MALARIAL DRUG TARGETS: CYSTEINE PROTEASES AS HEMOGLOBINASES

By

FORTUNATE MOKOENA

Submitted in partial fulfillment of the requirements of the degree

Magister Scientiae

In the

Faculty of Science

Department of Biochemistry, Microbiology and Biotechnology

Rhodes University

Grahamstown

SUPERVISOR:

Dr. A.Ö. Taştan Bishop

CO-SUPERVISOR:

Dr. B.J. Vorster

April 2012

Table of Contents

Acknowledgements

I would like to express my sincere words of gratitude and many appreciations to:

- The Lord God almighty, for being with me always.
- My parents, for their support, encouragement, and prayers.
- Dr. Özlem Taştan Bishop and Dr. Juan Vorster for allowing me to pursue this project, their kindness, patience, guidance, support and constructive criticism.
- My laboratory colleagues at Rhodes Bioinformatics Unit (RUBi).
- Students at the University of Pretoria Bioinformatics Unit.
- Prof. Fourie Joubert, for generously allowing me to work in his laboratory.
- Prof Greg Blatch, Dr T.A.P de Beer and Dr P.B. Burger for their guidance with challenges encountered throughout my studies.
- Acceryls Discovery studio team, for patiently answering my questions and assistance with software.
- National Research Foundation (NRF) and Rhodes University, for their financial support.

Typographical conventions

Residues are referred to using the standard three letter code and in some cases one letter code followed by residue number:

Amino Acid (AA)	Three-letter code	One-letter code
Alanine	ALA	A
Aspartic acid	ASP	D
Cysteine	CYS	C
Glutamic acid	GLU	E
Glycine	GLY	G
Histidine	HIS	H
Isoleucine	ILE	I
Lysine	LYS	K
Leucine	LEU	L
Methionine	MET	M
Phenylalanine	PHE	F
Asparagine	ASN	N
Proline	PRO	P
Glutamine	GLN	Q
Arginine	ARG	R
Serine	SER	S
Threonine	THR	T
Valine	VAL	V
Tryptophan	TRP	W
Tyrosine	TYR	Y

List of figures

List of tables

List of abbreviation

Å	Angstrom
AA/aa	Amino acids
ACT	Artemisinin-based combination therapies
CAPRI	Critical assessment of predicted Interactions
DE	Desolvation energy
DOPE	Discrete optimized protein energy
DS	Discovery studio
E64	L-*trans*-epoxysuccinyl-leucylamido(4-guanidino) butane
FFT	Fast fourier transform
GDT_TS	Global distance test total score
GSC	Grid-based shape complimentarity
HMM	Hidden markov model
ITNs	Insecticide-treated nets
IVM	integrated vector management
Kb	kilobase
kDa	kilo Dalton
MSA	Multiple sequence alignment
NMR	Nuclear magnetic resonance

PSC	Pairwise shape complimentarity
RMSD	Root mean square deviation
SDS-PAGE	Sodium dodecyl sulphate-polyacrylamide gel Electrophoresis

List of computer-related terms

BLAST	Basic local alignment search tool
CDD	Conserved domain database
CHARMm	Chemistry at Harvard molecular mechanics
ClustalX	Cluster alignment (for X windows)
COG	Clusters of orthologous groups of proteins
EBI	European Bioinformatics Institute
FASTA	Fast alignment
GONNET	Amino acid substitution matrix
HHpred	Homology detection prediction
MetaMQAP	Model quality assessment program
MODELLER	Homology modeling based on satisfaction of spatial restraints
MUSCLE	Multiple sequence comparison by log-expectation
NCBI	National Center for Biotechnology Information
PDB	Protein Data Bank
Pfam	Protein families database
PSI-BLAST	Position specific iterative BLAST
PSIPRED	Secondary structure prediction server

PROCHECK	Structure validation program
ProSA	Protein structure analysis program
PSIPRED	Secondary structure prediction server
SCOP	Structural classification of proteins
SMART	Simple modular architecture research tool

Abstract

Malaria has consistently been rated as the worst parasitic disease in the world. This disease affects an estimated 5 billion households annually. Malaria has a high mortality rate leading to distorted socio-economic development of the world at large. The major challenge pertaining to malaria is its continuous and rapid spread together with the emergence of drug resistance in *Plasmodium* species (vector agent of the disease). For this reason, researchers throughout the world are following new leads for possible drug targets and therefore, investigating ways of curbing the spread of the disease. Cysteine proteases have emerged as potential antimalarial chemotherapeutic targets. These particular proteases are found in all living organisms, *Plasmodium* cysteine proteases are known to degrade host hemoglobin during the life cycle of the parasite within the human host. The main objective of this study was to use various *in silico* methods to analyze the hemoglobinase function of cysteine proteases in *P. falciparum* and *P. vivax.* Falcipain-2 (FP2) of *P. falciparum* is the best characterized of these enzymes, it is a validated drug target. Both the three-dimensional structures of FP2 and its close homologue falcipain-3 (FP3) have been solved by the experimental technique X-ray crystallography. However, the homologue falcipain-2 (FP2')' and orthologues from *P.vivax* vivapain-2 (VP2) and vivapain-3 (VP3) have yet to be elucidated by experimental techniques. In an effort to achieve the principal goal of the study, homology models of the protein structures not already elucidated by experimental methods (FP2', VP2 and VP3) were calculated using the well known spatial restraint program MODELLER. The derived models, FP2 and FP3 were docked to hemoglobin (their natural substrate). The protein-protein docking was done using the unbound docking program ZDOCK. The substrate-enzyme interactions were analyzed and amino acids involved in binding were observed. It is anticipated that the results obtained from the study will help focus inhibitor design for potential drugs against malaria. The residues found in both the *P. falciparum* and *P. vivax* cysteine proteases involved in hemoglobin binding have been identified and some of these are proposed to be the main focus for the design of a peptidomimetric inhibitor

Chapter 1

1.Literature review

The disease malaria is discussed as well as its social and economic importance. The global frustrations caused by malaria parasite drug resistance are highlighted. The significance of the study is indicated by reviewing cysteine proteases which are involved in life cycle and pathogenicity of the parasite and therefore, promising targets enzymes for new anti-parasitic drugs.

1.1 Malaria

Malaria is one of the most prevalent and transmittable diseases contributing to global mortality and morbidity (Gaudalupe and Rodriquez, 2007). The greatest burden of malarial infections is borne by pregnant women and young children in sub-Saharan Africa, where the disease is increasingly implicated in social and economic impoverishment (Breman, 2001). Five species of *Plasmodium* are responsible for malaria in humans: *Plasmodium falciparum, P. vivax*, *P. malariae, P. ovale* and *P. knowlesi* (Schofield and Grau, 2005). Though the aforementioned species of *Plasmodium* infect human, the effects of *P. knowlesi* and *P. malaria* are less pronounced when compared to *P. falciparum* and *P. vivax.* Therefore, *P. falciparum* has received a lot of attention primarily because it is the most lethal and accounts for the most malarial infections (Gilles, 1985). Due to the severity of the disease and failure of malaria control strategies adopted in the past, there is a general consensus that significant reduction in the malaria burden will require the co-ordinated use of several strategies, including artemisinin-based combination therapies (ACTs), integrated vector management (IVM) including

insecticide-treated nets (ITNs) and better diagnostic and effective treatment (McKenzie *et al*., 2002).

1.1.1 Malaria control

Malaria has traditionally been managed in two ways: controlling anopheline mosquito vectors and effective case management (White, 2004). The former has been achieved by implementing approaches such as the removal of mosquito breeding sites, using insecticides and hindering mosquitoes from human contact (Trongtokit *et al*., 2005). The prevention of mosquito and human contact is established via the use of screens and bed nets, particularly those impregnated with insecticides (Zimmerman and Voorhman, 1997). Case management, on the other hand, has largely relied on antimalarial drugs (Huthmacher *et al*., 2010). The most used and widespread drugs are chloroquine and sulfadoxine-pyrimenthamine because they are cheap and slowly eliminated from the body (Lederman *et al*., 2006). The absence of vaccines and drug resistance has complicated the process of malaria control. There is therefore an urgent need for new antimalarial drugs (Kremsner and Krishna, 2004). Drug resistance has been reported in new areas and re-emerged in areas where the disease had previously been eradicated. The occurrence and harshness of malaria epidemics in certain parts of the world are attributed to antimalarial drug resistance (Boland, 2001).

1.1.2 The cause and life cycle of malaria

Human malaria is caused by infections from the intracellular parasite which belongs to the *Plasmodium* genus. The parasite is transmitted to human hosts via the *Anopheles* mosquito vector (Gardner *et al*., 2002). There are a few minor variations between various species but all *Plasmodium* species causing human malaria cases exhibit similar life cycles (Singh *et al*., 2004, Wellems *et al*., 2009). *P. falciparum* is the most virulent *Plasmodium* species as its infections are

associated with high levels of parasitemia. The other four *Plasmodium* species cause milder infections. There are cases where relapse has occurred within a few months to several years after transmissions from *P. vivax* and *P. ovale*, mostly due to the fact that appropriate treatment was not obtained (Michon *et al*., 2007). The *Plasmodium* life cycle occurs between two hosts: the mosquito vector and the human. The transition between the cold-blooded mosquito host and the warm-blooded human host occur during sexual and asexual stages respectively. The asexual phase is further divided into two stages: the liver (pre-erythrocytic) stage and the blood (erythrocytic) stage (Figure 1.1).

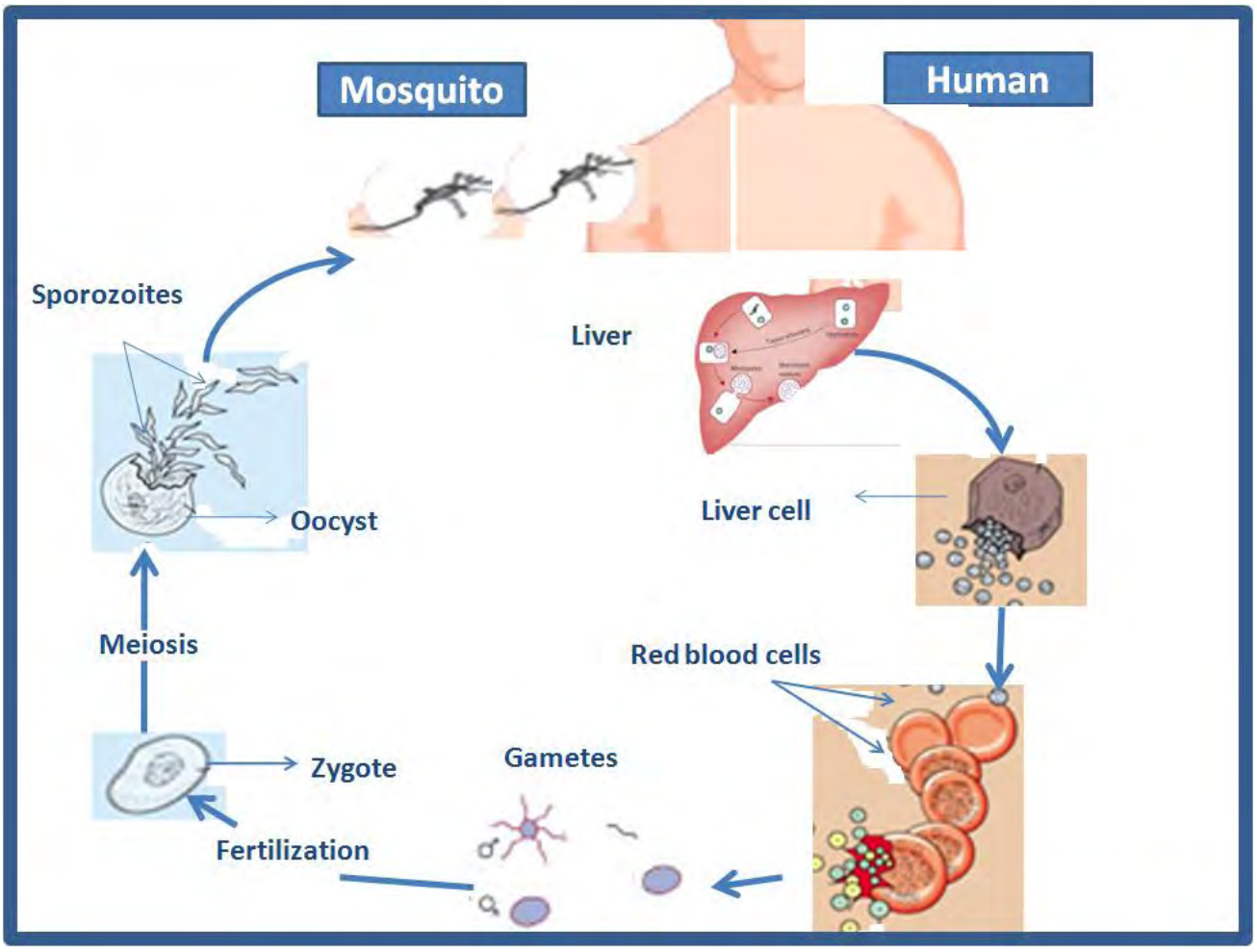

Figure 1. 1: The life cycle of the malaria parasite. The sexual stage takes place in the mosquito host (The exact life cycle details not shown in the figure) and asexual stage which takes in human host. Figure adapted from Mueller *et al.,* 2009 and Teixeira *et al*., 2011.

As indicated in Figure 1.1, the start of the malaria parasite life cycle is with a female mosquito biting the vertebrate host. As a consequence of the biting, sporozoites get transmitted into the host (Enayati *et al*., 2007). The sporozoites are then carried through the circulatory system where the infection of liver cells occurs. This stage is characterized by the intracellular parasite undergoing an asexual replication stage known as the exo-erythrocytic schizogony. Merozoites are then released into the blood circulation, where they infect the red blood cells and simultaneously undergo a trophic phase resulting in the enlargement of the parasite. The early trophozoite stage is usually called the ring form, a name pointing to the morphology of the parasite at this stage. Once trophozoites are enlarged, the parasite metabolism becomes activated; they ingest the host cytoplasm and cleave host hemoglobin into amino acids via a proteolytic process (Dorin-Semblat *et al*., 2008). Merozoites are released from the rupture of the infected erythrocyte (Fujioka and Aikawa, 2002). Erythrocyte invasion results in another round of blood stage replicative stage. Alternatively, the schizonts differentiate into male and female gametocytes. The gametocytes must be picked by another mosquito to complete the life cycle, in which they develop into a zygote. The zygote elongates and become motile (ookinetes). Ookinetes invades the midgut wall of the mosquito and develop into oocysts. Sporozoites develop from grown and ruptured oocysts (Vlachou *et al*., 2004; Baton and Ranford-Cartwright, 2005). The sporozoites migrate to the mosquito's salivary glands. The *Plasmodium* life cycle is then indefinitely continued by another infected mosquito biting a human during a blood meal. Thus the main focus of the study was to investigate the mechanism by which *Plasmodium* cysteine proteases degrade human host hemoglobin. In the following sections, cysteine proteases from *P. falciparum* and *P. vivax* are discussed (section **1.2**).

1.2 Cysteine proteases

1.2.1 Protease enzymes

Proteases refer to a group of enzymes which are found in all living organisms (Sajid and McKerrow, 2002). The primary role of proteases is to catabolically hydrolyze the peptide bonds which link amino acids in a protein molecule or a polypeptide chain. DNA replication, cell signalling, immunity and apoptosis (Burleigh and Soldati-Favre, 2008) are amongst the many roles proteases play in living organisms. Furthermore, proteases are involved in biologically important functions such as the activation of proenzymes, the liberation of physiologically active peptides, the inflammation and digestive system processes (Brown *et al*., 2000). These biologically important molecules range in sizes between 10 kilodalton (kDa) monomers to several thousand kDa multimeric complexes (Sajid and McKerrow, 2002).

The classification of proteases is largely based on where they cleave the peptides or proteins. Proteases are classified into two main groups: endopeptidases and exopeptidases. Endopeptidases cleave within a polypeptide whereas exopeptidases cleave the ends of a polypeptide chain. Exopeptidases cleaving the C-terminal and N-terminal of the substrate polypeptides are called carboxypeptidases and aminopeptidases respectively (Lecaille *et al*, 2002). Proteases belong to six major classes: metallo, serine, aspartic, threonine, cysteine and glutamic acid (Barret *et al*., 1998). The classification of proteases into specific groups is usually based on the amino acid residue at the active site, thus for aspartic proteases, ASP is used for catalytic activity. CYS, GLU, THR and SER are the amino acids used for the catalytic activities of cysteine, glutamic acid, threonine and serine proteases respectively. Substrate specificity and the catalytic mechanism of peptide hydrolysis is another basis on which proteases are classified. The essential amino acid residue at the active site, similarities in amino acid sequences, the optimum pH ranges for activity and inhibitor binding/similiarity (Bode and

Huber, 1992) are some of the other properties used for classifying and catergorizing proteases' in specific classes.

1.2.2 Cysteine proteases' characteristics and function

Interest in cysteine proteases goes back 10 years when their roles in life cycle and pathogenicity of several parasitic organisms were discovered. Cysteine proteases are adaptive enzymes; they adjust well to different biological environments and substrates. Parasite derived cysteine proteases play important roles in cell, tissue and immune evasions, activation of enzymes, hatching and molting (Sajid and McKerrow, 2002). There is a high level of similarity between cysteine and serine proteases, they share similar mechanical features principally because of the amino acid involved in catalysis. However, cysteine proteases have better nucleophiles than serine proteases because its sulphur containing amino acid (cysteine) offers a better center for catalysis. Thus, the catalytic activities of cysteine proteases are carried out by a cysteine residue which has an extra shell of electrons in the sulfur of the thiol group (Figure 1.2). Cysteine proteases are also known as thiol or sulfhydryl proteases, these names were derived following the property and activity of cysteine in the proteolysis of substrates and inhibitors.

The active site of cysteine proteases consist of the highly conserved cysteine, histidine and asparagine residues (papain numbering CYS^{25}, $HIS^{159}$and ASN^{175}). The proteolytic activity of cysteine proteases initializes with the formation of an ion pair between CYS and HIS, this ion pair is stabilized by hydrogen bond from ASN (Lecaille *et al*., 2002). The resulting close proximity between CYS and HIS enhances the nucleophillic attack of CYS which makes the cysteine residue stable even prior to substrate binding, therefore these proteases are regarded as priori activated enzymes (Polgar and Halasz, 1982). They interact with substrates and peptides (inhibitors) via three main processes (which are shown in Figure 1.2): hydrolysis (A), acylation (B) and deacylation (C). Upon interaction with a substrate or peptide, the nucleophilic thiolate cysteine attacks the carbonyl carbon of the substrate or peptide scissile bond (Lecaille

et al., 2002; Teixeira *et al*., 2011). This forms a tetrahedral intermediate which is stabilized by the oxyanion hole. The tetrahedral intermediate is then transformed into an acyl enzyme (enzyme-substrate thiol ester) and the C-terminal portion of the substrate or peptide is released in a process called acylation (B) (Lecaille *et al*., 2002). After this step, the acyl enzyme is hydrolyzed by water (A) and it forms a second tetrahedral intermediate which splits into the free enzyme and N-terminal protein of the substrate in a process called deacylation (C). The active site of cysteine proteases are generally known to be the representative prime targets for therapeutic intervention (Lecaille *et al*., 2002). Like other proteases, the substrate specificity of cysteine proteases is probed by their interaction with peptides and irreversible inhibitors. Falcipains are the most widely studied and best characterized cysteine proteases. The active site of falcipains and vivapains are located within the four substrate binding pockets: S1, S2, S3 and S1' (Sajid and McKerrow, 2002). The interaction of the substrate binding site with substrate or inhibitors is not merely based on affinity but also the sum of the contribution from all fragments of the protease (Turk *et al*., 1998). Much of the information of the protease substrate binding site is based on the kinetics and crystal structures of substrate mimicking inhibitors bound to the enzyme's active site (Turk *et al*., 1998).

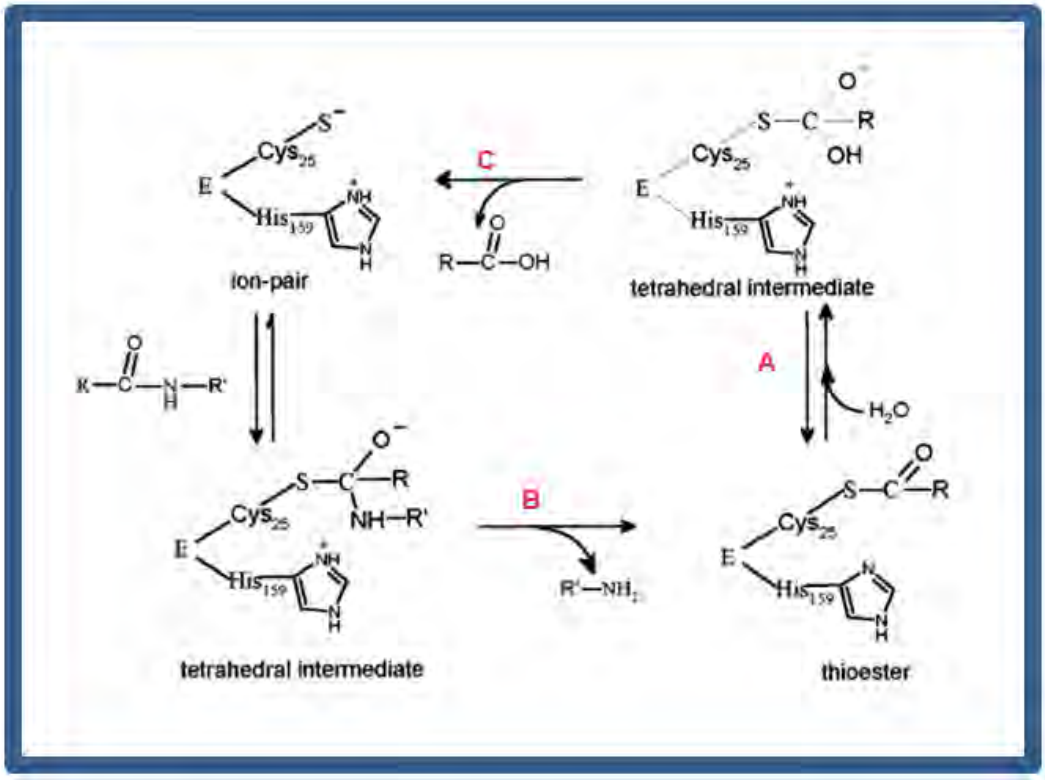

Figure 1. 2: Diagrammatic representation of the active site of the catalytic mechanism of cysteine proteases, Figure adapted from Lecaille *et al*., 2002. A shows the hydrolysis process, B the acylation, C-deacylation and E shows the peptide or substrate and R the side chain.

1.2.3 Classification and evolution of cysteine proteases

Cysteine proteases are divided into clans which are further differentiated into families. Clans are characterized by the fact that they are labeled by the letter C, followed by a letter (Rosenthal, 2004). One other feature of proteases within a specific clan is that they do not necessarily share sequence or structural identity and there is a likelihood that they arose independently. However, they share the function of cysteine residue in terms of hydrolytic cleavage of peptide bonds (Rosenthal, 2004). Cysteine proteases consist of clan A, B, C and D which are papain-like, viral and legumain-like respectively (Sajid and McKerow, 2002). Clan CA is the largest; it is often called papain-like (Lecaille *et al*., 2002).

Papain-like cysteine proteases derive their name from the papaya fruit (*Carica papaya*) which was the first protease to be purified and characterized. Since the identification of papain, other cysteine proteases with sequence similiarity to it have loosely been called papain-like (Barrett, 1994; Rawlings and Barrett, 1994; Sajid and McKerrow, 2002). Papain-like cysteine proteases are widely expressed in all organisms (Shindo and Van Der Hoorn, 2008) and they have been identified in animals, plants, viruses and bacteria (Figure 1.3). Mammalian papain-like cysteine proteases were neglected until recent years, when their importance in the pharmaceutical industry was recognized. Extracellular matrix turnover, antigen presentations and processing events are the roles of mammalian cysteine proteases which have made them drug targets. A study of the pathology and physiology of mammalian cysteine proteases has aided the design of selective therapeutic agents (Lecaille *et al*., 2002). Mamalian papain-like cysteine proteases have also been identified as viable drug targets for diseases such as osteoporosis, arthritis, immune-related diseases and cancer (Kempson *et al*., 1973).

Parasitic cysteine proteases are another family of papain-like proteases that have been widely studied and indeed the most characterized (Sajid and McKerrow, 2002). They are classified into family C1 (cathepsin B and cathepsin L-like) and family C2 (calpain-like) (Sajid and McKerrow, 2002; Rosenthal, 2004). As indicated by Figure 1.3., parasite-derived cysteine proteases are involved in the growth, development and replication of the parasite itself. The roles of parasitic cysteine proteases include involvement in tissue/skin penetration and host organism invasion. They induce diseases such as Chagas' diseases, malaria and other parasitic infections (Redzynia *et al*., 2009). The functions of many cysteine proteases have been identified using their inhibitors.

1.2.4 General features of papain-like cysteine proteases

All papain-like cysteine proteases are synthesized in the endoplasmic reticulum and are expressed as proenzymes. They all have the following features in common: a signal peptide, a propeptide (prodomain) and a catalytic domain which represents the mature proteolytically active enzyme (Lecaille *et al*., 2002). The signal peptide is responsible for translocating the peptide into the endoplasmic reticulum during protein expression. The length of the propeptide varies between different species: for example, it is about 36 amino acids long in human cathepsin X and 315 amino acids long in *P. falciparum* falcipain-1. Prodomains have three functions; they act as a scaffold protein folding into a catalytic domain (Wiederanders, 2000), they act as a chaperone transporting the proenzyme to the endosomal-lysosomal compartment (Schilling *et al*., 2001; Yamamoto *et al*., 1999) and prevent premature activation of the catalytic domain by high affinity reversible inhibition (Fox *et al*., 1992). The propeptide runs through the substrate binding cleft in reverse and is less structured when compared to the domain structure of papain-like cysteine proteases.

A typical papain-like cysteine proteases catalytic domain is about 220-260 amino acids in length. Other studies have shown that some parasitic cysteine proteases have other unique features which increases the length of their catalytic domain (Sajid and McKerrow, 2002; Rosenthal, 2004). *Plasmodium* papain-like cysteine proteases have an additional N-terminal extension (colored red) and C-terminal insert (colored green) (Figure 1.4). Papain-like cysteine proteases fold into two domains left (L) and right (R) (Grzonka *et al*., 2001). A well conserved active site is found between the two domains, CYS (located at the structurally conserved α-helix of the L-domain), HIS (located at the R-domain) and ASN residues (as shown in Figure 1.4). The active sites of papain-like cysteine proteases are found within the substrate binding pockets which are called subsites. Schechter and Berger (1967) was the first to describe enzyme subsites, stating that subsites in the N-terminal direction are named S1, S2, S3 and etc., while

S1', S2', S3' and Sn' are in the C-terminal direction, as indicated in Figure 1.4. The mechanism of hydrolysis for papain-family cysteine proteases is well documented. Their main mode of interaction with their substrates occurs at the subsites (Turk *et al*., 1998). Residues in the protease backbone and side-chain mainly bind through interaction with the S2, S1 and S1' binding pockets, while the S3 and S2' subsites are crucial in amino acid side-chain binding. Papain-like cysteine proteases are also known for their preference to bind LEU residue at P2 (Lecaille *et al*., 2002; Submanian *et al*., 2009). For the falcipains, the S1 subsite is the least characterized of the four grooves, it usually includes a glutamine for the oxyanion hole. Similar to other papain-like cysteine proteases, the S2 subsite is the most characterized, particularly its specificity towards substrates with a LEU-residue (Shenai *et al*., 2000, Sijiwali *et al*., 2004). The S1' subsite contains a highly conserved tryptophan which is known to interact with peptide from the substrate or inhibitor through hydrophobic interactions and the S3 groove contains a highly conserved glycine rich region. There are also other amino acids which are highly conserved in papain-like cysteine proteases: those forming the disulfide bridge and PRO2 (papain-numbering) whose role is to prevent premature activation of the mature proteases. PRO2 interacts with the aminopeptidases by truncating the N-terminal and this prevents inactivation of the mature protease as well (Lecaille *et al*., 2002). There is also the GLY-PRO motif which separates the α and β domains at the interface between L and R domains, this motif is highly conserved across the papain-like cysteine proteases (Lecaille *et al*., 2002). Most of the studies on cysteine proteases have focused on the interactions between the subsite residues and small ligands and not much work has been conducted on the protein-protein interactions. In the case of falcipain-2 (2OUL and 1YVB), protein-protein complex structures were co-crystallized with protein inhibitors (chagasin and cystatin respectively). The binding of LUE residues in the S2 subsite was observed.

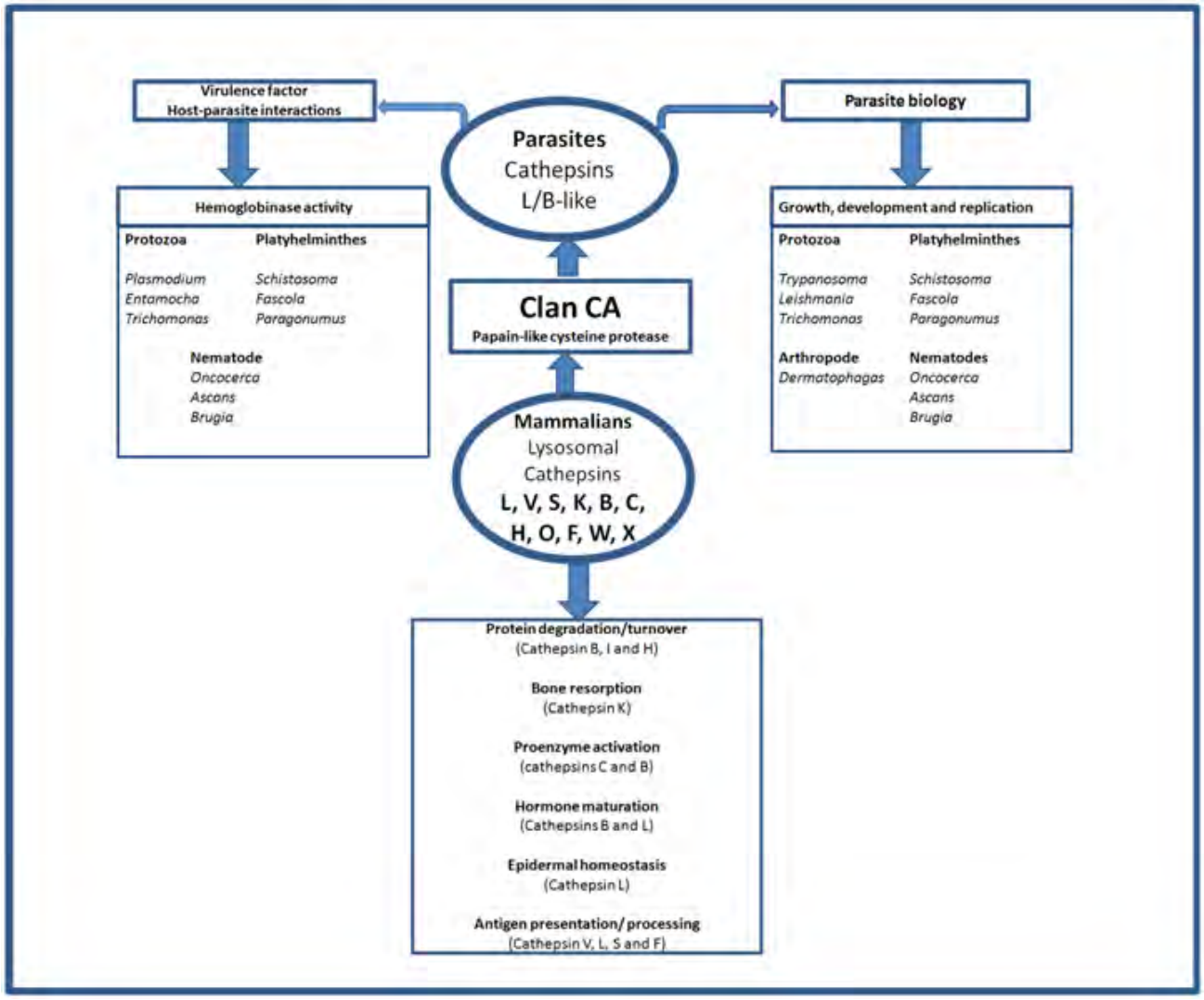

Figure 1. 3: The roles of human and parasitic cysteine proteases, figure adapted from Lecaille *et al*., 2002.

Clan CA proteases are also characterized by their sensitivity to general cysteine protease inhibitors. Small peptides, peptimimetics, isoquinolines, thiosemicarbones and chalcones are some of the papain-like cysteine proteases inhibitors (Ettari *et al*., 2009). These inhibitors are able to reversibly and irreversibly activate the enzymes. Generally all papain-like cysteine proteases are sensitive to E64 (L-*trans*-epoxysuccinyl-leucylamido (4-guanidino) butane) and they have a substrate specificity defined by the S2 pocket (Sajid and McKerrow, 2002; Rosenthal, 2004).

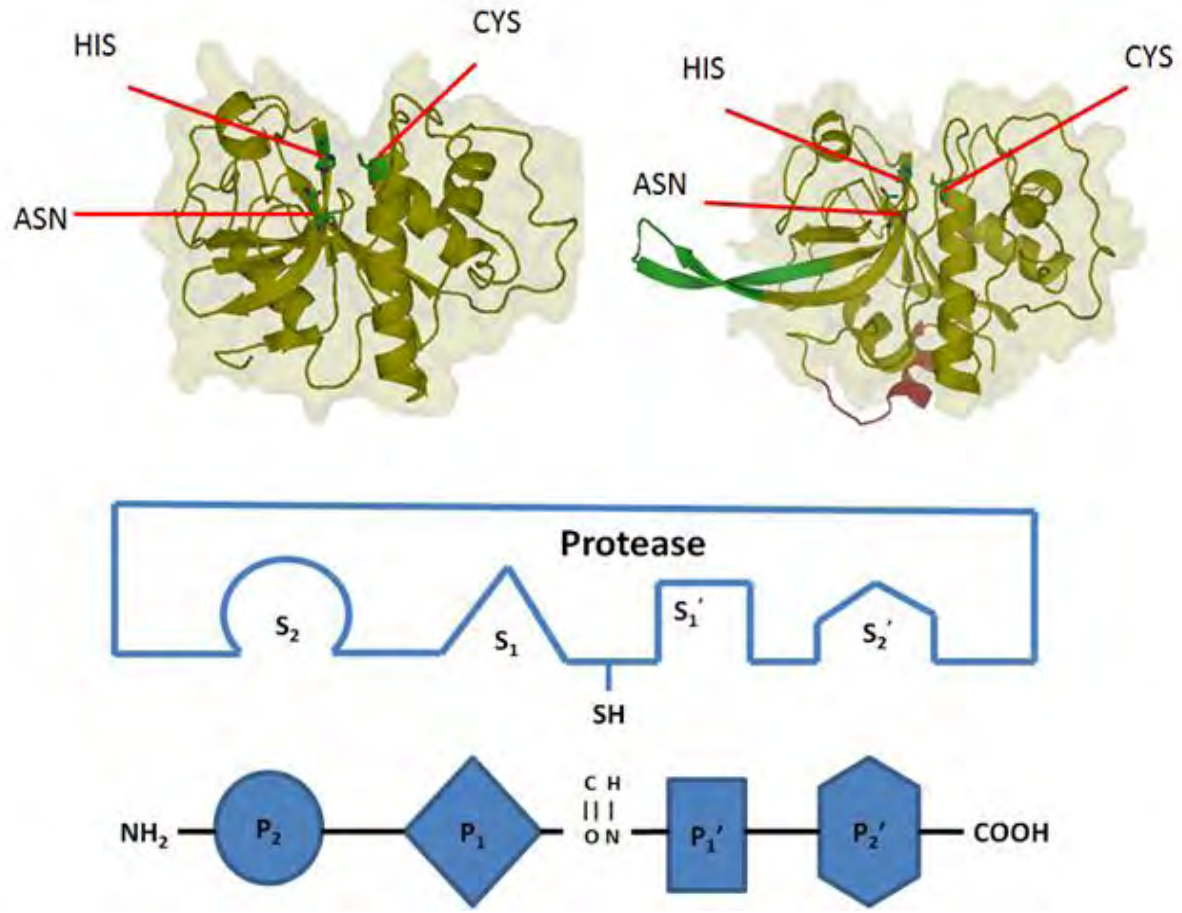

Figure 1. 4: The top part: Overall structures of papain (1PPP) on the left and falcipain-2 (1YVB) indicating the active site residues CYS, HIS and ASN. The unique features obtained in *Plasmodium* cysteine proteases are colored in red for the N-terminal extension and green for the C-terminal insert. Bottom: Representation of the interaction between cysteine proteases and substrate. Amino acids residues from the peptide are denoted "P" and the protease "S". Figure adapted from Sajid and McKerow, 2002.

1.3 *Plasmodium* cysteine proteases

There have been extensive *in vivo* and *in vitro* studies regarding the roles of *Plasmodium* cysteine proteases using their specific inhibitors. These studies have identified three key functions of *Plasmodium* cysteine proteases: hemoglobin hydrolysis, erythrocyte rupture and erythrocyte invasion by malaria parasites (Rosenthal, 2004). Analysis of the *P. falciparum*

genome has resulted in the identification of three papain-like cysteine proteases: falcipains, dipeptidyl peptidase, a calpain homolog, and serine-repeat antigens.

1.3.1 Roles of cysteine proteases from inhibitor studies

The scientific community and the world at large has come to appreciate protease inhibitors, largely because of their potential as drug targets for diseases like AIDS, malaria and cardiovascular related illnesses (Rosenthal *et al*., 2002). General cysteine protease inhibitors have been used to analyze their roles both *in vitro* and *in vivo*. For the *in vivo* studies, animal models of malaria were tested in order to confirm that inhibitors have potent antiparasitic activity, oral bioavailability, safety and pharmacokinetic properties (Rosenthal *et al*., 2002). The major challenge encountered from *in vivo* analysis is that *P. falciparum* can only be studied in a few primate species and therefore only murine models have been used so far. However, despite the limitations of murine models, cysteine protease inhibitors demonstrated *in vivo* antimalarial effects. Fluoromethyl ketone was needed in high doses but it cured malaria in 80% of *P. vinkei* infected mice (Rosenthal *et al*., 1993), while vinyl sulfone cured 40% of mice which were infected by *P. vinkei* (Palmer *et al*., 1995) These two compounds are relatively poor inhibitors of vinkepain-2, as compared to their inhibition of falcipain-2; therefore it can be assumed that improved inhibitors will generate better success (Rosenthal *et al*., 1993, Rosenthal *et al*., 2002). Peptidyl aldehyde and α-ketoamide are amongst the most promising cysteine protease inhibitors to be used for chemotherapeutic treatment. These inhibitors block the activity of falcipain-2 and falcipain-3 which prevents parasite development because it prevents hemoglobin degradation (Lee *et al*., 2003). Some of the small molecule inhibitors targeting the falcipains and other homologous cysteine proteases in other *Plasmodium* species include fluoromethyl ketones, vinyl sulfone, chalcones and phenothiazines. The compounds mentioned above have been argued to have inhibitory activity against the falcipains, the main support of this argument was provided by the evidence of undegraded hemoglobin and

parasite development was halted during *in vivo* antimalarial activity assays (Rosenthal *et al*., 2002, Lee *et al*., 2003).

It should be noted that during the process of inhibitor design, two critical factors must be taken into consideration: (1) the side effects of the inhibitor and (2) the complimentary of the inhibitor to the target protease active site. Therefore based on this approaches of inhibitor design it has been observed that despite the extensive *in vivo* and *in vitro* work that has been done on cysteine proteases inhibitors, some could not access the food vacuole and were therefore rendered ineffective (Singh and Rosenthal, 2001; Singh and Rosenthal, 2004). Some cysteine protease inhibitors have been found to be effective against five strains of *P. falciparum* that differ widely in their sensitivities towards standard antimalarial agents, this observation suggest that they will not result in multidrug resistant parasites (Singh and Rosenthal, 2001). In the next section the principal roles of falcipains and vivapains: in hemoglobin degradation will be discussed.

1.3.2 Falcipains

Falcipains are the best characterized cysteine proteases of *P. falciparum*. They share sequence identity and several features with papain-like cysteine proteases. There are four falcipains: falcipain-1 (FP1) (Sijiwali *et al*., 2004), two nearly identical proteases falcipain-2 (FP2) and falcipain-2' (FP2') also labeled FP2A and FP2B respectively (Shenai *et al*., 2000; Singh *et al*., 2006) and falcipain-3 (FP3) (Sijiwali *et al*., 2001). The falcipains are all expressed during the erthrocytic stage of the parasite (Rosenthal, 2004). FP1 is encoded on chromosome 14, while FP2, FP2' and FP3 are located on the 12 kilobase (kb) stretch of chromosome 11. FP2 shares 96% sequence identity with FP2', 68% identity with FP3 and 38% identity with FP1 (Rosenthal, 2004). And all four falcipains share less than 40% sequence identity with papain (Shown in Figure 1.5).

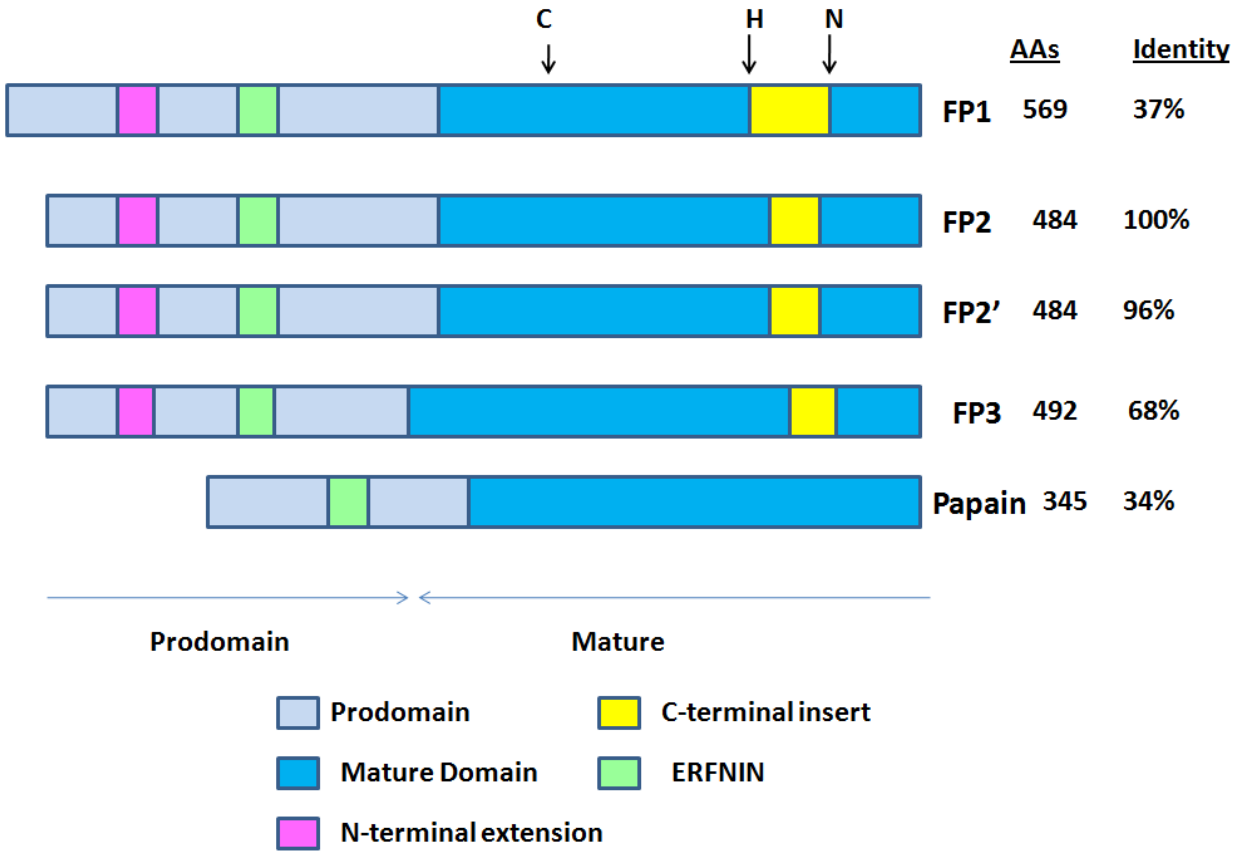

Figure 1. 5: Schematic representation of the falcipains. Their structural features, including the unusually large prodomain, mature domain, C-terminal and N-terminal insert together with the highly conserved ERFNIN motif are clearly labeled. The active site residues are labeled as C, H and N for cysteine, histidine and asparagines respectively. Figure adapted from Rosenthal, 2004.

FP1 is the least characterized of the four falcipains due to its low abundance and the lack of an efficient expression system to recombinatly produce this particular enzyme (Ettari *et al*., 2009). Earlier studies proposed that FP1 played a role in parasite invasion and was not involved in the degradation of hemoglobin (Greenbaum *et al*., 2001). However, a later study indicated that FP1 knockout had no effect on parasite development in the erythrocyte (Sijiwali *et al*., 2004). The findings from the latter study therefore suggested that FP1 is neither required for parasite invasion nor intracellular development within the erythrocyte. The role of FP1 was later revealed by gene disruption studies, which indicated its role in oocyst production during the development of the parasite in the mosquito gut (Eksi *et al*., 2004). However, FP1 does not play

an important role in asexual or gamete development (Eksi *et al*., 2004). Therefore, the current suggestion is that FP1 could be directly involved in the transition from gametocyte to oocyst by means of proteolytic processing which activates proteins (Sijiwali *et al*., 2004). If FP1 is secreted, it might be degrading the peritrophic matrix or midgut endothelium which facilitates the migration of the ookinete (Eksi *et al*., 2004).

FP2 and FP3 are the best characterized of the four falcipains and indeed the most studied. Hemoglobin degradation is an essential process for parasite survival within its host and one of the main focuses for drug development against malaria (Shenai *et al*., 2000). FP2 is the most abundant and widely studied of the two cysteine proteases (Hogg *et al*., 2005; Wang *et al*., 2006) and it often been called the principal hemoglobinase (Pandey *et al*., 2005). FP2 and FP3 share high sequence identity (68%), they both prefer substrates with LEU at P2 (Shenai *et al*., 2000; Sijiwali *et al*., 2001). However, they demonstrate different substrate specificities (Ramjee *et al*., 2006). Both FP2 and FP3 are localized in the food vacuole, where hemoglobin degradation occurs. FP2 is synthesized at the early trophozoite stage and is processed more quickly than FP3 which is expressed at the late trophozoite/early schizonts stage (Shenai *et al*., 2000; Sijiwali *et al*., 2001). Both cysteine proteases are synthesized as membrane-bound pro-enzymes whose activity is blocked by the prodomain and are further processed into soluble mature proteases (Sijiwali *et al*., 2002). FP2 and FP3 biochemical characterization led to the development of an efficient expression system for the two proteases. Both FP2 and FP3 have been biochemically expressed and purified in *E. coli* (Shenai *et al*., 2000; Sijiwali *et al*., 2001). This led to the ability to produce three dimensional crystal structures using X-ray crystallography of FP2 (Wang *et al*., 2006 (1YVB); Wang *et al*., 2007(2OUL), Hogg *et al* (2GHU), and 3BPF (Kerr *et al*., 2009) and FP3 (Kerr *et al*., 2009 (3BWK), and Kerr *et al*., 2009 (3BPM)). The falcipains have an acidic pH optimum, they require reducing conditions for optimal activity, and are inhibited by typical cysteine protease inhibitors (Rosenthal *et al*., 2002).

The cysteine protease inhibitor lactone antibiotic brefeldin A blocked the processing of FP2 and FP3, suggesting that it is trafficked to the food vacuole via the endoplasmic reticulum. FP2 and FP3 are also auto-hydrolyzed at a neutral pH before they arrive at the food vacuole (Dahl *et al*.,

2005). The functions of FP2 include degrading erythrocyte membrane skeletal proteins such as ankyrin and the band 4.1 protein (Hanspal *et al*., 2002). This occurs at pH optimum between 7.0 and 7.5, which is suspected to play a role in destabilizing the erythrocyte membrane, rupturing the host cell and releasing the merozoites (Shenai *et al*., 2004; Dahl *et al*., 2005). FP2 is a validated drug target as the disruption of its gene disrupted hemoglobin degradation (Sijiwali *et al*., 2004). However, the disruption of FP2 gene indicated that the loss of FP2 alone is not enough to kill the parasite, suggesting that other cysteine proteases are involved in parasite invasion and growth within the erythrocyte. Therefore, the parasite can be killed by deletion of multiple cysteine proteases. Unfortunately, the disruption of FP3 was not achieved, but replacement with a tagged copy was achieved more recently and indicated that FP3 played an important role as a hemoglobinase (Sijiwali *et al*., 2006).

FP2' was fully appreciated after the sequencing of *P. falciparum* genome in 2002 (Gardner *et al*., 2002), and is an almost identical copy of FP2 (96% sequence identity). This protease was expressed in a bacterial vector, and it has the same proteolytic activity as FP2 (Singh *et al*., 2006); except that recombinant FP2' only cleaves ankyrin but not band protein 4.1. FP2' has the same biochemical functions as FP2, such as pH optimum in the range of pH 5.5-7.0; requires reducing conditions and the same substrate preference (Shenai *et al*., 2000; Sijiwali *et al*., 2001, Singh *et al*., 2006). FP2' also has been predicted to have a hemoglobin degradation function (Jeong *et al*., 2006; Singh *et al*., 2006). Falcipains, except for FP1, have a hemoglobin degrading function and are involved in the conversion of proplasmepsins into active aspartic acid proteases (Drew *et al*., 2007). Since FP2 is expressed at the early trophozite stage, it might serve as a dominating maturase, as FP2' and FP3 are expressed at the late trophozoite/early schizont stage (Hogg *et al*., 2006).

- ***Structural features of the falcipains***

Falcipains have features which are specific to *Plasmodium* species and not present in other papain-like cysteine proteases. Also, the falcipains have a larger prodomain than other papain-

like cysteine proteases (a comparison of the prodomain length is indicated in Figure 1.6) which contains a membrane spanning sequence. The falcipains also have two unique motifs, an insertion of conserved residues near the carboxyl terminus and an extension on the N-terminus (Rosenthal and Nelson, 1992; Shenai *et al*., 2000; Sijwali *et al*., 2001b). The two unique motifs have been studied in more details in FP2, where the N-terminal extension has been labeled $FP2_{nose}$ and the C-terminal insert has been labeled $FP2_{arm}$ (Wang *et al*., 2006). The experimental studies which have been carried out to better characterize the two motifs have suggested that $FP2_{nose}$ could be involved in the activation of profalcipain into the mature falcipain enzymes (Sijiwali *et al*., 2002; Pandey *et al*., 2009). This is unique to the falcipains, as the activation of the pro-enzyme into mature proteases in papain-like cysteine proteases is done by the prodomain (Brown *et al*., 2000) and it appears that for the falcipains, the prodomain is not involved in this process. However, the mode and mechanism have yet to be elucidated (Wang *et al*., 2006). Deletion of 10 amino acids from the 14 amino acid C-terminal insert of FP2 resulted in negligible hemoglobinases activity (Pandey *et al*., 2005). In another study, a complex structure for the C-terminal motif bound to hemoglobin was generated, suggesting that this motif is involved in hemoglobin degradation by FP2, this is largely the reason why the C-terminal insert of FP2 is labeled $FP2_{arm}$ (Wang *et al*., 2006). However, the complex structure raises questions about the involvement of the active site in the hydrolysis of hemoglobin. The limitations of the two studies were that the mechanism by which the C-terminal insert might be degrading hemoglobin was not explained. In a study by Wang *et al*., 2006, where a protein-protein complex of FP2 and hemoglobin was generated, there was no interaction between hemoglobin and FP2 active site. However, the arm-like motif protruded far from the active site and this raises questions about how the degradation finally occurs at the active site. It is well known that in order for proteases to degrade their substrate, they must bind them at the active site (Sajid and McKerrow, 2002).

1.3.3 Vivapains

P.vivax is not as virulent as *P. falciparum* but it is the most wide spread. It causes 10-20% of all malarial deaths (Na *et al*., 2004; Desai and Avery, 2004). Vivapain-2 (VP2) shares 63% and 66% sequence identity to FP2 and FP3 respectively, while vivapain-3(VP3) shares 56% and 60% sequence identity with FP2 and FP3 respectively. Both these papain-like cysteine proteases have been recombinantly expressed in *E. coli* (Na *et al*., 2004). Vivapains share similar biochemical properties to the falcipains and require reducing conditions, have acidic pH optima and hydrolyze substrates with positively charged residues at P1 and LEU P2 positions (Na *et al*., 2004). Notably most functions of vivapains are inferred from falcipain studies due to their ability to hydrolyze hemoglobin at acidic conditions and erythrocytic membrane proteins (Na *et al*., 2004; Desai and Avery, 2004). Vivapains are potentially inhibited by several inhibitors of the falcipains, though VP2 is more sensitive to inhibitors (Na *et al*., 2004). The three dimensional structures (3D) of the vivapains have not yet been solved by experimental techniques, therefore 3D homology models have to be constructed in order to study these proteases.

1.4 Hemoglobin degradation

Hemoglobin, being the most abundant protein in the erythrocyte becomes completely degraded after parasite entry (Goldberg, 1990; Goldberg *et al*., 2005; Teixeira *et al*., 2011). Earlier studies have evidently demonstrated that the degradation of host hemoglobin provides a reservoir of amino acids for the synthesis of proteins by the parasite as the parasite has a limited capacity to synthesize its own amino acids (Sherman, 1979; Zarchin, 1986). Also, the amount of free amino acids within the erythrocyte is not sufficient for parasite survival (Rosenthal *et al*., 1988). Infected erythrocytes have a higher concentration of amino acids than uninfected erythrocytes, and the hemoglobin content of infected erythrocytes decreases significantly (25-75%) during the life cycle of the erythrocytic parasites. The composition of the amino acid pool in infected erythrocytes is similar to the amino acids of hemoglobin (Fulton *et*

al., 1956; Sherman and Tanigoshi, 1970). In other studies where the infected erythrocyte contained radiolabeled hemoglobin, the radiolabeled hemoglobin amino acids was found in the parasite's proteins (Fulton *et al*., 1956; Sherman and Tanigoshi, 1970; Theakston; 1970). It is apparent that the parasite relies on the host hemoglobin for its survival within the host, however, it appears that the degraded products of hemoglobin are not sufficient for the parasite's metabolic needs as hemoglobin is a poor source of amino acids such as methionine, cysteine, glutamine and glutamine and is not composed of isoleucine residues (Francis *et al.,* 1997). The degradation of hemoglobin releases the heme, which is detoxified by polymerization into a crystalline pigment called the hemozion (Goldberg *et al*., 1991). The hemoglobin degradation process occurs predominantly during the trophozoite stage of the erythrocytic parasite life cycle and is accompanied by erythrocyte cytoplasm ingestion (Rosenthal *et al*., 1998). In order to obtain the free amino acids from hemoglobin digestion, the parasites take up erythrocyte cystol via cytostome organelle and transport this material by vesicular trafficking to the food vacuole (Olliaro and Goldberg, 1995; Rosenthal and Meshnick, 1996). The food vacuole is an acidic lysosome-like organelle which contains three enzymes: aspartic, cysteine and metallo proteases, all of which are involved in hemoglobin degradation. These acidic proteases have been purified from the parasitized red blood extracts and some partially purified parasite extracts of the different *Plasmodium* species. The degradation of hemoglobin has been analyzed by Sodium dodecyl sulphate- polyacrylamide gel Electrophoresis (SDS-PAGE) in which the disappearance of the substrate (hemoglobin) has been monitored (Goldberg *et al*., 1990; Goldberg *et al*., 1991, Salas *et al*., 1995; Francis *et al*., 1996, Shenai *et al*., 2000). SDS-PAGE analyses have been a major breakthrough especially when characterizing the role of food vacuole proteases in the degradation of hemoglobin and also characterizing the roles of specific and non-specific inhibitors of each of the protease (Leung *et al.,* 2000; Bjelic *et al*., 2007). Native hemoglobin, digestive food vacuole lysate and inhibitors have been incubated and the effects of inhibitors have been analyzed by the appearance/disappearance of substrate on SDS-PAGE (Salas *et al*., 1995, Shenai *et al*., 2000, Hanspal *et al*., 2002).

The process of hemoglobin degradation is one that is controversial and currently there are two arguments suggested for the mechanism in which the digestion of hemoglobin in the food

vacuole occurs. The first argument proposes that the degradation of hemoglobin is a highly ordered process (Gluzman *et al*., 1994). The current understanding of this process is that hemoglobin is processed in the food vacuole where it is digested into small peptides. The small peptides are then transported to the cytosol, where additional processing of the globin fragment into free amino acids takes place (Kolakovich et al., 1997, Francis et al., 1997). Biochemical evaluation of the parasite biology has resulted in the observation that aspartic (Goldberg *et al*., 1991; Francis *et al*., 1997; Gluzman *et al*., 1994; Hill *et al*., 1994), cysteine (Rosenthal *et al*., 1988; Rosenthal and Nelson, 1992; Salas *et al*., 1995; Shenai *et al*., 2000; Sijiwali *et al*., 2001) and metallo (Eggleson *et al*., 1999; Gavigan *et al*., 2001) proteases are involved in the digestion of hemoglobin in an orderly fashion. Figure 1.6, points mainly to the current dogma of the ordered pathway suggestion for hemoglobin degradation. In this process the breakdown of hemoglobin is initiated by two aspartic acid proteases plasmepsin (PM) I and II. Secondary acidic proteases, PM IV and histo-aspartic protease (HAP) and cysteine proteases (FP2, FP2'and FP3) follow the cleavage of hemoglobin by PM I and II by cleaving the unraveled hemoglobin even further. Metallo-proteases then cleave the hemoglobin fragments from secondary cleavage into individual amino acids (Goldberg *et al*., 1990). The metallopeptidase, falcilysin, has been shown to have negligible activity against either native nor denatured hemoglobin but it readily destroys the peptide fragments of hemoglobin (Eggleson *et al*., 1999).

