Identification of possible natural compounds as potential inhibitors against *Plasmodium* M1 alanyl aminopeptidase

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

Master of Science in Bioinformatics

at

RHODES UNIVERSITY, SOUTH AFRICA

Department of Biochemistry and Microbiology

Faculty of Science

by

Omar Samir Abdel Ghaffar Soliman

2018

Abstract

Malaria is a major tropical health problem with a 29% mortality rate among people of all ages; it also affects 35% of the children. Despite the decrease in mortality rate in recent years, malaria still results in around 2000 deaths per day. Malaria is caused by *Plasmodium* parasites and is transmitted to humans via the bites from infected female *Anopheles* mosquitoes during blood meals. There are five different *Plasmodium* species that can cause human malaria, which include *Plasmodium falciparum, Plasmodium vivax, Plasmodium malariae, Plasmodium ovale* and *Plasmodium knowlesi*. Among these five species, the most pathogenic ones are *Plasmodium falciparum and Plasmodium vivax*. Malaria is usually hard to diagnose because the symptoms are not exclusive to malaria and very similar to flu, e.g., fever, muscle pain, and chills, which lead to the misdiagnosis of malaria cases. Malaria is lethal if not treated because it can cause severe complications in the respiratory tract, liver, metabolic acidosis, and hypoglycemia. The malaria parasite life cycle includes two types of hosts, i.e., a human host and female *Anopheles* mosquito host. Malaria continuously develops resistance to the available drugs, which is one of the major challenges in disease control. This situation confirms the need to develop new drugs that target virulence factors of malaria. The malarial parasite has three main life cycle stages, which include the host liver stage, host blood stage and vector stage. In the blood stage, parasites degrade hemoglobin to amino acids, which is important as these parasites cannot produce their own amino acids. Different proteases are involved in this hemoglobin degradation process. M1 alanyl aminopeptidase is one of these proteases involved at the end of hemoglobin degradation. This study focused on M1 alanyl aminopeptidase as a potential drug target. M1 alanyl aminopeptidase consists of four domains: N-terminal domain, catalytic domain, middle domain and C-terminal domain. The catalytic domain remains conserved among different *Plasmodium* species. Inhibition of this enzyme might prevent *Plasmodium* growth as it can't produce its own amino acids. In this study, sequence analysis was carried out in both human and *Plasmodium* M1 alanyl aminopeptidase to identify conserved and divergent regions between them. 3D protein models of the M1 alanyl aminopeptidase from *Plasmodium* species were built and validated. Then the generated models were used for virtual screening against 623 compounds retrieved from the South African Natural Compounds Database (SANCDB, https://sancdb.rubi.ru.ac.za/). Virtual screening was done using blind and targeted docking methods. Docking was used to identify compounds with selective high binding affinity to the active site of the parasite protein. In this study, one

SANCDB compound was selected for each protein: SANC00531 was selected against *P. falciparum* M1 alanyl aminopeptidase, SANC00469 against *P. knowlesi*, SANC00660 against *P. vivax*, SANC00144 against *P. ovale* and SANC00109 against *P. malariae.* It was found that *Plamsodium* M1 alanyl aminopeptidase can be used as a potential drug target as it showed selective binding against different inhibitor compounds. This result will be investigated in future work though molecular dynamic analysis to investigate the stability of protein-ligand complexes.

Declaration

I declare that this thesis is my own, unaided work, unless otherwise stated. It is being submitted for the degree of Master of science at Rhodes University. It has not been submitted before for any degree or examination in any other university.

Signature……

Date……

Dedication

This thesis is dedicated to the loving memory of

Prof. Ahmed Shokry

For always encouraging and motivating me to believe in myself and science

Acknowledgements

I would first like to thank Prof. Özlem Tastan Bishop. The door to Prof. Özlem's office was always open whenever I ran into a problem or had a question about my research or writing. She directs and advises me always in the right the direction. My sincere gratitude and appreciation goes to her for always supporting me.

I would like also to express my sincere gratitude to my advisor Dr. Vuyani Moses for the continuous support of my master study and research, for his patience, motivation, encouragement and explanation. His guidance helped me in all the time of research and writing of this thesis.

Furthermore, I would like to express my appreciation to all RUBi family for the continuous discussions and for the sleepless nights we were working together before deadlines. I am deeply especially grateful to Mr. Magambo Philip Kimuda, Mr. Olivier Sheik Amamuddy, Mr. Bakary N'tji Diallo, Dr. Natasha Sanabria and Afrah Khairallah.

Last but not the least, I would like to thank my family: My father Prof. Samir Abd El Ghaffar, my mother Prof. Elham El sayed and my wife Rofaida Saad. With their support, encouragement, sacrifice and prayer, this thesis has become possible.

Table of Contents

List of Figures

List of Tables

List of abbreviations

3D	3 dimensional
BLAST	Basic Local Alignment Search Tool
NCBI	National Centre For Biotechnology Information
MSA	Multiple Sequence Alignment
DOPE	Discrete Optimized Protein Energy
HMM	Hidden Markov Model
MAFFT	Multiple Alignment Using Fast Fourier Transform
MUSCLE	Multiple Sequence Comparison By Log-Expectation
PDB	Protein Data Bank
PSI-BLAST	Position Specific Iterated Blast
PROMALS3D	Profile Multiple Alignment With Predicted Local Structures And 3D Constraints
SANCDB	South African Natural Compound Database
WHO	World Health Organization

Chapter 1 - Literature Review

1.1 Introduction

Human malaria infection can be caused by any of the 5 different parasite species that belong to the *Plasmodium* species. These parasites include *Plasmodium falciparum, Plasmodium vivax, Plasmodium malariae, Plasmodium ovale* and *Plasmodium knowlesi.* The parasite is transmitted to the human body through the bite of an infected female *Anopheles* mosquito. The female *Anopheles* mosquito's saliva contains the parasite which is transmitted to human blood when the mosquito bites the human. The parasite matures and reproduces in the human liver before it infects and destroys red blood cells. The most pathogenic parasites in the *Plasmodium* genus are the *P. falciparum* and *P. vivax* species [1].

Around 1 million people are killed each year by malaria and in 2002, 515 million (range 300-600 million) were attributed to episodes of clinical *P. falciparum*. 90% of malaria cases and deaths occur in sub-Saharan Africa, but malaria is also a public health problem in South America and South East Asia [2]. *P. falciparum* is responsible for most deaths in humans, however, other malaria-causing parasites such as *P. vivax, P. ovale,* and *P. malariae* do cause a milder form of the disease [3].

1.2 Signs and Symptoms

Malaria symptoms usually appear after 10 – 15 days following the infective mosquito bite. The malaria symptoms can be delayed by using the appropriate antimalarial drugs [4]. The first symptoms are flu-like symptoms which make it difficult to diagnose malaria. These symptoms include: headaches, fever, chills, and vomiting. It is very important to treat malaria within 24 hours or it can progress to severe illness, which could lead to death [5]. The symptoms can develop into severe anemia, cause respiratory distress, cerebral complications, hypoglycemia, and glomerulonephritis [6].

1.3 Malaria life cycle

Malaria has a complex life cycle involving two different hosts, the first one being a female *Anopheles* mosquito while the second is the human host [4]. In general, it involves three main stages. Firstly, there is a human liver stage, followed by a human blood cell stagewhich finally ends in the mosquito stages. Malaria infection begins with a bite from an infected female *Anopheles* mosquito that transmits sporozoites to vertebrate host (e.g: human host). Once they enter the host, they travel through blood vessels and infect hepatocytes where the parasite grows and reproduces asexually to produce merozoites to infect red blood cells, as shown in

Figure 1-1. Some of this merozoites develop into a sexual form that arev transmitted later to another mosquito during mosquito blood feeding as shown in Figure 1-2 [7].

1.3.1 Liver stage

During *Anopheles* mosquito bite, parasite sporozoites are transmitted to the human dermis. A portion of sporozoites penetrates blood vessels by using gliding motility, which depends on the Trap-like protein (TLP) [8]. Then they invade hepatocytes by using a moving junction-independent process via cell traversal (CT) and a moving junction-dependent process, thus creating parasitophorous vacuoles (PVs). CT starts with the breakdown of hepatocyte cell membranes to move through the cell cytoplasm using proteins such as the Perforin-Like Protein 1 (PLP1), the sporozoite microneme protein essential for traversal (SPECT) [7], phospholipase (PL) and the gamete egress and sporozoite traversal protein (GEST). To avoid degradation by lysosomes, sporozoites use pH sensing and PLP1 [8].

To invade host hepatocytes, the surface of sporozoites are coated by a key protein called circumsporozoite protein (CSP),which consist of a type I thrombospondin repeat (TSR) and a highly repetitive region. CSP binds with heparin sulfate proteoglycans (HSPGs), which are located on the hepatocyte surface. These activate CSP and remove the N-terminus to expose the TSR domain. Sporozoites also contains important organelles for hepatocyte invasion, such as micronemes and rhoptries. In order to form the PV microneme, proteins P52 and P36 interact with each other and with the hepatocyte Ephrin A2 receptor (EphA2). Additionally, the hepatocyte receptor CD81 plays an important role in PV formation [9]. Once a sporozoite successfully infects a hepatocyte, it resides within the PV. The sporozoite remains in the liver stage from 2 – 10 days. The result of this stage is the development and release of up to 40000 merozoites per hepatocyte cell into the bloodstream in the form of merosomes, which are vesicles filled with parasites [8].

1.3.2 Intra-erythrocyte stage (Blood stage)

After infecting liver hepatocytes for 2 – 10 days, the merozoites are released into the bloodstream to infect erythrocytes via ligand-receptor interactions. For *P. falciparum,* basigin, red blood cell antigen, and *P. falciparum* reticulocyte binding protein homologue 5 (PFRh5) interact to form a complex. This complex consists of PfRh5, PfRh5-interacting protein (PfRipr) and Cysteine-rich protective antigen (CyRPA), which bind the basigin of erythrocyte cell. This leads to the invasion of this erythrocyte[10]. For *P. vivax* it requires the presence of the Duffy blood group antigen Fy^a or Fy^b. *P. vivax* cannot infect a host with the Duffy negative FyFy phenotype, and that explains why most people in West Africa are resistant to this species [11].

The parasite has two alternative methods of reproduction, namely asexual (Figure 1-1) and sexual (Figure 1-2) multiplication. An asexual cycle takes between 24 hours to 72 hours depending on the parasite species whereby *P. knowlesi* takes 24 hours, *P. falciparum* and *P. vivax* take 48 hours, and *P. malariae* takes 72 hours [11]. Most *Plasmodium* species take 48 hours to complete the sexual cycle while in *P. falciparum* it usually takes 10-12 days to complete a full cycle [12].

Each released merozoite invades an erythrocyte and begins the asexual cycle, which consumes the erythrocyte's contents [12]. Malaria cannot produce its own amino and acids and thus it needs to degrade erythrocyte hemoglobin. The degradation takes place in the parasite digestive vacuole at pH 5.2 and occurs during the blood stage [13]. Inside this vacuole, a massive proteolytic pathway degrades hemoglobin into amino acids [14]. Each asexual cycle produces 16-32 new merozoites, which invade new erythrocytes. As a result, the parasite population is enlarged by a factor of 6 to 20 times per cycle. The *Plasmodium* parasite selectively invades erythrocytes, for example young erythrocytes are usually infected by *P. vivax* [9].

The asexual cycle consists of ring stage, a trophozoite stage, and a schizont stage. The first stage that is established after entering the erythrocyte is the ring stage. This stage is characterized by a ring-like shape under the microscope. Then they enter the trophozoite stage, in which surface antigens are expressed, during which high metabolic activity is observed. The last step is the schizont stage, which produces around 16-32 merozoites through cell division to result in the rupture of the erythrocyte andin the invasion of new erythrocytes. These stages are classified under the asexual blood stage [15].

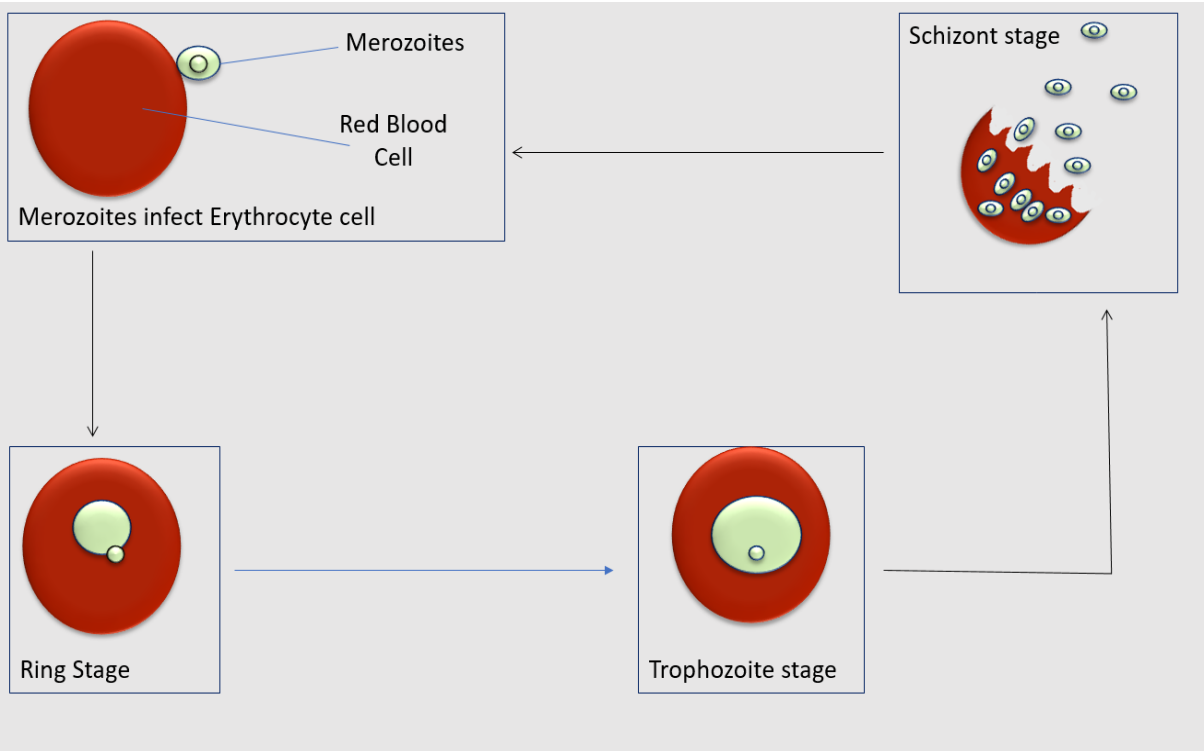


Figure 1-1: The malaria parasite asexual cycle. Merozoite invasion of erythrocyte cells and the asexual cycle result in the production of 16-32 merozoites, which invade new healthy erythrocytes to initiate the second wave of erythrocyte invasion [8].

The resulting merozoites cannot be transmitted to a mosquito; thus a small portion of merozoites - usually less than 10% - go through with sexual reproduction (gametocytogenesis) and develop into sexual form (gametocytes) of the parasite. This results in a male and female gametocyte, which can be transmitted to a female mosquito during a blood meals. The duration of a gametocyte of *P. vivax* after releasing merozoites from hepatocytes takes around one week. However, in *P. falciparum* the precise time of developing gametocyte is not fixed and is unclear as it depends on many factors. For example, if the parasite is exposed to an antimalarial drug, it will force the gametocyte to develop and survive. The same could happen if the human host is dying due to denaturation of red blood cells. At the same time, it could be affected with reproductive restraint such that the precise time of developing gametocyte is generally not clear and varies from one case to another and from one species to another [15].

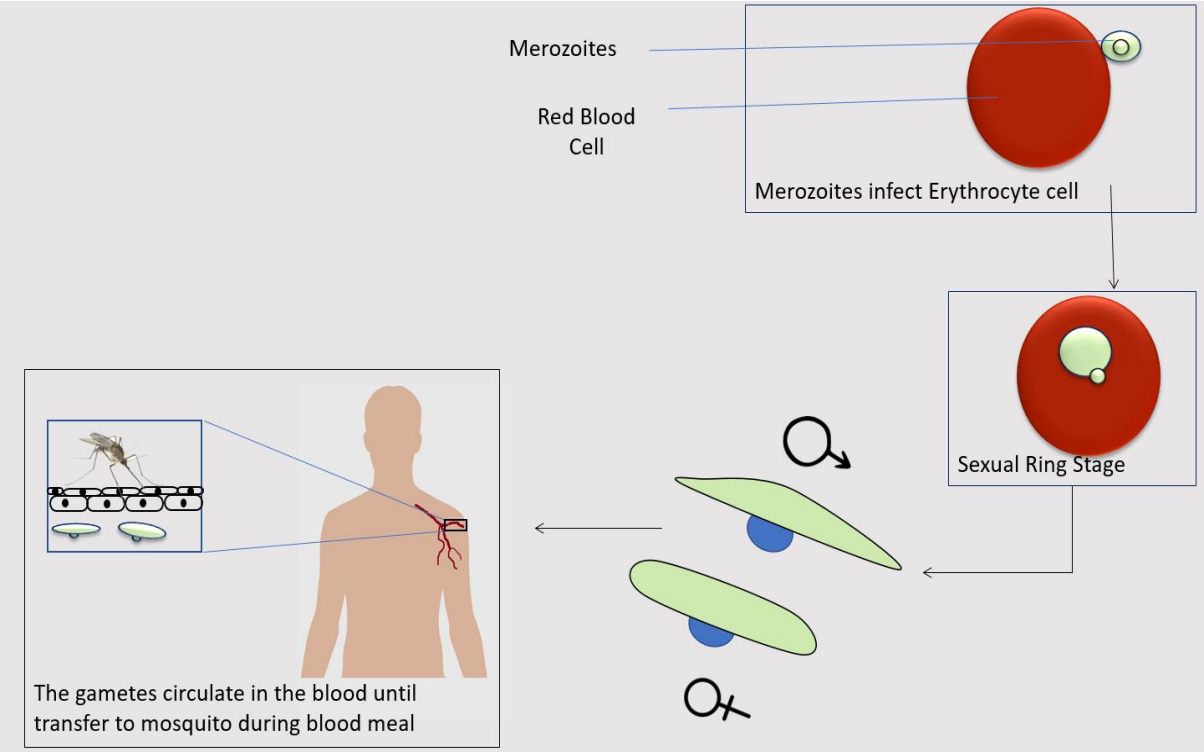


Figure 1-2: Malaria parasite sexual cycle. The sexual cycle of *Plasmodium* parasite which takes place in erythrocyte cell and results in the production of sexual ring then male and female gametocytes. After maturation, they transmit to another mosquito during a blood meal [8].

1.3.3 Mosquito stage

During the *Plasmodial* life cycle, the parasite undergoes one sexual reproduction, which takes place only in the mosquito stage. Ingestion of male and female gametocytes activates the gametocytes in the mosquito midgut [16]. This activation is caused by the temperature drop, pH change and xanthurenic acid. Thus, the gametocytes mature and develop into gametes. Male gametes form the octoploid nucleus so that it goes through three fast DNA replication events. Additionally, male gametes go through exflagellation, which results in the formation of eight flagella. The time needed to complete the maturation process differs from one *Plasmodium* species to another, as shown in Table 1-1. After completing the maturation step for both male (microgametes) and female gametes (macrogametes), the male gamete fertilizes female gamete to form a fertilized female gamete which will develop into an ookinete, as shown in Figure 1-3. Ookinetes go through the mosquito's midgut wall (epithelial cell wall) and form oocysts [15].

Each oocyst contains thousands of sporozoites. The sporozoite develops inside an oocyst until its rupture, resulting in the release of sporozoites into the body cavity. The sporozoites travel and migrate to the mosquito's salivary gland where they wait for the mosquito to take the next

blood meal. During this blood meal, they are transmitted to another human host and start the liver stage infection [15].

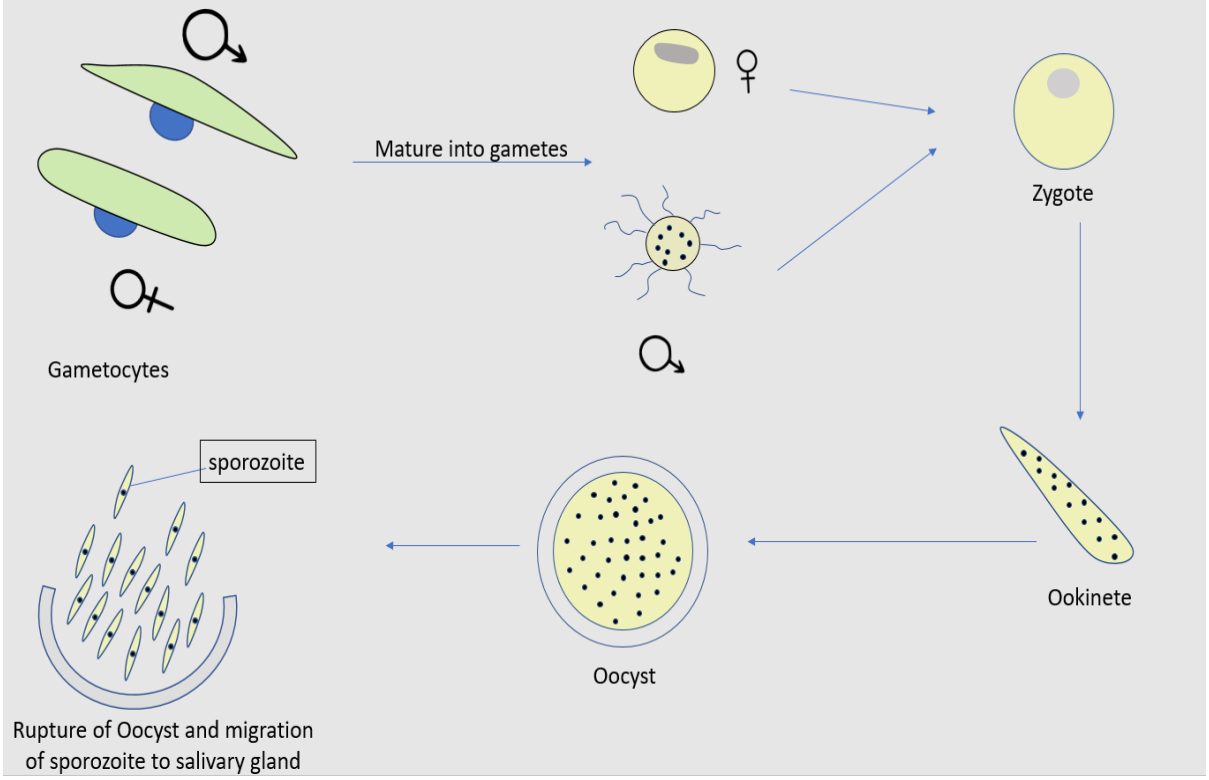


Figure 1-3: The mosquito stage, in which most of the steps take place inside mosquito midgut, to result in the production sporozoites. In the end, sporozoites migrate to the salivary gland where they stay until the next mosquito blood meal [17].

Table 1-1: Summary of needed time for complete maturation of gametocytes [15]

Plasmodium species	Time needed (days)
Plasmodium falciparum	8:10
P. malariae	6:8
P. vivax	3:4
P. ovale	3:4

1.4 Peptidases

According to the MEROPS database [18], there are different families of protease enzymes include Aspartic peptidases, Cysteine peptidases, Glutamic peptidases, Metallopeptidases, Asparagine peptidases, Mixed peptidases, Serine peptidases, Threonine peptidases and peptidase of unknown catalytic type (See Table 1-2). Each one of these families is identified

by a single letter representing the type of reaction of the protease enzyme and a unique number. For example, M1 belongs to metallopeptidase family [19].

Malaria peptidases have two main functions. These are invasion and rupture of erythrocytes, and hemoglobin degradation. Hemoglobin degradation involves different proteases such as aspartic proteases, falcilysin, plasmepsins, cysteine proteases, metalloproteases, dipeptidyl aminopeptidase 1 (DPAP1) and falcipains [20].

The first cleavage occurs between Phe at position 33 and Leu at position 34. Then falcipains and plasmepsins degrade the resulting molecule into small peptides. The enzymes DPAP1 and falcilysin degrade the small peptides into shorter oligopeptides or dipeptides which are transported to the parasite cytosol where they will be degraded into free amino acids by neutral aminopeptidase [14]

1.4.1 Metallopeptidases

Metallopeptidases are a set of homologous peptidases which need metal a ion for their catalytic mechanism. This metal is usually Zinc (Zn^{2+}), but could be Copper (Cu^{2+}) or Cobalt (Co^{2+}). Usually, three amino acid coordinate the metal ion in its position in the protein [19]. As shown in Table 1-2 and 1-3, there are over 50 metallopeptidase families and subfamilies, making them the largest peptidase enzyme family. Based on the cleavage site metallopeptidase are classified as end-opeptidase EC 3.4.21-25 and exo-peptidase EC 3.4.11-19 [21].

1.4.2 Exo-aminopeptidases

Exo-aminopeptidases can eliminate amino acids from N-termini of peptides. In *Plasmodium* parasites, in addition to providing free amino acids, they also have a role in re-invasion of erythrocytes [22]. *Plasmodium* parasites use nine different exo-aminopeptidases. Four of these enzymes are methionine aminopeptidases. The other enzymes are alanyl aminopeptidase, aspartic aminopeptidase, leucine aminopeptidase, prolyl aminopeptidase and post prolyl aminopeptidase. Exo-aminopeptidases have different functions depending on the catalytic activity of the enzyme. For example, they have the activity to remove the N-terminal methionine, which is the activity of methionine aminopeptidases. On the other hand, alanyl aminopeptidase and leucine aminopeptidase can digest dipeptides into free amino acids, which is very important for the parasite to grow. Inhibition of these enzymes can thus stop protein biosynthesis and as a result, inhibit the *Plasmodium* parasite growth [14].

M1 Aminopeptidases (EC 3.4.11) are enzymes that catalyze peptide bonds between amino acids from the amino terminal of proteins or polypeptide chains. M1 Aminopeptidase belongs to the metzincins clan, which are zinc-dependent metallopeptidases [23]. There are more than 10000 protein sequences that belong to the M1 aminopeptidase family and 25 PDB structures. M1 Alanyl aminopeptidase (EC 3.4.11.2) (P*f*M1-AAP) depend on single catalytic zinc ion, which is coordinated by two histidines and one glutamate. The optimum pH for the activity of this enzyme is 7.4. *P. falciparum* M1 Alanyl aminopeptidase has been detected in an asexual cycle of the erythrocyte stage during the trophozoite and schizont step, which makes it an ideal antimalarial drug target. There are some studies that have shown that Bestatin or quinolone-based inhibitors could be used to inhibit the activity of this enzyme. McGrown et al. [22] have reported the crystal structure of the empty form of this enzyme with PDB ID 3EBG [22].

A single gene encodes M1 Alanyl aminopeptidase which consists of 1095 amino acid arranged into 4 domains. These domains comprise the N-terminal, catalytic domain, middle domain and C-terminal domain. The enzyme shares ~70% identity across different *Plasmodium* species. The active region of M1 Alanyl aminopeptidase is conserved and the most divergent region is located at the N-terminal extension. The 3D structure shows that it contains 26 α-helices and 26 β-sheets. Five β-sheets and eight α-helices form the catalytic domain. The active site is located between β-sheet number 18 and α-helices number 2, 3 and 5. Putative substrate entry could be used to access the active site [24].

Due to the similarity between M1 and M17 aminopeptidases, a drug can be designed to potentially target both enzymes. Drinkwater *et al*. [25] developed (1H-pyrazole-1-yl)phenyl)(amino)methyl) phosphonic acid which can bind within the S1 socket of the active site. However, the molecular dynamic (MD) simulation performed for this enzyme with the drug did not take into account the correct geometry of the metal active site [25].

Human aminopeptidase homologs play an important role after protein hydrolysis by gastric and pancreatic proteases, whereby they digest the generated peptides to release an N-terminal amino acid [26].

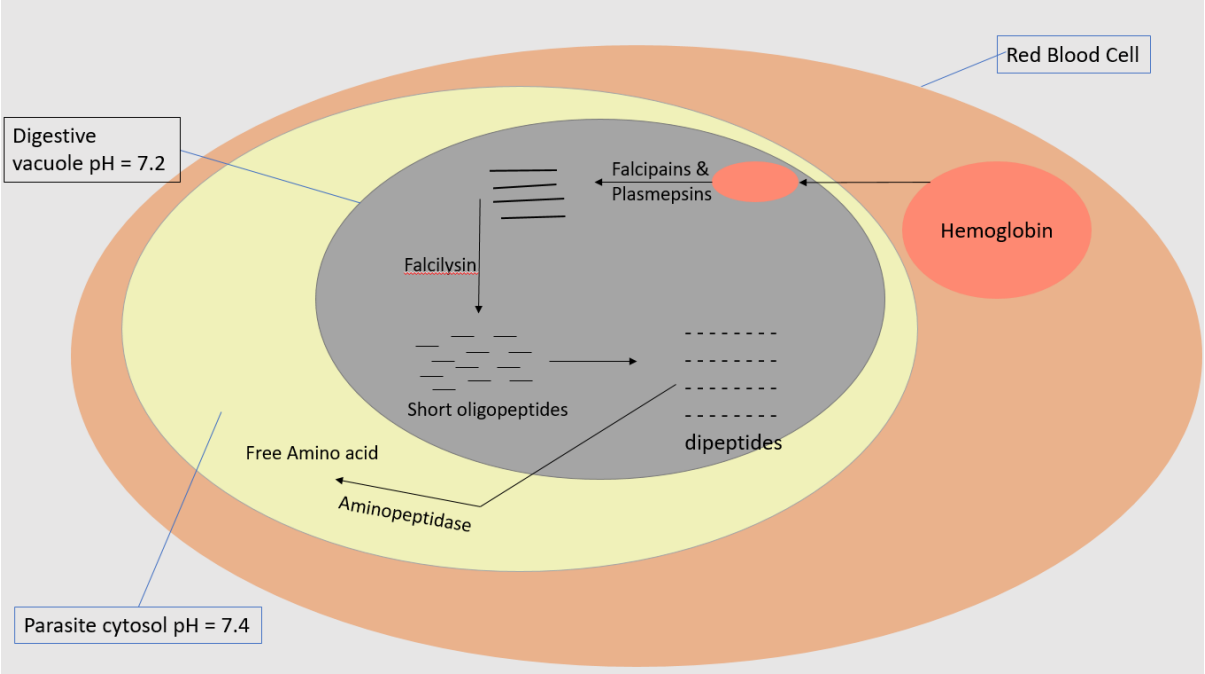


Figure 1-4: Hemoglobin digestion to release free amino acid in the erythrocyte stage during the sexual cycle of the *Plasmodium* parasite.

Table 1-2: Protease clans,families and sub-families, based on catalytic type (adapted from MEROPS) [19]

Catalytic type	Clan	Family	Sub-family
Aspartic peptidases	AA, AC, AD, AE and AF	A1, A2, A3, A5, A8, A9, A11, A22, A24, A25, A26, A28, A31, A32, A36 and A37	A1A, A1B, A2A, A2B, A2C, A2D, A3A, A3B, A11A, A11B, A22A, A22B, A24A, A28A and A28B
Cysteine peptidases	CA, CD, CE, CF, CL, CM, CN, CO, CP and CQ	C1:C28, C30, C31, C32, C33, C36, C37, C40, C41, C42, C44, C45, C46, C47, C48, C50, C51, C53, C54, C56, C57, C58, C59, C60, C62:C80, C82, C83, C84, C85, C86, C87, C89, C93, C95, C96, C97, C98, C99, C100, C101, C102, C104, C105, C107, C108, C110, C111, C113, C115 and C117	C1A, C1B, C2A, C3A, C3B, C3C, C3D, C3E, C3F, C3G, C3H, C11A, C11B, C14A, C14B, C16A, C16B, C58A, C58B, C60A, C60B, C82A, C85A and C85B
Mixed peptidases	PA, PB, PC, PD and PE	C3, C4, C24, C26, C30, C37, C46, C56, C62, C74, C99, C107, S1, S3, S6, S7, S29, S30, S31, S32, S39, S46, S55, S64, S65, S75, C44, P1 and P2	P2A and P2B
Serine peptidases	SB, SC, SE, SF, SH, SJ, SK, SO, SP, SR, SS and ST	S8, S53, S9, S10, S15, S28, S33, S37, S11, S12, S13, S24, S26, S21, S73, S77, S78, S80, S16, S50, S69, S14, S41, S49, S74, S59, S60, S66, S54, S48, S62, S68, S71, S72, S79 and S81	S1A, S1B, S1C, S1D, S1E, S1F, S8A, S8B, S9A, S9B, S9C, S9D, S26A, S26B, S26C, S39A, S39B, S41A, S41B, S49A, S49B and S49C
Metallopeptidases	MA, MC, MD ME, MF, MG, MH, MJ, MM, MN, MO, MP, MQ, MS and MT	M1:M13, M26, M27, M30, M32, M34, M35, M36, M41, M43, M48, M49, M54, M56, M57, M60, M61, M64, M66, M72, M76, M78, M80, M84, M85, M90, M91, M93, M95, M97, M98, M14, M86, M99, M15, M75 and M81	M3A, M3B, M9A, M9B, M10A, M10B, M10C, M12A, M12B, M14A, M14B, M14C, M14D, M15A, M15B, M15C, M15D, M16A, M16B, M16C, M20A, M20B, M20C, M20D, M20F, M23A, M23B, M24A, M24B, M28A, M28B and M28C.
Threonine peptidases		T1, T2, T3, T5, T7, and T8	T1A and T1B
Peptidase of unknown catalytic type		U32, U40, U49, U56, U57, U62, U69, U72, U73 and U74	

Table 1-3: Summary of MA metallopeptidase enzymes (adapted from MEROPS database)[19]

Clan	Family	Sub-family	Example (Organism name)
MA	M1		aminopeptidase N (*Homo sapiens*)
	M2		angiotensin-converting enzyme peptidase unit 1 (*Homo sapiens*)
	M3	M3A	thimet oligopeptidase (*Rattus norvegicus*)
		M3B	oligopeptidase F (*Lactococcus lactis*)
	M4		thermolysin (*Bacillus thermoproteolyticus*)
	M5		mycolysin (*Streptomyces cacaoi*)
	M6		immune inhibitor A peptidase (*Bacillus thuringiensis*)
	M7		snapalysin (*Streptomyces lividans*)
	M8		leishmanolysin (*Leishmania major*)
	M9	M9A	bacterial collagenase V (*Vibrio alginolyticus*)
		M9B	bacterial collagenase H (*Clostridium histolyticum*)
	M10	M10A	matrix metallopeptidase-1 (*Homo sapiens*)
		M10B	serralysin (*Serratia marcescens*)
		M10C	fragilysin (*Bacteroides fragilis*)
	M11		gametolysin (*Chlamydomonas reinhardtii*)
	M12	M12A	astacin (*Astacus astacus*)
		M12B	adamalysin (*Crotalus adamanteus*)
	M13		neprilysin (*Homo sapiens*)
	M26		IgA1-specific metallopeptidase (*Streptococcus sanguinis*)
	M27		tentoxilysin (*Clostridium tetani*)
	M30		hyicolysin (*Staphylococcus hyicus*)
	M32		carboxypeptidase Taq (*Thermus aquaticus*)
	M34		anthrax lethal factor (*Bacillus anthracis*)
	M35		deuterolysin (*Aspergillus flavus*)
	M36		fungalysin (*Aspergillus fumigatus*)
	M41		FtsH peptidase (*Escherichia coli*)
	M43	M43A	cytophagalysin (*Cytophaga* sp.)
		M43B	pappalysin-1 (*Homo sapiens*)
	M49		dipeptidyl-peptidase III (*Rattus norvegicus*)

	M54		archaelysin (*Methanocaldococcus jannaschii*)
	M56		BlaR1 peptidase (*Staphylococcus aureus*)
	M57		prtB g.p. (*Myxococcus xanthus*)
	M60		enhancin (*Lymantria dispar nucleopolyhedrovirus*)
	M61		glycyl aminopeptidase (*Sphingomonas capsulata*)
	M64		IgA peptidase (*Clostridium ramosum*)
	M66		StcE peptidase (*Escherichia coli*)
	M72		peptidyl-Asp metallopeptidase (*Pseudomonas aeruginosa*)
	M76		Atp23 peptidase (*Homo sapiens*)
	M78		ImmA peptidase (*Bacillus subtilis*)
	M80		Wss1 peptidase (*Saccharomyces cerevisiae*)
	M84		MpriBi peptidase (*Bacillus intermedius*)
	M85		NleC peptidase (*Escherichia coli*)
	M90		MtfA peptidase (*Escherichia coli*)
	M91		NleD peptidase (*Escherichia coli*)
	M93		BACCAC_01431 g.p. and similar (*Bacteroides caccae*)
	M95		selecase (*Methanocaldococcus jannaschii*)

1.5 Malaria diagnosis

It is important to diagnose malaria early to reduce the disease symptoms and prevent the complications which may lead to death. Different tools from different commercial kits are now available for accurate diagnosis of malaria in a short period of time. It is also important to identify the correct *Plasmodium* species as the choice of treatment options depends on the *Plasmodium* species [21].

