THE IDENTIFICATION OF NATURAL INHIBITORY COMPOUNDS AGAINST THE PLASMODIUM GTP CYCLOHYDROLASE I (GCH1) ENZYME

A thesis submitted in fulfillment of the requirements for the degree

MASTER OF SCIENCE OF RHODES UNIVERSITY

by

Coursework/Thesis

In

Bioinformatics and Computational Molecular Biology

Department of Biochemistry and Microbiology

Faculty of Science

by

AFRAH KHAIRALLAH

March 2018

ABSTRACT

Malaria is a disease caused by protozoan parasites that invade red blood cells causing an infection. Malaria remains a global health problem. The genus Plasmodium infects about a quarter of a billion people annually, resulting in over a million death cases. This can severely affect the public health and socioeconomic development especially in countries with limited resources. Malaria is transmitted by the female Anopheles mosquito. Five species within the Plasmodium genus are known to cause infection in humans; namely *Plasmodium falciparum*, *Plasmodium Ovale*, *Plasmodium knowlesi*, *Plasmodium vivax* and *Plasmodium malariae*. The increased resistance by the parasite to the majority of available anti-malarial drugs has raised a great challenge in anti-malarial drug discovery. With the problem of drug resistance on the rise, the need to develop new anti-malarial treatment strategies and identification of alternative metabolic targets for the treatment of malaria is crucial. This study is focused on the Guanosine triphosphate CycloHydrolase I (GCH1) enzyme as a potential drug target. GCH1 is important for the survival of malaria parasites as shown by failed attempts to generate knockout lines in *plasmodium falciparum.* In this study, sequence and evolutionary analysis were carried out in both the human host and parasite GCH1 enzyme. Accurate 3D models of the parasite GCH1 were built and validated. The resulting models were used for high throughput screening against 623 compounds from the South African Natural Compounds Database (SANCDB; https://sancdb.rubi.ru.ac.za/). The high throughput screening was done to identify possible binding sites as well as hit compounds with high selectivity and binding affinity towards the parasite enzyme, this is followed by molecular dynamics simulations to identify protein-ligand complexes and analyze their stability. In this study, a total of five SANCDB compounds were identified as potential inhibitors: SANC00317, SANC00335, SANC00368, SANC00106, SANC00103 and SANC00286. It was found that GCH1 protein can be a potential anti-malarial drug target as it showed selective binding with the inhibitor compounds. The identified inhibitors showed good selectivity and lower free energy of binding towards the parasite GCH1. Force field parameters of GCH1 active site metal were derived and validated. The development of these force field parameters was important for accurate MD simulations of the protein active site; which will allow for future investigation of interactions and stability of the GCH1 protein-ligand complexes.

DECLARATION

I, Afrah Khairallah, declare that this thesis submitted to Rhodes University is my own original work and has not previously been submitted for a degree or diploma at this or any other institution.

Signature: …… Date: ……27/03/2018……

DEDICATION

This thesis is dedicated to the loving memory of my father,

YOUSEF

For always encouraging me to believe in myself

ACKNOWLEDGEMENTS

First and foremost, I thank God for giving me the strength, knowledge, courage and opportunity to undertake this research study.

My sincere gratitude and appreciation goes to my supervisor Prof. Ozlem Tastan Bishop for her continuous rapid support, guidance and constants feedback to keep my progress on schedule.

I extend my gratitude to my supervisor Dr. Vuyani Moses for his valuable assistance, enthusiastic encouragement and for his useful critiques towards this work.

Furthermore, I would like to express my great appreciation to all RUBi members for their inspiring guidance and advice.

A very special gratitude goes to Magambo Phillip Kimuda for his valuable and constructive suggestions which contributed significantly in the development of this research and for his willingness to give his time so generously.

I must express my profound gratitude to my mother, for her endless support and prayers. To my sisters, for their constant love and motivation. To Wayde, your belief in me and love has motivated me all through.

Funding acknowledgment

This work would not have been possible without the financial support from:

- The Organization for Women in Science for the Developing World (**OWSD**)
- The Swedish International Development Cooperation Agency (**SIDA**)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF WEB SERVERS AND SOFTWARE TOOLS

ANOLEA
http://melolab.org/anolea/

AutoDock Vina

Discovery Studio Visualizer 4

GAUSSIAN 09

HHPred
https://toolkit.tuebingen.mpg.de/#/

InterPro server scan
http://www.ebi.ac.uk/Tools/pfa/iprscan/

Jalview

LigPlot+

MAFFT
https://mafft.cbrc.jp/alignment/server/

MEGA7

MEME
http://meme-suite.org/

MODELLER v.9.16

NCBI BLAST
http://blast.ncbi.nlm.nih.gov/

PlasmoDB
http://plasmodb.org/plasmo/

PROCHECK
http://services.mbi.ucla.edu/PROCHECK/

PROMALS3D
http://prodata.swmed.edu/promals3d/promals3d.php

ProSA
https://prosa.services.came.sbg.ac.at/prosa.php

PyMOL

QMEAN
https://swissmodel.expasy.org/qmean/

RCSB Protein Data Bank:
http://www.rcsb.org/pdb/home/home.do

LIST OF ABBREVIATIONS

Abbreviation	Description
3D	3-Dimensional
AMBER	Assisted Model building with Energy Refinement
BLAST	Basic Local Alignment Search Tool
BLOSUM	Blocks Substitution Matrix
CHARMM	Chemistry at HARvard Molecular Mechanics
DOPE	Discrete Optimised Protein Energy
GROMACS	GROningen MAchine for Chemical Simulations
GROMOS	GROningen MOlecular Simulation
GCH1	GTP CycloHydrolase I
HMM	Hidden Markov model
MAFFT	Multiple Alignment using Fast Fourier Transform
MD	Molecular dynamics
MM	Molecular mechanics
MQAPs	Model quality assessment programs
MSA	Multiple Sequence Alignment
NCBI	National Centre for Biotechnology Information
NMR	Nuclear magnetic resonance
PBC	Periodic boundary conditions
PDB	Protein Data Bank
PROMALS3D	PROfile Multiple Alignment with predicted Local Structures and 3D constraints
PSI-BLAST	Position-Specific Iterated BLAST
PSSMs	Position Specific Scoring Matrices
QMEAN	Qualitative Model Energy ANalysis
RESP	Restrained electrostatic potential

Rg	Radius of gyration
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
SANCDB	South African Natural Compound Database
WHO	World Health Organization
ZINC	Zinc Is Not Commercial

List of amino acids and their respective three letter codes (TLC) and single letter codes (SLC)

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

CHAPTER ONE

LITERATURE REVIEW AND STUDY BACKGROUND

1.1 Introduction

Malaria is a disease caused by protozoan parasites that invade red blood cells causing an infection. The main stage of this infection is the erythrocytic stage (red blood cells) which is associated with the pathology. However, the parasite is also able to infect the hepatocytes (liver cells) (Cowman and Crabb, 2006). Malaria is a major world health problem (World Health Organization, 2016b). The genus Plasmodium infects about a quarter of billion people annually. The rate of malaria morbidity and mortality is higher in pregnant women and children. Pregnant women infected with malaria exhibit more severe symptoms including miscarriage, premature delivery, severe anemia and maternal death (van Eijk *et al.*, 2015). In 2016 it was estimated that there were 216 million cases of malria infection worldwide with the majority (92%) occurring in Africa, making it an endemic region to malaria infection (WHO, 2017). Countries with limited resources are severely affected by malaria, mainly on its public health and socioeconomic development due to the lack of access to effective affordable treatments.

Malaria is transmitted by the female Anopheles mosquitoes. The female Anopheles transmits the malaria parasite to the human host through a bite during its blood meal. Five species within the Plasmodium genus are known to cause malria infection in humans; namely *Plasmodium falciparum* (*P. falciparum*), *Plasmodium ovale* (*P. ovale*), *Plasmodium knowlesi* (*P. knowlesi*), *Plasmodium vivax* (*P. vivax*) and *Plasmodium malariae* (*P. malariae*) (N. J. White, 2008). Other species of Plasmodium known to cause malaria in rodents include *Plasmodium berghei* (*P. berghei*) *Plasmodium yoelii* (*P. yoelii*) and *Plasmodium chabaudi* (*P. chabaudi*). The latter are used as models to understand the parasite biology and host- parasite interactions for anti-malarial drug development (Carter and Walliker, 1975).

P. flacipurm and *P. malariae* are spread worldwide with more existence in Africa whereas *P.vivax* is mostly found in Asia and Latin America. *P. ovale* is found mostly in West Africa and the islands of the western Pacific (Autino *et al.*, 2012). Malaria parasites and their respective mosquito vectors are generally selective to a specific host range. However, *P.knowlesi* has been recently reported to infect humans in South East Asia, this is after being

known as a malarial pathogen of the pig-tailed (*Macaca nemestrina)* and long tailed macaques (*Macaca fasicularis*). *P.knowlesi* has a short daily life cycle and it can rapidly reach a lethal status (Kantele and Jokiranta, 2011). *P.malariae* is characterized with a long-lasting, chronic infection that in some cases can last a lifetime and it can also cause serious complications such as the nephrotic syndrome. *P. vivax* is known to have dormant liver stages referred to as "hypnozoites" which remain in the human hepatocyte from months to years before it can activate and invade the blood cells, resulting in late malaria lapses (Campo *et al.*, 2015). Among the five species, *P. falciparum* is regarded as the most virulent and is responsible for the vast majority of death cases, more than 90% (Snow, 2015). This is followed by *P. vivax*, that is known to be responsible for 86% of death cases occurring outside sub-Saharan Africa (Autino *et al.*, 2012; Kevin Baird, 2013).

1.2 *Plasmodium falciparum* life cycle

Figure 1.1 shows the *P. falciparum* life cycle in humans. Malaria infection is transmitted through the bite of an infected female Anopheles mosquito. During its blood meal the mosquito inoculates the parasites (sporozoites) into the host blood stream (Cowman *et al*., 2012). The pre-erythrocytic cycle begins when the sporozoites are taken up by the liver cells (hepatocytes) and multiply asexually (Vaughan, Aly, & Kappe, 2008; Prudêncio, Rodriguez and Mota, 2006). Sporozoites are characterized as nucleated highly motile cells with a single mitochondrion, apicoplast and a single microtubule interconnected by tethering proteins. Sporozoites carry sporozoite surface proteins such as thrombospondin related anonymous protein (TRAP) and circumsporozoite protein (CSP) which are known to play a major role in the recognition and anchoring of the hepatocytes cells (Robson, 1995). During the pre-erythrocytic stage the disease does not show any clinical manifestations. Thus, the host immune defence mechanisms are not activated which allows the merozoites to survive (Soulard *et al.*, 2015).

The parasite erythrocytic cycle starts when the sporozoites develop to form a multinucleate stage of the cell known as schizonts. The schizonts contain many haploid spindle-shaped parasites referred to as merozoites. When the liver cells rupture, thousands of the merozoites are released into blood circulation invading the red blood cells. A group of proteins from both the merozoites and host red blood cells mediate the erythrocyte invasion process.

In addition, the apex of the merozoites contains rhoptries and micronemes (mixture of proteinases and metabolic enzymes) that contribute to the invasion process. The invasion process results in thousands of infected red-blood cells of the host circulatory system (Soulard *et al.*, 2015) (Figure 1.1).

After the erythrocyte invasion, merozoites undergo a trophic period in which they enlarge and lose their apical rings, conoid, and rhopteries structures and their nuclei become lobulated. This stage is referred to as the ring stage or early trophozoite stage (unfilled cytoplasm) (Tilley, Dixon and Kirk, 2011; Soulard *et al.*, 2015). Ring stage infected erythrocytes circulate freely in the peripheral blood and become easily identifiable in infected patients. During the post invasion the parasite speeds up its metabolic rate, increasing the uptake of haemoglobin from the host cells resulting in the late trophozoites stage. Parasites at the late trophozoites stage are rarely observed due to their adherence to the endothelial cells and separation from the blood circulation. This stage is characterized by the existence of hemozoin pigments (Prudêncio, Rodriguez and Mota, 2006) (Figure 1.1).

A small portion of the merozoites reproduce sexually and subsequently differentiate into male and female gametocytes that can be taken up by the mosquito during the blood meal (Figure 1.1). Inside the mosquito, the gametocytes mature into gametes. As they fuse, diploid zygotes are formed and become ookinetes. These ookinetes travel to the middle gut of the mosquito and pass through the gut wall to form the locusts. A meiotic division of the locusts follows, resulting in sporozoites being formed which then migrate to the salivary glands of the female Anopheles mosquito completing the cycle of transmission back to human. Attempts to fight the parasite by targeting the gametocyte are currently not available (Talman *et al.*, 2004; Josling and Llinás, 2015). The presence of the parasite in the host system causes inflammatory responses including high fever, severe anemia, unconsciousness, breathing difficulties, failure of a number of body organs, coma and death (peter Lam, 2004).

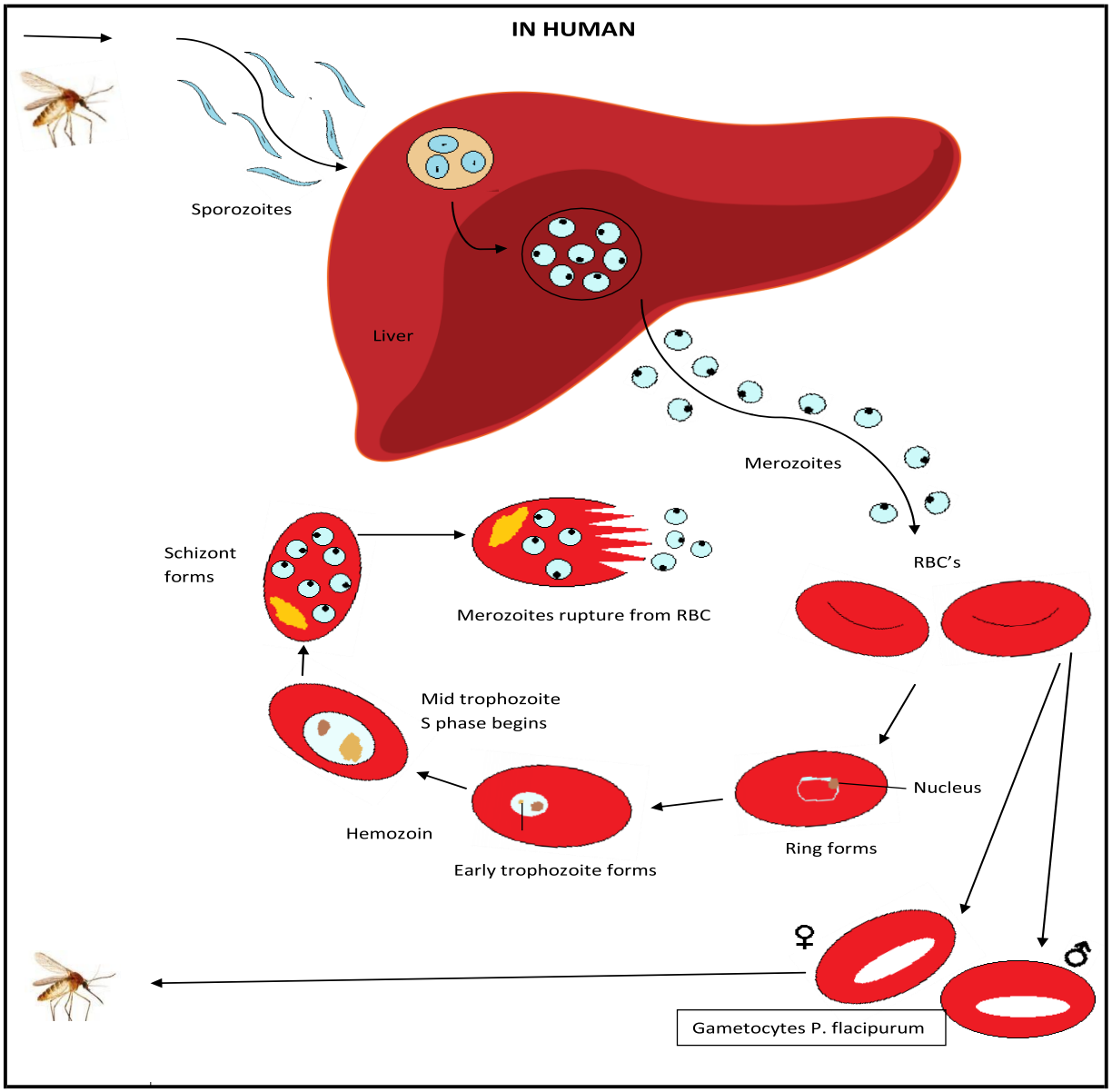


Figure 1.1: A schematic representation of *P. falciparum* life cycle in human. The cycle starts with the mosquito bite that inoculates sporozoites into the host blood stream. The sporozoites travel to invade the host hepatocytes then divide into haploid merozoites which are released back to the blood stream. The merozoites invade the red blood cells and start the asexual reproduction resulting in the release of thousands of merozoites which continue to invade uninfected erythrocytes. A small portion of merozoites in red blood cells develop to sexual gametocytes and get taken up by the mosquito during the blood meal; in which they differentiate and mature completing the cycle of transmission back to the host. Adapted from Hill (Hill, 2011).

1.3 Susceptibility to malaria infection

Individuals with sickle cell anemia and other hemoglobin related disorders are naturally immune to malaria infection. In addition, people with a negative Duffy blood group exhibit resistance to *P. vivax* infection however, *P. ovale* species has the ability to infect the Duffy-negative blood group (Langhi and Bordin, 2006). Folate-deficiency anemia is characterized by a shortage of folate, subsequently reducing the amount of red blood cells which can also provide protection against malaria. Variation in iron levels among different individuals is thought to offer some protection against malaria (Gwamaka *et al.*, 2012). Studies performed in children and pregnant women may indicate an association between iron deficiency and protection from malaria (Mockenhaupt *et al.*, 2000; Jonker *et al.*, 2012). A large-scale study in 2016 called the Pemba trial found a link between the supplementation of children with iron and folic acid and subsequent malaria infection and mortality (Sazawal *et al.*, 2006). The Pemba study was prematurely ended due to ethical reasoning. Following the Pemba trial, conflicting studies showed either no association or a protective effect when iron supplementations are administered (Zlotkin *et al.*, 2013).

1.4 Vector control

The environmental conditions in Africa provide an optimal state for the Anopheles mosquito vectors making malaria endemic to these regions (De Silva and Marshall, 2012). Several control methods have been used to control malaria vectors such as vaccination, chemotherapy and preventative regimens. The effective preventative regimens include rapid diagnosis, use of insecticide treated nets and effective insecticides with regular indoor spraying. These methods have been proven to be useful in saving lives, resulting in low morbidity and mortality rates over the past years (Beier *et al.*, 2008; Takken and Knols, 2009). However, increased parasite resistance to the available drugs and insecticides has adversely affected the recent progress and created a global problem to control and/or eradicate malaria.

1.5 Vaccine development

Currently there is no commercially available malaria vaccine. Around twenty vaccines are still under development or being clinically evaluated (World Health Organization, 2016a). Among the vaccine candidates, RTS, S/AS01 is considered as the most advanced. In July 2015, the European Medicines Agency approved the vaccine as the first malaria vaccine with a trade name Mosquirix (Wilby *et al.*, 2012; Singh and Mehta, 2016). In October 2015, a pilot implementation of RTS, S/ASO1, in parts of three to five sub-Saharan African countries, was recommended by the WHO and has been approved for a pilot trial. Due to the low protective efficacy (<50%) and limited target group shown by RTS, S/ASO1 this vaccine will be used as a complementary tool and will not replace the proven malaria preventive, diagnostic and treatment measures. Therefore, the discovery and development of more effective malarial vaccines that will meet all the required criteria is still far off (World Health Organization, 2016; Mahmoudi and Keshavarz, 2017).

1.6 Post genomic era

The complete genome sequence of Plasmodium species has been made available, this include *P. falciparum* (Gardner *et al.*, 2002), *Plasmodium malariae*, *Plasmodium ovale* (Rutledge *et al.*, 2017), *Plasmodium knowlesi* (Pain *et al.*, 2008), *Plasmodium yoelii* (Carlton *et al.*, 2002) and *Plasmodium chabaudi* (Janssen *et al.*, 2001). It is been established that the Plasmodium genus has two extra chromosomal genomes; the mitochondrion, which is a linear 6 kb DNA and the apicoplast, a circular 35 kb DNA (Wilson and Williamson, 1997). The genome size of *P. falciparum* is 23 megabases and consists of 14 chromosomes; it encodes about 5,300 genes of which the majority is employed for immune evasion and host–parasite interactions (Cowman and Crabb, 2002). The availability of complete assembled and annotated sequences provided a strong basis for comparative genomics, transcriptomic and proteomic studies in Plasmodium species. This allows for the identification of functional elements and understanding the parasite's metabolic pathways and mechanisms for future drug and vaccine discovery (Sims and Hyde, 2006; Greenwood and Owusu-Agyei, 2012). The resulting data has been made available in public databases such as Gen Bank, EMBL, UniProt, SCOP and the Protein Data Bank (Toomula, *et al.*, 2012).

1.7 Anti-Malarial drugs

Chemotherapy treatment is the primary method in controlling the malaria disease. The main anti- malarial drugs used in this treatment include chloroquine, quinine, sulfadoxinepyrimethamine and artemisinins. These drugs act against either schizonts or gametes at the tissue level (Delves *et al.*, 2012). The use of choloroquine (CQ) for malaria treatment has dropped due to the parasite strong resistance. As a result, greater use of sulfadoxine-pyrimethamine and artemisinin derived chemotherapies is encouraged. Artemisinin and its derivatives (ARTs) have proven to be a successful chemotherapy drug (Gopalakrishnan and Kumar, 2015). It acts against the trophozoite ring stage of the parasite infection cycle. The mechanism of action is believed to involve the cleavage of the Endoperoxide Bridge an active moiety of artemisinin derivatives by a source of Fe^{2+} or heme from the parasite, resulting in formation of oxy- radicals. Theses radicals are then re-arranged into primary or secondary carbon-cantered radicals that act as alkylating agents of the parasite heme and proteins, hence killing the parasite (Yang, Little and Meshnick, 1994). The selective toxicity of artemisinin derivatives is explained by the high level of the parasite intracellular heme, which also explains some of the adverse side effects of this drug (Gopalakrishnan and Kumar, 2015).

Artemisinin based combination therapy (ACT) has been used for roughly a decade. ACTs replaced the use of ARTs especially when the levels of parasite resistance increased against ARTs mono-therapies (Whitty *et al.*, 2008; Mbengue *et al.*, 2015). Sulfadoxinepyrimethamine drugs are the most chosen replacement for malaria treatment and regarded as safe and cost effective. Sulfadoxinepyrimethamine inhibit the formation of hemozoin from reactive heme cofactors, resulting in the accumulation of free toxic heme in the parasite cells and eventual death of the parasite (White, 1998). The efficacy of sulfadoxinepyrimethamine depends on its synergy: it has independent mechanisms of action in which two enzymes are targeted in the parasite folate-synthesis pathway (Chulay, atkins and Sixsmith, 1984; Kakar *et al.*, 2016).

The Malaria parasite has shown the highest resistance against anti-malarial drugs such as chloroquine (White, 2004). Countries in Asia have also reported parasite resistance to current ACTs (Grimberg and Mehlotra, 2011). This is considered as alarming especially since the anti-malarial control depends mostly on ACTs.

With the problem of drug resistance on the rise, the need to develop new anti-malarial treatment strategies has become crucial (White, 2004; Grimberg and Mehlotra, 2011). In addition to that, current chemotherapeutic treatment is associated with many obstacles such as government regulations, the absence of an approved vaccine, increased cost of insecticides and the resistance of the anopheles mosquitoes to the insecticides.

1.8 Introduction to the folate pathway

The folate pathway has been established to be an important pathway for the malaria parasite survival (Hyde, 2005). The enzymes of this pathway have been targeted for the treatment and prevention of malaria disease for over half a century (Yuthavong *et al.*, 2006). Folate is found naturally as conjugated pterins, composed of a heterocyclic pterin ring connected to paraAminoBenzoic Acid (pABA) and at least one glutamate residue (Figure 1.2). The products of the folate pathway supply one carbon unit for the metabolic pathways; biosynthesis of methionine, of purines and of pyrimidines, these pathways are important for DNA synthesis, cells division and growth. In addition, protein synthesis in mitochondria significantly depends on the carbon unit supplied by this pathway.

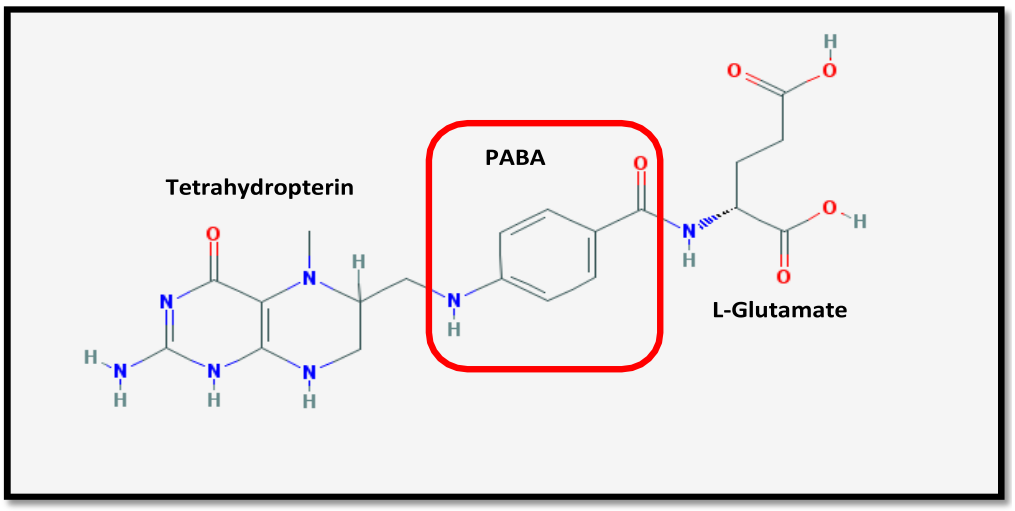


Figure 1.2: Folate simple precursor: Pterin ring that exists in tetrahydro form, p-AminoBenzoic Acid (pABA) and L-glutamate. Adapted from (Hyde, 2005).

Folate derivatives are important cofactors for DNA generation hence the malaria parasites depend on it for their survival (Salcedo-Sora and Ward, 2013). The inhibition of enzymes participating in this pathway can adversely affect the parasite.

Lower eukaryotes and some prokaryotes have the ability to synthesize folate from its simple precursors GTP, pABA and L-glutamate whereas higher organisms such as humans do not have the ability to do so and depend on dietary intake of pre-formed folate. Plasmodium species have the ability to exploit both of these means in which the folate is utilized either from a *de novo* pathway or exogenous folate salvage pathway, where the folate is taken up from the surrounding environment. Neither the detailed mechanism by which this occurs, nor the extent of the parasites ability to balance between the *de novo* and salvage two pathways, is known (Hyde, 2005).

The first reaction of the folate biosynthesis pathway is catalyzed by the enzyme GTP CycloHydrolase I (GCH1); it converts GTP of the basic folate moiety into dihydroneopterin triphosphate (DHNP) (reaction in Figure 1.3). In Plasmodium species DHNP is used for folate synthesis whereas in mammals it is used as a precursor for the synthesis of neopterin and biopterin derivatives. The reduction of biopterin produces tetrahydrobiopterin (THB) a cofactor for nitric oxide synthesis. Reduction of neopterin in Plasmodium species begins with the product 7, 8 dihydroneopterin triphosphate which is converted to 2-amino-4- hydroxy-6-hydroxymethyl dihydropterin in the presence of dihydroneopterin aldolase (DHNA).

Hydroxymethyldihydropterin pyrophosphokinase (PPPK) catalyzes the diphosphorylation of 2-amino-4-hydroxy-6-hydroxymethyl dihydropterin resulting in the formation of 6-hydroxymethyl-7, 8-dihydropterin pyrophosphate, an active pyrophosphorylated intermediate. Dihydropteroate synthase (DHPS) forms a carbon-nitrogen bond between the pterin ring moiety and pABA forming dihydropteroate. Dihydrofolate synthase then catalyzes the addition of glutamate to dihydropteroate forming dehydrofolate. The last step involves the conversion of dihydrofolate into tetrahydrofolate by dihydrofolate reductase (Nzila *et al.*, 2005; Müller and Hyde, 2013).

The folate derivative 5, 10 methylene-tetrahydrofolate plays a major role in the conversion of deoxyUridine MonoPhosphate (dUMP) to deoxyThymidine MonoPhosphate (dTMP) by providing the methyl group needed for the reaction; therefore, it is crucial for DNA replication and cell division. Each molecule of dTMP results in the oxidation of the THF molecule to DHF, and then recycled by dihydrofolate reductase (DHFR) back to the THF form (Hyde, 2005). The full pathway is illustrated in Figure 1.3.

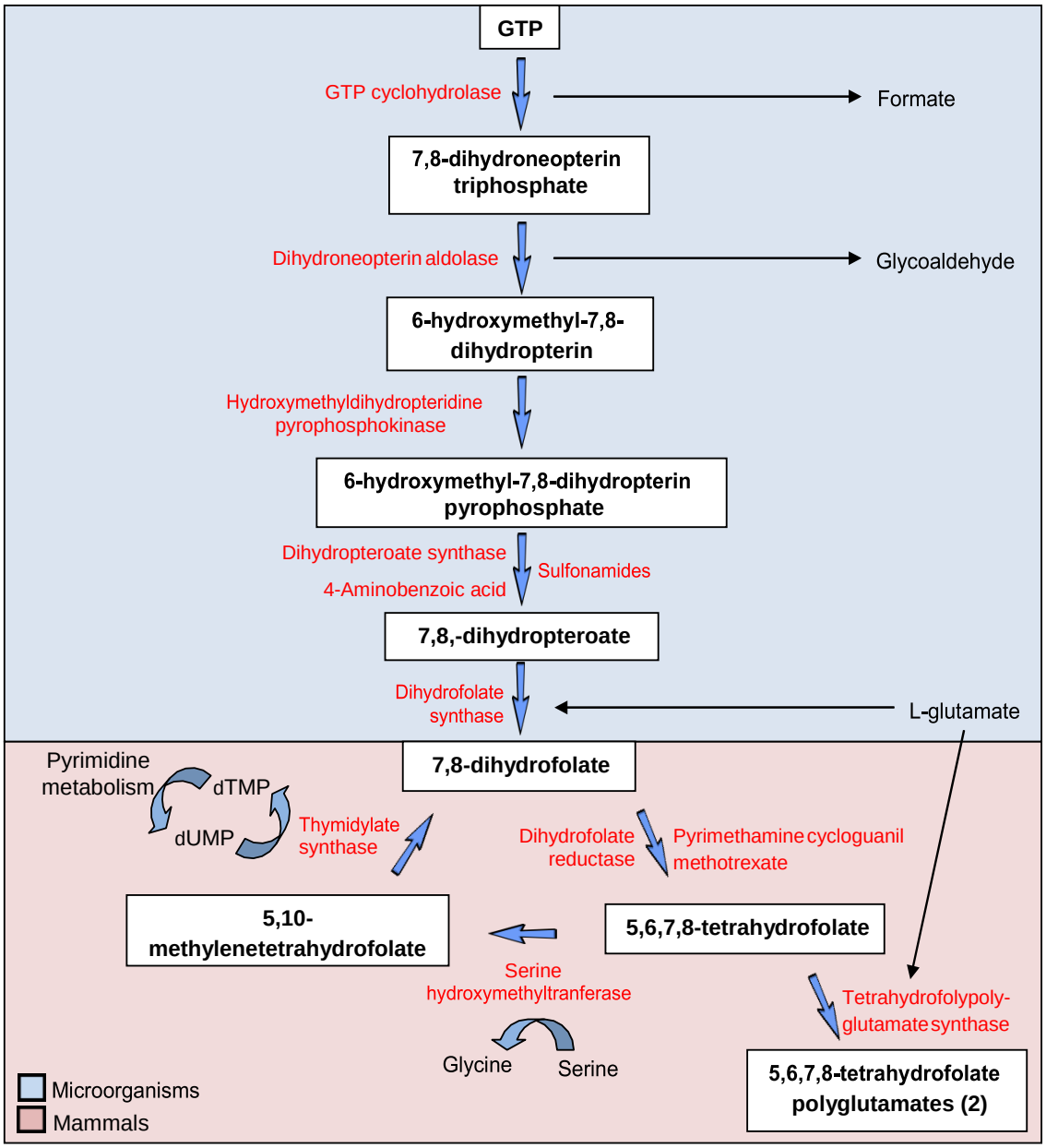


Figure 1.3: Biosynthesis of the folate moiety and enzymes catalyzing this pathway in Plasmodium spp. GTPCH, DHNA, PPPK, DHPS and DHFS. The Enzymes used in the one-carbon unit transfer reactions are DHFR, TS, SHMT and FPGS. Adapted from, Crider *et al.* (Crider *et al.*, 2012).

1.9 Antifolate drugs

Antifolate drugs have been effective in the inhibition of the *P. falciparum* liver stage. Because of the increased parasite resistance to chloroquine, antifolate drugs became the most common alternative with an affordable price. Antifolate drugs have shown some efficacy in cancers chronic inflammatory pathologies, viral and bacterial infections (Jarmuła, 2010).

Existing studies have shown that antifolate drugs pyrimethamine and cycloguanil are responsible for the inhibition of the DHFR enzyme activity of the parasite folate biosynthesis pathway, resulting in the interruption of dTMP and methionine synthesis, therefore killing the parasite (Basco, Ramiliarisoa and Le Bras, 1994). Proguanil, Sulfadoxine and Dapsone are also proven to inhibit the activity of the DHPS enzyme, using analogues of p-aminobenzoic acid, sulfonamides and sulfones as competitive inhibitors (Hammoudeh *et al.*, 2013). However, recent studies presented the parasite's developed resistance against available anti-folate drugs (Grimberg and Mehlotra, 2011).

1.10 GCH1: A search for new antifolates

GCH1 enzyme catalyzes the biosynthesis of several cofactors such as formic acid and dihydroneopterin triphosphate in prokaryotes and tetrahydrobiopterin (BH4) in mammals (Thöny, Auerbach and Blau, 2000). BH4 is known as a reducing cofactor for nitric oxide synthase and other enzymes in mammals. Partial purification of this enzyme was first done by Burg and Brown over 30 years ago, in which the name of the enzyme was also given (Burgs and Brown, 1968). This was followed by extensive studies of the enzyme biochemical properties in various species. The studies were limited by the instability of the GCH1 enzyme which also explains the low-resolution diffraction data obtained from GCH1 enzymes of different species. However, the availability of a thermostable protein from *Thermus.thermophilus (T.thermophilus)* reduced the encountered limitations and allowed for more structural and functional analyses (Omi *et al.*, 2003). The first solved structure of the GCH1 enzyme was of *Escherichia coli* (Nar, Huber, Meining, *et al.*, 1995). The *E.coli* enzyme structure was first solved in the absence of the coordinated zinc. Zinc coordinated structures were then solved for *E.coli* bacteria, as well as the human analogue five years later (Auerbach *et al.*, 2000).

The structure of GCH1 consists of two symmetrical pentamers; each has five chains making the protein structure homo-decameric. The centre of the protein exhibits a β-barrel flanked by α- helices (Figure 1.4a, b). GCH1 has ten zinc containing active sites; each active site is buried in a deep pocket of 10 Å between three adjacent subunits (Figure 1.4c). The zinc ion in the active site is coordinated to one His and two Cys residues (Nar, Huber, Auerbach, *et al.*, 1995) (Figure 1.4d). A fourth coordination was identified and proposed to be a water molecule (Yoko Tanaka *et al.*, 2005). The water molecule is situated near the zinc ion in a distance of 3.2 Å. It was also demonstrated that the water molecule changes its position and moves nearer to the ion during catalysis, getting activated by the zinc ion.

GCH1 catalytic residues are highly conserved; mutation of these residues results in loss of the protein activity (Rebelo *et al.*, 2003). A study by Tanaka *et al*., on the *T.thermophilus* GCH1 protein, proposed the protein reaction mechanism. The study established that the entrance of the enzyme active site pocket holds positively charged residues such as Arg64, Arg137 and Arg183 and one lysine residue Lys134 (Figure 1.4d). When the substrate binds to the active site it is usually surrounded by neutral and hydrophobic residues. Part of the substrate is in contact with the water molecules near the protein surface. This results in the formation of a hydrogen-bond network around the active site. The active site His residue participates in the bond-cleavage of the substrate and the zinc ion participates in the activation of the water molecule which is suggested to attack C-8 of the substrate (Rebelo *et al*. 2003; Tanaka *et al*. 2005).

Folate derivatives are essential for the parasite cell division and growth. A study on *P. falciparum* showed that transcription of the GCH1 enzyme peaks during the early trophozoite stage (Nirmalan *et al.*, 2002). In addition, gene knock out studies of GCH1 enzyme higlighted the importance of this enzyme in folate synthesis for the parasite growth and survival, as it failed to produce a knockout line (Witter *et al.*, 1996; Müller and Hyde, 2013).

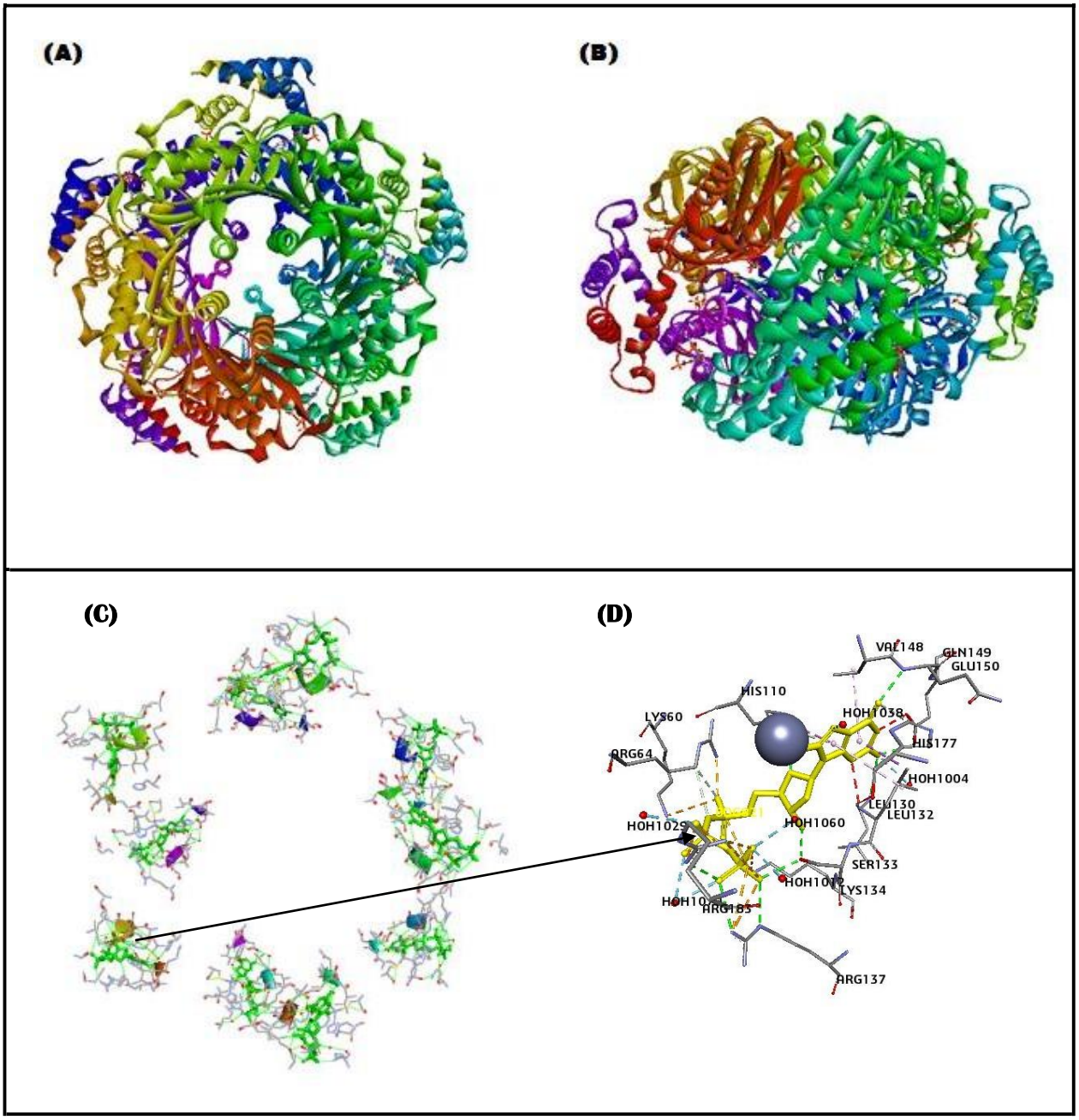


Figure 1.4: (A) top view and **(B)** side view of *T.thermophilus* GCH1 homo-decameric structure (PDB ID: 1WUR). The protein chains are coloured differently. The figure was generated using Discovery Studio. **(C)** A view of the *T.thermophilus* GCH1 protein ten active sites, substrates are shown in green at their corresponding binding site and the interacting residues shown as grey sticks. **(D)** Enlarged view of the GCH1 active site, the zinc ion is shown in grey, 8-oxoguanine derivative of GTP co-cystallized ligand is shown in yellow and the surrounding key residues shown as sticks. Adapted from Tanaka *et al.* (Yoko Tanaka *et al.*, 2005).

1.11 Project motivation

1.11.1 Problem statement and justification

Malaria is a worldwide health problem. The Malaria parasites have developed an increased resistance to the majority of available anti-malarial drugs, which has raised a great challenge in anti-malarial drug discovery (Grimberg and Mehlotra, 2011). However, the advancements in computational biology and availability of genomics and transcriptomics data have accommodated this dilemma, by allowing for the identification of thousands of new drugs with known molecular specificity and activity against the parasite metabolic enzymes (Xia, 2017). These approaches are considered to be time and cost effective.

Targeting the malaria parasite folate biosynthesis pathway has been proven to be a powerful strategy for malaria treatment (Sibley *et al.*, 2001). The GCH1 is proposed to be the rate limiting enzyme in folate synthesis (Kümpornsin *et al.*, 2014) yet; the enzyme is not well established as a drug target for the treatment of malaria. The Plasmodium GCH1 would be an ideal drug target. However, the presence of a human homolog along with a highly conserved active site creates some challenges in designing a drug specific to the Plasmodial GCH1 protein. Sequence and structural analysis allow for investigating the GCH1 protein and highlighting key residues that can distinguish the Plasmodial GCH1 enzyme for selective inhibition. Molecular Dynamics simulations, in combination with molecular docking, are used to assess the binding of the identified inhibitor compounds during structure based virtual screening. In addition, the absence of GCH1 zinc ion force field parameters is an obstacle; therefore, appropriate parameters should be developed and validated to obtain accurate consensus simulations.

1.11.2 Aim

The aim of this study was to use computational approaches to identify suitable compounds from the South African Natural Compounds Database (SANCDB) with potential inhibitory activity against the Plasmodium GCH1 enzyme and investigate their interactions.

1.11.3 Objectives

1. Sequence retrieval and alignment
2. To carry out phylogenetic analysis
3. To identify conserved and unique regions of the Plasmodial GCH1 sequences
4. To build a homology model of GCH1 protein 3D structures
5. Molecular Docking of compounds from the South African Natural Compounds Database
6. To identify potential inhibitor compounds
7. To perform molecular dynamics simulations

1.11.4 Overview of the Methodology

The sequence of GCH1 protein was retrieved from biological databases. This was followed by sequence analysis for the identification of sequence features and conserved regions of Plasmodium GCH1 enzyme that can highlight distinctive functional characteristics. The structural consequences of the unique sequence features were examined by homology modelling. Docking experiments were designed to perform a high throughput virtual screening on the compounds from the South African Natural Compounds Database (SANCDB) (Hatherley *et al.*, 2015). Molecular dynamics simulations were performed using the Chemistry at Harvard Macromolecular Mechanics (CHARMM) molecular dynamics simulation and analysis computer software package (Brooks *et al.*, 2009).

1.11.5 Sequence analysis

GCH1 protein sequence from mammals, prokaryotes and the Plasmodium genus were retrieved from the PlasmoDB (http://plasmodb.org/plasmo/) and UNIPROT (http://www.uniprot.org) databases. Multiple sequence alignment (MSA) was performed using different alignment tools such as PROMALS3D (Pei and Grishin, 2014), MAFTT (Katoh *et al.*, 2002) and MUSCLE (Edgar, 2004). The BLOSUM62 scoring matrix was used as a substitution matrix for the sequence alignment. The MSA outputs from different alignment tools were evaluated to determine their accuracy. For visualization of the sequence alignment, JalView software was used (Waterhouse *et al.*, 2009). Local re-alignment and sequence trimming was applied in order to increase the alignment accuracy.

1.11.6 Motif discovery

Motif discovery was performed using the MEME suite (Bailey *et al.*, 2015). The PyMOL molecular graphics system was used to map the identified motifs onto the protein structure (DeLano, 2014) and a heat map was generated to show the identified motifs of all sequences.

1.11.7 Phylogenetic analysis

Molecular Evolutionary Genetic Analysis (MEGA7.2) software was used for phylogenetic analysis (Kumar, Stecher and Tamura, 2016). This was important for the identification of the evolutionary relationship of the conserved domains of the GCH1 enzyme. The best three models were selected according to the lowest Bayesian information criterion (BIC) score. A phylogenetic tree was constructed for each of the selected models then compared to determine the robustness of the tree construction process.

1.11.8 GCH1 homology modelling

The Plasmodium GCH1 structure was identified from its sequence then analyzed using sequence and structural approaches. The MODELLER computer program for homology modelling was used to model the Plasmodium GCH1 structures (Šali, 2013). Final models were refined when necessary and validated before proceeding further.

1.11.9 Virtual screening

High throughput screening of SANCDB compounds was performed using the docking programs AutoDock Tools4 for the setup and AutoDock Vina for the docking (Morris *et al.*, 2009; Trott and Olson, 2010). SANCDB contains around 700 natural compounds from 166 different organisms of aquatic and land-based origin. 60% of SNACDB compounds met the conditions of Lipinski's rule of five. These natural compounds have not been screened yet for potential activity against GCH1 for malaria treatment. The virtual screening was carried out on the GCH1 protein of all the five human infective Plasmodium species as well as the human GCH1 protein. The compounds with low binding free energy to Plasmodium GCH1 protein were also docked to the human GCH1 protein. Compounds that show selective binding to the Plasmodium GCH1 were regarded to have potential use in drug development against malaria. Discovery Studio software was used to visualize the complexes and the docking pose of the ligands. Systemic diagrams of protein-ligand interactions were generated using the LigPlot tool (Laskowski and Swindells, 2011) and Discovery Studio software (San Diego: Accelrys Software Inc., 2012).

1.11.10 GCH1 Zn parameter determination

GCH1 protein contains a zinc ion, located in the active sites. It was important to develop appropriate force field parameters to obtain an accurate consensus simulation. Potential energy surface (PES) scans, using quantum mechanics were employed to generate the required force field parameters. These calculations were performed using Gaussian 09 (Frisch, M. J *et al.*,2009).

1.11.11 Force field parameter validation

The computational resources of the Centre for High Performance Computing (CHPC) was used to accommodate the computational cost associated with performing such calculations. A validation protocol for the force field parameters was developed to ensure the accuracy of the obtained results. The simulations were performed using CHARMM software package (Brooks *et al.*, 2009).

CHAPTER TWO

SEQUENCE ALIGNMENT AND ANALYSIS

GCH1 enzyme is well-conserved in bacteria, protozoa, plants and animals (Tatham *et al.*, 2009). The Plasmodium GCH1 is considered to be an ideal drug target for the treatment of malaria. However, it is not well established as a drug target and the 3D structure of the parasite's protein is still unknown. This chapter is focused on the identification and sequence analysis of the *P. falciparum* GCH1 protein sequence and its homologs in Plasmodium species, namely: *P.vivax, P. knowlesi, P.ovale P.malriae, P.berghei, P.yoelii and P.chabaudi* as well as other prokaryotes and mammals. *P.flaciparum* GCH1 protein sequence and its homologs were retrieved and analysed through multiple sequence alignment (MSA) and phylogenetic analysis. This was done to discover conserved motifs and key residues that are potentially involved in the protein function. Plasmodium sequences were compared against their human *homologs* in order to reveal potential structural characteristic differences between the parasite and the human protein which can aid in anti-malarial drug design.

2.1 Introduction

Proteins play a fundamental role in biological pathways of living organisms. This involves enzymatic catalysis, structural support, endocrine function and storage. Hence, it is important to understand their structures and mechanism of action (Alberts *et al*. 2002; Xiong 2006). Rapid development in sequencing technology has provided a fast and high throughput sequencing which allows for resourceful retrieval of amino acid sequences. The genomic and transcriptomic sequences have been studied extensively to understand the proteins encoded in several organisms, which can provide broad information about proteins' functionality (Tatusov, Koonin and Lipman, 1997).

Proteins that diverge from common ancestral genes share significant level of sequence similarity, structural and functional properties (Mindell and Meyer, 2001). The term homology is defined as the evolutionary relatedness of protein families which can be identified from the protein amino acid sequence. There are three types of *homologs*, namely: orthologs, paralogs and xenologs. Orthologs are homolog genes that result from speciation evolutionary events, while paralogs result from a duplication event and xenologs result from horizontal gene transfer between different species.

Orthologs are important in the prediction of protein structure and function because they maintain similar functional characteristics as they evolve (Mindell & Meyer 2001; Koonin 2005). The first step of characterising the structure of an unknown protein involves comparing its sequence against sequence databases to find other homologs that share some similarity. The identified homologs can then be used as a reference for sequence and structural analysis of the unknown protein (Iorio *et al.*, 2010).

Several biological databases have been developed and made publicly accessible. The biological databases store sequence and structural data of various organisms making it easier to retrieve the required data. Biological databases are divided into three main categories: primary, secondary and specialized databases. Primary databases contain raw sequence data and the secondary databases contain curated and annotated data (Xiong, 2006). Specialized databases focus on data of specific organisms of research interest such as PlasmoDB database (http://plasmodb.org/plasmo/). Examples of main primary biological databases are the National Centre for Biotechnology Information (NCBI) database (https://www.ncbi.nlm.nih.gov/) and the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB) (https://www.rcsb.org/).

2.1.1 Similarity based search tools

Sequence comparison is essential for functional and evolutionary relationship inference. This involves the comparison of a particular sequence of interest against the available sequences deposited in the biological databases (Kalaimathy, Sowdhamini and Kanagarajadurai, 2011). The availability of the biological databases has provided the opportunity to annotate, perform functional and structural analysis on homologs protein sequences that have not been characterised before. Sequence similarity search is based on sequence alignment and its alignment depends on the sequence similarity and identity of the matching residues that share similar physiochemical properties. In the end, sequence homology and structural similarity are deduced from a significant level of sequence similarity and identity (Xiong, 2006).

Basic Local Alignment Search Tool (BLAST) is one of the most popular tools used to search for homolog sequences. It employs a heuristic algorithm to rapidly find an optimal sequence alignment. The output of BLAST is displayed as a list of matching sequences ordered by their alignment score and expectation value (E-value).

The E-value measures the probability that the alignment is a random match from the database. Hence, a lower E-value represents high biological significance of the alignment; furthermore, a higher alignment score indicates a better sequence alignment (Altschul, 2005).

The BLAST tool can be accessed through different web servers, such as NCBI-BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and Uniprot (http://www.uniprot.org/blast/). The NCBI BLAST tool offers different databases search algorithms such as BLASTP, that screens protein databases using a protein query, BLASTN, which performs a nucleotide-nucleotide search to find more distant sequences, BLASTX to search a nucleotide query against a protein database while translating the query and TBLASTN, to search a protein query against a nucleotide database while translating the query (McGinnis and Madden, 2004). Some refinements have been applied to the BLAST tool in order to enhance its speed and sensitivity, this includes PSI-BLAST tool that makes use of position specific scoring matrices (PSSMs). When using PSSMs, amino acids substitution scores are allocated separately depending on their position in a protein multiple sequence alignment. As a result, this increases the sensitivity of PSI-BLAST in detecting remote homologs (Altschul, 2005).

The HHpred server (http://toolkit.tuebingen.mpg.de/hhpred) is another fast database search tool that provides a sensitive and speedy search. When compared to the PSSMs, HMM profiles are considered to have an increased sensitivity in the detection of homologs. HHpred server makes use of Hidden Markov Models (Hildebrand *et al.*, 2009). The search starts by querying the sequence against non-redundant databases using several PSI-BLAST iterations. This results in the generation of several sequence profiles, which are used for secondary structure prediction by PSIPRED (McGuffin, Bryson and Jones, 2000). The final alignment with the secondary structure prediction information is then used to create an HMM profile. This profile includes position specific amino acid, insertion and deletion probabilities (Söding, Biegert and Lupas, 2005).

In the alignment, each column is associated with an HMM profile holding the probabilities of a match, insert and delete state. The probabilities of an insertion and deletion are converted to specific gap penalties used in HMM alignments. The use of position specific gap penalties provides a more sensitive and accurate search. Finally, HMM-HMM comparison is performed between the consensus HMM profile and a selected database (Söding, 2005). The output of HHpred is displayed as the sequence alignment with its statistics, percentage identity, sequence similarity and the E-value.

2.1.2 Sequence alignment methods and algorithms

Sequence alignment is divided into global and/or local alignment. Global alignment spans the entire length of a query sequence in an attempt to align every residue to find the global optimum alignment, while local alignment aims to identify and align local regions of similarity between sequences that are often widely divergent (Altschul, 2005). Local alignment is often preferable for distantly related sequences that contain similar domains. Several computational algorithms have been developed and applied to improve sequence alignment (Notredame, 2007); such as dynamic programming which is based on solving a problem by breaking it down into sub-problems. This algorithm is effective in the optimization of sequence alignment. It assigns scores to matches and defines a certain cost for a residue being substituted by another or a gap then calculates the alignment score. In this case, a good alignment will have a high score (Nalbantoğlu, 2014). An example of an alignment algorithm that makes use of dynamic programming is the Needleman-Wunsch algorithm for global alignment (Needleman and Wunsch, 1970) and Smith-Waterman algorithm (Smith and Waterman, 1981) for local alignment.

Different scoring matrices are used to determine the probability of a residue being substituted by another in the sequence alignment. Examples of common scoring matrices used are the Point Accepted Mutations (PAM) (Dayhoff and Schwartz, 1978) and Blocks of Amino Acid Substitution Matrix (BLOSUM) (Henikoff and Henikoff, 1993). These matrices are very effective and have been shown to outperform matrices based purely on the physiochemical properties of amino acids. PAM matrices are based on the observation that amino acids sharing similar size, charge and hydrophobicity are more likely to substitute each other (Dayhoff & Schwartz 1978). Therefore, PAM matrices are used in aligning closely related sequences that share 70- 90% similarity.

Several PAM matrices have been developed such as PAM 50, PAM 100, PAM 160 and PAM 250. These matrices are based on real dataset observation and calculations of the likelihood of one substitution per 100 residues. For example, PAM 50 is derived by multiplying PAM 1 by itself 50 times, meaning that in 100 residues there are 50 substitutions. When the evolutionary distance between sequences increases, higher PAM matrices are used (Dayhoff & Schwartz 1978; Xiong 2006).

BLOSUM scoring matrix is based on the frequency of amino acid substitution. The matrix is derived from conserved blocks of sequence alignments. In BLOSUM matrices, sequences are clustered beyond a certain threshold. This is done to avoid similar sequences in a block to cluster together, thus producing bias in the alignment score. When aligning more distantly related sequences, this threshold is lowered to detect more distantly related sequences (Henikoff 1992). Several BLOSUM matrices have been developed including, BLOSUM62 for midrange similarity, BLOSUM80 for related sequences clustering at 80% similarity and BLOSUM45 for distantly related sequences.

2.1.3 Multiple sequence alignment

MSA alignment is important for the identification of conserved residues and motifs with functional importance. It also allows for understanding of ancestral relationships between organisms. Sequence alignment can be divided into local or global alignment. Local alignment finds local/sub regions of highest similarity between the sequences to find conserved sequence patterns, domains or motifs. Global alignment aligns the entire sequences end to end to identify the overall all similarity between the sequences and infer homology. Several algorithms and approaches have been developed to produce optimum and accurate sequence alignment. One of the algorithms used is the heuristic approach, which involves progressive alignments, iterative alignment and consistency-based alignment (Thompson *et al*. 1999). Progressive MSA is done in a stepwise manner by first aligning the closely related sequences and subsequently adding more distantly related sequences. The CLUSTAL MSA program builds the alignment progressively by performing an initial pairwise alignment between all possible sequences, then constructing a guide tree using either the Un-weighted pair group method with arithmetic mean (UPGMA) or the Neighbour- Joining (NJ) method. The guide tree is built based on the alignment scores and is used to build the MSA by aligning closely related sequences and adding sequences progressively according to their location in the guide tree (Chenna *et al*. 2003). T-COFFEE MSA program is known to have an increased accuracy. T-COFFEE is a consistency-based aligner that first creates a library containing local and global pairwise alignments. The pairwise alignments are generated by other alignment programs like CLUSTAL (Local alignments) and LALIGN (global alignments). The produced alignment is then evaluated by allocating weights to the alignments based on sequence identity. Both local and global alignment are combined and followed by library extension (Niun *et al*. 2006). This library is used to generate a position specific scoring matrix to construct a guide tree; the guide tree is then used to direct the MSA using the progressive approach (Notredame, Higgins and Heringa, 2000).

Profile multiple alignment with local structure (PROMALS3D) is an MSA alignment program that performs both sequence and structural alignments (Pei and Grishin, 2014). A pairwise alignment of similar sequence is first carried out to give groups of pre-aligned sequences. From this pre-alignment, a representative sequence is selected from each group. PSI-BLAST is then used to search its homologs from Uniprot non-redundant reference databases facts (UNIREF90) and Protein Structure Prediction server (PSIPRED) for secondary structure prediction (Pei and Grishin, 2014). Amino acid sequence profiles are derived from the PSI-BLAST and PSIPRED and used to come up with a consistency scoring function. The representative sequences are then aligned progressively using the consistency scoring function and all the pre-aligned sequences are incorporated to form a complete MSA.

Multiple sequence alignment based on Fast Fourier Transform (MAFFT) has the advantage of speed and production of accurate MSA. MAFFT uses fast Fourier transform (FFT) analysis to rapidly detect segments of sequence similarity; FFT algorithm aligns the sequences by progressive alignment based on a guide tree (Katoh and Standley, 2013). The algorithm of this alignment involves a preliminary alignment using the progressive method and then refines it iteratively to produce an optimal alignment. This is followed by calculating the distance between all pairs of sequences to be aligned which is used to build a guide tree. The guide tree is generated using UPGMA method. The alignment is further refined by generating a better distance matrix based on the pairwise alignments of the initial guide tree. The new distance matrix is used to generate a guide tree for progressive alignment; this process is iterated until no better scoring alignment is attained (Katoh and Standley, 2013).

Multiple Sequence Comparison by Log-Expectation (MUSCLE) MSA program is based on the refinement of the progressive method in order to produce an optimal alignment. It involves two sets of progressive alignments. First is a draft alignment based on a guide tree constructed using the sequences before they are aligned, then a more accurate guide tree is constructed to produce a second alignment. The alignment is then further refined using a profile function (Edgar, 2004).

2.1.4 Phylogenetic analysis

Phylogenetic analysis is important to understand the evolutionary relationship of protein sequences. This is done by studying their evolutionary divergence, which is usually represented by a phylogentic tree. Phylogenetic trees are constructed using multiple sequence alignment. These trees show the evolutionary divergence or similarity of sequences involved by providing a schematic view of how various species have evolved.

In order to approximate the evolutionary distances between sequences, different evolutionary /statistical models are used; this is followed by converting the calculated evolutionary distances into a distance matrix which is then used by an algorithm to generate the phylogenetic tree (Xiong 2006; Yang & Rannala 2012). Statistical models like Jones- Thornton-Taylor (JTT) and Dayhoff are considered to be accurate in calculating the evolutionary distance as they account for amino acid substitution between sequences and correcting for homoplasy (shared similarities among taxa which is not present in their common ancestor) (Le *et al*. 2008; Yang & Rannala 2012). One of the distance-based algorithms used in constructing the phylogentic tree is the Neighbour joining (NJ) algorithms. NJ algorithm joins all taxa into a single node forming a star-like tree, the most closely related pair forms the first node and the next closely related taxa is linked to the first node, this is done progressively until all taxa have been added creating a complete tree (Saitou and Nei, 1987).