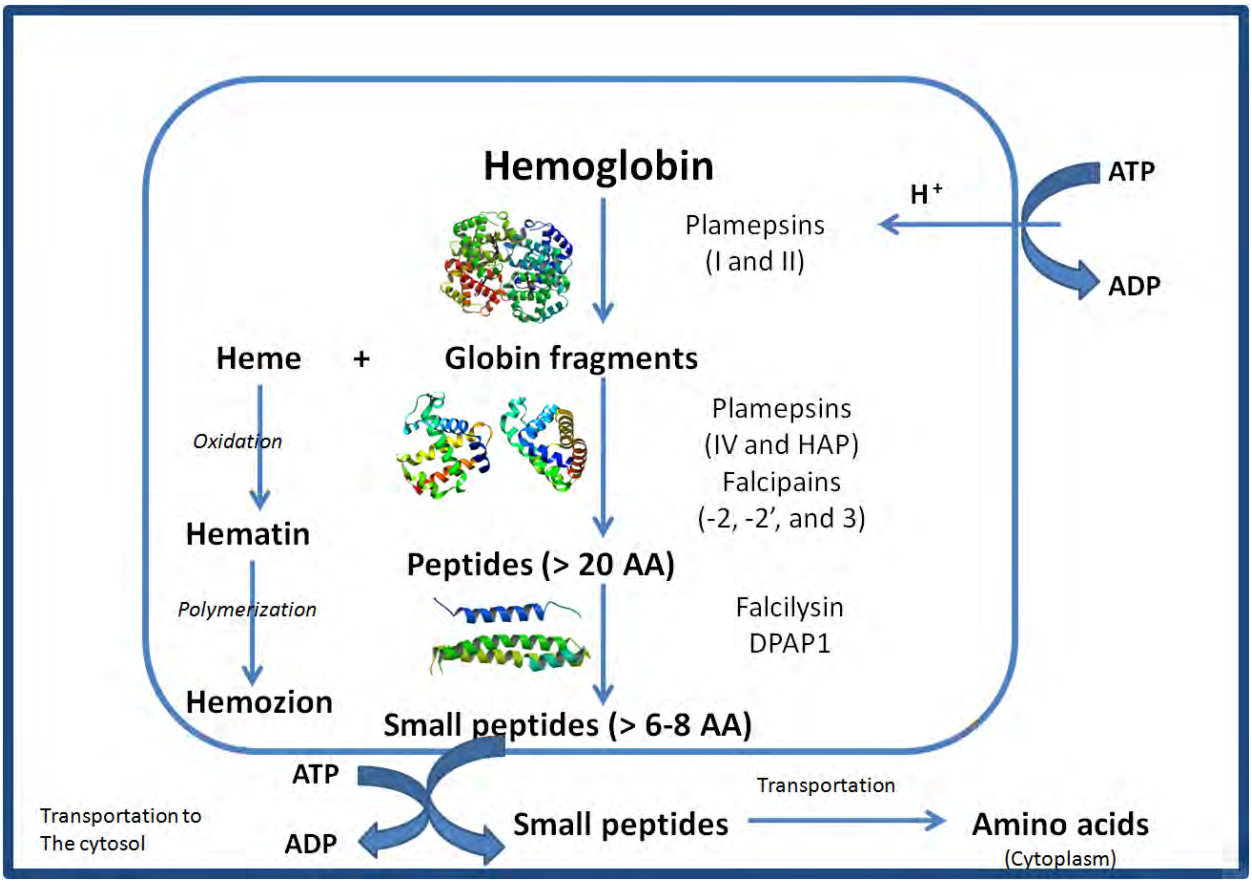

Figure 1. 6: The pathway for hemoglobin degradation initiated by aspartic acid proteases known as plasmepsins, once the globin is degraded into small peptides, it is cleaved by cysteine proteases known as falcipain-2, falcipain-2' and falcipain-3. The peptides are then further cleaved to small peptides of about 6-8 amino acids by metallo-proteases known as falcilysin. Figure adapted from Franscis Ettari *et al*., 2009

Analysis of the *Plasmodium* genome shows the presence of 10 aspartic proteases PM (Coombs *et al*., 2001), namely PM I, II, HAP, IV, V, VI, VII, VIII, IX and X. Only three (PM VI, VII and VIII) of the 10 aspartic proteases are not expressed at the intra-erythrocytic stage, and are thus likely to be involved in the insect or exo-erythrocytic stage of the parasite life cycle (Banerjee *et al*., 2002). The other 7 aspartic proteases are expressed during the erythrocytic stage of the parasite but PM V, IX and X do not seem to have any functional role in the food vacuole (Banerjee *et al*., 2002). Only two of the 10 aspartic proteases, PM I and II, initiate hemoglobin degradation (Coombs *et al*., 2001). PM I and II appear to have specificity towards native hemoglobin, making a single cleavage at the hinge region which maintains the quaternary structure, this unravels hemoglobin exposing it to further proteolysis (Goldberg *et al*., 1991;

Francis *et al*., 1996). Studies supporting this argument have provided results indicating that hemoglobin is initially clipped by PM I and II, which hydrolyze the peptide bond between PHE 33 and LEU 34 in the α-globin chain (Goldberg *et al*., 1990; Goldberg *et al*., 1991; Gluzman *et al*., 1994; Francis *et al*., 1997; Ettari *et al*., 2009). Though both proteases are involved in the initial degradation of hemoglobin, it appears that PM I readily cleaves native hemoglobin while PM II appears to prefer acid-denatured globin (Coombs *et al*., 2001). PM I and II are the most widely studied and best characterized aspartic proteases, and the other aspartic acid proteases histo-HAP and PM IV are said to be involved in cleaving the globin fragments obtained from cleaving hemoglobin at the hinge region (Banerjee *et al*., 2002). HAP and PM IV are closely related to each other (75% identical), they are both homologous to PM I and II and also share a high sequence identity. Although HAP has ~60% identity to PM I and II, it has several mutations, including the replacement of a catalytic aspartic acid with histidine and changes in the conserved flap region which lies over the binding cleft (Berry *et al*., 1999).

The second argument emphasizes that hemoglobin degradation is not at all a highly ordered process, rather redundant roles of acidic proteases are observed in this particular process (Liu *et al.,* 2006). This argument was supported by recent evidence which suggests that both aspartic and cysteine proteases cleave native hemoglobin (Liu *et al.,* 2006, Drew *et al*., 2007). Additionally, the latter argument proposes that cysteine proteases should be pursued as key drug targets for the epidemic malaria because they play a large role in hemoglobin degradation (Teixeira *et al.,* 2011; Shah *et al*., 2011). It is also suggested that cysteine proteases FP2, FP2' and FP3 are involved in the initial and subsequent cleavage of hemoglobin. The first evidence of this suggestion was backed up by experimental studies, where Salas *et al* (1995) showed that FP2 was able to cleave both denatured and native hemoglobin, and demonstrated that hemoglobin degradation was blocked by cysteine protease inhibitors and not by inhibitors from other classes of proteases. Also Liu *et al* (2006) demonstrated that the growth of hemoglobin-degrading enzymes (FP2 and PM or both) was impaired in a medium lacking isoleucine, the only amino acid absent from the hemoglobin molecule. Liu and co-workers (2006) also showed that blockage of plasmepsins using the potent inhibitor, pepstatin, had minimal effect on the

parasite, but FP2 knockout killed the parasite. The conclusion reached was that hemoglobin degradation uses dual protease families with overlapping function and that plasmepsins are not promising drug targets (Liu *et al.,* 2006, Drew *et al*., 2007). This conclusion was motivated by the fact that only cysteine protease inhibitors caused specific morphological abnomarlity on the parasite and accumulation of large quantities of undegraded hemoglobin (Rosenthal *et al*., 1988; Dluzewski *et al*., 1986; Bailly *et al*., 1992). The incubation of cysteine protease and its inhibitor L-transepoxy-succinyl-leucyl-amido-(4-guadino)-butane (E-64) resulted in accumulation of undigested erythrocyte cytoplasm in the parasite food vacuole *in vivo* (Rosenthal *et al*., 1998). The accumulation of intact hemoglobin in the parasite is an indication that indeed the cysteine protease inhibitor E-64 inhibited the hydrolysis of hemoglobin (Asawamahasakda *et al*., 1994). Leupeptin and E-64 are both non-specific inhibitors of cysteine proteases. Non-specific inhibitors of aspartic acid and cysteine proteases kill parasites both *in vivo* and *in vitro,* whereas inhibitors only specific to cysteine proteases prevented parasite maturation *in vivo* and *in vitro,* thus providing justification that only cysteine proteases should be pursued as promising drug targets (Liu *et al*., 2006). Other studies supporting the argument that cysteine proteases should be pursued as potential drug targets was the disruption of each individual gene of the plasmepsin proteases. The results obtained from the studies in which plasmepsin gene disruption was achieved confirmed that plasmepsins are important for parasite development, though not essential as plasmepsins knockouts possess the ability to compensate for the functions of individual plasmepsins (Omara-Opyene *et al*., 2004). Therefore, Individual knockout of plasmepsins and a combination knockout including plasmepsin IV/I only resulted in a slight impaired growth of the parasite (Omara-Opyene *et al*., 2004; Liu *et al*., 2005), whereas FP2 knockout markedly diminished the activity of cysteine proteases and blocked the hydrolysis of hemoglobin, which was indicated by a swollen, dark stained food vacuole (Shenai *et al.,* 2000; Sijiwali and Rosenthal, 2004). The effect of aspartic acid protease inhibitors is increased when used in combination with cysteine proteases inhibitors or in falcipain-2 knockout parasites (Sijiwali and Rosenthal, 2004: Liu *et al*., 2006). On the other hand, it is a well known fact that inhibition of cysteine proteases is lethal to the

parasite. Cysteine protease inhibitors irreversibly block the rupture of host cell membranes, thereby preventing the parasite from invading fresh erythrocytes (Glushakova *et al*., 2009).

Indeed, the determination of the precise sequence of events and the specific roles of the multiple proteases involved in hemoglobin degradation is a matter under much debate. It seems that hemoglobin degradation is not a neatly ordered process, but rather that redundant role of both aspartic and cysteine proteases are observed during the initial and subsequent cleavage of native hemoglobin and the globin fragments. However, both old and recent studies agree with each other pertaining to the pursuit of cysteine protease inhibitors as possible drug targets, therefore supporting the critical roles played by these proteases during hemoglobin hydrolysis.

Research problem statement

Cysteine proteases from different organism have been the cornerstone of pharmaceutically and industrially important studies for many years now. Although these studies have been carried out and major breakthroughs have been achieved, *Plasmodium* cysteine proteases are an interesting case. These particular proteases are different from cysteine proteases found in other organisms. They possess several unique features which is the main reason they were pursued in the study. *Plasmodium* cysteine proteases were also investigated because they have been identified as potential chemotherapeutic targets against the increasingly frustrating disease malaria. Biochemical characterization of *P. falciparum* and *P. vivax* cysteine proteases have indicated that these cysteine proteases play critical roles in hemoglobin degradation. Sequence and structural analyzes of these proteases have also shown that they are closely related to one another, evident by the high level of sequence identity and structural conservation. Also, *Plasmodium* cysteine proteases contain two extra features in addition to cathepsin-like cysteine proteases: the N-terminal insertion and C-terminal extension. In 2006, Wang and co-workers conducted a study using the principal cysteine protease falcpain-2 (FP2) of *P. falciparum*. The study found that hemoglobin (the natural substrate of cysteine proteases) binds to the C-terminal insert (also known as FP2 $_{arm}$). These findings suggest that papain-like cysteine proteases in *Plasmodium* species may have developed a novel mode of interaction with hemoglobin and therefore, this presents a different focus with regards to inhibitor design. The major limitation of the Wang and co-workers study was that they did not provide any details about the involvement of the protease active site in the degradation of substrate. This study was therefore conducted in order to understand the involvement of the active site in hemoglobin binding and also to identify residues most likely to interact with hemoglobin. We hypothesize that *P. falciparum* (FP2, FP2' and FP3) and *P.vivax* (VP2 and VP3) are typically papain-like family enzymes and should be able to cleave native hemoglobin at their active site. There is likelihood that these cysteine proteases have developed a novel mechanism of interaction with their natural substrate hemoglobin, however the protease active site cleaves

host hemoglobin thereby providing amino acids for parasite survival. This study investigates the mode of interaction between the falcipains, expect FP1 (as it lacks the hemoglobin degradation function), together with their orthologues vivapains and human hemoglobin We seek to investigate whether the findings obtained from the paper (Wang *et al*., 2006) suggesting the C-terminal insert binds hemoglobin are valid or whether the active site binds.

Aim and objectives

The aim of the study was to construct homology models which would be used to investigate the sequence variability between template(s) (structures used to model protein of interest) and target(s) (protein of interest). This could potentially lead to differences in the specificity of different cysteine proteases in their hemoglobin binding. Understanding the protease-substrate interaction would allow us to see where the essential binding takes place, therefore leading to the identification of the most likely regions that should be targeted for inhibition.

Based on the stated hypothesis the following objectives were carried out:

1. Homology models of *P. falciparum* falcipain-2' and *P. vivax* vivapain-2 and vivapain-3 were constructed separately to investigate the structural characteristics of the three enzymes. (**Chapter 2**)

2. The effects of sequence variability in the models generated and their template(s) were investigated. The possibility that sequence variability could lead to differences in the specifity of these enzymes to their natural substrate hemoglobin was investigated (**Chapter 2**)

3. The validity of the result obtained by Wang and co-workers (2006) were investigated by attempting to reproduce their data (**Chapter 3**).

4. The mechanism of enzymes-substrate interactions were investigated using the 5 cysteine proteases and hemoglobin as input files and binding them to the ARM-motif and the active site (**Chapter 3**)

5. The complexes obtained in 3 and 4 were refined by energy minimization, the change in their interaction energies before and after energy minimization would be analyzed (**Chapter 3**).

6. Forces driving complex formation and the residues likely to be involved in binding were identified (**Chapter 3**).

Chapter 2

2 Homology modeling of *P. falciparum* falcipain-2' and *P. vivax* vivapain-2 and vivapain-3

This chapter describes the homology modeling of *P. falciparum* and *P. vivax* cysteine proteases; falcipain-2' and vivapains (vivapain-2 and vivapain-3) respectively. The properties of the models will be discussed and analyzed. The credible models generated from this study will be used in the next chapter for the protein-protein docking studies.

2.1. Introduction

The properties of proteins are largely determined by their three-dimensional (3D) structure. The experimental determination of the 3D structures of proteins contributes towards overall characterization of the protein molecule. The information obtained enables researchers to unravel and understand the role that the protein of interest plays in the cell. Thus, the process of determining the 3D structure of a protein is vital. Protein structures facilitate numerous biological processes in living organisms; they are involved in processes such as enzymatic reactions and immune evasion by viruses (Parker, 2003). Two main techniques are used for the experimental determination of proteins 3D structure: X-ray crystallography and Nuclear magnetic resonance (NMR) spectroscopy.

X-ray crystallography is the most widely used technique for protein 3D structure determination; its success is attributed by the fact that it has the largest number of protein entries in PDB. This

method attempts to find a pure protein by growing it as a well-ordered crystal that can diffract light strongly. Crystals are basically 3D arrays consisting of a series of molecules (Bragg, 1913). In the case of protein crystals, some additional precipitants have to be used to enhance their growth. The pH, temperature, nature of solvent, precipitant (ammonium sulfate or polyethylene) and the presence of ions or ligand critically determines the crystallization of a protein, which may take several weeks to a few years. The crystallization of a protein is considered to be the most difficult task in X-ray crystallography as it is difficult to approximate how long it may take (Geerlof *et al*., 2005). Once the crystal has grown, the atoms within it are identified by striking it with a beam of X-ray which yields a diffraction pattern in which electrons are scattered in many directions (According and Boxes, 2006). This data produces a 2-dimensional picture which is converted to a 3D density map of the electrons within a crystal using the Fourier Transform (Hoffman, 1997). The electron density map is used to refine a crystal structure and computationally determine its chemical information (chemical bond, length and disorders). This is visualised in a 3D picture of the density of electrons within the crystal. Because the protein is crystallized, the dynamics, or motions, of the protein cannot be observed. Protein crystals used for diffraction studies are highly hydrated and studies have shown that structures determined from crystals are not much different from the structures of soluble proteins in aqueous solution. However, not all proteins can form crystals (Stryer *et al*, 2001; Birkholz, 2006; Parker, 2003). X-ray crystallography produces high-quality protein structures with resolutions as low as 0.54 Å. The disadvantage of this method of protein 3D structure determination is that it requires that a sufficient quantity of protein be isolated from its natural source. Proteins must also be overexpressed for X-ray crystallography, and this is difficult to almost impossible for some proteins (Jesch *et al*., 2000; Oksanen and Goldman, 2006).

The discovery of X-ray crystallography has been a major breakthrough in the pharmaceutical industry and in the field of molecular biology. This techniques has, since its discovery, been used to obtain high-quality protein structures with resolutions as low as 0.54 Å (Jesch *et al*., 2000). There are some limitations to X-ray crystallography: it requires proteins to be overexpressed and

this is not easily achieved especially for membrane proteins (Lundstrom, 2006).The *Plasmodium* genome is AT-rich which makes protein expression in *E. coli* difficult (Aravid *et al*, 2003).

NMR can only solve protein structures with a molecular size less than 80 kDa; hence it is the method of choice for small proteins which cannot be readily crystallized ((Bertini and Luchinat, 1998). This technique reveals the structure and dynamics of a protein in solution (Clore and Gronenborn, 1998; Tamm and Liang, 2006). It uses magnetic fields and electromagnetic radiation to detect magnetic shifts caused by interactions between atomic nuclei and electrons in the protein (Bertini and Luchinat, 1998; James, 1998).

Due to the limitations that X-ray crystallography and NMR present, researchers started to investigate alternative ways to solve protein 3D structure. Computational biology and bioinformatics aim to accelerate the determination of protein structures by providing computer models which aid the study of these structures. Computer modeling is more advantageous than experimental modeling because it speeds up the process of obtaining a protein model; however the quality of the structures obtained is highly dependent on the sequence identity between the target and template proteins.

2.2 Homology modeling

Homology modeling, also known as comparative modeling, refers to a process whereby a known crystal structure which has been solved by experimental techniques (template) is used to predict the atomic co-ordinate of an unknown (target) protein based on its amino acid sequence (Šali and Blundell, 1993; Sánchez and Šali, 1997). Homology modeling takes advantage of the well known biological concept which states that during evolution the overall fold (structures) of a protein is more conserved and changes less rapidly than its amino acid sequence (Chothial and Lesk, 1986; Hubbard and Blundell, 1987). There are four main steps which constitute the process of homology modeling (Figure 2.1):

- Template(s) identification
- Sequence alignment
- Model building
- Model evaluation

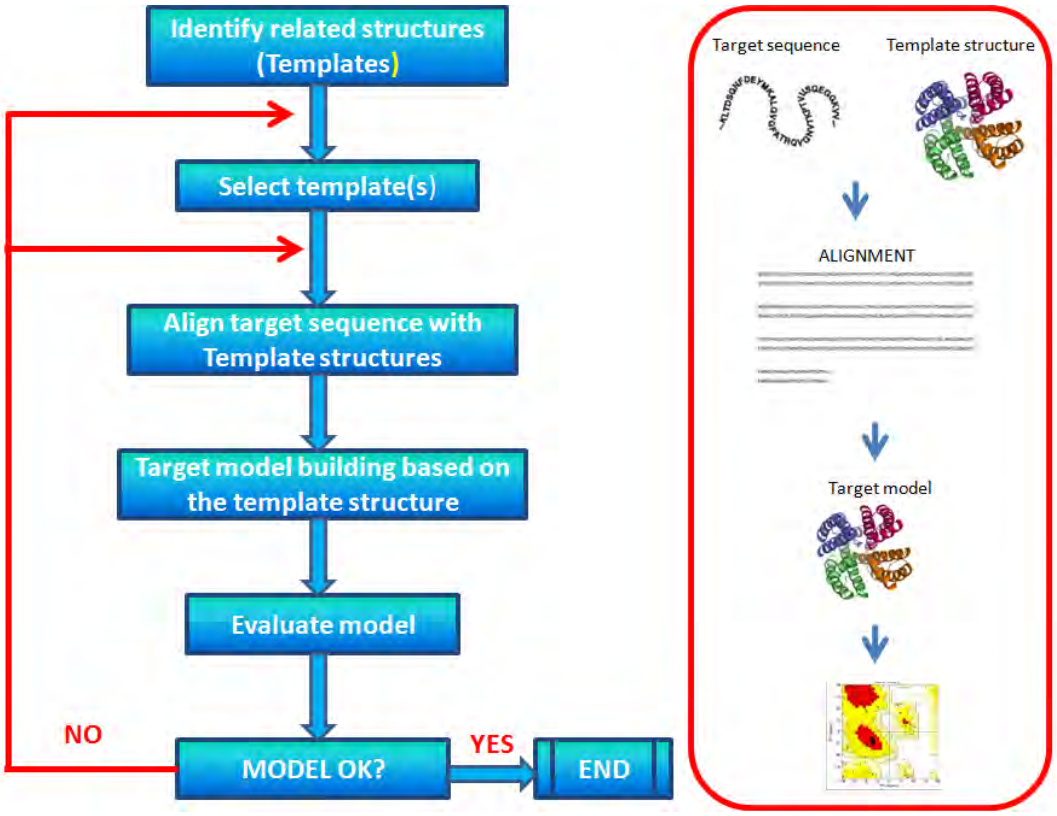

Figure 2. 1: Four basic steps followed in homology modeling, starting from template identification, sequence alignment, model building to model evaluation. Figure modified from Eswar *et al.*, (2006)

2.2.1 Template identification

Template identification is the most crucial step in homology modeling as it lays the foundation for the whole process. Basic Local Search Alignment Tool (BLAST) (Altschul *et al*., 1990) and Fast Alignment (FASTA) (Lipman and Pearson, 1985) are the two most popularly used programs for the detection of template (s). Both of these programs use pairwise alignment methods to

search PDB (Bernstein *et al*, 1977) for protein structures with detectable sequence identity to the query/target protein. The homology detection prediction (HHpred) interactive server (Söding *et al*., 2005) is also emerging as a popular and more efficient program for template identification. HHpred uses the query sequence to search various databases and outputs various template structures as hits.

- ***Template search***

BLAST and FASTA are useful tools, they detect weak and yet biological relevant similarity between query sequences and all the other structures in PDB. Both these programs are the most commonly used tools for biological analysis of protein and DNA sequences (Altschul *et al*., 1990). They use a rapid heuristic algorithm for obtaining pairwise alignment; however BLAST is more popular than FASTA (Henikoff and Henikoff, 1992; Oladele *et al*., 2008).

FASTA (Pearson and Lipman, 1988) is an improved derivative of the FASTP program (Lipman and Pearson, 1985); it is user-friendly, easily accessible and can be run online at http://fasta.bioch.virginia.edu/fastawww2/fastawww.cgi?rm=select&pgm=fap. The web interface at the European bioinformatics institute (EBI) makes provision for the user to run a FASTA search at: http://www.ebi.ac.uk/Tools/fasta/. The pairwise alignment in FASTA is carried out in four consecutive steps, which start with local alignment of two similar regions and bears no reference to gaps. The earlier versions of FASTA used PAM-250 scoring matrix (Dayhoff *et al*., 1978) to rescore similar regions, in which conservative replacements and substitution with identical amino acid increase and random chance substitutions is a negative score. However, with the latest versions of FASTA have implemented PAM-120, MDM-10, -20 and -40, BLOSUM-50, -62 and -80 scoring matrices for rescoring similar regions (Dayhoff *et al*., 1978).

BLAST was developed by Stephen Altschul of the National Center for Biotechnology Information (NCBI) in 1990; it is useful for the identification of conserved subsequences in the query to generate several distinct subsequences. Its algorithmic steps are implemented into three:

collecting a list of high scoring words, scanning database for hits and extending hits (Altschul *et al*., 1990). Position–Specific iterated BLAST (PSI-BLAST) can be used to detect distant homologs. PSI-BLAST iterates BLAST searches using a position-specific matrix (Altschul *et al*., 1997). It performs a database search by building a profile of sequences iteratively (Xiong, 2006), is more sensitive than BLAST and does not sacrifice the speed and accuracy of the algorithm. Both BLAST and PSI-BLAST are easily accessible at: http://blast.ncbi.nlm.nih.gov/Blast.cgi.

- ***HHPred server for identifying templates***

HHPred is found at http://protevo.eb.tuebingen.mpg.de/hhpred it is basically a sequence searching method which is just as easy to use much like BLAST, PSI-BLAST and FASTA. This server has higher sensitivity and is faster than the most popularly used programs for template identification. It implements pairwise comparisons of profile hidden markov models (HMMs) to make extensive homolog detection by searching through a variety of databases such as PDB (Bernstein *et al*, 1977), SCOP (Murzin *et al*., 1995; Hubbard *et al*., 1997), Pfam (Sonnhammer *et al*., 1998), SMART (Ponting *et al*., 1999), COG (Tatusov *et al*., 2003) and CDD (Marchler-Bauer *et al*., 2002).

Sequence identity between target and template(s) is a foundational basis for template identification. In addition to the sequence identity between the target protein and its template structure, there are also other factors about the latter which need to be considered. The experimental accuracy of the template to be used for modeling is vital. For a protein which was solved by X-ray crystallographically, its R-factor and resolution are used as a good guide to determine the accuracy of the elucidated structure. The accuracy of protein structures solved by NMR are indicated by the number of restraints per residue (Marti-Renom *et al*., 2000). The biological and environmental information in the template is also imperative for the modeling process, as it is correlated to what is required for the model. Thus, factors such as conservation of the active site, ligand bound/unbound, the pH, temperature and solvent should also be taken into consideration (Šali and Blundell, 1993; Sánchez and Šali, 1997).

2.2.2 Sequence alignment

In the previous step: template identification, a few hits of related protein structures and sequences can be obtained from several databases. The sequence and structures with a detectable sequence identity to the template can be used to construct either a structural and/or multiple sequence alignment (MSA). MSA are a basic extension of the sequence alignment between two proteins where all sequences in a specified set are used (Omar *et al*., 2005). Structural alignments are mainly used to validate MSA because of the high conservation of protein structures (Oladele *et al*., 2008).

In order to construct a good model, the sequence identity between the target and template(s) proteins must be high. The purpose of a MSA is to construct an accurate alignment, detect families which both the template and target protein belong to and organize them into subfamilies. Functionally conserved sites, phylogenetic analysis and functional predictions of amino acids can be identified using MSA (Barton, 1996). The construction of MSA is usually progressive, resulting in the alignment of closely related sequences first and other groups gradually aligned to the initial alignment (Tompson *et al*., 1997). This technique works exceptionally well for closely related sequences but becomes more difficult and less reliable for distantly related sequences outside the "twilight zone" (Figure 2.2). Thus, alignment of sequences with less than 30% sequences identity is problematic (Rost, 1999). The sequence identity and length of proteins in the safe homology modeling zone (Figure 2.2) adopt the same structures and alignment as expected , the alignment of proteins in this area is not meant to be problematic (Krieger *et al*., 2003).

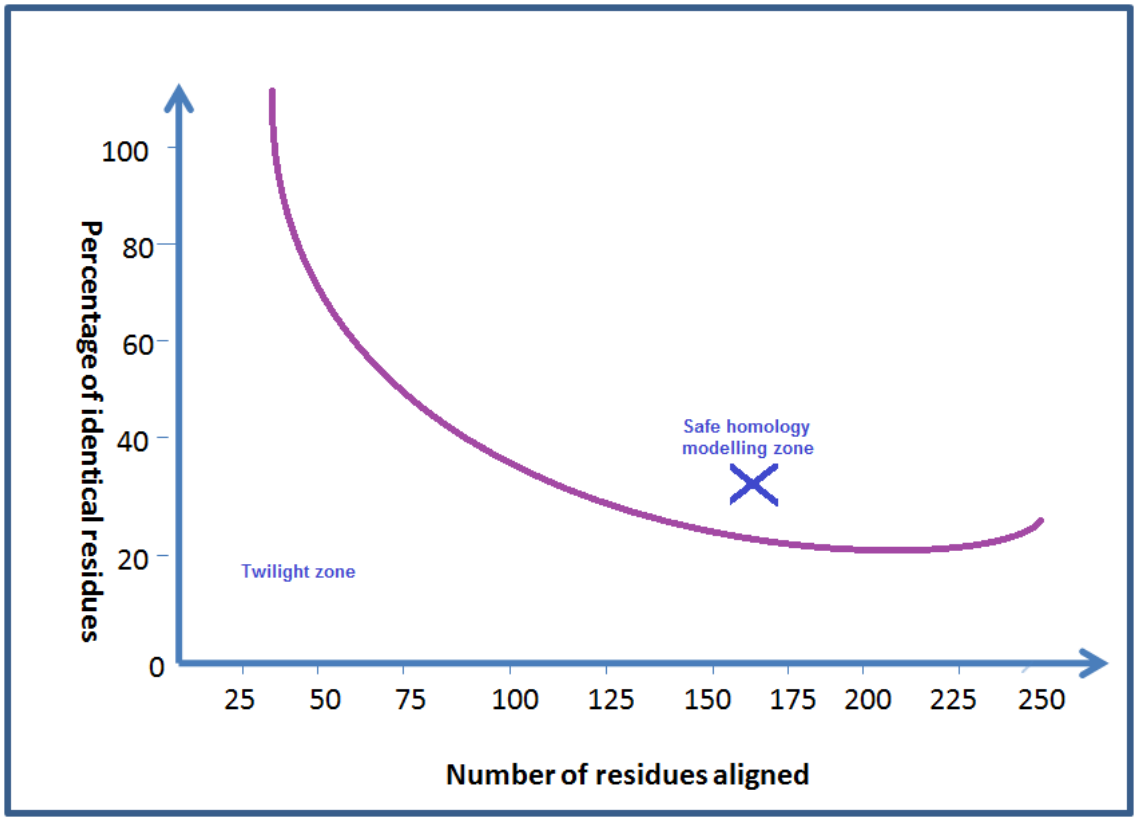

Figure 2. 2: Safe homology zone and twilight zones a (marked with a cross) for multiple sequence alignment confidence. Figure adapted from Krieger *et al*., 2003 page 508.

There are many programs used for the construction of MSA, and these programs employ different scoring matrixes to analyze various features in the sequences. The six most popular programs are ClustalW, DIALIGN-T, MAFFT, MUSCLE, PROBCONS and T-COFFEE (Golubchik *et al*., 2007). Most MSA programs, including the ones listed above use progressive alignment methods to achieve accurate results. However, they encounter problems when the sequences to be aligned are divergent (less than 30 % sequence identity), meaning they contain some deletions and insertions. Below is a brief description of the 3 MSA programs used in the homology modeling of FP2', VP2 and VP3. CLUSTALW2, T-COFFEE and PROMALS3D are the programs which were used in our study.

- ***ClustalW***

ClustalW is a fast and efficient progressive alignment approach (Feng and Doolittle, 1987) used for aligning sequences with high sensitivity. Its alignment for divergent sequences is improved by assigning individual weights in a partial alignment so that near-duplicate sequences are down-weighted by dynamic programming involving the use of PAM 250 and BLOSUM 62 matrixes (Tompson *et al*., 1994). More divergent sequences are assigned up-weights. ClustalW introduces residue specific gaps in hydrophilic and loop regions rather than in the regular secondary structures. This program is freely available online and portable for all computers platforms (Tompson *et al*., 1997). There are two versions of Clustal (ClustalX and ClustalW) which use the same principle, except that Clustal X window interface of ClustalW which is run on the terminal. Both these programs can be downloaded from: http://www.clustal.org/

- ***T-COFFEE***

T-COFFEE is an accurate MSA program that produces better results than many MSA programs in a modest speed. It is based on the popular progressive approach (Feng and Doolittle, 1987) for the generation of MSA and its greedy algorithm helps in avoiding errors in the alignment. T-COFFEE pre-processes data for pairwise alignment of sequences. This generates an alignment library with information that can be used to guide progressive alignments. It then follows the intermediate alignment which is based on all sequences to be aligned and how they align. The alignment information obtained here is derived from other alignment programs and structure superposition. This approach is powerful because it aligns sequences both locally and globally. The final alignment are more reliable and scientifically viable (Notredame *et al*., 2000). T-COFFEE is available for download and can be run on UNIX or UNIX-like platforms such as Linux, cygwin and MacOSX. This program can also be run online at the Swiss Bioinformatics web interface at: http://tcoffee.vital-it.ch/cgi-bin/Tcoffee/tcoffeecgi.

- ***PROMALS3D***

This program automatically identifies homologs with known 3D structures for the input sequences. It derives structural constraints through structure-based alignments and combines them with sequence constraints to consistence-based MSAs. PROMALS aligns similar sequences progressively using a relatively simple and fast algorithm. A more perfected technique is then applied to align more diverse sequences. The first alignment uses BLOSUM 62 scoring function to cluster pre-aligned sequences which are relatively distant from each other and this step occurs quickly (Pei *et al*., 2008). In the second alignment one representative sequence (target) is selected from each pre-aligned group and subjected to PSI-BLAST (Altschul *et al*., 1997) searches in order to detect additional homologs from UNIREF 90 database and the secondary structure prediction server (PSIPRED). Profile-profile alignment with secondary structure predictions using hidden markov models (Martin *et al*., 2006) are applied to pairs of representative sequences for obtaining posterior probabilities of residue matches. The probabilistic consistency scoring function is derived from a combination of sequence-based constraints and structure-based constraints which are obtained from homologs with 3D structures that have been experimentally determined. The purpose of the consistency scoring function is to align representative target sequence with the pre-aligned groups of sequence from the first alignment which are then merged to form the MSA (Pei *et al*., 2008). PROMALS3D can be run online at: http://prodata.swmed.edu/promals3d/promals3d.php.

2.2.3 Model building

The information contained in the target-template alignment obtained from the MSA is used to generate the 3D structural model of the target protein. The model is represented by a set of Cartesian co-ordinates for each atom of the protein (Baker and Šali, 2001). There are four major approaches for model generation: rigid-body assembly, segment matching, satisfaction of

spatial restraints and artificial evolution (Xiang, 2006).

In the rigid body assembly approach, small rigid pieces obtained from sequence alignments are used to build a protein model. These rigid pieces are based on the conserved core regions, loops and side-chain conformations of known proteins which are useful for constructing the backbone of the target protein (Sutcliffe *et al*., 1987). The other atoms of the target proteins are also obtained by superimposing it with template structure(s). Loop modeling is done by searching a library of similar structures and side chains are built using a combination of template-side chain conformation and preferred side-chains conformations. COMPOSER package implements a rigid body approach for homology modeling (Sutcliffe *et al*., 1987).

In the segment searching method, segments of the protein and not the entire length are aligned. The target protein is therefore divided into a series of short peptides which are matched to template peptides which are already available in PDB (Levitt, 1992). The short peptides are then put together and conformational restraint is applied to them. Segment matching is more advantageous than other methods as it constructs short insertions, deletions and side-chain atoms. This particular approach is applied in the SEGMOND package (Levitt, 1992).

Homology modeling by spatial restraints uses CHARMm forcefield to enforce proper stereochemistry into an objective function. The tolerance of extra optimization in the model through molecular dynamics is ensured by the inclusion of CHARMm forcefield. The most commonly used software in spatial restraint-based modeling is MODELLER (Šali and Blundell, 1993). MODELLER performs four spatial restraints, (I) homology based restraints based on distance and dihedral angles are derived from the target-template(s) alignment; (II) stereochemical restraints such as bond length and bond angles are attained using the molecular forcefield CHARMM-22 (Brooks *et al*., 1983; Brooks *et al*., 2007); (III) statistical preference of dihedral angles and non-bonded inter-atomic distances are used to derive a representative set of known structures in PDB;. (IV) there is also an option for manually curated preference, such

as rules for packing secondary structures, results obtained from cross-linking experiments, fluorescence spectroscopy, and reconstruction of images from electron microscopy, site-directed mutagenesis and intuition (Šali *et al*., 1990; Šali and Blundell, 1993; Šali and Overington, 1994). Spatial restraint calculations are expressed by mean of Probability density functions (pdfs). Restraints obtained from statistical analysis of the relationships between many pairs of homologous structures have been used to calculate Pdfs, which are also combined into an objective function that is optimized by a combination of conjugate gradients and molecular dynamics with simulated annealing (Eswar *et al*., 2006). Our study uses this method for the model construction.

Artificial evolution model building has been implemented in the NEST program, a module of JACKAL package. In this procedure the alignment between the target and the template is considered to be a list of operations such as residue mutation, insertion or deletion (Xiang, 2006). Building a target model is considered as a process of editing the template structure based on the alignment. Each operation: mutation, deletion or insertion, will disturb the template structure and thus involve an energy cost, either positive or negative. The model building starts from the operation with the least energy cost and so on. Each operation is followed by a slight energy minimization to remove atom clashes. The final structure is then subjected to more thorough energy minimization (Xiang, 2006).

2.2.4. Model validation

The accuracy and credibility of a model is the final and most essential step in homology modeling. This step mostly relies mostly on the sequence identity between target and template(s) (Chothial and Lesk, 1986). Models built from targets with less than 15% sequence identity to template(s) are less reliable and often derived from misled conclusions. The common errors in such models are speculated to be in the loop regions and side chains (Hilisch *et al*., 2004; Xiang, 2006). These speculations are made because at times in the initial stage of the modeling process, a BLAST search may fail to detect close homologues which may result in

a faulty alignment, therefore rendering the whole modeling process inaccurate. Also, sequences with a less than 15% identity often contain larger gaps (Hilisch *et al*., 2004). A Root Mean Square Deviation (RMSD) error up to 3.5 Å has been predicted for sequences with 30-40% identity (Bower *et al*., 1997). Models built with this sequence identity are used for structure-based drug targets, designing of mutagenesis and *in vitro* experiments (Figure 2.3). For proteins with >40% sequence identity, models are as reliable as the experimental structure, the alignment does not contain any gaps and is therefore straightforward. RMSD of about 1 Å can be expected (Sánchez and Šali, 1997). These models are useful for detailed prediction of structure-based design and preferred sites of metabolism of small molecules such as ligands

Relationship between target and template(s)	Sequence identity between target and template (s)	Expected RMSD	Method to be used for detecting sequence identity between target and template (s)	Application in drug discovery
	100%- 55%	+/- 0.5Å	Pairwise sequence alignment	•Structure-based drug design •Structure-based drug assessment •Site-directed mutagenesis •Assignment of protein function
	55%- 25%	0.5Å-1Å	Multiple sequence alignment	•Structure-based target assessment •Site-directed mutagenesis •Assignment of protein function
	25%-15%	1Å-3.5Å	Threading profile	•Site-directed mutagenesis •Assignment of protein function
	15%- 00%	>3.5Å	*Ab initio*	

Figure 2. 3: The relationship between sequence identity and model function. Arrows indicate the best method to proceed for model creation, and on the right side applicability of the model. Figure adapted from Hilisch *et al*., (2004) page 662.

Models built using any of the four modeling approaches (**section 2.2.3.**) have to be evaluated to

ensure that they are consistent with the physio-chemical rules of a protein molecule. The assessment of protein models is a difficult task and there is not a single method able to accurately and consistently predict the 3D structure of a protein (Xiang, 2006; Laskowski, 2003; Krieger *et al.,* 2003). In the same way, there is no method to predict the errors in the model protein structure, most programs used to assess the model were originally created for validating experimentally solved protein structures prior depositing them in PDB. There are two scoring function used for validating the credibility of a model, statistical effective energy function and physical energy function. The former employs an empirical method based on the observed residue-residue contact frequencies among proteins with known structures in the PDB. It assigns a probability or energy score to each possible pairwise interaction between amino acids and combines these pairwise interaction scores into a single score for the entire model (Sippl, 1995). Physical-based energy calculations aim to capture the inter-atomic interactions that are physically responsible for protein stability in solution, especially Van Der Waals and electrostatic interactions (Lazaridis and Karplus, 1998; Xiang, 2006). Model assessment programs are either based on statistically effective energy function, physical energy function and/or both to calculate errors (Sippl, 1993). Different programs use different approaches, they therefore complement each other and help raise confidence in the errors predicted for a specific region in a protein (Xiang, 2006). In this particular study, 3 model assessment programs: PROCHECK, ProSA and MetaMQAP II were used to identify potential errors in the models built.

- ***PROCHECK***

This is one of the most popularly used programs for assessing the stereochemistry of experimentally determined structures and protein models. PROCHECK compares the geometry of a given protein with that of well-defined, high-resolution structures. It looks at the phi (Φ) and psi (Ψ) angles, chirality, bond angles and bond length (Xiong, 2006). Unusual regions highlighted by PROCHECK are not necessarily errors, but may be distortions due to a ligand-binding site in the protein's active site (Laskowski *et al*., 1996).

PROCHECK takes PDB file containing the co-ordinates of the model proteins as input and outputs 10 PostScripts files, one of them is a Ramachandran plot. The Ramachandran plot is undoubtedly the best known and most powerful check for the stereochemical quality of a protein structure (Voet and Voet, 2006). It shows allowed regions, generously allowed regions and disallowed regions for the input file using the colour co-ordinations yellow, cream and white backgrounds respectively. Other colours on the Ramachandran plot are red for the most favourable core regions, black marker for most favourable regions and red markers for unfavourable regions. PROCHECK can be run online at http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/ and may also be downloaded to run on a local computer.

- ***Protein Structure Analysis (ProSA)***

This is a widely used tool for the detection of potential errors in the 3D models of protein structures. It is also used for error identifications in experimentally solved structures (prior to being deposited to PDB), theoretical models and protein engineering (Wiederstein and Sippl, 2007). ProSA estimates the probability of two residues being at a specific distance from each other (Wallner and Elofsson, 2007), for this it relies on the empirical energy derived from pairwise interactions as observed in high-resolution protein structures (Pawlowski *et al*., 2008). The input to ProSA is atomic co-ordinates of the protein model in PDB format or a four-letter PDB code of structures available in the database. It outputs a Z-Score plot, energy plot and 3D structure of the protein in a Jmol viewer (Sippl, 1993). ProSA-web is accessible at http://prosa.Services.came.sbg.ac.ta

- ***METAMQAP II***

MetaMQAP is a server which evaluates a model using model quality assessment programs

(MQAPs): VERIFY3D, ProSA 2003, PROVE, ANOLEA, BALA-SNAPP, TUNE, REFINER and PROQRES (Pawlowski *et al*, 2008). The basic idea behind VERIFY3D is to evaluate the environment of each residue in a model with respect to the expected environment as found in the high-resolution X-ray structures (Xiong, 2006). It is a statistical approach which basically divides the environment into 18 classes based on the secondary structure, areas buried and the fractions of polar contacts (Wallner and Elofsson, 2007). ANOLEA (Atomic Non-Local Environmental assessment) is used to assess protein chains (Melo and Feyman, 1998). ANOLEA calculates the energy of a protein chain by estimating the non-local environment (NLE) of each heavy chain atom within the molecule. NLE is defined as a set of all heavy atoms within a euclidean distance of 7Å and are not farther than 11 residues in the analysed polypeptide chain (Pawlowski *et al*., 2008). PROVE analyzes the packing in protein models, it evaluates the regularity of the atom volume (defined by the atom's radius) and planes separating it from other atoms. In BALA, the structure is basically evaluated by means of a four-body statistical potential, this application is aimed mainly at tetrahedral quadruplets and spatially neighbouring residues. Local quality of residue from a local and non-local contact of residues in the model is predicted by a neural network in TUNE. REFINER uses a statistical approach in the evaluation of a protein model, using terms such as: contacts potential, long distance potential, hydrogen bonds and burial pseudo energy. PROQRES is a protein evaluation method that has been specifically developed to detect local errors in a protein model. The user submits a protein model in a PDB file to a MetaMQAP II server: https://genesilico.pl/toolkit/. The server returns an output with absolute prediction deviations (in Å) of individual Cα atoms between the model and the unknown structure. It gives the global deviation which is expressed in (RMSD) and GlobalDistanceTest_TotalScore (GDT_TS) (Pawlowski *et al*., 2008).

2.3. Methodology

Because the target proteins (FP2, VP2 and VP3) for modeling derive from the same *Plasmodium* genus and share significant similiarity to one another, BLAST (Altschul *et al*., 1990) searches yielded the same results using each protease sequence as query. Therefore, homology or

comparative modeling steps for all three cysteine proteases were harmonized in the overall steps below. Also the model of human procathepsin K was generated as a control, the model was built using a distantly related template in order to show the reliability of MODELLER and to demonstrate the overall structure of cysteine proteases when bound to the prodomain.

2.3.1 Data retrieval

PDB is the single worldwide database containing structural data. The database is managed by Research Collaboratory for Structural Bioinformatics (RCSB), which aims to create a resource base on the modern use of technology involving the use and analysis of structural data for biological research (Breman *et al*., 2000). Protein structures solved by X-ray crystallography and NMR are deposited, processed and distributed in PDB data by RCSB. In the template identification step, the target protein sequence (each of the 4 cysteine proteases individually) was a query in the BLAST searches performed against PDB. The PDB-BLAST search resulted in the understanding that the query protein sequences (targets) had a significant sequence identity to other protease within the *Plasmodium* family. The target protein sequences had sequence identity higher than 50% to FP2 and FP3. Local similarity searches indicated that the alignment was at the regions where the prodomain was presumably cleaved off, as the alignment was at the last +241 residues of the proteins in PDB. The presence of a prodomain in papain-like cysteine proteases is a well known fact but the literature has indicated the evidence of an unusually large prodomain in FP2 subfamily proteases (Lecaile *et al*., 2002) and their *Plasmodium* homologues. Based on this information, the first +240 residues were presumed to be the prodomain and were cleaved off. The aim of the study was to built model mature protease so; FP2 and FP3 were used as a guide to cleave off the target sequences. Other than FP2 and FP3, some proteins with a detectable sequence identity to target proteins were included in the MSA, the purpose of which was to enhance the alignment and obtain an accurate structural alignment. *Plasmodium* cysteine protease sequences were retrieved from the Universal Protein Resource (UniProt) database, which can be found on:

http://www.uniprot.org (Bairoch *et al*., 2004; The UniProt Consortium, 2009). UniProt was the database of choice as it is non-redundant and its sole purpose is to provide the scientific community across the world with high quality, reliable and functionally annotated protein sequences (Bairoch *et al*., 2004). Therefore, sequences in UniProt can be considered to be highly accurate as they have been both manually and automatically curated. Using each target sequence (FP2', VP2 and VP3) at a time, BLAST searches were performed against the UniProt database and *Plasmodium* homologues with detectable sequence identity were retrieved. The Table (2.1.) below shows the PDB codes, UniProt accession numbers, resolutions of the structures, scientific organism in which proteases were expressed and sequence identities of all sequences used in the alignment.

The falcipains possess unique features which distinguish them from other papain-family cysteine proteases. The extra features present in the falcipains and their *Plasmodium* homologues include the arm-like motif (C-terminal extension) and the nose-like motif (N-terminal insert). The pdb files containing structural information on the proteases belonging to this family that have been elucidated by experimental techniques are complicated. In the two structures of falcipain-2 (PDB code: 2OUL and 1YVB) the presence of the nose-like motif and the arm-like motif appear to be the major features which have disrupted the normal papain-family cysteine protease numbering. For FP2 (2OUL), the first 17 residues (nose-like motif) are labeled from -16 to 0, and the first amino acid (GLN) in the mature domain has been labeled 1, then the numbering proceeds to the last amino acid GLU as residue no 224. The traditional papain-family cysteine protease numbering (where the catalytic cysteine is residue no:25) has been kept intact for the other structure of FP2 (1VYB). Tthe unique features of FP2 have been given unique numbers: the nose-like motif has been labeled as 0A-0L and the 14 amino acid arm-like motif has been labeled 169A-169N. For the purpose of simplicity in this study, we renumbered the PDB files of both FP2 and FP3 structures. In FP2 the first residues GLN is no. 1 and the last residue has been numbered 241. Also, it has often been observed that BLAST finds local optimal ungapped alignments to query (target) sequence using a BLOSUM 62 substitution matrix. The

major disadvantage of this approach is that at times it may not reflect the true global sequence identity between query and aligned sequences. Therefore, based on this observation, BioEdit (a sequence alignment editor) was used to re-calculate the sequence identities obtained from BLAST. BioEdit uses an all-against-all matrix for identity calculations, the sequence identities listed in Table 2.1 were initially calculated in BLAST and later re-calculated by BioEdit. Both BLAST and BioEdit indicated the same results due to the high sequence identity between targets and templates aswell as the removal of the prodomain.

Table 2. 1: All the protein structures (marked by PDB codes, row 1-7) and protein sequences (marked by UniProt accession numbers; row 8-15; which were included in the sequence alignment)

PDB code or UniProt Accession number	Molecule	Resolution	Scientific Name	Sequence Identity to FP2'	Sequence Identity to VP2	Sequence Identity to VP3
2OUL	Falcipain-2	2.20	*P. falciparum*	96	63	56
3BWK	Falcipain-3	2.42	*P. falciparum*	68	66	60
1PCI	Procaricain	3.20	*Carica papaya*	55	35	35
2FO5	Cysteineprotease EP-B2	2.20	*Hordeum vulgare*	67	41	37
1BY8	Procathepsin K	2.60	*Homosapiens*	40	37	36
2BDZ	Mexican	2.10	*Jacaritia mexicana*	38	36	40
1YAL	Chymopapain	1.70	*Carica papaya*	37	37	57
Q56CY9	Falcipain-2'	None	*P. falciparum*	100	49	46
Q7Z1W6	Vivapain-2	None	*P. vivax*	49	100	54
Q7ZOB2	Vivapain-3	None	*P. vivax*	46	54	100
Q8WQM4	Bergepain-2	None	*P. berghei*	43	41	41
Q7Z1Y8	Chabaupain-2	None	*P. chabaudi chabaudi*	45	42	42
Q8WQM3	Vinkepain-2	None	*P. Vinckei*	45	39	39
B3L4V5	*P.knowlesi* ortholog of falcipains	None	*P.knowlesi*	46	54	50

The sequence of human procathepsin K was used as query when performing a BLAST search against PDB. Several hits were obtained for potential template structure and 1XKG was used

because the objective was to generate a model from a template with a low sequence identity in order to demonstrate that MODELLER is not a side-chain substitution program.

2.3.2 Sequence alignment

As indicated in **section 2.2.2.** (Sequence alignment), MSA programs use different approaches to analyze the sequences. Therefore, relying on results from one MSA program may be misleading and often biased. It is advisable to use more than one program (Golubchik *et al*., 2007). All the sequences and structures listed in Table 2.1 were aligned using three MSA programs: ClustalW2 (Tompson *et al*., 1994), T-coffee (Notredame *et al*., 2000) and PROMALS3D (Pei *et al*., 2008) in order to remove any bias and increase our confidence in the alignment generated. The input to all three MSA programs was a FASTA file containing all the sequences listed in Table 2.1. The ClustalW2 alignment was run with no iterations set and performed by the BLOSUM62 scoring matrix. Default gap opening penalty and gap extension penalty which are 10.0 and 5.0 were not altered. Hydrophilic gaps and residue-specific gap penalties were set on. The scoring matrix for T-COFFEE was also BLOSUM 62, gap opening and gap extension were both set at 0 (default). PROMAL3D is a web-based program which was run online by simply providing the FASTA file of all sequences in Table 2.1.

2.3.3 Model building

The models of human procathepsin K, FP2', VP2 and VP3 were built in MODELLER 9v7, a spatial restraints homology modeling program. MODELLER is freely available online and was downloaded from: http//www.Šalilab.org/ modeller/download.installation.html. It uses three main files to execute the model building of target protein: atomic coordinates of the template(s) (pdb file), target-template(s) alignment file (in a PIR format) and script files (written in python programming language) (Marti-Renom *et al*., 2000). Template structures, FP2 (PDB

code: 2OUL) and FP3 (PDB code: 3BWK) were retrieved from PDB. FP2 was used as the only template for building the 3D model of FP2'because of the high sequence identity of the two proteases. Both the structures of FP2 and FP3 were used as templates for VP2 and VP3 model building in order to ensure a high degree of confidence in the results obtained. The 3D structure of recombinant proDER p1, a major house dust mite proteolytic allergen (Meno *et al*., 2005) was retrieved from PDB. The PDB (1XKG) was used as a template for the construction of human procathepsin K model. Target-template file for each target protein was provided to Mod9v7 in PIR format. The script file 1, 2, 3 and 4 (Appendix A1) were also provided for building the 3D structures of FP2', VP2, VP3 and procathepsin K respectively. In all cases, 100 models were built by optimizing the molecular pdfs from different initial conformations in order to select high quality models from mirrors of the whole molecule (Šali and Blundell, 1993).The outputs from MODELLER were all the models with non-hydrogen atoms and the models were evaluated using a statistical evaluation method. Discrete Optimized Protein Energy (DOPE) scores, based on atomic distance-dependant from native structures, assesses the reliability of a model (Shen and Šali, 2006). Three scripts were written for the model assessment, the first one calculating the DOPE score of the model, second one calculating the DOPE Z-scores of each model and last one sorting the DOPE Z-scores based on the accuracy. The scripts used for each calculation are provided in Appendix A: 1, Script 5, 6 and 7. The DOPE Z-score is used to evaluate the reliability of a model, thus models with DOPE Z-scores from -1 and below are considered to be reliable.

2.3.4. Model evaluation

The PDB file of the model or protein structure to be evaluated is usually the input for most model assessment programs, and this was the same for all the programs used in this study. The stereochemical properties of the models were evaluated by PROCHECK, which was downloaded from htttp://www.ebi.ac.uk/thornton-srv/software/PROCHECK/, ran on the terminal by providing the PDB file of the model, the PDB code and the resolution of the template structure(s). ProSA and MetaMQAP II were run online using the co-ordinates of the PDB file as

input, and Energy plots were returned for the former while PDB with GDT_TS scores were returned for the latter.

2.4. Results

2.4.1. Sequence retrieval

The protein sequences of our target proteins FP2', VP2 and VP3 were retrieved from UniProt. FP2 had a sequence length of 482 amino acids, VP2 and VP3 had 487 and 495 residues respectively. The first + 240 residues of the three proteases were the prodomain and were not used in the homology modeling. All other *Plasmodium* cysteine protease sequences (as indicated in the methodology section **2.3**) were retrieved from UniProt and the structural data of other cysteine proteases from different organisms were obtained by NCBI-BLAST search against PDB. The full length sequence of human procathepsin K was downloaded from PDB (it has the pdb entry: 1BY8) (Lalonde *et al*., 1999) and used to perform a BLAST search against PDB in order to retrieve a template with the least sequence identity.

2.4.2. Sequence alignment

Three MSA programs used for sequence analyzes were: ClustalW2, T-COFFEE and PROMALS3D, from which ClustalW2 indicates only the target-template alignments (Figure 2.4). FP2 was used as a reference structure and the numbering is indicated as it appears in the PDB file. For simplicity purposes the numbering was left at 1 to + 250 in the rest of the other alignment programs, T-COFFEE (Figure A: 1) and PROMALS3D (Figure A: 2). Figure 2.4 shows the active site for target-templates, the N-terminal insertion (Nose-like motif) and C-terminal extension (arm-

like motif) in arrows and black boxes respectively. The ClustalW2 alignment was used to generate PIR files for each of the target proteases and used for the model building stage.

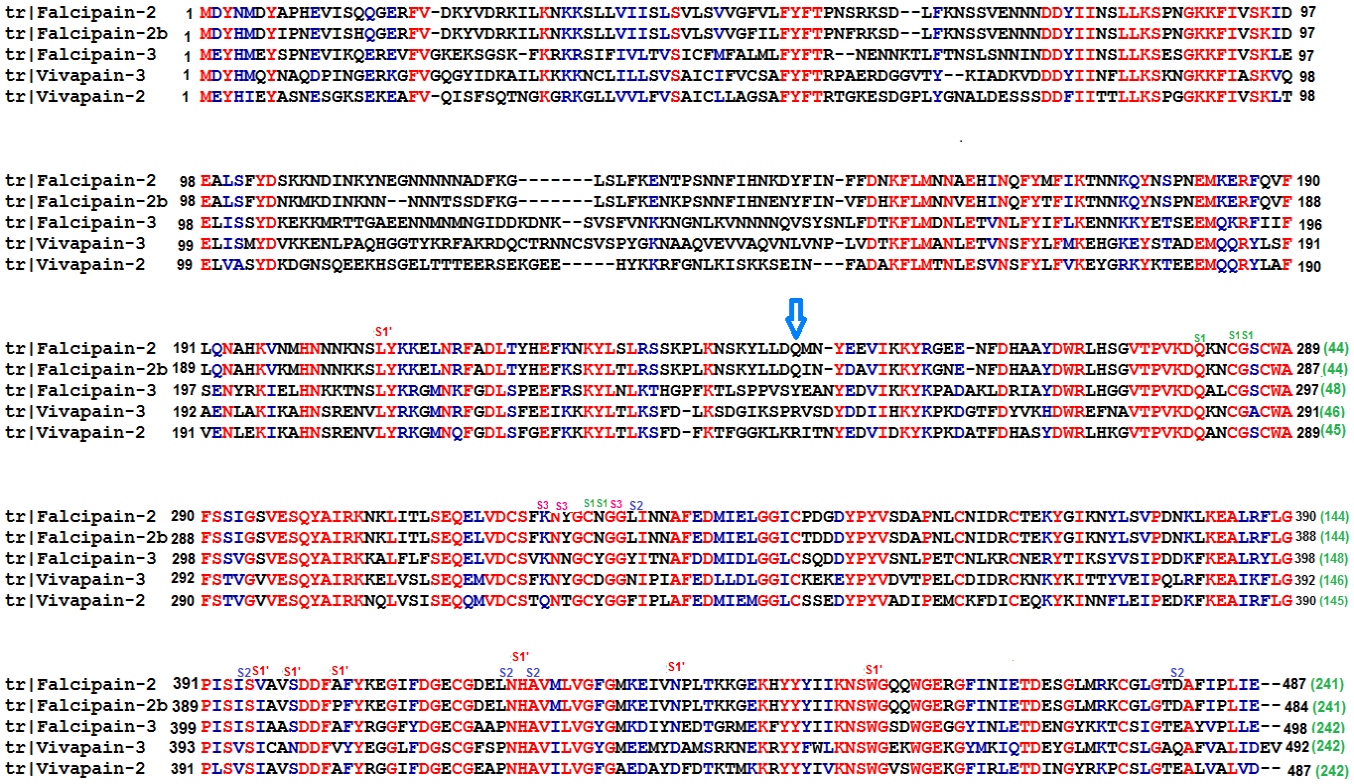

Figure 2. 4: Target-template alignments generated by ClustalW2, the numbering adjusted in Bio-edit using FP2 as a guide. The-N-terminal extension and C-terminal insert are clearly marked in red boxes. Residues in the substrate binding pockets are labeled as S1, S1', S2 and S3. The numbers are the actual residue numbers including the prodomain and the ones in green are the residue numbers when the prodomain is cleaved off

The target template alignment between human procathepsin K and major house dust mite proteolytic allergen was generated in all three programs. The structural alignment was generated from the protein sequences as obtained from PDB and it was used for model building.

