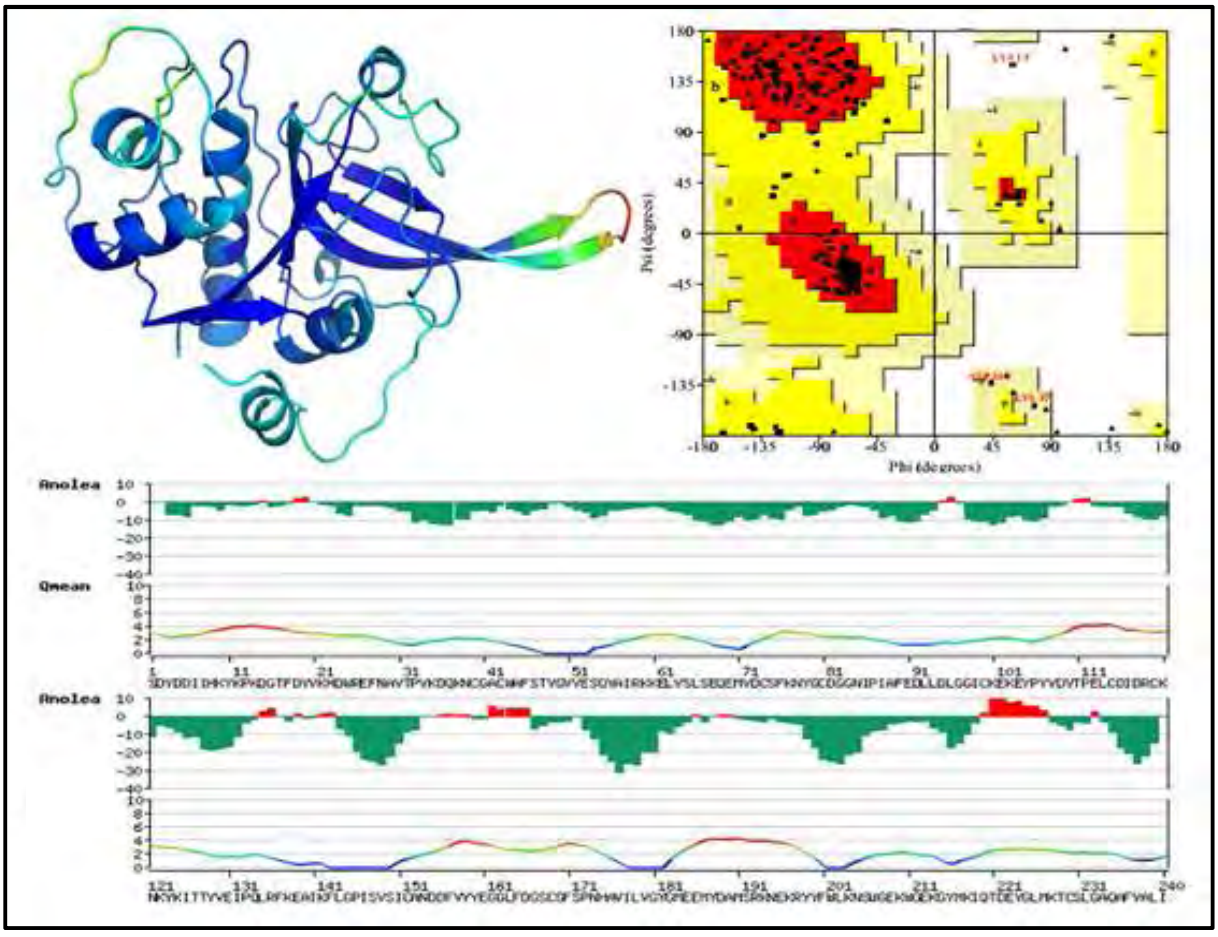
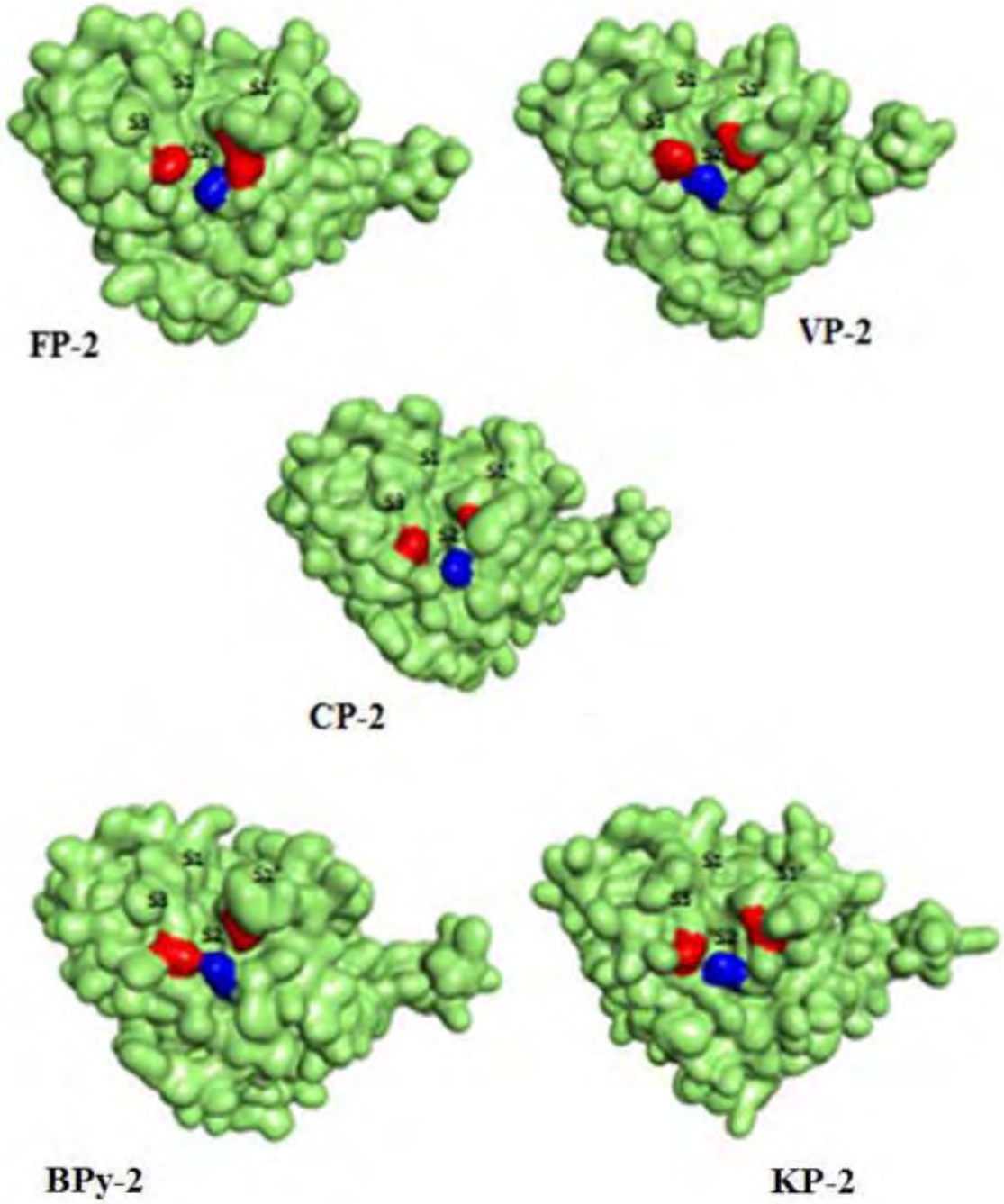


Figure 2.16: a) MetaMQAPII energy profile for VP-3. b) PROCHECK analysis and, c) ANOLEA evaluation and QMEAN6 energy profile. The MetaMQAPII and QMEAN scores are color coded from blue (stable) to red (unstable) regions.

2.3.4.6 Comparative structural analysis of active site size and volume

Substrate binding is affected my several factors among them the molecular accessible surface area which interacts with the substrate via intermolecular forces, size and volume. The S2 pocket as previously mentioned has been postulated as the major specificity determinant in papain-like cysteine proteases (Chapter 2 section 2.3.2.2, iii). From crystallography data, FP-2 sub site has been postulated to be larger than FP-3's. This has been attributed to Leu84 (327), Leu172 (415), Asp234 (477) and Tyr93 (335), Pro181 (423), Glu 243 (485) in FP-2 and FP-3 respectively. These residues sit at the entrance of the S2 pocket and have been referred to as "gatekeepers" in other cysteine proteases (Stack *et al*. 2008). The combination of Tyr93 (335) a bulky residue and Pro181 (423) a rigid amino acid causes narrowing of the S2 sub site of FP-3 (Kerr *et al*. 2009). In the multiple sequence analysis of the sub sites variation (Table 2.2), the *Plasmodium* homologs seemed to have residues with physiochemical characteristics similar to both FP-2 and FP-3. Figure 2.17 highlights these three residues on the S2 sub site entrance. The effect of these residues on the size S2 sub site has been previously described (Kerr *et al*. 2009). This observation can however be validated by assessing the binding characteristics of ligands docked to all the homolog proteins which was covered in Chapter 3. From a comparative molecular modeling study of FP-2 and FP-3, the volumes of these proteases were shown to be different with FP-2 having a larger S2 pocket (Sabnis *et al*. 2003). The effect of these variations on the proteases S2 and the binding pocket in general could be significant for inhibitor binding. From previous site mutation studies involving human cathepsin L and K revealed differences in both size and shape of the side chains at the S2 pocket at corresponding positions (Lecaille *et al*. 2007). In another FP-2 and FP-3 inhibitor screening study, it was shown that all inhibitors that could inhibit FP-3 could inhibit FP-2 as well while the reverse wasn't true for some of the inhibitors and this phenomenon was attributed to and somehow confirms the size differences in FP-2 and FP-3 S2 sub site previously suggested (Desai *et al*. 2006). Residue variations at the active site size observed in the human and *Plasmodium* homologs (Figure 2.2) in almost all the sub sites could help deduce important interactions in both cases. These variations could be used to maximize selectivity as well as ease *in vivo* test results interpretation.


FP-2
VP-2
CP-2
BPy-2
KP-2

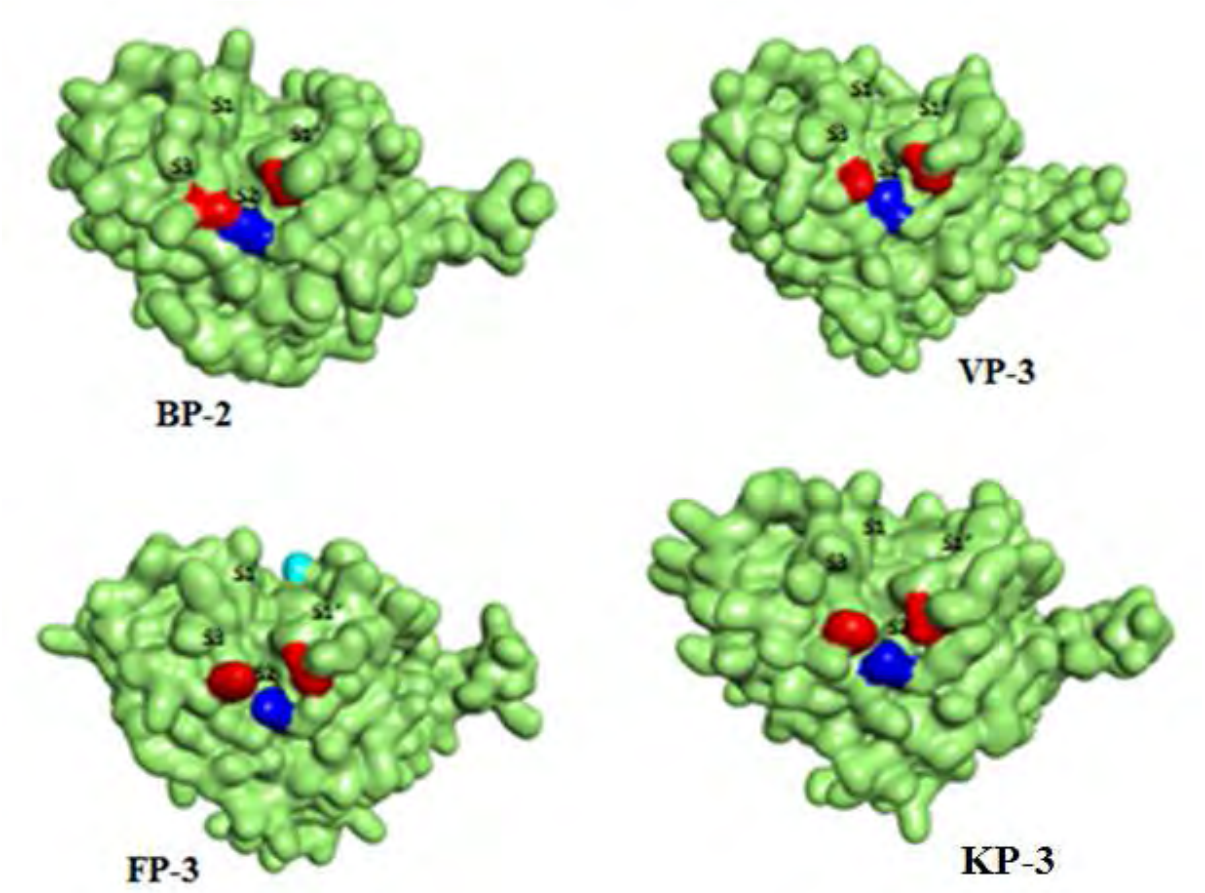


Figure 2.17: Surface presentations of FP-2, FP-3 and, *Plasmodium* homolog model structures. Colored in red are Leu84 (327), Leu172 (415) and blue D234 (477) of FP-2 and corresponding positions (Table 2.2) S2 sub site residues affecting the size and volume of the S2 sub site.

CHAPTER THREE

3. MOLECULAR DOCKING

Chapter 2, dealt with comparative analysis at the sequence and structural level of FP-2 and FP-3 homolog proteins. This chapter is aimed at finding out how residue variations at the active site among these homolog proteins may affect molecular binding to various known cysteine protease inhibitors from a structural perspective. These interactions could be important for deducing factors affecting molecular recognition by the FP-2, FP-3 *Plasmodium* and human homologs that could be exploited to enhance inhibitor selectivity. Malaria treatment is mainly based on drugs developed to target *P. falciparum* but drug resistance has been observed in both *P. falciparum* and *P. vivax* (Wongsrichanalai *et al*. 2002).

In vivo experiments on malarial drugs are conducted on rodent models (Carlton, 2002; Jambou *et al*. 2011; Stephens *et al*. 2012). Unfortunately, there is limited biochemical and structural information on these rodent infecting *Plasmodium* homologs. Differences observed between FP-2, FP-3 and the rodent *Plasmodium* homologs in terms of substrate utilization are of major concern. This is because they could complicate interpretation of *in vivo* results where inhibitors effective against FP-2 and FP-3 may not necessarily replicate the same inhibitory effect against the rodent models used (Rosenthal *et al*. 2002). It is therefore, important that the source of these variations both in *Plasmodium* and human homologs be studied in detail to inform *in vivo* inhibitor results interpretation thus improving inhibitor selectivity.

Currently, known inhibitors are not capable of having selective inhibition, and have been associated with susceptibility to degradation by host proteases and poor pharmacological profiles (Ettari *et al*. 2011; Ettari *et al*. 2009), thus, this has been the focus of FP inhibitor research (Turk, 2006). Due to the limitations mentioned in Chapter 1, section 1.9, molecular docking could provide an easier and effective way of analyzing these variations, as well as screen selected compounds to determine their binding affinities against FP-2 and FP-3 *Plasmodium* homologs.

3.1 Introduction

A clear understanding of structural details that govern recognition events between proteins or proteins and small molecules could accelerate the discovery of novel therapeutics to targets of interest (Rzychon *et al*. 2004). The success of small molecule inhibitors designed against protease of HIV, diabetes, hypertension and osteoporosis has spurred further interest in other proteases (Deacon, 2011; Stoch & Wagner, 2007). Cysteine protease inhibition can be achieved through molecular binding by partial substrate like binding, substrate like binding and backward binding all of which interfere with the substrate's ability to bind to the substrate with or without formation of a covalent bond (Rzychon *et al*. 2004). Experimental biochemical enzyme assays and X-ray crystallography have been used to deduce the structural information of the protein-inhibitor complexes important for successful binding and subsequent inhibition but has many limitations (Chapter 1, section 1.9). Structural details of receptor proteins could be obtained from comparative protein modeling which, in combination with docking results, could provide valuable information with regards to protein-inhibitor binding (Hillisch *et al*. 2004).

Molecular docking can be used to deduce important binding interactions of potential protein inhibitors as well as to validate homology models (Sabnis *et al*. 2003). Previously molecular docking has been used for virtual screening of inhibitor libraries targeting FP-2 and FP-3 (Desai *et al*. 2006; Desai *et al*. 2004; Kitchen *et al*. 2004; Laurie & Jackson, 2006; Shah *et al*. 2011).The objective of structure based drug design is to identify potential lead compounds from collections of compounds compiled in large libraries by accurately predicting their binding affinities for specific proteins/enzymes/receptors by docking (Kumar & Zhang, 2012; Saranya & Selvaraj, 2009).

Molecular docking is a computational method used to study protein-ligand interactions by searching the best conformation at which flexible ligands are able to bind in a fitting manner both geometrically and energetically to the binding site of a receptor protein (Gschwend *et al*. 1996; May & Zacharias, 2005). It is a common application used in the study of biomolecular complexes in structure and function analysis and rational structure based molecular drug design (May & Zacharias, 2005; Saranya & Selvaraj, 2011). The inhibitors, when bound to the receptor proteins exhibit conformations that are both geometrically and chemically complementary for

which molecular docking can be used to deduce these conformations. This can be done in two ways which include; one, blind docking and docking to predicted binding sites (Laurie & Jackson, 2006). Blind docking is applied to the whole protein and implicitly includes binding site prediction providing information about correct ligand binding orientation (Laurie & Jackson, 2006). Docking into the predicted binding site allows for virtual screening of large libraries of potential inhibitor compounds to be analysed to find which ones bind with the highest affinities (Kumar & Zhang, 2012; Laurie & Jackson, 2006). Docking into predicted binding sites is much faster and cheaper compared to experimental methods and has been used with significant outcomes of drugs being released into the market (Alvarez, 2004). As is in any other computational analysis, docking has its own challenges as well, one being the integration of protein flexibility into these algorithms that are being used in rational structure based drug design for screening libraries of possible therapeutic compounds. Several methods have been proposed and implemented in docking programs (Teodoro *et al*. 2001). There are two main types of protein docking, rigid and flexible.

3.1.1 Rigid docking

Rigid docking is based on the lock and key approach proposed by Fischer in 1890 where it is the ligand that undergoes changes in its 3D structure to find the best spatial and energetic conformation to fit into the proteins receptor's binding site (Sullivan & Holyoak, 2008). This method is however biased because it restricts conformational search in the case of ligands requiring certain conformational modification of the protein receptor before binding (Flick, Tristram, & Wenzel, 2012; Weikl & von Deuster, 2009). Improved performance in terms of accuracy of this method is mediated by allowing flexibility of the ligand side chains thus allowing for numerous conformational changes during docking (Lexa & Carlson, 2012). The presence of ligand bound structures (holo), in PDB which provide important structure activity relationship information makes this method more applicable to use with unbound (apo) structures for rational structure based drug design.

3.1.2 Flexible docking

Because of the limitations of rigid docking (Chapter 3, section 3.1.1), attempts have been made to implement the more feasible ligand protein binding approach (the induced-fit model) proposed by Koshland in 1958 (Lexa & Carlson, 2012). In this approach both ligand and protein change conformation during interaction forming a minimum energy protein-ligand complex (Murray *et al*, 1999; Weikl & von Deuster, 2009). Flexible docking increases the conformational search space for full protein flexibility. Flexible docking can be done partially in which case only a few residues in the receptor protein are selected to be flexible, or full flexibility (Morris *et al*. 2009) whereas the whole protein receptor is set to be flexible. Flexible docking has been proven to perform effectively (Bursulaya *et al*. 2003; Lexa & Carlson, 2012). Flexible docking however efficient, is computationally and time expensive than rigid docking, and is further complicated by cross docking difficulties when docking ligands from different ligand-receptor complexes thus, impractical in some instances e.g. when screening large compound libraries (Lexa & Carlson, 2012). There are many docking algorithms but only AutoDock4 and its applications will be discussed.

3.1.3 AutoDock4 as a docking tool

AutoDock4.2 is an automated scheme that performs conformational search for a given ligand to a particular receptor molecule. It is made of various suites which include AutoDock, AutoGrid and works well with AutoDock Tools (Morris *et al*. 2009). AutoGrid pre calculates the 3D grid of interaction from the receptor molecule. Each grid point is probed for interaction energy and the calculated energy values stored. The grid energies are used during the actual docking for rapid search of interacting energies for the ligand. AutoDock then does the actual Docking simulations (Morris *et al*. 2009). Docking allows for the analysis of how small molecules such as enzyme substrates or drug candidates bind to a receptor molecule of known three dimensional structures. It does so by simulations that explores spatial degrees of freedom obtained from rotational, translational and torsional degrees of freedom and can successfully reproduce crystallographically determined positions of ligands (Goodsell *et al*. 1996). AutoDock is carried out in three stages namely grid calculation, docking simulation and docking analysis.

3.1.3.1 Grids and grid maps

This is performed by a sub program of the AutoDock4 program, AutoGrid, which calculates the interaction energies based on the receptor macromolecule in this case FP-2, FP-3 and homolog model structures. The result is a three dimensional grid that is built surrounding the coordinates for the protein target. A probe atom is used to go through the grid points with the interaction energies calculated and stored separately. The grid provides a search table where a fast evaluation of the interaction energies can be done. Tables containing dispersion/repulsion energy and hydrogen bonding energy are calculated and stored separately as well. During simulation, the electrostatic interaction energy of the ligand is calculated as the product of local values from the grid and partial charge on the atoms (Morris *et al*. 2009).

3.1.3.2 Docking simulation

Docking simulations include the actual search for the best ligand conformation bound to the receptor macromolecule. The ligand explores six spatial degrees of freedom, rotation and translation and translational degrees of freedom (Goodsell, *et al*. 1996). There are a number of algorithms used to search for the best ligand conformational space which include; simulated annealing (SA), Monte Carlo method and Genetic Algorithm (GA) (Teodoro *et al*. 2001). Autodock4 uses the GA which performs global search for the best ligand conformation bound to the receptor but a more robust method, the Lamarckian Genetic Algorithm (LGA) has been implemented in the latest versions. LGA is a variant of the genetic algorithm which is one of the most used methods used to find docked ligands with calculated binding energies and constants (Morris *et al*. 1998). It is based in natural genetics of biological evolution where the ligand is represented as a set of values describing translation, orientation and conformation i.e. state variables with respect to the protein. Each state variable is conceived as a gene and the ligand state corresponds to the genotype. The atomic coordinates are the phenotypes.

The genetic algorithm (GA), as implemented in AutoDock4 has a chromosome which is comprised by a string of real valued genes, a Cartesian plane with 3 coordinates for ligand translation and a variable to define a quaternion specifying ligand orientations, one real value for each ligand torsion and in the order defined by AutoTors which is a preparatory program that defines the order of genes that encode torsion angles (Morris *et al*. 1998). The GA starts with a

specified population which refers to the number of random pairs of individuals (ligand state) to be mated to produce a generation in process called crossover. The term generations is used to refer to the number of energy interactions of the ligands (fitness) with the protein. In order to choose the best clusters, evaluations are applied which are a set of conditions applied to the selected generation to choose the best clusters of ligand conformations with the least binding energy. The number of energy evaluations needed for a docking depends on the number of torsions in the ligand (and protein receptor, if it is flexible). GA operates in a fashion similar to Darwinian evolution coupled with Mendelian genetics in which phenotypes are obtained from genotypes. However, the Lamarckian Genetic algorithm works in an inverse manner (Morris *et al*. 1998).

LGA combines global and local search algorithms to find mapped genotypes from the phenotype. The phenotypes are molecular transformations of molecule genotypic state variables into corresponding set of atomic coordinates (phenotypes) which are searched for using the global algorithm. A local search is then applied on the phenotypes to searches for the local minimum and an inverse function applied to convert phenotypes to corresponding genotypes, ligand state (Morris *et al*. 2009)

3.1.3.3 Energy scoring function

The importance of energy scoring in molecular docking cannot be overlooked. Popular traditional methods, used for energy evaluations in computational molecular biology/chemistry based on molecular dynamics and molecular mechanics include; AMBER (Assisted Model Building with Energy Refinement), CHARMM (Chemistry at HARvard Molecular Mechanics), among others(Morris *et al*. 1998). These methods calculate the energy scores of the ligand using the force field approach where factors such as dispersion, repulsion, hydrogen bonding, electrostatics, and deviation from the proper bond length and angles. However, these methods have limited usability to energy calculations between molecules with only a single atom charge performing poorly when ranking free energy of molecules that differing by more than a few atoms and are computationally expensive as well.