Light microscopy could be used to diagnose malaria by obtaining well-stained thick and thin films, whereby the thick film is used to improve diagnosis sensitivity while the thin film is better for species identification. The sample should be prepared for examination with light microscope immediately after collection. This should be done to minimize deformation of parasite and erythrocytes [27].

Rapid diagnostic tests (RDTs) can be used to detect *Plasmodium* parasites by using monoclonal antibodies specific to their antigens. Mainly RDTs should be used as an alternative to

microscopy diagnosis when high-quality microscope diagnosis cannot be done. The advantage of using RDTs includes simplicity, ease-of-understanding and interpretation; they do not require electricity and generate rapid results. Usually, it takes 15 minutes to get the result [28].

There are different *Plasmodium* parasite antigens available for use in RDTs, which include histidine-rich protein, parasite lactate dehydrogenase, and *Plasmodium* aldolase. Based on the antigen used, the RDTs can detect single species - usually *P. falciparum* or *P. vivax*. Other RDTs can detect all malaria parasites. Now in the market, there are more than 200 RDTs specific to malaria (the complete list of those RDTs can be found on http://www.who.int/malaria/news/2016/rdt-procurement-criteria/en/) [27].

The Polymerase chain reaction PCR has been used to detect *Plasmodium* species by targeting the 18s rRNA [27] and by using Nested PCR it is possible to distinguish between different *Plasmodium* species with high sensitivity and specificity [29].

It is recommended to use RDTs or PCR to diagnose malaria because the accuracy of diagnosis by microscopy depends on the level of the parasite in a blood sample. Moreover, now there is a wide range of commercially-available RDTs that offer higher accuracy and faster results, but they cannot detect how many parasites are in the host.

1.6 Malaria treatment

Antimalarial drugs have different goals, including (1) targeting the asexual cycle of the erythrocyte stage, (2) the prevention of recurrent infections and (3) the prevention of parasite transmission. The choice of a particular antimalarial drug is largely dependent on the *Plasmodium* species concerned. For example, *P. vivax* requires special treatment strategies because it can form dormant hypnozoites. Another factor to consider is the stage of infection - if it is complicated or severe, then a different treatment approach is required as opposed to early-diagnosed malaria. Hence no single drug can accomplish all goals while achieving antimalarial drug resistance. A solution is to use a combination of the different antimalarial drugs to achieve complete elimination of the *Plasmodium* parasite from the body. Drugs targeting the asexual cycle are called blood schizonticidal drugs, while those targeting the sexual cycle are called gametocytocidal [30].

There are three main groups of antimalarial drugs which include quinolines, antifolates, and artemisinin derivatives. Quinoline derivatives usually accumulate in the plasmodial digestive vacuole and prevent degradation of hemoglobin. Examples of quinoline derivative drugs

include chloroquine, quinine, mefloquine, and primaquine. The only drug that belongs to the quinoline derivatives but has a different mode of action is atovaquone, which interacts with the respiratory pathway of the parasite to inhibit parasite growth. Antifolate derivatives inhibit folate biosynthesis by different ways, including the inhibition of dihydropteroate synthetase or dihydrofolate reductase. Examples of antifolate derivative drugs are sulfadoxine and proguanil [31]. Artemisinin derivatives depend upon the production of carbon-centered free radicals. Artemisinin is toxic to malaria parasites because it targets hemoglobin molecule [32]. In addition to the previous main three antimalarial drug categories, there are antibiotics and other new antimalarial drugs. These include for example clindamycin, which inhibits the protein synthetic pathway [33].

Due to increasing levels of malarial parasite resistance to sulfadoxine/pyrimethamine and chloroquine, a combination of different antimalarial drugs with different modes of action is currently used, however there is still a high need for new drugs with new targets. World Health Organization (WHO) recommends artemisinin combination therapies as treatment for chloroquine-resistant *Plasmodium* parasites and uncomplicated malaria. In the case of severe malaria the recommended treatment includes a combination of artesunate, artemether, and quinine [30].

1.7 Antimalarial drug resistance

Aminoquinoline chloroquine was one of the favorable antimalarial drugs due to its efficacy and low side effects. However, since 1957 the *Plasmodium* parasite has started to develop resistance to this drug, and now the resistance has reached so many areas in the world that chloroquine is only effective in Central America [34]. In South East-Asia *P. falciparum* has started to develop resistance to the last available treatment which is artemisinin [35]. Another antimalarial drug, amodiaquine, which was more efficient than chloroquine has been used as an alternative where the parasite has already developed resistance to chloroquine. However, the *Plasmodium* parasite has later developed resistance to this drug as reported in Tanzania and Africa [36]. Currently artesunate-mefloquine is used as first-line treatment. To decrease the chance of developing resistance to this drug, WHO recommends using this drug with a combination of any other drug having a different mode of action. However, the failure rate for this combination is less than 10%, which raises global health concerns because the *Plasmodium* parasite that develops resistance to this combination could lead to a global outbreak[34].

Most drug resistance comes from a genetic mutation. It begins with a genetic mutation that gives the parasite the ability to survive in the presence of the drug. Then the resistant parasite multiplies and grows to lead to a parasite population resistant to the drug. These genetic mutations could be single point mutations or occur most commonly as multiple mutations. A complication happens when cross-resistance occurs. Cross-resistance means that if the parasite becomes resistant to a specific drug, it also becomes resistant to all drugs of the same chemical family or to those having the same mode of action, for example resistance against both halofantrine and mefloquine. Another factor that can lead to drug resistance is the drug half-life. As its half-life increases, the chance of developing drug resistance increases as the parasite encounters lower concentrations that are not enough to kill them, thus giving time for drug resistance to develop [34].

There are several reported mutations associated with antimalarial drug resistance. For example, mutations in *P. falciparum* chloroquine resistance transporter (*Pf*crt) have been associated with chloroquine resistance. The main mutation occurs in position 76 in which lysine changes to threonine; other mutations in the same protein include C72S, M74I, N75E, A220S, Q271E, N326S, I356T, and R371I. Those mutations are associated with the main mutation to give resistance to chloroquine [31].

1.8 Malaria vaccine

To control malaria, different vaccines have been developed to eliminate malaria and protect healthy humans. Based on the *Plasmodium* parasite life cycle stages, malaria vaccines can be divided into three main groups: pre-erythrocyte, erythrocyte, and other vaccines. In pre-erythrocytes, the goal is to prevent sporozoite from invading hepatocytes. This can be achieved with the help of both T-cells and the humoral response. Pre-erythrocyte vaccines target the circumsporozoite protein (CSP). The CSP antigen prevents sporozoites from invading hepatocytes. Due to its low immunogenicity, the vaccine RTS,S was developed. RTS,S was developed by PATH Malaria Vaccine Initiative (MVI) and GlaxoSmithKline (GSK) and is also commercially known as Mosquirix. RTS,S consists of hepatitis B surface antigen fused with CSP and a liposome-based adjuvant. RTS,S has reduced the number of infected children by almost 50% [37].

In the erythrocyte malaria vaccine, the goal is to prevent the merozoites from invading erythrocytes, and to prevent death and disease without complete prevention of infection. The targets are antigens expressed on the merozoites' surface or on that of infected erythrocytes.

These include the merozoite surface protein, glutamate-rich proteins and the apical membrane antigen 1 [37], [38].

1.9 Problem statement and hypothesis

It is important to develop new antimalarial drugs for alternative malaria targets due to the declining efficacy of available antimalarial drugs, as well as the development of drug resistance. The erythrocyte stage is mainly responsible for the symptoms of malaria, and it is the main source of amino acids for the *Plasmodium* parasite. Therefore, the erythrocyte stage has become the most targeted stage for antimalarial drug design. During this stage, especially during the asexual cycle, *Plasmodium* parasites use different proteases to degrade erythrocyte hemoglobin. About 65% to 75% of erythrocyte hemoglobin is digested, which results in the release of free amino acids. These proteases include aspartic proteases, falcilysin, plasmepsins, cysteine proteases, metalloproteases, dipeptidyl aminopeptidase 1 (DPAP1), falcipains and exo-aminopeptidases. One of the exo-aminopeptidases used by *Plasmodium* parasite is M1 Alanyl aminopeptidase. M1 Alanyl aminopeptidase is a zinc-dependent protease involved in the terminal stage of hemoglobin degradation and in the release of amino acids. Since the *Plasmodium* parasite cannot synthesize its own amino acids, inhibition of this enzyme has the potential to block *Plasmodium* parasite growth. M1 Alanyl aminopeptidase shares high sequence identity among different *Plasmodium* species, which makes it possible to use the same drug against different *Plasmodium* species.

1.10 Aim and objectives

The main aim of this study was to use structural bioinformatics tools to identify potential inhibitors against M1 alanyl aminopeptidase. To achieve this, homology modelling of *P. falciparum* M1 alanyl aminopeptidase and its homologs from other *Plasmodium* species was performed. To identify potential inhibitors, compounds from the South African National Compounds Database (SANCDB) and selected compounds from the ZINC and PubChem databaseswere screened *in silico* against these proteins. . Finally, top selected ligands were evaluated to ensure they selectively bind to the M1 alanyl aminopeptidase from *Plasmodium* species.

Chapter 2 – Sequence Analysis

Plasmodium M1 alanyl aminopeptidase could be considered a possible drug target against malaria. M1 alanyl aminopeptidase is present in different species including *Homo sapiens* and plays an essential role in the degradation of peptides, resulting in the release of free amino acids. Due to the presence of a human homolog, we need to analyze *Plasmodium* M1 alanyl aminopeptidase as well in order to highlight the difference between them. This chapter focuses on the analysis of the *H. sapiens* M1 alanyl aminopeptidase and its homologs in *Plasmodium* species, including *P. vivax, P. knowlesi, P. ovale, P. malriae,* and *P. falciparum.* These analyses include motif analysis, multiple sequence alignment and phylogenetic tree generation. The purpose of these analyses is to identify sequence and structural differences between the *Plasmodium* M1 alanyl aminopeptidase and human homologs, which may help in improving the specificity of the identified compounds against the *Plasmodium* M1 alanyl aminopeptidase protein.

2.1 Introduction

Sequence analysis is important to understand sequence features, conserved regions, motifs associated with functions, homology, sequence diversity between similar sequences and is an important part of structural analysis. Sequence analysis entails various techniques, such as motif analysis, multiple sequence analysis and phylogenetic analysis.

2.1.1 Motif Analysis

Motifs are short sequences with conserved patterns among different homologous sequences and through the evolution. Sequence motifs vary from DNA to amino acid short sequences depending on the sequence. Sequence motif lengths range from 3 letters to 50 letters depends on the motif type [39]. Motifs can be an indicator of a protein binding site and interaction domains, such as restriction enzyme binding sites, or transcription factor binding sites, regulatory regions on DNA, termination sites or active sites. Motifs can fall into two categories: they can be structural motifs or sequence motifs. A structural motif located in the exon region of a gene will also be in the encoded amino acid sequence [40], while a sequence motif would only be found in the intron region of a gene. All structural motifs are sequence motifs, but not all sequence motifs are structural motifs. Based on the toll used in motif analysis, motifs could be showed as sequence logo Figure 2-1. A sequence logo is a representation of a conserved region across analyzed sequences, in which the letter height corresponds to the amino acid conservation[40].

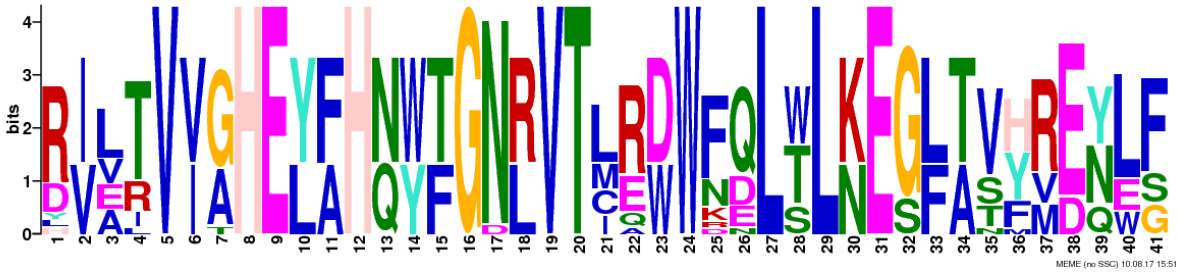


Figure 2-1: Example of a sequence logo. It shows the frequency of occurrence of each amino acid in the analyzed sequences on Y-axis against the amino acid letter, shown on the X-axis.

Currently, there are different motif analysis tools. Some are specialized in motif analysis only while others do additional sequence analysis. Examples include the Multiple Em for Motif Elicitation (MEME) [41] tool, the Regulatory Sequence Analysis Tools (RSAT) [42] and the Protein Family Database (Pfam) [43].

2.1.1.1 Multiple EM for Motif Elicitation suite

MEME is a toolkit that can be used either via its web server interface or by installing it locally for use as a command line tool. This software contains different tools covering different motif analysis types, including discovery and searching of motifs, comparing discovered motifs with known motifs and correlating previously known functions with discovered motifs [41].

For motif discovery, the user should input different sequences in unaligned (ungapped) format. These sequences should share some sequence similarity, for example, all the sequences should be orthologous, or they could have similar domains. Then MEME searches for motifs using different algorithms including the expectation maximization algorithm, the maximum likelihood and greedy search [44].

The ideal input sequence length should be less than 1000 bp, which means it is inefficient at analyzing large data sets. Both repetitive DNA elements and low information segments should be eliminated before submitting the sequences to motif analysis. It is easier and better to carry out motif analysis with protein sequences than with DNA sequences. This is because the protein alphabet consists of 20 amino acids while the DNA alphabet consists of 4 nucleic acids, which gives more significant results for motifs discovered from proteins. The same criterion applies to protein sequences as it should be free from low complexity regions. The MEME guide suggests using the SEG program to remove low complexity regions from protein sequences and the RepeatMasker program with DNA sequences [44].

As some motifs may contain insertions and deletions, the MEME suite includes a gapped local alignment of motifs (GLAM2) tool to discover gapped motifs [45]. It is highly recommended

to do an ungapped motif analysis after a gapped motif analysis and compare the results to avoid false positives[41].

Depending on whether you are working with DNA or protein sequences, there are different ways to analyze the motif results, as shown in Figure 2-2. It includes comparing the resulting motifs against known ones, such as known regulatory motifs, identifying corresponding GO (Gene Ontology) annotations and identifying additional motif occurrence for the desired motifs. Unfortunately comparing the resulting motifs with known ones is available for DNA motifs only [46].

The MEME suite became online in 1996, and has now become an essential tool for motif analysis, offering 13 different tools with different features including motif discovery and enrichment, and database comparison [46].

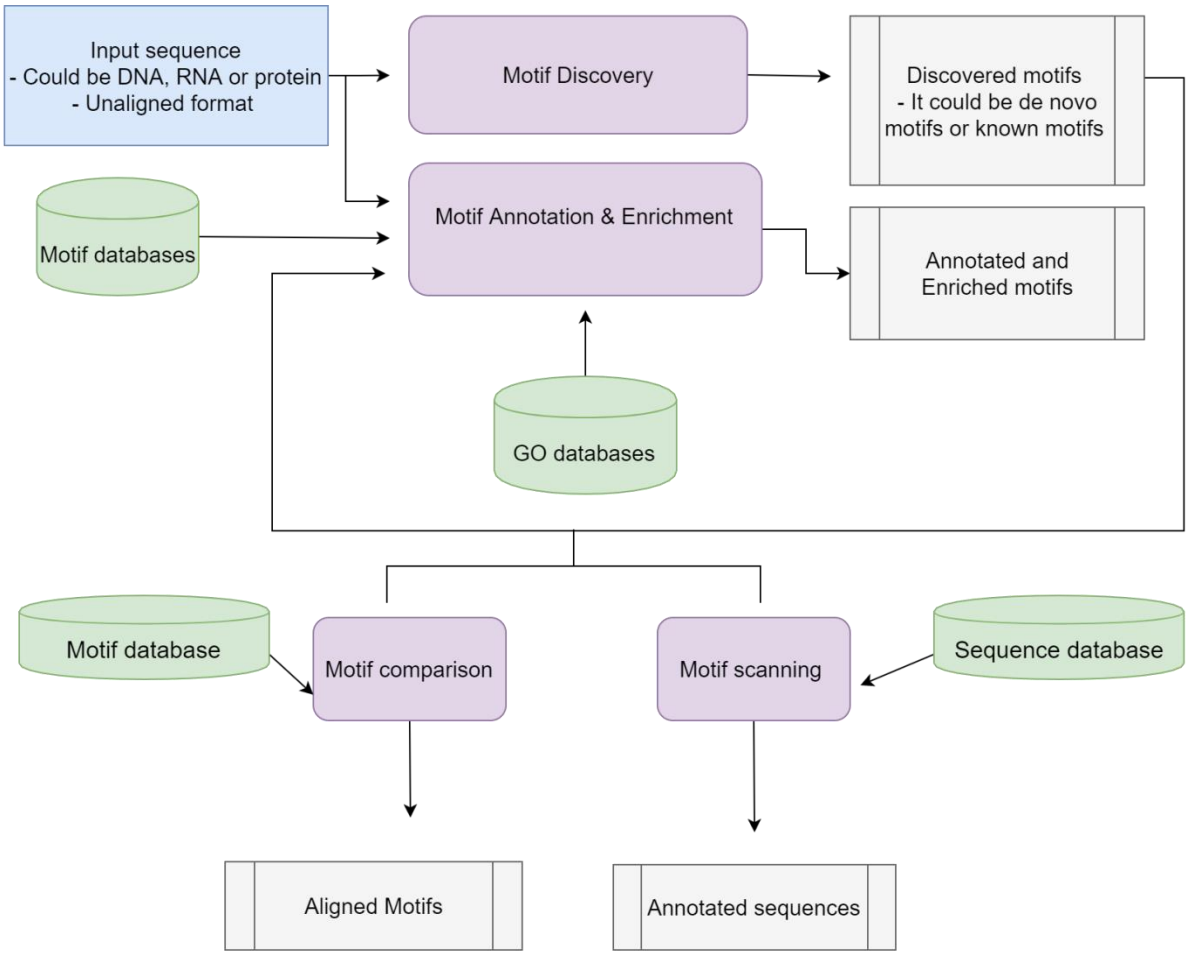


Figure 2-2: Summary of the MEME suite features and a suggested workflow with the output of each analysis.

For motif discovery, there are four different algorithms. The first and oldest one is MEME, which performs basic motif discovery from both DNA and protein sequences. The MEME algorithm is limited by being poor at finding short DNA motifs. The second algorithm is

implemented by the discriminative regular expression motif elicitation (DREME) tool, which was developed to produce more sensitive motifs, especially in the case of short motifs, as opposed to MEME [47]. Both MEME and DREME cannot discover ungapped motifs; to overcome this limitation GLAM2 developed. The last tool, termedMEME-chip, was developed to overcome the length limitation as a means to discover motifs from whole genome sequences. MEME, DREME and GLAM2 can only do motif discovery, while MEME-chip performs enrichment and comparative analyses as well as motif discovery, also giving a full report in comparison to the other tools [39].

2.1.1.2 Pfam

The first developed Pfam database was released in 1997 [48]. Currently, Pfam 31.0 was released on March 8, 2017, with 16712 protein families and 604 clans. Pfam is a multiple sequence alignment and a hidden Markov model representation of different protein families. It uses clans to organize its data, in which all related sequences are grouped as clans based on sequence similarity, sequence structure and profile. Pfam also takes advantage of protein domains in order to infer possible protein function. Initially a seed alignment is created for each protein family. This seed alignment used to train a hidden Markov model profile using the HMMER software. Clan quality heavily depends on seed quality. This model is then used to search against a large dataset to identify all possible homologous sequences [49]. Pfam can be used to identify the protein family of an input sequence by searching Pfam stored models, which help in identification of protein sequences and homologs sequences [50].

2.1.2 Sequence alignments

Sequence alignments are used to compare two or more nucleic acid or amino acid sequences to identify a conserved region in the sequences that may correspond to a function or an evolutionary relationship. Based on the number of aligned sequences, the alignment can be a pairwise alignment or a multiple sequence alignment [51]. Both approaches are performed using global alignment or local alignment. In global alignment, the aim is to create an end-to--end alignment, which includes the entire length of the sequences being aligned. Local alignments, on the other hand, aim to identify the most similar regions between aligned sequences as shown in Figure 2-3.

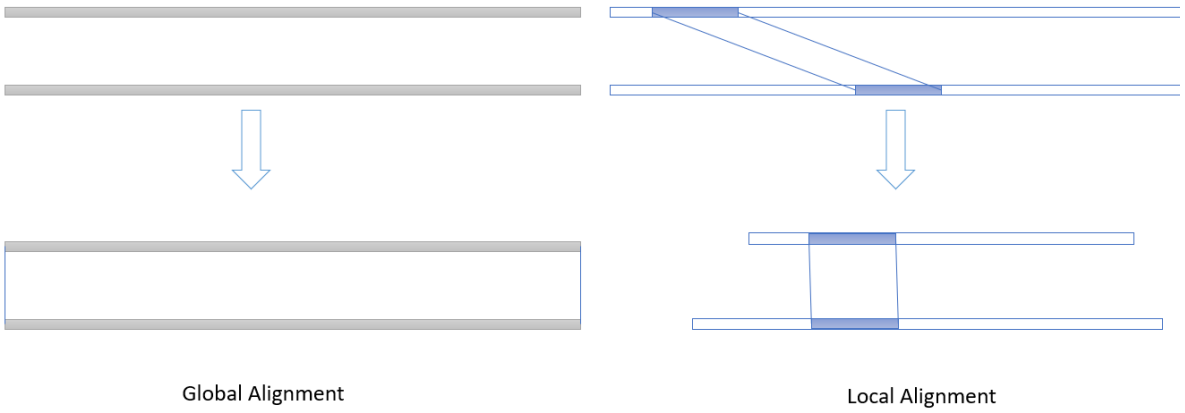


Figure 2-3: Global and local alignment. It shows the difference between a global alignment and a local alignment. In a global alignment, the entire first sequence aligned with the entire second sequence (end-to-end alignment), while in local alignment only includes the most similar parts [52].

2.1.2.1 Pairwise sequence alignment

A fundamental objective of bioinformatic analysis is used to find the best match between two sequences. There are many different methods that have been developed to perform pairwise alignment. The most common methods consist of dot matrices, dynamic programming and the word method [53]. Dynamic programming is very accurate but requires high computational power because it calculates all possible alignments between the query sequence to choose the one with the highest alignment score. This approach is highly impractical for very large genomic sequences [54].

The pairwise alignment algorithm uses comparison matrices to evaluate the significance of any match or mismatch. These matrices define a score for every possible match; the algorithm uses these scores to find the best total score for the alignment between aligned sequences. In DNA or RNA sequence alignment, the most common scoring matrix is the identity matrix. For protein sequence alignment, the most common matrices are the point (or percent) accepted matrix (PAM) [55] and block substitution matrix (BLOSUM) [56]. The identity matrix is very simple - it gives a value of one for a positive match and zero for a mismatch. The simplicity of this matrix lowers the computational cost needed for alignment calculation while at the same time it does not provide weights for insertions and deletions for the aligned sequences [53].

PAM matrices measure the likelihood of a mutation that occurs between homologous sequences, in which one amino acid changes to another specific amino acid during evolution. As a result, PAM matrices are based on the mutational model. BLOSUM matrices measure amino acid conservation and substitution probabilities in protein families (blocks) which are based on a starburst model. Therefore PAM matrices are very useful in evolutionary studies

while BLOSUM matrices are mainly used to find conserved domains. It is better to choose BLOSUM for local alignment [57].

2.1.2.2 Multiple sequence alignment (MSA)

A multiple sequence alignment (MSA) is an extension of pairwise alignments whereby three or more query sequences are aligned. An MSA is very important for evolution studies and phylogenetic tree construction. It improves the accuracy of the identified conserved residues and increases the ability to correctly identify insertions and deletions. MSA is important in many bioinformatics applications including secondary structure prediction, homology modelling, motif finding and phylogenetic analysis [54].

MSA is mainly performed using heuristic methods or exhaustive methods. The exhaustive methods like dynamic programming for MSA are highly impractical. Heuristic alignment methods have two common methods for MSA. These comprise the progressive alignment and iterative alignment Figure 2-4. Progressive alignments are usually fast but the accuracy is not guaranteed. In progressive alignment, errors that occur in any step of the algorithm are retained and carried over to the final step. The alignment starts by aligning the most similar sequences together, then the algorithm adds more sequences to this alignment until all query sequences are aligned [58]. The iterative method starts with a low-quality alignment then the algorithm iterates and improves the alignment until no improvement can be made to the alignment. The main idea of the iterative method is to continuously modify the alignment until an optimal alignment is produced [59].

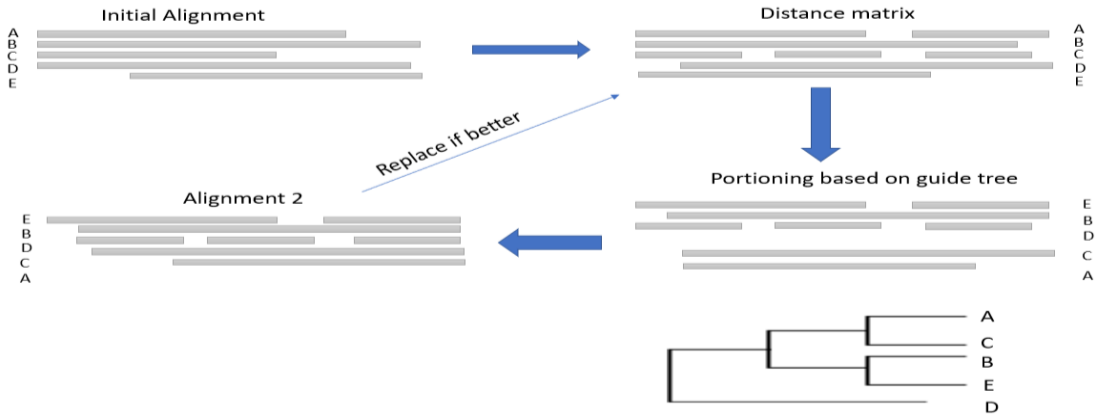


Figure 2-4: Steps for the iterative alignment method [60].

2.1.2.3 Phylogenetic analysis

Phylogenetic analysis is the representation of the evolutionary relationship between various species in the form of a branched diagram. The possible evolutionary relationship are constructed based on the physical or genetic differences and similarities [61]. The phylogenetic

tree could be a species tree or a gene tree. A species tree represents the evolutionary relationships between species or groups of the population while a gene tree measures the phylogenetic relationships between a group of homologous genes. Phylogenetic trees can be rooted or unrooted. A rooted phylogenetic tree is branched from a unique node that represents a common ancestor while in the unrooted tree there is no single common ancestor [62].

While a phylogenetic tree can be useful in understanding the history of evolutionary events, it can be biased if the input data is noisy or not accurate. Another limitation is using a small input set. For example, the construction of a species tree based on sequence similarity between conserved genes (ex: housekeeping genes) could be limited. This due to the fact that this tree is based on a single gene which may not reflect the complete organism genome [63]. The more genes used in the analysis, the more reliable the resulting phylogenetic tree. Using a small set of input genes results in a phylogenetic tree that requires further validation through techniques such as bootstrapping and the use of an outgroup [64]. An outgroup is group of distantly related sequences in a set of input genes. This outgroup acts as negative control which should appear near the root [64]. Bootstrapping includes pre-defined iterations meant to increase the confidence of the phylogenetic tree. In each iteration, the input MSA is randomly permuted then the phylogenetic tree is calculated. In the end, each branch of the final tree is labeled with a number. This number represents how many times the branch was recovered across all the iterations [65]. Different algorithms could be used to construct a phylogenetic tree. These include the Unweighted Pair Group Method with Arithmetic Mean (UPGMA), Neighbor-joining (NJ), Maximum parsimony (MP) and Maximum-likelihood (ML). UPGMA is based on a distance matrix calculated from an MSA and a constant evolution rate [66]. NJ is bottom-up clustering method suitable for large datasets since the algorithm is fast; however, accuracy is not guaranteed [66]. MP tries to produce a phylogenetic tree that minimizes the number of steps needed to reflect the variation between the sequences and the common ancestral sequence [67]. ML is based on a statistical approach and is very optimal for small input data of distantly-related sequences. However ML is not the best choice for large input data because it is computationally expensive [68].

2.2 Methods

2.2.1 Sequence retrieval

The M1 *P. falciparum* alanyl aminopeptidase sequence (accession number XP_001349846.1) was retrieved from the National Centre for Biotechnology Information (NCBI) using the NCBI global cross search tool. The retrieved sequence was submitted to pBLAST [69] (protein Basic Local Alignment Search Tool) to retrieve M1 alanyl aminopeptidase from other *Plasmodium* species including *P.vivax, P. knowlesi, P. ovale, P. malriae, P. gaboni, P. reichenowi and P. coatenyi* (the accession number for each species is shown in Table 2-1) using default BLAST parameters and restricting the organism search to *Plasmodium* species. The Ensembl genome browser [70] was used to retrieve a human sequence of M1 alanyl aminopeptidase (accession number NP_001141.2). The Ensembl orthologs finder was used to retrieve mammalian homologs. Bacterial homologs of the M1 *P. falciparum* alanyl aminopeptidase sequence were retrieved from UniProtKB [71] using UniRef [72] data available from the M1 *P. falciparum* alanyl aminopeptidase record found in the UniProtKB. At the end, 18 sequence were retrieved (shown in Table 2-1)

2.2.2 Motif analysis

2.2.2.1 Pfam

In order to investigate the relationship between human protein and *Plasmodium* sp. Homologs, the HMMER tool [73] was used to search for protein families and domains in all retrieved sequences. HMMER was accessed through https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan, and a FASTA-formatted file containing all the retrieved sequences was submitted to the HMMER tool using default parameters.

2.2.2.2 MEME suite

A locally-installed version of the MEME suite was used to discover motifs in all retrieved sequences. MEME version 4.12.0 was downloaded from the MEME official site and installed locally, after which a FASTA-formatted file containing all retrieved sequences was submitted to the MEME tool using default parameters. The generated files were submitted to the MAST tool [74]. The motif width was set to 6 as minimum and 50 as the maximum value. Moreover, repeated motifs were set to be skipped and 10000 was used as the maximum number of discovered motifs.

2.2.3 Sequence alignment

MSA was performed using MAFFT [75] for all retrieved sequences. MAFFT was accessed through the MAFFT online web server hosted on EBI servers https://www.ebi.ac.uk/Tools/msa/mafft/ and a FASTA file containing all sequences was submitted using MAFFT default parameters. A separate MSA for all *Plasmodium* species sequences was also created. Structural MSA was performed using T-Coffee expresso, which can be accessed through http://tcoffee.crg.cat/apps/tcoffee/do:expresso. A FASTA file containing all sequences and a 3D structure sequence of *Plasmodium* M1 alanyl aminopeptidase (PDB ID: 3Q43) were submitted using default parameters.

2.2.4 Phylogenetic analysis

MEGA 7 [76] was used to generate a phylogenetic tree representing the evolutionary relationships between *Plasmodium* sequences and their homologous sequences. The T-Coffee expresso alignment result was used as input to generate the phylogenetic tree using the Neighbor-joining algorithm. All gaps were eliminated and 1000 bootstrap iterations were used to increase the phylogenetic tree confidence. The evolutionary model was measured and selected based on the best BIC (Bayesian Information Criterion) score obtained using the MEGA goodness of fit test. The selected model was the “Le Gascuel” (LG) statistical model.

2.3 Result and Discussion

2.3.1 Sequence retrieval

Different *Plasmodium* species and homologs sequences were retrieved from different databases (NCBI, UniProtKB, and Ensembl). The retrieved sequences were confirmed using other databases, including PlasmoDB. The retrieved data was checked and compared against available data. All the retrieved sequences were submitted to BLAST to measure the percentage similarity between *P. falciparum (*XP_001349846*)* and all other retrieved sequences.

Table 2-1: Summary of M1 alanyl aminopeptidase *P. falciparum* sequence and its retrieved orthologs.

Name	Length (aa)	Accession number	Query Cover	Identity percentage
P. malariae	1100	SBS90191	100%	72%
Plasmodium gaboni	1080	XP_018639924.1	100%	97%
Plasmodium reichenowi	1087	CDO65912	100%	99%
P. knowlesi	1097	XP_002262014.1	99%	71%
P. ovale	1078	SBT47239	100%	73%
P. vivax	1097	SCO69705	99%	72%
Lactobacillus delbrueckii	843	WP_011544314.1	40%	25%
Klebsiella pneumoniae	870	CDK69214	81%	35%
Escherichia coli	870	WP_069905499.1	81%	35%
Shigella sp.	870	WP_094320956.1	81%	35%
Salmonella typhimurium	914	WP_069905499.1	82%	36%
Homo sapiens	967	NP_001141.2	42%	26%
Gorilla gorilla	967	XP_018866310	42%	27%
Macaca mulatta	968	XP_001093727.2	39%	27%
Sus scrofa	963	P15145.4	35%	26%
Trypanosoma grayi	869	XP_009314710	47%	25%
Trypanosoma theileri	868	ORC86065	51%	26%
Trypanosoma cruzi	870	XP_809697.1	39%	26%

M1 alanyl aminopeptidase 3D structures for both human and *P. falciparum* were retrieved from the Protein Databank (PDB) [77]. Currently, there are 17 different PDB records for *P. falciparum* and 4 human ones. By comparing resolution values, PDB ID: 3Q43 was selected as 3D structure for *P. falciparum* M1 alanyl aminopeptidase while 4FYT was selected for the human.

2.3.2 Motif analysis

2.3.2.1 Pfam

Pfam was used to confirm the relationship between retrieved sequences and also to search for conserved domains in retrieved sequences. The results are shown in Table 2-2.

Table 2-2: Summary of Pfam results shows the start and end position of founded domains.

Family		Clan	Description	Organism	Start	End
Name	Accession n.					
Peptidase_M1	PF01433.19	CL0126	Peptidase family M1 domain	Human and Gorilla	296	543
				M.a mulatta	296	543
				S. scrofa	291	539
				P. malariae	397	635
				P. falciparum	406	643
				Plasmodium gaboni	401	638
				P. knowlesi	414	652
				P. ovale	395	633
				P. vivax	414	652
				P. reichenowi	408	645
				L. delbrueckii	197	434
				E. coli and *Shigella* sp and *Klebsiella*	207	440
				S. typhimurium	251	484
ERAP1_C	PF11838.7	n/a	ERAP1-like C-terminal domain	*M. mulatta*	620	947
				Human and Gorilla	619	946
				S. scrofa	616	943
				L. delbrueckii	505	820
DUF3458	PF11940.7	n/a	Domain of unknown function (DUF3458) Ig-like fold	*P. malariae*	641	736
				P. falciparum	650	745
				P. gaboni	645	740
				P. reichenowi	652	645
				P. knowlesi	658	753
				P. ovale	639	734
				P. vivax	658	753
				E. coli and *Shigella* sp and *Klebsiella*	444	545
				S. typhimurium	488	589
DUF3458_C	PF17432.1	CL0020	Domain of unknown function (DUF3458_C) ARM repeats	*P. malariae*	739	1099
				P. falciparum	748	1083
				P. gaboni	743	1078
				P. reichenowi	750	1085
				P. knowlesi	756	1095
				P. ovale	756	1095
				P. vivax	737	1076
				S. typhimurium	592	914
				E. coli and *Shigella* sp and *Klebsiella*	548	870

The results from the Pfam analysis show that all retrieved sequences belong to metalloproteases ("zincins") superfamily with ID: 55486 and the aminopeptidases superfamily (ID: 63737). All the sequences were found to have the M1 aminopeptidase domain, which confirms they have the zinc coordinating active site needed for aminopeptidase activity. This domain position is almost in the same position in each closely-related group. All the *Plasmodium* species have this domain, starting from amino acid number 390 to amino acid number 650, while the

mammalian group (Human, *Macaca mulatta,* and Gorilla) has this domain exactly in the same region (296-543) with the exception of *Sus scrofa*. This confirms the importance of the M1 domain for this protein as it is conserved among different species of different evolutionary distance.