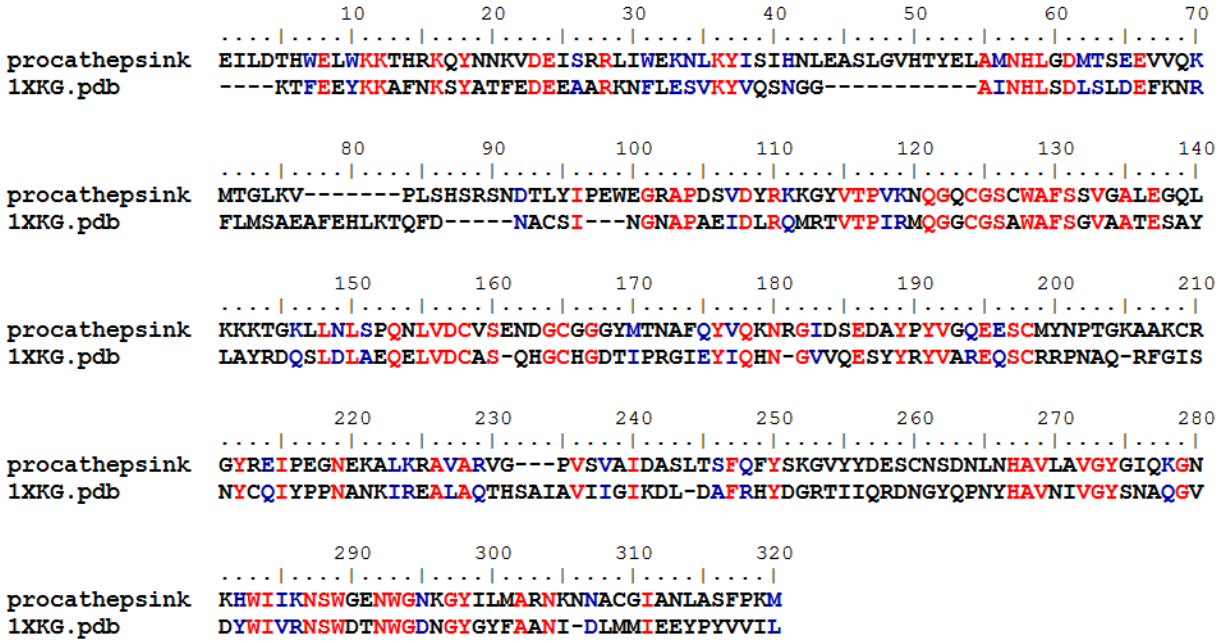

Figure 2. 5: Target-template alignment file that was used for human cathepsin K modeling

2.4.3 Homology models of FP2', VP2, VP3 and human procathepsin K

- ***FP2'***

FP2' models were built in MOD9v7and assessed based on their DOPE z-score. Five best models (with the lowest DOPE z-score) were superimposed on the template structure FP2 (Figure 2.7). The sequence identity between the target protein (FP2') and template (FP2) was 96%, which was the basis on which it was selected. These two proteases are also expressed in the same organism. Below is a Table with the DOPE z-score, RMSD and GDT_TS scores of the best five models.

Table 2. 2: FP2' five best models based on their DOPE z-score, RMSD and GDT_TS score

FP2' model No	DOPE z-score	RMSD (Å)	GDT_TS score
Model 77	-1.049	0.30	56.74
Model 42	-1.093	0.28	60.37
Model 19	-1.097	0.30	63.28
Model 54	-1.097	0.34	59.47
Model 70	-1.100	0.36	61.41

The models built had good DOPE z-scores ranging from -1.05 to -1.23. Model 42 was analyzed further with PROCHECK, ProSA and MetaMQAP II as it had the lowest RMSD and DOPE z-score of -1.09. Below in Figure 2.5, is the model 42, selected as the best model, and together with the other four best models superimposed on the template structure FP2.

Figure 2. 6: Model structure of FP2' generated by MODELLER 9v7 (left), the active site, N-terminal extension (FP2'_nose) and C-terminal insertion (FP2'_arm) are clearly labeled by arrows. (right) All five models of FP2' superimposed to the C_α of FP2 (green) and model 77(cyan), Model 42 (pink), model 19 (yellow), model 54 (blue) and model 70 (red).

PROCHECK

PROCHECK was used to assess the stereochemical quality of the FP2' model built. Amongst other plots, PROCHECK generates the Ramachandran plot as part of its output files. Figure 2.7 shows that the Ramachandran plots of FP2' and FP2 (template) compare well with each other.

Almost all (90.4%) of the residues in our model were in the most favoured region, 2% residues in the generously allowed regions, 8.2 % residues in additional allowed regions and only 1 residues (ASN 118) which accounts for 0.5% in the disallowed region. Except for ASN 118 in the disallowed region, both FP2 and FP2' have LYS and SER residue in the generously allowed region.

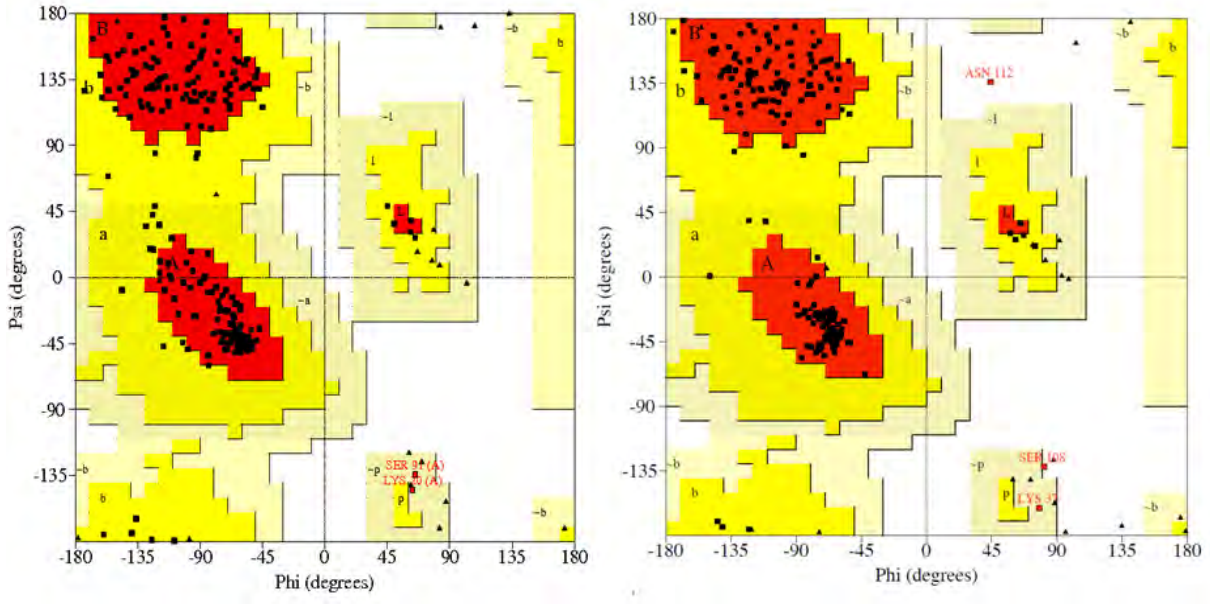

Figure 2. 7: Ramachandran plot for the template and target proteins FP2 (left) and FP2' (right) respectively. Both plots were generated by PROCHECK. Most sterically favoured region (red), additional allowed regions ((dark yellow), generously allowed regions (light yellow) and disallowed regions (white). α- helix (A), Left handed helix (L) and β-sheet (B)

ProSA

The overall fold of FP2' was also evaluated with ProSA, which generated Z-score and an energy plot which are consistent with template (FP2) values. The Z-scores of FP2 (A) and FP2' (B) are in the same range which is -6.57 and -7.19 respectively. The energy plots of FP2 (C) and FP2' (D) are also consistent with each other over both the 10 and 40 amino acid window size.

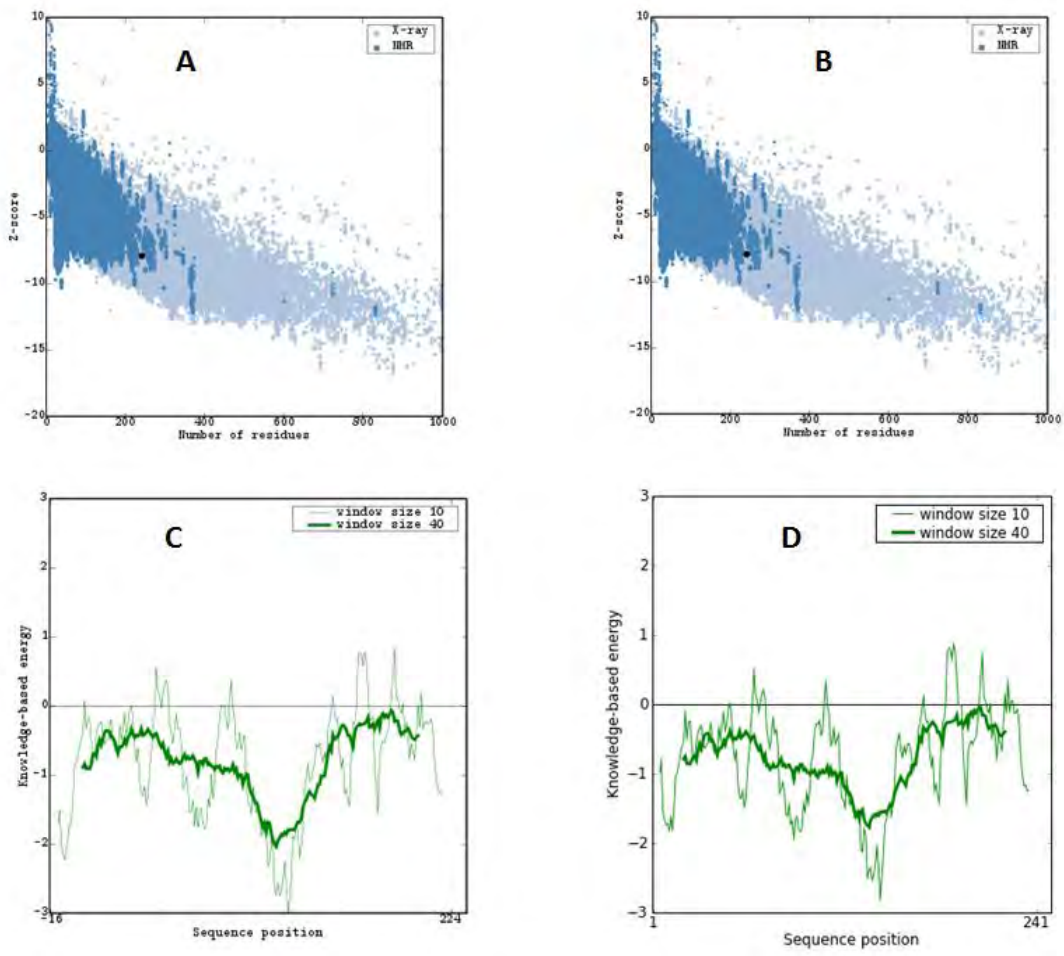

Figure 2. 8: ProSA analysis for the model structure of FP2' and the template structure used for modeling FP2. Z-scores of FP2 (A) and FP2' (B) with the light blue area indicating all protein structures in PDB that were solved by X-ray crystallography and Dark blue indicating all structures that were solved by NMR. Energy plots of FP2 (C) and FP2' (D) with light green indicating amino acid residues averaged over 10 windows and dark green average window size of 40.

MetaMQAP II

MetaMQAPII results revealed a high conservation at the active site which is shown by the blue colour in Figure 2.9; however there are also less conserved residues in the arm-motif. The N-terminal insert (nose-like motif) is moderately conserved. The overall quality of the model is good and can be considered an accurate prediction of FP2' protein structure.

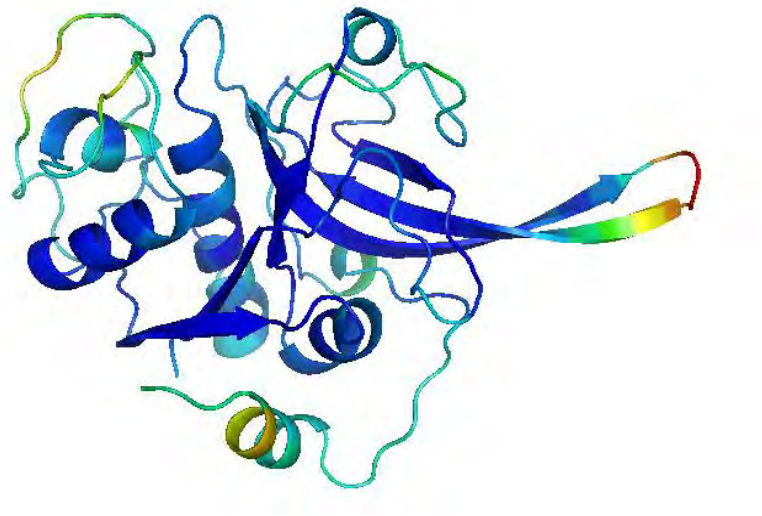

Figure 2. 9: FP2' model after final evaluation by MetaMQAP II which indicates statistically favourable residues in blue and non favourable ones in red

- ***VP2***

Vivapain-2 shares 63% sequence identity with FP2 and 66% with FP3, these two proteases were used as templates for building the model of VP2. One hundred models were generated using MODELLER and ranked by their DOPE z-scores. All the models were relatively good and predictions could be considered accurate, the models had DOPE z-scores ranging from -0.76 to -1.08. The first five models with the lowest DOPE z-scores, RMSD and GDT_TS score are indicated below in Table 2.3. .

Table 2. 3: VP2 five best models based on their DOPE z-score, RMSD and GDT_TS score

VP2 model No	DOPE z-score	RMSD (Å) FP2	RMSD (Å) FP3	GDT_TS score
Model 56	-0.757	0.42	0.43	62.72
Model 64	-0.826	0.47	0.43	59.96
Model 23	-0.835	0.44	0.41	64.05
Model 83	-0.832	0.39	0.44	65.02
Model 90	-0.846	0.44	0.44	66.01

The five best models of VP2 had slightly different RMSD values; model 83 was selected for further evaluation as it had the lowest RMSD deviation for the template structure(s). The DOPE z-score of model 83 was -0.84, showing that model was predicted accurately.

Figure 2. 10: Model structure of VP2 generated by MODELLER 9v7, with the α-helix in blue, β-strands in magenta and turns in light violet, The active site, C-terminal insert and N-terminal extension clearly marked by arrows (Top). At the right is five models of VP2 superimposed to the C_α of FP2 (green) and FP3 (cyan) and model 56 (pink), model 64 (yellow), model 23 (light pink), model 83 (light grey) and model 90 (blue).

PROCHECK

The models of VP2 were initially built using FP2 only as a template; however, the models contained many errors. Therefore, we resolved to use both FP3 and FP2 as templates for building one hundred models of VP2. All of the models were statistically assessed by their DOPE z-scores. The errors in the models were also reviewed by RMSD deviations to template structures. Model 2 was selected for further evaluation. Stereochemical assessment by PROCHECK showed that 91% of the model residues were in the most favourable regions. None of the VP2 residues were found in the disallowed region, the rest of the 8.1 % and 1.0% were in the additionally allowed region and generously allowed regions respectively (Figure 2.11 (middle). This model compares well with its template Ramachandran plots, which is a good indication of the accuracy of the model structure. There are two ALA residues found in the

additionally allowed regions for VP2, SER and LYS for FP2 and ASP and LYS in the case of FP3 , but this is only due to differences in the sequences of the structures.

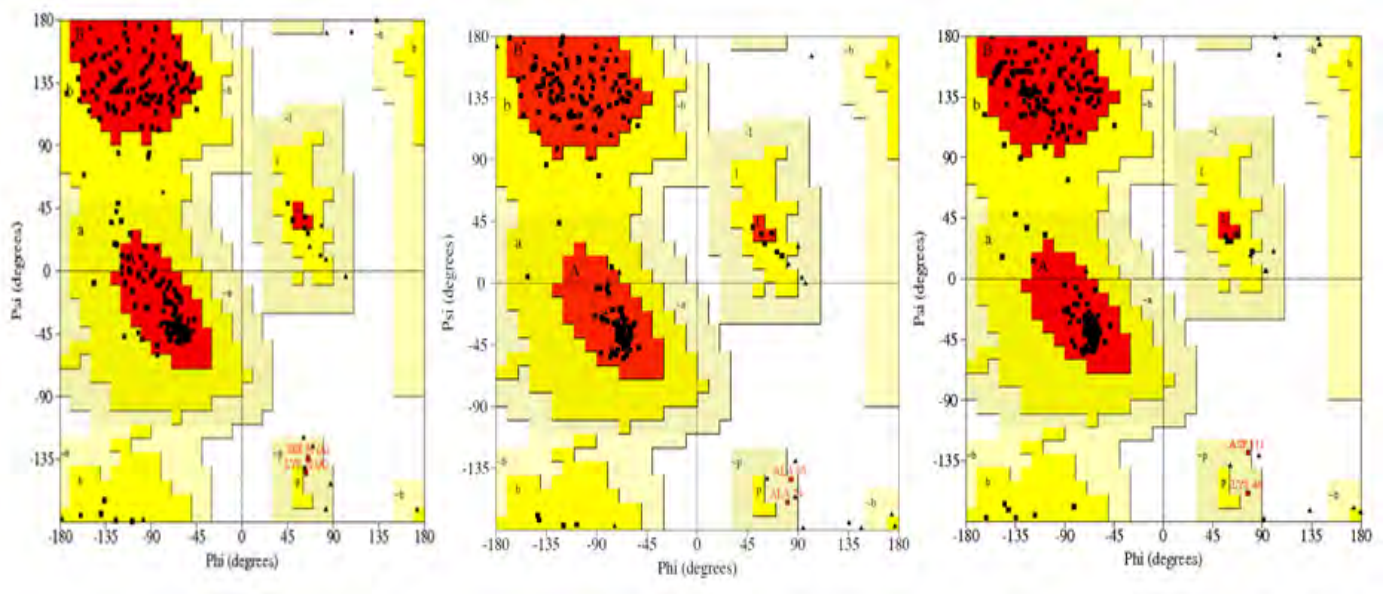

Figure 2. 11: Ramachandran plots of the templates [2OUL (left) and 3BWK (right)] and target protein VP2 (middle) generated by PROCHECK. . Most sterically favoured region (red), additional allowed regions (dark yellow), generously allowed regions (light yellow) and disallowed regions (white). α- helix (A), Left handed helix (L) and β-sheet (B).

ProSA

ProSA generated two plots: Z-score and an energy plot, of which VP2 is a perfect fit within the structures currently available in PDB. Figure 2.12 shows the Z-score plots of FP2 (A), VP2 (B) and FP3 (C) which are -6.57, -7.09 and -7.49 respectively; the difference between these scores are subtle and raise confidence in the predicted model. A high level of consistency across both the 10 window size and 40 window size energy plot was observed when comparing the model (E) with its templates (D and F). High energy residues in the model were also found to be high energy residues in templates.

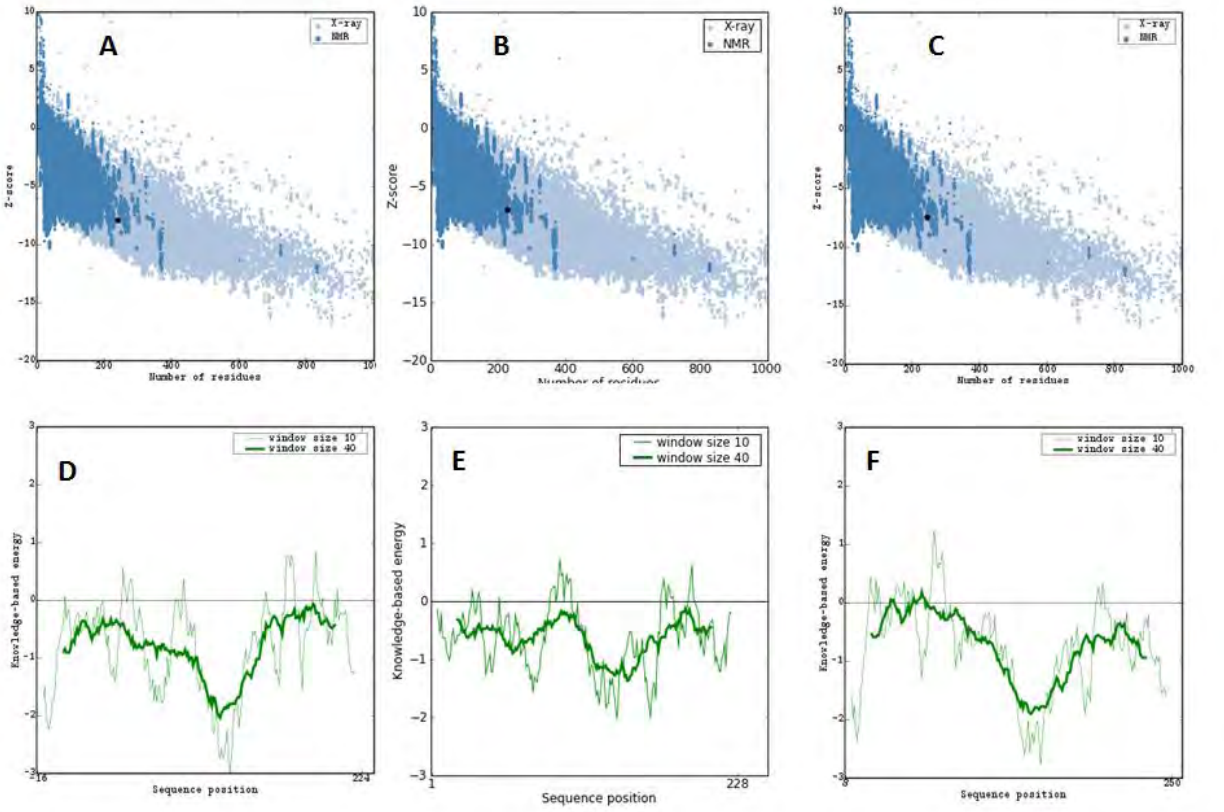

Figure 2. 12: ProSA analysis for the model structure of VP2 and the template structures used for modeling FP2 and FP3. Zscores of FP2 (A) , VP2 (B) and FP3(C) with the light blue area indicating all protein structures in PDB that were solved by X-ray crystallography and dark blue indicating all structures that were solved by NMR. Energy plots of FP2 (D), VP2 (E) and FP3 (F) with light green indicating amino acid residues averaged over 10 windows and Dark green average window size of 40.

MetaMQAP II

The final evaluation by MetaMQAP II revealed a rather well conserved structure, except for residue ALA-187 at the C-terminal insert. The red colour indicates unfavourable regions and this may be accounted for by the fact that this residue is substituted by ILE in both FP2 and FP3. Though these amino acids may have the same properties, the modeling of ALA may not have been optimal due to its small size.

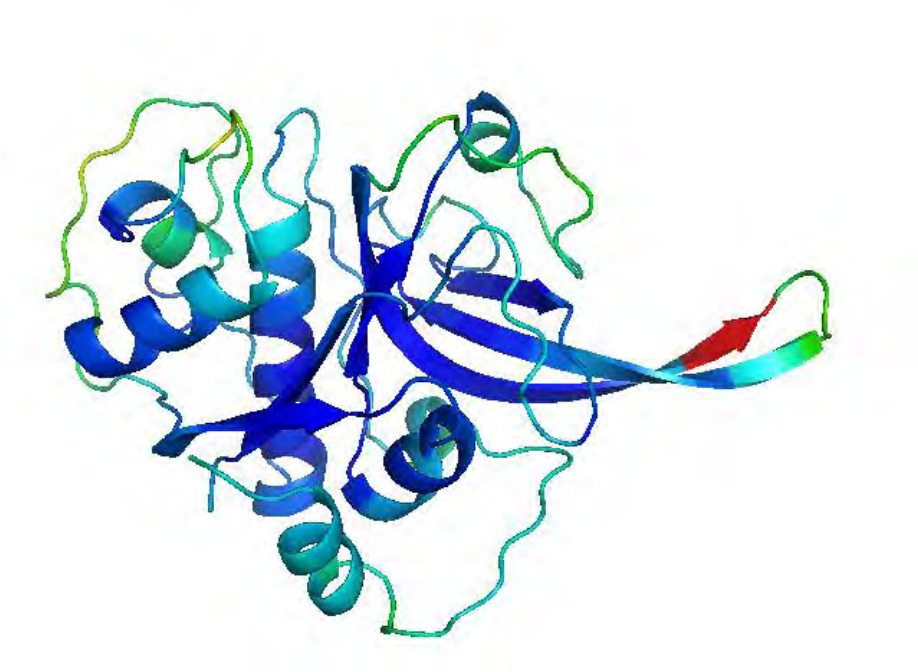

Figure 2. 13: Final evaluation of VP2 model by MetaMQAP II which indicates statistically favourable residues in blue and non favourable ones in red

The overall conclusion for the predicted model is that based on the results of the various programs, the sequence identity between the target and templates, this the best prediction possible for the 3D structure of VP2.

- ***VP3***

The model of vivapain-3 was built using the structures of FP2 and FP3 as templates, with sequence identity of 46% and 50 % respectively. MODELLER 9v7 was used to build the models which were evaluated by DOPE z-score. RMSD deviations based on the C_α backbone of the top five models and the templates were also determined in Pymol. DOPE z-score of the models ranged from -1.05 and -1.23, and were sorted in chronological order from low to high DOPE z-scores, which indicates very good models

Table 2. 4: VP3 five best models based on their DOPE z-score, RMSD and GDT_TS score

VP2 model No	DOPE z-score	RMSD (Å) FP2	RMSD (Å) FP3	GDT_TS score
Model 11	-0.929	0.45	0.44	67.15
Model 22	-0.930	0.50	0.37	61.34
Model 25	-0.942	0.53	0.39	61.05
Model 62	-0.960	0.53	0.36	63.22
Model 79	-0.963	0.46	0.49	62.34

Models of VP3 had good DOPE z-scores and their RMSD in Å were the same as those of structures already solved by experimental techniques. Model 11 was selected as the best representation of the typical fold of VP3. These elucidations were made based on the DOPE z-score (slightly more than -1 in Table 2.4) and the RMSD score of the models-templates superimposition. The GDT_TS score of model 11 was also better than the other four models.

Figure 2. 14: Model structure of VP3 generated by MODELLER 9v7, with the α-helix in red, β-strands in yellow and turns in green, The active site, C-terminal insert and N-terminal extension clearly marked by arrows (left). At the right is five models of VP3 superimposed to the C_α of FP2 (green) and FP3 (red) and model 11 (pink), model 22 (yellow), model 25 (blue) , model 62 (cyan) and model 79(orange)

PROCHECK

Stereochemical analysis observed in the Ramachandran plot generated by PROCHECK for the model of VP3 resulted in uncovering that 93% of the residues in the model are found within the most favourable region. There were no residues in the disallowed region, 6% and 1% were found in the additionally allowed region and generously allowed regions respectively (Figure 2.15: middle). ASP-108 and LYS-37 are the residues found within the generously allowed regions of the protein model, and these results compare well with templates particularly those of FP3, which has ASP and LYS also in this region. FP2 also has SER and LYS found within this region which proves that model is not deviant from its templates.

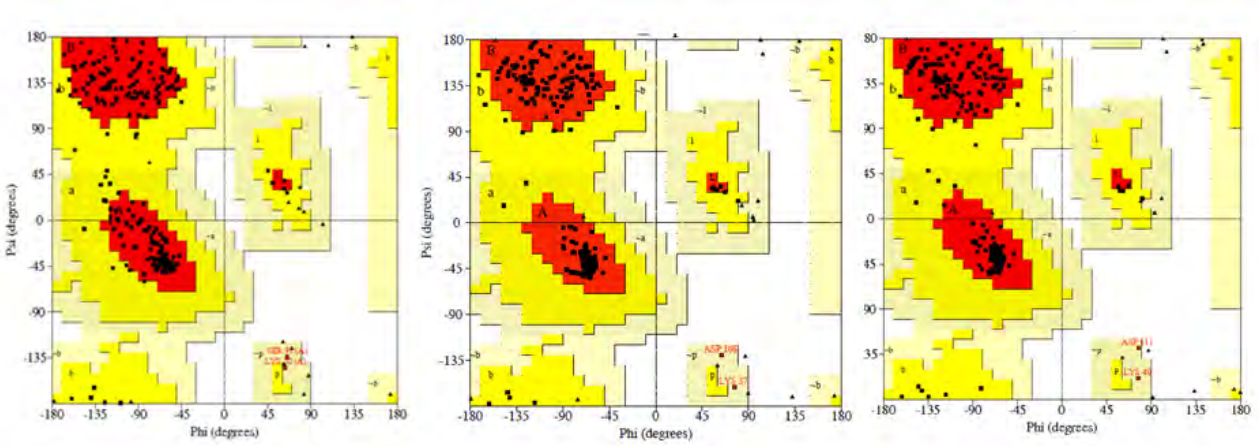

Figure 2. 15: Ramachandran plots of the templates [2OUL (left) and 3BWK (right)] and target protein VP3 (middle) generated by PROCHECK. . Most sterically favoured region (red), additional allowed regions (dark yellow), generously allowed regions (light yellow) and disallowed regions (white). α- helix (A), Left handed helix (L) and β-sheet (B).

ProSA

ProSA generated two plots: Z-score and energy plot, of which VP3 is a perfect fit within the structures currently available in PDB. Figure 2.16 shows the Z-score plots of FP2 (A), VP3 (B) and FP3 (C) which are -6.57, -7.87 and -7.49. The Z-score of VP3 is slightly higher than its templates, but it is a perfect fit within the structures in PDB. A high level of consistency across both the 10

window size and 40 window size energy plot was observed when comparing the model (E) with its templates (D and F). High energy residues in the model were also found to be high energy residues in templates.

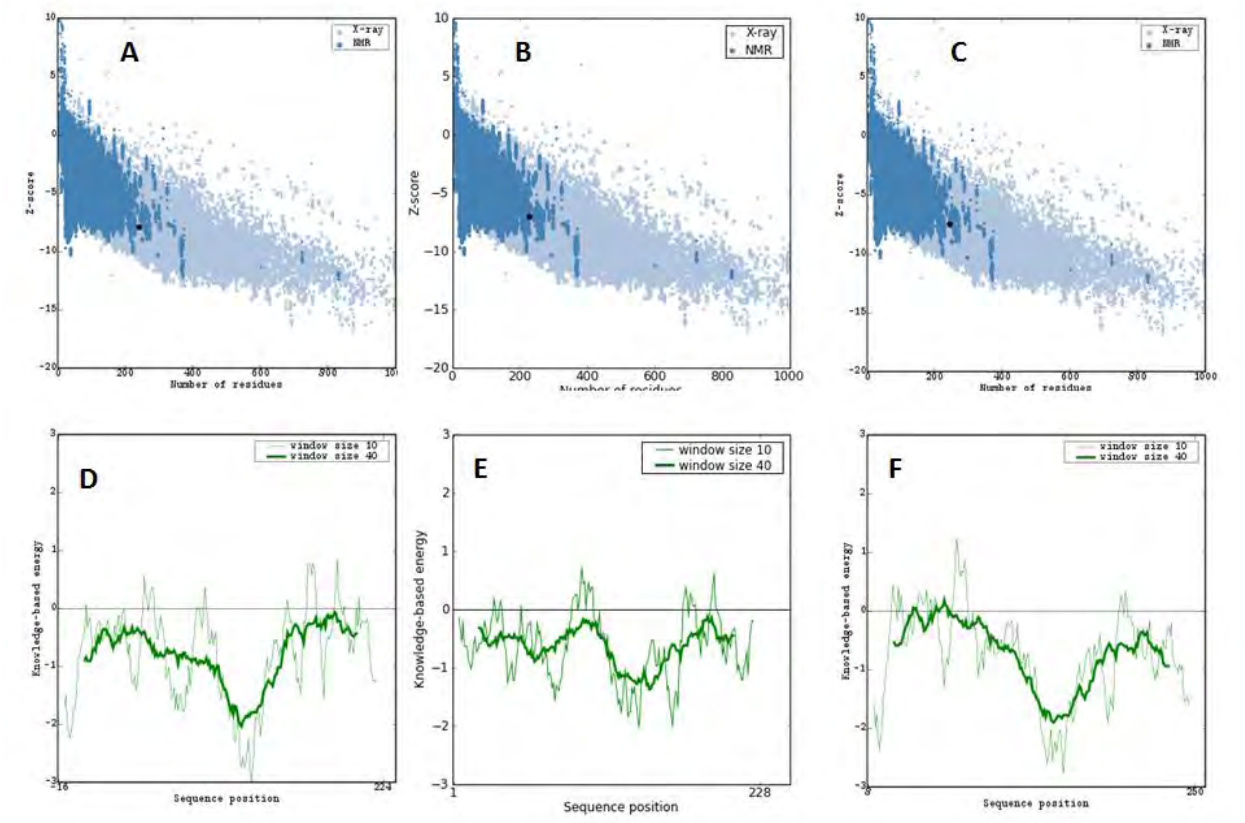

Figure 2. 16: ProSA analysis for the model structure of VP3 and the template structure used for modeling FP2 and FP3. Z-scores of FP2 (A) , VP3 (B) and FP3(C) with the light blue area indicating all protein structures in PDB that were solved by X-ray crystallography and Dark blue indicating all structures that were solved by NMR. Energy plots of FP2 (D), VP3 (E) and FP3 (F) with light green indicating amino acid residues averaged over 10 windows and dark green average window size of 40.

MetaMQAP II

The final evaluation by MetaMQAP II revealed a rather well conserved structure, except for the loop region right at the C-terminal insert. The red colour indicates an unfavourable region and this may be accounted for by the slightly low sequence identity of amino acids in that area of the protein as revealed by MSA. The overall conclusion for the predicted model is that based on

the results of the various programs, the sequence identity between the target and templates, this the best prediction possible for the 3D structure of VP3.

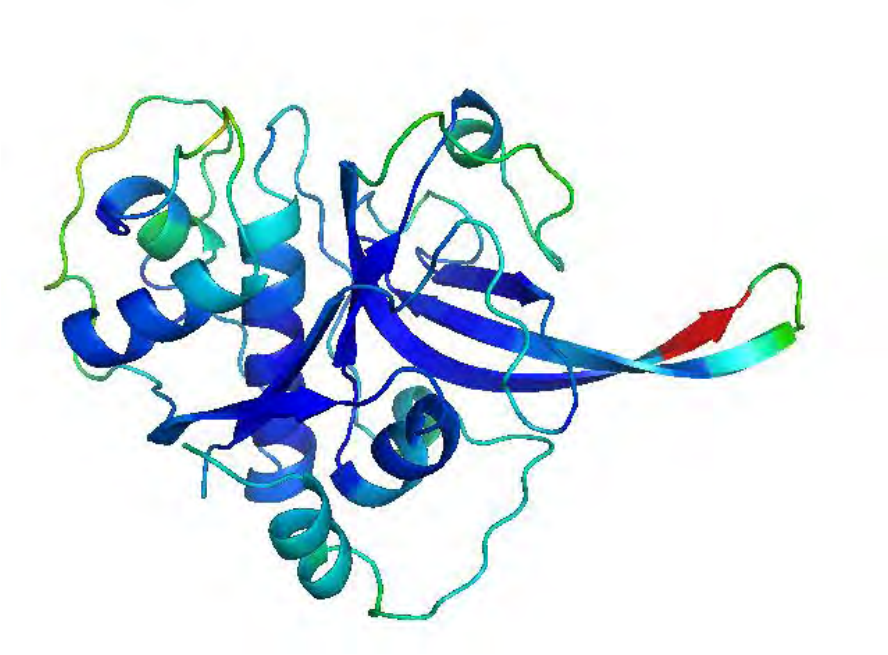

Figure 2. 17: Final evaluation of VP3 model by MetaMQAP II which indicates statistically favourable residues in blue and non favourable ones in red

Human procathepsin K

Human procathepsin K shares 26% sequence identity with major house dust mite proteolytic allergen. This particular template structure was used to built the model of human procathepsin K because it has a low sequence identity to it and therefore reflects the reliability of the comparative modeling program: MODELLER. All one hundred models were generated and ranked based on their DOPE z-scores. The RMSD and GDT_TS scores were also used to rank the models

Table 2. 5: Human procathepsin K five best models based on their DOPE z-score, RMSD and GDT_TS score

Human procathepsin K model No	DOPE z-score	RMSD (Å) 1BY8	RMSD (Å) 1XKG	GDT_TS score
Model 6	-0.37	2.12	0.70	65.17
Model 24	-0.71	1.80	0.64	67.36
Model 29	-0.64	2.06	0.76	65.36
Model 13	-0.51	2.56	0.71	62.09
Model 9	-0.48	2.32	0.86	61.85

The model with the best DOPE z-score was superimposed on the experimentally solved structure of human procathepsin K. The RMSD of the Cα-co-ordinates of the model to the template was found to be 1.8Å, which is good considering the sequence identity of the model and the template.

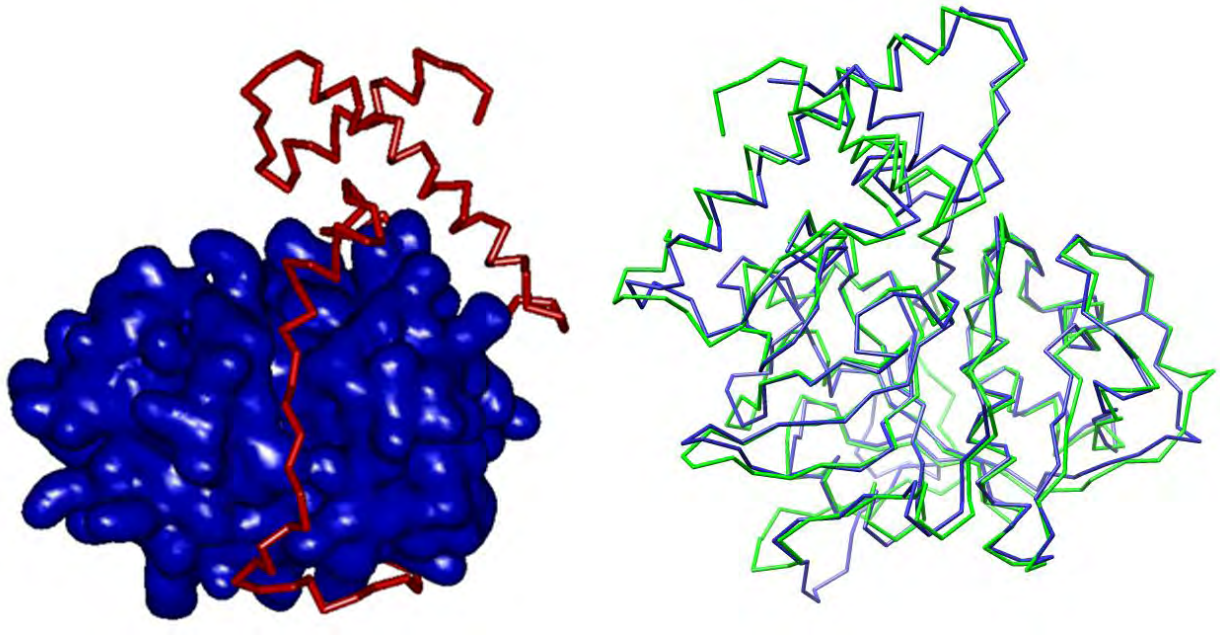

Figure 2.19: The model of human procathepsinK, also the model superimposed to its real structure. A- Model indicating the inhibition of cathepsin K inhibition by the prodomain. The mature domain in blue and the prodomain (red) binds cathepsin K like other cysteine proteases. B- Procathepsin K model (green) superimposed to its experimentally solved structure (Blue).

2.4 Discussion

Plasmodium cysteine proteases contain several unique features not found in human or papain-like cysteine proteases of other species (Verissimo *et al*., 2008). Their prodomain is much larger in length and there are two inserts. The uniqueness of *Plasmodium* cysteine proteases has made them an interesting case study and some of them have been identified and validated as potential drug targets (Lecaille *et al*., 2002) and chemotherapeutic targets for malaria (Rosenthal *et al*., 2002). Most of the studies looking into the possible drug targets against malaria parasites are mainly focused at the hemoglobin degradation stage of *Plasmodium* life cycle. Cysteine protease inhibitors have been demonstrated both *in vivo* and *in vitro* to block hemoglobin degradation, the process which has been correlated with blocking parasite development and indeed experimentally proved to cure mouse models (Rosenthal *et al*., 1998; Rosenthal *et al*., 1991; Rosenthal *et al*., 1993; Oslon *et al*., 1999). FP2 and FP3 are validated drug, therefore pursuing studies into their homologues and orthologues FP2', VP2 and VP3 will offer some insight into a synergic inhibitor. In depth studies of FP2', VP2 and VP3 have mostly been delayed by the fact their 3D structures have not yet been solved experimentally.

Homology modeling has been an excellent breakthrough in bioinformatics and molecular biology. This process helps predict the 3D structures of proteins not yet elucidated by experimental techniques and was valuable in the prediction of the structures of FP2', VP2 and VP3. PDB has 155 structural entries of papain-family cysteine proteases, only 6 of which are from *P. falciparum* FP2 (4) and FP3 (2). The crystal structure of FP2 (PDB code: 2GHU) correspond to the free FP2 (Hogg *et al*., 2006), and the crystal structures with PDB codes 1YVB (Wang *et al*., 2006), 2OUL (Wang *et al*., 2007) and 3BPF (Kerr *et al.,* 2009) correspond to FP2 in complex with inhibitors cystatin, chagasin and epoxysiccinate E-64 respectively. Although 3BPF was solved at a better resolution than the other 3 crystal structures of FP2, it is not complexed to a protein/ peptide inhibitor. Therefore the crystal structure of FP2 (20UL) was used as a template because it is complexed with an inhibitor, this structural distortion is important for

substrate interaction studies and identification of residues involved in binding which is useful for inhibitor design. Although 1YVB is also complexed with an inhibitor, 2OUL was solved at a better resolution (2.70 Å (1YVB) and 2.20 Å (2OUL). The two experimental structures of FP3 are in a complex form. The crystal structure of FP3 (PDB code: 3BPM) is in complex with aldehyde leupetin (Kerr *et al*., 2009) and a well known vinyl sulfone inhibitor K11017 is in complex with 3BWK (Kerr *et al*., 2009). For the homology modeling, the FP3 structure 3BWK was used as a template as it is the latest entry. An additional feature that was looked into when selecting FP2 and FP3 as templates was that they share significant sequence identity to target enzymes and also arise from the same ancestor (*Plasmodium*). It is well documented that cysteine proteases are synthesized as a zymogen with a prodomain preventing premature activation of the protease (Lecaille *et al*., 2002; Sajid and McKerrow, 2002). Similar to this observation, the sequences of our target proteins contained an unusually large prodomain of about 240 amino acids. All the target proteases and their templates had the prodomain (Figure 2.4).The interest of this study was to model the mature (functional state) cysteine proteases; therefore the prodomain was cleaved off using FP2 and FP3 as guides. However the inhibitory activity of the prodomain was illustrated in the model of human procathepsin K (Figure 2.19). As indicated in **section 2.2.3.**, there are several homology modeling programs and the reliability of MODELLER, the spatial restraint modeling program used in our study, was demonstrated by its ability to predict the 3D structure of human procathepsin K. Human procathepsin K (pdb code: 1BY8), which has been elucidated by experimental techniques, was built based on the structure of a distantly related protease which shares 26% identity. The generated model of human procathepsin K built from MODELLER was superimposed on its template and to the experimentally determined structure of human procathepsin K and found to have an RMSD of 0.64 Å and 1.80 Å respectively. Once the reliability of this most cited homology modeling program was confirmed, target proteases for our studies were predicted.

Prior to the model building step, the sequences of our target proteases and other papain-family cysteine proteases from different organism and their *Plasmodium* homologs were aligned using

three MSA programs. T-COFFEE, ClustalW2 and PROMALS3D were used for the sequence alignment and all three programs employ different approaches for sequence alignment. The level of sequence conservation in T-COFFEE (Notredame *et al*., 2000) is indicated by colours with orange being high sequence conservation and red less conserved sequences. Our results have shown general conservation across cysteine proteases and this is because of the high sequence identity in all the sequences in the alignment. PROMALS3D confirms T-COFFEE results, in the case of the former; the conservation is shown by numbers which are calculated by the statistically based program AL2CO. The conservation numbers range from 0 to 9, indicating low and high conservation respectively. The conservation scores in our aligned sequences are all above 5. The results from the three programs were consistent though they employ different algorithms. The most common errors encountered in homology modeling are mismatched sequence alignment and this is prominent when the sequence identities of the target protein(s) and template(s) are below 30 % (Eswar *et al*., 2003). The general pattern observed in the study was a high level of sequence conservation across papain-like cysteine proteases, especially the active site, therefore no mismatches were expected and none were incurred. An optimal alignment was generated between *Plasmodium* cysteine proteases and other cysteine proteases from different organisms. It is clear from the sequence alignment that the protease prodomains are weakly conserved compared to the mature protease. The sequences of the mature proteases on the other hand are highly conserved across all the proteases except for the C-terminal insert (which has been labeled hemoglobin binding site)(Wang *et al*., 2006). This unique motif is highly flexible; consists of two β-pleated sheets and protrudes far from the active site. The unique features found in *Plasmodium* cysteine proteases are also clearly observed in the sequence alignment as the absence of this motif is observed in other cysteine proteases. Studies have suggested that the N-terminal extension (nose-like) motif is involved in the folding of the mature protease without the presence of the prodomain (Sijiwali *et al*., 2002; Pandey *et al*., 2009). Based on the sequence alignment, this function can neither be rejected nor confirmed, however we observed that the N-terminal extension is also highly conserved in *Plasmodium* cysteine proteases. Although FP2 and FP2' have two deletions between the ASN3 and TYR 4, while the other three proteases contain polar

amino acids ASN, ASP and ASN for FP3, VP3 and VP2 respectively. Also, between GLU 15 and ASN 16 (numbering based on FP2), there is an insert, whereas FP3, VP3, VP2 contain small, non-polar, hydrophobic amino acids ALA GLY and ALA respectively. Therefore, due to the high sequence conservation of this motif, we can speculate that the functional role of this feature in these proteases is highly conserved.

All 3D models of the four target cysteine proteases were built in MODELLER, which generally begins the process of generating a 3D model of the target protein by generating many constraints or restraints on the protein of interest. The concept behind this method is similar to that used for NMR-derived restraints. Mainly, the restraints are obtained by making the assumption that the corresponding distance between aligned residues in the template and target structures are similar. The model is then obtained by minimizing the violations to these restraints (Šali and Blundell, 1993). Due to the fact that modeling by satisfaction of spatial restraint uses many different types of information, it remains one of the most promising techniques in comparative modeling. As the sequence similarity between target protein and template structure diverges, so does the accuracy of packing the side chains in the protein core. Often, at a low sequence identity (<30%), the conformation of the side chain is less conserved; this is a pitfall of many comparative modeling methods (Sanchez and Šali, 1997; Marti-Renom *et al*. 2000). The most critical errors in side-chain packing occur in the functional regions of the protein, such as active sites and ligand-binding site (Sanchez and Šali, 1997). The quality of side-chains in a model can be analyzed by looking at the RMSD for all atoms or detecting a fraction of the rotamers found (Waller and Elofsson, 2004). In the latter, it is advisable to use SCRWL, as it is well known to build better side chains than modeling programs. SCRWL is not a real comparative modeling program but as indicated by programs evaluated by Waller and Elofsson, (2004), it is a program best used for side-chain packing when the sequence identity between target and template is <30%. However, with an increase in sequence identity, the accuracy of this program is brought into question as the information about conserved rotamers is not used. Also, with an RMSD of 0.5Å the accuracy of side-chain modeling on a fixed backbone decreases

rapidly (Chung and Subbiah, 1996; Jacobson and Šali, 2004). In the case of our study, this extra measure was not necessary as the RMSD values obtained were all below 0.5Å (Table 2.2-2.4); therefore it can be assumed that side-chains were predicted accurately. However, the structural deviation between target and template which is indicated by the RMSD value was not the only parameter used to obtain confidence in the model generated.

DOPE is a statistical potential based tool used for assessing the protein prediction of homology models (Shen and Šali, 2006). DOPE score is useful for calculating the energy of the protein model generated through much iteration by the spatial restraint program MODELLER. Calculation of DOPE score is now implemented in the MODELLER program itself. The DOPE method is generally used to assess the quality of a structural model as a whole. One other thing DOPE does is to generate a residue-by-residue energy profile for the input model; therefore it is possible to spot the problematic regions within a model (Eramian *et al*., 2008). The DOPE score is unnormalized on the basis of the protein size and has an arbitrary scale, therefore, the scores of different proteins cannot be compared. Therefore, DOPE z-score which is the normalized DOPE is used to compare the scores of different models. DOPE z-score is a command within MODELLER that assesses the quality of the model using the normalized DOPE method. In the DOPE z-score, the models assigned a positive value are likely to be poor while those with a score -1 are likely to be near native. This is the reason for using DOPE z-score to compare the scores of the one hundred models which were generated for target proteins and select the best prediction. All the models generated for the enzymes which were used in these studies were within the near-native region. Thus, DOPE z-score was a good reflection of the reliability of the models constructed.

The role of FP2' (the last *P. falciparum* cysteine protease) to be discovered is still unknown (Singh *et al*., 2006). FP2' shares a significantly large sequence identity with FP2 (96%) and has been implicated in hemoglobin degradation. Therefore due to FP2' significant sequence identity to FP2, the latter protease model generated in this particular study was used to study its

interaction with cysteine proteases principal substrate hemoglobin. FP2' model was built using FP2 as a template and as expected the overall 3D fold of two cysteine proteases was the same, typical of papain-family cysteine proteases. Although the overall fold of FP2' and FP2 were the same, there were some notable differences between the two proteases especially at the amino acid level. Two amino acid point mutations were observed in FP2, substrate binding site residues were substituted from: VAL 150 (FP2) → ILE 150 (FP2') and ALA 157 → PRO 157. Both mutations were in the S1' subsite (Figure 2.21).

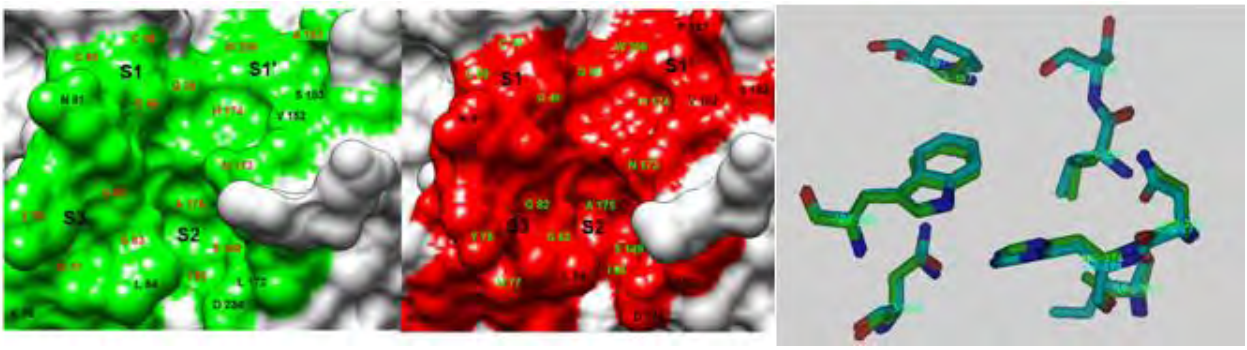

Figure 2. 21: FP2 (Left) and FP2' (right) subsite residues. On the right S1' subsite residues in stick representation, the only subsite with mutations.

Earlier studies have managed to achieve the expression of recombinant FP2' in bacterial cells (Singh *et al*., 2006). The findings of FP2' expression and purification studies have demonstrated its ability to bind to typical papain-like cysteine proteases substrates; however, it appears to have different substrate specificity to FP2. Based on the molecular modeling and amino acid analysis of the present study, we therefore, suggest that the difference in specificity could be attributed by the amino acid mutations. Although the amino acid properties of the substituted residues are not as significant, the orientation and the size of the substituted residues may be in position to bind differently to the substrates or inhibitors. The effect of these mutation on FP2'-hemoglobin binding were further explored in chapter 3, where protein-protein interaction studies were carried out.

Several studies have pointed out that *P.vivax* is the most neglected *Plasmodium* species, also that its infections are most widespread; it is the fastest and most rapid cause of malaria in many countries because it's infections from may relapse (Mendis *et al*., 2001; Baird, 2004; Barnwell *et al*., 2007). Both *P. vivax* and *P. falciparum* accounts for millions of deaths reported annually. VP2 and VP3 cysteine proteases from *P. vivax* have been identified and biochemically characterized (Na *et al*., 2004): These two proteases have also been shown to play a role in hemoglobin degradation consistent with their *P. falciparum* orthologues FP2, FP2' and FP3. However, the 3D structures of VP2 and VP3 have not been solved by experimental techniques due to the limitations of or lack of proper *in vivo* culture of *P. vivax*. Therefore, in order to fully appreciate and study the roles of VP2 and VP3, their structural information can only be obtained by the computational method homology modeling. Homology modeling was employed to construct the 3D structures of VP2 and VP3.

Typical papain-like cysteine protease features were observed in VP2 and VP3. They consist of two domains: left and right with an active site on either site of the domains (Lecaille *et al*., 2002; Sajid and McKerrow, 2002). Residues critical for enzymatic degradation of cysteine proteases: CYS, HIS and ASN were also mapped on the structures of VP2 and VP3. The N-terminal extension and C-terminal insert unique to *Plasmodium* cysteine proteases were also observed in the models of VP2 and VP3. Although the overall mature proteases have high sequence conservation, there are some significant differences in the substrate binding site residues between the falcipains and the vivapains. Figure 2.22 shows the different residues in the binding sites of FP2, VP2 and FP3, which were used as templates for homology modeling.

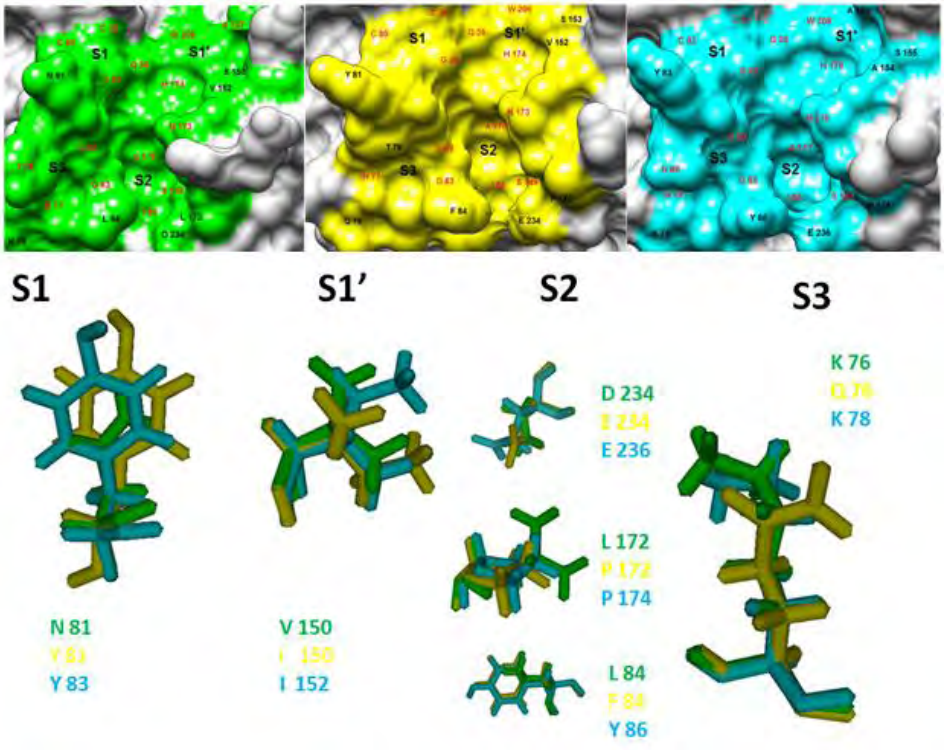

Figure 2. 22: FP2, VP2 and FP3 substrate binding pocket in colored in green, yellow and cyan respectively. S1, S1', S2 and S3 residues which have been mutated are indicated in line presentation.