AutoDock4.2 applies a semi empirical free energy force field function which is able to model free energy of binding of the ligand and in the process adding entropic and pair wise atomic terms; Van der Waals force, hydrogen bonds, electrostatic potential, conformation, torsion and solvation to the molecular mechanics equation. Solvation or hydration in this case is implicitly implied and in AutoDock4.2, a charge based solvation method is used (Huey *et al*. 2007). The empirical free energy force field provides functional definition for ligand-protein intermolecular interactions (Huey *et al*. 2007). After scoring the docked ligands in their best conformations, there has to be an effective means to evaluate and rank them accurately. Depending on the number of evaluations and GA runs set, the ligands are ranked in clusters of a given root mean square deviation (RMSD). Free energy of binding scores is calculated in kcal/mol with favorable energy being on the negative number scale. This energy is used to calculate the inhibition constant (Ki) which is dependent on the free energy of binding. A more positive free energy of binding corresponds to low inhibition constants and vice versa (Morris *et al*. 1998).

3.2 Methodology

The steps below outline the procedure followed for molecular docking. Figure 3.1 below gives an overview of the docking procedure.

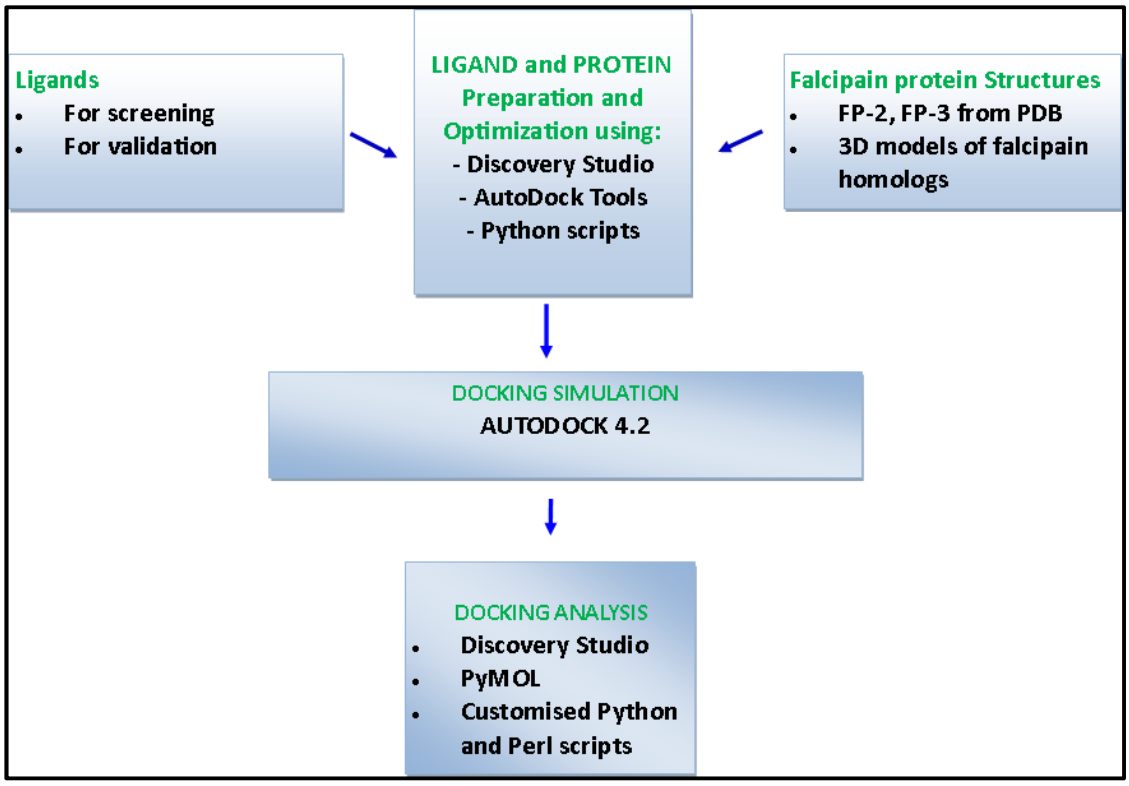


Figure 3.1: Summary of molecular docking methodology. Calculated 3D models and FP-2 and FP-3 protein structures and non-peptide inhibitors from literature and selected non-peptide compounds from natural sources were used.

3.2.1 Data preparation for molecular docking

In this study, crystal structures of FP-2 (2OUL), FP-3 (3BWK), Cathepsin L (3OF8), Cathepsin K (3OVZ), Cathepsin S (1NPZ) as well as 3D models FP *Plasmodium* homologs were used for molecular docking. Inhibitor ligands were recreated in pdb file format and protein (receptor) molecules prepared by removing crystallographic waters and other heteroatoms bound to the protein using Accelrys Discovery Studio 3.1 (Accelrys Software Inc. Discovery Studio Modeling Environment, Release 3.1, San Diego: Accelrys Software Inc. 2011).

The ligands used were divided into three groups; peptide-based inhibitors, non-peptide inhibitors and selected compounds from natural sources (Table 3.1).

The peptide and non-peptide inhibitors were used to evaluate the accuracy of the docking method used as well as deduce receptor-inhibitor binding interactions. There is no crystal structure of FPs with a non-peptide hence docking results were compared to respective experimental data.

Ligand type	**Inhibitor name**	**Structure (follow link)**	**Reference**
Peptide based	E-64	Appendix 2 A	(Kerr *et al*. 2009)
	leupeptin		(Kerr *et al*. 2009)
	Mu-Leu-homoPhe-VsPH		(Kerr *et al*. 2009b)
Peptidomimetic	Pyridone scaffold	Appendix 2 A	(Verissimo *et al*. 2008)
Non-peptide	2-cyanopyrimodine derivatives	Appendix 2 A	(Coterón *et al*. 2010)
	Isoquinoline derivatives	Appendix 2 A	(Batra *et al*. 2003)
	Chalcones	Appendix 2 A	(Liu *et al*. 2001; Li *et al*. 1995;Domínguez *et al*. 2005)
	Thiosemicarbazones derivatives	Appendix 2 A	(Chipeleme *et al*. 2007; Chiyanzu *et al*. 2003; Greenbaum *et al*. 2004)

Table 3.1: List of ligands used for evaluating docking method accuracy as well as elucidate protein-ligand/inhibitor interactions of calculated FP 3D models. The link provides full chemical structure description of the ligands.

3.2.1.1 Natural compounds for screening

A small compound library of natural compounds of South African natural sources was generated (*Supplementary_data/Chapter3/Screened Natural Compounds.docx*). This library consisted of24 selected compounds obtained from literature (Davies-coleman, 2005; Davies-coleman & Beukes, 2004). Among these compounds it was not obligatory that they were previously tested for antimalarial activity.

The main aim was to provide a basis for detecting favorably binding compounds which, with modifications could be used as lead compounds to generate an inhibitor library for further screening. The same could also be achieved through database search of natural compounds with the selected functional group. For this study, the ZINC database was to be used.

3.2.1.2 ZINC database search

Compounds with high predicted binding affinities to FP-2 and FP-3 and related *Plasmodium* homologs were used to search for other similar compounds available in existing databases. For this study, the ZINC free database for commercially available compound for virtual screening was used (http://zinc.docking.org/). ZINC has close to a million molecules, which are assigned biological protonation states, and is completely annotated has vendor purchasing information. The ZINC database was chosen because molecules are ready for docking using popular methods such as AutoDock4.2 (Irwin & Shoichet, 2006). The compounds were first converted to SMILES (Simplified Molecular Input Line Entry System) chemical notation using BABEL installed in the Linux cluster. The SMILE was then used as a query in a BLAST-like manner to search the ZINC database for compounds with similar basic chemical composition. Molecules in ZINC database were downloaded in mol2 file format and converted to pdb using BABEL (http://cds.dl.ac.uk/cds/interface_and_utilities/babel.html), inter file format converter program available in the local cluster.

3.2.2 Ligand and protein protonation

Protein crystal structures and 3D *Plasmodium* FP homologs as well as the ligands were converted into rigid (pdbqt) conformations using python scripts provided by AutoDock4.2 Tools. In this process, the protein and ligands are protonated by merging all hydrogens and adding polar hydrogens which is used to implicitly introduce solvation. This was particularly important in order to generate an optimized protein structure. This was followed by calculation of Gasteiger charges and assigning atom types (AutoDock4.2). In the case of ligands (inhibitor) the torsions were automatically set in the process (Morris *et al*. 2009). This step was automated by using

customised python scripts (*Supplementary_data/Chapter3/Docking_scripts*) modified from those written by Dr. Kevin Lobb, which were run on a Linux based cluster.

3.2.3 Grid calculation and docking parameter file preparation

3.2.3.1 Non-peptide ligands

The grids for the receptor molecules were calculated using AutoGrid4.2 to determine a 3D grid of interaction energies based on the macromolecular (receptor) coordinates (Goodsell *et al*. 1996). Grids from each atom types from the ligands and those for electrostatic interactions were chosen sufficiently large enough not just to cover the active site but also important surrounding areas. The grid points were therefore set at 70, 70, 65Å for all the ligands and a grid spacing of 0.3472Å. The cubic grids were centered on the Cys42 of FP-2 and corresponding positions in all homologs. The grid box spanned an area of residues around a 12Å radius (Desai *et al*. 2004). The grid parameter file also listed the interacting receptor and ligand atom types and their corresponding interaction maps. This procedure was automated for each docking experiment using a customised python script implementing the AutoGrid4 package of AutoDock4 (*supplementary_data/autogpf_creator.py*), and run on the cluster.

3.2.3.2 Peptide based ligands

Some of the known natural small molecule inhibitors of cysteine proteases for instance E-64 for FP-2, Leupeptin for FP-3, vinyl sulfone and peptidyl based inhibitors bind to the active site by forming a covalent bond with the sulfur atom of the catalytic cysteine (Kerr *et al*. 2009). The normal docking simulation routines are able to address hydrogen bonding, Van der Waals, hydrophobic interactions and solvation effects but not the covalent type bonding (Beavers *et al*. 2010). To overcome this problem, a special grid based approach in AutoDock4.2 was applied. A special map for the site attaching to the ligand covalently was calculated with a specified atom that forms the covalent bond with the receptor atom labeled as Z. A Gaussian function with zero energy at the binding atoms were calculated and high energy penalties assigned to surrounding areas (Morris *et al*. 2009). The covalent grid maps were created using a customised script (*Supplementary_data/Chapter3/Docking_scripts/autogpfcov_creator.py*).

3.2.4 Docking simulations

AutoDock4.2was used to perform the actual docking simulations using automated customised python scripts (*supplementary_data/autodpf_creator.py and final_autodocker.py*) run on a Linux based cluster with Autodock4.2 and AutoDock4.2 Tools (ADT) installed. A docking parameter file (DPF), was created for each ligand and the receptor proteins. In the using DPF was specified the parameters specific for each ligand to be docked with the different receptors. The docking parameters were set as follows: AutoDock4.2GA was selected to search for the conformational space in which the Lamarckian Genetic Algorithm variant was used for receptor-ligand conformational search. The population size was set at 150, GA evaluations at 450000 and the number of GA generations was 27000. Cluster analysis for docked results was done using a root mean square (RMS) tolerance of 2.0Å. The GA run was set at 100. All files containing the docking parameters (DPF) were saved accordingly in a designated folder. The docking experiment was carried out using the AutoDock4.2 program in the cluster.

3.2.5 Docking analysis

Docking analysis was performed using ADT implemented in customised python scripts (*supplementary_data/autodlg_analyzer.py and lowest_ligand.py*). The scripts were used to extract the best ligand conformations from the docking log file (DLG) as well as their corresponding estimated free energy of biding and estimated inhibition constant. Important aspects considered were: inhibitor interaction with enzyme active site residues, free binding energies, relative entropy, estimated inhibitor constant (Ki) and the energy values for the protein ligand complex. Ligands in the best scored conformation according to the energy function were converted to pdb format from the rigid pdbqt and analysed visually to identify interactions of specific interest using Accelrys Discovery Studio 3.2. Energy graphs and inhibition constant plots were created using the gnuplot program and Microsoft Excel spreadsheets to summarize the entire docking results. Docking validation was done using known FP-2 and FP-3 inhibitors exposed to the same parameters.

3.3 Results and Discussion

Molecular docking was performed in three categories. These categories were docking validation using known peptide inhibitors, non-peptide inhibitors, and docking for screening selected natural products. Below is each of these categories in details.

3.3.1 Docking validation

Crystal structures of FP-2 and FP-3 in complex with inhibitors were selected for validation. The structures were 3BPF (FP-2), 3BPM (FP-2), 3BWK (FP-3), in complex with peptide inhibitors E-64, leupeptin and Morpholine-Leucine-homoPhenyl-Vinylsulfone, respectively. The ligands were removed and re-docked to validate the docking protocol. These ligands form a covalent bond between the carbonyl carbon of the inhibitor and the catalytic Cys of FP-2 and FP-3. For the covalently bonding ligands, the covalent docking procedure described in section 3.2.4.2 was used. Unfortunately, the method was unable to replicate similar binding modes which could be explained by AutoDock4.2 having the inability to simulate covalent bonds in an explicit manner (Ouyang *et al*. 2012). However, one result (Figure 3.2) had the inhibitor posed in a similar conformation to Mu-Leu-homoPhe-VSPh docked to FP-2 but the carbonyl carbon was not close enough for covalent bond formation.

Despite this limitation, the method was however able to predict the Leu side chain docked at the S2 pocket the Phe interacting with S3 and the vinyl functional group positioned near the catalytic cysteine. The main aim of this study was focused on non-peptides but currently, there is no crystal structure of FP in complex with a non-peptide inhibitor thus, non-peptide inhibitors previously experimented on, were selected for further validation of the docking protocol.

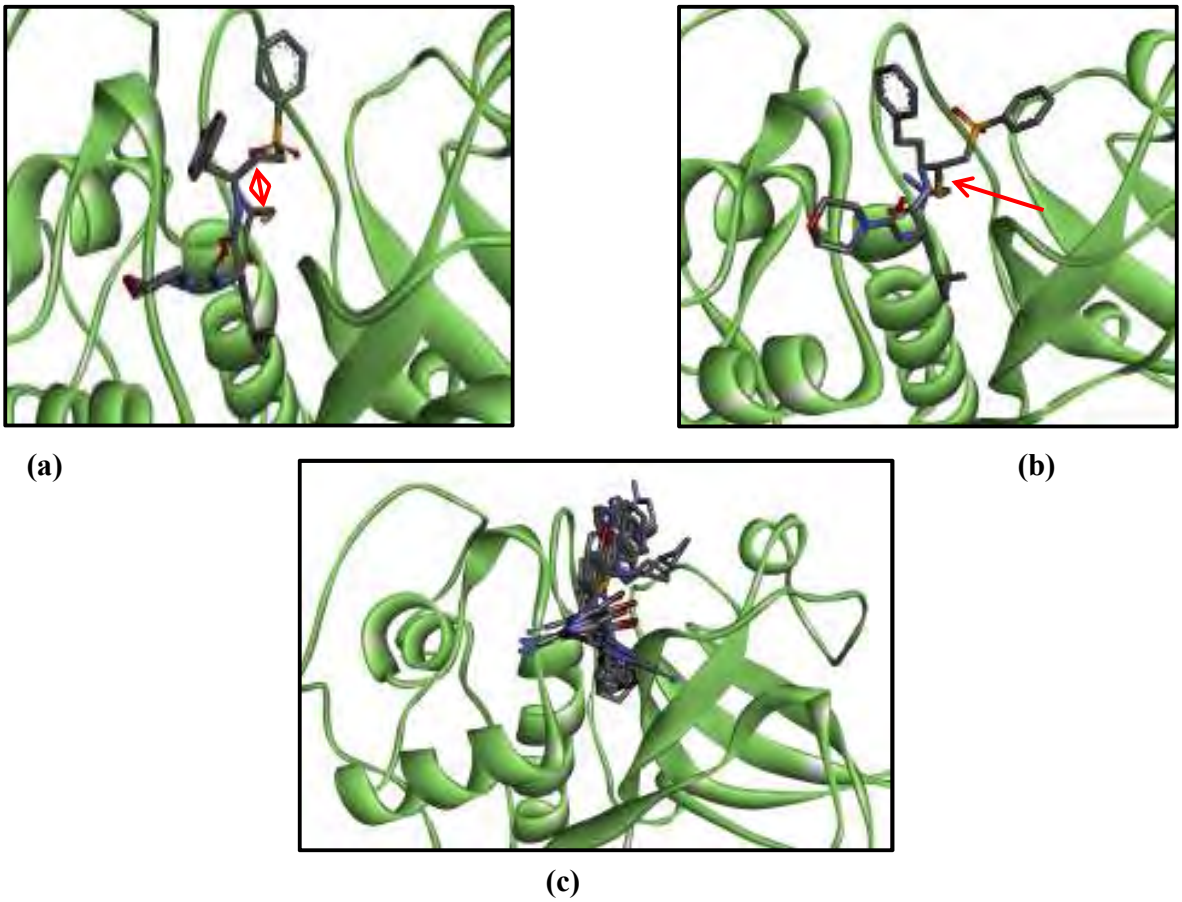


Figure 3.2: Mu-Leu-homoPhe-VsPH docked to FP-2. The arrow shows the carbonyl carbon and sulfur atom that involved in covalent bonding. b) Mu-Leu-homoPhe-VsPH docked to FP-3 (3BWK). The arrow point to the covalent bond linkage between the inhibitor and FP-3 and c) FP-2 with docked 2-cyanopyrimidine inhibitors

3.3.2 Non-peptide inhibitors

A total of 22 non-peptide inhibitors; isoquinolines, chalcones, thiosemicarbazones, pyridone based peptidomimetic and 2-cyanopyrimidine derivatives were docked to twelve cysteine proteases. The four peptide based inhibitors (Appendix 2A) were also docked as previously described. Out of these twelve, three cathepsin-L (3OF8), Cathepsin-S (1NPZ) and cathepsin-K (3OVZ) were human cysteine proteases which were used to assess inhibitor – protein interaction characteristics which could provide significant information with regard to selectivity. Two were crystal structures of FP-2 and FP-3 (2OUL and 3BWK respectively) and the remaining seven were the 3D models of FP-2 and FP-3 *Plasmodium* homologs.

From Figure 3.2 c, the non-peptides with the lowest free energy of binding were consistent in their binding conformations with FP-2 which proved the method used (rigid docking) quite accurate. Due to time constraints, the computational parameters were optimized (section 3.2.3) thus the whole conformational space was not completely sampled but result obtained were in agreement with the conformation observed in figure 3.2b.

From Figure 3.3 and 3.4 (a) and (b), 2-cyanopyrimidines had the lowest free energy of binding and were consistent across all *Plasmodium* homologs apart from KP-2, and VP-3. However, the estimated inhibition constants were all in the nanomolar range as was observed with FP-2 and FP-3 in the previous experimental data (Coterón *et al*. 2010).

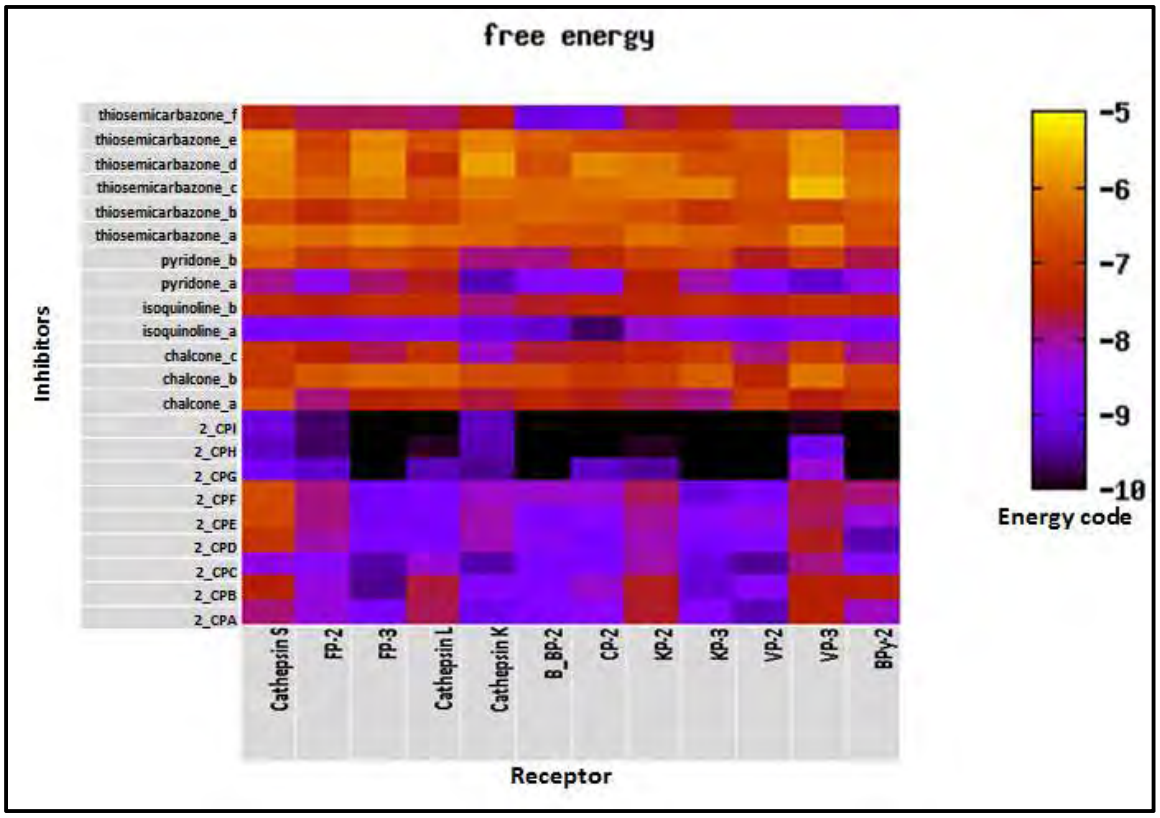


Figure 3.3: Estimated free energy of binding of best docked non-peptide inhibitors against FP-2, FP-3 and their *Plasmodium* homolog 3D models. The energy code ranges from high (yellow) to low (black). Inhibitors were renamed for study convenience (Appendix 2A).

Thiosemicarbazones and chalcones had high free energy of binding which implied lower estimated binding affinities and consequently, higher inhibition constants. These results were not in complete agreement with results obtained from experimental enzyme assays (Chiyanzu *et al*. 2003; Greenbaum *et al*. 2004; Liu *et al*. 2001). Further docking analysis might improve the results since only a small docking conformation space was searched.

The two Isoquinolines derivatives docked showed varied binding with one of them (Figure 3.3, Isoquinoline derivative, a) showing considerably good free energy of binding energies and subsequently binding affinity across all homolog proteins (Figure 3.3) which was comparable to earlier docking and experimental results were the inhibitor recorded inhibition at 3 μM against FP-2 *in vitro* (Batra *et al*. 2003). This particular result (Isoquinoline derivative (a)), was good in the context of *Plasmodium* homologs, with estimated inhibition constants in the nanomolar range. The inclusion of p-benzyloxyphenyl group was associated with the improved activity by docking at the S2 sub site (Batra *et al*. 2003). The almost uniform binding among the human homologs raises concerns of selectivity.

Of the two peptidomimetics based on a pyridone ring scaffold, only one seemed to have good binding energy associated with the addition of a vinyl sulfone (Verissimo *et al*. 2008). However, it had a poor binding against cathepsin L (3OF8) and KP-2. Chalcones in contrast to the previous experimental data (Li *et al*, 1995) had relatively low binding affinities compared to other docked inhibitors. A simililar observation was observed with thiosemicarbazones (Chiyanzu *et al*. 2003; Greenbaum *et al*. 2004). Only one of the thiosemicarbazones (Appendix 2A) had encouraging binding energies.

To analyse the interactions of the calculated 3D FP homologs with the non-peptide inhibitors, the 2-cyanopyrimidine derivative (2_CPI, Appendix 2A) was used. This inhibitor seemed to have almost uniform inhibition of all the Plasmodium FP homologs (Figure 3.3 and 3.4 a and 3.4 b)

3.3.2.1 Protein-inhibitor interactions

From the results obtained, non-peptide inhibitors analysed showed an array of interactions which when summed up resulted to high binding affinities (Appendix 2). Since none of these have been characterized in a crystal structure complex of FPs, biochemical data were used to supplement the results.

The most active compounds were 2-cyanopyrimidine derivatives whose free energy of binding was ranging between, -8 and -10 kcal/mol across all homologs (Figure 3.3). The results were consistent with the experimental data. Compounds with a cyclohexyl or cyclopentyl at the P2 position (Appendix 2A) and pyridinyl-phenyl combinations yielded anti parasitic activity at sub nanomolar and nanomolar levels against FP-2 and FP-3 respectively (Coterón *et al*. 2010). Docking results on FP-2 and FP-3 were quite accurate as was observed from Figure 3.5 where these inhibitors had high binding affinity for FP-3 than FP-2 comparable to the experimental data.