2.3.2.2 MEME suite

To confirm Pfam results and understand the conserved and divergent regions in this protein among different organisms, retrieved sequences were submitted to the MEME suite for motif discovery. The result was then submitted to the MAST tool to align the discovered motifs with retrieved sequences. The result was shown as a heat map in Figure 2-5. The MEME results shows that motif number 1 was conserved in all retrieved sequences, which is located in the peptidase family M1 domain retrieved from Pfam database. It was observed that all sequences have three different motifs located in the peptidase domain. Other domains are otherwise divergent between different organisms.

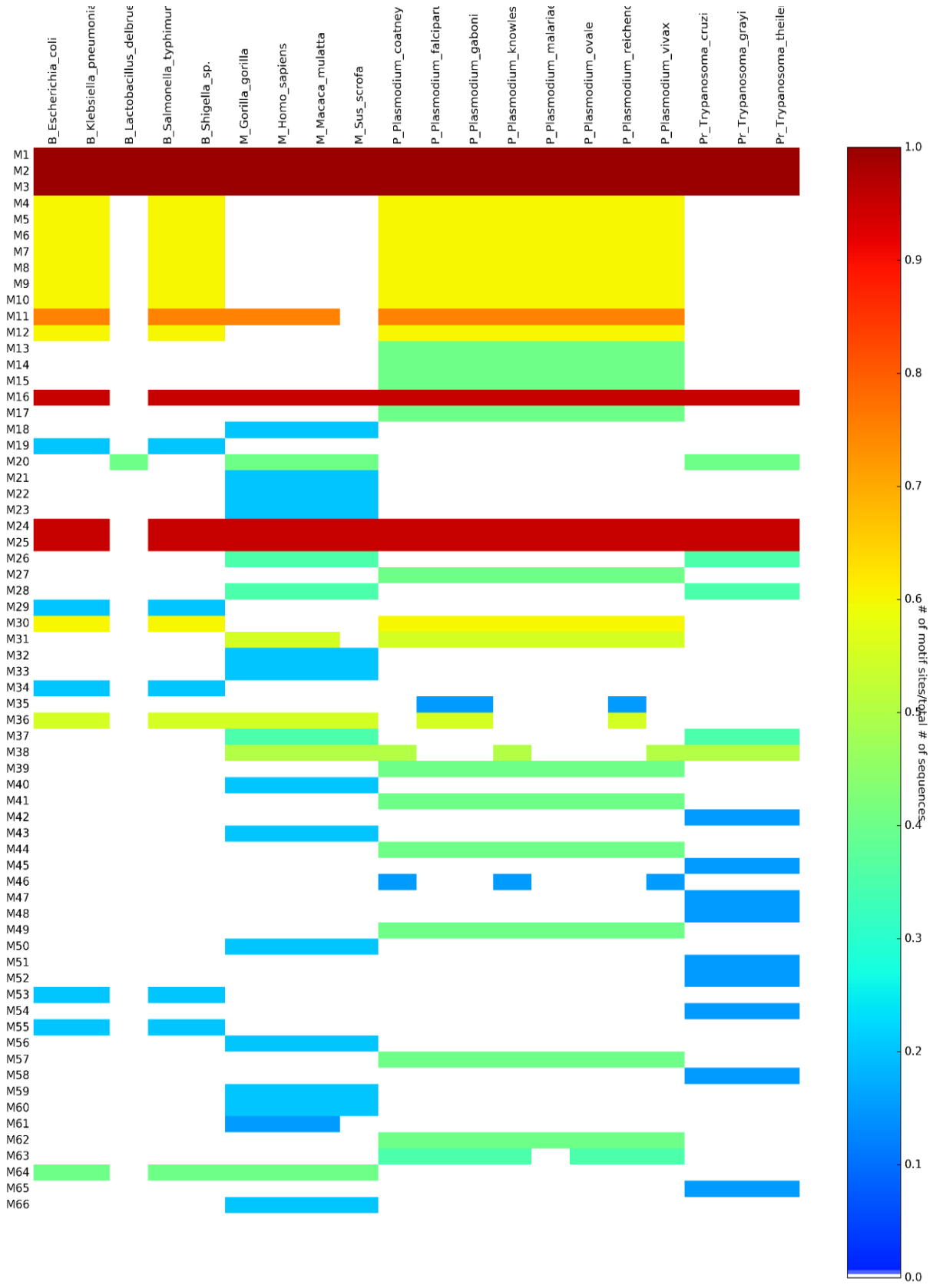


Figure 2-5: MEME heatmap, summarizing alanyl aminopeptidase motifs among all retrieved sequences.

2.3.3 Multiple sequence alignment

Different tools were used to perform MSA including MAFFT and T-Coffee expresso. 3D coffee takes the 3D structure of M1 alanyl aminopeptidase as an additional input, which increases the quality and accuracy of the resulting alignment. MSA was found to confirm the result obtained from motif analysis. This indicates that all sequences share a peptidase domain while the remaining sections of those sequences are more divergent in all the retrieved sequences, as indicated in Figure 2-7.

There are different degrees of conservation that can be observed from the sequence alignments. For *Plasmodium* species, there is high conservation between their aligned peptidase sequences, while there is more divergence in the N-terminal region with exception to *P. malariae*. The latter has an additional divergent part at the end of its sequence (Figure 2-6).

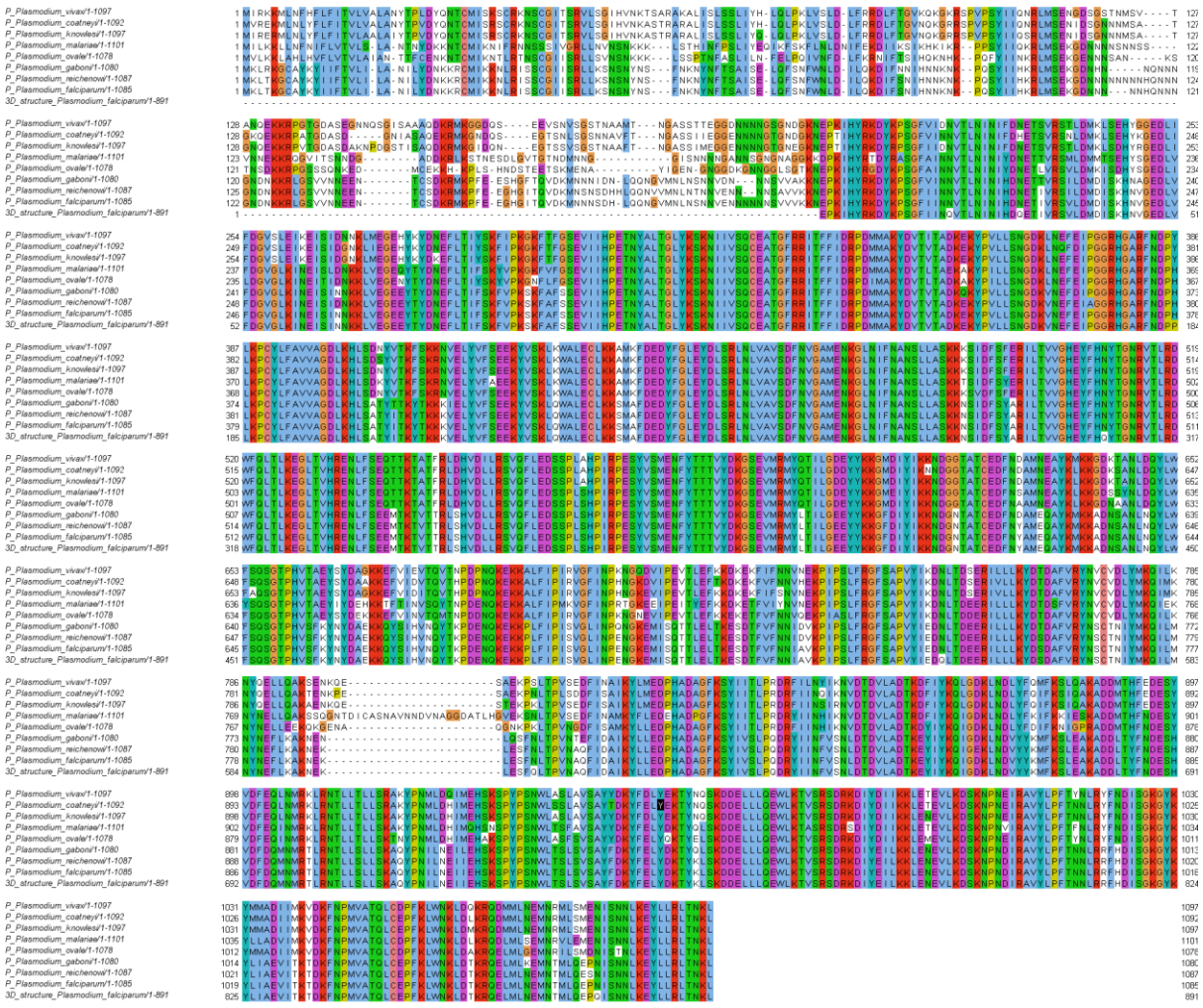


Figure 2-6: MUSCLE alignment result. The MSA alignment produced by MUSCLE for different *Plasmodium* sequences was viewed in the JalView software.

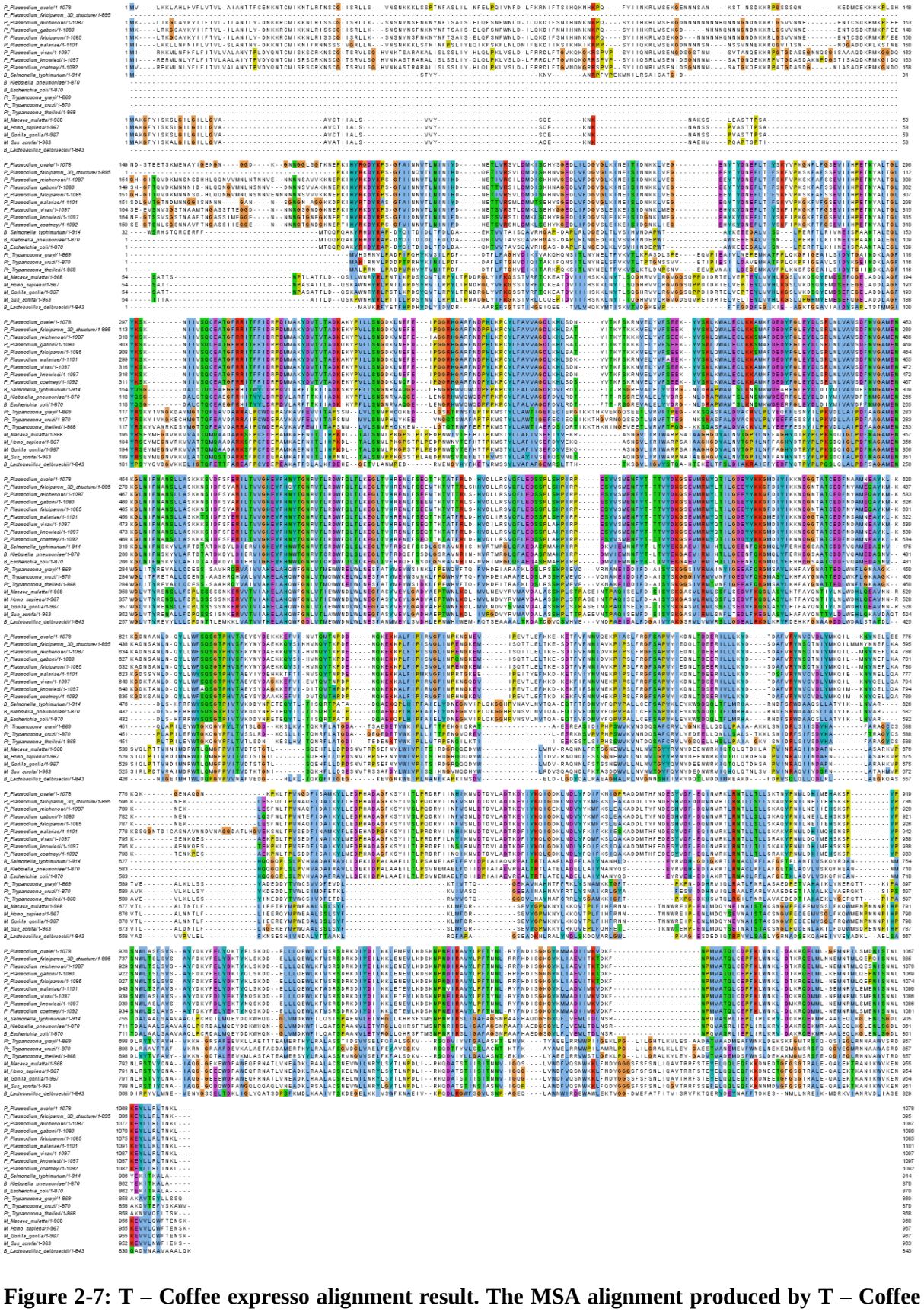


Figure 2-7: T – Coffee expresso alignment result. The MSA alignment produced by T – Coffee expresso for different *Plasmodium* sequences was viewed in the JalView software

Figure 2-8: The active site residues conserved in all retrieved sequences, are highlighted in violet.

As shown in Figure 2-8, the active site residues (histidine, histidine, and glutamine) are conserved in all organisms, while the flanking regions are similar between closely-related groups.

2.3.4 Phylogenetic tree

The phylogenetic tree was constructed to investigate the evolutionary relationship between *Plasmodium* M1 alanyl aminopeptidase and its homolog sequences. The model selection tool provided by MEGA was used to investigate the best evolutionary model according to BIC scores and the bootstrap consensus. As shown in Table 2-3, the top three models were all based on the "Le Gascuel" statistical model with different rates among sites.

Table 2-3: BIC scores of evolutionary models generated by the MEGA model selection tool.

Model	BIC score
LG+G+I	28762.72206
LG+G	28763.56086
LG+G+I+F	28871.46676
WAG+G+I	28906.10044
WAG+G	28932.00326
JTT+G+I	28943.83602
JTT+G+I+F	28951.95335
WAG+G+I+F	28955.33038
LG+G+F	28961.62294

The best phylogenetic tree was generated using the "Le Gascuel" statistical model combined with a gamma distribution with Invariant sites (LG + G + I), which is shown in Figure 2-9.

The generated phylogenetic tree showed clear species clustering in which all *Plasmodium* sequences clustered together as well as homologs from bacteria, fungi and mammals. Within the *Plasmodium* cluster the most similar sequence to *P. falciparum* was *P. richenowi,* followed by *P. gaboni* while *P. knowlesi, P. coatenyi and P. vivax* showed slight evolutionary distance to *P. falciparum.* This finding correlates with motif finding and MSA analysis as *P. knowlesi, P. coatenyi and P. vivax* share motifs 38 and 46, while *P. falciparum, P. richenowi* and *P. gaboni* did not have those motifs.

While the bacterial cluster was most similar to the *Plasmodium* cluster, the mammalian cluster was least similar to the *Plasmodium* cluster. This evolutionary difference points to the possibility of designing a drug with selective activity against *Plasmodium* species by targeting different regions.

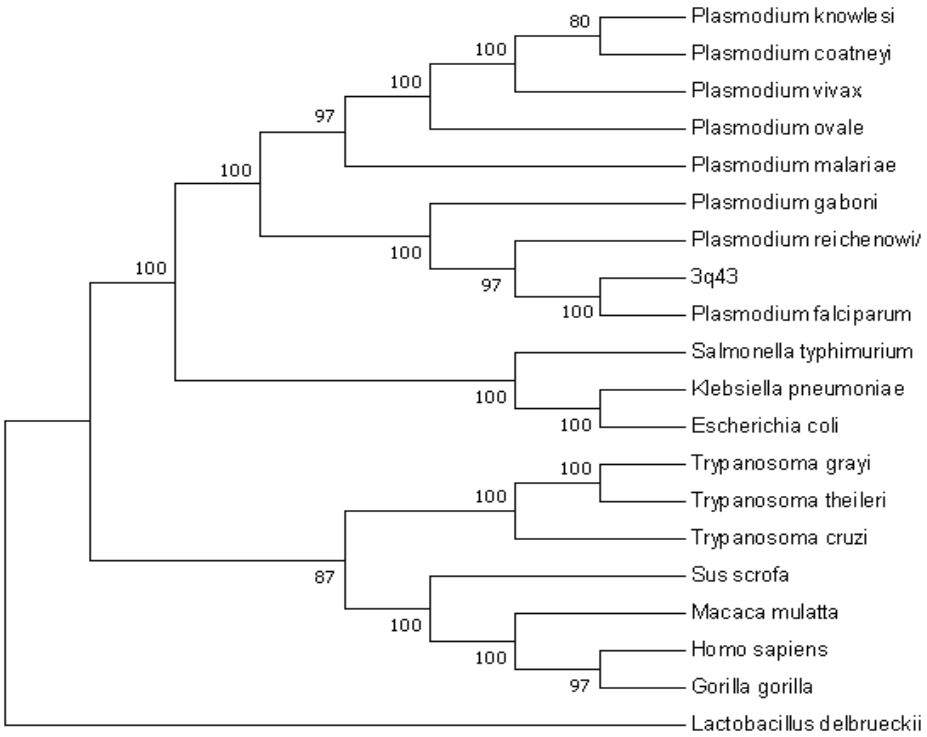


Figure 2-9: Molecular phylogenetic analysis by the Maximum Likelihood method generated by MEGA7. The generated tree based on the "Le Gascuel" 2008 model. A discrete gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 1.4895)). The rate variation model allowed for some sites to be evolutionarily invariable ([+I], 2.97% sites). All positions containing gaps and missing data were eliminated. There was a total of 767 positions in the final dataset.

2.4 Conclusion

This chapter includes an in-depth sequence analysis of *P. falciparum* M1 alanyl aminopeptidase and its homolog sequences from different organisms. These organisms include bacteria, fungi, and mammals. Seven M1 alanyl aminopeptidases from *Plasmodium* sequences, and 10 homolog sequences from other organisms were retrieved from NCBI, UniProtKB and Ensembl. All retrieved sequences were analyzed using different sequence analysis methods. This involved motif discovery, sequence alignments and phylogenetic tree calculations. Motif discovery shows that all sequences share the peptidase family M1 domain, which contains zinc coordinating residues. The domain position was mostly conserved within each species group. MEME motif analysis and MSA confirmed the conservation of metal-coordinating residues, including His 496, His 500 and Glu 519. Located near the entrance of the active site were conserved residues Glu 460, Ala 461, Met 462, Glu 463, Asn 464, Glu 466 and Leu 467. This highlights sequence diversity in the active sites, including Asn 501, Tyr 502, Thr 503, Arg 506, Arg 510, Asp 511 and Gln 514, which are conserved in all M1 alanyl aminopeptidases from *Plasmodium* species but are divergent when compared to human and other homologs. These sequence dissimilarities may indicate the presence of structural regions that may be exploited to obtain a selective drug against *Plasmodium* M1 alanyl aminopeptidase. An MSA was used to produce a phylogenetic tree to study the evolutionary relationships between the parasite and its host. The study shows the human protein and *Plasmodium* M1 alanyl aminopeptidase protein are distantly-related.

Chapter 3 - Homology Modelling

While the M1 alanyl aminopeptidase 3D structure of *P. falciparum* is available, the 3D structure of M1 alanyl aminopeptidase of other *Plasmodium* species is not yet determined. In this chapter, homology modelling techniques were used to generate 3D structures of the M1 alanyl aminopeptidase protein for other pathogenic *Plasmodium* species. The steps start with template identification and sequence alignment. Then model building is done based on the sequence alignment between the target sequence and template sequence for each species. At the end, the generated models underwent different refinement steps and validation tools were used to obtain the most accurate model. Then the generated models were used for the prediction of protein function and possible interactions with potential drugs.

3.1 Introduction

Homology modelling is a technique to calculate a 3D structure of a protein using related homolog proteins with experimentally-determined structures. Homology modelling could be done by using one (single template) or more (multiple templates) known structures. Thproduced obtained 3D structure can then be used in the determination of protein function, studies of disease-causing mutations and mutation impact on protein activity and in drug design experiments [77]. Currently, there are approximately 146,000 3D structure entries hosted on PDB [77], while there are over 550,000 protein sequences hosted on the UniProtKB/SwissProt database [78]. From this data, it is very clear the number of determined 3D structures is very low compared to the number of known sequences with unknown 3D structures, which emphasizes the need for using homology modelling. Homology modelling is based on the idea that homologous proteins share a similar 3D structural arrangement [79]. It starts with template identification and is followed by sequence alignment to highlight insertion, deletion, match and mismatch regions. Then the sequence alignment output is used to build the model. This model undergoes model refinement which includes loop refinement. Finally, the model is assessed for quality [80].

Homology modelling multi-steps:

3.1.1 Template identification

Template identification involves searching for all known structures using a query sequence to find its homologous structures. This includes a pairwise alignment between the query and a structure databases (e.g., PDB) by using alignment searching tools, for example the Basic Local Alignment Search Tool (BLAST) [81], [82]. BLAST gives a list of similar protein structures

based on sequence alignment. In order to get the optimal result, BLAST uses a residue exchange matrix and an alignment-matrix based on the latter. This is because we need to give a better score for residues that are easily exchanged for example, in the case of a Ile to Leu mutation, these residues should get a better score than residues that have different properties, while conserved residues with a specific function get the best score [83].

Template retrieval is an important step in homology modelling and to increase the sensitivity of template identification, evolutionary models and profiles can be used. Commonly-used profiles include sequence profiles and those based on Hidden Markov Models (HMM). The most common tools implementing these methods are PSI-BLAST [84] and HHpred [85], [86]. Once the final list of potential templates is obtained, it is necessary to select one or more templates. In order to filter the obtained list, the template with the highest sequence similarity to the input sequence is selected. Then the template sharing the same conditions as the input sequence is selected, for example, they might have the same solvent, pH, ligands and quaternary interactions. Finally, the resolution and R-factor of the template should be considered. It is preferred to choose more than one template to improve the model accuracy [87].

Depending on the identity percent between template and target, the best model is when the identity is greater than 90%. In this case, we can compare the model structure against experimentally-determined structures. If it is between 90% and 50%, it is considered to contain larger local errors. If it drops to 25%, it turns out to be the main bottleneck for homology modelling, which can often lead to very large errors [88].

3.1.2 Sequence alignment

Alignment errors are the main cause of deviations in homology modelling. Even when the correct template is chosen, alignment error can result from an incorrect insertion or deletion. Therefore, there is huge need to improve alignment result sensitivity. One of the suggested methods to improve the sensitivity is by using an iterative method to identify the template and generate the final sequence alignment to guide the model building process. It is also recommended to use an MSA to correct the alignment and highlight the features of the protein family and conservation degree. A change of Ala to Glu is possible, but unlikely to happen in a hydrophobic core, so this Ala and Glu cannot be aligned. By using an MSA program such as MUSCLE [89], Clustal Omega [90] or MAFFT [59], the residues and properties that must be conserved can be found [91].

In order to improve alignment quality, structural alignment can be used. This includes programs such as 3DCoffee [92] or PROMALS3D [93]. The idea of using a structural alignment tool is based on the conservation of structural configuration across homologous sequences. It is also suggested to manually optimize the final alignment to avoid any possible alignment errors. For example, a gap in structural element should be avoided. [80].

3.1.3 Model building

The main aim of this step to build the model based on the 3D template structure, such that the best models rely on alignments with the fewest possible errors. Based on the alignment, the model building tool copies the coordinates of the template residues to the residues in the input sequence if there is a match. In this case, it can include the side chain positions as well. While in the case of a mismatch, only backbone coordinates are copied. In the case of using multiple templates structures, errors can be fixed if the error is present in one template. as Additionally, the insertion and deletion present in one of the templates can be fixed by using the structural information of another template [86].

A variety of methods can be used to build a protein model for the target. Generally, rigid-body assembly, segment matching, spatial restraint, and artificial evolution are used for model building. Rigid-body assembly relies on the assembly of a complete model from conserved structural fragments identified from closely-related solved structures. Model accuracy is based on the template selection and alignment accuracy. Segment matching is based on dividing the target into the short part, then each part will be matched to its own template in PDB database. Modelling by the satisfaction of spatial restraints is based on the generation of many constraints or restraints on the structure of a target sequence, using its alignment to related protein structures as a guide. The most common tool used for this step is MODELLER [94], [95].

3.1.4 Structural refinement

This step includes improving and refining the alignment; loops and side chains are also modeled. It is very important to correct the alignment because it is the main source of error which results from misalignments. The correction and refinement of the alignment lead to changes in the backbone structure of the homology model. The changed backbone affects side chain which also leads to other changes on the backbone [96].

Gaps in sequence alignment can occur in the template sequence or target sequence. In both cases, this leads to conformational changes and problems in the modeled structure. Knowing that secondary structure elements remain conserved between homologous sequences, it is

preferable to insert any gaps in turn or loop structures. Loop modelling could be used to solve this problem. Loop modelling can be done using knowledge-based or energy-based methods [97]. The knowledge-based method depends on previously known structures hosted on PDB databases, in which the PDB database is searched for matched loops that have the same length and similar geometries. Then, the coordination of the best-matched loop is copied to model structure [98].The energy-based method depends on applying an energy function to assess the quality of the loop and modify its conformation to find the best conformation. In some cases, the produced loop could not fit properly to the modeled protein structure [99]. Side-chain modeling uses a combination of knowledge-based methods and energy functions to improve side chain conformation quality. The knowledge-based method is used to identify commonly known side chain conformations. Then an energy function is applied to select the best conformation. This could be computationally expensive in case of low-level similarity side chains [100].

3.1.5 Model validation

After the model building and refinement process is completed, it is important to check for errors in the model. Model quality depends on different factors including the percentage identity between the template and target sequence, and the alignment quality. Also, errors in the template itself should be considered [100]. Different tools are available to assist and validate model quality. These tools can be used to either validate the whole protein in addition to other tools that validate individual regions of proteins. These tools assist in the evaluation of protein stereochemistry, z-DOPE score estimation, geometry and residue fitness. Examples of commonly-used tools are PROCHECK, WHATIF, VERIFY3D, GRASP2, ANOLEA, and PROSAII. Also, it is very important to manually inspect any error present in the model [101].

3.2 Methodology

3.2.1 Template identification

HHpred (https://toolkit.tuebingen.mpg.de/#/tools/hhpred) and BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) were used to identify the best available template. A sequence retrieved from NCBI (accession number XP_001349846) was used as the M1 alanyl aminopeptidase *P. vivax* input for both tools. The PDB proteins were used as database search set for BLAST. The other BLAST parameters included an E-value threshold of ten, a word size of six, use of the BLOSUM62 scoring matrix and a gap penalty of 11 for new gaps and a gap extension value of 1. The HHpred parameters employed HHblits uniprot20_2016_02 as the MSA generation method, a maximal number of three MSA generation steps, an E-value

threshold of 1e-3, the local alignment mode and a MAC realignment threshold of 0.3. The same step with the same parameters were used for other *Plasmodium* species, where sequence accession number and the species names are shown in Table 3-1.

Table 3-1: Summary of *Plasmodium species* and their corresponding accession number

Name	**Accession number**
P. malariae	SBS90191
P. knowlesi	XP_002262014.1
P. ovale	SBT47239
P. vivax	SCO69705

3.2.2 Sequence alignment

The best-selected templates combined with the target sequence and homologous sequences were submitted for structural alignment using 3D-Coffee with default parameters. The final MSA output was refined manually for alignment errors by eliminating gaps in functional regions of the M1 alanyl aminopeptidase enzyme using Jalview. The resulting alignment profile was compared with the that generated from MODELLER using the align2d() function implemented in MODELLER. The best alignment profile was selected based on gap positions and mismatches between target and template sequences.

3.2.3 Model building and refinement

The MSA output was used to generate a .pir file for each of *P. malariae, P. knowlesi, P. ovale and P. vivax* sequences. The generated .pir files were submitted to MODELLER to generate the models. Modelling was done on local Linux machine using a locally-installed MODELLER (version 9.17) to produce 100 models for each *Plasmodium* sequence. The refinement method used was the "Refine.very_slow" MODELLER function, used to provide the highest refinement level (very slow). The generated models were sorted using the Discrete Optimized Protein Energy (DOPE) assessment method which generates z-DOPE scores. The best three models were selected for model validation and evaluation.

3.2.4 Model evaluation

The best three models for each *Plasmodium* target were submitted to different evaluation tools. Those tools include Verify3D, PROCHECK, and ProSA. The best model among each of the three tools was selected. This selected model was visualized using PyMol and superimposed with its corresponding template. The different regions were investigated to check if they affect the active site and protein function.

3.3 Result and Discussion

3.3.1 Template identification

As mentioned in the methodology section, both BLAST and HHpred were used to identify templates. The most similar templates were first selected based on percentage similarity and E-value. Templates with an E-value close to zero were selected to eliminate the chance of getting a random result. Also, templates with similarity and percentage identities higher than 30 % were selected to match the safe alignment zone [102]. The second step was to filter the selected templates based on the total number of gaps and gaps position. Finally, each template was assessed based on its R-value, number of missing residues and the position of these missing residues.

Results retrieved from BLAST are shown in Figure 3-1 and Table 3-2. BLAST returned 100 templates for each target sequence because BLAST parameters were adjusted to show the first 100 query result based on E-value. From those 100 templates, 11 were found to have an E-value equal to 0. Unfortunately, there is no template covering the first 150 residues of target sequences. The template PDB ID 5DLL was found to have the lowest percentage identity and the highest percentage mismatch, which exclude this template from the possible template list. Further, templates with PDB ID: 4R5X, 4J3B, 4K5L, 4R5T and 4R5V had unaligned tails. This exclude them from the possible template list because other remaining templates had similar percentage identities, percentage query coverage values and E-values. Only one expression for 4J3B as it covers position where other templates have mismatches or gaps, which suggest using this template when performing multiple template alignment. The five remaining possible templates are those with PDB IDs 3Q43, 3EBG, 3EBI, 4K5N and 3T8V. As shown in Table 3-3, all possible templates have missing residues or atoms. Fortunately, all the missing residues are found at the N-terminal or C-terminal, which have no effect on important functional positions of the protein. According to R-free values, the best templates were 3T8V and 3Q43. Template 3Q43 had a higher alignment score and query coverage. The template 3T8V was found to have eight mutations while 3Q43 has seven mutations.

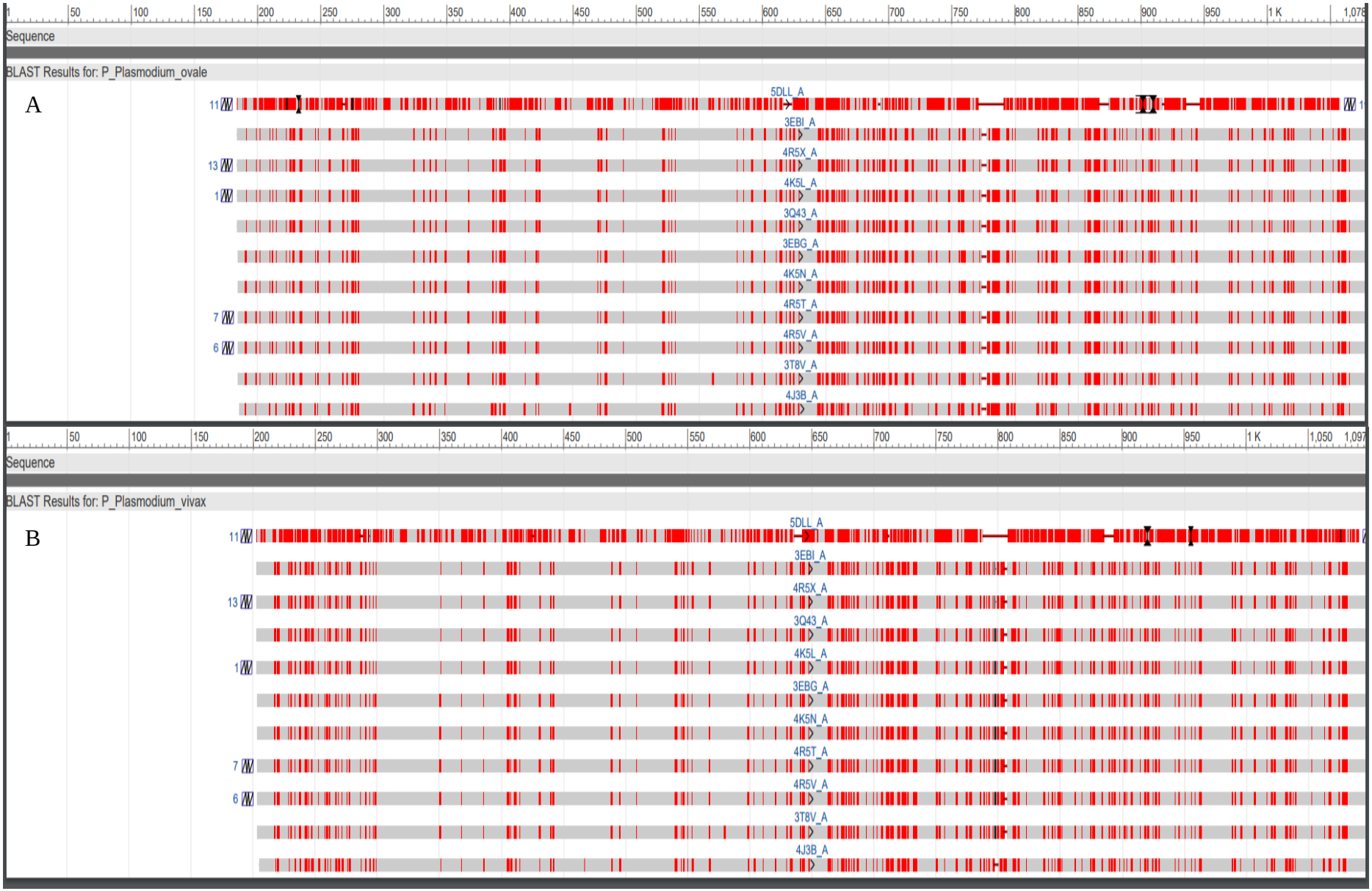
As shown in Table 3-4, the HHpred result had lower E-values, percentage sequence similarities and identities in comparison to BLAST. The HHpred best result was retrieved for *Escherichia coli*, while the BLAST result was used in the case of *P. falciparum*, which is more evolutionary related to other *Plasmodium* species compared to *E. coli.*

Table 3-2: Summary of templates retrieved from BLAST with e-value = 0, showing the identity percent and query cover percent among

Accession number	E-Value	Identity percent	Query cover	Target organism sequence
4R5X	0.0	78%	83%	*P. malariae*
	0.0	80%	81%	*P. knowlesi*
	0.0	80%	81%	*P. vivax*
	0.0	80%	82%	*P. ovale*
3Q43	0.0	87%	83%	*P. malariae*
	0.0	80%	81%	*P. knowlesi*
	0.0	80%	81%	*P. vivax*
	0.0	80%	83%	*P. ovale*
4J3B	0.0	78%	83%	*P. malariae*
	0.0	80%	81%	*P. knowlesi*
	0.0	81%	82%	*P. vivax*
	0.0	81%	82%	*P. ovale*
3T8V	0.0	78%	83%	*P. malariae*
	0.0	80%	81%	*P. knowlesi*
	0.0	80%	81%	*P. vivax*
	0.0	80%	83%	*P. ovale*
3EBG	0.0	78%	83%	*P. malariae*
	0.0	80%	81%	*P. knowlesi*
	0.0	80%	81%	*P. vivax*
	0.0	80%	82%	*P. ovale*
3EBI	0.0	78%	83%	*P. malariae*
	0.0	80%	81%	*P. knowlesi*
	0.0	80%	81%	*P. vivax*
	0.0	80%	82%	*P. ovale*
4K5N	0.0	78%	83%	*P. malariae*
	0.0	80%	81%	*P. knowlesi*
	0.0	80%	81%	*P. vivax*
	0.0	80%	82%	*P. ovale*
5DLL	0.0	37%	82%	*P. malariae*
	0.0	37%	80%	*P. knowlesi*
	0.0	37%	80%	*P. vivax*
	0.0	37%	81%	*P. ovale*

Table 3-3: Possible templates without unaligned tails sorted from left to right according to resolution then number of missing residues.