An optimal tree topology is ensured by generating several trees each with a different initial pair. Then the optimal tree that fits the evolutionary distance is selected. For validation, the resulted phylogenetic tree is statistically evaluated using a bootstrap test. Bootstrapping test for robustness by iteratively altering the data set to generate trees which are compared against the original tree. Bootstrap values are indicative of the confidence levels of the topology (Hillis and Bull, 1993). One of the common programs used for phylogenetic analysis is MEGA7 (Kumar, Stecher and Tamura, 2016). It includes tools for DNA and protein sequence alignment, evolutionary distance calculation and phylogentic tree construction. Sequence alignment can be performed using a built in MUSCLE or CLUSTAL alignment program and it can also accept aligned data. MEGA7 provides different evolutionary/statistical models that are used to estimate the evolutionary distances between the protein sequences. It also allows for visualisation of the constructed tree via a Tree Explorer (Kumar, Stecher and Tamura, 2016).

2.1.5 Motif analysis

Sequence motif is a distinctive pattern of nucleotide or amino acid sequences with a biological significance. Structural motifs occur in the exon region of the gene while others occur outside, expressing regulatory/recognition function over the sequences. Several computational methods have been developed for motifs discovery. This includes the Multiple Expectation Maximisation for Motif Elicitation (MEME) tool, which is used for the identification of biologically functional motifs. MEME software discovers motifs using expectation maximisation which allows for parameter estimation in probabilistic models with incomplete data to fit a two-component finite mixture model (Bailey *et al.*, 2015).

MEME software returns sequence motifs as a sequence logo with their corresponding scores, it also provides the E-value of each motif. The E-value indicates the statistical significance of the discovered motif; it is based on the log likelihood ratio of the returned motif that has the same width and site count (Bailey and Charles, 1994). Each motif sequence is shown as a line overlaid with block diagrams in colours. This shows the motif number and its scores according to the positional p-value. The p-value indicates the probability that a random sequence has an equivalent match score or higher (Bailey *et al.*, 2015).

MEME sequence logo contains a stack of letters at every position in the motif. The height of the letters represents the probability (in bits) of the letter occurring at that position multiplied by the number of times that residue occurs within that site in each motif site in the total dataset (Bailey *et al.*, 2015). When the residue site is not well conserved at a particular position in the motif the height of the stack is reduced (Kyte and Doolittle, 1982). The colours of the individual letters in the motif are based on the biochemical properties of the amino acids (Table 2.1).

MAST calculation is done in parallel to MEME jobs within the MEME suite (Bailey *et al.*, 2015). It calculates the pairwise correlations between each pair of motifs; this is done to determine the probability that two motifs are significantly different. After trying all possible motif pairs, the sum of Pearson's correlation coefficients for the aligned columns is divided by the width of the shortest motif in the pair. When the correlation value of the motif pairs is high they are considered similar and will not be treated as separate motifs (Bailey *et al.*, 2015).

Table 2.1: MEME amino acid colour codes for sequence logos, from (Kyte & Doolittle 1982).

Amino acids	Colour	Colour Properties
A C F I L V W M		Hydrophobic
N Q S T		Polar, non-charged, non-aliphatic
D E		Acidic
K R		Positively charged
H		Positively charged, cyclic
G		Non-polar
P		Cyclised
Y		Non-polar, aromatic

2.2 Methods

2.2.1 Sequence retrieval

P. falciparum GCH1 protein sequence was retrieved from PlasmoDB database. *P. falciparum* sequence was used as a query to identify other Plasmodium homologs in PlasmoDB database and other homolog sequences from Uniprot database (PlasmoDB, 2001; Bateman *et al.*, 2017). Sequences of Plasmodium genus: *P.vivax* (PVX_123830), *P.knowlesi* (PKNH_1443200), *P.ovale* (PocGH01_14049700), *P.malariae* (PmUG01_14058300), *P.berghei* (PBANKA_1438900), *P.yoelii* (PY17X_1441400), *P.chabaudi* (PCHAS_1440900), *P.reichenowi* (PRCDC_1223300) and *P.gaboni* (PGSY75_1224000) were retrieved from PlasmoDB. Uniprot BLAST tool was used for the search, with default alignment parameters of BLOSUM-62 scoring matrix and an expectation value (E) threshold of ten. The most probable GCH1 homolog sequences with significantly low E-values, from the three kingdoms animalia, fungi, prokaryote and the Plasmodium genus, were downloaded and saved as a FASTA format files for use in sequence alignment.

2.2.2 Sequence alignments

Multiple sequence alignment was performed using the selected sequences, followed by comparative analysis of Plasmodium sequences against their human homolog. This was important for the identification of key residues that were conserved within the Plasmodial GCH1 protein and to provide distinctive structural/functional information.

Multiple sequence alignment of all retrieved sequences was carried out using alignment tools such as PROMALS3D (http://prodata.swmed.edu/promals3d/promals3d.php), T-COFFEE (http://tcoffee.crg.cat/), MAFFT (http://toolkit.tuebingen.mpg.de/mafft/) and MUSCLE (https://www.ebi.ac.uk/Tools/msa/muscle/). The alignment tools were accessed via web servers and sequence alignment was performed using the default parameters. The fasta file containing all sequences was uploaded into the MSA programs and run. One 3D structure sequence of GCH1 protein of *E.coli* (PDB ID: 1WM9) was added to the alignment file. The sequence alignment was performed in two stages: Firstly, the entire sequences were used to give a view of the local regions of similarity. Secondly, the alignment was performed using the catalytic domain sequences to obtain an optimal sequence alignment for the region of interest. Visualization of the alignment was done using Jalview alignment viewer (Waterhouse *et al.*, 2009). The residues were coloured by their physiochemical properties: green-hydrophobic, red-polar acidic, blue-polar basic, black-polar uncharged. Sequence identity between aligned sequences was captured and represented in a heat map generated by MATLAB (The Mathworks Inc., 2016).

2.2.3 Phylogenetic analysis

MEGA7 software was used to generate a Neighbour Joining phylogenetic tree representing the evolutionary relationships between the Plasmodium sequences and its orthologs. PROMALS3D alignment was used for the tree generation. All positions containing gaps were completely eliminated. The evolutionary model used to calculate the distances was selected based on MEGA7 goodness of- fit test, in which the data is measured by BIC score (Beaumont & Rannala, 2004). The appropriate model usually appears among the top three (Tamura *et al.*, 2011). The parameters used were Le Gascuel statistical model and 1000 replicates of bootstrapping analysis.

2.2.4 Whole protein motif analysis

MEME suite was used to discover motifs of the selected sequences. The search was done with a distribution expectation of zero to one occurrence per sequence and a maximum number of 36 motifs. The width of the motifs was set to a minimum of six and maximum of 50. Search parameters were set to skip repeated matching of motifs.

2.3 Results and Discussion

2.3.1 Sequence retrieval

Conserved regions in a sequence hold potential functional and structural importance. Comparing the protein sequence against a non-redundant protein sequence database provide necessary information to understand the relationship between sequences and to characterize their structural/functional properties. *P. falciparum* sequence was first used as a query to retrieve *homolog* sequences of Plasmodium genus then used to search other homologs sequences of non-Plasmodium. The sequence identity of the Plasmodium species ranged from 45% to 98% relative to *P.falicparum*, this point towards the relatedness of these sequences. It was observed that *P.reichenowi* shares the highest similarity with *P. falciparum* sequence and *P.vivax* being the least similar. Bacterial and fungal sequences shared a sequence identity between 40 % and 48 % relative to *P.falicparum*. Mammalian sequence were the least similar sharing a sequence identity between 35 % and 36 %. It was also observed that the sequence identity in mammals was consistent (Table 2.2). In general, 30% sequence identity is required to report homolgy (Özlem Tastan Bishop *et al*. 2008; Schmidt *et al*. 2014). This is subjected to the length of aligned sequences. According to the sequence identity >30% over more than 100 residues and an E-values < 1.60E-27 we deduced that the retrieved sequences were homologs.

Table 2.2: Summary of *P. falciparum* GCH1 sequence and its homologs retrieved from PlasmoDB and Uniprot data.

Species name	Accession number	E-Value	Identity
P. falciparum	PF3D7_1224000.1-p1	0.00	100.00%
P. vivax	PVX_123830	2.00×10^{-82}	45.00%
P. malariae	PmUG01_14058300	$2.00\text{x}10^{-85}$	68.00%
P. ovale	PocGH01_14049700	$6.00\text{x}10^{-95}$	70.00%
P .knowlesi	PKNH_1443200	$7.00\text{x}10^{-87}$	69.00%
P. chabaudi	PCHAS_1440900	$2.00\text{x}10^{-85}$	70.00%
P. yoelii	PY17X_1441400	$3.00\text{x}10^{-87}$	70.00%
P. gaboni	PGSY75_1224000	0.00	86.00%
P. berghei ANKA	PBANKA_1438900.1-p1	$4.00\text{x}10^{-88}$	68.00%
P. reichenowi	PRCDC_1223300	0.00	98.00%
Colletotrichum orchidophilum	A0A1G4BKI0	$6.80\text{x}10^{-30}$	44.10%
Blumeria graminis f. sp. tritici 96224	A0A061HDS3	$1.60\text{x}10^{-31}$	43.20%
Pseudogymnoascus sp. VKM F-4517 (FW-2822)	A0A094GQ82	$4.30\text{x}10^{-30}$	43.20%
Fusarium oxysporum f. sp. vasinfectum 25433	X0M7M4	$1.90\text{x}10^{-30}$	44.40%
Rhodothermus profundi	A0A1M6TV79	$5.40\text{x}10^{-35}$	47.80%
Gramella sp. LPB0144	A0A1L3J3H7	$3.10\text{x}10^{-35}$	45.00%
Candidatus Kaiserbacteria bacterium	A0A1F6FZQ3	$3.80\text{x}10^{-36}$	48.40%
Mucilaginibacter	A0A142HP81	$1.60\text{x}10^{-38}$	41.00%
Pan troglodytes (Chimpanzee)	H2RBI2	$3.40\text{x}10^{-28}$	35.40%
Oryctolagus cuniculus (Rabbit)	G1SIY3	$1.60\text{x}10^{-27}$	35.40%
Mus musculus (Mouse)	Q05915	$1.70\text{x}10^{-29}$	35.30%
Rattus norvegicus (Rat)	P22288	$5.30\text{x}10^{-28}$	36.00%
Homo sapiens (Human)	P30793	$3.40\text{x}10^{-28}$	35.40%

All sequences retrieved were from the GCH1 family and T-folds super-family, this was confirmed via NCBI database and InterProScan server. InterProScan takes the protein sequences as an input query and uses predictive models, known as signatures. The signatures are derived from diverse databases referred to as a member database and made into a single searchable resource (Zdobnov and Apweiler, 2001). The functional domains corresponding to *P. falciparum* GCH1 were identified and highlighted by InterProScan server. The functional domain sequence started at position 200 to 378 representing the GCH1 domain. Other matches were obtained from an unintegrated signature including the superfamily of tetrahydrobiopterin biosynthesis enzymes-like; starting at position 171 to 375 (Figure 2.1). The pre-domain and catalytic domain regions of the sequences are indicated in Table 2.3.

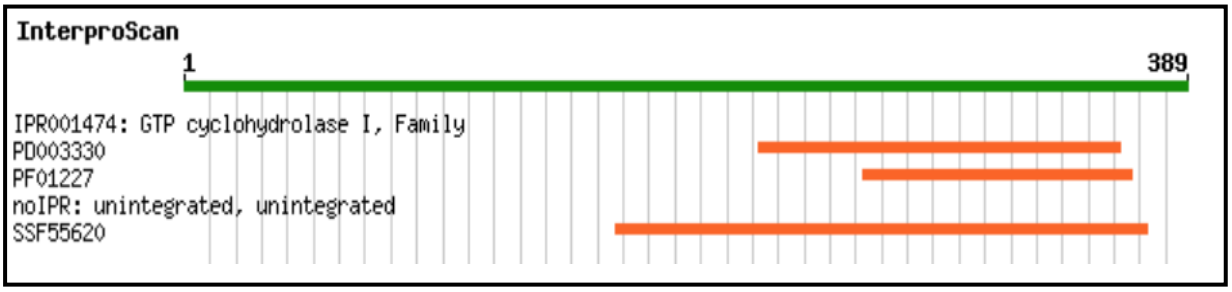


Figure 2.1: Predicted functional domains, protein family and important sites of *P. falciparum* GCH1 protein from InterProScan server (http://www.ebi.ac.uk/Tools/pfa/iprscan/).

Table 2.3: Positions of the catalytic domain within the whole protein sequences of different *P. falciparum* homolog sequences.

Species name	Protein sequence length	Catalytic domain
P.falciparum	1-389	200-378
P.vivax	1-423	226-389
P.ovale	1-390	191-351
P.knowlesi	1-451	256-416
P.malariae	1-459	266-426
P.chabaudi	1-275	90-242
P.berghei	1-293	108-260
P.gaboni	1-372	189-365
P.reichenowi	1-389	206-378
P.yoelii	1-315	130-299

2.3.2 Multiple sequence alignment and analysis

Multiple sequence alignment was carried out using MUSCLE, MAFFT, T-COFEE and PROMALS3D alignment programs. Due to the fact that the alignment programs are not entirely error free, different programs were used to obtain a consensus with regards to the alignment. It was observed that PROMALS-3D program provided the most accurate /optimal alignments (contiguous region of the alignment with minimal gaps) followed by MAFFT alignment (Figure 2.2& 2.3). This is due to the fact that PROMLAS3D incorporates sequence and structural information from the sequences included in the alignment. A significant level of conservation was observed in the functional domains of Plasmodium sequences and high variation in the N terminal and C terminal regions. This variation is explained by the unique features of the terminal regions among these sequences (Appendix 1, Figure A1.3).

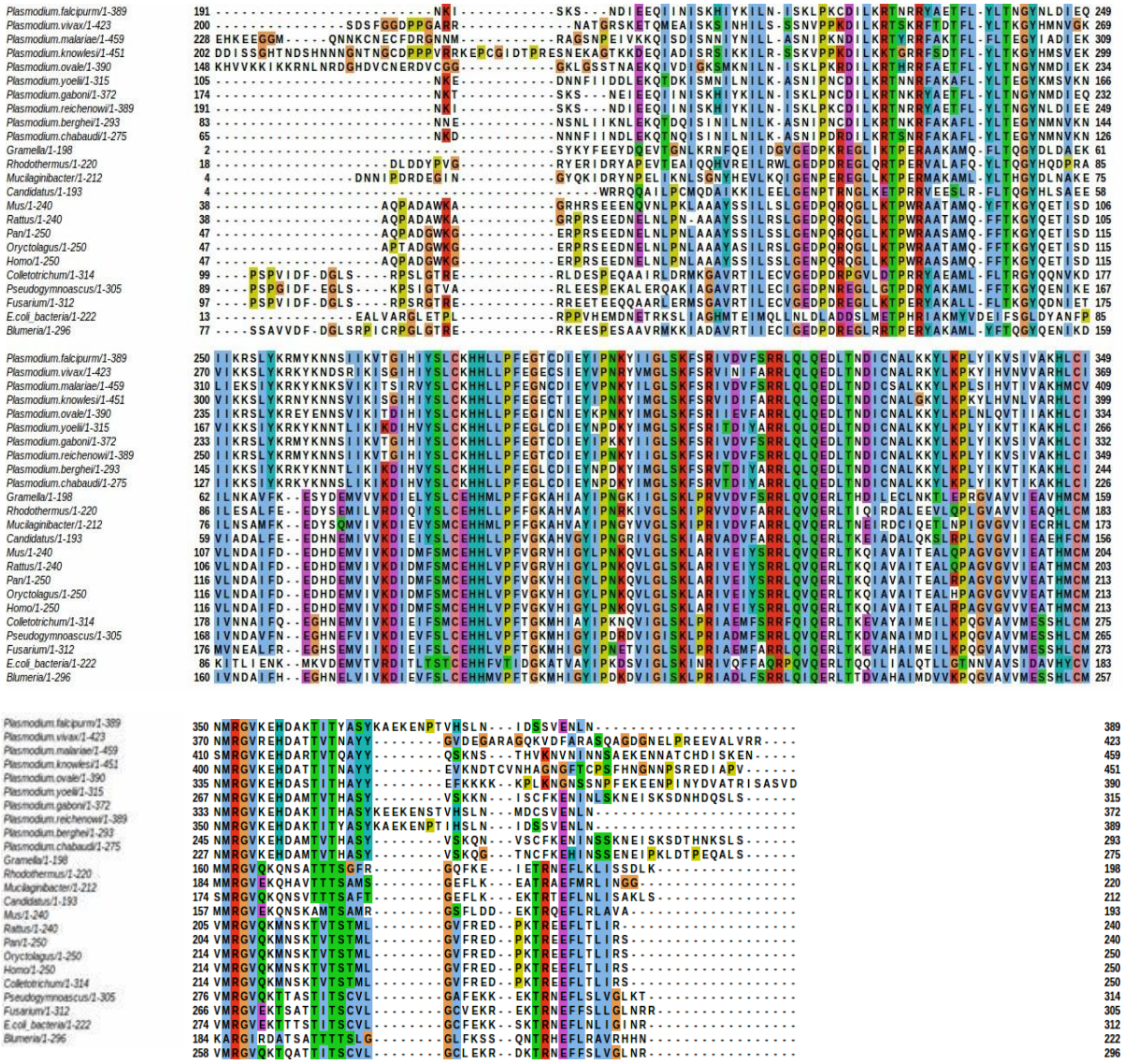


Figure 2.2: PROMALS3D multiple sequence alignment of the GCH1 complete sequence (The N terminal region was trimmed). Residues are shaded by conservation using Jalview alignment editor tool (Waterhouse *et al*. 2009). High sequence variation was observed from the N and C-terminals of all GCH1 sequences. The catalytic domain is fairly conserved among all species. Plasmodium genus showed a significant level of residues conservation. This can differentiate the Plasmodium sequences from the rest of bacterial, fungal and mammal sequences.

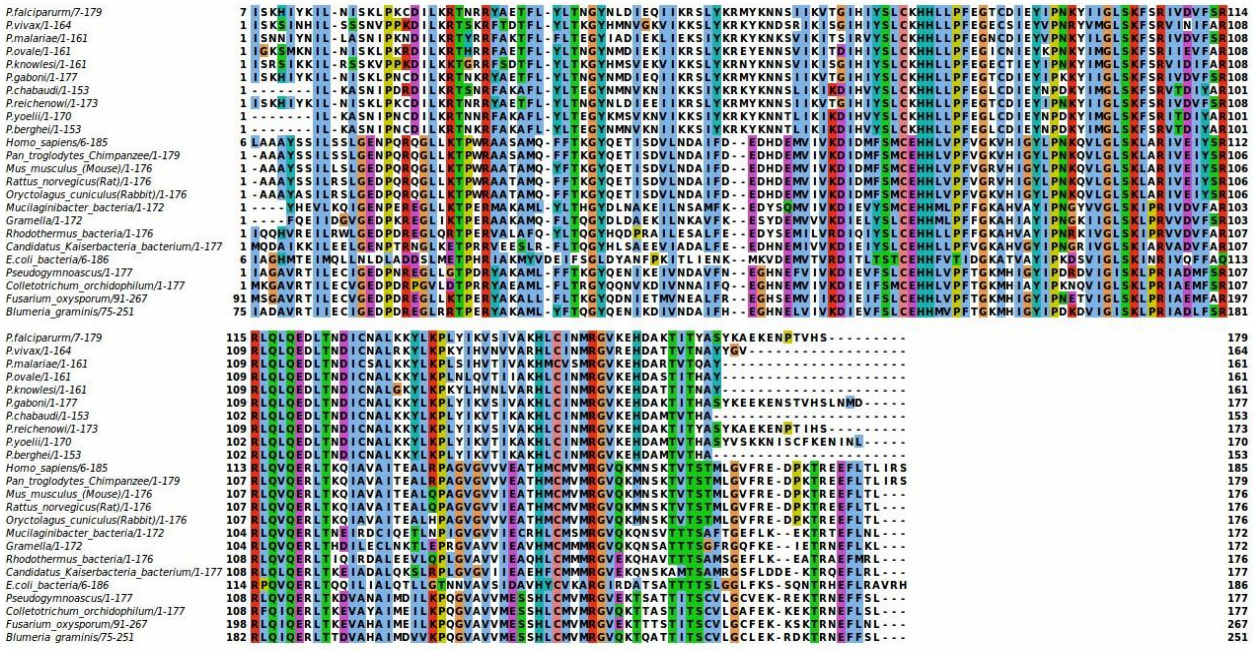


Figure 2.3: MAFTT Multiple sequence alignment of GCH1 protein catalytic domain only. Residues are shaded by conservation using Jalview alignment editor tool (Waterhouse *et al*. 2009). The catalytic domain was well conserved in all species however; Plasmodium genus showed a considerable level of residues conservation.

For the sequence alignment, a contiguous region of the alignment with minimal gaps was selected as the dataset. Structural information obtained from PROMALS3D and T-COFEE was mapped onto the catalytic domain sequences, in which the conserved residues were closely examined and differences between *Plasmodium* and human homologs were identified (Figure 2.5). Residues conservation was observed among all sequences more specifically, in the active site catalytic residues (Figure 2.6).

The function of proteins is often described by its spatial configuration and type of amino acids at a particular site (Pils, Copley and Schultz, 2005). The conserved residues in *P. falciparum* were mapped onto the protein structure in order to identify their location within the structure (Figure 2.4). The conserved residues were found to be in the N terminal α_1, α_2, α_3 helices and in the anti-parallel β sheets of the protein body as well as the α_4 and α_5 helices flanking the β sheets. Additionally, residues were relatively conserved among all species in the α_6 helix. This helix is located at the centre of the protein surface and known to form the entry point for the substrates. (Chapter 3, Figure 3.10). The properties of these conserved residues are explained in the next section.

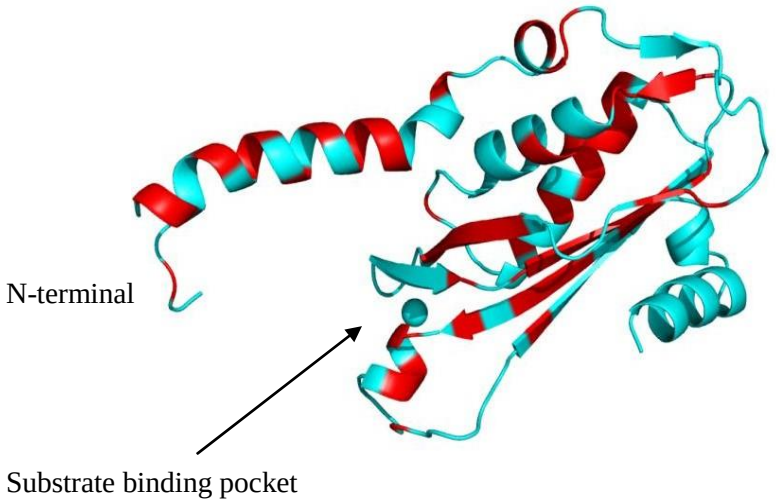


Figure 2.4: Model structure of *P. falciparum* GCH1 protein (Chain A) with its conserved residues mapped onto the structure and marked in (red).

MSA is important to underline structural and functional position related characteristics of a protein in a visual format (Do and Katoh, 2008). The sequence alignment output was viewed using Jalview. Aligned residues were coloured using different options to identify certain features of the GCH1 protein. (Appendix 1, Figure A1.3) shows the first panels until position 200; it was observed that this region relating to the GCH1 protein N terminus possesses the least conservation and it was the least well aligned as it holds the most insertions and gaps sites. It was also observed that Plasmodium species have significantly extended N-terminal sequences. The N terminal region was found to be highly divergent and variable in length. This region is proposed to be involved in regulatory activity such as for species-specific protein–protein interactions (Witter *et al*. 1996).

The catalytic domain exhibited high conservation of hydrophobic residues which are known to provide structural integrity (Appendix 1, Figure A1.2). *P. falciparum* catalytic residues His 346, His 279, His 280, Cys 277, Cys 348 (polar), Pro 283, Phe 284 and Gly 286 (hydrophobic) were conserved in all sequences. His 280, Cys 277 and Cys 348 were strongly conserved and demonstrated to coordinate the protein zinc metal (Yoko Tanaka *et al.*, 2005).

Lys 303, Arg 306, Arg 352 and Arg 231 were also conserved in all sequences and known of their interactions with the substrate phosphate groups (Kümpornsin *et al.*, 2014). Charged amino acids were fairly conserved in all sequences as they are often exposed to the solvent and considered to be important for the stabilization of the protein three-dimensional structure.

MSA results in the identification of up to 70 amino acid residues that were unique to all Plasmodium sequences. These residues were highlighted in green (Figure 2.6). A number of these residues were substituted by residues sharing similar properties such as Tyr 332, Leu 317, Ile 247, Phe 304 (hydrophobic) being substituted with Ala, Ile, Gly and Leu (hydrophobic) respectively. Other residues were substituted by dissimilar residues such as charged residues that were conserved in all Plasmodium species being substituted by polar residues in other species; this includes Arg 230 substituted by Trp (polar), Asp 246 substituted by Ile (Hydrophobic), Asp 320 substituted by Gln (polar), Glu 285 being substituted by Val (hydrophobic). Lys 330 was substituted by Thr (polar) in mammals, Met (hydrophobic) in fungi and Gln or Asn (polar) in bacteria with exception to *Rhodothermus* bacteria having Glu (charged) at that position (Figure 2.6). Lys 345 was found in all Plasmodium sequences except for *P.vivax* and *knowlesi* which have Arg in that position; these were substituted by Thr (polar) in mammals, Ser (polar) in fungi and in bacteria substituted with Gln, Val and Glu. Lys 355 (charged) was substituted by Gln (polar).

Plasmodium species also hold conserved polar residues such as Ser 305 which was substituted by Ala and Pro (hydrophobic) in bacteria and fungi sequences whereas E coli kept an Asn (polar) at that position. His 272 was substituted by Asp (charged) and Cys 288 was substituted by hydrophobic residues; Val in mammals, Ala in bacteria and Met in fungi sequences. Tyr 332 was also conserved in Plasmodium sequences and substituted by Ala in mammals, Val in bacteria and Ile in fungi. Cys 326 was substituted by Alanine (hydrophobic). Val 343 and Ile 154 hydrophobic residues were found in all Plasmodium sequences except for (*P.chabudi, P.gaboni and P.yolii)* in which the latter had a conserved Lys (charged) residue. The conserved hydrophobic residues of Plasmodium sequence at that position were substituted by Asp and Glu (charged). This analysis was done at the protein sequence level and it would be important to visualise the residues interaction structurally.

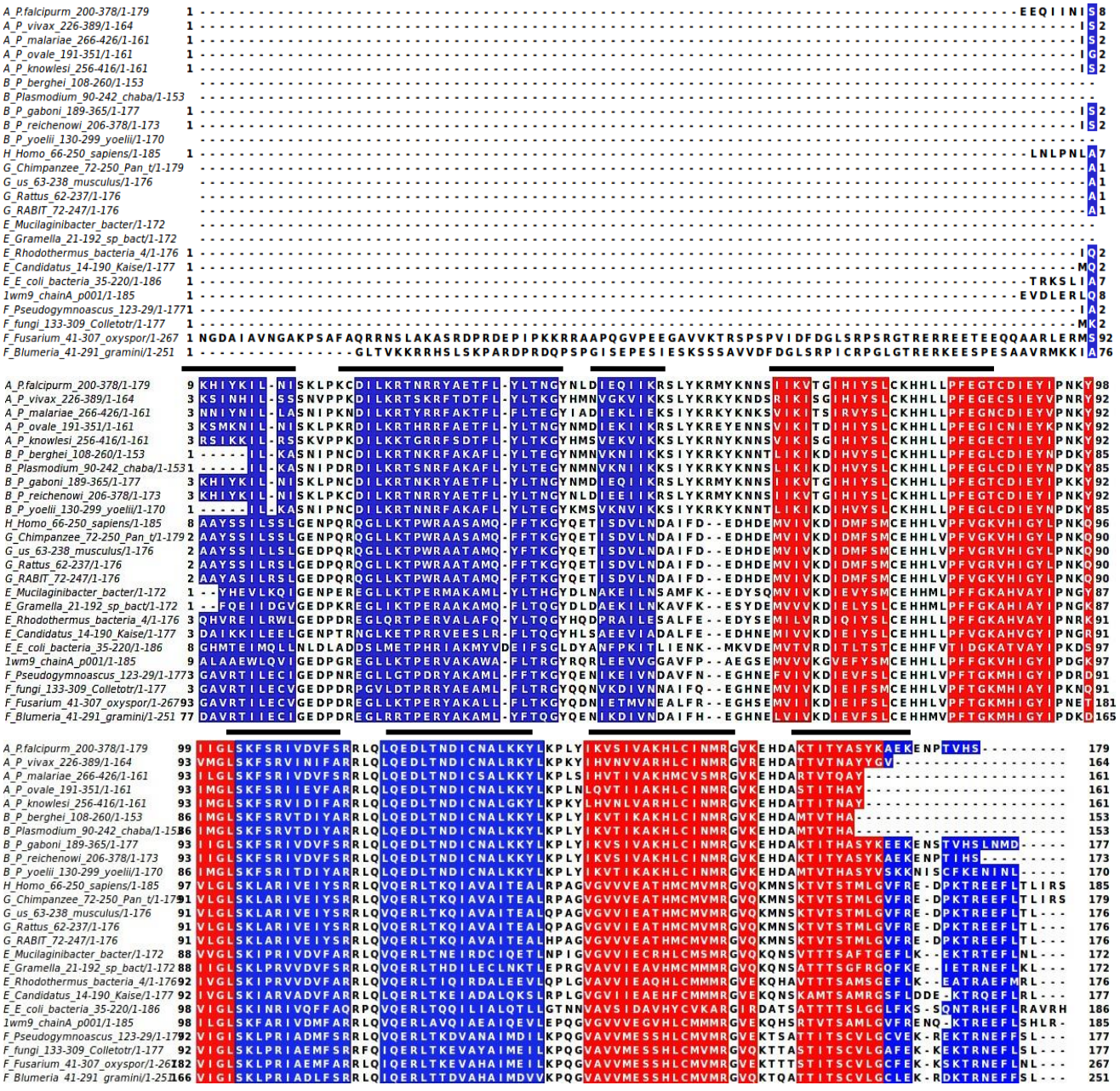


Figure 2.5: Structural information mapped onto the mature domain sequence. Helices are shown in Blue and beta-sheets in red.

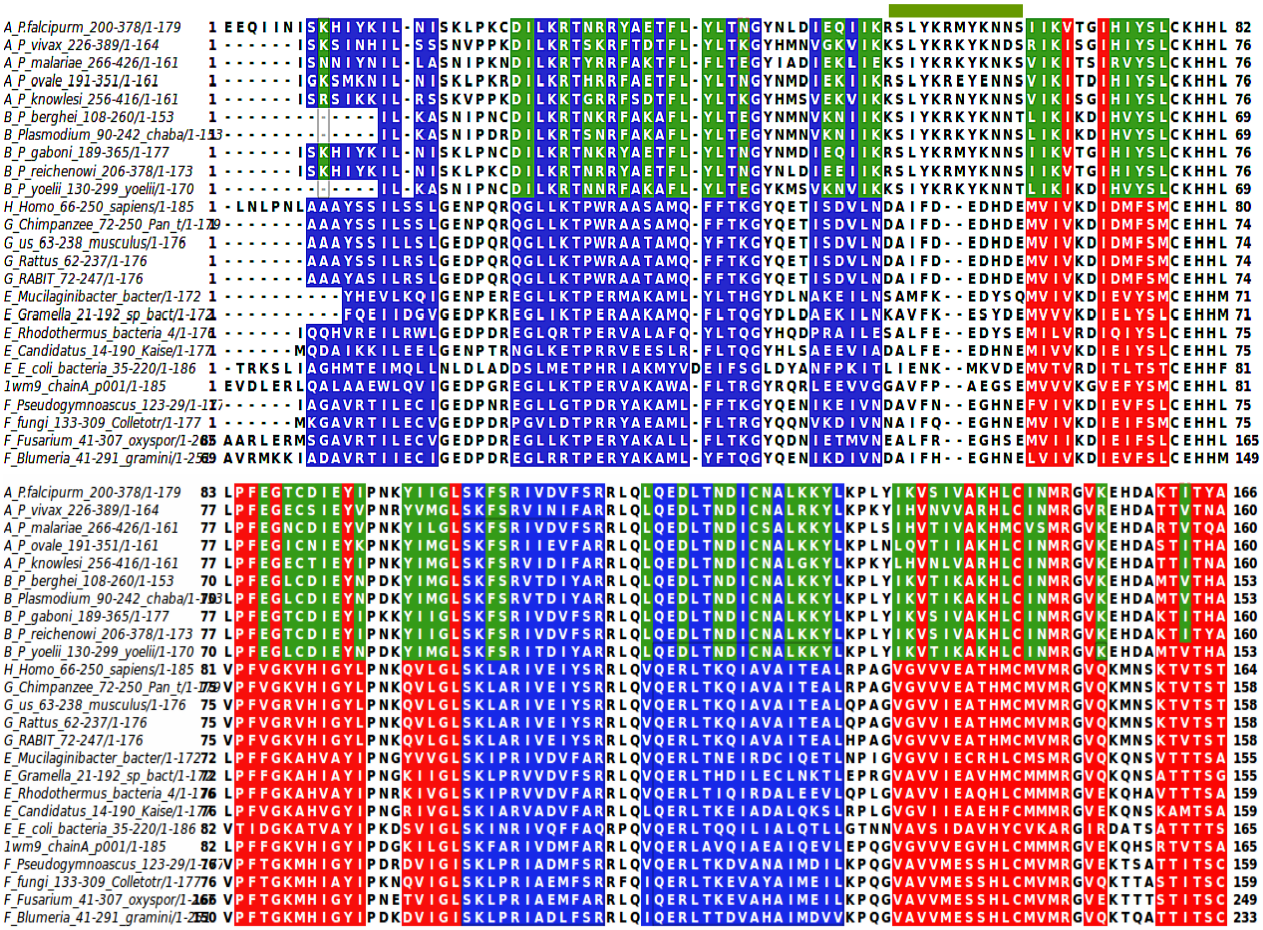


Figure 2.6: Conserved residues of Plasmodium species marked (green) in comparison to their homologs. To identify the location of the conserved residues within the structure sequence α helices are shown in blue and β sheets in red.

The sequence identities were captured and presented in a heat-map (Figure 2.7). The heat-map provided an overview of the relationship between Plasmodium sequences and other sequences of prokaryotes, fungi and mammals. Plasmodium species GCH1 protein sequences were significantly diverged from other sequences. This can point towards sequence and functional properties that can distinguish the Plasmodium GCH1 protein for future selective inhibition.

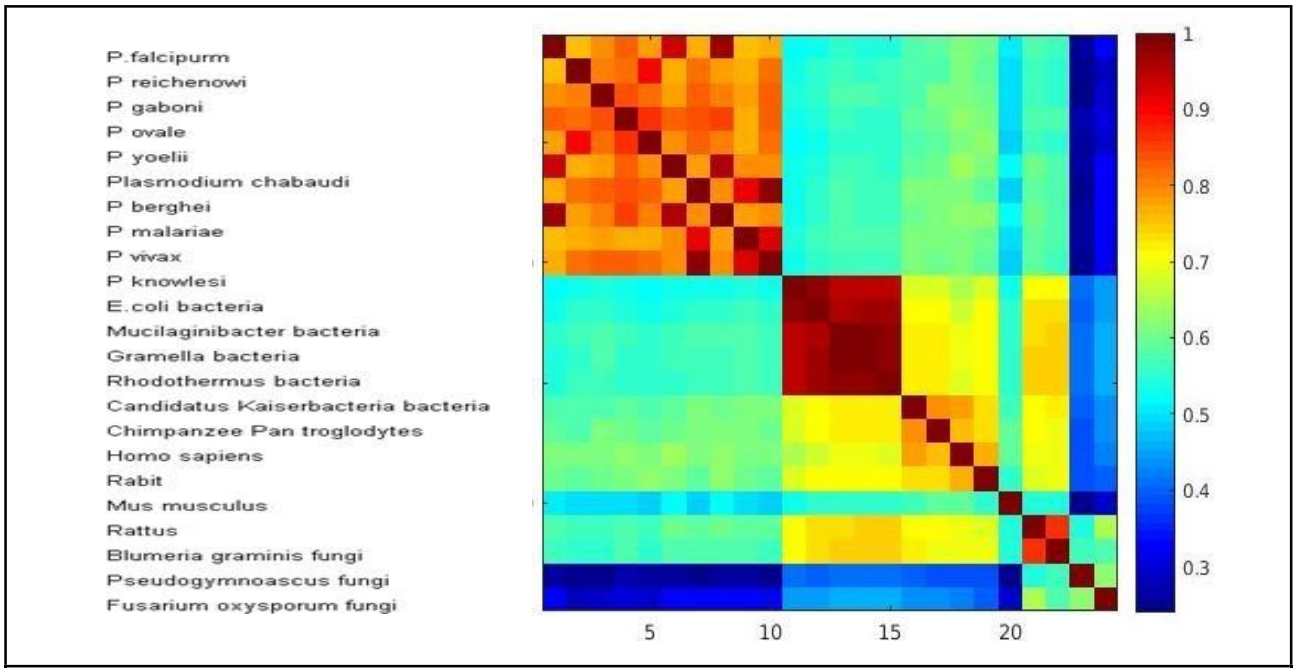


Figure 2.7: Sequence identity heat map generated using MATLAB. The colour of each element shows the level of sequence identity among the sequences. The most similar sequences are shown in red and the least similar in blue.

2.3.3 Motif analysis

The motif search resulted in the identification of 36 motifs; the discovered motifs had low E-values indicating their statistical significance. Motif one and two occurred in all sequences. Motif three occurred in all sequences except for *E. coli* GCH1 protein (Figure 2.8). Plasmodium species exhibited conserved motifs which were lacking in other sequences. The following motifs only occurred in the Plasmodium species, motif 7, 8, 9, 11, 13, 14, 15, 16, 17 and 19 (Figure 2.8). *P. falciparum* conserved motifs are tabulated in Table 2.4. It was observed that majority of the discovered motifs occur in the N terminal region. This region was not modelled in the Plasmodium species, therefore could not be mapped onto the structure. Only motifs 1, 2, 3 and 17 were mapped onto the structure (Figure 2.9). As majority of the discovered motifs were found in the N terminal region this implies the uniqueness of this region which hold potential regulatory functions (Y. Tanaka *et al.*, 2005). Motif 17 was located in the catalytic domain near the C terminal region in the last α helix of the protein structure. This motif has distinguished *P. falciparum* GCH1 protein from its human homolog.

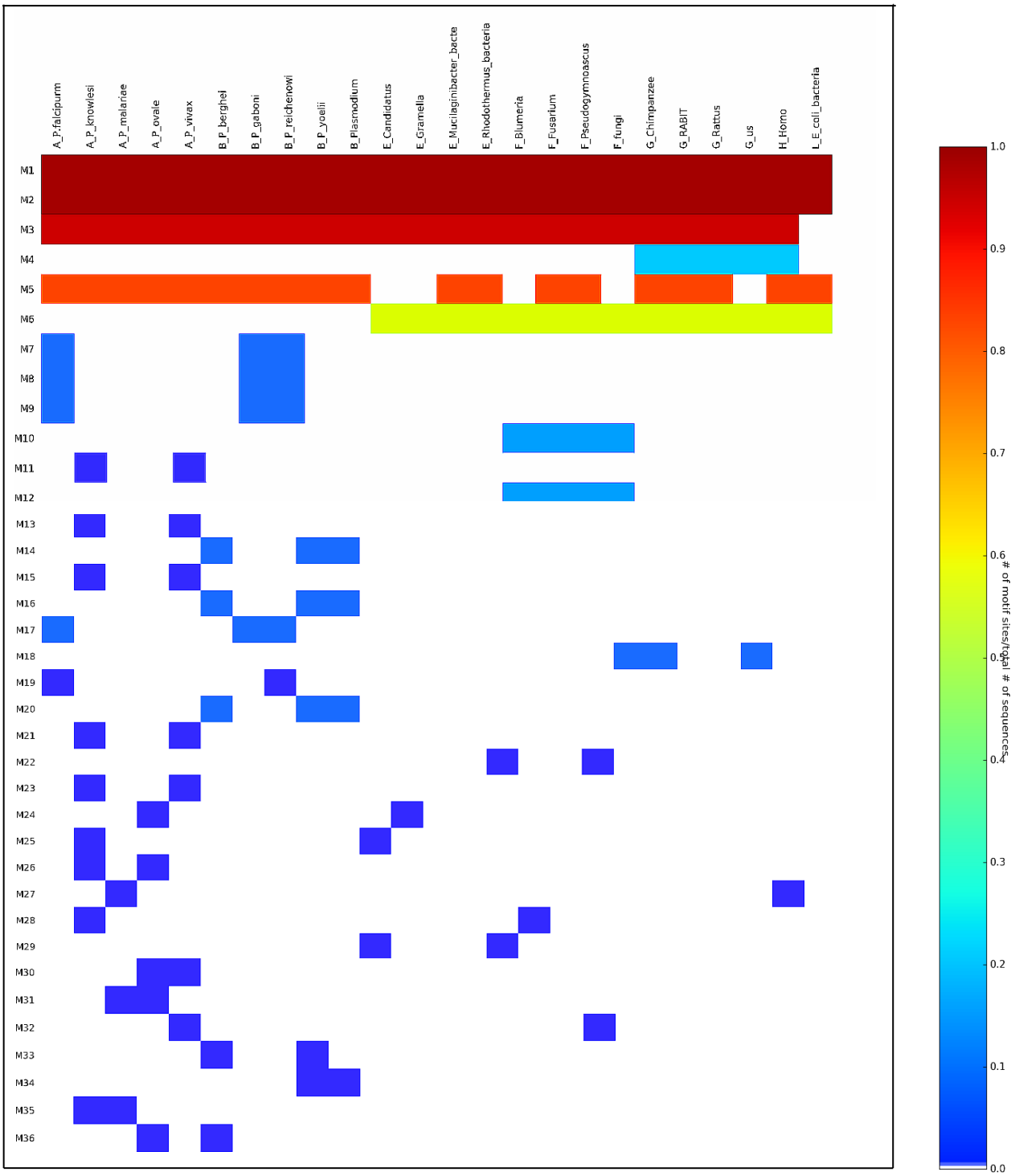
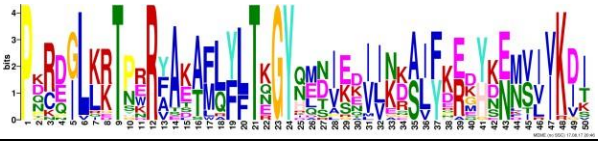
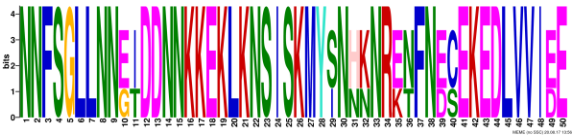
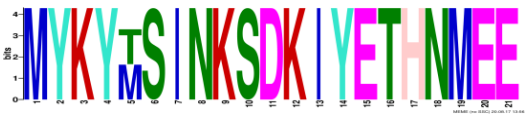


Figure 2.8: MEME heat map summarizing motif information for group of GCH1 proteins. The white regions show sequence lacking a motif. The level of conservation increases from blue to red.

Table2.4: Sequence logo of motifs found in full length GCH1 protein using MEME.

Motifs	Regular expression	E-value
Motif 1		$3.6x10^{-830}$
Motif 2		$6.3x10^{-593}$
Motif 3		$4.2x10^{-542}$
Motif 5 **(Terminal region)**		$9.6x10^{-053}$
Motif 7 **(Terminal region)**		$1.8x10^{-035}$
Motif 9 **(Terminal region)**		$5.4x10^{-016}$
Motif 8 **(Terminal region)**		$6.4x10^{-025}$
Motif 17		$3.3x10^{-003}$
Motif 19 **(Terminal region)**		$4.3x10^{-002}$

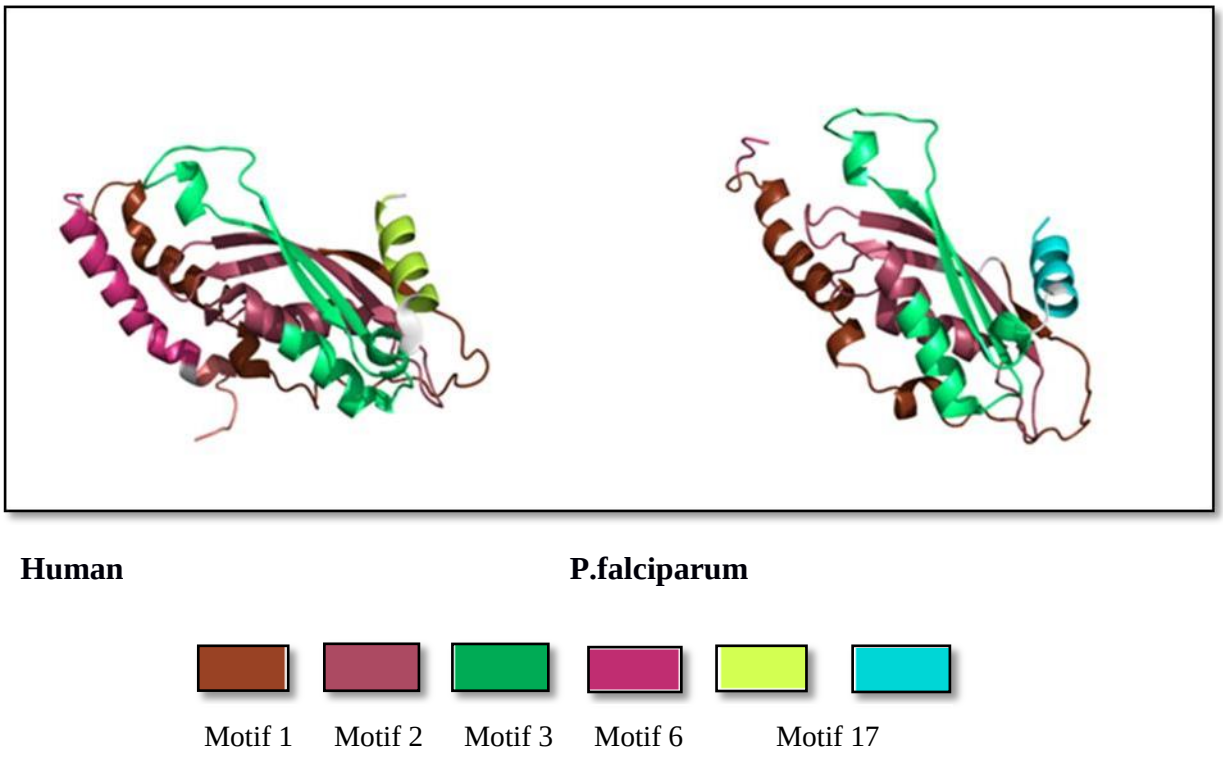


Figure 2.9: Motifs mapped onto the respective protein structures of the human GCH1 (PDB ID: 1FB1) and *P. falciparum* model using PyMOL. Motif 6 occurred only in the human structure. This region was not modelled in *P. falciparum* structure.

2.3.4 Phylogenetic analysis

The evolutionary relationship between the Plasmodium and human GCH1 protein sequences was studied. Several trees were constructed using different evolutionary models. The best tree was selected based on the model BIC scores, bootstrap consensus and consensus with other models. In future, a gene tree of all selected proteins can be built to investigate the reliability of the generated tree, this was not feasible due to time constrains. The generated phylogenetic tree showed five distinct groupings of Plasmodium, *E.coli*, bacteria, fungi and mammals. Within the Plasmodium species; *P. vivax* and *P. Knowlesi* were seemingly divergent from the other Plasmodium species while *P. falciparum* showed more relation to *P. richenowi.* This correlates with the sequence identities observed earlier (Figure 2.7). It was also observed that *P. berghei, P.chapudi* and *P.yoelii* branch from a common ancestor indicating their relatedness (Figure 2.10). The phylogenetic tree also showed that *E.coli* is distinctly different from other bacteria and the mammalian sequences were grouped together, being the least similar to the Plamsodial sequences. This may point out for evolutionary differences that could be important in designing inhibitors for selectively targeting the Plasmodial protein.

Results obtained from the phylogenetic analysis agreed with the sequence analysis results. Therefore, we can conclude that the Plasmodial GCH1 sequences are related by structural and evolutionary attributes and considerably different from their mammalian homologs.

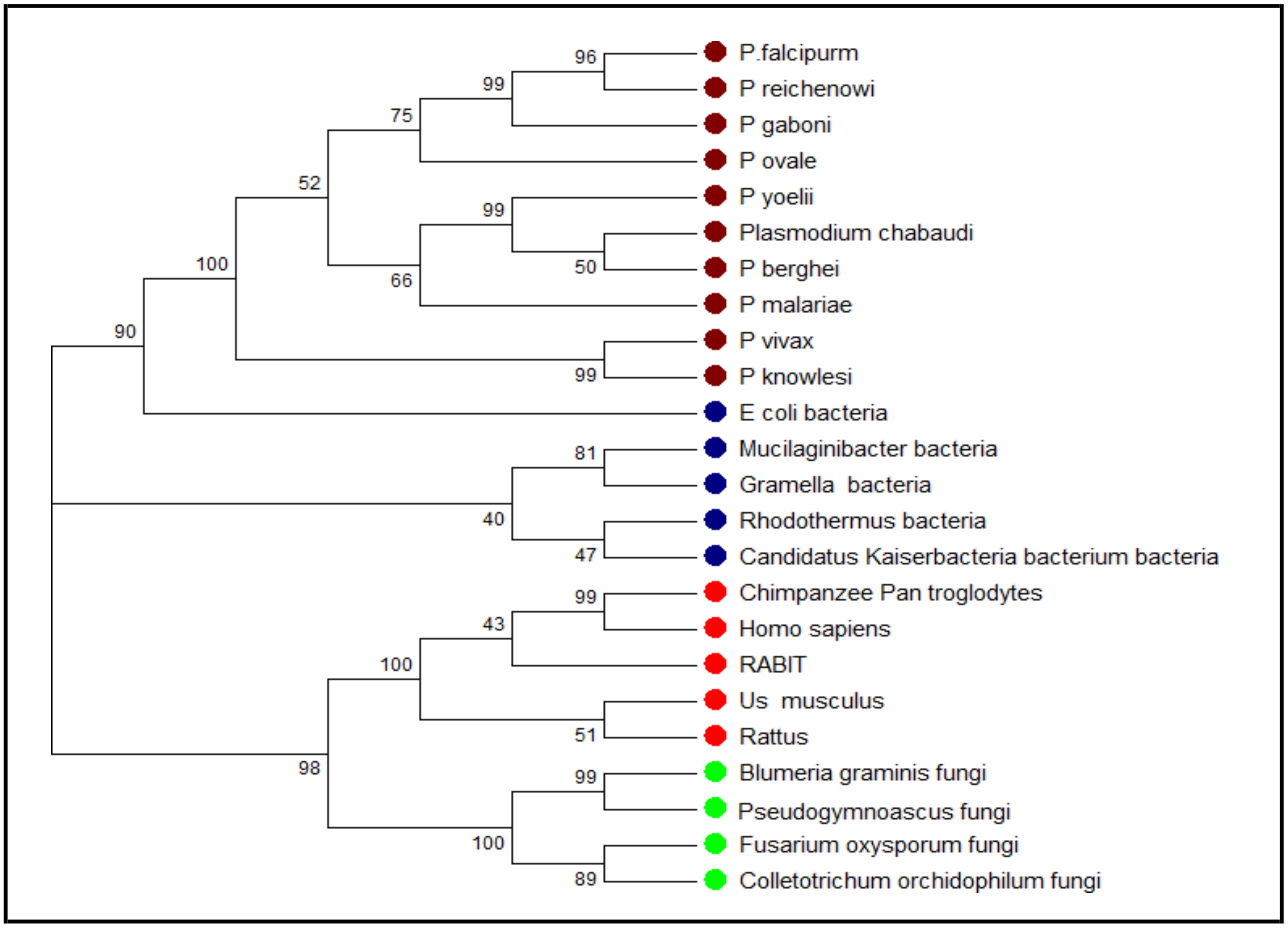


Figure 2.10: Phylogenetics tree of *P. falciparum* and its orthologs based on a PROMALS-3D alignment. The Neighbour joining tree was generated using maximum likelihood method based on the Le_Gascuel_2008 model (Quang, Gascuel and Lartillot, 2008). A bootstrap phylogenetic tree is shown in [Appendix 1, Figure A1.7]. The scale bar represents the number of amino acids substitutions per site. All positions containing gaps were eliminated.

Chapter summary

This chapter presented an in-depth sequence analysis of *P. falciparum* GCH1 protein and its homologs. A total of nine plasmodial homologs were retrieved from PlasmoDB database. MSA resulted in the identification of key residues that distinguished Plasmodial homologs from bacterial, fungal and mammalian sequences. Plasmodium sequences were found to have conserved sequence motifs. Sequence analysis showed that these motifs were significantly different in residue composition to the human GCH1 protein. The catalytic key residues were conserved between all species. The residues coordinating the GCH1 zinc ion; His 280, Cys 277 and Cys 348 were highly conserved. This has also extended to some key catalytic residues which were also highly conserved such as Glu 319, Ser 302, and His 346. Other highly conserved residues were located near the entrance of the active site pocket including Lys 303, Arg 352 and Arg 231. Sequence alignment also revealed some sequence variation, the highlighted differences were found to be in the active site and substrate binding pocket regions. The identified conserved residues of the Plasmodium species include Leu 276, Leu 317, Leu 225, Leu 239, Leu 282 Lys 278, Ile 349, Asn 350, Ile 224 and Tyr 232. Other conserved residues were found in the α4 and α5 helices flanking the β sheets at the body of the protein structure. These sequence dissimilarities can reflect functional characteristics important for inhibitor's interaction. The evolutionary relationship between the parasites and their mammalian hosts has been established through sequences analysis, phylogeny construction and motif discovery. The parasites sequences were conserved across other species and were distantly related to the human sequence; hence we can propose selective inhibition. This analysis was done at the protein sequence level and it would be important to visualise the interaction of the residues. Therefore, there is a need to predict the 3D structure of *P. falciparum* and its homologs to authenticate the results derived from the sequence analysis.

CHAPTER THREE

STRUCTURAL ANALYSIS: HOMOLGY MODELLING

GCH1 enzyme catalyzes the first reaction of the folate biosynthesis pathway. It is responsible for the breakage and rearrangement of the guanine and ribose rings in GTP and form 7, 8-dihydroneopterin triphosphate. This substrate is essential for the subsequent reaction in the folate pathway (Gräwert, Fischer and Bacher, 2013). In all species, GCH1 enzyme has a homo-decameric barrel like structure with ten zinc-containing active sites, each formed between every three adjacent chains (Auerbach *et al.*, 2000). The Plasmodium GCH1 is important for the malaria parasite survival; hence it is considered to be a potential drug target for the treatment of malaria. In this study homology modelling is employed to predict the 3D structures of the parasite GCH1 enzyme. The availability of the 3D structure can reveal biologically important properties and features such as protein mechanism of function and interactions which can be valuable in drug discovery.

3.1 Introduction

Protein tertiary structure is more conserved than its amino sequence. This observation forms the basis for homology modelling (Xiong 2006). Protein function is based on its tertiary structures which underline essential key function such as binding sites and domain interactions. Structural analysis of the protein 3D structure can allow for the identification of its function and biological significance. This plays a major role in modern drug discovery practice (Hillisch, Pineda and Hilgenfeld, 2004; Xiong 2006). The number of available protein sequences in the biological databases has grown at an approximate exponential rate. This massive growth is a result of the advances in Next Generation Sequencing Technology (NGS) (Mardis 2008; Koboldt *et al*. 2013). The numbers of protein structures that have been identified experimentally are minimal when compared to the available sequence data. This forms the need for *in silico* approaches to produce the 3D structure of proteins in order to complement the experimental techniques. In this regard, computational biology aims to fill this gap by predicting unknown structures of proteins (Dorn *et al.*, 2014).

3.1.1 Protein structure determination

The 3D structure of proteins underlines essential functional properties; it allows for understanding the protein function at an atomic level (Nayeem *et al*. 2006; Gherardini *et al*. 2008). Several experimental techniques have been developed to determine the protein 3D structure; this includes X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy and Electron Microscopy (EM), however these methods are considered to be laborious, expensive and time consuming. To overcome these obstacles, computational biology approaches such as homology modelling are used to determine the protein 3D structure using protein structures that have been characterized experimentally.

When using X-ray crystallography technique, the protein is first purified, crystallized then exposed to X-ray beams. The X-ray beam is diffracted by the electrons in the protein with different intensities and directions; this information is captured and referred to as an electron density map. The electron density map is then used to identify the location of each atom in the protein molecule, this resulting in building of the protein 3D structure. The final step involves the refinement and validation of the crystal structure; this is often assessed by the protein crystal structure resolution which is referred to as the level of detail seen at atomic level (Xiong 2006; Wlodawer *et al*. 2008).

Another technique used is the Nuclear Magnetic Resonance (NMR) spectroscopy in which the protein structure is determined in solution by utilizing the magnetic properties of the atomic nuclei. The significance of this is to identify the protein structure in conditions close to its native environment (Wüthrich 2003; Kwan *et al*. 2011). The protein is first solvated, then exposed to a magnetic field and probed with radio waves. The resulted resonance frequency is then captured. Nuclear Overhauser effect (NOE) caused by nearby hydrogen atoms in the protein are used to assign each hydrogen signal in the NMR spectrum. Distances between pairs of the scores of hydrogen atoms are then calculated allowing the determination of the 3D structures (Di Luccio *et al*. 2011).

Electron Microscopy (EM) method is based on electron diffraction. It involves the preparation of samples by cryo-freezing (cryoEM) or negative staining with heavy metals. This is followed by exposing the sample to electron beams; then captured in an image. The image is used to construct the 3D structure computationally, then refined and verified to attain an optimum image (Zhou, 2008).

Experimental techniques have the benefit of reliability and accuracy. However, it is still unfeasible to obtain all protein structures as yet. Therefore, homology modelling becomes the tool of choice with better cost and time effectiveness in attaining the protein 3D structures until experimentally structures are made available.

3.1.2 Homology modelling

Computational protein structure prediction can be done by using three main approaches; *ab-initio* folding, threading and homology modelling. In homology modelling the protein 3D structure is predicted based on its sequence similarity to a template protein structure that has been experimentally characterized (Sander and Schneider, 1991). Structural similarity can be determined based on the level of sequence similarity between the target and template protein sequences. This similarity increases with higher sequence identity. Generally, 30% sequence identity is required to generate a good model (Baker & Sali 2001; Özlem Tastan Bishop *et al*. 2008; Schmidt *et al*. 2014).

Modelling resolves the 3D structure of proteins in cases where the structures were not experimentally identified. This allows for understanding the protein-inhibitor complexes, functional specificity and interactions (Vyas *et al.*, 2012). Several computational software packages have been made available for homology modelling as well as web-based servers such as SWISS MODEL (Biasini *et al.*, 2014). Stand-alone programs include MODELLER (Webb and Sali, 2016) and WHATIF (Vriend, 1990). Modelling the 3D structure of a protein (target protein) involves a series of steps: (i) template identification, (ii) template–target alignment (iii) model building and refinement and (iv) model validation (Xiong 2006; Vyas *et al*. 2012).

3.1.3 Template selection

Selecting a suitable structural template is an essential step for generating quality models. This is done by first searching the Protein Data Bank (PDB) for a template that meets anumber of criteria including, high sequence similarity to the target protein, structure resolution and coverage. Templates with a sequence identity > 30% to the target protein are acceptable and likely to have a common 3D structure (Hillisch *et al*. 2004; Özlem Tastan Bishop *et al*. 2008; Du *et al*. 2015). Sequence identity between 30-50% of target and template protein is ideal to generate high quality models.

Models generated at 15-30% identity can still be used however; this must be examined carefully and accurately aligned to identify homology between sequences. Sequence identity below 15% can result in the production of unreliable protein structure (Hillisch *et al*. 2004; Nayeem *et al*. 2006). When selecting from multiple possible templates, the template that has the highest sequence identity, highest resolution and best coverage of the target protein is selected. If one template does not sufficiently represent the target protein, a combination of multiple templates can be used. Sequence similarity search tools such as HHpred, PSI- BLAST, and PDB sequence search can be used to search for a template structure.

PSI-BLAST employs PSSM profiles, this allows for the identification of distant homologs with biologically significant sequence similarity. PSI- BLAST is considered to be more sensitive than the usual BLAST search. However, it may not effectively detect homologs of known structure because it is based on sequence comparison. To overcome this, protein structure prediction methods like HHpred are used.

HHpred (http://toolkit.tuebingen.mpg.de/hhpred) is a server for structure prediction and detection of structural homologs. It is considered to be fast and accurate (Söding *et al*. 2005). HHpred search is carried out using PSIPRED structural predictions and HHsearch (HMM-HMM comparison). This search provides high sensitivity in the detection of homologs (for further explanation see section 2.1.1). The alignment produced by HHpred is considered to be accurate due to the incorporation of the protein structure prediction in the process. HHpred output results are displayed in a list with all possible templates ordered by their E-values (Söding, Biegert and Lupas, 2005).