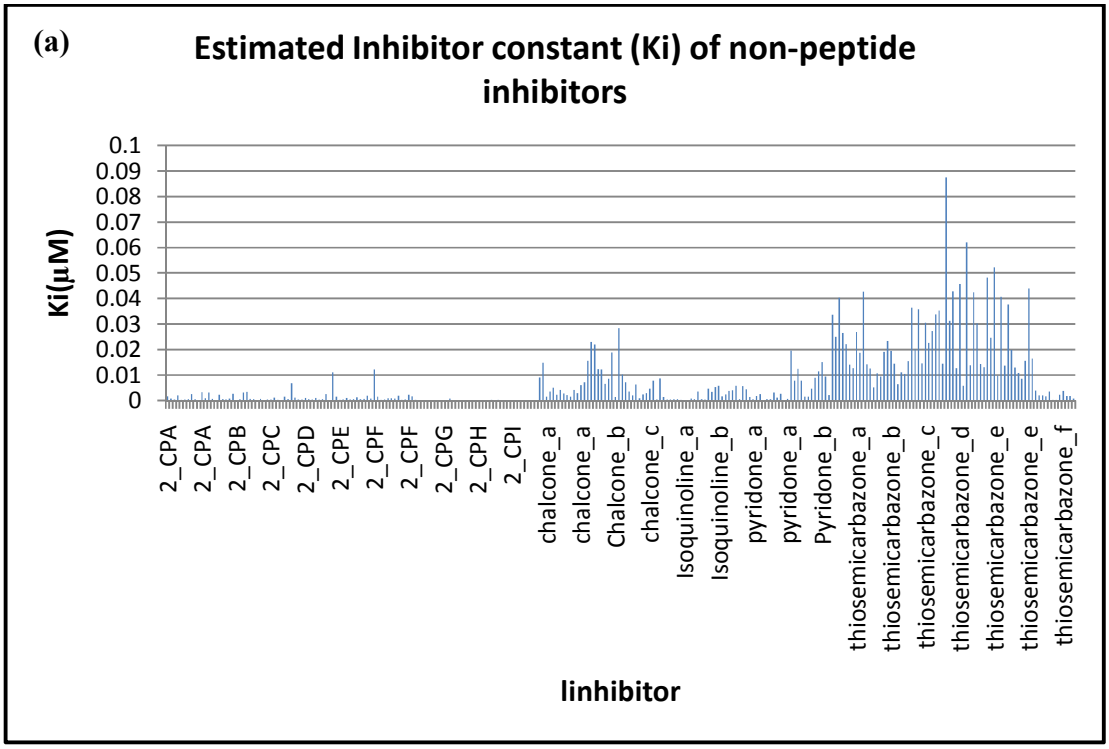
Previous biochemical FP-2 and FP-3 inhibitor enzyme assays, structure activity relationship studies and structural information obtained from co crystalised FP-2 and FP-3 with small molecule inhibitors, showed that S2 of FP-2 and FP-3 preferred inhibitors with a Leu at the P2 position (Kerr *et al*. 2009; Kerr, *et al*. 2009; Na, Shenai, *et al*. 2004; Sijwali *et al*. 2001). However, it has also been shown that inhibitors with Phe such as Z-Phe-Arg-CH2F at the same position could have an equally potent activity (Rosenthal *et al*. 1991; Sijwali *et al*. 2001). 2-cyanopyrimidine inhibitors had either cyclopentyl or cyclohexyl at the P2. 2-cyanopyrimidine derivatives that had cyclohexyl which is similar to phenyl ring of Phe had the highest predicted inhibition constants predicted (Appendix 2B and Figure 3.3) and the best energy scores as was evident in Figures 3.3 consistent with the experimental data. The human cysteine proteases of 1NPZ (cathepsin S), 3OF8 (cathepsin L), and 3OVZ (cathepsin K) seemed to have higher energy scores apart from a few instances which could be good in terms of selectivity. Structural analysis of their interactions with the inhibitors could perhaps explain the observation.

One notable difference was the conformation of the docked inhibitor between FP-3 and the rest of the homologous proteases (Figure 3.6). In FP-3, the cyclohexyl ring seemed to not fit in the S2 pocket. This may support the observation made by Kerr *et al.*, 2009 that the FP-3 S2 is narrower than that of FP-2 and probably limiting access to the FP-3 S2 sub site.

The 2-cyanopyrimidine inhibitors differed with other studies where it is the FP-2 inhibitors that have been reported to perform poorly against FP-3. 2-cyanopyrimidine performed better in FP-3 than FP-2 which was evident in Figure 3.3 and Figure 3.4b and the other *Plasmodium* FP homologs. Often, this observation has been attributed to the binding site size and volume differences (Desai *et al*. 2004). Against the other *Plasmodium* homologs, 2–cyanopyrimidine inhibitors were quite effective showing consistent high energies of binding apart from VP-3 which had poor binding with most of the inhibitors.

Binding affinities were high for 2-cyanopyrimidine with a cyclohexyl and pyridinylphenyl P2 and P3 substituents and isoquinolone derivative (a) as was observed in Figure 3.3. Pyridone derivative (a), a pyridone based peptidomimetic with a vinyl sulfone R group (Appendix 2 A) had binding affinity that was quite uniform against all FP homologs including the human homologs.

To get further insights on the kind of interactions resulting to the predicted low energy scores and low inhibition constants, all the FP-2 and FP-3 *Plasmodium* homologs were analysed visually. Apart from FP-3, all the homologous proteins had the cyclohexyl ring neatly docked at the S2 pocket. The interacting residues in respective homologs as shown in Figures 3.5 – 3.14 were put together in Table 3.2. Specific interactions are further depicted in 2D diagrams cross referenced in (Appendix 2 A). From these analyses, all homologs formed a hydrogen bond between the Gly83 (326) at the S3 sub site of FP-2 and corresponding positions of the *Plasmodium* homolog proteases with the carbonyl or amine of the 2-cyanopyrimidine derivative apart from FP-3, which formed a hydrogen bond between Gln38 (287) of the S1 sub site. Gly83 and Asn173 have previously been shown to form important hydrogen bonds for stabilizing the inhibitor molecules in FP-2 crystal structure complex with a small molecule inhibitor (Kerr *et al*. 2009). FP-3 had few interactions with the S3 sub site which was attributed to its narrow S2 pocket which seemed not to accommodate the cyclohexyl ring perfectly hence pushing the inhibitor slightly upwards and further from S3 residues. Surprisingly 2-cyanopyrimidines were more effective against FP-3 than FP-2, unlike a previous FP-2 and FP-3 inhibitor screening study where FP-3 was found not to always favorably bind some FP-2 inhibitors which was attributed to FP-3 narrow S2 sub site (Desai *et al*. 2006).



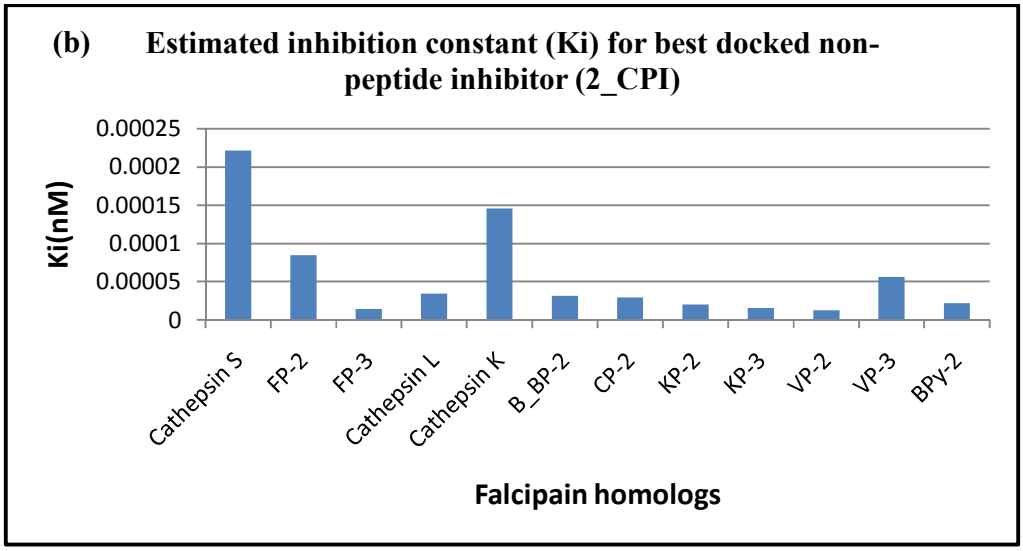


Figure 3.4: a) Inhibition constants (Ki) of docked non-peptides as predicted by AutoDock4.2. Inhibition constants were derived from the estimated free energy of binding. Each inhibitor was docked to all the 12 proteins (Figure 3.3) 2-cyanopyrimidines had the lowest estimated inhibitor constants observed. b) Estimated inhibition constant (Ki) for best docked non-peptide inhibitor, 2_CPI (zoomed in from Fig 3.4a). The inhibitor had higher estimated inhibition constant against cathepsin S and K compared to the *Plasmodium* homologs indicative of possible binding selectivity.

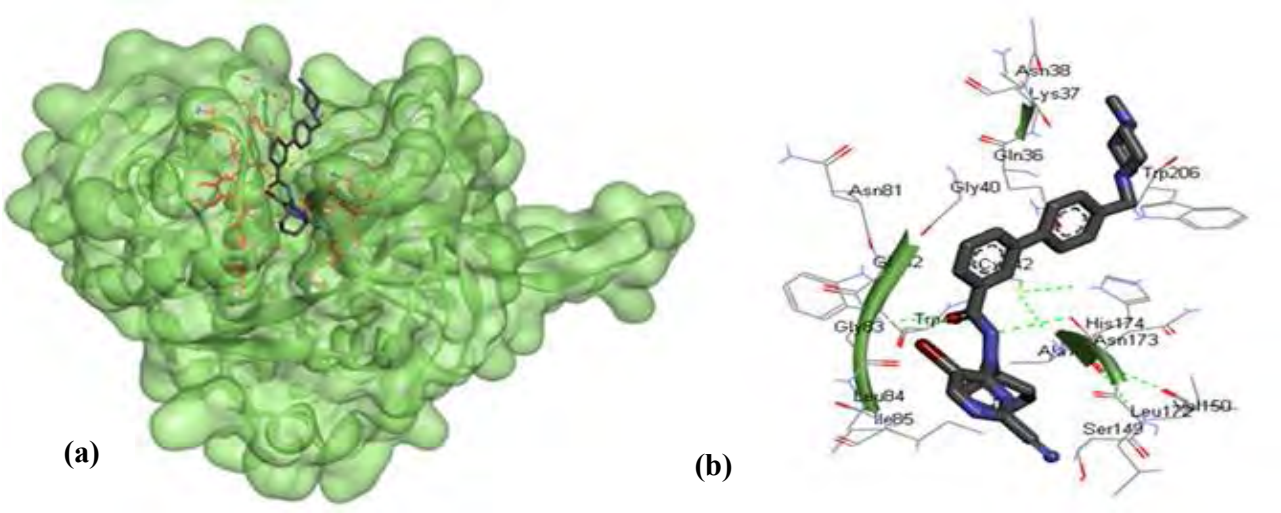


Figure 3.5: FP-2 surface presentation with best docked 2-cyanopyrimidine derivative (2_CPI). Red lines are active site residues. b) FP-2 and ligand binding site atoms interactions.

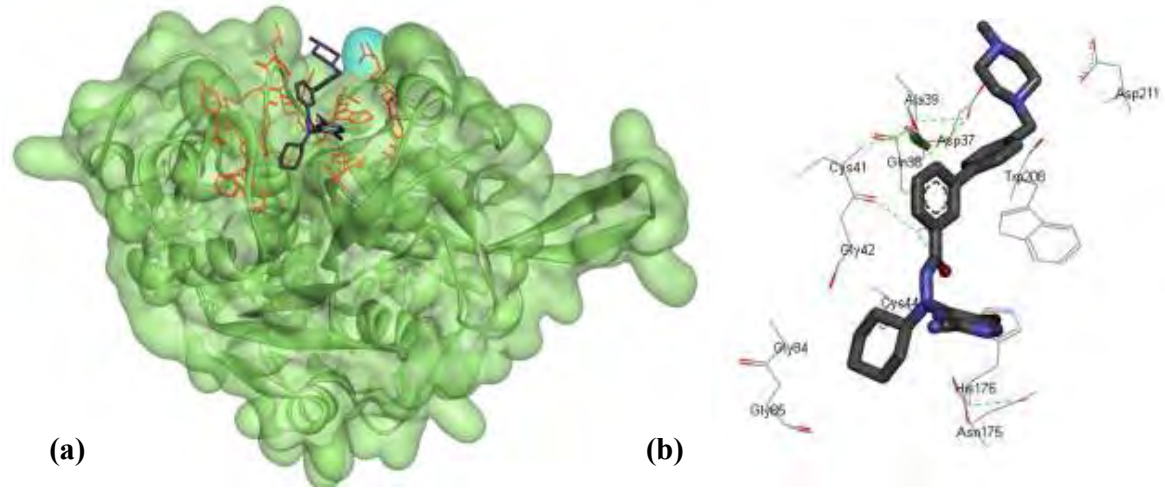


Figure 3.6: a) FP-3 surface presentation with best docked 2-cyanopyrimidine derivative (2_CPI). Red lines are active site residues. b) FP-3 and ligand binding site atoms interactions.

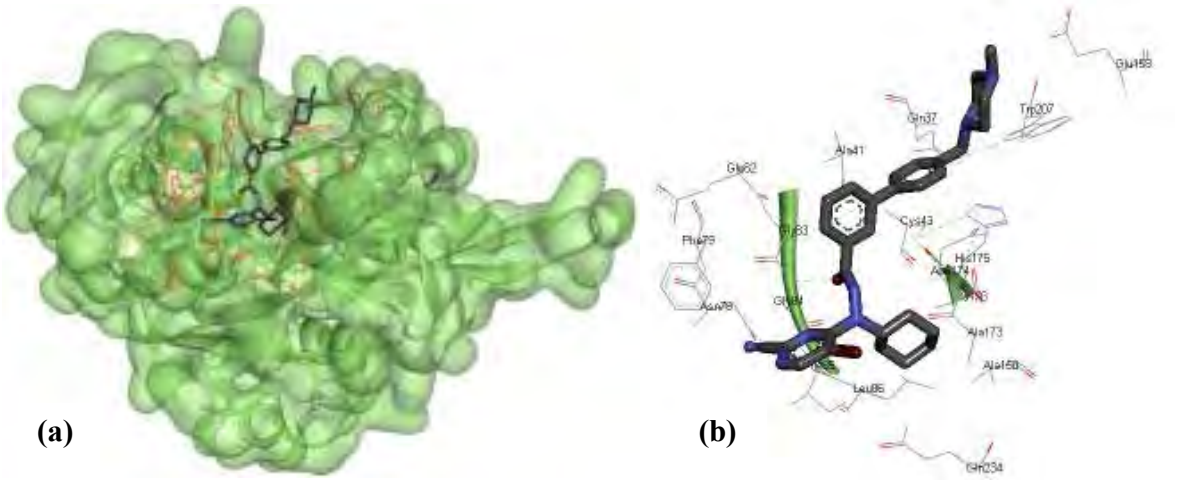


Figure 3.7: a) BP-2 (*P. Berghei*) surface presentation with best docked 2-cyanopyrimidine derivative (2_CPI). Red lines are active site residues. b) BP-2and ligand binding site atoms interactions.

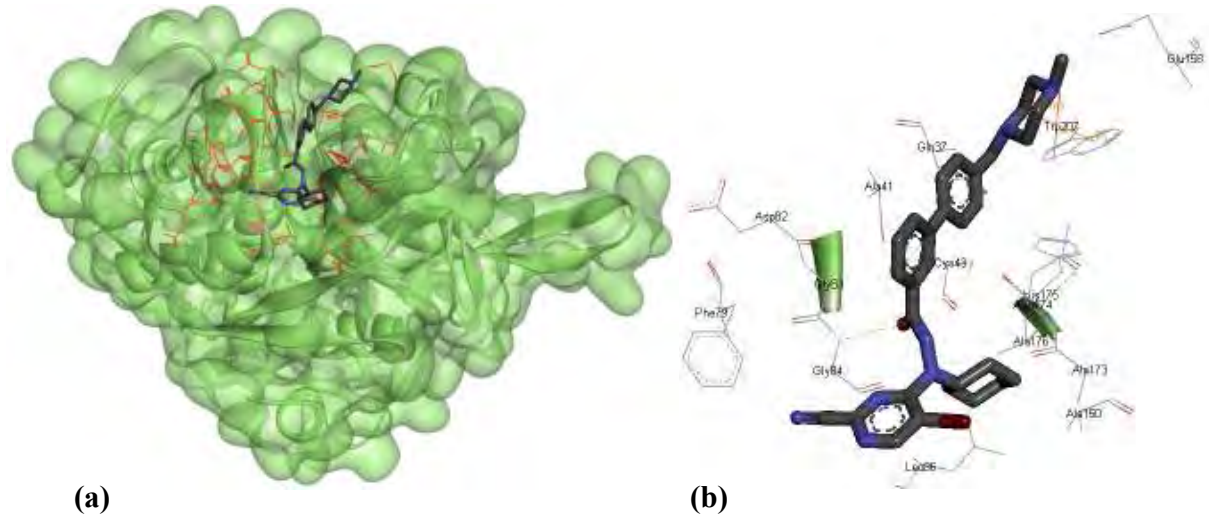


Figure 3.8: a) BP-2 (*P. Yoelii yoelii*) surface presentation with best docked 2-cyanopyrimidine derivative (2_CPI). Red lines are active site residues. b) BP-2 and ligand binding site atoms interactions.

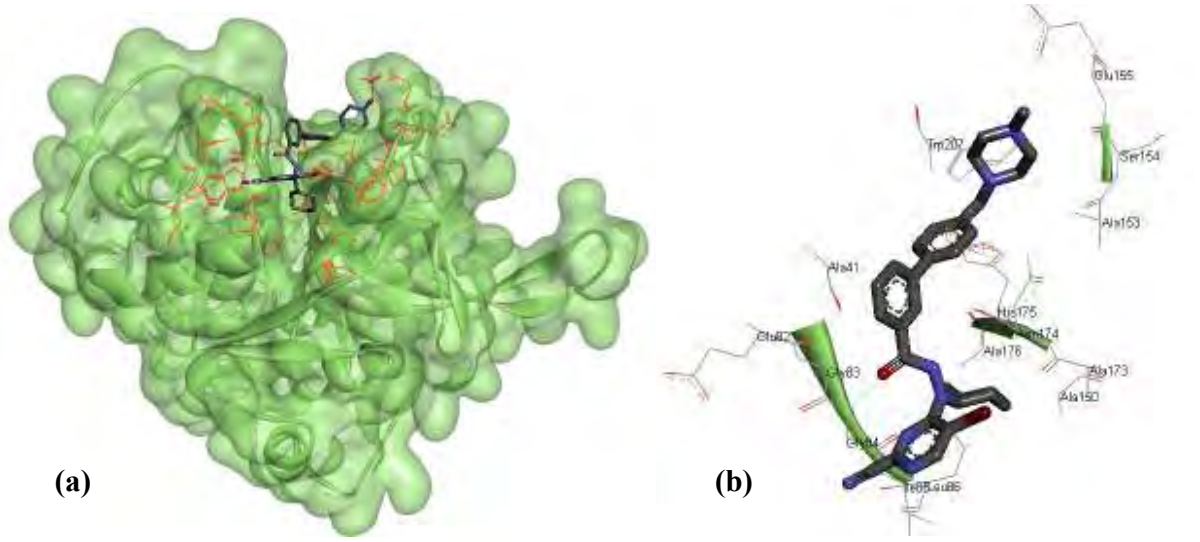


Figure 3.9: a) CP-2 surface presentation with best docked 2-cyanopyrimidine derivative (2_CPI). Red lines are active site residues. b) CP-2 and ligand binding site atoms interactions.

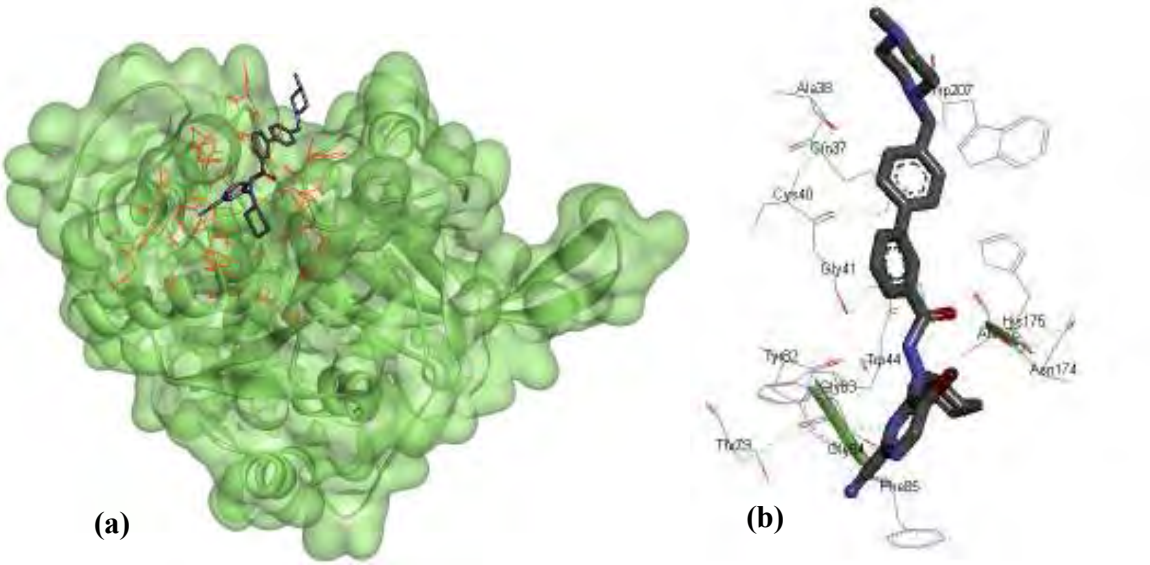


Figure 3.10: a) VP-2 surface presentation with best docked 2-cyanopyrimidine derivative (2_CPI). Red lines are active site residues. b) VP-2 and ligand binding site atoms interactions.

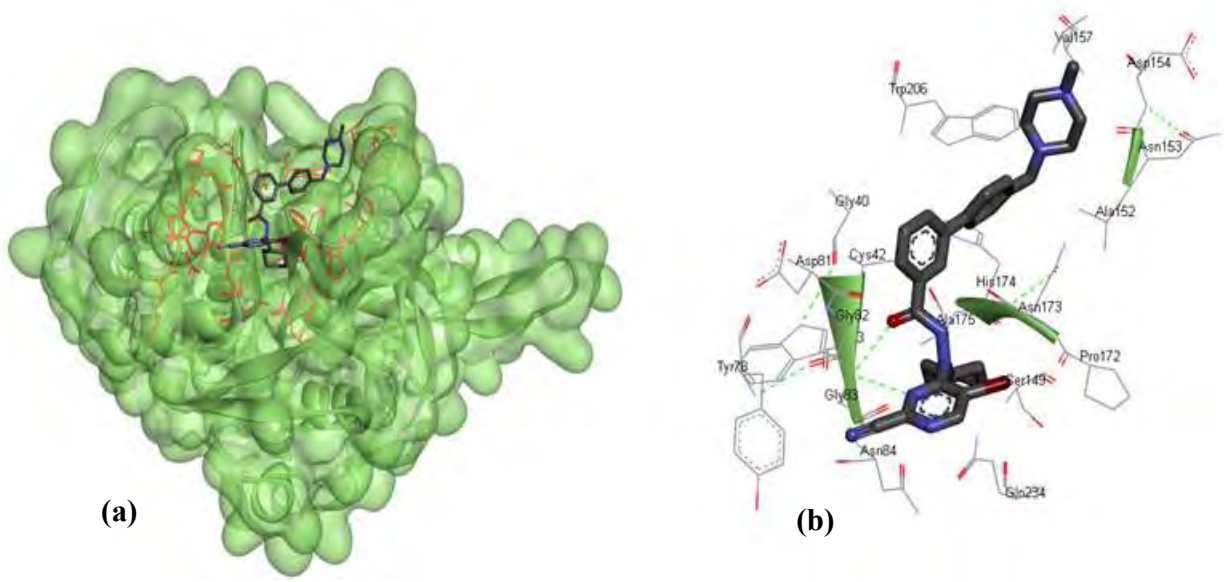


Figure 3.11: a) VP-3 surface presentation with best docked 2-cyanopyrimidine derivative (2_CPI). Red lines are active site residues. b) VP-3and ligand binding site atoms interactions.

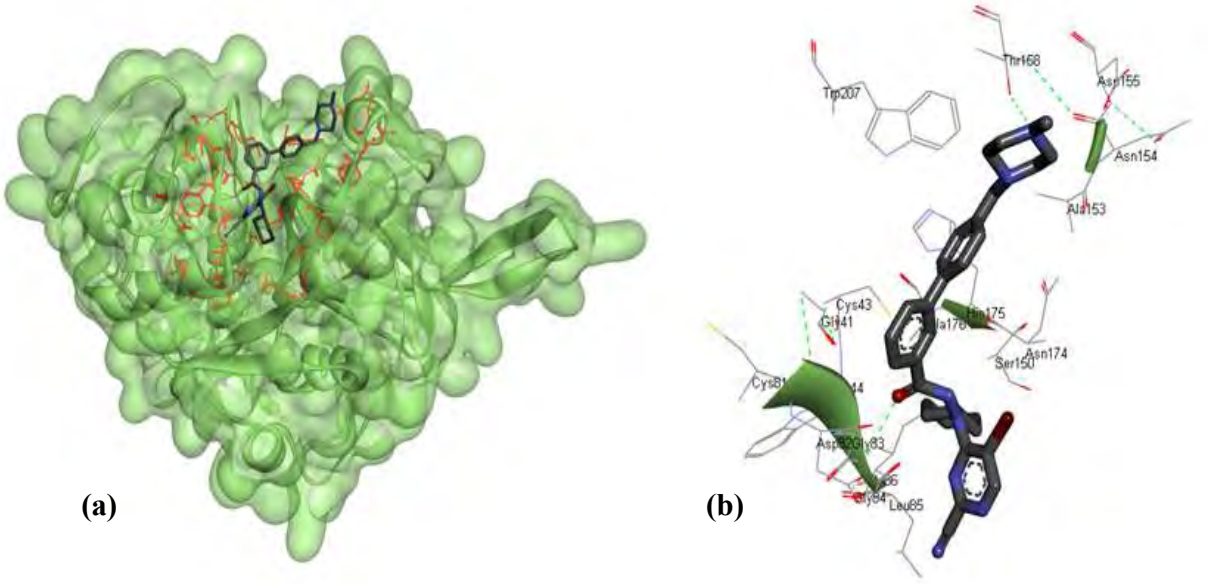


Figure 3.12: a) KP-2 surface presentation with best docked 2-cyanopyrimidine derivative (2_CPI). Red lines are active site residues. b) KP-2 and ligand binding site atoms interactions.