PDB ID	**3Q43**	**3T8V**	**4K5N**	**3EBI**	**3EBG**	**4J3B**
Length	891	895	895	890	889	889
N. Chains	1	1	1	1	1	1
Resolution (A)	1.8	1.8	1.91	2.0	2.1	2.2
Number of missing residues	2	6	6	1	1	1
Number of missing atoms	30	29	30	4	38	32


Sequence
BLAST Results for: P_Plasmodium_ovale
A
5DLL_A
3EBI_A
4R5X_A
4K5L_A
3Q43_A
3EBG_A
4K5N_A
4R5T_A
4R5V_A
3T8V_A
4J3B_A
Sequence
BLAST Results for: P_Plasmodium_vivax
B
5DLL_A
3EBI_A
4R5X_A
3Q43_A
4K5L_A
3EBG_A
4K5N_A
4R5T_A
4R5V_A
3T8V_A
4J3B_A

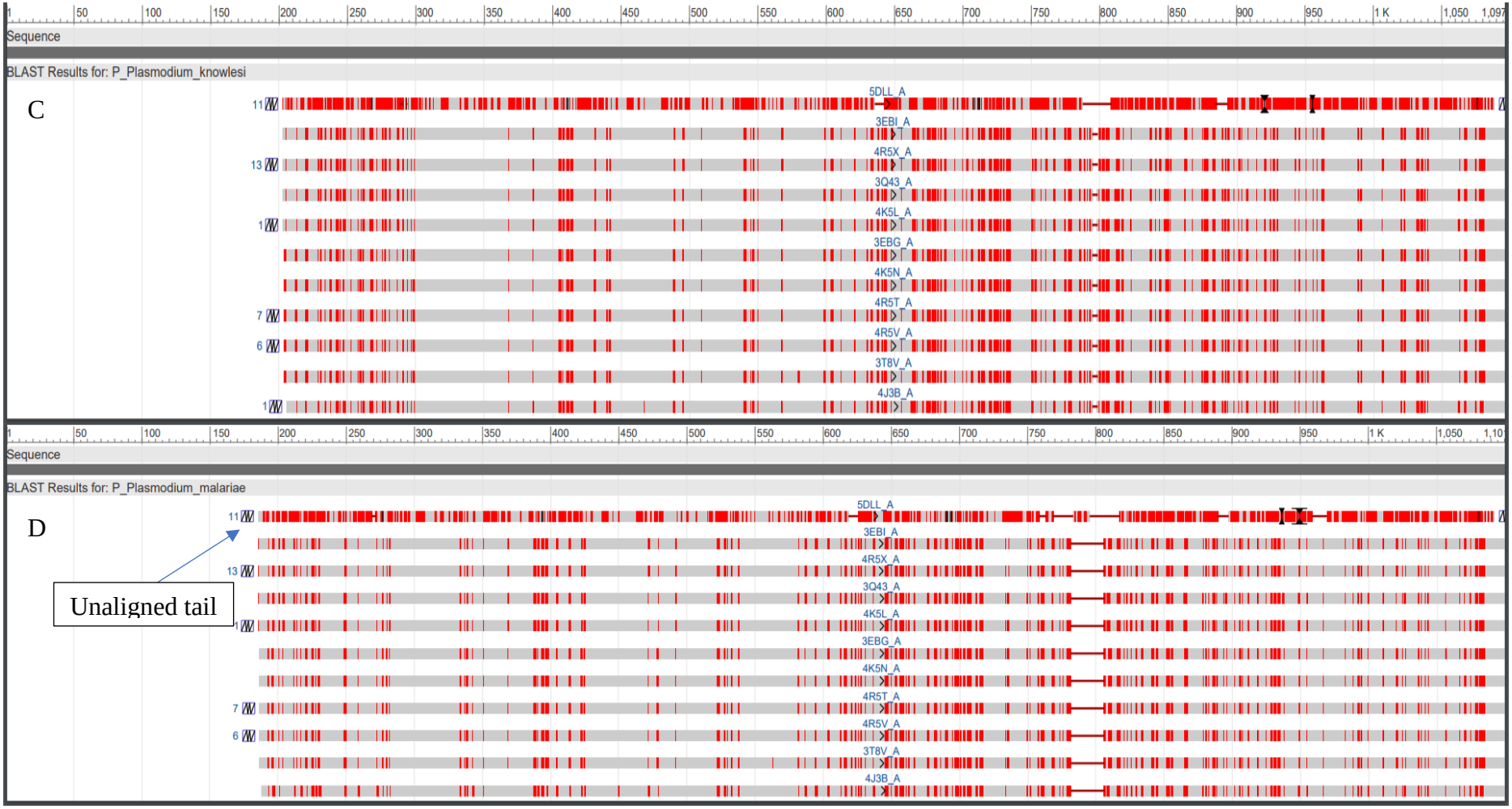


Figure 3-1: Summary of target *Plasmodium* sequences with the best 10 possible templates. A: shows the alignment of *P. ovale* sequence against the best ten templates, B: shows the alignment of *P. vivax* sequence against the best ten templates, C: shows the alignment of *P. knowlesi* sequence against the best ten templates, D: shows the alignment of *P. malariae* sequence against the best ten templates. The red vertical lines show mismatch positions while the grey bars show the matched positions between the target and each possible template.

Table 3-4: Summary of the best template for each sequence retrieved from HHpred with the lowest E-value, showing the percentage identity and percentage query coverage among

Target organism sequence	PDB ID	E-value	Identity percent	Similarity percent
P. malariae	4XO5	2.3E-127	35%	60%
P. knowlesi	4XO5	2.2E-126	34%	60%
P. vivax	4XO5	3.6E-125	34%	59%
P. ovale	4XO5	3.2E-124	34%	59%

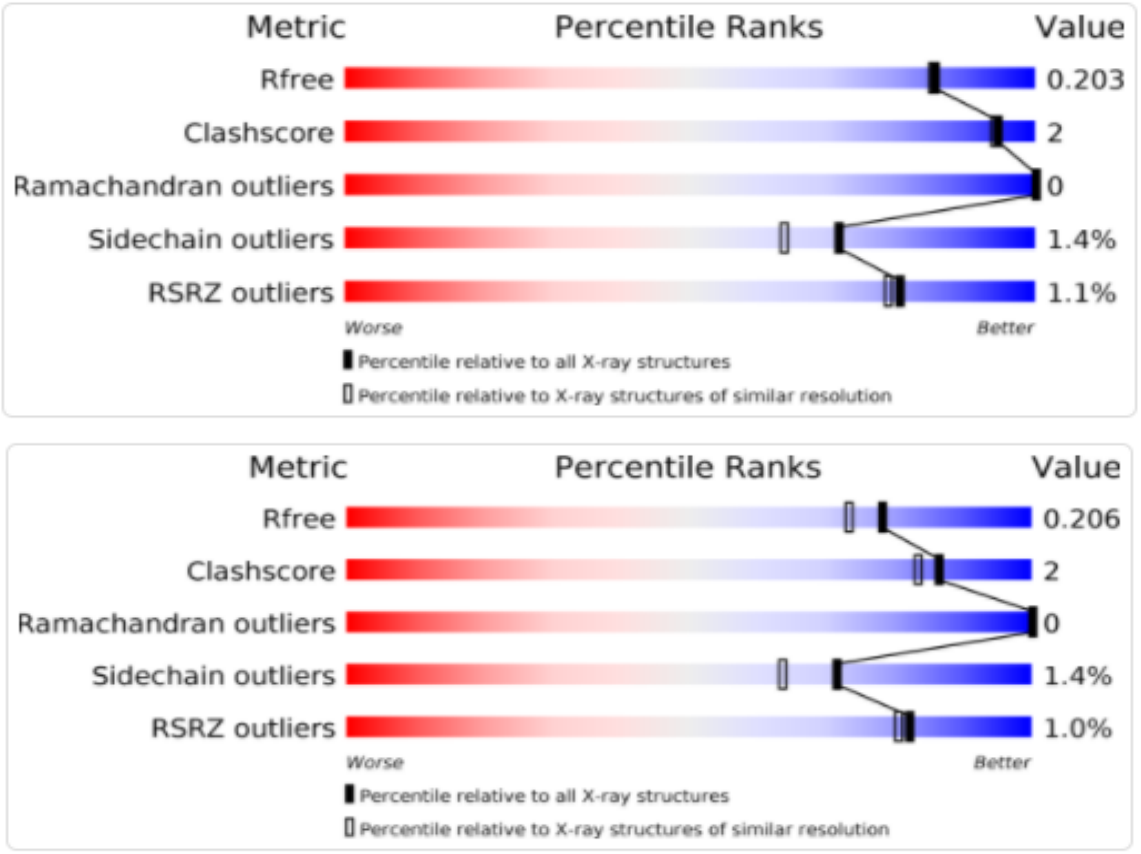


Figure 3-2: wwPDB validation, representing the overall structure quality for A: 3T8V and B: 3Q43

Homology modelling is based on transferring the 3D coordinates of amino acid positions to those of the template, which is why the template quality and suitability are evaluated to ensure the best template is selected. Both potential templates 3Q43 and 3T8V were submitted to QMEAN and verify 3D. From verify 3D both have 3D-1D Averaged Scores higher than zero. Additionally, 3Q43 has 97.19% residues with an averaged 3D-1D score >= 0.2, while 3T8V has 97.08%. In 3Q43, the active site residue scores were 0.52 for HIS number 301, 0.53 for HIS number 305 and 0.3 for GLN number 324, while in 3T8V these were 0.52 for HIS number 301, 0.47 for HIS number 305 and 0.29 for GLN number 324. Also, 3Q43 showed higher scores than 3T8V, as shown in Figure 3-3.

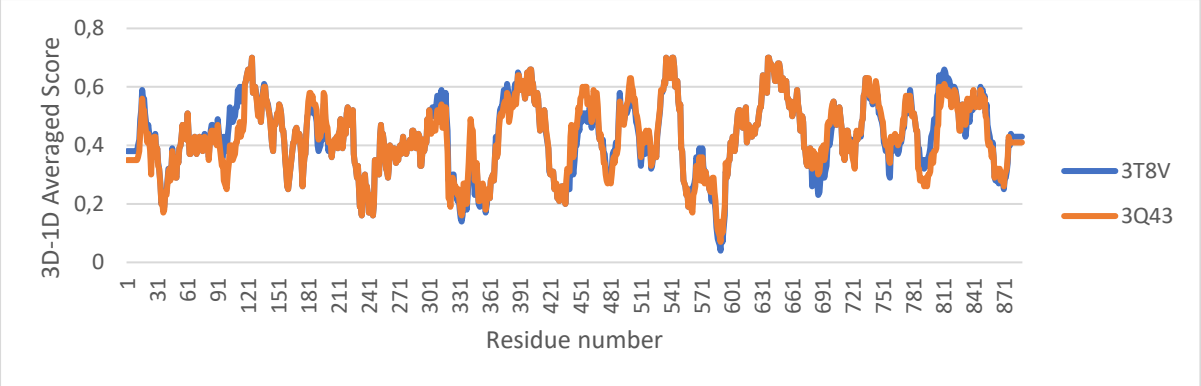


Figure 3-3: Graphical representation of 3D-1D averaged scores per residue number (blue for 3T8V and Orange for 3Q43) The lowest value for 3Q43 was 0.04 while for 3T8V was 0.07.

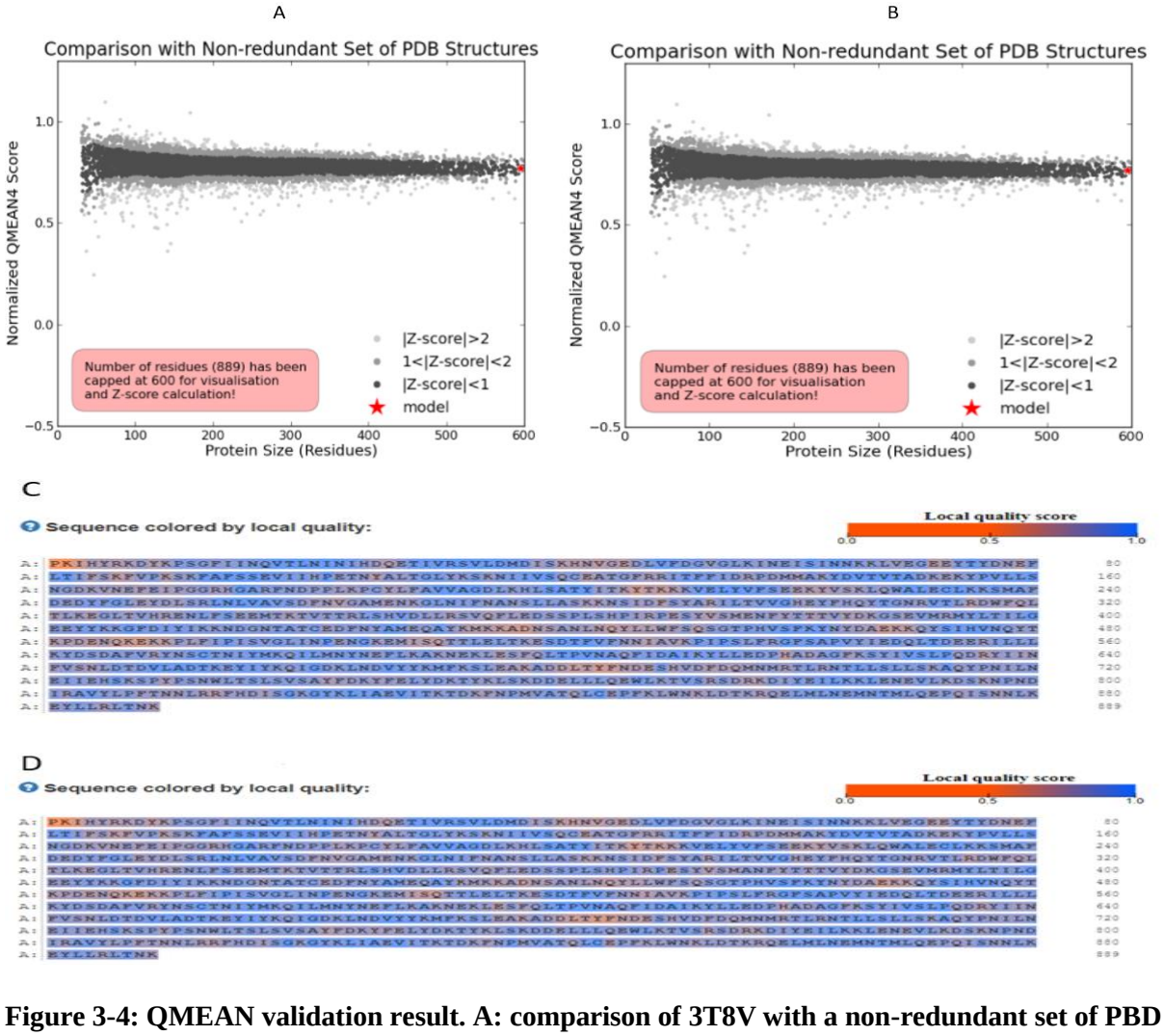


Figure 3-4: QMEAN validation result. A: comparison of 3T8V with a non-redundant set of PBD entries. B: comparison of 3Q43 with a non-redundant set of PBD entries. C: 3T8V sequence coloured by local quality (Orange low quality – blue high quality). C: 3Q43 sequence coloured by local quality (Orange low quality – blue high quality)

In the QMEAN result, the QMEAN4 score for 4T8V was 0.08 while being 0.06 for 3Q43. While both models were in the safe zone when compared to other non-redundant sets of PDB entries, as both get normalized Z-scores lower than 1. For local quality, both templates have

bad residue qualities for the first three residues while the active site residues have good local quality scores. Based on QMEAN and verify 3D results, the 3Q43 template was selected for homology modelling.

3.3.2- Sequence alignment

Two different methods were used to generate sequence alignments between the template and the target sequence. The first method was an MSA using 3D-Coffee [103]. The purpose of doing the MSA was to use the structural alignment tool to perform multiple sequence alignment with a focus on evolutionary distance and changes between the template and target sequence, as well as a structural element in aligned sequences [104], [105]. The second method was using align2d implemented function of MODELLER which automatically generates a pairwise alignment or multiple alignments depending on the number of input templates [106]. The generated alignment was in MODELLER-compatible format, which doesn't require an additional step to prepare the alignment output for modelling. In both methods, 3Q43 was used as the template sequence. As shown in the alignment retrieved from 3D-Coffee (Figure 3-5), the active site from the template matched the active site from the target sequence, while it mismatched in the alignment retrieved from align2d() function.

In align2d the alignment starts from the first residue, which later introduces gaps in functional positions, while in 3D-coffee it inserts gaps in the N-terminal positions, which improve the overall alignment quality. Hence 3D-coffee alignment was used for all target sequence. The alignment was used to create a .PIR file for each template, as shown in Figure 3-6. Each alignment was manually curated and edited if needed.

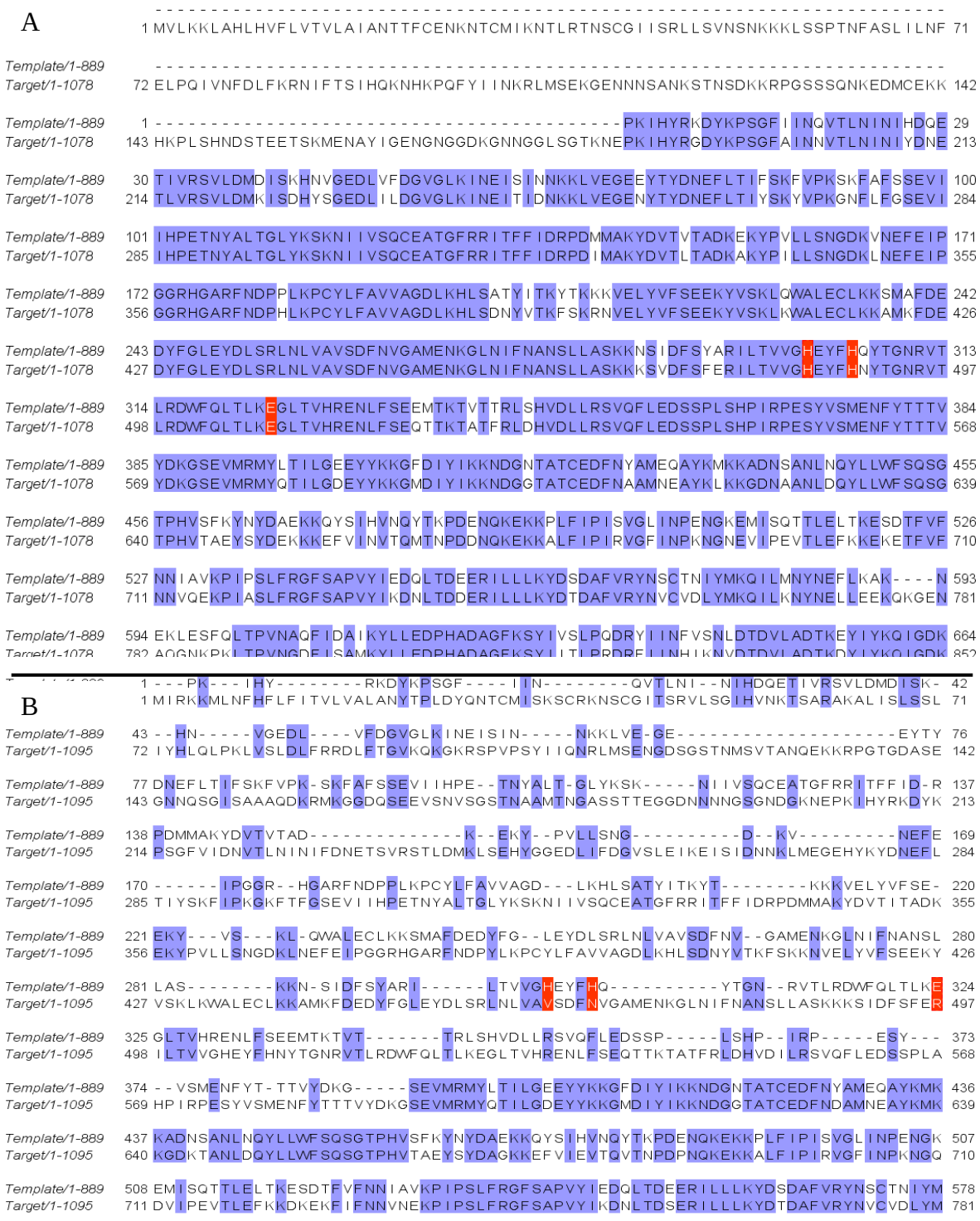


Figure 3-5: Alignment between the template (PDB ID 3Q43) and the target M1 alanyl aminopeptidase sequence from *P. vivax*. A. The alignment produced by the align2d function. B. The alignment produced from 3D-Coffee. The active site is highlighted in red boxes.

A- *P. ovale*

B- *P. knowlesi*

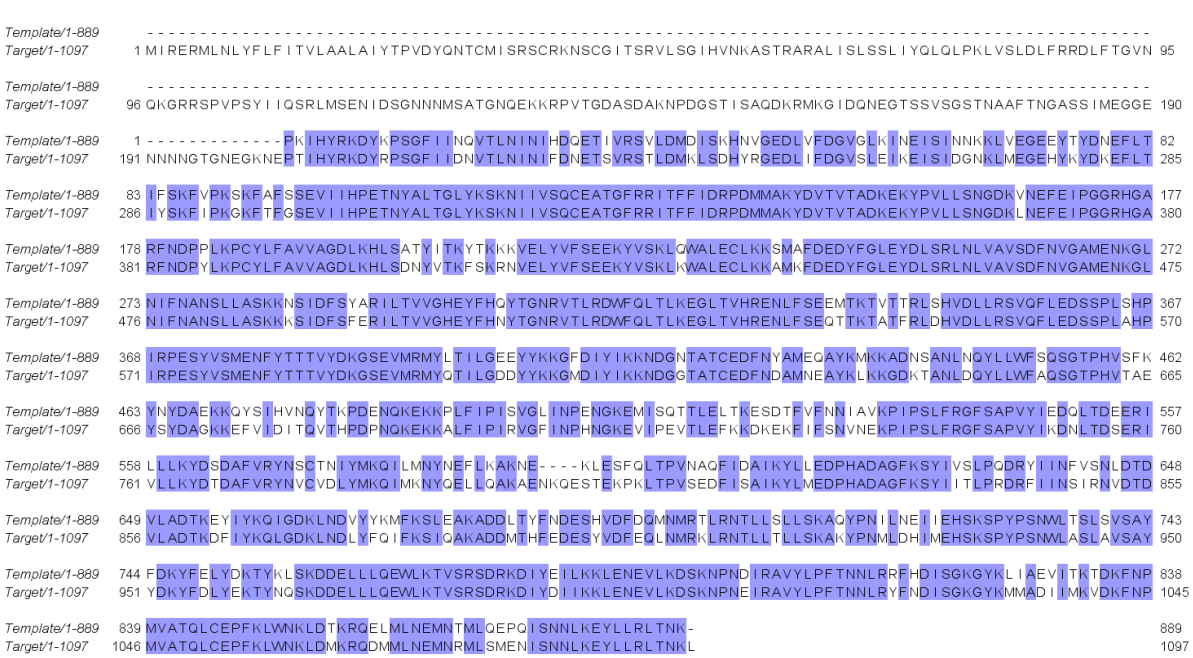


Figure 3-6: Template-target alignment generated by 3D-coffee. A, B, C and D: Graphical representations of alignment used to prepare a .pir file for each target sequence using a single template (PDB ID 3Q43). Each alignment was generated by 3D-Coffee. Matched residues are highlighted in blue.

3.3.3- Model Building

The generated .PIR files were modified to include zinc metal ions in the model building process. This was done by adding "/." at the end of each .PIR file because the zinc atom was the first atom after the last amino acid residue. Once all files were prepared, MODELLER was used to calculate 100 3D-models for each target. 100 models were created based on the 3Q43 template. All MODELLER runs were done using slow refinement. All generated 3D models were assessed using the z-DOPE score [107]. The z-DOPE scores for the best three models for each MODELLER run are shown in Table 3-5.

Table 3-5: Summary of DOPE-Z score and RMS score of best three models for each run.

Source organism	Model name	RMS score	z-DOPE score
P. malariae	Model 0024	0.106	-1.5775
	Model 0052	0.100	-1.5872
	Model 0084	0.100	-1.6159
P. knowlesi	Model 0008	0.107	-1.7606
	Model 0083	0.099	-1.7660
	Model 0086	0.093	-1.7493
P. vivax	Model 0038	0.103	-1.7925
	Model 0041	0.098	-1.8403
	Model 0094	0.094	-1.7666
P. ovale	Model 0044	0.110	-1.7295
	Model 0070	0.100	-1.7114
	Model 0079	0.115	-1.7100

The overall quality of produced structures was good, and the homology models had z-DOPE scores close to those of the template (PDB ID 3Q43), which was -2.0740. All three models were superimposed with the template, as shown in Figure 3-7.

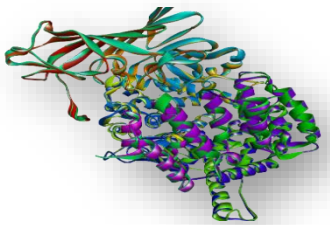
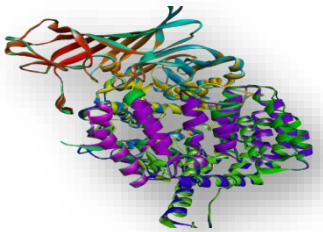
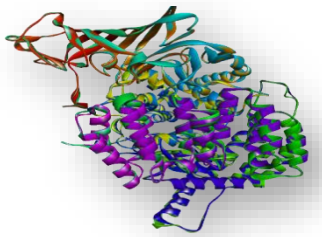
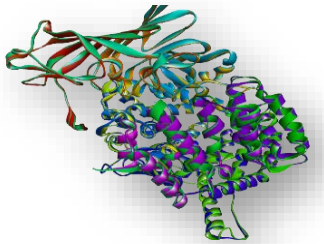
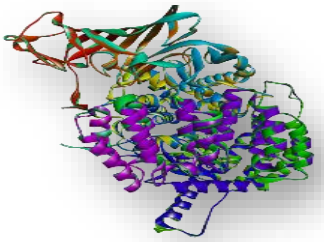
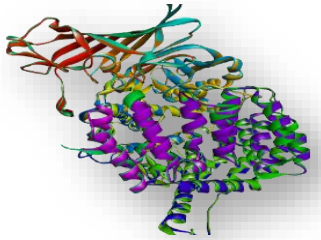
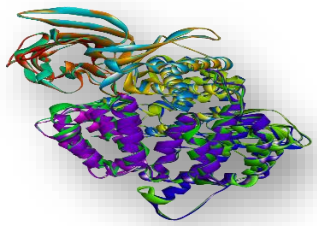
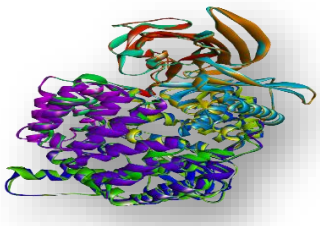
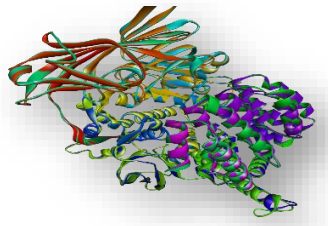
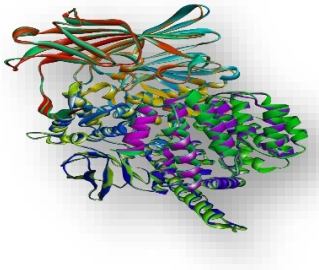
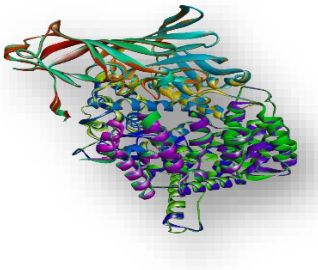
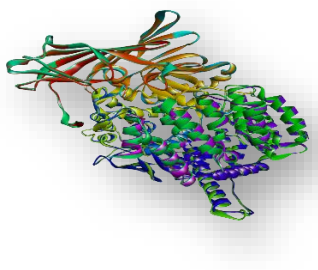
	P. knowlesi models	
		
Model 0008	Model 0083	Model 0086
	P. malariae models	
		
Model 0024	Model 0052	Model 0084
	P. ovale models	
		
Model 0044	Model 0070	Model 0079
	P. vivax models	
		
Model 0038	Model 0041	Model 0094

Figure 3-7: Top three model for each run superimposed onto the original template (PDB ID: 3Q43).

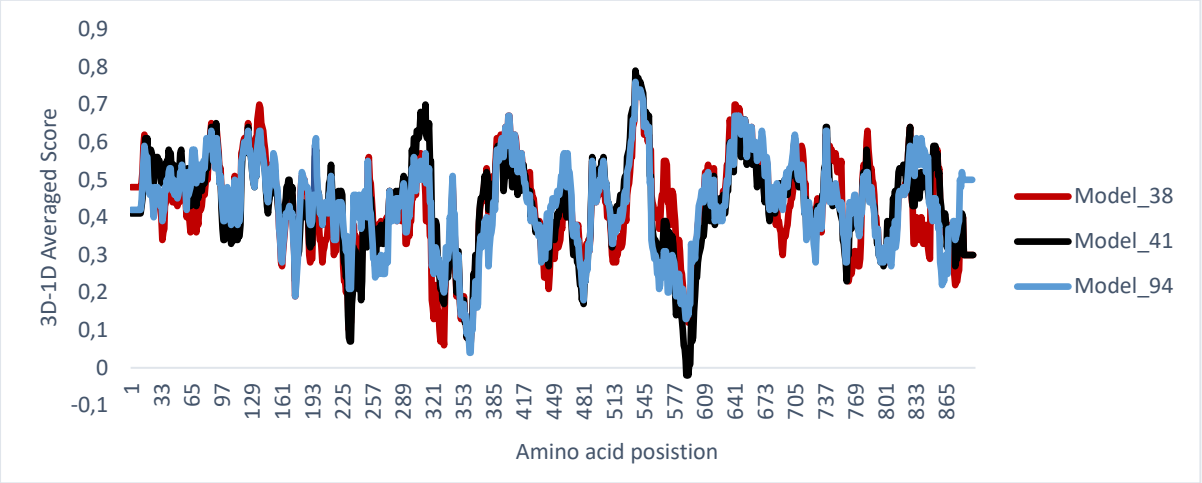
3.3.4- Model Evaluation

The accuracy of homology modelling structures is largely dependent on the inputs and upstream steps, including template selection and template-target alignment. As a result, any problem happening during any step will generate a model with errors. Possible errors could result from gaps in wrong places or errors in the template 3D structure. Due to the complexity of loop regions, loop modelling may also introduce errors. As such, the generated model should pass through different model evaluation tools to evaluate the accuracy of the models. In this work, different tools were used to evaluate the produced top three models for each source organism. The tools used include PROCHECK, ANOLEA, QMEAN and verify 3D. The purpose of using different tools is to evaluate different criteria in which PROCHECK was used to assess model stereochemistry. PROCHECK results are shown in Table 3-6. ANOLEA was used to evaluate the energy of the protein chain, including all non-local interactions of all heavy atoms in the evaluated model. QMEAN was used to evaluate the protein structure through a different scoring function that evaluates the entire protein as well as each residue. Finally, verify 3D was used to measure the relationship between the 3D structure and its amino acid sequence, based on amino acid favorable geometries and good known structures.

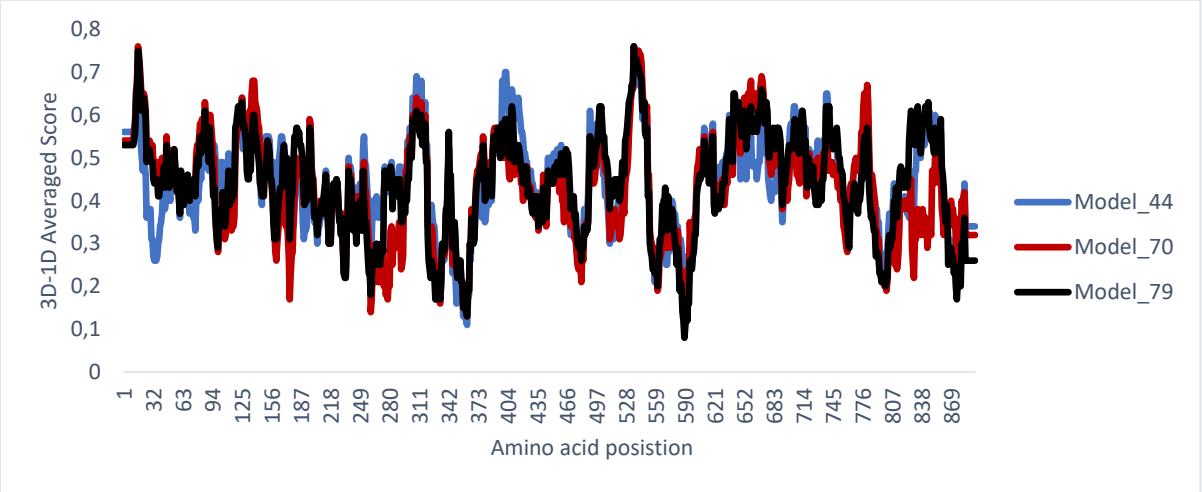
Table 3-6: PROCHECK local quality assessment scores. It represents each model with its corresponding QMEAN 4 score and percentage of residues in the most favored regions, residues in additional allowed regions, residues in generously allowed regions and residues in disallowed regions.