Another search tool is the PDB, in which the target protein sequence is used as a query to search the PDB database (Prlić *et al.*, 2010). The search gives a list of protein structures that have significant similarity to the target protein sequence; one can then choose a suitable template from this list.

3.1.4 Template-target alignment

After obtaining a suitable template structure, sequence alignment of the target-template structure should follow. An accurate sequence alignment is essential, because any errors made in the alignment of residues can adversely distort the structure of the generated model (Xiong, 2006).

During target template alignment, attention should be given to both the matching of similar and identical residues and structural correctness of the alignment. Thus, it is important to use alignment programs that incorporate protein structure prediction such as HHpred and PROMALS-3D (Söding *et al*. 2005; Pei *et al*. 2008).

3.1.5 Model building and refinement

Modelling is based on the alignment of the target-template structure, in which the templates main-chain and side-chain atom coordinates are copied onto the model. When the aligned residues differ, only the main-chain atom coordinates are copied. In gap regions of the alignment loops are modelled. These loops can be modelled using two approaches; database search methods and *ab initio* method (Xiong, 2006). Once the main-chain has been modelled, the side chains of residues are subjected to a conformation or rotamer search from libraries that contain experimentally derived structures. The most favourable conformations with minimal steric clashes and energy scores are then selected (di Luccio and Koehl, 2012).

Modelling can be performed using MODELLER (Andrej Šali, 1993); a script-based program. It requires the structural alignment file in Protein Identification Resource (PIR) format, template atom coordinate files and a script file. It automatically builds the 3D models of the target by going through the backbone, loop and side-chain building steps and finally refining the model (Webb and Sali, 2014). MODELLER has the advantage against other web-based modelling programs of providing the user with control over the whole modelling procedure. The outputs of MODELLER are atom coordinate files of the generated models in a PDB format that can be visualized using visualization programs such as PyMol (DeLano, 2002). MODELLER can generate several models of the same target; these models are ranked by their Discrete Optimised Protein Energy (DOPE) scores.

The DOPE score is an atomic-distance-dependent, knowledge-based potential. It is derived from a sample of native structures. DOPE uses an enhanced reference state corresponding to non-interacting atoms in a heterogeneous sphere and it is calculated from the number of restraints acting on each residue, giving a z-DOPE normalised score. The lowest normalised z-DOPE score represents an accurate model structure, usually scores below -1 are acceptable (Shen and Sali, 2006).

3.1.6 Model refinement

The complete model should be refined and optimised in both geometric and energetic aspects so as to attain a structure of stable native conformation. This requires an effective sampling of the conformational space to identify near native-like model conformation which can be achieved through energy minimization of bond lengths and angles without distorting its structural conformation (Levitt & Lifson 1969). Model refinement can be either local or global. Local refinement involves loops and side chains of the protein, in which energy scores are used to access local quality. Global refinement involves resolving structural irregularities in the entire protein structure. These methods are usually incorporated into model building programs like MODELLER (Webb and Sali, 2014).

3.1.7 Model validation

Computational models are not as accurate as the experimentally identified structures; therefore, it is important to validate the generated models by assessing their quality in terms of their stereo-chemical properties, protein folding quality and model compatibility with its amino acid sequence. Several Model Quality Assessment Programs are made available (MQAPs) including ProSA, VERIFY3D, ANOLEA and QMEAN (Benkert, Biasini and Schwede, 2011). Some web servers such as MetaMQAPII (https://genesilico.pl/toolkit/unimod?method=MetaMQAPII) incorporates several validation programs. This grouping of different MQAPs combines the strengths of each program and makes it possible to verify many qualities of a model at once.

PROCHECK (http://www.ebi.ac.uk/thornton-srv/software/PROCHECK/) evaluates the stereo-chemical quality of a protein structure such as the bond length, chirality and ψ and ϕ torsion angles. The output of PROCHECK is presented in a plot analyzing overall and residue by residue geometry. This plot is referred to as a Ramachandran plot (Ramachandran, Ramakrishnan and Sasisekharan, 1963). The colouring of the plot represents different regions; where red represents the most favoured region (Laskowski *et al.*, 1993).

ProSA (https://prosa.services.came.sbg.ac.at/prosa.php) evaluates models' quality by calculating the surface energy of the protein providing a Z-score. **ProSA** employs a knowledge-based force field as an energy function that is derived from statistical analysis of all experimentally determined protein structures in databases (Sippl 1995; Ferrada & Melo 2009). The Z-score of a protein represents the overall quality of the model and measures the deviation of the overall energy of the model with respect to random conformations of experimentally determined structures. A Z score that is not within the range of characteristics for native proteins symbolises bad structural model (Wiederstein and Sippl, 2007). This score determines the global model quality and it can be used to determine if the structure is in the desired range of native structures. ProSA also gives a local model quality energy plot, which plots the energy scores against the amino acid positions over a 10 and a 40-residue window. Residues with positive energy values show erroneous parts of the structure (Wiederstein and Sippl, 2007).

3.2 Methods

3.2.1 Template selection

The target GCH1 protein sequences of *P.falciparum*, *P.knowlesi*, *P.ovale*, *P.vivax* and *P.malriae* were used as queries to search for template structures using PRIMO server with default parameters and HHpred server. T-COFEE alignment tool was used for the search by the mode of local alignment. A list of possible templates was generated and ordered by their E-value. The criteria for selecting the best templates were; percentage similarity to the target sequence, coverage and template structure resolution (Table 3.1). The selected template was retrieved from PDB database and validated before its use. The template structure was also viewed via PyMOL (DeLano, 2002) and Discovery Studio visualization programs (San Diego: Accelrys Software Inc., 2012).

3.2.2 Generation of GCH1 Biological Unit Assembly

PyMOL simple script (Jason Vertrees, 2010) was used to build the protein biological unit from the available crystal structure of *T.thermophilus* (PDB ID: 1WUR) [Appendix 2]. The resulting biological unit was used as a template to generate the Plasmodium biological unit models. Sequences of all ten chains were used as an input (Appendix 2, Figure A2.6).

3.2.3 Template-target alignment

Alignment of the target and template sequences was carried out using T-COFFEE alignment tool. The sequence of the selected template structures was retrieved from the Protein Data Bank (PDB) then aligned to the target protein sequences. Inspections and minor manual adjustments were applied in order to obtain an accurate alignment. The resulting alignment was viewed using Jalview (Waterhouse *et al*., 2009) and saved as a PIR format for homology modelling.

3.2.4 Homology modelling

Homology modelling was performed using MODELLER9v7. Alignment in the PIR format as well as the template protein coordinate file was provided as an input for MODELLER to generate models using a python script. 100 models of the protein chain A and the biological units were generated for each of the five Plasmodium species.

The modelling process was carried out with slow refinement to the model building process. The refinement involves repacking of the side chains and energy minimization of the entire structure in order to regulate the alignment and modelling loops and side chains. The slow refinement is a predefined function of MODELLER refine module, which allows for controlling the degree of MD refinement. Models generated were saved in PDB format containing the atomic coordinates. The global quality of the models was assessed by using MODELLER to calculate their z-DOPE energy scores and sort them accordingly. The models of each target with the lowest energy scores were then selected for further validation.

Loop refinement was performed using MODELLER via a python script as well as the models coordinates file. 100 loop refined models were generated. After the refinement, an improvement in the z-Dope score was observed. The best three models having the lowest energy scores were selected and validated. For the validation process of the best models attained ANOLEA, QMEAN, PROCHECK and ProSA servers were used. The servers mentioned were accessed via web servers in which the PDB files were uploaded and automatically validated.

3.3 Results and Discussion

3.3.1 Template selection

One template was selected for modelling each of the Plasmodium GCH1 proteins. The template was selected based on its sequence identity to the target sequence, resolution and coverage (Table 3.1).

Table 3.1: Tabulated result of candidate templates, the selection was based on the sequence identity to the target, resolution and completeness of the structure target sequence. The selected template (1WUR) showed the highest similarity, sequence coverage and resolution.

Organism	***T.thermophilus***	***Homo sapiens***	***T.thermophilus***	***Escherichia coli***
PDB ID	1WUR	1FB1	1WM9	1A8R
E-value	3.50E-54	1.60E-56	3.50E-54	4.40E-59
R-Value Free	0.238	0.293	0.261	0.246
R-Value Work	0.206	0.204	0.208	0.200
Structure resolution	1.82 Å	3.1 Å	2.2 Å	2.1 Å
z-DOPE	-1.15	-0.48	-1.1	-0.85

The template structure PDB ID: 1WUR, which is the crystallographic structure of GCH1 protein from *T.thermophilus* bacteria, showed the highest similarity to most of the target protein. It showed 30% - 34 % similarity to the target sequences, high resolution of 1.82 Å and significant E-values close to zero (Table 3.2). The template PDB ID: 1WM9 was similar to the selected template PDB ID: 1WUR however, it had lower resolution and it did not contain a co crystallized ligand.

The E-values showed the likelihood that the sequence alignment result between the target and template sequences occurred by chance. Thus, lower E-values point towards a high probability that the two proteins are similar. The template also showed relatively good sequence coverage of the targets with no missing residues or gaps except for the N and C terminal regions. The target-template sequence identities for all five Plasmodium species were above 30% which falls in the "safe zone" (Krieger, Nabuurs and Vriend, 2005) (Table

3.2). Therefore, we can deduce that the target-template match can be considered as one having close homology (Xiong 2006). Due to the high variation and unique features of the N and C-terminal regions of the GCH1 sequences the template did not sufficiently represent the target sequences. These sequences were absent in the template sequence and no other second template could cover it. These regions were trimmed to obtain an optimum alignment and better quality of the generated models.

Table 3.2: A summary of the target-template sequence alignment results.

Target sequence	Template	Organism	Sequence similarity%	Score	E-value
P.falciparum	1WUR	*T.thermophilus*	34%	394.89	2.5E-53
P.vivax	1WUR	*T.thermophilus*	34%	390.60	5.8E-52
P.ovale	1WUR	*T.thermophilus*	34%	377.40	8.0E-51
P.malariae	1WUR	*T.thermophilus*	36%	378.01	3.3E-50
P.knowlesi	1WUR	*T.thermophilus*	32%	385.14	6.1E-51

3.3.2 Template validation

Different validation programs were used to assess the quality of the selected template such as PROCHECK (Laskowski *et al.*, 1993), ANOLEA (Melo and Feytmans, 1998), ProSA (Wiederstein and Sippl, 2007) and QMEAN (Benkert, Biasini and Schwede, 2011). The selected template was assessed to establish its suitability for homology modelling. The template validation report as well as PDB file was inspected for completeness. The template R-value of 0.238 indicates a fair match between the simulated diffraction pattern and the observed experimental diffraction pattern. The slider graph metrics of global quality indicators shown in Figure 3.1 displayed better percentile scores when compared to other X-ray structures.

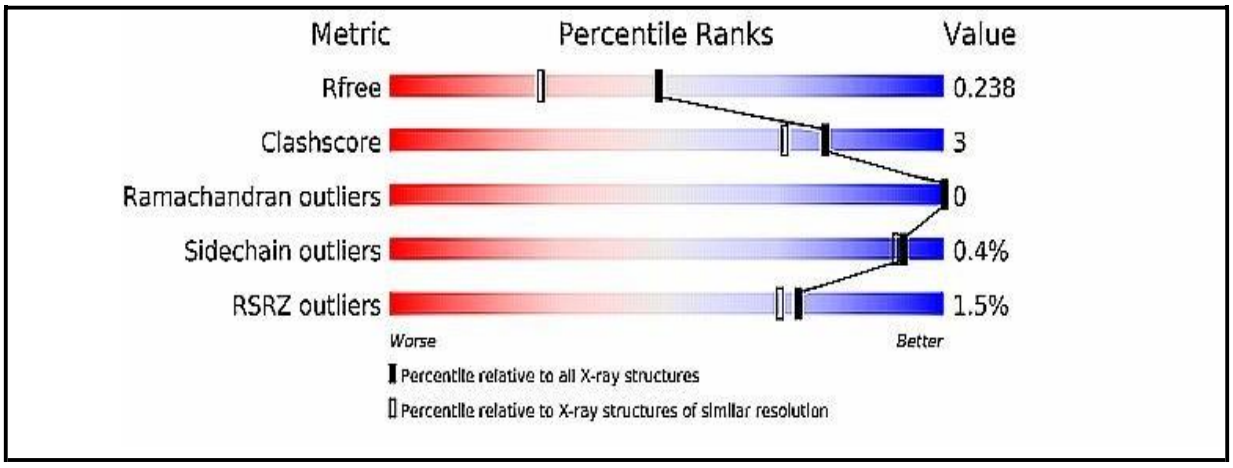


Figure 3.1: Slider graph metrics of global quality indicators of the template structure 1WUR. Percentile scores for global validation metrics identified from structure validation report. The template had no outliers. Overall, the template structure had better global scores than other candidate templates. Source (Y. Tanaka *et al.*, 2005).

QMEAN scoring function consists of three statistical potential terms and two additional simple terms used to illustrate the agreement of the predicted and observed secondary structure and solvent accessibility (Benkert, Tosatto and Schomburg, 2008). QMEAN statistical potentials are extracted from a non-redundant set of high-resolution protein structures. QMEAN quality assessment is derived from six different structural features including C-beta interaction energy, all-atom pair wise energy, solvation energy, torsion angle energy, secondary structure agreement and solvent accessibility agreement. The local geometry is analyzed by identifying torsion angle potential over three consecutive amino acids. Long range interactions are assessed by using secondary structure specific distance dependent pair wise residue level potential and the burial status of the residues is described by the solvation potential (Benkert, Tosatto and Schomburg, 2008).The output scores were expressed as Z-score and compared to scores derived from the evaluation of high resolution experimental data. Outliers are often identified by lower agreement with QMEAN Z-score; if the deviation from the mean value is greater than 5 or less than -5 then it is considered as alarming and it means that something is wrong with the structure (Benkert, Biasini and Schwede, 2011). The slide graph provided by QMEAN showed the level of deviation from the mean score to be acceptable (Figure.3.2) as the Z-score values were within the range. Structures with QMEAN values between 0 and 1 are considered to be reliable and error free (Benkert, Tosatto and Schomburg, 2008). The QMEAN score of the template structure was 0.694. This score deviates by less than 1 standard deviation from the mean score hence it is considered to be within the expected quality range and indicative of the structure reliability.

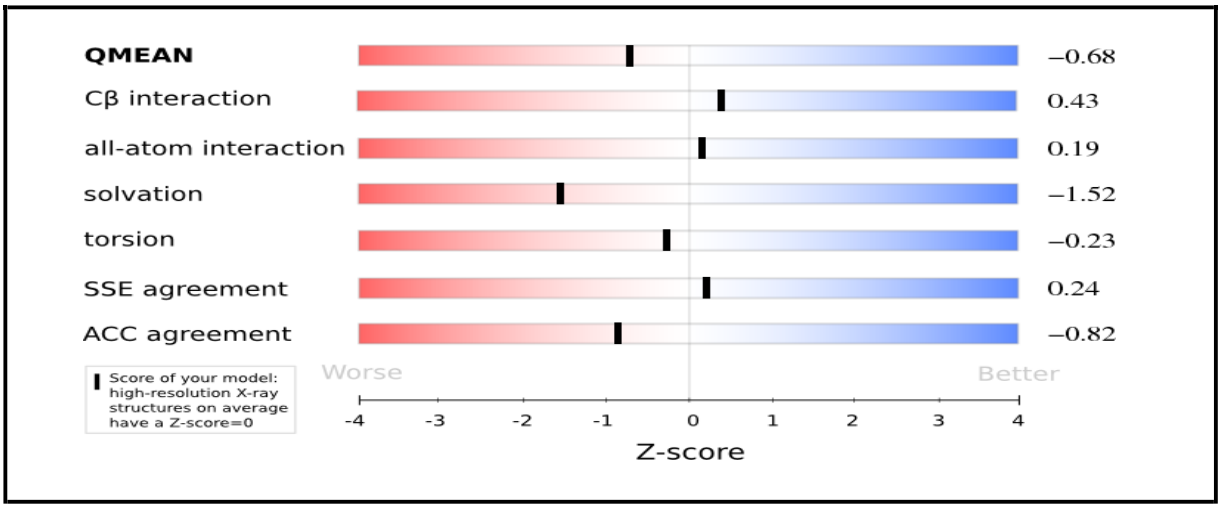


Figure 3.2: QMEAN quality assessment derived from the six different structural features descriptors. A slight deviation was observed on the solvation energy score which indicate lower agreement with QMEAN Z-score.

The stereochemistry of the template was examined using PROCHECK server. The server provided a Ramachandran plot and a list of residue-residue values. Results are shown in Figure 3.3. Ideally 90% of the residues should be in the most favoured regions for a good structure (Hollingsworth and Karplus, 2010). The template structure was found to have 95.6% of its residues in the most favoured regions, 4.4% residues in the additional allowed regions and 0.0% residues in the disallowed regions. From this we can conclude that 1WUR template structure is of good quality.

ProSA validation showed a surface energy Z-score of -4.7. This was plotted along-side Z-scores of all the experimentally derived structures in the PDB (Figure 3.4). The plot showed that the selected template structure was within the desired range of native conformations. ProSA local model quality plot showed the energy values per residue. Most residues in the structure seemed to be having appropriately low energy scores (below the zero line).

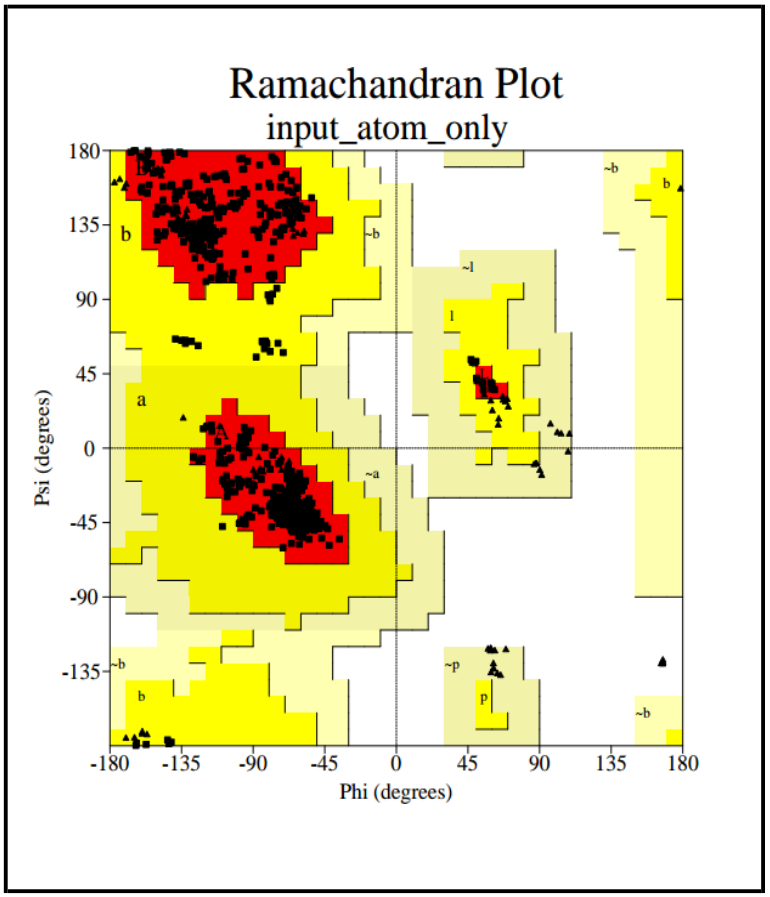


Figure 3.3: Ramachandran plot of the template structure 1WUR. Red indicates most sterically favoured region, dark-yellow indicates the additional allowed regions, light yellow shows the generously allowed regions and the disallowed regions are shown in white.

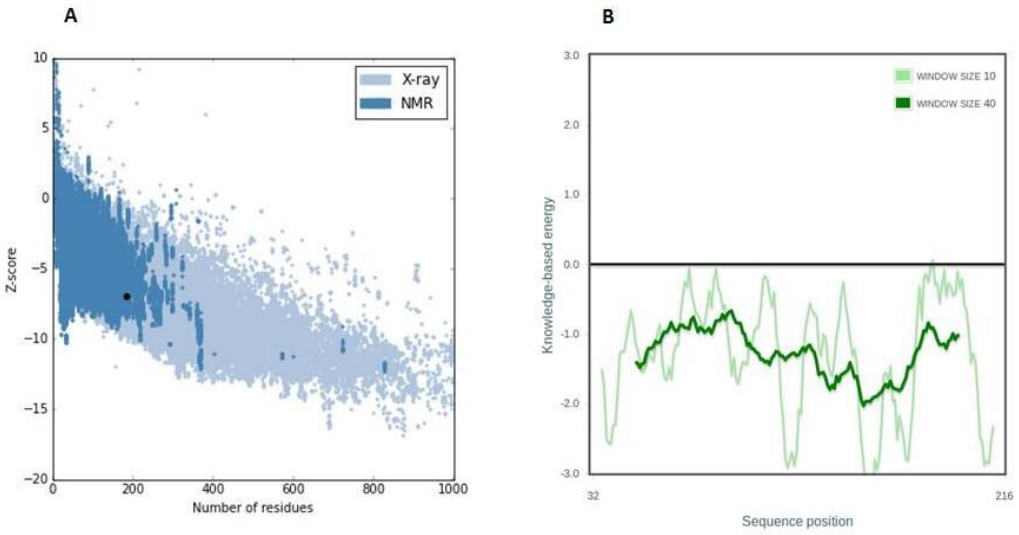


Figure 3.4: ProSA validation results of the template structure (PDB ID: 1 WUR). ProSA global energy plot of *T.thermophilus* structure plotted as a black dot, other PDB structures are shown in blue and light blue dots **(A)**. ProSA local quality plot in a 10 and 40 residue window **(B)**, most residues are below the zero-line indicating low energy scores.

ANOLEA (Atomic Non-Local Environment Assessment) assessment was also carried out. It performed energy calculations on a protein by analysing the "Non-Local Environment" of each heavy atom in the molecule using a very accurate and sensitive Atomic Mean Force Potential (AMFP) to calculate the non-local energy profile of a protein. The template structures had majority of its residues with energy values on the negative side, indicating a favourable energy environment; therefore, the template structure was of good quality (Figure 3.5).

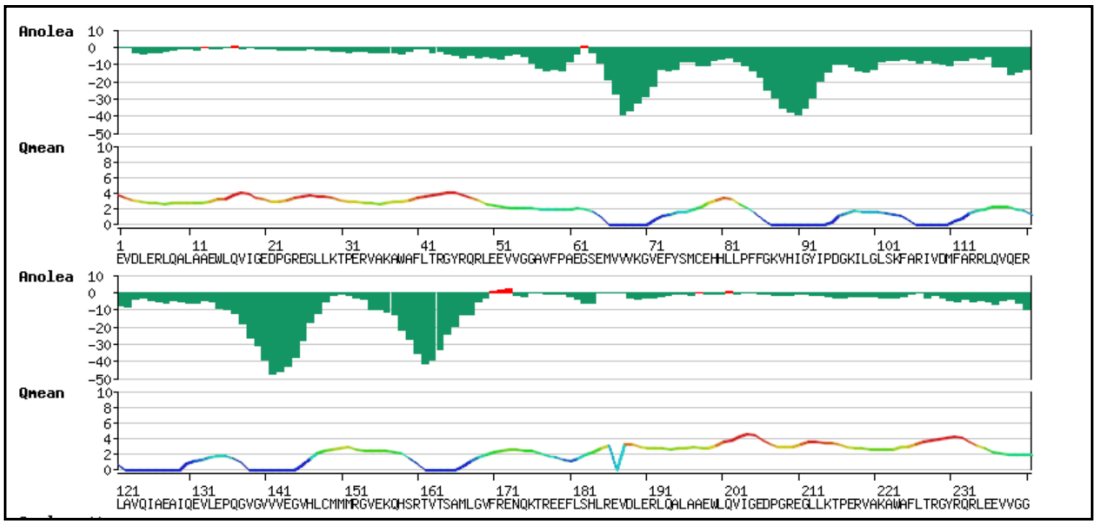


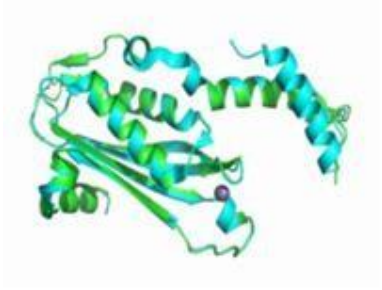
Figure 3.5: ANOLEA local quality assessment of the template structure. Negative energy values in green indicate a favourable energy environment. Positive values in red indicate unfavourable energy environment. In general, most residues are in the favourable energy environment.

3.3.3 Template-target sequence alignment

Sequence alignment of the selected targets and the template structure sequence was carried out using T-COFFEE alignment tool. T-COFFEE alignment was chosen because of its accuracy; as it incorporates both sequence and structural information in the alignment. The template structure sufficiently covered the target sequences except for the terminal region due to its sequence uniqueness (See chapter 2).

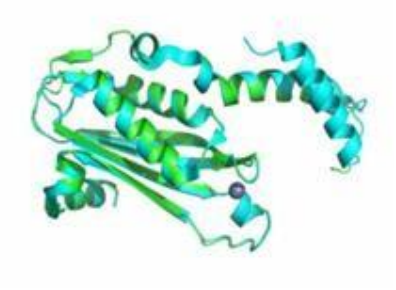
3.3.4 Model building and refinement

The 3D structures of Plasmodium GCH1 protein (Chain A) for *P.falciparum*, *P.knowlesi*, *P.ovale*, *P.vivax* and *P.malriae* were built and refined using MODELLER. For each of the Plasmodium proteins the best model was selected and validated. (Table 3.3) gives a detailed summary of the most optimal models attained and their corresponding quality assessment scores. The 3D structure of the generated models is shown in figure 3.6. The RMSD between the model and the template structure was given (as calculated by superimposing the structures using PyMol). The RMSD values were below 1, indicating the reliability of the generated models.



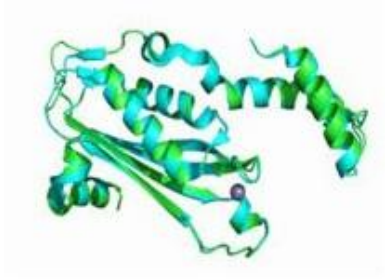
P.falciparum

RMSD 0.252



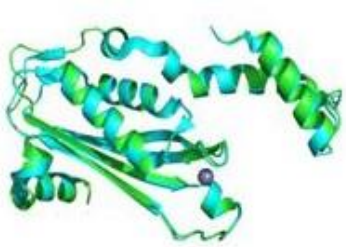
P.malriae

RMSD 0.250



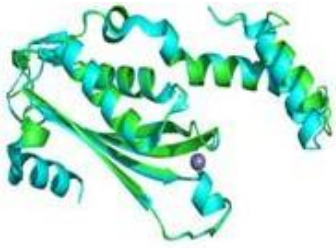
P.ovale

RMSD 0.246



P.vivax

RMSD 0.253



P.knowlesi

RMSD 0.248

Figure 3.6: Top generated models of GCH1 protein (Chain A) superimposed on their original templates, 1WUR. The original template is shown in blue and the models in green. RMSD values are illustrated on each protein. The RMSD value between the template and the generated models was below 1 indicating higher similarity between the two.

3.3.5 Model validation

The best 3D models out of 100 from each of the five Plasmodium species were selected for validation (Figure 3.6). These models were attained based on the template structure of GCH1 protein from *T.thermophilus* (PDB: ID 1WUR). Using DOPE, the global quality of the structures was assessed and their normalized z-DOPE-scores were ranked using a Python script. The structures with the lowest z-DOPE-score correspond to the best quality model and share the highest similarity to the native structure. Negative scores of -1 or below usually indicate the accuracy and reliability of the generated models (Eramian *et al.*, 2008). z- DOPE-score of the top three models from each species are presented in Table 3.3. A boxplot of calculated z-DOPE-scores was created using R software; it showed that most values were on the negative side, suggesting that the modelling procedure was good (Figure 3.7). Overall, the structures were of good quality with best models having z-DOPE-scores close to that of the crystallized structure. Model validation was carried out using ProSA, QMEAN and PROCHECK. The MQAP evaluated different properties of the modelled structures and detect problematic regions including miss-oriented side chains, alignment errors and incorrect chain conformation. From the validation results, it was observed that most errors were found to occur in the N-terminal and C-terminal loop regions.

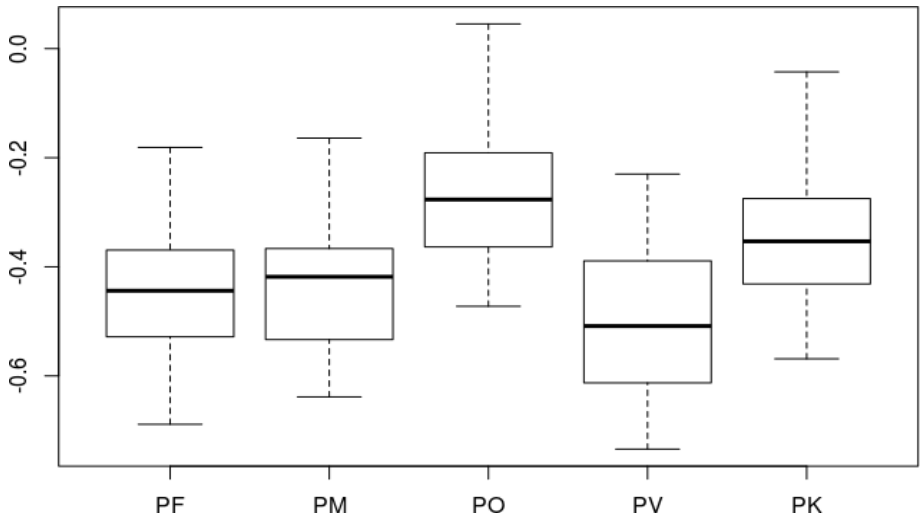


Figure 3.7*:* Calculated Z-DOPE-scores of all models summarized in a boxplot.

Table 3.3: Summary of the top three model-quality assessments values for each of plasmodial GCH1 protein. Models were evaluated based on their z-DOPE-score, QMEAN Z- score and ProSA Z-scores.

ORGANISM	BEST THREE MODELS	PROSA SCORE	QMEAN SCORE	QMEAN Z-SCORE	MODELLER Z-DOPE SCORE
	Model 1	-4.85	0.603	-1.617	-0.69
P.falciparum	Model 2	-3.31	0.556	-2.098	-0.67
	Model 3	-4.70	0.626	-1.392	-0.62
	Model 1	-4.73	0.623	-1.41	-0.64
P.malriae	Model 2	-4.68	0.621	-1.436	-0.63
	Model 3	-4.56	0.656	-1.079	-0.62
	Model 1	-4.38	0.567	-1.995	-0.47
P.ovale	Model 2	-4.53	0.606	-1.600	-0.47
	Model 3	-4.38	0.566	-2.003	-0.46
	Model 1	-5.39	0.541	-2.335	-0.73
P.vivax	Model 2	-5.44	0.555	-2.193	-0.71
	Model 3	-5.44	0.495	-2.808	-0.69
	Model 1	-3.70	0.496	-2.701	-0.57
P.knowlesi	Model 2	-3.62	0.481	-2.861	-0.53
	Model 3	-3.79	0.499	-2.671	-0.52
Template	X-ray structure	-4.70	0.694	-0.678	-1.40

Table 3.4: PROCHECK local quality assessments (Laskowski *et al*. 1993): Percentage values indicate the number of residues in the most favoured, additional allowed, generously allowed and disallowed regions. All models had at least 90% residues in the most favoured region.

Organism	**Model**	**PROCHECK Validation report**			
		Number of residues in favoured region/Core (Percentage value)	Number of residues in additional allowed region (Percentage value)	Number of residues in generously allowed region (Percentage value	Number of residues in disallowed region (Percentage value)
	Model 1	93.7% core	5.7% allow	0.0% gener	0.6% disall
P.falciparum	Model 2	91.8% core	7.0% allow	0.6% gener	0.6% disall
	Model 3	92.4% core	6.3% allow	1.3% gener	0.0% disall
	Model 1	92.5% core	6.9% allow	0.6% gener	0.0% disall
P.malariae	Model 2	93.1% core	6.3% allow	0.6% gener	0.0% disall
	Model 3	91.8% core	7.5% allow	0.6% gener	0.0% disall
	Model 1	91.9% core	7.5% allow	0.0% gener	0.6% disall
P.ovale	Model 2	93.1% core	6.2% allow	0.0% gener	0.6% disall
	Model 3	91.9% core	7.5% allow	0.0% gener	0.6% disall
	Model 1	94.2% core	5.3% allow	0.0% gener	0.6% disall
P.vivax	Model 2	93.6% core	6.4% allow	0.0% gener	0.0% gener
	Model 3	94.7% core	4.7% allow	0.6% gener	0.0% disall
	Model 1	90.8% core	7.8% allow	0.7% gener	0.7% disall
P.knowlesi	Model 2	92.2% core	7.2% allow	0.7% gener	0.0% disall
	Model 3	91.5% core	7.2% allow	0.7% gener	0.7% disall
Template	X-ray structure	95.6% core	4.4% allow	0.0% gener	0.0% disall

3.3.6 ProSA validation

ProSA surface energy Z-score of the generated models was attained and tabulated in Table 3.3. It was observed that these scores were close to that of the crystallized structure. ProSA Z-scores of the top models were plotted along-side the Z-scores of all the experimentally derived structures in the PDB (Appendix 2, Figures A2.3 to A2.5). The plot showed that the Plasmodium model structures were within the desired range of native conformations. ProSA local model quality plot showed the energy value per residue. The knowledge-based energy values were below zero in the 40-residue window which is less sensitive than the 10-residue window. Most residues of the generated models seemed to be appropriately modelled having energy scores below the zero line, except for the N and C terminal regions in which some peaks were observed, indicating positive energy values. These regions include a significant level of sequence variation and it was difficult to model as it was absent in the template structure 1WUR used for modelling. This region may not be accurate however, it does not lie in the structure catalytic domain area and it will not affect further study of protein interaction.

3.3.7 ANOLEA

ANOLEA (Atomic Non-Local Environment Assessment) was used for model assessment. Overall the structures had more residues with energy values on the negative side, representing a favourable energy environment, therefore the packing quality of predicted models is considered as good (Appendix 2, Figure A2.4). ANOLEA also showed that the N and C terminal were less accurate. This agrees with the results obtained from ProSA validation. Predicted structure in this region may not be fully accurate. Once more, the terminal regions do not lie in the area of interest of the structure thus it will not affect the observation of the desired interactions in the structure (Appendix 2, Figure A2.4).

3.3.8 PROCHECK

PROCHECK was used to assess the stereochemistry of the generated models. PROCHECK statistics were captured and tabulated (Table 3.4). PROCHECK Ramachandran plots are shown in Appendix 2, Figures A2.7 to A2.9. All models had at least 90% of their residues in the most favoured region. *P. falciparum* top model was found to have 93.7% of its residues in most favoured regions and the remaining 5.7% residues in the additional allowed regions. One residue was found in the disallowed region Lys 153. *P. malriae* top models had 92.5 % of its residues in the most favoured region, 6.9% residues in the additional allowed regions and 0.0% residues in the disallowed regions. This model was of good quality having no residues in the disallowed regions. *P. ovale* top model was found to have 91.5 % of its residues in the most favoured region, 7.5% residues in the additional allowed regions and 0.6% residues in the disallowed regions. One residue was found in the disallowed region Glu

156. *P.vivax* top model was found to have 94.2 % of its residues in the most favoured region, 5.3. % residues in the additional allowed regions and 0.6% residues in the disallowed regions. One residue was found in the disallowed region Lys 60. *P. knowlesi* top had 90.8 % of its residues in the most favoured region, 7.8 % residues in the additional allowed regions and 0.7% residues in the disallowed regions. Two residues were found in the disallowed regions Glu 151 and Tyr 43. The residues found in the disallowed regions of the models were far away from the active site and would not have an effect on the region of interest. Judging from the results obtained, the structures were considered fit for use in structural analysis.

3.3.9 GCH1 Biological Unit Assembly

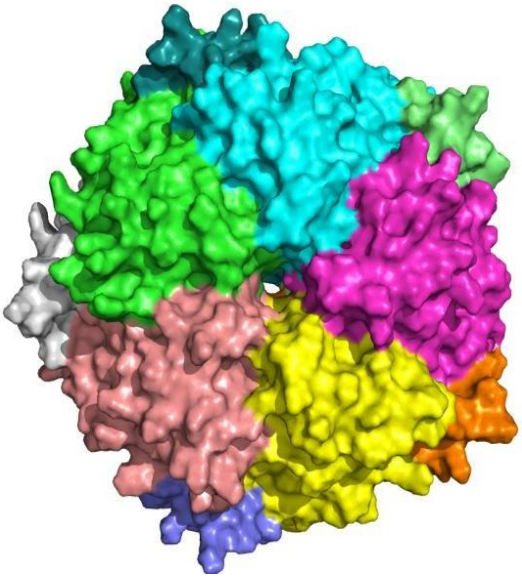


Figure 3.8: *P. falciparum* biological unit assembly. GCH1 monomer exists as a single fold that is copied and assembled. The structure is shown as a surface, with each of the ten symmetrical chains coloured differently for the purpose of classification. The figure was generated using PyMOL (DeLano, 2014).

The protein biological unit also, referred to as the biological assembly, is the functional quaternary structure of the protein (Jefferson, Walsh and Barton, 2006). The biological assembly of GCH1 protein consists of two pentamers; each has five chains which makes the protein homo-decameric (Figure 3.8). Each chain folded into α helices, flanking anti parallel β sheets, in the body of the protein. The N terminal helix was remote from the protein centre. Each of the protein ten active sites was formed at the interface of three adjacent chains, two chains from one pentamer and one from the other. The active site is only accessible through the opening formed by the last helices of each chain in the centre of the pentamers (Thöny, Auerbach and Blau, 2000) (Figure 3.10).

The biological unit assembely of the template structure (PDB ID: IWUR) was used to model the Plasmodium models shown in Figure 3.10. Quality improvement of the generated models was observed, espicallay when compared to the previous one chain models. A boxplot of calculated z-DOPE scores was created using R software. Majority of the z-DOPE-score values from the generated models were on the negative side, having z-DOPE-scores close to that of the crystallized structure. Some outliers were observed with positive scores but overall, the structures were of good quality (Figure 3.9).

The resultant models were validated using their z-DOPE score from MODELLER, ProSA, QMEAN and PROCHECK. The quality assessment scores were tabulated in (Table 3.5 and 3.6); the scores were within the range of reliable native crystal structures and close to that of the crystallized template structure. ProSA validation gave a surface energy Z-score, this was plotted along-side Z-scores of all the experimentally derived structures in the PDB (Appendix 2, Figure A2.8). The plot showed that the generated model structures are within the desired range. ProSA local quality plot showed that most inaccuracy was found to occur in the N and C terminal regions. In some cases, the problematic regions were due to the misalignment of residues, hence manual adjustments were applied to the sequence alignment.

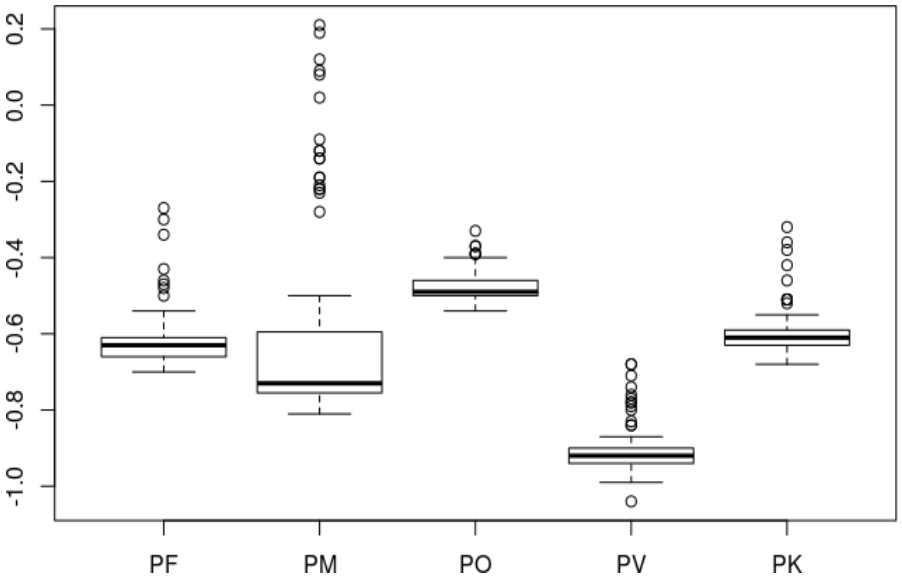


Figure 3.9: Calculated z-DOPE-scores of all models summarized in a boxplot.

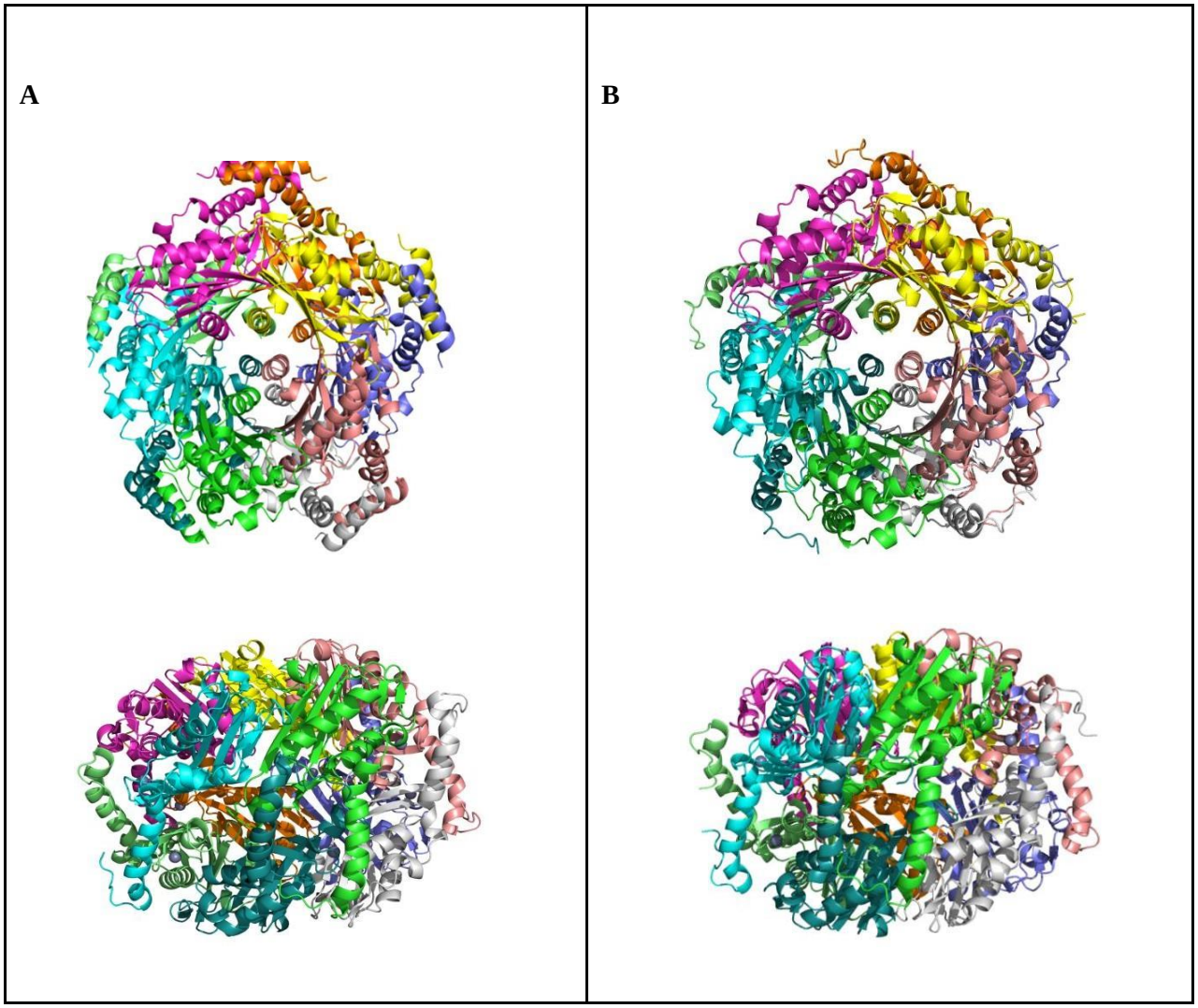


Figure 3.10: Top 3D models of Plasmodium species complete biological unit. **A**: *P. falciparum* and **B**: *P.vivax*, each chain is labelled differently. Figure shows top and side view of each generated homo-decameric model structure.

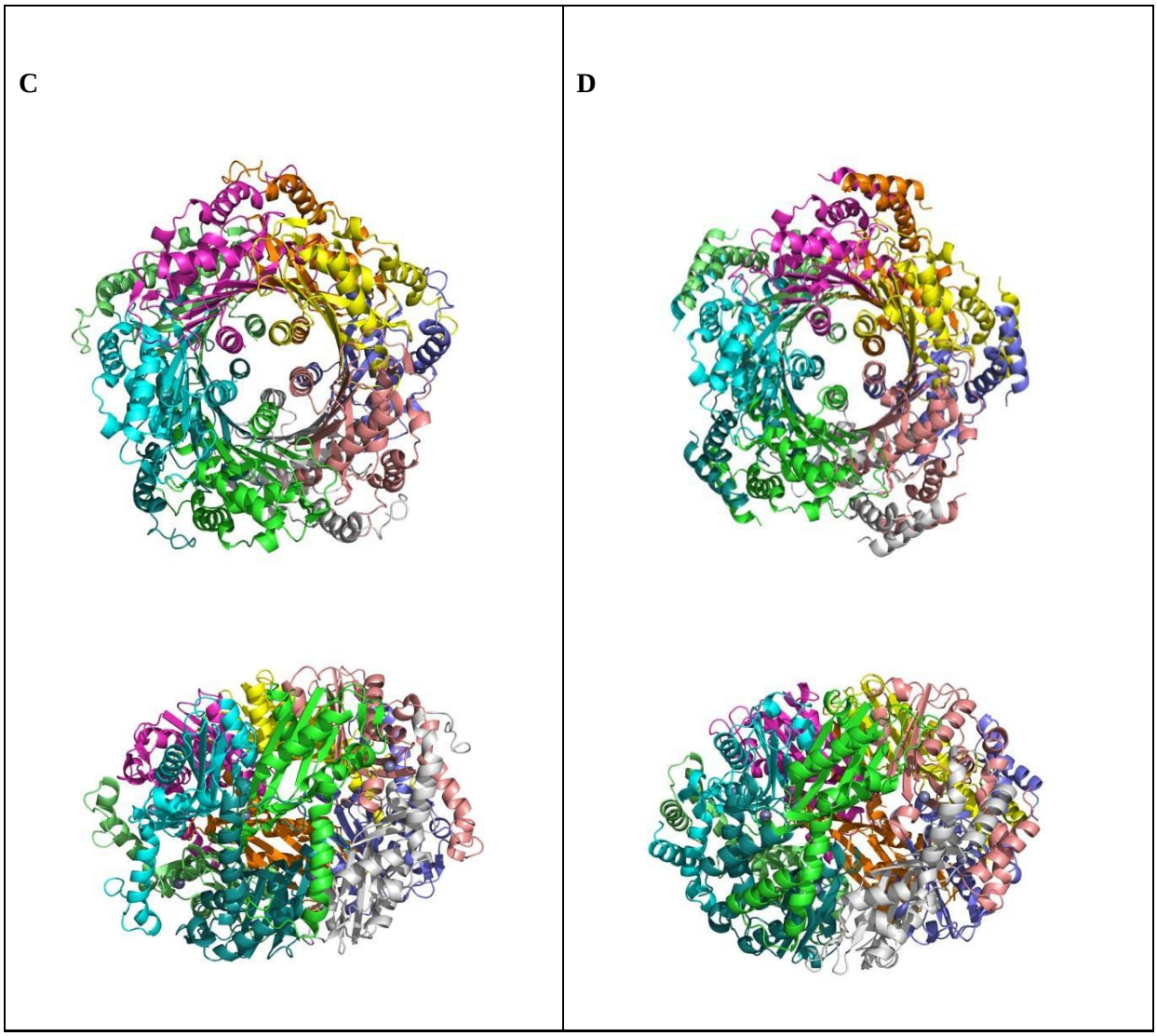


Figure 3.11: Top 3D models of Plasmodium species complete biological unit. **C**: *P.malariae* and **D**: *P.ovale*, each chain is labelled differently. Figure shows top and side view of each generated homo-decameric model structure.

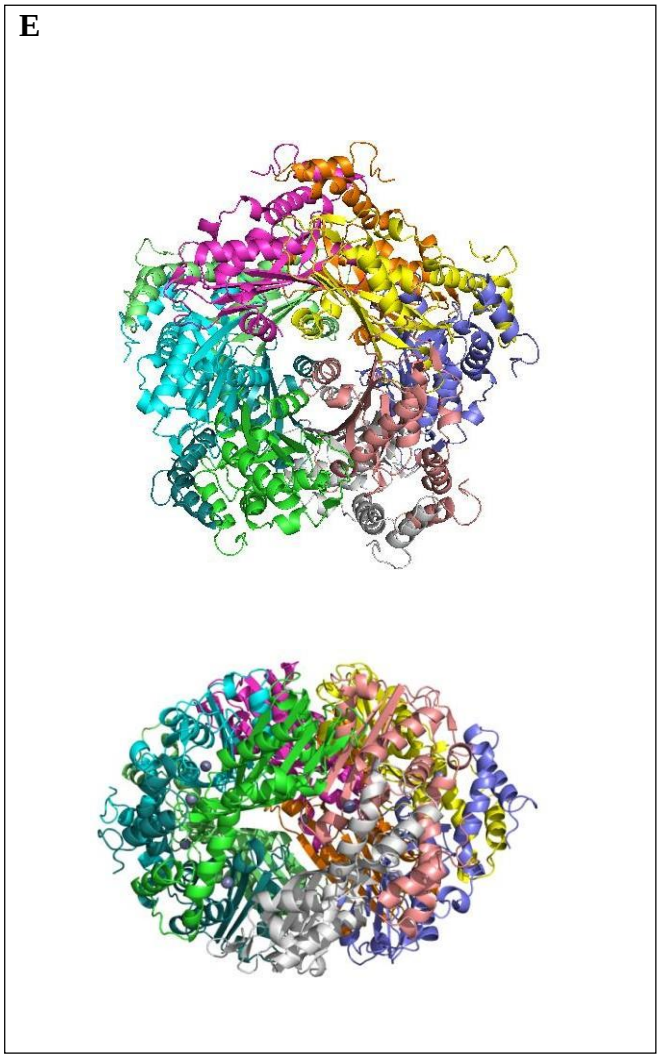


Figure 3.12: Top 3D model of Plasmodium species complete biological unit. **E**: *P.knowlesi*, each chain is labelled differently. Figure shows top and side view of the generated homo-decameric model structure.

Table 3.5: Summary of the top three model-quality assessment scores for each Plasmodial GCH1 protein. Models were evaluated using their z-DOPE-score, QMEAN Z-score and ProSA Z-scores

ORGANISM	BEST THREE MODELS	PROSA SCORE	QMEAN SCORE	QMEAN Z-SCORE	MODELLER Z-DOPE SCORE
P.falciparum	Model 1	-4.50	0.647	-1.248	-0.70
	Model 2	-4.42	0.650	-1.213	-0.70
	Model 3	-4.90	0.653	-1.180	-0.70
	Model 1	-4.70	0.669	-0.995	-0.81
P.malriae	Model 2	-4.56	0.676	-0.901	-0.81
	Model 3	-4.57	0.674	-0.924	-0.79
	Model 1	-4.35	0.597	-1.841	-0.54
P.ovale	Model 2	-4.32	0.597	-1.843	-0.54
	Model 3	-4.31	0.599	-1.819	-0.54
	Model 1	-5.44	0.643	-1.177	-1.04
P.vivax	Model 2	-5.38	0.661	-1.081	-0.99
	Model 3	-5.36	0.635	-1.264	-0.99
	Model 1	-3.71	0.572	-1.971	-0.68
P.knowlesi	Model 2	-3.85	0.566	-2.042	-0.67
	Model 3	-4.02	0.568	-2.016	-0.67
Template	Crystal structure	-4.70	0.694	-0.678	-1.40

Table 3.6: PROCHECK local quality assessment. Percentage values indicate the number of residues in the most favoured, additional allowed, generously allowed and disallowed regions. All models had at least 90% residues in the most favoured region

Organism	Model	PROCHECK validation report			
		Number of residues in favoured region/Core (Percentage value)	Number of residues in additional allowed region (Percentage value)	Number of residues in generously allowed region (Percentage value)	Number of residues in disallowed region (Percentage value)
P.falciparum	Model 1	90.6% core	7.0% allow	2.2% gener	0.2% disall
	Model 2	90.6% core	7.3% allow	% gener	0.2% disall
	Model 3	91.0% core	6.7% allow	% gener	% disall
P.malariae	Model 1	92.3% core	6.0% allow	1.5% gener	% disall
	Model 2	91.6% core	6.3% allow	1.8% gener	0.3% disall
	Model 3	92.5% core	5.6% allow	1.8% gener	0.1% disall
P.ovale	Model 1	92.2% core	6.2% allow	0.8% gener	0.8% disall
	Model 2	92.6% core	6.0% allow	0.7% gener	0.8% disall
	Model 3	91.4% core	7.2% allow	0.6% gener	0.8% disall
P.vivax	Model 1	92.8% core	5.2% allow	1.5% gener	0.5% disall
	Model 2	90.3% core	7.3% allow	1.9% gener	0.4% gener
	Model 3	91.9% core	5.6% allow	1.8% gener	0.7% disall
P.knowlesi	Model 1	89.3% core	7.5% allow	2.4% gener	0.7% disall
	Model 2	89.5% core	7.4% allow	2.6% gener	0.6% disall
	Model 3	90.2% core	7.4% allow	2.1% gener	0.3% disall
Template	Crystal structure	95.6% core	4.4% allow	0.0% gener	0.0% disall

Chapter summary

In this chapter, high quality models of GCH1 protein from the five human-infective Plasmodium species were generated. A template structure was successfully retrieved and validated. The template structure from *T.thermophilus* GCH1 protein (PDB ID: 1WUR) was shown to be suitable by sharing high percentage similarity of more than 30% to the target sequences (Table 3.2). The template structure also showed high sequence coverage with minimal gaps and high structure resolution of 1.82 Å. Target-template sequence alignment was performed using the T-COFFEE alignment tool. Manual adjustments were applied on the sequences by trimming the N and C terminal regions as they were not covered sufficiently by the template sequence. GCH1 have a homo-decameric structure; initially one chain (Chain A) of the protein was modelled then it was found that the protein active site is formed between three adjacent chains, thus modelling one chain was not sufficient to represent the protein active site. Accordingly, the complete biological unit of the protein was constructed. The generated models showed consistency in their structures. The two face to face pentamers formed a homo-decameric functional unit of the protein. The N terminal region which is known for its regulatory function was unique, thus it was excluded from the sequence alignment and the models. Results of quality assessments showed that the predicated models were within the range of reliable experimental native crystal structures and were suitable for use in subsequent docking and simulation assays. The generated models provided an opportunity to study the protein biological unit and screening of all five human-infective Plasmodium species proteins against SANCDB compounds; which had not been done before.

CHAPTER FOUR

MOLECULAR DOCKING

Malaria is a parasitic infection disease transmitted to humans through the bite of an infected female Anopheles mosquito. The human-infective Plasmodium species are *P.falciparum*, *P.ovale*, *P.knowlesi*, *P.vivax* and *P.malariae.* The parasites have developed high resistance against available anti-malarial drugs. This resistance has resulted in reduced efficacy of anti-malarial drugs, causing a delay in clearing the parasites from the infected host blood system. As a result, the need to develop new treatment strategies and identifying other alternative metabolic targets for the treatment of malaria has become essential (Nicholas J. White, 2008). Advancements in computational biology and the availability of sequence data has accommodated this problem by allowing for the screening of thousands of new drugs with known molecular specificity and activity against the malaria parasite enzymes (Rao and Srinivas, 2011). This chapter aimed to detect favourably binding compounds from a collection of compounds deposited in SANCDB database. Lead compounds should exhibit selective binding towards the Plasmodium species. This was deduced by accurately identifying their binding affinities and interactions with the GCH1 enzymes via molecular docking. The identified compounds can then be regarded as lead compounds and validated further.

4.1 Introduction

Computational models combined with molecular docking have become important in drug discovery as they underline key binding interactions of potential protein inhibitors as well as the validation of homology models (Hillisch *et al*. 2004; Du *et al*. 2015). Molecular docking is used to predict the binding mode of the protein-ligand complexes by searching all possible conformations to find the best conformation at which the ligand is bound in a fitted pose, both geometrically and energetically (May and Zacharias, 2005). When compared to the available experimental methods, molecular docking is considered to be time and cost effective (Teodoro *et al*. 2001; Morris & Lim-Wilby 2008).

Molecular docking can be sub-divided into two main categories; blind and targeted docking. Blind docking is applied to the whole protein whereas targeted docking targets a subsite of the protein, usually the active site (Laurie & Jackson 2006; Cavasotto & Orry 2007). Furthermore, protein docking can be either rigid or flexible. In rigid docking, the ligand is made flexible; undergoing changes in its 3D structure while the protein is kept in a rigid state (Guedes, de Magalhães and Dardenne, 2014). The lock and key mechanism of protein-ligand binding was first proposed by Fischer in 1890. This proposed mechanism has been effectively applied in computational molecular docking approaches. In 1958 a modification of the lock and key mechanism was proposed by Koshland, which allows both the ligand and protein to change conformation during the interaction, forming a minimum energy protein-ligand complex, thereby introducing flexible docking (Koshland, 1995; Cramer, 2007). Flexible docking is more accurate. However, when compared to rigid docking, flexible docking is more computationally and time expensive. This is because of the increased conformational search space for full protein flexibility. Flexible docking is also further complicated by cross docking, in particular docking ligands from different ligand-receptor complexes of large compound libraries (Teague 2003; Bonvin 2006; Forrey *et al*. 2012).

4.1.1 Molecular Docking

Molecular docking can be performed using different programs such as AutoDock Tools (Morris G.M. and Dallakyan S., 2013), AutoDock Vina (Oleg Trott and Olson, 2010), MolDock (Thomsen and Christensen, 2006) and Gold (Verdonk *et al.*, 2003). These programs perform conformational searches by exploring all possible binding conformations of potential inhibitor compounds within the receptor binding sites. AutoDock implements an automated scheme that performs ligand-protein binding mode searches. It explores all possible spatial degrees obtained from rotational, translational and torsional degrees of freedom (Dias and de Azevedo Jr., 2008).

Autogrid is used for 3D calculations of the receptor molecule interaction on a predefined grid, in which a probe atom is used to go through the grid points while the interaction energies are calculated and stored separately. The grid energies are then used during the actual docking for rapid evaluation of the interaction energies. The product of this is a 3D grid that is built surrounding the coordinates of the protein target. During the simulation, electrostatic interaction energy of the ligand is calculated as the product of local values from the grid and partial charge on the atoms (Sallem, De Sousa and E Silva, 2007).

4.1.2 Docking simulation

Docking simulations involve the actual search for the best ligand conformation bound to the receptor. Several algorithms have been employed to search for the best ligand-protein binding mode, based on Simulated Annealing (SA) (Goodsell and Olson, 1990), Monte Carlo method (Hart and Read, 1992) and Genetic Algorithms (GAs). A Lamarckian Genetic Algorithm (LGA) has been implemented in the latest versions of AutoDock Tools (Wang *et al.*, 2008). LGA algorithm is based on natural genetics of biological evolution in which the ligand is represented as a set of values describing translation, orientation and conformation. Each state variable is conceived as a gene and the ligand state corresponds to the genotype. The atomic coordinates are the phenotypes (Morris *et al.*, 1998). LGA combines global and local search algorithms to find mapped genotypes from the phenotype. The local search is applied on the phenotypes to search for the local minimum, this followed by using a function to invert the phenotype into its corresponding genotype/ligand conformation (Morris *et al*. 1998; Dias & de Azevedo Jr. 2008; Morris G.M. & Dallakyan S. 2013)

4.1.3 Energy scoring function

The energy scoring function is important in molecular docking as it evaluates the predicted ligand conformations. Ligand conformations with low binding free energies are considered to have suitable binding modes thus, represent a favourable interaction. The estimation of binding energies considers several physicochemical properties such as bonded and non- bonded interactions, thermodynamic quantities, solvation and de-solvation energies (Jain 2006; Lionta *et al*. 2014). Scoring functions are categorised into force-field based, empirical, and knowledge-based functions (Lionta *et al.*, 2014). The force-field based scoring estimates the binding free energy by adding up the energy contributions of non-covalent interactions such as electrostatic interactions and intermolecular van der Waals (Huang, Grinter and Zou, 2010). AutoDock tool makes use of force-field based scoring function whereas AutoDock Vina makes use of a hybrid scoring function that combines knowledge-based potentials and empirical scoring functions. The use of this hybrid scoring function showed a significant improvement in both speed and accuracy of the docking (Trott and Olson, 2010).

