Using Bioinformatics tools to Screen for Trypanosomal Cathepsin B Cysteine Protease Inhibitors from the SANCDB as a Novel Therapeutic Modality against Human African Trypanosomiasis (HAT)

A research thesis submitted in fulfilment of the requirements for the degree of

MASTER OF SCIENCE

IN BIOCHEMISTRY

of

RHODES UNIVERSITY, SOUTH AFRICA

Department of Biochemistry and Microbiology

Department of Chemistry

FACULTY OF SCIENCE

By

GAONE MOKHAWA

December 2015

Abstract

Human African Trypanosomiasis (HAT), also known as sleeping sickness, is a fatal chronic disease that is caused by flagellated protozoans, *Trypanosoma brucei gambiense* and *Trypanosoma brucei rhodesiense*. HAT is spread by a bite from an infected tsetse fly of the Glosina genus. Up to 60 million people in 36 countries in sub-Saharan Africa are at a risk of infection from HAT with up to 30 000 deaths reported every year. Current chemotherapy for HAT is insufficient since the available drugs exhibit unacceptable side effects (toxicity) and parasite resistance. Novel treatments and approaches for development of specific and more potent drugs for HAT are therefore required.

One approach is to target vital proteins that are essential to the life cycle of the parasite. The main interest of this study is to explore *Trypanosoma brucei* cathepsin B-like protease (TbCatB) structural and functional properties with the primary goal of discovering non peptide small molecule inhibitors of TbCatB using bioinformatics approaches. TbCatB is a papain family C1 cysteine protease which belongs to clan CA group and it has emerged as a potential HAT drug target. Papain family cysteine proteases of Clan CA group of *Trypanosoma brucei* (rhodesain and TbCatB) have demonstrated potential as chemotherapeutic targets using synthetic protease inhibitors like Z-Phe-Ala-CHN2 to kill the parasite *in vitro* and *in vivo*. TbCatB has been identified as the essential cysteine protease of *T. brucei* since mRNA silencing of TbCatB killed the parasite and resulted in a cure in mice infected with *T. brucei* while mRNA silencing of rhodesain only extended mice life. TbCatB is therefore a promising drug target against HAT and the discovery and development of compounds that can selectively inhibit TbCatB without posing any danger to the human host represent a great therapeutic solution for treatment of HAT.

To understand protein-inhibitor interactions, useful information can be obtained from high resolution protease-inhibitor crystal structure complexes. This study aims to use bioinformatics approaches to carry out comparative sequence, structural and functional analysis of TbCatB protease and its homologs from *T. congolense, T, cruzi, T. vivax* and *H. sapien* as well as to identify non-peptide small molecule inhibitors of TbCatB cysteine proteases from natural compounds of South African origin. Sequences of TbCatB (PDB ID: 3HHI) homologs were retrieved by a BLAST search. Human cathepsin B (PDB ID: 3CBJ) was selected from a list of templates for homology modelling found by HHpred. MODELLER version 9.10 program was used to generate a hundred models for *T. congolense, T, cruzi and T. vivax* cathepsin B like proteases using 3HHI and 3CBJ as templates. The best models were chosen based on their low DOPE Z scores before validation using MetaMQAPII, ANOLEA, PROCHECK and QMEAN6. The DOPE Z scores and the RMSD (RMS) values of the calculated models indicate that the

models are of acceptable energy (stability) and fold (conformation). Results from the different MQAPs indicate the models are of acceptable quality and they can be used for docking studies. High throughput screening of SANCDB using AutoDock Vina revealed nine compounds, SANC00 478, 479, 480, 481, 482, 488, 489, 490 and 491, having a strong affinity for *Trypanosoma spp.* cathepsin B proteases than HsCatB. SANC00488 has the strongest binding to *Trypanosoma spp.* cathepsin B proteases and the weakest binding to HsCatB protease. Molecular dynamics (MD) simulations show that the complexes between SANC00488 and TbCatB, TcCatB, TcrCatB and TvCatB are stable and do not come apart during simulation. The complex between this compound and HsCatB however is unstable and comes apart during simulation. Residues that are important for the stability of SANC00488-TbCatB complex are Gly328 of the S2 subsite, Phe208, and Ala256. In conclusion SANC00488 is a good candicate for development of a drug against HAT.

Declaration

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

Signature: _Date: _

Dedication

To my family

Thank you for understanding

Thank you for your support

And

Thank you for your patience!

Acknowledgement

I would like to express my gratitude to my supervisors, Dr Kevin Lobb and Professor Ö. Tastan Bishop who patiently and generously made it possible for me to complete this project.

I thank Rhodes University for granting me the opportunity to study in its friendly academic environment.

A also thank the National Food Technology Research Centre (NFTRC) in Botswana for granting me the time to study and most for the financial support during the second year of study.

I also thank the Research Unit in Bioinformatics (RUBi) for being helpful and supportive.

And special thanks to my mother Ms Seile Mokhawa for believing and for her support.

Table of Contents

Abbreviations and acronyms

WHO	World Health Organisation
PDB	Protein Data Base
NCBI	National Centre for Biotechnology Information
3D	3-Dimentional
HAT	Human African Trypanosomiasis
Z-Phe-Ala-CH2	Benzyloxycarbonyl-Phenylalanine-Alanine- diazomethane
DNA	Deoxy Ribonucleic Acid
RNA	Ribonucleic Acid
MEGA	Molecular Evolutionary Genetic Analysis
MSA	Multiple Sequence Alignment
CLN	Cervical Lymph Node
SANCDB	South African Natural Compounds Database
NIN	Nearest-Neighbour Interchange
TbCatB	*Trypanosoma brucei* cathepsin B
HsCatB	*Homo sapien* cathepsin B
TcCatB	Trypanosoma congolense cathepsin B
TcrCatB	*Trypanosoma cruzi* cathepsin B
TvCatB	*Trypanosoma vivax* cathepsin B
CATT	Card Agglutination Test for Trypanosomiasis
CNS	Central Nervous System
E64	(2S,3S)-trans-epoxysuccinyl-L-leucyl-agma-tine
CA074	N-(l-3-trans-propylcarbamoyloxirane-2-carbonyl)-l-isoleucyl-l-proline
APC	Antigen Presenting Cells

HTS	High Throughput Screening
EST	Expressed Sequence Tag
GSS	Genome Survey Sequences
SNP	Short Genetic Variations
GEO	Gene Expression Omnibus
NMR	Nuclear Magnetic Resonance
SANCDB	South African Natural Compound Database
EMBL	European Molecular Biology Laboratory
DDBJ	DNA Data Bank of Japan
OMIM	Online Mendelian Inheritance of Man
PAM	Point Accepted Mutation

Symbols used

Å	-Angstrom
β	-Beta
α	-Alpha
π	-Pie bond

List of amino acids

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

List of Figures

List of tables

Chapter 1.

1.1. Introduction

Human African Trypanosomiasis (HAT), also known as sleeping sickness, is a fatal chronic disease that is caused by flagellated protozoans, *Trypanosoma brucei gambiense* (*T. b. gambiense*) and *Trypanosoma brucei rhodesiense* (*T. b. rhodesiense*) [1]. HAT is spread by a bite from an infected tsetse fly of the Glossina genus [2], [3]. An infection by *T. b. gambiense* causes a chronic disease that develops over months or even years before it reaches an advanced stage, where it affects the central nervous system while *T. b. rhodesiense* causes a form of the disease that can kill within weeks if not treated [4]. Another form of the disease which is found in Central and South America is American Trypanosomiasis or Chagas' disease which is caused by the protozoa *Trypanosoma cruzi* (*T. cruzi*) [5], [3]. Chagas' disease is the leading cause of heart disease in Latin America [6]. Other African trypanosomes are *T. b. brucei*, *T. congolese* and *T. vivax*, which are responsible for 'Nagana' disease in cattle and for causing huge economic damages every year with up to 46 million cattle threatened with Nagana in active foci [7].

HAT occurs in 36 countries in sub-Saharan Africa. The form of the disease caused by *T. b gambiense* is found in west and eastern Africa while the *T. b rhodesiense* form is found in eastern and southern Africa [4]. The disease is found in rural areas where the tsetse fly can easily transmit the parasite between people and both domestic and wild animals. According to the World Health Organisation (WHO), close to 10 000 new cases of the disease were recorded in 2013. The actual number of infected individuals may be much higher since some people live in remote areas where the disease has not been monitored but it still poses a danger. Up to 60 million people are estimated to be exposed to HAT in active foci but only up to 5 million live in areas where the disease is monitored [2] and it causes up to 30 000 deaths per year [8].

Although, HAT is transmitted by a bite from an infected tsetse fly [2], [9], it can also be transmitted from mother to child during pregnancy as well as through the use of contaminated needles or by exposure to any other sharp objects [4]. The parasite infects the body in two stages; during the first stage (haemolymphatic phase), the parasite multiply in the blood, the lymph and the subcutaneous tissues [4]. Symptoms include headaches, fever and pains in the joints and itching [10]. During the second stage of the disease, the parasite infects the central nervous system by crossing the blood-brain barrier. During this phase of the disease, the infected person experiences changes in behaviour, confusion and poor coordination [10]. The

patient's normal sleep cycle is also disturbed, which gives the disease its name. This phase of the disease is also known as the neurological or meningoencephalic phase. If not treated, the infected person may die [4],[9].

Diagnosis of HAT is carried out in three steps; (i) screening, (ii) diagnostic confirmation, and (iii) staging [4]. The Card Agglutination Test for Trypanosomiasis *(CATT/T. b. gambiense)* is used for screening [4], [10]. Other methods used are the cervical lymph node (CLN), palpain and puncture method [2], [4]. Diagnostic confirmation relies on screening results. Disease staging is carried out to determine the stage of the disease in the patient. In order to obtain direct evidence that the infection is from a trypanosome, blood, lymph node aspirate, or cerebrospinal fluid samples can be examined under a microscope.

A variety of drugs are available for the treatment of HAT and they are administered according to the stage of the disease. Drugs used during treatment of the first stage of the disease are pentamidine and suramin [9], [10]. Both of these drugs are toxic and they have undesirable side effects on patients. The second stage of the disease is treated using melarsopropol and eflornithine [9], [10]. Treatment with melarsopropol has undesirable side effects like encephalopathy (brain dysfunction) [9], which can be fatal. There has also been an increase in resistance to the drug. Treatment with eflornithine is less toxic but difficult to administer and it is only effective against *T. b. Gambiense* [10]. Nifurtimox and benznidazole are used for the treatment of American Trypanosomiasis [10]. These drugs also have poor efficacy and they produce some serious side effects. These difficulties and dangers associated with administering currently available drugs make them inefficient and unattractive to use, therefore hampering the fight against the disease. This situation has motivated the search for new drugs that are more effective and more tolerable to administer without posing any danger to patients [11].

1.2. The biology and lifecycle of *Trypanosoma brucei*

T. brucei is a unicellular parasitic protozoan belonging to the *Trypanosoma* genus in the Trypanosomatidae family and the order *Kinetoplastida* [2]. Kinetoplastids are unicellular flagellated eukaryotic protozoans that can exist as free-living microorganisms or as parasites of invertebrate, vertebrate and plant species [12]. Kinetoplastids that are pathogenic to human beings include *T. b. gambience* and *T. b. rhodesience* (cause HAT), *T. cruzi* (causes Chargas' disease) and *Leishmania* (causes Leishmaniasis) [10]. They are spindle shaped and have one mitochondrion flagellum which branches throughout the cell. The mitochondrion contains a specialized part called the kinetoplast, which contains the mitochondrion DNA (kDNA) of the organism. During the life cycle of the parasite, it goes through a trypomastigote form and an

amastigote form (in the mammalian host), and an epimastigote and a trypomastigote form (in the insect host) [2], [10]. The metacyclic trypomastigotes are injected into the mammalian host by the tsetse fly before it feeds on blood. The trypomastigotes first multiply at the site of the bite before they enter the bloodstream and the lymphatic system where they continue to replicate by binary fission. In the late-stage of the disease, the trypomastigotes enter the central nervous system (CNS) and occupy the cerebrospinal fluid and intercellular spaces. The trypomastigotes exist in two forms in the mammalian host: the long and proliferative bloodstream trypomastigote form which transforms into the short non proliferative intracellular amastigote. The amastigotes replicates within the cells before they transform back in to the infectious trypomastigotes and rupture the cell to release the trypomastigotes back into the bloodstream. In the gut of the tsetse fly the parasite exists as the replicative epimastigote and the infective metacyclic trypomastigote [2], [10],[9]. The trypomastigotes re-enter the tsetse fly when it feeds on blood containing the protozoan from an infected host. Figure 1.1 summarises the lifecycle of *Trypanosoma brucei* parasite as presented by the Centre for Disease Control and Prevention.

1.3. Proteases

Proteases (peptidases) are enzymes that catalyse the hydrolytic cleavage of peptide bonds [13] and they play an important part in vital biological processes in living organisms. Protease activity is dependent on their specificity, ability to access the scissile bond of the substrate, the activation of precursor enzymes and regulation of enzymes by endogenous protease inhibitors [14]. Proteases can be classified as exopeptidases (those that cleave at the end of a polypeptide) and endopeptidases (those that cleave within a polypeptide chain). Proteases that cleave polypeptide at the N-terminus are called aminopeptidases, and those that cleave at the C-terminus are known as carboxypeptidases [15]. Some proteases, for example cathepsin B-like proteases, can perform both endopeptidase and exopeptidase activity [16].

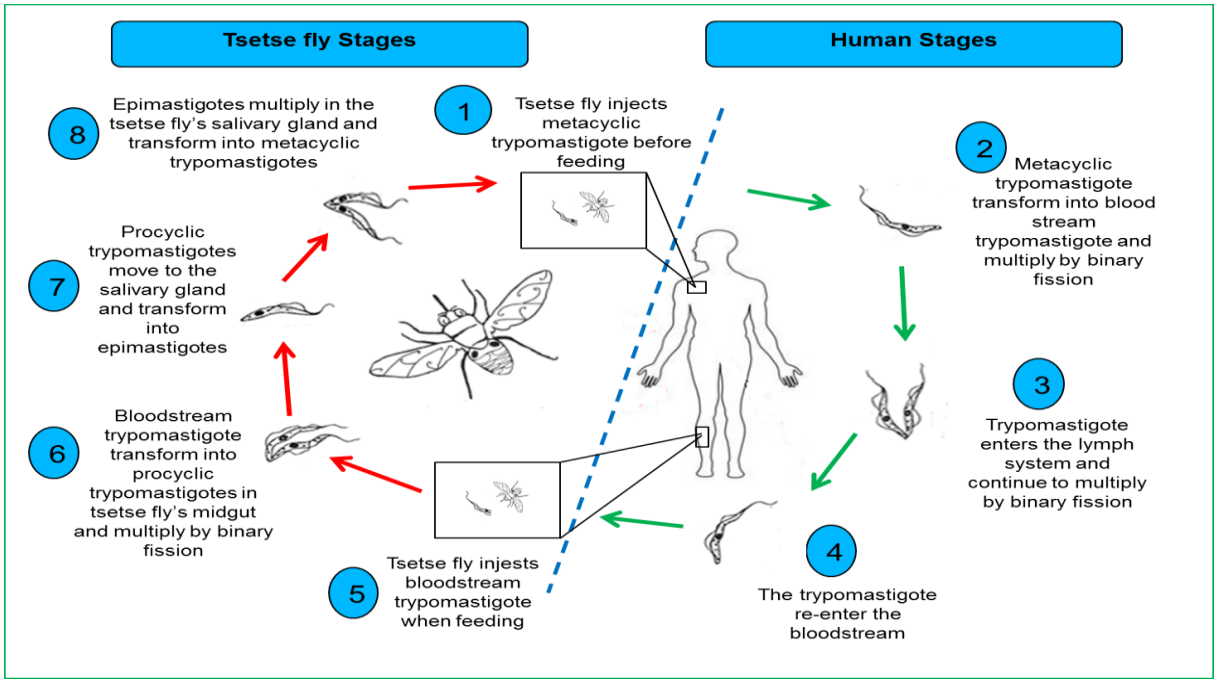


Figure 1.1: Diagrammatic representation of the lifecycle of *T. brucei* in humans (right) and the tsetse fly (left). *T. brucei* cathepsin B (TbCatB) which is essential for parasite survival is expressed in the bloodstream trypomastigote.