Table 2. 5: The overall cavity sizes of *Plasmodium* cysteine proteases

Enzyme	Cavity	Area	Volume
FP2	1	123.1	242.1
	2	17.7	16.3
FP3	1	169.8	369.4
	2	10.9	9.2
FP2'	1	115.5	193.
	2	17.3	11.8
VP2	1	128.9	220
	2	13.7	16.2
VP3	1	88.2	192.8
	2	39.2	42.9

The amino acids mutations of the substrate binding sites resulted in the different sized enzyme cavities in the different cysteine proteases, as shown in Table 2.5 for the cavity areas and volumes obtained from the enzyme cavity program CASTp: http: // sts. Bioengr .uic .edu /castp/calculation.php. The VP2 substrate binding pocket is slightly smaller than both FP2 and FP3, based on both the volume and area size. The reduced substrate binding pocket of VP2 can be especially noted at the S2 subsite. Residues in the S1 and S2 subsites are the ones which are different and therefore confer some structural differences and substrate specificity. VP2 subsites amino acid composition is highly similar to FP3 subsite residues than FP2 (as also accounted by a higher sequence identity between the two proteases). The amino acid TYR in the S1 subsite which is known to bind to substrate through hydrophobic interaction is one of the residues similar in both FP3 and VP2, while the same residue is substituted by ASN in FP2. The major amino acid substitution is found in the S2 subsite where L 84 (FP2) is F 84 (VP2) and Y 86 (FP3). This amino acid substitution makes VP2 substrate binding pocket narrower than the rest.

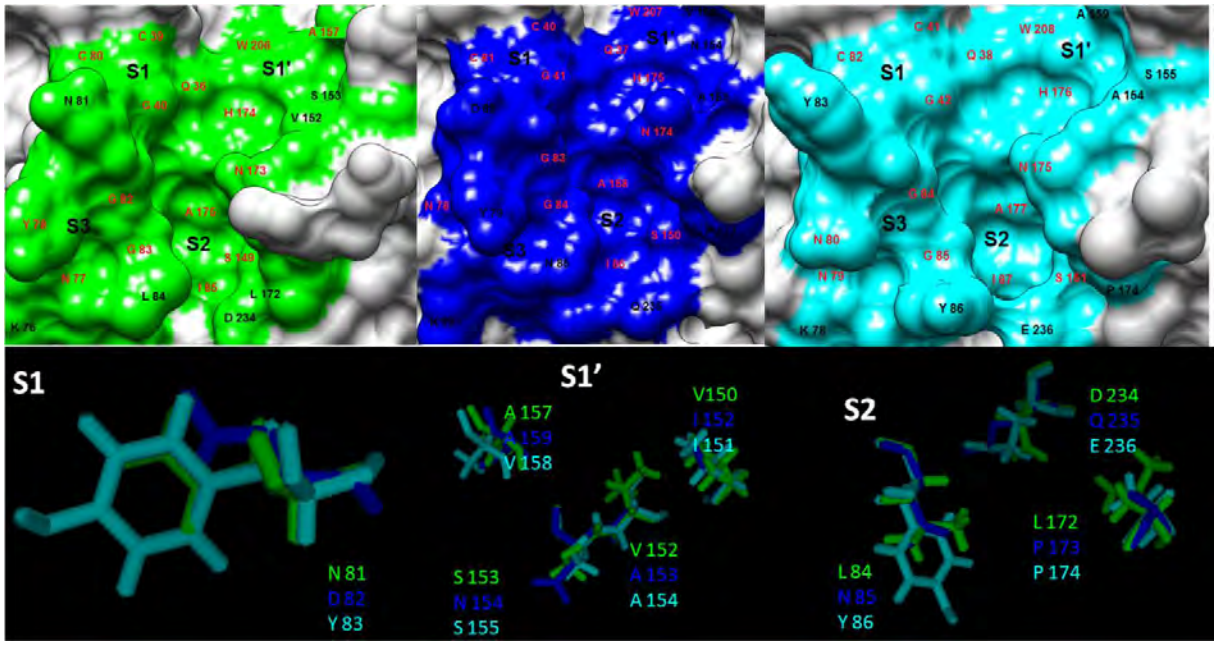

Figure 2. 23: FP2, VP3 and FP3 substrate binding pocket in green, blue and cyan respective.

In the case of VP3, an earlier study conducted by Desai *et al* (2004) has shown that the region between S1' and S2 binding pockets in VP3 are folded more inwards than that of VP2. While the VP2 binding pocket S2 appears to be narrower than that of VP3. However, it appears that the important subsites residues involved in catalytic hydrolysis by the proteases are well conserved and therefore the less conserved residues may account for different reactivities of the proteases. FP2 and FP3 have also been reported to have reacted differently towards different substrates. Other studies have also reported that FP3 hydrolyses hemoglobin twice as rapidly as FP2, and this may be due to the biochemical difference of the residues in the binding pockets. Since VP2 and VP3 have more similar residues in the binding pocket, it is likely that they may also hydrolyze hemoglobin rapidly. Below is a table that summarizes the substrate binding pockets residues in comparisons with each other.

Table 2. 6: Substrate binding pocket residues of the 5 cysteine proteases and the residues highlighted were specified for the ligand binding site (In chapter 3) and residues that are not conserved in all 5 are bolded.

Subsites	Falcipain-2	Falcipain-2'	Falcipain-3	Vivapain-2	Vivapain-3
S1	Q 36, C 39, G 40, C 80 and **N 81**	Q 36, C 39, G 40, C 80 and **N 81**	Q 38, C 41, G42, C 82 and **Y 83**	Q 36, C 39, G 40, C 80 and **Y 81**	Q 37, C 40, G 41, C 81 and **D 82**
S2	**L 84**, I 85, S 149, **L 172**, N 173, A 175 and **D 234**	**L 84,** I 85, S 149, **L 172,** N 173, A 175 and **D 234**	**Y 86**, I 87, S 151, **P 174**, N 175, A 177 and **E 236**	**F 84**, I 85, S 149, **P 172**, N 173 , A 175 and **E 234**	**N 85**, I 86, S 150, **P 173**, N 174, A 176 and **Q 235**
S3	**K 76**, N 77, Y 78 , G82 and G 83	**K 76**, N 77, Y 78, G 82 and G 83	**K 78**, N 79, N 80, G 84 and G 85	**Q 76**, N 77, T 78, G 82 and G 83	**K 77**, N 78, Y 79, G 83 and G 84
S1'	**V 150**, **V 152**, **S 153**, **A 157**, H 174 , N 204 and W 206	**I 150**, **V152**, **S 153**, **P 157**, H 174, N 204 and W 206	**I 152**, **A 154**, **S 155**, **A 159**, H 176, N 206 and W 208	**I 150**, **V 152, S 153**, **A 157**, H 174, N 204 and W 206	**I 151**, **A 153**, **N 154**, **V 158**, H 175, N 205 and W 207

Homology models often have some disadvantages that experimentally determined structures do not have. The accuracy of homology models depends on the following factors: (i) the sequence identity between the target and template usually indicated by RMSD; (ii) alignment mistakes which may lead to errors in the packing of side-chains, core protein being distorted and loop modeling (Sanchez and Šali, 1997; Baker and Šali, 2001) and (ii) using an incorrect template or errors in the template structures. Different model assessment programs use different approaches, therefore it is more advisable to use more than one program as this increases the confidence in the model obtained. The models generated in the present study were also subjected to a few model assessment programs in order to determine the accuracy at which they were predicted. For all 3 cysteine proteases: FP2', VP2 and VP3, all the model assessment programs employed in the study showed insignificant errors and also supported the fact that the models were predicted with high accuracy, as they were within the range of high-quality structures. The models built in the study were analyzed by PROCHECK, ProSA and MetaMQAP II. PROCHECK analysis of a model provides the user with an idea of the overall stereochemical quality of all the chains in the input file. In addition to analyzing the whole structure of the protein, PROCHECK also highlights regions of the protein which may appear to have unusual geometry (Laskowski *et al*., 1993). PROCHECK outputs various plots which indicate that it analyzes the input protein structure in a lot of detail. The outputs of PROCHECK include plots such as the main Ramachandran plot, all-residue Ramachandran plot, all-residue Chi1-Chi2 plots, main-chain parameters, side-chain parameters, residue properties plot, main-chain bond lengths, main-chain bond angles RMS distance from planarity and distorted geometry (Laskowski *et al*., 1996). PROCHECK is the most popularly used program for protein structures and model assessment with over one hundred citations in pubmed database (http://www.ncbi.nlm.nih.gov/pmc/), and over thirty publications between 2010-2011, which mean it is still being used currently. ProSA also confirmed that the overall folds of the model generated were correct. Ideally, ProSA uses knowledge-based mean fields to analyze the energy distribution of atoms during protein folding. It uses the energies encountered for individual sequence and transforms them into z-scores (Sippl, 1993). The favourable conformations are therefore given a negative z-score and the least favourable a positive z-score. The models built

in our study all had negative z-scores and were perfectly fitted with the crystal structures already in PDB, which is consistent with favourable conformation. The energy plots generated by ProSA show the steric violations of the model built, in which case, the models violating the basic steric requirements generally have high energy peaks. For FP2', VP2, VP3 and human procathepsin K were consistent with their template structures and corresponded well with the basic interactions that those atoms in a native protein must make. Based on the results obtained from model validation check, we assumed that the models built complement experimentally determined structures. The models built were used with a high degree of confidence as they compare well with high-resolution structures which have been elucidated by experimental techniques. The differences will be examined in the next chapter where we look at the weakly conserved arm-like motif (C-terminal insert.) binding to hemoglobin. We also look at the active site binding to hemoglobin and possibly see the effect of the different substrate binding pockets.

Chapter 3

3 Protein-protein docking between *P. falciparum* and *P. vivax* cysteine proteases and human hemoglobin

This chapter describes an unbound docking study that was carried out between papain-family cysteine proteases of *P. falciparum* (falcipains) and *P. vivax* (vivapains) and their natural substrate human hemoglobin. Substrate-protease complex structures were predicted by docking experiments which was specifically targeting the protease active sites and arm-motifs for binding. Best predictions were filtered from incorrect ones using a scoring function and selected docked structures were refined by energy minimization. The total energy and interaction energies of the best prediction were calculated prior and after energy minimization. Mode of interactions of the complex structures and forces mediating the interactions were also identified. The docking was validated by reproducing the co-crystal structure of FP2-cystatin protein complex (PDB code: 1VYB).

3.1. Introduction

Proteins are fundamental components of all living organisms. They carry out the biological activities of most molecules. Often times, proteins do not single-handedly carry out their function, they require one or more interaction partner (Matthew *et al*., 2007). Examples of protein-protein complexes include enzyme-inhibitors, antibody-antigen and hormone-hormone receptors. Protein-protein complexes are a vital component of molecular biology, they yield

insightful knowledge about the functions of component proteins and this may guide the design of novel molecules regulating the protein interaction network (Vakser *et al*., 1999). Many diseases can be traced to undesirable or malfunctioning protein-protein interactions; this signifies the study of such interactions (Singh *et al*., 2006). Several techniques have been implemented and they are employed in the experimental determination of protein-protein interaction. Proteomic techniques such as mass spectrometry, genome-scale yeast 2-hybrid and display cloning are being used to solve many protein-protein complexes (Uetz *et al*., 2000; Ito *et al*., 2000; Weng and DeLisi, 2002). Despite the success in uncovering experimental techniques aimed at solving the structures of protein-protein complexes, there are still some challenges in this field. For instance, the experimental techniques that are employed for protein-protein complex determination have progressed slowly. And therefore computational approaches in which the protein-protein complex structures are predicted are becoming more significant and popular.

The *in silico* prediction of protein-protein complex structures using co-ordinates of individual structures is called protein-protein docking. This computational prediction of protein-protein complexes are divided into two, depending on the algorithms used: bound docking and unbound docking (as indicated below in Figure 3.1). Bound docking is when a complex is pulled apart and re-assembled into individual proteins. This is a fairly easy technique and its success has been complimented by the excellent results obtained (Norel *et al*., 1995; Fischer *et al*., 1995; Meyer *et al*., 1996; Ackerman *et al*., 1998). Unbound docking is when experimentally determined or computationally modelled protein structures are used to generate a protein-protein complex structure (Chen and Weng, 2002). This technique is more challenging and more difficult than bound docking due to the fact that upon complex structure formation proteins undergo conformational changes especially at their side chains (Betts and Sternberg, 1999). The complexity of this unbound docking makes it an interesting case of study and also presents scientists with opportunities to refine the technique. The problem of unbound docking drew a lot of attention in the structural biology area and as such several algorithms were developed to address it.

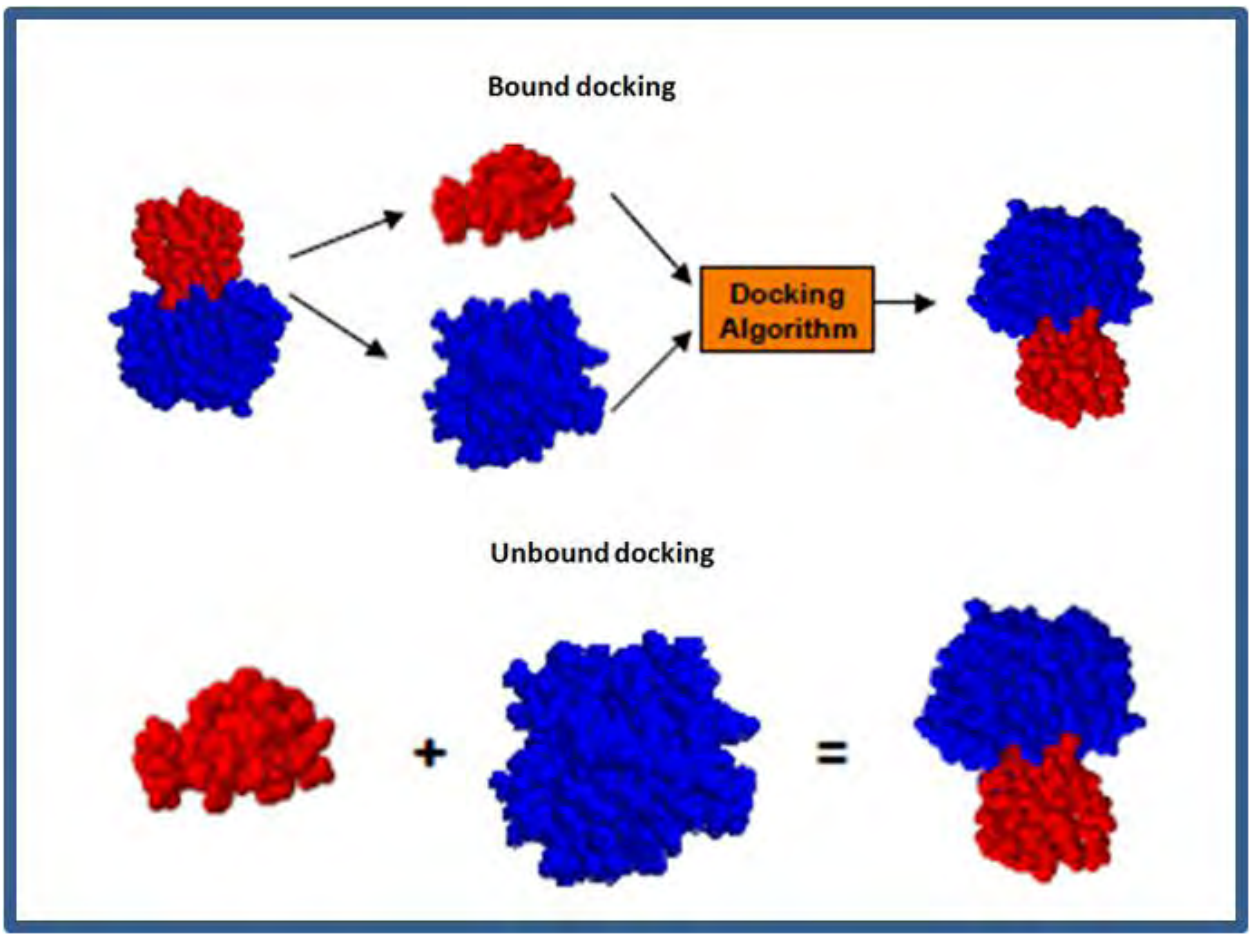

Figure 3. 1: Computational determination of protein complex structures, indicating bound docking approach and the unbound docking

Two main algorithms have been developed and are mainly used for unbound docking: flexible body docking and rigid body docking. The former takes advantage of the biological concept which states that during complex formation or interaction with one another, protein side chains tend to change in conformation. Therefore, flexible docking allows conformational changes. The main set-back about this approach however, is that it is a computational cost and occurs over several hours. Rigid body docking treats the input proteins as rigid bodies; this approach assumes that the component of the proteins (bond angles, bond lengths and torsion angle) are not modified at any stage during complex formation. Rigid body docking has a high success rate for the initial stage of unbound docking. Most of the protein-protein docking programs are based on the rigid body docking approach including ZDOCK (Chen and Weng, 2002; Chen *et al*.,

2003a), AUTODOCK (Morris *et al*., 1995), ROSETTA (Lyskov *et al*., 2008) and many other programs.

Most of the protein-protein docking programs use four steps to successfully carry out their docking predictions: representation of the system, conformational space search, scoring and ranking potential solutions and refinement of accepted solutions. The principle behind representation of a system is that protein interactions are transmitted by amino acids at their surface. And the protein surface is simply the atomic representation of exposed residues (Francis-Lyon *et al*., 2010). Most docking programs describe the protein surface by mathematical models which offer sparse distribution of surface points while simultaneously storing as much information as possible. Protein surfaces are usually described by Fourier Transform and their geometric features. The conformational space search step explores all the possible orientations of the two individual proteins. The 3D structure of a protein complex shows a close geometric and chemical complimentarity between two parts of the molecular surfaces in contact. All docking programs search the conformational space for structures which reveal a high correlation complimentary to adjacent surfaces. In most cases the conformational spaces of the two protein structures is searched by using two algorithms: Fast Fourier Transform (FFT) docking and Geometric harshing (Kaapro and Ojanen, 2002). For the latter, the algorithm scans groups of surface dots (or atoms) and detects optimally matched surfaces. In the FFT docking, the conformation space is searched for conformations where 3D grids representing proteins overlap. Programs such as CKORDO, MOLFIT, DOT, ZDOCK, BDOCK, GRAMM and FTDock use the FFT docking algorithm in their conformational space search.

The scoring and ranking of potential solutions step uses a set of generated complex conformations. From those, the complex structures showing the highest similarity to native complex conformation are selected. Scoring functions are usually based on physcio-chemical complementarities and correlations (Lee, 2008). Therefore, a numerical value is assigned to each of the proposed conformations according to their Electrostatics, hydrogen bonds and hydrophobic interactions. The free energy obtained from the interactions of proposed complex structures are used as a reliable discrimination of native conformations from non-native ones.

The last step in docking involves refinement of accepted solutions. In this step the understanding that proteins are not rigid bodies and they are likely to undergo conformational changes when associating with other proteins is accounted for. It has been observed that the changes in proteins usually occur at the side chains level and in rare cases, movement of flexible loop regions. The refinement algorithms allows for fine re-ranking of near-native structures from initial stage docking. Popular methods for the refinement of accepted solutions are molecular simulations, energy minimization and the use of rotamer libraries. Below, we discuss the docking as carried out by ZDOCK algorithm, which was used for the prediction of complex structures in our study. All the calculations and refinements were carried out in Discovery Studio 2.5 [DS 2.5: http://www.accelrys.com/dstudio (Acceryls, California)]

3.2 ZDOCK

ZDOCK algorithm (Chen *et al*., 2003a) provides a rigid body docking of two protein structures. It clusters the predicted protein poses based on their ligand positions and allows for the user to filter protein poses by specifying residues at the binding interface or blocking residues not involved in binding. Docked protein poses are rescored and re-ranked by a ZRANK scoring function (Pierce and Weng, 2007).

Initial stage unbound protein-protein docking using the ZDOCK algorithm (Chen and Weng, 2002) was the program used to make all the protein-protein complex structure predictions in this study. Protein-protein complexes of *P. falciparum* and *P. vivax* papain-family cysteine proteases and hemoglobin were generated by ZDOCK. This algorithm integrates Pairwise Shape Complimentarity (PSC) with Desolvation (DE) and Electrostatic (ELEC) energy terms to create a powerful and best performing scoring function (Chen *et al*., 2003a). Once predictions are made by ZDOCK, a scoring function needs to be used to discriminate correct predictions from incorrect ones. Therefore, the development of a scoring function is of critical importance in all protein-protein docking protocols, shape complimentarity is the most basic ingredient and a very fundamental tool geared towards this purpose. Protein surfaces at the binding interface

are generally complimentary to each other, a geometric descriptor arising from this observation is referred to as shape complimentarity (Chen and Weng, 2003). Most docking algorithms use grid-based shape complimentarity (GSC) as their scoring function (Chen and Weng, 2003). GSC approach does not provide any clear information regarding the surface curvature; instead it uses a surface descriptor to compute grid points on the receptor protein and ligand protein. The GSC score is therefore calculated by identifying a total layer of grid points surrounding and not overlapping the receptor and the total number of grid points in the layer corresponding any ligand grid points, minus the clash penalty (Chen *et al*., 2003a). The scoring function for ZDOCK is not GSC; it is a rather abstractly simple and easy to compute Pairwise Shape Complimentarity (PSC). Unlike GSC, surface areas or surface curvatures are not the main basis of this particular scoring function. PSC uses a specific distance cut-off to reward all close contacts between the ligand and receptor protein, minus a clash penalty (Chen and Weng, 2003). There are two term constituting PSC: the favourable and penalty terms. The latter counts the total number of interface atom pairs between the receptor and ligand proteins at a specific distance cut-off. The latter is linearly proportional to the number of grid points overlapping the receptor and ligand proteins. PSC has also been shown to rank more near-native protein complex structures than GSC (Chen and Weng, 2003). ZDOCK uses FFT (Katchalski-Katzir *et al*., 1992) to better calculate the PSC of receptor and ligand proteins. It uses two complex functions receptor PSC (R_{psc}) and ligand PSC (L_{psc}) to map the geometric characteristics of the receptor and ligand protein on the grids. FFT has been shown to be successful, making predictions with a Root Mean Square Deviation (RMSD) of 1.1. Å to the native structures (Gabb *et al*., 1997). It performs a complete translational and rotational search in Fourier space, and then selects binding geometries with high surface correlations as the best predictions (Katchalski-Katzir *et al*., 1992). FFT offers another advantage in that it is computationally fast and mathematically elegant (Figure for FFT docking is indicated below).

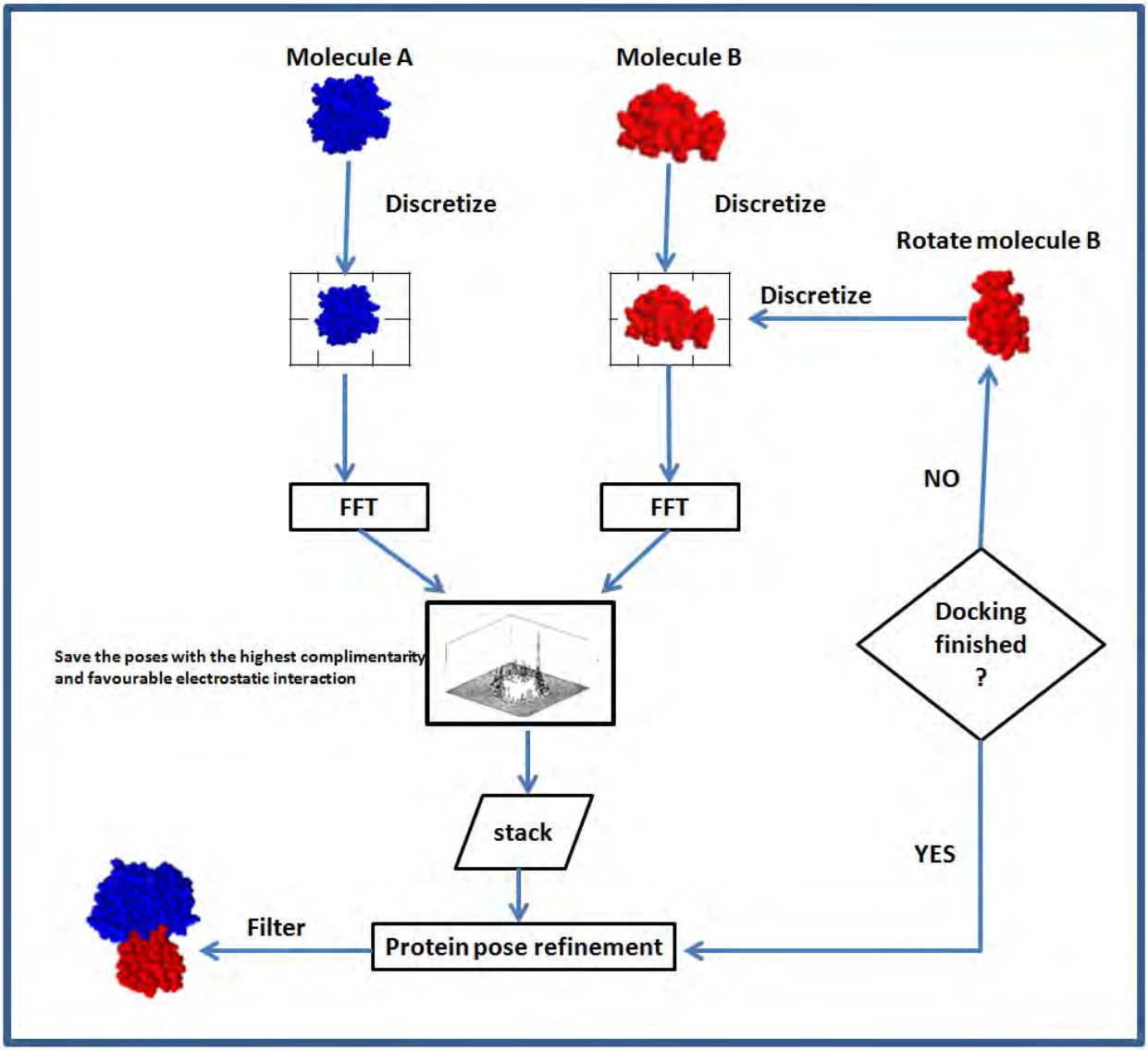

Figure 3. 2: Typical docking experiment implementing the Fourier Transform (Katchalski-Katzir *et al*., 1992). Figure adapted from Gabb *et al* (1997) page 1

Chen and Weng (2002) demonstrated that the inclusion of DE term improves the performance of docking significantly. ZDOCK uses the Atom Contact Energy (ACE) to estimate the desolvation term (Zhang *et al*., 1997). The free energy necessary for replacing two proteins atom-water contact with corresponding protein atom-protein atom and water-water contact, is termed as ACE. Non-pairwise ACE, contact of a protein atom of a specific ACE type with a protein atom of a non-specific therefore averaged type, speeds up the calculations (Chen *et al*., 2003b).

Coulombic formula is expressed as the correlation between partial charges of the ligand atom and the electrostatic potential created by the receptor atoms (Gabb *et al*., 1997). This coulombic formula is the basis in which ZDOCK uses to account for electrostatic contribution to the docking scores (Chen *et al*., 2003b). Pierce and Weng (2007) developed a more detailed scoring function called ZRANK, which significantly improve the success rate of ZDOCK predictions. ZRANK uses more detailed electrostatics, van der Waals and desolvation to rescore predictions initially made by ZDOCK. ZRANK can be used as a refinement stage on its own or as a preprocessing stage well ranked poses prior further refinements (Pierce and Weng, 2007).

3.3. Methods

The entire protein-protein docking experiment followed in this study has been summarized in the flowchart diagram below (Figure 3.3.). Three major protocols were followed: data retrieval, protein docking and protein simulations in order to predict favourable protein poses.

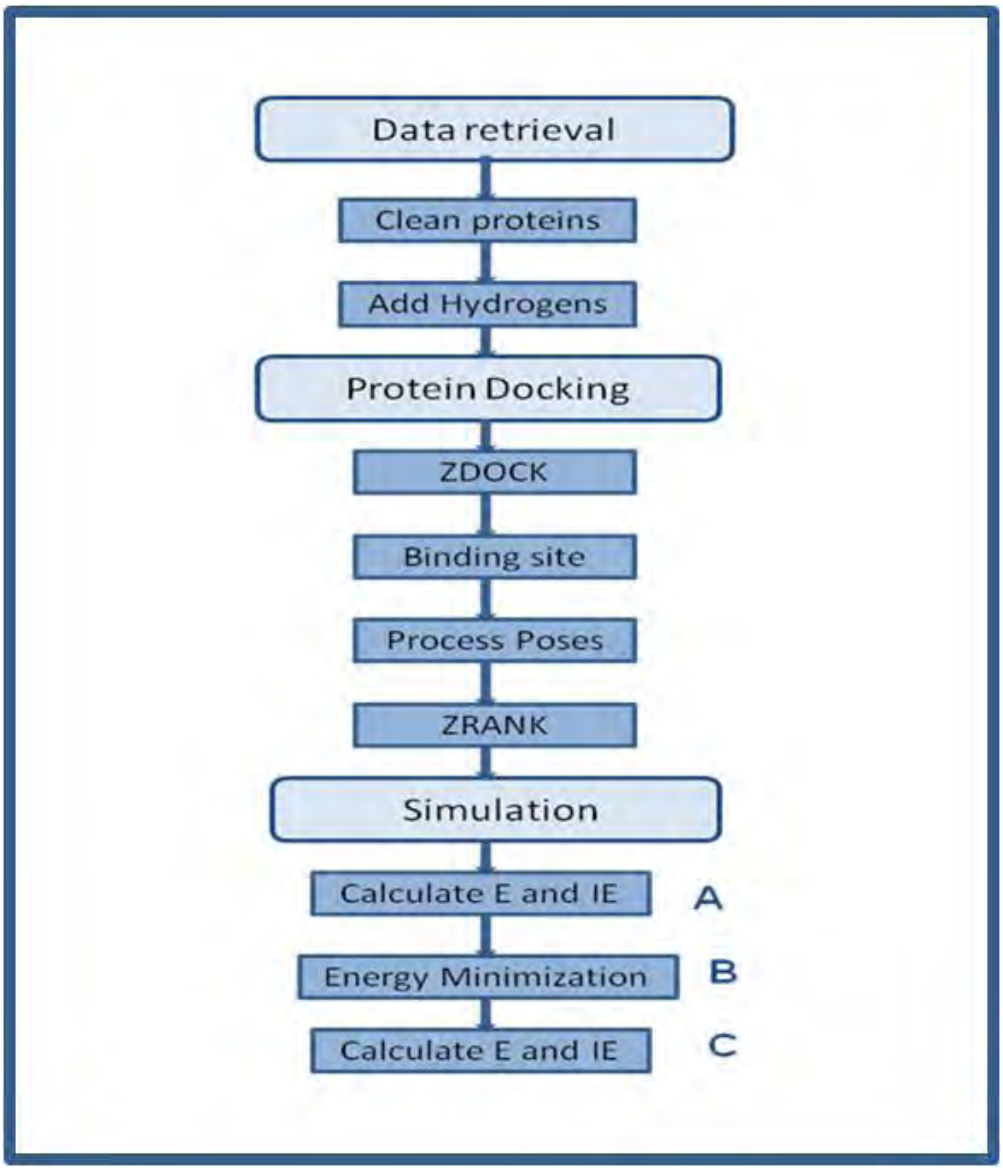

Figure 3. 3: Flowchart diagram indicating the steps followed when predictions of the protein complex structures of *P. falciparum* and *P. vivax* cysteine proteases and human hemoglobin. The main process data retrieval, protein docking and simulation are in light blue and sub processes within each protocol are in dark blue

3.3.1. Data retrieval

Human hemoglobin (PDB code: 1BZ0) was used as an input receptor for enzyme-substrates protein complex predictions made in this study. The 3D structure of hemoglobin was retrieved from PDB. Currently there are 476 structural files of the hemoglobin molecules in PDB, most of which was solved by X-ray crystallography. The hemoglobin structure chosen to generated predictions in our experiment was solved more than a decade ago by X-ray crystallography, resolution 1.50 Å (Kavanaugh *et al*., 1993). Also, this particular hemoglobin structure was the one used in a parallel study where the model of FP2-hemoglobin complex structure (Wang *et al*., 2006) was generated, so it became the obvious choice for accurately comparing our results with already published work. For the ligand proteins: cysteine proteases of *P. falciparum* and *P. vivax* were used. Only FP2 and FP3 have been experimentally elucidated and FP2', VP2 and VP3 were derived by homology modeling (**chapter 2**)

PDB contains six structural entries for *P. falciparum* cysteine proteases, four FP2 and two FP3. FP2 structures, PDB codes: 1YVB and 2OUL were both solved by X-ray crystallography at resolutions of 2.70 Å and 2.20 Å respectively. Both the structural files of FP2 were used for the protein-protein docking study, in the case of 1YVB to accurately compare our results to published work. Also, 2OUL was included in order to determine the consistency of the results predicted and make comparisons between the residues involved in binding for the two FP2. FP3 structures were also solved by X-ray crystallography at resolutions 2.42 Å and 2.50 Å for the structural files with PDB codes 3BWK and 3BPM respectively. FP3 (PDB code: 3BWK) was used for the docking experiment as it was the latest entry (released date: 2009) and was solved at a slightly better resolution compared the other two FP3 protein structures. For the docking validation, the complex structure of FP2-cystatin was disassembled to two individual protein molecules. And each of the cysteine protease (FP2) and cystatin (inhibitor) were once again used as inputs for the prediction of the co-crystal structure.

3.3.2. Proteins preparations

Before any docking experiments, the input structures (whether experimentally solved or generated by homology modeling) must be cleaned. Both the ligand and receptor proteins were prepared for docking using the clean protein function and add hydrogen atoms in Discovery Studio 2.5 [DS 2.5: http://www.accelrys.com/dstudio (Acceryls, California)].

- **Clean proteins**

Structures retrieved from PDB and those computationally determined were cleaned in DS 2.5. In each case Hemoglobin was treated as the receptor, falcipains and vivapains as ligand, most docking algorithm have been trained to treat the larger protein as receptor (Chen and Weng, 2002).

- **Add hydrogen**

Hydrogen atoms are absent from most protein structure files as a resolution of 1.0 Å or less is needed to determine the exact positions of hydrogen atoms. Hydrogen atoms were added to the input structures by typing the proteins with a CHARMm forcefield (Brooks *et al*., 1983). The Add Hydrogens function predicts and adds missing hydrogens to the selected protein molecule. This ensures that the correct electrostatic potential is applied, and this helps facilitate protein-protein docking experiments with minimal errors. Once the hydrogen atoms were added to the input proteins using CHARMm forcefield, initial stage unbound docking studies were performed using the ZDOCK algorithm (Chen and Weng, 2003).

3.3.3. Protein-protein docking

- **ZDOCK**

Default parameters were used for the initial stage unbound docking predictions of the complex structure of enzymes with substrate and in the case of FP2-cystatin, enzyme inhibitor. The angular step size (sampling size) was set to 6°. None of the ligand and receptor residues were blocked. No protein poses were filtered because no residues were specified for receptor or ligand binding site. Therefore a blind ZDOC-K protocol was ran, with ZRANK set to true and parallel processing also set to true.

- **Process protein poses**

The process protein poses (ZDOCK) protocol allows the user to select a subset from a set of docked protein poses generated from ZDOCK. The selection can be made according to pose rank or specifying residues at the binding interface. Process protein poses also re-rank protein poses with a ZRANK scoring function (Pierce and Weng, 2007). The selection of residues at the binding interface can be done in two ways: either block residues not involved in binding or specify residues involved in binding. Two set of protein-protein complex structures were generated: ones in which the proteases C-terminal insert was bound to hemoglobin and ones in which the active site was bound to hemoglobin. For the latter residues were specified as the ligand binding site. And for the former, with FP2 both 2OUL and 1YVB residues from both hemoglobin and FP2 were specified as there was published work about the residues involved in binding, therefore LYS 196 and SER 223 were specified for ligand (FP2) binding site and ASP 85 (chain C) and LYS 11 (chain A) for receptor binding site. These residues were obtained from studies already published by Wang *et al*., (2006).

All protein poses obtained from ZDOCK were used as input docked protein poses for process protein poses protocol, and the arm-bound residues were filtered by the residues indicated above and the active site were filtered in the same manner. The ZRANK score was set to true for process protein poses and protein poses were also set to be clustered based on ligand position and distance.

3.3.4. Protein simulation

The primary requirement for all simulation protocols is that the simulated system may be typed with an appropriate forcefield. CHARMm was the forcefield of choice in all the protein simulation protocols because it has been designed to give good results in molecular mechanics and molecular dynamics (Brooks *et al*., 1983). CHARMm achieves this by using a complete empirical energy function to calculate the geometries, interactions and conformation energies of small molecules, solvated complexes and modelled systems (Momany and Rone, 1992). Local minima, barriers to rotation and free energies are also calculated by this particular forcefield. This protocol is used to calculate the energy, interaction energy and protein complex structure minimization.

- **Calculate Energy**

The potential energies of the input complex structures were evaluated by the calculate energy protocol. This protocol basically confirms that a given structure or system does not exhibit serious distortions from typical equilibrium molecular geometries or posses substantial atomic overlap. The calculate energy protocol was ran using default parameters The protein complexes obtained from process protein poses and the minimized protein complex structures were the input molecules for calculate energy protocol. The protocol for calculate energy uses CHARMm

forcefield to perform the calculations on the input structures, Non-bond list radius set at 10 Å, estimate entropy set to false and electrostatics were at a spherical cutoff.

- **Calculate interaction energy**

The calculate interaction energy protocol was used to calculate the non-bonded interactions such as Van Der Waals and electrostatic energy of the input structures (complex structures before and after minimizations). Prior running the calculate interaction energy protocol, residues involved in binding from both the receptor proteins and the ligand protein were defined. Following this the interaction energies of the residues within the complex structure (the input) file were calculated using the algorithm's default parameter, aswell as the providence of residues involved in binding. None dielectric model and non-bond list radius set to 14.00 Å

- **Energy Minimization**

The refinement step is critical to every protein-protein docking experiment. The protein complexes obtained from the process poses protocol (under **section 3.3.3.** in protein-protein docking), showed that the predicted poses were unstable. This was deduced from their high total and interaction energy (positive values in Kcal/mol); therefore these poses had to be refined. Energy minimization protocol was used as a refinement step to the protein-protein docking experiment. It was ran using the smart minimize algorithm, with an RMS gradient of 0.1 Å and the maximum cycle of minimization were set to 2000. No implicit solvent model, non-bond list radius set at 14.0 Å and spherical cut-off for the electrostatics.

3.3.5. Interacting Residues

Protein Interaction Calculator (PIC) found on: http://crick.mbu.iisc.ernet.in/~PIC/ was used to identify residues involved in binding and the nature of forces driving the interactions during complex structure formation. PIC uses standard published criteria to calculate various interactions involved in protein structure prediction and/or protein assembly stabilization. This program considers and calculates the strength of disulphide bonds, hydrophobic interactions, ionic interactions, hydrogen bonds and other similar interactions (Tina *et al*., 2007). Forces contributing to complex formation were identified using the PIC server. The hydrophobic interactions, hydrogen bonds and charge-charge interactions of the protein complex structures between proteases and substrate were calculated. Default distance in Å was used for the hydrogen bonds and 10 Å for hydrophobic interactions and charge-charge interactions were used in the determination of residues involved in binding.

3.4. Results and discussion

3.4.1. Protein-protein docking

Table 3.1: Initial stage unbound docking scores of the complex structures:

DOCKED POSE NO:	COMPLEX STRUCTURE	ZDOCK	ZRANK
Pose 52	$FP2_{arm}$-hemoglobin (1YVB)(Pub)	11.70	32.22
Pose 3	$FP2_{arm}$-hemoglobin (1YVB)	17.80	-78.42
Pose 1	$FP2_{activesite}$- hemoglobin (1YVB)	23.54	-106.77
Pose 26	$FP2_{arm}$-hemoglobin (2OUL)(Pub)	13.32	108.22
Pose 1	$FP2_{arm}$-hemoglobin (2OUL)	16.92	-69.02
Pose 19	$FP2_{activesite}$- hemoglobin (2OUL)	23.10	-93.88
Pose 43	$FP2'_{arm}$-hemoglobin (Pub)	17.00	-34.46
Pose 31	$FP2'_{arm}$-hemoglobin	17.04	-40.84
Pose 13	$FP2'_{activesite}$- hemoglobin	22.24	-82.04
Pose 2	$FP3_{arm}$-hemoglobin	19.28	-85.07
Pose 9	$FP3_{activesite}$- hemoglobin	19.94	-86.32
Pose 1	$VP2_{arm}$-hemoglobin	18.92	-41.56
Pose 1	$VP2_{activesite}$- hemoglobin	21.18	-87.36
Pose 23	$VP3_{arm}$-hemoglobin	16.42	-14.90
Pose 2	$VP3_{activesite}$- hemoglobin	23.34	-104.29
Pose 1	FP2_cystatin	19.00	-115.71

The protein-protein docking experiment followed in our study incorporated a two stage docking approach for the prediction of the protease-substrate (experimental) and protease-inhibitor (validation) complex structures. The first stage was an unbound docking experiment in which the rigid body algorithm ZDOCK was used to generate protein-protein complex structures. In the second approach a more flexible energy minimization smart minimize algorithm was used for complex structure re-ranking and refinement stage. The unbound docking approach generated a total of 38 000 protein poses (hits) for the entire study as ZDOCK outputs 2000 protein poses per job (good and bad predictions) and 19 protein –protein complex structures were generated. Once ZDOCK has made its predictions, the 2000 protein poses must be re-filtered by a process protein poses protocol which discriminates false positives from near

accurate protein pose prediction. The main advantage of the process pose protocol is not only the fact that it can be used as refinement stage but also that it allows the user to make a biased search. If any experimental evidence data for binding in specific area is available or computational tools have been used to determine the residues involved in binding they can therefore be selected. In Table 3.2, a summary of the results yielded through the filtering and yet refinement step: process poses protocol which determines the level of pose accuracy by their ZRANK score is indicated.

Table 3.2: Process poses protocol ZRANK scores of the complex structures obtained

Complex	LOWEST ZRANK SCORE	HIGHEST ZRANK SCORE	Protein poses
$FP2_{arm}$-hemoglobin (1YVB)(Pub)	32.22	108.72	26
$FP2_{arm}$-hemoglobin (1YVB)	-81. 23	63.67	268
$FP2_{activesite}$- hemoglobin (1YVB)	-106.77	180.95	595
$FP2_{arm}$-hemoglobin (2OUL)(Pub)	47.03	171.56	20
$FP2_{arm}$-hemoglobin (2OUL)	-80.18	106.14	296
$FP2_{activesite}$- hemoglobin (2OUL)	-98.64	200.96	610
$FP2'_{arm}$-hemoglobin	-87. 55	182.07	309
$FP2'_{activesite}$- hemoglobin	-92.24	152.77	600
$FP3_{arm}$-hemoglobin	-89.59	173.90	295
$FP3_{activesite}$- hemoglobin	-92.24	152.77	716
$VP2_{arm}$-hemoglobin	-87.36	104.96	70
$VP2_{activesite}$- hemoglobin	-88.00	142.90	190
$VP3_{arm}$-hemoglobin	-85.55	120.00	217
$VP3_{activesite}$- hemoglobin	-98.20	144.75	543
FP2_cystatin	-115.71	1.24	1560

All 2000 protein poses generated from ZDOCK were used as input for the process poses protocol and a summary of the results obtained therein is shown in Table 3.3. The table above shows the ZRANK scores of every protein-protein complex structure prediction made in the study. Based on the program's ranking system negative ZRANK scores shows near-native predictions and positive ZRANK scores are likely for false positives.

The main interest of the study was to explore cysteine proteases binding to hemoglobin at the C-terminal insert (arm-like motif) and active site, therefore the process poses protocol was biased. In case of the former, residues within the C-terminal insert which will henceforth be referred to as the arm-like motif for convenience, were selected for ligand binding site and none from hemoglobin were selected. In terms of the active site, although most protein poses from ZDOCK were bound to the active site residues listed in Table 3.1 were selected for ligand binding site. The output from process pose protocol was a more reduced number of protein poses than those from ZDOCK. In general the active site bound complex structures had good ZRANK scores and a larger number of protein poses than the rest. This raises the likelihood that the predictions maybe correctly made. Also, the arm-like motif proposed mode of hemoglobin binding seem to have favourable ZRANK scores whereas the protein poses resembling published data do not, therefore they are likely to be false positives. The results obtained using FP2 structures (1YVB and 2OUL) are consistent with each other. For FP2 co-crystals, two 3D structures (PDB codes: 1YVB and 2OUL) and hemoglobin were used as ligand and receptor proteins respectively. FP2 (1VYB) was used as the first one because this structure was not merely used for complex structure generation but also to compare with the already published data from Wang *et al* (2006). The same was done for FP2 (2OUL)-hemoglobin complex structures, whereby three set of protein complex structures were generated, $FP2_{arm}$-hemoglobin resembling published data, the proposed $FP2_{arm}$-hemoglobin and $FP2_{activesite}$-hemoglobin complex structures were generated (indicated in Table 3.3). When generating complex structures in which the arm-like motif of FP2 (1YVB) was bound to hemoglobin, 268 protein poses were filtered from ZDOCK protocol input protein poses. The ZRANK scores of the 268 protein poses were from -81.23 to 63.67 and 142 of these were negative. These protein poses were grouped into 32 clusters. There were 595 protein pose outputs from process poses still using the 2000 blind ZDOCK predictions as input and by specifying residues listed in Table 3.2 for active site bound structures. The ZRANK scores of active site bound protein poses were ranging from -106.77 to 180.95 and 382 of these protein poses were energetically favourable (negative ZRANK scores). From these predictions, three sets of protein poses were analyzed, a

cluster representing most energetically favourable ZRANK scores for the active site bound, C-terminal (arm-like motif) insert, and the pose resembling published data for FP2 (1YVB) hemoglobin complex structures. Once again using FP2 (2OUL) as the ligand protein and hemoglobin as the receptor protein, ZDOCK generated 2000 protein poses with ZRANK scores between from -98.97 to 157.14, and 1174 poses of which had negative ZRANK scores for the complexes of FP2 (2OUL) and hemoglobin. Protein poses in which the arm-like motif was bound to hemoglobin were 296 grouped into 40 clusters. The scoring function of protein poses were in the range of -80.18 and 106.14 for the ZRANK score of all protein poses. Half of the protein poses of the $FP2_{arm}$-hemoglobin had a negative ZRANK score, which may suggest that these complexes may be physically improbable or unlikely to be stable. For the active site bound to hemoglobin complex structures, 610 protein poses were generated by the process poses protocol. The protein poses were grouped into 41 clusters and had ZRANK scores between -98.64 to 200.96. In the case of both FP2 proteases, two sets of protein complexes of the FP2arm bound to hemoglobin were generated. Ones resembling published work (Wang *et al*., 2006) and those which based on their ZDOCK and ZRANK score are energetically favourable and could infer correct conformation of the native complex structure. Therefore, Pose 52 and Pose 26 shows the scores of poses resembling published data, while Pose 3 and Pose 1(Table 3.2) for 1YVB and 20UL shows preferable binding based on the scoring functions. ZDOCK scores are based on shape complimentarity and are positive. Therefore, the higher the ZDOCK score the more complimentary is the interaction between the proteins of interest. Thus, based on the results tabulated above, the active site and arm bound protein poses shows better complimentarity than poses resembling published work. Even their ZRANK scores are of the active site and arm bound complexes are lower than that of poses resembling published work which might otherwise have been filtered out as faulty poses, however further refinements were carried out on them. In a similar manner as FP2, docking studies were carried out on FP2' whereby three sets of protein poses were generated. Protein poses of FP2' binding in a manner resembling the pose published, where the arm motif was bound at better complimentarity and the active site bound to hemoglobin. The results obtained for FP2' are not significantly different from that of FP2. ZDOCK protocol for FP2' and hemoglobin yielded 1092 poses with good shape

complimentary and a range of -98.42 and 191.46 for ZRANK scores of all poses. FP2'$_{arm}$ – hemoglobin complex structures which had ZRANK scores between -87. 55 to 182.07, grouped into 42 clusters, a total of 309 protein poses. The protein poses of FP2'active site bound to hemoglobin were 600 and they were grouped into 51 clusters. Their ZRANK score was between -92.24 to 152.77. The three groups of protein poses were subjected to further analysis through energy calculations, interaction energy calculations and energy minimizations. For FP3, VP2 and VP3 poses resembling published data were not found in the 2000 docked structures obtained from ZDOCK. Even filtering poses in the process poses protocol employing a biased search where poses of the C-terminal insert and hemoglobin were specified, none of the protein poses slightly resembled FP2$_{arm}$-hemoglobin which was published. Therefore for these three proteases, only two protein poses were analyzed: arm bound and active site bound. For FP3-hemoglobin protein poses, 295 and 716 docked structures were generated for the arm-motif and active site bound complexes respectively. The FP3$_{arm}$-hemoglobin protein poses were grouped into 36 clusters, with their ZRANK scores in the range of -89.59 to 173.90. Poses that were energetically favourable (negative ZRANK) score were 105. FP3 $_{active\ site}$ –hemoglobin protein poses were grouped into 100 clusters with their ZRANK scores between -94.22 to 185.09. Less than 50% of the total protein poses generated from process poses of FP3 active site bound to hemoglobin were energetically favoured (350 protein poses). Process poses filtering stage generated 70 and 190 protein poses for VP2$_{arm}$-hemoglobin complex structure and VP2$_{active\ site}$-hemoglobin complex structures respectively. For VP2$_{arm}$-hemoglobin complex structure, protein poses had ZRANK scores between -87.36 and 104.96 and they were grouped into 7 clusters. Cluster 4 contained protein poses with more near-native structures than the rest which were bound to the active site, and was therefore selected for further refinements. As per Table 3.2, pose 7 was the most stable of 10 other poses in this cluster (Table A5). FP2 $_{active\ site}$-hemoglobin protein poses were grouped into 23 clusters and ZRANK scores of -88.00 to 142.90. The first pose (most stable) did not belong to any cluster; therefore it was not bound in the same manner as any of the 190 protein poses, which shows that it may be a false positive. Therefore, pose 5 belonging to cluster 3 together with the rest of 10 protein poses in this cluster were evaluated by energy minimization below. Clearly, Pose 5 is the most stable of all

the poses in this cluster (lowest ZRANK score), however ZDOCK analysis of the other poses are shown in Table B5: complex 11-22. They were 1140 protein poses with favourable ZRANK scores for VP3-hemoglobin complex structures. The score of the rest of the 2000 post were ranging from most stable to least stable: -102.27 to 174. 11. Using the Biased search indicated in Table 3.1, VP3 $_{arm}$-hemoglobin and VP3 $_{active\ site}$-hemoglobin were filtered from the rest by process poses protocol. Complex structure in which hemoglobin was bound to the arm motif filtered leaving 217 poses with ZRANK scores of -85.55 to 120.00 and 150 poses with favourable ZRANK. For FP3active site bound to hemoglobin, 543 protein poses were filtered with ZRANK scores between -98.20 to 144.75, and 309 of this protein poses had favourable ZRANK scores. Most of the top predictions belonged to unassigned clusters, therefore pose 4 in cluster 2 (Table 3.2) shows the most stable conformation of the complex structure for VP3 $_{arm}$-hemoglobin and the rest of the protein poses with these conformation are in Table B6 (complex 1-12). Docked structure between VP3 active site and hemoglobin in cluster 2 were selected for further refinements. Pose 2 (Table 3.2) is the most stable of all the docked structures in cluster 2 (details of which are indicated in Table B6: complex 13-25)

The docking was validated by reproducing a co-crystal structure of FP2 bound to its inhibitor (cystatin) (Wang *et al*., 2006).A total of 2000 protein poses were generated and almost 65% of the predictions from initial stage unbound docking resembled the already published complex structure of FP2 and cystatin. The protein poses had ZDOCK scores between 13 and 23, and ZRANK scores between -115.71 to 1.24 in a 100 clusters. The 2000 protein poses were filtered by running a ZRANK protocol in which the same subsite residues as listed in Table 3.1 for FP2 were selected. This step was carried out in order to be consistent with method that had been employed throughout the study but was not necessary as most protein poses were already docked in the conformation that resembled published complex. The process poses protocol generated 1560 poses with the largest cluster (1) containing 105 protein poses. The ZDOCK and ZRANK scores of the first 20 protein poses are listed in Table B7.

3.4.2. Pose Refinements

Table 3.3: The total energy and interaction energies of the protease-substrate complexes before and after minimization

Complexes	Before Minimization				After Minimization			
	A1	A2	A3	B1	B2	C1	C2	C3
$FP2_{arm}$-hemo (1yvb) (P)	1.08×10^{08}	1.08×10^{08}	-86.03	1.08×10^{08}	-5.55×10^{04}	-577.28	-52.03	-525.26
$FP2_{arm}$(1yvb)-hemo	5.32×10^{08}	5.33×10^{08}	244.96	5.32×10^{08}	-5.69×10^{04}	-1747.36	-133.67	-1613.69
$FP2_{arm}$(1yvb)-hemo (AC)	2.24×10^{09}	2.24×10^{09}	-85.10	2.24×10^{09}	-5.70×10^{04}	-2099.87	-124.94	-1974.93
$FP2_{arm}$-hemo (2oul) (P)	0.31×10^{14}	0.31×10^{14}	0.29×10^{03}	0.31×10^{14}	-5.54×10^{04}	-662.42	-69.87	-592.55
$FP2_{arm}$(2oul)-hemo	1.05×10^{09}	1.05×10^{09}	-339.60	1.06×10^{09}	-5.74×10^{04}	-1718.25	-107.50	-1610.74
$FP2_{arm}$(2oul)-hemo (AC)	0.65×10^{12}	0.65×10^{12}	0.00	0.65×10^{12}	-5.74×10^{04}	-2412.08	-191.73	-2220.34
$FP2'_{arm}$-hemo (P)	1.42×10^{06}	1.43×10^{06}	-41.11	1.42×10^{06}	-5.62×10^{04}	-874.68	-80.98	-793.69
$FP2'_{arm}$-hemo	1.89×10^{08}	1.89×10^{08}	71.20	1.89×10^{08}	-5.68×10^{04}	-1248.95	-95.85	-1154.10
$FP2'_{arm}$ -hemo (AC)	3.16×10^{06}	3.16×10^{06}	-44.51	3.16×10^{06}	-5.70×10^{04}	-2525.98	-200.19	-125.79
$FP3_{arm}$-hemo	3.42×10^{08}	3.42×10^{08}	-3.10×10^{04}	3.42×10^{08}	-5.74×10^{04}	-1012.32	-99.21	-913.11
FP3- hemo (AC)	5.62×10^{06}	5.64×10^{06}	-3.12×10^{04}	5.62×10^{06}	-5.75×10^{04}	-1716.57	-98.09	-1618.49
$VP2_{arm}$-hemo	5.30×10^{08}	5.30×10^{08}	-59.51	5.30×10^{08}	-5.69×10^{04}	-1614.58	-138.66	-1475.92
VP2- hemo (AC)	7.78×10^{08}	7.77×10^{08}	131.68	7.78×10^{08}	-5.73×10^{04}	-2103.37	-191.09	-2294.46
$VP3_{arm}$-hemo	1.15×10^{10}	1.15×10^{10}	-3.12×10^{04}	1.15×10^{10}	-5.62×10^{04}	-1306.26	-104.15	-1202.11
VP3- hemo (AC)	6.56×10^{09}	6.56×10^{09}	-3.14×10^{04}	6.56×10^{09}	-5.81×10^{04}	-2274.24	-168.72	-2105.52
FP2-cystatin	1.27×10^{08}	1.27×10^{08}	-127.13	1.27×10^{08}	-2.42×10^{04}	-2316.99	-255.07	-2062.92

***A1- Total interaction energy (Kcal/mol) before minimization**
***A2- Total Van Der Waals interaction energy (Kcal/mol) before minimization**
***A3- Total Electrostatic interactions (Kcal/mol) before minimization**
***B1- Potential energy before minimization**
***B2- Potential energy after minimization**
***C1- Total interaction energy (Kcal/mol) after minimization**
***C2- Total Van Der Waals interaction energy (Kcal/mol) after minimization**
***C3- Total Electrostatic interactions (Kcal/mol) after minimization**
***P- Pose resembling published work and AC- active site**

Biochemist and many other scientists interested in studying the 3D structures of proteins and their interactions have come to the understanding that the most correctly bound complex structure has the lowest energy state (Crippen, 1991; Goldstein *et al*., 1992; Goldstein *et al*., 1992). This feature was exploited when analyzing the nature of the complexes that were generated, in which case, the potential energy (both electrostatic and Van Der Waals) of the determined complex structures were calculated before and after energy minimization. The Table above provides the summary of the energy values obtained for the representative (most correct) protein pose from the complex structures generated. A1 and B1 (Table 3.3) shows the same values for the total interaction energy before minimization and the overall energy of the structures. It can also be observed that before minimization, the complexes driving force or rather the predominant energy was contributed by Van Der Waals interaction (A2), charge-charge interaction which plays an important role driving the interaction was stable (energetically favourable) but it was compromised by the highly unstable structures. An alteration of the complexes conformation occurred after energy minimization. The overall potential energies of the structures are more or less the same (Table 3.3: B2), except for the complex structures resembling published pose, which have a significantly higher potential energy. The interaction energy between residues at the binding interface was no longer equated to the total potential energy. Energy minimization essentially played a key role in the refinement but could not yield more information about the residues at the binding interface. Once the complex structures were energy minimized, they reached local minima at more or less the same potential energy. Therefore, the interactions energy looking specifically into the total energy of residues at the binding interface was used to select representative complex structures. Complexes with the lowest interaction energy were selected as best representative of the near-native structure of the protease and substrate. As indicated in Table 3.3: C1, complexes in which the active site is bound to hemoglobin are more stable than the rest, though the differences are not significantly large. This suggests that the complimentarity to the active site is great and affirms that the degradation of hemoglobin by the cysteine proteases occurs at the active site.

3.4.3. Protein-protein interactions

Numerous interactions, both strong and weak, play a role in rendering the stability of a protein structure or an assembly. Therefore, understanding the interactions within protein structure as well as those between proteins in a complex structure is essential towards gaining insight into the molecular basis of stability of proteins and their assemblies (Tina *et al*., 2007). The protein-protein interactions at the amino acid level were determined using the PIC server. Below, is an indication of the comparisons with published data and other complexes cysteine protease – hemoglobin complex structures.