The force-field based scoring is limited by its inaccuracy in estimating entropic contributions. On the other hand, knowledge-based and empirical scoring functions are limited by their dependence on the accuracy of the data set used to adjust its scoring functions. The free energy of binding scores is calculated in kcal/mol with favourable energy being on the negative number scale. This energy is used to calculate the inhibition constant (*K*i) which is dependent on the free energy of binding. A more positive free energy of binding corresponds to low inhibition constants and vice versa (Tanchuk *et al.*, 2015).

4.2 Methods

4.2.1 Data preparation for molecular docking

Molecular docking was carried out using the crystal structure of *Thermus. thermophilus* bacteria (PDB ID: 1WUR), the crystal structure of GCH1 from *Homo sapiens* (PDB ID: 1FB1) and the 3D generated models of the human-infective Plasmodium species: *P. falciparum*, *P.knowlesi*, *P.vivax*, *P.ovale* and *P.malarie*. The South African Natural Compounds database was used for screening (Hatherley *et al.*, 2015). The SANCDB library consists of 623 minimised compounds. These compounds were obtained from literature and have never been tested for anti-malarial activity. The compounds were recreated in pdb file format. The protein receptors were prepared by removing the crystallographic water molecules and adding hydrogen atoms using Discovery Studio visualization software (San Diego: Accelrys Software Inc., 2012).

4.2.2 Ligands and protein protonation

GCH1 crystal structures of the *T.thermophilus*, *Homo.sapiens*, 3D model structures of Plasmodium species and ligands were converted into the rigid (pdbqt) conformations using a python script provided by AutoDock tools (Appendix 3, script A3.1). The proteins and ligands were prepared by merging all hydrogen atoms and adding polar hydrogen atoms. This was followed by the calculation of Gasteiger charges and assigning atom types. The Gasteiger charge of the zinc atom was assigned manually to 1.125 (see section 5.3.2 for details). The preparation step of the ligands and receptors was automated using the python script run on a Linux based cluster.

4.2.3 Grid calculation and docking parameter file preparation

AutoGrid was used to calculate the grids for each receptor. This allows for the determination of the 3D grid of interaction energies which is based on the receptor coordinates (Goodsell *et al*. 1996). The grid points were set at 70, 70, and 70 Å for the cubic box and the grid spacing at 0.3522 Å. The grid was chosen to sufficiently cover the whole protein receptor. The cubic grid centre was set at the centre of the protein. The interacting receptor and ligand atom types and their corresponding interaction maps were also listed in the grid parameter file.

4.2.4 Docking simulations

Docking simulations were carried out using AutoDock Vina via an automated, customized Python script (Appendix 3, script A3.2). A Vina parameter file was created for each ligand and its receptor protein. The Lamarckian Genetic Algorithm was used for receptor-ligand conformational search in which the exhaustiveness was set to 192. The parameter file contained the cubic box size values and the coordinates of the central atom. Vina files containing the parameters were saved accordingly in a folder. To accommodate the computational cost docking simulations were done on the Centre of High-Performance Computing (CHPC) cluster.

4.2.5 Docking validation

In order to validate the docking protocol, a validation step was performed. This validation step is important to check for the ability of the docking protocol to reproduce the correct ligand poses. Docking validation was done using a known co crystallized ligand of the *T.thermophilus* GCH1 protein 8-oxoguanine derivative of GTP (8-oxo-GTP) (Tanaka *et al.*, 2005). The docking simulation was performed using same parameters mentioned above. The validation was done on three categories of the receptor, first on one chain of the protein receptor, three chains that fully represent the active site and finally the complete biological unit of the protein which contains ten chains. The re-docking was done using AutoDock Vina. The original ligand was re-docked to *T.thermophilus* GCH1 crystal structure and the docking pose was compared to that in the crystal structure.

4.2.6 Docking analysis

A customised python script was used to perform docking analysis. The script extracted the best ligand conformation of the population and the corresponding free energy of binding from the log files. The extracted ligands were then converted from the rigid (pdbqt) format to a pdb format and saved in a designated folder. Accelrys Discovery Studio was used to visually identify ligand interactions. The free energy of binding values were exported to a Microsoft Excel spreadsheet and plotted as an energy graph and a heat map to summarize the docking results. The ligands were then ranked based on their free energy of binding. The top ligands that were bound to the active site were selected and analysed further. LigPlot and Accelrys Discovery Studio were used to identify the protein-ligand interactions.

4.3 Results and Discussion

Molecular docking was performed in two stages of three categories. First stage was the docking validation using the crystal structure of *Thermus. Thermophilus* bacteria (PDB ID: 1WUR) and its co crystallised ligand 8-oxo-GTP. The second stage involved screening of the SANCDB compounds against the 3D modelled receptors of *P. falciparum*, *P. knowlesi*, *P.viavx*, *P.ovale* , *P.malriae* and the crystal structure of GCH1 of *Homo sapiens* (PDB ID: 1FB1). Docking validation and simulations were performed on one chain of GCH1 structure (Chain A) of all receptors, three chains (Chain A, B and J) of *P. falciparum* and human receptors only and the complete biological unit structures of *P. falciparum* and human receptors. The three chains represented a complete active site unit whereas the biological unit receptors represented all ten zinc-containing active sites. All protein receptors were prepared by removing the crystallographic water molecules and hydrogen atoms were added.

4.3.1 Docking validation

The co-crystallized ligand was removed and re-docked to the receptor. The presence of a ligand bound structure in the PDB provided important structure activity relationship information, which made this method more reliable to use. The co-crystallized ligand can be easily used for assignment of the screening area and validation of the docking protocol reproducibility.

The validation showed that the docking protocol was able to reproduce the binding of the co-crystallized ligand. Docking validation and simulations were performed on one chain of the protein receptor (Figure 4.1), three chains (Figure 4.3) and the complete structure of GCH1 protein (Figure 4.5), this was important to monitor the ligand's interaction mode over different courses of the structure. Docking validation of the three chains aimed to confirm the binding interaction towards a fully shaped active site, which was not defined in the one chain structure; however, the one chain validation/simulations was important to mimic the co-crystallized ligand's interaction mode when the protein monomer is first formed at a molecular level. The complete structure was used to recognize the reproducibility of the crystal structure pose among the ten identified active sites.

Liglplot software was used to generate a 2D diagram showing the interactions of the re- docked ligand with GCH1 protein chain A (Figure 4.2), Chain A, B and J (Figure 4.4) and the complete biological unit (Figure 4.6). The validation against the protein monomer (chain A) did not reproduce the ligand pose accurately. This can be reasoned to be due to the absence of the full active site shape in which the ligand interacts with residues from each of the three chains and not just one chain. Figure 4.4 illustrates the additional interacting residues which were missing in the validation of the one chain. This includes Lys 134, Ala 37, Arg 137, Glu 150, Leu 132, Leu130 and Ser 133 from Chain B and Arg 64 from chain J. The additional residues can cause stability of the re-docked ligand providing more accurate reproduction of the original ligand pose.

In Figure 4.6 it can be seen that His 177 in the active site forms a non-bonded interaction with the ligand. As illustrated in literature (Kümpornsin *et al.*, 2014) this interaction is responsible for the protonation of N-7 of the ligand and the cleavage of N7/C8 at the guanine ring. The ligand also formed a hydrogen bond with His 111. His 111 is known to coordinate to the metal ion at the active site as well as causing the protonation of oxygen at the ribose moiety of the ligand, which results in cleavage of the ribose ring (Yoko Tanaka *et al.*, 2005).

Glu 149 and Glu 150 formed two hydrogen bonds with the N1 and N2 atoms of the guanine ring, respectively. A non-covalent interaction was observed between Ser 133 and the hydroxyl groups of the ligand ribose ring. Arg 137, Arg 183, Arg 64 and Lys 134, also formed non-covalent interactions with the phosphate groups of the ligand. Ala 87, Leu 133, Val 148, Gly 131 and Ile 129 formed non-bonded interactions with the ligand. The above stated interactions were identified from the re-docking and they were all similar to the co- crystallized ligand interactions. All observed interactions were in confirmation with the interactions stated in previous studies (Yoko Tanaka *et al.*, 2005; Kümpornsin *et al.*, 2014). This proved that the docking protocol used is fairly accurate, since it can reproduce the correct conformational poses.

Docking validation results of the protein biological unit structure were similar to the ones obtained from the three-chain protein (Figure 4.4 and Figure 4.6). Based on the reproducibility of the crystal structure pose of Chain A, B and J and the complete protein structure we can deduce that the best overlay between the re-docked and original ligand was observed on both; the three chains and the complete protein structure. Thus, working on the protein three chains will be sufficient to describe the protein functionality and it will save the computational cost and time.

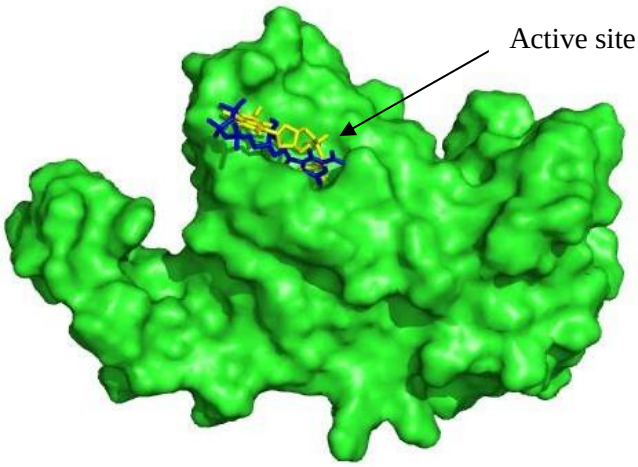


Figure 4.1: Docking validation results of GCH1 crystal structure (Chain A). The re-docked ligand is shown in yellow and the co crystallized ligand is shown in dark blue. Significant overlap between the re-docked and co-crystallized ligand was observed.

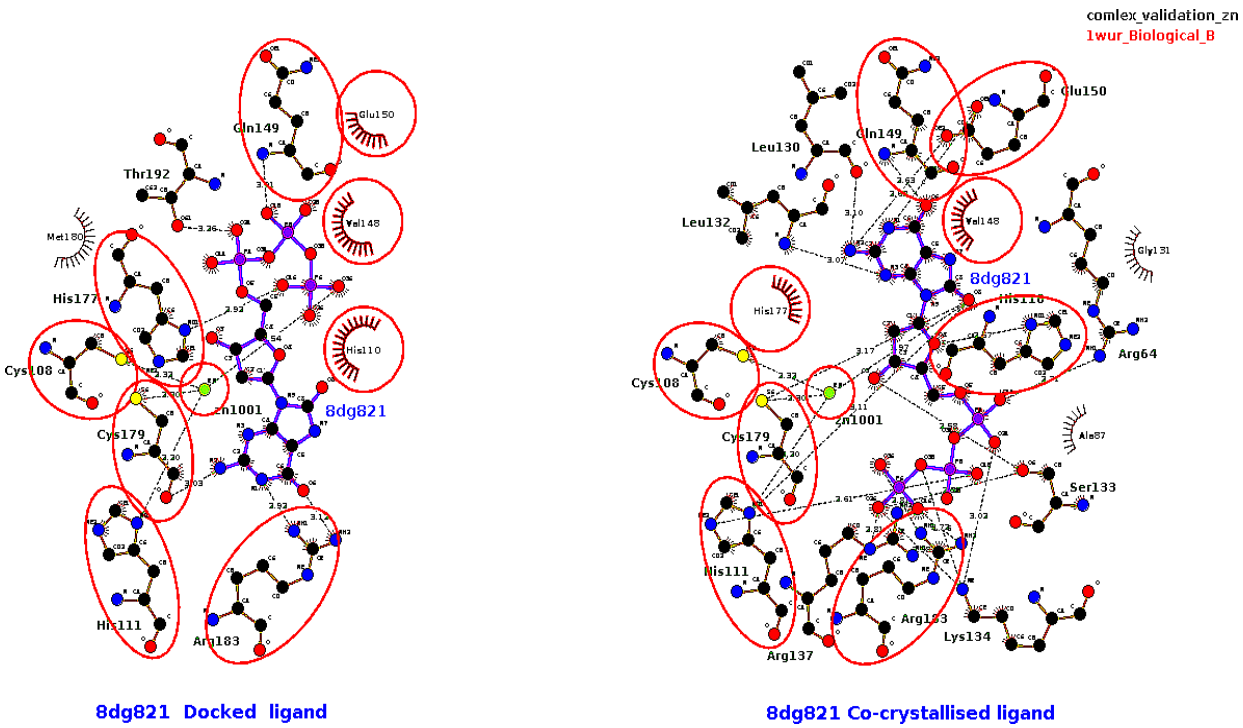


Figure 4.2: 2D diagram of GCH1 protein (Chain A) and its co-crystallized ligand interactions generated by Liglplot. Interacting residues of the ligand and protein are shown in balls and sticks. Hydrogen bonds are shown in green dashed line with a specification of their length in Å. The purple solid lines represent the ligand bonds and the spoked arches represent the protein residues involved in non-bonded interaction. The red circles show the residues of both docked and co-crystallized ligands that share similar/common binding to the receptor.

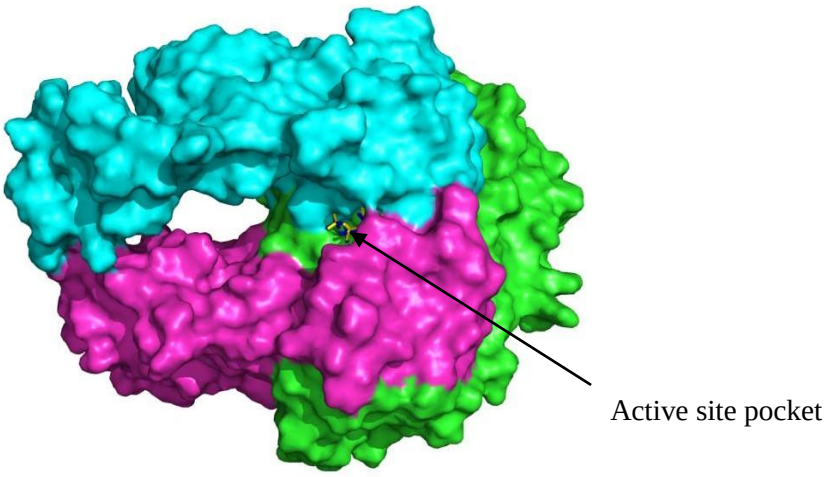


Figure 4.3: Docking validation results of GCH1 crystal structure (Chain A, B and J). The re-docked ligand is shown in yellow and the co crystallized ligand is shown in dark blue. Both ligands were docked similarly in the active site pocket.

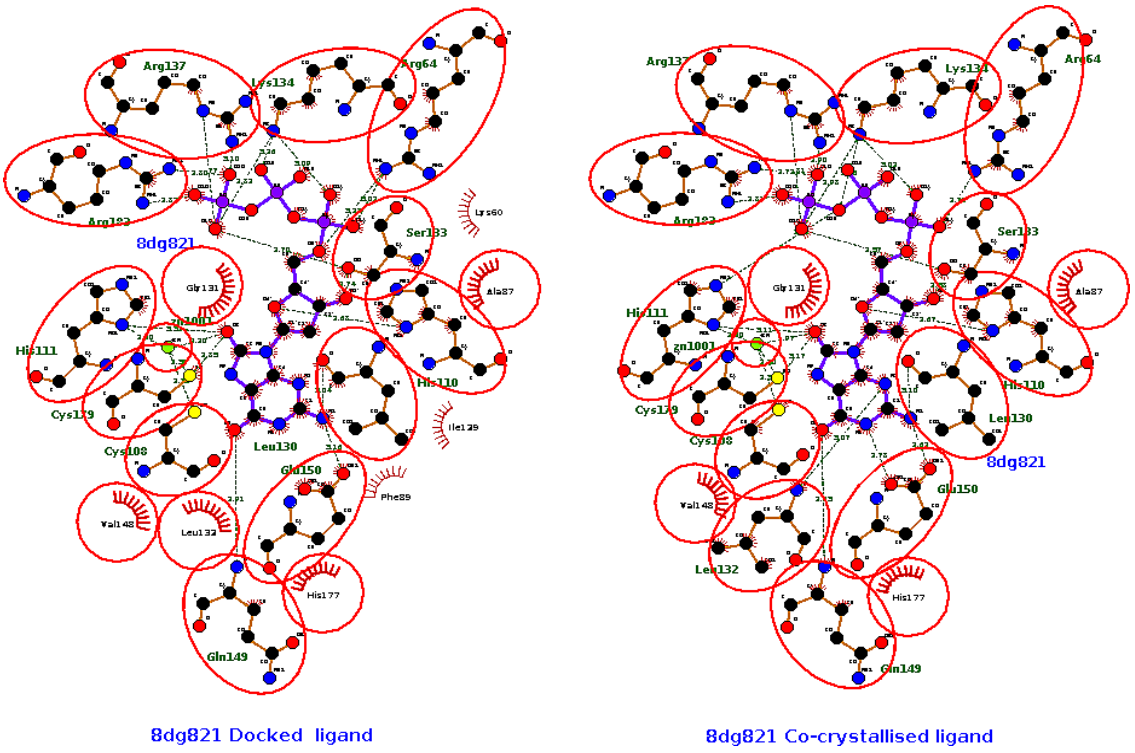


Figure 4.4: 2D diagram of the GCH1 protein (Chain A, B and J) and its co-crystallized ligand, interactions were generated by LigPlot. More residue interactions were observed due to the availability of the protein three chains forming the active site pocket.

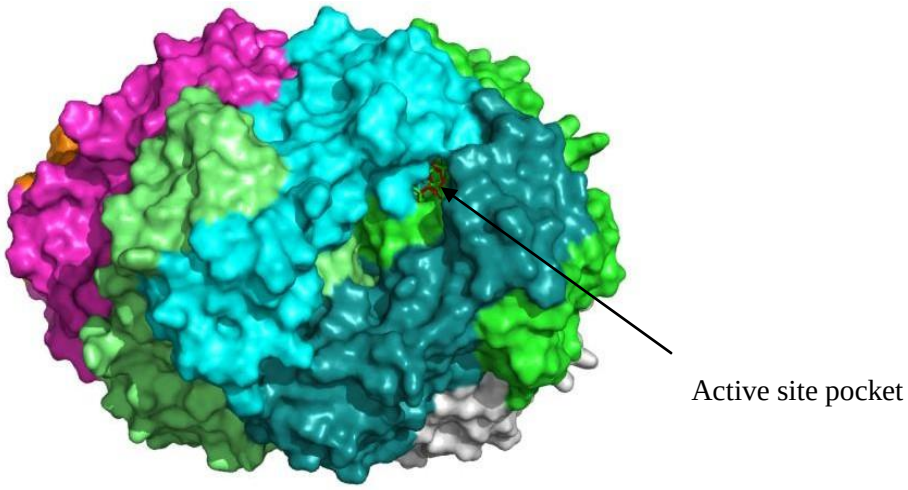


Figure 4.5: Docking validation results of the GCH1 complete biological unit. The re-docked ligand is shown in yellow and the co crystallized ligand is shown in red. Both ligands were docked similarly in one of the GCH1 active sites.

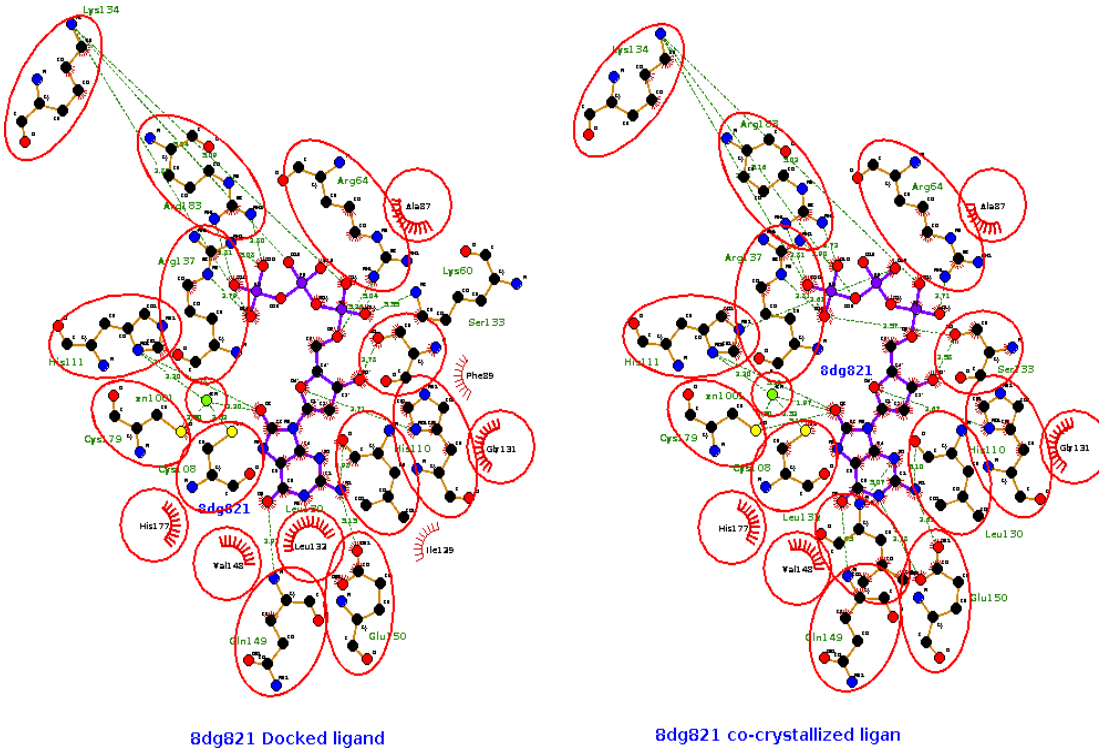


Figure 4.6: 2D diagram of the GCH1 complete protein structure and its co-crystallized ligand, interaction generated by LigPlot. More residue interactions were observed due to the availability of the ten chains that build the complete functional unit of the protein. The ligand interacts with residues from two chains of the same pentamer and one chain from the other pentamer.

4.3.2 SANCDB screened compounds

Virtual screening was carried out using 623 selected compounds from SANCDB database; these compounds were optimised in order to obtain their lowest energy structures. The selected compounds have not been tested for anti-malarial activity before. The protein-ligand complexes from Plasmodium species and the human protein were visualised using PyMOL visualisation tool (Figure 4.7 - 4.11). From this, it was observed that the Plasmodium proteins exhibited similar binding with their inhibitor compounds. The screened compounds were bound to two distinct sites: the active site pocket and the front shallow tunnel of the receptor proteins. In contrast, none of the screened compounds were bound to the human GCH1 protein active site rather to its surface.

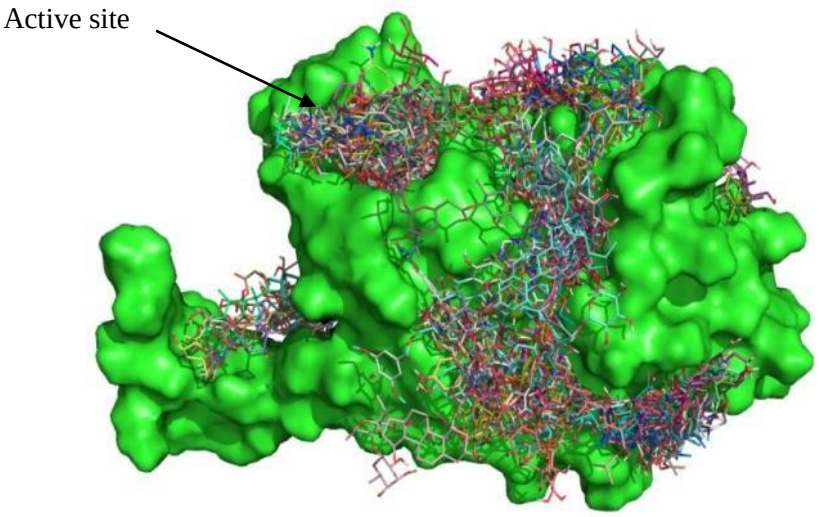


Figure 4.7: *P. falciparum* protein-ligand complexes. The arrow indicates the active site pocket. Ligand molecules are shown as sticks and the protein as a solid surface.

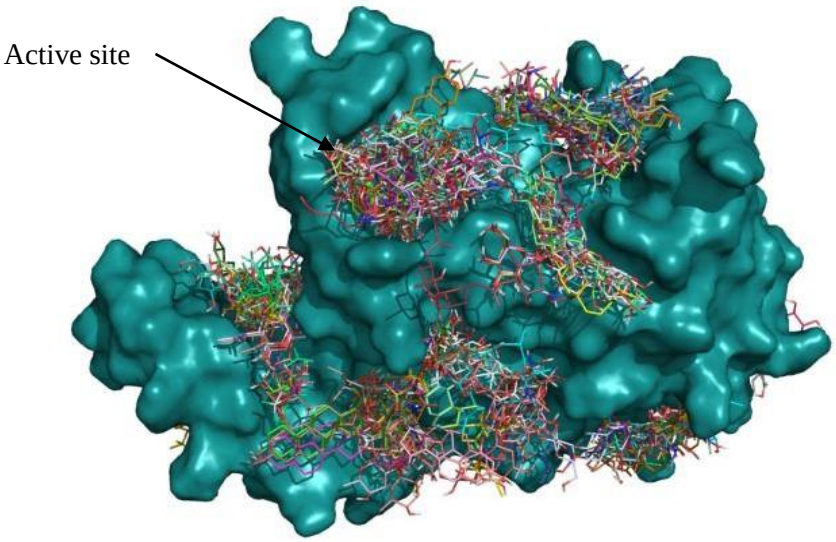


Figure 4.8: *P.ovale* protein-ligand complexes. Ligand molecules are shown as sticks and the protein as a solid surface.

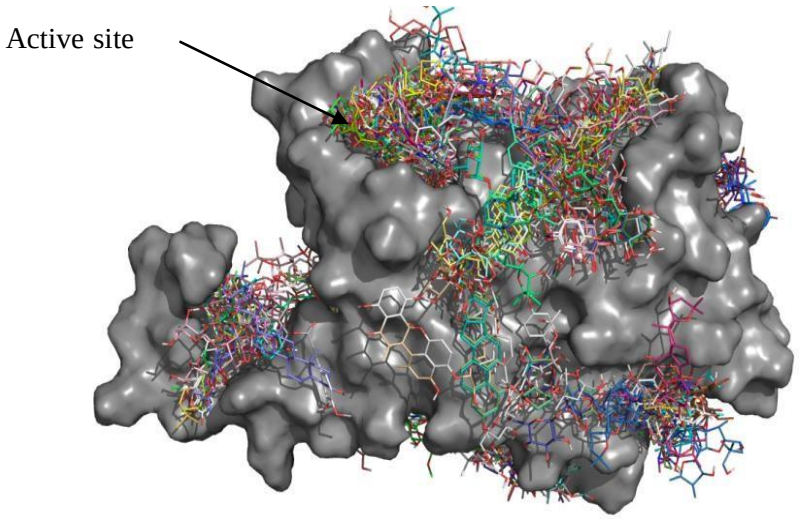


Figure 4.9: *P.malariae* protein-ligand complexes. Ligand molecules are shown as sticks and the protein as a solid surface.

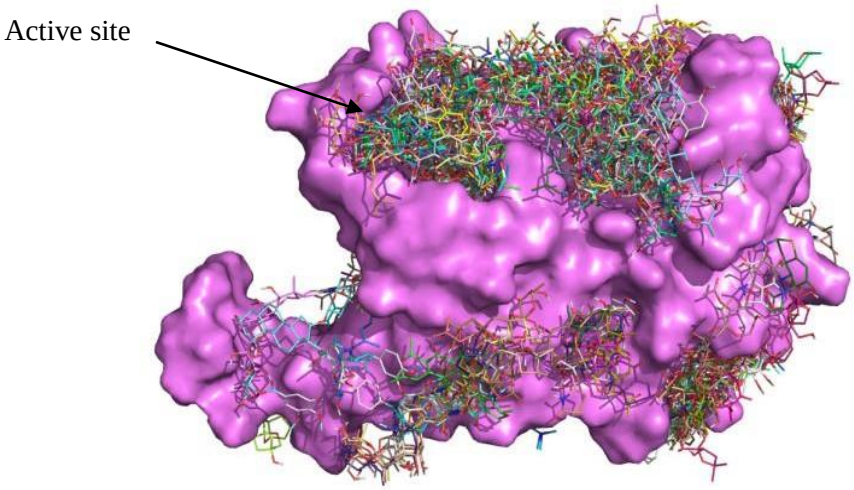


Figure 4.10: *P.knowlesi* protein-ligand complexes. Ligand molecules are shown as sticks and the protein as a solid surface.

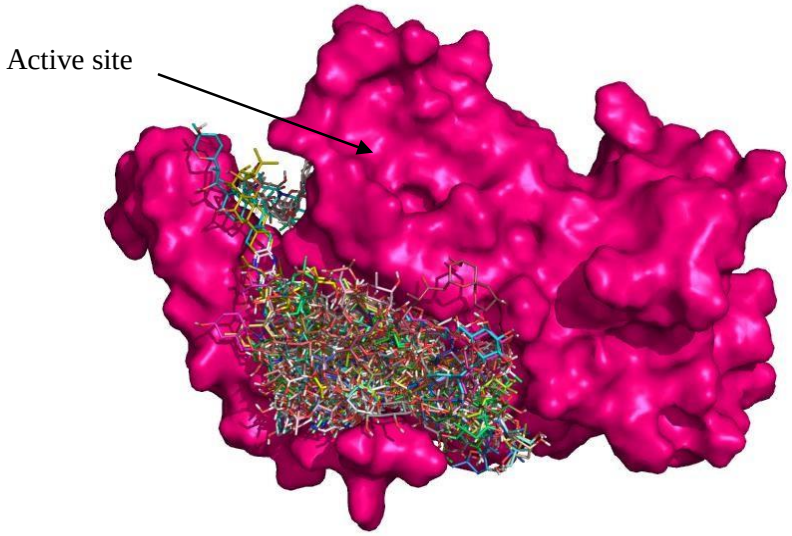


Figure 4.11: Human GCH1 protein-ligand complexes. Ligand molecules are shown as sticks and the protein as a solid surface. None of the screened compounds were bound to the active site residues.

Protein-ligand association can be assessed by the magnitude of the binding free energy. Negative values of the binding free energy signify a spontaneous binding to the receptor. Whereas, positive values indicate that the binding is consuming energy to occur. Therefore, lower free binding energy means higher affinity towards the receptor and vice versa. The free energy of binding of the screened SANCDB compounds was captured and exported into a Microsoft Office spreadsheet, then sorted based on their lowest energy. This was done for each receptor protein. Each of the Plasmodial screened compounds was compared against its corresponding compound bound to the human protein, by calculating the difference in the free energy of binding. This was followed by generating a heat map to represent the binding free energy of all screened compounds and to compare the binding free energy between the Plasmodium species and the human GCH1 protein monomers (Figure 4.12). From the heat map, it was observed that some of the screened compounds have relatively low free energies of binding which was consistent across all the Plasmodium homologs; this indicates a selective binding based on the free energy of binding. It was also observed that the screened compounds exhibited lower binding free energy towards the human GCH1 protein. Docking results against one chain structure showed that the majority of the compounds have high binding affinity towards the human receptor rather than the Plasmodium's. When these compounds were investigated further, it was found that they were not bound to the active site pocket of the human receptor but to the surface instead (Figure 4.11). This serves the purpose of selectivity towards Plasmodium species. Additionally, the absence of a complete active site shape reduced the binding affinity towards the active sites of the Plasmodium proteins.

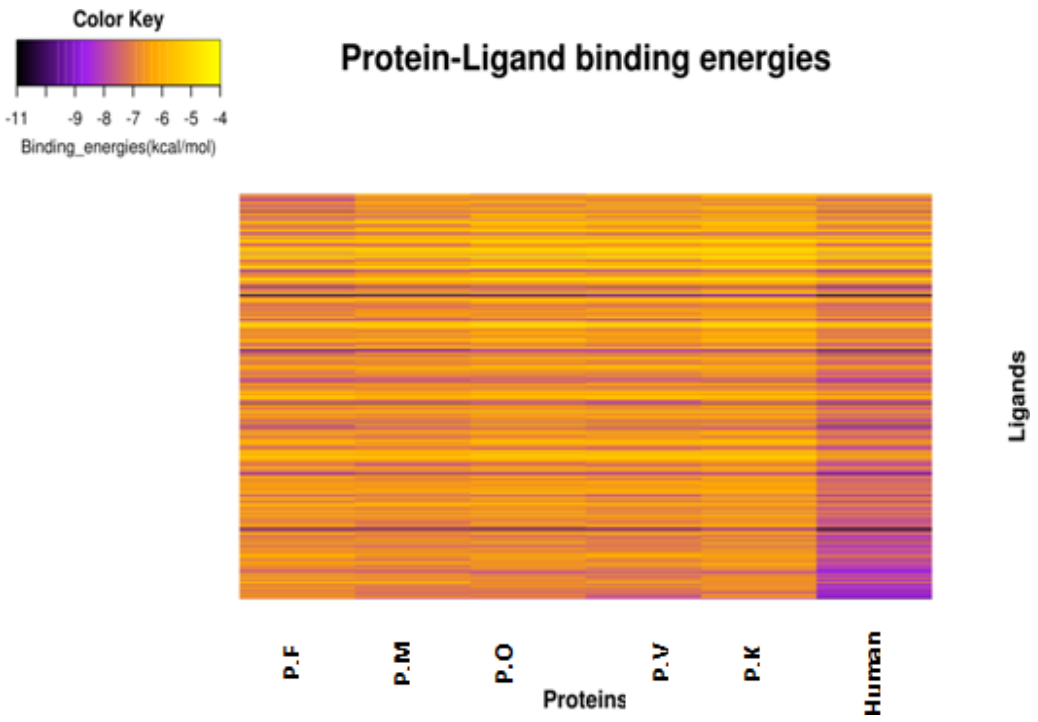


Figure 4.12: Heat map of binding energies of all docked SANCDB compounds against GCH1 protein (Chain A). Low binding free energy scores are shown in black indicating high biding affinity, whereas yellow indicate high binding-energy scores (lower binding affinity). Screened compounds exhibited lower binding affinity towards the human GCH1 protein; these compounds were not bound in the protein active site. In addition, the absence of the full active site shape in the one chain of the protein caused the compounds to have higher free energy of binding energies as some important interactions were not formed.

Docking results against GCH1 chain A, B and J resulted in an increase of the binding affinity towards the Plasmodium proteins due to the presence of the active site full shape (Figure 4.14). The screened compounds were visualized to identify their binding sites using Discovery studio software. 80 compounds were docked to the active site pocket of the Plasmodium proteins; this explains their low free energy of binding scores while others were bound to the surface of the protein. From the docking results against the human GCH1 chain A, B and J it was observed that most of the screened compounds were bound to the protein surface (Figure 4.13). Molecular docking against the complete biological unit of the Plasmodium and human GCH1 proteins resulted in a significant increase of the binding affinity towards the Plasmodium proteins, showing more selective binding based on the free energy of binding (Figure 4.16). It also showed uniform binding among the ten chains of the protein (Figure 4.15).

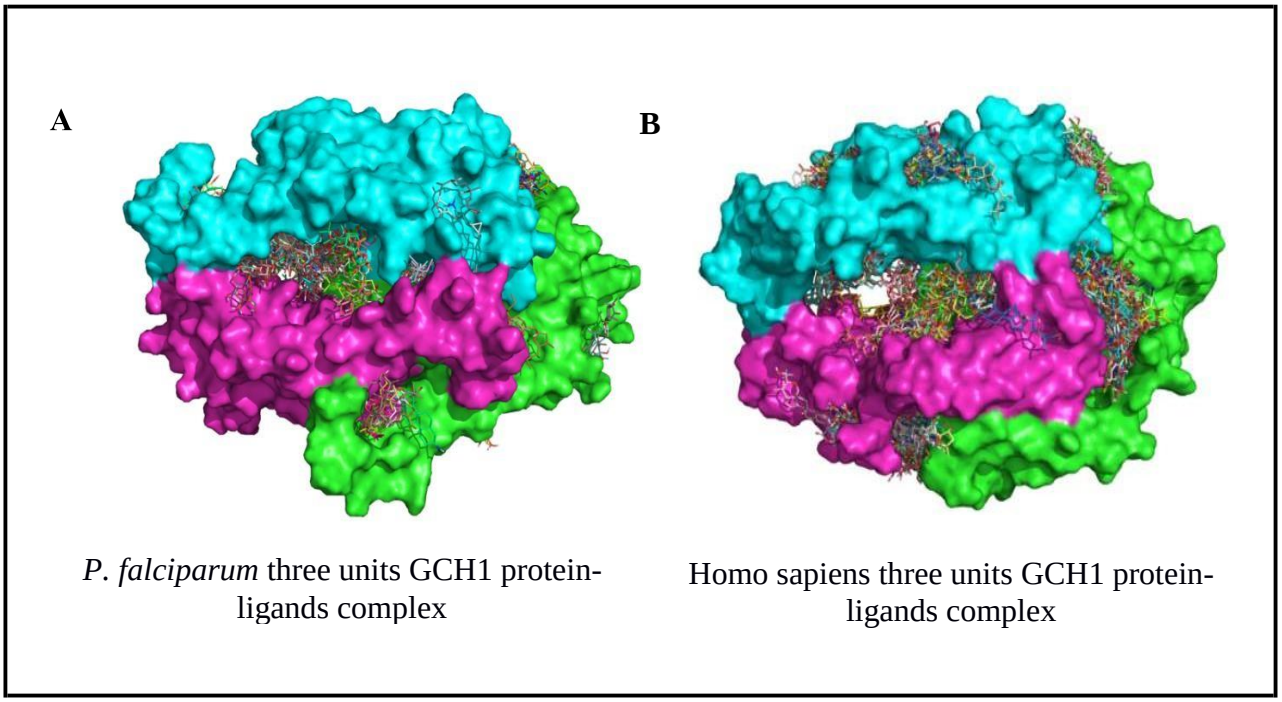


Figure 4.13: GCH1 protein-ligand complexes. The protein consists of chain A, B and J representing the protein first active site. Each chain is coloured differently. The ligand molecules are shown as sticks and the proteins as surfaces. The figures were generated using PyMOL. Ligands were bound to two distinct sites: the active site pocket and the surface. The majority of the SANCDB compounds were bound to the human GCH1 protein surface. Fewer compounds are observed on the *P. falciparum* surface.

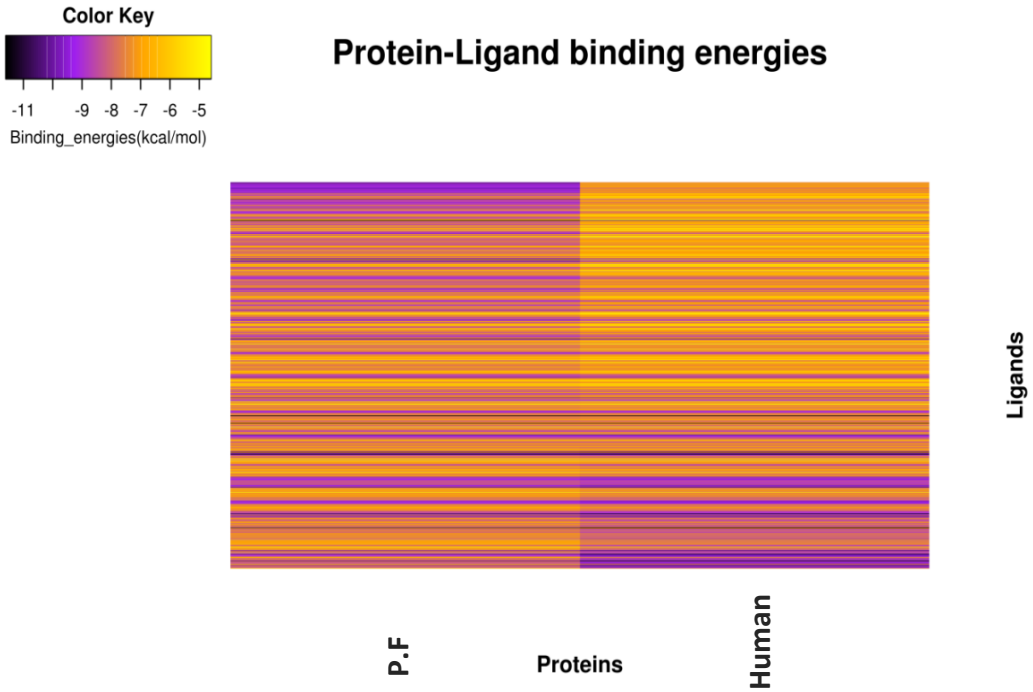


Figure 4.14: Heat map of binding energies of all docked SANCDB compounds against GCH1 protein (Chain A, B and J). Low binding free energy scores are shown in black/purple. Yellow indicates high binding energy scores (lower binding affinity). Ligands binding with low energy scores towards the Plasmodium proteins only are considered to be target specific. The availability of the full active site shape between the three chains resulted in a decrease of the binding energies of the compounds bound to the active site, hence selectivity was observed towards the Plasmodium proteins.

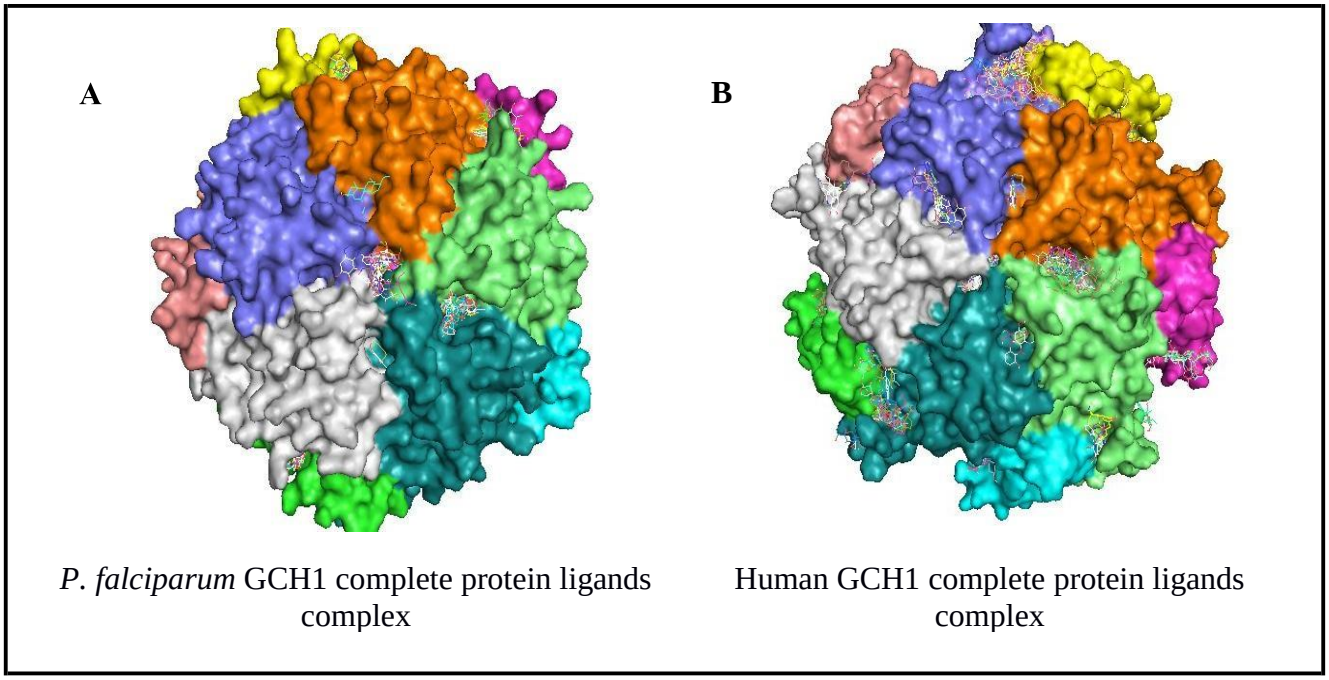


Figure 4.15: GCH1 protein-ligand complexes. The complete protein structure consists of ten chains (A-J), representing the ten active sites. Each chain is labelled differently. The ligand molecules are shown as sticks and the proteins as surfaces. Ligands were bound to two distinct sites: the active site pocket and the surface. The majority of the SANCDB compounds were bound to the human GCH1 protein surface. Fewer compounds are observed on the *P. falciparum* surface.

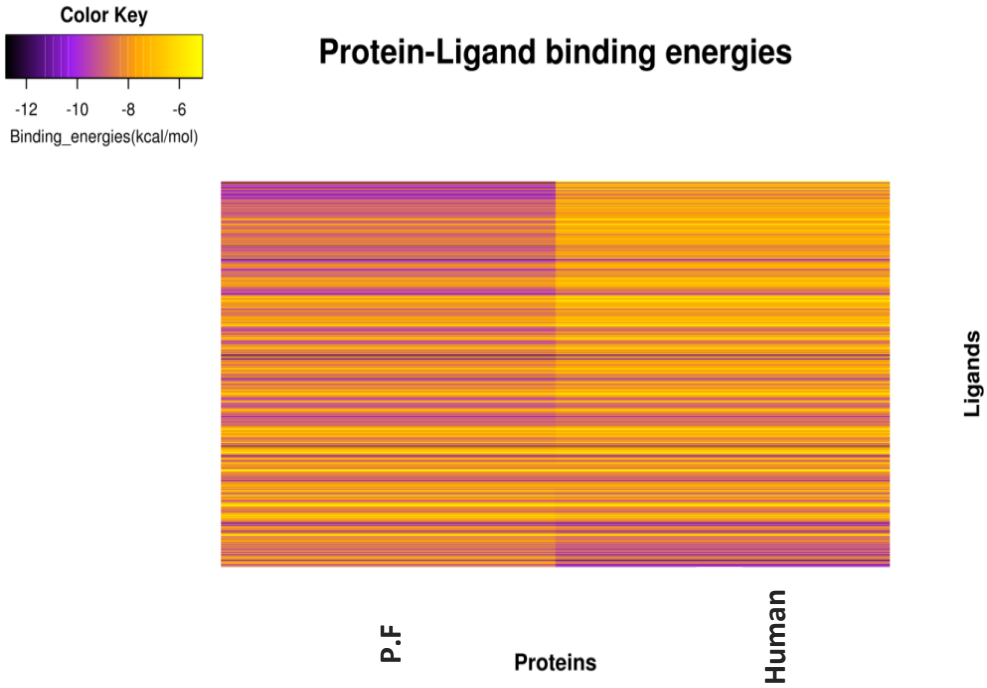


Figure 4.16: Heat map of binding energies of all docked SANCDB compounds against GCH1 protein complete structure. Low binding free energy scores are shown in black/purple which indicates high biding affinity. Yellow indicates high binding-energy (lower binding affinity).

Compounds bound to the *P. falciparum* protein active site were selected and their corresponding binding energies were compared to the human GCH1 protein. It was found that these compounds have lower binding free energies for *P. falciparum* protein. A heat map was generated to compare the binding free energy of the compounds bound only to the *P. falciparum* active site pocket and their corresponding free energy of binding in human protein. An obvious lower free energy of binding was observed towards the *P. falciparum* GCH1 protein (Figure 4.17).

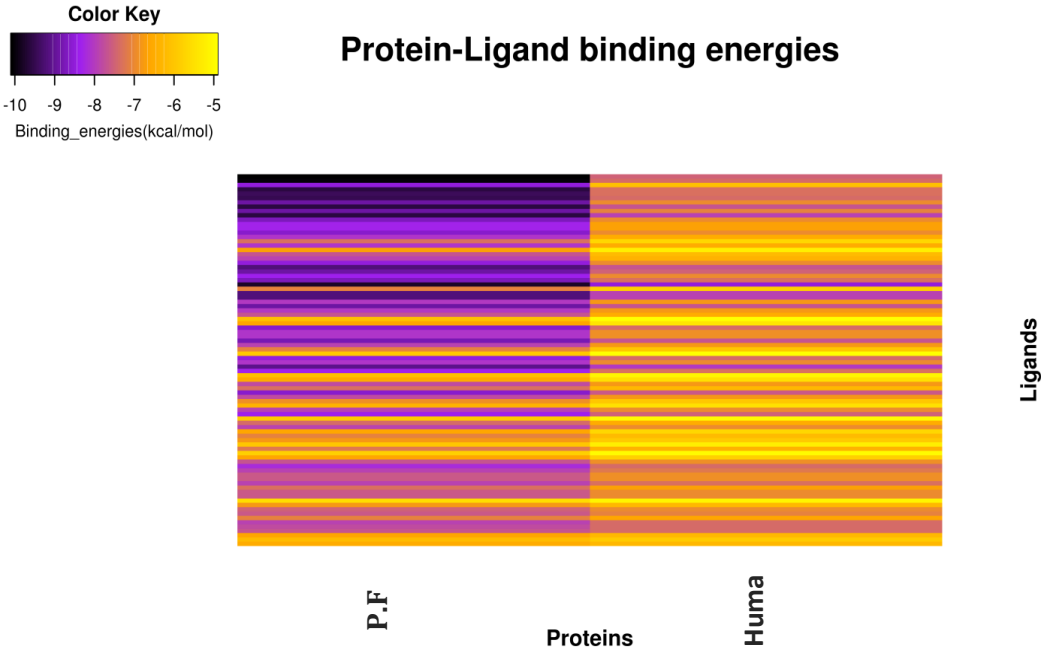


Figure 4.17: Heat map of the estimated free energy of binding of the best docked compounds bound to the active site of *P. falciparum* GCH1 protein and their corresponding binding free energy in the human GCH1 protein. The energy code ranges from high (yellow) to low (black). Higher selectivity was observed towards the *P. falciparum* active site.

A bar graph was created showing the binding free energy of the top 80 compounds bound to the *P. falciparum* GCH1 protein active site pocket (Appendix 3, Figure A3.1**).** Binding free energy of the top 20 compounds was also captured and represented in a bar graph (Figure 4.18). The bars demonstrate the binding free energy difference between the *P. falciparum* and human GCH1 protein. A significant difference in the binding free energy was observed, in particular lower binding free energy towards *P. falciparum* GCH1 protein. Compounds with a binding free energy less than -9.0 kcal/mol were selected and tabulated in Table 4.1. The difference of the compound's binding free energy between the *P. falciparum* and Human GCH1 proteins was calculated and represented by ΔG. The more negative ΔG values are, the lower the binding free energy in *P.falciparum*, which indicates favourable selective binding for the *P. falciparum* GCH1 protein.

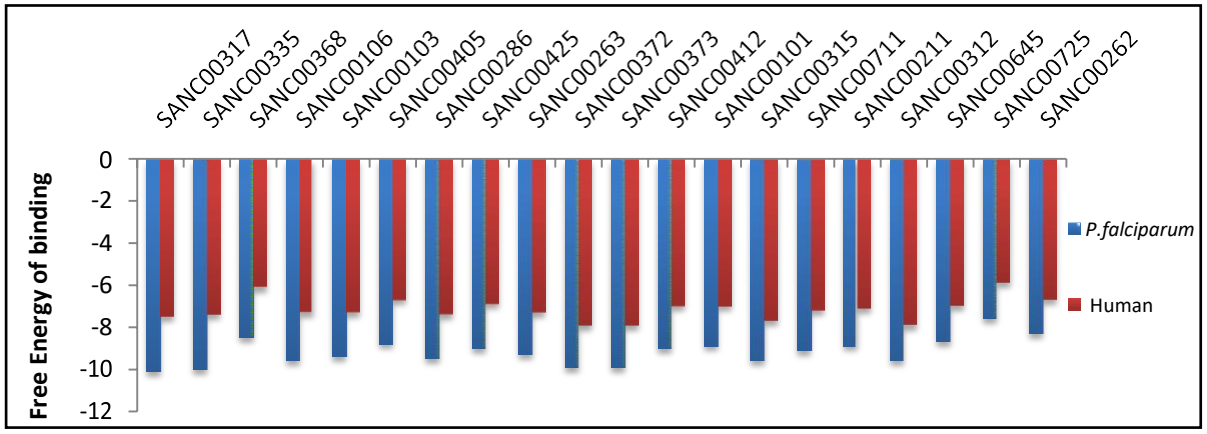


Figure 4.18: Bars graph displaying the free energies of binding of SANCDB compounds against the *P. falciparum* GCH1 protein (active site) versus its human homolog.

Table 4.1: Best hit compounds of SANCDB database against *P. falciparum* GCH1 protein with free energy of binding < -9.0 kcal/mol. The top compounds were all bound to the *P. falciparum* active site pocket. ΔG shows the difference of the compounds binding free energy between the *P. falciparum* and human GCH1 proteins. Lower ΔG indicates more selective binding towards the Plasmodium GCH1 protein.

Ligand	*P. falciparum*	Human	ΔG
SANCDB00317	-10.1	-7.5	-2.6
SANCDB00335	-10	-7.4	-2.6
SANCDB00106	-9.6	-7.3	-2.3
SANCDB00315	-9.6	-7.7	-1.9
SANCDB00286	-9.5	-7.4	-2.1
SANCDB00103	-9.4	-7.3	-2.1
SANCDB00101	-8.9	-7	-1.9
SANCDB00211	-8.9	-7.1	-1.8
SANCDB00368	-8.5	-6.1	-2.4

SANCDB00103

SANCDB00103 exhibited high binding affinity in the protein active site pocket explained by the low free energy of binding of -9.4 kcal/ mol. This indicates higher estimated binding affinity consequently, higher inhibition constants. Discovery Studio was used to investigate the protein-inhibitor interaction and to generate a 2D representation of these interactions (Figure 4.19).

The inhibitor interactions included hydrogen bonds with Glu 319, His 280 (chain A), Lys 303 and Tyr 256 (chain B). His 280 and His 346 formed a π-π T-shaped interaction with the ligand aromatic ring; this interaction is known to be favourable and frequent in proteins catalytic sites (Cauët *et al.*, 2005). A hydrogen bond was formed between the carbonyl of the inhibitor and Gly 300 (chain B). One unfavourable interaction was formed with Arg 352 (chain A). The above mentioned interacting residues were located in the GCH1 active site pocket; these residues were identified in literature and known to be highly conserved in all species (Kümpornsin *et al.*, 2014). However, the interacting residues Tyr 256 and Leu 317 were found to be conserved in *P. flaciparum* only and it was substituted with Ile 121 and Val 181 in the human GCH1 protein. Tyr side chain contains a reactive hydroxyl group that can interact with non-carbon atoms, whereas the Ile side chain is non-reactive, such substitutions can have functional implementation.

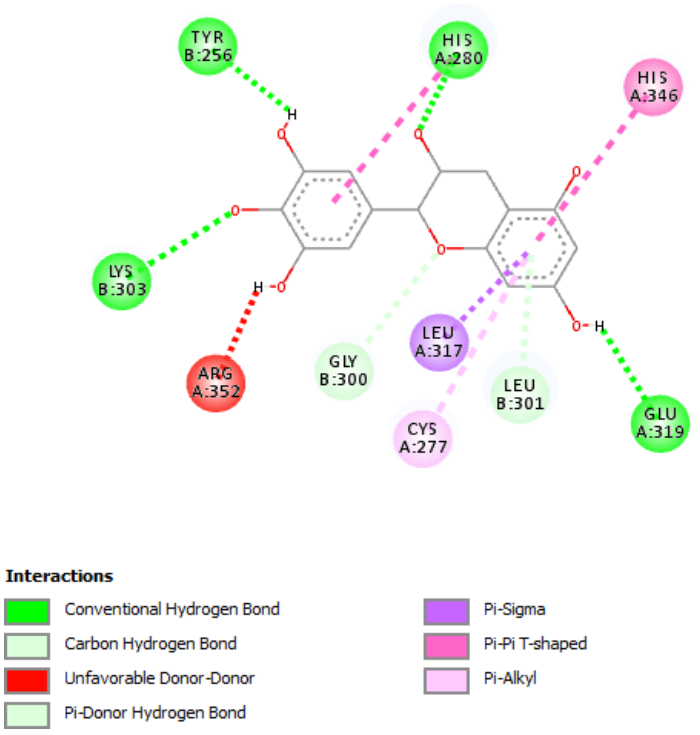


Figure 4.19: SANC00103 protein-ligand interactions in 2D. Each interacting amino acid is labelled and coloured based on the interaction type.

SANCDB00106

SANCDB00106 exhibited low free energy of binding of -9.6 kcal/ mol. In the human GCH1 protein this compound was not bound to the active site. 2D representation of the ligand-protein interactions is shown in Figure 4.20. The observed interactions of the ligands included three hydrogen bonds with His 280 (chain A). Non-covalent interactions were also formed between the ligand aromatic rings and Cys 277, His 279, His 280 Leu 317 (chain A) and Leu 301 (chain B). The identified interacting residues are known to be highly conserved (Kümpornsin *et al.*, 2014). This compound showed the least number of interactions when compared to the other compounds; it also had a low molecular weight.

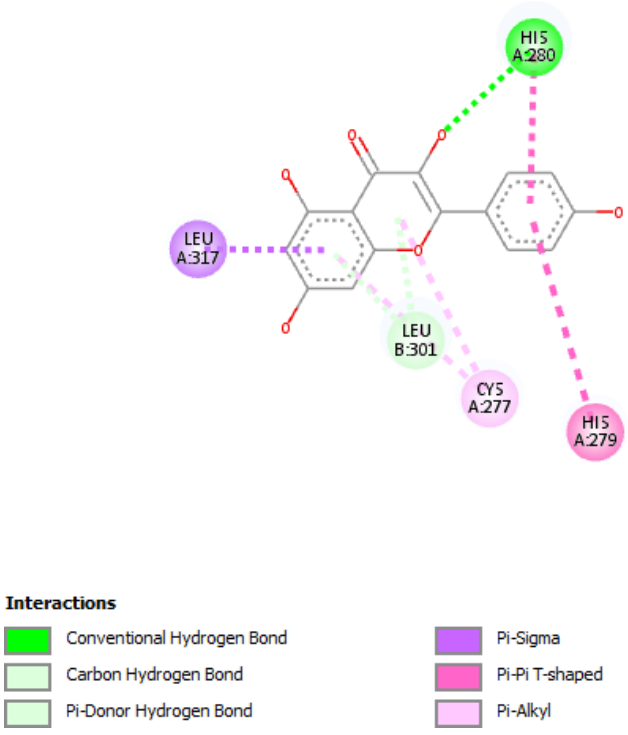


Figure 4.20: SANC00106 protein-ligand interactions in 2D. Each interacting amino acid is labelled and coloured based on the interaction type.

SANCDB00286

SANCDB00286 showed high binding affinity to the GCH1 protein active site pocket. It exhibited low free energy of binding of -9.5 kcal/ mol. 2D representation of the ligand-protein interactions is shown in Figure 4.21. The ligands interactions included hydrogen bonds with Glu 319, His 280 (chain A). His 279, His 280 and His 346 (chain A) formed a π-π T-shaped interaction with the ligand aromatic ring. More non-covalent interactions were formed between Cys 277, Leu 317 (chain A) and LEU 301 (chain B). A hydrogen bond was formed between the carbonyl of the inhibitor and Gly 300 (chain B). One unfavourable interaction was formed with Gln 318 (chain A) (Figure 4.21). Tyr 256 and Leu 317 were found to be conserved in *P.falciaprum* and were substituted by Ile 121 and Val 81 respectively in the human GCH1 protein sequence; this can point towards functional relevance.

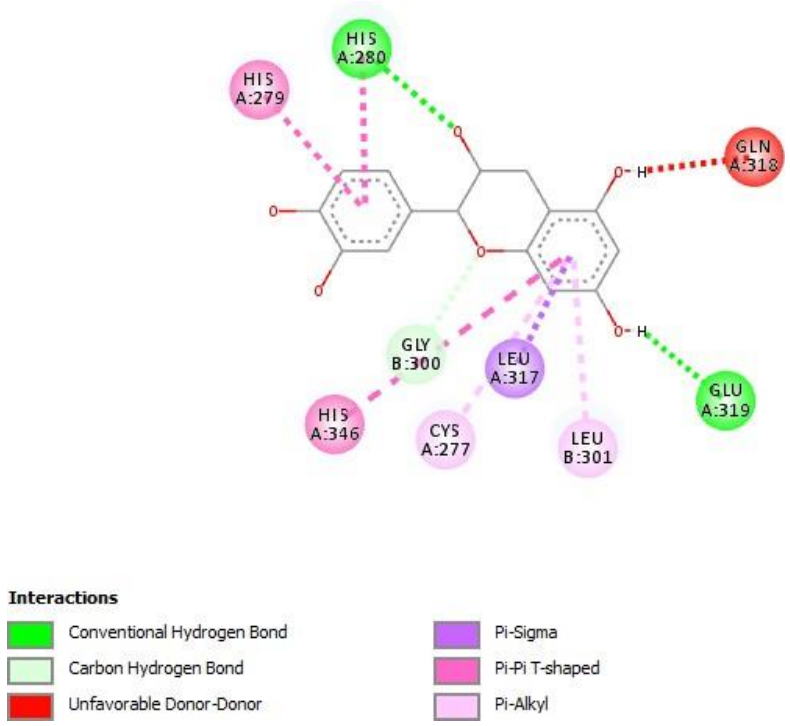


Figure 4.21*:* **SANC00286** protein-ligand interactions and bond types in 2D. Each interacting amino acid is labelled and coloured based on the interaction type.