(http://www.cdc.gov/parasites/sleepingsickness/biology.html).

Proteases can also be categorized into classes according to the basis of their mechanism of peptide hydrolysis. These include the five major protease classes; cysteine proteases (CPs), serine peptidases (SPs), metalloproteases (MPs), threonine and aspartic proteases [15]. Each class is characterised by a set of amino acid residues arranged in a particular configuration to form the active site [16]. These classes are sub divided in to clans and families according to protein sequence similarity at the active site. Any one of these proteases is expected to exist in multiple species as orthologous forms of one protease. Orthologous forms of a protease from different species are recognised by having the same biochemical specificity to substrates, they have the same optimum pH range and they show the same sensitivity to inhibitors. For this research, focus shall be on papain family C1 cysteine proteases which belong to Clan CA group since they are important during the lifecycle of *T. brucei* and to human health [12], [17],[18].

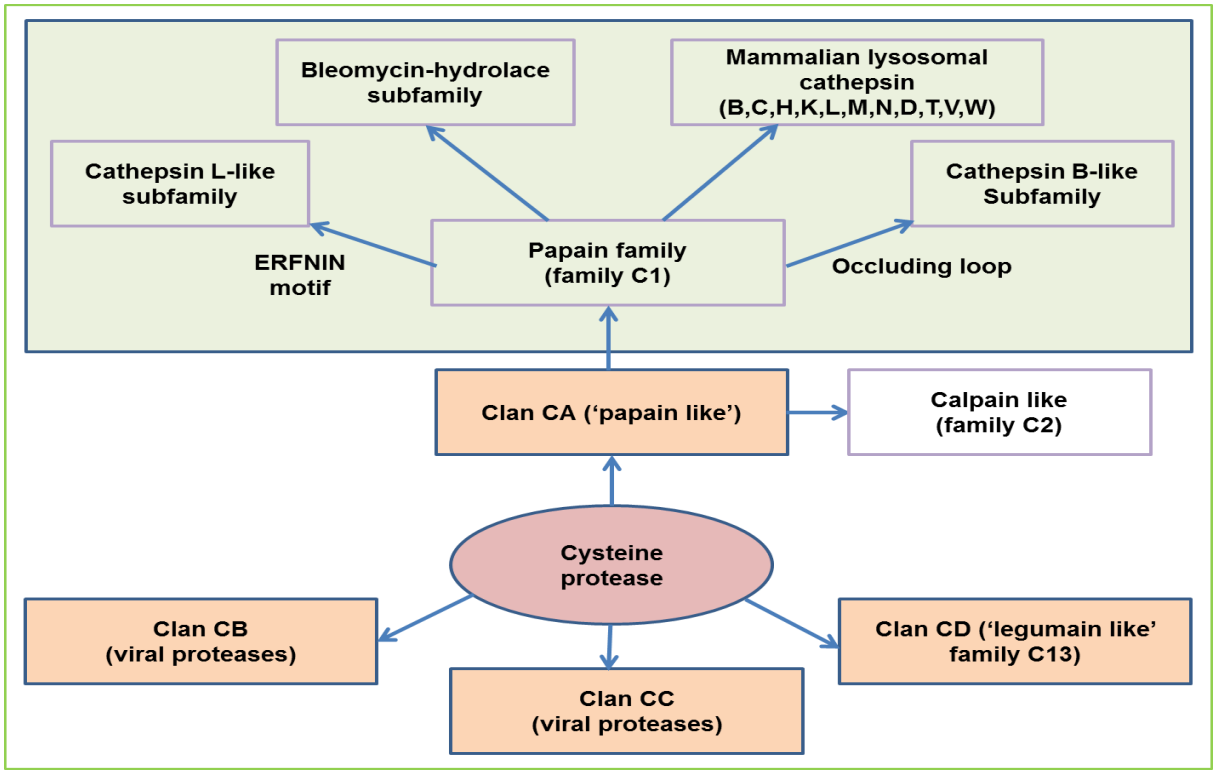


Figure 1.2: Schematic diagram showing cysteine protease family and how cysteine proteases are related. Subfamilies are determined by sequence homology around the active site amino acid residues.

1.3.1. Cysteine Proteases (Cps)

Cysteine proteases (sometimes called thiol proteases), were first purified and characterized in 1879 from *Carica papaya*, the papaya fruit [12]. CPs are found in all living organisms and they are comprised of six major families: the papain family, calpains, clostripains, streptococcal, viral and caspases/apopains [16]. The majority of discovered cysteine proteases are from viruses. Many are found in bacteria (e.g clostripain in *clostridium*) and fungi (cathepsin B in *Aspergillus flavus*). There are two main groups of cysteine proteases in mammals; cytosolic calpains (calpain type I, calpain type II) and lysosomal cathepsins (B, C, H, K, L, M, N, S, T, V, and W). There are ten clans recognised in the cysteine class (CA, CD, CE, CF, CH, CL, CM, CN, CO and C- a family not assigned to a particular clan). Classification of cysteine proteases into clans and families is based on sequence similarity, biochemical specificity to substrates and on possessing an inserted peptide loop as demonstrated in Figure 1.2 [12], [15].

1.3.2. Structure and hydrolysis mechanism

Cysteine proteases have a molecular mass of about 21-30kDa. They have an optimum pH of 4-6.5 and they need to be in an environment which contains a reducing agent to avoid oxidation of the thiol group [19].

Cysteine proteases like papain, cruzain, and cathepsin are mainly made up of antiparallel β sheets with segregated α and β protein subunits. The papain is the best characterized cysteine protease. The papain has 212 amino acid residues and its structure consist of two domains. The left hand (L-domain) consists of residues 10-108 and 207-212 while the right hand (R-domain) consists of the remaining residues [20], in accordance with the standard view. The L-domain consists of three α-helices. The central helix is the longest and it is oriented vertically. The R-domain is mostly made of β-barrels with α-helix at the bottom. The 'V' like shaped active site is located between the domains on top of the enzyme structure. The two catalytic residues Cys25 and His159 (papain numbering), each from the N-terminal of the central helix of the L-domain and the β-barrel residues of the R-domain respectively, are located in the enzyme active centre. Figure 1.3 shows the crystal structure of papain in complex with the cysteine protease inhibitor E64 as adopted from PDB ID: 1PE6 [20]. The catalytic residues form the thiolate-imidazolium ion pair responsible for the proteolytic activity of all cysteine proteases and they are highly conserved in all cysteine proteases. These conserved regions can therefore be used to classify proteases and to clone orthologous genes. They can also be used to identify selective anti-protease inhibitors and to predict natural compounds .The active site of papain like proteases is complemented with Asn175 to keep the His imidazole ring in optimal orientation during hydrolysis circles [21].

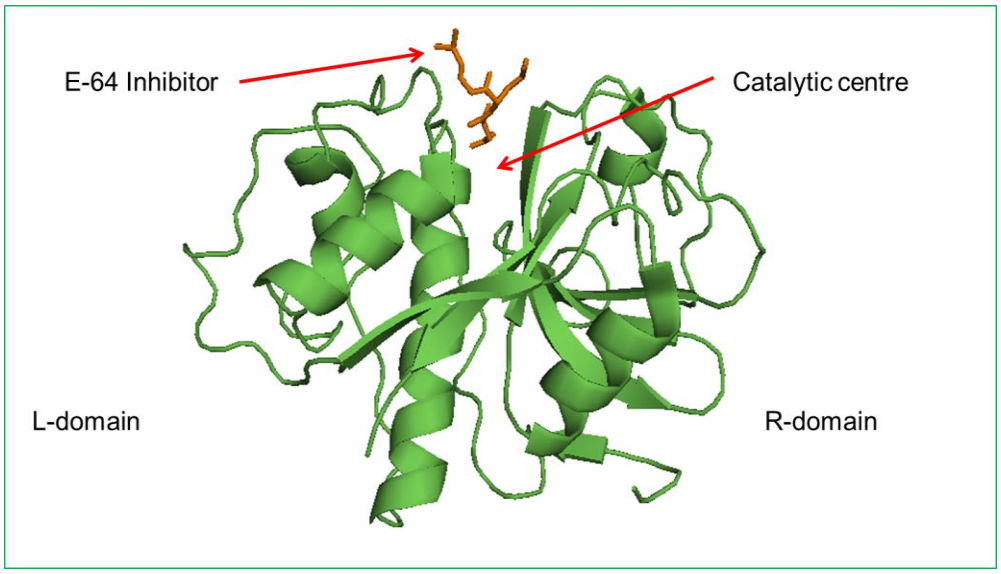


Figure 1.3: The crystal structure of the cysteine peptidase papain, PDB ID 1PE6 (green) in complex with its covalent inhibitor E-64 (orange). The Cys25 and His129 residues of the catalytic centre are located in the groove between the L and the R domains.

The papain family cysteine proteases catalyse the hydrolysis of peptide, ester, thiol ester and thiono ester bonds [16]. The large binding site of papain allows for multiple interactions between the enzyme and the substrate which binds along the active site cleft in an extended conformation. These interactions are important during hydrolysis at the active site since they contribute to stabilization of intermediates that are formed. An essential cysteine residue is required in the active site for hydrolysis. During hydrolysis, the Cys25 residue is used as a nucleophile and the His residue is used as a general base [21]. The nucleophilic thiolate cysteine attacks the carbonyl carbon of the scissile bond of the bound substrate breaking the double bond between the carbon and oxygen to a single bond (Figure 1.3. A). This forms a tetrahedral intermediate which is stabilised by the oxyanion hole [16], [21]. Hydrogen bonding to the NH group of Gln19 side chain and Cys25 backbone stabilizes the oxyanion hole. The tetrahedral intermediate converts into an acyl enzyme (enzyme-substrate thiol ester) when protons are transferred from the imidazolium cation to the nitrogen of the peptide bond being hydrolysed, resulting in cleavage. Hydrogen bonds are formed between the His159 and the new substrate amide while the carboxylic part of the substrate is bonded to Cys25 by a thioester bond (acylation) (Figure 1.3. B). The amide part of the substrate dissociates and it is replaced by a water molecule. The polarized water molecule attacks the carbonyl carbon of the acyl enzyme, releasing the free enzyme and the N-terminal fragment of the substrate (deacylation) (Figure 1.3. C). The second tetrahedral intermediate is them formed. The last step involves the thioester

deacylation which results in the reconstruction of the carboxyl group of the hydrolysed peptide, at the same time readying the enzyme molecule for a new catalytic cycle (Figure 1.3. D). The hydrolysis mechanism of cysteine proteases as shown by the papain cysteine protease and as described by Rzychon et al is shown in Figure 1.3.

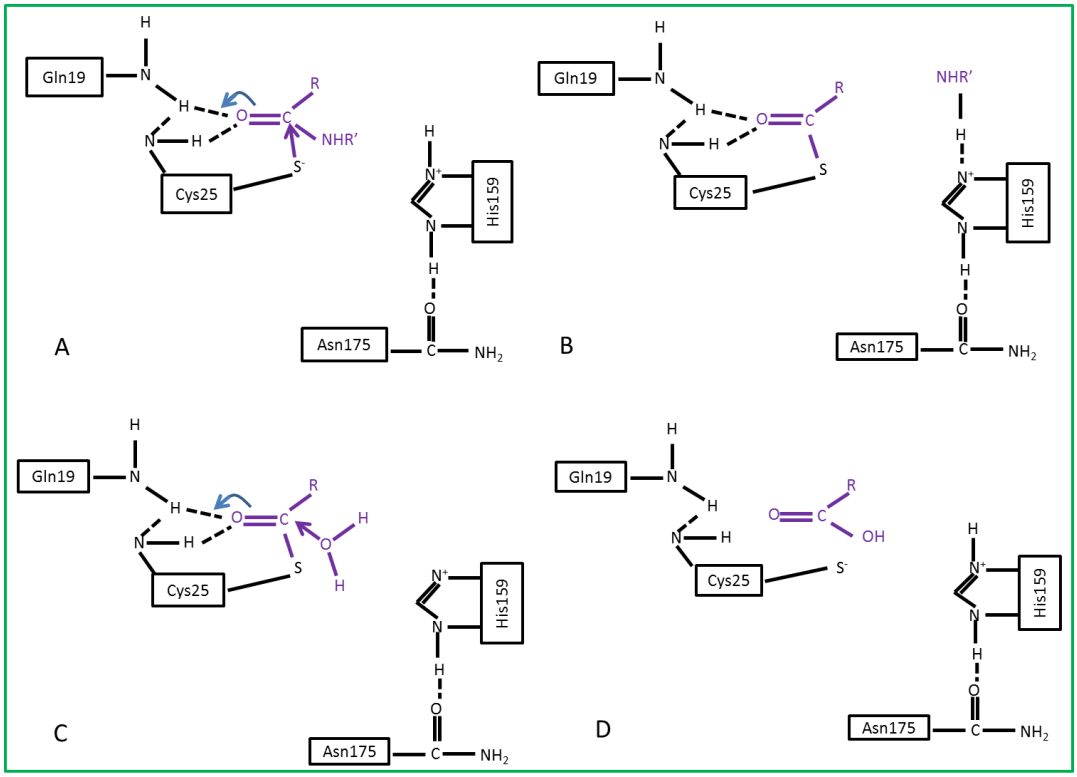


Figure 1.4: The mechanism of hydrolysis of cysteine proteases as shown by papain (description is in text).

The binding site of cysteine proteases has binding pockets called subsites. These subsites interact with substrate amino acids in the N-terminal and the C-terminal direction from the scissile bond. Subsites in the N-terminal (non-prime) direction are labelled S1, S2 and S3 and those in the C-terminal (prime) direction are named S1', S2' and S3'[8]. The substrates or inhibitor amino acids that bind in the subsites are named P1, P2, and P3 (those on the amino acid residue side of the scissile bond) and P1', P2', and P3' (on the carboxyl-terminal side of the scissile bond) [15]. Active site residues are located in four loops located in both the L-domain (two short loops) and the R-domain (two large loops). The main interactions with the substrate are thought to occur at the S1, S3 and S2' subsites of the L-domain loops and S2 and S1' subsites of the R-domain loop in papain family cysteine proteases. It has been shown that the E64 inhibitor binds to papain by forming hydrogen bonds with residues of the S subsites and the catalytic sites [20]. All papain family proteases share similarities in hydrolysis mechanism, optimum pH range, molecular mass, enzyme activity and at the regions near the active site [16], [22].