- **Comparisons with published data**

As already indicated the purpose of the present study was two-fold, comparing the generated protein-protein complex structures with published data and investigating the limitations from the observations by Wang *et al* (2006). Figure 3.4 shows how the data generated from the study compares with the work already published. These comparisons were made based on eye inspection looking at the manner in which the arm-like motif binds to hemoglobin. Furthermore, the amino acids involved in binding were comparable to those from published data.

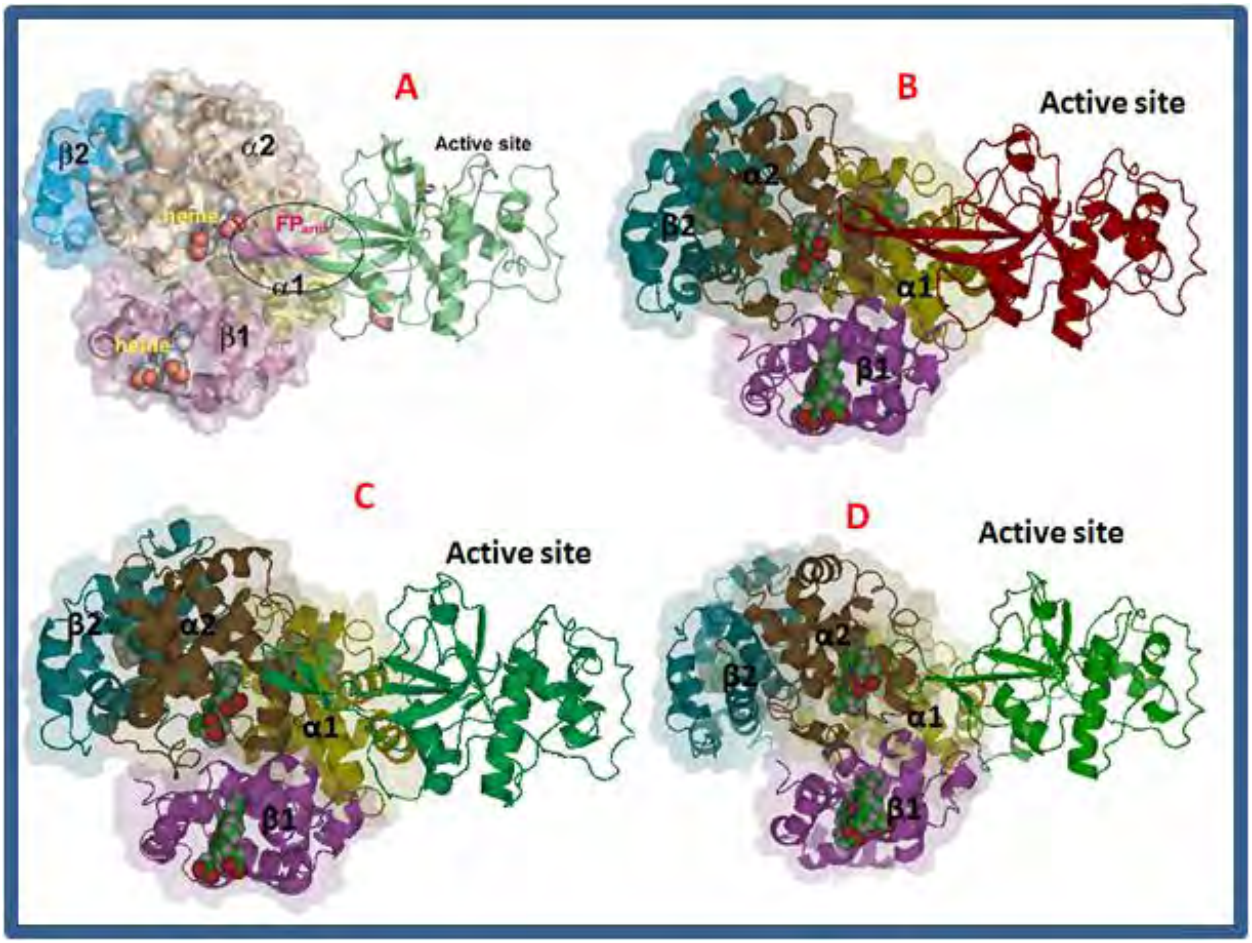

Figure 3. 4: Complex structures of FP2 and FP2' bound to hemoglobin resembling published data. (A) Indicating the complex structure from Wang *et al.,* 2006. (B) FP2' $_{arm}$ **–hemoglobin complex and (C) FP2 (1YVB) arm-hemoglobin and (D) FP2 (2OUL) arm-hemoglobin complex**

The residues involved in binding for the complex structures generated were determined using the PIC server and for complex (A), the amino acids involved in binding were obtained from (http://www.pnas.Org/content/103/31/11503/suppl/ DA1).

Table 3.4: Comparison of the $FP2_{arm}$-hemoglobin and $FP2'_{arm}$-hemoglobin complex structures

Complex	Type of interaction	Percentage
COMPLEX A	Hydrophobic	57%
	Hydrogen bonds	29%
	Charge-charge	14%
COMPLEX B	Hydrophobic	55%
	Hydrogen bonds	15%
	Charge-charge	30%
COMPLEX C	Hydrophobic	54%
	Hydrogen bonds	8%
	Charge-charge	38%
COMPLEX D	Hydrophobic	57%
	Hydrogen bonds	10%
	Charge-charge	33%

Table 3.5 shows the type of interaction (represented by percentages) involved in protein-protein complex structure formation. The forces mediating complex formation compares well with each other, indicating that hydrophobic interactions are the predominant force driving the interaction. This suggests that complex structures were predicted with a high degree of confidence and accuracy. Once it was established that the complex structures generated compares well with each other, the amino acids at the binding interface of the four complex structures build were analyzed. Amino acid analysis showed that chain A and C of hemoglobin are mostly bound to the arm-like motif of FP2', FP2 (1YVB) and FP2 (2OUL) in complex B, C and D respectively. It was also observed that residues from the highly conserved acidic motif found near the arm-like motif in both FP2 and FP2' were involved in hemoglobin binding.

The major drawback about these particular complex structures is that despite the fact that they have been predicted with a high degree of confidence (generating comparable results between published and experimental data), their docking scores were unfavourable. Their ZDOCK scores were much lower than the other protein poses and their ZRANK score were significantly higher,

suggesting that these protein poses are incorrect predictions or their complimentarity at the binding interface is unfavourable. Also comparison of complex B, C and D (Figure 3.4 and Table 3.5) interaction energies with the active site bound protein complexes and the proposed mode in which the arm-like motif binds hemoglobin (shown in Figure 3.5) protein-protein complex structures shows the former to have significantly lower interaction energy than the latter. For all three cysteine proteases (Two FP2 and FP2'), the complex structures had unfavourable ZRANK scores. The likelihood that the complex structures in Figure 3.4 could be false positives is therefore raised, moreover the mode of interaction between the proteases active site and its natural substrate hemoglobin cannot be established observing these complex protein structures. They do not indicate how the degradation of hemoglobin at the active site of cysteine proteases occurs. Additionally, the active site is far from the arm-like motif and residues within the substrate binding pockets are also not found within the binding interfaces (Refer to Table D1, D2 and D3 in the supplimentary data).

The results obtained from the protein-protein docking experiments indicates two possible interpretations: (1) The predicted protein-protein complexes in Figure 3.4 could be false positives and (2) The arm-like motif could be involved in hemoglobin binding perhaps not in the structures generated in Figure 3.4. The protein complexes in Figure 3.4 have two disadvantageous factors: unfavourable ZRANK scores (scoring functions in ZDOCK and process poses do not favor the protease-substrate associations) and the protease (arm-like motif) hemoglobin interaction do not indicate how the substrate degradation will eventually occur at the active site. However, experimental studies have proven the importance of the arm-like motif in hemoglobin degradation (Pandey *et al*., 2005). Therefore, the results obtained in the present study have not conclusively proven that the arm-like motif is not involved in hemoglobin binding. This is principally because *in silico* studies have been designed to guide and be validated by experimental studies, and past experimental studies have shown the importance of the arm-like motif in hemoglobin binding. Pandey *et al* (2005) conducted a study in which they have shown that deletion of 10 amino acids from FP2 arm-like motif results in negligible hemoglobin degradation. The parallel study conducted by Wang *et al* (2006) suggested that it is not the same hemoglobin molecule which binds to the arm-like motif and

active site, even so, how the other hemoglobin binds to the active site still remains unknown. The possibility of the arm-like motif acting as an exo-site in the binding of hemoglobin binding has also been suggested (Wang *et al*., 2006). Therefore, as informed by the results obtained herein, additional experiments were performed. These experiments included binding cysteine proteases active site to hemoglobin and analyzing their docking scores, mode of interaction and residues essential in protease-substrate association. Also, there were several protein poses from the process poses protocol (search biased to the arm-like motif) with favourable ZRANK score, and these protein poses were analyzed as explained above for the active site bound protease-substrate complex structures. These experiments were conducted in order to establish whether or not the *P .falciparum* and *P. vivax* cysteine proteases active site is involved in hemoglobin binding. Additionally, as the initial experiments did not provide much insight about the involvement of the arm-like motif in hemoglobin, the following experiments were designed to follow up on the involvement of this unique motif in the protease-substrate complex structure formation.

FP2-hemoglobin complexes

FP2 has been labeled the principal hemoglobinase and is a validated malarial drug target (Sijiwali *et al*., 2004) and it is the only *Plasmodium* cysteine proteases with hemoglobin degradation data published so far. Therefore, both FP2 structures were used in the docking experiments as it is essential to get consistent results with high reproducibility although the structures were not solved at the same resolution or with the same inhibitor bound (Wang *et al*., 2006 and Wang *et al*., 2007). Figure 3.5 shows that there is consistency in the mode of hemoglobin binding interaction between the two FP2 structures.

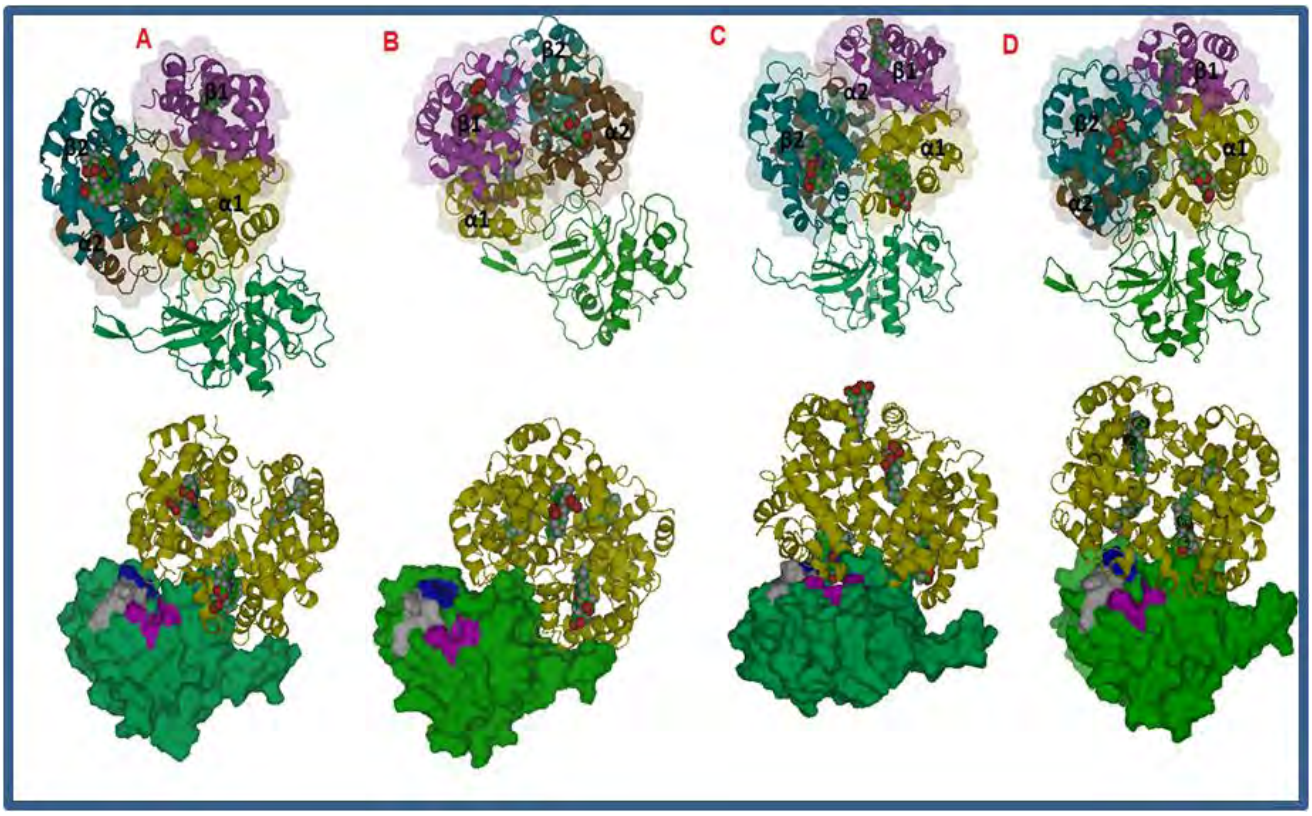

Figure 3. 5: FP2- hemoglobin complex, FP2 subsites S1, S2, S3 and S1' in blue, magenta, white and cyan respectively and FP2 (2OUL) in green and FP2 (1YVB) in limegreen

In the four protein-protein complexes shown above in Figure 3.5, there is a particular preference for hemoglobin binding at the alpha1 chain (colored yellow).

Table 3.5: Forces involved in the FP2-hemoglobin complex structures.

Complex	Type of interaction	Percentage
COMPLEX A and B	Hydrophobic	50%
	Hydrogen bonds	16%
	Charge-charge	34%
COMPLEX C and D	Hydrophobic	35%
	Hydrogen bonds	27%
	Charge-charge	38%

The $FP2_{arm}$-hemoglobin complex structures (shown in Figure 3.5 A and B) are dominated by polar amino acids at the binding interface surface. Table D4 and D6 (supplimentary data) show that the same amino acids and the type of force (Table 3.6) driving substrate-enzyme association are consistent with each other. The $FP2_{activesite}$-hemoglobin complex structure (shown in Figure 3.5 C and D) have an equal number of polar and non-polar amino acids at the enzyme (FP2) and substrate (hemoglobin) at the binding interface, refer to Table D5 and D7 in the supplimentary data. The type on interactions driving protein complex structure formation and the contribution of each has been summarized in Table 3.6 for $FP2_{activesite}$-hemoglobin. The protein complex structures shown in Figure 3.5 were selected based on their ZDOCK and ZRANK scores. The scores of the $FP2_{arm}$-hemoglobin complex structures compared well with the FP2 active site bound hemoglobin, however, the $FP2_{activesite}$-hemoglobin complex structures showed a more pronounced complimentarity. The $FP2_{activesite}$-hemoglobin complex structures had lower ZRANK scores and the lowest interaction energies.

As hemoglobin is a bulky molecule, we show in the $FP2_{arm}$-hemoglobin complex the arm-motif residues and some substrate binding site residues can be bound to hemoglobin. Table D4 and D6 show the in the complexes of arm-motif bound to hemoglobin, they are also some residues within the substrate binding sites of the proteases. Residues from S1', S1 and S2 subsites are found within the binding interface in these complexes, also there is a conserved In Table: D 4 and 6. Residues in S1', S1 and S2 subsites were found within the binding interface in the $FP2_{arm}$-hemoglobin complex structures. Also the conserved acidic motif is also found within the binding interface; therefore we suggest that it mediates the charge-charge interactions driving hemoglobin to the active site.

$FP2_{activesite}$-hemoglobin shows a binding interface consisting of closely an equal number of polar and non-polar amino acids. These particular complex structures are associated by 35% hydrophobic interactions, 27% hydrogen bonds and 38% charge-charge interactions (Table D 5 and 7). Subramanian *et al*., 2009 suggested the FP2 subsites S1 preferably cleaves ARG, LYS, GLN, THR and MET, this observations were consistent with the results obtained from our study

which show that ASN 81 of the S1 subsites binds to LYS 61 (chain A) from hemoglobin through hydrogen bond. There is a high preference for LEU at the S2 subsites residues especially by LEU 84 and LEU 172 residues. We also observed that ILE 85 has a high preference for ALA residues and ASP binds basic charged amino acid through charge-charge interactions. The slight preference of VAL by S2 subsites proposed by Submanian *et al*., 2009 was also observed. Preference for hydrophobic amino acids was also observed at S3 subsites indicating that TYR 78 binds with three ALA residues by hydrophobic interactions and GLY residues forms hydrogen bonds with other hydrophobic amino acids. S1' subsite binds to HIS residues as observed by Submanian *et al*., 2009. $FP2_{activesite}$-hemoglobin suggests that hydrolysis occurs at both the alpha (chain A) and beta (chain D) globins. We observed that even the heme group from chain A is right at the very hydrolytic site, inferring that it is also sequestrated into free hemozion (Singh *et al*., 2000).The pronounced preference for LEU at S2 subsites was also observed. Also by visual inspection, it is clear how hemoglobin is moved into the active site through the charge-charge interactions. Therefore, we propose the complexes below as the ones best representative of the mode of interaction between FP2 and hemoglobin. In which the arm-motif is bound to hemoglobin complex A and B, and through charge-charge interaction and the unstable nature of the arm motif which may also be moving as this occurs in an aqueous environment may drive hemoglobin to the active site where it is being degraded.

FP2'-hemoglobin complexes

Due to the similarity between FP2 and FP2', the latter protein 3D model was used to compare the results obtained with the former and also observe if a similar mode of interaction with mostly with the α-globin chains will be obtained. FP2' $_{active\ site}$ and FP2 $_{active\ site}$ binding to hemoglobin was also slightly similar except that more residues from β-globin seem to be at the binding interface for the former much more than the latter. $FP2'_{arm}$-hemoglobin complexes is driven by 51% hydrophobic interactions, 23% hydrogen bonds and 26% charge-charge interactions (Table D 9), this compared well with $FP2_{arm}$-hemoglobin complexes. Both

complexes affirmed that hydrophobic interactions seem to be driving complex structure stabilizations, although there was a slight difference in the charge-charge interaction of the two complexes. Previous studies of FP2' have also observed a difference in the specificity of the two proteases (Singh *et al*., 2006), which may account for this slight difference. We have also identified the negatively charged motif, extending into the active site at the binding interface of FP2', this also suggests that indeed charge-charge interactions leads hemoglobin binding into the active site. Our complex structure for FP2'arm-hemoglobin (Figure 3.6) also indicated some subsites (Table D9) residues at the binding interface, mostly S1 and S3.

FP2'$_{activesite}$-hemoglobin indicated a cleavage preference at both α- and β-globin (chain A, C and D: Figure 3.6). Preferred cleavage was observed at the S1', S2 and S3 pockets. Comparison with hydrolysis data as observed by Subramanian *et al* (2009) still confirmed the pronounced preference for LEU at the S2 subsites. The slight preference of VAL was also observed and previously unreported preference of ALA was also observed in our complex structure of FP2'active site-hemoglobin. S1 subsite pockets were not found in the binding interface, suggesting the residues from this cleavage site do not play a role in hemoglobin hydrolysis. Contrary to the observed preference of positively charged amino acids in the S3 subsite binding pocket, we found a preferred cleavage to ASP specifically by GLY 81 and GLY 83 through hydrogen bonds. Hydrophobic amino acids were found to be in the cleavage site of S1' pocket, although the specificities is less pronounced

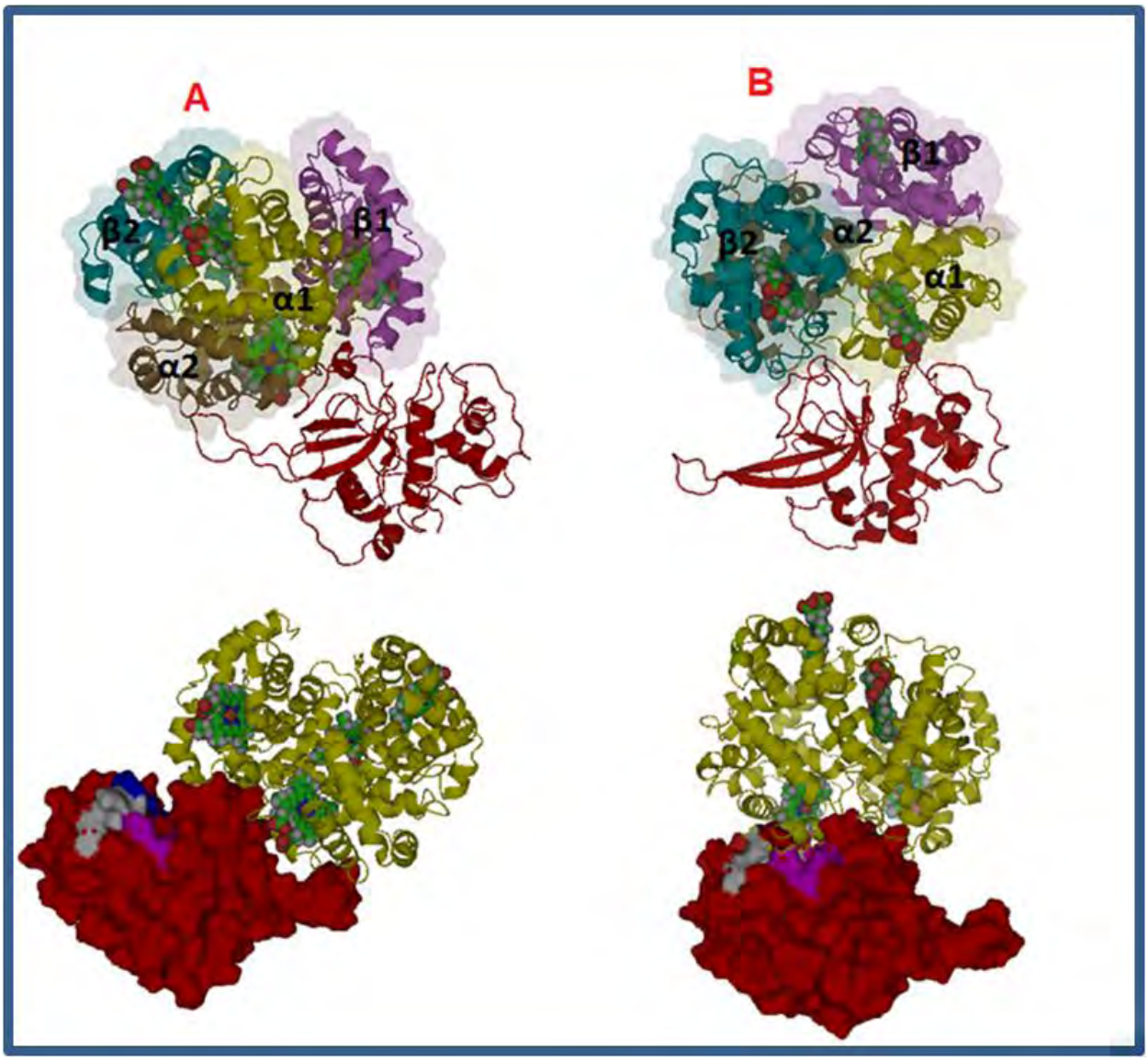

Figure 3. 6: Complex structures of falcipain-2' bound to hemoglobin. Red indicates falcipains-2', hemoglobin chain A, B, C and D are represented in green, yellow, pink and cyan respectively (B) FP2' subsites S1, S2, S3 and S1' in blue, magenta, white and cyan respectively

FP3-hemoglobin complexes

In Figure 3.7, the $FP3_{arm}$ binding hemoglobin (A) and complex B is the prediction of the active site binding mode. The amino acids within 10 Å of the binding interface were identified and the forces mediating the interactions observed. The amino acids at the binding interface of complex A consist of 16% hydrogen bonds and 43% charge-charge interaction. Details of each amino acid and type of interactions it drives are provided in Table D6. As for complex B, there is 63%, 25% and 12% hydrophobic interactions, hydrogen bonds and charge-charge interaction at the binding interface respectively.

$FP3_{arm}$-hemoglobin complex showed a preferred cleavage at the two α-globin chains whereas; FP3 $_{active\ site}$-hemoglobin shows cleavage preference at the two α-globin chains and also a slight preference for chain B (β-globin). For $FP3_{arm}$-hemoglobin complex, there were some residues from the binding pockets at the binding interfaces. S1 (TYR 90), S1' (ALA 166) and S2 (PHE 172) played a role in mediating the hydrophobic interaction between the proteases and its substrate (Figure 3.7A). The FP3 active site bound complex has the S1, S1', S2 and S3 at the binding interface playing major roles in the hydrolysis of hemoglobin (Figure 3.7B). Our data once again confirms Subramanian *et al* (2009) on the binding preference of LEU residues at the S2 binding pockets. The preference of LYS and ARG was not observed at the S1 binding pocket of the $FP3_{active\ site}$-hemoglobin complex. Hydrogen bonds between GLY 84 and GLY 85 with chain C residues ASP 74 and ASN 78 showed the contribution of the S3 subsite in hemoglobin cleavage.

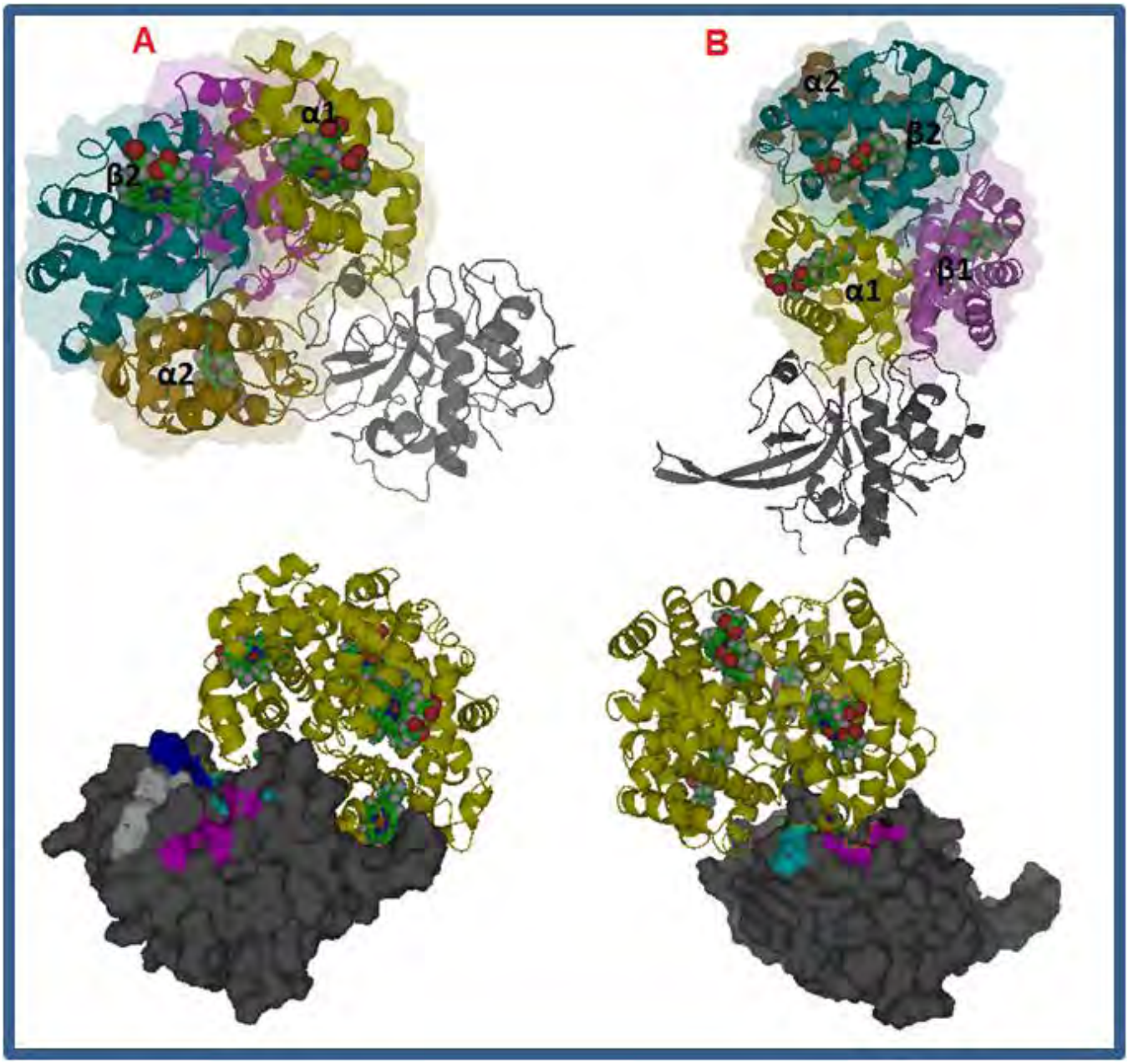

Figure 3. 7: (A) FP3-hemoglobin complex structures generated by ZDOCK. $FP3_{arm}$-hemoglobin complex and $FP3_{active\ site}$-hemoglobin complex, with FP3 in gray and hemoglobin chain A, B, C and D in cyan, hotpink, blue and magenta respectively. (B) FP3 subsites S1, S2, S3 and S1' in blue, magenta, white and cyan respectively

VP2-hemoglobin complexes

Docked protein structures between VP2 and hemoglobin as listed in Table A5 were refined by energy minimization and after inspections based on the total energy and total interaction energies, the most stable complexes of $VP2_{arm}$ and $VP2_{active\ site}$ were selected and they are shown in Figure 3.8. The forces mediating complex formation were determined and for $VP2_{arm}$-hemoglobin complex, 66%, 15% and 19% hydrophobic interactions, hydrogen bonds and charge-charge interactions respectively were found at the binding interface of the complex structure. $VP2_{active\ site}$-hemoglobin complex structure, the contribution of 45 % hydrophobic interaction, 25 % hydrogen bond and 30% charge-charge interactions towards hemoglobin hydrolysis was observed.

Figure 3.8 (below) indicates the cleavage graphical representation for both the VP2 arm and VP2 active site complex structures. The cleavage site preference for the arm-motif bound to hemoglobin was found to be at the two α-chains (Chain A and C) from hemoglobin. The subsite S1' residues are also found as the binding interface of $VP2_{arm}$-hemoglobin complex. In the complex structure of $VP2_{active\ site}$-hemoglobin, the protease seems to be cleaving both the α- and β-globin chains of hemoglobin (chain A, B and C). Data also corresponds well with FP2 and FP3 hydrolysis work about the preference of LEU residues at the S2 subsite.

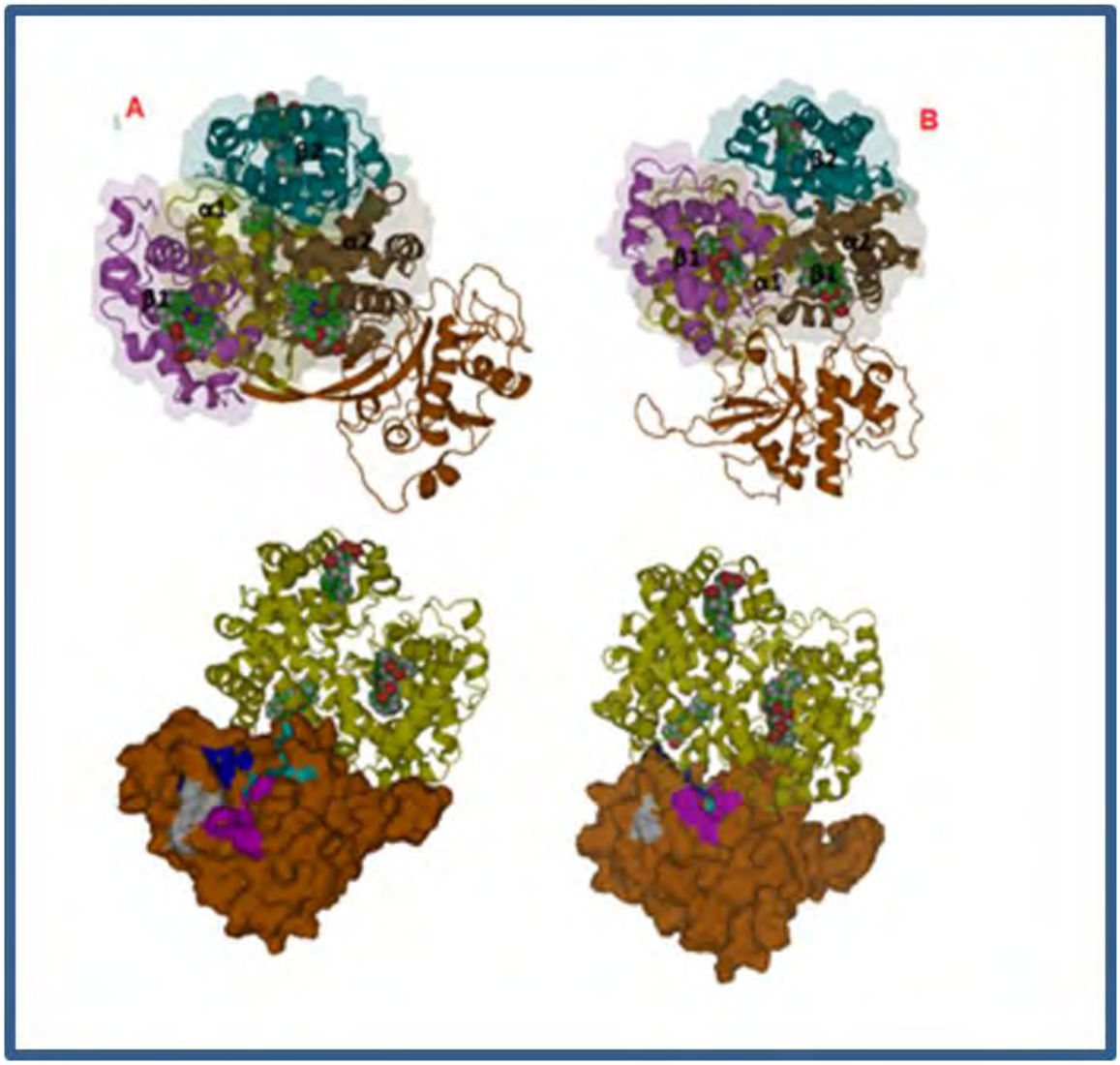

Figure 3. 8: (A) VP2-hemoglobin complex structures generated by ZDOCK. $VP2_{arm}$-hemoglobin complex and $VP2_{active\ site}$-hemoglobin complex, with VP2 in orange and hemoglobin chain A, B, C and D in sand, green, yellow and limegreen respectively. (B) VP2 subsites S1, S2, S3 and S1' in blue, magenta, white and cyan respectively.

VP3-hemoglobin complexes

The modeled structure of VP3 and hemoglobin complex structures, with hemoglobin bound to the arm-motif (complex A) and active site (complex B) were used to identify forces mediating

the interactions (Figure 3.7).The forces driving the formation of complexes were found to be 52% hydrophobic interactions, 20% hydrogen bonds and 28% charge-charge interactions for $VP3_{arm}$-hemoglobin complex structure and 43% hydrophobic interactions, 27% hydrogen bonds and 27% charge interaction for $VP3_{active\ site}$ -hemoglobin complex structure.

$VP3_{arm}$ had cleavage preference at the α–globin (chain A and C) while VP3 $_{active\ site}$ preferred α- and β-globin (chain A, B and D) for hydrolytic cleavage. Refer to the supplementary data Table B:10 for $VP3_{arm}$-hemoglobin complex structure and Table B:11 for VP3 $_{active\ site}$-hemoglobin complex structures .Once again with like all the papain-like family cysteine proteases, VP3 had a preferred cleavage preference of LEU at S2 subsite.

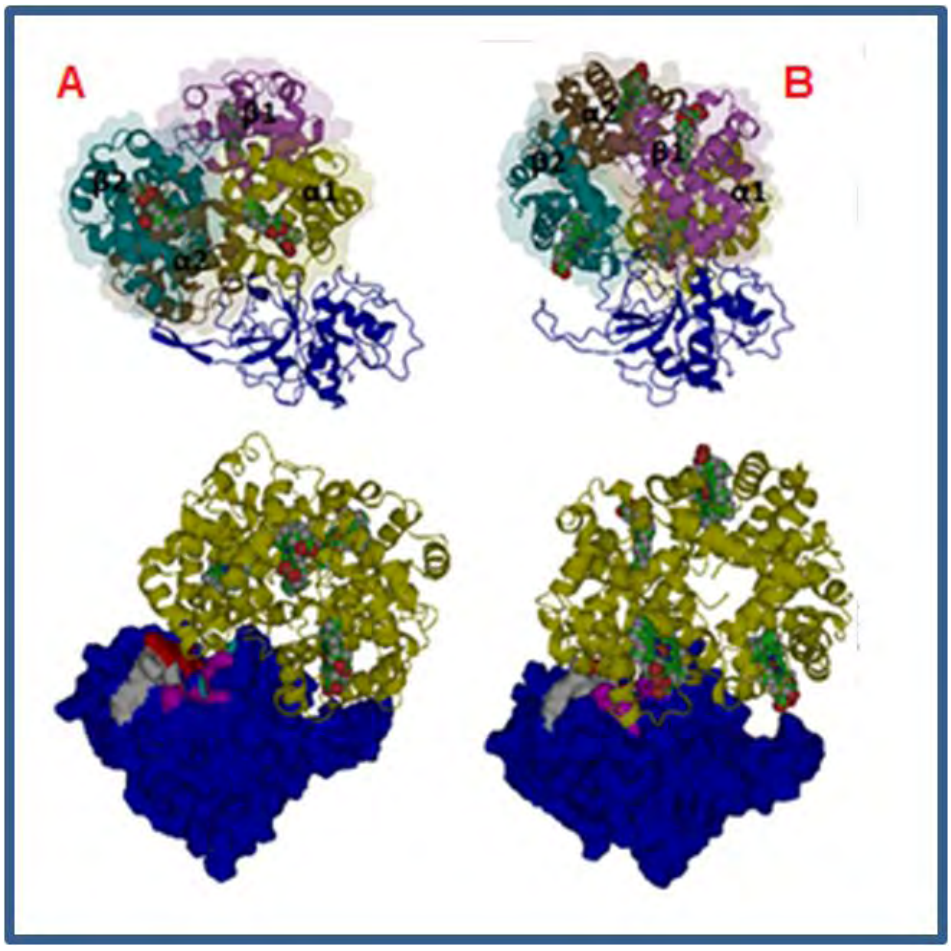

Figure 3. 9: (A) Complex structures of VP3 bound to hemoglobin. Blue indicates VP3, hemoglobin chain A, B, C and D are represented in red, yellow, magenta and cyan respectively. (B) VP3 subsites S1, S2, S3 and S1' in blue, magenta, white and cyan respectively

FP2-cystatin complexes

The docking was validated by reproducing the complex structure of FP2 and cystatin. The complex built was superimposed to the co-crystal structure already in PDB and yielded an RMSD deviation of 0.79 Å. The residues involved in binding were also observed and once again the preference of LEU at the S2 binding pocket was observed. In the FP2-cystatin complex this particular affinity for LEU by the well defined S2 pocket was observed as cystatin contains more LEU residues that hemoglobin.

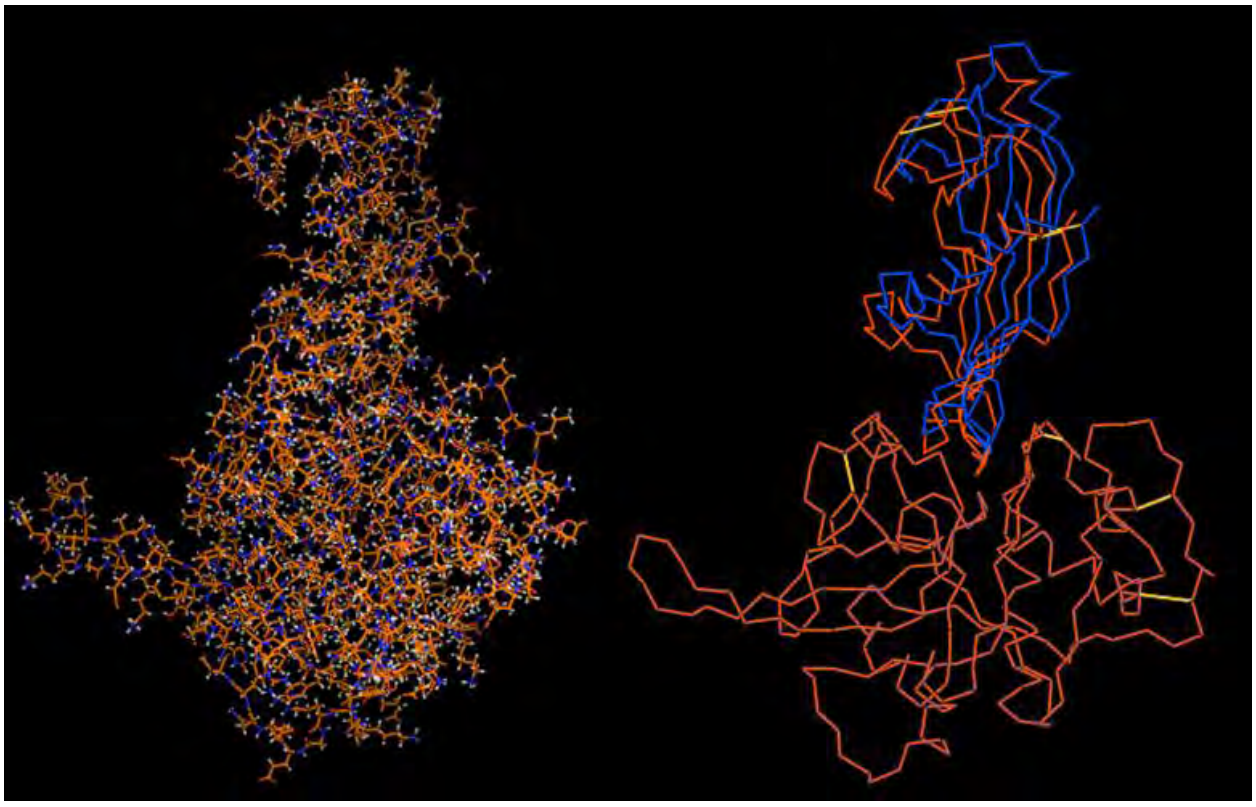

Figure 3. 10: Complex structure predicted for FP2-cystatin complex in line graphic representation and The complex generated (orange) superimposed to 1YVB (blue), the experimental elucidated co-crystal.

3.5. Summary

The main purpose of this research was to analyze the protease-substrate interaction between 5 papain-family cysteine proteases and their natural substrate human hemoglobin. The study was performed purely by bioinformatics analysis and no laboratory study was conducted. There has only been one study similar to this which was conducted by Wang and his co-researchers (2006). The predictions of protein-protein complex structures were done by protein-protein docking. Prior to initial stage docking, both proteins to be used need to be prepared. In DS 2.5., there is a provision to prepare the protein structural files using the clean function which corrected problems in the protein structure file occurring due to nonstandard naming, structures with alternate conformations, protein residue connectivity and bond orders as well as corrected missing side-chain or backbone atoms. This function does not attempt to optimize bond angles or bond lengths. The clean function was also used to remove crystallographic waters, therefore the docking was carried out *in vacuo*.

In structure of complexes, water molecules and ions are often present at the protein binding pocket along with ligands. However, during docking the ligand can displace waters and ions, the volume of a receptor site will be explored more completely if the waters and ions are removed. For example, water molecules play an important role in the catalytic activity of aspartic proteases. They act as nucleophiles by attacking the scessile peptide bond in a substrate (Coombs *et al*., 2001). The two aspartic acid residues at the active site of proteases hold the water molecules in positions and this activates them to cleave the substrate. However, inhibitors targeting the active site of aspartic proteases usually possess a hydroxyl moiety which displaces the catalytic water molecules. Displacing water molecules with a hydroxyl moiety is the same mechanism which pepstatin, a well-known extremely potent inhibitor of aspartic protease uses (Coombs *et al*., 2001). Cysteine proteases do not use the same mechanism of catalysis as aspartic acid; they employ the appropriate amino acid side chains for nucleophillic attack. It is undeniable that water molecules are important in biological systems, the scientific community has had several discussions as to whether these molecules should or should not be included in computational approaches for structure-based drug design , mainly for molecular

docking simulations. However, no consensus has been reached about the roles water molecules play in such systems simulations (Zhong *et al*., 2007). In a study conducted by Huang and co-workers (2008), they demonstrated that the docking accuracy was improved by including water molecules in 12 targets that they had used in their study. On the contrary Birch *et al*. (2002) described the opposite in a study of neuraminidase structures. Water molecules are usually included in structure-based drug design methods as this approaches mimics the binding of a ligand to a protein and mainly concentrate on the a cavity of a protein surface that is involved in docking. However, protein-protein docking programs mainly focus on poses (the binding mode), they geometry of the ligand in the binding site and including water molecules in the docking is a computationally expensive process. Some of the leading servers for the initial stage of protein-protein docking include ClusPro (Comeau *et al*., 2004), GRAMM-X (Tovchigrechko and Vakser, 2005; Tovchigrechko and Vakser, 2006), HEX (Ritchie and Kemp, 1999; Ritchie and Kemp, 2000), ZDOCK, PatchDock and SymmDock (Duhovy *et al*., 2002; Schneidman-Duhovy *et al*., 2005). All these protein docking programs are well-known for their success in the Critical Assessment of Predicted Interactions (CAPRI) experiment, and they require water molecules and heteroatoms to be removed from the input files before running the docking. The program may remove the water molecules itself (HEX) or the user may have to do so before the experiment. In our study we used ZDOCK which also requires water molecules to be removed. The strength of ZDOCK lies in the fact that it has consistently been rated as one of the most accurate docking programs in CAPRI experiments (Wiehi *et al*., 2008), a competition where the scientific community makes blind predictions of protein-complexes which are then compared to experimental solved protein complexes. Also, the excellency of ZDOCK is indicated by the fact that it has also been implemented in ClusPro, another program well-known for its success in CAPRI experiments. Like all good scientific problems, the protein docking problem is easy to state but difficult to solve (Ritchie, 2008). The difficulty with adding water molecules was that ZDOCK does not recognize them. ZDOCK removes hydrogens from the waters and then approximates the remaining water atoms (the oxygens) as having the same radii as carbon with zero charge. The presence of water affects the protein surface calculation in ZDOCK and this changes the scoring of the poses in terms of complimentarity and the electrostatic interaction.

The program includes a desolvation term for the desolvation of the proteins when they bind to each other. Also, the study was carried out in order to determine whether the finding from Wang and co-workers (2006), which suggest that hemoglobin binds to $FP2_{arm}$ were valid. In that study, computational models of FP2-hemoglobin complexes were initially generated by chimera (www.cgl.uscf.edu/chimera). The docking method employed a low-resolution, rigid-body, Monte Carlo search which was followed by simultaneous optimization of the backbone and side-chain conformations using Monte Carlo minimization with RosettaDock (www.rosettacommons.org). The Wang and co-workers (2006) did not provide any evidence of the docking being carried out in vacuo. We used ZDOCK instead of chimera for the initial docking because it is more reliable than chimera, whose success in docking is not indicated anywhere in the proceedings of the CAPRI experiments.

Once the protein structures were cleaned, hydrogen atoms were added on to each of the protein structural files, and then proteins were ready for the protein-protein docking. Protein-protein docking algorithms were trained to generate a numerous possibilities of complex structures, with a few of them being near-native. In the case of our study, ZDOCK generated 2000 protein poses for all predictions. These poses were ranked by ZDOCK score and ZRANK scoring function. ZDOCK score is based on shape complimentarity and it is positive. Therefore, the higher the ZDOCK score the more likely is the prediction to be near-native. Results from ZDOCK for all 5 set of complex structures predicted had ZDOCK score from 13.00-24.00. The complexes selected for analysis were not necessarily those with the highest ZDOCK scores but had lowest ZRANK score All initial stage ZDOCK predictions were filtered using the process poses protocol. The process poses (ZDOCK) protocol allows the user to select a subset from a set of docked protein poses generated from ZDOCK. The selection can be made according to pose rank or specifying residues at the binding interface. Process poses also re-rank poses with a ZRANK scoring function (Pierce and Weng, 2007).

The selection of residues at the binding interface can be done in two ways: either block residues not involved in binding or specify residues involved in binding. For the former, residues

specified are marked as a special type and only the penalty part of PSC score is applied to them during calculations. Therefore, complex conformations with blocked residues obtain a lower score than other conformations. If it is known that some residues should be at the protein-protein binding interface, a procedure is used to filter poses that do not include the specified residues. Blocking residues or including residues for docking reduces the number of hits for refinement stage and increases the confidence of the final prediction.

The complex structures of arm-bound to hemoglobin and active site- bound to hemoglobin were not significantly different as per initial stage unbound docking analysis. However, a major difference was observed for the complex structure resembling published data. This particular complex was preceded by an energetically unfavourable ZRANK score which suggest that it may be a false-positive.

The overall energy of the predictions made by ZDOCK was calculated and complexes exhibited highly unstable conformations. van der Waals forces were the main forces rendering the instabilities of the complex structures; this may be due to the inclusion of β-factors in ZDOCK (Chen *et al*., 2003a) which filter out the electrostatic interactions. CAPRI experiments have indicated the success of ZDOCK in initial stage unbound docking predictions; however complexes generated are to be refined to remove steric contact and bad conformation through energy minimization. Therefore, energy minimization employing CHARMm forcefield was used for the refinement stage. The effect of energy minimization was clearly observed by the drastic change in the free energy stage (Total energy) of the complexes structures. Bearing in mind that at equilibrium, most protein are at their global minimimum free energy state (Crippen, 1990), this principle was used as to select the most stable conformation. However, complex structures reached almost the same energy state after several cycles of minimization. This is explained by the similarity in the structure of *P. falciparum* and *P. vivax* (same amino acid , molecular weight) and therefore their complexes with the same molecule should reach the same energy state (local minima).The interaction energies of the complexes was calculated and the lowest energy was assumed to represent near-native conformation of the structures.

Chapter 4

Concluding discussion and future recommendations

Malaria still remains a major health concern due to the parasite *P. falciparum*, which has developed resistance to most currently available drugs. Its continual mortality and morbidity rates are escalating and on a continual rise. Cysteine proteases have been validated as drug targets and are being studied for that purpose. *Plasmodium* cysteine proteases (the 5 which were studied in **chapter 3**) degrade hemoglobin, which in turn provides the parasite with amino acids ensuring its survival within the human host. On a recent study conducted by Ch'ng *et al*., 2010, a cysteine protease inhibitor was used to demonstrate Clan CA papain-like cysteine proteases of *P. falciparum* are involved in chloroquine mediated programmed cell death (Ch'ng *et al*., 2010). FP2 being the most abundant and well studied protease shares high sequence identity with FP2', FP3, VP2 and VP3. Only FP2 and FP3 3D structures were readily available in PDB and could be easily retrieved. The structure of FP2' has not yet been solved by experimental techniques, due to the fact that the existence of this protease was not known until the sequencing of *P. falciparum* genome. FP2' shares significantly high sequence identity with FP2, it also shares similar biochemical features with FP2. Therefore, FP2 was used as a template to generate a credible homology model structure of FP2'.

VP2 and VP3 has received very little interest over the years in terms of studying these proteases, even though *P. vivax* is the most widely spread malarial parasite and the second most lethal *Plasmodium* species. VP2 and VP3 functions can be easily inferred from FP2 and FP3, as they have similar physiological and biochemical functions. Therefore, FP2 and FP3 structures were used as templates to model the structures of VP2 and VP3. Evaluation of these models with various validation programs raised the confidence in the models to ensure that they are the best possible representation of the target proteases. The results obtained for the models of FP2', VP2 should help guide experimental determination of these structures, this

models further offer insight into the structural features of this protease and even their biochemical attributes. The models generated from this study were predicted with reasonable accuracy. The differences in the substrate binding pockets have been mapped out and it was shown that the sequence variability between the target proteins and template protein structures resulted in minor distortion on the structures. In the future, the models generated may go a long way to help identify an efficient bacterial expression system suitable for VP2 and VP3 as it has been difficult to do so.

The structural models (FP2', VP2 and VP3), FP2 and FP3 were docked to hemoglobin. This was done for the purpose of analyzing the protein-protein interaction occurring between these proteases and hemoglobin. Protease-substrate complexes were generated for each protease using ZDOCK and the predicted complex structures re-filtered by process poses. Unstable complexes generated from this initial docking stage were refined using the minimization algorithm. The total energy and interaction energies of the refined complexes were calculated using the "calculate energy" and "calculate interaction energy" protocols in DS 2.5. The amino acids and forces mediating the interactions at the binding interface were identified. The ZDOCK algorithm has been consistently ranked in the top 10 for CAPRI experiments (Wiehi *et al*., 2008), and had been commended for its accuracy in initial stage unbound docking. Results obtained using ZDOCK algorithm represents near-native complex structures of the protease-substrate complexes.

The present study suggests that the *Plasmodium* cysteine proteases used as ligands for the protein-protein docking study could possibly bind hemoglobin in more ways than one. The results obtained are consistent with the findings by Subramanian *et al.,* 2009, which suggests that hemoglobin hydrolysis is not carried out in a highly ordered process, but rather it is preceded by falcipains rapid cleavage at multiple sites (Subramanian *et al.,* 2009). , the importance of the arm-like motif was also demonstrated in a study where Pandey *et al.,* 2005, demonstrated that mutating the amino acids residues constituting the arm-like motif results in negligible hemoglobin hydrolysis. Therefore, this study suggests that the arm-like motif may be binding certain amino acids on the hemoglobin molecule and due to hemoglobin 's bulky size

some of its amino acid already establishes contact with the subsite residues especially through the highly conserved acidic residues found near the active site of all 5 cysteine protease used in the experiments. For all 5 proteases, the complex structures in which the arm-like motif was bound to hemoglobin, there is a conserved acidic motif present at the binding interface. This acidic motif: ASP 154, ASP 155, GLU 167 and ASP 170 (FP2 numbering but it is present in all cysteine proteases) extends from the arm-motif into the active site. Thus we support the proposal initially made by Pandey *et al* (2005) that hemoglobin binds with the acidic motif through charge-charge interaction (Refer to supplimentary data: Appendix D), thereby coming closer to the protease active site where it is degraded. Another aspect to this is that the arm-like motif is weakly conserved across all 5 cysteine protease and also unstable. Therefore, due to this unstable nature of the arm-motif it might move hemoglobin to the active site where the degradation occurs. However, the critical role that this arm-like motif plays in hemoglobin binding cannot be ruled out, as the Pandey *et al* (2005) showed that a peptide encoding this motif blocked the hydrolysis of hemoglobin but not the casein, suggesting that it is essential for hemoglobin binding (Pandey *et al* 2005). The results in this study offer a great breakthrough in terms of understanding how the arm-like motif from the falcipains and vivapains might play a role in the binding of hemoglobin, and how the degradation eventually occurs at the active site. Throughout the study, we observed good complimentarity between cysteine proteases active site and hemoglobin. The energies obtained are not the same across all 5 cysteine proteases but there is consistency in the results obtained indicating that the cysteine proteases arm-like motif hemoglobin complex structures have stable energy values after minimization but their energies are always lower than the active site bound complexes. This observation also provides strong evidence that the actual degradation of hemoglobin occurs at the protease active site. This interpretation is in agreement with the structure of FP2 as solved by Wang *et al*., (2006) indicating that the arm-like motif is surrounded by a predominant negative charges and it may be these residues that are functioning as exo-sites for hemoglobin binding.

Amino acid analysis of the residues involved in binding has also offered us with an interesting discovery; we observed that there is a high level of specificity in the amino acids involved in binding for all 5 proteases. The amino acids involved in binding are mostly conserved across

these five cysteine proteases and they are clustered in the same area. These residues are structurally at the same position, and this information will be useful for future studies. .

The aims of our study were successfully achieved, and some interesting discoveries about the *Plasmodium* cysteine proteases were uncovered. The 3 target protein models of interest were generated with a reasonable accuracy. The study showed that the overall fold of *Plasmodium* cysteine proteases including their unique features is present in both the *falciparum* and *vivax* species. The study also has managed to propose from *in silico* analysis that indeed the active site of cysteine proteases is involved in the digestion of hemoglobin. The major motivation for this proposal is the complimentary of the substrate-enzyme complexes as obtained from the protein-protein docking studies, and the presence of highly charged residues at the arm-like motif. Therefore, it may be the charged residues in arm-like motif that are crucial for interaction and not digestion with hemoglobin. The current proposal made is that hemoglobin (highly dominated by charged surfaces) first binds the arm-like motif and through charge-charge interaction the hemoglobin is brought closer to the active site where hydrolysis occurs. The study has managed to propose and suggest an answer to the question of hemoglobin degradation at the active site. The interaction of *Plasmodium* cysteine proteases with the biologically relevant substrate hemoglobin is an important starting point for the development of an effective drug against the endemic malaria. This is the first study to do so, and also, the results obtained from this study will guide the design of inhibitors that should interfere with hemoglobin binding and digestion. Inhibition of these cysteine proteases goes a long way to discovering a chemotherapeutic target of malaria. The roles of these cysteine proteases in parasite life cycle have been outlined and prevention of their role could significantly reduce the death rates and global frustration posed by malaria.

Due to time constraints molecular dynamics studies were not carried out where the proposed mode of interaction of cysteine protease with hemoglobin was to be observed were not done. Therefore, in future we propose that this analysis should be the starting point. The residues identified to be involved in binding should be compared to those of papain-like cysteine

proteases inhibitors, although sequence analysis has shown that human cathepsin L-like subsite 2 contains residues different from that of *Plasmodium* therefore, this region could be the main focus for inhibitor design. The more specific the inhibitor the better the success of such a study, as they are cysteine proteases called cathepsins (Mason *et al*., 1985; Ritonja *et al*., 1985; Chapman *et al*., 1997) within the human liver which are involved in the degradation of old blood cells. Therefore, non-specific inhibitors may also prevent the activities of such proteases. Alternatively, a further *in silico* study can be carried out between human hemoglobin and human cysteine protease to identify the mode of interaction, residues involved in binding and comparisons between human and parasitic cysteine proteases binding to hemoglobin. This will also gain more insight into the inhibitor to be designed and the effect it would have.