SANCDB00317

SANCDB00103 exhibited low free energy of binding of -10.1 kcal/ mol. In the human GCH1 protein this compound was not bound to the protein active site pocket. 2D representation of the ligand-protein interactions is shown in Figure 4.22. The ligand's observed interactions were two hydrogen bonds with His 280 (chain A) and Lys 303 (chain B) and metal interaction with the protein zinc atom. Non-covalent interactions were also observed between the ligand's aromatic rings and protein residues: Cys 277, His 279, His 280, Leu 317 (chain A), Gly 300, Leu 301 (chain B). These interactions are known to play a major role in the protein catalytic function (Kümpornsin *et al.*, 2014).

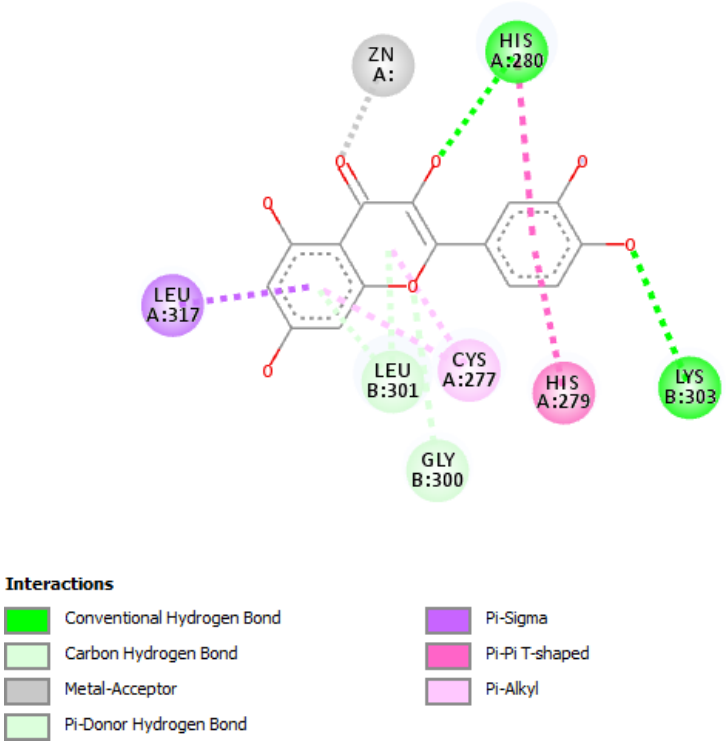


Figure 4.22: SANC00317 protein-ligand interactions and bond types in 2D. Each interacting amino acid is labelled and coloured based on the interaction type.

SANCDB00335

SANCDB00335 exhibited a high free energy of binding of -10 kcal/ mol. This indicates high binding affinities and consequently higher inhibition constants. 2D representation of the ligand-protein interactions are shown in Figure 4.23. The interactions observed were hydrogen bonds with His 346, His 280 (chain A) and Tyr 256 (chain B). Non- covalent interactions were also observed between the ligand's aromatic rings and protein residues: Leu 317, Cys 277, His 279 (chain A), Gly 300, Leu 301 (chain B) and Arg 231 (chain J). The identified interactions were considered to be favourable and similar to what stated in literature (Kümpornsin *et al.*, 2014). The interacting residues Tyr 256 and Leu 317 (Chain A) were conserved in *P. falciparum* GCH1 protein sequence.

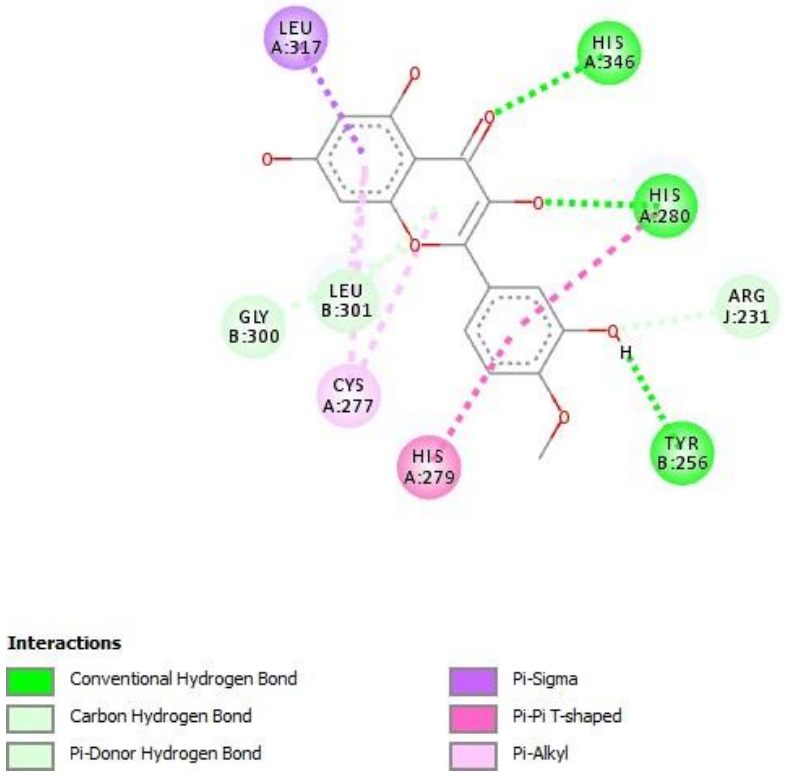


Figure 4.23: **SANC00335** protein-ligand interactions and bond types in 2D. Each interacting amino acid is labelled and coloured based on the interaction type.

The top selected SANCDB compounds exhibited similar and desirable binding. This might explain the high conservation at the active site and can raise the possibility of a broad range inhibitor targeting the Plasmodium GCH1 protein. Figure 4.24 to 4.26 shows 3D views of top selected SANCDB compounds. The majority of the compounds screened against the human homolog exhibited low energy of binding to other sites than the active site. This can be explained by the variation of the residues surrounding the active site (See chapter two).

The active site of *P. falciparum* showed a significant level of conservation in the residues lining the active site pocket, some which interacted directly with the selected ligands such as Tyr 256 and Leu 317. These dissimilarities can underline structural properties of the Plasmodium GCH1 protein. In this case the residues lining the active site pocket of the GCH1 included Leu 276, Leu 317, Leu 225, Leu 239, Leu 282, Lys 278, Ile 349, Asn 350, Ile 224 and Tyr 232 of *P.falciparum*. (See chapter 2). From this we can suggest that the residues variation around the active site has functional implementations.

Docking against the receptors that were composed of three chains and the complete biological unit of the protein resulted in an increased binding affinity of the screened compounds and obvious selectivity towards the Plasmodium homologs; this can be explained by the availability of the complete shape of the active site which allowed the compounds to enter the active site pocket formed by the three adjacent chains, hence taking their best poses.

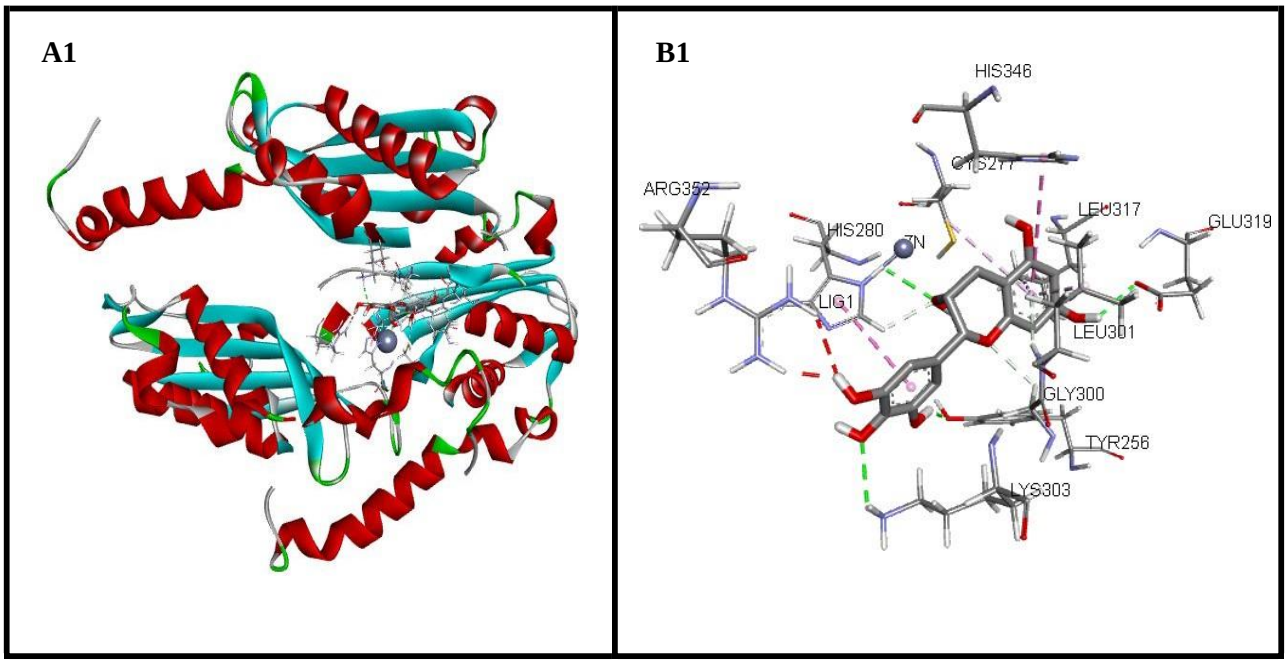


Figure 4.24: **A1** show the protein-ligand interactions of SANCDB00103. The protein is shown in a cartoon representation. The zinc metal in the active site is shown as a grey sphere. **B1**: 3D view of GCH1 residues interacting with SANCDB00103. Amino acid side chains are shown as sticks and the dotted lines correspond to interaction bonds.

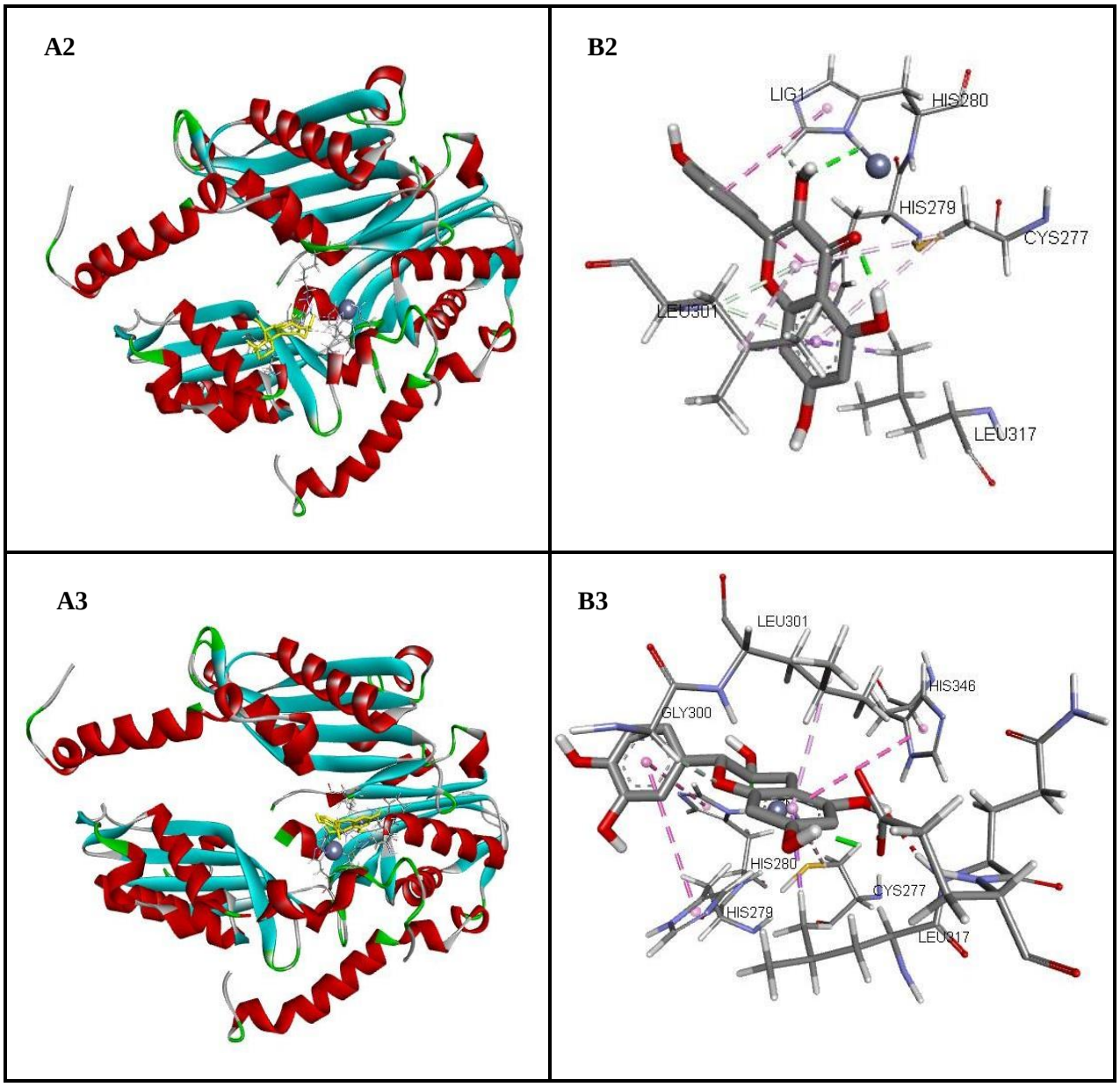


Figure 4.25: **A2** and **A3** show the protein-ligand interactions of SANCDB00106 and SANCDB00286, respectively. The proteins are shown in a cartoon representation. The zinc metal in the active site is shown as a grey sphere. **B2** and **B3**: 3D view of GCH1 residues interacting with SANCDB00106 and SANCDB00286. Amino acid side chains are shown as sticks and the dotted lines correspond to interaction bonds.

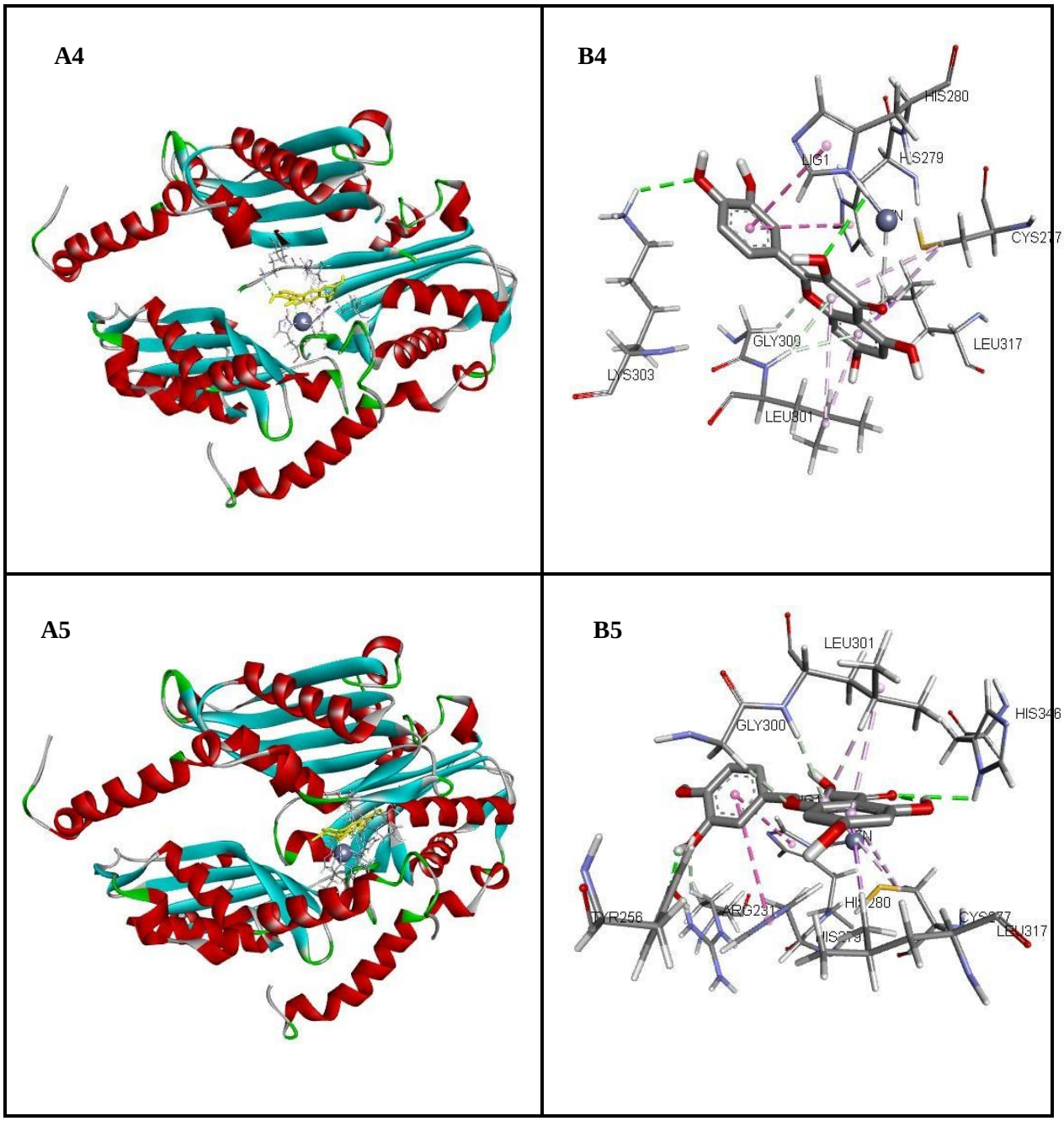


Figure 4.26: **A4** and **A5** show the protein-ligand interactions of SANCDB00317 and SANCDB00335 respectively. The proteins are shown in a cartoon representation. The zinc metal in the active site is shown as a grey sphere. **B4** and **B5**: 3D view of GCH1 residues interacting with SANCDB00317 and SANCDB00335, respectively. Amino acid side chains are shown as sticks and the dotted lines correspond to interaction bonds.

4.3.3 Druggability properties (LIPINSKI rule of five)

The Lipinski rule of five (Lipinski, 2004) was used as a general guideline to test for the suitability of potential inhibitors for oral formulations. Lipinski's rules suggest that a desirable drug candidate should pass at least four of the proposed five rules of drug-likeness. Based on these rules a desirable drug should have less than 5 H-bond donors, less than 10 H-bond acceptors, a molecular weight less than 500 and a calculated Log P value less than 5. The Lipinski rule of five should be used as a general guide and not to cast aside potential compounds that did not pass. Further biochemical analysis should follow to complement/support the outcome of this rule (McKerrow and Lipinski, 2017). The Supercomputing Facility for Bioinformatics and Computational Biology (SCFBio) web-server was used to perform the test (http://www.scfbio-iitd.res.in/) (Jayaram*et al*. 2012). The input pH value was set to 7.

SANCDB00335, SANCDB00368 and SANCDB00106 compounds passed all the above stated rules while SANCDB00317, SANCDB00103 and SANCDB00286 violated only one rule by having the number of hydrogen bond donors marginally out of range (Table 4.2). From this we can conclude that all five compounds may be ideal drug candidates for further lead optimization experiments and *in vitro* investigations. These compounds should be validated further by the use of molecular dynamics simulations, to assess their bindin in a solution environment. Each of them can then be re-docked and their binding properties can be re-assessed before proceeding to molecular dynamics simulations.

Table 4.2: Tabulated results of Lipinski test for ‒drug-likeness‖. SANC00335, SANC00368 and SANC00106 compounds passed all the stated rules while SANC00317, SANC00103 and SANC00286 violated only one rule by having the number of hydrogen bond donors only out of range.

Ligand	Molecular mass<500 Dalton	Lipophilicity (LogP) <5	Hydrogen bond donors <5	Hydrogen bond acceptors <10	Molar refractivity between (40-130)
SANC00317	302.00000	2.010899	5	7	74.050476
SANC00335	316.00000	2.313899	4	7	78.937675
SANC00368	288.00000	2.505900	2	5	76.752075
SANC00106	286.00000	2.305299	4	6	72.385681
SANC00103	306.00000	1.251699	6	7	74.287773
SANC00286	290.00000	1.546099	5	6	72.622971

Chapter summary

In this chapter, a total of 623 compounds from the SANCDB database were screened against the parasite and human GCH1 proteins. Molecular docking showed that some of these compounds can selectively bind with the parasitic proteins. Potential binding sites for inhibitor compounds were identified, making GCH1 a promising drug target. The compounds bound to these sites were ranked and evaluated based on their binding free energy and binding sites. Molecular docking resulted in the identification of desirable hit compounds that had a selective high binding affinity towards the parasitic GCH1 protein. These compounds were SANC00317, SANC00335, SANC00368, SANC00106, SANC00103 and SANC00286. Inhibitors that had many reactive atoms and groups such as carbonyls and aromatic rings had more interactions consequently an increased binding affinity. Inhibitors were able to interact with *P. falciparum* GCH1 active site key residues, such as Cys 277, His 279, His 280, Arg 231 and Arg 352, Lys 303, Gly 300 and Leu 301. The interactions observed were hydrogen bonds, polar, hydrophobic, van der Waal and electrostatic interactions. The interacting residues were in confirmation with ones stated in literature and established to be important.

Furthermore, it was found that the residues lining the Plasmodial GCH1 active site were different to those in the human GCH1. This can propose different mechanism of substrate recognition and conformational change which has distinguished the Plasmodium protein from its human homolog. The next step will be the validation of the selected compounds by the use of molecular dynamics simulations, to assess their binding.

CHAPTER FIVE

DEVELOPMENT AND VALIDATION OF ZN^{+} FORCE FIELD PARAMETERS OF THE GCH1 ENZYME

Computational approaches play a major role in modern drug discovery and its development; it helps in understanding molecules interaction and dynamics. In order to describe molecule interaction, several force fields and force field parameters have been developed and validated for proteins, lipids and nucleotides. However, these force fields lack the accuracy to describe metals in metallo-enzymes. This lack of accuracy may introduce incorrect predictions of structural dynamics and it can adversely affect drug design efforts. One of the commonly used force fields is CHARMM force field (Brooks *et al.,* 1983). CHARMM also facilitates the development of a molecular force field and generates parameters that can be used to study the protein-ligand complexes. The work presented in this chapter aimed to study the GCH1 metal site in order to provide insight into the development and validation of its force field parameters. Quantum mechanics (QM) calculations were used to derive the GCH1 metal force field parameters. The validation of the generated parameters aimed to find a good agreement between the experimental and computed values; this was also examined by Molecular Dynamics (MD) simulations.

5.1 Introduction

Metal ions are known to be essential in many biological processes. They are considered to be important in reduction/oxidation reactions; this is due to their electron transfer properties (Stadtman, 1990). Metal ions also provide structural support for the protein active sites, because of their high charges in a natural environment (Lu *et al.*, 2015). GCH1 is a metallo-enzyme containing a zinc ion in its active site. The zinc serves as a catalytic centre and provides the GCH1 protein with structural stability (Auerbach *et al.*, 2000). The enzyme forms a homo-decameric barrel-like structure with ten equivalent zinc-containing active sites each of them formed between three adjacent subunits (Thöny, Auerbach and Blau, 2000).

Enzymes' zinc ions are known to have different coordination geometries including, trigonal bipyramidal, square pyramidal and octahedral (Patel, Kumar and Durani, 2007). In the GCH1 protein, the zinc ion is known to be tetrahedrally coordinated by one water molecule from the surrounding solvent, a histidine residue and two cysteine residues (Yoko Tanaka *et al.*, 2005).

5.1.1 Potential Energy Surface (PES)

The energy of a system or a molecule is important to understand its chemical reactivity. Chemical reactions are often described on a coordinate system that is composed of a potential energy axis and a reaction coordinate axis; this creates a Potential Energy Surface (PES). A transition state occurs when the energy of the system increases, while a product state occurs when the system relaxes to a lower energy. The difference in free energy between the transition state and product state demonstrate the chemical equilibrium and thermodynamic stability. This stability is characterized by energy minimum (Lewars, 2011). Different methods have been developed for the calculation of potential, such as semi empirical methods, quantum mechanics (QM) and molecular mechanics (MM). These methods use different approaches to describe the energy of a system with different levels of detail (Collins, 2002).

5.1.2 Protein dynamics study

Molecular Mechanics

In molecular mechanics atoms are treated as small entities and molecular bonds between them are described as springs by Newtonian mechanics (Vanommeslaeghe *et al.*, 2014). Molecular mechanics is computationally more efficient and it can represent the molecular structure very well in its equilibrium state. However, the MM PES is known to not capture bond breaking and forming very well. In addition, most MM force fields do not consider the electronic degree of freedom, which creates more difficulty in describing metal ions of bio-molecules. An MM force field expresses the system energy as a sum of energy terms, which accounts for bonded and non-bonded interactions (Bowen and Allinger, 2007).

Molecular Dynamics (MD)

Molecular dynamics was first introduced in the 1950s by Alder and Wainwright (Alder and Wainwright, 1957, 1959). This was followed by Rahman and Stillinger's first molecular dynamics simulations of liquid water (Stillinger and Rahman, 1974); then the first protein simulations of the bovine pancreatic trypsin inhibitor (BPTI) in 1977 (McCammon, Gelin and Karplus, 1977).

Experimental methods such as X-ray diffraction (XRD) and nuclear magnetic resonance (NMR) are used for the characterization of proteins crystal structures. NMR can be used to determine solution structures; however, XRD cannot be used to study the protein structure in solution. Computer simulations are used to solvate such structures. MD is one of the major computer simulation methods used in computational chemistry. It determines trajectories of atoms from an initial state and describes individual particle motions as a function of time. MD simulations have a wide range of applications including the characterization of protein dynamics, prediction of protein mechanism of function, studying of protein folding and unfolding, force field parameter validation and finally drug design and development (Alder & Wainwright 1959; Allen 2004).

MD simulations start with the protein 3D structure, then uses an empirical force field to approximate the potential energy surface of the protein. Once the force field has been specified, forces are calculated and acceleration determined according to the Newton's equation of motions. For metal ions, classical molecular dynamics is unable to describe quantum mechanical effects such as charge transfer and polarization. Thus, other computational methods are used to accurately simulate the behaviour of the metal ion (Allen 2004; Durrant & McCammon 2011).

$$\boldsymbol{F = ma}$$

In Newton's equation of motion, given above F is the force exerted on an atom, m is mass and a is acceleration. Integration of the equation yields trajectories that describe the positions, velocities and accelerations of the particles over time. The forces are then calculated using a potential energy function. Collectively this equation and the associated parameters are referred to as a force field (Boas & Harbury 2007).

Quantum Mechanics (QM)

QM methods are based on electronic structure theory. This method is based on solving the Schrödinger equation in order to calculate the energy of the system.

$$\hat{H}\Psi = E\Psi$$

$\hat{H}$ is the Hamiltonian operator that describes the potential and kinetic energy. Once the energy and vibrational frequencies are known, the rigid rotor-harmonic oscillator approximation can be used to calculate thermodynamic properties from the ideal gas partition function. In general, the QM method is considered to be more accurate for calculating structural and thermodynamic properties compared to classical mechanics/force fields, but it is much more computationally demanding. Because of this limitation QM calculations are only feasible for small systems. In addition, static QM does not capture the dynamics of proteins, but methods such as *ab initio* molecular dynamics (AIMD) are able to perform dynamics on such systems, but can only be used with tens of atoms over periods of under one ns (Dominik Marx and Jurg Huutter, 2009).

Hartree-Fock (HF) and Density Functional theories (DFT) are examples of two methods that can be used to describe the system energy. DFT derives the energy from the system electron density (Runge and Gross, 1984). This method is recommended for studying metal containing molecules due to its reasonable accuracy. The functionals used in DFT to calculate the electronic energy include electron-electron Coulomb energy, electron-nucleus Coulomb energy, electron kinetic energy, and combined exchange-correlation energy (Cohen, Mori-Sánchez and Yang, 2012). One of the most widely used exchange-correlation functional is the B3LYP functional. It is recognized for being accurate in describing molecular geometries and energies. The B3LYP functional combines a Generalized Gradient Approximation (GGA) hybrid exchange functional from the work of Becke (Becke, 1988) with the correlation functional of Lee, Yang and Parr (LYP) (Lee, Yang and Parr, 1988).

In HF theory, the system energy is obtained using an antisymmetric wavefunction by using a Hamiltonian operator which consists of electron-electron and electron-nucleus Coulomb interaction as well as electron kinetic energy. This method does not consider the correlation energy between electrons, that is defined as the difference between the real energy and the HF energy. Correlation energy is considered to be an essential aspect in describing chemical systems accurately.

QM/MM (Hybrid method)

The Quantum Mechanics and Molecular Mechanics (QM/MM) hybrid method treats a subset of atoms with quantum mechanics and the rest of the atoms using molecular (classical) mechanics (Harrison, 1999). The two methods are linked together with an additional term in the Hamiltonain that describes the interaction between the QM and MM regions. Atoms described by each are linked together using link atoms that are treated both ways and are attached to atoms from both sets. QM/MM hybrid methods are established as efficient however, it is not yet feasible for generating trajectories of protein simulations (tens of thousands of atoms) for a sufficient length (several ns) to determine the thermodynamic properties (Lin and Truhlar, 2007).

5.1.3 Potential energy function

Classical MD simulations are based on using a molecular mechanics force field. The force field is a combination of mathematical functions and their associated parameters used to describe the energy of the protein as a function of its atomic coordinates (Boas & Harbury 2007). Force field functions and parameter sets are usually derived by either fitting to experimental data or high-level quantum mechanical calculations. The potential energy function involves the energetic contributions of bond stretching, angle bending, dihedral rotation and non-bonded energetic contributions due to van der Waals interactions and electrostatic interactions.

$$E_{total} = \sum_{bonds} k_b\left(r - r_{0,b}\right)^2 + \sum_{angles} k_a\left(\theta - \theta_{0,a}\right)^2 + \sum_{dihedrals} k_{d,n}[1 + \cos\left(n\chi - \delta_{d,n}\right)] + \sum_{non-bonded} \epsilon\left[\left(\frac{R_{min_{ij}}}{r_{ij}}\right)^{12} - \left(\frac{R_{min_{ij}}}{r_{ij}}\right)^{6}\right]$$

Equation 1

Where k_b is the bond force constant, r is the bond length; r_0 is the equilibrium bond distance. k_a is the valence angle force constant, θ is the valence angle, and θ_0 is the equilibrium angle. For dihedrals, χ is the dihedral or torsion angle, $k_{d,\,n}$ are the dihedral force constants, n is the multiplicity and δ is the phase angle. The non-bonded interaction calculations are considered computationally expensive. Several attempts have been made to develop algorithms to make this part as efficient as possible.

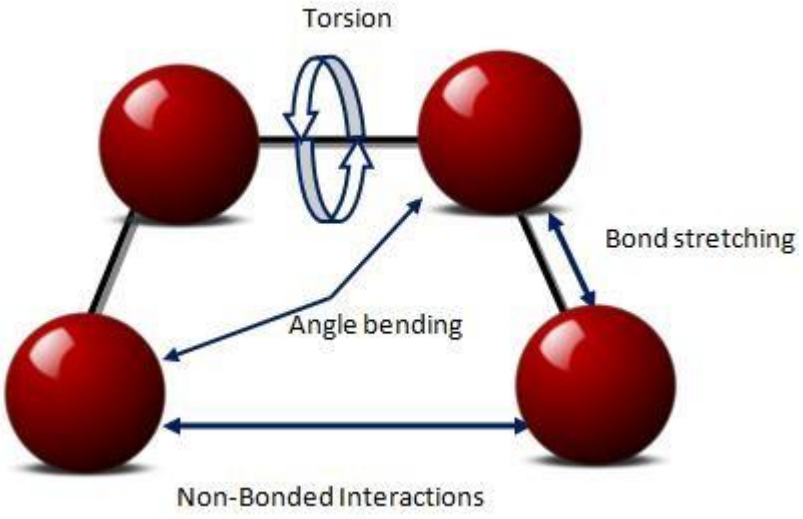


Figure 5.1: According to the molecular mechanics model, atoms are described as spheres and bonds as springs. This can be used to describe the ability of bonds to stretch, angles to bend, and dihedrals to twist. The total energy is obtained by the equation, Energy = Stretching Energy + Bending Energy +Torsion Energy +Non-Bonded Interaction Energy. Adapted from (Fundamentals of Molecular Dynamics for Nano-technology Applications Mario Blanco Materials and Process Simulation Center California Institute of Technology, no date)

Currently, there are a number of force field that have been made available for different types of molecules such as CHARMM (MacKerell *et al.*, 1998), AMBER (Cornell *et al.*, 1995), GROMOS (Oostenbrink *et al.*, 2004) and OPLS (Jorgensen, Maxwell and Tirado-Rives, 1996). These force fields have been subjected to improvement over decades in terms of speed, scalability, compatibility with different computer software and hardware, applicability to different biological systems and compatibility with different force fields. These improvements made them popular for different applications. The main difference between these force fields is the parameters used in describing the non-bonded and bonded terms and the approaches taken to derive the parameters.

5.2 Methods

5.2.1 Subset selection

A subset of the active site residues was created from the initial crystallographic structures by selecting only the zinc metal and residues bonded to it within 5 Å. This process involved a set of assumptions, as the coordination shell of the zinc ion in some of the structures could not be unequivocally assigned. Literature points towards key residues in the active site that are connected to the zinc atom, these residues were His 280, Cys 277 and Cys 348 (Gräwert, Fischer and Bacher, 2013; Kümpornsin *et al.*, 2014). Other studies refer to the existence of one water molecule coordinating the zinc (Yoko Tanaka *et al.*, 2005). However, this was considered carefully since not all crystallized waters are catalytic or conserved. After selecting an initial representative subset, a scan was carried out in order to explore the PES of the subset coordinates.

5.2.2 Geometry optimization

The geometry of the selected subset system was optimised using quantum mechanics. DFT calculations were performed via Gaussian 09 (Frisch, M. J et al., 2009) with the Becke three-parameter hybrid exchange functional and the Lee-Yang- Parr (B3LYP) correlation functional (Becke 1993, Lee, Yang & Parr 1988, Vosko, Wilk & Nusair 1980). The LanL2DZ pseudopotential and associated basis functions were used to describe the zinc atom and the 6-31G (d) basis set was used for the organic atoms (Ditchfield, Hehre & Pople 1971, Hehre, Ditchfield & Pople 1972, Hariharan, Pople 1973, Hariharan, Pople 1974, Gordon 1980, Rassolov et al. 2001). This method is considered to be reliable in terms of accuracy and computational cost. To accommodate the computational cost, the optimization step was performed on the CHPC cluster. The resultant optimised structure was superimposed onto the initial crystallographic structure to ensure its stability and to check if any bonds were broken.

5.2.3 RESP charge

Partial atomic charges of the selected active site sub system were calculated using the Restrained ElectroStatic Potential (RESP) charge-fitting method (Cieplak et al. 1995), at the B3LYP level of theory using 6-31G (d) basis set for the organic atoms and LanL2DZ pseudopotential and associated basis functions for the zinc atom (Dunning, Hay 1977). Charges and their corresponding atom types were illustrated on the subset structure via PyMOL.

5.2.4 Force field parameters

From the optimised geometry equilibrium bond lengths, angles and dihedrals were extracted. The force constants were derived using QM method at the B3LYP/6-31G(d)/LanL2DZ level of theory and redundant internal coordinates scans of selected bonds, angles and dihedrals values. Each bond was stretched by 0.05 Å each way in ten steps, each angle was bent by 1 Å each way and one dihedral angle was rotated by 1 Å each way in ten steps. The energy profiles from the PES scans were manually fitted to the bonded terms in Equation 1 using the least squares method. Based on the results obtained modifications to the CHARMM 36 force fields were implemented. This allowed the use of the CHARMM force field to model the behaviour of the solution phase of the subset system.

5.2.5 Validation of force field parameters with MD simulations

After obtaining the required parameters, a validation step was done to test for the stability of the subset system. MD simulations were carried out using CHARMM 36 force field. Due to the fact that non-standard biological molecules (zinc ion in this case) do not have built-in files to match the force field in CHARMM, the generated force field parameters had to be implemented manually thus, adding an extra step of preparation before MD simulations. The system was minimized in vacuum using 100 steps of steepest descent, then solvated using TIP3P water model and neutralized within a cubic box a of suitable dimensions. The system was then minimized further with another 200 steps of steepest descent, followed by heating for 100000 steps. Finally, the dynamics equilibration and production runs were carried out. After MD simulations, various properties like the root mean square deviation (RMSD) and radius of gyration of the protein can be extracted from these simulations in order to characterize the atomic-level behaviour of the subset system.

5.3 Results and Discussion

5.3.1 Geometry optimization

Initial geometry optimisation was performed on the created subset shown in Figure 5.2. This subset was created to reduce the computational cost. Geometry optimization was carried out to find the atomic arrangement that makes the molecule most stable. It was observed that one of the hydrogen atoms of the water molecule was slightly tilted or twisted; which made the two hydrogen atoms non-identical (Figure 5.2). After optimization the geometry was cleaned and this was resolved (Figure 5.3). The PES scans are used to describe the potential energy surface around a local minimum so that the force can be fitted (Appendix 5, Figure A4.2).

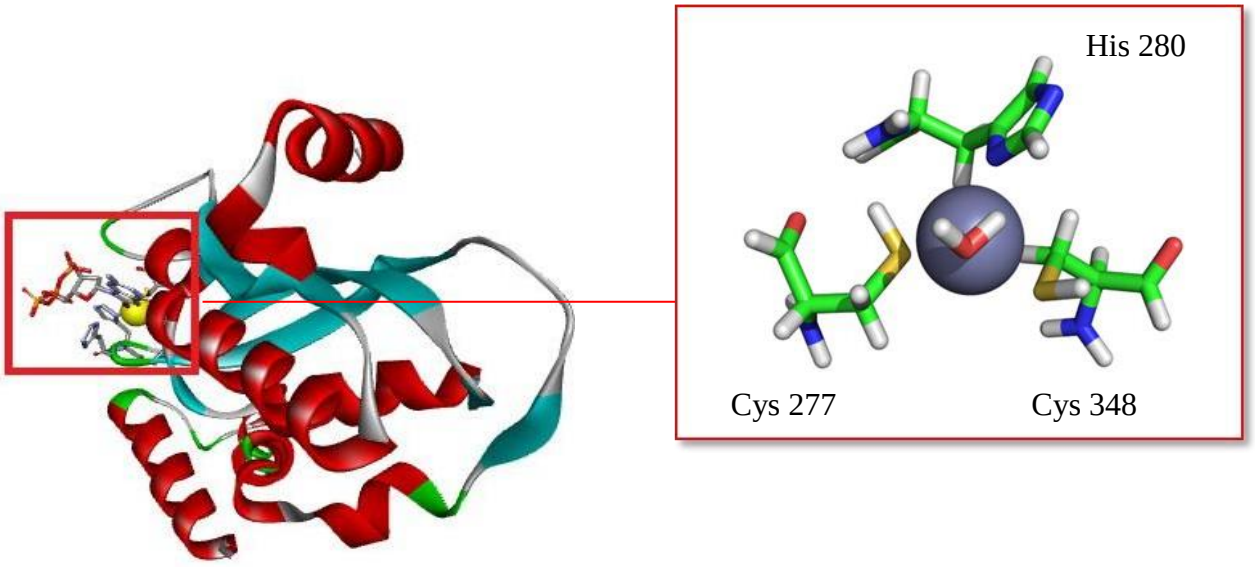


Figure 5.2: Active site subset selected from the X-ray structure. The zinc metal in the centre is coordinated by His 280, Cys 277 and Cys 348 and one water molecule.

The optimised structure is shown in Figure 5.3 where the zinc atom exhibited a tetrahedral geometry. The optimised subset was superimposed onto the initial X-ray structure (Figure5.3); this indicates that the optimised subset structure was stable and it was not distorted. To support this finding, the optimised values of the bonds distance and angles were captured and are shown in Table 5.1, with comparison to the initial X-ray structure. It was observed that these values were fairly similar, demonstrating that the optimised structure was not distorted.

Table 5.1: Optimised values of bond distance (Å), angle **(°)** compared to the initial X-ray structure

Parameter	Crystal structure	Optimised structure
Bonds	**Distance(Å)**	**Distance(Å)**
Zn-SG	2.318	2.203
Zn-ND1	2.204	2.102
Zn-SG	2.301	2.205
Angles	**Angle (°)**	**Angle (°)**
SG-ZN-ND1	113.157	113.141
SG_ZN_SG	119.293	119.201
SG_ZN_ND1	107.340	107.182

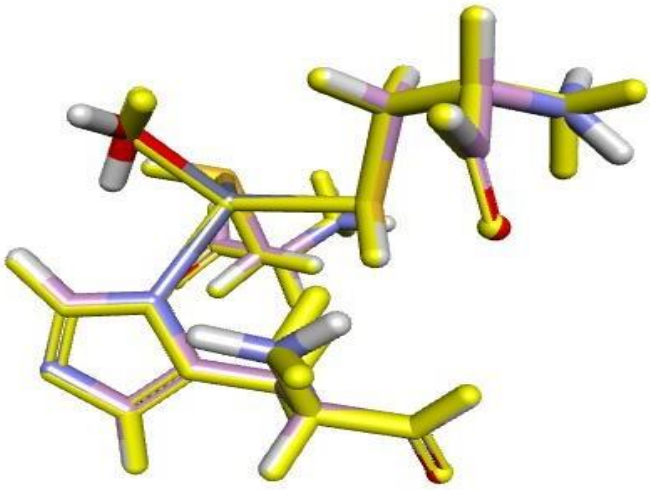


Figure 5.3: The optimised structure shown in yellow superimposed onto the initial subset from the crystal structure. No bonds were broken and the geometry of the zinc was maintained.

5.3.2 RESP charges

Partial atomic charges were obtained by fitting to the electrostatic potential of the molecules obtained from QM calculations. The calculated charge of zinc was +1.125, the rest of the residues charges were mapped onto the subset system (Figure 5.4). It was observed that the partial charges were subjected to variation depending on the type of calculation implemented. Further investigation of this can results in improving the calculation accuracy.

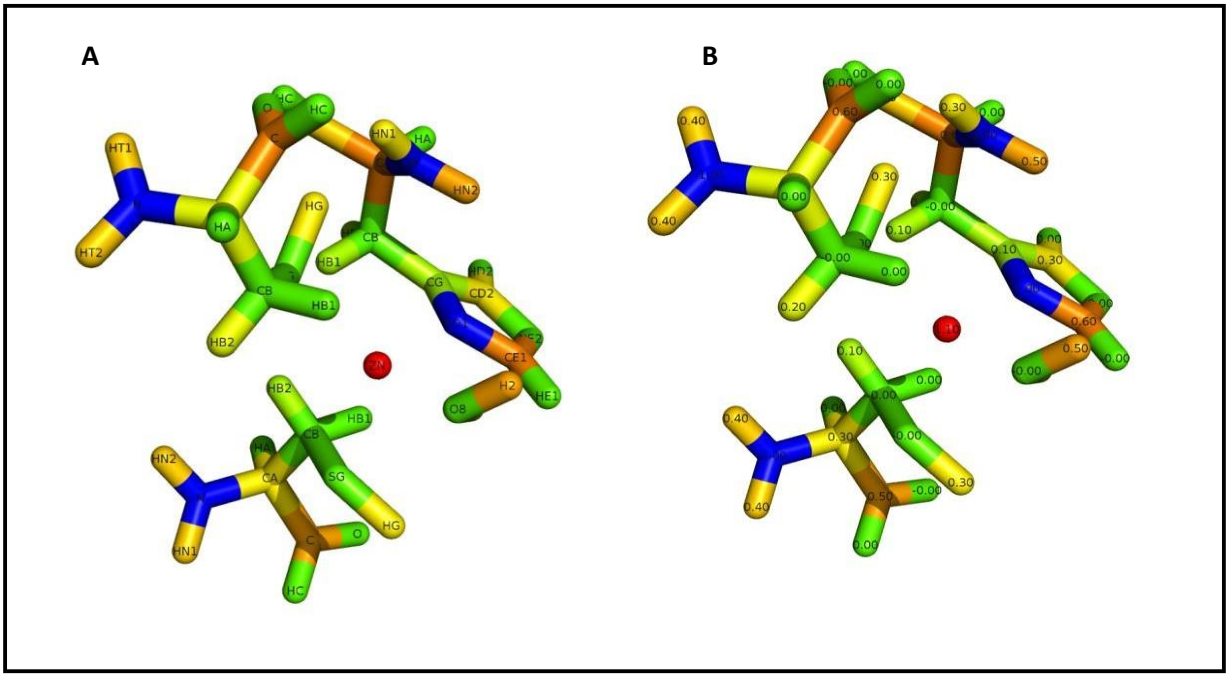


Figure 5.4: Representation of the GCH1 active site subset. (**A**) Atom types and (**B)** RESP charges

5.3.3 Force field parameters

Force field parameter development is considered to be computationally demanding and a labour-intensive process. Therefore, a small optimised subset was used to derive the force field parameters. The approaches used to derive these parameters dictate the applicability and quality of the force field (Hu and Ryde, 2011). QM calculations were used at the DFT/B3LYP level of theory; this resulted in the generation of energy profiles for bond stretching, angle bending and dihedral rotation. The energy profiles exhibited a harmonic profile. The force field was fitted to the QM curve. (Table 5.2) summarizes the fitted force field parameters.

The calculated and fitted PES' for bond stretching, angle bending and dihedral rotations around the zinc are presented in Figure 5.5, 5.6, and 5.7. The solid line represents the fitted force field values and the dots correspond to the QM values. A good fit was obtained as the data points of QM values were reasonably following the trend of the theoretical (MM) data.

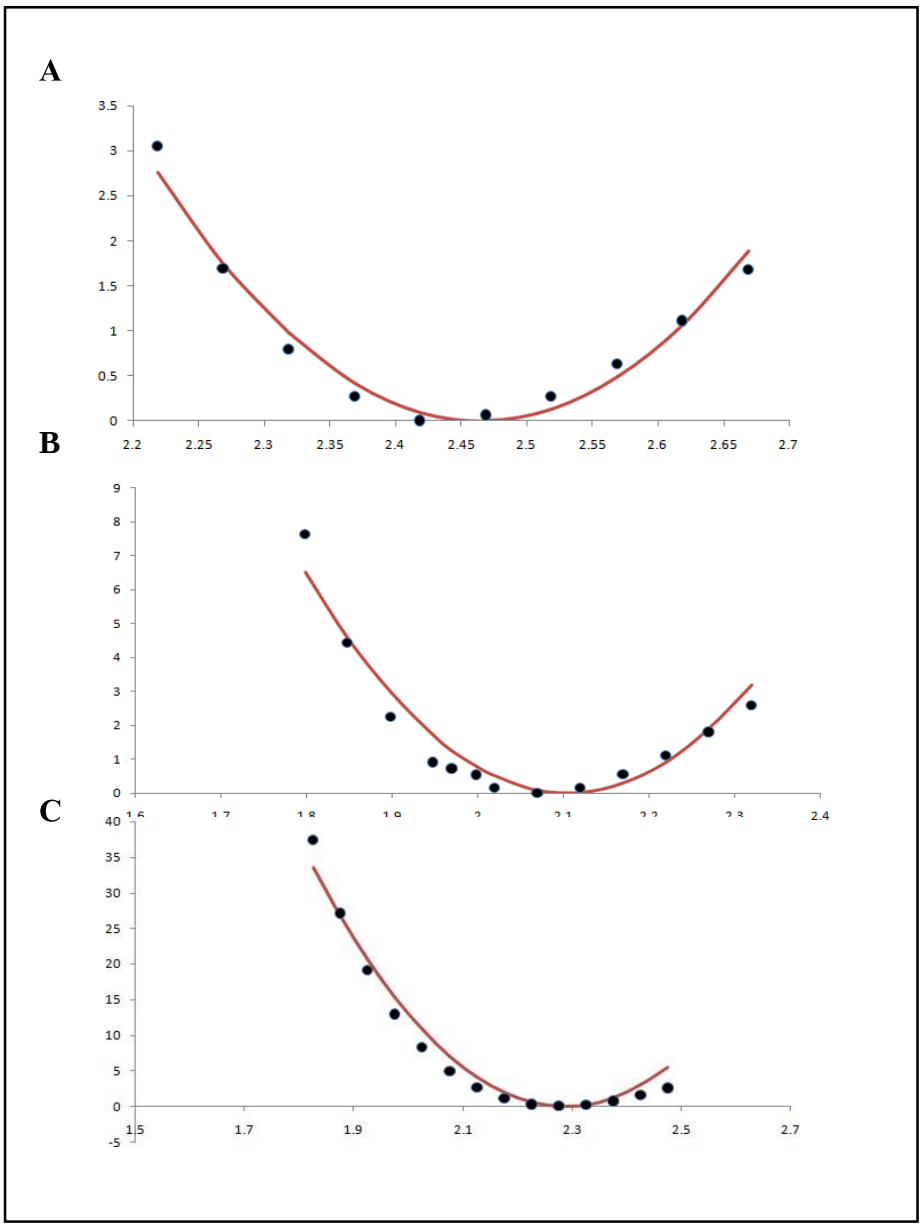


Figure 5.5: Energy profiles of the GCH1 active site subset. The fitting curves are shown in red lines. The energy values for bond stretching of **A**: Zn-SG, **B**: Zn-ND1and **C:** Zn-SG are shown as black dots. The energy profiles exhibited a harmonic potential for bond-stretching. This shows good reproduction of the corresponding calculated PES data with some slight deviation in some values.

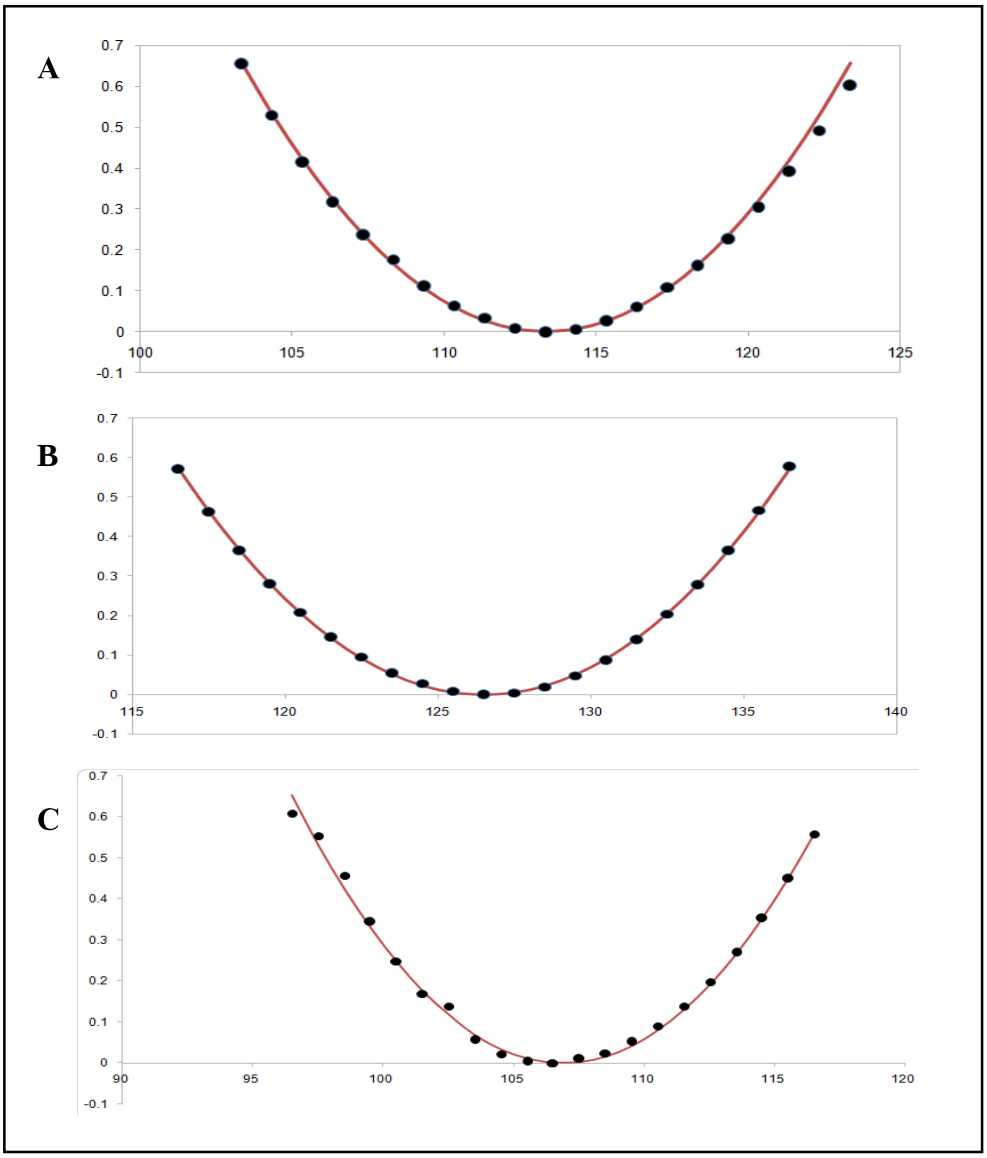


Figure 5.6: Energy profile of the GCH1 active site subset. The fitting curves are shown in red lines. The energy values of angles bending (degrees) of **A**: SG_ZN_ND1, **B**: SG_ZN_SG and **C**: SG_ZN_ND1 are shown as black dots. The energy profiles exhibited a harmonic potential for angle bending. PES scans showed well reproduction of the corresponding calculated data with some slight deviation in some values in **C**.

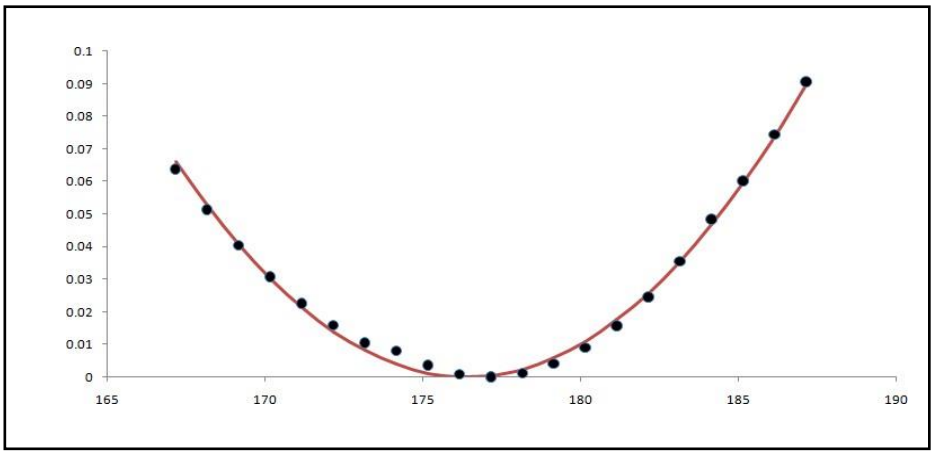


Figure 5.7: Energy profile of GCH1 active site subset. The fitting curves are shown in red lines. The energy values for torsions rotation (degrees) of SG_ZN_ND1_CE1 are shown as black dots. The energy profiles exhibited a harmonic potential for torsion rotation. This shows well reproduction of the corresponding calculated PES data.

Table 5.2: Fitted force field parameters of GCH1 active site by DFT/B3LYP calculations

Bonds	**K r (kcal mol $^{-1}$ Å $^{-2}$)**	**r $_{eq}$ (A)**	
Zn-SG	45.447	2.268	
Zn-ND1	69.308	2.105	
Zn-SG	157.192	2.289	
Angles	**K $_{\theta}$ (kcal mol $^{-1}$ rad $^{-2}$)**	**θ $_{eq}$ (degrees)**	
SG-ZN-ND1	19.943	106.899	
SG_ZN_SG	18.779	126.497	
SG_ZN_ND1	21.561	113.354	
Dihedral	**V $_{n}$ (kcal mol $^{-1}$)**	**n**	**γ**
SG_ZN_ND1_CE1	8.629	1.329	3.083
SG-ZN- ND1 CG	9.129	1.057	0.111

5.3.4 Molecular Dynamic simulations

Molecular dynamics has become an important tool to study the dynamic motion of biomolecules at an atomic level, which is difficult to characterize experimentally. The dynamic of the bio- molecules is essential to recognize their structure function relationship (Nair and Miners, 2014). MD simulations were performed using the parameters derived above for the zinc atom and the associated bonded terms. The protein was prepared without including the ligand.

To save on the computational time and cost the simulations were performed on a single chain which can still provide information about the validation of the generated force field parameters. However, in future we intend to perform MD simulations on three chains of the protein to represent the full active site shape. A total of 20 ns MD simulations were carried out (Figure 5.9). The results of MD simulations were trajectories containing the atomic coordinates, velocities, forces and potential energy. This information can be used to calculate the root mean square deviation (RMSD) and the radius of gyration (Rg). RMSD measures the deviation of the protein structure with respect to a particular conformation whereas Rg measures the protein compactness. Steady Rg values indicate that the protein is stably folded whereas unstable values point to unfolding of the protein. Trajectory analysis can allow the study of the protein time dependent conformational properties relevant to its function (Pirolli *et al.*, 2014).

A change in coordination distance was observed (Figure 5.8). The result showed minimal fluctuation in bond distances (from 2.1 to 2.8 Å) but the mean distance of the three coordinating residues was maintained throughout the simulations. This finding suggests that the evaluated force field parameters were sufficient in adequately describing the coordination of the GCH1 metal in the protein active site.

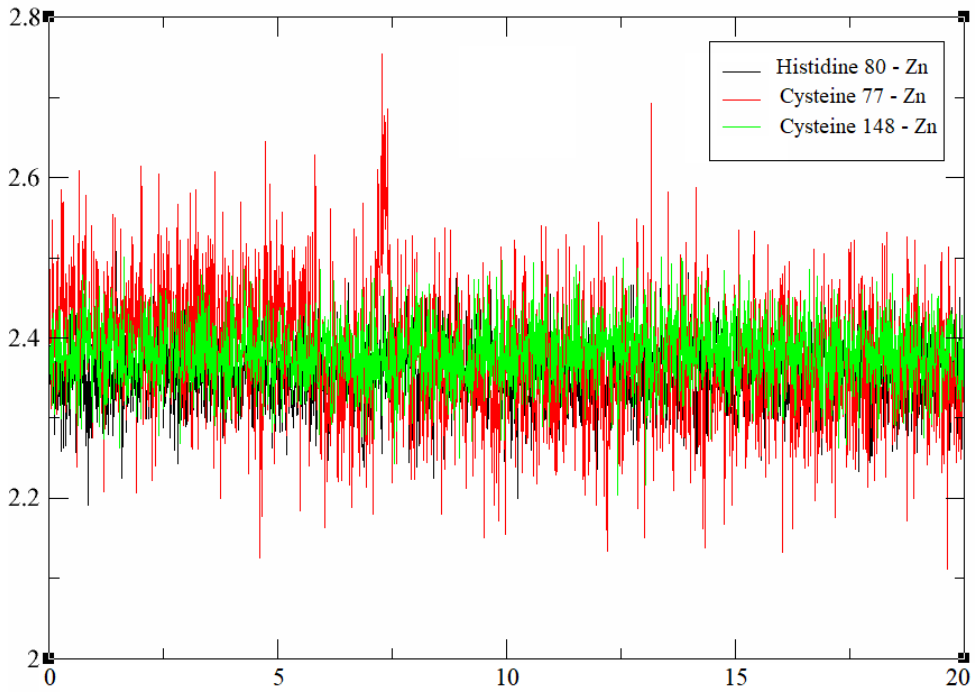


Figure 5.8: Coordination bond distance fluctuation during MD simulation. The black line represents the bond distance fluctuation of His 80; the red line represents the bond distance fluctuation of Cys 77 and the green line represents the bond distance fluctuation of Cys 148. The mean distance of the zinc atom to the three coordinating residues was maintained.