1.4. Cysteine proteases of parasitic organisms

1.4.1. Classification and evolution

Cysteine proteases of parasitic organisms are divided into clan CA and clan CD proteases. Crucial parasite proteases are located in the papain family C1 cysteine proteases of clan CA group (Cathepsin-L and Cathepsin-B like) [23] and in the C2 family (calpain-like). Cysteine proteases of pathogenic organisms, specifically papain family C1 proteases belonging to clan CA group have been implicated in the virulence-associated with the organisms [17]. They contain the plant papain cysteine protease, and the mammalian lysosomal Cathepsins B, C, K, L and S. In human, C1 proteases Cathepsin-L and Cathepsin B proteases are important in the immune system for protein degradation/turnover (cathepsins B, L and H), and for bone resorption (cathepsin K). In parasites, they are important for host entry, feeding and suppression of the host immune responses, for example brucipain is involved in parasite migration across a model of the blood-brain barrier [18]. Papain- family C1 proteases of Clan CA, are crucial for parasitic diseases and many of them have been identified as promising drug targets, like falcipain-2 of *P.falciparum.* Other cysteine protease of Trypanosomes are papain-like family C13 belonging to Clan CA (GPI:protein transamidase) and C50 (separase) of clan CD. Other clans and families of parasitic organisms include clan CB and CC (viruses) and legumain-like family C13 proteases belonging to Clan CD. Cysteine proteases of parasitic organisms are products of independent evolutionary events. This is demonstrated by the order of catalytic residues Cys/His (as in clan CA) or His/Cys (as in clan CD) in their protein sequences [12] .

1.4.2. Papain-like family proteases

The papain like family C1 cysteine proteases of Clan CA makes the majority of parasitic cysteine proteases [24]. Their catalytic activities includes endopeptidases (papain and glycyl), and aminopeptidases (Cat H). They also include peptidases with both endopeptidase and exopeptidase activity like Cathepsin B and H [16]. A common feature of all papain-like cysteine proteases is that they are made up of a peptide, a propeptide (prodomain) and a catalytic domain. The catalytic domain represent the mature active enzyme [15]. The function of the 10-20 amino acid long peptides is translocation into the endoplasmic reticulum during ribosomal protein expression. The prodomain is responsible for protein folding of the catalytic domain. The prodomain also plays a role in the transport of the proenzyme to the endosomal-lysosomal compartment. Thirdly it acts as a high-affinity reversible inhibitor to prevent the catalytic domain from being activated prematurely. The catalytic domain of papain-like proteases is 220-260 amino acids long [15]. An exception is made for cysteine proteases from some parasites whose cysteine proteases contain a C-terminal extension of unknown function. The catalytic

domain of papain like proteases has the most highly conserved regions when compared to the other two domains. As mentioned previously, the conserved active site of cysteine proteases consist of a cysteine, histidine and asparagine residues. Another feature of clan CA proteases is that they are inhibited by E64 (L-*trans*-epoxysuccinyl-leucyl-amido (4-guanidino) butane) and that their substrate specificity is define by the S2 pocket [12], and their specificity is restricted to members of the chathepsin sub families.

1.4.3. Lysosomal cysteine proteases

Lysosomal cysteine proteases, also known as cysteine cathepsins (Cats) are crucial in many biological processes in all living organisms as they are involved in degradation of extracellular and intracellular material [22], [25]. Cysteine cathepsins have pH-optima of 5.0 – 6.5, but they can be stable at different pH conditions, for example cathepsin S which is stable and active at a neutral and a slightly alkaline pH. Human cysteine cathepsins include cathepsins B, C, F, H, K, O, L, S, V, X, and W [12], [14], [22]. Cathepsins share the same amino acid sequences and mechanism of action as other members of the papain family. They however display a broad and distinct substrate specificity and regulation, preferring to cleave the substrate after basic or hydrophobic residues [16]. Cysteine cathepsins are mostly endopeptidases (except Cat C) and they can work as both endopeptisase and exopeptidase enzymes (Cat B) and as aminopeptidases (Cat H). They are glycosylated and phosphorylated in the golgi apparatus as precursor proteins. The majority of chathepsins are monomeric proteins of Mr ~ 30-50 kDa, an exception is made for Cat C which is an oligomeric enzyme of Mr ~ 200kDa [14]. Most of them are ubiquitously expressed (B, H, L, C, F, O and V) and they are involved in normal protein degradation and turn over. The most abundant of these is cathepsin B which has been implicated with cancer and is capable of degrading extra-lysosomal matrix in diseases such as muscular dystrophy, and rheumatoid arthritis [25]–[27]. Other cathepsins, (K, W, S) are expressed only in specific tissues or cells which indicates a more specific role. Cathepsin S is restricted to the major histocompatibility complex in antigen presenting cells (APCs) derived from the bone marrow. Cathepsin K, which plays a role in bone resorption, is mainly expressed in the osteoclasts, in epithelial cells and in the synovial fibroblasts in rheumatoid arthritis joints. Cathepsin W is expressed in the CD8+ T-lymphocytes and in natural killer cells. Other tissue-specific cathepsins are Cat V which is expressed in the thymus, testis and cornea [14], [22].

The expression of lysosomal proteases is regulated in order to allow the cell to respond to changing physiological situations. An imbalance in their enzymatic activity has been found to be involved in pathological conditions. The involvement of cathepsins in diseases seems to be restricted to their enzymatic activity outside the lysosome as a result of an imbalance between

their catalytic activity and their natural inhibitors. Cysteine cathepsins are associated with arthritis, neurodegenerative as well as cardiovascular diseases. Genetic diseases like severe bone abnormalities (pycnodysostosis) is linked to a mutation that results in a loss of function of Cat K, while a loss of function mutation of Cat C gene is linked to Papillon-Lefevre syndrome [14].

1.5. The role of cysteine proteases in Trypanosomes

Trypanosomes contain cysteine proteases that are regulated at different stages during their life cycle. Through gene manipulation and the use of specific inhibitors, studies have revealed the role of enzymes in parasite pathogenicity, in manipulation of the host immune system and in parasite replication [28]. The enzymes with the highest activity in trypanosomes are Type I cysteine proteases. Type I cysteine proteases of trypanosomes have a distinct carboxy-terminal extension that sets them apart from other cysteine proteases. The mRNA of TbCatB is mostly expressed in the blood stream form of the parasite [29]. Cruzipain and gp57/51, cysteine proteases from *Trypanosoma cruzi* are expressed throughout the life cycle of the parasite. They are most abundant in the replicating forms and in the epimastigote stage in the insect host. Previous analysis of trypanosomes has shown that Type I cysteine proteases are encoded by multicopy genes arranged in tanden arrays. During the development of the parasite, proteases are synthesized to serve different functions.

1.5.1. Nutrition

Cysteine proteases have evolved to hydrolyse proteins early on in their evolution. This is confirmed by their existence in parasites and other cellular organisms that represent the earliest forms of eukaryotic cells. They are capable of carrying both endogenous and exogenous protein degradation. Their exogenous activity is best exemplified by the hydrolysis of haemoglobin by cathepsin-B1 (SmCBI) of *S. mansoni* and falcipain 2 of *P. falciparum* (Fp2) [12]. The blood stream form of *Trypanosoma brucei* lacks cytochromes, so they acquire iron from the host by degradation of transferrin using TbCatB in the lysosomes [29].

1.5.2. Tissue and cell invasion

Cysteine proteases are involved in cellular process and they are thought to play a role in cell invasion in *T.cruzi*. According Sajid & McKerrow, parasite invasion and development were reported to be reduced by peptidyl diazomethane inhibitors *in vitro*. Brucipain is thought to facilitate in disruption of the blood-brain barrier [18].

1.5.3. Encystment and hatching

An infection by a parasite is usually followed by the formation of a cyst where the insect (tsetse fly) bit the host and deposited the parasite. The parasites first multiply in this protective cyst before they hatch and invade other parts of the host. Proteases are required during the formation of a cyst and during the cyst rupture for infection to occur [12].

1.5.4. Immunoevasion

During invasion of the host cells, parasites need to find a way to evade the host immune system. It is hypothesized that parasites cysteine proteases are involved in evading the immune system of the host. African trypanosomes seem to release Cysteine proteases into the host bloodstream, and these released enzymes are suspected of causing platelet aggregation thereby contributing to the pathogenicity of the disease [28]. Cysteine protease inhibitors of *T. cruzi* have been documented hydrolysing the host antibodies [12]. Cruzain, the major cysteine protease of *T. cruzi*, has been connected to blood plasma leakage in veins. Cruzain is also suspected of recruiting macrophages for invasion of host cells. Cathepsin-L like proteases have been linked to the reduction of secretory leukocyte protease inhibitor, (a protective inhibitor found in saliva, blood, tears and vaginal fluid).

1.5.5. Non Erythrocytic parasite stages

There is limited data on the role of cysteine proteases in the non erythrocytic parasite stages. Allergic responses have been reported for *T. cruzi* and papain active enzyme in mice.

1.6. Cysteine protease inhibitors mechanism

Cysteine proteases produced by parasitic organisms play a role in the pathogenicity of the organisms and their effect in the host may lead to an imbalance in endogenous protease activity, which in turn may lead to formation of diseases. Precise regulation of protease activity is essential for homeostatic cell activity and organism survival. Biological systems have developed natural ways (regulated expression, secretion and activation of the pro-proteases) to protect the organism from unwanted protease activity [30]. Protease inhibition is one way of regulating protease activity and contribution to protection against exogenous protease activity. An understanding of the inhibitory mechanisms employed by natural protease inhibitors may provide prospects into application of selective inhibitors in chemotherapy. Cysteine protease inhibitors act by blocking access to the active centre from the substrate. Natural inhibitors have developed effective mechanisms to archive this [21].

1.6.1. The propeptide backward binding mechanism

Cysteine proteases are synthesized as inactive precursors [26], activation of newly synthesized cysteine protease precursors requires proteolytic cleavage of the N-terminal prodomain that inhibits the mature enzyme [8]. The mechanism of propeptide inhibition was revealed by crystallographic studies of procathepsins B, L and K. Most cysteine protease prodomains are made up of two domains; the N-terminal domain which is made up of two α-helices and an extended β-strand and the C-terminal domain which interacts with the "proregion binding loop" of the mature protease. During inhibition, the C-terminal segment blocks the substrate binding site using its backbone and cuts access to the enzyme active site by binding between the two domains that make up the enzyme [8], [21]. The proenzyme covers most of most of the enzyme's hydrophobic surface behind the S1' subsite by exposing its own hydrophilic surface to the solvent [26].

1.6.2. The pSpeB mechanism (profragment that distorts the enzyme catalytic centre)

Streptococcal pyrogenic exotoxin B (SpeB) is a papain like protease isolated from Streptococcus pyogenesis [31]. Its profragment is made up of four-stranded antiparallel β-sheet flanked by α-helices. To archive inhibition of the SpeB enzyme, the catalytic His195 residue is pushed out from the active centre, so that it does not interact with the catalytic Cys47 residue. This is accomplished when the Asn89 residue penetrates the substrate binding site of the mature enzyme in a position similar to the S1' site in papain-like proteases [21].

1.6.3. The serpins mechanism (covalent interaction and catalytic centre distortion)

According to Rzychon et al. 2004, some serine protease inhibitors can block both serine and cysteine protease activity. These serpins are distinguished by a surface exposed reactive site loop that is a target for proteases. Serpins inhibit an enzyme by partial denaturing and disruption of its catalytic centre [32].

1.6.4. The p35 mechanism (covalent inhibition and steric hindrance)

The p35 is a virus cysteine protease covalent inhibitor which is produced to suppress the host immune response by forming a thioester bond with caspases [33]. The p35 protease inhibitors can inhibit almost all caspases but have no activity towards other protease families. The p35 inhibitor blocks access to the caspace catalytic His317 residue to inactivate the enzyme [33], [21].

1.6.5. The cystatins mechanism

Cystatins are the largest and most well described natural inhibitors of cysteine proteases. Cystatins inhibit the activity of papain superfamily members in viruses, bacteria, plants and animals. They are divided in to stefins (family I), cystatins (family II) and kininogens (family III) according to sequence homology. They block access of the substrate by binding adjacent to the protease active site without directly interacting with the catalytic centre. Cystatin inhibitory domain is made up of five-stranded antiparallel β-pleaded sheets surrounding α-helix. The domain forms a chisel like shape that fits in the papain active site. The N-terminus is distinguished by Gly8 and Ala10 residues and two hairpin loops carrying conservative motifs QVVAG and PW. During inhibition, the two hairpin loops interact with the protease surface from S1' to the S4' binding sites while the N-terminal of the cystatin interacts with the S3-S1 subsites. This leaves the polypeptide chain pointing away from the enzyme active site at the P1 position, avoiding cleavage [21], [30].

1.6.6. The thyropins and chagasins mechanism

Thyropins mechanism of inhibition has been described as efficient as and more selective (to cathepsin L) than that of cystatins owing to its structure which allows extra contact with protease surfaces [34]. Chagasins were first identified in *Trypanosoma cruzi* and they inhibit papain like proteases. Their inhibition mechanism is comparable to that of cystatins [21].

1.6.7. The IAP mechanism

Inhibitors of the apoptosis protein family (IAP) are endogenous cysteine protease inhibitors that directly inhibit caspases. They have a characteristic subunit structure with one or more BIR (baculovirul IAP repeat) domain. Their mode of inhibition works by sterically blocking the substrate access to the enzyme catalytic centre [21], [35].

1.6.8. Staphostatins

Staphostatins are highly specific towards bacterial papain-like cysteine proteases called staphopains. They are able to inhibit protease activity my preventing the stabilisation of the tetrahedral intermediate during proteolysis [21].

Mammalian homologs of parasite proteases are currently being targeted by pharmaceutical companies and this has produced a group of inhibitors that are suitable drug targets against parasitic diseases. However, for these inhibitors to be useful for clinical purposes, they have to be selective for parasite proteases only and not affect parasite protease homologs in human. Several compounds that can target parasite proteases without posing danger to the host have

been demonstrated; Vinyl sulfone-derivatised pseudopeptides have been shown to cure mice infections of *T. cruzi* [12]. This indicates that cysteine proteases inhibitors can be developed to make effective, safe and orally administered drugs against parasitic diseases.

1.7. *Trypanosoma brucei* Cathepsin B like proteases (TbCatB)

TbCatB is a papain family C1 cysteine protease which belongs to Clan CA group. It is secreted together with another cysteine protease, rhodesain, a cathepsin L-like protease which was previously thought to be the essential cysteine protease of the parasite [23]. Research has shown that RNA interference of rhodesain did not eliminate the parasite from cell cultures while RNA interference of TbCatB killed cultured parasite [18], and cured infected mice [29]. So, TbCatB is a more promising drug target than rhodesain.