The propeptide of cysteine proteases inhibit the selectivity and specificity of these enzymes. (Chapman *et al*., 1997). Therefore, the prodomain of FP2 (since it is the most widely studied) can be evaluated for its inhibitory effect on the other cysteine proteases. In a more recent study Pandey *et al*., (2009) showed that the C-terminal part of the prodomain is required for inhibiting the mature domain. The study entailed the expression of constructs encoding different portions of the prodomain and it was found that the C-terminal segment (LEU 155-ASP 243) is critical for mature protease inhibition. Two other motifs (ERNIN and GNFD) within the prodomain which are conserved across all cathepsin L-like cysteine proteases have been demonstrated to have inhibitory activity (Pandey *et al*., 2009). Therefore we propose the synthesis of a peptidomimetric inhibitor including prodomain residues critical for inhibition as a possible drug target. The study must use human cathepsin L-like cysteine protease as a control and if the prodomain of FP2 inhibits the other *Plasmodium* cysteine proteases and none from human, this segment should be pursued for drug design.

As cysteine proteases are synthesized between the early trophozoite and late trophozoite stage, it is expected that any ideal inhibitor preventing their activity should block late-stage malaria parasite development in the red blood cells. Also experimental studies using the designed inhibitors should be carried out. *P. falciparum* can only be cultured in human blood so that will help test the effect of inhibitor on human cysteine proteases. Much work still needs to

be done on malarial cysteine proteases, but the development of inhibitors of these proteases stands to play an important role in managing malaria throughout the world.

References

Accordingly, V. and Boxes, I. **(2006)**. Principles of X-ray Diffraction. *Analys*. 3:1-42.

Abraham, Z., Martinez, M., Carbonero, P. and Diaz, I. **(2006)**. Structural and functional diversity within the cystatin gene family of *Hordeum vulgare*. *Access*. 57(15):4245-4255.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipmanl D.J. **(1990)**. Basic Local Alignment Search Tool. *J. Mol. Biol.*215:403-410.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Madden, T. L., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. **(1997)**. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25(17):3389-3402.

Aravind, L., Iyer L.M., Wellems, T.E. and Miller, L.H. **(2003)**. *Plasmodium* Biology: Genomic Gleanings The highly A-T rich genomes of human and rodent. *Cell*. 115:771-785.

Baird, J.K. **(2004)**. Minireview Chloroquine Resistance in *Plasmodium vivax*. *Health*. 115:771-785.

Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N. and Yeh, L.L. **(2004)**. The universal protein resource (UniProt). *Nucl. Acids Res.* (33):D154-D159.

Bairoch, A. **(2005)**. The universal protein resource (UniProt). *Nucleic. Acids Res*. 33:D154-D159.

Baker, D. and Sali, A. **(2001)**. Protein structure prediction and structural genomics. *Science.* 294:93-96.

Banerjee, R., Liu, J., Beatty, W., Pelosof, L., Klemba, M. and Goldberg,D. E. **(2002)**. Four plasmepsins are active in the *Plasmodium falciparum* food vacuole, including a protease with an active-site histidine. *Proc. Natl. Acad. Sci. USA* 99: 990-995.

Barnwell, J., Carlton, J., Collins, W., Mullikin, J. and Saul, A. **(2007)**. Neglected Burden of Human vivax Malaria: Comparative Analysis of *Plasmodium vivax* and Key Related Species. *Plasmo. Comp*.1: 1-26.

Barrett A.J. **(1994)**. Classification of peptidases. *Methods Enzymol*.244:1–15.

Barrett, A.J., Rawlings, N.D. and Woessner, J.F. **(1998)**.Handbook of proteolytic enzymes. *Academic Press, San Diego, Calif*. (Kennedy JF, Tarun N, eds.). 1: 338-360.

Baton, L.A. and Ranford-cartwright L.C. **(2005)**. Do malaria ookinete surface proteins P25 and P28 mediate parasite entry into mosquito midgut epithelial cells ? *Malaria J*. 8:1-8.

Bernauer, J., Bahadur, R.P., Rodier, F. and Poupon, A. **(2008)**. Structural bioinformatics DiMoVo : a Voronoi tessellation-based method for discriminating crystallographic and biological protein – protein interactions. *Bioinfo*. 24(5):652-658.

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.E., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. **(1977)**. The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol* 112, 535-542.

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

Berry, C., Humphreys, M. J., Matharu, P., Granger, R., Horrocks, P., Moon,R. P., Certa, U., Ridley, R. G., Bur, D. & Kay, J. **(1999).** A distinct member of the aspartic proteinase gene family from the human malaria parasite Plasmodium falciparum. *FEBS Lett.* 447,149–154

Bertini, I. and Luchinat, C. **(1998)**. NMR of paramagnetic proteins NMR of paramagnetic proteins. *Protein Sci*. 1-37.

Betts, M.J. and Sternberg, M.J.E. **(1999)**. An analysis of conformational changes on protein – protein association: implications for predictive docking. *Protein Eng*.12(4):271-283.

Birch, L.; Murray, C. W.; Hartshorn, M. J.; Tickle, I. J, and Verdonk,M. L. **(2002)** Sensitivity of molecular docking to induced fit effects ininfluenza virus neuraminidase. *J. Comput.-Aided Mol. Des.* 16 (12), 855–869

Bjelic, S., Nervall, M., Gutiørrez-de-terµn, H., Ersmark, K. and Hallberg, A. **(2007)**. Review Computational inhibitor design against malaria plasmepsins. *Cell and Mol Life Sci*.64:2285 - 2305.

Blake, J.D. and Cohen, F.E. **(2001)**. Pairwise sequence alignment below the twilight zone. *J. Mol. Biol*. 307:721-735

Bode, W. and Huber, R. **(1992)**. Natural protein proteinase inhibitors and their interaction with proteinases. *Euro journal of biochem*. 204(2):433-451.

Bower, M.J., Cohen, F.E. and Dunbrack, R.L. **(1997)**. Prediction of Protein Side-chain Rotamers from a Backbone-dependent Rotamer Library: A New Homology Modeling Tool. *Mol Pharmac*. 1268-1282.

Bragg, A.W.L. **(1913)**. The Structure of Some Crystals as Indicated by Their Diffraction of X-rays The Structure of Some Crystals as Indicated by their Diffraction of X-rays . *Soc*. 89(610):248-277.

Breman, J.G. **(2001)**. The ears of the hippopotamus manifestations, determinants and estimates of the malaria burden. *Tropic. Med*. 64:1-11.

Brock, K., Talley, K., Coley, K., Kundrotas, P. and Alexov, E. (**2007**). Optimization of Electrostatic Interactions in Protein-Protein Complexes. *Biophysic J.* 93(10):3340-3352.

Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M. **(1983)**. CHARMM: a program for macromolecular energy, minimization and dynamics calculations. *J. Comput. Chem*. 4: 187-217.

Brooks, C.L., Chen, J. and Im, W. **(2007**). Balancing solvation and intramolecular interactions: toward a consistent generalized born force field (CMAP opt. for GBSW) . *J Am Chem Soc* 128 (11): 3728–3736.

Brown, M.S., Ye, J., Rawson, R.B. and Goldstein, J.L. **(2000)**. Regulated Intramembrane Proteolysis : A Control Mechanism Conserved from Bacteria to Humans. *Cell*. 100:391-398.

Ceusters, W. and Smith, B. **(2009)**. Malaria Diagnosis and the Plasmodium Life Cycle : the BFO Perspective. *Bioinfo and life scie*. 1-10.

Ch'ng JH, Kotturi SR, Chong AG, Lear MJ and Tan KS-W **(2010)**. A programmed cell deathpathway in the malaria parasite *Plasmodium falciparum* has general features of mammalian apoptosis but is mediated by clan CA cysteine proteases. *Cell Death and Disease*, 126; 1038

Chen, R. and Weng, Z. **(2002)**. Docking Unbound Proteins Using Shape Complementarity , Desolvation , and Electrostatics. *Biomedic Eng*.; 294(09):281-294.

Chen, R., Li, L. and Weng, Z. **(2003a)**. ZDOCK: An Initial-Stage Protein-Docking Algorithm. *Performance Eval*. 11:80 -87.

Chen, R., Mintseris, J. and Weng, Z. **(2003b)**. A Protein – Protein Docking Benchmark. *Biochem*. 91(10):88 -91.

Chen, R., Tong, W., Mintseris, J., Li, L. and Weng, Z. **(2003c)**. ZDOCK Predictions for the CAPRI Challenge. *Sys*.73(10):68 -73.

Chen, S.W. and Pellequer, J.l. **(2004)**. Identification of Functionally Important Residues in Proteins Using Comparative Models. *Curr*. 595-605.

Cheryl, C., Ling, G.L. and Tiow-suan, S. **(2005)**. Differences in biochemical properties of the Plasmodial falcipain-2 and berghepain-2 orthologues : Implications for *in vivo* screens of inhibitors. *FEMS Microbiol Letters*. 249:315-321.

Chothial, C. and Lesk A.M. **(1986)**. The relation between the divergence of sequence and structure in proteins. *EMBO Journal*. 5(4):823-826.

Chung, S.Y. and Subbiah, S. **(1996)**. A structural explanation for the twilight zone of protein sequence homology. *Struct*. 4:1123 - 1127.

Clore, G.M. and Gronenborn, A.M. **(1998)**. NMR structure determination of proteins and protein complexes larger than 20 kDa. *Curr Opinion in Chem Biol*. 564-570.

Cogswell, F.B. **(1992)**. The Hypnozoite and Relapse in Primate Malaria. *Microbiol*. 5(1):26-35.

Comeau, S.R., Gatchell, D.W. and Vajda, S. **(2004)**. ClusPro : an automated docking and discrimination method for the prediction of protein complexes. *Bioinfo*. 20(1):45-50.

Coombs, G. H., Goldberg, D. E., Klemba, M., Berry, C., Kay, J. & Mottram,J. C. **(2001)** *Trends Parasitol.* 17, 532**–**537.

Cooper, R.A., Hartwig, C.L. and Ferdig, M.T. **(2005)**. pfcrt is more than the *Plasmodium falciparum* chloroquine resistance gene: a functional and evolutionary perspective. *Acta Tropica*. 94: 170-180.

Crippen, G.M. **(1991)**. Prediction of Protein Folding from Amino Acid Sequence over Discrete Conformation Spaces. 4232-4237.

Cunha, C.A., Romão, M.J., Sadeghi, S.J., Gilardi, F.V.G. and Soares, C.M. **(1999)**. Effects of protein-protein interactions on electron transfer : docking and electron transfer calculations for complexes between flavodoxin and c-type cytochromes. *Polar*. 360-374.

Dahl, E.L. and Rosenthal, P.J. **(2005)**. Biosynthesis, localization, and processing of falcipain cysteine proteases of *Plasmodium falciparum*. *Parasit*. 139:205-212.

Dalton, J.A.R. and Jackson, R.M. **(2007)**. Structural bioinformatics An evaluation of automated homology modeling methods at low target – template sequence similarity. *Bioinfo*. 23(15):1901-1908.

Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. **(1978)**. A Model of Evolutionary Change in Proteins. *Amino Acids*. 10: 345-352.

Desai, P.V., Patny, A., Sabnis, Y., Gut, J.,Rosenthal, P.J., Tekwani, B. and Srivastava, A. **(2004)**. Identification of Novel Parasitic Cysteine Protease Inhibitors Using Virtual Screening . 1. *The ChemBridge Database*. 6609-6615.

Dibrov, A., Myal, Y. and Leygue, E. **(2009)**. Computational Modeling of Protein Interactions: Energy Minimization for the Refinement and Scoring of Association Decoys. *Acta Biot*. 419-428.

Dorin-semblat, D., Sicard, A., Doerig, C., Ranford-cartwright, L. and Doerig, C. **(2008)**. Disruption of the PfPK7 Gene Impairs Schizogony and Sporogony in the Human Malaria Parasite *Plasmodium falciparum* . *Soc.* 7(2):279-285.

Dowse, T.J., Koussis, K., Blackman, M.J. and Soldati-Favre, D. **(2009)**. Roles of Proteases during Invasion and Egress by *Plasmodium* and *Toxoplasma*. *Mol. Mechanisms of parasite Invasion*. 121-139.

Drew, M.E., Banerjee, R., and Uffman, E.W. **(2008)**. *Plasmodium* Food Vacoule Plasmepsins Are Activated by Falcipains. *Journal of Biol Chem*.283(19):12870 -12876.

Eggleson, K.K., Duffin, K.L., and Goldberg, D.E. **(1999)**. Identification and Characterization of Falcilysin, a Metallopeptidase Involved in Hemoglobin Catabolism within the Malaria Parasite *Plasmodium falciparum*. *Biochem*. 274(45):32411-32417.

Enayati, A., Hemingway, J. and Garner, P., Eggleson, K.K., Duffin, K.L. and Goldberg, D.E. **(1999)**. Identification and Characterization of Falcilysin, a Metallopeptidase Involved in Hemoglobin Catabolism within the Malaria Parasite *Plasmodium falciparum*. *Biochem*. 274(45):32411-32417.

Enayati, A., Hemingway, J. and Garner, P. **(2010)**. Electronic mosquito repellents for preventing mosquito bites and malaria infection. *Env Health*. (3):1-16.

Eswar, N., Eramian, D., Webb, B., Shen, M. and Sali, A. **(2006)**. Protein Structure Modeling With MODELLER. *Notes*.:1-25.

Ettari, R., Bova, F., Zappala, M., Grasso, S. and Micale, N. **(2009)**. Falcipain-2 Inhibitors. *Med Res Rev*. 30(1):136-167.

Feng, D.F. and Doolittle, R.F. **(1987)**. Progressive sequence alignment as a prerequisite to correct phylogenetic tress. *J Mol Evol*. (25):351-360.

Fernandez-fuentes, N., Rai, B.K., Madrid-aliste, C.J. and Fajardo, J.E. **(2007)**. Structural bioinformatics Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments. *Bioinfo*. 23(19):2558-2565.

Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S., Mackerell, A.D.J., Bashford, D., Bellott, R.L., Dunbrack, R.L. and Jr,. E. **(1998)**. All-Atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Physical. Chem* 102:3586-3616.

Fischer, D., Lin, S.L., Wolfson, H.L., Nussinov, R. and Aviv, T. **(1995)**. A Geometry-based Suite of Molecular Docking Processes. *Mol Biol*. 459-477.

Florens, L. Washburn, M.P. and Raine, J.D. **(2002)**. A proteomic view of the *Plasmodium falciparum* life cycle. *Nature*. 419(October):520-526.

Forrest, L.R., Tang, C.L. and Honig, B. **(2006)**. On the Accuracy of Homology Modeling and Sequence Alignment Methods Applied to applied to membrane proteins *Biophysic J*. 91:508-517.

Fox, T., Miguel, E.D. and Mort, J.S. **(1992)**. Potent Slow-Binding Inhibition of Cathepsin B by Its Propeptide. *Biochem* 31:12571-12576.

Francis, S.E., Sullivan, D.J. and Goldberg, D.E. **(1997)**. Hemoglobin metabolism in the malaria parasite *Plasmodium falciparum*.*Mol Microbiol*. 97-123.

Francis-lyon, P., Gu, S., Hass, J., Amenta, N. and Koehl, P. **(2010)**. Sampling the conformation of protein surface residues for flexible protein docking. *BMC Bioinformatics*. 11(1):575-586.

Fujioka, H. and Aikawa, M. **(2002)**. Structure and Life Cycle. *Chem immunol*. 80:1-7.

Fulton, J. D., and P. T. Grant. **(1956).** The sulphur requirements of the erythrocytic form of Plasmodium knowlesi. *Biochem. J*.63:274-282.

Gabb. H.A., Jackson, R.M. and Sternberg, M.J.E. **(1997)**. Modeling Protein Docking using Shape Complementarity , Electrostatics and Biochemical Information. *J mol biol*. 272:106-120.

Gardner, M.J., Hall, N., Fung, E., White, O.B., Matthew ,H., Richard, W.C., Jane, M.P., Arnab, N., Karen, E.B., Sharen, P.T., James, K.E., Jonathan, A., Rutherford, K., Salzberg, S.L., Craig, A.K., Sue, C., Man-suen, N.S., Shamira, J.S., Bernard, P., Jeremy, A.S., Pertea, M.A., Jonathan, S., Jeremy, H., Daniel, M., Michael, W., Vaidya, A.B., David, M. A., Fairlamb, A.H., Fraunholz, M.J., Roos, D.S., Ralph, S.A., Mcfadden, G.I., Cummings, L.M., Subramanian, G.M., Mungall, C., Venter, J., Craig, C., Daniel, J., Hoffman, S.L., Newbold, C., Davis, R.W., Fraser, C.M. and Barrell, B. **(2002)**. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*. 419-433.

Geerlof, A., Brown, J., Courtard, B., Egloff, M.P., Enguita, F.J., Fogg, M.J., Gilbert, R.J.C., Groves, M.R., Haouz, A., Nettleship, J.E., Nordlund, P., Owens, R.J., Ruff, M. and Sainsbury, S. **(2005)**. The impact of protein characterization in structural proteomics research papers. *Acta Cryst*. (62):1125-1136.

Gilles, H. M. **(1985)**. The malaria parasites. *In Bruce – chwatt's Essential Malar*. Edward Arnold, London. 12 – 34 pp.

Gluzman IY, Francis SE, Oksman A, Smith CE, Duffin KL, Goldberg DE**.(1994)** Order and specificity of the *Plasmodium falciparum* hemoglobin degradation pathway. *J Clin Invest*;93:1602–1608.

Goldberg, D.E., Slater, A.F.G., Cerami, A. and Henderson, G.B. **(1990)**. Hemoglobin degradation in the malaria parasite *Plasmodium falciparum*: An ordered process in a unique organelle. *Biochem*. 87(04):2931-2935.

Goldberg, B.D.E., Slater, A.F.G., Beavis, R., Cerami, A. and Henderson, G.B. **(1991)**. Hemoglobin Degradation in the Human Malaria Pathogen *Plasmodium falciparum*: A Catabolic Pathway Initiated by a Specific Aspartic Protease. *Exp med*. 173: 961-969.

Golubchik, T., Wise, M.J., Easteal, S. and Jermin, L.S. **(2007)**. Mind the Gaps: Evidence of Bias in Estimates of Multiple Sequence Alignments. *Mol Biol*. 24 (11):2433-2442.

Greenbaum, D.C., Baruch, A. and Grainger, M. **(2002)**. A Role for the Protease Falcipain 1 in Host Cell Invasion by the Human Malaria Parasite. *Sci* 298(12):2002-2006.

Grzonka, Z., Jankowska, E., Kasprzykowski, F., Kasprzykowska, R., Wiczk, W., Wieczerzak, E., Ciarkowski, J., Drabik, P., Janowski, R., Kozak, M., Jaskólski, M. and Grubb, A. **(2001)**. Structural studies of cysteine proteases and their inhibitors . *Rev Lit And Arts Of The Americas*. 48(1):1-20.

Guex, N., Peitsch, M.C. and Schwede, T. **(2003)**. SWISS-MODEL: an automated protein homology-modeling server. *Comp*. 31(13):3381-3385.

Halperin, I., Ma, B., Wolfson, H. and Nussinov, R. **(2002)**. Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions. *In Prac*. 443(1):409 - 443.

Hanspal, M., Dua, M., Takakuwa, Y., Chishti, A.H. and 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.

Hay, S.I., Guerra, C.A., Tatem, A.J., Noor, A.M. and Snow, R.W. **(2004)**. Reviews The global distribution and population at risk of malaria: past, present and future. *The Lanc*. 24(6):327-336.

Henikoff, S. and Henikoff, J.G. **(1992)**. Amino acid substitution matrices from protein blocks. *Biochem*. 89(11):10915-10919.

Hillisch, A., Pineda, L.F. and Hilgenfeld, R. **(2004)**. Utility of homology models in the drug discovery process. *Rev.* 9(15):659-669.

Hoffman, F.M. **(1997)**. An Introduction to Fourier Theory. *Biophysics*. (7):1-9.

Hogg, T., Nagarajan, K., Herzberg, S., Chen, L., Shen, X., Jiang, H., Wecke, M., Blohmke, C., Hilgenfeld, R. and Schmidt, C.L. **(2006)**. Structural and Functional Characterization of Falcipain-2, a Hemoglobinase from the Malarial Parasite *Plasmodium*. *J Biol Chem*. 281(35):25425-25437.

Holt, R.A., Subramanian, G.M., Halpern, A.S., Granger, G.C., Rosane, N., Deborah, R., Wincker, P., Clark, A.G., Ribeiro, M.C., Wides, R., Salzberg, S.L., Loftus, B., Yandell, M., Majoros, W.H., Rusch, D.B., Zhongwu, K., Chyl, L.A., Josep, F., Anthouard, V., Arensburger, P., Atkinson, P.W., Baden, H.B., Veronique, D,B., Danita, B., Vladimir, B., Jim, B., Claudia, B., Randall, B., Didier, L., Jhy-jhu, L., Neil, F.L., John, R., Malek, J.A., Tina, C., Mongin, E., Murphy, S.D., David, A.O., Shao, H., Sharakhova, M.V., Sitter, C.D., Kalush, F.M., Richard, J., Myers, E.W., Adams, M.D., Fraser, C.M., Birney, E., Bork, P. and Brey, P.T. **(2010)**. The Genome Sequence of the Malaria Mosquito *Anopheles gambiae*. 129:129-150.

Huang, N. and Shoichet, B. K. **(2008)** Exploiting Ordered Waters in Molecular Docking. *J. Med. Chem.* 51 (16), 4862–4865.

Hubbard, T.J.P. and Blundell, T.L. **(1987)**. Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modeling. *Prot Eng*. 1(3):159-171.

Hubbard, T.J.P., Murzin, A.G., Brenner, S.E. and Chothia, C. **(1997)**. SCOP: a Structural Classification of Proteins database. *Nucleic Acids Res*. 25(1):236-239.

Huthmacher, C., Hoppe, A., Bulik, S. and Holzhütter, H.G. **(2010)**. Antimalarial drug targets in Plasmodium falciparum predicted by stage-specific metabolic network analysis. *BMC sys biol*. 4:120-147.

Ito T, Tashiro K, Muta S, Ozawa, R., Chiba, T., Nishizawa, M., Yamamoto, K., Kuhara, S. and Sakaki, Y. **(2000)**. Toward a protein – protein interaction map of the budding yeast: A comprehensive system to examine two-hybrid interactions in all possible combinations. *Genetics*. 97(3):1143-1147.

Jacobson, M. and Sali, A. **(2004)**. Comparative Protein Structure Modeling and its Applications to Drug Discovery. *Comp and Gen Pharmacol*. 39(04): 259-276.

James, T.L. **(1998)**. Fundamentals of NMR. *Rev. Series*. (2):1-31.

Jelsch ,C., Teeter, M.M., Lamzin, V.,. Pichon-pesme, V., Blessing, R.H., Lecomte, C. and Poincare, H. **(2000)**. Accurate protein crystallography at ultra-high resolution: Valence electron distribution in crambin. *PNAS*. 97(7):3171-3176.

Jeong, J.J.; Kumar, A.; Hanada, T.; Seo, P.-S.; Li, X.; Hanspal,M. and Chishti, A. H. **(2006)**. Cloning and characterization of *Plasmodium falciparum* cysteine protease, falcipain-2B. *Blood Cells Mol. Dis.* 36(3): 429-435.
Kaapro, A. and Ojanen, J. **(2002)**. Protein docking. *Biophysics*. 11:1-18.

Kappe, S.H.I., Vaughan, A.M., Boddey, J.A. and Cowman, A.F. **(2010)**. That Was Then But This Is Now: Malaria Research in the Time of an Eradication Agenda. *Sci*. 862(1): 1-10.

Katchalski-katzirtt, E., Shariv, I., Eisenstein, M. and Friesem, A.A. **(1992)** Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Biophysics*. 89(3):2195-2199.

Kavanaugh, J.S., Moo-Penn, W.F. and Arnone, A. **(1993)**. Accommodation of insertions in helices: the mutation in hemoglobin Catonsville (Pro 37 alpha-Glu-Thr 38 alpha) generates a 3(10)->alpha bulge. *Biochem.* 32(10):2509-13.

Kempson, G. E.; Muir, H.; Pollard, C. and Tuke, M. **(1973)**. Cysteine proteases *Biochem.Biophys. Acta* 297: 456-463.

Kerr, I.D., Lee, J.H., Farady, C.J., Marlon, R., Rickert, M., Sajid, M., Pandey, K.C., Caffrey, C.R., Legac, J., Hansell, E., McKerrow, J.H., Craik, C.S., Rosenthal, P.J. and Brinen, L.S. **(2009)**. Vinyl Sulfones as Antiparasitic Agents and a Structural Basis for Drug Design *J. Biol Chem*. 284 (38) :25697-25703.

Kerr ID, Lee JH, Pandey KC Harrison, A., Sajid, M., Rosenthal, P.J. and Brinen, L.S. **(2009)**. Structures of Falcipain-2 and Falcipain-3 Bound to Small Molecule Inhibitors : Implications for Substrate Specificity. *J. Med. Chem.* (52):852-857.

Kilama, W.L. **(2003)**. Malaria vaccines in Africa. *Acta Tropica*. 88:153-159.

Klemba, M., Gluzman, I. and Goldberg, D.E. **(2004)**. A Plasmodium falciparum Dipeptidyl Aminopeptidase I Participates in Vacuolar Hemoglobin Degradation *. *Biol chem*. 279(41):43000 -43007.

Krieger, E., Nabuurs, S.B. and Vriend, G. **(2003)**. Homology modeling. In *Structural Bioinformatics,* pp. 507-521. Wiley-Liss, New York.

Laskowski R A, MacArthur M W, Moss D S, Thornton J M **(1993)**. PROCHECK - a program to check the stereochemical quality of protein structures. *J. App. Cryst.*, **26**, 283-291

Laskowski R A, Rullmannn J A, MacArthur M W, Kaptein R, Thornton J M **(1996)**. AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. *J Biomol NMR*, 8, 477-486

Laskowski, R.A., Macarthurt, M.W. and Thornton, J.M. **(1996)**. Validation of protein models derived from experiment. *Biophysic Met.* 631-639.

Laskowski, R.A. **(2003)**. Structural quality assurance. (Weissing PE, Helge BA, eds.). *Struc Bioinfo* 271-301.

Lazaridis, T. and Karplus, M. **(1999)**. Discrimination of the Native from Misfolded Protein Models with an Energy Function Including Implicit Solvation. *J. Mol biol*. 288: 477-487.

Liñares-García, E.G. and Rodriguez, J.B. **(2007)**. Current Status and Progresses Made in Malaria Chemotherapy. *Curr. Med. Chem*. 14:289-314.

Lecaille, F., Kaleta, J. and Bro, D. **(2002)**. Human and Parasitic Papain-Like Cysteine Proteases: Their Role in Physiology and Pathology and Recent Developments in Inhibitor Design. *Chem Rev*. 102:4459-4488.

Lee, B.J., Singh, A., Chiang, P., Kemp, S.J., Goldman, E.A., Weinhouse, M.I., Vlasuk, G.P. and Rosenthal, P.J. **(2003)**. Antimalarial Activities of Novel Synthetic Cysteine Protease Inhibitors. *Soc*. 47(12):3810-3814.

Lee, K. **(2008)**. Computational Study for Protein-Protein Docking Using Global Optimization and Empirical Potentials. *Bioinfo.* 65-77.

Levitt, M. **(1992)**. Accurate Modeling of Protein Conformation by Automatic Segment Matching. *J. Mol. Biol*. 226:507-533.

Lew, V.L., Tiffert, T. and Ginsburg, H. **(2003)**. Excess hemoglobin digestion and the osmotic stability of Plasmodium falciparum – infected red blood cells. *Blood*. 101(10):4189-4194.

Liu, J., Gluzman, I.Y., Drew, M.E. and Goldberg, D.E. **(2005)**. The Role of *Plasmodium falciparum* Food Vacuole Plasmepsins. *Biochem*. 280(2):1432-1437.

Liñares, G.E.G. and Rodriguez, J.B. **(2007)**. Current Status and Progresses Made in Malaria Chemotherapy. *Curr Med Chem*. 14:289-314.

Lundstrom, K. **(2006)**. Structural genomics for membrane proteins. *Cell and Mol Life Sci*. 63:2597-2607.

Lyskov, S. and Gray, J.J. **(2008)**. The RosettaDock server for local protein – protein docking. *Access*. 36(4):233-238.

Maguire, J.D., Marwoto, H., Richie, T.L., Fryauff, D.J. and Baird, J.K. **(2006)**. Mefloquine Is Highly Efficacious against Chloroquine-Resistant *Plasmodium vivax* Malaria and *Plasmodium falciparum* Malaria in Papua , Indonesia. *Clinic infect diseases*. 2197:1067-1072.

Marchler-bauer, A., Panchenko, A., Shoemaker, B., Thiessen, P., Geer, L. and Bryant, S. **(2002)**. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res*. 30:281-283.

Martin, J., Gibrat, J.F. and Rodolphe, F. **(2006)**. Analysis of an optimal hidden Markov model for secondary structure prediction. *BMC Struct Biol.* 13:6-25.

Marti-renom, C.M.A., Fiser, A., Madhusudhan, M.S.,John, B., Stuart, A,, Eswar, N., Pieper, U., Shen, M. and Sali, A. **(2003)**. Modeling Protein structure from its sequence. *Bioinfo*. 1-42.

Marti-renom, M.A., Madhusudhan, M.S. and Sali, A. **(2004)**. Alignment of protein sequences by their profiles. *Prot.Sci*. 14:1071-1087.

Matthew, C.K., van Holde, K.E. and Ahern, K.G. **(2007)**. Biochemistry. *John Wiley & Son*s 3rd ed. pp 1-124. University of Iowa, Iowa City

Mckerrow, J.H. **(1999)**. Development of cysteine protease inhibitors as chemotherapy for parasitic diseases : insights on safety , target validation , and mechanism of action. *Int J. Parasit.* 29:833-837.

Mckerrow, J.H., Engel, J.C. and Ca, C.R. **(1999)**. Cysteine Protease Inhibitors as Chemotherapy for Parasitic Infections. *Bioorganic & Med Chem*. 7:639-644.

McKenzie, F.E., Baird, J. K., Beier, J. C., Lal, A. A., and Bossert, W. H. **(2002)**. A biologic basis for integrated malaria control. *The USA J. Tropic Med and Hygiene.* 67:571–577.

Melo, F., Sanchez, R. and Sali, A. **(2002)**. Statistical potentials for fold assessment. *Prot Sci*. 11: 430-448.

Meno, K., Thorsted, P.B., Ipsen, H., Kristensen, O., Larsen, J.N., Spangfort, M.D., Gajhede, M, and Lund, K. **(2005)**. The Crystal Structure of Recombinant proDer p 1, a MajorHouse Dust Mite Proteolytic Allergen. *J Immunol* 175;3835-3845.

Meyer, M., Wilson, P. and Schomburg, D. **(1996)**. Hydrogen Bonding and Molecular Surface Shape Complementarity as a Basis for Protein Docking. *J. Mol Biol*. 199-210.

Mills, A., Lubell, Y. and Hanson, K. **(2008)**. Malaria eradication: the economic, financial and institutional challenge. *Bio. Med Central*. 7Suppl1:S11-S18.

Momany, F.A. and Rone, R. **(1992)**. Validation of the General Purpose QUANTA3.2 / CHARMm. 13(7):888-900.

Mueller, I.; Galinski, M.R.; Baird, J.K.; Carlton, J.M.; Kochar, D.K.; Alonso, P.L. and Portillo, H.A. **(2009)**. Key gaps in the knowledge of Plasmodium vivax, a neglected human malaria parasite. *Lancet Infect Dis, 9* (9), 555-66.

Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. **(1995)**. SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *Mol. Biol*. 536-540.

Na, B., Shenai, B.R., Sijiwali, P. Puran, S., Choe, Y., Pandey, K.C., Singh, A., Craik, C.S. and Rosenthal, P.J. **(2004)**. Identification and biochemical characterization of vivapains , cysteine proteases of the malaria parasite *Plasmodium vivax*. *Enzy*. 538:529-538.

Norel, R., Lin, S.L., Wolfson, H.J. and Nussinov, R. **(1995)**. Molecular Surface Complementarity at Protein-Protein Interfaces : The Critical Role Played by Surface Normals at Well Placed , Sparse , Points in Docking. *J. Mol. Biol*. 263-273.

Oladele, T.O., Bamigbola, O.M. and Bewaji, C.O. **(2009)**. On efficiency of sequence alignment algorithms. *Scient*. 10(1):9-14.

Omar, M.F., Salam, R.A., Abdullah, R. and Rashid, N.A. **(2005)**. Multiple Sequence Alignments Using Optimization Algorithms. *Comp Intell*. 1(2):81-89.

Pandey, K.C., Sijwali, P.S., Singh, A., Na, B. and Rosenthal, P.J. **(2004)**. Independent Intramolecular Mediators of Folding, Activity, and Inhibition for the *Plasmodium falciparum* Cysteine Protease Falcipain-2. *Biochem*. 279(5):3484 -3491

Pandey, K.C., Wang, S.X., Sijwali, P.S., McKerrow, J.H. and Rosenthal, P.J. **(2005)**. The *Plasmodium falciparum* cysteine protease falcipain-2 captures its substrate , hemoglobin , via a unique motif. *PNAS*. 102(26).

Pandey, K.C., Barkan, D.T., Sali, A. and 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):1-9.

Parker, W. **(2003)**. Protein Structure from X-Ray Diffraction. *J. Biol. Physics.* (29):341-362.

Pawlowski, M., Gajda, M.J., Matlak, R. and Bujnicki, J.M. **(2008)**. MetaMQAP: A meta-server for the quality assessment of protein models. *BMC Bioinformatics*. 20:1-20.

Pearson, W.R. and Lipman, D.J. **(1988)**. Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*. 85:2444-2448.

Pei, J., Kim, B. and Grishin, N.V. **(2008)**. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Access*. 36(7):2295-2300.

Pierce, B. and Weng, Z. **(2007)**. ZRANK: Reranking Protein Docking Predictions With an Optimized Energy Function. *Bioinfo*. 1086(10):1078-1086.

Pongsumpun, P. and Tang, I.M. **(2008)**. *Plasmodium Vivax* Malaria Transmission in a Network of Villages. *Eng and Tech*. 333-337.

Ponting, C.P., Schultz, J., Milpetz, F. and Bork, P. **(1999)**. SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res*. 24:229-232.

Ramjee, M.K., Flinn, N.S., Pemberton, T.P., Quibell, M., Wang, Y. And Watts, J.P. **(2006)**. Substrate mapping and inhibitor profiling of falcipain-2, falcipain-3 and berghepain-2: implications for peptidase anti-malarial drug discovery *Biochem J.* 1: 1-29.

Rawlings, N.D. and Barrett, A.J. **(1993)**. Evolutionary families of peptidases. *Soc*. 218:205-218.

Rawlings N.D., Tolle, D.P. and Barrett, A.J. **(2004)**. Evolutionary families of peptidase inhibitors. *Soc* 716:705-716.

Redzynia, I., Ljunggten, A., Abrahamson, M., Mort, J.S., Krupa, J.C., Jaskolsi, M. and Bujacz, G. **(2009)**. Displacement of the occluding loop by parasite protein chagasin results in efficient inhibition of human cathepsin B. *Biol. Chem.* 283:22815-22823.

Rosenthal, P.J., Mckerrow, J,H., Aikawa, M., Nagasawa, H. and Leech, J.H. **(1988)**. A Malarial Cysteine Proteinase Is Necessary for Hemoglobin Degradation by *Plasmodium falciparum*. *J. Clin.Invest.* 82(11):1560-1566.

Rosenthal, P.J., Wollish, W.S., Palmer, J.T. and Rasnick, D. **(1991)**. Antimalarial Effects of Peptide Inhibitors of a *Plasmodium falciparum* Cysteine Proteinase. *J.Clin.Invest*. 88(11):1467-1472.

Rosenthal, P.J. and Nelson, R.G. **(1992)**. Isolation and characterization of a cysteine protease gene of *Plasmodium falciparum*. *Parasite*. 51:143-152.

Rosenthal, P.J., Lee, G.K. and Smitht, R.E. **(1993)**. Inhibition of a *Plasmodium vinckei* Cysteine Proteinase Cures Murine Malaria. *J.Clin.Invest*. 1052-1056.

Rosenthal, P.J., Oslon, J.E., Lee, G.K., Palmer, J.T., Klaus, J.L. and Rasnick, D. **(1996)**. Antimalarial Effects of Vinyl Sulfone Cysteine Proteinase Inhibitors. *Microbiol*. 40(7):1600-1603.

Rosenthal, P.J. **(1998)**. Proteases of Malaria Parasites: New Targets for Chemotherapy. *Emerg Infect Dis*. 4(1):49-57.

Rosenthal, P.J., Sijwali, P.S., Singh, A. and Shenai, B.R. **(2002)**. Cysteine Proteases of Malaria Parasites: Targets for Chemotherapy. *Curr.* 1659-1672.

Rosenthal PJ. **(2003)**. Review Antimalarial drug discovery: old and new approaches. *Journal of Experimental Biology*. 3735-3744.

Rosenthal, P.J., Press, H. and Chiodini, P.L. **(2003)**. Antimalarial chemotherapy.Mechanism of action, resistance and new directions in drug discovery. *J. Antimicrob. Chemo*. 1053: 3736-3741.

Rosenthal, P.J. **(2004)**. Cysteine proteases of malaria parasites. *Int. J.Parasit.* 34:1489-1499.

Rost, B. **(1999)**. Twilight zone of protein sequence alignments. *Prot. Eng.* 12:85-94.

Sabnis, Y.A., Desai, P.V., Rosenthal, P.J. and Avery, M.A. **(2003)**. Probing the structure of falcipain-3, a cysteine protease from *Plasmodium falciparum*. *Prot Sci*. 12:501-509.

Schechter, I. and Berger, A. **(1967)**. On the size of the active site in proteases I. Papain. *Biochem. Biophys. Res. Commun*. 27:157-167

Schilling, K., Pietschmann, S., Fehn, M., Wenz, I. and Wiederanders, B. **(2001)**. Folding incompetence of cathepsin L-like cysteine proteases may be compensated by the highly conserved, domain-building N-terminal extension of the proregion. *Biol. Chem.* 382(5):859-65.

Shah, F., Mukherjee, P., Gut, J., Legac, J., Rosenthal, P.J., Tekwani, B.L., and Avery, M.A. (2011) Identification of Novel Malarial Cysteine Protease Inhibitors Using Structure-Based Virtual Screening of a Focused Cysteine Protease Inhibitor Library. *J. Chem. Inf. Model*. 51, 852–864

Sherman, I. W., and L. Tanigoshi. **(1970)**. Incorporation of '4Camino-acids by malaria (Plasmodium lophurae). IV. In vivo utilization of host cell hemoglobin. *Int. J. Biochem*. 1:635-637 **.**
Sherman, I. 1979. Biochemistry of P*lasmodium* . *Microbiol Rev*.43:453 .

Shindo, T and Van Der Hoorn, R.A.L. **(2008)**. Papain-like cysteine proteases : key players at molecular battlefields employed by both plants and their invaders. *Mol. Plant Path.*. 9:119-125.

Schofield, L., and Grau, G. E. **(2005).** Immunological processes in malaria pathogenesis. *Nature reviews,* 5; 722-735.

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

Sali, A. and Blundell T. **(1993)**. Comparative Protein Modeling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* 234:779-815.

Sanchez, R. and Sali, A. **(1997)**. Advances in comparative protein-structure modeling. *Curr. Opin. Struct. Biol*. 206-214.

Shen, M. and Sali, A. **(2006)**. Statistical potential for assessment and prediction of protein structures Statistical potential for assessment and prediction of protein structures. *Prot. Sci*. 2507-2524.

Shenai, B.R., Sijwali, P.S., Singh, A. and Rosenthal, P.J. **(2000)**. Characterization of Native and Recombinant Falcipain-2, a Principal Trophozoite Cysteine Protease and Essential Hemoglobinase of Plasmodium falciparum . *Biochem*. 275(37):29000 -29010.

Sijiwali, P.S., Shenai, B.R., Gut, J., Singh, A. and Rosenthal, P.J. **(2001)**. Expression and characterization of *Plasmodium falciparum* hemoglobinase falcipain-3. *Biochem J*. 489:481-489.

Sijwali, P.S., Shenai, B.R. and Rosenthal, P.J. **(2002)**. Folding of the *Plasmodium falciparum* Cysteine Protease Falcipain-2 Is Mediated by a Chaperone-like Peptide and Not the Prodomain. *Biochem.* 277(17):14910 -14915.

Sijwali, P.S., Kato, K., Seydel, K.B., Gut, J., Lehman, J., Klemba, M., Goldberg, D.E., Miller, L.H. and Rosenthal, P.J. **(2004)**. *Plasmodium falciparum* cysteine protease falcipain-1 is not essential in erythrocytic stage malaria parasites. *PNAS*. 101(23):8721-8726.

Sijwali, P.S. and Rosenthal, P.J. **(2004)**. Gene disruption confirms a critical role for the cysteine protease falcipain-2 in hemoglobin hydrolysis by Plasmodium falciparum. *PNAS*. 101(13):4384-4389.

Singh, A, and Rosenthal, P.J. **(2001)**. Comparison of Efficacies of Cysteine Protease Inhibitors against Five Strains of *Plasmodium falciparum*. *Soc*. 45(3):949-951

Singh, A. and Rosenthal, P.J. **(2004)**. Selection of Cysteine Protease Inhibitor-resistant Malaria Parasites Is Accompanied by Amplification of Falcipain Genes and Alteration in Inhibitor Transport . *Biochem*. 279(34):35236 -35241.

Singh, N., Sijwali, P.S., Pandey, K.C. and Rosenthal, P.J. **(2006)**. *Plasmodium falciparum*: Biochemical characterization of the cysteine protease falcipain-2 .*Exp. Parasit.*. 112:187-192.

Sternberg, M.J., Gabb, H.A. and Jackson, R.M. **(1998)**. Predictive docking of protein-protein and protein-DNA complexes. *Curr. Opin. Struct. Biol.* 8:250-6.

Sippl, M.J. **(1993)**. Recognition of Errors in Three-Dimensional Structures of Proteins. *Prot*. 17:355-362.

Soding, J., Biegert, A. and Lupas, A.N. **(2005)**. The HHpred interactive server for protein homology detection and structure prediction. *Comp*. 33:244-248.

Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A. and Durbin, R. **(1998)**. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acid Res*. 26(1):320-322.

Stucliffe, M.J., Haneef, I., Carney, D. and Blundell, T.L. **(1987)**. Knowledge based modeling of homologous proteins, Part I: Three dimensional frameworks derived from simultaneous superposition of multiple structures. *Prot. Eng*. 1:377-384.

Subramanian, A.R., Weyer-menkhoff, J. and Kaufmann, M. **(2005)**. Morgenstern B. DIALIGN-T: An improved algorithm for segment-based multiple sequence alignment. *BMC Bioinformatics*. 13: 1-13.

Subramanian, S., Hardt, M., Choe, Y., Legac, J., Gut, J., Kerr, I.D., Craik, C.S. and Rosenthal, P.J. **(2009)**. Hemoglobin Cleavage Site-Specificity of the *Plasmodium falciparum* Cysteine Proteases Falcipain-2 and Falcipain-3. *Plus one*. 4(4): e5156.

Theakston, R. D. G., S. A. Fletcher, and B. G. Maegraith. **(1970)**. The use of electron microscope autoradiography for examining the uptake and degradation of haemoglobin by Plasmodium berghei. *Ann.Trop. Med. Parasitol*. 64:63-71 **.**

Tamm, L.K. and Liang, B. **(2006)**. NMR of membrane proteins in solution. *Prog Nucl Magn Res Spect*. 48:201-210.

Turk, D., Guncar, G., Podobnik, M. and Turk, B. **(1998)**. Revised definition of substrate binding sites of papain-like cysteine proteases. *Biol Chem.* 379(2):137-47.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., Smirnov, S., Sverdlov,, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J. and Natale, D.A. **(2003)**. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*. 14:1-14.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. **(1994)**. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. 22(22):4673-4680.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. **(1997)**. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*. 25(24):4876-4882.

Tina, K.G., Bhadra, R. and Srinivasan, N. **(2007)**. PIC: Protein Interactions Calculator. *Nucleic Acids Res.* 5:473-476.

Trongtokit, Y., Curtis, C.F. and Rongsriyam, Y. **(2005)**. Efficancy of repellent products against caged and free flying anopheles stephensi mosquitoes.*Eng*. 36(6):1423-1431.

Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emill, A., Brian Godwin, Y.L., Conover, D., Kalbfleisch, T., Vijadamodar, G., Yang, M., Johnston, M, Fields, S. and Rothberg, J.M. **(2000)**. A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature*. 403(2): 623-630.

Vakser, I.A., Matar, O.G., Lam, C.F., **(1999)**. A systematic study of low-resolution recognition in protein-protein complexes, *Proc. Natl. Acad. Sci. USA* 96:8477-8482.

Voet, D. and Voet, J.G. **(2006)**. Fundamentals of Biochemistry *John Wiley & Son*s 2nd ed. pp 124-159.

Wallner, B. and Elofsson, A. **(2005)**. All are not equal: a benchmark of different homology modeling programs. *Prot Sci*. 14:1315-1327.

Wang, S.X., Pandey, K.C., Scharfstein, J., Whisstock, J., Huang, R.K., Jacobelli, J., Fletterick, R.J., Rosenthal, P.J., Abrahamson, M., Brinen, L.S., Rossi, A., Sali, A. and Mckerrow, J.H. **(2006)**. Structural basis for unique mechanisms of folding and hemoglobin binding by a malarial protease. *PNAS*. 103(31):11503-11508.

Wang, S.X., Pandey, K.C., Scharfstein, J. Whisstock, J., Huang, R.K., Jacobelli, J., Fletterick, R.J, Rosenthal, P.J., Abrahamson, M., Brinen, L.S., Rossi, A., Sali, A. and Mckerrow, J.H. **(2007)**. The Structure of Chagasin in Complex with a Cysteine Protease Clarifies the Binding Mode and Evolution of an Inhibitor Family. *Struct*. 5:535-543.

Wegscheid-gerlach, C., Gerber, H. and Diederich, W.E. **(2010)**. Proteases of *Plasmodium falciparum* as Potential Drug Targets and Inhibitors Thereof. *Curr.topics in med. chem.* 346-367.

Wellems, T.E., Hayton, K. and Fairhurst, R.M. **(2009)**. Review series The impact of malaria parasitism : from corpuscles to communities. *Rev. Series*. 119(9):2496-2506.

Weng, Z. and Delisi, C. **(2002)**. Protein therapeutics: promises and challenges for the 21st century. *Trends in biotech*. 20(1):29-35.

Wiederanders, B. **(2000)**. The function of propeptide domains of cysteine proteinase *Adv. Exp. Med. Biol.* 477:261-269.

Wiederstein, W. and Sippl M.J. (2007) ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins.*Nucleic Acids Research* 35, W407-W410.

Yamamoto, Y., Watabe, S., Kageyama, T. and Takahashi, S. Y. **(1999)**. *Bombyx* acid cysteine proteases (BCP): Hormonal regulation of biosynthesis and accumulation in the ovary. *Arch.Insect Biochem. Physiol.* 46:783-791.

Xiang, Z. **(2006)**. Advances in Homology Protein Structure Modeling. *Curr. Prot. Pep. Sci*. 217-227.

Xiong, J. **(2006).** Protein tertiary structure prediction. In *Essential Bioinformatics,* pp. 214-230. Cambridge University Press, Cambridge

.

Zarchin, S., M. Krughak, and H. Ginsburg . **(1986)** . Digestion of the host erythrocyte by malaria parasites is the primary target for quinolone-containing antimalarials. Biochem . Pharmacol.35 :2435 .

Zhang, C., Vasmatzis, G., Cornette, J.L. and Delisi, C. **(1997)**. Determination of Atomic Desolvation Energies From the Structures of Crystallized Proteins. *J. Mol. Biol.* 267:707-726

Appendix A

Supplimentary data for chapter 2

Clustalw2 protein colouring:

RED (Small, hydrophobic amino acids)	ALA, VAL, PHE, PRO, MET, ILE, LEU and TRP
BLUE (Acidic amino acids)	ASP and GLU
MAGENTA (Basic amino acids)	ARG and LYS
GREEN (Hydroxyl, amine and basic aa)	SER, THR, TYR, HIS, CYS, ASN, GLY and GLN
GRAY	Others

T-coffee protein colouring:

Bad 

Average 

Good 

Bio-edit coloring:

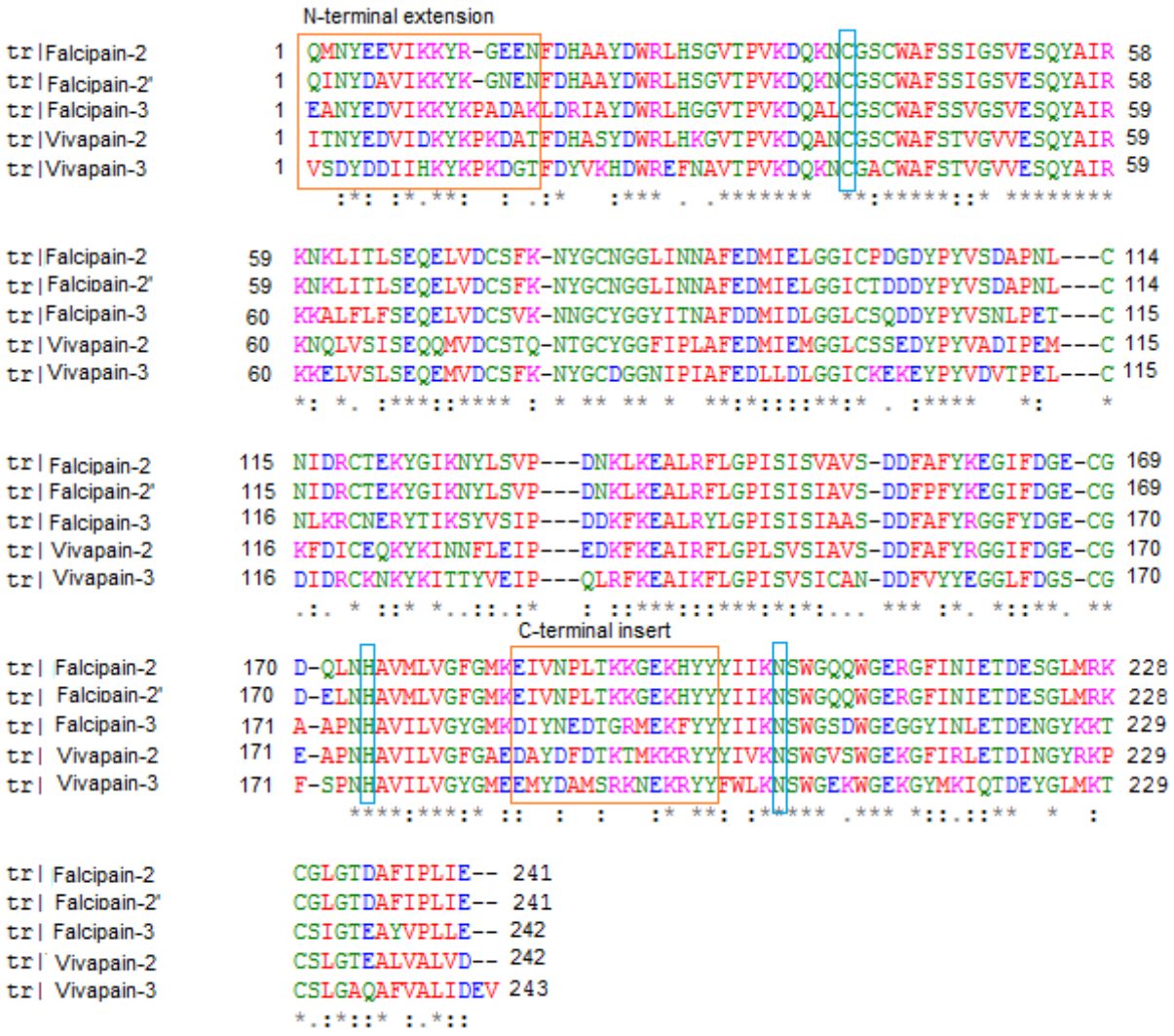

Figure A. 1: ClustalW2 alignment of the target proteases (FP2' and VP2, VP3) and their templates. The nose-like and arm-like motifs are highlighted in an orange rectangle. Active site residues: CYS, HIS and ASN are highlighted in blue.

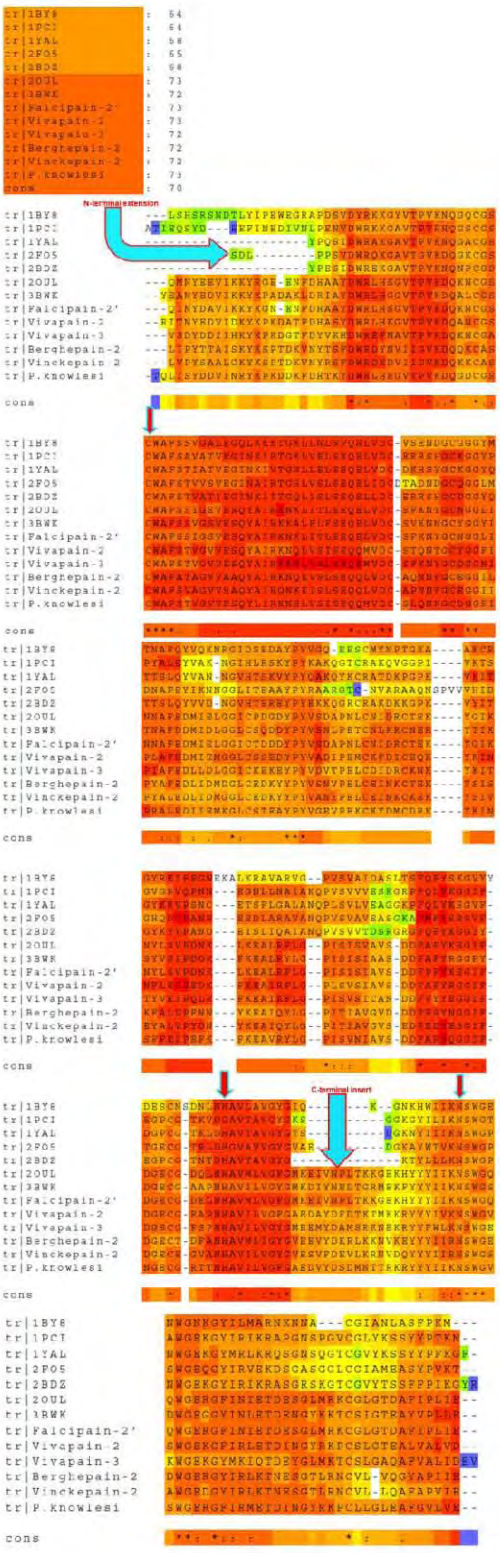

Figure A. 2: Multiple sequence alignment, indicating all sequences that were included for studying the relationship between CPs. N-terminal extension and C-terminal insert marked by cyan arrows and active site residues C, H and N are marked with red. Default colour settings from T-COFEE indicating the level of sequence conservation.

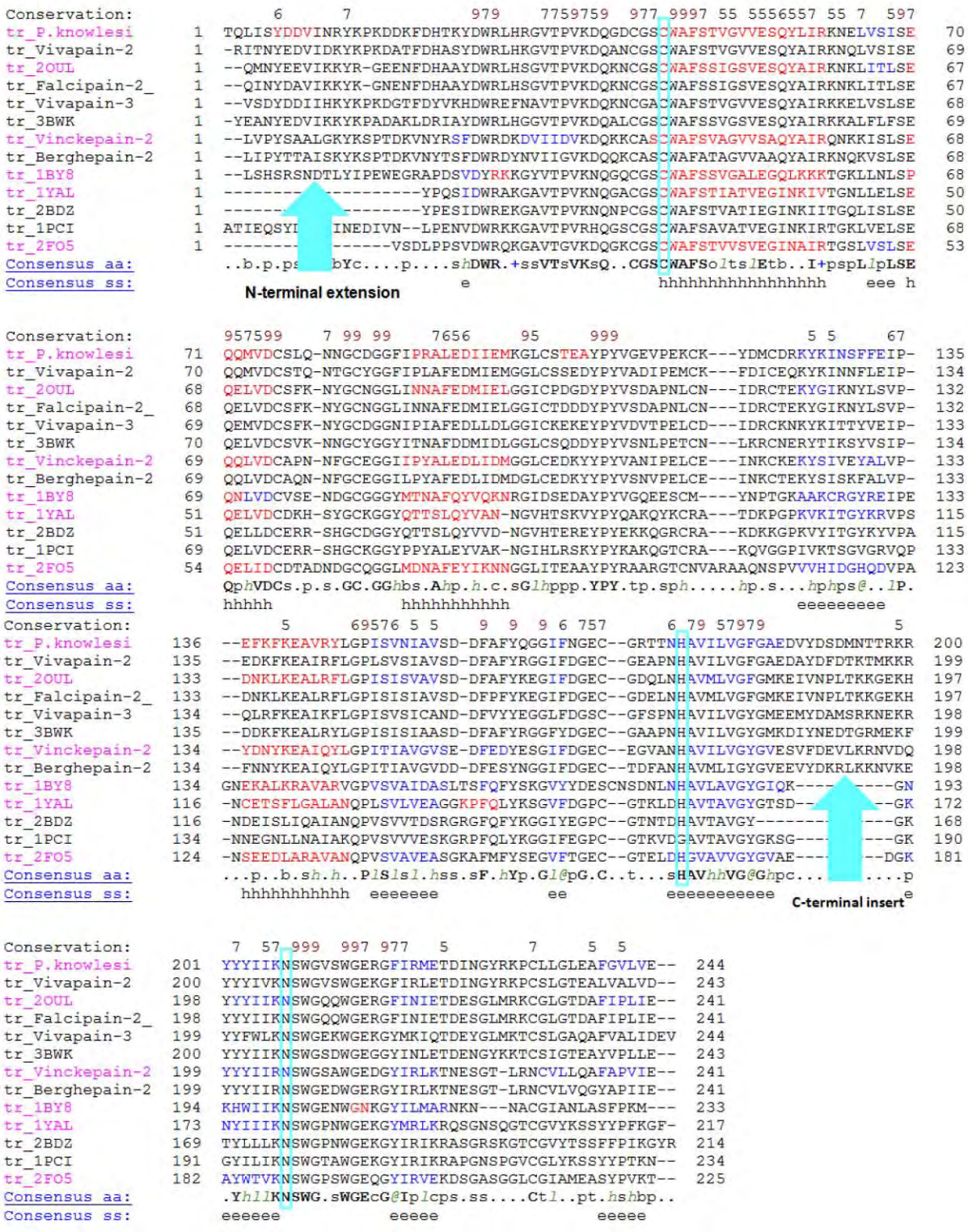

Figure A. 3: PROMALS3D alignment with 9 Plasmodium CPs and 6 other structures of CPs from other organisms, the level of conservation, consensus amino acids and secondary structures are indicated. The highly conserved active site, arm-like and nose-like motif labeled as N-terminal extension and C-terminal insert are marked in cyan.