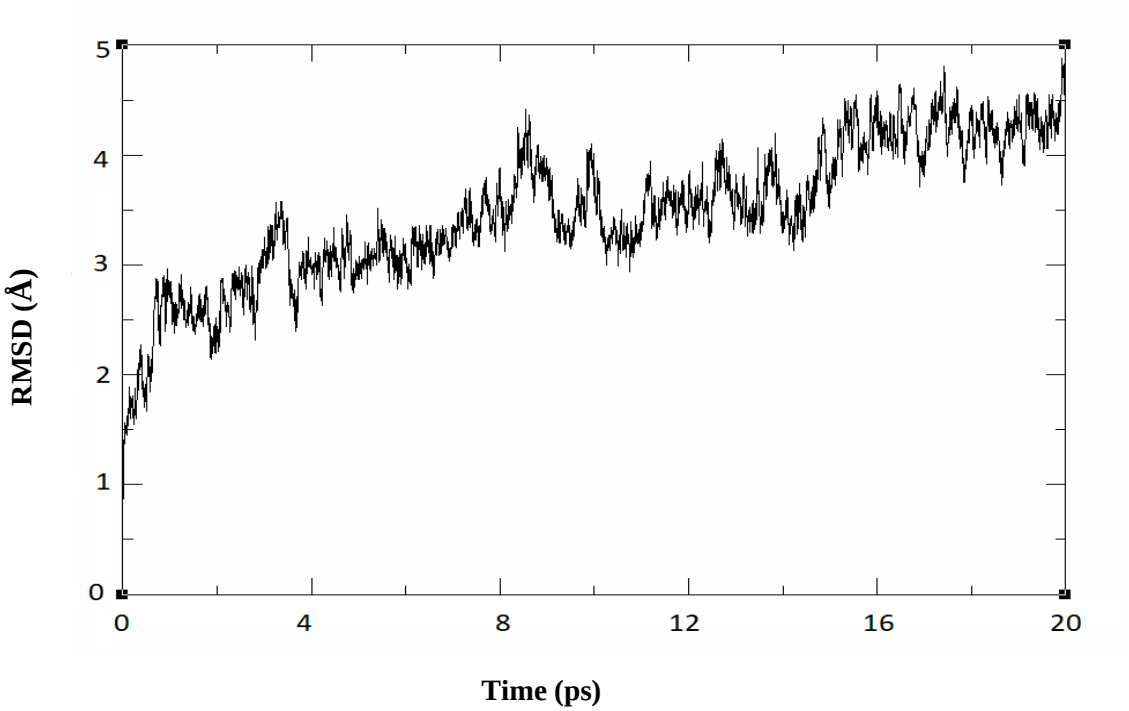


Figure 5.9: Coordination of the GCH1 zinc atom during the MD simulations over 20 ns. The line in (Black) moves from starting structure around 1.5 Å to an average ensemble around 3.5 Å from the original. Figure is generated via xmgrace software (*Xmgr: Introduction*, no date).

The stability of GCH1 protein was assessed using the RMSDs of the backbone atoms relative to the initial structure. The MD trajectories of 20 ns were analysed to establish the protein stability and to deduce the validity of the integrated force field parameters. The line in Figure 5.9 moved from a starting value around 1.5 Å to around 3.5 Å. The MD simulation had failed to reach convergence during the first 15 ns. This was attributed to sustained changes in GCH1 structure. However, in the last 5 ns of the simulation, a fluctuation around 1 Å was maintained until the end of the simulations. This suggests that the structure was starting to stabilize. However, from this result the MD simulation had not reached equilibrium and as such, an extension of the MD simulations is required to conclude certainly if the protein has reached to an equilibrium conformation, which can be observed with a steady line with less fluctuation.

Chapter summary

The work presented in this chapter aimed to produce information to enable future molecular dynamics simulations of GCH1 protein. GCH1 is known to have a metal ion in its active site that plays a major rule in the enzyme catalysis. Force field parameters of the GCH1 zinc ion were not available thus it was important to generate these parameters and validate it through MD simulations. The generated force field parameters of the metal site will allow future study of the protein properties in solution, providing accurate simulations. Standard harmonic QM force field parameters were derived. The parameters used within this force field were developed by fitting the force field to the QM values. The derived force constants were captured and tabulated (Table 5.2). Validation of the generated force field parameters was done via MD simulations. A total of 20 ns of MD simulations were performed. The time evolutions of the backbone RMSD showed that the structure started at around 1.5 Å and progressed to an average around 3.5 then a maintained 1 Å fluctuation in the last 5 ns. This suggests that the protein structure is starting to reach an equilibrated state, however due to the short nature of the simulation longer simulation runs will be performed to verify this. Full MD trajectory analyses will be performed in future work. The generated force field in this work can be valuable in future MD simulations for accurate protein ligand interactions simulations.

Concluding remarks and future work

GCH1 protein sequences from Plasmodium, prokaryote, bacteria and mammals were successfully retrieved from the PlasmoDB and Uniprot databases. Comparative analysis was carried out via multiple sequence alignment. Results obtained from sequence and phylogenetic analysis showed ancestral relationship of Plasmodium sequences, it also showed that mammals' homologs were the least similar. Multiple sequence alignment revealed some residues variation between all homologs in the active site region and surrounding area. However, the key catalytic residues were all conserved in all homologs. The identified sequence features allowed distinguishing the Plasmodium homologs from human protein. Phylogeny reconstruction revealed the evolutionary relationship between the plasmodial GCH1 proteins as the sequences were clustered together showing a distinct clade in the phylogenetic tree and possessing motifs that were lacking in mammalian sequences. This finding can point towards the possibility of selective inhibitation of the Plasmodial GCH1 protein.

Quality structural models of Plasmodium homologs were generated then validated. Homolog models showed a high level of conservation. Yet, some structural variations were observed in the active site surrounding region and terminal regions. The structural variation is attributed to residue variations observed previously at the sequence analysis level. The resulting models were shown to be of good quality both locally and globally.

Molecular docking resulted in the identification of compounds with selective binding towards the Plasmodium species. Even though some variations were observed at the sequence analysis level, potential inhibitors displayed uniform binding across all Plasmodium homologs; this implies the possibility of a broad-spectrum inhibitor in future. The human homolog, which has showed sequence and structural variation, seemed to have poor binding affinities for the docked inhibitors; particularly within the protein active site pocket.

A number of SANCDB compounds showed desirable inhibition properties such as SANC00317, SANC00335, SANC00368, SANC00106, SANC00103 and SANC00286. Biochemical assays for the screened compounds as well as the modification of these compounds can add more evidence of their inhibitory activity. In future, analysis of compounds from the ZINC database will follow, to identify points for improvement and possibly create a library for modified compounds for further screening.

In this study, only rigid docking was performed, this was attributed to the computational time cost; in future, flexible docking for these compounds can be performed for optimum ligand protein interaction. QM calculations were performed to derive parameters describing the metal ion and its bonded terms to the surrounding residues; these parameters were implemented in the CHARMM force field and validated by MD for 20 ns. In future we intend to extend the simulations further and perform full MD trajectory analysis. Ligand stability and interaction could be further investigated through molecular a dynamics simulation which was not performed due to time constraints. A detailed study of GCH1 dynamics will allow for future identification of the protein interaction with the selected SANCDB compounds resulting in lead compounds that may be taken for further analysis and wet laboratory tests.

References

Alberts, B. et al. (2002) _Protein Function‘. Garland Science. Available at: https://www.ncbi.nlm.nih.gov/books/NBK26911/ (Accessed: 24 January 2018).

Alder, B. J. and Wainwright, T. E. (1957) _Phase Transition for a Hard Sphere System‘, J. Chem. Phys., 27(5), pp. 1208–1209. doi: 10.1063/1.1743957.

Alder, B. J. and Wainwright, T. E. (1959) _Studies in Molecular Dynamics. I. General Method‘, The Journal of Chemical Physics, 31(2), pp. 459–466. doi: 10.1063/1.1730376. Allen, M. (2004) _Introduction to molecular dynamics simulation ‘, Computational Soft Matter: From Synthetic Polymers to …, 23(2), pp. 1–28. doi: 10.1016/j.cplett.2006.06.020. Altschul, S. F. (2005) _BLAST Algorithm ‘, in Encyclopedia of Life Sciences. doi: 10.1038/npg.els.0005253.

Anderson, A. C. (2005) _Targeting DHFR in parasitic protozoa ‘, Drug Discovery Today, pp. 121–128. doi: 10.1016/S1359-6446(04)03308-2.

Andrej Šali (1993) _MODELLER A Program for Protein Structure Modeling ‘, Comparative protein modelling by satisfaction of spatial restraints., pp. 779–815.

Auerbach, G. et al. (2000) _Zinc plays a key role in human and bacterial GTP cyclohydrolase I. ‘, Proceedings of the National Academy of Sciences of the United States of America, 97(25), pp. 13567–72. doi: 10.1073/pnas.240463497.

Autino, B. et al. (2012) _Epidemiology of malaria in endemic areas ‘, Mediterranean Journal of Hematology and Infectious Diseases. doi: 10.4084/MJHID.2012.060.

Bailey, T. L. et al. (2015) _The MEME Suite ‘, Nucleic Acids Research, 43(W1), pp. W39–W49. doi: 10.1093/nar/gkv416.

Baker, D. and Sali, a (2001) _Protein structure prediction and structural genomics. ‘, Science, 294(5540), pp. 93–96. doi: 10.1126/science.1065659.

Basco, L. K., Ramiliarisoa, O. and Le Bras, J. (1994) ‘In vitro activity of pyrimethamine, cycloguanil, and other antimalarial drugs against African isolates and clones of Plasmodium falciparum’, American Journal of Tropical Medicine and Hygiene, 50(2), pp. 193–199. doi: 10.4269/ajtmh.1994.50.193.

Bateman, A. et al. (2017) ‘UniProt: the universal protein knowledgebase’, Nucleic Acids Research. Oxford University Press, 45(D1), pp. D158–D169. doi: 10.1093/nar/gkw1099.

Becke, A. D. (1988) _Density-functional exchange-energy approximation with correct asymptotic behavior ‘, Physical Review A, 38(6), pp. 3098–3100. doi: 10.1103/PhysRevA.38.3098.

Beier, J. C. et al. (2008) _Integrated vector management for malaria control. ‘, Malaria journal, 7 Suppl 1, p. S4. doi: 10.1186/1475-2875-7-S1-S4.

Benkert, P., Biasini, M. and Schwede, T. (2011) _Toward the estimation of the absolute quality of individual protein structure models ‘, Bioinformatics, 27(3), pp. 343–350. doi: 10.1093/bioinformatics/btq662.

Benkert, P., Tosatto, S. C. E. and Schomburg, D. (2008) ‘QMEAN: A comprehensive scoring function for model quality assessment’, Proteins-Structure Function and Bioinformatics, 71(1), pp. 261–277. doi: 10.1002/prot.21715.

Biasini, M. et al. (2014) _SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information ‘, Nucleic Acids Research, 42(W1). doi: 10.1093/nar/gku340.

Boas, F. E. and Harbury, P. B. (2007) _Potential energy functions for protein design ‘, Current Opinion in Structural Biology, pp. 199–204. doi: 10.1016/j.sbi.2007.03.006.

Bonvin, A. M. (2006) _Flexible protein-protein docking ‘, Current Opinion in Structural Biology, pp. 194–200. doi: 10.1016/j.sbi.2006.02.002.

Bowen, J. P. and Allinger, N. L. (2007) _Molecular Mechanics: The Art and Science of Parameterization ‘, Reviews in Computational Chemistry, 2, pp. 81–97. doi: 10.1002/9780470125793.ch3.

Brooks, B. R. et al. (1983) _CHARMM: A program for macromolecular energy, minimization, and dynamics calculations ‘, Journal of Computational Chemistry, 4(2), pp. 187–217. doi: 10.1002/jcc.540040211.

Brooks, B. R. et al. (2009) _CHARMM: The biomolecular simulation program ‘, Journal of Computational Chemistry, 30(10), pp. 1545–1614. doi: 10.1002/jcc.21287.

Burgs, A. W. and Brown, G. M. (1968) _The Biosynthesis of Folic Acid VIII. PURIFICATION AND PROPERTIES OF THE ENZYME THAT CATALYZES THE PRODUCTION OF FORMATE FROM CARBON ATOM 8 OF GUANOSINE TRIPHOSPHATE*‘, THE JOURNAL OF BIOLOGICAL CHEMISTRY, 243(9), pp. 2349–
2358. Available at: http://www.jbc.org/content/243/9/2349.full.pdf (Accessed: 30 January 2018).

Campo, B. et al. (2015) _Killing the hypnozoite – drug discovery approaches to prevent relapse in Plasmodium vivax ‘, Pathogens and Global Health, 109(3), pp. 107–122. doi: 10.1179/2047773215Y.0000000013.

Carlton, J. M. et al. (2002) _Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. ‘, Nature, 419(6906), pp. 512–519. doi: 10.1038/nature01099.

Carter, R. and Walliker, D. (1975) _New observations on the malaria parasites of rodents of the Central African Republic— Plasmodium vinckei petteri subsp. nov. and Plasmodium chabaudi

Landau, 1965‘, Annals of Tropical Medicine & Parasitology. Taylor & Francis, 69(2), pp. 187–196. doi: 10.1080/00034983.1975.11687000.

Cadet, E. et al. (2005) _Histidine-aromatic interactions in proteins and protein-ligand complexes: Quantum chemical study of X-ray and model structures ‘, Journal of Chemical Theory and Computation, 1(3), pp. 472–483. doi: 10.1021/ct049875k.

Cavasotto, C. N. and Orry, A. J. W. (2007) _Ligand docking and structure-based virtual screening in drug discovery. ‘, Current topics in medicinal chemistry, 7(10), pp. 1006–1014. doi: 10.2174/156802607780906753.

Chulay, J. D., atkins, W. M. and Sixsmith, D. G. (1984) _Synergistic antimalarial activity of pyrimethamine and sulfadoxine against Plasmodium falciparum in vitro ‘, American Journal of Tropical Medicine and Hygiene, 33(3), pp. 325–330.

Cohen, A. J., Mori-Sánchez, P. and Yang, W. (2012) _Challenges for density functional theory ‘, Chemical Reviews, pp. 289–320. doi: 10.1021/cr200107z.

Collins, M. A. (2002) _Molecular potential-energy surfaces for chemical reaction dynamics ‘, Theoretical Chemistry Accounts, 108(6), pp. 313–324. doi: 10.1007/s00214-002-0383-5.

Cornell, W. D. et al. (1995) _A Second-Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules ‘, Journal of the American Chemical Society, 117(19), pp. 5179–5197. doi: 10.1021/ja00124a002.

Cowman, A. F., Berry, D. and Baum, J. (2012) _The cell biology of disease: The cellular and molecular basis for malaria parasite invasion of the human red blood cell ‘, The Journal of Cell Biology, 198(6), pp. 961–971. doi: 10.1083/jcb.201206112.

Cowman, A. F. and Crabb, B. S. (2002) _The Plasmodium falciparum genome--a blueprint for erythrocyte invasion. ‘, Science (New York, N.Y.), 298(5591), pp. 126–128. doi: 10.1126/science.1078169.

Cowman, A. F. and Crabb, B. S. (2006) _Invasion of red blood cells by malaria parasites ‘, Cell, pp. 755–766. doi: 10.1016/j.cell.2006.02.006.

Cramer, F. (2007) ‘Emil Fischer’s Lock-and-Key Hypothesis after 100 years-Towards a Supracellular Chemistry’, in. Wiley-Blackwell, pp. 1–23. doi: 10.1002/9780470511411.ch1.

Crider, K. S. et al. (2012) _Folate and DNA Methylation: A Review of Molecular Mechanisms and the Evidence for Folate ‘s Role ‘, Advances in Nutrition: An International Review Journal, 3(1), pp. 21–38. doi: 10.3945/an.111.000992.

Dayhoff, M. and Schwartz, R. (1978) _A Model of Evolutionary Change in Proteins ‘, In Atlas of protein sequence and structure, pp. 345–352. doi: 10.1.1.145.4315.

DeLano, W. L. (2002) _The PyMOL Molecular Graphics System, Version 0.99 Schrödinger, LLC. ‘, Schrödinger LLC, p. http://www.pymol.org. doi: citeulike-article-id:240061.

DeLano, W. L. (2014) _The PyMOL Molecular Graphics System, Version 1.8‘, Schrödinger LLC, p. http://www.pymol.org. doi: 10.1038/hr.2014.17.

Delves, M. et al. (2012) _The activities of current antimalarial drugs on the life cycle stages of plasmodium: A comparative study with human and rodent parasites ‘, PLoS Medicine, 9(2). doi: 10.1371/journal.pmed.1001169.

Dias, R. and de Azevedo Jr., W. (2008) _Molecular Docking Algorithms ‘, Current Drug Targets, 9(12), pp. 1040–1047. doi: 10.2174/138945008786949432.

Do, C. B. and Katoh, K. (2008) _Protein multiple sequence alignment ‘, Methods in Molecular Biology, pp. 379–413. doi: 10.1007/978-1-59745-398-1_25.

Dominik Marx and Jurg Huutter (2009) Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods, Annals of Physics. doi: 10.1017/CBO9781107415324.004.

Dorn, M. et al. (2014) _Three-dimensional protein structure prediction: Methods and computational strategies ‘, Computational Biology and Chemistry, pp. 251–276. doi: 10.1016/j.compbiolchem.2014.10.001.

Du, H. et al. (2015) _Protein structure prediction provides comparable performance to crystallographic structures in docking-based virtual screening ‘, Methods, 71(C), pp. 77–84. doi: 10.1016/j.ymeth.2014.08.017.

Durrant, J. D. and McCammon, J. A. (2011) _Molecular dynamics simulations and drug discovery ‘, BMC Biology, 9(1), p. 71. doi: 10.1186/1741-7007-9-71.

Edgar, R. C. (2004) _MUSCLE: Multiple sequence alignment with high accuracy and high throughput ‘, Nucleic Acids Research, 32(5), pp. 1792–1797. doi: 10.1093/nar/gkh340.

van Eijk, A. M. et al. (2015) _Prevalence of malaria infection in pregnant women compared with children for tracking malaria transmission in sub-Saharan Africa: A systematic review and meta-analysis ‘, The Lancet Global Health, 3(10), pp. e617–e628. doi: 10.1016/S2214- 109X (15)00049-2.

Eramian, D. et al. (2008) _How well can the accuracy of comparative protein structure models be predicted? ‘, Protein science: a publication of the Protein Society, 17(11), pp. 1881–93. doi: 10.1110/ps.036061.108.

Ferrada, E. and Melo, F. (2009) _Effective knowledge-based potentials ‘, Protein Science, 18(7), pp. 1469–1485. doi: 10.1002/pro.166.

Forrey, C., Douglas, J. F. and Gilson, M. K. (2012) _The fundamental role of flexibility on the strength of molecular binding ‘, Soft Matter, 8(23), p. 6385. doi: 10.1039/c2sm25160d.

Fundamentals of Molecular Dynamics for Nano-technology Applications Mario Blanco Materials and Process Simulation Center California Institute of Technology. - ppt download (no date). Available at: http://slideplayer.com/slide/6972574/ (Accessed: 23 January 2018).
Gardner, M. J. et al. (2002) _Genome sequence of the human malaria parasite Plasmodium falciparum. ‘, Nature, 419(6906), pp. 498–511. doi: 10.1038/nature01097.
Goodsell, D. S. and Olson, A. J. (1990) _Automated docking of substrates to proteins by simulated annealing ‘, Proteins: Structure, Function, and Bioinformatics, 8(3), pp. 195–202. doi: 10.1002/prot.340080302.
Frisch, M. J.; Trucks, G.W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.;Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenber, D. J. (2009) ‘Gaussian 09’, Gaussian, Inc. Wallingford CT, pp. 2–3. doi: 111.
Gopalakrishnan, A. M. and Kumar, N. (2015) _Antimalarial action of artesunate involves DNA damage mediated by reactive oxygen species ‘, Antimicrobial Agents and Chemotherapy, 59(1), pp. 317–325. doi: 10.1128/AAC.03663-14.
Gräwert, T., Fischer, M. and Bacher, A. (2013) _Structures and reaction mechanisms of GTP cyclohydrolases ‘, IUBMB Life, pp. 310–322. doi: 10.1002/iub.1153.
Greenwood, B. and Owusu-Agyei, S. (2012) _Malaria in the Post-Genome Era ‘, Science, 338, pp. 49–50. doi: 10.1126/science.1229177.
Grimberg, B. T. and Mehlotra, R. K. (2011) _Expanding the antimalarial drug arsenal-now, but how? ‘, Pharmaceuticals, pp. 681–712. doi: 10.3390/ph4050681.
Guedes, I. A., de Magalhães, C. S. and Dardenne, L. E. (2014) _Receptor-ligand molecular docking ‘, Biophysical Reviews, pp. 75–87. doi: 10.1007/s12551-013-0130-2.
Gwamaka, M. et al. (2012) _Iron deficiency protects against severe Plasmodium falciparum malaria and death in young children. ‘, Clinical infectious diseases: an official publication of the Infectious Diseases Society of America, 54(8), pp. 1137–44. doi: 10.1093/cid/cis010.
Hammoudeh, D. I. et al. (2013) _Replacing sulfa drugs with novel DHPS inhibitors ‘, Future Medicinal Chemistry, 5(11), pp. 1331–1340. doi: 10.4155/fmc.13.97.
Harrison, R. W. (1999) _Integrating quantum and molecular mechanics ‘, Journal of Computational Chemistry, 20(15), pp. 1618–1633. doi: 10.1002/(SICI)1096- 987X (19991130)20:15<1618: AID-JCC3>3.0.CO;2-V.
Hart, T. N. and Read, R. J. (1992) _A multiple-start Monte Carlo docking method ‘, Proteins: Structure, Function, and Bioinformatics, 13(3), pp. 206–222. doi: 10.1002/prot.340130304.
Hatherley, R. et al. (2015) _SANCDB: A South African natural compound database ‘, Journal

of Cheminformatics, 7(1). doi: 10.1186/s13321-015-0080-8.

Henikoff, S. and Henikoff, J. G. (1993) _Performance evaluation of amino acid substitution matrices. ‘, Proteins, 17(1), pp. 49–61. doi: 10.1002/prot.340170108.

Hildebrand, A. et al. (2009) _Fast and accurate automatic structure prediction with HHpred ‘, Proteins: Structure, Function and Bioinformatics, 77(SUPPL. 9), pp. 128–132. doi: 10.1002/prot.22499.

Hill, A. V. S. (2011) _Vaccines against malaria. ‘, Philosophical transactions of the Royal Society of London. Series B, Biological sciences. The Royal Society, 366(1579), pp. 2806– 14. doi: 10.1098/rstb.2011.0091.

Hillis, D. M. and Bull, J. J. (1993) _An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis ‘, Systematic Biology, 42(2), pp. 182–192. doi: 10.1093/sysbio/42.2.182.

Hillisch, A., Pineda, L. F. and Hilgenfeld, R. (2004) _Utility of homology models in the drug discovery process ‘, Drug Discovery Today, pp. 659–669. doi: 10.1016/S1359- 6446(04)03196-4.

Hollingsworth, S. A. and Karplus, P. A. (2010) _A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins ‘, Biomolecular Concepts, pp. 271–283. doi: 10.1515/bmc.2010.022.

Hu, L. and Ryde, U. (2011) _Comparison of methods to obtain force-field parameters for metal sites ‘, Journal of Chemical Theory and Computation, 7(8), pp. 2452–2463. doi: 10.1021/ct100725a.

Huang, S.-Y., Grinter, S. Z. and Zou, X. (2010) _Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. ‘, Physical chemistry chemical physics: PCCP, 12(40), pp. 12899–908. doi: 10.1039/c0cp00151a. Hyde, J. E. (2005) _Exploring the folate pathway in Plasmodium falciparum ‘, Acta Tropica, 94(3 SPEC. ISS.), pp. 191–206. doi: 10.1016/j.actatropica.2005.04.002.

Iorio, F. et al. (2010) _Discovery of drug mode of action and drug repositioning from transcriptional responses ‘, Proceedings of the National Academy of Sciences, 107(33), pp. 14621–14626. doi: 10.1073/pnas.1000138107.

Jain, A. N. (2006) _Scoring functions for protein-ligand docking. ‘, Current protein & peptide science, 7(5), pp. 407–20. doi: 10.2174/138920306778559395.

Janssen, C. S. et al. (2001) _Gene discovery in Plasmodium chabaudi by genome survey sequencing ‘, Molecular and Biochemical Parasitology, 113(2), pp. 251–260. doi:

10.1016/S0166-6851(01)00224-9.

Jarmuła, A. (2010) _Antifolate inhibitors of thymidylate synthase as anticancer drugs ‘, Mini. Rev. Med. Chem., 10(13), pp. 1211–1222. doi: 10.2174/13895575110091211.

Jefferson, E. R., Walsh, T. P. and Barton, G. J. (2006) _Biological Units and their Effect upon the Properties and Prediction of Protein-Protein Interactions ‘, Journal of Molecular Biology, 364(5), pp. 1118–1129. doi: 10.1016/j.jmb.2006.09.042.

Jonker, F. A. M. et al. (2012) _Iron status predicts malaria risk in Malawian preschool children ‘, PLoS ONE, 7(8). doi: 10.1371/journal.pone.0042670.

Jorgensen, W. L., Maxwell, D. S. and Tirado-Rives, J. (1996) _Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids ‘, Journal of the American Chemical Society, 118(45), pp. 11225–11236. doi: 10.1021/ja9621760.

Josling, G. a and Llinás, M. (2015) _Sexual development in Plasmodium parasites: knowing when it ‘s time to commit. ‘, Nature reviews. Microbiology, 13(9), pp. 573–87. doi: 10.1038/nrmicro3519.

Kakar, Q. et al. (2016) _Efficacy of artemisinin-based combination therapies for the treatment of falciparum malaria in Pakistan (2007–2015): In vivo response and dhfr and dhps mutations ‘, Acta Tropica, 164, pp. 17–22. doi: 10.1016/j.actatropica.2016.08.006.

Kalaimathy, S., Sowdhamini, R. and Kanagarajadurai, K. (2011) _Critical assessment of structure-based sequence alignment methods at distant relationships ‘, Briefings in Bioinformatics, 12(2), pp. 163–175. doi: 10.1093/bib/bbq025.

Kantele, A. and Jokiranta, T. S. (2011) _Review of cases with the emerging fifth human malaria parasite, Plasmodium knowlesi ‘, Clin. Infect. Dis., 52(11), pp. 1356–1362. doi: 10.1093/cid/cir180.

Katoh, K. et al. (2002) _MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. ‘, Nucleic acids research, 30(14), pp. 3059–3066. doi: 10.1093/nar/gkf436.

Katoh, K. and Standley, D. M. (2013) _MAFFT multiple sequence alignment software version 7: Improvements in performance and usability ‘, Molecular Biology and Evolution, 30(4), pp. 772–780. doi: 10.1093/molbev/mst010.

Kevin Baird, J. (2013) _Evidence and implications of mortality associated with acute plasmodium vivax malaria ‘, Clinical Microbiology Reviews, 26(1), pp. 36–57. doi: 10.1128/CMR.00074-12.

Koboldt, D. C. et al. (2013) _The next-generation sequencing revolution and its impact on genomics ‘, Cell. doi: 10.1016/j.cell.2013.09.006.

Koonin, E. V. (2005) _Orthologs, Paralogs, and Evolutionary Genomics ‘, Annual Review of Genetics, 39(1), pp. 309–338. doi: 10.1146/annurev.genet.39.073003.114725.

Koshland, D. E. (1995) ‘The Key–Lock Theory and the Induced Fit Theory’, Angewandte Chemie International Edition in English. Wiley-Blackwell, 33(2324), pp. 2375–2378. doi: 10.1002/anie.199423751.

Krieger, E., Nabuurs, S. B. and Vriend, G. (2005) _Homology Modeling ‘, in Structural Bioinformatics, pp. 509–523. doi: 10.1002/0471721204.ch25.

Kumar, S., Stecher, G. and Tamura, K. (2016) _MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets ‘, Molecular biology and evolution, 33(7), pp. 1870–1874. doi: 10.1093/molbev/msw054.

Kümpornsin, K. et al. (2014) _Biochemical and functional characterization of Plasmodium falciparum GTP cyclohydrolase i ‘, Malaria Journal, 13(1). doi: 10.1186/1475-2875-13-150.

Kwan, A. H. et al. (2011) _Macromolecular NMR spectroscopy for the non-spectroscopist ‘, FEBS Journal, pp. 687–703. doi: 10.1111/j.1742-4658.2011.08004. x.

Kyte, J. and Doolittle, R. F. (1982) _A simple method for displaying the hydropathic character of a protein ‘, Journal of Molecular Biology, 157(1), pp. 105–132. doi: 10.1016/0022-2836(82)90515-0.

Langhi, D. M. and Bordin, J. O. (2006) _Duffy blood group and malaria. ‘, Hematology, 11(5), pp. 389–398. doi: 10.1126/science.1257752.

Laskowski, R. A. et al. (1993) _PROCHECK: a program to check the stereochemical quality of protein structures ‘, Journal of Applied Crystallography, 26(2), pp. 283–291. doi: 10.1107/S0021889892009944.

Laskowski, R. A. and Swindells, M. B. (2011) ‘LigPlot+: Multiple ligand-protein interaction diagrams for drug discovery’, Journal of Chemical Information and Modeling, 51(10), pp. 2778–2786. doi: 10.1021/ci200227u.

Laurie, A. T. R. and Jackson, R. M. (2006) _Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. ‘, Current protein & peptide science, 7(5), pp. 395–406. doi: 10.2174/138920306778559386.

Le, S. Q., Lartillot, N. and Gascuel, O. (2008) _Phylogenetic mixture models for proteins ‘, Philosophical Transactions of the Royal Society B: Biological Sciences, 363(1512), pp. 3965–3976. doi: 10.1098/rstb.2008.0180.

Lee, C., Yang, W. and Parr, R. G. (1988) _Development of the Colle-Salvetti correlation-

energy formula into a functional of the electron density ‘, Physical Review B, 37(2), pp. 785–789. doi: 10.1103/PhysRevB.37.785.

Lewars, E. G. (2011) _The Concept of the Potential Energy Surface ‘, in Computational Chemistry, pp. 9–43. doi: 10.1007/978-90-481-3862-3_2.

Lexa, K. W. and Carlson, H. A. (2012) _Protein flexibility in docking and surface mapping ‘, Quarterly Reviews of Biophysics, pp. 301–343. doi: 10.1017/S0033583512000066.

Lin, H. and Truhlar, D. G. (2007) _QM/MM: what have we learned, where are we, and where do we go from here? ‘, Theoretical Chemistry Accounts. Springer-Verlag, 117(2), pp. 185– 199. doi: 10.1007/s00214-006-0143-z.

Lionta, E. et al. (2014) _Structure-based virtual screening for drug discovery: principles, applications and recent advances. ‘, Current Topics in Medicinal Chemistry, 14(16), pp. 1923–1938. doi: 10.2174/1568026614666140929124445.

Lipinski, C. A. (2004) _Lead- and drug-like compounds: The rule-of-five revolution ‘, Drug Discovery Today: Technologies, pp. 337–341. doi: 10.1016/j.ddtec.2004.11.007.

Lu, Z. et al. (2015) _Behavior of metal ions in bioelectrochemical systems: A review ‘, Journal of Power Sources, pp. 243–260. doi: 10.1016/j.jpowsour.2014.10.168.

di Luccio, E. and Koehl, P. (2012) _The H-factor as a novel quality metric for homology modeling. ‘, Journal of clinical bioinformatics, 2(1), p. 18. doi: 10.1186/2043-9113-2-18.

MacKerell, A. D. et al. (1998) _All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins † ‘, The Journal of Physical Chemistry B, 102(18), pp. 3586–3616. doi: 10.1021/jp973084f.

Mahmoudi, S. and Keshavarz, H. (2017) _Efficacy of phase 3 trial of RTS, S/AS01 malaria vaccine: The need for an alternative development plan ‘, Human Vaccines & Immunotherapeutics. Taylor & Francis, pp. 1–4. doi: 10.1080/21645515.2017.1295906.

Mardis, E. R. (2008) _The impact of next-generation sequencing technology on genetics ‘, Trends in Genetics, pp. 133–141. doi: 10.1016/j.tig.2007.12.007.

May, A. and Zacharias, M. (2005) _Accounting for global protein deformability during protein-protein and protein-ligand docking ‘, in Biochimica et Biophysica Acta - Proteins and Proteomics, pp. 225–231. doi: 10.1016/j.bbapap.2005.07.045.

Mbengue, A. et al. (2015) _A molecular mechanism of artemisinin resistance in Plasmodium falciparum malaria ‘, Nature, 520(7549), pp. 683–687. doi: 10.1038/nature14412.

McCammon, J. A., Gelin, B. R. and Karplus, M. (1977) _Dynamics of folded proteins ‘, Nature, 267(5612), pp. 585–590. doi: 10.1038/267585a0.

McGinnis, S. and Madden, T. L. (2004) _BLAST: At the core of a powerful and diverse set of

sequence analysis tools ‘, Nucleic Acids Research, 32(WEB SERVER ISS.). doi: 10.1093/nar/gkh435.

McGuffin, L. J., Bryson, K. and Jones, D. T. (2000) _The PSIPRED protein structure prediction server ‘, Bioinformatics, 16(4), pp. 404–405. doi: 10.1093/bioinformatics/16.4.404. McKerrow, J. H. and Lipinski, C. A. (2017) _The rule of five should not impede anti-parasitic drug development ‘, International Journal for Parasitology: Drugs and Drug Resistance, 7(2), pp. 248–249. doi: 10.1016/j.ijpddr.2017.05.003.

Melo, F. and Feytmans, E. (1998) _Assessing protein structures with a non-local atomic interaction energy ‘, Journal of Molecular Biology, 277(5), pp. 1141–1152. doi: 10.1006/jmbi.1998.1665.

Mindell, D. P. and Meyer, A. (2001) _Homology evolving ‘, Trends in Ecology and Evolution, pp. 434–440. doi: 10.1016/S0169-5347(01)02206-6.

Mockenhaupt, F. P. et al. (2000) _Anaemia in pregnant Ghanaian women: importance of malaria, iron deficiency, and haemoglobinopathies ‘, Transactions of the Royal Society of Tropical Medicine and Hygiene, 94, pp. 477–483. doi: http://dx.doi.org/10.1016/S0035-9203(00)90057-9.

Morris, G. M. et al. (1998) _Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function ‘, Journal of Computational Chemistry, 19(14), pp. 1639–1662. doi: 10.1002/(SICI)1096-987X (19981115)19:14<1639: AID-JCC10>3.0.CO;2-B.

Morris, G. M. et al. (2009) _Software news and updates AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility ‘, Journal of Computational Chemistry, 30(16), pp. 2785–2791. doi: 10.1002/jcc.21256.

Morris, G. M. and Lim-Wilby, M. (2008) _Molecular docking ‘, Methods Mol Biol, 443, pp. 365–382. doi: 10.1002/prot.

Morris G.M. and Dallakyan S. (2013) _AutoDock — AutoDock ‘, 02-27, 1(1), pp. 15–45.

Müller, I. B. and Hyde, J. E. (2013) _Folate metabolism in human malaria parasites—75 years on ‘, Molecular and Biochemical Parasitology, 188(1), pp. 63–77. doi: 10.1016/j.molbiopara.2013.02.008.

Nair, P. C. and Miners, J. O. (2014) _Molecular dynamics simulations: from structure function relationships to drug discovery. ‘, In silico pharmacology, 2(4), pp. 1–4. doi: 10.1186/s40203-014-0004-8.

Nalbantoğlu, Ö. U. (2014) ‗Dynamic programming. ‘, Methods in molecular biology (Clifton, N.J.), 1079, pp. 3–27. doi: 10.1007/978-1-62703-646-7_1.

Nar, H., Huber, R., Auerbach, G., et al. (1995) ‗Active site topology and reaction mechanism of GTP cyclohydrolase I. ‘, Proceedings of the National Academy of Sciences of the United States of America, 92(26), pp. 12120–12125. doi: 10.1073/pnas.92.26.12120.

Nar, H., Huber, R., Meining, W., et al. (1995) ‗Atomic structure of GTP cyclohydrolase I. ‘, Structure (London, England: 1993), 3(5), pp. 459–466.

Nayeem, A., Sitkoff, D. and Krystek, S. (2006) ‗A comparative study of available software for high-accuracy homology modeling: from sequence alignments to structural models. ‘, Protein science: a publication of the Protein Society, 15(4), pp. 808–824. doi: 10.1110/ps.051892906.

Needleman, S. B. and Wunsch, C. D. (1970) ‗A general method applicable to the search for similarities in the amino acid sequence of two proteins ‘, Journal of Molecular Biology, 48(3), pp. 443–453. doi: 10.1016/0022-2836(70)90057-4.

Nirmalan, N. et al. (2002) ‘Transcriptional analysis of genes encoding enzymes of the folate pathway in the human malaria parasite Plasmodium falciparum’, Molecular Microbiology, 46(1), pp. 179–190. doi: 10.1046/j.1365-2958.2002.03148. x.

Notredame, C. (2007) ‗Recent evolutions of multiple sequence alignment algorithms ‘, PLoS Computational Biology, pp. 1405–1408. doi: 10.1371/journal.pcbi.0030123.

Notredame, C., Higgins, D. G. and Heringa, J. (2000) ‗T-Coffee: A novel method for fast and accurate multiple sequence alignment. ‘, Journal of Molecular Biology, 302(1), pp. 205–17. doi: 10.1006/jmbi.2000.4042.

Nzila, A. et al. (2005) ‗Comparative folate metabolism in humans and malaria parasites (part II): Activities as yet untargeted or specific to Plasmodium ‘, Trends in Parasitology, pp. 334–339. doi: 10.1016/j.pt.2005.05.008.

Omi, R. et al. (2003) ‗Crystal structures of threonine synthase from Thermus thermophilus HB8: conformational change, substrate recognition, and mechanism. ‘, The Journal of biological chemistry, 278(46), pp. 46035–45. doi: 10.1074/jbc.M308065200.

Oostenbrink, C. et al. (2004) ‗A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6‘, Journal of Computational Chemistry, 25(13), pp. 1656–1676. doi: 10.1002/jcc.20090.

Özlem Tastan Bishop, A., de Beer, T. A. P. and Joubert, F. (2008) ‗Protein homology modelling and its use in South Africa ‘, South African Journal of Science, 104(February), pp. 2–6.

Pain, A. et al. (2008) ‗The genome of the simian and human malaria parasite Plasmodium

knowlesi.‘, Nature, 455(7214), pp. 799–803. doi: 10.1038/nature07306.

Patel, K., Kumar, A. and Durani, S. (2007) _Analysis of the structural consensus of the zinc coordination centers of metalloprotein structures‘, Biochimica et Biophysica Acta - Proteins and Proteomics, 1774(10), pp. 1247–1253. doi: 10.1016/j.bbapap.2007.07.010.

Pei, J. and Grishin, N. V (2014) _PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information.‘, Methods in molecular biology (Clifton, N.J.), 1079, pp. 263–71. doi: 10.1007/978-1-62703-646-7_17. peter Lam (2004) _Malaria: Causes, Symptoms and Treatments‘, Malaria: Causes, Symptoms and Treatments, (2016).

Pils, B., Copley, R. R. and Schultz, J. (2005) _Variation in structural location and amino acid conservation of functional sites in protein domain families‘, BMC Bioinformatics, 6. doi: 10.1186/1471-2105-6-210.

Pirolli, D. et al. (2014) _Insights from molecular dynamics simulations: Structural basis for the V567D mutation-induced instability of zebrafish alpha-dystroglycan and comparison with the murine model‘, PLoS ONE, 9(7). doi: 10.1371/journal.pone.0103866.

Prlić, A. et al. (2010) _Pre-calculated protein structure alignments at the RCSB PDB website‘, Bioinformatics, 26(23), pp. 2983–2985. doi: 10.1093/bioinformatics/btq572.

PlasmoDB (2001) ‘PlasmoDB: An integrative database of the Plasmodium falciparum genome. Tools for accessing and analyzing finished and unfinished sequence data. The Plasmodium Genome Database Collaborative.’, Nucleic acids research, 29(1), pp. 66–9. doi: 10.1093/nar/29.1.66.

Prudêncio, M., Rodriguez, A. and Mota, M. M. (2006) _The silent path to thousands of merozoites: the Plasmodium liver stage.‘, Nature reviews. Microbiology, 4(11), pp. 849–56. doi: 10.1038/nrmicro1529.

Quang, L. S., Gascuel, O. and Lartillot, N. (2008) _Empirical profile mixture models for phylogenetic reconstruction‘, Bioinformatics, 24(20), pp. 2317–2323. doi: 10.1093/bioinformatics/btn445.

Ramachandran, G. N., Ramakrishnan, C. and Sasisekharan, V. (1963) ‘Stereochemistry of polypeptide chain configurations’, Journal of Molecular Biology, pp. 95–99. doi: 10.1016/S0022-2836(63)80023-6.

Rao, V. S. and Srinivas, K. (2011) _Modern drug discovery process : An in silico approach‘, Journal of Bioinformatics and Sequence Analysis, 2(June), pp. 89–94.

Rebelo, J. et al. (2003) _Biosynthesis of pteridines. Reaction mechanism of GTP cyclohydrolase I‘, Journal of Molecular Biology, 326(2), pp. 503–516. doi: 10.1016/S0022- 2836(02)01303-7.

Robson, K. J. H. (1995) ‗Thrombospondin-related adhesive protein (TRAP) of Plasmodium falciparum: Expression during sporozoite ontogeny and binding to human hepatocytes', Parasitology Today, p. 410. doi: 10.1016/0169-4758(95)80020-4.

Runge, E. and Gross, E. K. U. (1984) ‗Density-functional theory for time-dependent systems', Physical Review Letters, 52(12), pp. 997–1000. doi: 10.1103/PhysRevLett.52.997. Rutledge, G. G. et al. (2017) ‗Plasmodium malariae and P. ovale genomes provide insights into malaria parasite evolution', Nature, 542(7639), pp. 101–104. doi: 10.1038/nature21038. Saitou, N. and Nei, M. (1987) ‗The neighbour-joining method: a new method for reconstructing phylogenetic trees', Mol Biol Evo, 4(4), pp. 406–425. doi: citeulike-article- id:93683.

Salcedo-Sora, J. E. and Ward, S. A. (2013) ‗The folate metabolic network of Falciparum malaria', Molecular and Biochemical Parasitology, pp. 51–62. doi: 10.1016/j.molbiopara.2013.02.003.

Šali, a (2013) ‗MODELLER: A Program for Protein Structure Modeling Release 9.12, r9480', Rockefeller University, pp. 779–815.

Sallem, M. A. S., De Sousa, S. A. and E Silva, F. J. D. S. (2007) ‗AutoGrid: Towards an autonomic grid middleware', in Proceedings of the Workshop on Enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE, pp. 223–228. doi: 10.1109/WETICE.2007.4407158.

San Diego: Accelrys Software Inc. (2012) Discovery Studio Modeling Environment, Release 3.5, Accelrys Software Inc. Sander, C. and Schneider, R. (1991) ‗Database of homology???derived protein structures and the structural meaning of sequence alignment', Proteins: Structure, Function, and Bioinformatics, 9(1), pp. 56–68. doi: 10.1002/prot.340090107.

Sazawal, S. et al. (2006) ‗Effects of routine prophylactic supplementation with iron and folic...', The Lancet, 367(9505), pp. 133–143.

Schmidt, T., Bergner, A. and Schwede, T. (2014) ‗Modelling three-dimensional protein structures for applications in drug design', Drug Discovery Today, pp. 890–897. doi: 10.1016/j.drudis.2013.10.027.

Shen, M. and Sali, A. (2006) ‗Statistical potential for assessment and prediction of protein structures', Protein Science, 15(11), pp. 2507–2524. doi: 10.1110/ps.062416606.

Sibley, C. H. et al. (2001) ‗Pyrimethamine-sulfadoxine resistance in Plasmodium falciparum: what next?', Trends Parasitol, 17(12), pp. 582–588. doi: 10.1016/S1471-4922(01)02185-7.

Sibley, C. H. et al. (2001) ‗Pyrimethamine-sulfadoxine resistance in Plasmodium falciparum:

What next?‘, Trends in Parasitology, pp. 582–588. doi: 10.1016/S1471-4922(01)02085-2.
De Silva, P. M. and Marshall, J. M. (2012) _Factors contributing to urban malaria transmission in sub-saharan Africa: A systematic review‘, Journal of Tropical Medicine. doi: 10.1155/2012/819563.
Sims, P. F. G. and Hyde, J. E. (2006) _Proteomics of the human malaria parasite Plasmodium falciparum.‘, Expert review of proteomics, 3(1), pp. 87–95. doi: 10.1586/14789450.3.1.87.
Singh, K. and Mehta, S. (2016) _The clinical development process for a novel preventive vaccine: An overview‘, Journal of Postgraduate Medicine, 62(1), pp. 4–11. doi: 10.4103/0022-3859.173187.
Sippl, M. J. (1995) _Knowledge-based potentials for proteins‘, Current Opinion in Structural Biology, 5(2), pp. 229–235. doi: 10.1016/0959-440X(95)80081-6.
Smith, T. F. and Waterman, M. S. (1981) _Identification of common molecular subsequences‘, Journal of Molecular Biology, 147(1), pp. 195–197. doi: 10.1016/0022- 2836(81)90087-5.
Snow, R. W. (2015) _Global malaria eradication and the importance of Plasmodium falciparum epidemiology in Africa.‘, BMC medicine, 13(1), p. 23. doi: 10.1186/s12916-014- 0254-7.
Söding, J. (2005) _Protein homology detection by HMM-HMM comparison‘, Bioinformatics, 21(7), pp. 951–960. doi: 10.1093/bioinformatics/bti125.
Söding, J., Biegert, A. and Lupas, A. N. (2005) _The HHpred interactive server for protein homology detection and structure prediction‘, Nucleic Acids Research, 33(SUPPL. 2). doi: 10.1093/nar/gki408.
Soulard, V. et al. (2015) _Plasmodium falciparum full life cycle and Plasmodium ovale liver stages in humanized mice.‘, Nature communications, 6(May), p. 7690. doi: 10.1038/ncomms8690.
Stadtman, E. R. (1990) _Metal ion-catalyzed oxidation of proteins: Biochemical mechanism and biological consequences‘, Free Radical Biology and Medicine, 9(4), pp. 315–325. doi: 10.1016/0891-5849(90)90006-5.

Stillinger, F. H. and Rahman, A. (1974) _Improved simulation of liquid water by molecular dynamics‘, J. Chem. Phys., 60(4), pp. 1545–1557. doi: 10.1063/1.1681229.
Takken, W. and Knols, B. G. (2009) _Malaria vector control: current and future strategies‘, Trends Parasitol, 25(3), pp. 101–104. doi: 10.1016/j.pt.2008.12.002.
Talman, A. M. et al. (2004) _Gametocytogenesis: the puberty of Plasmodium falciparum.‘, Malaria journal, 3, p. 24. doi: 10.1186/1475-2875-3-24.
Tanaka, Y. et al. (2005) _Novel reaction mechanism of GTP cyclohydrolase I. High- resolution

X-ray crystallography of Thermus thermophilus HB8 enzyme complexed with a transition state analogue, the 8-oxoguanine derivative‘, Journal of Biochemistry, 138(3), pp. 263–275. doi: 10.1093/jb/mvi120.

Tanchuk, V. et al. (2015) ‗A New Scoring Function for Molecular Docking Based on AutoDock and AutoDock Vina‘, Current Drug Discovery Technologies, 12(3), pp. 170–178. doi: 10.2174/1570163812666150825110208.

Tatham, A. L. et al. (2009) ‗GTP cyclohydrolase I expression, protein, and activity determine intracellular tetrahydrobiopterin levels, independent of GTP cyclohydrolase feedback regulatory protein expression‘, Journal of Biological Chemistry, 284(20), pp. 13660–13668. doi: 10.1074/jbc.M807959200.

Tatusov, R. L., Koonin, E. V. and Lipman, D. J. (1997) ‗A genomic perspective on protein families‘, Science, 278(5338), pp. 631–637. doi: 10.1126/science.278.5338.631.

Teague, S. J. (2003) ‗Implications of protein flexibility for drug discovery‘, Nature Reviews Drug Discovery, 2(7), pp. 527–541. doi: 10.1038/nrd1129.

Teodoro, M. L., Phillips, J. and Kavraki, L. E. (2001) ‗Molecular docking: A problem with thousands of degrees of freedom‘, in Proceedings - IEEE International Conference on Robotics and Automation, pp. 960–965. doi: 10.1109/ROBOT.2001.932674.

The Mathworks Inc. (2016) MATLAB - MathWorks, www.mathworks.com/products/matlab. doi: 2016-11-26.

Thomsen, R. and Christensen, M. H. (2006) ‗MolDock: A new technique for high-accuracy molecular docking‘, Journal of Medicinal Chemistry, 49(11), pp. 3315–3321. doi: 10.1021/jm051197e.

Thöny, B., Auerbach, G. and Blau, N. (2000) ‗Tetrahydrobiopterin biosynthesis, regeneration and functions.‘, The Biochemical journal, 347 Pt 1, pp. 1–16. doi: 10.1042/0264-6021:3470001.

Tilley, L., Dixon, M. W. a and Kirk, K. (2011) ‗The Plasmodium falciparum-infected red blood cell.‘, The international journal of biochemistry & cell biology, 43(6), pp. 839–42. doi: 10.1016/j.biocel.2011.03.012.

Toomula, N. et al. (2012) ‗Biological Databases- Integration of Life Science Data‘, Journal of Computer Science & Systems Biology, 4(5). doi: 10.4172/jcsb.1000081.

Trott, O. and Olson, A. J. (2010) ‗AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading‘, Journal of

Computational Chemistry, 31(2), pp. 455–61. doi: 10.1002/jcc.

Vanommeslaeghe, K. et al. (2014) _Molecular mechanics.‘, Current pharmaceutical design, 20(20), pp. 3281–92. doi: 10.1016/j.biotechadv.2011.08.021.Secreted.

Vaughan, A. M., Aly, A. S. I. and Kappe, S. H. I. (2008) _Malaria Parasite Pre-Erythrocytic Stage Infection: Gliding and Hiding‘, Cell Host and Microbe, pp. 209–218. doi: 10.1016/j.chom.2008.08.010.

Trott, O. and Olson, A. (2010) ‘AutoDock Vina: inproving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading’, Journal of Computational Chemistry, 31(2), pp. 455–461. doi: 10.1002/jcc.21334.AutoDock.

Verdonk, M. L. et al. (2003) _Improved protein-ligand docking using GOLD.‘, Proteins, 52(4), pp. 609–23. doi: 10.1002/prot.10465.

Vriend, G. (1990) _WHAT IF: A molecular modeling and drug design program‘, Journal of Molecular Graphics, 8(1), pp. 52–56. doi: 10.1016/0263-7855(90)80070-V.

Vyas, V. K. et al. (2012) _Homology Modeling a Fast Tool for Drug Discovery: Current Perspectives‘, Indian Journal of Pharmaceutical Sciences, 1, pp. 1–17. doi: 10.4103/0250-474X.102537.

Wang, L. et al. (2008) _Design and implementation of parallel lamarckian genetic algorithm for automated docking of molecules‘, in Proceedings - 10th IEEE International Conference on High Performance Computing and Communications, HPCC 2008, pp. 689–694. doi: 10.1109/HPCC.2008.50.

Waterhouse, A. M. et al. (2009) _Jalview Version 2-A multiple sequence alignment editor and analysis workbench‘, Bioinformatics, 25(9), pp. 1189–1191. doi: 10.1093/bioinformatics/btp033.

Webb, B. and Sali, A. (2014) _Comparative protein structure modeling using MODELLER‘, Current Protocols in Bioinformatics, 2014, p. 5.6.1-5.6.32. doi: 10.1002/0471250953.bi0506s47.

Webb, B. and Sali, A. (2016) _Comparative protein structure modeling using MODELLER‘, Current Protocols in Protein Science, 2016, p. 2.9.1-2.9.37. doi: 10.1002/cpps.20. Weiner, P. K. and Kollman, P. A. (1981) _AMBER: Assisted model building with energy refinement. A general program for modeling molecules and their interactions‘, Journal of Computational Chemistry, 2(3), pp. 287–303. doi: 10.1002/jcc.540020311.

White, N. J. (1998) _Preventing antimalarial drug resistance through combinations‘, Drug Resistance Updates, 1(1), pp. 3–9. doi: 10.1016/S1368-7646(98)80208-2.

White, N. J. (2004) ‗Antimalarial drug resistance', J.Clin.Invest, 113(0021–9738 (Print)), pp. 1084–1092. doi: 10.1172/JCI200421682.1084.

White, N. J. (2008) ‗Plasmodium knowlesi: The Fifth Human Malaria Parasite', Clinical Infectious Diseases, 46(2), pp. 172–173. doi: 10.1086/524889.

White, N. J. (2008) ‗The role of anti-malarial drugs in eliminating malaria', Malaria Journal. doi: 10.1186/1475-2875-7-S1-S8.

Whitty, C. J. M. et al. (2008) ‗Deployment of ACT antimalarials for treatment of malaria: challenges and opportunities.', Malaria journal, 7 Suppl 1(Suppl 1), p. S7. doi: 10.1186/1475-2875-7-S1-S7.

Wiederstein, M. 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(SUPPL.2). doi: 10.1093/nar/gkm290.

Wilby, K. J. et al. (2012) ‗Mosquirix (RTS,S): a novel vaccine for the prevention of Plasmodium falciparum malaria.', The Annals of pharmacotherapy, 46(3), pp. 384–93. doi: 10.1345/aph.1AQ634.

Wilson, R. J. and Williamson, D. H. (1997) ‗Extrachromosomal DNA in the Apicomplexa.', Microbiology and molecular biology reviews : MMBR, 61(1), pp. 1–16.

Witter, K. et al. (1996) ‗Cloning, sequencing and functional studies of the gene encoding human GTP cyclohydrolase I', Gene, 171(2), pp. 285–290. doi: 10.1016/0378- 1119(95)00886-1.

Wlodawer, A. et al. (2008) ‗Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures', FEBS Journal, pp. 1–21. doi: 10.1111/j.1742-4658.2007.06178.x.

World Health Organization (2016a) Malaria vaccine: WHO position paper-January 2016, Weekly Epidemiological Record. doi: 10.1371/jour.

World Health Organization (2016b) World Malaria Report 2016, World Health Organization. doi: 10.1071/EC12504.

Wüthrich, K. (2003) ‗NMR studies of structure and function of biological macromolecules (Nobel Lecture)', Journal of Biomolecular NMR, pp. 13–39. doi: 10.1023/A:1024733922459.

Xia, X. (2017) 'Bioinformatics and Drug Discovery.', Current topics in medicinal chemistry. Bentham Science Publishers, 17(15), pp. 1709–1726. doi: 10.2174/1568026617666161116143440.

Xiong, J. (2006) Essential bioinformatics, Essential Bioinformatics. doi: 10.1017/CBO9780511806087.

Xmgr: Introduction (no date). Available at: http://plasma-

gate.weizmann.ac.il/Xmgr/doc/intro.html#copyright (Accessed: 2 February 2018). Yang, Y.-Z., Little, B. and Meshnick, S. R. (1994) ‗ALKYLATION OF PROTEINS BY ARTEMISININ EFFECTS OF HEME, pH, AND DRUG STRUCTURE', Biochemical Pharmacology, 48(3), pp. 569–573. Available at: https://deepblue.lib.umich.edu/bitstream/handle/2027.42/31394/0000308.pdf?sequence=1&isAllowed=y (Accessed: 6 February 2018).

Yang, Z. and Rannala, B. (2012) ‗Molecular phylogenetics: Principles and practice', Nature Reviews Genetics, pp. 303–314. doi: 10.1038/nrg3186.

Yuthavong, Y. et al. (2006) ‗Folate metabolism as a source of molecular targets for antimalarials.', Future microbiology, 1(1), pp. 113–25. doi: 10.2217/17460913.1.1.113.

Zdobnov, E. M. and Apweiler, R. (2001) ‗InterProScan - An integration platform for the signature-recognition methods in InterPro', Bioinformatics, 17(9), pp. 847–848. doi: 10.1093/bioinformatics/17.9.847.

Zhou, Z. H. (2008) ‗Towards atomic resolution structural determination by single-particle cryo-electron microscopy', Current Opinion in Structural Biology, pp. 218–228. doi: 10.1016/j.sbi.2008.03.004.

Zlotkin, S. et al. (2013) ‗Effect of iron fortification on malaria incidence in infants and young children in Ghana: a randomized trial.', JAMA : the journal of the American Medical Association, 310(9), pp. 938–47. doi: 10.1001/jama.2013.277129.

Appendices

Appendix 1 – sequence alignment and analysis

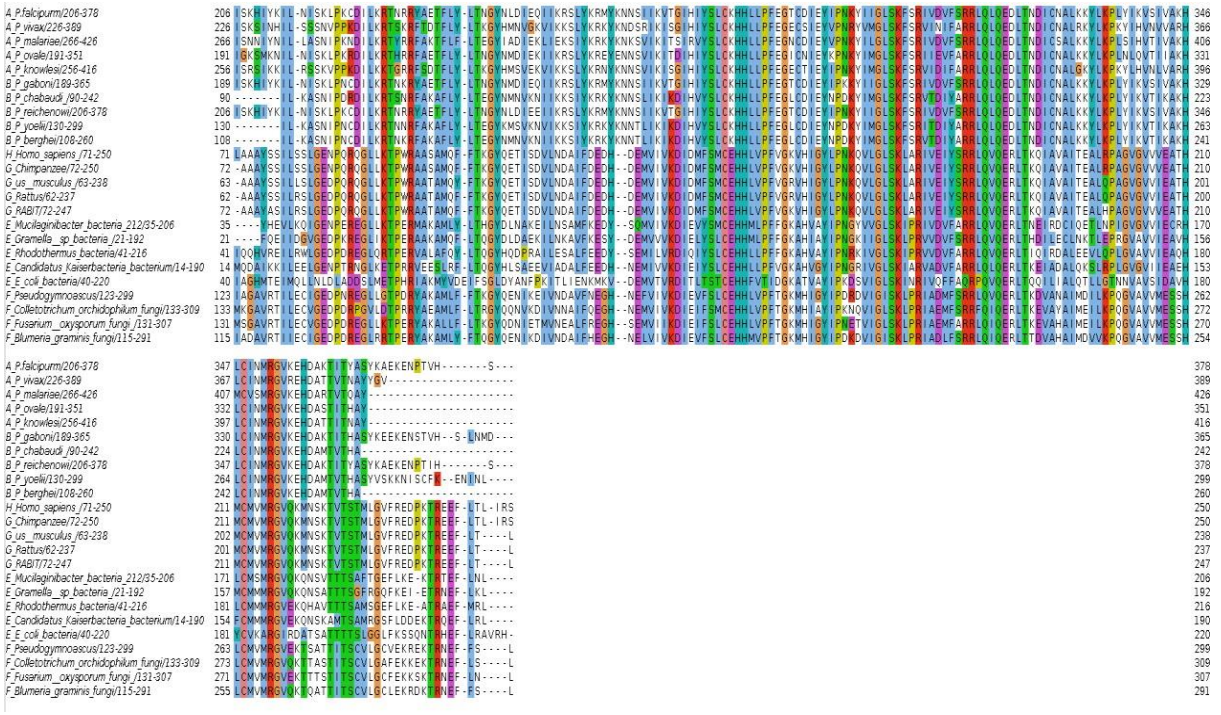


Figure A1.1: MAFFT multiple sequence alignment results. Residues are shaded by conservation using Jalview alignment editor tool.