1.7.1. Expression by parasites

In the mammalian host TbCatB and rhodesain are both produced by bloodstream *T. brucei* although it is produced in smaller amounts [23]. During the lifecycle of the parasite, the mRNA of TbCatB is expressed in larger quantities in the metacyclic trypomastigote (bloodstream) stage than in the procyclic trypomastigote (tsetse fly) stage [29]

1.7.2. Biochemical characterisation

Sequence analysis of TbCatB has shown it to be a 341 amino acid polypeptide with a predicted Mr of 37.223. Its open reading frame is made up of the same motifs identified in the active sites of lysosomal cathepsins. TbCatB is a carboxypeptidase of both endopeptidase and exopeptidase activity. In common with all the other Clan CA cysteine proteases, the catalytic triad, Cys122(29), His282(199), and Asn302(219) (TbCatB numbering with HsCatB numbering in brackets) is present. The open reading frame of TbCatB also encodes Gly-Cys-Xaa-Gly-Gly motifs, which are found in cathepsin B family proteases. In human cathepsin B (HsCatB), this motif is similar to residues 70-74. In addition to all the features common to Clan CA cysteine proteases, the cathepsin B-like enzyme has an ‘occluding loop’. The occluding loop of TbCatB contains histidine residues (His194 and His195) which are used to dock the C-terminal carboxylic group of peptidyl substrates [29]. A unique feature of cathepsin B proteases is the presence of Glu245 residue as part of the S2 subsite [22] while TbCatB contanins Gly328 at the same position [23].

1.7.3. Structure and function of TbCatB domains

The crystal structure of TbCatB in complex with CA074 has been reported at a refined resolution of 1.60 Å and having an R-free of 17.8 % and an R-factor of 14.7 % [23]. TbCatB shares the papain like fold that is characteristic of cathepsin B-like proteases [8], [36]. Its structure is made up of the L-domain and the R-domain. The L-domain consists of three α-helices with the central helix arranged vertically while the R-domain is made up of six antiparallel strands that form a twisted β-pleated sheet with α-helix at the bottom and a helical loop segment on top. It also has an 'occluding loop' which is a unique characteristic of cathepsin B-like proteases. The occluding loop covers the prime side of the substrate binding site and it is thought to give cathepsin B-like proteases their exopeptidase activity [29] by the removal of dipeptide units from the C-terminus end of the substrate [37]. The occluding loop of TbCatB has a rigid and a flexible (mobile) region. As observed in HsCatB, the occluding loop in TbCatB can be oriented in a "closed" conformation which results in the flexible region covering the S1' and S2' subsites at the substrate cleft or "open" conformation which exposes the hydrophobic subsites. The S1' and S2' subsites are highly conserved between human and trypanosome cathepsin B proteases. The occluding loop in TbCatB is found to be more rigid than that of HsCatB. This results from the differences in the number of hydrogen bonds in the flexible part of the occluding loop. When the occluding loop of TbCatB is in the closed conformation, four hydrogen bonds from residues (His189, His190, His194 and His195) restricts the flexible part of the loop to four residues which results in an opening of 8.5Å in the occluding loop crevice. In HsCatB, one hydrogen bond is involved in the closed conformation, the flexible part of the loop then has ten residues which results in an opening of 11.9Å in the crevice [8]. The occluding loop in TbCatB has three residues (Lys197, Try202 and Phe208) that are not available in mammalian homologs and it has a different motif ("FNFD") to that of mammalian cathepsin B ("GEGB"). The two phenylalanine residues in the "FNFD" motif of TbCatB results in a more stable opening around the S1' subsite while glycine residues in the "GEGD" motif in HsCatB results in the flexibility that allows movement of the Glu residue in and out of the S1' subsite. The mature domain of TbCatB is also reported to have a longer N-terminus when compared to that from rhodesain, HsCatB and papain. It shares this feature with malarial proteases falcipain-2 (FP-2) and falcipain-3(FP-3)[23]. The structure of TbCatB is shown in Figure 1.3.

1.7.4. Structural basis of TbCatB inhibition

The specificity of a protease results from interactions between the substrate and the enzyme substrates at the active site [22]. Understanding these interactions and how they differ between HsCatB and TbCatB can aid in the design of parasite specific inhibitors. According to Turk et

al, only three substrates binding sites (S2, S1 and S1'), are well defined for substrate residues to interact with the enzyme by main and side chain interactions while the S3 and the S4 binding sites are not real sites but areas in which the substrate residues find their most favourable binding position. Vertebrate cathepsin B proteases were described as having an acidic residue at the bottom of the S2 subsite to assist in binding with basic P2 residues. The S2 of TbCatB has a Gly328 in this position while that of HsCatB has Glu245. The small TbCatB Gly residue allows for a large and deep S2 pocket which can accommodate lager P2 residues whereas the large Glu residue results in a smaller and shallower S2 pocket in HsCatB [23]. Acidic residues (Asp166, Asp168, Asp258) line the sides while Asp327 lines the bottom of the S2 subsite making TbCatB acidic in this position [23]. The occluding loop of TbCatB contains His194 and His195 residues which are used to dock the C-terminal (carboxylic group) of peptityl substate during hydrolysis. Another difference between the occluding loop of TbCatB and the HsCatB protease is the possession of "FNFD"motiff by TbCatB, which corresponds to "GEGB" in HsCatB. Differences between the HsCatB protease and the TbCatB protease could be exploited to design additional specificity to parasite specific inhibitors [8].

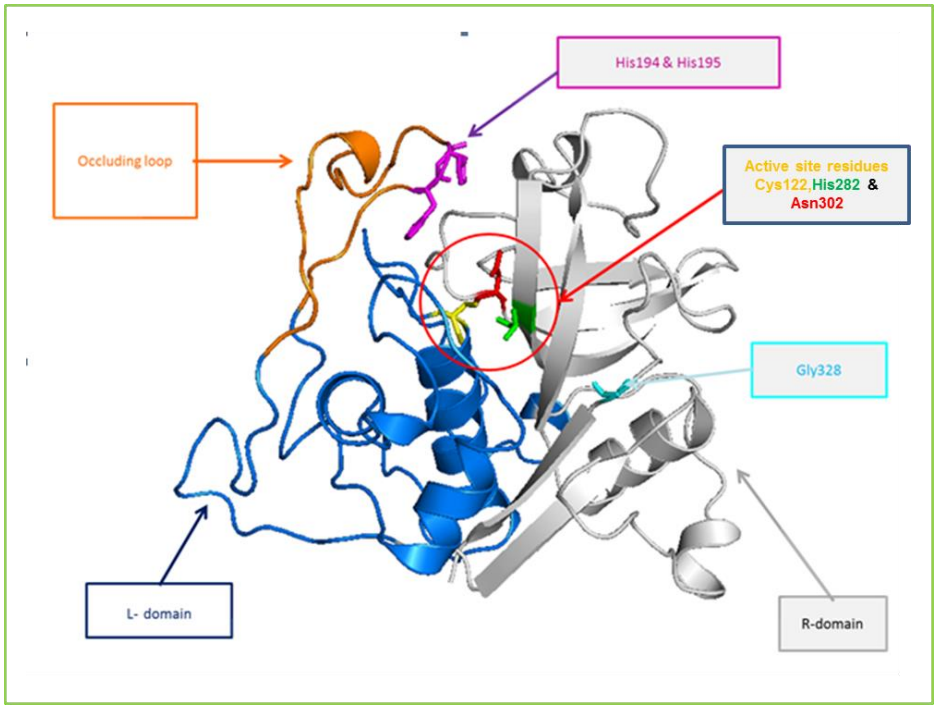


Figure 1.5: Cartoon plot of the TbCatB protease (PDB ID 3HHI) showing the typical papain-like fold of cathepsin B-like proteases. The L domain (Navy Blue), R domain (grey) and the occluding loop (orange) are highlighted. The active site residues (Cys122, His282 and Asn302) are high highlighted yellow, green and red respectively. Occluding loop His194 and His195 residues are shown as pink sticks. The Gly328 residue of the S2 subsite is shown in magenta.

1.7.5. Peptide based inhibitors of TbCatB

Peptide based inhibitors of proteases are characterised by a peptide segment in order to be recognized by enzymes. This peptide segment is an electrophilic moiety (warhead), which can undergo nucleophilic attack of the cysteine residue of the active site. They usually inactivate the enzyme irreversibly. To avoid reactivity of inhibitors with non-target host proteases, focus has been on developing papain like selective inhibitors that can identify parasite proteases from mammalian proteases by taking advantage of the amino acid residues around the cleavage site in the catalytic pocket. Electrophilic moieties that can be used as warheads incudes aldehydes and ketone derivatives, examples include chloromethyl, fluoromethyl, and diazomethyl ketones. Methyl ketone derivatives tested against *T.b. brucei* cultures have shown they can inhibit cysteine proteases. Another class of cysteine protease peptidic inhibitors contain the vinyl sulfone moiety which forms covalent bonds with the active Cysteine residue [38].

1.7.6. Non peptide inhibitors

Non peptidyl TbCatB inhibitors belonging to different classes of chemicals including acylhydrazides, ureas and thioureas, thiosemicarbazones and triazine nitriles have been identified by virtual or high throughput screening followed by rational drug design by molecular modelling studies [38]. A series of purine derived nitriles that display selectivity against TbCatB have been developed using homology modelling methods [39].

1.7.7. Peptidomimetic TbCatB inhibitors

Peptide based inhibitors have limitations when it comes to their clinical application. They can be degraded by endogenous proteases and they have low selectivity. Another limitation of peptide based inhibitors is that their absorption through the cell membrane and the blood brain barrier is poor. Peptidomimetic drug design has therefore become attractive in improving pharmacokinetic and pharmacodynamics drug properties. Research has been carried out on the development of peptidomimetic cysteine protease inhibitors based on the 1, 4 benzodiazepine scaffold as a β-turn mimetic. The β-turn is proposed as the structural motif for biologically active linear peptides. The benzodiazepine (BZD) nucleus is an excellent mimetic of the β-turn types, and it has both oral bioavailability and tolerability as attested to its use in medication (as muscle relaxants, anxiolytics, hypnotics and anticonvulsants) [38]. Reported effective inhibitors are aldehydes, α-halomethyl ketone, nitriles, epoxides, aziridines, and Michael-type acceptors. Vinyl sulfones with added functional groups that can react with the protease sulfur atom have also been reported as cysteine protease inhibitors [40].

1.7.8. Inhibition by endogenous macromolecules

Cysteine proteases are synthesized as inactive protease precursors. The N-terminal propeptide function as a selective inhibitor until the enzyme reaches the lysosome. In the lysosome, the propeptide is released and the mature active enzyme is formed. This mode of inhibition can be used to develop inhibitors that are species specific, but structural information on protease and inhibitor interaction is needed. More information is also needed to understand the extent of structural conservation between the mammalian and the trypanosome cathepsin B active site [8].

1.8. Problem statement and research justification

HAT is an example of a neglected disease and this has affected the development of new diagnostic tests and drugs [2]. Currently available drugs were developed based on their non-parasite properties without knowledge of the biochemical pathways of the parasite and therefore they are accompanied by lack of specificity to the target organism, poor efficacy and high levels of toxicity [40], [8]. Due to undesirable side effects, and the dangers associated with administering currently available drugs against HAT, it is necessary to develop new and more effective drugs against the parasites that cause the disease. Another reason to develop new drugs is because the parasites are growing resistance to some of the available drugs.

The process of drug discovery historically involved high throughput screening (HTS) of compounds in order to identify hits or biologically active compounds. This process is not only labour intensive and expensive, it also does not always end with the identification of a potent compound. To enhance this process, and potentially reduce time and cost, *in silico* methods like structure based drug design provide an alternative approach. These methods use information from high resolution 3D protein structures to explore the connection between structure and function, identify and select drug targets, study residues that are involved in protein-ligand interactions and characterize the binding pockets, develop a library of compounds that are specific for the targets, identify hits by docking experiments, and then make an optimisation of the lead compounds [41]. The *in silico* approach has gained significant attention in drug development for the treatment of parasitic diseases in the development of compounds that selectively inhibit proteases (enzymes) that are crucial for the survival of the parasites. These enzymes should be significantly different within the mammalian host to avoid accidental targeting of host proteases [38], [42]. Parasite proteases present a good target because they play an important role in replication, metabolism, survival and pathology of the organism [40].

Genome sequencing projects have shown that proteases make up approximately 2% of all expressed genes and their sequences does not differ a lot between organisms [12]. The main catalytic protease groups are serine, threonine, aspartate, metallo and cysteine proteases. The cysteine (thiol or sulfhydryl) proteases have gained interest as drug targets against parasitic diseases since they have been recognised to be critical to the life cycle or virulence of many parasites. Cysteine proteases of parasites are important in pathogenicity, tissue and cellular invasion, immunoevasion, enzyme activation, excystment, hatching and moulting. Papain family C1 cysteine proteases which belong to Clan CA group are important for metabolic pathways during the life cycle of parasitic organism so they present a valuable drug target for the treatment of parasitic diseases [23] and many of them have been identified as promising drug targets, like falcipain-2 of *P.falciprum*. Targeting of essential enzymes has also been employed in *mycobacterium tuberculosis* research in which peptidoglycan biosynthesis was a targets for drug development [43]. Parasites rely on essential cysteine proteases for survival; e.g cruzipain is an essential cysteine protease for *T. cruzi* and TbCatB and rhodesain are essential cysteine proteases for *T. brucei* [44]. Inhibition of the activity of essential cysteine proteases has been shown to be lethal to parasites, for example the cysteine protease of *T. brucei* is a validated drug target [29]. This was demonstrated by the *in vitro* and *in vivo* killing of the parasite with the cysteine protease inhibitor benzyloxycarbonyl-phenylalanine-alanine- diazomethane (Z-Phe-Ala-CH2). Although both rhodesain and TbCatB have been identified as essential cysteine proteases for *T. brucei* mRNA interference of TbCatB was demonstrated to cure mice from a lethal dose of *T. brucei* while interference of rhodesain only extended mouse life [23]. This therefore marks TbCatB as a better candidate than rhodesain as a drug target for treatment of HAT. The discovery and development of compounds that can selectively inhibit TbCatB without posing any danger to the human host represent a great therapeutic solution for treatment of HAT. An understanding of cysteine protease properties is important towards the development of inhibitors for pharmaceutical and agricultural applications.

To design potent and selective inhibitors, it is necessary to know the high-resolution structure of a target protease. Useful information about the binding interactions of inhibitors with protease structures can be obtained from crystal structure complexes with inhibitors [5]. Experimental methods like X-ray crystallography and nuclear magnetic resonance (NMR) can be used to determine high resolution crystal structures [45]. However, difficulties associated with purification and crystallization of 3D protein structures has slowed down the determination of high resolution crystal structures by X-ray crystallography and NMR techniques [41]. Crystal structure determination (x-ray and NMR) of proteases is far behind sequence determination [16] and so only a few crystal structures of selected proteins can be found in the Protein Data Bank.

In the absence of an experimentally determined crystal structure, when the amino acid sequence of a protein is known, homology modelling methods can be used to predict a reliable three-dimensional model of that protein by using coordinates of proteins with a known structure that share 30% or more sequence identity with the protein of interest [46]. Bioinformatics approaches like homology modelling for protein inhibitor docking provide a solution for solving and understanding protein interactions [43], [47]. The general principle used in homology modelling is that when given a homologous protein with known crystal structure, the crystal structure can be used as a template to model the 3D structure of the protein of interest [48]. The availability of the crystal structure of TbCatB (PDB ID: 3HHI) in complex with CA074 makes it possible to use bioinformatics approaches to carry out comparative sequence, structural and functional analysis of TbCatB protease and homologs.

The focus of this study is to explore the properties TbCatB protease and homologs, with the primary goal of identifying nonpeptidic small molecule inhibitors of TbCatB using bioinformatics tools.