Model building scripts and alignment files

Falcipain-2'

The following alignment file and modeller scripts were used to built the model of FP2'

A

```
>P1;falcipain2
sequence:falcipain2:  1: : 241: :: : 0.00: 0.00
QINYDAVIKKYKGNENFDHAAYDWRLHSGVTPVKDQKNCGSCWAFSSIGSVESQYAIRKNKLITLSEQELDCSFKNYGCNG
GLINNAFEDMIELGGICTDDDYPYVSDAPNLCNIDRCTEKYGIKNYLSVPDNKLKEALRFLGPISISIAVSDDFPFYKEGIFDGE
CGDELNHAVMLVGFGMKEIVNPLTKKGEKHYYYIIKNSWGQQWGERGFINIETDESGLMRKCGLGTDAFIPLIE*
>P1;2oul.pdb
structureX:2oul.pdb: -16:A:224 :A:undefined: undefined: -1.00: -1.00
QMNYEEVIKKYRGEENFDHAAYDWRLHSGVTPVKDQKNCGSCWAFSSIGSVESQYAIRKNKLITLSEQELVDCSFKNYGC
NGGLINNAFEDMIELGGICPDGDYPYVSDAPNLCNIDRCTEKYGIKNYLSVPDNKLKEALRFLGPISISVAVSDDFAFYKEGIF
DGECGDQLNHAVMLVGFGMKEIVNPLTKKGEKHYYYIIKNSWGQQWGERGFINIETDESGLMRKCGLGTDAFIPLIE*
```

B

```
# Homology modelling by the automodel class

from modeller.automodel import *    # Load the automodel class

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

# directories for input atom files
env.io.atom_files_directory =
'\Users\fortunate\Documents\Write_up_2010\July.Aug.2010\Chapter2\Aug2010_finalmodelfiles\Vivapain2\Twotemplates'

a = automodel(env,
        alnfile  = 'vp2.pir',     # alignment filename
        knowns   = ('2oul.pdb', '3bwk.pdb'),   # codes of the templates
        sequence = 'vivapain2', assess_methods=(assess.DOPE, assess.GA341))   # code of the target
a.starting_model= 1                 # index of the first model
a.ending_model  = 100               # index of the last model
                    # (determines how many models to calculate)
a.final_malign3d = True             # generate superimposed templates and model (*_fit.pdb files)
a.make()                    # do the actual homology modelling
ok_models = filter(lambda x: x['failure'] is None, a.outputs) # Get a list of all successfully built models from a.outputs
```

Script 1: Alignment file and script used for modeling FP2' labeled A and B respectively. MODELLER code written in python used for building 100 models of FP2' adopted from MODELLER manuals (Šali and Blundell, 1993) and the alignment generated by CLUSTALX2.

Vivapain-2

The alignment file and the model script used for building the model of VP2 contained the following co-ordinates:

```
>P1;vivapain2                                                    A
sequence:vivapain2:  1: : 228: :: : 0.00: 0.00
DATFDHASYDWRLHKGVTPVKDQANCGSCWAFSTVGVVESQYAIRKNQLVSISEQQMVDCSTQNTGCYGGFIPLAFEDMIE
MGGLCSSEDYPYVADIPEMCKFDICEQKYKINNFLEIPEDKFKEAIRFLGPLSVSIAVSDDFAFYRGGIFDGECGEAPNHAVILVG
FGAEDAYDFDTKTMKKRYYYIVKNSWGVSWGEKGFIRLETDINGYRKPCSLGTEALVALVD*
>P1;2oul.pdb
structureX:2oul.pdb: -3:A: 224:A:undefined: undefined: -1.00: -1.00
EENFDIIAAYDWRLISGVTPVKDQKNCGSCWAFSSIGSVESQYAIRKNKLITLSEQELVDCSTKNYGCNGGLINNAFEDMIELG
GICPDGDYPYVSDAPNLCNIDRCTEKYGIKNYLSVPDNKLKEALRFLGPISISVAVSDDFAFYKEGIFDGECGDQLNHAVMLVGF
GMKEIVNPLTKKGEKHYYYIIKNSWGQQWGERGFINIETDESGLMRKCGLGTDAFIPLIE*
>P1;3bwk.pdb
structureX:3bwk.pdb: 23:A: 249:A: undefined: undefined: -1.00: -1.00
DAKLDRIAYDWRLHGGVTPVKDQALCGSCWAFSSVGSVESQYAIRKKALFLFSEQELVDCSVKNNGCYGGYITNAFDDMIDL
GGLCSQDDYPYVSN_PETCNLKRCNERYTIKSYVSIPDDKFKEALRYLGPISISIAASDDFAFYRGGFYDGECGAAPNHAVILVGY
GMKDIYNEDTGRMIEKFYYYIIKNSWGSDWGEGGYINLETDENGYKKTCSIGTEAYVPLL-*

                                                                 B
# Homology modelling by the automodel class

from modeller.automodel import *    # Load the automodel class

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

# directories for input atom files
env.io.atom_files_directory =
'\Users\fortunate\Documents\Write_up_2010\July.Aug.2010\Chapter2\Aug2010_finalmodelfiles\Vivapain2\Twotemplates'

a = automodel(env,
        alnfile  = 'vp2.pir',     # alignment filename
        knowns   = ('2oul.pdb', '3bwk.pdb'),   # codes of the templates
        sequence = 'vivapain2',assess_methods=(assess.DOPE, assess.GA341))   # code of the target
a.starting_model= 1                 # index of the first model
a.ending_model = 100                # index of the last model
                    # (determines how many models to calculate)
a.final_malign3d = True             # generate superimposed templatesand model (*_fit.pdb files)
a.make()                    # do the actual homology modelling
ok_models = filter(lambda x: x['failure'] is None, a.outputs) # Get a list of all successfully built models from a.outputs
key = 'DOPE score' # Rank the models by DOPE score
```

Script 2: A: VP2 alignment against template structures FP2 and FP3 which were used for model generation, the pir was generated in CLUSTALX2. B. shows the MODELLER code written in python used for building 100 models of VP2 adopted from MODELLER manuals (Šali and Blundell, 1993

Vivapain-3

A

```
>P1;vivapain3
sequence:vivapain3:  1: : 242: :: : 0.00: 0.00
SDYDDIIHKYKPKDGTFDYVKHDWREFNAVTPVKDQKNCGACWAFSTVGVVESQYAIRKKELVSLSEQEMVDCSFKNYGCDGGN
IPIAFEDLLDLGGICKEKEYPYVDVTPELCDIDRCKNKYKITTYVEIPQLRFKEAIKFLGPISVSICANDDFVYYEGGLFDGSCGFSPNHA
VILVGYGMEEMYDAMSRKNEKRYYFWLKNSWGEKWGEKGYMKIQTDEYGLMKTCSLGAQAFVALIDE*
>P1;2oul.pdb
structureX:2oul.pdb:-14 :A: 224:A:undefined: undefined: -1.00: -1.00
-NYEEVIKKYRGEENFDIIAAYDWRLIISGVTPVKDQKNCGSCWAFSSIGSVESQYAIRKNKLITLSEQELVDCSFKNYGCNGGLINN
AFEDMIELGGICPDGDYPYVSDAPNLCNIDRCTEKYGIKNYLSVPDNKLKEALRFLGPISISVAVSDDFAFYKEGIFDGECGDQLNHA
VMIVGFGMKFIVNPITKKGFKHYYYIIKNSWGQQWGFRGFINIFTDFSGLMRKCGLGTDAFIPLIF-*
>P1;3bwk.pdb
structureX:3bwk.pdb: 10:A: 249:A:undefined: undefined: -1.00: -1.00
ANYEDVIKKYKPADAKLDRIAYDWRLHGGVTPVKDQALCGSCWAFSSVGSVESQYAIRKKALFLFSEQELVDCSVKNNGCYGGYIT
NAFDDMIDLGGLCSQDDYPYVSNLPETCNLKRCNERYTIKSYVSIPDDKFKEALRYLGPISISIAASDDFAFYRGGFYDGECGAAPNH
AVILVGYGMKDIYNEDTGRMEKFYYYIIKNSWGSDWGEGGYINLETDENGYKKTCSIGTEAYVPLL--*
```

B

```
# Homology modelling by the automodel class

from modeller.automodel import *    # Load the automodel class

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

# directories for input atom files
env.io.atom_files_directory =
'\Users\fortunate\Documents\Write_up_2010\July.Aug.2010\Chapter2\Aug2010_finalmodelfiles\Vivapain3'

a = automodel(env,
        alnfile = 'vp3.pir',    # alignment filename
        knowns  = ('2oul.pdb', '3bwk.pdb'),   # codes of the templates
        sequence = 'vivapain3',assess_methods=(assess.DOPE, assess.GA341))   # code of the target
a.starting_model= 1             # index of the first model
a.ending_model  = 100           # index of the last model
                        # (determines how many models to calculate)
a.final_malign3d = True         # generate superimposed templates and model (*_fit.pdb files)
a.make()                    # do the actual homology modelling
ok_models = filter(lambda x: x['failure'] is None, a.outputs) # Get a list of all successfully built models from
a.outputs
```

Script 3: A is the alignment file containing the co-ordinates of the sequence and structures to be used for the model construction. B MODELLER code written in python used for building 100 models of VP3, adopted from MODELLER manuals (Šali and Blundell, 1993)

```
>P1;procathepsink
sequence:procathepsink:  1: : 310: :: : 0.00: 0.00
EILDTHWELWKKTHRKQYNNKVDEISRRLIWEKNLKYISIHNLEASLGVHTYELAMNHLGDMTSEEVVQKMTGLKVPLSHSRSNDTLYIPEWE
GRAPDSVDYRKKGYVTPVKNQGQCGSCWAFSSVGALEGQLKKKTGKLLNLSPQNLVDCVSENDGCGGGYMTNAFQYVQKNRGIDSEDAYP
YVGQEESCMYNPTGKAAKCRGYREIPEGNEKALKRAVARVGPVSVAIDASLTSFQFYSKGVYYDESCNSDNLNHAVLAVGYGIQKGNKHWIIK
NSWGENWGNKGYILMARNKNNACGIANLASFPKM*
>P1;1XKG.pdb
structureX:1XKG.pdb: 6:A:302 :A:undefined: undefined: -1.00:
KTFEEYKKAFNKSYATFEDEEAARKNFLESVKYVQSNGGAINHLSDLSLDEFKNRFLMSAEAFEHLKTQFDNACSINGNAPAEIDLRQMRTVTPI
RMQGGCGSAWAFSGVAATESAYLAYRDQSLDLAEQELVDCASQHGCHGDTIPRGIEYIQHNGVVQESYYRYVAREQSCRRPNAQRFGISNYC
QIYPPNANKIREALAQTHSAIAVIIGIKDLDAFRHYDGRTIIQRDNGYQPNYHAVNIVGYSNAQGVDYWIVRNSWDTNWGDNGYGYFAANID
LMMIEEYPYVVIL*
```

```
Homology modelling by the automodel class                    B

from modeller.automodel import *    # Load the automodel class

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

# directories for input atom files
env.io.atom_files_directory =
'\Users\fortunate\Documents\MSc\All_data\Corrections\modelling\1by8\procath_1by8'

a = automodel(env,
        alnfile  = 'alignmentfile.pir',    # alignment filename
        knowns   = ('1XKG.pdb'),    # codes of the templates
        sequence = 'procathepsink', assess_methods=(assess.DOPE, assess.GA341))   # code of the target
a.starting_model= 1                 # index of the first model
a.ending_model  = 100               # index of the last model
                    # (determines how many models to calculate)
a.final_malign3d = True             # generate superimposed templatesand model (*_fit.pdb files)
a.make()                    # do the actual homology modelling
ok_models = filter(lambda x: x['failure'] is None, a.outputs) # Get a list of all successfully built models
from a.outputs
```

Script 4: Alignment file and script used for modeling procathepsin K labeled A and B respectively. MODELLER code written in python used for building 100 models of procathepsin K adopted from MODELLER manuals (Šali and Blundell, 1993) and the alignment generated by CLUSTALX2

Model Evaluation

```
import sys

filename = sys.argv[1]                    # Example for: model.assess_normalized_dope()
from modeller import *
from modeller.scripts import complete_pdb
env = environ()
env.libs.topology.read(file='$(LIB)/top_heav.lib')
env.libs.parameters.read(file='$(LIB)/par.lib')   # Read a model previously generated by Modeller's
automodel class
mdl = complete_pdb(env, filename)
zscore = mdl.assess_normalized_dope()
wfile = open("zdope_scores.txt","a")
wfile.write(str(zscore)+" "+filename+"\n")
wfile.close()
```