Source Organism	Model name	PROCHECK - Ramachandran Plot				QMEAN 4
		Residues in most favored regions	Residues in additional allowed regions	Residues in generously allowed regions	Residues in disallowed regions	
P. malariae	Model 0024	800 (94.5%)	45 (5.3%)	2 (0.2%)	0 (0.0%)	- 0.70
	Model 0052	802 (94.7%)	43 (5.1%))	2 (0.2%)	0 (0.0%)	- 0.86
	Model 0084	804 (94.9%)	41 (4.8%)	2 (0.2%)	0 (0.0%)	- 0.76
P. knowlesi	Model 0008	783 (94.7%)	41 (5.0%)	3 (0.4%)	0 (0.0%)	- 0.70
	Model 0083	786 (95.0%)	39 (4.7%)	2 (0.2%)	0 (0.0%)	- 0.92
	Model 0086	788 (95.3%)	37 (4.5%)	2 (0.2%)	0 (0.0%)	-0.60
P. vivax	Model 0038	784 (94.9%)	39 (4.7%)	3 (0.4%)	0 (0.0%)	-0.61
	Model 0041	781 (94.6%)	42 (5.1%)	3 (0.4%)	0 (0.0%)	- 0.76
	Model 0094	781 (94.6%)	43 (5.2%)	2 (0.2%)	0 (0.0%)	- 0.71
P. ovale	Model 0044	782 (95.1%)	38 (4.6%)	2 (0.2%)	0 (0.0%)	- 0.83
	Model 0070	779 (94.8%)	41 (5.0%)	2 (0.2%)	0 (0.0%)	- 0.74
	Model 0079	781 (95.0%)	38 (4.6%)	3 (0.4%)	0 (0.0%)	- 0.78

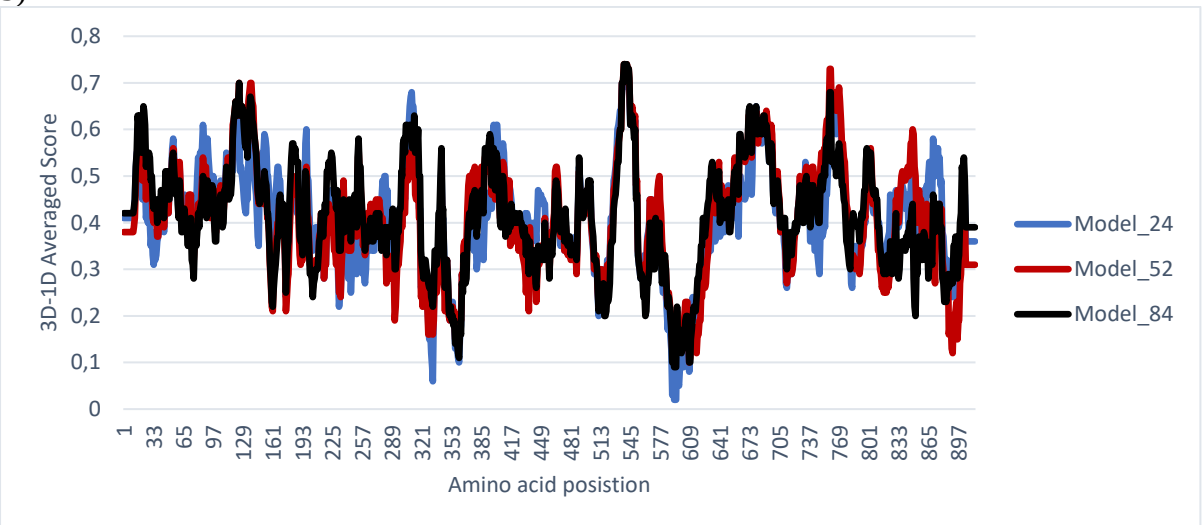
1)


0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
-0,1
3D-1D Averaged Score
1
33
65
97
129
161
193
225
257
289
321
353
385
417
449
481
513
545
577
609
641
673
705
737
769
801
833
865
Amino acid posistion
Model_38
Model_41
Model_94

2)


0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
3D-1D Averaged Score
1
32
63
94
125
156
187
218
249
280
311
342
373
404
435
466
497
528
559
590
621
652
683
714
745
776
807
838
869
Amino acid posistion
Model_44
Model_70
Model_79

3)


0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
3D-1D Averaged Score
1
33
65
97
129
161
193
225
257
289
321
353
385
417
449
481
513
545
577
609
641
673
705
737
769
801
833
865
897
Amino acid posistion
Model_24
Model_52
Model_84

4)

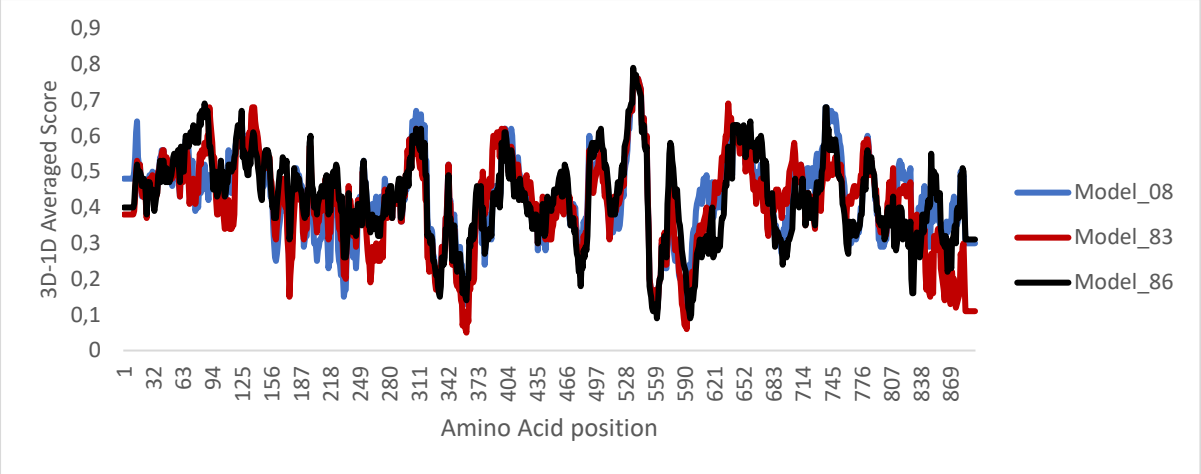


Figure 3-8: verify 3D results for the top selected three models for M1 alanyl aminopeptidase from 1) *P. vivax*, 2) *P. ovale*, 3) *P. malariae* and 4) *P. knowlesi*.

Table 3-7: Verify 3D quality assessment score for each model.

Source Organism	Model name	Verify 3D score
P. malariae	Model 0024	95.41%
	Model 0052	94.54%
	Model 0084	96.72%
P. knowlesi	Model 0008	96.65%
	Model 0083	91.62%
	Model 0086	96.31%
P. vivax	Model 0038	95.30%
	Model 0041	94.97%
	Model 0094	96.64%
P. ovale	Model 0044	98.43%
	Model 0070	97.54%
	Model 0079	97.43%

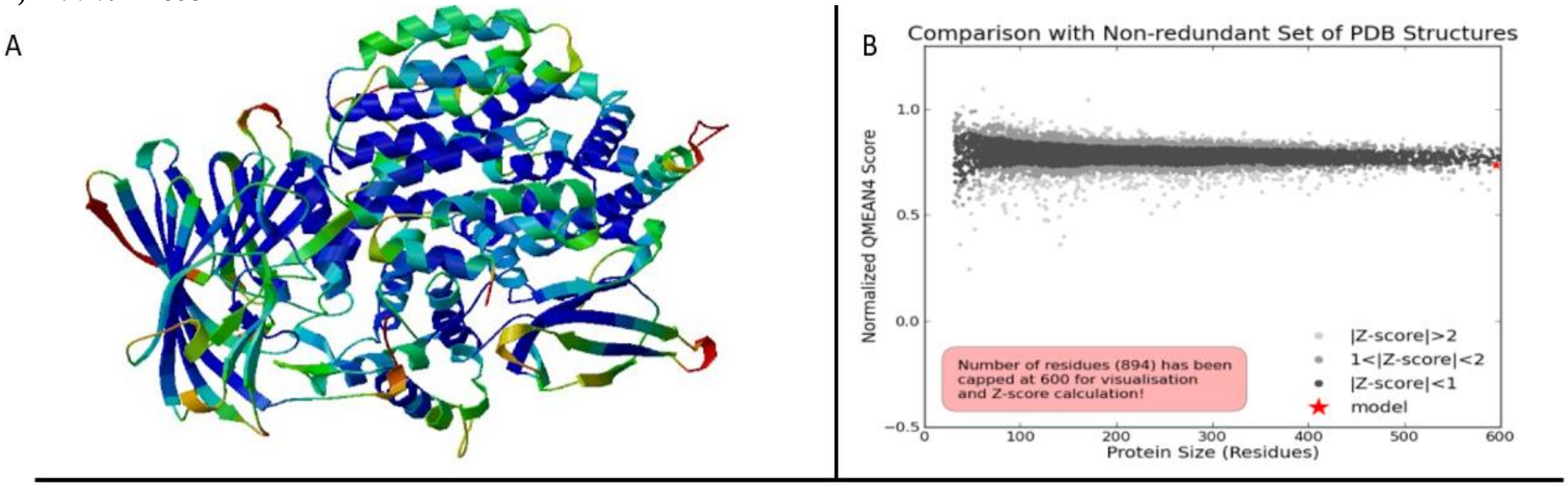
As shown in Figure 3-8 and Table 3-7, all the 12 models (3 models for each *Plasmodium* species) pass the verify 3D assessment analysis. In *P. vivax*, Model 0041 has been eliminated as it has negative values for the residue position 592. Then Model 0094 was selected as the best model because its active site residues have higher scores than those of Model 0038. For *P. ovale*, Model 0079 show score lower than that of the two other models. While the difference between the remaining models was very low, Model 0044 was selected as it has a higher overall score and a higher score for active site residues. In *Plasmodium malaraie*, Model 0024 was eliminated as it shows a lower score. The Model 0084 was selected because it has a higher 3D-1D average score for active site residues. Finally, for *P. knowlesi*, the difference between the three models was very low, while Model 0083 showed lower overall quality compared to the

other models. In the end, Model 0008 was chosen because it has a higher score than Model 0086.

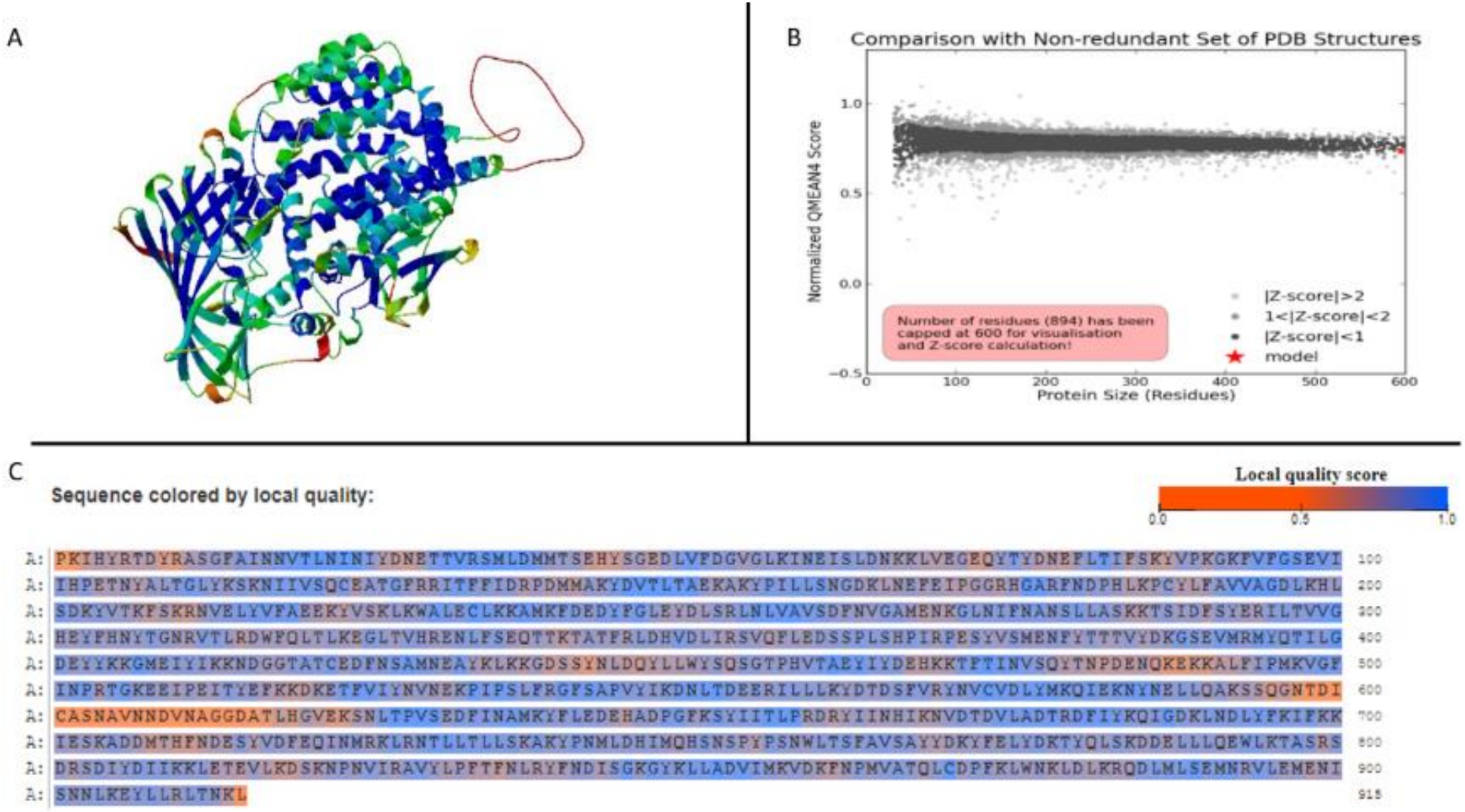
Table 3-8: Top selected model with the corresponding *Plasmodium species.*

Organism name	Top selected model
P. malariae	Model 0024
P. knowlesi	Model 0008
P. vivax	Model 0094
P. ovale	Model 0044

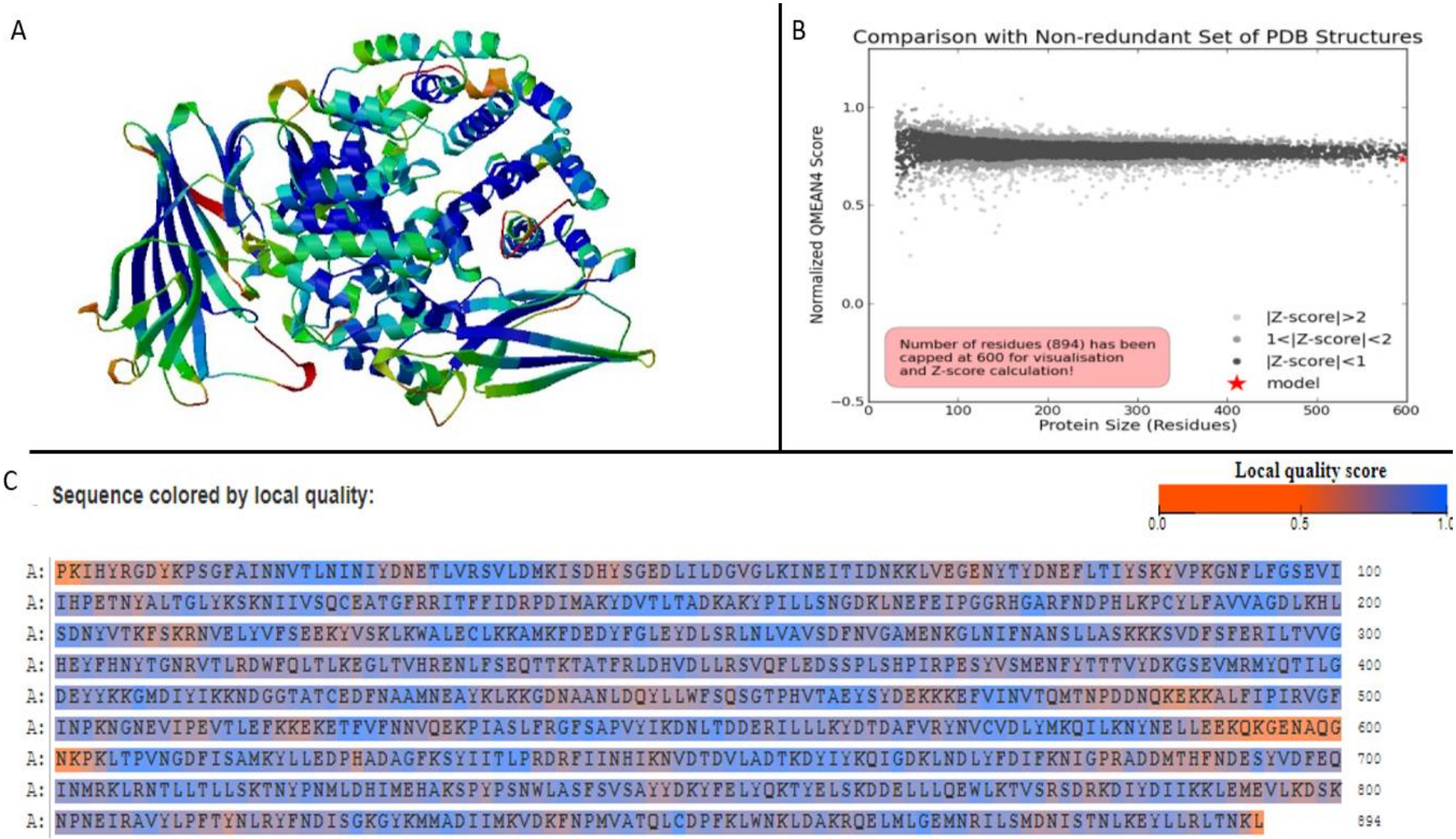
1) *P. vivax* model



2) *P. malariae* model



3) *P. ovale* model



4) *P. knowlesi* model

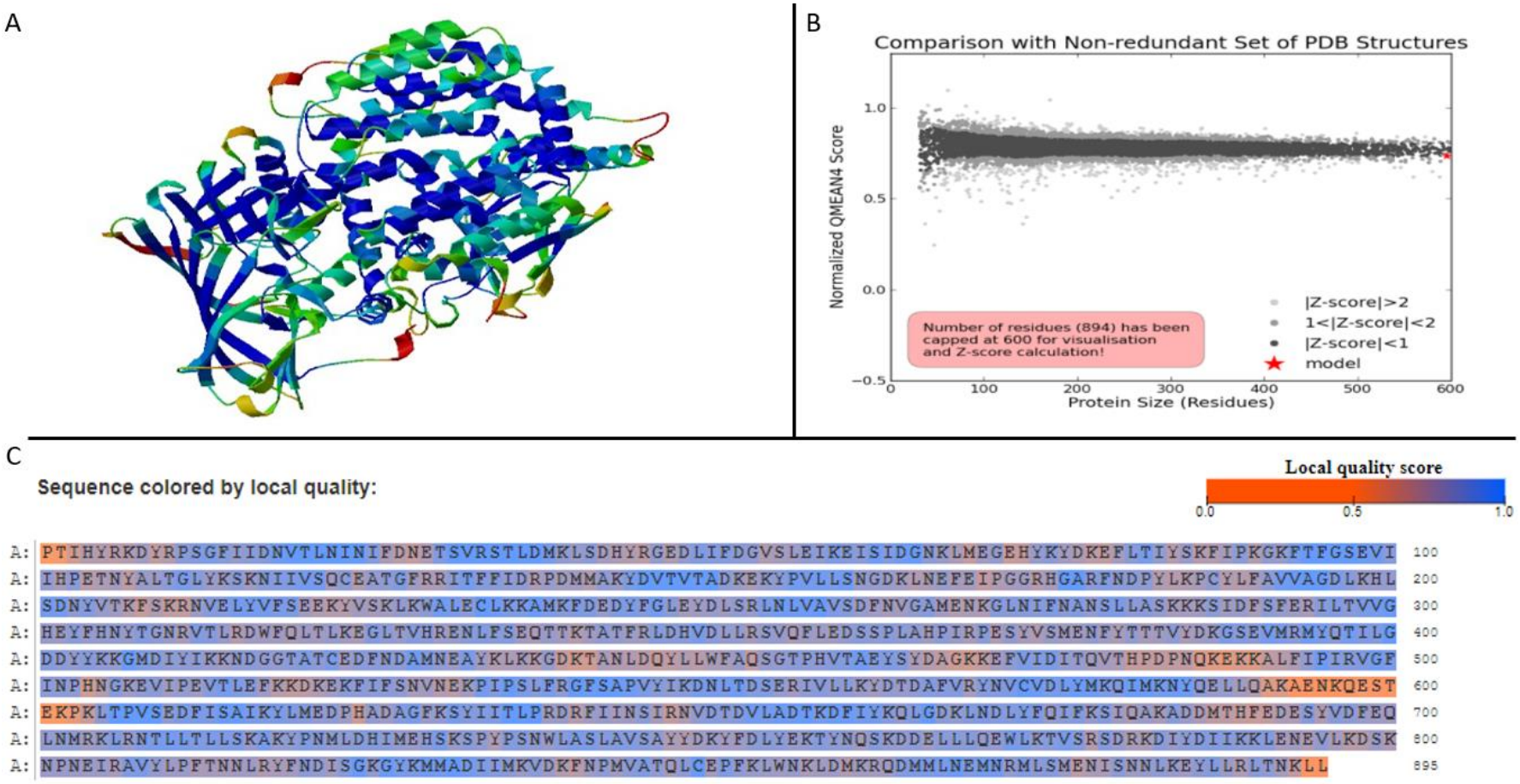


Figure 3-9: QMEAN analysis result. A) Residues coloured by residues error, representing the estimated residue inaccuracy where blue corresponds to the more accurate regions while red represents the inaccurate regions. B) Estimated absolute model quality generated by QMEAN, where the model is highlighted in red. C) Model amino acid sequence coloured according to local quality score in which the lowest scores are in red while the blue colour represents the highest scores.

As shown in Figure 3-9, all top selected models for the *Plasmodium* species passed QMEAN analysis, in which the QMEAN6 value for *P. vivax* was 0.962, 0.691 for *P. ovale*, 0.664 for *P. malariae* and 0.712 for *P. knowlesi*. The problematic residues were located mainly in the loop regions, while the active site residues (histidine 301, histidine 305 and glutamine 324) in all selected models had a high local quality score. All models were found to have Z-scores lower than one, which is considered a good Z-score [108].

All top selected models show low local quality scores for residues located in the N-terminal. This was attributed to the missing residues in the N-terminal of the used template, the mismatch and gaps located at the start region of template-target alignment used in homology modelling.

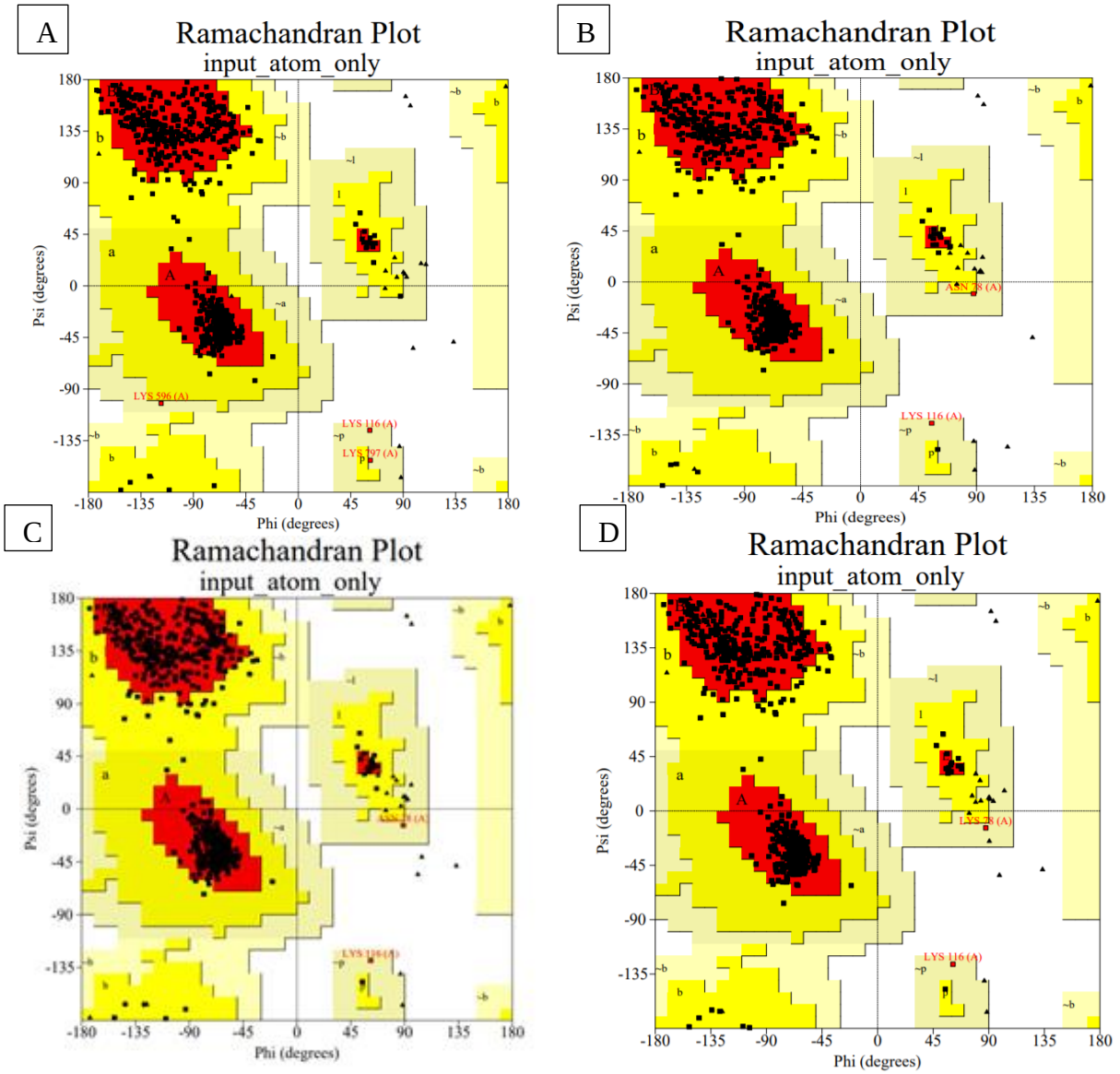


Figure 3-10: The PROCHECK results, showing Ramachandran plots for the top selected models. A) *P. vivax,* B) *P. ovale*, C) *P. malariae and* D) *Plasmodium* knowlesi.

For each *Plasmodium* species, a Ramachandran plot was generated using PROCHECK for the top selected models. As shown in Figure 3-10, no model was found to have residues in the disallowed regions. The lowest quality model was that of *P. vivax*, which has three residues in the generously allowed regions. The other models were found to have only two residues in the generously allowed regions. However, overall this does not cause a structural problem [109].

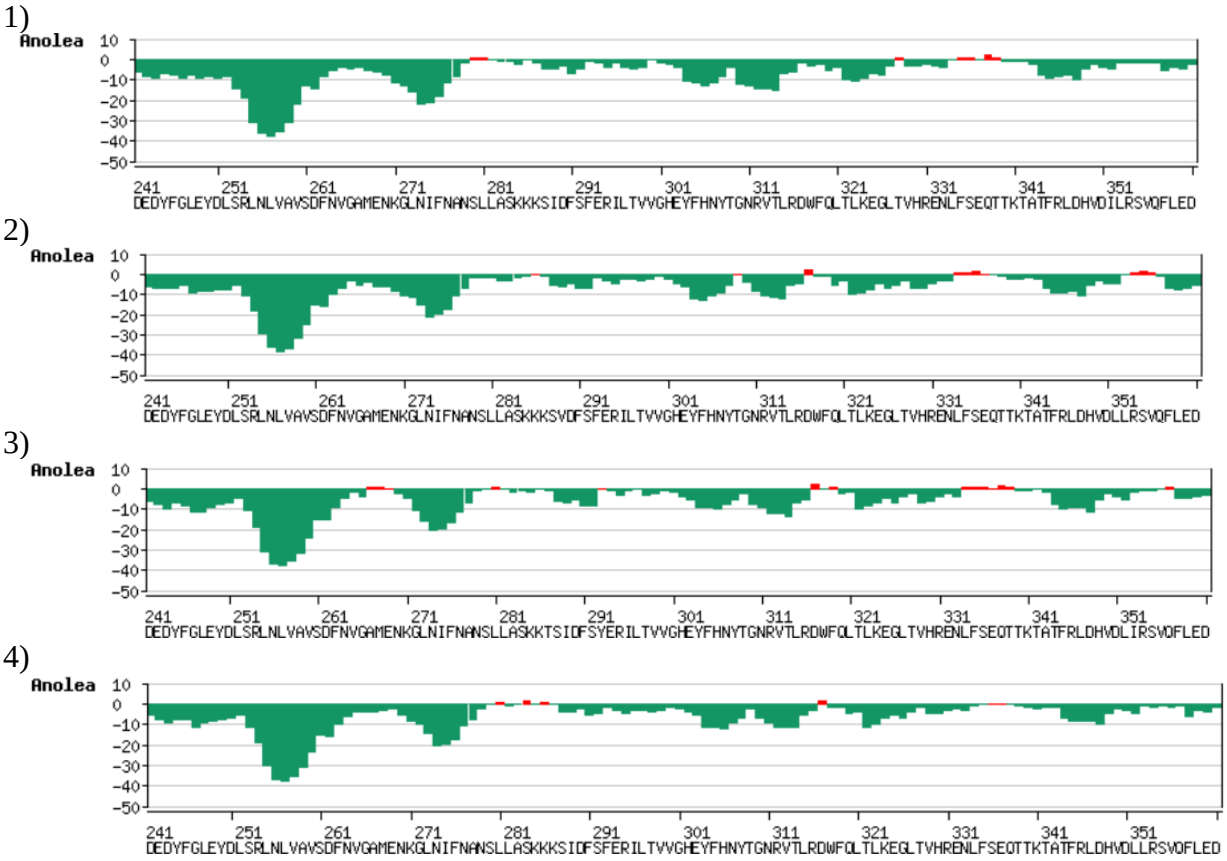


Figure 3-11: ANOLEA result for the active site region. ***1) P. vivax. 2) P. ovale. 3) P. malariae. 4) Plasmodium knowlesi.*** **The green part shows the favourable energy parts while red corresponds to the unfavourable parts.**

All top selected models passed ANOLEA analysis. As shown in Figure 3-11, all the active site residues have energy values below zero, which means that they are located in the favorable energy regions, while the positive energy regions do not consist of large regions and do not have a big impact on the overall structure quality.

3.4 Conclusion

This chapter presents the use of homology modelling to build a 3D representative structure for **M1 alanyl aminopeptidase from** pathogenic *Plasmodium* species including *P. malariae, P. knowlesi, P. ovale and P. vivax.* Homology modelling starts with template identification. Different 3D structures for **M1 alanyl aminopeptidase from** *P. falciparum* were retrieved from

the PDB. Then all retrieved structures were assessed to choose the best quality templates that are most similar to the target sequences. The 3D structure with PDB ID 3Q43 was selected as the best template. The selected structure showed high sequence coverage with high-resolution. Template–target alignment was done using the 3D-coffee alignment tool. The produced alignment was manually trimmed to remove the N terminal region. Then, 100 models were generated for **M1 alanyl aminopeptidase from** each *Plasmodium* species. The top model was selected based on different quality assessment tools. The selected models were of best local and global quality.

Chapter 4 - Virtual Screening

All generated models from homology modelling and the 3D structure of M1 alanyl aminopeptidase from *P. falciparum* and the human protein were submitted to virtual screening. Each protein was screened against 623 compounds retrieved from the SANCDB. This chapter aims to identify potential inhibitors against *Plasmodium* M1 alanyl aminopeptidases, which involve eliminating ligands with inhibition properties against the human protein. As a result, the final selected ligands would be selective against the **M1 alanyl aminopeptidase from** *Plasmodium* species.

4.1 Introduction

4.1.1 Computation docking

3D structures of proteins and ligands make it possible to study the interaction between different proteins involved in vital pathways and also enable the study of protein-ligand interactions, protein inhibition and activation [110]. Currently, there are different ways to study protein-ligand interactions. One of them is computational docking., which is a process involved in testing different orientations and conformations of a small molecule (ligand) until it finds the best orientation and conformation upon binding the target protein structure to form a stable protein-ligand complex. The process of selecting the best orientation and conformation is done by using a mathematical function that calculates the binding free energy. Then the lowest binding energy corresponds to the best orientation and conformation. Hence it corresponds with the best complex stability [111]. The process starts with selecting a protein with an available or a generated 3D structure. This protein mainly corresponds to a medical disease. Then a small molecule (ligand) library is optimized for the screening process. Finally, a docking tool calculates the binding free energy of ligands with respect to the target protein [112].

4.1.2 Virtual screening

Virtual screening involves selecting the best binding compound from possible ligand compounds databases by using different *in silico* tools [113]. There are two main strategies to perform virtual screening, namely and ligand-based and structure-based methods. Ligand-based virtual screening (LBVS) is very useful when the 3D structure of target protein is unknown. LBVS is involved in different techniques, such as molecular representation, data mining methods, similarity searching and pharmacophore mapping [114]. Structure-based virtual screening (SBVS) is usually used when the 3D structure is known. SBVS involves

docking of ligands from a database against a selected target site. Additionally, SBVS applies a different scoring function; the generated scores could be used to rank the docked molecules [115].

4.1.3 Structural based virtual screening

SBVS consists of four different steps: 1) Target preparation, 2) ligand database selection, 3) molecular docking and 4) analysis of docking results. SBVS starts with the preparation of a target 3D structure., which involves adding any missing atoms, the protonation of the target structure by adding hydrogen atoms [116], the removal of water molecules (with exception of water molecules that coordinate active site or are involved in important interactions) and the choice of the correct protonation state for each amino acid - especially for active site residues [117]. After preparation of the target 3D structure, it is time to select ligand database. Currently, there are many databases to be considered. Those includes ZINC [118], PubChem [119], DrugBank [120], Binding DB [121], [122], SANCDB [123], ChEMBL [124], [125] and ChemBank [126]. Most databases include a query engine to search and select compounds that meet predetermined chemical characteristics. Selected compounds should be prepared to match the correct stereochemistry and ionization states. The third step is to perform molecular docking, in which the prepared subset is docked into a previously selected target site in the 3D structure of target protein [127].

Currently, there are different software applications that can be used to perform molecular docking depending on the docking strategy. The most common docking strategies comprise the rigid body docking and flexible docking [128]. The most common tools that apply rigid body docking are FRODOCK [129], ZDOCK [130] and MEGADOCK [131], while for flexible docking the most common tools are: AutoDock [132], AutoDock Vina [133], ParaDockS [134] and GOLD [135].

The main difference between rigid body docking and flexible docking is the flexibility of both target protein and ligand, whereby the former allows for ligand flexibility and treats the protein as a rigid body, which means that bond, angles and the dihedral lengths between protein atoms are fixed during the docking experiment [136]. The purpose of rigid protein is to minimize the search space. However, ignoring flexibility of target protein reduce the accuracy of the docking result [137], while in flexible docking both ligand and target protein are flexible. This allows for the inclusion of conformational changes (backbone and side chain) in the docking experiment. However, incorporating this degree of flexibility increases the search space, which

increases the running time and can lead to an increase in the number of the false positive results [138].

AutoDock Vina is a newly developed version of AutoDock. The main difference was removal of an empirical scoring function to implement a sophisticated method with the Monte Carlo sampling technique during the local optimization procedure. This scientifically increases the prediction accuracy and decreases the docking time, especially when using multithreading [139]. Molecular docking using AutoDock or AutoDock vina require the identification of the grid box size. This is used to define the search space and docking regions to identify low energy binding pose regions. The grid box size is usually calculated based on the 3D position of active site residues [140]. Depending on the grid box (search box), docking can be blind or targeted. In blind docking, the grid box includes the entire protein surface, which allows for the detection of possible binding sites. In targeted docking, the grid box size is selected to include only part of the target protein, usually the active site or a cofactor binding site [141]. Finally, the last step is to analyze the docking result, which includes validation of docking experiment, geometric analysis and consensus scoring. Also, it is very important to visually inspect the produced result and check the bonds between the ligand and target protein [142]. In this study, structure-based virtual screening was used by applying a flexible docking strategy in both case blind docking and targeted docking by using AutoDock Vina.

4.2 Methodology

4.2.1 Target and ligand preparation

Six structures were prepared for molecular docking. These include the M1 alanyl aminopeptidase 3D structure of *P. falciparum* (PDB ID: 3Q43), the top selected model for other *Plasmodium* species (*P. malariae, P. knowlesi, P. ovale and P. vivax* (as shown in Chapter 3)) and a *Homo sapiens* structure. Target preparation starts with the removal of all water and ligand/inhibitor atoms. Then, hydrogen atoms were added to protein residues. The protonation state of protein residues was manually checked, especially for the active site residues. In the end, each target structure had all residue atoms and one zinc atom. This step was performed using discovery studio 2016.

A ligand dataset was retrieved from the SANCDB. All the retrieved ligands were in minimized form. The compound Bestatin was retrieved from the ZINC database since it has been used as protein inhibitor against M1 alanyl aminopeptidase [20]. Also, the human ligand and the *Plasmodium* ligand were isolated from the protein 3D structure which was used later in the docking validation step. In the end, the ligand dataset contained 626 compounds.

All target and ligand structures were in PDB format. However, AutoDock Vina requires the input structures go be in .pdbqt format. The Python scripts prepare_receptor4.py and prepare_ligand4.py were used to convert PDB files to .pdbqt format for the proteins and ligands respectively. Both these scripts are provided by AutoDock MGL tools. The Python scripts merge non-polar hydrogens and add polar hydrogens. The scripts also change hydrogen atom names to match the AutoDock atom type symbols. The Python scripts also identify aromatic carbons and automatically adds Gasteiger charges [132]. As shown in the Figure, this problem occurs due to the presence of zinc atom in all target structures. ESP charge calculation was used to overcome this problem. Finally, the zinc Gasteiger charge was manually assigned a value of 1.125.

4.2.2 Grid box calculation and parameter file generation

The Grid box calculation was performed by using Pymol v 1.8 and the AutoDock plugin. In the blind docking experiment, the grid box center for the human protein was set to 108.7, 20.81, 19.27 Angstroms and the grid spacing set to 46.50, 40, 46.50. For all *Plasmodium* species, the center of grid box was set to 20.002, 15.945, 3.313 and the grid box size was set to 60, 60, 60. The parameter files were generated for each target and ligand. A total of 3756 parameter files were generated by AutoDock Vina. These files were generated by python script (Appendix 1).

Each parameter file had specific information including the target and ligand pdbqt file names, x, y and z coordinates and size, and the exhaustiveness value. In both blind and targeted docking, the exhaustiveness value was set to 576.