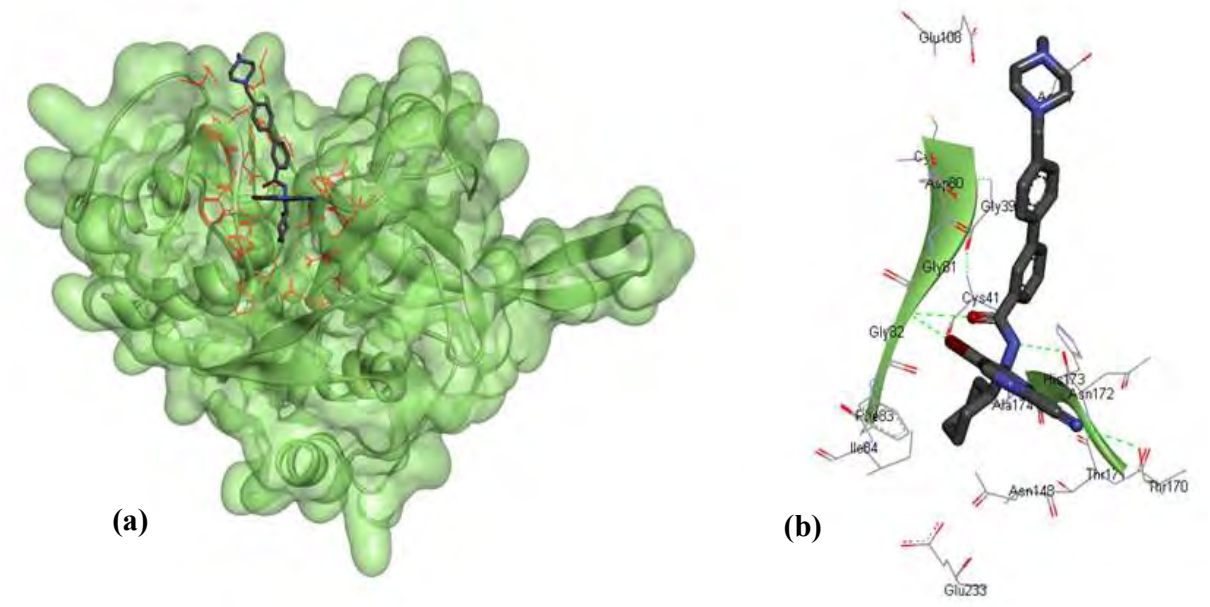


Figure 3.13: a) KP-3 surface presentation with best docked 2-cyanopyrimidine derivative (2_CPI). Red lines are active site residues. b) KP-3 and ligand binding site atoms interactions.

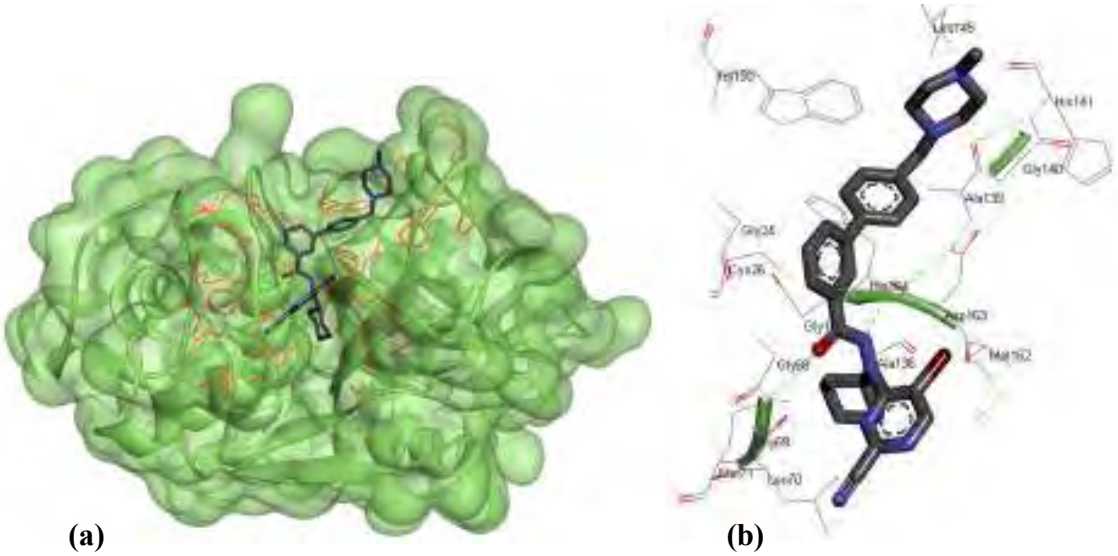


Figure 3.14: a) Cathepsin-L surface presentation with best docked 2-cyanopyrimidine derivative (2_CPI). Red lines are active site residues. b) Cathepsin L and ligand binding site atoms interactions.

Protein/homolog	Catalytic residues	S1	S2	S3	S1'
FP-2	C42-H174	Q36-G40	I85-S149-L84-L172	G83-	W206
FP-3	C44-H176	Q38-G42	-	G84-G85	W208
BP-2 (*P. yoelii*)	C43-H175	Q37-A41	L86- A150-A173	F79-G83-G84	W207
BP-2(*P. berghei*)	C43-H175	Q37-A41	I85-L86-A150-A173-Q234	N78-F79-G83-G84	W207
CP-2	H175	A41	I85-L86-A150-A173	G83-G84	E155-W207
VP-2	C43-H175	G41-Q37-Y82	F85	G83-G84-T79	W207
VP-3	C42-H174	G40	S149-P172-Q234	Y78-G82-G83	A152-N153-V157-W206
KP-2	C43-H175	G41-C81-	L85-I86-A150	G83-G84	A153-N154-T158-W207
KP-3	C41-H173	G39-C79- D80	F83-I84-N148-T171-E233	G81-G82	-
Cathepsin L	C26-H164	G24	L70-M71-M162	G68-G69	A139-G140-L145-W150
Cathepsin K	C25-H162	G23	Y67-M68-L160-L209	N60-G65-G66-D61	-
Cathepsin S	H164	Q19-G23-C66-N67	-	-	F146

Table 3.2: Active site amino acid residues interacting with the best docked ligand (2-cyanopyrimidine derivative, 2_CPI). Interactions observed included; Vander Waals, hydrogen bonding, π-π and π-δ interactions, electrostatic and hydrophobic interactions. FP-3 seemed not to have any interactions with the S2 sub site but had better binding affinity for the inhibitor than FP-2 (Figure 3.3).

Other residues that contributed to polar, electrostatic and Van der Waals interactions included Asn173 (416), Cys42 (285), Trp206 (449) in FP-2. Trp208 (437) and Ala174 (403) formed π and π-δ interactions as observed in CP-2 and BP-2 (*P. yoelii yoelii*.) (Appendix 2, Figure 4.5 and 4.6). CP-2 had a π interaction between the phenyl rings of His175 (405) and one of the aromatic rings in the P3 substituent. BP-2 (*P. yoelii yoelii*) had a π-δ (sigma) interaction between the pyridinyl ring and Trp 207 (437). π stacking is a type of chemical interaction formed when two aromatic rings overlap but are weaker than π-δ interactions (Meyer *et al*. 2003). π-δ interactions are quite common with cyclic compounds. These additional interactions clearly missing in FP-2 could explain the lower free energy of binding of these two proteases compared to FP-2.

From the crystal structure of FP-2 and FP-3 with their respective small peptide inhibitors E-64 and leupeptin, the conserved Gly83 (326)/Gly92 (334) has been shown to hold tight the inhibitor to the main chains of FP-2 and FP-3 respectively (Kerr *et al*. 2009). Similar interactions were observed where the S3 Gly 83 (326) in FP-2 and corresponding positions in the *Plasmodium* homologs, formed hydrogen bonds with the amine or carbonyl groups of the inhibitors. This observation might explain the high conservation at the S3 sub site. The pyridinyl–phenyl substituent was designed to increase interaction with the P3 of FP-2 and FP-3 (Coterón *et al*. 2010). The docking results confirmed this observation in all *Plasmodium* homologs except FP-3.FP-2 formed another hydrogen bond between Asn173 (416) and the amine connecting the P3 substituent to P2 and the inhibitors functional group (Appendix 2 B).

An analysis of one of the closest human FP-2, FP-3 homolog, cathepsin L (3OF8), revealed a similar mode of binding (Figure 3.14). There were two hydrogen bonds, one formed between the carbonyl of the inhibitor with Gly69 (mature domain numbering) and the other between the amide connecting the P3 substituent to the inhibitors functional group with Asp163. In cathepsin K Gly66 formed a hydrogen bond with the amine in the 2-cyanopyrimidine functional group. Cathepsin S S1 sub site residues Gln19 and Gly20 (mature domain numbering) had polar interactions with the nitrile of the 2-cynopyrimidine functional group (Appendix 2 B -11, 12)

There were no explicit covalent bonds observed between the inhibitor and the proteases observed but other types of interactions were observed apart from the already discussed hydrogen bonds (Appendix 2 A). The interactions included Van der Waals forces, hydrophobic and electrostatic interactions. If all these occur in high frequencies, the sum total could imply significant biological activity *in vivo* at levels equitable to the well-known covalent bonds that are formed between peptide based inhibitors and cysteine proteases (Beavers *et al*. 2010; Morris *et al*. 2009). The presence of many aromatic rings in the 2-cyanopyrimidine inhibitors could also explain higher binding affinities. Aromatic rings have been shown to increase molecular recognition in biological complexes, thus attracting great attention in the quest for improved drug design and lead optimization (Meyer *et al*. 2003).

From the above results, the possibility of a broad spectrum inhibitor targeting the *Plasmodium* FP homologs seemed feasible.

3.3.2.2 Effect of active site residue variation, size and volume on inhibitor binding

A comparison was done for the best scoring 2-cyanopyrimidines, 2_CPI (Appendix 2 A), docked to the FP-2 and FP-3 *Plasmodium* and human homologs based on the free energy of binding to establish if residue variations previously observed at the active site could have an effect on inhibitor binding and subsequent inhibition. Although the data presented may not be absolute in terms of exhaustive docking, it could provide substantial information to establish the effects of these variations.

From the free energy of binding energy plot (Figure 3.5), the human FP homologs of cathepsin S (1NPZ), and cathepsin K (3OVZ) seemed to have lower free energy of binding across all the ligands which translate to lower binding affinities. This could be partially explained by their distant homology and high residue variations observed at the S2 sub site (Chapter 2, Table 2.2). Cathepsin L (3OF8) the closest human homolog had varying binding affinities against 2-cyanopyrimidines. 2-cyanopyrimidines with a single aromatic ring in the P3 substituent seemed to bind poorly but improved with the introduction of a cyclohexyl at the P2 position. Among the

FP homologs, only VP-3 and KP-2 had similar binding outcomes. This observation could be significant for improving inhibitor selectivity against the human homologs.

Table 2.2 in Chapter 2 listed residue variations at S1-S3 and S1' sub sites of all retrieved homologs. Previous biochemical analysis of Vinkepain-2 and FP-2 (Rosenthal *et al*. 1996; Rosenthal *et al*. 2002) showed differences which are also reflected from comparative multiple sequence analysis of FP-2 *Plasmodium* homologs (Chapter 2 Table 2.2 and Figure 2.2) as well. Rodent and murine models are used for *in vivo* antimalarial drug experiments (Carlton, 2002; Chan *et al*. 2005; Jambou *et al*. 2011; Stephens *et al*. 2012). Differences previously observed are of concern as they may complicate result interpretation. This is because certain inhibitors effective against FP-2 may not be in these homologs. Results in the energy graph (Figure 3.3) showed consistency of the free energy of binding of among FP-2 and FP-3 and the rodent *Plasmodium* homologs which was also confirmed by the pattern observed in the corresponding inhibition constants (Figure 3.4) which were all ranging in the nanomolar levels. Whether these variations have significant impact on inhibitor binding activity of the *Plasmodium* FP homologs can only be fully described from enzyme assays.

VP-2 had seemed to have poor binding characteristics as evidenced by the lower binding energies by most of the 2-cyanopyrimidine inhibitors this outcome in FP-2 was attributed to the size of its active site which seemed smaller than the other FP-2,-3 *Plasmodium* homologs (Chapter 2, Figure 2.22). However, against 2_CPG, 2_CPH and especially 2_CPI, the inhibitors had comparable free energy of binding to the other *Plasmodium* homologs.

A clear relationship between active site residue composition, size and volume as well as the ligand and chemical composition seemed to affect binding affinity. For instance, the smaller 2-cyanopyrimidine inhibitors, 2_CPA-2CPF, which also had fewer chemically reactive groups or atoms in their structures had lower binding affinities across all homologs compared to the larger 2-cyanopyrimidines, 2_CPG-I which had additional aromatic and pyridinyl rings that consequently increased the number of intermolecular interactions hence higher binding affinities (Figure 3.5). Similar kinds of observation involving protein and ligand volumes and ligand chemical compositions have been achieved elsewhere (Goh & Sim, 2005; Saranya & Selvaraj, 2009; Saranya & Selvaraj, 2011).

3.3.3 Screened natural compounds

Figure 3.15 summarizes the free energies of binding obtained from the screened small library of 24 selected natural compounds compiled from literature (Davies-Coleman & Beukes, 2004). The majority of the compounds had high free energy of binding which was consistent across all homologs. Surprisingly, one compound, 5α-pregna-1,20-dien-3-one (Figure 3.17 and 3.18)a pregnadiene sterol which is a xenicane diterpene obtained from *Capnella thyrsoidea* found in soft corals, had lower binding energy across all homologs with a pattern consistent with the non-peptide inhibitors docking results in Figure 3.3. This compound was not previously tested for antimalarial activity but was shown to stimulate superoxide production in rabbit cell neutrophils which was attributed to the compounds cytotoxicity (Davies-coleman & Beukes, 2004). Whether the compound can inhibit FPs and express cytotoxicity to *Plasmodium* parasite is yet to be determined.

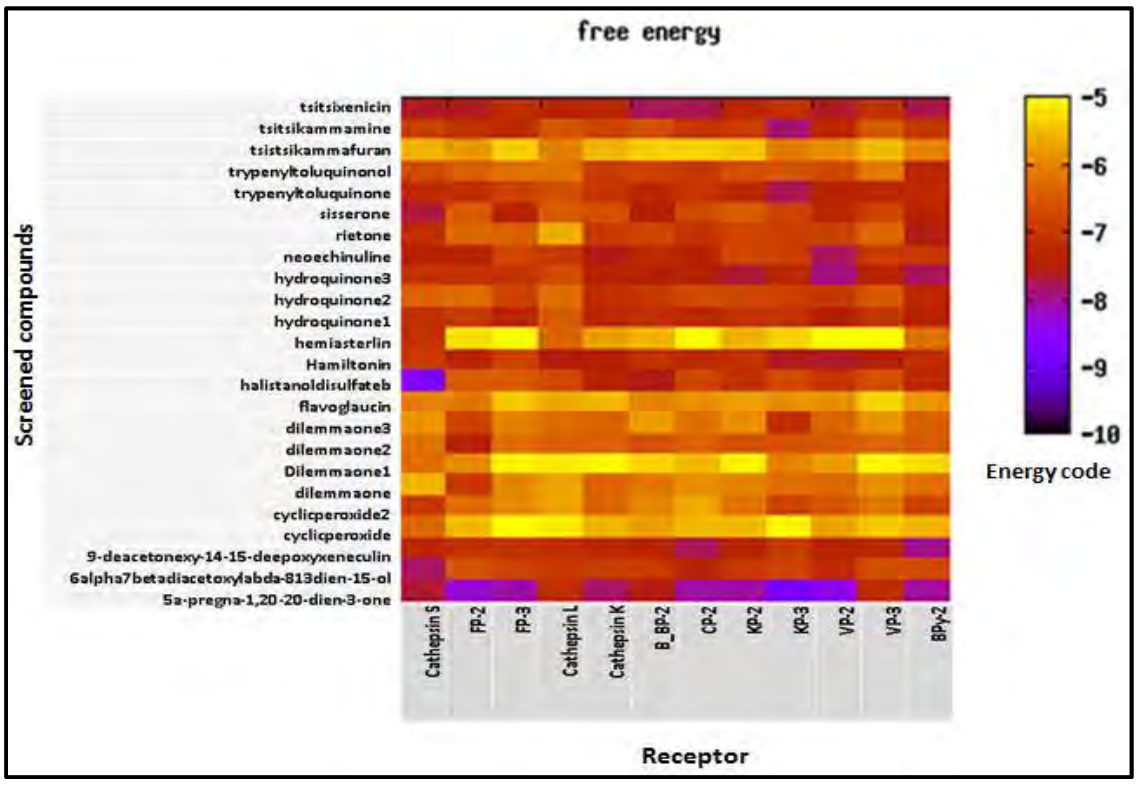


Figure 3.15: Estimated free energy of binding map for screened natural compounds. Energy scores are colored from the highest (yellow) to the lowest (black). Low energy scores correspond to high binding affinities. Only one compound (5α-pregna-1,20-dien-3-one) had low binding energy.

5α-pregna-1,20-dien-3-one had high predicted binding affinities to most of the FP *Plasmodium* homologs except for VP-3. FP-2, KP-3 and VP-2 had the highest estimated binding affinities with inhibition constants in the nanomolar range (Figure 3.16). The compound had low binding affinities to the human homolog of cathepsin S and L but had higher affinity for cathepsin K. This observation was deemed significant since selectivity of FP inhibitor still remains to be a problem (Ettari *et al*. 2009).

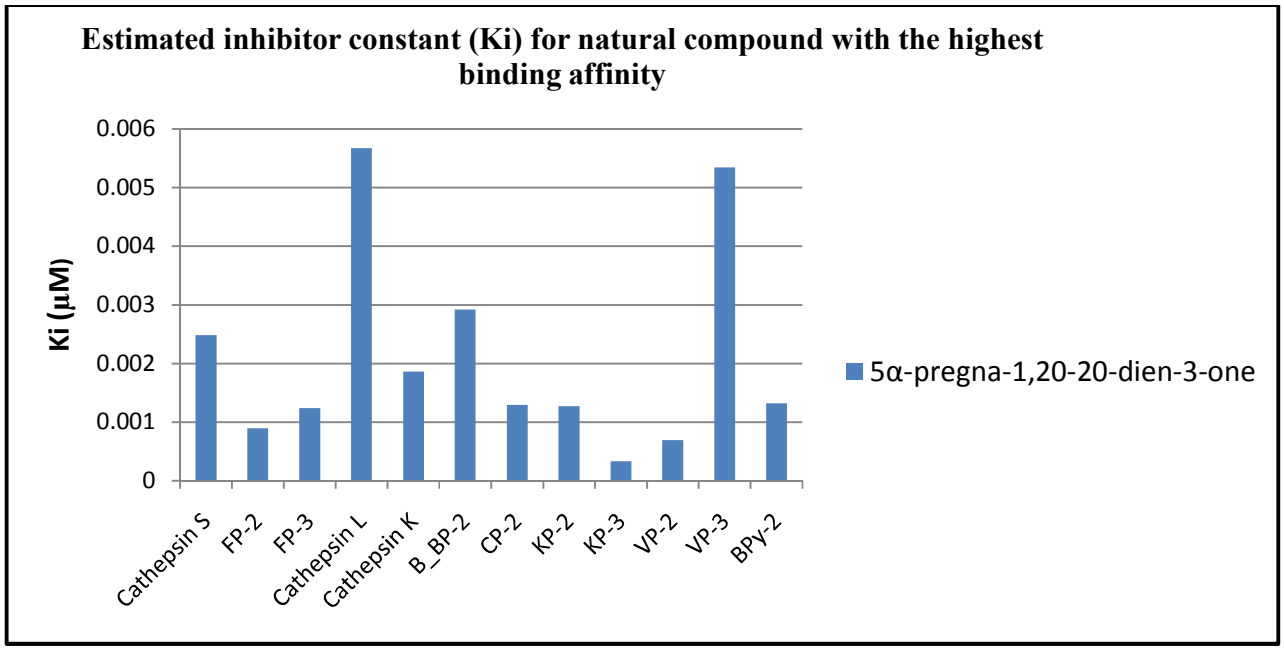


Figure 3.16: Inhibitor constant of 5α-pregna-1, 20-dien-3-one against all FP homologs. The compound had estimated inhibitor constant at nanomolar range against FP-2, KP-3 and VP-2. The compound had low binding affinities for cathepsin S, cathepsin L and VP-3.

The compound was docked perfectly to the S2 sub site which could perhaps explain the low free energy of binding scores. However, on further analysis of the interactions involved, only residues around the S2 sub site were involve as the compound lacked side chains to interact with the S1, S3 and S1' sub sites (Figure 3.17a and Appendix 2B -13). This result was however encouraging considering the small sample size hence. The high binding affinity observed with this compound was attributed to the hydrophobic interactions with the S2 sub site of *Plasmodium* FP homologs.

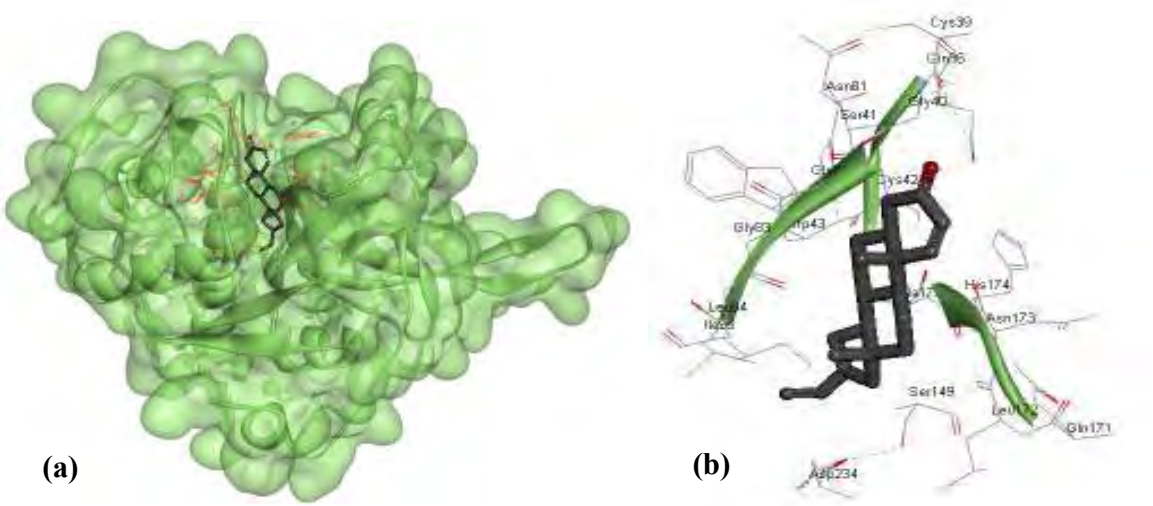


Figure 3.17: a) Surface presentation of FP-2 with best docked natural compound (5α-pregna- 1, 20-dien-3-one). Red lines indicate active site residues within 6Å of the ligand. b) Ligand binding site residues.

5α-pregna-1,20-dien-3-one however, lacked interactions such as hydrogen bonds and other important intermolecular interactions such as Van der Waals and electrostatic which was due to lack of reactive atoms to interact with the other sub sites as previously mentioned. However, already having a scaffold to start with, the compound could have its activity increased by synthetically adding more reactive groups. An alternative method would be to search for similar natural compounds from compound libraries such as the ZINC database (Irwin & Shoichet, 2006).

In other studies, xenicane diterpene obtained from algae in the Red Sea of Egypt were shown to possess anti-tumor activity against human lung and liver carcinomas *in vitro* (Awad *et al*, 2008).

Some pregnadiene derivatives are already being used in medicine e.g. corticosteroids for the treatment of various conditions thus the pharmacological properties are already known (Beek *et al*. 2007; Manzur *et al*. 2004).

3.3.4 ZINC database search

The ZINC database search for compounds similar to the best binding natural compound, 5α-pregna-1, 20-dien-3-one, produced 186 hits. All the 186 compounds were docked to the FP human homologs as well as 3D models of FP *Plasmodium* homologs. Out of these 186 some showed consistency in their binding to FP-2, FP-3 and homologs as well. Compounds with the relatively good free energy of binding scores were between compound numbers 5 to about 15 and 120 to 186 (Figure 3.18).