Script 4: The DOPE z-score of each model was calculated using this code which was written in python programming language adopted from MODELLER manuals (Šali and Blundell, 1993)

```
import subprocess

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

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

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

Script 5: Code for calculating the DOPE z-score of each model, written in python programming language by Matthys Kroon, 2010

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

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

Script 6: Program also written in python language by Matthys Kroon for sorting the DOPE z-scores based from the lowest to the highest scores.

Appendix B

Supplimentary Data for chapter 3

Table B 1: ZDOCK scores for FP2 (1YVB)-hemoglobin complex structures

Docked Pose no:	Complex	Cluster centre	Best Energy Rep	ZDOCK	ZRANK
Cluster 8					
Pose 14	Complex 1	Yes	1	10.02	18.32
Pose 49	Complex 2	No	0	14.37	30.33
Pose 52	Complex 3	No	0	11.70	32.22
Pose 61	Complex 4	No	0	15.60	41.52
Cluster 6					
Pose 3	Complex 5	Yes	1	17.80	-78.42
Pose 9	Complex 6	No	0	18.80	-60.28
Pose 12	Complex 7	No	0	18.24	-57.09
Pose 15	Complex 8	No	0	17.68	-46.80
Pose 18	Complex 9	Yes	1	18.52	-32.35
Pose 19	Complex 10	Yes	1	19.24	-27.46
Pose 26	Complex 11	Yes	1	17.44	-16.88
Pose 30	Complex 12	Yes	1	16.98	-13.91
Pose 34	Complex 13	Yes	1	17.58	-7.52
Cluster 1					
Pose 1	Complex 14	Yes	1	23.54	-106.77
Pose 3	Complex 15	Yes	1	24.58	-102.30
Pose 4	Complex 16	Yes	1	21.05	-100.85
Pose 5	Complex 17	Yes	1	20.04	-95.81
Pose 6	Complex 18	No	0	21.58	-95.80
Pose 8	Complex 19	No	0	22.10	-94.14
Pose 10	Complex 20	Yes	1	18.40	-91.93
Pose 12	Complex 21	No	0	23.54	-90.42
Pose 17	Complex 22	Yes	1	20.62	-88.58
Pose 20	Complex 23	Yes	1	21.16	-88.24
Pose 21	Complex 24	Yes	1	19.20	-87.38
Pose 22	Complex 25	Yes	1	19.94	-66.18
Pose 56	Complex 26	Yes	1	20.00	-65.29
Pose 58	Complex 27	Yes	1	17.84	-62.63
Pose 64	Complex 28	Yes	1	21.38	-60.94
Pose 70	Complex 30	No	0	18.30	-60.14
Pose 73	Complex 31	No	0	17.92	-59.89
Pose 76	Complex 32	No	0	20.84	-56.73
Pose 89	Complex 33	Yes	1	21.00	-55.37
Pose 93	Complex 34	No	0	21.12	-55.21
Pose 94	Complex 35	Yes	1	19.82	-54.81
Pose 95	Complex 36	Yes	1	17.48	-54.37

Table B 2: ZDOCK scores for FP2 (2OUL)-hemoglobin complex structures

Docked Pose no:	Complex	Cluster centre	Best Energy Rep	ZDOCK	ZRANK
Cluster 5					
Pose 10	Complex 1	Yes	1	14.02	22.65
Pose 14	Complex 2	No	0	14.92	37.82
Pose 26	Complex 3	No	0	13.32	108.22
Cluster 3					
Pose 1	Complex 4	No	0	16.92	-69.02
Pose 2	Complex 5	No	0	16.78	-68.13
Pose 3	Complex 6	Yes	1	16.42	-67.85
Pose 6	Complex 7	No	0	16.64	-61.87
Pose 11	Complex 8	Yes	1	16.54	-55,87
Pose 29	Complex 9	No	0	16.50	-51.54
Pose 69	Complex 10	Yes	1	16.48	-41.56
Pose 73	Complex 11	Yes	1	16.36	-40.64
Pose 104	Complex 12	No	0	16.34	-32.76
Pose 105	Complex 13	No	0	16.38	-31.92
Pose 114	Complex 14	Yes	1	16.72	-28.03
Pose 129	Complex 15	Yes	1	16.56	-22.08
Pose 160	Complex 16	Yes	1	16.46	-17.72
Pose 166	Complex 17	Yes	1	16.86	-10.23
Pose 178	Complex 18	Yes	1	16.70	-5.67
Cluster 1					
Pose 19	Complex 19	Yes	1	23.10	-93.88
Pose 58	Complex 20	Yes	1	20.48	-75.82
Pose 62	Complex 21	Yes	1	20.34	-72.23
Pose 65	Complex 22	Yes	1	20.06	-70.85
Pose 70	Complex 23	No	0	18.64	-70.00
Pose 83	Complex 24	No	0	19.50	-55.68
Pose 84	Complex 25	Yes	1	19.24	-55.24
Pose 91	Complex 26	Yes	1	19.48	-54.34
Pose 105	Complex 27	No	0	19.84	-51.08
Pose 123	Complex 28	No	0	19.28	-50.73
Pose 130	Complex 29	Yes	1	18.54	-49.31
Pose 138	Complex 30	No	0	21.28	-47.28
Pose 139	Complex 31	No	0	18.32	-47.00
Pose 141	Complex 32	Yes	1	19.88	-46.78
Pose 158	Complex 33	Yes	1	18.12	-40.54
Pose 169	Complex 34	Yes	1	19.86	-40.41
Pose 186	Complex 35	No	0	21.28	-36.10
Pose 190	Complex 36	No	0	19.72	-33.86
Pose 192	Complex 37	No	0	17.20	-30.77
Pose 193	Complex 38	Yes	1	17.92	-29.74
Pose 204	Complex 39	Yes	1	17.54	-29.28

Table B 3: ZDOCK scores for FP2'-hemoglobin complex structures

Docked Pose no:	Complex	Cluster centre	Best Energy Rep	ZDOCK	ZRANK
Cluster 16					
Pose 42	Complex 1	No	0	16.62	-35.83
Pose 43	Complex 2	No	0	17.00	-34.46
Pose 47	Complex 3	No	0	17.16	-30.11

Docked Pose no:	Complex	Cluster centre	Best Energy Rep	ZDOCK	ZRANK
Pose 50	Complex 4	Yes	1	16.64	-23.70
Pose 58	Complex 5	No	0	17.84	5.52
Pose 64	Complex 6	No	0	16.88	18.08
		Cluster 13			
Pose 21	Complex 7	Yes	1	17.06	-45.19
Pose 29	Complex 8	No	0	17.44	-41.28
Pose 30	Complex 9	Yes	1	16.84	-41.14
Pose 31	Complex 10	No	0	17.04	-40.84
Pose 34	Complex 11	No	0	16.88	-39.70
Pose 54	Complex 12	Yes	1	17.28	-29.97
Pose 153	Complex 13	No	0	16.68	-5.67
Pose 166	Complex 14	No	0	18.62	3.55
		Cluster 4			
Pose 21	Complex 7	Yes	1	17.06	-45.19
Pose 29	Complex 8	No	0	17.44	-41.28
Pose 30	Complex 9	Yes	1	16.84	-41.14
Pose 31	Complex 10	No	0	17.04	-40.84
Pose 34	Complex 11	No	0	16.88	-39.70
Pose 54	Complex 12	Yes	1	17.28	-29.97
Pose 153	Complex 13	No	0	16.68	-5.67
Pose 166	Complex 14	No	0	18.62	3.55
Pose 21	Complex 7	Yes	1	17.06	-45.19
Pose 29	Complex 8	No	0	17.44	-41.28
Pose 30	Complex 9	Yes	1	16.84	-41.14
Pose 31	Complex 10	No	0	17.04	-40.84
Pose 34	Complex 11	No	0	16.88	-39.70
Pose 54	Complex 12	Yes	1	17.28	-29.97
Pose 153	Complex 13	No	0	16.68	-5.67
Pose 166	Complex 14	No	0	18.62	3.55
Pose 21	Complex 7	Yes	1	17.06	-45.19
Pose 29	Complex 8	No	0	17.44	-41.28
Pose 30	Complex 9	Yes	1	16.84	-41.14
Pose 31	Complex 10	No	0	17.04	-40.84
Pose 34	Complex 11	No	0	16.88	-39.70
Pose 54	Complex 12	Yes	1	17.28	-29.97
Pose 153	Complex 13	No	0	16.68	-5.67
Pose 166	Complex 14	No	0	18.62	3.55

Table B 4: ZDOCK scores for FP3-hemoglobin complex structures

Docked Pose no:	Complex	Cluster centre	Best Energy Rep	ZDOCK	ZRANK
		Cluster 2			
Pose 2	Complex 1	Yes	1	19.28	-85.07
Pose 15	Complex 2	No	0	16.90	-67.24
Pose 19	Complex 3	Yes	1	17.26	-64.15
Pose 26	Complex 4	Yes	1	17.78	-61.52
Pose 27	Complex 5	Yes	1	15.84	-61.17
Pose 28	Complex 6	Yes	1	16.02	-61.04
Pose 41	Complex 7	No	0	18.22	-56.62
Pose 47	Complex 8	Yes	1	16.34	-54.41
		Cluster 1			
Pose 2	Complex 9	Yes	1	22.90	-94.22

Docked Pose no:	Complex	Cluster centre	Best Energy Rep	ZDOCK	ZRANK
Pose 3	Complex 10	Yes	1	21.96	-89.85
Pose 5	Complex 11	Yes	1	20.32	-89.14
Pose 8	Complex 12	Yes	1	21.72	-87.35
Pose 9	Complex 13	No	0	19.94	-86.32
Pose 17	Complex 14	No	0	18.65	-69.60
Pose 30	Complex 15	No	0	18.36	-65.92
Pose 35	Complex 16	Yes	1	17.50	-63.03
Pose 38	Complex 17	Yes	1	17.00	-62.70
Pose 42	Complex 18	Yes	1	17.50	-62.48
Pose 44	Complex 19	Yes	1	16.38	-61.35
Pose 63	Complex 20	Yes	1	15.58	-61.27
Pose 69	Complex 21	Yes	1	16.14	-60.38
Pose 80	Complex 22	No	0	15.30	-47.43
Pose 105	Complex 23	No	0	15.20	-40.99
Pose 109	Complex 24	Yes	1	15.38	-40.25

Table B 5: ZDOCK scores for VP2-hemoglobin complex structures

Docked Pose no:	Complex	Cluster centre	Best Energy Rep	ZDOCK	ZRANK
Cluster 12					
Pose 48	Complex 1	Yes	1	18.92	-41.56
Pose 63	Complex 2	No	0	16.74	-29.88
Pose 77	Complex 3	No	0	17.24	-20.44
Pose 79	Complex 4	No	0	18.92	-19.78
Pose 89	Complex 5	No	0	18.20	-16.39
Pose 96	Complex 6	Yes	1	18.30	-14.10
Pose 105	Complex 7	Yes	1	18.96	-0.58
Pose 115	Complex 8	No	0	18.06	2.29
Pose 131	Complex 9	No	0	17.76	7.39
Pose 181	Complex 10	Yes	1	18.22	81.08
Cluster 1					
Pose 1	Complex 11	Yes	1	21.18	-87.36
Pose 7	Complex 12	Yes	1	20.58	-77.76
Pose 10	Complex 13	Yes	1	20.68	-68.51
Pose 11	Complex 14	No	0	20.10	-66.13
Pose 15	Complex 15	No	0	18.18	-65.14
Pose 16	Complex 16	Yes	1	17.98	-58.42
Pose 21	Complex 17	Yes	1	17.92	-57.60
Pose 27	Complex 18	Yes	1	17.54	-49.63
Pose 30	Complex 19	Yes	1	17.60	-42.62
Pose 31	Complex 20	Yes	1	19.82	-39.73
Pose 35	Complex 21	No	0	18.24	-30.64
Pose 46	Complex 22	Yes	1	16.58	-29.21
Pose 54	Complex 23	Yes	1	17.72	-26.68
Pose 55	Complex 24	Yes	1	18.62	-25.26
Pose 58	Complex 25	Yes	1	18.56	-2.84

Table B 6: ZDOCK scores for VP3-hemoglobin complex structures

Docked Pose no:	Complex	Cluster centre	Best Energy Rep	ZDOCK	ZRANK
Cluster 12					
Pose 23	Complex 1	No	0	16.50	-36.69
Pose 43	Complex 2	No	0	17.80	-34.24
Pose 65	Complex 3	Yes	1	16.48	-27.95
Pose 70	Complex 4	Yes	1	16.42	-14.90
Pose 74	Complex 5	No	0	17.12	-6.32
Pose 83	Complex 6	Yes	1	17.82	1.05
Cluster 2					
Pose 2	Complex 7	Yes	1	23.34	-104.29
Pose 11	Complex 8	Yes	1	20.34	-91.94
Pose 23	Complex 9	Yes	1	18.92	-83.07
Pose 25	Complex 10	Yes	1	19.84	-80.86
Pose 33	Complex 11	Yes	1	20.26	-78.33
Pose 53	Complex 12	No	0	20.84	-74.33
Pose 70	Complex 13	No	0	19.26	-67.46
Pose 71	Complex 14	Yes	1	19.36	-62.72
Pose 81	Complex 15	Yes	1	19.76	-62.41
Pose 105	Complex 16	Yes	1	21.11	-59.58
Pose 120	Complex 17	No	0	17.42	-54.82
Pose 124	Complex 18	Yes	1	18.58	-50.93
Pose 156	Complex 19	No	0	18.14	-50.41
Pose 158	Complex 20	Yes	1	17.74	-49.41
Pose 181	Complex 21	Yes	1	19.72	-44.47

Table B 7: ZDOCK scores for FP2-cystatin complex structures

Docked Pose no:	Complex	Cluster centre	Best Energy Rep	ZDOCK	ZRANK
Cluster 1					
Pose 1	Complex 1	Yes	1	19.00	-115.71
Pose 2	Complex 2	Yes	1	18.47	-114.07
Pose 3	Complex 3	Yes	1	17.28	-112.74
Pose 4	Complex 4	Yes	1	16.78	-107.19
Pose 5	Complex 5	Yes	1	16.92	-106.17
Pose 6	Complex 6	Yes	1	15.30	-101.75
Pose 7	Complex 7	No	0	15.38	-100.29
Pose 8	Complex 8	No	0	16.54	-96.68
Pose 9	Complex 9	Yes	1	17.06	-94.37
Pose 10	Complex 10	Yes	1	17.00	-92.01
Pose 11	Complex 11	Yes	1	15.42	-90.29
Pose 12	Complex 12	Yes	1	15.02	-85.67
Pose 13	Complex 13	Yes	1	16.54	-80.22
Pose 14	Complex 14	Yes	1	15.44	-78.14
Pose 15	Complex 15	Yes	1	16.00	-74.11
Pose 16	Complex 16	Yes	1	16.98	-70.12
Pose 17	Complex 17	Yes	1	17.72	-61.74
Pose 18	Complex 18	Yes	1	16.32	-55.22
Pose 19	Complex 19	No	0	16.96	-52.09
Pose 20	Complex 20	Yes	1	17.20	-50.11

Appendix C

Table C. 1: Energies of FP2 (1YVB)-hemoglobin complex structures before and after minimization

FP2(1YVB)-hemogloin complex	Before Minimization				After Minimization			
	A1	A2	A3	B1	B2	C1	C2	C3
Complex 1	1.38×10^{07}	1.38×10^{07}	-393.47	1.38×10^{07}	-5.56×10^{04}	-373.9	-63.34	-310.56
Complex 2	6.83×10^{07}	6.83×10^{07}	29.23	6.83×10^{07}	$-5.54.x 10^{04}$	-349.47	-65.95	-283.52
Complex 3	1.08×10^{08}	1.08×10^{08}	-86.03	1.08×10^{08}	-5.55×10^{04}	-577.28	-52.03	-525.26
Complex 4	0.56×10^{12}	0.56×10^{12}	17.21	0.56×10^{12}	-5.54×10^{04}	-455.09	-58.13	-396.96
Complex 5	5.32×10^{08}	5.33×10^{08}	244.96	5.32×10^{08}	-5.69×10^{04}	-1747.36	-133.67	-1613.69
Complex 6	2.66×10^{07}	2.66×10^{07}	-366.29	2.66×10^{07}	-5.69×10^{04}	-1792.93	-110.20	1682.73
Complex 7	7.20×10^{09}	7.20×10^{09}	-109.90	7.20×10^{09}	-5.68×10^{04}	-1732.73	-99.26	-1633.47
Complex 8	0.12×10^{13}	0.13×10^{13}	187.51	0.12×10^{13}	-5.70×10^{04}	-1763.92	-127.14	-1636.78
Complex 9	7.85×10^{09}	7.86×10^{09}	78.31	7.85×10^{09}	-5.75×10^{04}	-1433.19	-123.52	-1309.67
Complex 10	1.15×10^{09}	1.15×10^{09}	-9.97	1.15×10^{09}	-5.76×10^{04}	-1430.79	-100.18	-1330.61
Complex 11	1.76×10^{09}	1.77×10^{09}	0.00	1.76×10^{09}	-5.74×10^{04}	-1053.59	-103.78	-949.81
Complex 12	1.00×10^{09}	1.00×10^{09}	0.00	1.00×10^{09}	-5.73×10^{04}	-1140.04	-103.47	-1036.57
Complex 13	2.24×10^{09}	2.24×10^{09}	-85.10	2.24×10^{09}	-5.70×10^{04}	-2099.87	-124.94	-1974.93
Complex 14	1.64×10^{09}	1.65×10^{09}	85.06	1.64×10^{09}	-5.70×10^{04}	-2016.62	-131.02	1885.60
Complex 15	1.19×10^{09}	1.20×10^{09}	125.21	1.19×10^{09}	-5.72×10^{04}	-2052.35	-128.28	-1924.07
Complex 16	2.19×10^{09}	2.19×10^{09}	-73.41	2.19×10^{09}	-5.71×10^{04}	-1944.72	-124.10	-1820.62
Complex 17	1.29×10^{09}	1.30×10^{09}	0.00	1.29×10^{09}	-5.71×10^{04}	-1840.60	-135.87	-1704.73
Complex 18	0.10×10^{13}	0.10×10^{13}	-0.44	0.10×10^{13}	-5.72×10^{04}	-1810.04	-110.29	-1699.74
Complex 19	0.93×10^{13}	0.93×10^{13}	0.72×10^{02}	0.93×10^{13}	-5.70×10^{04}	-1803.71	-118.15	1685.55
Complex 20	2.17×10^{09}	2.17×10^{09}	-47.11	2.17×10^{09}	-5.74×10^{04}	-1821.04	-117.31	-1703.73
Complex 21	1.49×10^{09}	1.49×10^{09}	108.47	1.49×10^{09}	-5.72×10^{04}	-1803.18	-136.53	-1666.64
Complex 22	1.56×10^{09}	1.55×10^{09}	-68.00	1.56×10^{09}	-5.73×10^{04}	-1690.31	-116.19	-1574.12
Complex 23	9.16×10^{09}	9.16×10^{09}	12.0	9.16×10^{09}	-5.70×10^{04}	-1526.78	-105.11	-1421.66
Complex 24	7.74×10^{09}	7.74×10^{09}	15.45	7.74×10^{09}	-5.72×10^{04}	-1561.32	-104.32	1457.00
Complex 25	8.17×10^{09}	8.17×10^{09}	74.90	8.17×10^{09}	-5.75×10^{04}	-1535.67	-108.06	-1427.61
Complex 26	1.02×10^{09}	1.02×10^{09}	0.00	1.02×10^{09}	-5.68×10^{04}	-1479.42	-105.23	-1347.19
Complex 27	1.94×10^{09}	1.94×10^{09}	91.20	1.94×10^{09}	-5.69×10^{04}	-1467.80	120.97	-1346.83
Complex 28	1.63×10^{09}	1.63×10^{09}	-156.83	1.63×10^{09}	-5.67×10^{04}	-1448.31	-95.32	-1352.79

Table C. 2: Energies of FP2 (2OUL)-hemoglobin complex structures before and after minimization

FP2(2OUL)-hemoglobin complex	Before Minimization				After Minimization			
	A1	A2	A3	B1	B2	C1	C2	C3
Complex 1	7.36×10^{08}	7.36×10^{07}	-102.71	7.40×10^{08}	-5.54×10^{04}	-379.61	-70.93	-308.68
Complex 2	7.36×10^{08}	7.40×10^{08}	-38.29	7.36×10^{08}	-5.53×10^{04}	-480.98	-69.79	-411.19
Complex 3	0.31×10^{14}	0.31×10^{14}	0.29×10^{03}	0.31×10^{14}	-5.54×10^{04}	-662.42	-69.87	-592.55
Complex 4	1.05×10^{09}	1.05×10^{09}	-339.60	1.06×10^{09}	-5.74×10^{04}	-1718.25	-107.50	-1610.74
Complex 5	2.47×10^{08}	2.47×10^{08}	-388.08	2.47×10^{08}	-5.71×10^{04}	-1845.13	-104.49	-1740.65
Complex 6	7.85×10^{07}	7.85×10^{07}	-380.63	7.85×10^{07}	-5.71×10^{04}	-1685.77	-95.79	-1589.79
Complex 7	8.06×10^{07}	8.06×10^{07}	-297.18	8.06×10^{07}	-5.69×10^{04}	-1756.68	-111.03	-1645.65
Complex 8	2.26×10^{07}	8.06×10^{07}	-357.38	2.27×10^{07}	-5.72×10^{04}	-1577.06	-88.26	-1488.80

Complex 9	1.66×10^{08}	1.66×10^{08}	-489.77	1.66×10^{08}	-5.74×10^{04}	-1493.94	-121.83	-1372.11

FP2(2OUL)-hemoglobin complex	Before Minimization				After Minimization			
	A1	A2	A3	B1	B2	C1	C2	C3
Complex 10	1.42×10^{10}	1.42×10^{10}	0.00	1.42×10^{10}	-5.72×10^{04}	-1645.59	-133.10	-1512.49
Complex 11	0.16×10^{13}	0.16×10^{13}	-0.39×10^{03}	0.16×10^{13}	-5.73×10^{04}	-1411.44	-93.94	-1317.51
Complex 12	1.71×10^{08}	1.71×10^{08}	-192.47	1.72×10^{08}	-5.69×10^{04}	-1618.38	-102.79	-1515.59
Complex 13	2.77×10^{09}	2.77×10^{09}	-329.40	2.77×10^{09}	-5.73×10^{04}	-1617.32	-119.25	-1498.07
Complex 14	9.92×10^{10}	9.92×10^{10}	-393.00	9.92×10^{10}	-5.74×10^{04}	-1328.53	-105.10	-1223.43
Complex 15	1.77×10^{09}	1.77×10^{09}	-269.60	1.77×10^{09}	-5.72×10^{04}	-1351.69	-108.58	-1243.11
Complex 16	1.75×10^{08}	1.75×10^{08}	-317.68	1.76×10^{08}	-5.70×10^{04}	-1449.75	-103.92	-1345.83
Complex 17	2.11×10^{10}	2.11×10^{10}	-453.00	2.11×10^{10}	-5.74×10^{04}	-1584.91	-112.05	-1472.86
Complex 18	7.75×10^{07}	7.75×10^{07}	-482.18	7.75×10^{07}	-5.72×10^{04}	-1503.37	-124.19	-1379.17
Complex 19	0.65×10^{12}	0.65×10^{12}	0.00	0.65×10^{12}	-5.74×10^{04}	-2412.08	-191.73	-2220.34
Complex 20	0.61×10^{12}	0.61×10^{12}	56.75	0.61×10^{12}	-5.68×10^{04}	-2398.93	-116.07	-2282.86
Complex 21	0.61×10^{12}	0.61×10^{12}	-193.10	0.61×10^{12}	-5.70×10^{04}	-2254.73	-115.43	-2139.30
Complex 22	0.61×10^{12}	0.61×10^{12}	-184.45	0.61×10^{12}	-5.73×10^{04}	-2005.10	-133.87	-1871.22
Complex 23	0.62×10^{12}	0.62×10^{12}	-160.90	0.62×10^{12}	-5.72×10^{04}	-2056.65	-127.69	-1928.96
Complex 24	0.60×10^{12}	0.60×10^{12}	0.20	0.60×10^{12}	-5.70×10^{04}	-2185.08	-130.61	-2054.46
Complex 25	0.59×10^{12}	0.59×10^{12}	-129.53	0.59×10^{12}	-5.70×10^{04}	-1929.94	-125.58	-1804.35
Complex 26	0.61×10^{12}	0.61×10^{12}	41.89	0.61×10^{12}	-5.73×10^{04}	-1881.04	-104.39	-1776.65
Complex 27	0.60×10^{12}	0.60×10^{12}	0.74×10^{02}	0.60×10^{12}	-5.70×10^{04}	-1921.49	-113.47	-1808.03
Complex 28	0.61×10^{12}	0.61×10^{12}	245.17	0.61×10^{12}	-5.72×10^{04}	-1784.57	-132.730	-1651.85
Complex 29	0.59×10^{12}	0.59×10^{12}	-488.55	0.59×10^{12}	-5.74×10^{04}	-1866.39	-116.53	-1749.86
Complex 30	0.63×10^{12}	0.63×10^{12}	-0.88×10^{02}	0.63×10^{12}	-5.72×10^{04}	-1821.21	-126.17	-1695.04
Complex 31	0.61×10^{12}	0.61×10^{12}	215.03	0.61×10^{12}	-5.67×10^{04}	-1802.01	-110.42	-1691.59
Complex 32	0.61×10^{12}	0.61×10^{12}	136.12	0.61×10^{12}	-5.69×10^{04}	-2046.32	-120.97	-1925.35
Complex 33	0.62×10^{12}	0.62×10^{12}	0.00	0.62×10^{12}	-5.74×10^{04}	-1753.52	-111.37	-1642.15
Complex 34	0.60×10^{12}	0.60×10^{12}	24.10	0.60×10^{12}	-5.75×10^{04}	-1771.47	-102.32	-1669.15
Complex 35	0.68×10^{12}	0.68×10^{12}	120.90	0.68×10^{12}	-5.70×10^{04}	-1711.92	-111.78	-1600.14
Complex 36	0.60×10^{12}	0.60×10^{12}	0.43×10^{01}	0.60×10^{12}	-5.72×10^{04}	-1414.11	-121.02	-1293.09

Table C. 3: Energies of FP2'-hemoglobin complex structures before and after minimization

FP2'-hemoglobin complex	Before Minimization				After Minimization			
	A1	A2	A3	B1	B2	C1	C2	C3
Complex 1	3.94×10^{06}	3.93×10^{06}	144.71	3.94×10^{06}	-5.63×10^{04}	-695.58	-109.64	-676.84
Complex 2	1.42×10^{06}	1.43×10^{06}	-41.11	1.42×10^{06}	-5.62×10^{04}	-874.68	-80.98	-793.69
Complex 3	3.38×10^{07}	3.38×10^{07}	33.98	3.38×10^{07}	-5.61×10^{04}	-795.59	-119.59	-676.00
Complex 4	1.12×10^{06}	1.12×10^{06}	-104.66	1.12×10^{06}	-5.63×10^{04}	-753.45	-112.68	-640.77
Complex 5	6.65×10^{07}	6.66×10^{07}	20.60	6.65×10^{07}	-5.62×10^{04}	-252.48	-84.64	-167.84
Complex 6	1.89×10^{06}	1.89×10^{06}	-1368.81	1.89×10^{06}	-5.66×10^{04}	-250.53	-82.70	-116.83
Complex 7	2.06×10^{06}	2.06×10^{06}	100.73	2.06×10^{06}	-5.60×10^{04}	-221.93	-80.81	-141.11
Complex 8	4.53×10^{06}	4.54×10^{06}	-280.35	4.53×10^{06}	-5.64×10^{04}	-336.19	-119.30	-216.89
Complex 9	1.89×10^{08}	1.89×10^{08}	71.20	1.89×10^{08}	-5.68×10^{04}	-1248.95	-95.85	-1154.10
Complex 10	8.11×10^{08}	8.11×10^{08}	0.00	8.11×10^{08}	-5.67×10^{04}	-1453.36	-111.68	-1341.68
Complex 11	6.19×10^{07}	6.19×10^{07}	-154.50	6.19×10^{07}	-5.66×10^{04}	-1156.29	-100.09	-1056.19
Complex 12	9.84×10^{07}	9.85×10^{07}	216.00	9.84×10^{07}	-5.68×10^{04}	-1423.25	-117.20	-1306.05
Complex 13	0.49×10^{12}	0.49×10^{12}	0.25×10^{03}	0.49×10^{12}	-5.73×10^{04}	-1587.17	-103.36	-1483.81

Complex 14	3.45 x 10 ^{06}	3.46 x 10 ^{06}	-79.76s	3.45 x 10 ^{06}	-5.69 x 10 ^{04}	-1365.23	-87.16	-1278.08
Complex 15	3.16 x 10 ^{06}	3.16 x 10 ^{06}	-44.51	3.16 x 10 ^{06}	-5.70 x 10 ^{04}	-2525.98	-200.19	-125.79
Complex 16	2.49 x 10 ^{06}	2.49 x 10 ^{06}	-161.00	2.49 x 10 ^{06}	-5.71 x 10 ^{04}	-2018.56	-118.36	-1900.20
Complex 17	8.06 x 10 ^{06}	8.06 x 10 ^{06}	0.00	8.06 x 10 ^{06}	-5.69 x 10 ^{04}	-1927.34	-130.18	-1797.16
Complex 18	5.22 x 10 ^{06}	5.22 x 10 ^{06}	-63.82	5.22 x 10 ^{06}	-5.69 x 10 ^{04}	-1990.35	-128.28	-1862.07
Complex 18	5.22 x 10 ^{06}	5.22 x 10 ^{06}	-63.82	5.22 x 10 ^{06}	-5.69 x 10 ^{04}	-1990.35	-128.28	-1862.07
Complex 19	1.53 x 10 ^{06}	1.54 x 10 ^{06}	42.15	1.53 x 10 ^{06}	-5.70 x 10 ^{04}	-1870.76	-143.17	-1727.59
Complex 20	6.76 x 10 ^{06}	6.76 x 10 ^{06}	21.31	6.76 x 10 ^{06}	-5.70 x 10 ^{04}	-1858.04	-133.09	-1724.95
Complex 21	3.53 x 10 ^{06}	3.53 x 10 ^{06}	49.11	3.53 x 10 ^{06}	-5.69 x 10 ^{04}	-1863.16	-149.00	-1714.15
Complex 22	4.78 x 10 ^{06}	4.78 x 10 ^{06}	144.80	4.78 x 10 ^{06}	-5.72 x 10 ^{04}	-1717.05	-91.98	-1725.07
Complex 23	1.52 x 10 ^{06}	1.52 x 10 ^{06}	0.00	1.52 x 10 ^{06}	-5.71 x 10 ^{04}	-1727.50	-113.47	-1614.04
Complex 24	7.27 x 10 ^{06}	7.27 x 10 ^{06}	58.98	7.27 x 10 ^{06}	-5.70 x 10 ^{04}	-1516.18	-114.77	-1401.41
Complex 25	5.19 x 10 ^{06}	5.19 x 10 ^{06}	165.10	5.19 x 10 ^{06}	-5.74 x 10 ^{04}	-1851.16	-146.09	-1705.07
Complex 26	7.80 x 10 ^{06}	7.80 x 10 ^{06}	278.80	7.80 x 10 ^{06}	-5.72 x 10 ^{04}	-1777.46	-102.97	-1674.49
Complex 27	9.63 x 10 ^{06}	9.63 x 10 ^{06}	-21.25	9.63 x 10 ^{06}	-5.70 x 10 ^{04}	-1712.14	-131.28	-1580.86
Complex 28	1.40 x 10 ^{06}	1.40 x 10 ^{06}	-228.00	1.40 x 10 ^{06}	-5.74 x 10 ^{04}	-1728.22	-117.62	-1610.60
Complex 29	2.24 x 10 ^{08}	2.23 x 10 ^{08}	69.15	2.24 x 10 ^{08}	-5.71 x 10 ^{04}	-1550.63	-92.18	-1458.44
Complex 30	5.08 x 10 ^{08}	5.08 x 10 ^{08}	96.00	5.08 x 10 ^{08}	-5.69 x 10 ^{04}	-1709.59	-117.65	-1591.94
Complex 31	1.12 x 10 ^{08}	1.12 x 10 ^{08}	80.00	1.12 x 10 ^{08}	-5.71 x 10 ^{04}	-1626.19	-96.21	-1529.98
Complex 32	0.15 x 10 ^{13}	0.15 x 10 ^{13}	0.39 x 10 ^{03}	0.15 x 10 ^{13}	-5.70 x 10 ^{04}	-1769.79	140.92	-1628.86

Table C. 4: Energies of FP3-hemoglobin complex structures before and after minimization

FP3-hemoglobin complex	Before Minimization				After Minimization			
	A1	A2	A3	B1	B2	C1	C2	C3
Complex 1	3.42 x 10 ^{08}	3.42 x 10 ^{08}	-3.10 x 10 ^{04}	3.42 x 10 ^{08}	-5.74 x 10 ^{04}	-1012.32	-99.21	-913.11
Complex 2	1.39 x 10 ^{07}	1.39 x 10 ^{07}	-3.10 x 10 ^{04}	1.39 x 10 ^{07}	-5.70 x 10 ^{04}	-1020.36	-61.02	-959.34
Complex 3	3.64 x 10 ^{06}	3.67 x 10 ^{06}	-3.10 x 10 ^{04}	3.64 x 10 ^{06}	-5.76 x 10 ^{04}	-1001.05	-97.13	-903.92
Complex 4	1.21 x 10 ^{07}	1.22 x 10 ^{07}	-3.08 x 10 ^{04}	1.21 x 10 ^{07}	-5.71 x 10 ^{04}	-1001.58	-86.21	-915.37
Complex 5	5.43 x 10 ^{05}	5.64 x 10 ^{05}	-3.11 x 10 ^{04}	5.43 x 10 ^{05}	-5.71 x 10 ^{04}	-971.81	-84.95	-886.86
Complex 6	1.05 x 10 ^{07}	1.05 x 10 ^{07}	-3.09 x 10 ^{04}	1.05 x 10 ^{07}	-5.65 x 10 ^{04}	-981.02	-92.51	-888.51
Complex 7	1.31 x 10 ^{08}	1.31 x 10 ^{08}	-3.10 x 10 ^{04}	1.31 x 10 ^{08}	-5.71 x 10 ^{04}	-973.71	-81.48	-892.23
Complex 8	4.98 x 10 ^{04}	7.17 x 10 ^{04}	-3.12 x 10 ^{04}	4.98 x 10 ^{04}	-5.70 x 10 ^{04}	-918.26	-72.31	-845.96
Complex 9	8.14 x 10 ^{08}	8.14 x 10 ^{08}	-3.09 x 10 ^{04}	8.14 x 10 ^{08}	-5.67 x 10 ^{04}	-987.52	-87.38	-900.14
Complex 10	1.45 x 10 ^{07}	1.46 x 10 ^{07}	-3.09 x 10 ^{04}	1.45 x 10 ^{07}	-5.70 x 10 ^{04}	-885.03	-79.00	-806.02
Complex 11	5.62 x 10 ^{06}	5.64 x 10 ^{06}	-3.12 x 10 ^{04}	5.62 x 10 ^{06}	-5.75 x 10 ^{04}	-1716.57	-98.09	-1618.49
Complex 12	6.31 x 10 ^{06}	6.31 x 10 ^{06}	-3.15 x 10 ^{04}	6.31 x 10 ^{06}	-5.74 x 10 ^{04}	-1718.61	-108.29	-1610.32
Complex 13	5.56 x 10 ^{05}	5.77 x 10 ^{05}	-3.08 x 10 ^{05}	5.56 x 10 ^{05}	-5.66 x 10 ^{04}	-1896.58	-190.91	-1705.67
Complex 14	2.20 x 10 ^{08}	2.22 x 10 ^{08}	-3.15 x 10 ^{04}	2.20 x 10 ^{08}	-5.72 x 10 ^{04}	-1844.26	-122.68	-1721.54
Complex 15	3.78 x 10 ^{04}	5.96 x 10 ^{04}	-3.10 x 10 ^{04}	3.78 x 10 ^{04}	-5.70 x 10 ^{04}	-1814.69	-185.41	-1729.26
Complex 16	3.54 x 10 ^{06}	3.56 x 10 ^{06}	-3.15 x 10 ^{04}	3.54 x 10 ^{06}	-5.75 x 10 ^{04}	-1753.26	-110.94	-1642.31
Complex 17	2.90 x 10 ^{07}	2.91 x 10 ^{07}	-3.14 x 10 ^{04}	2.90 x 10 ^{07}	-5.72 x 10 ^{04}	-1681.45	-114.51	-1566.94
Complex 18	1.07 x 10 ^{05}	1.29 x 10 ^{05}	-3.10 x 10 ^{04}	1.07 x 10 ^{05}	-5.72 x 10 ^{04}	-1546.19	-135.16	-1511.03
Complex 19	1.16 x 10 ^{08}	1.16 x 10 ^{08}	-3.09 x 10 ^{04}	1.16 x 10 ^{08}	-5.71 x 10 ^{04}	-1575.53	-105.33	-1470.20
Complex 20	5.68 x 10 ^{05}	5.89 x 10 ^{05}	-3.09 x 10 ^{04}	5.68 x 10 ^{05}	-5.71 x 10 ^{04}	-1431.50	-100.16	-1431.34
Complex 21	3.49 x 10 ^{09}	3.49 x 10 ^{09}	-3.10 x 10 ^{04}	3.49 x 10 ^{09}	-5.72 x 10 ^{04}	-1396.51	-109.08	-1287.44
Complex 22	2.35 x 10 ^{07}	2.35 x 10 ^{07}	-3.10 x 10 ^{04}	2.35 x 10 ^{07}	-5.71 x 10 ^{04}	-1316.66	-103.28	-1213.38
Complex 23	5.38 x 10 ^{06}	5.41 x 10 ^{06}	-3.10 x 10 ^{04}	5.38 x 10 ^{06}	-5.71 x 10 ^{04}	-1236.62	-98.35	-1138.27
Complex 24	2.95 x 10 ^{10}	2.95 x 10 ^{10}	-3.09 x 10 ^{04}	2.95 x 10 ^{10}	-5.73 x 10 ^{04}	-1196.08	-109.08	-1287.44

Complex 25	2.15 x 10^{10}	2.15 x 10^{10}	-3.09 x 10^{05}	2.15 x 10^{10}	-5.71 x 10^{04}	-1135.39	-82.00	-1053.38

Table C.5: Energies of VP2-hemoglobin complex structures before and after minimization

VP2-hemoglobin complex	Before Minimization				After Minimization			
	A1	A2	A3	B1	B2	C1	C2	C3
Complex 1	5.30 x 10^{08}	5.30 x 10^{08}	-59.51	5.30 x 10^{08}	-5.69 x 10^{04}	-1614.58	-138.66	-1475.92
Complex 2	2.98 x 10^{08}	2.99 x 10^{08}	-500.64	2.98 x 10^{08}	-5.68 x 10^{04}	-1481.36	-199.29	-1282.07
Complex 3	4.46 x 10^{06}	4.47 x 10^{06}	137.44	4.46 x 10^{06}	-5.68 x 10^{04}	-1321.85	-94.32	-1227.52
Complex 4	3.33 x 10^{07}	3.33 x 10^{07}	23.58	3.33 x 10^{07}	-5.68 x 10^{04}	-1171.36	-99.29	-1062.07
Complex 5	1.00 x 10^{09}	1.00 x 10^{09}	-28.11	1.00 x 10^{09}	-5.68 x 10^{04}	-1319.63	-109.85	-1209.78
Complex 6	3.28 x 10^{09}	3.28 x 10^{09}	-331.20	3.28 x 10^{09}	-5.61 x 10^{04}	-1306.21	-112.21	-1194.00
Complex 7	8.22 x 10^{09}	8.23 x 10^{09}	-451.90	8.22 x 10^{09}	-5.68 x 10^{04}	-1107.77	-112.19	-995.58
Complex 8	3.97 x 10^{06}	3.97 x 10^{06}	-208.40	3.97 x 10^{06}	-5.66 x 10^{04}	-976.63	-93.60	-883.04
Complex 9	0.17 x 10^{12}	0.17 x 10^{12}	-0.86x 10^{02}	0.17 x 10^{12}	-5.69 x 10^{04}	-672.16	-77.48	-594.67
Complex 10	1.94 x 10^{08}	1.94 x 10^{08}	67.30	1.94 x 10^{08}	-5.75 x 10^{04}	-2383.54	-221.79	-2161.75
Complex 11	7.78 x 10^{08}	7.77 x 10^{08}	131.68	7.78 x 10^{08}	-5.73 x 10^{04}	-2103.37	-191.09	-2294.46
Complex 12	5.62 x 10^{06}	5.64 x 10^{06}	-3.12 x 10^{04}	5.62 x 10^{06}	-5.75 x 10^{04}	-1716.57	-98.09	-1618.49
Complex 13	3.26 x 10^{06}	3.28 x 10^{06}	-3.16 x 10^{04}	3.26 x 10^{06}	-5.73 x 10^{04}	-1576.08	-107.52	-1468.56
Complex 14	6.31 x 10^{06}	6.31 x 10^{06}	-3.15 x 10^{04}	6.31 x 10^{06}	-5.74 x 10^{04}	-1718.61	-108.29	-1610.32
Complex 15	3.54 x 10^{06}	3.56 x 10^{06}	-3.15 x 10^{04}	3.54 x 10^{06}	-5.75 x 10^{04}	-1753.26	-110.94	-1642.31
Complex 16	2.90 x 10^{07}	2.91 x 10^{07}	-3.14 x 10^{04}	2.90 x 10^{07}	-5.72 x 10^{04}	-1681.45	-114.51	-1566.94
Complex 17	2.20 x 10^{08}	2.22 x 10^{08}	-3.15 x 10^{04}	2.20 x 10^{08}	-5.72 x 10^{04}	-1844.26	-122.68	-1721.54
Complex 18	9.53 x 10^{09}	9.53 x 10^{09}	-3.15 x 10^{04}	9.53 x 10^{09}	-5.73 x 10^{04}	-1590.31	-106.40	-1483.92
Complex 19	1.33 x 10^{09}	1.33 x 10^{09}	-3.14 x 10^{04}	1.33 x 10^{09}	-5.70 x 10^{04}	-1694.23	-104.45	-1589.78
Complex 20	2.03 x 10^{07}	2.03 x 10^{07}	-3.07 x 10^{04}	2.03 x 10^{07}	-5.70 x 10^{04}	-1894.07	-131.56	-1762.51
Complex 21	2.29 x 10^{06}	2.31 x 10^{06}	-3.04 x 10^{04}	2.29 x 10^{06}	-5.69 x 10^{04}	-1783.49	-121.21	-1662.28
Complex 22	0.13 x 10^{10}	0.13 x 10^{10}	-0.32 x 10^{05}	0.13 x 10^{10}	-5.70 x 10^{04}	-1709.59	-117.65	-1591.94
Complex 23	0.49 x 10^{15}	0.49 x 10^{15}	-0.31 x 10^{05}	0.49 x 10^{15}	-5.70 x 10^{04}	-1626.19	-96.21	-1529.98
Complex 24	0.60 x 10^{15}	0.61 x 10^{15}	-0.31 x 10^{05}	0.60 x 10^{15}	-5.70 x 10^{04}	-1769.79	140.92	-1628.86
Complex 25	0.40 x 10^{12}	0.40 x 10^{12}	-3.08 x 10^{05}	0.40 x 10^{12}	-5.66 x 10^{04}	-1756.98	-112.49	-1644.49

Table C.5: Energies of VP3-hemoglobin complex structures before and after minimization

VP3-hemoglobin complex	Before Minimization				After Minimization			
	A1	A2	A3	B1	B2	C1	C2	C3
Complex 1	3.69 x 10^{09}	3.69 x 10^{09}	-72.69	3.69 x 10^{09}	-5.60 x 10^{04}	-976.63	-93.60	-883.04
Complex 2	0.11 x 10^{13}	0.11 x 10^{13}	-0.18 x 10^{03}	0.11 x 10^{13}	-5.62 x 10^{04}	-672.16	-77.48	-594.67
Complex 3	2.33 x 10^{09}	2.34 x 10^{10}	-3.10 x 10^{04}	2.33 x 10^{07}	-5.64 x 10^{04}	-1189.88	-92.28	-1097.60
Complex 4	1.15 x 10^{10}	1.15 x 10^{10}	-3.12 x 10^{04}	1.15 x 10^{10}	-5.62 x 10^{04}	-1306.26	-104.15	-1202.11
Complex 5	2.58 x 10^{06}	2.60 x 10^{06}	-3.09 x 10^{04}	2.58 x 10^{06}	-5.66 x 10^{04}	-1221.50	-98.03	-1123.47
Complex 6	3.70 x 10^{07}	3.71 x 10^{07}	-3.11 x 10^{04}	3.70 x 10^{07}	-5.69 x 10^{04}	-1255.98	-121.55	-1134.83
Complex 7	6.56 x 10^{09}	6.56 x 10^{09}	-3.14 x 10^{04}	6.56 x 10^{09}	-5.81 x 10^{04}	-2274.24	-168.72	-2105.52
Complex 8	0.13 x 10^{12}	0.13 x 10^{12}	-0.31 x 10^{05}	0.13 x 10^{12}	-5.74 x 10^{04}	-2211.45	-161.93	-2049.52
Complex 9	2.66 x 10^{09}	2.66 x 10^{09}	-3.12 x 10^{04}	2.66 x 10^{09}	-5.77 x 10^{04}	-2178.82	-150.78	-2028.04
Complex 10	2.59 x 10^{07}	2.60 x 10^{07}	-3.10 x 10^{04}	2.59 x 10^{07}	-5.74 x 10^{04}	-2166.48	-158.48	-2008.00
Complex 11	1.04 x 10^{07}	1.05 x 10^{07}	-3.09 x 10^{04}	1.04 x 10^{07}	-5.75 x 10^{04}	-2154.26	-165.37	-1988.89
Complex 12	0.23 x 10^{13}	0.23 x 10^{13}	-0.31 x 10^{05}	0.23 x 10^{13}	-5.78 x 10^{04}	-2161.86	-170.27	-1991.59
Complex 13	1.79 x 10^{09}	1.80 x 10^{09}	-3.11 x 10^{04}	1.79 x 10^{09}	-5.75 x 10^{04}	-1938.93	-154.88	-1784.04
Complex 14	1.12 x 10^{06}	1.14 x 10^{06}	-3.08 x 10^{04}	1.12 x 10^{06}	-5.77 x 10^{04}	-1881.98	-143.34	-1738.64
Complex 15	2.88 x 10^{08}	2.89 x 10^{08}	-3.10 x 10^{04}	2.88 x 10^{08}	-5.78 x 10^{04}	-1969.31	-145.18	-1824.13
Complex 16	0.10 x 10^{12}	0.10 x 10^{12}	-0.31 x 10^{05}	0.10 x 10^{12}	-5.74 x 10^{04}	-1838.06	-128.69	-1709.37

Complex 17	2.84 x 10 ^{07}	2.85 x 10 ^{07}	-3.10 x 10 ^{04}	2.84 x 10 ^{07}	-5.72 x 10 ^{04}	-2115.17	-162.07	-1953.11
Complex 18	0.14 x 10 ^{15}	0.14 x 10 ^{15}	-0.31 x 10 ^{05}	0.14 x 10 15	-5.78 x 10 04	-1661.41	-134.73	-1526.68
Complex 19	1.82 x 10 ^{09}	1.82 x 10 ^{09}	-3.11 x 10 ^{04}	1.82 x 10 09	-5.81 x 10 04	-1735.65	-134.17	-1601.48
Complex 20	5.28 x 10 ^{09}	5.29 x 10 ^{09}	-3.13 x 10 ^{04}	5.28 x 10 09	-5.78 x 10 04	-1686.87	-122.39	-1564.48
Complex 21	2.36 x 10 ^{08}	2.36 x 10 ^{08}	-3.10 x 10 ^{04}	2.36 x 10 ^{08}	-5.82 x 10 ^{04}	-1737.63	-131.64	-1605.99

Table C. 6: Energies of FP2 (1YVB)-hemoglobin complex structures before and after minimization

FP2-cystatin (1YVB)	Before Minimization				After Minimization			
	A1	A2	A3	B1	B2	C1	C2	C3
Complex 1	1.27 x 10 ^{08}	1.27 x 10 ^{08}	-127.13	1.27 x 10 ^{08}	-2.42 x 10 ^{04}	-2316.99	-255.07	-2062.92
Complex 2	1.26 x 10 ^{08}	1.26 x 10 ^{08}	-128.68	1.26 x 10 ^{08}	-2.40 x 10 ^{04}	-2143.73	-250.56	-1893.17
Complex 3	2.51 x 10 ^{07}	2.50 x 10 ^{07}	-145.65	2.51 x 10 ^{07}	-2.39 x 10 ^{04}	-2130.02	-213.07	-1916.95
Complex 4	1.26 x 10 ^{08}	1.26 x 10 ^{08}	-128.99	1.26 x 10 ^{08}	-2.43 x 10 ^{04}	-1933.92	-208.06	-1825.87
Complex 5	1.28 x 10 ^{08}	1.28 x 10 ^{08}	-128.36	1.28 x 10 ^{08}	-2.45 x 10 ^{04}	-1950.28	-197.39	-1752.89
Complex 6	1.24 x 10 ^{08}	1.24 x 10 ^{08}	-128.00	1.24 x 10 ^{08}	-2.27 x 10 ^{04}	-1855.63	-198.28	-1657.35
Complex 7	1.25 x 10 ^{08}	1.25 x 10 ^{08}	-126.01	1.25 x 10 ^{08}	-2.44 x 10 ^{04}	-1733.92	-108.06	-1625.87
Complex 8	1.26 x 10 ^{08}	1.26 x 10 ^{08}	-127.95	1.26 x 10 ^{08}	-2.47 x 10 ^{04}	-1531.94	-102.79	-1329.15
Complex 9	1.27 x 10 ^{09}	1.28 x 10 ^{09}	-127.20	1.27 x 10 ^{09}	-2.41 x 10 ^{04}	-1555.57	-106.81	-1348.76
Complex 10	1.48 x 10 ^{09}	1.47 x 10 ^{09}	-124.92	1.48 x 10 ^{09}	-2.48 x 10 ^{04}	-1133.92	-108.06	-1025.87
Complex 11	1.31 x 10 ^{08}	1.31 x 10 ^{08}	-127.04	1.31 x 10 ^{08}	-2.47 x 10 ^{04}	-1031.94	-102.79	-929.15
Complex 12	1.24 x 10 ^{08}	1.25 x 10 ^{08}	-128.45	1.24 x 10 ^{08}	-2.40 x 10 ^{04}	-1055.57	-106.81	-948.76
Complex 13	1.35 x 10 ^{08}	1.34 x 10 ^{08}	-123.85	1.35 x 10 ^{08}	-2.43 x 10 ^{04}	-750.60	-93.23	-657.37
Complex 14	1.28 x 10 ^{08}	1.28 x 10 ^{08}	-127.64	1.28 x 10 ^{08}	-2.44 x 10 ^{04}	-1047.74	-101.67	-946.07
Complex 15	1.43 x 10 ^{09}	1.43 x 10 ^{09}	-126.94	1.43 x 10 ^{09}	-2.41 x 10 ^{04}	-1016.99	-255.07	-761.92
Complex 16	3.01 x 10 ^{09}	3.01 x 10 ^{09}	-127.47	3.01 x 10 ^{09}	-2.44 x 10 ^{04}	-931.11	-88.80	-842.30
Complex 17	4.27 x 10 ^{09}	4.27 x 10 ^{09}	-128.40	4.27 x 10 ^{09}	-2.46 x 10 ^{04}	-930.64	-93.30	-837.34
Complex 18	5.40 x 10 ^{09}	5.40 x 10 ^{09}	-127.55	5.40 x 10 ^{09}	-2.43 x 10 ^{04}	-1100.95	-95.22	-1005.73
Complex 19	2.63 x 10 ^{09}	2.64 x 10 ^{09}	-127.71	2.63 x 10 ^{09}	-2.40 x 10 ^{04}	-950.28	-97.39	-852.89
Complex 20	0.22 x 10 ^{14}	0.23 x 10 ^{14}	-0.12 x 10 ^{05}	0.22 x 10 ^{14}	-1.97 x 10 ^{04}	-955.63	-98.28	-857.34

Appendix D

Table D. 1: FP2$_{arm}$ (Pub) (1YVB) hemoglobin complex structure

FP2 (1YVB)	Hemoglobin	Chain	Type of interactions
VAL 134	ALA 71	A	Hydrophobic interactions
ALA 135	ALA 71	A	Hydrophobic interactions
PHE 141	ALA 71	A	Hydrophobic interactions
ILE 148	ALA 19, ALA 21 and ALA 63	A	Hydrophobic interactions
PHE 149	ALA 71	A	Hydrophobic interactions
MET 168	ALA 53	A	Hydrophobic interactions
ILE 169	ALA 12, ALA 19, TYR 24, ALA 26, PHE 46, ALA 53 and ALA 111	A	Hydrophobic interactions
TYR 171	ALA 82	A	Hydrophobic interactions
ILE 189	ALA 71	A	Hydrophobic interactions
LEU 196	ALA 65, ALA 79, LEU 80, ALA 83, LEU 83, LEU 86 and LEU 91	A	Hydrophobic interactions
MET 197	TRP 14, ALA 21, VAL 62, ALA 63, ALA 65, LEU 66, ALA 69, VAL 70, ALA 71, ALA 79, LEU 80, ALA 82, LEU 83 and LEU 86	A	Hydrophobic interactions
LEU 201	ALA 71, ALA 79 and ALA 82	A	Hydrophobic interactions
ASN 134	ALA 71 and ASP 74	A	Hydrogen Bonds
ASN 188	ALA 79	C	Hydrogen Bonds
GLU 195	SER 81 and ALA 82	C	Hydrogen Bonds
ARG 12	ASN 9	A	Hydrogen Bonds
GLU 5	ASP 6, LYS 7, LYS 11 and HIS 72	A	Charge-charge interactions
GLU 6	LYS 16	A	Charge-charge interactions
GLU 15	GLU 116 and HIS 122	A	Charge-charge interactions
ASP 133	LYS 7, LYS 11, HIS 72 and ASP 74	A	Charge-charge interactions
GLU 138	LYS 11	A	Charge-charge interactions
GLU 185	ASP 85, HIS 89, LYS 90 and LYS 139	C	Charge-charge interactions
GLU 195	ASP 85 and LYS 139	C	Charge-charge interactions
ASP 221	LYS 7	A	Charge-charge interactions
ASP 221	LYS 139	C	Charge-charge interactions
GLU 222	ASP 6, LYS 7 and LYS 127	A	Charge-charge interactions
GLU 222	ASP 85, HIS 89, LYS 139 and ARG 141	C	Charge-charge interactions

Table D. 2: FP2$_{arm}$ (Pub) (2OUL) hemoglobin complex structure

FP2 (2OUL)	Hemoglobin	Chain	Type of interactions
TYR 4	PRO 4, ALA 5 and ALA 12	A	Hydrophobic interactions
ILE 8	ALA 5 and ALA 12	A	Hydrophobic interactions
PHE 17	ALA 120	A	Hydrophobic interactions
PHE 17	PRO 51	B	Hydrophobic interactions
ALA 21	PRO 4	A	Hydrophobic interactions
VAL 131	PRO 4	A	Hydrophobic interactions
PRO 132	ALA 71	A	Hydrophobic interactions
PHE 142	ALA 120 and VAL 121	A	Hydrophobic interactions
LEU 178	ALA 5	A	Hydrophobic interactions
VAL 187	ALA 82	C	Hydrophobic interactions
LEU 190	ALA 65, ALA 71, ALA 79, LEU 80 and LEU 83	C	Hydrophobic interactions
TYR 198	PRO 4	A	Hydrophobic interactions
TYR 198	ALA 88	C	Hydrophobic interactions
LEU 225	VAL 1, LEU 2, PRO 4, VAL 73, MET76, PRO 77, ASN 78 and VAL 135	A	Hydrophobic interactions
MET 226	PRO 4	A	Hydrophobic interactions
MET 226	PRO 77 and ASN 78	C	Hydrophobic interactions
ASN 134	ALA 71 and ASP 74	A	Hydrogen Bonds
ASN 188	ALA 79	C	Hydrogen Bonds
GLU 195	SER 81 and ALA 82	C	Hydrogen Bonds
ARG 12	ASN 9	A	Hydrogen Bonds
GLU 5	ASP 6, LYS 7, LYS 11 and HIS 72	A	Charge-charge interactions
GLU 6	LYS 16	A	Charge-charge interactions
GLU 15	GLU 116 and HIS 122	A	Charge-charge interactions
ASP 133	LYS 7, LYS 11, HIS 72 and ASP 74	A	Charge-charge interactions
GLU 138	LYS 11	A	Charge-charge interactions
GLU 185	ASP 85, HIS 89, LYS 90 and LYS 139	C	Charge-charge interactions
GLU 195	ASP 85 and LYS 139	C	Charge-charge interactions
ASP 221	LYS 139	C	Charge-charge interactions
GLU 222	ASP 6, LYS 7 and LYS 127	A	Charge-charge interactions
GLU 222	ASP 85, HIS 89, LYS 139 and ARG 141	C	Charge-charge interactions

Table D. 3: FP2'$_{arm}$ -hemoglobin complex structure

FP2'	Hemoglobin	Chain	Type of interactions
TYR 4	PRO 4, ALA 12	A	Hydrophobic interactions
ILE 8	PRO 4, VAL 10, ALA 12 and ALA 13	A	Hydrophobic interactions
PHE 17	PRO 4, ALA 5 and PRO 51	A	Hydrophobic interactions
VAL 33	ALA 5	A	Hydrophobic interactions
VAL 131	PRO 4	A	Hydrophobic interactions
PRO 132	PRO 4 and VAL 73	A	Hydrophobic interactions
LEU 136	PRO 4	A	Hydrophobic interactions
LEU 140	PRO 4	A	Hydrophobic interactions
PHE 142	PRO 4, ALA 5 and PRO 51	A	Hydrophobic interactions
LEU 143	PRO 4	A	Hydrophobic interactions
PHE 181	PRO 4	A	Hydrophobic interactions
ILE 186	ALA 82	C	Hydrophobic interactions
VAL 187	ALA 82	C	Hydrophobic interactions
PHE 190	ALA 82, LEU 83	C	Hydrophobic interactions
TYR 198	VAL 1, ALA 88	A	Hydrophobic interactions
TYR 200	PRO 4	A	Hydrophobic interactions
LEU 225	VAL 1, PRO 4, VAL 73, MET 76 and PRO 77	A	Hydrophobic interactions
MET 226	VAL, LEU 2 and PRO 4	A	Hydrophobic interactions
MET 226	PRO 77	C	Hydrophobic interactions
LYS 9	GLU 116	A	Hydrogen Bonds
LYS 12	ASN 9	A	Hydrogen Bonds
ASN 14	ASP 52	B	Hydrogen Bonds
HIS 19	SER 3	A	Hydrogen Bonds
GLU 138	ALA 5 and ASN 9	A	Hydrogen Bonds
LYS 196	ASP 85 and HIS 89	C	Hydrogen Bonds
LYS 12	ASP 6	A	Charge-charge interactions
ASP 133	LYS 7	A	Charge-charge interactions
LYS 135	ASP 6 and ASP 74	A	Charge-charge interactions
LYS 137	ASP 74	A	Charge-charge interactions
GLU 138	ASP 6, LYS 7 and ASP 74	A	Charge-charge interactions
ARG 141	ASP 47 and ASP 52	B	Charge-charge interactions
LYS 184	ASP 85	C	Charge-charge interactions
GLU 185	ASP 85	C	Charge-charge interactions
LYS 192	ASP 64	C	Charge-charge interactions
GLU 195	ASP 85	C	Charge-charge interactions
LYS 196	ASP 85	C	Charge-charge interactions
HIS 197	ASP 85	C	Charge-charge interactions
GLU 222	ASP 6	A	Charge-charge interactions
ARG 227	ASP 74	A	Charge-charge interactions

Table D. 4: FP2$_{arm}$ (1YVB) hemoglobin complex structure

FP2 (1YVB)	Hemoglobin	Chain	Type of interactions
TRP 43	PRO 4	A	Hydrophobic interactions
TYR 78 (S3)	PRO 4 and ALA 5	A	Hydrophobic interactions
LEU 84 (S2)	LEU 2 and PRO 4	A	Hydrophobic interactions
VAL 150 (S1')	PRO 77	A	Hydrophobic interactions
ALA 151	VAL 1 and PRO 77	A	Hydrophobic interactions
ALA 151	PRO 77	C	Hydrophobic interactions
VAL 152 (S1')	VAL 1, MET 76 and PRO 77	A	Hydrophobic interactions
PHE 156	VAL 1 and PRO 77	A	Hydrophobic interactions
PHE 158	VAL 1, LEU 80, ALA 82, LEU 83, ALA 88 and TYR 140	A	Hydrophobic interactions
TYR 159	VAL 1	A	Hydrophobic interactions
ILE 163	PRO 4	A	Hydrophobic interactions
PHE 164	VAL 1 and LEU 2	A	Hydrophobic interactions
LEU 172 (S2)	LEU 2 and PRO 4	A	Hydrophobic interactions
ALA 175	PRO 4	A	Hydrophobic interactions
ILE 186	PRO 4	A	Hydrophobic interactions
PRO 189	PRO 4 and ALA 12	A	Hydrophobic interactions
LEU 190	ALA 12	A	Hydrophobic interactions
TRP 206	VAL 73 and ALA 79	A	Hydrophobic interactions
TRP 210	ALA 79	A	Hydrophobic interactions
MET 226	ALA 71, MET 76 and PRO 77	C	Hydrophobic interactions
GLN 36 (S1)	ASP 74	A	Hydrogen Bonds
ASN 38	LYS 11, VAL 70 and ALA 71	A	Hydrogen Bonds
CYS 80 (S1)	LEU 83	A	Hydrogen Bonds
ASN 81 (S1)	THR 8	A	Hydrogen Bonds
ASP 155	VAL 1	C	Hydrogen Bonds
ASP 165	LEU 2	C	Hydrogen Bonds
ASP 165	LYS 7	C	Hydrogen Bonds
GLU 167	VAL 1 and SER 131	C	Hydrogen Bonds
ASP 170	VAL 1	A	Hydrogen Bonds

FP2 (1YVB)	Hemoglobin	Chain	Type of interactions
HIS 174 (S1')	ASP 74	A	Hydrogen Bonds
MET 226	ASP 75	C	Hydrogen Bonds
LYS 228	ASP 74	C	Hydrogen Bonds
ASP 35	HIS 72	A	Charge-charge interactions
LYS 37	ASP 74 and ASP 75	A	Charge-charge interactions
ASP 109	HIS 72	A	Charge-charge interactions
LYS 228	ASP 74	C	Hydrogen Bonds
ASP 35	HIS 72	A	Charge-charge interactions
LYS 37	ASP 74 and ASP 75	A	Charge-charge interactions
ASP 109	HIS 72	A	Charge-charge interactions

Table D. 5: FP2$_{\text{active site}}$ (1YVB) hemoglobin complex structure

FP2 (1YVB)	Hemoglobin	Chain	Type of interactions
TRP 43	ALA 82 and LEU 83	A	Hydrophobic interactions
ALA 44	ALA 82	A	Hydrophobic interactions
PHE 45	ALA 82	A	Hydrophobic interactions
VAL 71	ALA 65, ALA 82, LEU 83	A	Hydrophobic interactions
TYR 78 (S3)	MET 76, PRO 77, ALA 79, LEU 80, ALA 82, LEU 83 and VAL 135	A	Hydrophobic interactions
LEU 84 (S2)	PRO 77, LEU 80 and LEU 83	A	Hydrophobic interactions
TYR 106	LEU 86	A	Hydrophobic interactions
VAL 150 (S1')	PRO 4	C	Hydrophobic interactions
ALA 151	VAL 1 and PRO 4	C	Hydrophobic interactions
VAL 152 (S1')	ALA 88	A	Hydrophobic interactions
VAL 152 (S1')	PRO 36	D	Hydrophobic interactions
PHE 156	ALA 88	A	Hydrophobic interactions
PHE 156	PRO 36 and LEU 48	D	Hydrophobic interactions
PHE 158	ALA 88	A	Hydrophobic interactions
PHE 158	TRP 37, PHE 41, PHE 42, PHE 45, ALA 53 and VAL 54	D	Hydrophobic interactions
TYR 159	LEU 48 and ALA 53	D	Hydrophobic interactions
PHE 164	LEU 48, PRO 51, ALA 53 and VAL 54	D	Hydrophobic interactions
LEU 172 (S2)	PRO 4	C	Hydrophobic interactions
TYR 206	LEU 86	A	Hydrophobic interactions
ILE 218	PRO 4	C	Hydrophobic interactions

FP2 (1YVB)	Hemoglobin	Chain	Type of interactions
LEU 225	PRO 4	C	Hydrophobic interactions
MET 226	ALA 5	C	Hydrophobic interactions
LEU 231	PRO 4	C	Hydrophobic interactions
GLN 36 (S1)	LYS 90	A	Hydrogen Bonds
ASN 81 (S1)	ASN 68, SER 81 and ALA 82	A	Hydrogen Bonds
GLY 82 (S1)	ALA 79	A	Hydrogen Bonds
GLY 83 (S1)	ASN 78 and SER 81	A	Hydrogen Bonds
TYR 78 (S3)	HIS 72	A	Hydrogen Bonds
ASP 109	LYS 61	A	Hydrogen Bonds
ASP 154	ARG 92	A	Hydrogen Bonds
ASP 154	TRP 37	D	Hydrogen Bonds
ASP 155	GLN 39	D	Hydrogen Bonds
PHE 158	ASP 47	D	Hydrogen Bonds
GLU 167	SER 49	D	Hydrogen Bonds
CYS 168	SER 3	C	Hydrogen Bonds
GLY 169	ASP 6 and LYS 127	C	Hydrogen Bonds
ASP 170	VAL 1	C	Hydrogen Bonds
ASP 170	LYS 127	C	Hydrogen Bonds
LEU 172 (S2)	LYS 139	A	Hydrogen Bonds
ASN 173	HIS 89	A	Hydrogen Bonds
ASP 35	LYS 90	A	Charge-charge interactions
ASP 109	HIS 58 and LYS 61	A	Charge-charge interactions
ASP 154	HIS 87 ,HIS 89 and ARG 92	A	Charge-charge interactions
ASP 154	LYS 127	C	Charge-charge interactions
ASP 155	HIS 89, LYS 127 and ARG 141	C	Charge-charge interactions
LYS 160	ASP 47	D	Charge-charge interactions
GLU 167	HIS 122, ASP 126 and LYS 127	C	Charge-charge interactions
GLU 167	ARG 30 and ASP 52	D	Charge-charge interactions
GLU 219	ASP 52	D	Charge-charge interactions
ASP 234 (S2)	LYS 139	A	Charge-charge interactions

1 **Table D. 6: FP2$_{arm}$ (2OUL) hemoglobin complex structure**

FP2 (2OUL)	Hemoglobin	Chain	Type of interactions
ALA 151	PRO 4	C	Hydrophobic interactions
VAL 152 (S1')	PRO 4	C	Hydrophobic interactions
PHE 156	VAL 1	C	Hydrophobic interactions
ALA 157 (S1')	LEU 2, ALA 79, LEU 80 and PHE 128	C	Hydrophobic interactions
PHE 158	VAL 1	A	Hydrophobic interactions
PHE 158	VAL 1 and LEU 2	C	Hydrophobic interactions
TYR 159	VAL 1	A	Hydrophobic interactions
ILE 163	VAL 1, VAL 73, MET 76 and PRO 77	A	Hydrophobic interactions
PHE 164	VAL 1 and ALA 79	A	Hydrophobic interactions
LEU 172 (S2)	LEU 2, PRO 4	C	Hydrophobic interactions
ILE 186	PRO 4	A	Hydrophobic interactions
VAL 187	ALA 12 and TRP 14	A	Hydrophobic interactions
PRO 189	VAL 10, ALA 12, TRP 14, VAL 17, ALA 19, ALA 19, ALA 63 and ALA 65	A	Hydrophobic interactions
LEU 190	TRP 14, VAL 17, ALA 19, ALA 21, LEU 66, VAL 70 and ALA 71	A	Hydrophobic interactions
TRP 206	LEU 2, MET 76 and PRO 77	C	Hydrophobic interactions
TRP 210	MET 76 and ALA 79	C	Hydrophobic interactions
ASN 38	ALA 71 and HIS 72	C	Hydrogen Bonds
ASP 154	VAL 1 and SER 131	C	Hydrogen Bonds
GLU 161	LYS 7	A	Hydrogen Bonds
ASP 165	VAL 1	A	Hydrogen Bonds
GLU 167	LYS 139	A	Hydrogen Bonds
VAL 187	LYS 11	A	Hydrogen Bonds
GLN 209	ASN 78	C	Hydrogen Bonds
GLU 219	ASN 78	A	Hydrogen Bonds
LYS 228	ASN 78	A	Hydrogen Bonds
ASP 35	ASP 74	A	Charge-charge interactions
LYS 37	ASP 74	C	Charge-charge interactions
ASP 109	GLU 23 and HIS 72	A	Charge-charge interactions
ASP 154	LYS 7 and LYS 127	C	Charge-charge interactions
ASP 155	LYS 7 and LYS 139	C	Charge-charge interactions
GLU 161	ASP 74	A	Charge-charge interactions
ASP 165	ASP 74	A	Charge-charge interactions
ASP 170	LYS 7 and ARG 141	C	Charge-charge interactions
GLU 195	HIS 72	A	Charge-charge interactions
GLU 219	HIS 72	A	Charge-charge interactions

2 **Table D. 7: FP2$_{active\ site}$(2OUL) hemoglobin complex structure**

FP2 (2OUL)	Hemoglobin	Chain	Type of interactions
TRP 43	ALA 82	A	Hydrophobic interactions
VAL 71	ALA 82	A	Hydrophobic interactions
TYR 78 (S3)	ALA 71 and ALA 79	A	Hydrophobic interactions
LEU 84	PRO 77 and LEU 80	A	Hydrophobic interactions
VAL 150 (S1')	VAL 1 and PRO 4	C	Hydrophobic interactions
ALA 151	VAL 1 and PRO 4	C	Hydrophobic interactions
VAL 152 (S1')	ALA 88 and TYR 140	A	Hydrophobic interactions
VAL 152 (S1')	VAL 1	C	Hydrophobic interactions
PHE 156	ALA 88	A	Hydrophobic interactions
PHE 156	VAL 1	C	Hydrophobic interactions
ALA 157 (S1')	PRO 36, LEU 48	D	Hydrophobic interactions
PHE 158	LEU 32, PRO 36, TRP 37, LEU 48 and PRO 51		Hydrophobic interactions
PHE 164	VAL 1, PRO 4	C	Hydrophobic interactions
LEU 172 (S2)	LEU 2	C	Hydrophobic interactions
LEU 172 (S2)	PRO 77 and LEU 80	A	Hydrophobic interactions
TYR 200	PRO 4	C	Hydrophobic interactions
TRP 206	ALA 88	A	Hydrophobic interactions
ILE 218	LEU 2 and PRO 4	C	Hydrophobic interactions
LEU 225	PRO 4	C	Hydrophobic interactions
MET 226	PRO 4	C	Hydrophobic interactions
LEU 231	LEU 2 and PRO 4	C	Hydrophobic interactions
CYS 80 (S1)	LEU 83	A	Hydrogen Bonds
ASN 81 (S1)	SER 81	A	Hydrogen Bonds
GLY 83 (S1)	ASN 78	A	Hydrogen Bonds
SER 153	TYR 140	A	Hydrogen Bonds
ASP 154	HIS 89	A	Hydrogen Bonds
ASP 154	PRO 36	D	Hydrogen Bonds
LYS 160	ASP 47	D	Hydrogen Bonds
GLU 167	ASP 6 and SER 124	C	Hydrogen Bonds
CYS 168	VAL 1, SER 3 and LYS 127	C	Hydrogen Bonds
ASP 170	SER 138	A	Hydrogen Bonds
GLN 171	SER 138	A	Hydrogen Bonds
ASN 173	ASP 85 and LYS 139	A	Hydrogen Bonds
ASP 35	LYS 90	A	Charge-charge interactions
LYS 37	ASP 85 and LYS 90	A	Charge-charge interactions
ASP 109	LYS 61	A	Charge-charge interactions

FP2 (2OUL)	Hemoglobin	Chain	Type of interactions
ASP 154	ARG 40	D	Charge-charge interactions
ASP 154	ASP 85 and HIS 87	A	Charge-charge interactions
ASP 155	ASP 6	C	Charge-charge interactions
ASP 155	GLU 43, ASP 47 and ASP 52	D	Charge-charge interactions
GLU 161	ASP 47	D	Charge-charge interactions
ASP 165	ASP 6	C	Charge-charge interactions
GLU 167	ASP 6, LYS 7, ASP 126 and LYS 127	C	Charge-charge interactions
ASP 170	ASP 74 and ASP 85	A	Charge-charge interactions
ASP 170	ASP 6 and ARG 141	C	Charge-charge interactions
HIS 174 (S1')	ASP 85	A	Charge-charge interactions
GLU 219	ASP 6 and LYS 7	C	Charge-charge interactions
ARG 227	ASP 74	C	Charge-charge interactions
LYS 228	ASP 6	C	Charge-charge interactions

Table D. 8: FP2'$_{arm}$ hemoglobin complex structure

FP2'	Hemoglobin	Chain	Type of interactions
ALA 151	LEU 48	B	Hydrophobic interactions
VAL 152 (S1')	LEU 48, PRO 51 and ALA 53	B	Hydrophobic interactions
PHE 156	LEU 48, PRO 51, ALA 53 and VAL 54	B	Hydrophobic interactions
PRO 157 (S1')	LEU 48, ALA 53 and VAL 54	B	Hydrophobic interactions
PRO 157 (S1')	ALA 5	A	Hydrophobic interactions
PHE 158	VAL 1, PRO 4 and ALA 5	A	Hydrophobic interactions
PHE 158	VAL 34, PRO 36, PRO 51 and VAL 54	B	Hydrophobic interactions
TYR 159	VAL 1 and ALA 5	A	Hydrophobic interactions
ILE 163	VAL 1	A	Hydrophobic interactions
PHE 164	VAL 1	A	Hydrophobic interactions
PHE 164	PRO 36	B	Hydrophobic interactions
LEU 190	VAL 1 and LEU 2	A	Hydrophobic interactions
LEU 190	LEU 80	C	Hydrophobic interactions
TRP 206	PRO 51 and ALA 53	B	Hydrophobic interactions
TRP 210	ALA 5 and PRO 51	A	Hydrophobic interactions
MET 226	ALA 82 and LEU 86	A	Hydrophobic interactions
SER 153	THR 50	B	Hydrogen Bonds

FP2'	Hemoglobin	Chain	Type of interactions
ASP 154	LEU 48	B	Hydrogen Bonds
PHE 165	VAL 1	A	Hydrogen Bonds
GLU 167	PRO 36 and TRP 37	B	Hydrogen Bonds
GLU 167	ARG 92	C	Hydrogen Bonds
CYS 168	HIS 89 and ARG 92	C	Hydrogen Bonds
ASP 170	GLN 39, PHE 42 and GLU 43	B	Hydrogen Bonds
GLU 171	PHE 45, GLY 46 and LYS 59	B	Hydrogen Bonds
ASN 173	GLY 46, ASP 47 and SER 49	B	Hydrogen Bonds
ASN 188	ASP 75 and HIS 89	C	Hydrogen Bonds
GLU 195	ASN 78	C	Hydrogen Bonds
LYS 228	ASP 85 and HIS 89	C	Hydrogen Bonds
ASP 154	ASP 6	A	Charge-charge interactions
ASP 154	ARG 30 and ASP 52	B	Charge-charge interactions
ASP 155	ASP 6 and ARG 141	A	Charge-charge interactions
LYS 160	ASP 6 and ASP 74	A	Charge-charge interactions
GLU 161	LYS 7 and ASP 74	A	Charge-charge interactions
ASP 165	ASP 6 and LYS 7	A	Charge-charge interactions
GLU 167	ASP 6	A	Charge-charge interactions
GLU 167	ARG 40 and HIS 89	B	Charge-charge interactions
ASP 170	ARG 40, GLU 43 and LYS 59	B	Charge-charge interactions
GLU 171	GLU 43, ASP 47 and LYS 59	B	Charge-charge interactions
HIS 174	ASP 47	B	Charge-charge interactions
LYS 192	ASP 74 and ASP 75	C	Charge-charge interactions
LYS 193	ASP 75	C	Charge-charge interactions
GLU 195	ASP 75	C	Charge-charge interactions
GLU 219	ASP 85, HIS 87 and HIS 89	C	Charge-charge interactions
ASP 221	ASP 85	C	Charge-charge interactions
ARG 227	ASP 85	C	Charge-charge interactions
LYS 228	ASP 85	C	Charge-charge interactions

Table D. 9: FP2'$_{active\ site}$ hemoglobin complex structure

FP2'	Hemoglobin	Chain	Type of interactions
TRP 43	ALA 82	A	Hydrophobic interactions
ALA 44	ALA 82	A	Hydrophobic interactions
PHE 45	ALA 82	A	Hydrophobic interactions

FP2'	Hemoglobin	Chain	Type of interactions
VAL 71	ALA 82	A	Hydrophobic interactions
PHE 75	ALA 71	A	Hydrophobic interactions
TYR 78 (S3)	ALA 65, ALA 69, ALA 71, VAL 73 and LEU 80	A	Hydrophobic interactions
LEU 84 (S2)	PRO 77, ALA 79, LEU 80, LEU 83 and LEU 86	A	Hydrophobic interactions
PRO 111	ALA 65	A	Hydrophobic interactions
ALA 151	VAL 1 and PRO 4	C	Hydrophobic interactions
VAL 152 (S1')	ALA 88	A	Hydrophobic interactions
PRO 157 (S1')	PRO 36, PHE 41, PHE 42, PHE 45 and LEU 48	D	Hydrophobic interactions
PHE 158	LEU 32, VAL 33, PRO 36, PHE 42 and LEU 48	D	Hydrophobic interactions
PHE 164	LEU 48, PRO 51 and ALA 53	D	Hydrophobic interactions
LEU 172 (S2)	PRO 77 and LEU 80	A	Hydrophobic interactions
LEU 172 (S2)	LEU 2 and PRO 4	C	Hydrophobic interactions
TRP 206	LEU 86 and ALA 88	A	Hydrophobic interactions
GLN 36 (S1)	LYS 90	A	Hydrogen Bonds
CYS 80 (S1)	LEU 83	A	Hydrogen Bonds
ASN 81 (S1)	LEU 80, SER 81 and ALA 82	A	Hydrogen Bonds
ASP 109	LYS 61	A	Hydrogen Bonds
VAL 152 (S1')	HIS 89	A	Hydrogen Bonds
ASP 154	HIS 89 and ARG 92	A	Hydrogen Bonds
ASP 155	GLN 39 and SER 49	D	Hydrogen Bonds
ASP 165	SER 49	D	Hydrogen Bonds
GLY 166	THR 50	D	Hydrogen Bonds
VAL 170	VAL 1	C	Hydrogen Bonds
GLU 171	VAL 1	C	Hydrogen Bonds
LEU 172 (S2)	LYS 139	A	Hydrogen Bonds
ASN 173	ASP 85	A	Hydrogen Bonds
ASP 109	HIS 58, LYS 60, LYS 61 and HIS 87	A	Charge-charge interactions
ASP 154	HIS 87 and ASP 94	A	Charge-charge interactions
ASP 154	ARG 40	D	Charge-charge interactions
ASP 155	ARG 40 and ASP 52	D	Charge-charge interactions
LYS 160	ASP 47	D	Charge-charge interactions
GLU 161	ASP 47	D	Charge-charge interactions
ASP 165	ASP 47 and ASP 52	D	Charge-charge interactions
GLU 167	ASP 6, LYS 127, LYS 139 and ARG 141	C	Charge-charge interactions
ASP 170	ASP 6 , LYS 7	C	Charge-charge interactions
GLU 171	ASP 85	A	Charge-charge interactions
GLU 171	LYS 7, LYS 127, LYS 139, ARG 141	C	Charge-charge interactions
HIS 174 (S1')	ASP 85	A	Charge-charge interactions
ASP 234 (S2)	LYS 139	C	Charge-charge interactions

Table D. 10: FP3$_{arm}$ hemoglobin complex structure

FP3	Hemoglobin	Chain	Type of interactions
PHE 160	ALA 12	A	Hydrophobic interactions
TYR 161 (S1')	VAL 70, ALA 71, VAL 73	A	Hydrophobic interactions
PHE 165	VAL 1, LEU 2, PRO 4, ALA 5, VAL 73	A	Hydrophobic interactions
TYR 166	PRO 4	A	Hydrophobic interactions
ILE 188	VAL 1, LEU 2, PRO 4, ALA 5	A	Hydrophobic interactions
TYR 189	VAL 1, ALA 123	A	Hydrophobic interactions
TYR 189	VAL 33, PRO 36, TRP 37	B	Hydrophobic interactions
TYR 189	TYR 140	C	Hydrophobic interactions
MET 196	VAL 1	A	Hydrophobic interactions
MET 196	PRO 77, TYR 140	C	Hydrophobic interactions
PHE 199	VAL 1, LEU 2, PRO 4, ALA 5	A	Hydrophobic interactions
TYR 200	PRO 4	A	Hydrophobic interactions
TYR 201	PRO 4	A	Hydrophobic interactions
ARG 162	LYS 11	A	Hydrogen Bonds
GLY 163	LYS 11	A	Hydrogen Bonds
PHE 165	ASP 74	A	Hydrogen Bonds
LYS 186	ASP 74	A	Hydrogen Bonds
GLY 194	SER 35	C	Hydrogen Bonds
ARG 27	ASP 74, ASP 75	A	Charge-charge interactions
ARG 162	ASP 74	A	Charge-charge interactions
PHE 165	ASP 74	A	Charge-charge interactions

FP3	Hemoglobin	Chain	Type of interactions
HIS 178	THR 8, LYS 11, HIS 72	B	Charge-charge interactions
LYS 186	ASP 6, LYS 7, HIS 72, ASP 74	A	Charge-charge interactions
ASP 187	ASP 74, LYS 139	C	Charge-charge interactions
TYR 189	VAL 1	A	Charge-charge interactions
GLU 191	HIS 122, ASP 126, LYS 127	A	Charge-charge interactions
ASP 192	ASP 6, LYS 127	A	Charge-charge interactions
GLY 194	HIS 89	C	Charge-charge interactions

Table D. 11 $FP3_{active\ site}$ hemoglobin complex structure

FP3	Hemoglobin	Chain	Type of interactions
LEU 40	LEU 48, PRO 51 and ALA 53	D	Hydrophobic interactions
TRP 45	VAL 1 and PRO 4	C	Hydrophobic interactions
VAL 77	VAL 1	C	Hydrophobic interactions
VAL 77	MET 76, PRO 77, LEU 80 and VAL 135	A	Hydrophobic interactions
TYR 83 (S1)	VAL 1,LEU 2, ALA 123 and TYR 140	A	Hydrophobic interactions
TYR 83 (S1)	VAL 33, PRO 36, TRP 37 and PRO 51	D	Hydrophobic interactions
LEU 86 (S2)	LEU 2, PRO 4	A	Hydrophobic interactions
TYR 83 (S1)	LEU 2, PRO 4, VAL 73 and PRO 77,	C	Hydrophobic interactions
ILE 87 (S2)	LEU 2 and PRO 4	C	Hydrophobic interactions
ALA 90	PRO 4	C	Hydrophobic interactions
LEU 97	PRO 77, ALA 79 and ALA 82	A	Hydrophobic interactions
PRO 113	ALA 82, LEU 86 and ALA 88	A	Hydrophobic interactions
LEU 118	ALA 79 and ALA 82	A	Hydrophobic interactions
ILE 152	PRO 4	C	Hydrophobic interactions
ALA 153	PRO 4 and ALA 12	C	Hydrophobic interactions
ALA 156 (S1')	PRO 4 and ALA 5	C	Hydrophobic interactions
ALA 172	ALA 12, ALA 13, VAL 121	C	Hydrophobic interactions
ALA 173	PRO 4, VAL 70 and VAL 73	C	Hydrophobic interactions
PRO 176(S2)	PRO 4, ALA 5, ALA 12 and VAL 73	C	Hydrophobic interactions
ALA 173	PRO 4, ALA 5	C	Hydrophobic interactions
TRP 208	ALA 5	C	Hydrophobic interactions
ASP 74	LYS 139	A	Hydrogen Bonds
CYS 84 (S1)	SER 81	A	Hydrogen Bonds
ASN 81 (S3)	HIS 89, SER 138 and LYS 139	A	Hydrogen Bonds
TYR 83 (S1)	VAL 73	C	Hydrogen Bonds
ASP 96	ASN 78	A	Hydrogen Bonds
GLU 114	LYS 90	A	Hydrogen Bonds
CYS 116	ASP 85	A	Hydrogen Bonds
ALA 172	THR 8	C	Hydrogen Bonds
PRO174(S2)	THR 8	C	Hydrogen Bonds
ASN 175 (S2)	THR 8	C	Hydrogen Bonds
ASP 74	ASP 85, HIS 89 and LYS 139	A	Charge-charge interactions
LYS 80 (S3)	ASP 75	A	Charge-charge interactions
ASP 93	LYS 139	A	Charge-charge interactions
ASP 96	ASP 75	A	Charge-charge interactions
GLU 114	ASP 85 and LYS 90	A	Charge-charge interactions
ARG 120	ASP 85	A	Charge-charge interactions
GLU 236 (S2)	LYS 7, LYS 11, ASP 74 and ASP 75	C	Charge-charge interactions

Table D. 12: $VP2_{arm}$ hemoglobin complex structure

VP2	Hemoglobin	Chain	Type of interactions
PRO 33	PRO 77	C	Hydrophobic interactions
VAL 34	VAL 73, MET 76 and PRO 77	C	Hydrophobic interactions
ALA 38	PRO 4	C	Hydrophobic interactions
ALA 109	PRO 4	C	Hydrophobic interactions
ALA 158 (S1')	VAL 1	C	Hydrophobic interactions
PHE 159	VAL 1, PRO 77 and VAL 135	C	Hydrophobic interactions
TYR 160	VAL 1, MET 76, PRO 77 and VAL 135	A	Hydrophobic interactions
ILE 164	VAL 73, PRO 77	A	Hydrophobic interactions
ALA 184	PRO 4	A	Hydrophobic interactions
ALA 187	PRO 4, VAL 10, ALA 71 and VAL 73	A	Hydrophobic interactions
TYR 188	PRO 4, ALA 12 and VAL 7O	A	Hydrophobic interactions
PHE 190	VAL 10, ALA 13, TRP 14, VAL 17, ALA 19, VAL 70, ALA 71 and VAL 73	A	Hydrophobic interactions

VP2	Hemoglobin	Chain	Type of interactions
MET 195	PRO 4, ALA 12 and TRP 14	A	Hydrophobic interactions
TYR 200	PRO 4	A	Hydrophobic interactions
TRP 207	VAL 1	C	Hydrophobic interactions
VAL 209	VAL 1, LEU 2, VAL 73, PRO 77 and VAL 135	C	Hydrophobic interactions
TRP 211	VAL 1 and LEU 2	C	Hydrophobic interactions
ILE 217	PRO 77	A	Hydrophobic interactions
ASP 36	LYS 7	C	Hydrogen Bonds
ASP 156	ASN 78	A	Hydrogen Bonds
ARG 161	PRO 77	A	Hydrogen Bonds
GLY 163	ASP 74	A	Hydrogen Bonds
ILE 164	ASP 74 and ASP 75	A	Hydrogen Bonds
ASP 166	HIS 72 and ASP 75	A	Hydrogen Bonds
GLU 185	SER 3 and LYS 7	A	Hydrogen Bonds
TYR 188	THR 8	A	Hydrogen Bonds
ARG 218	ASP 74	A	Hydrogen Bonds
ASP 36	LYS 7	C	Charge-charge interactions
GLU 103	HIS 72	A	Charge-charge interactions
ARG 161	ASP 74	A	Charge-charge interactions
ASP 166	LYS 7, HIS 72 and ASP 74	A	Charge-charge interactions
ARG 185	ASP 6, ASP 74 and LYS 127	A	Charge-charge interactions
ASP 186	ASP 74	A	Charge-charge interactions
ASP 189	LYS 11 and HIS 72	A	Charge-charge interactions
LYS 196	ASP 74	A	Charge-charge interactions
ARG 198	ASP 74	A	Charge-charge interactions
GLU 213	LYS 7	A	Charge-charge interactions
ARG 218	ASP 74	A	Charge-charge interactions

Table D. 13: VP2$_{\text{active site}}$ hemoglobin complex structure

VP2	Hemoglobin	Chain	Type of interactions
ALA 38	LEU 86	C	Hydrophobic interactions
TRP 44	ALA 82	C	Hydrophobic interactions
TYR 82	ALA 65, ALA 69, ALA 71, MET 76, PRO 77, VAL 135 and LEU 136	C	Hydrophobic interactions
PHE 85	PRO 77, ALA 79 and VAL 135	C	Hydrophobic interactions
ILE 151 (S1')	PRO 4	A	Hydrophobic interactions
ALA 152	VAL 1, LEU 2 and ALA 5	A	Hydrophobic interactions
VAL 153 (S1')	PRO 36	B	Hydrophobic interactions
PHE 157	ALA 5	A	Hydrophobic interactions
PHE 157	PRO 36	B	Hydrophobic interactions
PHE 159	LEU 32, PRO 36, PHE 42 and LEU 48	B	Hydrophobic interactions
TYR 160	PRO 51 and ALA 53	B	Hydrophobic interactions
PHE 165	ALA 5 and LEU 48	A	Hydrophobic interactions
ALA 172	VAL 1, LEU 2 and PRO 4,	A	Hydrophobic interactions
ALA 172	LEU 80 and TYR 140	C	Hydrophobic interactions
PRO 173 (S2)	LEU 2 and PRO 4	A	Hydrophobic interactions
TYR 201	PRO 4	A	Hydrophobic interactions
LEU 219	PRO 4	A	Hydrophobic interactions
PRO 229	ALA 5	A	Hydrophobic interactions
LEU 232	LEU 2 and PRO 4	A	Hydrophobic interactions
ASN 37 (S1)	LYS 90	C	Hydrogen Bonds
THR 79	HIS 72	C	Hydrogen Bonds
CYS 81 (S1)	LEU 2	A	Hydrogen Bonds
TYR 82	ALA 82	C	Hydrogen Bonds
GLY 83 (S3)	ASN 78	C	Hydrogen Bonds
ASP 155	PRO 36	B	Hydrogen Bonds
PHE 159	ASP 47	B	Hydrogen Bonds
GLU 168	VAL 33	B	Hydrogen Bonds
CYS 169	SER 3	A	Hydrogen Bonds
GLY 170	LYS 127	A	Hydrogen Bonds
GLU 171	VAL 1 and SER 3	A	Hydrogen Bonds
GLU 171	SER 138	C	Hydrogen Bonds
ASP 110	LYS 61 and LYS 90	C	Charge-charge interactions
GLU 134	LYS 7	A	Charge-charge interactions
ASP 155	ARG 40 and GLU 43	B	Charge-charge interactions
ASP 156	ASP 6	A	Charge-charge interactions
ASP 156	ASP 47	B	Charge-charge interactions
ASP 166	ASP 47 and ASP 52	B	Charge-charge interactions
GLU 168	ASP 6	A	Charge-charge interactions
GLU 168	ASP 52	B	Charge-charge interactions

VP2	Hemoglobin	Chain	Type of interactions
GLU 171	ASP 6, LYS 7	A	Charge-charge interactions
GLU 171	ASP 85, HIS 87, LYS 127 and ARG 141	C	Charge-charge interactions
HIS 175 (S1')	ASP 85	C	Charge-charge interactions
ARG 227	ASP 6 and ASP 74	A	Charge-charge interactions
LYS 228	ASP 74	A	Charge-charge interactions

Table D. 14: $VP3_{arm}$ hemoglobin complex structure

VP3	Hemoglobin	Chain	Type of interactions
TRP 43	VAL 73	A	Hydrophobic interactions
TYR 78	PRO 4, VAL 73	A	Hydrophobic interactions
ILE 150	VAL 1, PRO 77	A	Hydrophobic interactions
ALA 152	VAL 1, MET 76, PRO 77 and VAL 135	A	Hydrophobic interactions
PHE 156	VAL 1 and PRO 77	C	Hydrophobic interactions
VAL 157	VAL 1 and MET 76	C	Hydrophobic interactions
TYR 158	VAL 1 and MET 76	C	Hydrophobic interactions
TYR 159	VAL 1, LEU 2 and PRO 4	C	Hydrophobic interactions
LEU 163	VAL 1, LEU 2 and PRO 4	C	Hydrophobic interactions
TRP 164	VAL 1, LEU 2, PRO 4, ALA 5 and MET 76	C	Hydrophobic interactions
PHE 170	LEU 2, MET 76, ALA 130 and VAL 135	A	Hydrophobic interactions
PRO 172	LEU 2, PRO 4 and LEU 83	C	Hydrophobic interactions
MET 186	VAL 10, ALA 12 and VAL 70	C	Hydrophobic interactions
TYR 187	ALA 12	C	Hydrophobic interactions
ALA 189	VAL 10, ALA 12, TRP 14 and VAL 70	C	Hydrophobic interactions
MET 190	VAL 10, ALA 12, ALA 13, TRP 14, VAL 17, LEU 66 and PHE 128	C	Hydrophobic interactions
TRP 206	MET 76	A	Hydrophobic interactions
TRP 206	VAL 1 and PRO 77	C	Hydrophobic interactions
TRP 210	VAL 1	C	Hydrophobic interactions
MET 216	VAL 1	C	Hydrophobic interactions
MET 226	ALA 71 and PRO 77	C	Hydrophobic interactions
GLN 36 (S1)	ASP 75	A	Hydrogen Bonds

VP3	Hemoglobin	Chain	Type of interactions
LYS 37	ASN 78	A	Hydrogen Bonds
ASN 38	ASN 68 and ALA 79	A	Hydrogen Bonds
CYS 80 (S1)	HIS 72	A	Hydrogen Bonds
ASP 81	LYS 11	A	Hydrogen Bonds
ASN 153	VAL 1	A	Hydrogen Bonds
VAL 157	LYS 139	A	Hydrogen Bonds
TYR 158	VAL 135 and LEU 136	A	Hydrogen Bonds
PHE 164	LYS 7 and ASP 74	C	Hydrogen Bonds
ALA 189	LYS 11	C	Hydrogen Bonds
ARG 192	GLY 15,	C	Hydrogen Bonds
ARG 197	ALA 71	C	Hydrogen Bonds
GLN 219	HIS 72	C	Hydrogen Bonds

Table D. 15: $VP3_{active\ site}$ hemoglobin complex structure

VP3	Hemoglobin	Chain	Type of interactions
ALA 41	PRO 5	B	Hydrophobic interactions
TRP 43	ALA 10	B	Hydrophobic interactions
PHE 45	PRO 5	B	Hydrophobic interactions
TYR 78	ALA 13, LEU 14, PRO 124 and VAL 126	B	Hydrophobic interactions
ILE 85 (S2)	LEU 34	A	Hydrophobic interactions
ILE 85 (S2)	LEU 125	B	Hydrophobic interactions
PRO 86	PRO 124	B	Hydrophobic interactions
ILE 87	PRO 124	B	Hydrophobic interactions
VAL 109	PRO 5 and ALA 13	B	Hydrophobic interactions
PRO 132	LEU 48 and ALA 53	A	Hydrophobic interactions
PHE 136	LEU 48	A	Hydrophobic interactions
ALA 152 (S1')	PRO 5	B	Hydrophobic interactions
PHE 156	PRO 5	B	Hydrophobic interactions
VAL 157 (S1')	LEU 3	B	Hydrophobic interactions
TYR 158	PRO 5	B	Hydrophobic interactions
PHE 170	PHE 33, PHE 36, PRO 37 and LEU 100	A	Hydrophobic interactions
PRO 172	LEU 34, LEU 43 and PRO 125,	B	Hydrophobic interactions

1

VP3	Hemoglobin	Chain	Type of interactions
TRP 206 (S1')	PRO 5	B	Hydrophobic interactions
TRP 210	PRO 5	B	Hydrophobic interactions
MET 226	LEU 91	A	Hydrophobic interactions
ALA 233	LEU 34	A	Hydrophobic interactions
ALA 233	PHE 33	A	Hydrophobic interactions
PHE 236	LEU 48	A	Hydrophobic interactions
TYR 78 (S3)	LYS 120, GLU 121 and THR 123	B	Hydrogen Bonds
CYS 80 (S1)	SER 9	B	Hydrogen Bonds
ASP 81 (S1)	SER 9 and ALA 10	B	Hydrogen Bonds
ASN 84 (S2)	GLU 30	A	Hydrogen Bonds
GLN 133	ASP 47	A	Hydrogen Bonds
CYS 151	LYS 40	A	Hydrogen Bonds
ASN 153	HIS 2	B	Hydrogen Bonds
SER 167	HIS 143	D	Hydrogen Bonds
CYS 168	TYR 145	D	Hydrogen Bonds
PHE 170	TYR 145	D	Hydrogen Bonds
SER 171	LEU 34	A	Hydrogen Bonds
HIS 174 (S3)	GLU 6 and GLU 7	B	Hydrogen Bonds
ALA 233	SER 49	A	Hydrogen Bonds
GLN 234	HIS 50	A	Hydrogen Bonds
ASP 2	LYS 56	A	Charge-charge interactions
LYS 37	GLU 6	B	Charge-charge interactions
ASP 72	LYS 17	B	Charge-charge interactions
ASP 81	GLU 7 and LYS 17	B	Charge-charge interactions
GLU 130	ASP 47	A	Charge-charge interactions
ARG 135	ASP 47	A	Charge-charge interactions
ASP 154	HIS 2, LYS 8 and LYS 17	B	Charge-charge interactions
ASP 155	HIS 2	B	Charge-charge interactions
ASP 165	GLU 90	D	Charge-charge interactions
HIS 174 (S3)	GLU 6, ASP 79 and LYS 82	B	Charge-charge interactions
ASP 221	ASP 94	D	Charge-charge interactions
LYS 227	ASP 94	D	Charge-charge interactions

2

3

4 **Table D. 16: FP2-cystatin interacting residues**

Falcipain-2	Cystatin	Type of interactions
VAL 33	TRP 104	Hydrophobic interactions
TRP 43	LEU 7, LEU 8, ALA 10, PRO 11 and VAL 55	Hydrophobic interactions
ALA 44	LEU 8	Hydrophobic interactions
PHE 45	LEU 8 and VAL 55	Hydrophobic interactions
TYR 78 (S3)	LEU 7, LEU 8, ALA 10 and PRO 11	Hydrophobic interactions
LEU 84 (S2)	LEU 7, LEU 8 and ALA 10	Hydrophobic interactions
ILE 85 (S2)	LEU 7 and LEU 8	Hydrophobic interactions
ALA 88	LEU 7 and LEU 8	Hydrophobic interactions
PHE 89	LEU 8	Hydrophobic interactions
TYR 106	VAL 55 and TYR 100	Hydrophobic interactions
ILE 148	LEU 8	Hydrophobic interactions
VAL 150 (S1')	LEU 8 and LEU 54	Hydrophobic interactions
ALA 151	LEU 8 and LEU 54	Hydrophobic interactions
VAL 152 (S1')	VAL 12 and VAL 55	Hydrophobic interactions
PHE 156	PRO 103 and TRP 104	Hydrophobic interactions
ALA 157 (S1')	VAL 55, TYR 60 ,TRP 100, ILE 102, PRO 103 and TRP 104,	Hydrophobic interactions
PHE 158	ALA 24, ILE 58, TYR 60, ILE 102, PRO 103 and TRP 104	Hydrophobic interactions
TYR 159	ILE 58, PRO 103 and TRP 104	Hydrophobic interactions
LEU 172 (S2)	LEU 7, LEU 8, ALA 10, VAL 12 and LEU 54	Hydrophobic interactions
ALA 17 (S1')5	LEU 7, LEU 8, ALA 10 and LEU 54	Hydrophobic interactions
VAL 176	LEU 8, LEU 54	Hydrophobic interactions
TRP 206 (S1')	VAL 55, TYR 100, ILE 102, PRO 103, TRP 104 and LEU 105	Hydrophobic interactions
TRP 210	PRO 103 and TRP 104	Hydrophobic interactions
ILE 216	LEU 54 and TRP 104	Hydrophobic interactions
ALA 235	LEU 8	Hydrophobic interactions
PHE 236	LEU 8	Hydrophobic interactions
ASP 35	SER 56	Hydrogen Bonds
LYS 37	SER 56	Hydrogen Bonds
GLY 40 (S1)	GLN 53	Hydrogen Bonds
CYS 42	LEU 8 and GLY 9	Hydrogen Bonds
CYS 80	GLN 53	Hydrogen Bonds

Falcipain-2	Cystatin	Type of interactions
GLY 83 (S3)	LEU 8	Hydrogen Bonds
ASP 154	ASP 18	Hydrogen Bonds
ASN 173 (S2)	GLY 9	Hydrogen Bonds
TRP 206	VAL 55, SER 56 and TRP 104	Hydrogen Bonds
ASP 35	LYS 109	Charge-charge interactions
ASP 109	LYS 59, LYS 109 and GLU 112	Charge-charge interactions
ASP 154	ASP 15, ASP 18, GLU 19, ARG 23, ARG 52 and LYS 59	Charge-charge interactions
ASP 155	GLU 19 and ARG 52	Charge-charge interactions
GLU 167	GLU 19	Charge-charge interactions
ASP 170	ASP 15, GLU 16, ASP 18, GLU 19 and ARG 52	Charge-charge interactions
ASP 234 ()	ARG 6	Charge-charge interactions

1