Example of a parameter file:

```
receptor = ../Target/receptor_name.pdbqt
ligand = ../Ligand/ligand_name.pdbqt
center_x =  2
center_y =  6
center_z = -7
size_x = 25
size_y = 25
size_z = 25
exhaustiveness= 4
```

4.2.3 Molecular docking

Molecular docking was performed using AutoDock Vina. Since AutoDock Vina accepts one ligand and one target per run, a customized python script (Appendix 2) was used to automate the docking process. After running all docking experiments, another customized python script was used to submit each output file to the vina_split tool, which split the output grouped conformations as separate structures, according to their binding energy score (The lowest energy corresponds to the first produced conformational structure). The Python script then extracts the best ligand conformation with the corresponding binding energy (Appendix 3).

4.2.4 Docking validation

The ability of AutoDock Vina to reproduce the same ligand conformation for the original ligand was evaluated. The original ligands for all target structures were included in the docking experiments. Then the poses produced from docking were compared with the original poses of the ligand before docking. Discovery studio and LigPlot were used to validate and compare the ligand-target bonds and confirm that they were the same before and after docking.

4.2.5 Docking analysis

Ligands structures were converted from pdbqt to PDB format to prepare the docking result for analysis using the following command:

cut -c-66 input.pdbqt > output.pdb

Then each ligand name was stored in a text file with the corresponding binding energy. This file was used to draw a heat map using Microsoft Excel and an R-script. All ligands were analyzed to choose the best ligands. First, all docked ligands were sorted according to their binding energy. Then ligands with binding energies better than the original ligand were selected. The selected ligands were submitted to X-Score to calculate their binding affinities. The ligands were then submitted to Discovery studio, and the bonds between each ligand and target structure were counted using a Discovery studio script and a customized Python script. This python script counts the number of bonds and the bond type. The selected ligands were filtered according to the number of bonds between them and their target structure as well as those that bind the active site of the *Plasmodium* structure but not the human structure. The best ten ligands were manually visualized using Discovery studio and LigPlot.

4.3 Result and Discussion

Molecular docking was done at the Center for High-Performance Computing (CHPC) using 240 cores and 14 computing hours to dock all ligands against all target structures.

4.3.1 Grid box calculation

The zyx points of original co-crystalized ligand were identified using Discovery studio to calculate the grid box for blind docking. Then PyMol and the Autodock plugin were used to identify the grid box size. In blind docking, it was necessary to include all target protein residues inside the grid box Figure 4-1 and 4-2.

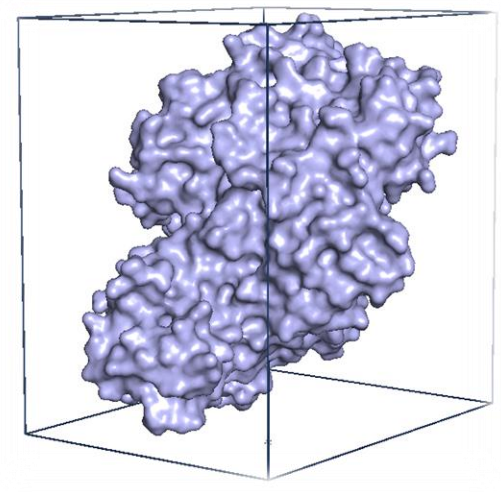


Figure 4-1: The human structure and the grid box in which all human residues were included.

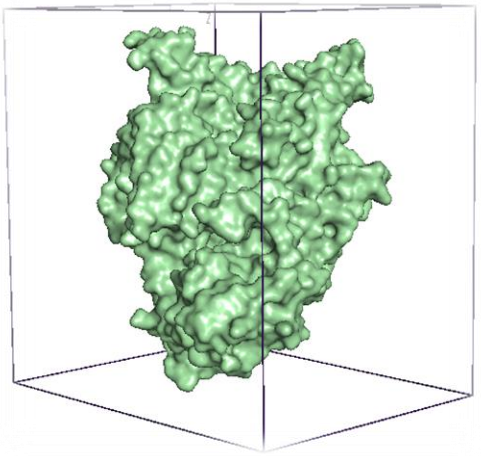


Figure 4-2: The *P. falciparum* structure and the grid box in which all human residues were included.

4.3.2 Docking validation

The docking experiment was validated by redocking the original ligands with their respective original structure. In *Plasmodium* species and human 3D structure (PDB ID 4FYR), the original co-crystalized ligand was bestatin. In all cases, the original co-crystalized ligand was removed and re-docked using AutoDock Vina. As shown in Figure 4-3 and 4-4, AutoDock Vina was able reproduce the same conformation and bonds between the ligand and the target *Plasmodium*

protein, while in the human target structure AutoDock Vina produced a similar conformation and bonds as observed from the original ligand. Thus, the ability of AutoDock Vina to produce the correct conformation poses was validated.

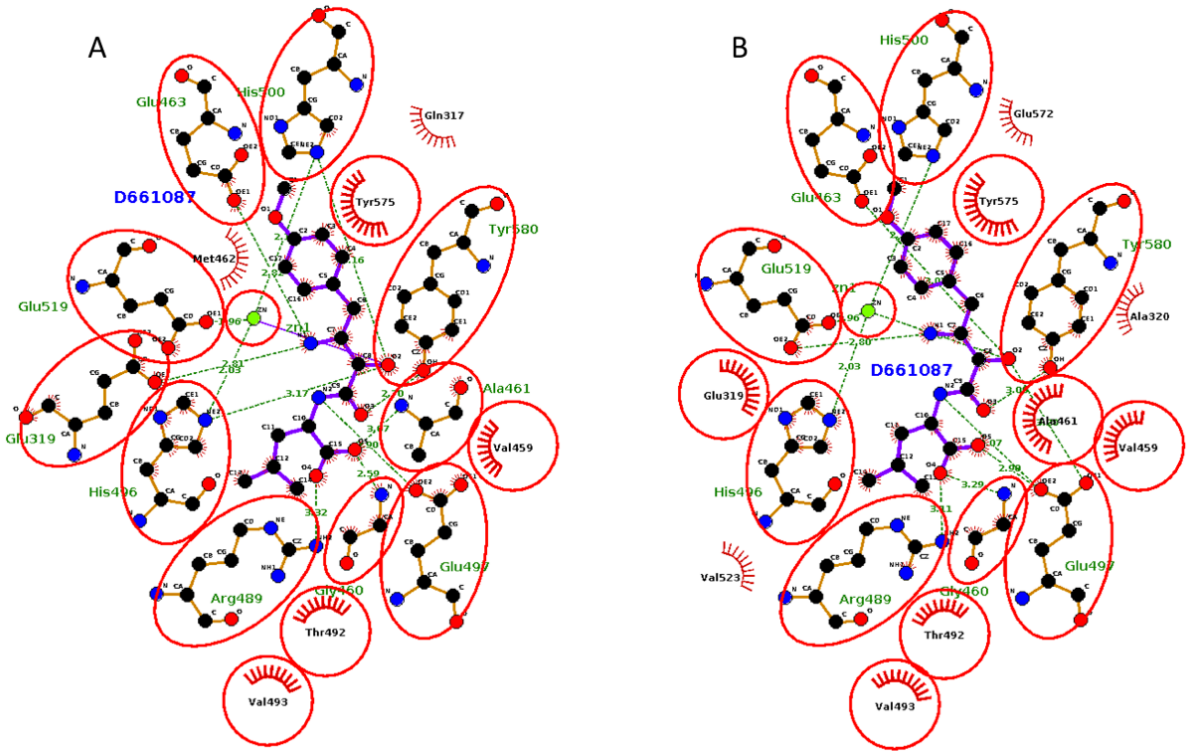


Figure 4-3: Ligand-Target 2D interactions, created by LigPlot for *P. falciparum* target, where compound D661087 is the original co-crystalized ligand. A: original interactions between the ligand and its target before docking. B: the original co-crystalized ligand with the target after redocking.

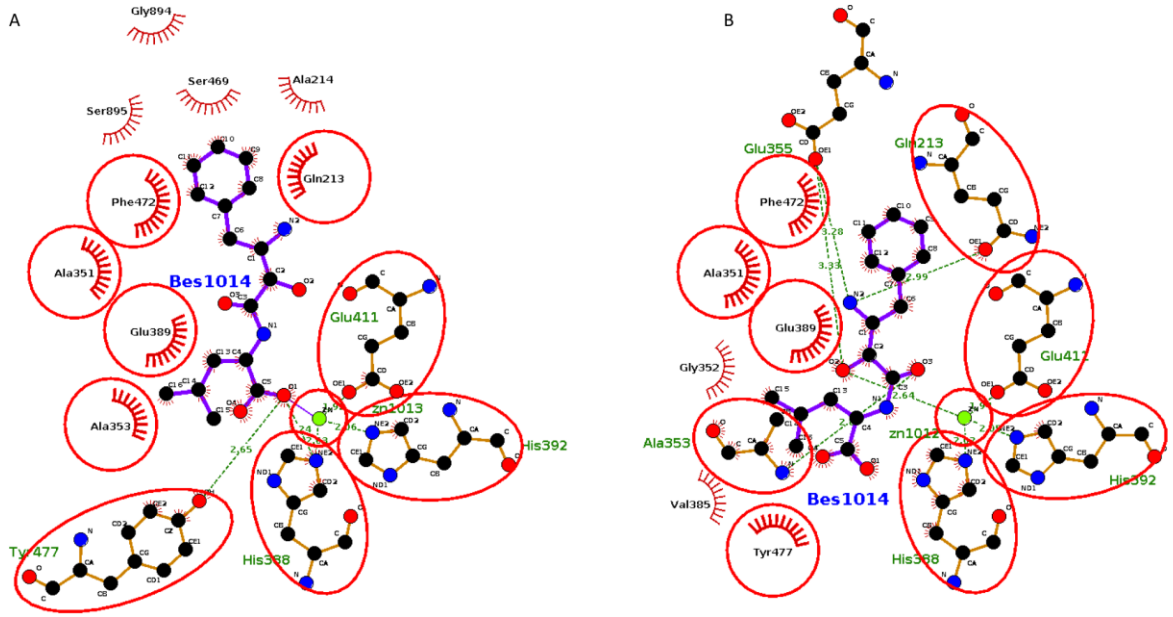


Figure 4-4: Ligand-Target 2D interactions, created by LigPlot for the human target. Compound Bes1014 is the original co-crystalized ligand. A: original interactions between ligand and target

before docking. B: the original co-crystalized ligand with its target after redocking. While the conformation slightly changed. but in both cases the active site residues bind with the ligand.

4.3.3 Docking analysis

All retrieved SANCDB compounds were in an energy-minimized form so, they were submitted directly for virtual screening. Each AutoDock Vina run produces two different files. The first file is a log file, which is used to capture the binding energy. The second file is a pdbqt file that contains different conformation poses for the same ligand. Vina_split was used to split these poses. Then the lowest binding energy poses were selected among them. All ligand names with their corresponding binding energy were captured in a Microsoft Excel file. This file was used to produce a heat map (Figure 4-5). As shown in the heat map, there is a huge difference in binding energy between the human and the *Plasmodium* species. The difference ranges from 5.5 Kcal/mol to 0.4 Kcal/mol. To select the best binders, all ligands having a binding energy lower than that of the original ligand were selected. The docked ligands in *Plasmodium* species which bind to the active site as well as other allosteric sites in the structure are shown in Figure 4-6. In the human protein (as shown in Figure 4-7) the majority of the ligands bound to allosteric regions but not the active site. Therefore, ligand selection based on binding energy difference was not used to select the best ligand.

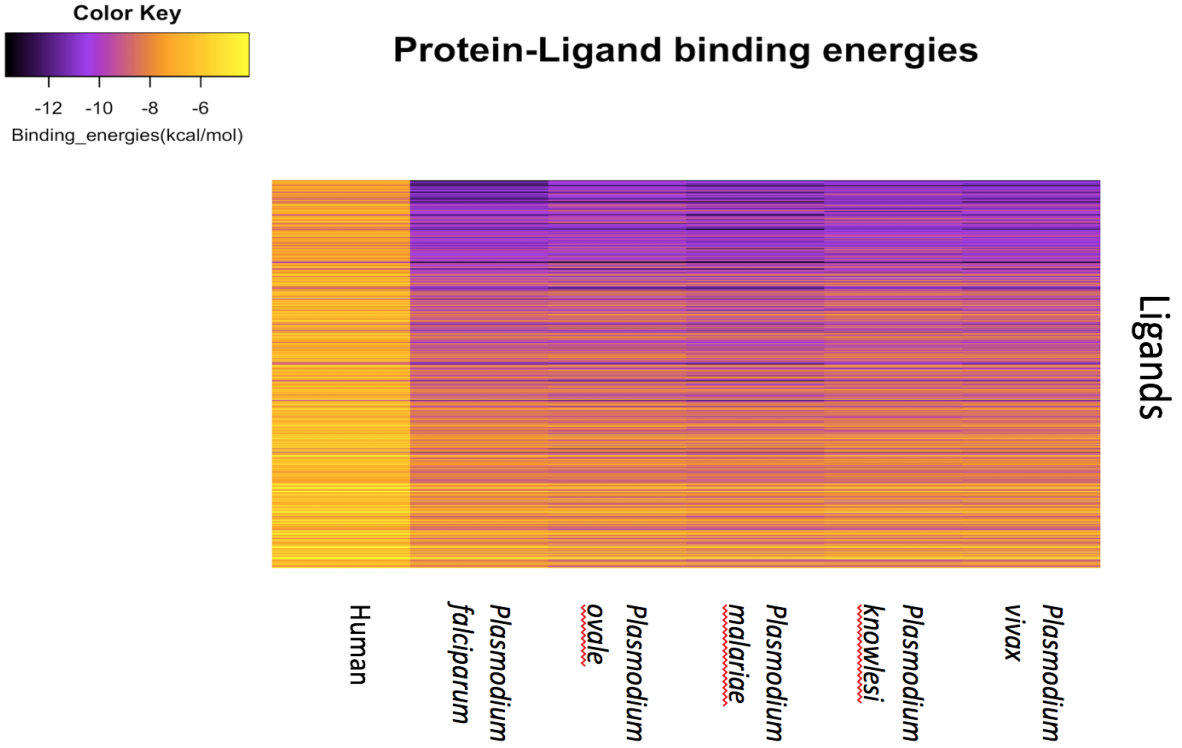
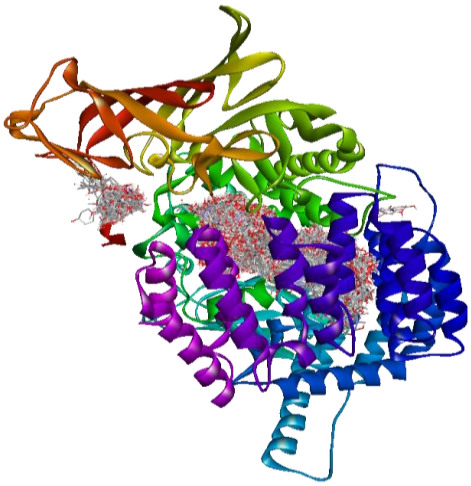
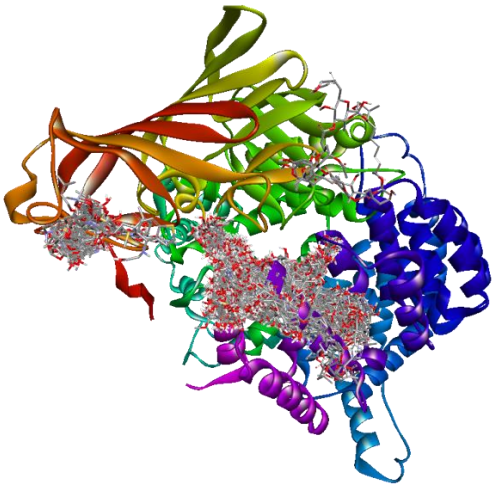


Figure 4-5: Heatmap for all docked compounds against the M1 alanyl aminopeptidase of human and *Plasmodium species*. The dark violet color corresponds to a high binding affinity and a low binding energy, while the yellow corresponds to the low binding affinity and high binding energy.

A) *P. falciparum*

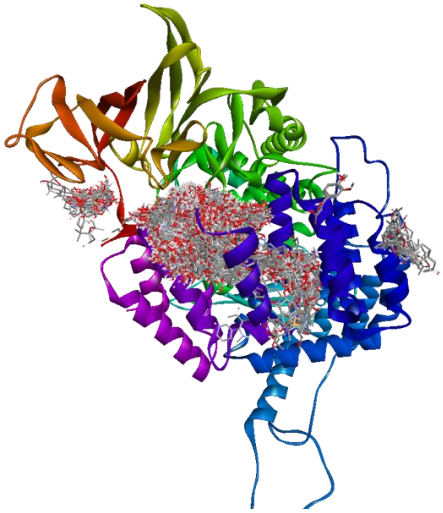


B) *P. knowlesi*



C) *P. vivax*

D*) P. malariae*



E*) P. ovale*

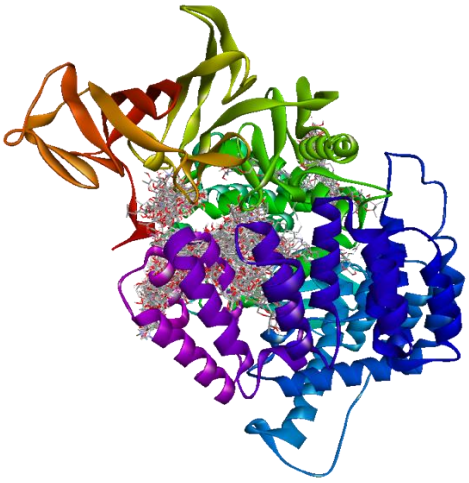


Figure 4-6: Protein-ligand complexes. A) M1 Alanyl aminopeptidase *P. falciparum* structure and all ligand complexes. B) A:M1 Alanyl aminopeptidase *P. knowlesi* structure complexed to all ligands. C) A: M1 Alanyl aminopeptidase *P. vivax* structure complexed to all ligands. D) A: M1 Alanyl aminopeptidase *P. malariae* structure complexed to all ligands. D) A: M1 Alanyl aminopeptidase *P. ovale* structure complexed to all ligands.

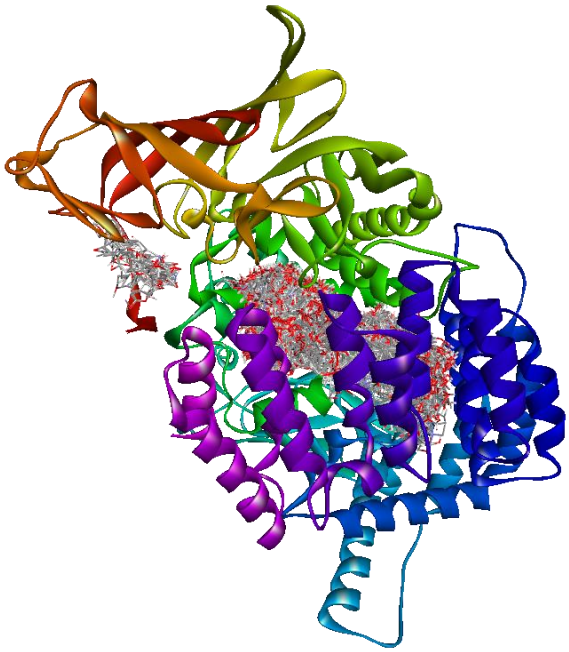


Figure 4-7: M1 Alanyl aminopeptidase human structure complexed to all ligands.

The first selection process involved selecting ligands which have binding energies lower than original ligand. 265 ligands were selected in the case of M1 *P. falciparum* alanyl aminopeptidase. Other organisms are shown in Table 4-1. (Figure 4-8).

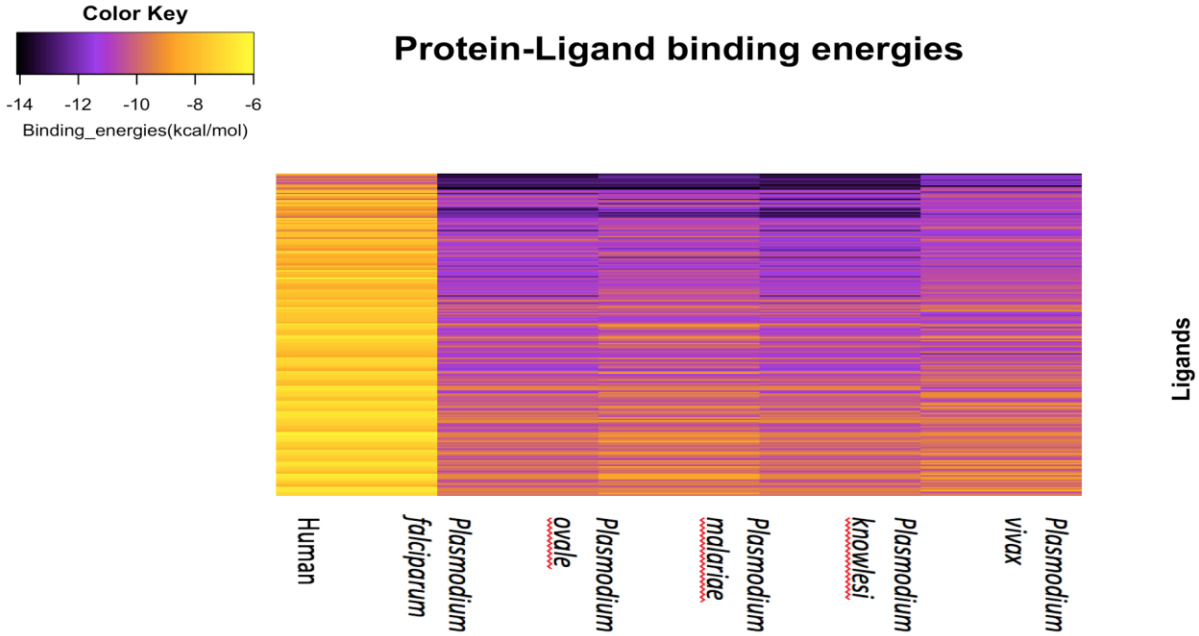


Figure 4-8: Heatmap for ligands with binding energies higher than that of the original ligand against the human and plasmodial M1 alanyl aminopeptidases. The dark violet color corresponds

to a high binding affinity and a low binding energy, while the yellow stripes correspond to a low binding affinity and a high binding energy.

The next step was selecting ligands that bind the active site of *Plasmodium* species. A protein-ligand interaction script implemented in Discovery studio and was used to calculate the bond between ligand and active site residues automatically. This generated a text file containing the ligand file name and the found interactions. An example of generated output is shown below:

plasmo_falci.pdbqt_SANC00170_minRM1.vinaall_ligand_1.pdbqt
Found 2 non-bond interactions (total):
2 of these are favorable interactions (such as H-bonds)
0 of these are unfavorable interactions (such as bumps).
Analyze all non-bond interaction:
The NonbondTypes property can be used to identify all interaction types of a non-bond.
- A:THR896:OG1 (H-Donor) and :LIG1:O (H-Acceptor):conventionalHBondType
- A:ASN899:ND2 (H-Donor) and :LIG1:O (H-Acceptor):conventionalHBondType

This text file was analyzed using a customized python script to select ligands which interact with the active site residues and the zinc atom. This selection process resulted in 58 ligands for the M1 alanyl aminopeptidase of *P. falciparum*, as shown in Table 4-1. All ligands that bind using at least one hydrogen bond to one of the active site residues were selected. The result generated from X-Score was similar to binding energy generated by AutoDock Vina, as shown in Figure 4-9.

Table 4-1: Number of selected ligands in ligand selection steps for each target organism.

Target organism name	First selection step (Ligand with binding energy lower than original ligand)	Second selection step (Ligand bind with active site residues)
Plasmodium falciparum	265	58
P. knowlesi	263	57
P. ovale	265	58
P. vivax	261	54
P. malariae	263	57

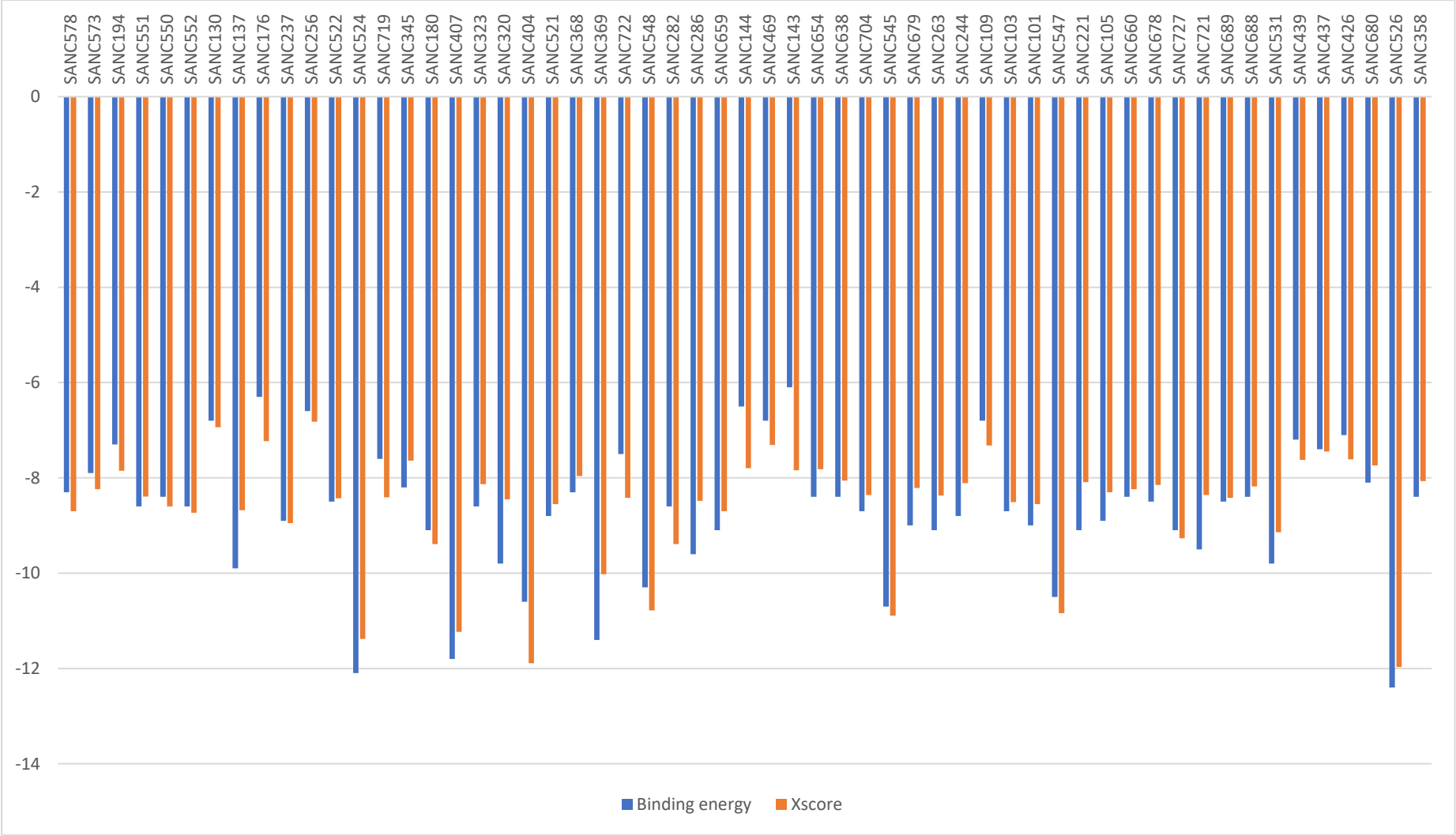


Figure 4-9: Graphical representation of X-Score result and binding energy for each ligand docked against the M1 alanyl aminopeptidase of *Plasmodium falciparum*. X-Score values are in orange and binding energy scores are in blue.

The next step was aimed at selecting the best ten ligands based on hydrogen bond between the ligand and the corresponding target protein. Additionally, other bonds between the ligand and target protein were taken into consideration, together with ligand efficiency. Durability properties based on Lipinski's rule of five [143] and possible unfavorable bond interactions were also considered. The first step comprised the elimination of ligands with unfavorable bonding interactions with the target protein. For example, in *P. falciparum* out of 58 ligands, 19 ligands were eliminated because they had at least one unfavorable interaction, which consisted of bumps. (Table 4-2).

Table 4-2: Summary of eliminated ligands representing the number of unfavorable bonding interactions between each ligand and their *P. falciparum* target protein structure.

Ligand name	Number of unfavorable interactions
SANC00545	4
SANC00548	4
SANC00286	3
SANC00244	2
SANC00263	2
SANC00320	2
SANC00680	2
SANC00526	2
SANC00369	1
SANC00368	1
SANC00323	1
SANC00547	1
SANC00282	1
SANC00426	1
SANC00407	1
SANC00404	1
SANC00521	1
SANC00180	1
SANC00137	1

The remaining 39 ligands were analyzed and the number of favorable interactions were captured. In *P. falciparum* the number of interactions ranges from 18 to 5, as shown in Figure 4-10.

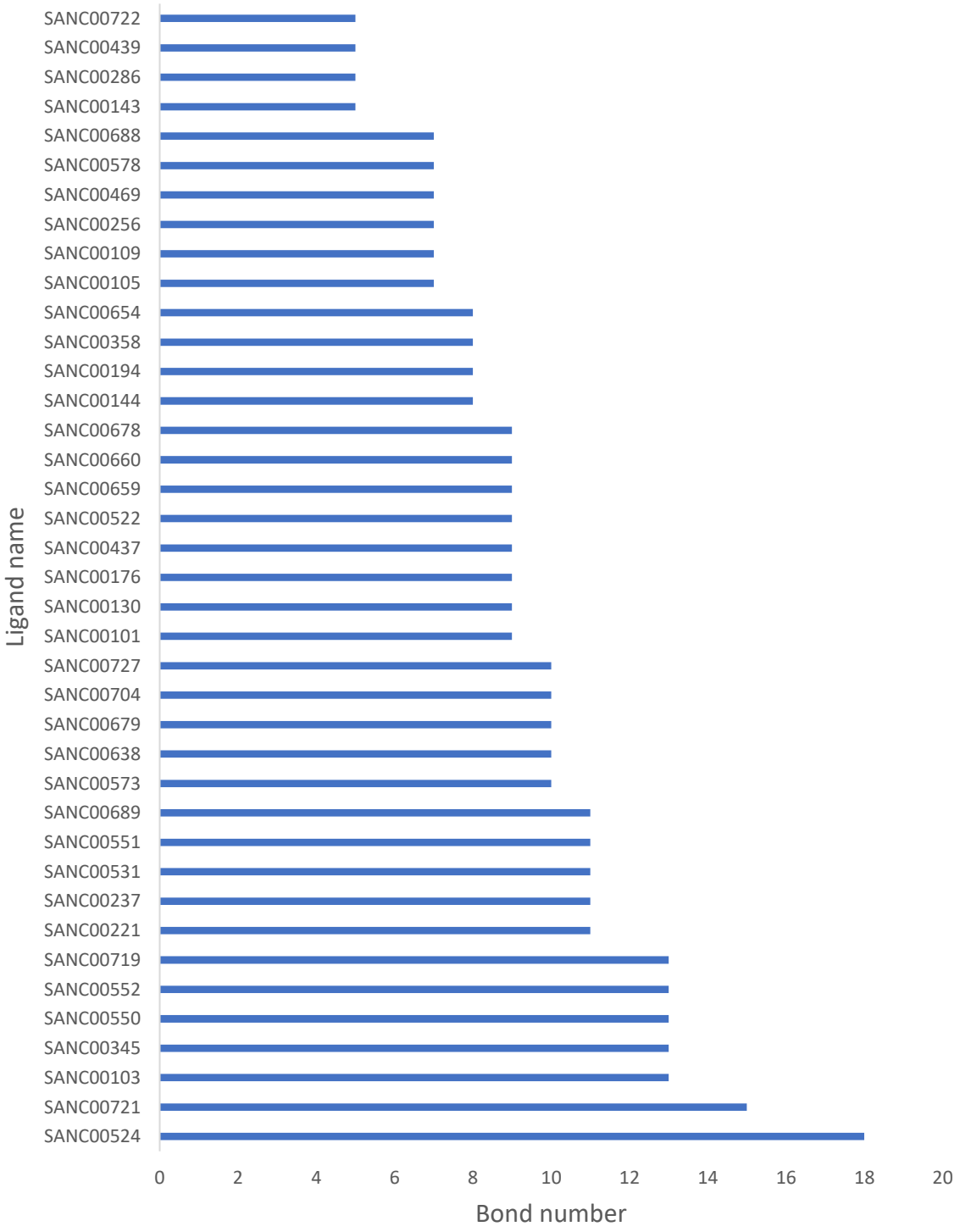


Figure 4-10: Graphical representation of the number of bonding interactions between each ligand and their M1 alanyl aminopeptidase in the case of *P. falciparum*, sorted in ascending order.

Ligands with more than ten favorable interactions were selected and submitted to FAF-Drugs4, which measure the durability characteristics to apply Lipinski's rule of five. Lipinski's rule tests 4 properties including molecular weight, lipophilicity, hydrogen bond donors and hydrogen bond acceptors. A ligand is considered acceptable if it passes three of the properties.

Ligands that fail in two or more properties were considered poorly-absorbed[143]. As shown in Table 4-3, SANC00524 was eliminated, while SANC00103, SANC00719, and SANC00237 were accepted as they failed in only one parameter.

Table 4-3: The tabulated result of Lipinski's test for the best ten ligands against M1 alanyl aminopeptidase from *Plasmodium falciparum*. The acceptable values for each parameter are: molecular weight ≤ 500, lipophilicity ≤ 5, hydrogen bond donors ≤ 5 and hydrogen bond acceptors ≤ 10

Ligand Name	molecular weight	lipophilicity	hydrogen bond donors	hydrogen bond acceptors	Status
SANC00524	848.75	-2.61	13	21	Fail
SANC00721	284.26	3.04	2	5	Pass
SANC00103	306.27	0.15	6	7	Accepted
SANC00345	329.35	1.34	0	6	Pass
SANC00550	456.49	3.42	0	8	Pass
SANC00552	442.5	4.39	0	7	Pass
SANC00719	304.42	6.66	2	3	Accepted
SANC00221	314.29	1.38	2	6	Pass
SANC00237	392.53	5.48	0	5	Accepted
SANC00531	354.4	4.64	2	5	Pass

The next step was to manually analyze the best six filtered ligands. This step was done using Discovery studio and LigPlot. The aim of this step was to select for the best ligands based on residues which interact with the ligand and the number of hydrogen bonds.

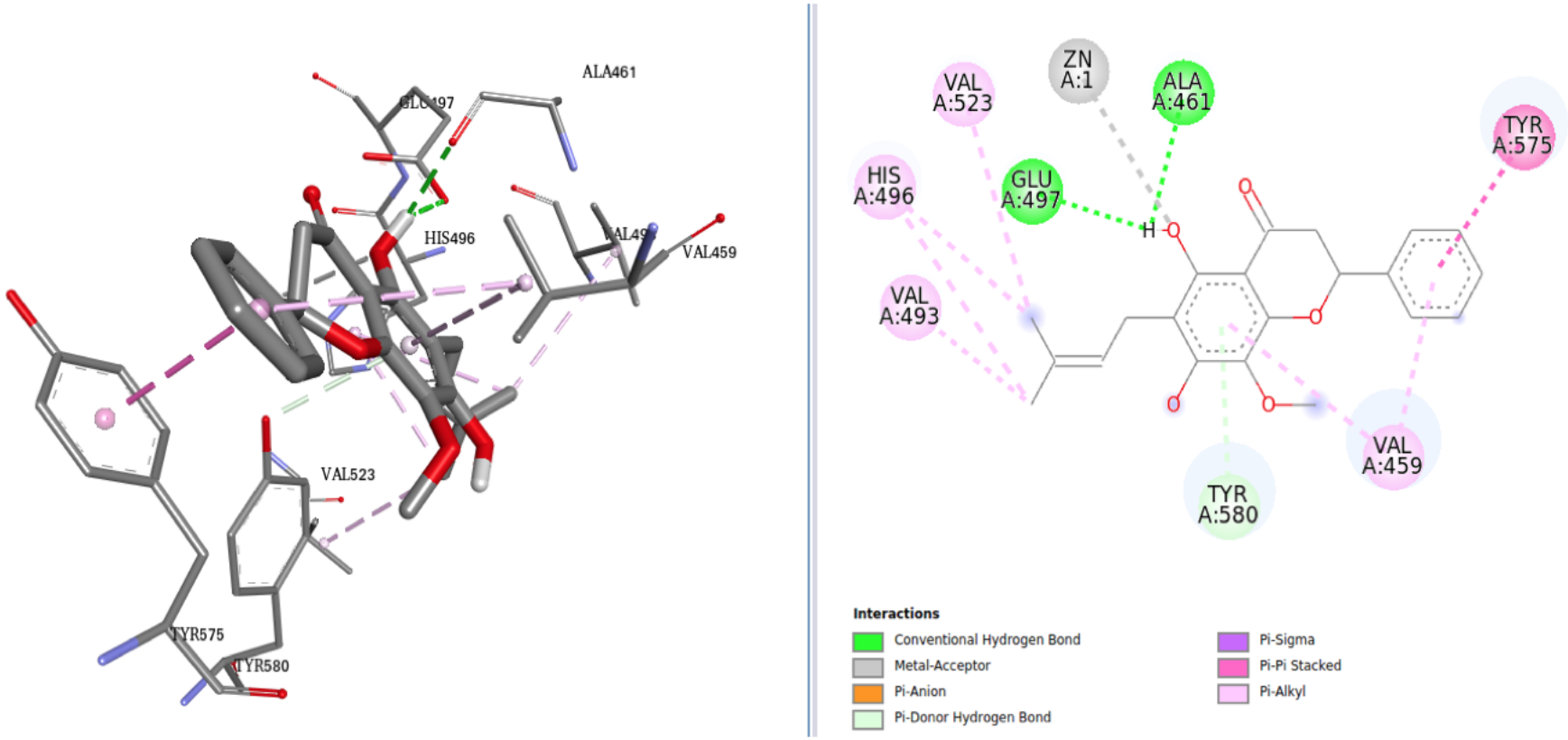


Figure 4-11: Graphic representation showing the interactions between Ligand SANC0531 and M1 alanyl aminopeptidase from the *P. falciparum* protein.