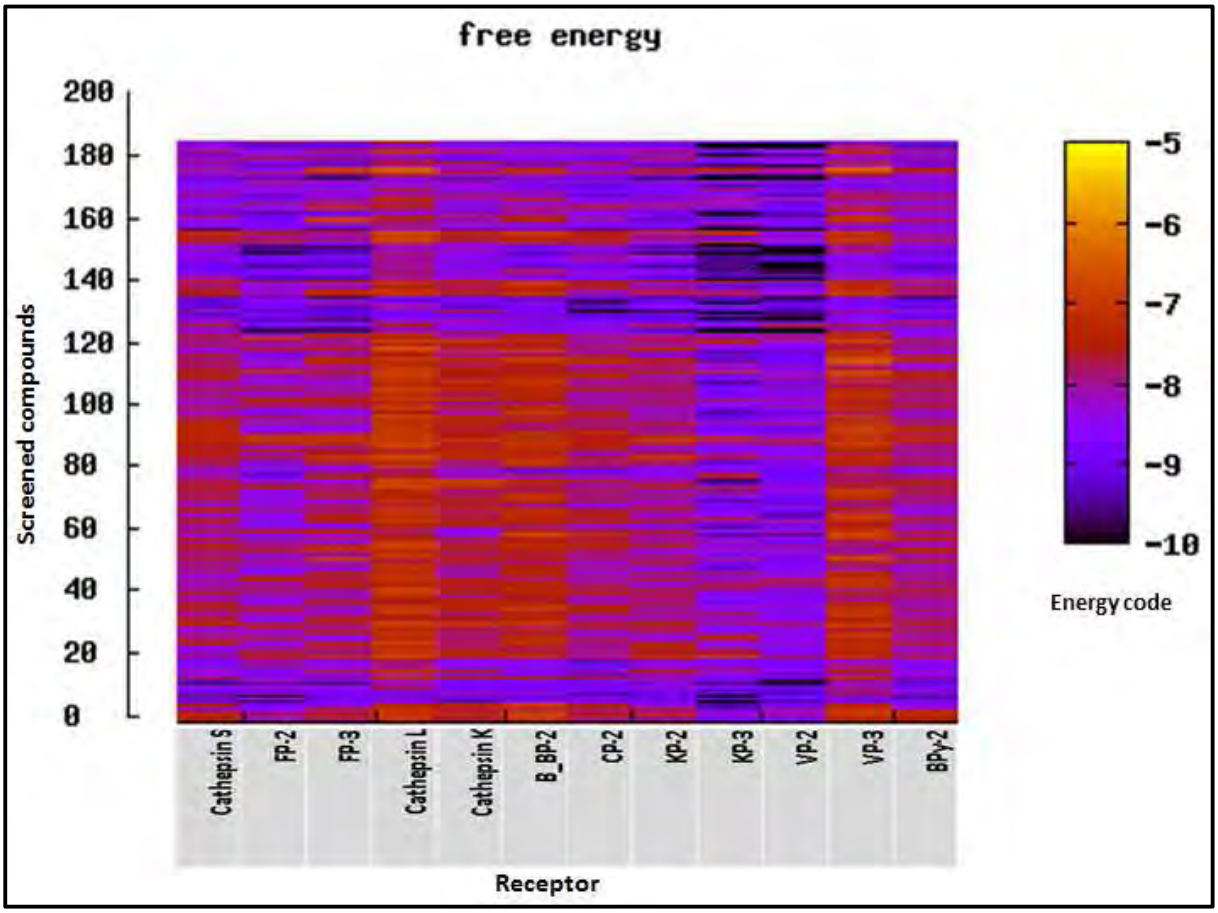


Figure 3.18: Estimated free energy of binding map for screened compounds (0001-186) from the ZINC database search. Low energy scores were observed in compounds between compound numbers 130 to 186. The energy code is colored from high energy (yellow) to low energy (black). Low energy scores correspond to high binding affinity.

Compounds shown in Figure 3.19 had the lowest predicted estimated free energy of binding. They had the best energies when docked to FP-2, FP-3, VP-2, KP-2 and KP-3. Of interest was that these favorable binding ligands were mainly docked to VP-2, KP-2 and KP-3 with a few bound to FP-2 and FP-3. Only one of the rodent *Plasmodium* homologs CP-2 was observed in this range of estimated free energy. None of these compounds (Figure 3.19) was observed docked to the human FP homologs with estimated free energy of binding of lower than -9.00 kcal/mol, which was a good indication of selectivity. There was consistency in binding across all *Plasmodium* homologs were estimated free energy of binding was ranged from -9.45 kcal/mol and above.

This observation seems to agree with experimental enzyme inhibitor assays where not all effective FP-2 or FP-3 inhibitors express the same kind of activity among the rodent *Plasmodium* homologs (Rosenthal *et al*. 2002). The compound with the lowest estimated free energy of binding was 000174 docked to VP-2, FP-3 and KP-2. Interestingly from the multiple sequence analysis and phylogenetic analysis (Table 2.1, Figure 2.4), these three homologs share high percentage sequence identity with VP-2 clustering with KP-3 and were the closest group to FP-2 and FP-3. Based on these observations, the docking outcome seems to predict the binding correctly. A similar observation was made with compound 000150 which docked with low estimated free energy of binding to FP-2, VP-3, KP-2 and KP-3 as well as compound 000126 with FP-2, FP-3, VP-2 and KP-3.

Compounds listed in Figure 3.19 did not bind with the human cathepsins with high affinities. From all the docking that was carried out (Figure 3.3, 3.15 and 3.18), there was a consistency in the way the human homologs were binding. Cathepsin K and S seemed to have better binding than cathepsin L but not comparable to the *Plasmodium* homologs which could be an indication that it is possible to achieve inhibitors selective to *Plasmodium* homologs only. Cathepsin L consistently had poor energy scores which improved whenever the inhibitors/ligands had high energy scores observed in the rest of the homologs.

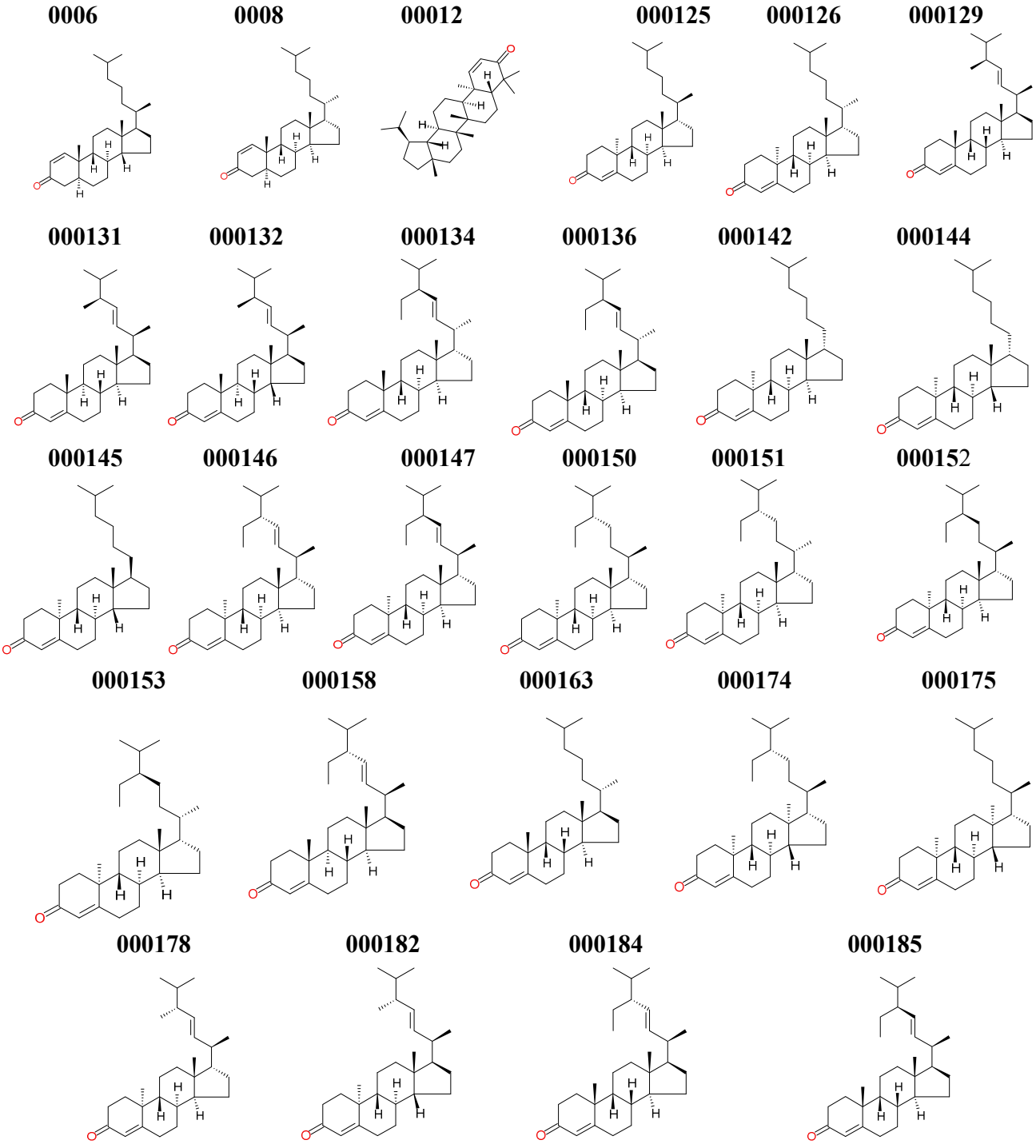


Figure 3.19: Docked ZINC database hits with lowest estimated free energy of binding in the range -9.45 kcal/mol and below as predicted by AutoDock4.2.

The improved energy scores among these compounds (Figure 3.21), was attributed to the additional atoms in the main chain of the inhibitors that were interacting with the S3 and S1 sub sites. This observation proves that these sub sites are important as well in determining the ligand/inhibitor binding which motivated the ZINC database search.

Analysis of 18 compounds from Figure 3.21 that docked to VP-2 with low free energy of binding revealed that it may have been favored by the narrow S2 pocket of VP-2 seemed to favor binding of these compounds therefore was possibly increasing the number of intermolecular interactions Figure 3.20). Variations observed at the S2 of the analysed FP *Plasmodium* homologs could have influenced intermolecular distances between the ligands and receptor which would definitely reduce the number of interactions between the ligand and receptor, thus the various binding affinity outcomes.

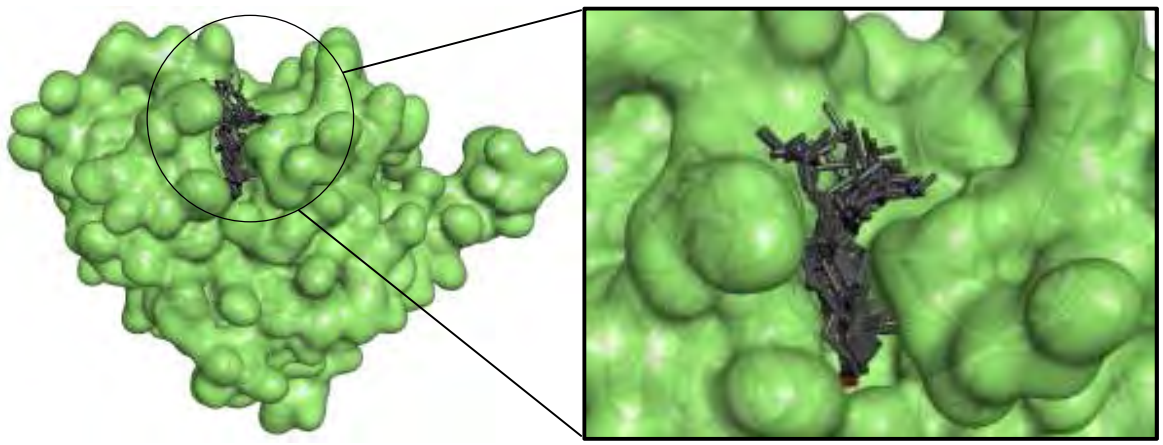


Figure 3.20: VP-2 with docked compounds; 000125-126, 000129, 000131, 000142, 000144-145, 000150-152, 000158, 000174-175, 000178, 000184-185. The compounds were consistent in their docking poses.

Further analysis shows that there were more of hydrophobic interactions involved than any type of interactions. Other protein-ligand interactions observed included polar interactions, ionic, electrostatic and Van der Waals interactions. Looking at the composition of the active site residues, the residues have ability to form strong non covalent bonding interactions of the previously stated kind.

To improve binding capability of these compounds, one may need to introduce more reactive groups such as amine groups, halogens, carbonyl, aromatic rings, and nucleophilic groups such as nitriles to induce interactions with the catalytic cysteine. This was according to observation from the docking with 2-cyanopyrimidine (Figure 3.3), whose binding affinities with the proteases increased with introduction of more reactive groups specifically aromatic rings, nitrogen rich aromatic rings e.g. pyridinyl, and halogens (Coterón *et al*. 2010). As has been discussed previously, the size of the ligand and active pocket should also be put to consideration in order to achieve maximum binding affinities (Goh & Sim, 2005; Saranya & Selvaraj, 2009; Saranya & Selvaraj, 2011).

CHAPTER FOUR

4. CONCLUSION AND FUTURE PROSPECTS

A total of five *Plasmodium* and four human FP-2/FP-3 homologs were retrieved (Table 2.1) and comparative analysis carried out via multiple sequence analysis (Figure 2.2). Multiple sequence analysis and phylogenetic analysis showed VP-2, 3 and KP-2 and KP-3 homologs were closer to FP-2 than the rodent *Plasmodium* homologs with the human homologs the least similar.

Comparative multiple sequence analysis carried out with emphasis placed on the proteases' active site of the proteases which is involved in substrate binding and the main focus for inhibitor development revealed that residue variations did exist in all the sub sites (S1-S1') among all the homologs. However, the catalytic residues were all conserved consistent with the typical C1A cysteine protease characteristics (Rosenthal, 2004; Sajid & Mckerrow, 2002).

Quality models for FP-2 and FP-3 *Plasmodium* homologs were created and from the structural analysis, the homolog proteins showed high fold conservation (Figure 2.9) at the active site but with few conformational differences which were attributed to residue variations observed previously at the analysis sequence level. Variations at the S2 as discussed in section 2.3.2.2 have been shown to cause size, volume and chemical composition differences.

Human homologs which appeared different from the phylogenetic analysis (Figure 2.4) seemed to have poor binding affinities to docked inhibitors compared to the *Plasmodium* counterparts. The *Plasmodium* FP homologs that were grouped together in the phylogenetic tree showed similar binding characteristics e.g. VP-2 and VP-3 KP-2 and KP-3, CP-2 BP-2 and BPy-2. Despite the variations observed, some inhibitors displayed uniform inhibition across all *Plasmodium* homologs which imply the possibility of a broad spectrum inhibitor (Figure 3.5).

Docking analysis results confirmed this observation where the binding site size, volume and ligand chemical composition seemed to dictate the kind of ligand pose and intermolecular interactions too as was observed in CP-2 and BP-2 with some inhibitors such as 2-cyanopyrimidine derivatives.

Smaller ligands/inhibitors seemed to have fewer intermolecular interactions compared to bigger ones, thus, when different sizes and volumes of the receptor/proteins are involved, ligand volume and binding pocket size and volume complementarity becomes important for increased binding affinities. Further insights on the effect of the binding site variations could be obtained through molecular dynamics which unfortunately, due to time constraints was not done.

The chemical composition of the ligand was important as was observed with the 2-cyanopyrimidines and isoquinoline derivatives and pyridone based peptidomimetic with a vinyl sulfone war head. Inhibitorsthat had many reactive atoms and groups e.g. carbonyls, nitriles, amines, and aromatic rings had more intermolecular interactions hence increased binding affinity. Inhibitors that were able to interact with Asn73, Gly83, Cys42, Trp206, Gln37, in FP-2 and corresponding positions in the *Plasmodium* homologs had increased binding affinities attributed to the hydrogen bonds, polar, hydrophobic, Van der Waal and electrostatic interactions formed. These interactions were deemed important since no explicit covalent bonds were formed as is the case in peptide based inhibitors.

From the small library of natural compounds, one compound, 5α-pregna-1,20-dien-3-one a xenicane diterpene had a high binding affinity consistent across the *Plasmodium* homologs. A ZINC database search returned 186 hits, with 27 showing improved binding affinities to FP-2, FP-3, VP-2, KP-2 and KP-3.

Despite the small number of natural compounds screened, there seems to a great a potential of discovering more compounds binding to *Plasmodium* FP homologs with higher affinities than observed thus the need to create a larger library of natural compounds for screening. In the future, we intend to analyse compounds from ZINC database search individually to identify points for improvement by identifying chemical groups that could be added to these compounds to increase intermolecular interactions at the active site and possibly create a library for these modified compounds for further screening. Due to computational time costs, only rigid docking was carried out, but better protein-ligand interactions could be inferred from both rigid and flexible docking for these compounds after the above recommendations. Enzyme inhibitor assays for the screened compounds as well as the modified compounds would add more evidence on the inhibitory activity of these compounds.

REFERENCES

Aguiar, A. C. C., Rocha, E. M. D., Souza, N. B. D., França, T. C., & Krettli, A. U. (2012). New approaches in antimalarial drug discovery and development: a review. *Memórias do Instituto Oswaldo Cruz*, *10* (7), 831-45.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403-10.

Altschul, S. F., Madden, T. L., Schäffer, 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*, *25*(17), 3389-402.

Alvarez, J. C. (2004). High-throughput docking as a source of novel drug leads. *Current Opinion in Chemical Biology*, *8*(4), 365-70.

Aly, A. S. I., & Matuschewski, K. (2005). A malarial cysteine protease is necessary for Plasmodium sporozoite egress from oocysts. *The Journal Of Experimental Medicine*, *202*(2), 225-30.

Arnold, K., Bordoli, L., Kopp, J., & Schwede, T. (2006). The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics (Oxford, England)*, *22*(2), 195-201.

Baker, D., & Sali, (2001). Protein structure prediction and structural genomics. *Science (New York, N.Y.)*, *294*(5540), 93-6.

Banerjee, R., Liu, J., Beatty, W., Pelosof, L., Klemba, M., & Goldberg, D. E. (2002). Four plasmepsins are active in the *Plasmodium falciparum* food vacuole, including a protease with an active-site histidine. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(2), 990-5.

Batra, S., Sabnis, Y. a, Rosenthal, P. J., & Avery, M. (2003). Structure-based approach to F-2 inhibitors: synthesis and biological evaluation of 1,6,7-Trisubstituted dihydroisoquinolines and isoquinolines. *Bioorganic & Medicinal Chemistry*, *11*(10), 2293-2299.

Beavers, M. P., Myers, M. C., Shah, P. P., Purvis, J. E., Scott, L., Cooperman, B. S., Huryn, D. M., *et al*. (2010). Molecular Docking of Cathepsin L inhibitors in the Binding Site of Papain. *J Chem Inf Model.*, *48*(7), 1464-1472.

Benkert, P., Biasini, M., & Schwede, T. (2011). Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics (Oxford, England)*, *27*(3), 343-50.

Benkert, P., Tosatto, S. C. E., & Schomburg, D. (2008). QMEAN: A comprehensive scoring function for model quality assessment. *Proteins*, *71*(1), 261-77.

Brady, R. L., & Cameron, A. (2004). Structure-based approaches to the development of novel anti-malarials. *Current Drug Targets*, *5*(2), 137-149.

Bursulaya, B. D., Totrov, M., Abagyan, R., & Brooks, C. L. (2003). Comparative study of several algorithms for flexible ligand docking. *Journal of Computer-Aided Molecular Design*, *17*(11), 755-63.

Caldeira, R. L., Gonçalves, L. M. D., Martins, T. M., Silveira, H., Novo, C., Rosário, V. D., & Domingos, A. (2009). Plasmodium chabaudi: expression of active recombinant chabaupain-1 and localization studies in Anopheles sp. *Experimental Parasitology*,*122*(2), 97-105. Elsevier Inc.

Carlton, J. M. (2002). Genome sequence and comparative analysis of model rodent malaria parasite *Plasmodium yoelii yoelii.Nature*, *Vol. 419*(3 Oct. 2002).

Cavasotto, C. N., & Phatak, S. S. (2009). Homology modeling in drug discovery: current trends and applications. *Drug Discovery Today*,*14*(13-14), 676-83.

Chan, C., Goh, L. L., & Sim, T. S. (2005). Differences in biochemical properties of the Plasmodial falcipain-2 and berghepain-2 orthologues: implications for *in vivo* screens of inhibitors. *FEMS Microbiology Letters*, *249*(2),

Chavain, N., Davioud-Charvet, E., Trivelli, X., Mbeki, L., Rottmann, M., Brun, R., & Biot, C. (2009). Antimalarial activities of ferroquine conjugates with either glutathione reductase inhibitors or glutathione depletors via a hydrolyzable amide linker. *Bioorganic & Medicinal Chemistry*,*17*(23), 8048-59.

Cheng, J. (2008). A multi-template combination algorithm for protein comparative modeling. *BMC structural biology*, *8*(2), 18.

Chipeleme, A., Gut, J., Rosenthal, P. J., & Chibale, K. (2007). Synthesis and biological evaluation of phenolic Mannich bases of benzaldehyde and (thio)semicarbazone derivatives against the cysteine protease falcipain-2 and a chloroquine resistant strain of *Plasmodium falciparum*. *Bioorganic & Medicinal Chemistry*, *15*(1), 273-82.

Chiyanzu, I., Hansell, E., Gut, J., Rosenthal, P. J., McKerrow, J. H., & Chibale, K. (2003). Synthesis and evaluation of isatins and thiosemicarbazone derivatives against cruzain, falcipain-2 and rhodesain. *Bioorganic & Medicinal Chemistry Letters*, *13*(20), 3527-3530.

Coterón, J. M., Catterick, D., Castro, J., Chaparro, M. J., Dı, B., Fern, E., Ferrer, S., *et al*. (2010). Falcipain Inhibitors□: Optimization Studies of the 2-Pyrimidinecarbonitrile Lead Series †. *Synthesis*, *Journal of Medicinal Chemistry* . 53 (21): 7885-6.

Cowman, A. F., & Kappe, S. H. I. (2006). Microbiology. Malaria's stealth shuttle. *Science (New York, N.Y.)*, *313*(5791), 1245-6.

Dahl, E. L., & Rosenthal, P. J. (2005). Biosynthesis, localization, and processing of falcipain cysteine proteases of *Plasmodium falciparum*. *Molecular and Biochemical Parasitology*, *139*(2), 205-212.

Dalal, S., & Klemba, M. (2007). Roles for two aminopeptidases in vacuolar hemoglobin catabolism in *Plasmodium falciparum*. *The Journal Of Biological Chemistry*,*282*(49), 35978-87.

Dastidar, E. G., Dayer, G., Holland, Z. M., Dorin-Semblat, D., Claes, A., Chêne, A., Sharma, A., *et al*. (2012). Involvement of *Plasmodium falciparum* protein kinase CK2 in the chromatin assembly pathway. *BMC biology*, *10*(1), 5.

Davies-coleman, M. T. (2005). Bioactive Natural Products From Southern African Marine Invertebrates. *Studies in Natural Products Chemistry*, *32*, 61-107.

Davies-coleman, M. T., & Beukes, D. R. (2004). Ten years of marine natural products research at Rhodes University. *South African Journal of Science*, *100*(December), 539-544.

Dayhoff, M. O., & Schwartz, R. M. (1978). 22 A Model of Evolutionary Change in Proteins. *Atlas Of Protein Sequence And Structure*,*Vo.l 5*(No. suppl. 3), pp345-351.

Deacon, C. F. (2011). Diabetes: A comparative review. *Diabetes, Obesity and Metabolism*, *13*(7), 7-18.

Desai, P V, & Avery, M. A. (2004). Structural characterization of vivapain-2 and vivapain-3, cysteine proteases from *Plasmodium vivax*: comparative protein modeling and docking studies. *Journal Of Biomolecular Structure Dynamics*, *21*(6), 781-790.

Desai, Prashant V, Patny, A., Gut, J., Rosenthal, P. J., Tekwani, B., Srivastava, A., & Avery, M. (2006). Identification of novel parasitic cysteine protease inhibitors by use of virtual screening. 2. The available chemical directory. *Journal of Medicinal Chemistry*, *49*(5), 1576-84.

Desai, Prashant V, Patny, A., Sabnis, Y., Tekwani, B., Gut, J., Rosenthal, P., Srivastava, A., *et al*. (2004). Identification of novel parasitic cysteine protease inhibitors using virtual screening. 1. The ChemBridge database. *Journal Of Medicinal Chemistry*,*47*(26), 6609-15.

Do, C. B., Mahabhashyam, M. S. P., Brudno, M., & Batzoglou, S. (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research*,*15*(2), 330-40.

Domínguez, J. N., León, C., Rodrigues, J., Gamboa de Domínguez, N., Gut, J., & Rosenthal, P. J. (2005). Synthesis and Evaluation of New Antimalarial Phenylurenyl Chalcone Derivatives. *Journal of Medicinal Chemistry*, *48*(10), 3654-3658.

Dorin-Semblat, D., Schmitt, S., Semblat, J.-P., Sicard, A., Reininger, L., Goldring, D., Patterson, S., *et al*. (2011). *Plasmodium falciparum* NIMA-related kinase Pfnek-1: sex specificity and assessment of essentiality for the erythrocytic asexual cycle. *Microbiology (Reading, England)*, *157*(Pt 10), 2785-94.

Edgar, R. C., & Batzoglou, S. (2006). Multiple sequence alignment. *Current opinion in structural biology*, *16*(3), 368-73.