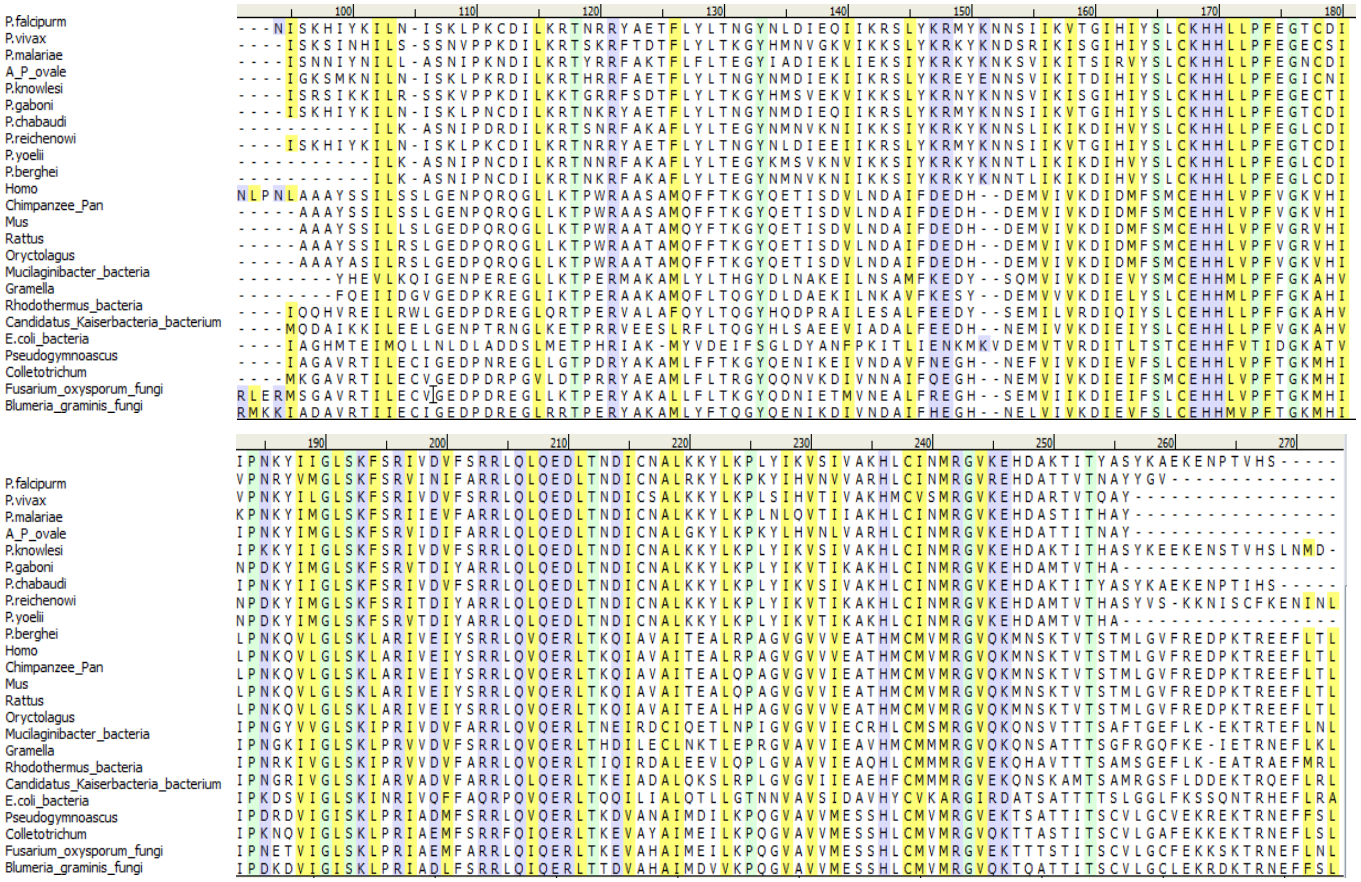


Figure A1.2: Multiple sequence alignment coloured by hydrophobicity similarity using Discovery studio visualization tool. This highlights differences and similarities based on amino acids hydrophobicity.

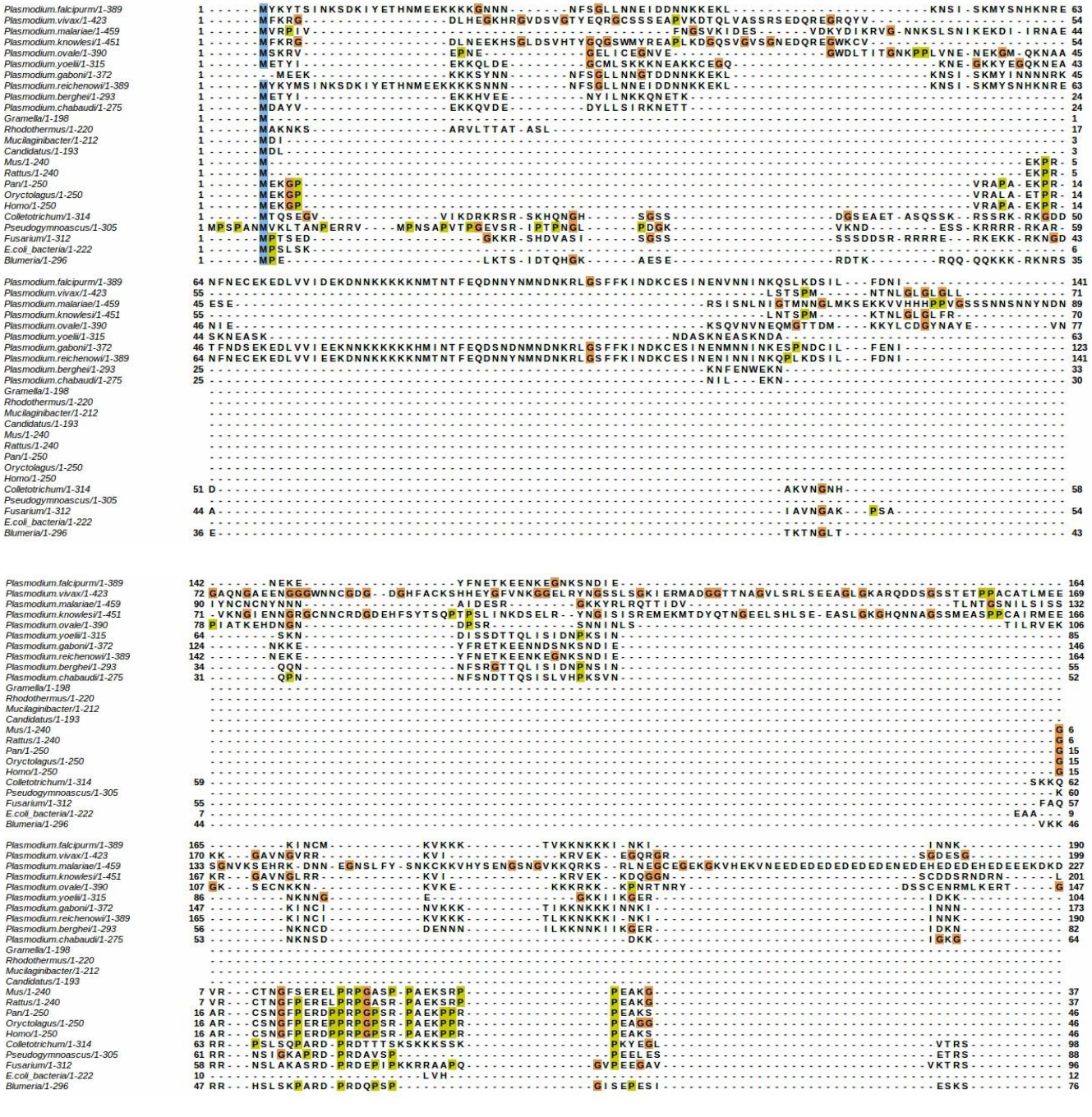


Figure A1.3: Multiple sequence alignment of the GCH1 protein complete sequence. This shows the high level of variation in the N-terminal region. Residues are shaded by conservation using Jalview alignment editor tool (Waterhouse *et al*. 2009). High sequence variation is observed from the N and C-terminals of all GCH1 sequences indicating the exclusivity of these regions among all species.

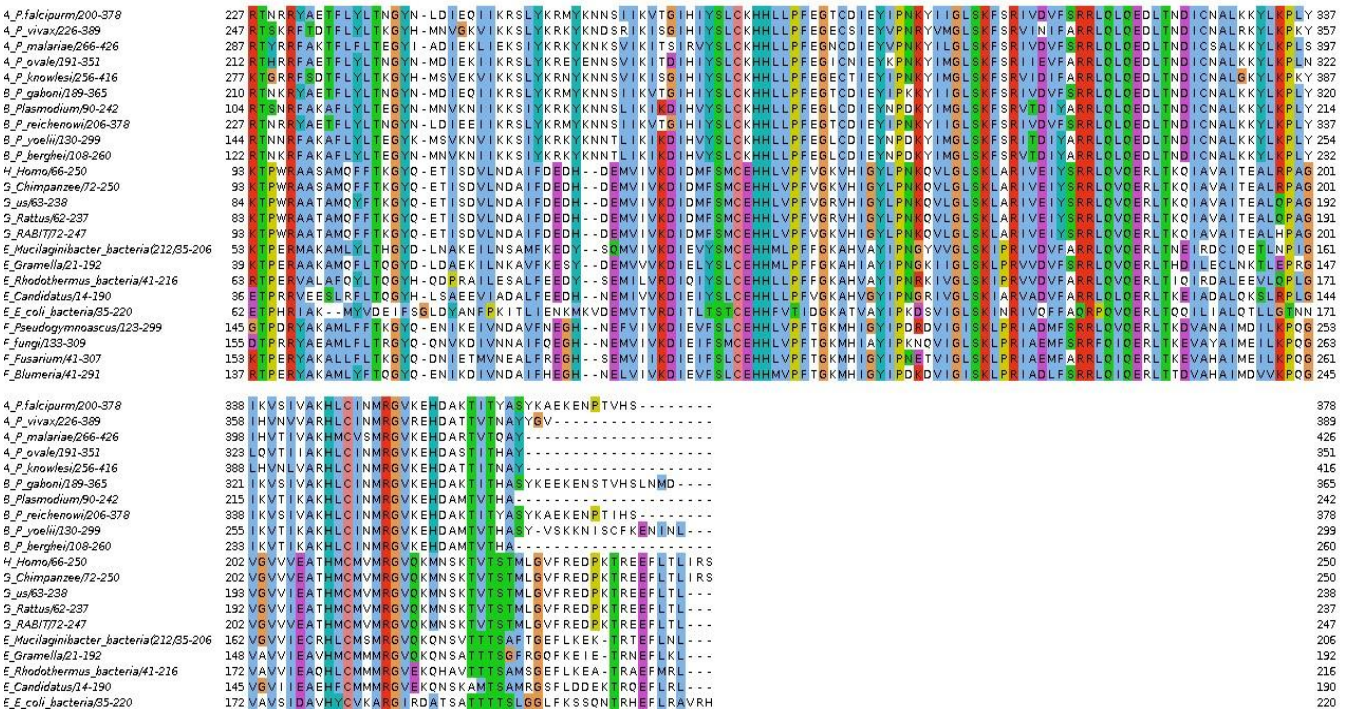


Figure A1.4: MUSCLE multiple sequence alignment of GCH1 catalytic domain results shaded by residue conservation using Jalview alignment editor tool.

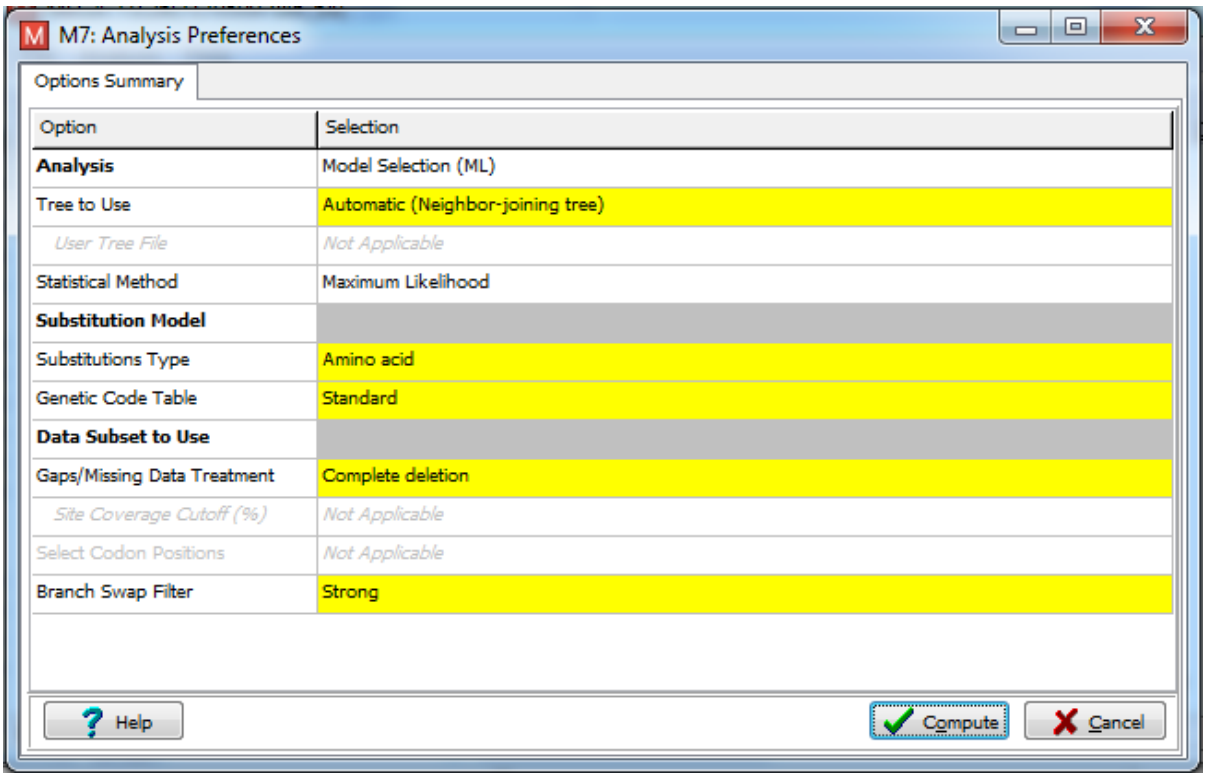


Figure A1.5: MEGA7 phylogenetic analysis preferences.

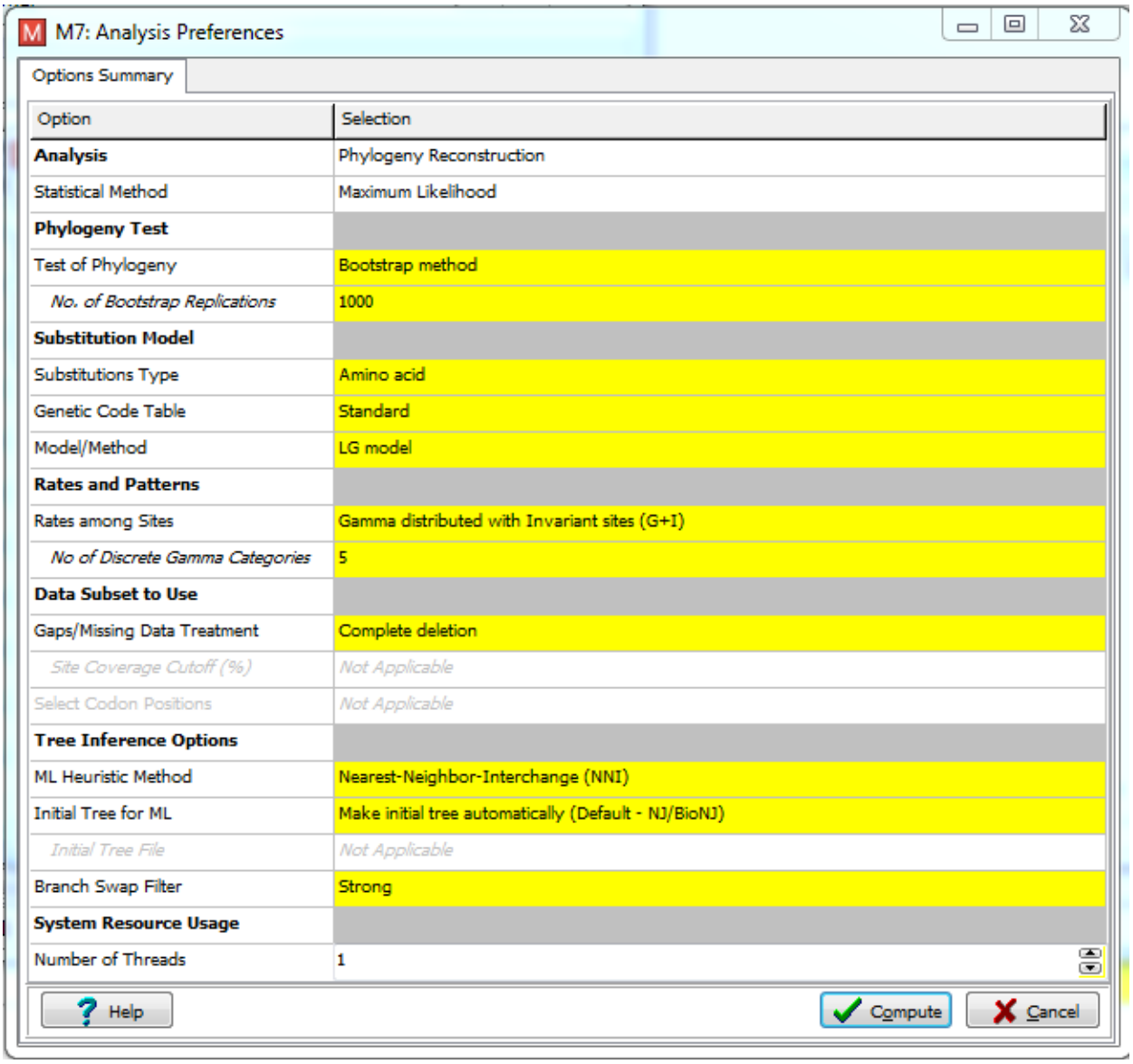
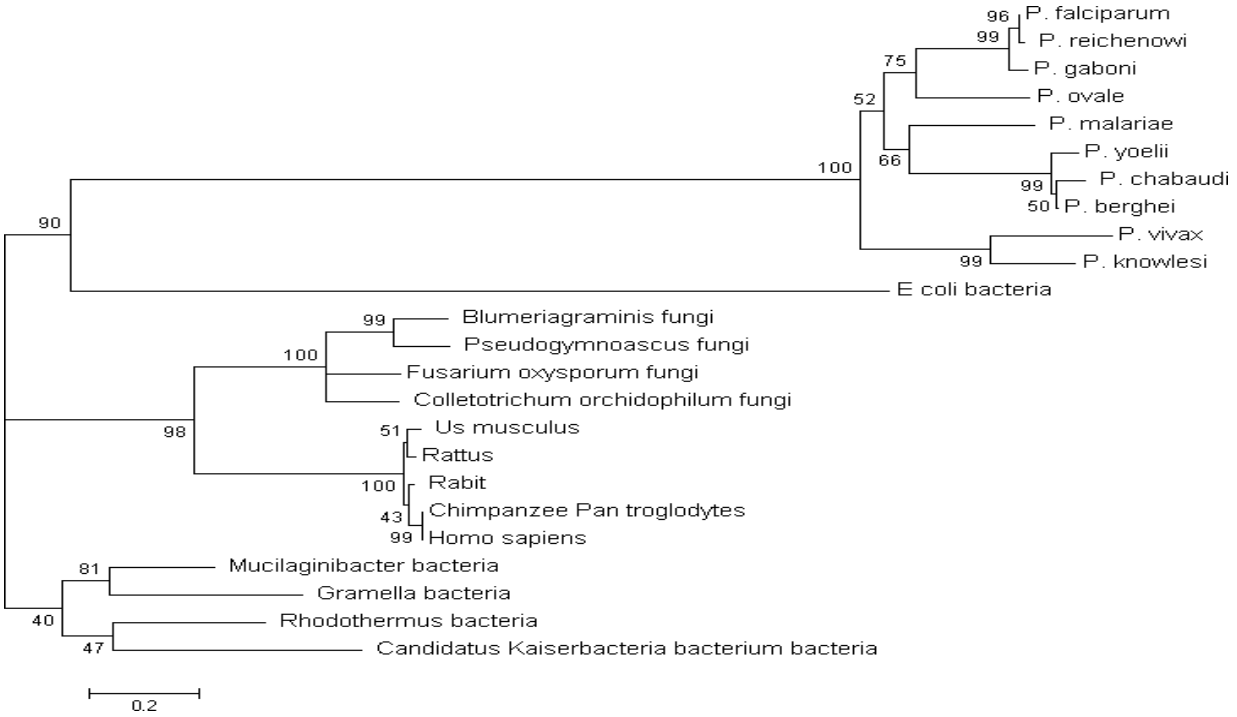


Figure A1.6: Phylogenetic tree calculation parameters.

(A)



(B)

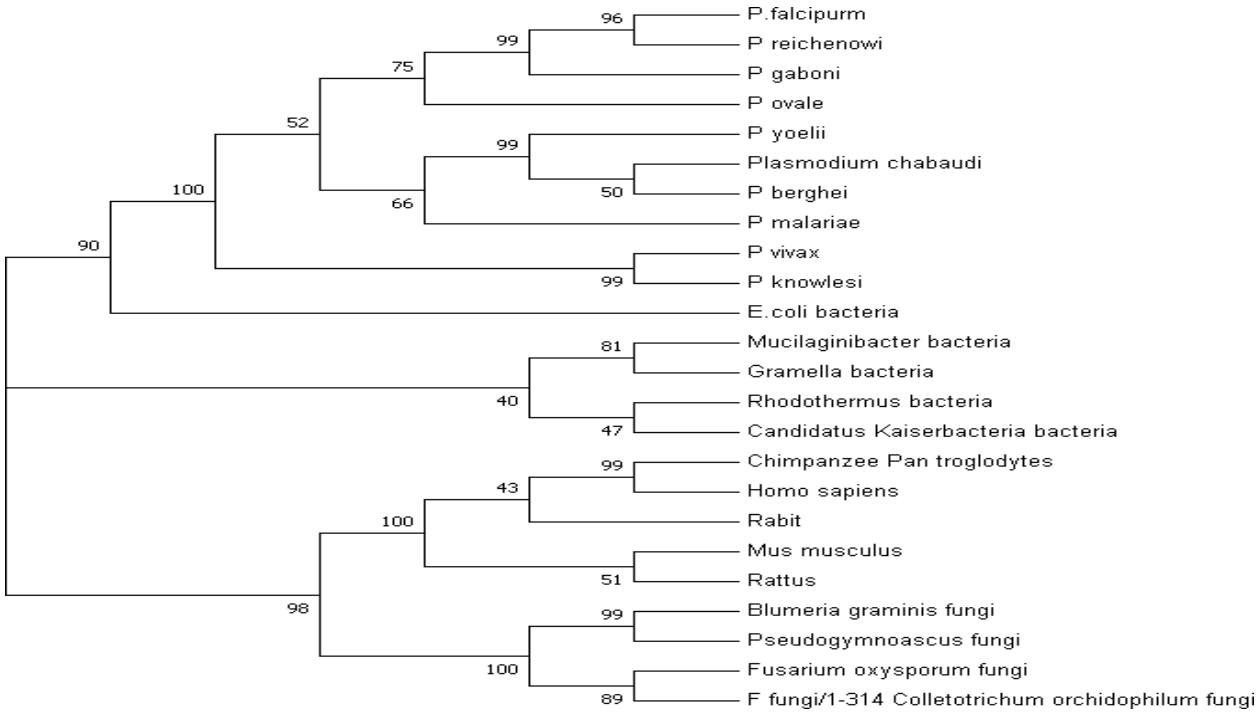


Figure A1.7: A: Original and B: Bootstrap phylogenetic trees shown in a rectangular style.

Appendix 2 – Structural analysis: Homology Modelling

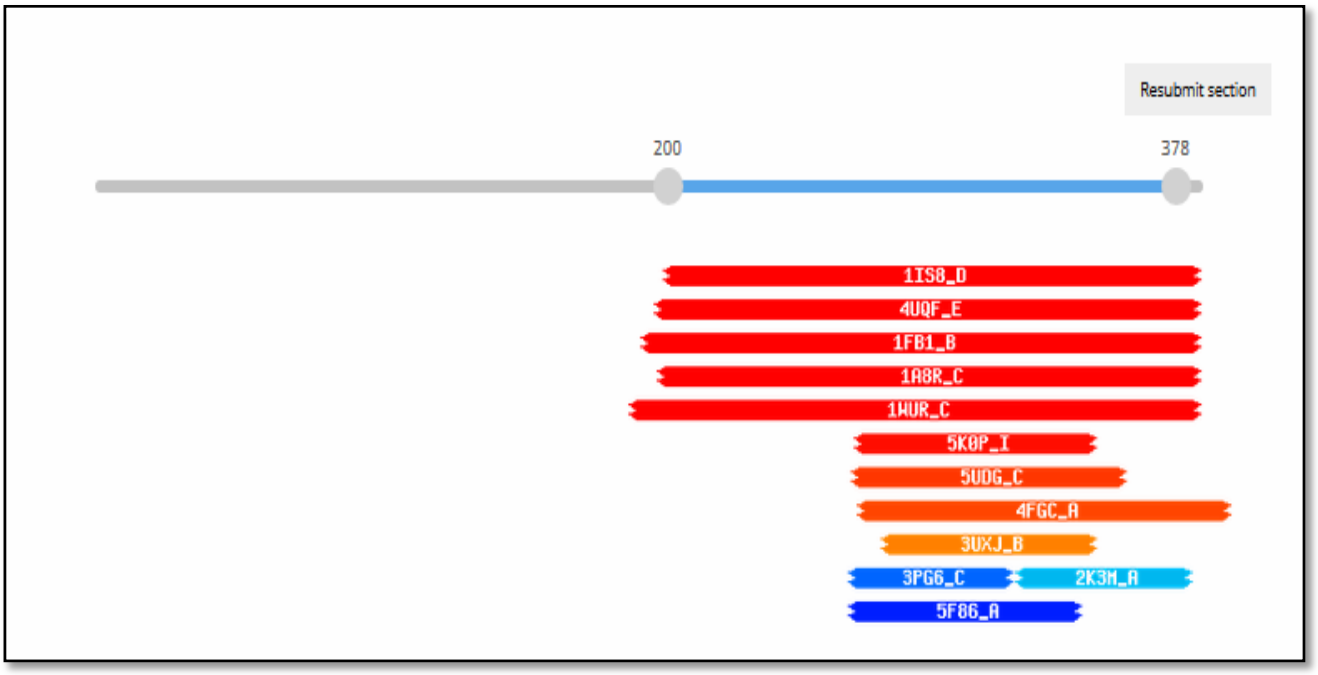
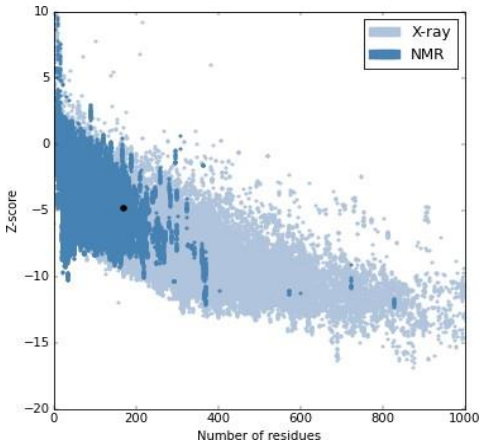


Figure A2.1: HHpred search database; colour coded bar graphs shows the positions of top matches. Red bars indicate hits with very significant match. The target sequence of *P. falciparum* GCH1 protein was covered by a number of potential templates.

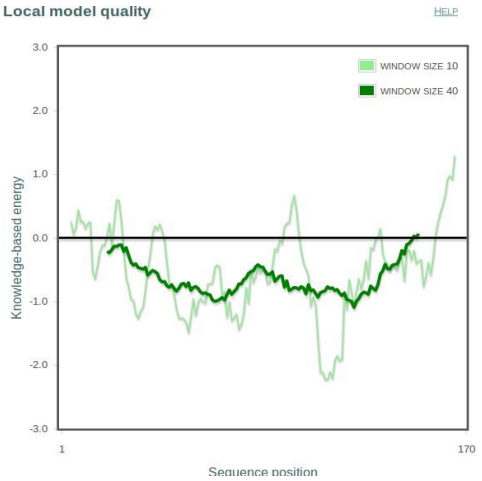
Figure A2.2: HHpred search query-template alignment: column-column residue match is coherent and the predicted template secondary structure is consistent with that of the query secondary structure.

P.falciparum

A1

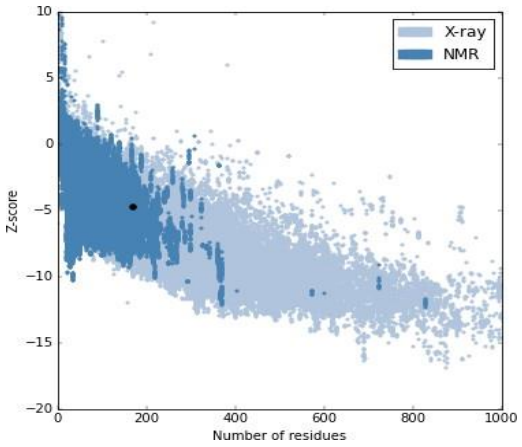


B1



P.malariae

A2



B2

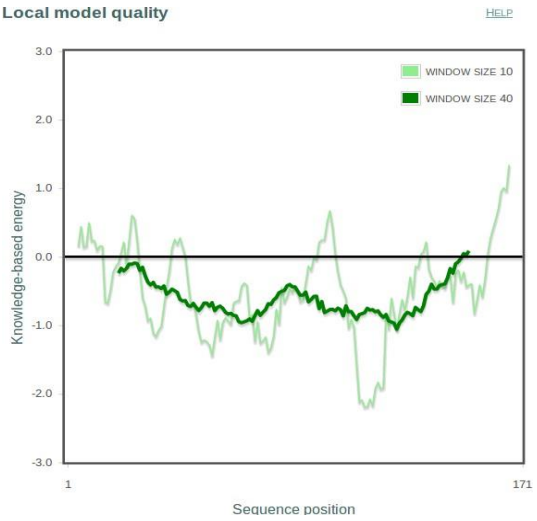


Figure A2.3: ProSA model quality assessment: Validation of top selected models of each plasmodium protein. **A1** and **A2** show *P. falciparum* and *P.malariae* models quality against all experimentally known structures in the PDB; ProSA graphically represents the Z-score for the model against all known structures in the PDB. The plots indicate that the generated models were of reasonable quality as they fall within the desired range. **B1** and **B2**: PROS local model quality energy plot showing the plotted energy scores against the amino acid positions over a 10 and a 40 residue window.

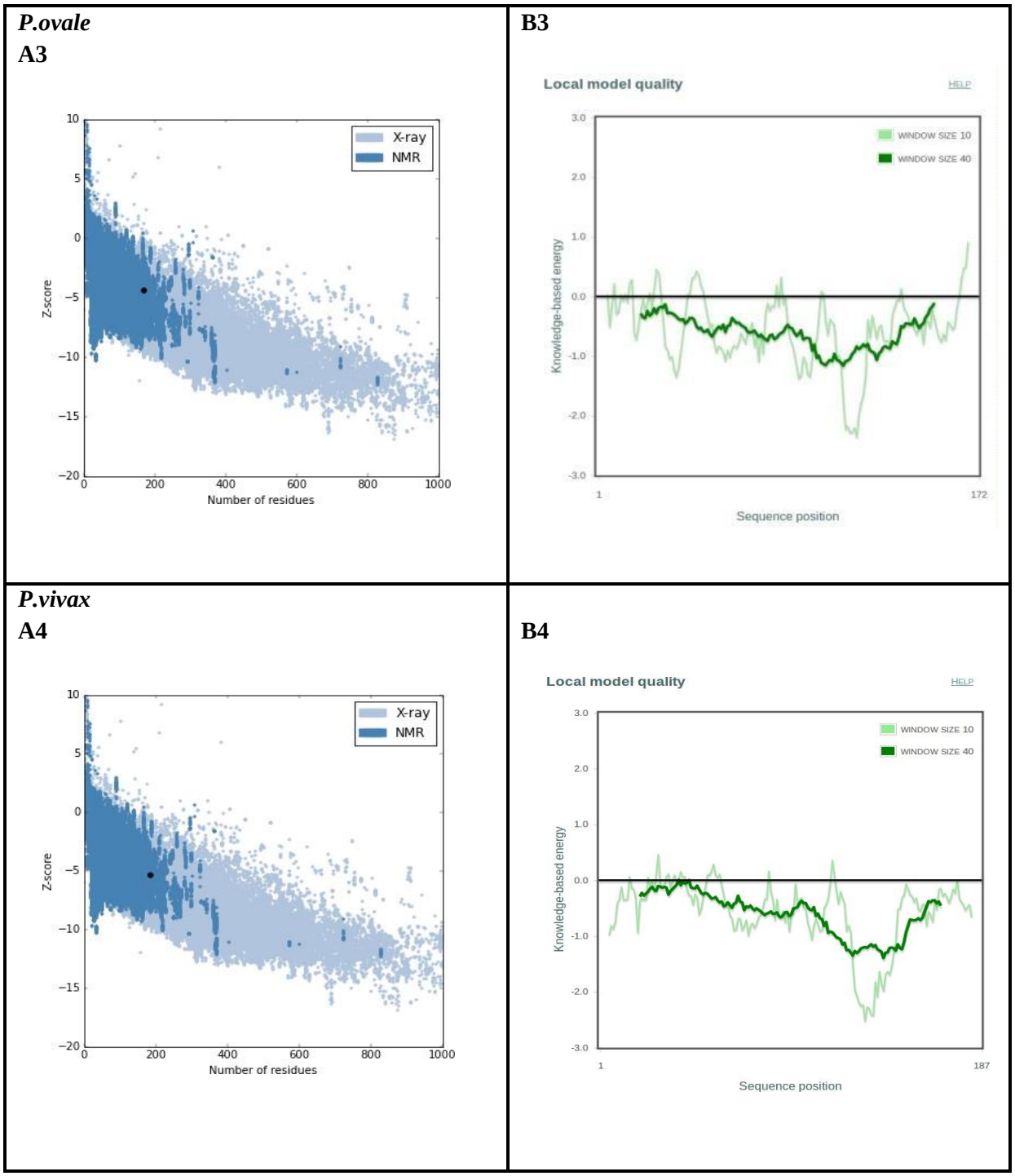


Figure A2.4: ProSA model quality assessment: Validation of top selected models of each Plasmodium protein. **A3** and **A4** show *P.ovale* and *P.vivax* models quality against all experimentally known structures in the PDB; ProSA graphically represents the Z-score for the model against all known

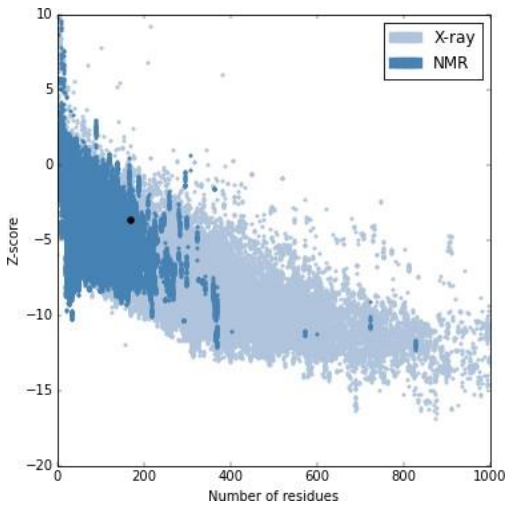
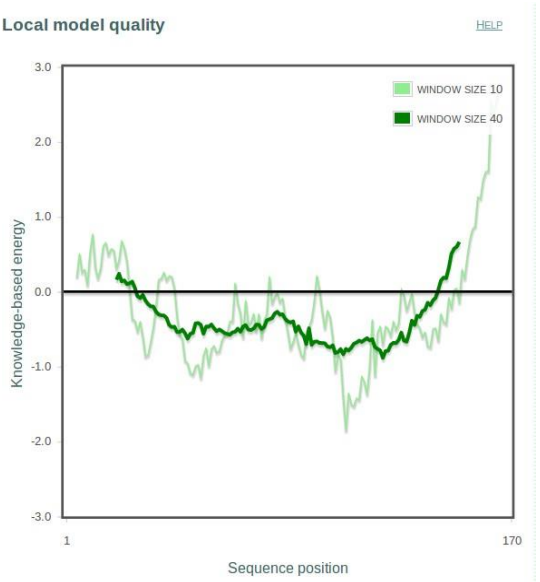
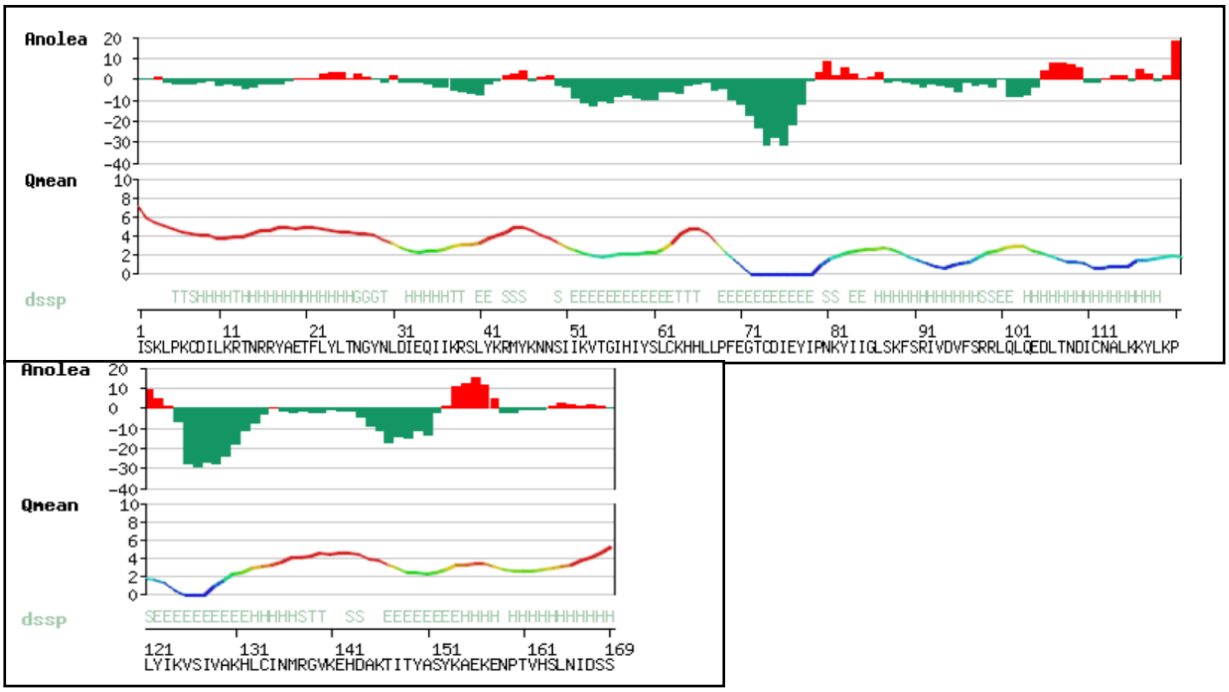
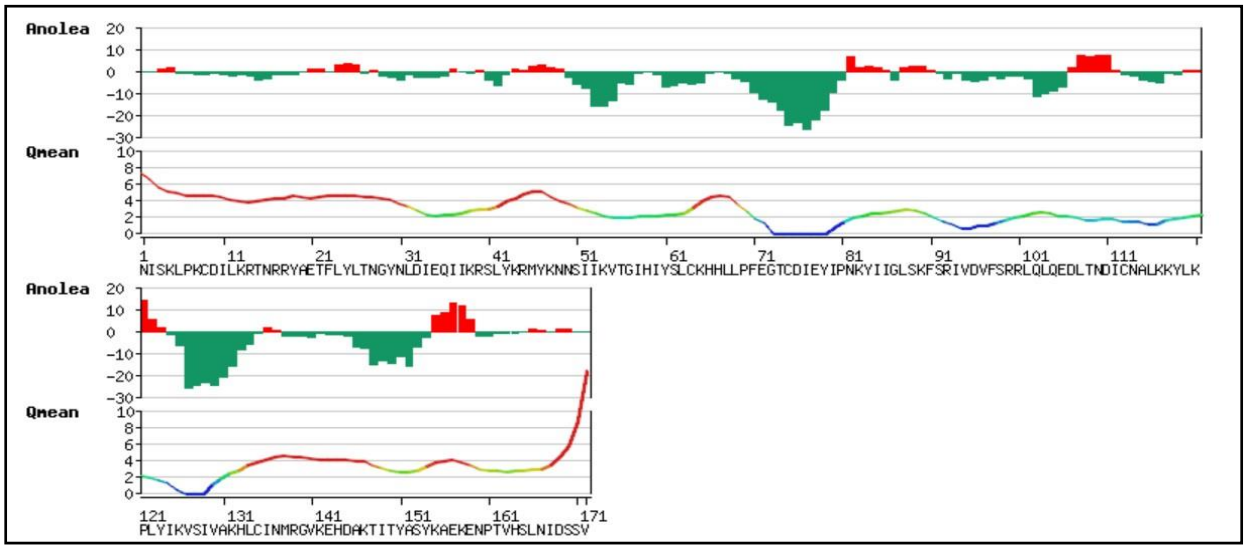
P.knowlesi	
A5	**B5**
	

Figure A2.5: ProSA model quality assessment: Validation of top selected models of each Plasmodium protein. **A5** shows *P.knowlesi* model quality against all experimentally known structures in the PDB; ProSA graphically represents the Z-score for the model against all known structures in the PDB. The plots indicate that the generated models were of reasonable quality as they fall within the desired range. **B5**: ProSA local energy plot showing the plotted energy scores against the amino acid positions over a 10 and a 40 residue window.

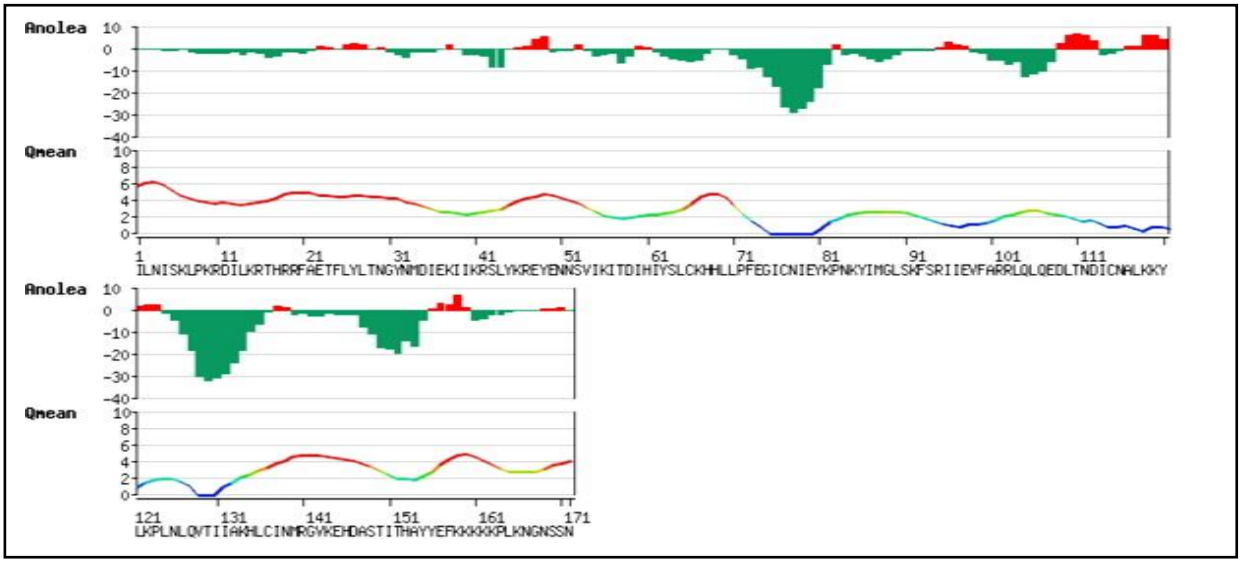
A


Anolea
Qmean
dssp
TTSHHHHTHHHHHHHHHHHHGGGT HHHHHTT EE SSS S EEEEEEEEEEETTT EEEEEEEEEEE SS EE HHHHHHHHHHHSSEE HHHHHHHHHHHHHHH
1 11 21 31 41 51 61 71 81 91 101 111
ISKLPKCDILKRTNRRYAETFLYLTNGYNLDIEQIIKRSLYKRMYKNNSIIKVTGIHIYSLCKHHLLPFEGTCDIEYIPNKYIIGLSKFSRIVDVFSRRLQLQEDLTNDICNALKKYLKP
Anolea
Qmean
dssp
SEEEEEEEEEEHHHHSTT SS EEEEEEEEHHHH HHHHHHHHHH
121 131 141 151 161 169
LYIKVSIVAKHLCINMRGVKEHDAKTITYASYKAEKENPTVHSLNIDSS

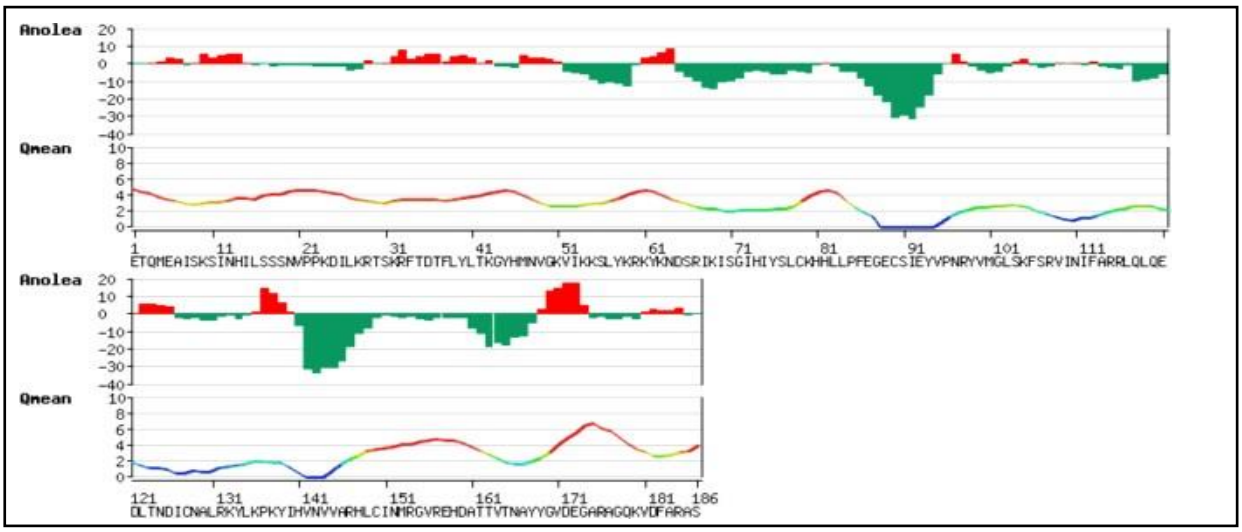
B


Anolea
Qmean
1 11 21 31 41 51 61 71 81 91 101 111
NISKLPKCDILKRTNRRYAETFLYLTNGYNLDIEQIIKRSLYKRMYKNNSIIKVTGIHIYSLCKHHLLPFEGTCDIEYIPNKYIIGLSKFSRIVDVFSRRLQLQEDLTNDICNALKKYLK
Anolea
Qmean
121 131 141 151 161 171
PLYIKVSIVAKHLCINMRGVKEHDAKTITYASYKAEKENPTVHSLNIDSSV

C


Anolea
Qmean
ILNISKLPKRDILKRTHRRFAETFLYLTNGYNMDIEKIIKRSLYKREYENNSVIKITDIHIYSLCKHHLLPFEGICNIEYKPNKYIMGLSKFSRIIEVFARRLQLQEDLTNDICNHLKKY
LKPLNLQVTIIAKHLCINMRGVKEHDASTITHAYYEFKKKKKKPLKNGNSSN

D


Anolea
Qmean
ETQMEAISKSINHILSSSNVPPKDILKRTSKRFTDTFLYLTKGYHMNVGKVIKKSLYKRKYKNDSRIKISGIHIYSLCKHHLLPFEGECSIEYVPNRYVMGLSKFSRVINIFARRLQLQE
DLTNDICNALRKYLKPKYIHVNVVARHLCINMRGVREHDATTVTNAYYGVDEGARHGQKVDFARRS

E

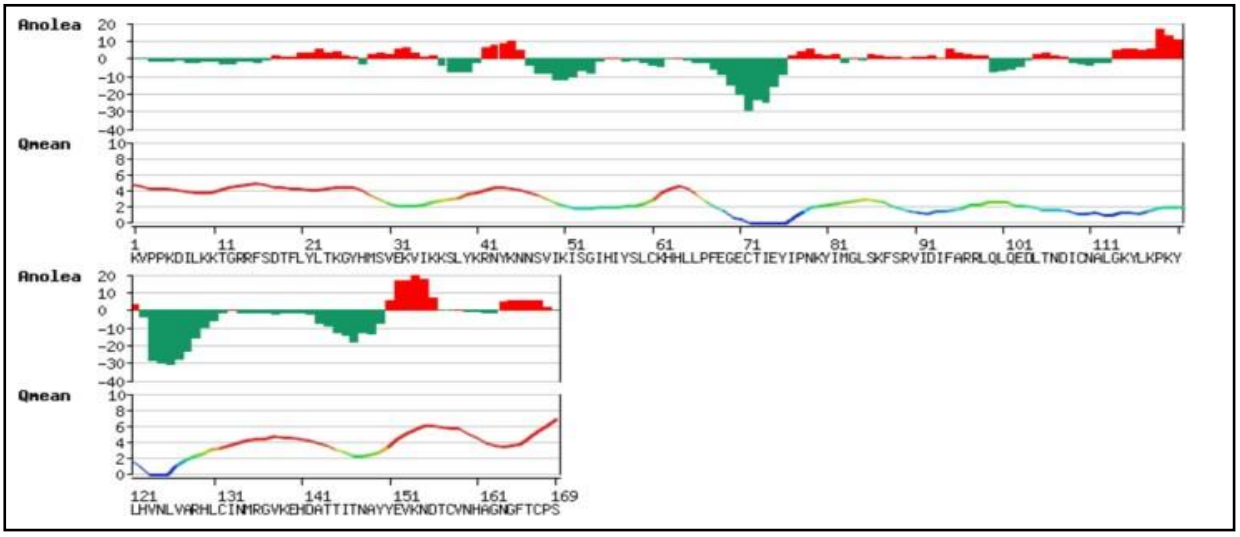
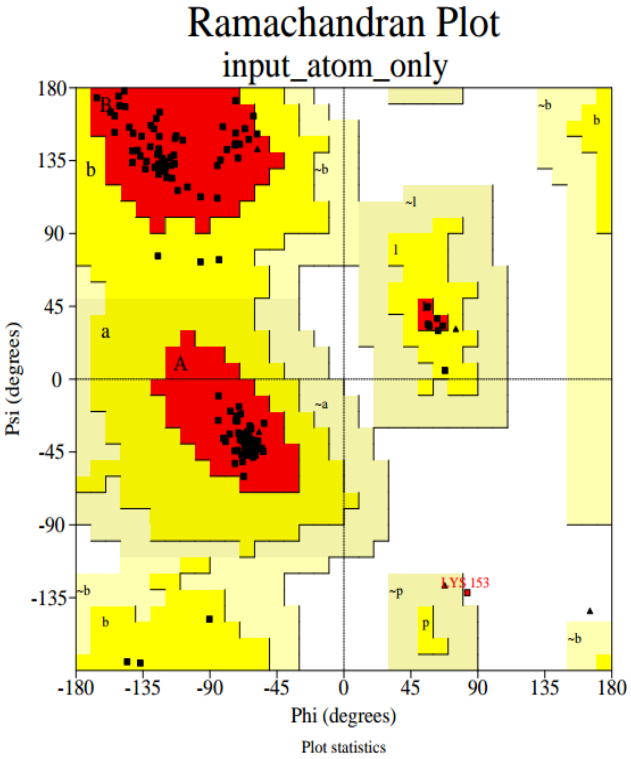


Figure A2.6: ANOLEA local quality assessment of the top selected models (chain A) **A**: *P.flaciparum*, **B**: *P.malriae*, **C**: *P.ovale*, **D**: *P.vivax* and **E**: *P.knowlesi*. Negative energy values indicates favourable energy environment (green) whereas positive values represent unfavourable energy environment (red). Overall most residues are in the favourable energy environment. Most problematic regions (in red) were identified to occur within the N, C terminal regions of the structures.

A



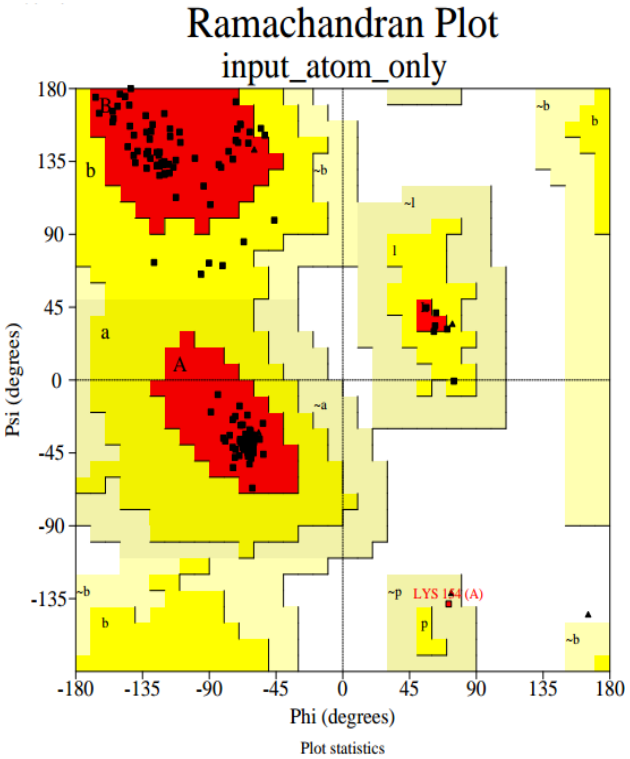
Plot statistics

Plot statistics

Residues in most favoured regions [A,B,L]	148	93.7%
Residues in additional allowed regions [a,b,l,p]	9	5.7%
Residues in generously allowed regions [~a,~b,~l,~p]	0	0.0%
Residues in disallowed regions	1	0.6%
Number of non-glycine and non-proline residues	158	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	5	
Number of proline residues	5	
Total number of residues	170	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

B



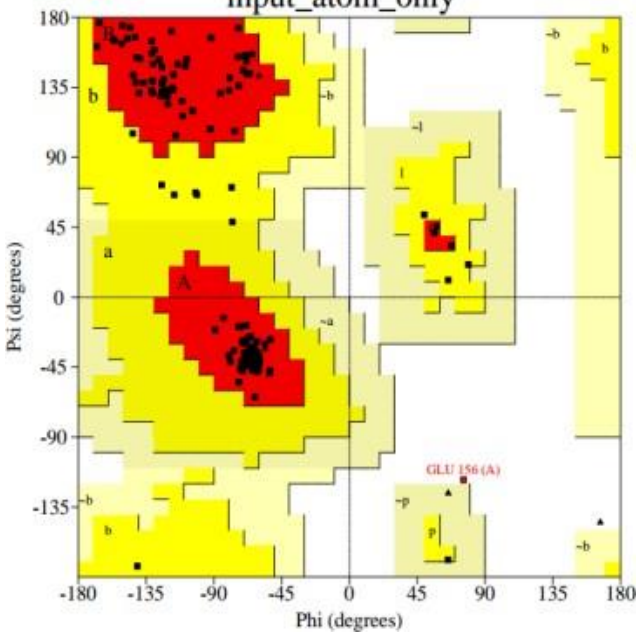
Plot statistics

Plot statistics

Residues in most favoured regions [A,B,L]	147	92.5%
Residues in additional allowed regions [a,b,l,p]	11	6.9%
Residues in generously allowed regions [~a,~b,~l,~p]	1	0.6%
Residues in disallowed regions	0	0.0%
Number of non-glycine and non-proline residues	159	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	5	
Number of proline residues	5	
Total number of residues	171	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

C

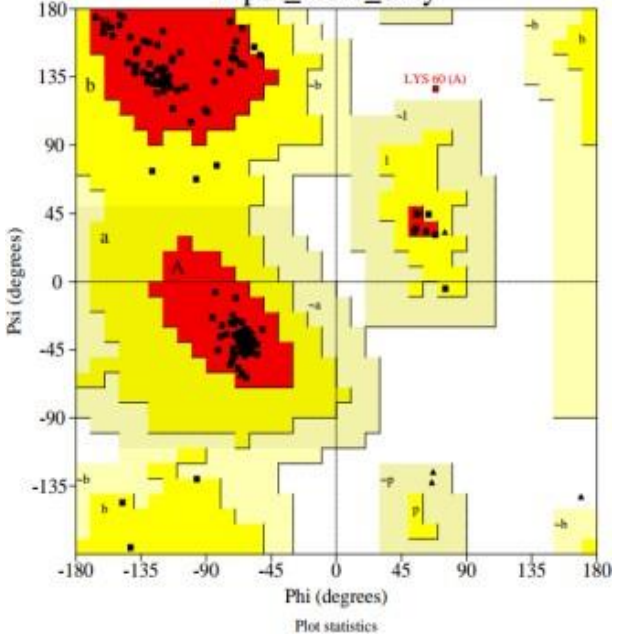


Plot statistics

Residues in most favoured regions [A,B,L]	147	91.9%
Residues in additional allowed regions [a,b,l,p]	12	7.5%
Residues in generously allowed regions [~a,~b,~l,~p]	0	0.0%
Residues in disallowed regions	1	0.6%
Number of non-glycine and non-proline residues	160	100.0%
Number of end-residues (excl. Gly and Pro)	1	
Number of glycine residues (shown as triangles)	5	
Number of proline residues	6	
Total number of residues	172	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

D

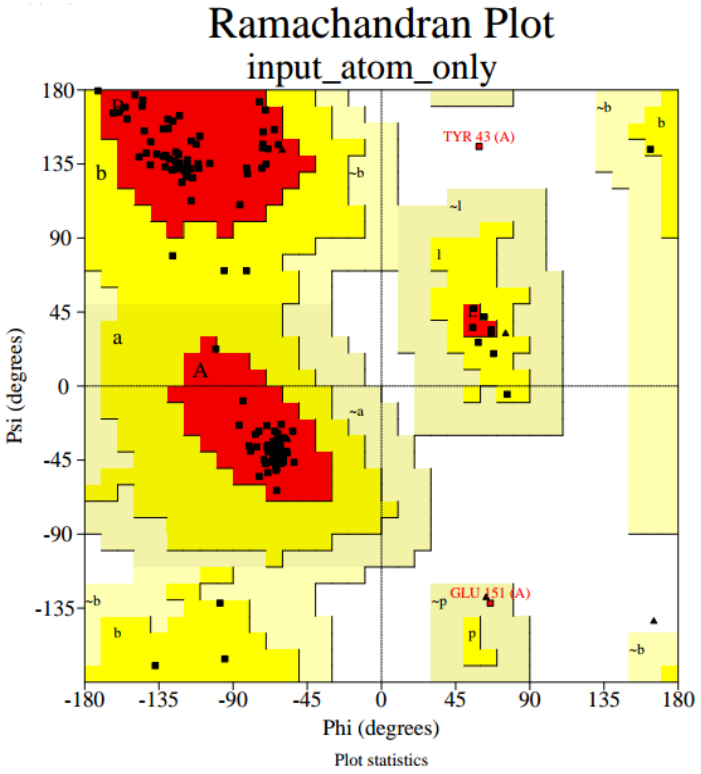


Plot statistics

Residues in most favoured regions [A,B,L]	161	94.2%
Residues in additional allowed regions [a,b,l,p]	9	5.3%
Residues in generously allowed regions [~a,~b,~l,~p]	0	0.0%
Residues in disallowed regions	1	0.6%
Number of non-glycine and non-proline residues	171	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	9	
Number of proline residues	5	
Total number of residues	187	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

Figure A2.8: PROCHECK validation Ramachandran plot of **C**: *P.ovale* and **D**: P.vivax. Red indicates the most sterically favoured region, dark-yellow indicates the additional allowed regions; light yellow shows the generously allowed regions and the disallowed regions are shown in white. Percentage values indicate the number of residues in the most favoured, additional allowed, generously allowed and disallowed regions as plotted on the Ramachandran plot.



Plot statistics

Residues in most favoured regions [A,B,L]	139	90.8%
Residues in additional allowed regions [a,b,l,p]	12	7.8%
Residues in generously allowed regions [~a,~b,~l,~p]	1	0.7%
Residues in disallowed regions	1	0.7%
Number of non-glycine and non-proline residues	153	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	9	
Number of proline residues	6	
Total number of residues	170	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

Figure A2.9: PROCHECK validation Ramachandran plot of **E**: *P.knowlesi*. Red indicates the most sterically favoured region, dark-yellow indicates the additional allowed regions; light yellow shows the generously allowed regions and the disallowed regions are shown in white. Percentage values indicate the number of residues in the most favoured, additional allowed, generously allowed and disallowed regions as plotted on the Ramachandran plot.