1.8.1. Hypothesis

The project hypothesis is that the active site of TbCatB protease is different from that of HsCatB protease and so TbCatB protease can be used as a drug target against HAT.

1.8.2. Aims

The aim of this project is to use bioinformatics approaches to perform comparative sequence, structural and functional analysis of TbCatB cysteine protease and its homologs.

The project also aims to use molecular docking experiments and the South African natural compound database (SANCDB)[49] to screen for small molecule inhibitors (hits) of cathepsin B proteases and identify natural compounds that can lead to development of a drug against HAT.

1.8.3. Objectives

To use BLASTP to search and retrieve TbCatB protease sequence homologs from *T, cruzi, T. b. brucei, T. vivax, T. congolese* and to use HHpred to retrieve the HsCatB protease template for homology modelling.

To carry out multiple sequence alignment of TbCatB and its homologs to determine residue characteristics whose effect on the substrate will be analysed by docking experiments.

To carry out phylogenetic analysis of aligned protein sequences in order to determine their evolutionary relationship.

To build homology models of protein tertiary structures to determine structural properties of TbCatB and its homologs.

To perform inhibitor docking experiments on the protein model structures of TbCatB and homologs using natural compounds of South African origin to determine inhibitor and protein interaction characteristics.

Identify small molecule TbCatB inhibitors (hits) that can lead to development of a drug against HAT from the South African Natural Compounds Database (SANCDB).

To carry out molecular dynamics of the selected lead compounds to determine the stability of the ligand-protein complex models.

Chapter 2.

Homology Modelling and Protein Structure Analysis

In this chapter, bioinformatics approaches and databases shall be used for sequence, structural and functional analysis of TbCatB protease and its homologs. The sequences shall be identified from sequence and structural databases and then analysed using sequence and structural approaches. Protein function and structure is encoded in sequences [50],[51] it is therefore important to get information about the 3D structure of a protein in order to understand how it performs its function [45]. Due to difficulties associated with experimental determination of 3D protein structure [41], only a few 3D crystal structures of cathepsin B protease are available. Available cathepsin B-like protease crystals structures that are relevant to this project are those from *T. brucei*, *T. b. brucei*, and *H. sapiens*. The availability of these structures makes it possible to calculate 3D models of homologous proteases using bioinformatics methods such as homology modelling. As mentioned in chapter 1, HAT is caused by the protozoan parasite *Trypanosoma brucei* which is spread by tsetse flies [8]. Another form of the disease is Chagas' disease which is caused by *T.b. cruzi* [5]. Other trypanosomes include those that cause nagana disease in livestock. These are *T.b.brucei, T. congolense* and *T. vivax* [7]. Homology models shall be calculated for cathepsin B-like proteases from *H.sapien, T. cruzi, T. congolense* and *T. vivax*. Comparative structural analysis shall then be carried out to identify residues that are involved in protein and ligand interactions at the active site.

Once the sequence of interest also known as the target sequence is identified, four major steps are usually followed in homology modelling. The first step is to identify the template from which it will be calculated. A template and target sequence alignment is then made to assign residue correspondence before a model is calculated. If necessary the model is refined and then validated before it can be used for further application [50], [41]. In this project, the validated models shall be used for protease-inhibitor docking studies of natural compounds of South African origin. Compounds that inhibit TbCatB protease more than they do the human homolog shall be identified as leads compounds that can be used for development of drugs for HAT. Homology modelling and docking studies have been used by other researchers to screen for inhibitors with therapeutic applications [43], [45].

2.1. Introduction

Methods for predicting the 3D structure of a protein from sequence information includes; (i) *ab-initio* methods which are based on physical chemical principles, (ii) homology methods that are reliant on information available in sequence and structural databases and (iii) threading of fold recognition methods which rely on finding a template structure that closely resembles the structure of the query sequence [52]. Genome projects are producing sequences faster than X-ray crystallography and NMR laboratories can solve 3D-structures [41]. The rapid expansion in sequence databases, coupled with new sequence and structure analysis algorithms have increase the accuracy of homology models, making homology modelling a method of choice for predicting 3D coordinates of proteins for most researchers. The CASP modelling competition has also revealed that homology modelling results can be verified and the models can be used in designing projects since biologically important regions are generally accurately modelled [53] ,[54]. Proteins, RNA and DNA sequences are archived in biological databases. To make it easy for the user to search through the immense data available in databases, database management systems like BLAST are used to perform a search. During a BLAST search, the query sequence is compared with other sequences in the database to identify similar or matching sequences. Scoring matrices are used to give a statistical value to ensure the accuracy of the selected sequence [55], [56].

Structural and functional information of nucleotides (DNA/RNA) and amino acid sequences are a result of evolution [51]. Due to mutations (insertions, deletions) and other evolutionary changes, sequences undergo changes in the arrangement of their residues. During this period, regions that code for properties that are important for the survival of the species are conserved by natural selection. Examples of residues that are conserved by natural selection include residues that code for functional and structural roles. These conserved regions can therefore be exploited by sequence alignment comparison to study their evolutionary relationships and to determine their degree of similarities from which their functionality can be extrapolated [55].

Sequences whose residues are arranged in a similar way are concluded to be from the same family and to be coding for the same function, if the structure and function of one of the aligned sequences is known, its properties can be used to predict the structure and function of its homologs [57].

The number of possible ways in which a protein can fold in nature seems to be limited and the 3D structure of proteins is more conserved that their sequences. So if a sequence of known structure shares a high (30 %) degree of residue similarity with another sequence of unknown

structure and function, the sequence with a known structure can be used as a template to guide in the calculation of a model structure for the sequence with unknown structure [41], [45], [50]. In this way sequence alignment can be used to characterise protein structure and function as well as to infer phylogeny.

Detailed structural information on protein residues can be obtained from x-ray diffraction and NMR spectroscopy [45]. Obtaining protein structures using these methods can be time consuming and difficult [41], as a result only a few crystal structures have been developed and archived in the PDB. On the other hand sequence databases are expanding rapidly [53]. Since protein structure and function is encoded in their protein sequences, for a given protein sequence, bioinformatics tools like homology modelling can be used to calculate its 3D protein model by using coordinates of a protein with a known structure that shares more than 30% sequence identity [45], [41]. Functional properties of these models can then be characterised using docking studies.

2.2. Databases

Biological data is stored in computerized archives. This information in organised in such a way that information can be retrieved easily using a search criteria. A sophisticated search tool such as BLASTX [58] is usually used. Different types of data include primary data or raw data, secondary data and tertiary data. The primary data can be DNA, or amino acid sequences. The sequence information that comes from sequencing experiments is uploaded into primary databases by researchers. Examples include the GenBank (http://www.ncbi.nlm.nih.gov/genbank/) [59] made up of sequence data from authors, expressed sequence tag (EST), genome survey (GSS), SNP and GEO. Databases are crucial to bioinformatics research since they are a library of resources of information that is need for bioinformatics research. A newly acquired sequence (DNA or amino acid) can be uploaded into the database to compare it to other sequences in the database and gather information about it from its homologs [60]. So databases can be used for discovering new information. Some databases contain only DNA sequences while other contains only protein sequences. In protein databases, amino acids can be arranged to make domains which can be further arranged into motifs which are structural units of proteins. These constitute the secondary data which gives information on the protein secondary structure like alpha-helices or beta-strands. This information is stored in secondary databases which are curated and their content is controlled by the database developers. Examples include the NCBI (http://www.ncbi.nlm.nih.gov/) [61] which maintains the GenBank and also provides data retrieval and analysis systems, Protein, Refseq and RefSNP e.t.c. The tertiary data is related to the tertiary protein structure and it can be used

to predict the tertiary structure of the protein of interest. Databases are therefore divided into nucleotide and protein databases. The main nucleotide databases include EMBL (The European Molecular Biology Laboratory) [62], GenBank (USA) and the DDBJ (DNA Data Bank of Japan) [63]. They are collectively called International Nucleotide Sequence Database (INSDC) [64]. Protein databases are composed of sequence and structural databases. Protein sequence databases include UniProt, PIR and SwissProt. Structural databases include the Protein Data Bank (PDB) [65], CATH and SCOPE. Other databases include protein interaction databases like the Biogrid and STRING. There are also some whole genome databases like ENSEMBL as well as specialised data bases like OMIM (Online Mendelian Inheritance of Man). Although they are independently made, databases are interconnected in such a way that a search in one can get you information from another. Sequence alignment is applied during a search [55], [61].

2.3. Sequence analysis

Comparing sequences is the first step in structural and functional analysis as well as in phylogenetic inference of protein sequences. Similarity between protein sequences indicates similarity in structure and it provides an opportunity for structural modelling when one of the sequences has a known 3D structure [57]. To clearly visualize regions of similarity and differences in protein sequences, they have to be aligned in such a way that matching residue to residue correspondence is established so that residue variations and similarities can be visualised easily. The maximum match is reached when the largest number of amino acids of one protein are matched with those from another protein while allowing for all possible deletions [66]. One of the most popular alignment approach for nucleotide and amino acid sequences is the dynamic programming developed in 1970 [66]. During database similarity search and sequence retrieval, pairwise sequence alignment is employed to retrieve homologous sequences. The same process is also used in multiple sequence analysis to align more than two sequences based on the similarity of their residues. Sequence alignment is carried out using sequence alignment algorithms. There are two types of sequence alignment methods; the local sequence alignment and global sequence alignment method. In local sequence alignment, two sequences are aligned in such a way that the highest number of local similar residues (segments) is aligned without trying to align the whole sequence. In global sequence alignment, two sequences are aligned from beginning to end, searching their entire lengths for an arrangement that will result in the maximum number of matching residues [56], [66]. These methods can be used in different situations. Since global alignment method compares full length sequences, it is appropriate for comparing short sequences that are closely related and are almost equal in length. When it comes to comparing distantly related sequences in which only short portions are related, the

local alignment method is preferred. This method can compare short conserved regions in sequences without aligning the rest of the sequence, so it can pick and align conserved areas and motifs in sequences that are distantly related and are not of the same length [55].

Once the sequences are aligned, the best alignment has to be chosen. To guide in choosing the best aligned sequence and quantify the degree of similarity between aligned sequences, scoring or substitution matrices are used to assign scores for residue matches and substitutions [56]. A special score or penalty is usually given to account for deletions and insertions in the aligned sequences. Some scoring matrix take into consideration the frequency of amino acid substitution as well as the frequency of each amino acid that has occurred in the evolution of the sequence homologs [55]. The most common scoring matrices used in scoring multiple sequence alignments are BLOSUM (blocks amino acid substitution matrix) and PAM (point accepted mutations).

The BLOSUM was developed as a series consisting of BLOSUM45, BLOSUM52, BLOSUM60, BLOSUM80 and BLOSUM90. BLOSUM series scores are based on comparison of blocks or local segment arrangement of residues of aligned sequences [56]. They were developed to account for distantly related sequences. Each series was developed using multiple alignments of sequences with a percent identity matching the relative BLOSUM number, for example, BLOSUM45 was developed using sequences of a sequence identity of about 45%. The series with the lowest BLOSUM number is more suitable for scoring alignment of sequences that are distantly related, while the series with the highest BLOSUM number is more suitable for scoring closely related sequences.

PAM matrices are a series that was developed using very closely related sequence homologs. They were developed based on the replacement of an amino acid residue by another in a way that is acceptable by natural selection. The replacement of a single nucleotide by another is known as a point mutation. The PAM matrices were named with a number representing the number of mutations per 100 amino acid residues. In other words, this can be taken as a percentage of mutations. The lowest, PAM1, represents a 1% mutation per 100 amino acid residues. The PAM matrix followed by a high number is therefore suitable for scoring alignments that are distantly related, while the one followed by a low number is more suitable for scoring sequences that are closely related. Available PAM series range from PAM100, PAM120, PAM160, PAM200 and PAM 250 [55].

2.4. Database Similarity Search and Sequence Retrieval

Protein sequences represent a large part of data that is stored in biological databases. These sequences have to be retrieved from databases before they can be analysed. To retrieve the sequences of interest, the sequence in question, which is called a query, is submitted to the database of choice to search for homologous sequences. Pairwise sequence alignment is mainly employed during database sequence retrieval. A sophisticated search engine or database management system is usually used to carry out the search. These database management systems are required to be sensitive, selective and fast to be able to go through the amount of data stored in databases and carry out the required computation accurately. The Basic Local Alignment Tool (BLAST; https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch) [67] is usually the tool of choice for database similarity search and sequence retrieval [55] and it is based on pair wise sequence alignment. BLAST uses a heuristic word method to quickly produce a pairwise sequence alignment. During a BLAST search, high scoring un-gapped sequence regions are found. These sequences are then weighed above a certain threshold at which pairwise alignment cannot occur by random change. BLAST results include percent sequence similarity and coverage as well as an E-value. The E-value is a statistical value that calculates the probability of a pairwise sequence alignment happening by chance. Databases that can be accessed by a BLAST search include the National Centre for Biotechnology Research (NCBI). BLAST program includes;

BLAST-N : for nucleotide query to nucleotide database search,

BLAST-P : for protein query to protein database search,

BLAST-X : which uses nucleotide as a query which is translated into six open reading frames to produce translated protein sequences. The translated sequences are then used as a query to search a protein database,

TBLAST-N : which uses a protein sequence query to search a nucleotide database in which the nucleotide sequences have been translated into all the six open reading frames and finally the

TBLAST-X : which accepts nucleotide sequences as a query. The sequences are then translated into all six open reading frames to search a nucleotide database that has sequences that have been translated into all open reading frames.

Position-specific profile search method, PSI-BLAST is a BLAST variant which is able to detect distant homologs [67].

When choosing the type of sequence to use in homology detection, it is advantageous to use protein sequences since their scoring matrix take into consideration physicochemical properties between the residues in addition to residue substitution and gap penalties [56], while nucleotide scoring matrix consider residue substitution and gap penalties only.

The HHpred server (http://protevo.eb.tuebingen.mpg.de/hhpred) is also another program that can be used to identify template sequences. In addition to identification of sequences, it can also be used to calculate models in MODELLER using one of the templates it has identified. HHpred can identify distantly related sequences by searching a wide range of databases like the PDB, SCOP, Pfam, SMART, COG and CDD in a very short time. A pairwise sequence alignment search using a profile hidden Markov models (HMMs) is employed. The input is a query sequence or alignment with options to search in global or local alignment mode. An option for searching in PSI-BLAST iterations is also available. The output results include sequences (templates) with secondary structure annotation, an E-value and true probabilities. The probability value represents a principle measure of statistical significance at which the retrieved sequence (template) is a true positive homolog of the query sequence [48].

Depending on the evolutionary relationship of the query sequence to the target homologs, one or both of the two programs can be used during sequence retrieval. Another program that can be used for searching for DNA and amino acid sequences is the LFAST for local similarity analyses [68].

2.5. Multiple Sequence Alignment (MSA)

MSA are an important step in protein structure and functional analysis as well as in phylogenetic studies since biological information can be revealed from conservation or differences within aligned positions [69]. In a MSA, homologous residues are represented in a given column and they can be super imposable in structure and they share a functional role. MSA also provide an advantage over pairwise sequence alignment since they make it easy to identify conserved regions, motifs, predict functional sites and protein function in the whole sequence family [41]. Different programs are available for multiple sequence analysis. These include PROMALS3D, MAFFT, CLUSTALW and T-COFFEE. In this project PROMALS3D and MAFFT shall be used for multiple sequence analysis and to calculate a phylogenetic tree with the best bootstrap values.