Eggleson, K. K., Duffin, K. L., & Goldberg, D. E. (1999). Identification and characterization of falcilysin, a metallopeptidase involved in hemoglobin catabolism within the malaria parasite *Plasmodium falciparum*. *The Journal Of Biological Chemistry*, *274*(45), 32411-7.

Ehmke, V., Kilchmann, F., Heindl, C., Cui, K., Huang, J., Schirmeister, T., & Diederich, F. (2011). Peptidomimetic nitriles as selective inhibitors for the malarial cysteine protease falcipain-2. *MedChemComm*, *2*(8), 800.

Eksi, S., Czesny, B., Greenbaum, D. C., Bogyo, M., & Williamson, K. C. (2004). Targeted disruption of *Plasmodium falciparum* cysteine protease, falcipain 1, reduces oocyst production, not erythrocytic stage growth. *Molecular microbiology*, *53*(1), 243-50.

Elmar Krieger, Sander B. Nabuurs, and G. V. (2003). Homology modeling. *Structural Bioinformatics*, 507-521.

Eswar, N. (2003). Tools for comparative protein structure modeling and analysis. *Nucleic Acids Research*, *31*(13), 3375-3380.

Eswar, Narayanan, Webb, B., Marti-Renom, M. a, Madhusudhan, M. S., Eramian, D., Shen, M.-Y., Pieper, U., *et al*. (2007). Comparative protein structure modeling using MODELLER. *Current protocols in protein science / editorial board, John E. Coligan [et al.]*, *Chapter 2*, Unit 2.9.

Ettari, R., Zappalà, M., Micale, N., Grazioso, G., Giofrè, S., Schirmeister, T., & Grasso, S. (2011). Peptidomimetics containing a vinyl ketone warhead as falcipain-2 inhibitors. *European Journal of Medicinal Chemistry*, *46*(6), 2058-65.

Fernandez-Fuentes, N., Rai, B. K., Madrid-Aliste, C. J., Fajardo, J. E., & Fiser, A. (2007). Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments. *Bioinformatics (Oxford, England)*, *23*(19), 2558-65.

Flick, J., Tristram, F., & Wenzel, W. (2012). Modeling loop backbone flexibility in receptor-ligand docking simulations. *Journal Of Computational Chemistry*, *33*(31), 2504-15.

Florent, I., Lecaille, F., Montagne, J.-J., Gauthier, F., Schrével, J., & Lalmanach, G. (2005). Labelling of four distinct trophozoite falcipains of *Plasmodium falciparum* by a cystatin-derived probe. *Biological Chemistry*, *386*(4), 401-406.

Forrest, L. R., Tang, C. L., & Honig, B. (2006). On the Accuracy of Homology Modeling and Sequence Alignment Methods Applied t *Biophysical Journal*, *91*(2), 508-517.

Fujishima, A., Imai, Y., Nomura, T., Fujisawa, Y., Yamamoto, Y., & Sugawara, T. (1997). The crystal structure of human cathepsin L complexed with E-64. *FEBS Letters*, *407*(1), 47-50. Federation of European Biochemical Societies.

Ghosh, A., & Edwards, M. J. (2000). The Journey of the Malaria Parasite in the Mosquito□: *Parasitology Today*, *16*(5), 196-201.

Ginsburg, H., & Deharo, E. (2011). A call for using natural compounds in the development of new antimalarial treatments - an introduction. *Malaria Journal*,*10* (Suppl 1), S1.

Gluzman, I. Y., Francis, S. E., Oksman, a, Smith, C. E., Duffin, K. L., & Goldberg, D. E. (1994). Order and specificity of the *Plasmodium falciparum* hemoglobin degradation pathway. *The Journal Of Clinical Investigation*,*93*(4), 1602-8.

Goh, L. L., & Sim, T. S. (2005). Characterization of amino acid variation at strategic positions in parasite and human proteases for selective inhibition of falcipains in *Plasmodium falciparum*. *Biochemical and Biophysical Research Communications*, *335*(3), 762-770.

Goldberg, D. E., Slater, a F., Beavis, R., Chait, B., Cerami, a, & Henderson, G. B. (1991). Hemoglobin degradation in the human malaria pathogen *Plasmodium falciparum*: a catabolic pathway initiated by a specific aspartic protease. *The Journal Of Experimental Medicine*,*173*(4), 961-9.

Golubchik, T., Wise, M. J., Easteal, S., & Jermiin, L. S. (2007). Mind the gaps: evidence of bias in estimates of multiple sequence alignments. *Molecular Biology and Evolution*, *24*(11), 2433-42.

Goodsell, D. S., Morris, G. M., & Olson, J. (1996). Automated docking of flexible ligands: applications of AutoDock. *Journal of Molecular Recognition□: JMR*, *9*(1), 1-5.

Greenbaum, D. C., Baruch, A., Grainger, M., Bozdech, Z., Medzihradszky, K. F., Engel, J., DeRisi, J., *et al*. (2002). A role for the protease falcipain 1 in host cell invasion by the human malaria parasite. *Science (New York, N.Y.)*, *298*(5600), 2002-6.

Greenbaum, D. C., Mackey, Z., Hansell, E., Doyle, P., Gut, J., Caffrey, C. R., Lehrman, J., *et al*. (2004). Synthesis and Structure−Activity Relationships of Parasiticidal Thiosemicarbazone Cysteine Protease Inhibitors against *Plasmodium falciparum*, Trypanosoma brucei, and Trypanosoma cruzi. *Journal of Medicinal Chemistry*, *47*(12), 3212-3219.

Greenwood, B. M., Fidock, D. A., Kyle, D. E., Kappe, S. H. I., Alonso, P. L., Collins, F. H., & Duffy, P. E. (2008). Review series Malaria□: progress , perils , and prospects for eradication, *118*(4).

Gschwend, D. a, Good, a C., & Kuntz, I. D. (1996). Molecular docking towards drug discovery. *Journal of molecular recognition□: JMR*, *9*(2), 175-86.

Gupta, D., Yedidi, R. S., Varghese, S., Kovari, L. C., & Woster, P. M. (2010). Mechanism-based inhibitors of the aspartyl protease plasmepsin II as potential antimalarial agents. *Journal of Medicinal Chemistry*, *53*(10), 4234-4247.

Hansen, G., Heitmann, A., Witt, T., Li, H., Jiang, H., Shen, X., Heussler, V. T., *et al*. (2011). Structural basis for the regulation of cysteine-protease activity by a new class of protease inhibitors in Plasmodium. *Structure 1993*, *19*(7), 919-929.

Hanspal, M., Dua, M., Takakuwa, Y., Chishti, A. H., & Mizuno, A. (2002). *Plasmodium falciparum* cysteine protease falcipain-2 cleaves erythrocyte membrane skeletal proteins at late stages of parasite development. *Blood*, *100*(3), 1048-1054.

Harbut, M. B., Velmourougane, G., Dalal, S., Reiss, G., Whisstock, J. C., Onder, O., Brisson, D., *et al*. (2011). Bestatin-based chemical biology strategy reveals distinct roles for malaria M1- and M17-family aminopeptidases. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(34), E526-34.

Hartjes, L. B. (2012). Preventing and Detecting Malaria Infections. *Nurse Pract.*, *36*(6), 45-53.

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(22), 10915-9.

Hillisch, A., Pineda, L. F., & Hilgenfeld, R. (2004). Utility of homology models in the drug discovery process. *Drug Discovery Today*,*9*(15), 659-69.

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. *Journal of Biological Chemistry*, *281*(35), 25425-25437.

Huey, R., Morris, G. M., Olson, A. J., & Goodsell, D. S. (2007). Software News and Update A Semiempirical Free Energy Force Field with Charge-Based Desolvation. *Journal of Computational Chemistry*. Vol 28(6), 1145-52.

Irwin, J. J., & Shoichet, B. K. (2006). ZINC _ A Free Database of Commercially Available Compounds for Virtual Screening. *Journal of Chem Inf Model.*, *45*(1), 177-182.

Jacobson, M., & Sali, A. (2004). Comparative Protein Structure Modeling and its Applications to Drug Discovery. *Annual Reports in Medicinal Chemistry*, *39*(04), 259-275.

Jambou, R., El-Assaad, F., Combes, V., & Grau, G. E. (2011). *In vitro* culture of *Plasmodium berghei*-ANKA maintains infectivity of mouse erythrocytes inducing cerebral malaria. *Malaria Journal*,*10*(1), 346.

Jenko, S., Dolenc, I., Gunčar, G., Doberšek, A., Podobnik, M., & Turk, D. (2003). Crystal Structure of Stefin A in Complex with Cathepsin H: N-terminal Residues of Inhibitors can Adapt to the Active Sites of Endo- and Exopeptidases. *Journal of Molecular Biology*, *326*(3), 875-885. doi:10.1016/S0022-2836(02)01432-8

Jongwutiwes, S., Putaporntip, C., Iwasaki, T., Sata, T., & Study, T. (2004). Naturally Acquired knowlesi Malaria. *Emerging Infectious Diseases*, *10*(12), 2211-2213.

Katoh, K., Misawa, K., Kuma, K.-ichi, & Miyata, T. (2002). MAFFT□: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acid Research*,*30*(14), 3059-3066.

Kerr, I. D., Lee, J. H., Farady, C. J., Marion, R., Rickert, M., Sajid, M., Pandey, K. C., *et al*. (2009). Vinyl sulfones as antiparasitic agents and a structural basis for drug design. *The Journal of Biological Chemistry*,*284*(38), 25697-703.

Kerr, I. D., Lee, J. H., Pandey, K. C., Harrison, A., Sajid, M., Rosenthal, P. J., & Brinen, L. S. (2009). Structures of Falcipain-2 and Falcipain-3 Bound to Small Molecule Inhibitors: Implications for Substrate Specificity. *Journal of Medicinal Chemistry*, *52*(3), 852-857.

Khan, J. M., & Ranganathan, S. (2009). A multi-species comparative structural bioinformatics analysis of inherited mutations in alpha-D-mannosidase reveals strong genotype-phenotype correlation. *BMC Genomics*, *10 Suppl 3*, S33.

Kitchen, D. B., Decornez, H., Furr, J. R., & Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews. Drug discovery*, *3*(11), 935-49.

Korde, R., Bhardwaj, A., Singh, R., Srivastava, A., Chauhan, V. S., & Bhatnagar, R. K. (2008). A Prodomain Peptide of *Plasmodium falciparum* Cysteine Protease (Falcipain-2) Inhibits. *Journal of Medicinal Chemistry*, *51*, 3116-3123.

Kosiol, C., Bofkin, L., & Whelan, S. (2006). Phylogenetics by likelihood□: Evolutionary modeling as a tool for understanding the genome. *Journal of Biomedical Informatics*, *39*, 51-61.

Kumar, Amit, Kumar, K., Korde, R., Puri, S. K., Malhotra, P., & Singh Chauhan, V. (2007). Falcipain-1, a *Plasmodium falciparum* cysteine protease with vaccine potential. *Infection and Immunity*, *75*(4), 2026-34.

Kumar, Ashutosh, & Zhang, K. Y. J. (2012). Computational fragment-based screening using RosettaLigand: the SAMPL3 challenge. *Journal of computer-Aided Molecular Design*, *26*(5), 603-16.

Kumar, K. A., & Mishra, S. (2012). Plasmodium Pre-Erythrocytic Stages: Biology, Whole Parasite Vaccines And Transgenic Models. *American Journal of Immunology*. pg. 88-100

Laurie, A. T. R., & Jackson, R. M. (2006). Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. *Current Protein & Peptide Science*, *7*(5), 395-406.

Lecaille, F., Brömme, D., & Lalmanach, G. (2008). Biochemical properties and regulation of cathepsin K activity. *Biochimie*, *90*(2), 208-26.

Lecaille, F., Chowdhury, S., Purisima, E., Lalmanach, G., & Bro, D. (2007). The S2 sub sites of cathepsins K and L and their contribution to collagen degradation,*Protein Science, 16(4): 662-670.* 662-670.

Lecaille, F., Kaleta, J., & Brömme, D. (2002). Human and parasitic papain-like cysteine proteases: their role in physiology and pathology and recent developments in inhibitor design. *Chemical Reviews*, *102*(12), 4459-88.

Lew, V. L., Tiffert, T., & Ginsburg, H. (2003). Excess hemoglobin digestion and the osmotic stability of *Plasmodium falciparum*-infected red blood cells. *Blood*, *101*(10), 4189-94. doi:10.1182/blood-2002-08-2654

Lexa, K. W., & Carlson, H. (2012). Protein flexibility in docking and surface mapping. *Quarterly Reviews Of Biophysics*,*45*(3), 301-43.

Li, S. C., Bu, D., Xu, J., & Li, M. (2011). Finding nearly optimal GDT scores.*Journal Of Computational Biology□: A Journal Of Computational Molecular Cell Biology*,*18*(5), 693-704.

Li, X., Jacobson, M. P., & Friesner, R. A. (2004). High-Resolution Prediction of Protein Helix, *382 PROTEINS: Structure, Function, and Bioinformatics 55:368–382 (2004)* (September 2003), 368-382.

Liu, J., Istvan, E. S., Gluzman, I. Y., Gross, J., & Goldberg, D. E. (2006). *Plasmodium falciparum* ensures its amino acid supply with multiple acquisition pathways and redundant proteolytic enzyme systems. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(23), 8840-5.

Liu, M., Wilairat, P., & Go, M.-L. (2001). Antimalarial Alkoxylated and Hydroxylated Chalones□ Structure−Activity Relationship Analysis. *Journal of Medicinal Chemistry*, *44*(25), 4443-4452. American Chemical Society.

Liñares, G. E. G., & Rodriguez, J. B. (2007). Current Status and Progresses Made in Malaria Chemotherapy. *Current Medicinal Chemistry*, 289-314.

Manzur, A. Y., Kuntzer, T., Pike, M., & Swan, A. (2004). Glucocorticoid corticosteroids for Duchenne muscular dystrophy. *Cochrane Database Of Systematic Reviews Online*, (2), CD003725. John Wiley & Sons, Ltd.

May, A., & Zacharias, M. (2005). Accounting for global protein deformability during protein-protein and protein-ligand docking. *Biochimica et Biophysica Acta*, *1754*(1-2), 225-31.

McCormack, J. E., Huang, H., & Knowles, L. L. (2009). Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. *Systematic biology*, *58*(5), 501-8.

McKerrow, J. H., Engel, J. C., & Caffrey, C. R. (1999). Cysteine protease inhibitors as chemotherapy for parasitic infections. *Bioorganic & Medicinal Chemistry*,*7*(4), 639-44.

Mcguffin, L. J., Bryson, K., & Jones, D. T. (2000). The PSIPRED protein structure prediction server. *Bioinformatics*, *16*(4), 404-405.

Melo, F., Devos, D., Depiereux, E., & Feytmans, E. (1997). ANOLEA: a www server to assess protein structures. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology□; ISMB. International Conference on Intelligent Systems for Molecular Biology*, *5*, 187-90.

Mendis, K., Sina, B. J., Marchesini, P., & Carter, R. (2001). The neglected burden of *Plasmodium vivax* malaria. *The American Journal Of Tropical Medicine And Hygiene*,*64*(1-2 Suppl), 97-106.

Meyer, E. A., Castellano, R. K., & Diederich, F. (2003). *Interactions with Arenes Interactions with Aromatic Rings in Chemical and Biological Recognition Angewandte* (pp. 1210-1250).

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., Olson, A. J., *et al*. (1998). Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function, *J ournal of Computational Chemistry.19*(14), 1639-1662.

Morris, G. M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S., & Olson, A. J. (2009). Software News and Updates AutoDock4 and AutoDockTools4□: Automated Docking with Selective Receptor Flexibility. *Journal of Computational Chemistry*, *30*, 2785-2791.

Mukherjee, S., Ukil, A., & Das, P. K. (2007). Immunomodulatory peptide from cystatin, a natural cysteine protease inhibitor, against leishmaniasis as a model macrophage disease. *Antimicrobial Agents And Chemotherapy*, *51*(5), 1700-7. doi:10.1128/AAC.01555-06

Murray, C. W., Baxter, C. a, & Frenkel, D. (1999). The sensitivity of the results of molecular docking to induced fit effects: application to thrombin, thermolysin and neuraminidase. *Journal of Computer-Aided Molecular Design*, *13*(6), 547-62.

Na, B.-K., Kim, T.-S., Rosenthal, P. J., Lee, J.-K., & Kong, Y. (2004). Evaluation of cysteine proteases of *Plasmodium vivax* as antimalarial drug targets: sequence analysis and sensitivity to cysteine protease inhibitors. *Parasitology Research*,*94*(4), 312-7.

Na, B.-K., Shenai, B. R., Sijwali, P. S., Choe, Y., Pandey, K. C., Singh, A., Craik, C. S., *et al*. (2004). Identification and biochemical characterization of vivapains, cysteine proteases of the malaria parasite *Plasmodium vivax*. *The Biochemical Journal*, *378*(Pt 2), 529-38.

Newman, D. J., & Cragg, G. M. (2007). Natural Products as Sources of New Drugs over the Last 25 Years□. *Journal of Natural Products*, *70*(3), 461-477.

Olson, J. E., Lee, G. K., Semenov, A., & Rosenthal, P. J. (1999). Antimalarial E □ ects in Mice of Orally Administered Peptidyl Cysteine Protease Inhibitors. *Bioorganic and Medical Chemistry*, *7*, 633-638.

Omara-Opyene, a L., Moura, P. a, Sulsona, C. R., Bonilla, J. A., Yowell, C. a, Fujioka, H., Fidock, D., *et al*. (2004). Genetic disruption of the *Plasmodium falciparum* digestive vacuole plasmepsins demonstrates their functional redundancy. *The Journal of Biological Chemistry*, *279*(52), 54088-96.

Ouyang, X., Zhou, S., Su, C. T. T., Ge, Z., Li, R., & Kwoh, C. K. (2012). CovalentDock: Automated covalent docking with parameterized covalent linkage energy estimation and molecular geometry constrains. *Journal of Computational Chemistry*, 1-11.

Palmer, J. T., Rasnick, D., Klaus, J. L., & Brömme, D. (1995). Vinyl sulfones as mechanism-based cysteine protease inhibitors. *Journal Of Medicinal Chemistry*,*38*(17), 3193-6.

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*, *4*(5), e5694.

Pandey, K. C., & Dixit, R. (2012). Structure-Function of Falcipains□: Malarial Cysteine Proteases. *Tropical Medicine*, *2012*. doi:10.1155/2012/345195

Pandey, K. C., Sijwali, P. S., Singh, A., Na, B.-K., & Rosenthal, P. J. (2004). Independent intramolecular mediators of folding, activity, and inhibition for the *Plasmodium falciparum* cysteine protease falcipain-2. *The Journal Of Biological Chemistry*,*279*(5), 3484-91.

Pandey, K. C., Singh, N., Arastu-kapur, S., Bogyo, M., & Rosenthal, P. J. (2006). Falstatin , a Cysteine Protease Inhibitor of *Plasmodium falciparum* , Facilitates Erythrocyte Invasion. *PLoS Pathogens*, *2*(11), 1031-1041.

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*, *102*(26), 9138-43.

Pawlowski, M., Gajda, M. J., Matlak, R., & Bujnicki, J. M. (2008). MetaMQAP: a meta-server for the quality assessment of protein models. *BMC Bioinformatics*, *9*, 403.

Pei, J., Kim, B.-H., & Grishin, N. V. (2008). PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Research*, *36*(7), 2295-300. doi:10.1093/nar/gkn072

Puran S. Sijwali, Bhaskar R. Shenai, Jiri Gut, A. S. and P. J. R. (2001). Expression and characterization of *Plasmodium falciparum*haemoglobinase falcipain-3. *Biochemistry Journal*, *489*, 481-489.

Pérez, B. C., Teixeira, C., Figueiras, M., Gut, J., Rosenthal, P. J., Gomes, J. R. B., & Gomes, P. (2012). Novel Cinnamic Acid/4-Aminoquinoline Conjugates Bearing Non-Proteinogenic Amino Acids: Towards The Development Of Potential Dual Action Antimalarials. *European Journal Of Medicinal Chemistry*, *54*, 887-99.

Raval, A., Piana, S., Eastwood, M. P., Dror, R. O., & Shaw, D. E. (2012). Dynamics Simulations. *Proteins*, *80*(February), 2071-2079.

Redzynia, I., Ljunggren, A., Bujacz, A., Abrahamson, M., Jaskolski, M., & Bujacz, G. (2009). Crystal Structure Of The Parasite Inhibitor Chagasin In Complex With Papain Allows Identification Of Structural Requirements For Broad Reactivity And Specificity Determinants For Target Proteases. *The FEBS Journal*, *276*(3), 793-806.

Reiser, J., Adair, B., & Reinheckel, T. (2010). Science In Medicine Specialized Roles For Cysteine Cathepsins In Health And Disease. *The Journal of Clinical Investigation*, *120*(10), 24-26.

Rosenthal, P J. (1998). Proteases Of Malaria Parasites: New Targets For Chemotherapy. *Emerging Infectious Diseases*, *4*(1), 49-57.

Rosenthal, P J, McKerrow, J. H., Aikawa, M., Nagasawa, H., & Leech, J. H. (1988). A Malarial Cysteine Proteinase is Necessary for Hemoglobin Degradation By *Plasmodium Falciparum*. *The Journal of Clinical Investigation*,*82*(5), 1560-6.

Rosenthal, P J, Olson, J. E., Lee, G. K., Palmer, J. T., Klaus, J. L., & Rasnick, D. (1996). Antimalarial effects of vinyl sulfone cysteine proteinase inhibitors. *Antimicrobial Agents And Chemotherapy*, *40*(7), 1600-3.

Rosenthal, P J, Wollish, W. S., Palmer, J. T., & Rasnick, D. (1991). Antimalarial effects of peptide inhibitors of a *Plasmodium falciparum* cysteine proteinase. *The Journal of clinical investigation*, *88*(5), 1467-72.

Rosenthal, P. J. (2003). Antimalarial drug discovery: old and new approaches. *Journal of Experimental Biology*, *206*(21), 3735-3744.

Rosenthal, Philip J. (2004). Cysteine proteases of malaria parasites. *International Journal for Parasitology*, *34*, 1489-1499.

Rosenthal, Philip J. (2011). Falcipains and other cysteine proteases of malaria parasites. *Advances In Experimental Medicine And Biology*, *712*, 30-48.

Rosenthal, Philip J, Sijwali, P. S., Singh, A., & Shenai, B. R. (2002). Cysteine proteases of malaria parasites: targets for chemotherapy. *Current pharmaceutical design*, *8*(18), 1659-72.

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein engineering*, *12*(2), 85-94.

Rzychon, M., Chmiel, D., & Stec-Niemczyk, J. (2004). Modes of inhibition of cysteine proteases. *Acta biochimica Polonica*, *51*(4), 861-73. doi:045104861

Sabnis, Y. a, Desai, P. V., Rosenthal, P. J., & Avery, M. (2003). Probing the structure of falcipain-3, a cysteine protease from *Plasmodium falciparum*□: Comparative protein modeling and docking studies Probing the structure of falcipain-3 , a cysteine protease from *Plasmodium falciparum*□: Comparative protein modeling a. *Protein Science*, *12*(3), 501-509.

Sajid, M., & Mckerrow, J. H. (2002). Cysteine proteases of parasitic organisms. *Science*, *120*, 1-21.

Salas, F., Fichmann, J., Lee, G. K., Scott, M. D., & Rosenthal, P. J. (1995). Functional expression of falcipain, a *Plasmodium falciparum* cysteine proteinase, supports its role as a malarial hemoglobinase. *Infection And Immunity*,*63*(6), 2120-5.

Šali, A. (2011). MODELLER A Program for Protein Structure Modeling. *(http://salilab.org/modeller/manual/)*.

Santos, C. C., Scharfstein, J., & Lima, A. P. C. D. (2006). Role of chagasin-like inhibitors as endogenous regulators of cysteine proteases in parasitic protozoa. *Parasitology Research*,*99*(4), 323-4.

Saranya, N, & Selvaraj, S. (2009). Variation of protein binding cavity volume and ligand volume in protein-ligand complexes. *Bioorganic & Medicinal Chemistry Letters*,*19*(19), 5769-72.

Saranya, Nallusamy, & Selvaraj, S. (2011). Role of interactions and volume variation in discriminating active and inactive forms of cyclin-dependent kinase-2 inhibitor complexes. *Chemical Biology & Drug Design*,*78*(3), 361-9.

Schuldt, N. J., & Amalfitano, A. (2012). Malaria vaccines: focus on adenovirus based vectors. *Vaccine*, *30*(35), 5191-8. Elsevier Ltd. doi:10.1016/j.vaccine.2012.05.048

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*, *51*(4), 852-864.

Shen, M.-Y., & Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Science□: a publication of the Protein Society*, *15*(11), 2507-24.

Shenai, B R, Sijwali, P. S., Singh, a, & Rosenthal, P. J. (2000). Characterization of native and recombinant falcipain-2, a principal trophozoite cysteine protease and essential hemoglobinase of *Plasmodium falciparum*. *The Journal of Biological Chemistry*, *275*(37), 29000-10.

Shenai, Bhaskar R, Lee, B. J., Alvarez-hernandez, A., Chong, P. Y., Emal, C. D., Neitz, R. J., Roush, W. R., *et al*. (2003). Structure-Activity Relationships for Inhibition of Cysteine Protease Activity and Development of *Plasmodium falciparum* by Peptidyl Vinyl Sulfones. *Antimicrobial Agents and Chemotherapy*, *47*(1), 154-160.