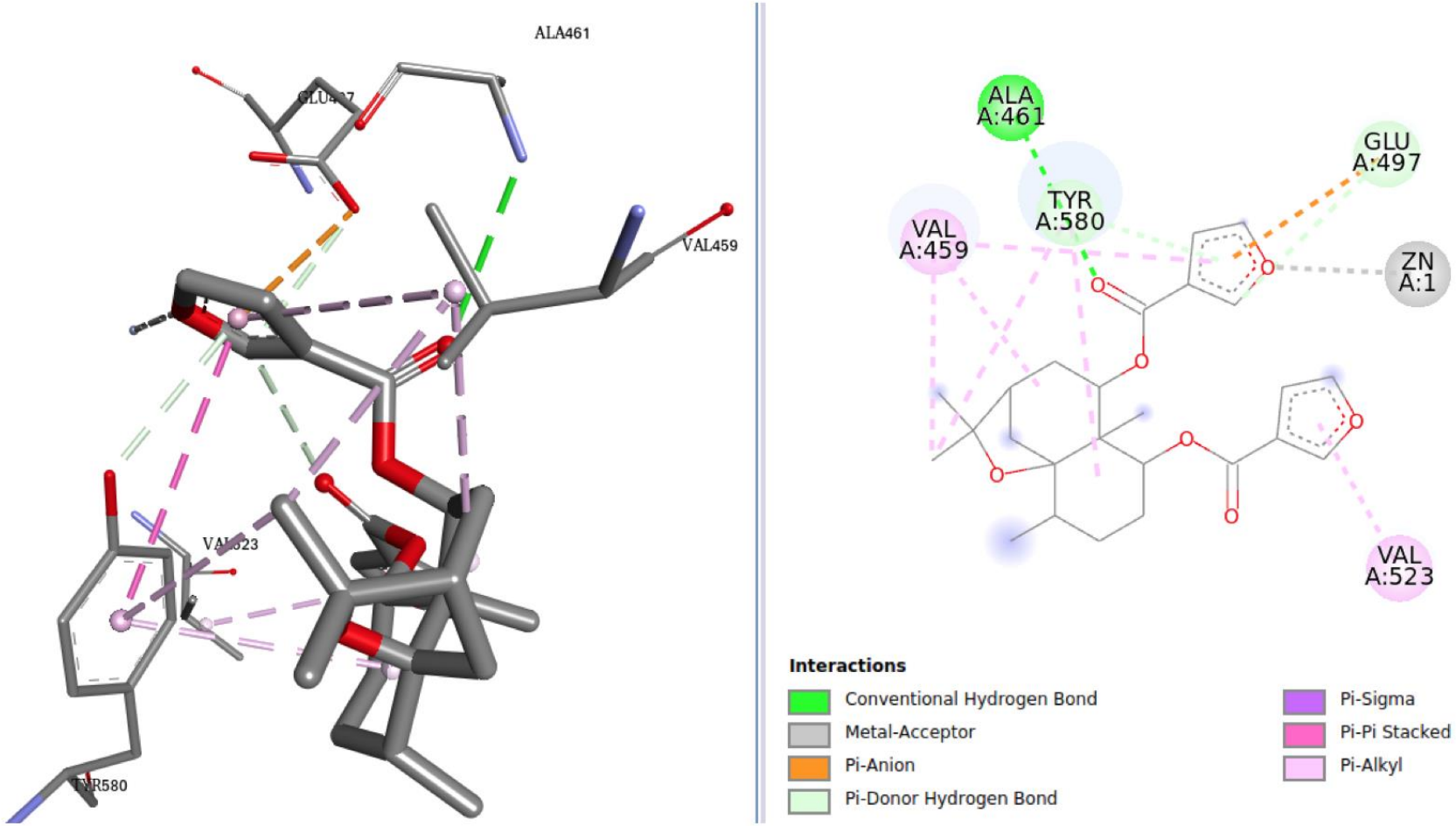


Figure 4-12: Graphic representation shows the interactions between Ligand SANC0552 and the M1 alanyl aminopeptidase from the *P. falciparum* protein.

From the selected ten ligands, SANC0531 was selected as the best ligand for the M1 alanyl aminopeptidase of *P. falciparum.* As shown in Figure 4-11, SANC0531 interacts with histidine number 496 which is one of the active site residues. Also, it has a hydrogen bond with the zinc metal ion. The next ligand after SANC0531 was SANC0552, which has a hydrogen bond with alanine number 461, but this residue is not located in the active site as well as another hydrogen bond with the Zinc metal ion as shown in Figure 4-12. In Figure 4-13 a hydrogen bond can be seen between the ligand, the zinc metal ion and glutamine number 497, which increases the bond stability between the ligand and the target protein [144].

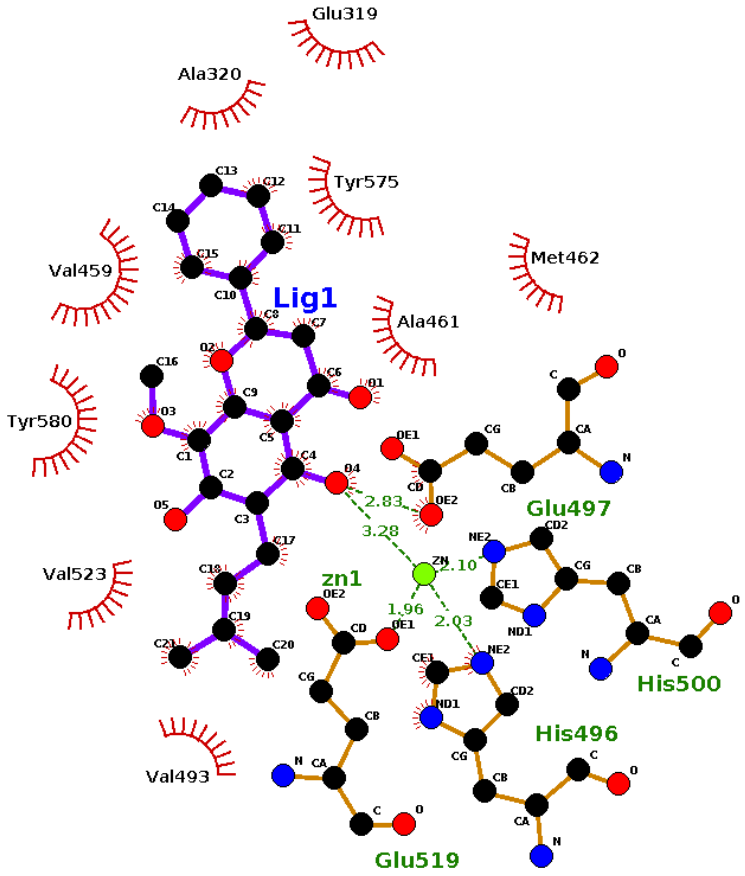


Figure 4-13: Graphical representation created by LigPlot for SANC00531 and the M1 alanyl aminopeptidase of *P. falciparum* the protein. Hydrogen bonds are shown in green.

The top 10 ligands for *P. knowlesi* (Table 4-4), *P. ovale* (Table 4-5), *P. vivax* (Table 4-6) and *Plasmodium malariae* (Table 4-7) were selected and submitted to FAF-Drugs4.

Table 4-4: The tabulated result of Lipinski's test for the best ten ligands against M1 alanyl aminopeptidase of *P. knowlesi.*

Ligand Name	molecular weight	Lipophilicity	hydrogen bond donors	hydrogen bond acceptors	Status
SANC00137	317.34	2.28	2	5	Pass
SANC00143	311.46	4.95	1	4	Pass
SANC00719	304.42	6.66	2	3	Accepted
SANC00176	168.23	2.03	1	2	Pass
SANC00654	331.36	0.36	2	6	Pass
SANC00659	348.39	1.48	0	6	Pass
SANC00469	164.16	1.46	2	3	Pass
SANC00404	955.13	3.92	9	18	Fall
SANC00407	594.78	3.92	5	9	Accepted
SANC00426	230.26	0.8	2	5	Pass

Table 4-5: Tabulated result of Lipinski's test for the best ten ligands against M1 alanyl aminopeptidase of *P. ovale.*

Ligand Name	molecular weight	Lipophilicity	hydrogen bond donors	hydrogen bond acceptors	Status
SANC00144	295.42	4	1	4	Pass
SANC00323	330.33	2.74	2	6	Pass
SANC00426	230.26	0.8	2	5	Pass
SANC00638	289.33	0.75	2	5	Pass
SANC00130	194.19	0.06	2	5	Pass
SANC00524	848.75	-2.61	13	21	Fall
SANC00578	352.47	2.38	2	5	Pass
SANC00526	980.87	-4.69	15	25	Fall
SANC00550	456.49	3.42	0	8	Pass
SANC00547	650.84	2.57	6	10	Fall

Table 4-6: Tabulated result of Lipinski's test for the best ten ligands against the M1 alanyl aminopeptidase of *P. vivax.*

Ligand Name	molecular weight	Lipophilicity	hydrogen bond donors	hydrogen bond acceptors	Status
SANC00660	306.35	0.91	1	5	Pass
SANC00282	476.6	2.84	2	7	Pass
SANC00286	290.27	0.51	5	6	Pass
SANC00320	288.25	2.02	4	6	Pass
SANC00521	256.25	3.18	3	4	Pass
SANC00680	304.34	3.34	2	5	Pass
SANC00704	244.24	2.48	1	4	Pass
SANC00137	317.34	2.28	2	5	Pass

SANC00130	194.19	0.06	2	5	Pass
SANC00176	168.23	2.03	1	2	Pass

Table 4-7: Tabulated result of Lipinski's test for the best ten ligands against the M1 alanyl aminopeptidase of *Plasmodium malaraie*

Ligand Name	molecular weight	Lipophilicity	hydrogen bond donors	hydrogen bond acceptors	Status
SANC00101	290.27	0.51	5	6	Pass
SANC00103	306.27	0.15	6	7	Accepted
SANC00105	256.25	2.88	2	4	Pass
SANC00407	594.78	3.92	5	9	Accepted
SANC00426	230.26	0.8	2	5	Pass
SANC00551	454.47	2.91	0	8	Pass
SANC00552	442.5	4.39	0	7	Pass
SANC00689	298.29	3.29	2	5	Pass
SANC00722	324.41	3.69	0	5	Pass
SANC00109	239.31	1.32	1	4	Pass

In order to select the best ligand for each *Plasmodium* species, Discovery studio and LigPlot were used to manually investigate each ligand interaction with the target protein and to determine which residues interact with the ligand and hydrogen bond between the ligand and the target protein.

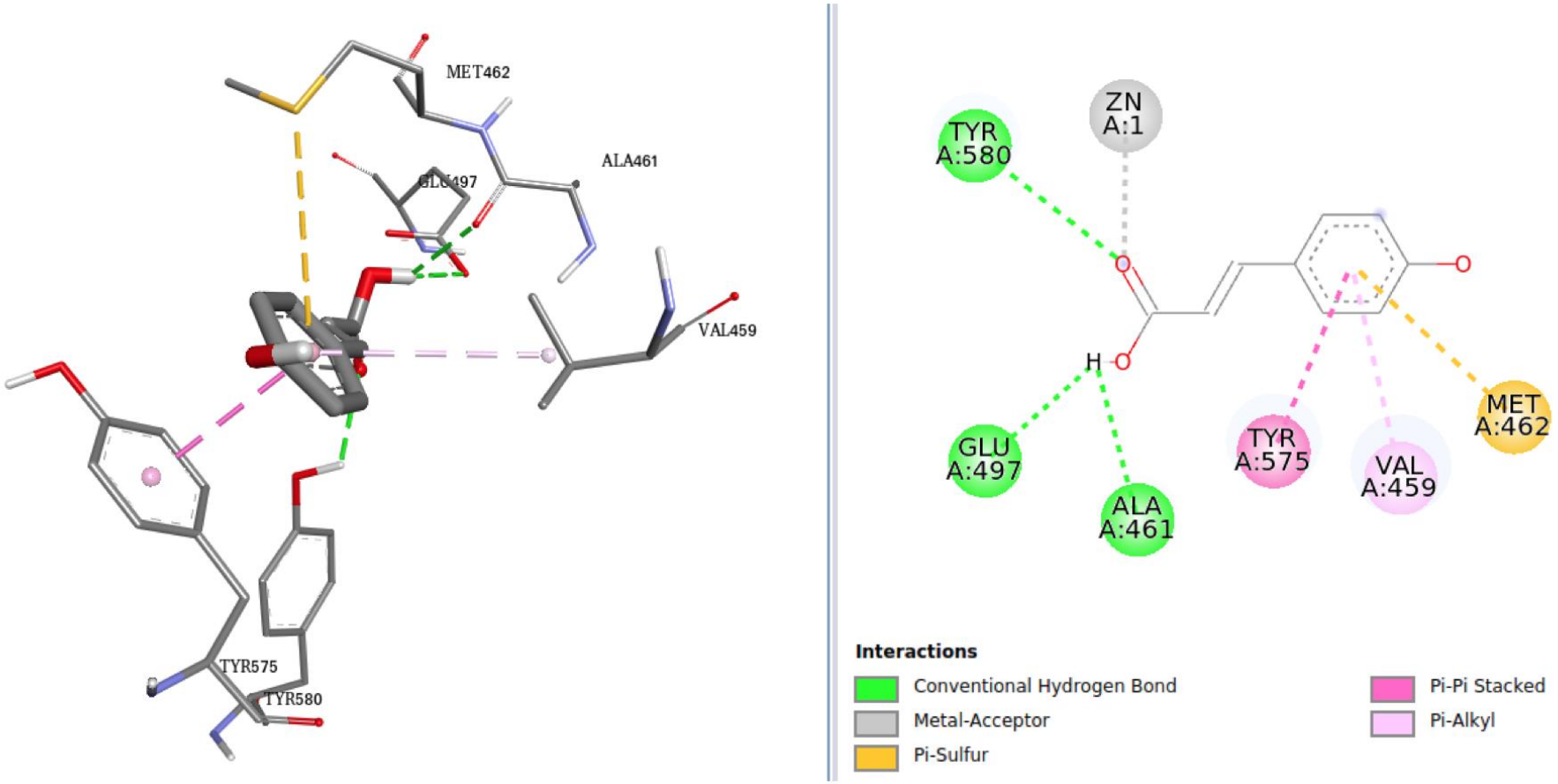


Figure 4-14: Graphical representation showing the interactions between Ligand SANC0469 and the M1 alanyl aminopeptidase from the *P. knowlesi* protein.

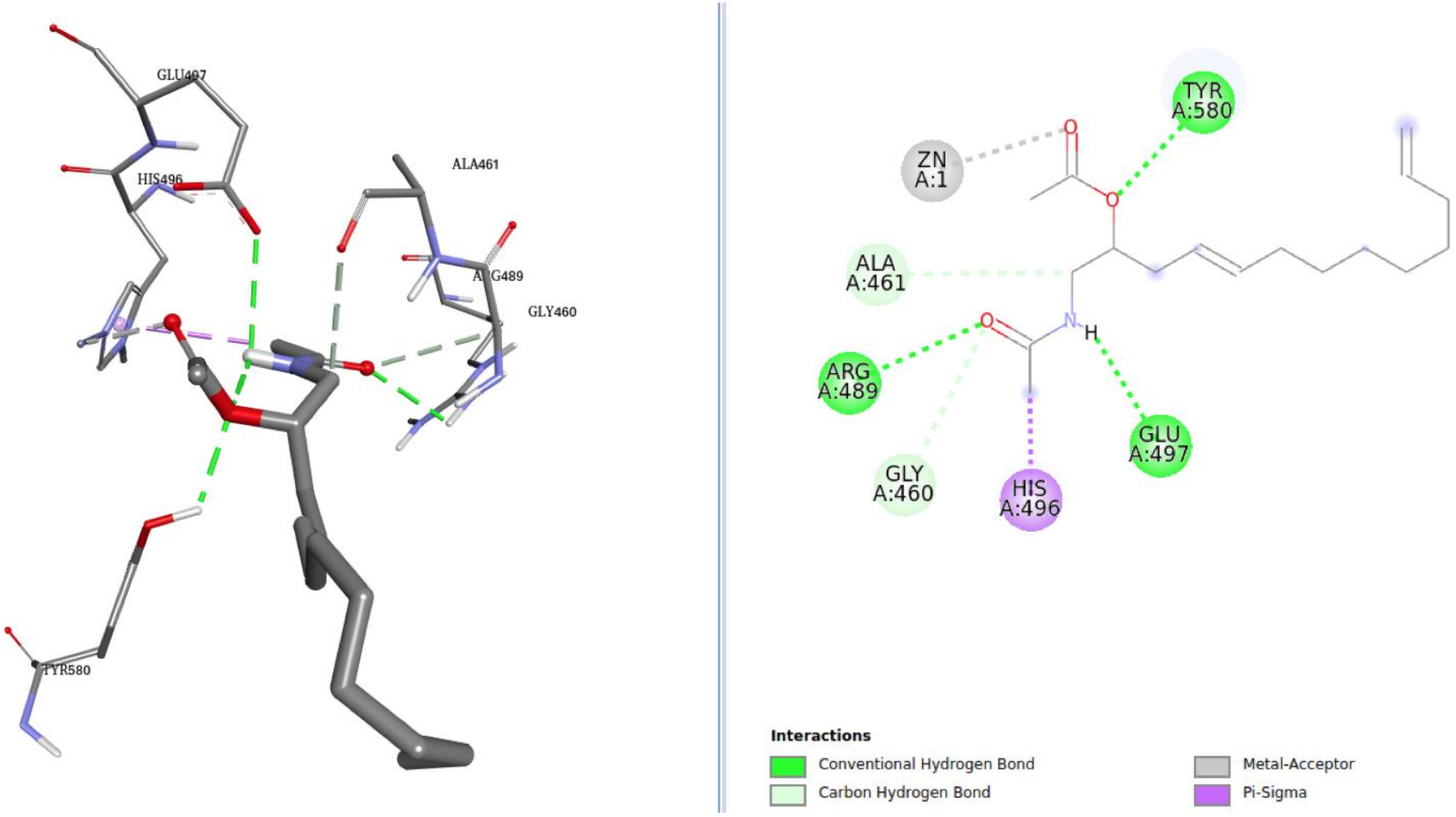


Figure 4-15: Graphical representation showing the interactions between Ligand SANC0144 and the M1 alanyl aminopeptidase from the *P. ovale* protein.

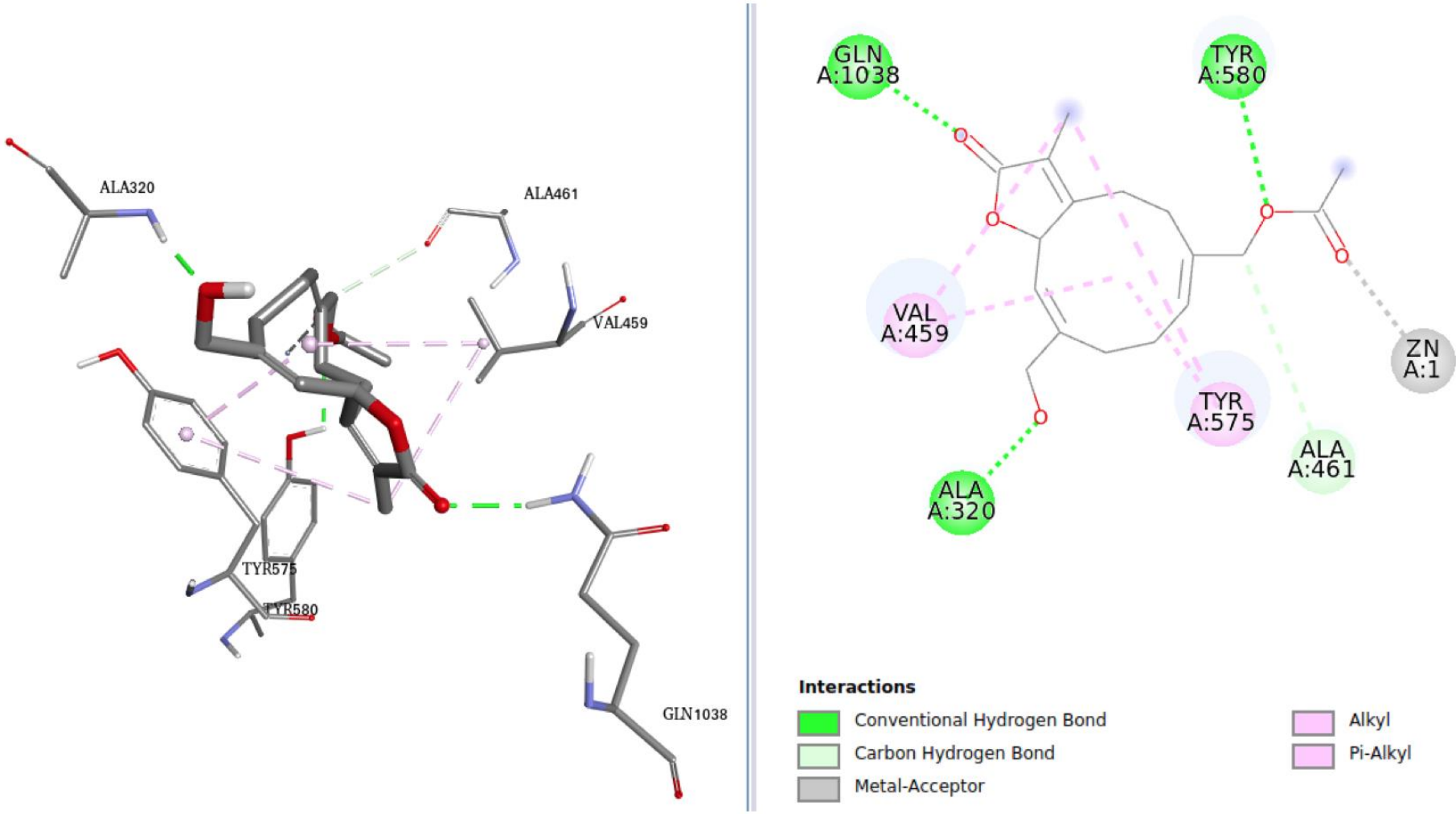


Figure 4-16: Graphical representation showing the interactions between Ligand SANC0660 and M1 alanyl aminopeptidase of *P. vivax* protein.

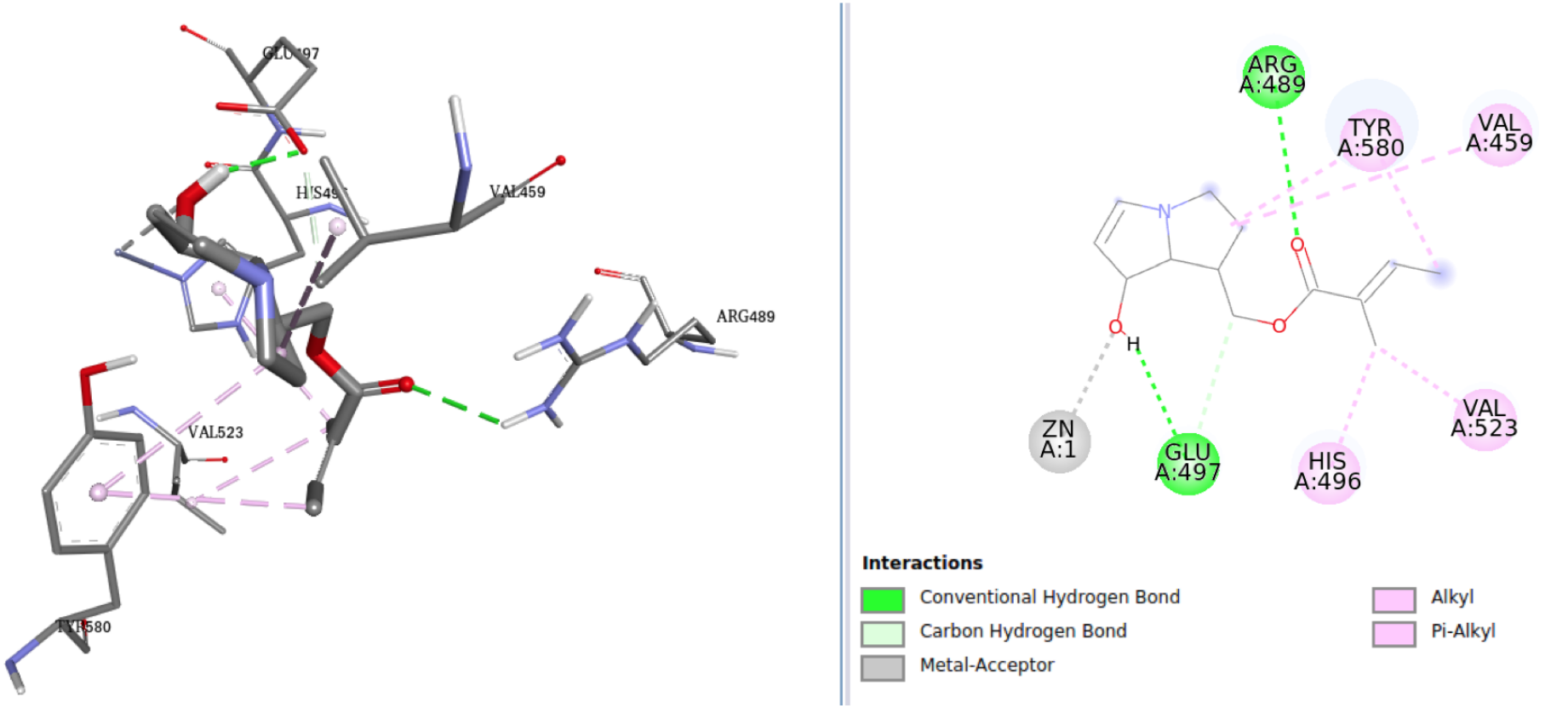


Figure 4-17: Graphical representation showing the interactions between Ligand SANC0109 and M1 alanyl aminopeptidase of *Plasmodium malariae* protein.

Ligand SANC00469 was selected for the M1 alanyl aminopeptidase of *P. knowlesi* as it interacts with the target protein with three hydrogen bonds, while other ligands had one hydrogen bond with the target protein. In *P. ovale*, SANC00144 was selected because it interacts with histidine number 496, which located in the active site. SANC00144 has three hydrogen bonds with the target protein. For the M1 alanyl aminopeptidase of *P. vivax*, Ligand SANC00660 was selected because it passes all of Lipinski's tests in addition to having more hydrogen bonds compared to the other ligands. Finally, SANC00109 was selected as the best ligand for M1 alanyl aminopeptidase of *P. malariae* because it binds to histidine number 496, which is located in the active site. SANC00109 also forms a hydrogen bond with glutamine number 497 and arginine number 489. To analyze top selected ligand interactions with the human M1 alanyl aminopeptidase, LigPlot was used. For all ligands, there is no interaction with the active site residues of the human protein. As shown in Figures 4-17,18,19,20 and 21, none of the top selected ligands bind the active site of the human protein.

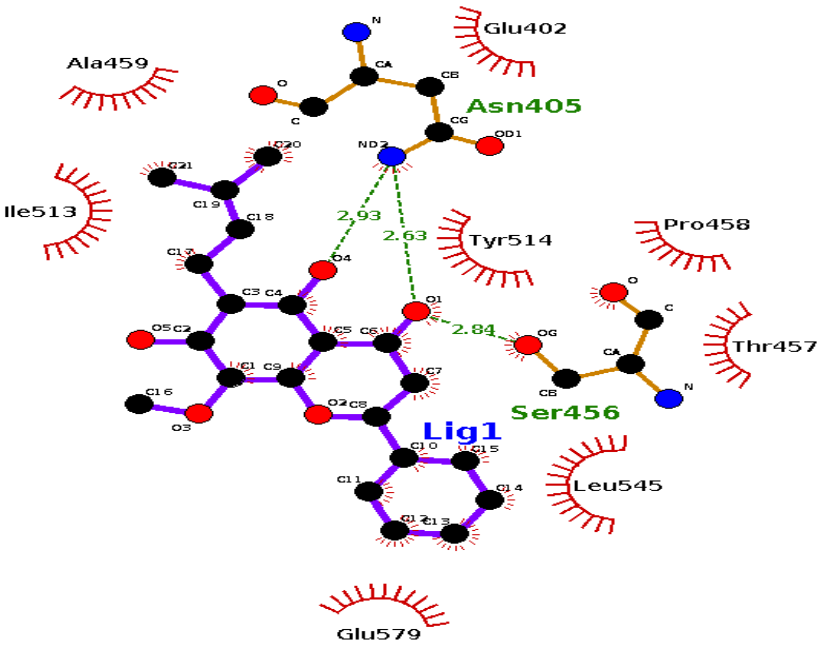


Figure 4-18: Graphical representation created by LigPlot for SANC00531 and M1 alanyl aminopeptidase of *Homo sapiens* protein. Hydrogen bonds are coloured green.

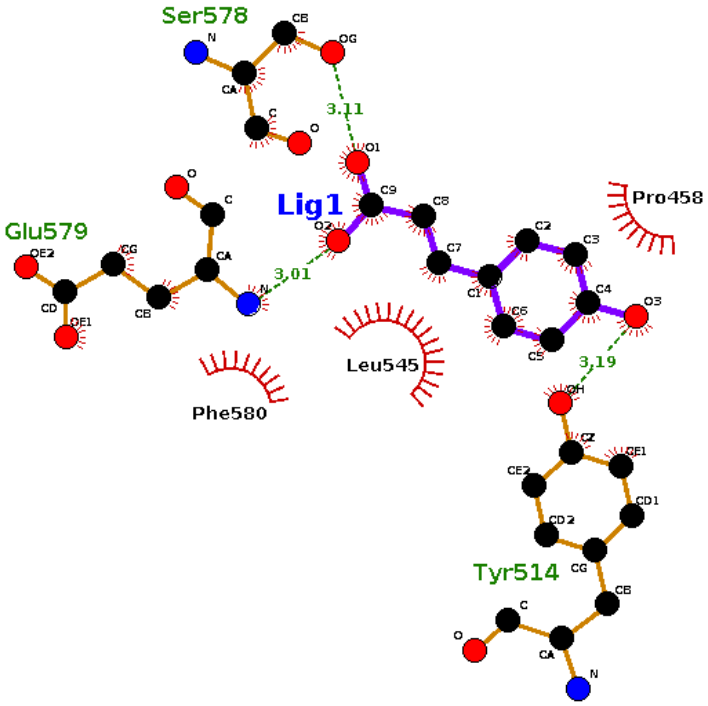


Figure 4-19: Graphical representation created by LigPlot for SANC00469 and M1 alanyl aminopeptidase of *Homo sapiens* protein. Hydrogen bonds are coloured green

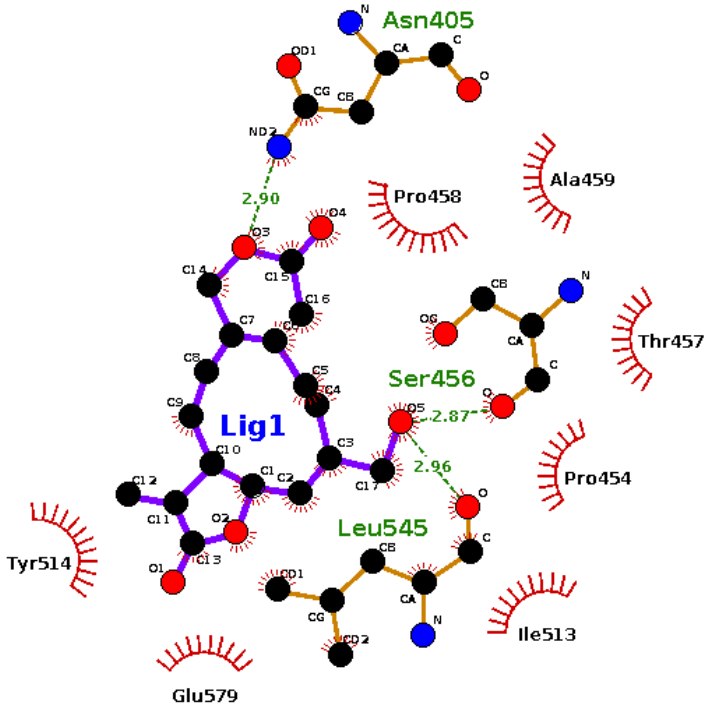


Figure 4-20: Graphical representation created by LigPlot for SANC00660 and M1 alanyl aminopeptidase of Homo sapiens protein. Hydrogen bonds are coloured green

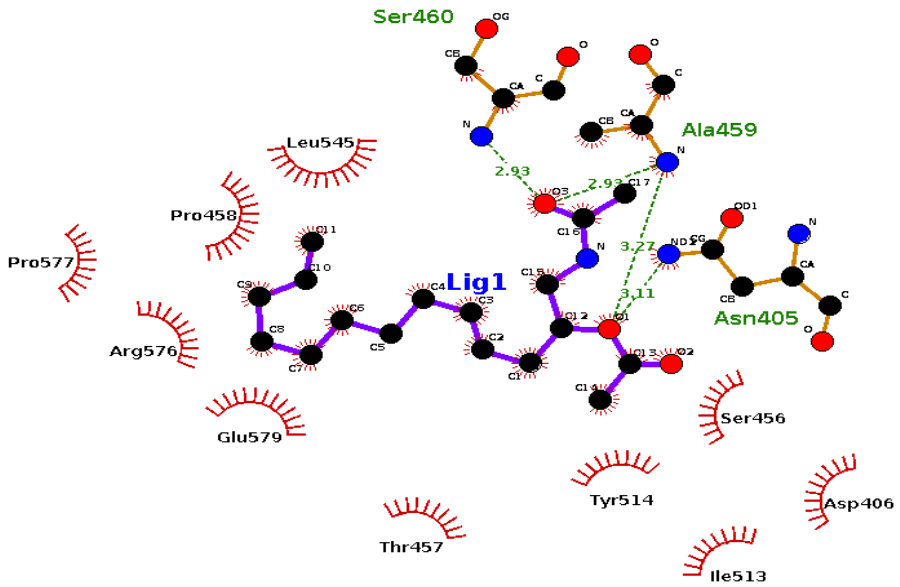


Figure 4-21: Graphical representation created by LigPlot for SANC00144 and M1 alanyl aminopeptidase of Homo sapiens protein. Hydrogen bonds are coloured green

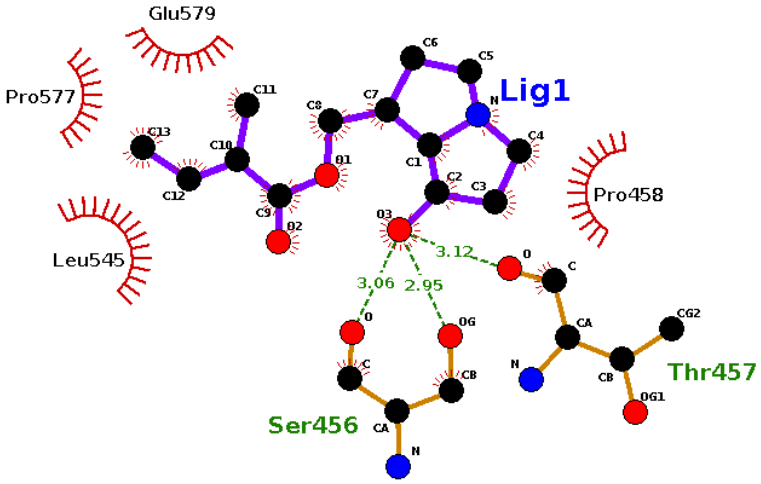


Figure 4-22: Graphical representation created by LigPlot for SANC00109 and M1 alanyl aminopeptidase of Homo sapiens protein. Hydrogen bonds are coloured green

4.4 Conclusion

In this chapter, 623 compounds were retrieved from SANCDB in energy minimized form. These compounds were virtually screened against *Plasmodium* parasite proteins and a human protein. The molecular virtual screening retrieved compounds with selective inhibition against the parasite protein and not its human counterpart. Blind docking was used to perform the virtual screening on all the retrieved compounds. Then the compounds were curated according to their binding energy, hydrogen bonding and binding to the active site or metal-coordinating residues. All selected ligands pass the Lipinski's rule of 5. The selected ligands were chosen such that they interact with the target protein, including active site residues and their ability to hydrogen bond the zinc metal ion. None of the ligands bound the active site of the human protein. In the end ligand, SANC00531 was selected against *P. falciparum,* SANC00469 against *P. knowlesi*, SANC00660 against *P. vivax*, SANC00144 against *P. ovale* and SANC00109 against *P. malariae.*

Chapter 5 - Summary and future perspectives

M1 alanyl aminopeptidase protein sequences from *Plasmodium* sp., bacteria, fungus, human and mammals were retrieved from the NCBI nucleotide, UniProt and Ensembl databases. Retrieved sequences went through different sequence and comparative analysis techniques, starting with motif and domain identification, followed by multiple sequence analysis and phylogenetic analysis. Domain analysis showed the presence of the Peptidase family M1 domain at almost the same position in *Plasmodium* M1 alanyl aminopeptidase. Also, it showed presence of different domains, if we compare human and *Plasmodium* M1 alanyl aminopeptidase domains. Motif analysis showed many common motifs between different M1 alanyl aminopeptidase retrieved from different *Plasmodium* species, while it showed few common motifs between mammals (including humans) and *Plasmodium* species, using protein sequences. Multiple sequence alignment confirms motif and domain analysis findings in which all M1 alanyl aminopeptidases from *Plasmodium* sequences shared a high similarity, which significantly decreased when *Plasmodium* M1 alanyl aminopeptidase sequences were compared to mammalian alanyl aminopeptidase or other retrieved sequences. Also, multiple sequence alignment showed a slight sequence variation in the protein N-terminus. However, the active site residues remain conserved in all *Plasmodium* M1 alanyl aminopeptidase sequences. These comparative analyses allowed the identification of key differences between the human sequence and *Plasmodium* alanyl aminopeptidase sequences, which were used later in virtual screening. Phylogenetic analysis showed the evolutionary relationship between all retrieved *Plasmodium* M1 alanyl aminopeptidase sequences - all the sequences were clustered together while mammalian sequences clustered together but far from the *Plasmodium* M1 alanyl aminopeptidase cluster. All these findings prove the possibility of selective inhibition of *Plasmodium* M1 alanyl aminopeptidase.