PROMALS3D is a web server (http://prodata.swmed.edu/promals3d/promals3d.php) for construction of multiple sequence alignments for proteins and/ or structures [70]. It produces highly accurate alignments that have both sequence and protein structure consistency. The

server automatically identifies known 3D structure homologs for the query sequence and then combines both structure based constraints (from structural alignment) with sequence constraints to produce an alignment that has both sequence and structural information.

MAFFT is also a web based (http://www.ebi.ac.uk/Tools/msa/mafft) MSA method that is based on the Fast Fourier Transformation (FFT). The FFT is an algorithm which increases the speed at which homologous sequences are detected by converting an amino acid sequence to a sequence made up of volumes and polarity values (electronegativity) of its residues. MAFFT also employs two different heuristic methods for reducing CPU time by using a simplified scoring system which also increases accuracy of sequence alignments that have multiple variations like large insertions or extensions and sequences of diverse evolutionary origin of matching length. These methods are the progressive method (FFT-NS-2) and the interactive refinement method (FFT-NS-i). Compared with other MSA methods like CLUSTALW and T-COFFEE, MAFFT was demonstrated to be faster without losing accuracy [71]

2.6. Phylogenetic analysis

Products of evolution provide revelations to sources of diseases and they can be used to understand developing drug resistance in pathogenic microorganisms [72]. Homologous sequences frequently share the same structure and consequently the same function. Due to evolution, some homologous sequences have evolved as a result of speciation. Some homologous sequences have evolved to carry out the same function while others have evolved to carry out different functions. Homologous sequences that share a similar function in different species are called orthologs, and those that carry different functions are called paralogs and they are a result of gene duplication. To explore the structure and function of a sequence, the first step is usually to compare the query sequence with that of an evolutionarily related protein of known structure. Comparison of evolutionary related sequences has been used in identification of functional structures in genomes and in detecting homologs within and between genomes [73]. This comparison can be best assessed by using phylogenetic analysis to calculate an evolutionary tree that shows relatedness of species or certain genes that make the tree branches. From phylogenetic trees, relationships among copies of a gene or among loci of a multigene family can be inferred [74]. The separate branches represent the sources of sequences and they are called taxa. The point at which the taxa meet is referred to as a node and it represent a hypothetical common ancestral origin or a point at which a mutation was introduced and the species separated and evolved differently. The relative timing of species divergence or the lengths of internal branches determine the accuracy of the tree [75]. Branches from the same node form a clade representing descendants from a hypothetical common ancestor. Sequences

that are different from the rest of the test population form an out group. Phylogenetic results can reveal which sequences share more similarities and which sequences are different in the population [76]. Methods for estimating phylogenetic trees include neighbour joining (NJ), maximum parsimony (MP), maximum likelihood (ML) and Bayesian methods [74], [77].

2.7. Homology Modelling

Homology modelling is the calculation of the full 3D protein atomic model of unknown structure from its sequence by using the 3D structure (s) of a closely related (homologous) protein sequence as a template [76] and it has been used for structural and functional studies of enzymes to identify drug targets [43], [45]. The biological function of a protein is determined by its structure, so construction of protein structures is important in gaining information about their function. Prediction of protein tertiary structure from their sequences is based on the observation that protein three dimensional structures tend to be more conserved than their sequence [78]. The most reliable method for predicting protein 3D structure is homology modelling which is also known as comparative modelling. The principle behind homology modelling is that proteins that share a high sequence identity often adopt the same 3D structure [41]. If an experimentally solved protein structure (template) is available, models can be calculated for homologous sequences (targets) with up to 30% sequence similarity or those that show structural similarity [41]. Steps involved in homology modelling are (i) template selection/identification of a known 3D protein structure; (ii) sequence alignment of target and template proteins; (iii) 3D model construction of the target using the coordinates of 3D structure of the templates; (iv) model refinement, model evaluation and validation. These steps can be repeated until a good quality model is built [41].

2.7.1. Template Selection

Template selection is the first and crucial step towards the calculation of a high quality model [79]. Currently the best way to predict the 3D structure of a model is to search for a homologous protein of known structure from protein structure databases like the PDB. Programs that can be used for template selection are BLAST and HHpred. More sophisticated methods for searching for a template include PSI-BLAST and profile-profile alignment. A sequence identity of at least 30% between the template and target is considered high enough to give a reliable model during template selection [50]. If template and target sequence identity is below 40%, HHpred method has been shown to identify good templates but for template and target sequence identity above 40% the BLAST method can be used for selecting the template [79]. Once a template has been

identified, a template and target sequence alignment can be made in preparation for homology modelling.

2.7.2. Template and Target Sequence Alignment

The accuracy and quality of a 3D protein model depends on the template and target sequence alignment. The quality of the template and target sequence alignment is therefore the most critical step in determining the quality of the final 3D model since it provides a framework from which backbones can be copied from the template to the target model structure [78]. Multiple sequence alignments of homologous sequences can be used to increase the accuracy and quality of the template and target sequence alignment. To improve the quality of the alignment, it may be necessary to manually edit the alignment [55]. Multiple sequence alignments also provide the additional option of using multiple templates in the alignment (if they are available) to improve the 3D model quality.

2.7.3. Modelling

The process of modelling involves copying the template's backbone residue coordinates into the target sequence and calculating the actual 3D model of the target. Identical and aligned residues will assume the same side chain and main chain atoms while differences in aligned residues results in only the back bone atoms copied leaving the side chains to be added on a later step [55].

There are different programs available for use in homology modelling. Some of them are web based like the SWISS-MODEL (http://swissmodel.expasy.org/SWISS-MODEL.html) and HHpred (http://toolkit.tuebingen.mpg.de/hhpred). When using web based programs, the user inputs the query sequence and the program searches for the right template then constructs a model. Other model building programs are standalone programs like MODELLER which shall be employed in this research.

The accuracy of the calculated model is usually compared to that of the template(s) by superimposition of the model and template structures. This process gives an RMSD (root-mean-square-deviation) value. This value represents the root mean square deviation distance between corresponding Cα atoms and it can be ~1-2 [41]. Model quality assessment is usually carried out to determine the suitability of the calculated model to its intended applications.

2.7.4. Model Evaluation and Validation

The use of protein models for docking studies and other applications depends on their quality [45]. The application of the model determines the extent of the model quality and regions that

need to be modelled accurately [53]. There is currently no method available to consistently and accurately calculate protein 3D structure and there is also no method available to evaluate and calculate all the errors in a protein 3D model [80]. Models are therefore evaluated with a variety of protein model assessment methods to increase credibility of the model from complimentary results. Model evaluation results will determine if the model is good enough for the desired application or if further refinement process is necessary.

2.7.5. Model Refinement

Depending on the model evaluation results, some regions of interest may be highlighted as inaccurate and therefore may need further refinement. Sources of error occur in the template and target sequence alignment mainly due to deletions and insertions which result in gaps in the aligned sequences [57]. Since no residue information is available in the gaps, they cannot be modelled and they form loops in the model. To deal with this problem a loop modelling procedure may be necessary. The challenge with loop modelling is that there are currently no methods for loop modelling except to use techniques that search databases for parts that match the loop regions and try to fix the model. In the event that only the main chain atoms have been copied to the target, the missing residues must be calculated in those regions. Residues are important in determining protein and ligand interaction at the active site and so their geometry and energy constraints must be considered [52], [55] when loops are generated in this way.

2.8. Methodology

An overview of the methods and respective tools used during the different steps (sequence retrieval, multiple sequence alignment, phylogenetic analysis, homology modelling and model quality assessment and validation) leading to homology models is shown in Figure 2.0.

Homologous sequences were retrieved from the NCBI and PDB database by a BLAST search while the HsCatB template was selected from a list of sequences produced by searching in the HHpred server. To determine which will give the best phylogenetic tree (based on bootstrap values), MAFFT and PROMALS3D MSA tools were both used to make MSAs. The MSA from PROMALS3D was used in calculation of homology models because PROMALS3D produces accurate alignments that have both sequence and protein structure information. A hundred models were calculated and from them the best models were chosen based on their low DOPE Z score using python scrips. The models were then evaluated using different model quality programs before they could be accepted for further application.

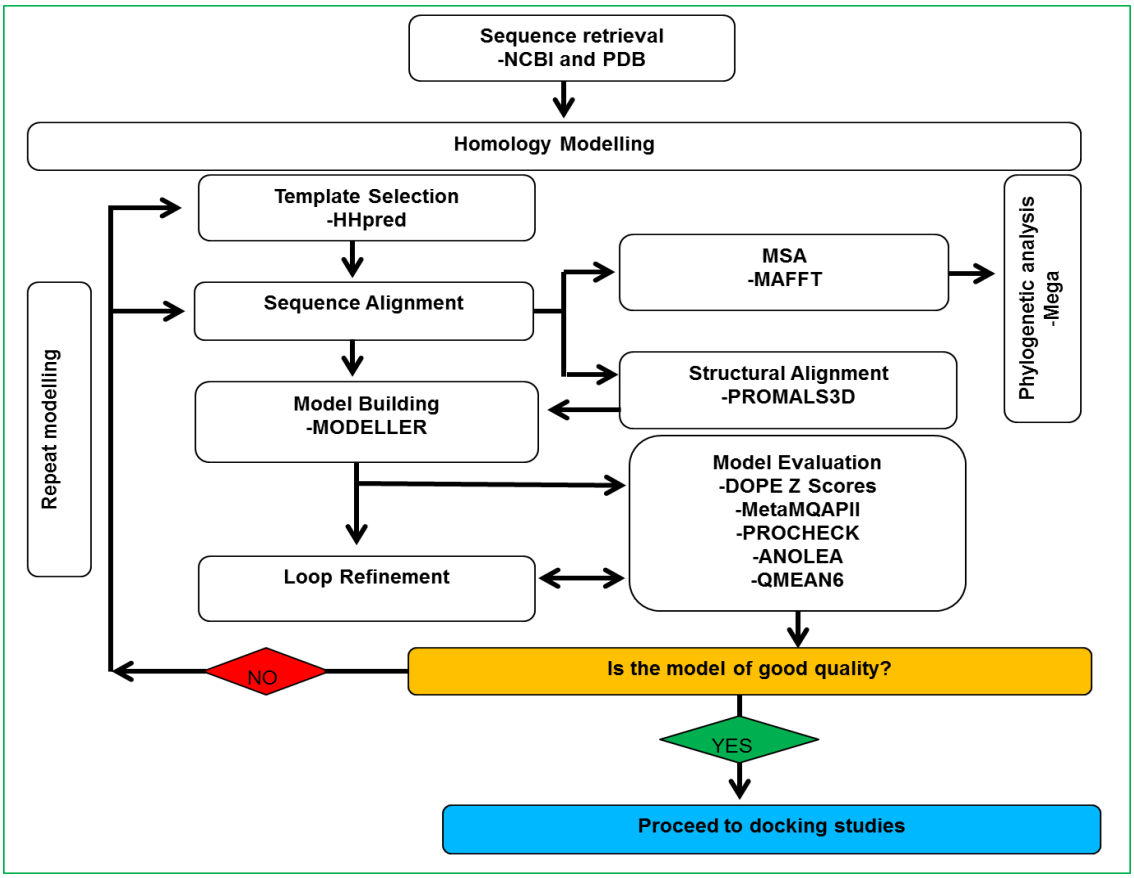


Figure 2.0: An overview of the methods used for sequence analysis and homology modelling. Both sequence and homology modelling were used for comparative analysis.

2.8.1. Database similarity search and sequence retrieval

A search *for Trypanosoma brucei* cathepsin B cysteine protease in the NCBI produced the crystal structure of *Trypanosoma brucei* procathepsin B (PDB ID 4HWY) as the latest addition to the list of results. Using default settings, the 4HWY sequence was used as a query to search for homologous sequences in a protein BLASTP search of the NCBI database. The main reason for using the 4HWY sequence as a query is that as a procathepsin, it contained both the full mature enzyme and the N-terminal prodomain. Since it contains a full mature enzyme sequence, it improved the search for finding other full mature enzyme sequence homologs. Trypanosoma cathepsin B-like (TbCatB) protease homologs from *T. congolense*, *T. cruzi*, *T. vivax* and *T. brucei* (PDB ID: 3HHI) were retrieved from the NCBI and PDB databases from the BLAST search. The HsCatB protease homolog (PDB ID: 3CBJ) was retrieved from a list of templates found by HHpred search using default settings. E-values lower than $1.0e^{-5}$ we considered to be significant when choosing sequences. Sequence coverage was also taken into consideration.

Sequences that were used in this project include two crystal structures (PDB ID: 3HHI and 3CBJ) which were used as templates, and three target sequences (*T. congolense, T. cruzi* and *T. vivax*) whose 3D models were calculated. Since the occluding loop of Cathepsin B proteases can occur in a closed conformation and block access to the active site, it was important that the two selected crystal structures be in complex with inhibitors to ensure and open occluding loop. The two crystal structures were used individually as templates to make homology models for the target sequences. They were also used together (double templates) to make homology models for the same target sequences. This provided an opportunity to compare models from a single template and from a double template. Altogether, a total of 9 models were calculated.

2.8.2. Multiple sequence alignment

MSA of sequences was carried out using MAFFT and PROMALS3D programs. The PROMALS3D alignment was used in selecting sequences for template and target alignment prior to homology modelling (making PIR files) as well as in phylogenetic analysis, while the MAFFT alignment was used in phylogenetic analysis only. Default settings were used for MSA. The use of two alignment tools was carried out to determine which alignment will produce the best phylogenetic tree.

2.8.3. Phylogenetic analysis in MEGA

Phylogenetic studies were carried out using Molecular Evolutionary Genetic Analysis (MEGA) version 5.03 [81]. Molecular Evolutionary Genetics Analysis (MEGA) is software that provides tools for statistical analysis of DNA and protein sequences from an evolutionary point of view. These tools include sequence alignment tools, phylogenetic tree construction and viewing, evolutionary hypothesis tests, sequence divergence estimates, database sequence retrieval and options for generating data. When carrying out phylogenetic analysis in MEGA5, options for phylogenetic tree inference include the Maximum Likelihood (ML), Neighbour Joining (NJ), Minimum Evolution (ME) and Maximum Parsimony (MP) methods. Phylogeny tests can be carried out in bootstrap or branch-length tests. Options for substitution model (substitution type and model/method) are available to evaluate the fit of major models of nucleotide and amino acid substitutions. The user also chooses Rates and Patterns; gamma distributed (G), has invariant sites (I), and gamma distributed with invariant sites (G+I). More options include data subset to use (choose to use all sites, complete deletion of gaps, or partial deletion of gaps) and tree inference options [81].

The maximum likelihood with a thousand boot strap tests method was used during phylogenetic tree construction. The model of substitution was set to amino acid, WAG+G substitution model.

The coverage cut off was set to 95 % for the Nearest-Neighbour Interchange (NIN) inference with a bootstrap value of 1000. Other parameters were left as default setting. Trees were estimated for both the MAFFT and the PROMALS3D alignment.