Shenai, Bhaskar R, & Rosenthal, P. J. (2002). Reducing requirements for hemoglobin hydrolysis by *Plasmodium falciparum* cysteine proteases. *Molecular And Biochemical Parasitology*,*122*(1), 99-104.

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*, *277*(17), 14910-14915.

Sijwali, Puran 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*, *150*(1), 96-106.

Sijwali, Puran 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*, *101*(13), 4384-9.

Singh, A., & Rosenthal, P. J. (2001). Comparison of Efficacies of Cysteine Protease Inhibitors against Five Strains of *Plasmodium falciparum*, *Antimicrobial Agents And Chemotherapy*,Mar. 2001. *45*(3), 949-951

Singh, N., Sijwali, P. S., Pandey, K. C., & Rosenthal, P. J. (2006). *Plasmodium falciparum*: biochemical characterization of the cysteine protease falcipain-2'. *Experimental parasitology*, *112*(3), 187-92.

Stack, C. M., Caffrey, C. R., Donnelly, S. M., Seshaadri, A., Lowther, J., Tort, J. F., Collins, P. R., *et al*. (2008). Structural and functional relationships in the virulence-associated cathepsin L proteases of the parasitic liver fluke, Fasciola hepatica. *The Journal of biological chemistry*, *283*(15), 9896-908.

Steinbuechel, M., & Matuschewski, K. (2009). Role for the Plasmodium sporozoite-specific transmembrane protein S6 in parasite motility and efficient malaria transmission. *Cellular microbiology*, *11*(2), 279-88.

Stephens, R., Culleton, R. L., & Lamb, T. J. (2012). The contribution of Plasmodium chabaudi to our understanding of malaria. *Trends in Parasitology*, *28*(2), 74-83.

Stoch, S. A., & Wagner, J. A. (2007). Cathepsin K Inhibitors: A Novel Target for Osteoporosis Therapy. *Clin Pharmacol Ther*, *83*(1), 172-176.

Sturm, A., Amino, R., van de Sand, C., Regen, T., Retzlaff, S., Rennenberg, A., Krueger, A., *et al*. (2006). Manipulation of host hepatocytes by the malaria parasite for delivery into liver sinusoids. *Science (New York, N.Y.)*, *313*(5791), 1287-90.

Subramanian, S., Hardt, M., Choe, Y., Niles, R. K., Johansen, E. B., Legac, J., Gut, J., *et al*. (2009). Hemoglobin Cleavage Site-Specificity of the *Plasmodium falciparum* Cysteine Proteases Falcipain-2 and Falcipain-3. (M. M. Rodrigues, Ed.)*PLoS ONE*, *4*(4), 10. Public Library of Science.

Subramanian, S., Sijwali, P. S., & Rosenthal, P. J. (2007). Falcipain cysteine proteases require bipartite motifs for trafficking to the *Plasmodium falciparum* food vacuole. *The Journal of biological chemistry*, *282*(34), 24961-9.

Sullivan, S. M., & Holyoak, T. (2008). Enzymes with lid-gated active sites must operate by an induced fit mechanism instead of conformational selection. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(37), 13829-34.

Söding, J., Biegert, A., & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, *33*(Web Server issue), W244-8.

Takayuki Shindo and Reiner A.L Van der Hoorn. (2008).*Molecular Plant Pathology*, *9*(1), 119-125.

Talman, A. M., Domarle, O., McKenzie, F. E., Ariey, F., & Robert, V. (2004). Gametocytogenesis: the puberty of *Plasmodium falciparum*. *Malaria Journal*, 3:24

Tamm, L. K., & Liang, B. (2006). NMR of membrane proteins in solution. *Progress in Nuclear Magnetic Resonance Spectroscopy*, *48*(4), 201-210.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., & Kumar, S. (2011). MEGA5□: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood , Evolutionary Distance , and Maximum Parsimony Methods Research resource.*Molecular Biology and Evolution*. *28*(10), 2731-2739.

Tastan Bishop, OzlemA., T. A. P. de B. and F. J. (2008). Protein homology modelling and its use in South Africa. South African Journal of Science 104, (February), 2-6.

Teixeira, C., Gomes, J. R. B., & Gomes, P. (2011). Falcipains, *Plasmodium falciparum* cysteine proteases as key drug targets against malaria. *Current Medicinal Chemistry*, *18*(10), 1555-1572.

Teodoro, M. L., Phillips, G. N., & Kavraki, L. E. (2001). Molecular docking: a problem with thousands of degrees of freedom. *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No.01CH37164)*, *1*, 960-965.

Tina, K. G., Bhadra, R., & Srinivasan, N. (2007). PIC: Protein Interactions Calculator. *Nucleic Acids Research*,*35*(Web Server issue), W473-6.

Turk, B. (2006). Targeting proteases: successes, failures and future prospects. *Nature Reviews. Drug discovery*, *5*(9), 785-99.

Turk, B., Turk, D., & Salvesen, G. S. (2002). Regulating cysteine protease activity: essential role of protease inhibitors as guardians and regulators. *Current Pharmaceutical Design*,*8*(18), 1623-37.

Van De Beek, D., De Gans, J., McIntyre, P., & Prasad, K. (2007). Corticosteroids for acute bacterial meningitis. *Cochrane Database Of Systematic Reviews Online*,*357*(1), CD004405.

Vaughan, A. M., Aly, A. S. I., & Kappe, S. H. I. (2009). Malaria parasite pre-erythrocytic stage infection: Gliding and Hiding.*Cell Host Microbe.4*(3), 209-218.

Venclovas, C. (2012). Homology Modeling Methods and Protocols. (A. J. W. Orry & R. Abagyan, Eds.), *857*. *Methods in Molecular Biology*, 857. pg. 55- 82.

Verissimo, E., Berry, N., Gibbons, P., Cristiano, M. L. S., Rosenthal, P. J., Gut, J., Ward, S., *et al*. (2008). Design and synthesis of novel 2-pyridone peptidomimetic falcipain 2/3 inhibitors. *Bioorganic & Medicinal Chemistry Letters*, *18*(14), 4210-4214.

Wang, S. X., Pandey, K. C., Scharfstein, J., Whisstock, J., Huang, R. K., Jacobelli, J., Fletterick, R. J., *et al*. (2007). The Structure of Chagasin in Complex With A Cysteine Protease Clarifies The Binding Mode And Evolution Of An Inhibitor Family. *Structure (London, England□: 1993)*, *15*(5), 535-43.

Wang, S. X., Pandey, K. C., Somoza, J. R., Sijwali, P. S., Kortemme, T., Brinen, L. S., Fletterick, R. J., *et al*. (2006). Structural Basis For Unique Mechanisms of Folding And Hemoglobin Binding By A Malarial Protease. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(31), 11503-8.

Weikl, T. R., & von Deuster, C. (2009). Selected-Fit Versus Induced-Fit Protein Binding: Kinetic Differences And Mutational Analysis. *Proteins*, *75*(1), 104-10.

Weissman, K. J., & Leadlay, P. F. (2005). Combinatorial Biosynthesis of Reduced Polyketides. *Nat Rev Micro*, *3*(12), 925-936.

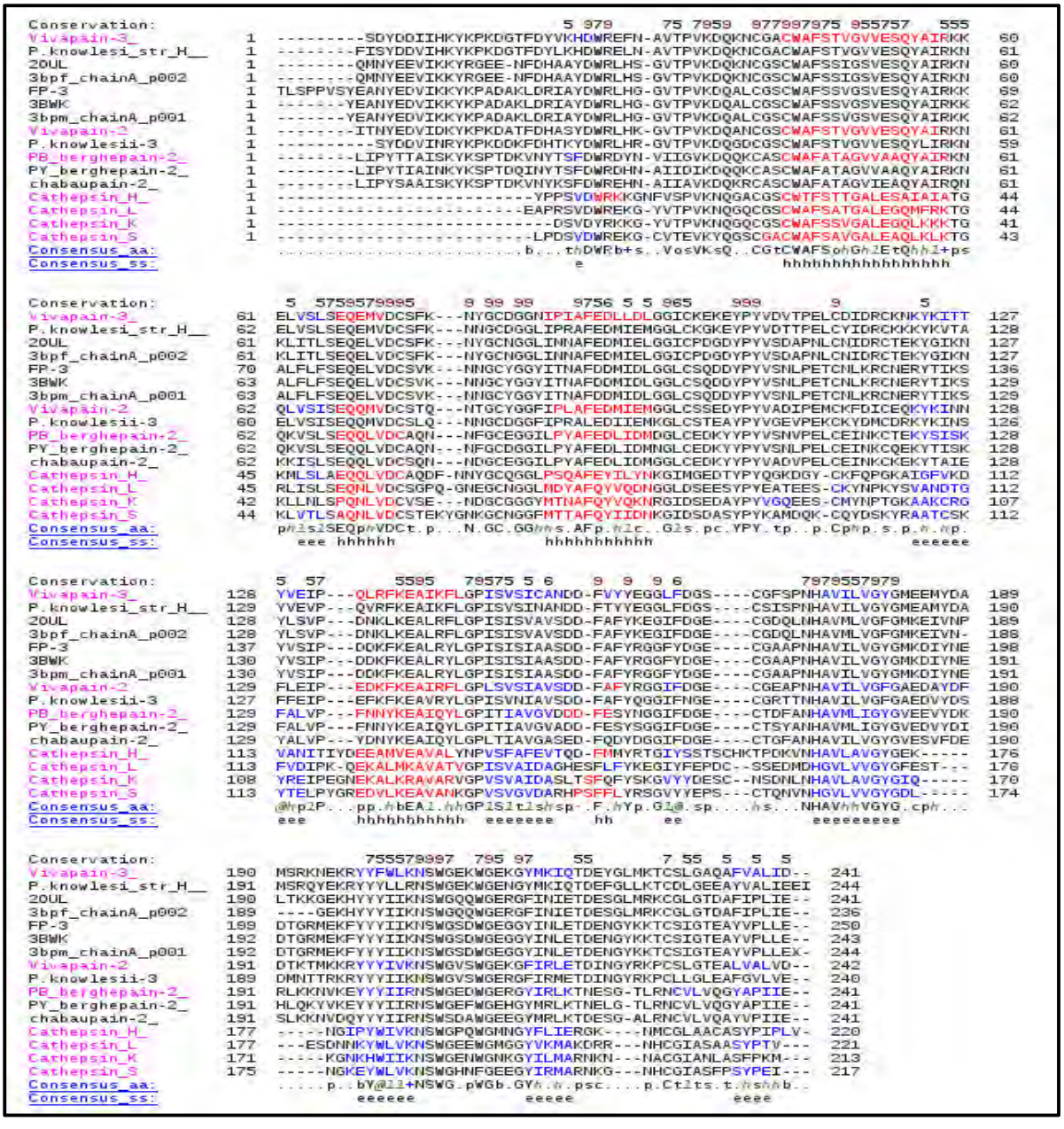
Wirth, D. F. (2002). Biological Revelations. *Nature*, *419*(6906), 495-6. doi:10.1038/419495a

Wongsrichanalai, C., Pickard, A. L., Wernsdorfer, W. H., & Meshnick, S. R. (2002). Epidemiology of drug-resistant malaria, *Lancet Infect Dis.* 2002 Apr;2(4):209-218.

Xiang, Z. (2006). Advances in homology protein structure modeling. *Current protein & peptide science*, *7*(3), 217-27.

APPENDICES

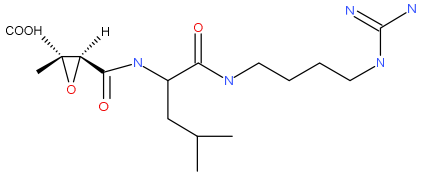
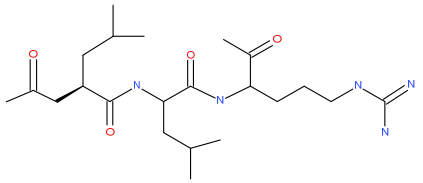
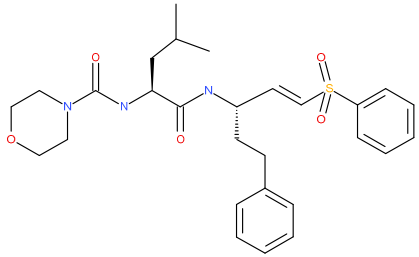
Appendix 1

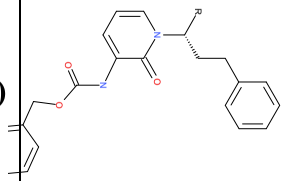
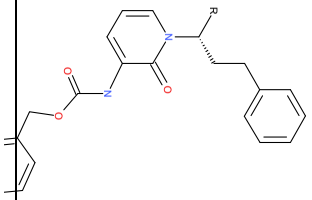
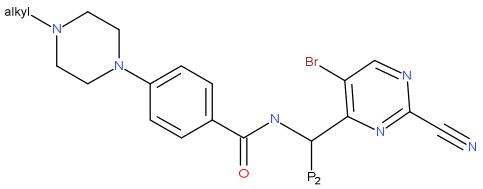
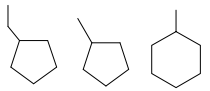
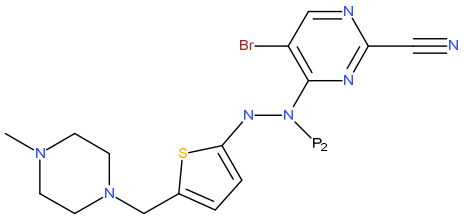
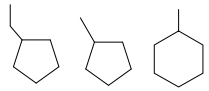


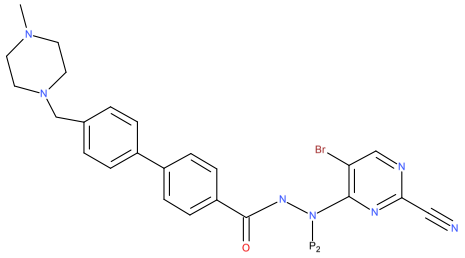
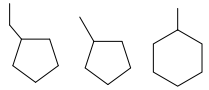
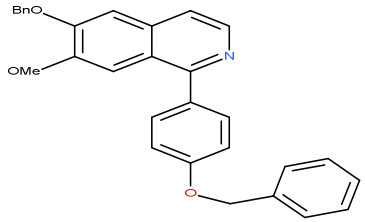
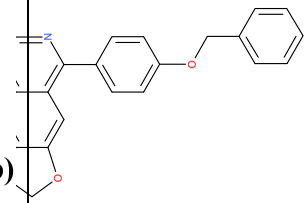
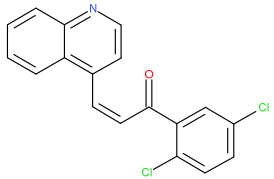
Appendix 1-1: Multiple sequence alignment as predicted by PROMALS3D.

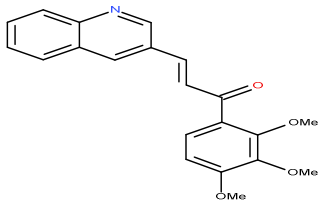
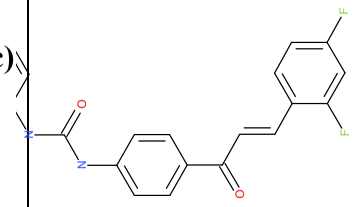
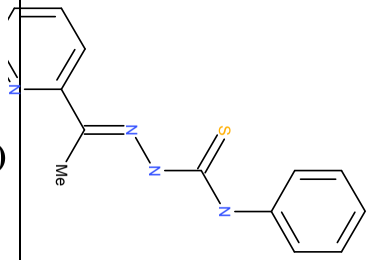
Appendix 2A

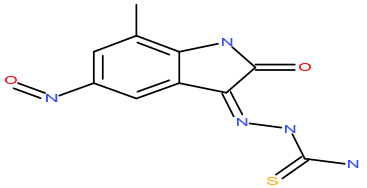
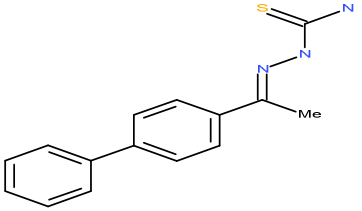
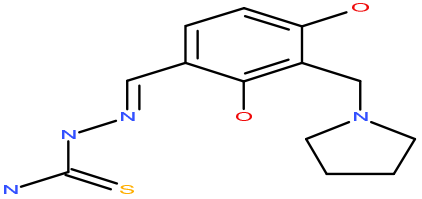
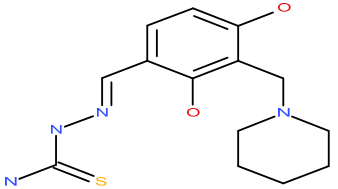
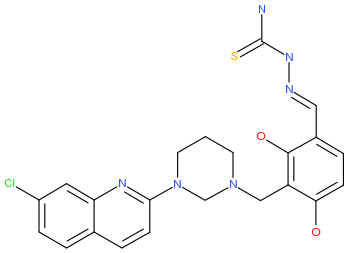
Chemical structures of non-peptide ligands used for molecular docking

Ligand type	Inhibitor name	Structure
Peptide based	**E-64**	
	leupeptin	
	Mu-Leu-homoPhe-VsPH	

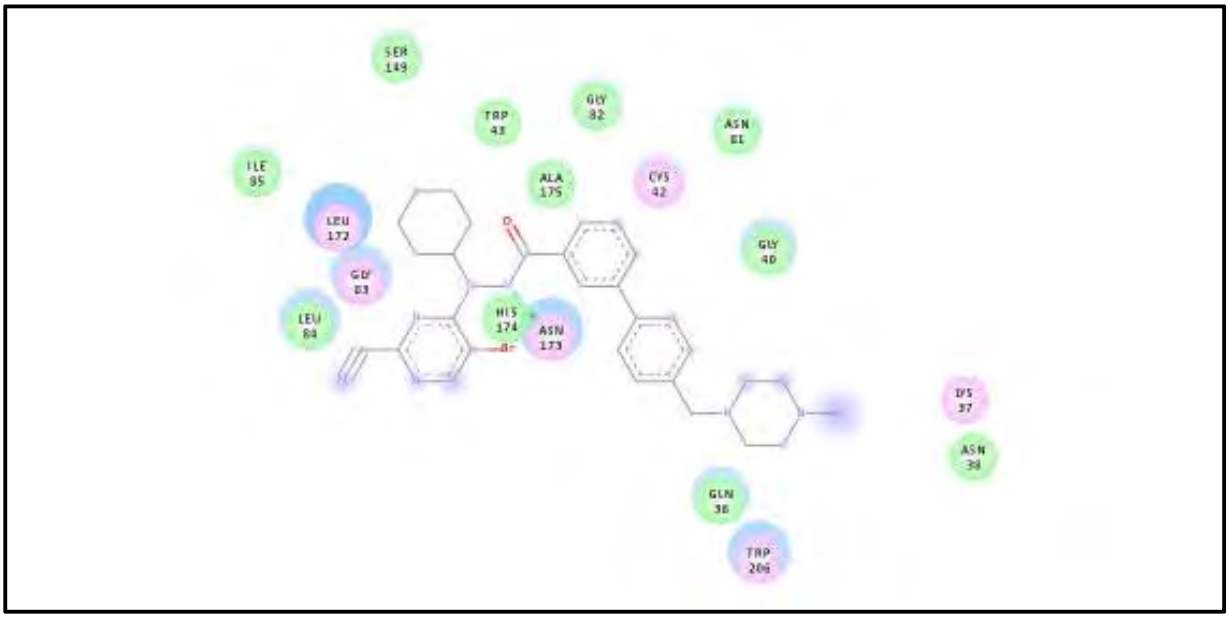
Peptidomimetic	**Peptidomimetics based on a pyridone ring scaffold**	**a)**  R=CH=CHSO2Ph
		b)  R=CHO
Non-peptide	**2-cyanopyrimodine derivatives** For experimental convenience, derivatives were renamed; 2_CPA, 2_CPB, 2_CPC based on the P2 substituents respectively	 P2= 
	For experimental convenience, derivatives were renamed; 2_CPD, 2_CPE, 2_CPF based on the P2 substituents respectively	 P2= 

	For experimental convenience, derivatives were renamed; 2_CPG, 2_CPH, 2_CPI based on the P2 substituents respectively	 P2= 
	Isoquinoline derivatives **a)**	
	Isoquinoline derivatives **b)**	
	Chalcones **a)**	

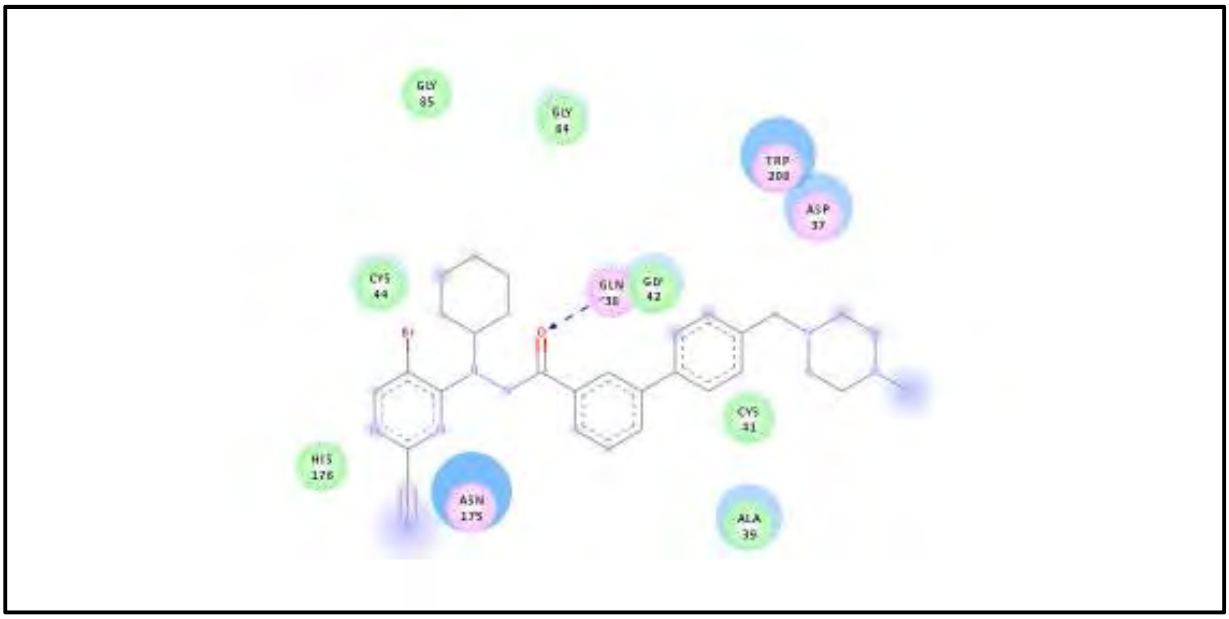
	b)	
	c)	
	Thiosemicarbazones derivatives **a)**	

	Thiosemicarbazones derivatives **b)**	
	c)	
	d)	
	e)	
	f)	

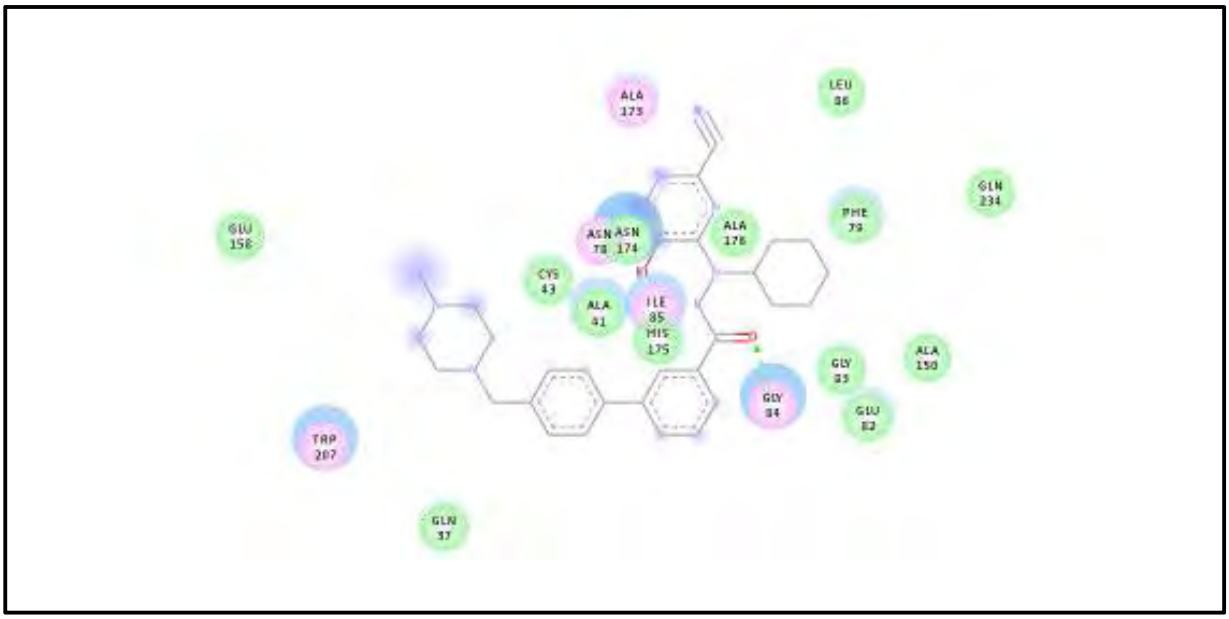
Appendix 2B



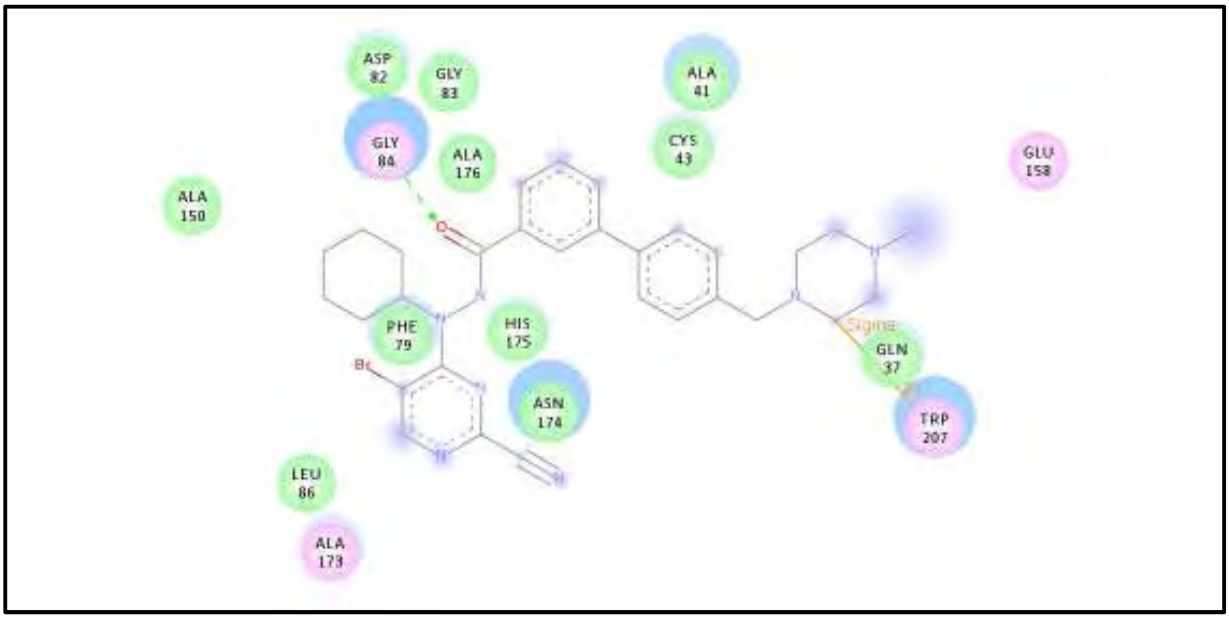
Appendix 2B-1: FP-2 active site residues interacting with 2-cyanopyrimidine inhibitor derivative 2_CPI.