3D structures of *P. falciparum* M1 alanyl aminopeptidase and human homologues proteins were retrieved from the Protein Data Bank, while those of the M1 alanyl aminopeptidase for the remaining *Plasmodium* species were not available. To overcome this problem, homology modelling was used to generate the missing structures. The quality of generated models was evaluated through different model validation tools. The resulting models have good local and global quality.

After getting 3D structures for all M1 alanyl aminopeptidases for the *Plasmodium* and human proteins, virtual screening was used to identify possible compounds with selective binding activity against the M1 alanyl aminopeptidase from different *Plasmodium* species. Blind docking and targeted docking were used to identify compounds with high binding affinity to the *Plasmodium* alanyl aminopeptidase protein. Human homolog proteins showed low binding affinity against the best-selected compounds. The best ligand selection criteria started with selecting any ligand that binds to the active site residues. Then any ligand with unfavorable interactions were eliminated. The best ligands were selected based on hydrogen bonding between the ligand and the target protein. Other bonds between the ligand and the targeted protein were also taken into consideration. Ligand efficiency, as well as durability properties based on Lipinski's rule of five were also used.

SANC00531 was selected against the *P. falciparum* M1 alanyl aminopeptidase, SANC00469 against the *P. knowlesi* M1 alanyl aminopeptidase, SANC00660 against the *P. vivax* M1 alanyl aminopeptidase, SANC00144 against the *P. ovale* M1 alanyl aminopeptidase and SANC00109 against the *P. malariae* M1 alanyl aminopeptidase. In future analysis of these compounds and their similar compounds from the ZINC and BioChem databases will be done to improve protein inhibition. as Additionally, molecular dynamic simulations of selected ligands will be performed to investigate the protein-ligand complexes and their stability.

Reference list

[1] J. Recht, A. M. Siqueira, W. M. Monteiro, S. M. Herrera, S. Herrera, and M. V. G. Lacerda, “Malaria in Brazil, Colombia, Peru and Venezuela: current challenges in malaria control and elimination.,” *Malar. J.*, vol. 16, no. 1, p. 273, Jul. 2017.

[2] R. W. Snow, C. A. Guerra, A. M. Noor, H. Y. Myint, and S. I. Hay, “The global distribution of clinical episodes of Plasmodium falciparum malaria,” *Nature*, vol. 434, no. 7030, pp. 214–217, Mar. 2005.

[3] “WHO | Malaria,” *WHO*, 2017.

[4] B. Nadjm and R. H. Behrens, “Malaria: An Update for Physicians,” *Infect. Dis. Clin. North Am.*, vol. 26, no. 2, pp. 243–259, Jun. 2012.

[5] “WHO | Malaria,” *WHO*, 2017.

[6] G. Ruiz Lopez del Prado *et al.*, “Malaria in developing countries,” *J. Infect. Dev. Ctries.*, vol. 8, no. 01, pp. 001–004, Jan. 2014.

[7] V. Risco-Castillo *et al.*, “Malaria Sporozoites Traverse Host Cells within Transient Vacuoles,” *Cell Host Microbe*, vol. 18, no. 5, pp. 593–603, Nov. 2015.

[8] A. F. Cowman *et al.*, “Malaria: Biology and Disease,” *Cell*, vol. 167, no. 3, pp. 610–624, Oct. 2016.

[9] A. M. Vaughan and S. H. I. Kappe, “Malaria Parasite Liver Infection and Exoerythrocytic Biology,” *Cold Spring Harb. Perspect. Med.*, vol. 7, no. 6, p. a025486, Jun. 2017.

[10] J. C. Volz *et al.*, “Essential Role of the PfRh5/PfRipr/CyRPA Complex during Plasmodium falciparum Invasion of Erythrocytes,” *Cell Host Microbe*, vol. 20, no. 1, pp. 60–71, Jul. 2016.

[11] N. J. White, S. Pukrittayakamee, T. T. Hien, M. A. Faiz, O. A. Mokuolu, and A. M. Dondorp, “Malaria,” *Lancet*, vol. 383, no. 9918, pp. 723–735, Feb. 2014.

[12] M. Tibúrcio, R. Sauerwein, C. Lavazec, and P. Alano, “Erythrocyte remodeling by Plasmodium falciparum gametocytes in the human host interplay,” *Trends Parasitol.*, vol. 31, no. 6, pp. 270–278, 2015.

[13] S. Kapishnikov *et al.*, “Oriented nucleation of hemozoin at the digestive vacuole membrane in Plasmodium falciparum,” *Proc. Natl. Acad. Sci.*, vol. 109, no. 28, pp. 11188–11193, Jul. 2012.

[14] K. K. Roy, “Targeting the active sites of malarial proteases for antimalarial drug discovery: approaches, progress and challenges,” *Int. J. Antimicrob. Agents*, Jun. 2017.

[15] G. A. Josling and M. Llinás, “Sexual development in Plasmodium parasites: knowing when it’s time to commit,” *Nat. Rev. Microbiol.*, vol. 13, no. 9, pp. 573–587, Aug. 2015.

[16] A. S. I. Aly, A. M. Vaughan, and S. H. I. Kappe, “Malaria parasite development in the mosquito and infection of the mammalian host.,” *Annu. Rev. Microbiol.*, vol. 63, pp. 195–221, 2009.

[17] M. A. Phillips, J. N. Burrows, C. Manyando, R. H. van Huijsduijnen, W. C. Van Voorhis, and T. N. C. Wells, “Malaria,” *Nat. Rev. Dis. Prim.*, vol. 3, p. 17050, Aug. 2017.

[18] N. D. Rawlings, A. J. Barrett, P. D. Thomas, X. Huang, A. Bateman, and R. D. Finn, “The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database,” *Nucleic Acids Res.*, 2018.

[19] N. D. Rawlings, A. J. Barrett, and R. Finn, “Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors,” *Nucleic Acids Res.*, vol. 44, 2016.

[20] P. M. Jones, M. W. Robinson, J. P. Dalton, A. M. George, and O. Keskin, “The Plasmodium falciparum Malaria M1 Alanyl Aminopeptidase (PfA-M1): Insights of Catalytic Mechanism and Function from MD Simulations,” *PLoS One*, vol. 6, no. 12, p. e28589, Dec. 2011.

[21] R. D. Finn *et al.*, “InterPro in 2017-beyond protein family and domain annotations.,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D190–D199, Jan. 2017.

[22] D. L. Gardiner, K. R. Trenholme, T. S. Skinner-Adams, C. M. Stack, and J. P. Dalton, “Overexpression of leucyl aminopeptidase in Plasmodium falciparum parasites. Target for the antimalarial activity of bestatin.,” *J. Biol. Chem.*, vol. 281, no. 3, pp. 1741–5, Jan. 2006.

[23] A. Mucha, M. Drag, J. P. Dalton, and P. Kafarski, “Metallo-aminopeptidase inhibitors,” *Biochimie*, vol. 92, no. 11, pp. 1509–1529, Nov. 2010.

[24] P. M. Jones, M. W. Robinson, J. P. Dalton, and A. M. George, “The Plasmodium falciparum malaria M1 alanyl aminopeptidase (PfA-M1): insights of catalytic mechanism and function from MD simulations.,” *PLoS One*, vol. 6, no. 12, p. e28589, 2011.

[25] N. Drinkwater *et al.*, “Potent dual inhibitors of Plasmodium falciparum M1 and M17 aminopeptidases through optimization of S1 pocket interactions,” *Eur. J. Med. Chem.*, vol. 110, pp. 43–64, Mar. 2016.

[26] Y. Watanabe, S. Iwaki-Egawa, H. Mizukoshi, and Y. Fujimoto, “Identification of an

Alanine Aminopeptidase in Human Maternal Serum as a Membrane-Bound Aminopeptidase N," *Biol. Chem. Hoppe. Seyler.*, 1995.

[27] R. Kundu, "Diagnosis and management of malaria in children: Recommendations and IAP plan of action," *Pediatr. Infect. Dis.*, vol. 6, no. 1, pp. 7–14, Jan. 2014.

[28] M. L. Wilson, "Malaria Rapid Diagnostic Tests," *Clin. Infect. Dis.*, vol. 54, no. 11, pp. 1637–1641, Jun. 2012.

[29] P. Ranjan and U. Ghoshal, "Utility of nested polymerase chain reaction over the microscopy and immuno-chromatographic test in the detection of Plasmodium species and their clinical spectrum," *Parasitol. Res.*, vol. 115, no. 9, pp. 3375–3385, Sep. 2016.

[30] J. K. Baird, N. Valecha, S. Duparc, N. J. White, and R. N. Price, "Diagnosis and Treatment of Plasmodium vivax Malaria.," *Am. J. Trop. Med. Hyg.*, vol. 95, no. 6 Suppl, pp. 35–51, Dec. 2016.

[31] H. A. Antony and S. C. Parija, "Antimalarial drug resistance: An overview.," *Trop. Parasitol.*, vol. 6, no. 1, pp. 30–41, 2016.

[32] S. R. Meshnick, "Artemisinin: mechanisms of action, resistance and toxicity.," *Int. J. Parasitol.*, vol. 32, no. 13, pp. 1655–60, Dec. 2002.

[33] B. Lell and P. G. Kremsner, "Clindamycin as an antimalarial drug: review of clinical trials.," *Antimicrob. Agents Chemother.*, vol. 46, no. 8, pp. 2315–20, Aug. 2002.

[34] S. Sinha, B. Medhi, and R. Sehgal, "Challenges of drug-resistant malaria.," *Parasite*, vol. 21, p. 61, 2014.

[35] F. Nosten, "[Elimination in South-East Asia? The role of antimalarial drugs].," *Bull. Acad. Natl. Med.*, vol. 200, no. 3, pp. 467-75; discussion 475–6, 2016.

[36] J. M. Sa *et al.*, "Geographic patterns of Plasmodium falciparum drug resistance distinguished by differential responses to amodiaquine and chloroquine," *Proc. Natl. Acad. Sci.*, vol. 106, no. 45, pp. 18883–18889, Nov. 2009.

[37] A. Ouattara and M. B. Laurens, "Vaccines against malaria.," *Clin. Infect. Dis.*, vol. 60, no. 6, pp. 930–6, Mar. 2015.

[38] M. A. Thera *et al.*, "A Field Trial to Assess a Blood-Stage Malaria Vaccine," *N. Engl. J. Med.*, vol. 365, no. 11, pp. 1004–1013, Sep. 2011.

[39] W. Ma, W. S. Noble, and T. L. Bailey, "Motif-based analysis of large nucleotide data sets using MEME-ChIP," *Nat. Protoc.*, vol. 9, no. 6, pp. 1428–1450, 2014.

[40] P. D'Haeseleer, "What are DNA sequence motifs?," *Nat Biotech*, vol. 24, no. 4, pp. 423–425, 2006.

[41] T. L. Bailey *et al.*, "MEME Suite: Tools for motif discovery and searching," *Nucleic*

Acids Res., vol. 37, no. SUPPL. 2, 2009.

[42] M. Thomas-Chollier *et al.*, "RSAT: regulatory sequence analysis tools.," *Nucleic Acids Res.*, vol. 36, no. Web Server issue, 2008.

[43] R. D. Finn *et al.*, "The Pfam protein families database: Towards a more sustainable future," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D279–D285, 2016.

[44] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li, "MEME: Discovering and analyzing DNA and protein sequence motifs," *Nucleic Acids Res.*, vol. 34, no. WEB. SERV. ISS., 2006.

[45] M. C. Frith, N. F. W. Saunders, B. Kobe, and T. L. Bailey, "Discovering sequence motifs with arbitrary insertions and deletions," *PLoS Comput. Biol.*, vol. 4, no. 5, 2008.

[46] T. L. Bailey, J. Johnson, C. E. Grant, and W. S. Noble, "The MEME Suite," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W39–W49, 2015.

[47] T. L. Bailey, "DREME: Motif discovery in transcription factor ChIP-seq data," *Bioinformatics*, vol. 27, no. 12, pp. 1653–1659, 2011.

[48] E. L. L. Sonnhammer, S. R. Eddy, and R. Durbin, "Pfam: A comprehensive database of protein domain families based on seed alignments," *Proteins Struct. Funct. Genet.*, vol. 28, no. 3, pp. 405–420, 1997.

[49] R. D. Finn *et al.*, "Pfam: The protein families database," *Nucleic Acids Research*, vol. 42, no. D1. 2014.

[50] R. D. Finn, "Pfam: clans, web tools and services," *Nucleic Acids Res.*, vol. 34, no. 90001, pp. D247–D251, 2006.

[51] R. C. Edgar and S. Batzoglou, "Multiple sequence alignment," *Current Opinion in Structural Biology*, vol. 16, no. 3. pp. 368–373, 2006.

[52] V. O. Polyanovsky, M. A. Roytberg, and V. G. Tumanyan, "Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences," *Algorithms Mol. Biol.*, 2011.

[53] L. Mullan, "Pairwise sequence alignment - It's all about us!," *Brief. Bioinform.*, vol. 7, no. 1, pp. 113–115, 2006.

[54] S. Batzoglou, "The many faces of sequence alignment," *Briefings in Bioinformatics*, vol. 6, no. 1. pp. 6–22, 2005.

[55] M. Dayhoff and R. Schwartz, "A Model of Evolutionary Change in Proteins," *Atlas protein Seq. Struct.*, pp. 345–352, 1978.

[56] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks.," *Proc. Natl. Acad. Sci.*, vol. 89, no. 22, pp. 10915–10919, 1992.

[57] D. W. Mount, “Comparison of the PAM and BLOSUM amino acid substitution matrices,” *Cold Spring Harb. Protoc.*, vol. 3, no. 6, 2008.

[58] D. W. Mount, “Using progressive methods for global multiple sequence alignment,” *Cold Spring Harb. Protoc.*, vol. 4, no. 7, 2009.

[59] K. Katoh and D. M. Standley, “MAFFT: Iterative refinement and additional methods,” *Methods Mol. Biol.*, vol. 1079, pp. 131–146, 2014.

[60] J. D. Thompson, F. Plewniak, and O. Poch, “A comprehensive comparison of multiple sequence alignment programs.,” *Nucleic Acids Res.*, 1999.

[61] J. Waikagul, U. Thaenkham, J. Waikagul, and U. Thaenkham, “Methods of Molecular Study,” in *Approaches to Research on the Systematics of Fish-Borne Trematodes*, Elsevier, 2014, pp. 77–90.

[62] S. Choudhuri, *Bioinformatics for beginners : genes, genomes, molecular evolution, databases and analytical tools*. .

[63] Y. Hao, Z. Pei, and S. M. Brown, “Bioinformatics in Microbiome Analysis,” *Methods Microbiol.*, vol. 44, pp. 1–18, Jan. 2017.

[64] Y. Hao, Z. Pei, and S. M. Brown, “Bioinformatics in Microbiome Analysis,” *Methods Microbiol.*, vol. 44, pp. 1–18, Jan. 2017.

[65] B. Efron, E. Halloran, and S. Holmes, “Bootstrap confidence levels for phylogenetic trees.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 93, no. 23, pp. 13429–34, Nov. 1996.

[66] B. Efron, E. Halloran, and S. Holmes, “Bootstrap confidence levels for phylogenetic trees.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 93, no. 23, pp. 13429–34, Nov. 1996.

[67] D. W. Mount, “Maximum parsimony method for phylogenetic prediction.,” *CSH Protoc.*, vol. 2008, no. 4, p. pdb.top32, Apr. 2008.

[68] A. Stamatakis, “RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models,” *Bioinformatics*, vol. 22, no. 21, pp. 2688–2690, Nov. 2006.

[69] W. Gish and D. J. States, “Identification of protein coding regions by database similarity search,” *Nat Genet*, 1993.

[70] D. R. Zerbino *et al.*, “Ensembl 2018,” *Nucleic Acids Res.*, 2018.

[71] T. UniProt Consortium, “UniProt: the universal protein knowledgebase,” *Nucleic Acids Res.*, 2018.

[72] B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, and C. H. Wu, “UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches,” *Bioinformatics*, 2015.

[73] N. G. Traylor-Knowles *et al.*, “HMMER web server: interactive sequence similarity searching,” *Biochem Biophys Res Commun*, 2012.

[74] T. L. Bailey and M. Gribskov, “Combining evidence using p-values: Application to sequence homology searches,” *Bioinformatics*, 1998.

[75] T. Nakamura, K. D. Yamada, K. Tomii, and K. Katoh, “Parallelization of MAFFT for large-scale multiple sequence alignments,” *Bioinformatics*, 2018.

[76] S. Kumar, G. Stecher, and K. Tamura, “MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets,” *Mol. Biol. Evol.*, 2016.

[77] S. K. Burley, H. M. Berman, G. J. Kleywegt, J. L. Markley, H. Nakamura, and S. Velankar, “Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive,” Humana Press, New York, NY, 2017, pp. 627–641.

[78] A. Bateman *et al.*, “UniProt: The universal protein knowledgebase,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D158–D169, 2017.

[79] K. Ginalski, “Comparative modeling for protein structure prediction,” *Current Opinion in Structural Biology*, vol. 16, no. 2. pp. 172–177, 2006.

[80] A. Özlem Tastan Bishop, T. A. P. de Beer, and F. Joubert, “Protein homology modelling and its use in South Africa,” *S. Afr. J. Sci.*, vol. 104, no. February, pp. 2–6, 2008.

[81] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, “Basic local alignment search tool,” *J. Mol. Biol.*, vol. 215, no. 3, pp. 403–410, 1990.

[82] G. M. Boratyn, A. A. Schäffer, R. Agarwala, S. F. Altschul, D. J. Lipman, and T. L. Madden, “Domain enhanced lookup time accelerated BLAST,” *Biol. Direct*, vol. 7, no. 1, p. 12, 2012.

[83] A. Fiser, “Comparative Protein Structure Modelling,” in *From Protein Structure to Function with Bioinformatics*, Dordrecht: Springer Netherlands, 2017, pp. 91–134.

[84] S. F. Altschul *et al.*, “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.,” *Nucleic Acids Res.*, vol. 25, no. 17, pp. 3389–402, 1997.

[85] J. Söding, A. Biegert, and A. N. Lupas, “The HHpred interactive server for protein homology detection and structure prediction,” *Nucleic Acids Res.*, vol. 33, no. SUPPL. 2, 2005.

[86] V. K. Vyas, R. D. Ukawala, M. Ghate, and C. Chintha, “Homology Modeling a Fast Tool for Drug Discovery: Current Perspectives,” *Indian J. Pharm. Sci.*, vol. 1, pp. 1–17, 2012.

[87] A. Szilagyi and Y. Zhang, “Template-based structure modeling of protein-protein interactions,” *Current Opinion in Structural Biology*, vol. 24, no. 1. pp. 10–23, 2014.

[88] A. Fiser, “Template-based protein structure modeling.,” *Methods Mol. Biol.*, vol. 673, pp. 73–94, 2010.

[89] R. C. Edgar, “MUSCLE: Multiple sequence alignment with high accuracy and high throughput,” *Nucleic Acids Res.*, vol. 32, no. 5, pp. 1792–1797, 2004.

[90] F. Sievers *et al.*, “Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega,” *Mol. Syst. Biol.*, vol. 7, no. 1, pp. 539–539, 2014.

[91] S. W. Robinson, A. M. Afzal, D. P. Leader, and D. P. Leader, “Bioinformatics: Concepts, Methods, and Data,” in *Handbook of Pharmacogenomics and Stratified Medicine*, Elsevier, 2014, pp. 259–287.

[92] J.-F. Taly *et al.*, “Using the T-Coffee package to build multiple sequence alignments of protein, RNA, DNA sequences and 3D structures,” *Nat. Protoc.*, vol. 6, no. 11, pp. 1669–1682, 2011.

[93] J. Pei, B.-H. Kim, and N. V Grishin, “PROMALS3D: a tool for multiple protein sequence and structure alignments.,” *Nucleic Acids Res.*, vol. 36, no. 7, pp. 2295–300, Apr. 2008.

[94] A. Fiser and A. Šali, “MODELLER: Generation and Refinement of Homology-Based Protein Structure Models,” *Methods in Enzymology*, vol. 374. pp. 461–491, 2003.

[95] A. Hillisch, L. F. Pineda, and R. Hilgenfeld, “Utility of homology models in the drug discovery process,” *Drug Discovery Today*, vol. 9, no. 15. pp. 659–669, 2004.

[96] A. C.R.Martin, K. Toda, H. J. Stirk, and J. M. Thornton, “Long loops in proteins,” *Protein Eng. Des. Sel.*, vol. 8, no. 11, pp. 1093–1101, 1995.

[97] N. Koga *et al.*, “Principles for designing ideal protein structures,” *Nature*, vol. 491, no. 7423, pp. 222–227, 2012.

[98] M. Jamroz and A. Kolinski, “Modeling of loops in proteins: a multi-method approach.,” *BMC Struct. Biol.*, vol. 10, p. 5, 2010.

[99] S. Hongmao and S. Hongmao, “Homology Modeling and Ligand-Based Molecule Design,” in *A Practical Guide to Rational Drug Design*, Elsevier, 2016, pp. 109–160.

[100] E. Krieger, S. B. Nabuurs, and G. Vriend, “Homology modeling.,” *Methods Biochem. Anal.*, vol. 44, pp. 509–23, 2003.

[101] N. Soni and M. S. Madhusudhan, “Computational modeling of protein assemblies,” *Current Opinion in Structural Biology*, vol. 44. pp. 179–189, 2017.

[102] B. Rost, “Twilight zone of protein sequence alignments,” *Protein Eng.*, vol. 12, no. 2, pp. 85–94, 1999.

[103] J. F. Taly *et al.*, “Using the T-Coffee package to build multiple sequence alignments of

protein, RNA, DNA sequences and 3D structures," *Nat. Protoc.*, vol. 6, no. 11, pp. 1669–1682, 2011.

[104] L. R. Forrest, C. L. Tang, and B. Honig, "On the Accuracy of Homology Modeling and Sequence Alignment Methods Applied to Membrane Proteins," *Biophys. J.*, vol. 91, no. 2, pp. 508–517, 2006.

[105] J. A. Cuff and G. J. Barton, "Application of multiple sequence alignment profiles to improve protein secondary structure prediction," *Proteins Struct. Funct. Genet.*, vol. 40, no. 3, pp. 502–511, 2000.

[106] M. S. Madhusudhan, M. A. Marti-Renom, R. Sanchez, and A. Sali, "Variable gap penalty for protein sequence-structure alignment," *Protein Eng. Des. Sel.*, vol. 19, no. 3, pp. 129–133, 2006.

[107] D. Eramian, N. Eswar, M.-Y. Shen, and A. Sali, "How well can the accuracy of comparative protein structure models be predicted?," *Protein Sci.*, vol. 17, no. 11, pp. 1881–93, 2008.

[108] A. R. J. Young, M. Narita, and M. Narita, *Homology Modeling*, vol. 857, no. 1. 2012.

[109] P. Alejster, W. Jurkowski, and I. Roterman-Konieczna, "Structural information involved in the interpretation of the stepwise protein folding process," in *Protein Folding in Silico*, Elsevier, 2012, pp. 39–54.

[110] H. Park, H. Lee, and C. Seok, "High-resolution protein-protein docking by global optimization: Recent advances and future challenges," *Current Opinion in Structural Biology*, vol. 35. pp. 24–31, 2015.

[111] M. A. Khamis, W. Gomaa, and W. F. Ahmed, "Machine learning in computational docking," *Artificial Intelligence in Medicine*, vol. 63, no. 3. pp. 135–152, 2015.

[112] S. Forli, R. Huey, M. E. Pique, M. F. Sanner, D. S. Goodsell, and A. J. Olson, "Computational protein-ligand docking and virtual drug screening with the AutoDock suite," *Nat. Protoc.*, vol. 11, no. 5, pp. 905–919, 2016.

[113] P. D. Lyne, "Structure-based virtual screening: An overview," *Drug Discovery Today*, vol. 7, no. 20. pp. 1047–1055, 2002.

[114] H. Geppert, M. Vogt, and J. Bajorath, "Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation," *J. Chem. Inf. Model.*, vol. 50, no. 2, pp. 205–216, 2010.

[115] E. Lionta, G. Spyrou, D. K. Vassilatis, and Z. Cournia, "Structure-based virtual screening for drug discovery: principles, applications and recent advances.," *Curr. Top. Med. Chem.*, vol. 14, no. 16, pp. 1923–1938, 2014.

[116] A. N. Jain and A. Nicholls, “Recommendations for evaluation of computational methods,” *J. Comput. Aided. Mol. Des.*, vol. 22, no. 3–4, pp. 133–139, 2008.

[117] Y. Yan, W. Wang, Z. Sun, J. Z. H. Zhang, and C. Ji, “Protein-Ligand Empirical Interaction Components for Virtual Screening,” *J. Chem. Inf. Model.*, vol. 57, no. 8, pp. 1793–1806, 2017.

[118] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, and R. G. Coleman, “ZINC: A free tool to discover chemistry for biology,” *Journal of Chemical Information and Modeling*, vol. 52, no. 7. pp. 1757–1768, 2012.

[119] PubChem, “PubChem Compound,” *National Center for Biotechnology Information, U.S. National Library of Medicine*, 2016. [Online]. Available: https://www.ncbi.nlm.nih.gov/pccompound.

[120] D. S. Wishart, “DrugBank: a comprehensive resource for in silico drug discovery and exploration,” *Nucleic Acids Res.*, vol. 34, no. 90001, pp. D668–D672, 2006.

[121] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, “BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities,” *Nucleic Acids Res.*, vol. 35, no. SUPPL. 1, 2007.

[122] M. K. Gilson, T. Liu, M. Baitaluk, G. Nicola, L. Hwang, and J. Chong, “BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology,” *Nucleic Acids Res.*, vol. 44, no. D1, pp. D1045–D1053, 2016.

[123] R. Hatherley *et al.*, “SANCDB: a South African natural compound database,” *J. Cheminform.*, vol. 7, no. 1, p. 29, Dec. 2015.

[124] A. P. Bento *et al.*, “The ChEMBL bioactivity database: An update,” *Nucleic Acids Res.*, vol. 42, no. D1, 2014.

[125] A. Gaulton *et al.*, “ChEMBL: A large-scale bioactivity database for drug discovery,” *Nucleic Acids Res.*, vol. 40, no. D1, 2012.

[126] K. P. Seiler *et al.*, “ChemBank: A small-molecule screening and cheminformatics resource database,” *Nucleic Acids Res.*, vol. 36, no. SUPPL. 1, 2008.

[127] L. Ferreira, R. dos Santos, G. Oliva, and A. Andricopulo, *Molecular Docking and Structure-Based Drug Design Strategies*, vol. 20, no. 7. 2015.

[128] N. S. Pagadala, K. Syed, and J. Tuszynski, “Software for molecular docking: a review,” *Biophysical Reviews*, vol. 9, no. 2. pp. 91–102, 2017.

[129] J. I. Garzon *et al.*, “FRODOCK: A new approach for fast rotational protein-protein docking,” *Bioinformatics*, vol. 25, no. 19, pp. 2544–2551, 2009.

[130] R. Chen, L. Li, and Z. Weng, “ZDOCK: An initial-stage protein-docking algorithm,”

Proteins Struct. Funct. Genet., vol. 52, no. 1, pp. 80–87, 2003.

[131] M. Ohue, T. Shimoda, S. Suzuki, Y. Matsuzaki, T. Ishida, and Y. Akiyama, "MEGADOCK 4.0: An ultra-high-performance protein-protein docking software for heterogeneous supercomputers," *Bioinformatics*, vol. 30, no. 22, pp. 3281–3283, 2014.

[132] G. Morris and R. Huey, "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility," *J. …*, vol. 30, no. 16, pp. 2785–2791, 2009.

[133] O. Trott and A. J. Olson, "AutoDock Vina," *J. Comput. Chem.*, vol. 31, pp. 445–461, 2010.

[134] R. Meier, M. Pippel, F. Brandt, W. Sippl, and C. Baldauf, "ParaDockS: A framework for molecular docking with population-based metaheuristics," *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 879–889, 2010.

[135] G. Jones, P. Willett, R. C. Glen, A. R. Leach, and R. Taylor, "Development and validation of a genetic algorithm for flexible docking," *J. Mol. Biol.*, vol. 267, no. 3, pp. 727–748, 1997.

[136] V. S. Rao, K. Srinivas, G. N. Sujini, and G. N. S. Kumar, "Protein-Protein Interaction Detection: Methods and Analysis," *Int. J. Proteomics*, vol. 2014, pp. 1–12, 2014.

[137] I. Halperin, B. Ma, H. Wolfson, and R. Nussinov, "Principles of docking: An overview of search algorithms and a guide to scoring functions," *Proteins: Structure, Function and Genetics*, vol. 47, no. 4. pp. 409–443, 2002.

[138] N. Andrusier, E. Mashiach, R. Nussinov, and H. J. Wolfson, "Principles of flexible protein-protein docking," *Proteins: Structure, Function and Genetics*, vol. 73, no. 2. pp. 271–289, 2008.

[139] O. Trott and A. Olson, "AutoDock Vina: inproving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading," *J. Comput. Chem.*, vol. 31, no. 2, pp. 455–461, 2010.

[140] W. P. Feinstein and M. Brylinski, "Calculating an optimal box size for ligand docking and virtual screening against experimental and predicted binding pockets," *J. Cheminform.*, vol. 7, no. 1, 2015.

[141] C. Hetényi and D. Van Der Spoel, "Blind docking of drug-sized compounds to proteins with up to a thousand residues," *FEBS Lett.*, vol. 580, no. 5, pp. 1447–1450, 2006.

[142] D. Seeliger and B. L. De Groot, "Ligand docking and binding site analysis with PyMOL and Autodock/Vina," *J. Comput. Aided. Mol. Des.*, vol. 24, no. 5, pp. 417–422, 2010.

[143] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and

development settings," *Advanced Drug Delivery Reviews*, vol. 64, no. SUPPL. pp. 4–17, 2012.

[144] D. Chen, N. Oezguen, P. Urvil, C. Ferguson, S. M. Dann, and T. C. Savidge, "Regulation of protein-ligand binding affinity by hydrogen bond pairing," *Sci. Adv.*, vol. 2, no. 3, p. e1501240, 2016.

Appendices

Appendix 1

```
import os

Ligand_files = os.listdir('../Ligand')

PDB_files = os.listdir('../Target')

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("/mnt/lustre/users/osamir/Docking_v2/Vina/" +
vina_name, "w") as vw:

          vw.writelines(["receptor=/mnt/lustre/users/osamir/Docking_v2/Target/"+PDB+"qt"
,"\n"])

          vw.writelines(["ligand=/mnt/lustre/users/osamir/Docking_v2/Ligand/"+ligand_name+".pd
bqt", "\n"])

          vw.writelines(["out=/mnt/lustre/users/osamir/Docking_v2/Out/"+vina_name+"all.pdbqt",
"\n"])

          vw.writelines(["log=/mnt/lustre/users/osamir/Docking_v2/Log/"+vina_name+"all.log",
"\n"])
                                        vw.writelines(["center_x=20.002", "\n"])
                                        vw.writelines(["center_y=15.945", "\n"])
                                        vw.writelines(["center_z=3.313", "\n"])
                                        vw.writelines(["size_x=58.88", "\n", "size_y=53.62"])
                                        vw.writelines(["\n", "size_z=57.38", "\n"])
                                        vw.writelines(["cpu=8", "\n", "exhaustiveness=192"])
```

Appendix 2

```
import os
vina_files = os.listdir('../Vina')
gnu_w = open("gnu_parallel.jobs", "w")
for vina in vina_files:
          gnu w.writelines(["/home/osamir/lustre/Docking/Script/vina --config " +
"/home/osamir/lustre/Docking/Vina/"+vina+"\n"+"\n"])

gnu_w.close()
```

Appendix 3

```
Out = os.listdir("./Out")

for pdbqt in Out:
    os.system("vina_split --input Out/" + pdbqt)
```

```
    ligand_1 = ligand_1 + "Out/" + pdbqt[:-6] + "_ligand_1.pdbqt "

os.system("babel -ipdbqt " + ligand_1 + " -osdf all.sdf")

with open("all.sdf", "r") as sdf:
    lines = sdf.readlines()
temp = 0
for line in range(len(lines)):
    if "VINA RESULT" in lines[line]:
        temp = line
        lines[temp + 4] = "\n" + ">  <Score> \n" + lines[line].split()[2] +
"\n"
        lines[temp + 5] = "\n" + lines[temp + 5]
with open("news.sdf", "w") as ss:
    ss.writelines(lines)

log = os.listdir("./Log")

names = {}

for files in log:
    tmp = files[5:]
    tmp = tmp[:-10]
    with open("Log/"+files, "r") as tmpr:
        lines = tmpr.readlines()
        for line in lines:
            if line.startswith("   1"):
                names[tmp] = line.split()[1]

with open("Output.csv", "w") as tmpw:
    for i in names:
        tmp = i,",", names[i], "\n"

        tmpw.writelines(tmp)
```