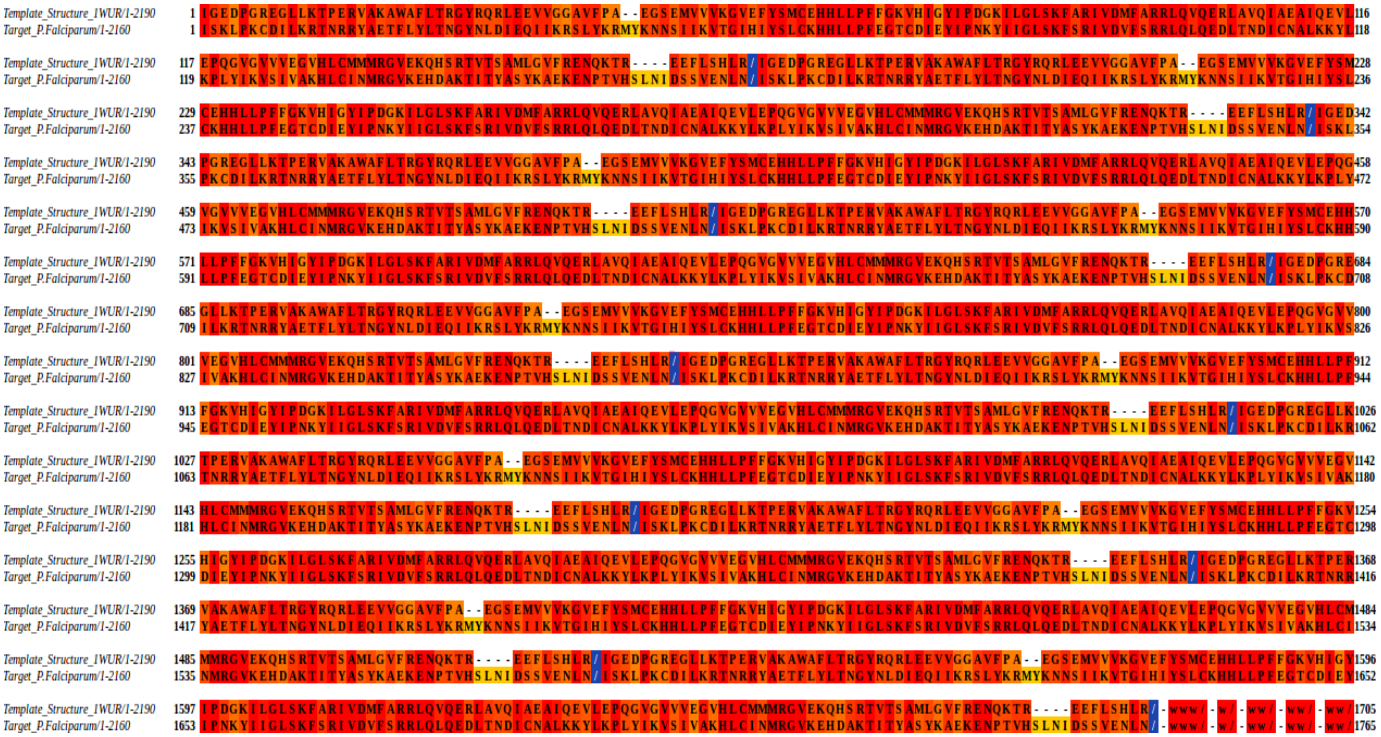
```
# Jason Vertrees <Jason-dot-Vertrees-at-schrodinger_dot_com>, 2010.
#
import pymol
from pymol import cmd

def readSymmetry(inFile, verbose=None):
  """
  This function will read "inFile" and glean the
  symmetry operations, if any, from it.

  PARAMS
    inFile
      (string) path to PDB file

    verbose
      (boolean) if verbose is not None, print more

  RETURNS
    matrix
      Array of lists.  One 16-element list per symmetry operation.  Feed this matrix
      into manualSymExp in order to make the other symmetry mates in the biological unit
  """
  # a remark-350 lines has:
  # REMARK 350 BIOMTn TAGn X Y Z Tx
  REM, TAG, BIOMT, OPNO, X, Y, Z, TX = range(8)

  thePDB = open(inFile, 'rb').readlines()
```

```
  matrices = []
  curTrans = -1

  # The transformation is,
  # output = U*input + Tx

  for l in thePDB:
    tokens = l.split()
    if len(tokens)!=8:
      continue
    if tokens[REM]=="REMARK" and tokens[TAG]=="350" and tokens[BIOMT].startswith("BIOMT"):
      if tokens[OPNO]!=curTrans:
        # new transformation matrix
        matrices.append([])

      matrices[-1].append( map( lambda s: float(s), tokens[X:]))
      curTrans = tokens[OPNO]

  if verbose!=None:
    print "Found %s symmetry operators in %s." % (len(matrices), inFile)
  return matrices


def biologicalUnit(prefix, objSel, matrices ):
  """
  Manually expands the object in "objSel" by the symmetry operations provided in
"matrices" and
  prefixes the new objects with "prefix".
```

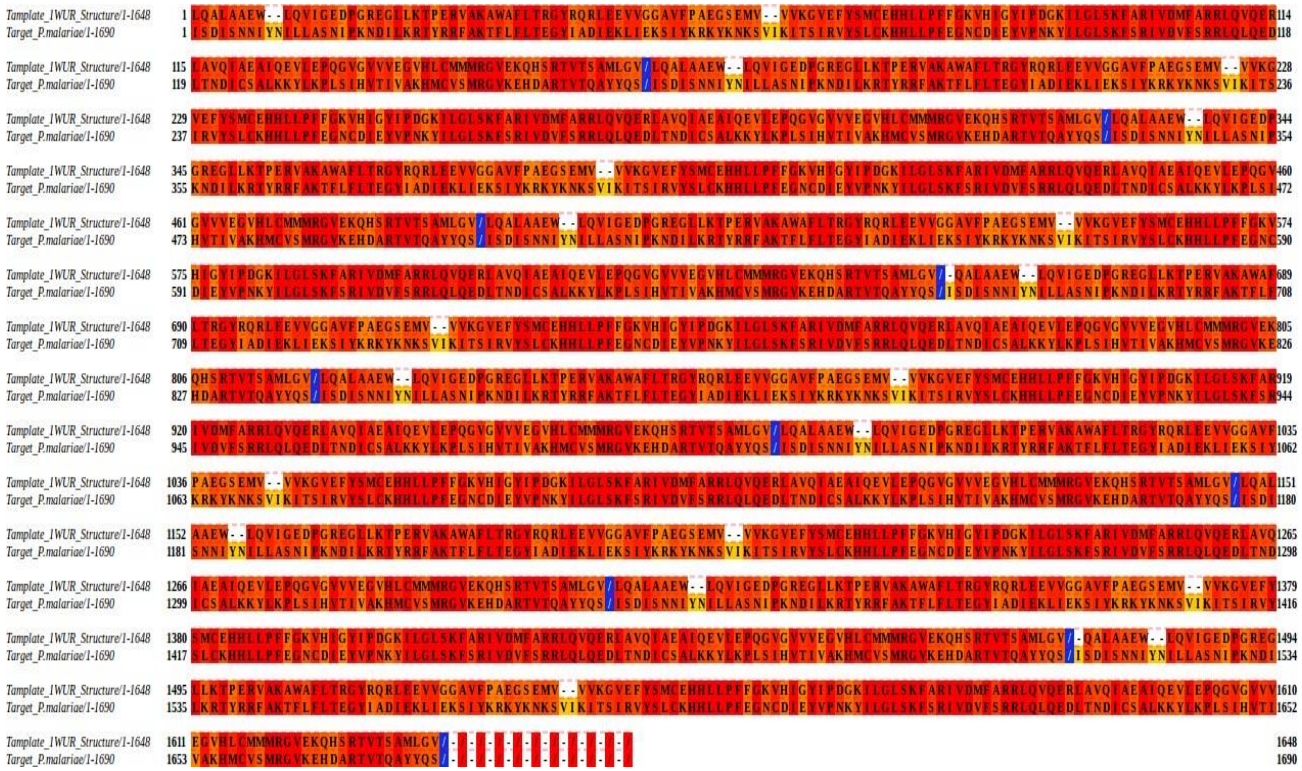
```
  PARAMS
    prefix
      (string) prefix name for new objects

    objSel
      (string) name of object to expand

    matrices
      (list of 16-element lists) array of matrices from readSymmetry

    RETUNRS
      None

    SIDE EFFECTS
      Creates N new obects each rotated and translated according to the symmetry
operators, where N
      equals len(matrices).
  """
  for m in matrices:
    n = cmd.get_unused_name(prefix)
    cmd.create(n, objSel)
    s1 = "%s + (x*%s + y*%s + z*%s)" % (m[0][3], m[0][0], m[0][1], m[0][2])
    s2 = "%s + (x*%s + y*%s + z*%s)" % (m[1][3], m[1][0], m[1][1], m[1][2])
    s3 = "%s + (x*%s + y*%s + z*%s)" % (m[2][3], m[2][0], m[2][1], m[2][2])
    cmd.alter_state(1, n, "(x,y,z) = (%s, %s, %s)" % (s1, s2, s3) )
```

Script A2.1: Pymol script used to build the protein biological unit from the crystal structure of *Thermus thermophilus*.

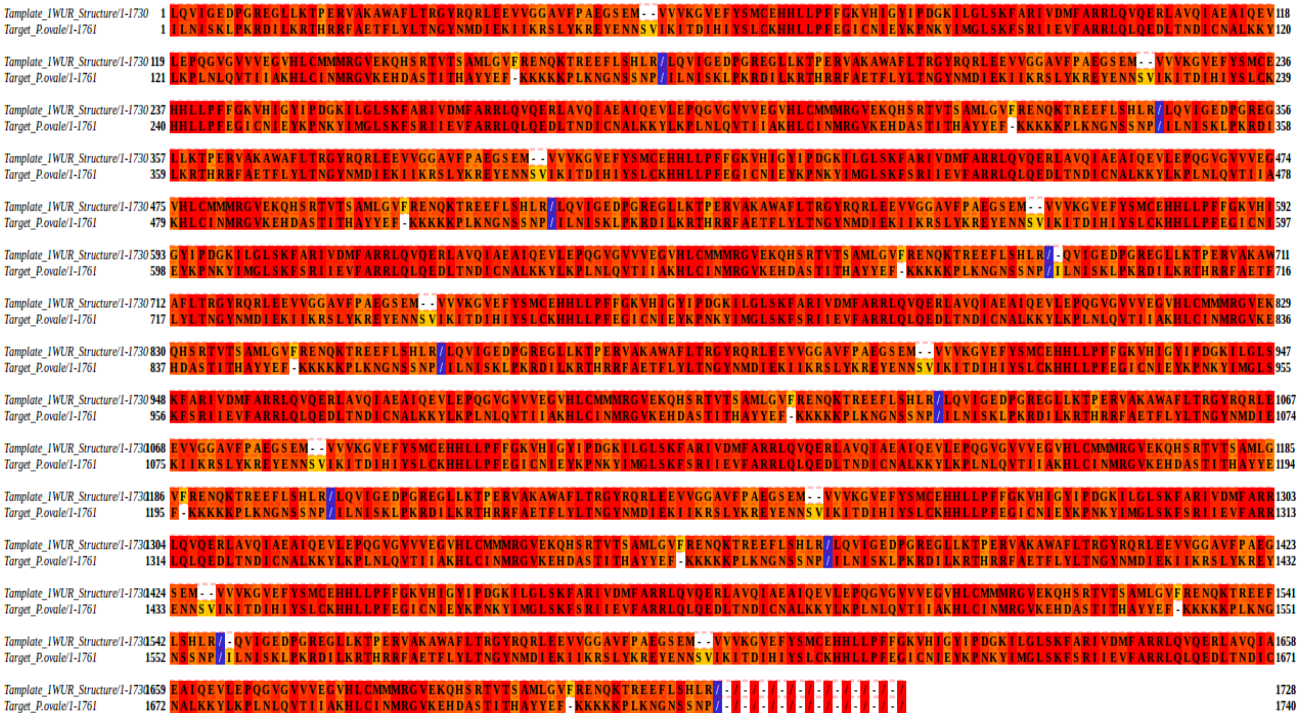
(A) P. falciparum



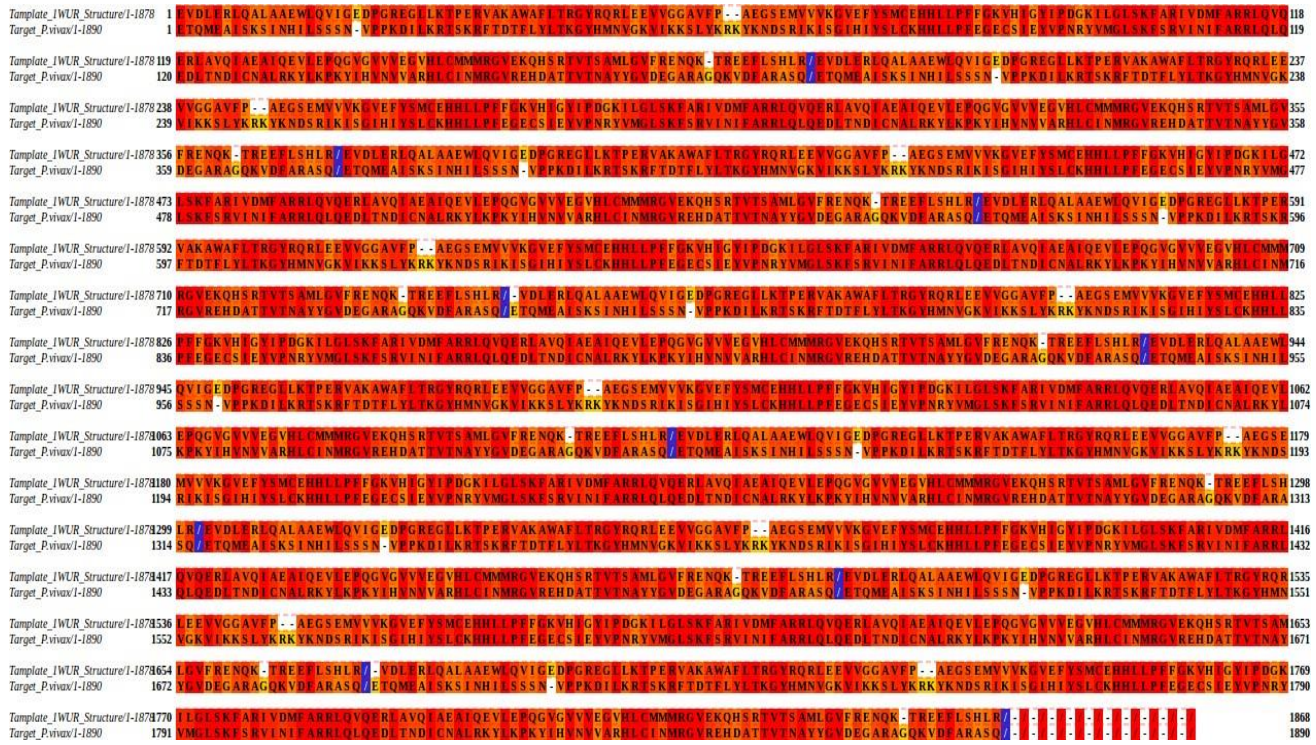
(B) P.malariae



(C) P.ovale



(D) P.vivax



(E) P.knowlesi

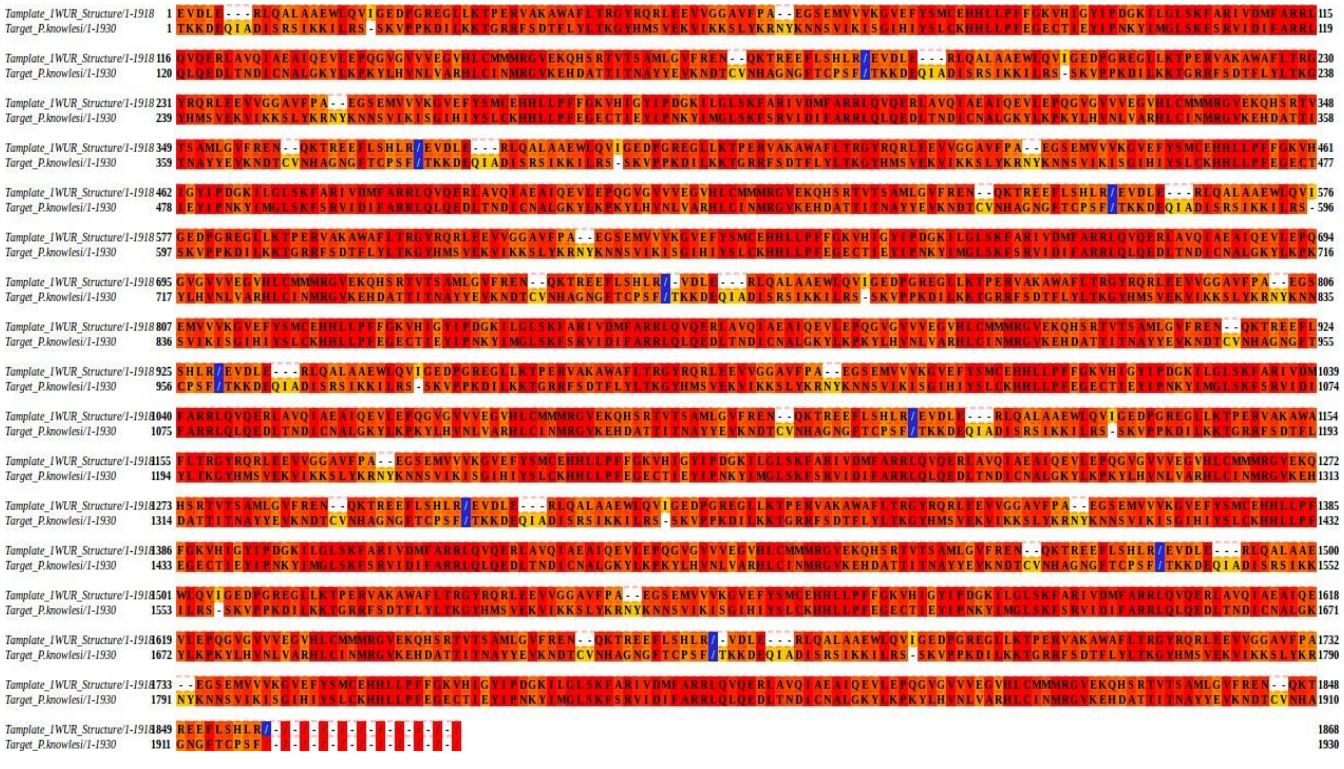


Figure A2.10: A, B, C, D and E: Sequence alignment files prepared for each of the plasmodium proteins against the template sequence (Top). Colour intensity varies with residue-residue similarity match. Dark red shows identical residues whereas white indicate gaps region.

```
>P1;1wur_Biological_B.pdb
structureX:1wur_Biological_B.pdb: 1 :A:2190:J:::-1.00:-1.00
IGEDPGREGLLKTPERVAKAWAFLT
RGYRQRLEEVVGGAVFPA--EGSEMVVVKGVEFYSMCEHHLLPFFGKVHIGYIPDGKILG
LSKFARIVDMFARRLQVQERLAVQIAEAIQEVLEPQGVGVVVEGVHLCMMMRGVEKQHSR
TVTSAMLGVFRENQKTR----EEFLSHLR
/
IGEDPGREGLLKTPERVAKAWAFLT
RGYRQRLEEVVGGAVFPA--EGSEMVVVKGVEFYSMCEHHLLPFFGKVHIGYIPDGKILG
LSKFARIVDMFARRLQVQERLAVQIAEAIQEVLEPQGVGVVVEGVHLCMMMRGVEKQHSR
TVTSAMLGVFRENQKTR----EEFLSHLR
/
IGEDPGREGLLKTPERVAKAWAFLT
RGYRQRLEEVVGGAVFPA--EGSEMVVVKGVEFYSMCEHHLLPFFGKVHIGYIPDGKILG
LSKFARIVDMFARRLQVQERLAVQIAEAIQEVLEPQGVGVVVEGVHLCMMMRGVEKQHSR
TVTSAMLGVFRENQKTR----EEFLSHLR
/
IGEDPGREGLLKTPERVAKAWAFLT
RGYRQRLEEVVGGAVFPA--EGSEMVVVKGVEFYSMCEHHLLPFFGKVHIGYIPDGKILG
LSKFARIVDMFARRLQVQERLAVQIAEAIQEVLEPQGVGVVVEGVHLCMMMRGVEKQHSR
TVTSAMLGVFRENQKTR----EEFLSHLR
/
IGEDPGREGLLKTPERVAKAWAFLT
RGYRQRLEEVVGGAVFPA--EGSEMVVVKGVEFYSMCEHHLLPFFGKVHIGYIPDGKILG
LSKFARIVDMFARRLQVQERLAVQIAEAIQEVLEPQGVGVVVEGVHLCMMMRGVEKQHSR
TVTSAMLGVFRENQKTR----EEFLSHLR
/
```

```
IGEDPGREGLLKTPERVAKAWAFLT
RGYRQRLEEVVGGAVFPA--EGSEMVVVKGVEFYSMCEHHLLPFFGKVHIGYIPDGKILG
LSKFARIVDMFARRLQVQERLAVQIAEAIQEVLEPQGVGVVVEGVHLCMMMRGVEKQHSR
TVTSAMLGVFRENQKTR----EEFLSHLR
/
IGEDPGREGLLKTPERVAKAWAFLT
RGYRQRLEEVVGGAVFPA--EGSEMVVVKGVEFYSMCEHHLLPFFGKVHIGYIPDGKILG
LSKFARIVDMFARRLQVQERLAVQIAEAIQEVLEPQGVGVVVEGVHLCMMMRGVEKQHSR
TVTSAMLGVFRENQKTR----EEFLSHLR
/
IGEDPGREGLLKTPERVAKAWAFLT
RGYRQRLEEVVGGAVFPA--EGSEMVVVKGVEFYSMCEHHLLPFFGKVHIGYIPDGKILG
LSKFARIVDMFARRLQVQERLAVQIAEAIQEVLEPQGVGVVVEGVHLCMMMRGVEKQHSR
TVTSAMLGVFRENQKTR----EEFLSHLR
/
IGEDPGREGLLKTPERVAKAWAFLT
RGYRQRLEEVVGGAVFPA--EGSEMVVVKGVEFYSMCEHHLLPFFGKVHIGYIPDGKILG
LSKFARIVDMFARRLQVQERLAVQIAEAIQEVLEPQGVGVVVEGVHLCMMMRGVEKQHSR
TVTSAMLGVFRENQKTR----EEFLSHLR
/
IGEDPGREGLLKTPERVAKAWAFLT
RGYRQRLEEVVGGAVFPA--EGSEMVVVKGVEFYSMCEHHLLPFFGKVHIGYIPDGKILG
LSKFARIVDMFARRLQVQERLAVQIAEAIQEVLEPQGVGVVVEGVHLCMMMRGVEKQHSR
TVTSAMLGVFRENQKTR----EEFLSHLR/.www/.w/.ww/.ww/.ww/.www/.w/.ww/.ww/.ww/*
```

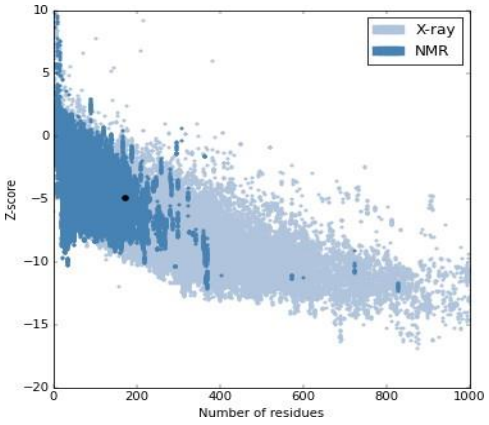
```
>P1;PFALCIPARUM.cha
sequence:P.cha:1 :A:+2160 :J:M18AAP:Plasmodium: :
ISKLPKCDILKRTNRRYAETFLYLT
NGYNLDIEQIIKRSLYKRMYKNNSIIKVTGIHIYSLCKHHLLPFEGTCDIEYIPNKYIIG
LSKFSRIVDVFSRRLQLQEDLTNDICNALKKYLKPLYIKVSIVAKHLCINMRGVKEHDAK
TITYASYKAEKENPTVHSLNIDSSVENLN
/
ISKLPKCDILKRTNRRYAETFLYLT
NGYNLDIEQIIKRSLYKRMYKNNSIIKVTGIHIYSLCKHHLLPFEGTCDIEYIPNKYIIG
LSKFSRIVDVFSRRLQLQEDLTNDICNALKKYLKPLYIKVSIVAKHLCINMRGVKEHDAK
TITYASYKAEKENPTVHSLNIDSSVENLN
/
ISKLPKCDILKRTNRRYAETFLYLT
NGYNLDIEQIIKRSLYKRMYKNNSIIKVTGIHIYSLCKHHLLPFEGTCDIEYIPNKYIIG
LSKFSRIVDVFSRRLQLQEDLTNDICNALKKYLKPLYIKVSIVAKHLCINMRGVKEHDAK
TITYASYKAEKENPTVHSLNIDSSVENLN
/
ISKLPKCDILKRTNRRYAETFLYLT
NGYNLDIEQIIKRSLYKRMYKNNSIIKVTGIHIYSLCKHHLLPFEGTCDIEYIPNKYIIG
LSKFSRIVDVFSRRLQLQEDLTNDICNALKKYLKPLYIKVSIVAKHLCINMRGVKEHDAK
TITYASYKAEKENPTVHSLNIDSSVENLN
/
ISKLPKCDILKRTNRRYAETFLYLT
NGYNLDIEQIIKRSLYKRMYKNNSIIKVTGIHIYSLCKHHLLPFEGTCDIEYIPNKYIIG
LSKFSRIVDVFSRRLQLQEDLTNDICNALKKYLKPLYIKVSIVAKHLCINMRGVKEHDAK
TITYASYKAEKENPTVHSLNIDSSVENLN
/
```

```
ISKLPKCDILKRTNRRYAETFLYLT
NGYNLDIEQIIKRSLYKRMYKNNSIIKVTGIHIYSLCKHHLLPFEGTCDIEYIPNKYIIG
LSKFSRIVDVFSRRLQLQEDLTNDICNALKKYLKPLYIKVSIVAKHLCINMRGVKEHDAK
TITYASYKAEKENPTVHSLNIDSSVENLN
/
ISKLPKCDILKRTNRRYAETFLYLT
NGYNLDIEQIIKRSLYKRMYKNNSIIKVTGIHIYSLCKHHLLPFEGTCDIEYIPNKYIIG
LSKFSRIVDVFSRRLQLQEDLTNDICNALKKYLKPLYIKVSIVAKHLCINMRGVKEHDAK
TITYASYKAEKENPTVHSLNIDSSVENLN
/
ISKLPKCDILKRTNRRYAETFLYLT
NGYNLDIEQIIKRSLYKRMYKNNSIIKVTGIHIYSLCKHHLLPFEGTCDIEYIPNKYIIG
LSKFSRIVDVFSRRLQLQEDLTNDICNALKKYLKPLYIKVSIVAKHLCINMRGVKEHDAK
TITYASYKAEKENPTVHSLNIDSSVENLN
/
ISKLPKCDILKRTNRRYAETFLYLT
NGYNLDIEQIIKRSLYKRMYKNNSIIKVTGIHIYSLCKHHLLPFEGTCDIEYIPNKYIIG
LSKFSRIVDVFSRRLQLQEDLTNDICNALKKYLKPLYIKVSIVAKHLCINMRGVKEHDAK
TITYASYKAEKENPTVHSLNIDSSVENLN
/
ISKLPKCDILKRTNRRYAETFLYLT
NGYNLDIEQIIKRSLYKRMYKNNSIIKVTGIHIYSLCKHHLLPFEGTCDIEYIPNKYIIG
LSKFSRIVDVFSRRLQLQEDLTNDICNALKKYLKPLYIKVSIVAKHLCINMRGVKEHDAK
TITYASYKAEKENPTVHSLNIDSSVENLN/.www/.w/.ww/.ww/.ww/.www/.w/.ww/.ww/.ww/*
```

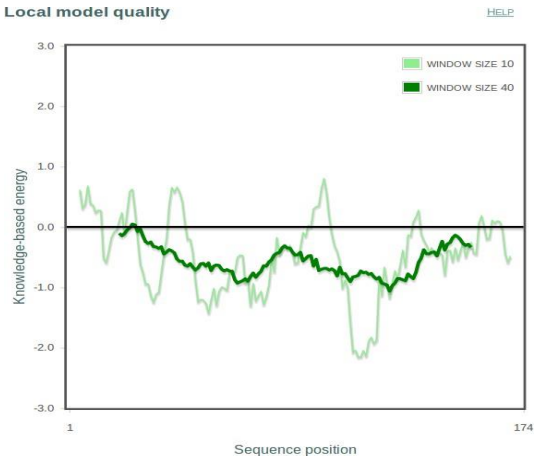
Figure A2.11: Example of the alignment PIR file used in modelling the *P. falciparum* 3D structure.

P.falciparum

A1

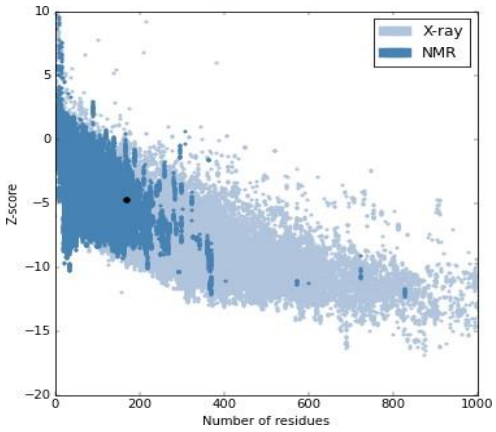


B1



P.malriae

A2



B2

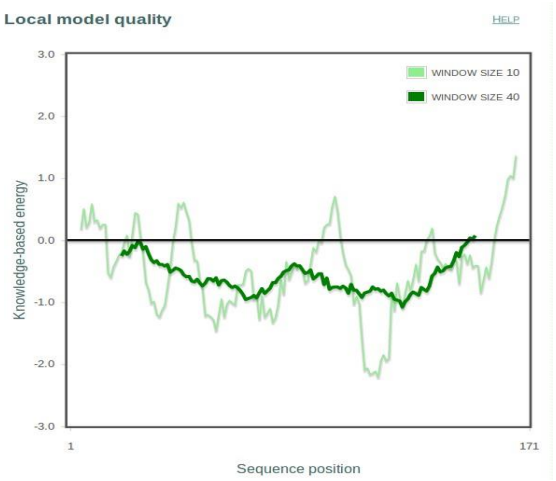
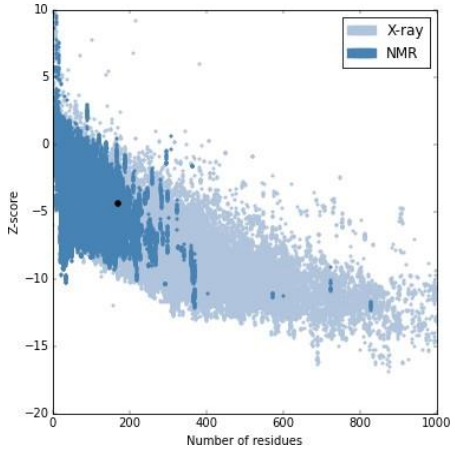


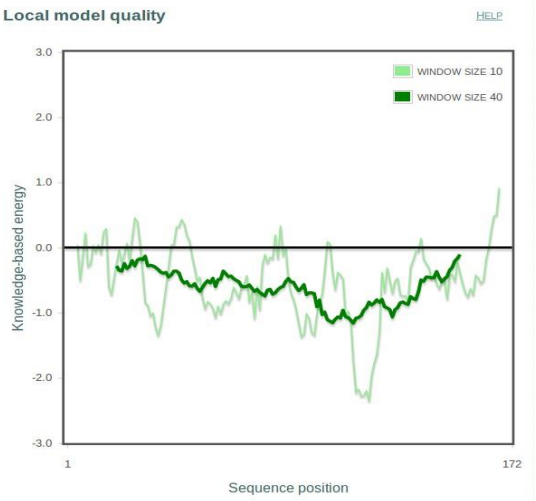
Figure A2. 12: ProSA model quality assessment of GCH1 biological unit assembly top selected models from each Plasmodium protein. **A1** and **A2** shows the models global quality compared against all experimentally determined structures in the PDB; The plots indicate that the generated models were of reasonable quality as they fall within the desired range. **B1** and **B2**: ProSA local model quality energy plot.

P.ovale

A3

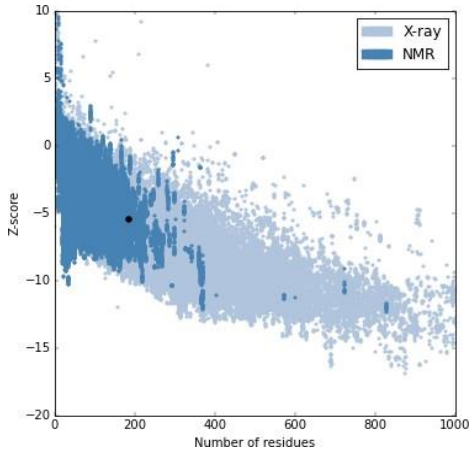


B3

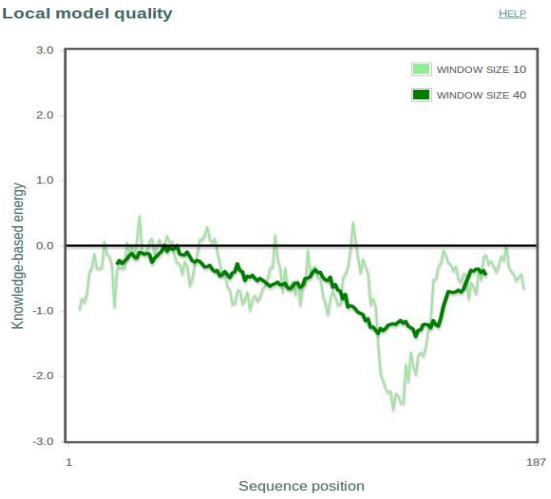


P.vivax

A4



B4



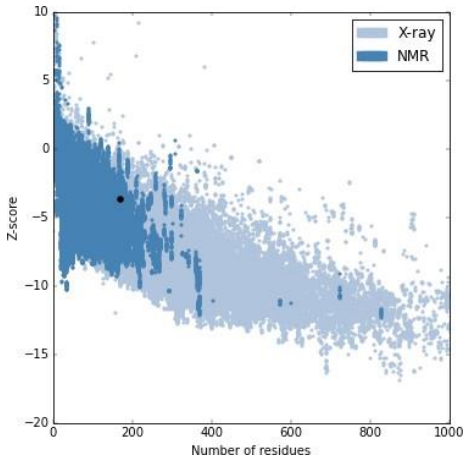
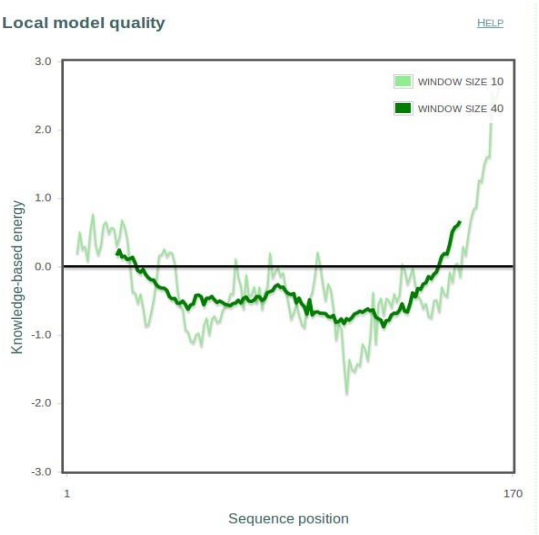
P.knowlesi	
A5	**B5**
	

Figure A2. 13: ProSA model quality assessment of GCH1 biological unit assembly top selected models from each Plasmodium protein. **A3**, **A4** and **A5** shows the models global quality compared against all experimentally determined structures in the PDB; The plots indicate that the generated models were of reasonable quality as they fall within the desired range. **B3**, **B4** and **B5**: ProSA local quality energy plot showing the plotted energy scores against the amino acid positions over a 10 and a 40 residue window.

Appendix 3 – Molecular docking

```
import os
Ligand_files = os.listdir('/home/afrah/Desktop/WORK/DOCKING/Ligand')
for ligand in Ligand_files:
        if ligand.endswith(".pdb"):
                ligand_name = ligand[:-4]
                os.system("prepare_ligand4.py -l "+"../Ligand/"+ligand_name
+".pdb")
```

Script A3.2: Photo-snap of the script used to prepare all the ligands and convert them into the rigid (pdbqt) format.

```
import os
Ligand_files = os.listdir('../Ligand')
PDB_files = ['P.cha.falciparum.pdbqt']
for ligand in Ligand_files:
        if ".pdb" in ligand:
                ligand_name = ligand[:-6]
                for PDB in PDB_files:
                        vina_name = PDB+"_"+ligand_name+".vina"
                        with open("/home/akhairallah/lustre/afrah/P.falciparum/
Vina/"+vina_name, "w") as vw:
                                vw.writelines("receptor=/home/akhairallah/lustre/
afrah/P.falciparum/Target/"+ PDB + "\n")
                                vw.writelines("ligand=/home/akhairallah/lustre/afrah/
P.falciparum/Ligand/"+ligand_name+".pdbqt" + "\n")
                                vw.writelines("out=/home/akhairallah/lustre/afrah/
P.falciparum/Out/"+vina_name+"all.pdbqt" + "\n")
                                vw.writelines("log=/home/akhairallah/lustre/afrah/
P.falciparum/Log/"+vina_name+"all.log"+vina_name+"all.log" + "\n")
                                vw.writelines("center_x=29.9183121934" + "\n")
                                vw.writelines("center_y=-41.232414506" + "\n")
                                vw.writelines("center_z=23.1270550105" + "\n")
                                vw.writelines("size_x=48.75" + "\n")
                                vw.writelines("size_y=48.75" + "\n")
                                vw.writelines("size_z=84.75" + "\n")
                                vw.writelines("energy_range=4" + "\n")
                                vw.writelines("cpu=8" + "\n")
                                vw.writelines("exhaustiveness=192" + "\n")
```

Script A3. 3: Photo-snap of the script used for docking simulation of all ligands against the receptor.

Table A3.1: Best hits against the *P. falciparum* GCH1 protein identified from SANCDB with low interaction energy and high binding affinity towards the *P. falciparum* protein. The identified compounds (Total of 80) were all bound to the *P. falciparum* active site pocket.
ΔG shows the difference of compounds binding free energy between the *P. falciparum* and human protein. Lower ΔG indicates more selective binding towards the Plasmodium GCH1 protein.

LIGAND	FALCIPARUM	HUMAN	ΔG
SANC00317	-10.1	-7.5	-3
SANC00335	-10	-7.4	-3
SANC00368	-8.5	-6.1	-2
SANC00106	-9.6	-7.3	-2
SANC00103	-9.4	-7.3	-2
SANC00286	-9.5	-7.4	-2
SANC00101	-8.9	-7	-2
SANC00315	-9.6	-7.7	-2
SANC00211	-8.9	-7.1	-2
SANC00312	-9.6	-7.9	-2
SANC00645	-8.7	-7	-2
SANC00262	-8.3	-6.7	-2
SANC00611	-8.3	-6.7	-2
SANC00242	-8.6	-7	-2
SANC00627	-8.1	-6.5	-2
SANC00656	-7.3	-5.7	-2
SANC00727	-8.1	-6.5	-2
SANC00142	-6.7	-5.2	-2
SANC00439	-7.7	-6.2	-2
SANC00628	-7.9	-6.4	-2
SANC00643	-8.5	-7	-2
SANC00642	-9.2	-7.7	-2
SANC00209	-8.9	-7.5	-1
SANC00610	-8.4	-7	-1
SANC00523	-8.7	-7.3	-1
SANC00414	-9.8	-8.5	-1
SANC00605	-7.1	-5.8	-1
SANC00105	-9.2	-7.9	-1
SANC00609	-9.2	-7.9	-1
SANC00522	-8	-6.8	-1
SANC00562	-8.9	-7.7	-1
SANC00564	-8	-6.8	-1
SANC00629	-7.7	-6.5	-1
SANC00302	-6.1	-4.9	-1
SANC00461	-6.6	-5.4	-1
SANC00464	-8.6	-7.4	-1
SANC00521	-8.1	-6.9	-1
SANC00678	-8.1	-6.9	-1
SANC00320	-8.8	-7.7	-1
SANC00472	-7.9	-6.8	-1
SANC00109	-7.1	-6	-1
SANC00206	-6	-4.9	-1
SANC00229	-8.5	-7.4	-1

SANC00475	-8.1	-7	-1
SANC00520	-9.1	-8	-1
SANC00513	-8.3	-7.3	-1
SANC00216	-6.1	-5.1	-1
SANC00304	-6.5	-5.5	-1
SANC00437	-7.8	-6.8	-1
SANC00606	-7.3	-6.3	-1
SANC00655	-8.6	-7.6	-1
SANC00677	-7.8	-6.8	-1
SANC00129	-6.9	-6	-1
SANC00176	-6.4	-5.5	-1
SANC00210	-7.9	-7	-1
SANC00249	-8.4	-7.5	-1
SANC00259	-5.9	-5	-1
SANC00267	-7.4	-6.5	-1
SANC00675	-7.9	-7	-1
SANC00115	-6.6	-5.7	-1
SANC00128	-7.1	-6.2	-1
SANC00130	-6.8	-5.9	-1
SANC00303	-6	-5.2	-1
SANC00426	-7.2	-6.4	-1
SANC00462	-5.8	-5	-1
SANC00469	-6.6	-5.8	-1
SANC00679	-7.7	-6.9	-1
SANC00240	-8.2	-7.4	-1
SANC00638	-7.9	-7.2	-1
SANC00337	-7.6	-6.9	-1
SANC00608	-7.6	-6.9	-1
SANC00634	-7.9	-7.3	-1
SANC00266	-7.3	-6.7	-1
SANC00471	-7.6	-7	-1
SANC00683	-7.6	-7	-1
SANC00202	-5.6	-5.1	-1
SANC00256	-6.8	-6.3	-1
SANC00332	-7.5	-7	-1
SANC00353	-7.6	-7.1	-1
SANC00532	-7.1	-6.6	-1
SANC00721	-7.9	-7.4	-1
SANC00329	-7.8	-7.4	-0
SANC00231	-7.7	-7.4	-0
SANC00468	-6.6	-6.3	-0
SANC00630	-6.2	-5.9	-0
SANC00467	-6.5	-6.3	-0

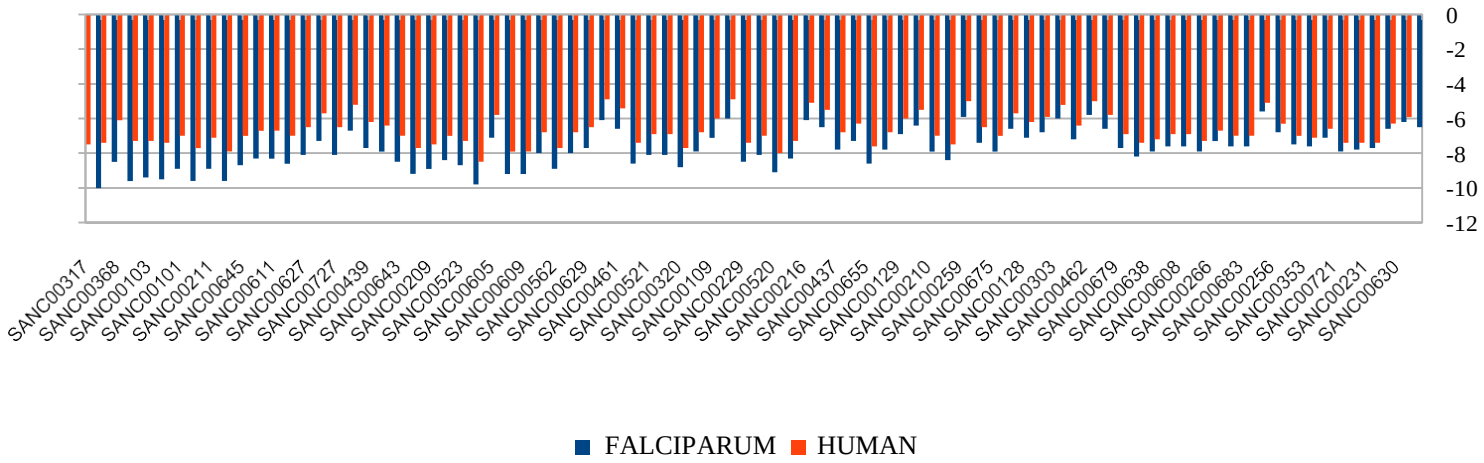


Figure A3.4: Bar graph representation showing the free energies of binding of SANCDB compounds bound to the *Plasmodium. falciparum* GCH1 protein active site versus its human.

Appendix 4 – Calculations and validation of zn+ force field parameters of the GCH1 enzyme

```
%nprocshared=24
%lindaworkers=cnode0858.cm.cluster
%mem=50GB
%chk=subset.chk
# opt b3lyp/6-31g(d) iop(6/50=1) scf=(qc,Nosymm) geom=connectivity

Title Card Required

3 2
 Zn(PDBName=ZN,ResName=2,ResNum=2)                  -22.49100000
-20.29800000     2.40300000
 C(PDBName=C,ResName=2,ResNum=2)                    -18.28300000
-22.86100000     2.05600000
 C(PDBName=C,ResName=2,ResNum=2)                    -19.25900000
-21.79600000     2.53700000
 S(PDBName=S,ResName=2,ResNum=2)                    -21.02100000
-22.21200000     2.05200000
 N(PDBName=N,ResName=2,ResNum=2)                    -28.22400000
-25.58000000     5.38900000
 C(PDBName=C,ResName=2,ResNum=2)                    -28.51900000
-26.12200000     4.06000000
 C(PDBName=C,ResName=2,ResNum=2)                    -27.37200000
-25.74700000     3.14900000
 C(PDBName=C,ResName=2,ResNum=2)                    -29.86300000
-25.65900000     3.45800000

 N(PDBName=N,ResName=2,ResNum=2)                    -26.40600000
-25.05300000     3.63900000
 C(PDBName=C,ResName=2,ResNum=2)                    -25.12500000
-24.70200000     2.99700000
 C(PDBName=C,ResName=2,ResNum=2)                    -24.55500000
-23.32400000     3.41300000
 C(PDBName=C,ResName=2,ResNum=2)                    -25.11100000
-22.12100000     2.71600000
 C(PDBName=C,ResName=2,ResNum=2)                    -26.50300000
-21.86100000     2.32400000
 N(PDBName=N,ResName=2,ResNum=2)                    -24.40600000
-21.02100000     2.37200000
 C(PDBName=C,ResName=2,ResNum=2)                    -25.34800000
-20.17600000     1.81500000
 N(PDBName=N,ResName=2,ResNum=2)                    -26.60900000
-20.66800000     1.77700000
 N(PDBName=N,ResName=2,ResNum=2)                    -21.26200000
-15.73300000    -0.10000000
 C(PDBName=C,ResName=2,ResNum=2)                    -21.53800000
-16.99000000    -0.91500000
 C(PDBName=C,ResName=2,ResNum=2)                    -21.26800000
-18.25500000    -0.10100000
 S(PDBName=S,ResName=2,ResNum=2)                    -22.28700000
-18.23800000     1.44700000
 O(PDBName=O,ResName=2,ResNum=2)                    -22.00100000
```

```
-19.82200000     4.35400000
 H(PDBName=H,ResName=2,ResNum=2)                    -17.28200000
-22.59300000     2.41100000
 H(PDBName=H,ResName=2,ResNum=2)                    -18.24300000
-22.91900000     0.96400000
 H(PDBName=H,ResName=2,ResNum=2)                    -18.52900000
-23.84900000     2.45600000
 H(PDBName=H,ResName=2,ResNum=2)                    -19.30200000
-21.74000000     3.62800000
 H(PDBName=H,ResName=2,ResNum=2)                    -19.03100000
-20.80300000     2.14300000
 H(PDBName=H,ResName=2,ResNum=2)                    -20.87100000
-22.22300000     0.71100000
 H(PDBName=H,ResName=2,ResNum=2)                    -28.13100000
-26.31200000     6.09000000
 H(PDBName=H,ResName=2,ResNum=2)                    -28.96000000
-24.95200000     5.70900000
 H(PDBName=H,ResName=2,ResNum=2)                    -28.52000000
-27.22400000     4.07200000
 H(PDBName=H,ResName=2,ResNum=2)                    -27.34700000
-26.06300000     2.10600000
 H(PDBName=H,ResName=2,ResNum=2)                    -30.00300000
-26.04300000     2.44300000

 H(PDBName=H,ResName=2,ResNum=2)                    -29.93000000
-24.56600000     3.43200000
 H(PDBName=H,ResName=2,ResNum=2)                    -30.67900000
-26.04400000     4.07600000
 H(PDBName=H,ResName=2,ResNum=2)                    -26.59500000
-24.88100000     4.66000000
 H(PDBName=H,ResName=2,ResNum=2)                    -24.40500000
-25.46400000     3.31200000
 H(PDBName=H,ResName=2,ResNum=2)                    -25.24700000
-24.78600000     1.91400000
 H(PDBName=H,ResName=2,ResNum=2)                    -24.68700000
-23.19400000     4.49900000
 H(PDBName=H,ResName=2,ResNum=2)                    -23.47300000
-23.36500000     3.25100000
 H(PDBName=H,ResName=2,ResNum=2)                    -27.35600000
-22.51600000     2.45600000
 H(PDBName=H,ResName=2,ResNum=2)                    -25.08100000
-19.19600000     1.43700000
 H(PDBName=H,ResName=2,ResNum=2)                    -20.26300000
-15.62400000     0.11600000
 H(PDBName=H,ResName=2,ResNum=2)                    -21.56300000
-14.88400000    -0.59700000
 H(PDBName=H,ResName=2,ResNum=2)                    -22.58600000
-16.92500000    -1.21600000
```

```
 H(PDBName=H,ResName=2,ResNum=2)                    -20.90200000
-16.93800000    -1.80200000
 H(PDBName=H,ResName=2,ResNum=2)                    -21.53200000
-19.09800000    -0.74500000
 H(PDBName=H,ResName=2,ResNum=2)                    -20.20300000
-18.33800000     0.14200000
 H(PDBName=H,ResName=2,ResNum=2)                    -21.77200000
-15.78500000     0.79700000
 H(PDBName=H,ResName=2,ResNum=2)                    -21.57200000
-18.95300000     4.45800000
 H(PDBName=H,ResName=2,ResNum=2)                    -22.44200000
-20.02000000     5.19900000

 1 14 1.0 20 1.0
 2 23 1.0 3 1.0 22 1.0 24 1.0
 3 4 1.0 25 1.0 26 1.0
 4 27 1.0
 5 6 1.0 28 1.0 29 1.0
 6 8 1.0 30 1.0 7 1.0
 7 9 2.0 31 1.0
 8 32 1.0 33 1.0 34 1.0
 9 10 1.0 35 1.0
 10 36 1.0 37 1.0 11 1.0
 11 12 1.0 38 1.0 39 1.0
 12 13 1.0 14 1.5

 13 16 2.0 40 1.0
 14 15 1.0
 15 16 1.5 41 1.0
 16
 17 48 1.0 18 1.0 42 1.0 43 1.0
 18 44 1.0 19 1.0 45 1.0
 19 20 1.0 46 1.0 47 1.0
 20
 21 50 1.0 49 1.0
 22
 23
 24
 25
 26
 27
 28
 29
 30
 31
 32
 33
 34
 35
 36
 37
```

```
 38
 39
 40
 41
 42
 43
 44
 45
 46
 47
 48
 49
 50



C H O N 0
6-31G(d)
****
Zn 0
LANL2DZ
****

Zn 0
LANL2DZ
```

Figure A4.1: Photo-snap of Gaussian input file for geometry optimization of the subset system

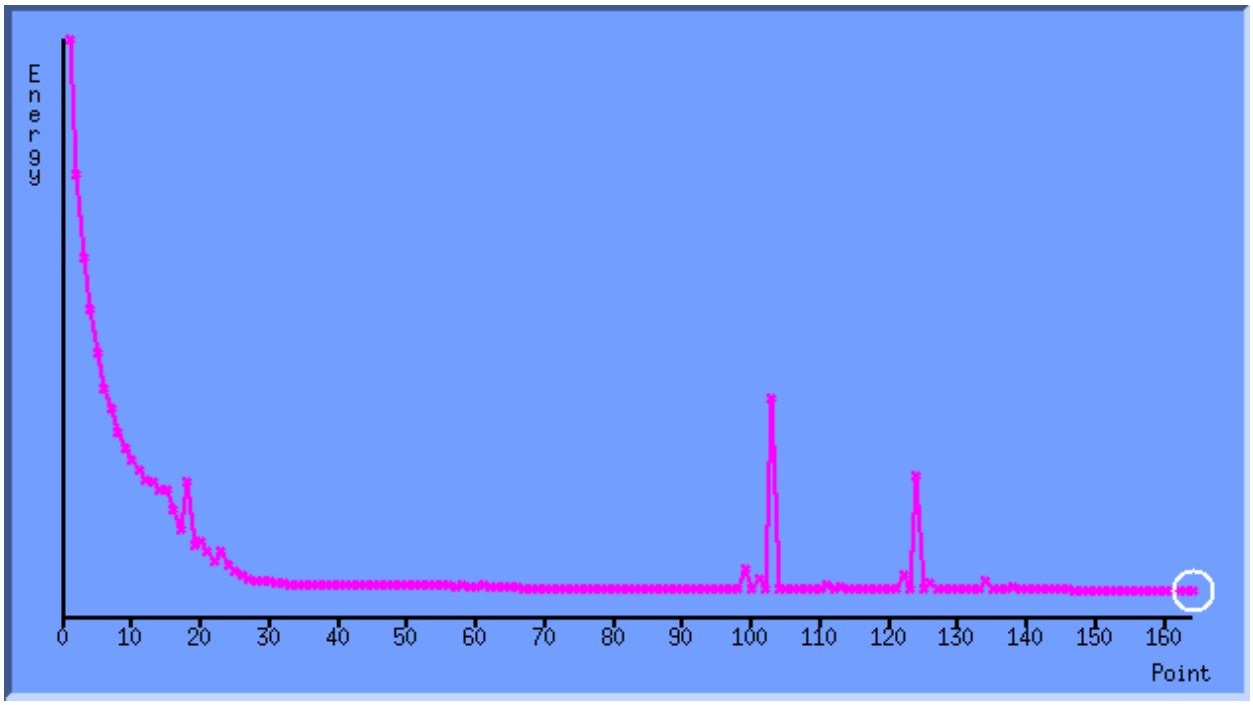


Figure A4.2: MOLDEN geometry convergence graph. It shows a plot of energy against the geometry point of the subset system. The final geometry is donated by a white circle. The saddle point in the graph corresponds to a minimum energy structure, therefore the stability of the optimised structure in terms of energy was presumed.

```
%nprocshared=24
%lindaworkers=cnode0315.cm.cluster
%mem=50GB
%chk=1_4cys_h_bond+.chk
# opt=modredundant b3lyp/6-31g(d) iop(6/50=1) geom=connectivity scf=(qc,Nosymm)

Title Card Required

3 2
 Zn(PDBName=ZN,ResName=2,ResNum=2)                  -23.48700000  -20.75300000
2.82800000
 C(PDBName=C,ResName=2,ResNum=2)                    -20.47200000  -23.98100000
4.43100000
 C(PDBName=C,ResName=2,ResNum=2)                    -21.11600000  -22.63500000
4.09800000
 S(PDBName=S,ResName=2,ResNum=2)                    -22.09400000  -22.59100000
2.58700000
 N(PDBName=N,ResName=2,ResNum=2)                    -24.53100000  -24.95500000
4.93000000
 C(PDBName=C,ResName=2,ResNum=2)                    -25.81200000  -24.46100000
4.44100000
 C(PDBName=C,ResName=2,ResNum=2)                    -25.99500000  -24.45700000
2.92400000
 C(PDBName=C,ResName=2,ResNum=2)                    -26.08500000  -23.06300000
4.99400000
 N(PDBName=N,ResName=2,ResNum=2)                    -24.92500000  -24.19400000
2.17700000
 C(PDBName=C,ResName=2,ResNum=2)                    -25.01200000  -24.14100000
0.71600000
 C(PDBName=C,ResName=2,ResNum=2)                    -24.80100000  -22.70600000
```

```
B 1 4 S 10 0.050000


C H O N 0
6-31G(d)
****
Zn 0
LANL2DZ
****

Zn 0
LANL2DZ
```

Figure A4.3: Example of Gaussian input file for Coordinates scan of the subset system bonds. Bonds were stretched by 0.05 Å each way in ten steps.

```
%nprocshared=24
%lindaworkers=cnode0640.cm.cluster
%mem=50GB
%chk=ANGLES_1_14_10_N+.chk
# opt=modredundant b3lyp/6-31g(d) iop(6/50=1) scf=(qc,Nosymm) geom=connectivity

Title Card Required

3 2
 Zn(PDBName=ZN,ResName=2,ResNum=2)                  -22.49100000  -20.29800000
2.40300000
 C(PDBName=C,ResName=2,ResNum=2)                    -18.28300000  -22.86100000
2.05600000
 C(PDBName=C,ResName=2,ResNum=2)                    -19.25900000  -21.79600000
2.53700000
 S(PDBName=S,ResName=2,ResNum=2)                    -21.02100000  -22.21200000
2.05200000
 N(PDBName=N,ResName=2,ResNum=2)                    -28.22400000  -25.58000000
5.38900000
 C(PDBName=C,ResName=2,ResNum=2)                    -28.51900000  -26.12200000
4.06000000
 C(PDBName=C,ResName=2,ResNum=2)                    -27.37200000  -25.74700000
3.14900000
 C(PDBName=C,ResName=2,ResNum=2)                    -29.86300000  -25.65900000
3.45800000
 N(PDBName=N,ResName=2,ResNum=2)                    -26.40600000  -25.05300000
3.63900000
 C(PDBName=C,ResName=2,ResNum=2)                    -25.12500000  -24.70200000

A 4 1 14 S 10 -1.000000


C H O N 0
6-31G(d)
****
Zn 0
LANL2DZ
****

Zn 0
LANL2DZ
```

Figure A4.4: Example of Gaussian input file for Coordinates scan of the subset system angles. Angles were stretched by 1 Å each way in ten steps.

```
%nprocshared=24
%lindaworkers=cnode0552.cm.cluster
%mem=50GB
%chk=Dihedral_4_12_1_19_P+.chk
# opt=modredundant b3lyp/6-31g(d) iop(6/50=1) scf=(qc,Nosymm) geom=connectivity

Title Card Required

3 2
 Zn(PDBName=ZN,ResName=2,ResNum=2)                  -22.49100000  -20.29800000
2.40300000
 C(PDBName=C,ResName=2,ResNum=2)                    -18.28300000  -22.86100000
2.05600000
 C(PDBName=C,ResName=2,ResNum=2)                    -19.25900000  -21.79600000
2.53700000
 S(PDBName=S,ResName=2,ResNum=2)                    -21.02100000  -22.21200000
2.05200000
 N(PDBName=N,ResName=2,ResNum=2)                    -28.22400000  -25.58000000
5.38900000
 C(PDBName=C,ResName=2,ResNum=2)                    -28.51900000  -26.12200000
4.06000000
 C(PDBName=C,ResName=2,ResNum=2)                    -27.37200000  -25.74700000
3.14900000
 C(PDBName=C,ResName=2,ResNum=2)                    -29.86300000  -25.65900000
3.45800000
 N(PDBName=N,ResName=2,ResNum=2)                    -26.40600000  -25.05300000
3.63900000
 C(PDBName=C,ResName=2,ResNum=2)                    -25.12500000  -24.70200000



D 4 1 20 19 S 10 1.000000


C H O N 0
6-31G(d)
****
Zn 0
LANL2DZ
****

Zn 0
LANL2DZ
```

Figure A4.5: Example of Gaussian input file for coordinates scan of the subset system dihedrals; each were stretched by 1 Å each way in ten steps.

```
* Run Segment Through CHARMM
*



!read rtf form name ffield/top_all36_prot.rtf


read rtf card name ffield/top_all36_prot.rtf
read param card name ffield/par_all36_prot.prm

! Read sequence from the PDB coordinate file
open unit 1 card read name protein/1wur.pdb
read sequ pdb unit 1

generate A setup warn first none last none

rewind unit 1

! set bomlev to -1 to avois sying on lack of hydrogen coordinates
bomlev -1
read coor pdb unit 1
! them put bomlev back up to 0
bomlev 0



close unit 1

!patch disu A 163 A 43 setup sort warn

! prints out number of atoms that still have undefined coordinates.
define test select segid A .and. ( .not. hydrogen ) .and. ( .not. init )
show end

ic para
ic fill preserve
ic build
hbuild sele all end

! ZN ++ ions
open unit 3 card read name protein/ZN.pdb
read sequ pdb unit 3
!read sequ cu 1
generate Z setup warn first none last none
rewind unit 3
read coor pdb unit 3
close unit 3
```

```
!patching bonds



patch hison Z 186 A 80
patch cyson Z 186 A 77
patch cyson Z 186 A 148

!patching angles

patch hisok A 77 Z 186 A 80
patch hisop A 77 Z 186 A 148
patch hisoq A 148 Z 186 A 80

!patching dihedrals
patch hisox A 148 Z 186 A 80 A 80
patch hisos A 148 Z 186 A 80 A 80

! write out the protein structure file (psf) and
! the coordinate file in pdb and crd format.

write psf card name 01_1wur_prepare.psf
* PSF
*


write coor pdb name 01_1wur_prepare.pdb
* PDB
*

write coor card name 01_1wur_prepare.crd
* Coords
*

stop
```

Figure A4.6: Photo-snap of CHARMM preparation file.inp; it involves reading the residues topology files (RTF) and the residues parameter file (PAR) and the PDB file sequence and coordinates.