2.8.4. Homology modelling

Models were calculated for *T.congolense, T.cruzi and T.vivax* cathepsin B-like protease sequences since they did not have crystal structures. These models shall be evaluated with different model evaluation tools until satisfactory models are acquired before they can be used for further applications.

The following sections explain activities that were carried out during homology modelling to calculate the 3D protein models.

2.8.4.1. Template selection and multiple sequence alignment

Two templates were used in the calculation of *T.congolense, T.cruzi and T.vivax* cathepsin B-like protease models. One template was the *T. brucei* protease (PDB ID: 3HHI) template which was retrieved by the BLAST search during sequence retrieval. This is the structure of TbCatB in complex with the cysteine protease inhibitor CA074 [23]. The other structure was the HsCatB protease (PDB ID: 3CBJ) template which was chosen from a list of templates retrieved from searching in HHpred server (http://toolkit.tuebingen.mpg.de/hhpred). This is the structure of cathepsin B in complex with the *T. cruzi* inhibitor, chagasin [37]. It was important to choose the template in complex with an inhibitor to make sure that the occluding loop is modelled in the open position to allow for access to the active site by ligands during docking studies.

2.8.4.2. Model building

The MODELLER version 9.10 [82] program was used to carry out the calculation of the 3D protease models. Scripts that were used were obtained from the MODELLER manual and then edited to suit the available data and project needs. A hundred models were calculated for *T. congolense, T, cruzi and T. vivax* cathepsin B-like proteases using 3HHI and 3CBJ as templates. The best models were chosen based on their low DOPE Z scores before validation using MetaMQAPII, ANOLEA, PROCHECK and QMEAN6. The 3D model structures were visualized using PyMOL version 3.5.

2.8.4.3. Model evaluation

Before the models could be used for docking experiments, their quality had to be assessed and validated using different model quality assessment programs. The models were assessed and validated on the basis of their geometry using PROCHECK, and energy aspects using DOPE Z

scores, MetaMQAPII, ANOLEA and QMEAN6. The root mean square deviation (RMSD) between the main-chain atom of model and template was calculated by structural superimposition of templates and each model.

2.8.4.4. Model quality evaluation programs

1. DOPE Z scores

It is a common practice to calculate a number of alternative models during structure prediction. From these models, the most accurate model is selected for subsequent assessments before it can be used for the desired application. The best models from MODELLER are chosen based on their DOPE Z score. The DOPE Z score is a measure of the model energy which is used to determine the stability of the model. A positive DOPE Z is likely to be a poor model while scores lower than -1 are likely to be close to the native like state of the protease. MODELLER provides a script that can be customised to calculate the DOPE-Z score.

2. Root Mean Square Diviation (RMSD)

To measure the similarity between two alternative conformations of a globular protein the root-mean-square deviation (RMSD) of virtual backbone Cα atomic coordinates is determined after optimal rigid body superposition. The RMSD measures the difference between the Cα atom positions between two proteins and so a smaller deviation means that the proteins are more spatially equivalent. In the macromolecule viewer PyMol [83], structural alignment can be used to determine the RMSD of two structures by superposition. To determine the RMSD, PyMol first carries out a sequence alignment and then tries to align the structures to minimise the RMSD. The output contains an executive RMSD for the two structures. The calculation can be restricted to a certain number of residues or the alignment of just the backbone atoms [84].

3. MetaMQAPII

MetaMQAPII program is available as a free webserver (https://genesilico.pl/toolkit/) for protein model quality assessment. While most model quality assessment programs give a global assessment of the model, MetaMQAPII was developed to give the local residue assessment in the model. The server uses results from eight other model quality assessment programs (MQAP), (VERIFY3D, PROSA, BALA, ANOLEA, PROVE, PROQRES, REFINER, and TUNE) together with local residue features to assess a model's local residues in relation to similar residues in native structures. MetaMQAPII can highlight single residue deviations in a model without the need for supplementary information. The input to the server is a protein model in a PDB format. The server outputs three files which are then sent by e-mail. The first file confirms the use of all the eight MQAPs for assessment. The second file contains raw data produced by the MQAPs and the MetaMPQAP predicted deviations together with the predicted

GDT_TS for the model. Last the model with its B-factor fields replaced by MetaMQAPs scores is produced in PDB format. This model can be visualised by a macromolecule viewer like PyMol and then coloured according to B-factor values. The resultant coloured model shows a spectrum of colours from blue, which correspond to regions of high accuracy, to red, which corresponds to low accuracy regions [85].

4. ANOLEA (Atomic Non Local Environment Assessment)

ANOLEA is a server for protein structure assessment which calculates the energy of a protein chain. For each atom in the molecule it evaluates all the heavy atoms that are within a diameter of 7 Å, and belonging to an amino acid that is more than 11 residues away in the same amino acid chain or from a different chain. These are the "non-local environment (NLE)" of that particular heavy atom. The energy of each pairwise interaction in the NLE is taken from a distance dependent knowledge based mean force potential derived from a database of 147 non-redundant amino acid sequences with a sequence identity below 25%. The amino acid sequences that make up the database were solved by X-ray crystallography and their resolution is lower than 3 Å. [80]. The input of the server is a PDB file with coordinates for each atom in the molecule together with identity of the protein chain to be assessed. Default setting can be set to user specifications to determine the threshold and the window average to perform the energy profile. The output includes a non-local energy profile plot for each amino acid of the molecule showing low energy (below zero and in green) representing favourable energy regions and high energy (above zero and in red) representing unfavourable energy regions. High energy regions correspond to errors and in some cases interacting protein regions. The ANOLEA server used in this project is part of the SWISS-MODEL work space for model structure assessment (http://swissmodel.expasy.org/workspace).

5. QMEAN6

The QMEAN (Qualitative Model Energy Analysis) server is a gateway to two model quality estimation scoring functions; QMEAN6 which is a composite scoring function for major geometrical aspects of a protein structure and QMEANclust which is a clustering based function [86].

The QMEAN6 function gives an estimate of both the global and local quality of the model; it can therefore be used in selection of models and for prediction of regions that might need further examination in models [86]. It estimates the global quality of the model from a linear combination of six structural properties. These properties include statistical potential of mean force, the solvation potential, the correlation of the predicted and calculated secondary structure and the solvent accessibility. The input model can be a single PDB format or multiple models as

zip or *tar.gz* file(s). The result gives an insight into the contribution of each property to the quality of the models in a table containing the QMEAN score and the values of the six properties. The output models are ranked from 0 to 1 based on the predicted global model reliability. An energy profile plot is also produced with scores ranging from zero (green) indicating stable regions, to 10 (red) indicating unstable regions. The output of QMEAN results also includes a model structure that predicts expected errors on a per residue basis. The model is coloured according to the QMEAN score where blue represent stable or reliable regions and red represent potentially unstable or unreliably modelled regions [87].

The QMEAN used in this project is part of the SWISS-MODEL work space for model structure assessment and it can be found at; (http://swissmodel.expasy.org/qmean.)

2.8.4.5. Model stereochemistry evaluation programs

Protein structures are bound to have both experimental and result interpretation errors. It is therefore necessary to have control measures and be able to assess the quality of the resultant structure or model [88]. In addition to values that measure the overall quality of the structure, (resolution, the R-factor, the R-indices and the ‘free R-value), stereochemistry measures that give information on the different regions of the structure can also be assessed.

1. PROCHECK

The PROCHECK is a suit of five programs that assesses the stereochemistry of a protein structure and also indicated residues in regions that might be wrong and need some attention [88]. The main input to the program is a structure file containing the coordinates. The stereo chemical information, bond lengths and bond angles used in protein assessment are derived from different research projects. The output of the PROCHECK program includes a Ramachandran plot. The Ramachandran plot shows the torsion angles for all residues in the structure with the exception of those at the chain end. PROCHECK produces a Ramachandran plot with a list of residues in the structure. The list of residues gives detailed information on the stereo chemical parameters and shows values that are wrong. A statistical plot of a percentage of residues in allowed regions, generously allowed regions and disallowed regions is produced. The program can be used in assessment of published structures, structures that are being solved and calculated model structures

The PROCHECK program used in this project is part on the SWISS-MODEL work space for model structure assessment (http://swissmodel.expasy.org/workspace).

2.9. Results and Discussion

2.9.1. Sequence retrieval

TbCatB (PDB ID: 3HHI) protease sequence had a 99 % sequence identity to the TbCatB (PDB ID: 4HWY) and a 93% sequence coverage although they are from the same organism. The 93 % sequence coverage was due to 4HWY procathepsin protease sequence being longer because the mature protein was still attached to its propeptide. The HsCatB sequence scored the lowest sequence identity of 50% as a result of being distantly related to the trypanosome species. All the sequences had a significant E-value lower than $10e^{-5}$. A summary of the properties of the retrieved sequences is presented in Table 2.0 below.

2.9.2. MSA and structural analysis

Although PROMALS3D was used for homology modelling alignments (Appendix 1A), MAFFT program was used in the phylogenetic analysis alignment and in sequence comparison. The MSA (Figure 2.1 & 2.3) were viewed in Jalview [89] and they were coloured by percent identity. The mature form of cathepsin B polypeptide share the same sequence and structural features with other clan CA cycteine proteases and it is folded into two domains that form a V-shaped active site cleft that contains the conserved catalytic triad Cys122(29), His282(199) and Asn302(219) (TbCatB numbering with HsCatB numbering in brackets), [23], [26], [29], [37].

Accession number	Organism	Abbreviation	% sequence identity	% coverage	E-value
3HHI	T. brucei	TbCatB	99	93	0.0
3CBJ	H. sapien	HsCatB	50	*	1e-53
gb|ABY78821.1	T. congolense	TcCatB	64	99	1e-157
emb|CCC48215.1	T. vivax	TvCatB	65	91	3e-145
ref|XP_816569.1	T. cruzi	TcrCatB	58	98	7e-136

Table 2.0: Listed are retrieved TbCatB protease sequence homologs**.** Trypanosome homologs were obtained from NCBI using a Blast search while the human homolog was obtained using HHpred. Percent coverage and E-values are based on TbCatB structure 4HWY as the query sequence. * Similarity=1.049 (HHpred has no % coverage values)

For this project, chain A of TbCaB protease sequence was used. This chain is made up of residues 78-335 which makes 257 amino acids of mature TbCatB [23]. Since there is no standard mature domain numbering, for clarity the TbCatB and other trypanosomatids sequences shall be numbered according to TbCatB numbering with HsCatB numbering in brackets. The crystal structure of the HsCatB protease sequence used in MSA and structural analysis was determined for an inactive recombinant enzyme variant. Mutations were introduced to change the catalytic Cys29 to Ala29 and the occluding loop Ser115 to Ala115 and H110 to Ala110 to prevent auto catalytic degradation during crystallization [37]. The sequence of the HsCatB starts with an Asp60p residue which is part of the propetide attached to the mature enzyme sequence. The mature domain of HsCatB consists of residues Leu1 to Asp254 [26]. The L-domain of HsCatB is made up of residues from the amino terminal end (excluding the first 10 residues) up to Tyr148 (cathepsin B numbering), and by about the last four residues of the carboxy terminal while the R-domain is made up of the first 10 residues of the amino terminal and the carboxy terminal (excluding the last four residues) of the polypeptide chain [25]. HsCatB protease contains an occluding loop made up of 22 residues from Ile105 to Pro126 in the L domain of the enzyme while the occluding loop of TbCatB is made up of residues from

Phe189 to Pro213 and has three extra residues [8], [23], [37]. The structure and function of TbCatB and HsCatB in the MSA shall be used to predict the structure and function of the aligned homologs. The focus of this alignment analysis shall be on regions that contribute to the protein and inhibitor interactions; the occluding loop, the active site residues and the catalytic residues.

2.9.3. Inserts

The MSA (Figure 2.1) shows that there is conservation of residues among the aligned homologs. There is also notable difference in the size of the sequences caused by residue insertions which have resulted in structural differences. Most notably the trypanosomatids have a longer N-terminal than the HsCatB, a feature which was also observed by Kerr et al. The L-domain of the HsCatB has two insertions (Val50 and Met66) which are absent in the trypanosomatids cathepsin B proteases. The region from Ser90 to Arg101 forms a hairpin loop [25] in the HsCatB protease consisting of residues Gly91, Gly92, Leu93, Tyr94, Glu94, Ser96, His97, and Val98 that are not available in trypanosomatids cathepsin B. The occluding loop of TbCatB has three extra residues, Lys175, Tyr80 and Phe86 (TbCatB numbering) that are absent in HsCatB protease. The HsCatB protease also has three residues Lys141, Glu142 and Asp143 in the L domain hairpin loop which are not in any of the trypanosomatids proteases. There is also Val153 insert in HsCatB which only corresponds to a Leu residue insert in *T.congolense* cathepsin B protease.

2.9.4 The occluding loop

The occluding loop region is made up of residues from Phe189(Ile105) to Pro213(Pro126) [8], [23], [26], [37]. From the MSA, the occluding loop of TbCatB has Lys197, Try202 and Phe208 residues which are not available in the HsCatB protease. The Lys197 residue is unique to TbCatB while the Try202 corresponds to Lys202, Asn202 and Leu202 in *T. congolense, T.vivax* and *T. cruzi* Cathepsin B like proteases respectively. The Phe208 residue of TbCatB respectively corresponds to Try208 and Met208 residues in *T. congolense* and *T. vivax* cathepsin B like proteases. The His194(110) and His-195(111) which contribute to the carboxypeptidase activity [23], [8] of cathepsin B proteases is highly conserved in all the sequences except in *T. vivax* cathepsib B like protease where His195(111) is substituted by a glycine residue.

In HsCatB, a circular structure is formed when the occluding loop chain crosses over itself at Cys108 and Cys119 to form a disulphide bridge [25], [37]. The region leading to and from the occluding loop circle is flanked by conserved regions consisting of Pro106, Pro107 and Cys108 leading to the circle and Pro117, Pro118 and Cys119 leading from the circle. These Pro-Pro-Cys

regions together with hydrogen bonds close to the disulphide bridge are suspected to contribute to the stability of the occluding loop [25], [37]. Between residues Ser206(Thr120) to Phe210(Gly123), the occluding loop of TbCatB contains a "FNFD" motif which corresponds to "GEGD" motif in HsCatB. In HsCatB the Gly121 and Gly123 on both sides of Glu122 makes this region more flexible than the corresponding Phe208 and Phe210 residues which come together with Phe189 to create a more stable S1' opening in TbCatB. This difference may be may exploited to design more specific inhibitors for the TbCatB enzyme [23].

2.9.5. Active site residues

The active site of cathepsin B cysteine proteases is made of a catalytic triad consisting of a cysteine residue that acts as a nucleophile and a histidine residue that acts as a general base and an asparagine residue that helps in the orientation of the histidine residue and also neutralizes the histidine charge during the intermediate state of hydrolysis of peptides [21]. The active site residues that make up this catalytic triad are Cys122(29) from the L-domain, His282(199) and Asn302(219) from the R-domain and it is situated in a cleft between the two domains that make up the protease [29], [23], [37]. The Asn302(219) side chain amide group is supported by N-H-π interactions from Trp304(221) and Trp308(225) [37]. Although they are not in the active site cleft, the His194(110) and His195(111) residues in the occluding loop are important for the endopeptidase activity of the enzyme since they are used for docking the C-terminal carboxylic group of the peptide substrate [22] by providing a positively charged anchor for the substrate carboxylic group [26].