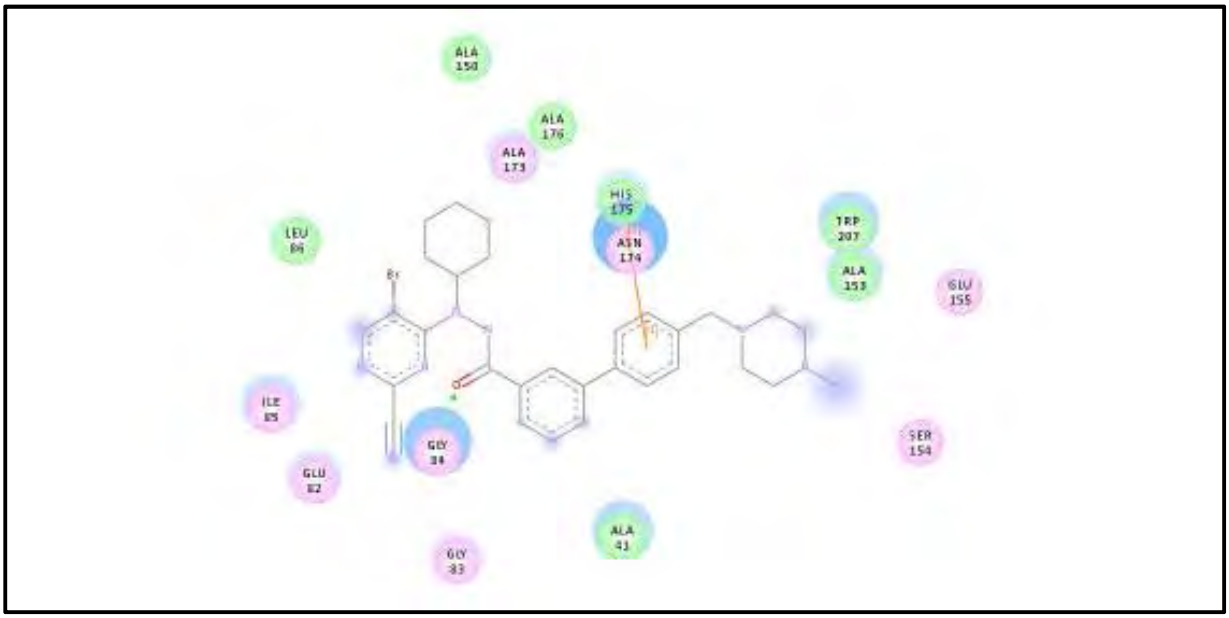
Appendix 2B-2: FP-3 active site residues interacting with 2-cyanopyrimidine inhibitor derivative 2_CPI.



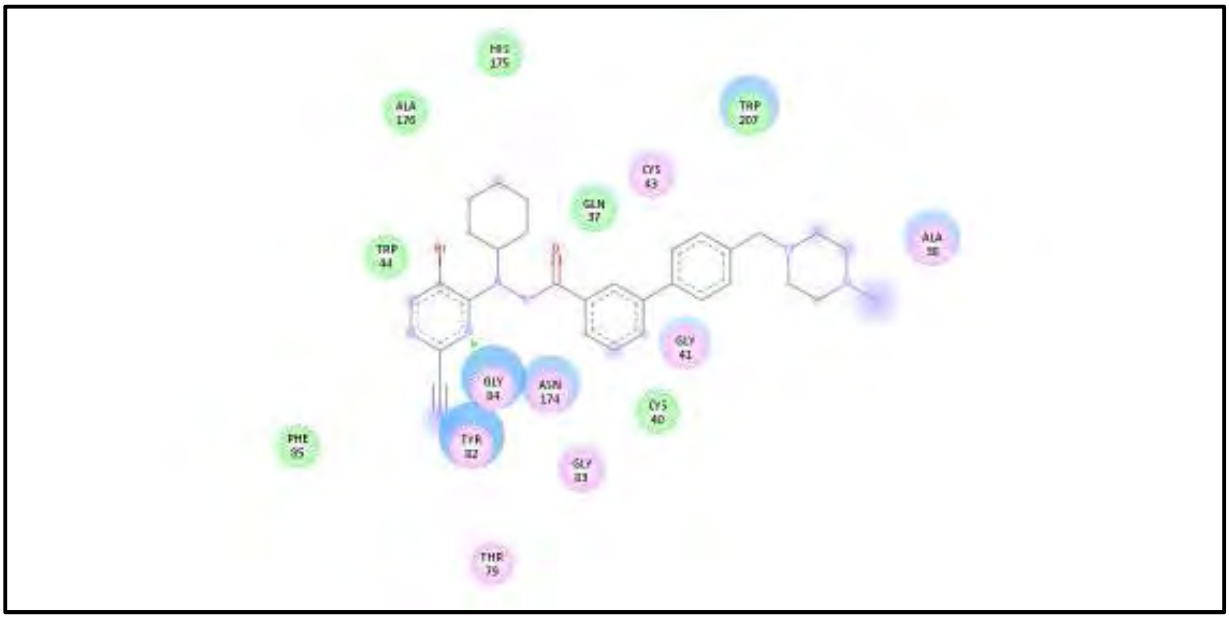
Appendix 2B-3: BP-2 active site residues interacting with 2-cyanopyrimidine derivative 2_CPI.



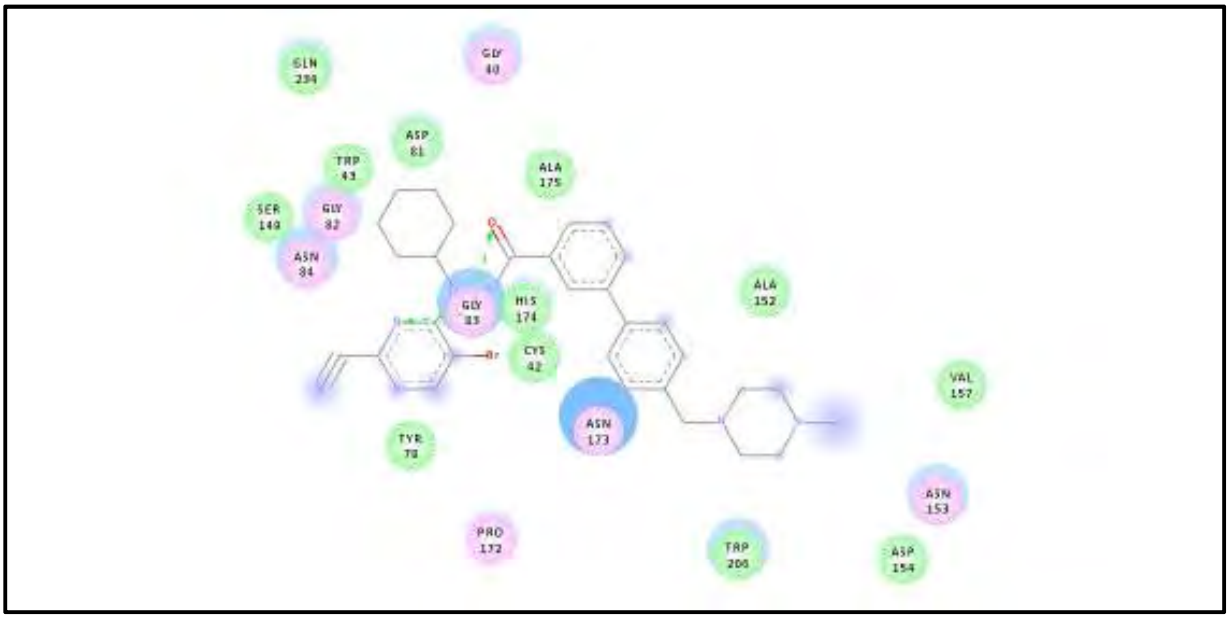
Appendix 2B-4: BPy-2 active site residues interacting with 2-cyanopyrimidine inhibitor derivative 2_CPI.



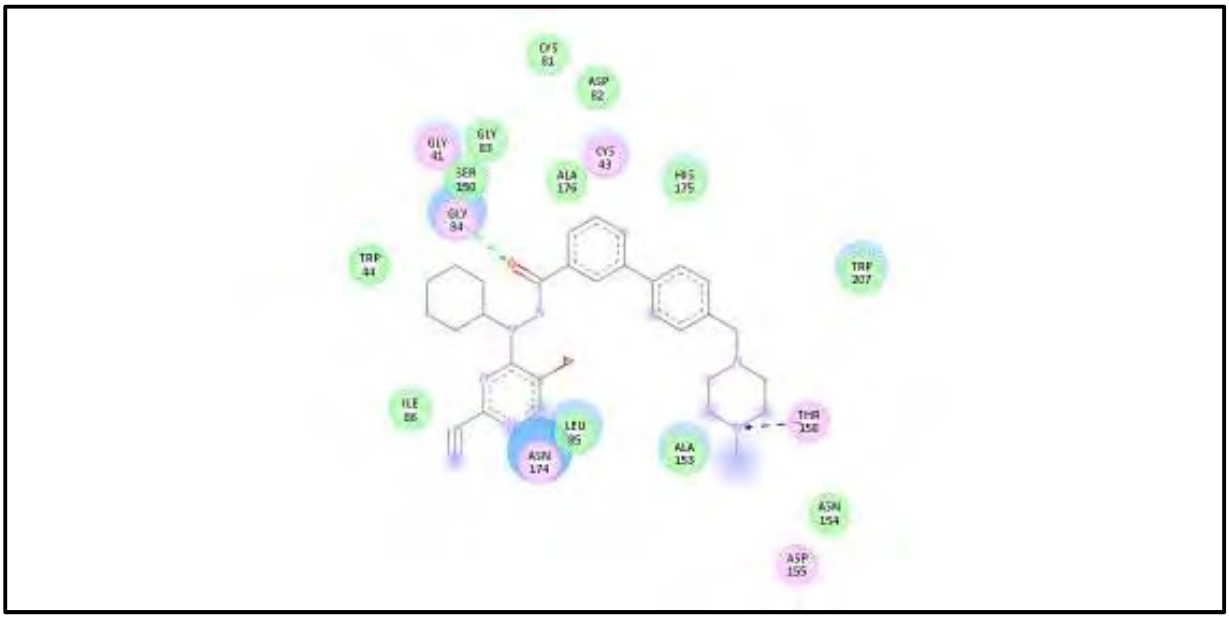
Appendix 2B-5: CP-2 active site residues interacting with 2-cyanopyrimidine inhibitor derivative (2_CPI).



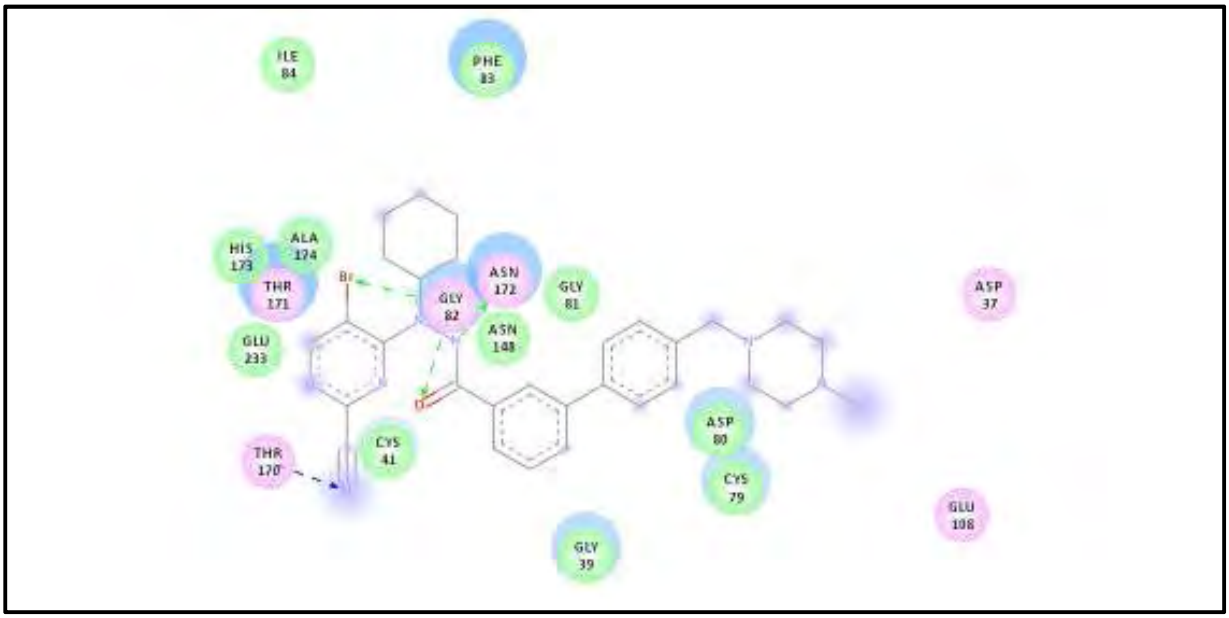
Appendix 2B-6: VP-2 active site residues interacting with 2-cyanopyrimidine derivative 2_CPI.



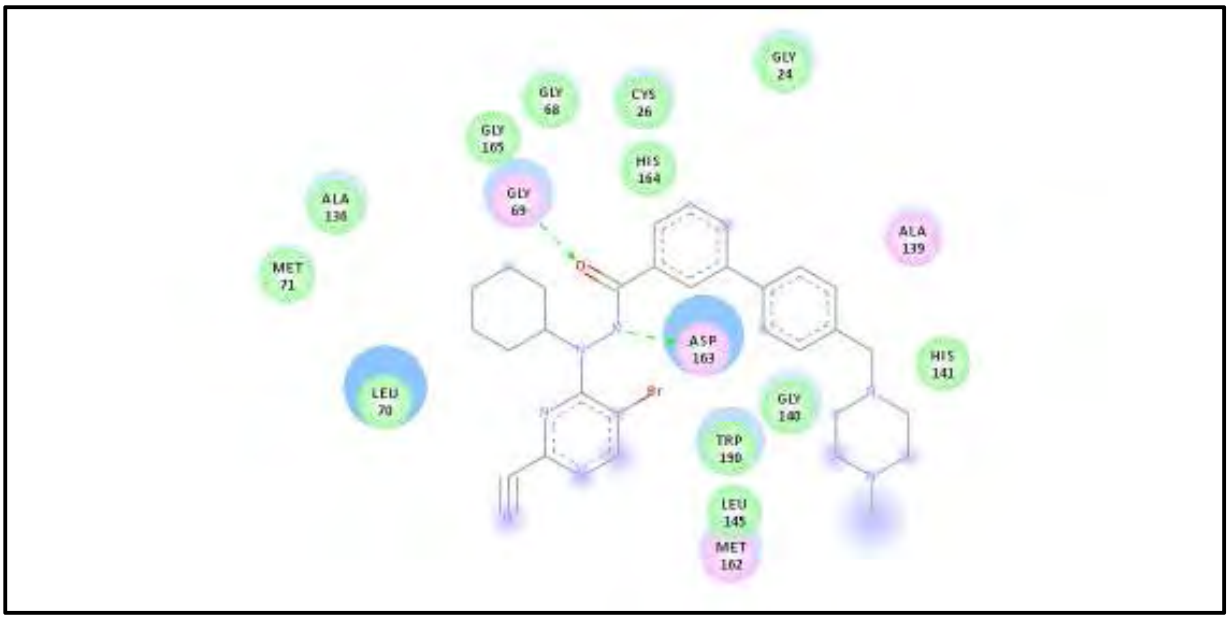
Appendix 2B-7: VP-3 active site residues interacting with 2-cyanopyrimidine derivative 2_CPI.



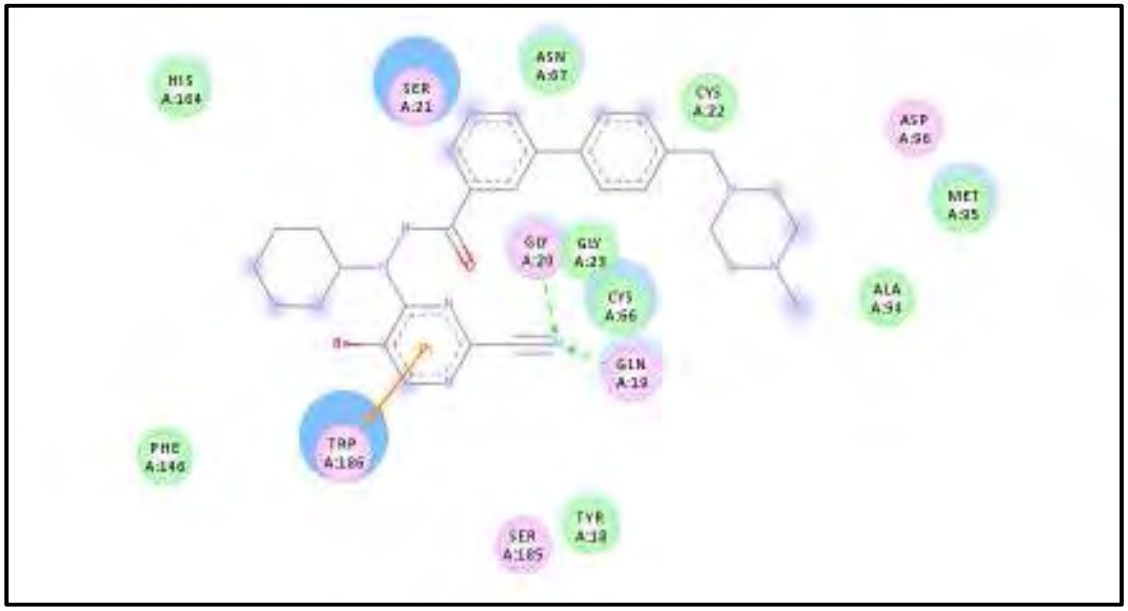
Appendix 2B-8: KP-2 active site residues interacting with 2-cyanopyrimidine derivative 2_CPI.



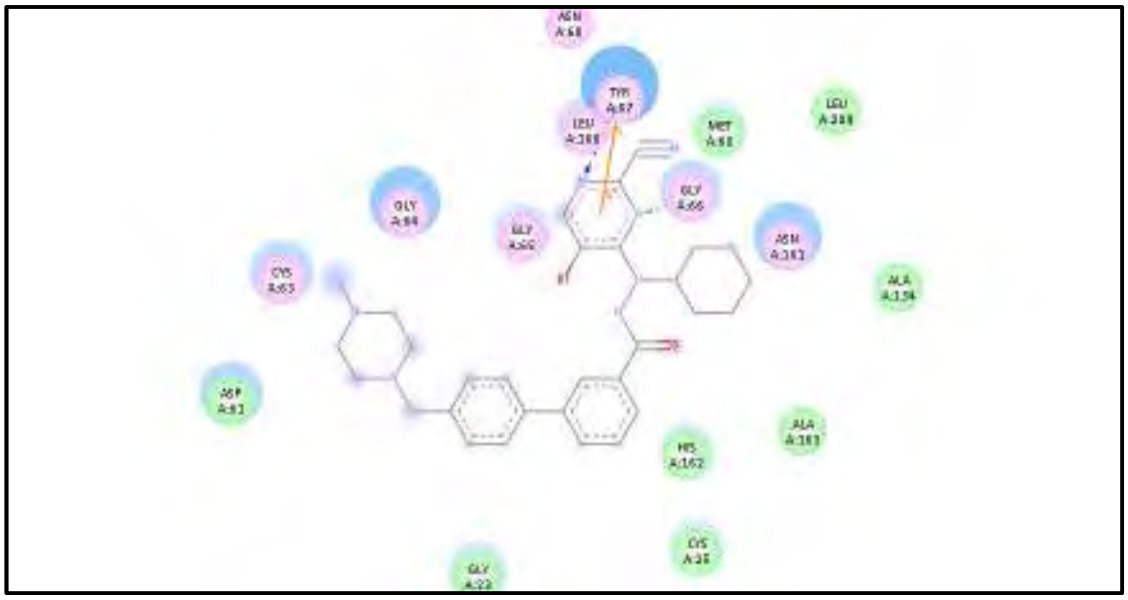
Appendix 2B-9: KP-3 binding site residues interacting with 2-cyanopyrimidine derivative 2_CPI.



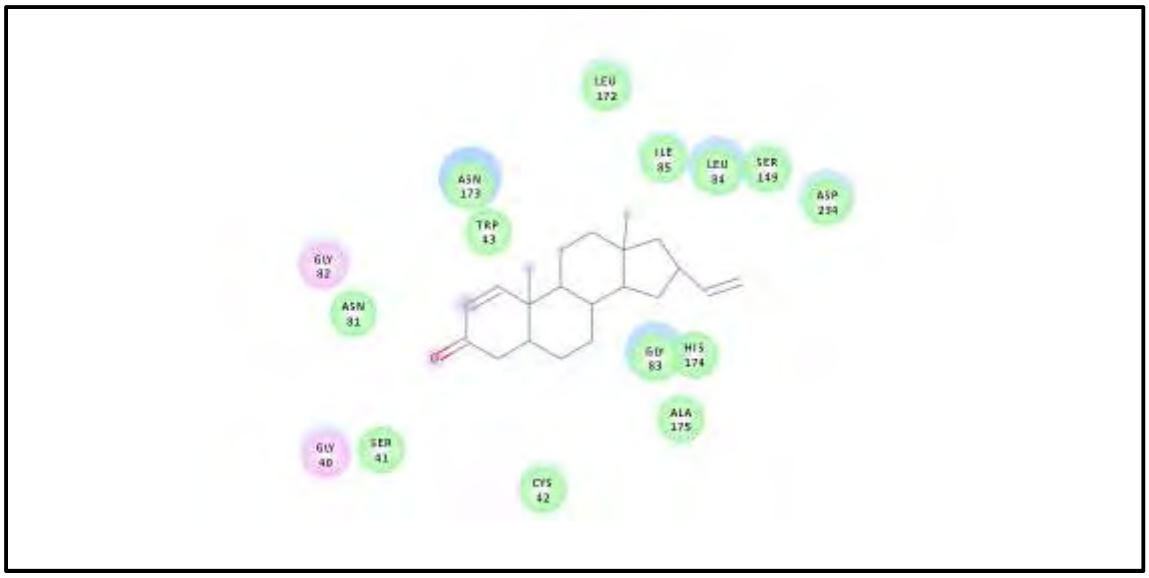
Appendix 2B-10: Cathepsin L active site residue interactions with 2-cyanopyrimidine derivative 2_CPI.



Appendix 2B-11: Cathepsin S active site interactions with 2-cyanopyrimidine derivative 2_CPI.



Appendix 2B-12: Cathepsin K active site interactions with 2-cyanopyrimidine derivative 2_CPI.



Appendix 2B-13: FP-2 with best docked compound 5α-pregna-1,20-dien-3-one. The compound had fewer Vander Waals, polar, ionic and electrostatic interactions compared to the validation set of 2-cyanopyrimidine.