1. S2 subsite

The S3 and S4 binding site of cathepsin B cysteine proteases are described as not real subsites but areas in which substrate residues find their favourable binding position [22]. The 3D complexes of TbCatB and epoxysuccinyl-based inhibitors CA030, CA074 and NS134 that were used during subsite determination revealed three subsites, S2, S1 and S1' in which the substrate binds in an extended conformation[27], [22]. In HsCatB protease the S2 subsite is made up of residues Asp166(Tyr75), Pro167(76), Ala256(173), and Ala283(200) [25], and it determines specificity of the enzyme. In vertebrates, cathepsin B family members were defined as having an acidic residue at the bottom of the S2 subsite to accommodate basic P2 residues [23]. HsCatB protease has Glu245 at this position while TbCatB has Gly328. The corresponding position is occupied by Thr328, Gln328 and Ser328 in *T. congolense, T. cruzi and T. vivax* cathepsin B like proteases respectively.

The area around the S2 subsite of TbCatB protease is occupied by Asp166(Tyr75), Asp168(Ala77), Asp258(Ser175), as well as Asp327(Ser244) and Gly328(Glu245) at the bottom of the subsite [23]. The HsCatB Tyr75 residue is conserved in *T. cruzi and T. vivax* cathepsin B-like proteases, but corresponds to Ala166 residue in *T. congolense* cathepsin B-like protease. Asp168(Ala77) residues correspond to Glu168 residue in *T. cruzi* and to Asp168 residues in *T. congolense* and *T. vivax* homologs. TbCatB and HsCatB residues Asp258(Ser175) corresponds to Gly258, Thr258 and Ser258 residues in *T. congolense, T. vivax* and *T. cruzi* protease homologs respectively. At the bottom of the S2 subsite, the Ser244 residue is conserved in *T. congolense* and *T. vivax* while Gly244 occupies the corresponding residue in *T. cruzi.*

2. S1 subsite

Interactions of peptidy substrate with cathepsin B-like enzyme subsites S1, S1' and S2' has not been fully predicted. The substrate P1 carbonyl group is predicted to bind to Cys122(29) and Gln116(23) in HsCatB [25], [27]. The MSA shows that both the catalytic Cys and the Gln residues are conserved in all the sequences.

3. S1' subsite

The S1' subsite is predicted to be around Val259(176), Phe263(180), Leu264(181), His282(199), and Trp304(221) [25], [27]. These residues are conserved in all the cathepsin B proteases used in this project.

4. S2' subsite

The S2' subsite is a shallow hydrophobic depression around the side chains of His194(110) and His195(111) [25], [27] and residues Trp308(225) and Phe263(180)[26]. These residues are highly conserved in all the sequences.

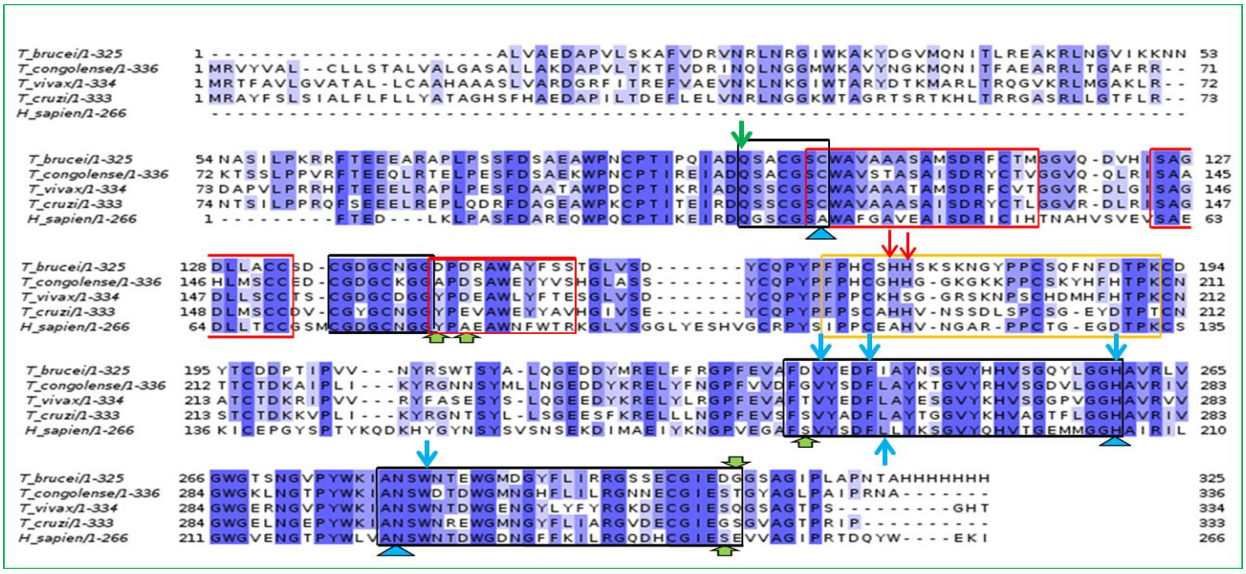


Figure 2.1: MSA of TbCatB protease homologs as predicted by MAFFT. Amino acids are highlighted according to percent agreement, > 80% in mid blue, > 60% in Light blue, > 40 % in light grey, and <= 40 % in white.

2.9.6. Phylogenetic analysis

To establish the evolutionary relationship of *T. brucei* cathepsin B-like protease, a phylogenetic tree (Figure 2.4) was calculated. The phylogenetic tree shows cathepsin B protease forming an out group and the Trypanosomes clustering together in one branch. This indicates that there are sequence differences between the HsCatB proteases and the trypanosome cathepsin B-like proteases. This observation is consistent with the sequence identity of 48.06%, 49.62%, 50.00% and 49.23% (Table 2.2) observed between the *H. sapiens'* and *T. brucei*, *T. congolense*, *T. cruzi* and *T. vivax* respectively; while a sequence identity higher than 63% was observed among the trypanosomes. These differences can be exploited to identify potential inhibitors that are specific for trypanosomal cathepsin B proteases. Among the trypanosome taxa, *T, cruzi* forms a branch of its own, indication early evolutionary divergence from the other trypanosomes.

Protein ID	S2 Subsite	S1 Subsite	S1' Subsite	S2' Subsite
Human cathepsin B	Y75-P76-A173-A200-E245	C29-Q23	H199-V176-F180-L181-W221	H110-H111-Phe180-Trp225
T. brucei cathepsin B	D166-D168-D258-D327-G328	C122-Q116	H282-V259-F263-L264-W304	H194-H195-Phe263-Trp308
T. congolense cathepsin B	A166-D168-G258-S244-T328	C122-Q116	H282-V259-F263-L264-W304	H194-H195-Phe263-Trp308
T. cruzi cathepsin B	Y166- E168-S258-G327-Q328	C122-Q116	H282-V259-F263-L264-W304	H194-H195-Phe263-Trp308
T. vivax cathepsin B	Y166- A168-T258-S327-S328	C122-Q116	H282-V259-F263-L264-W304	H194-H195-Phe263-Trp308

Table 2.1: Shown are the Subsite residues for the S2, S1, S1'and S2' subsites of the homologs. For clarity the trypanosomatidae homologs were numbered according to TbCatB numbering.

2.9.7. Homology Modelling

Homology models were calculated for *T. congolense, T. cruzi* and *T. vivax* cathepsin B-like proteases. The best models were selected based on their model evaluation results from different model evaluation methods. The following steps demonstrate the procedure followed during homology modelling.

2.9.8. Template selection

As previously stated, template selection is a very important step in homology modelling since the template (s) determine (s) the quality and accuracy of the 3D model. Two protease crystal structures were identified and used as templates for calculation of the models.

These are TbCatB, (PDB ID: 3HHI) and the HsCatB protease, (PDB ID: 3CBJ). The identified 3HHI template is reported as the first structure of TbCatB in complex with the cathepsin B selective inhibitor CA074 [23]. This structure was determined by X-ray Crystallography at

resolution of 1.60Å. The HsCatB protease structure is in complex with the *T. cruzi* inhibitor chagasin. This structure was solved by X-ray diffraction at 1.80 Å [37].

It was important to use templates in complex with inhibitors because it ensured that the resulting model will be calculated with the occluding loop in an open conformation. The occluding loop of TbCatB proteases can occur in both open and closed conformation. When in open conformation, it allows access of substrates to the S2 subsite which determines specificity of the protease.

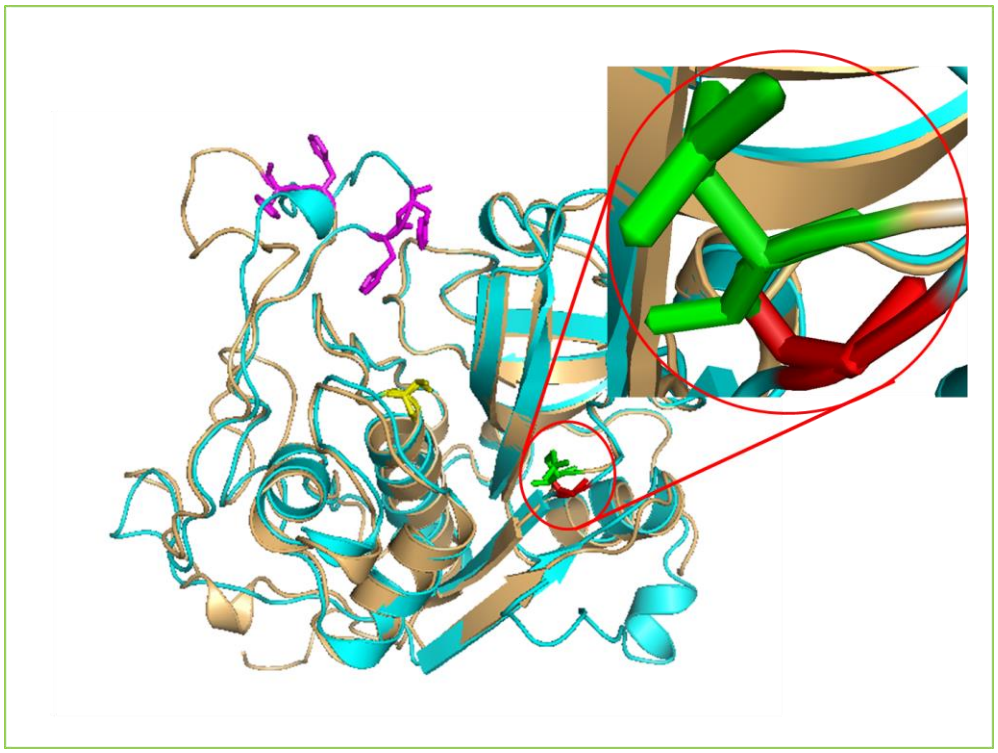


Figure 2.2: Showing the superimposed TbCatB (Cyan blue) and HsCatB (light orange) protease structures. The active site Cys122(29) residues are shown as yellow sticks and the His194(110) and His195(111) are shown in purple. Residues Gly328(Glu245) are shown in red and green respectively. Gly residue is much smaller than the Glu residue, creating a deeper S2 pocket.

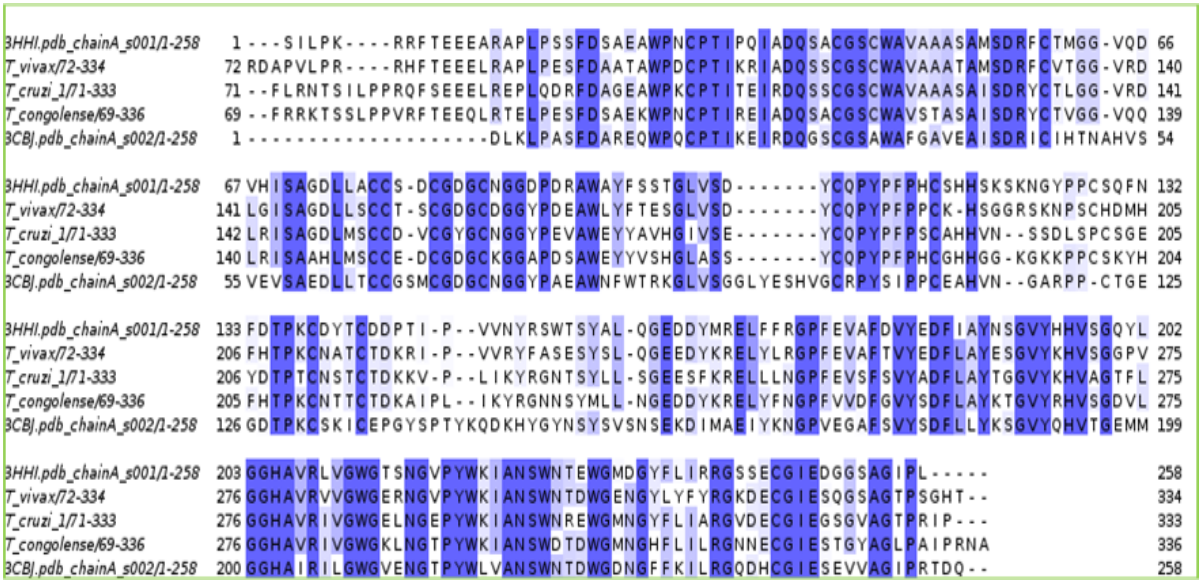


Figure 2.3: MSA of TbCatB protease homologs as predicted by PROMALS3D**.** Amino acids are highlighted according to percent agreement, > 80% in mid blue, > 60% in Light blue, > 40 % in light grey, and <= 40 % in white**.** PIR alignments were made from this MSA for homology modelling.

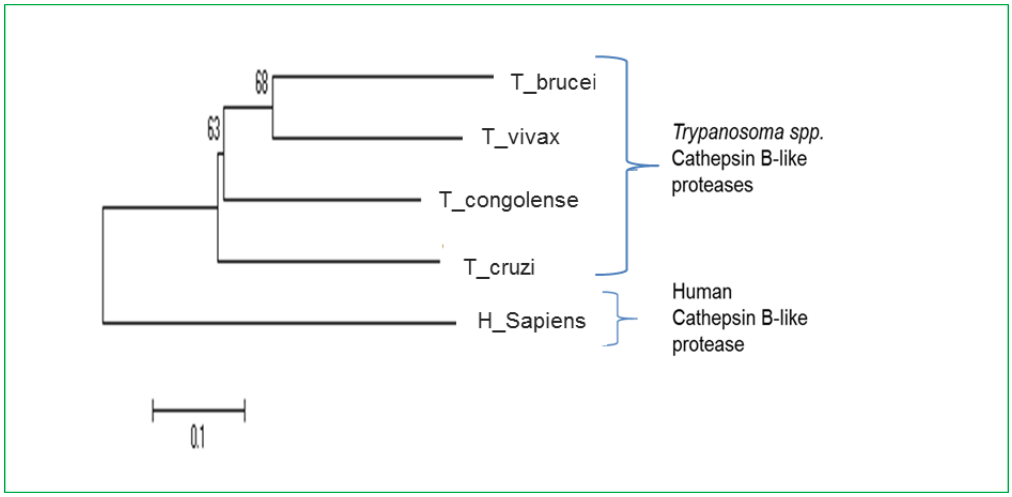


Figure 2.4 Phylogenetic analyses of TbCatB homologs. The tree shows distinct clustering of trypanosomatidae family homologs from the HsCatB protease homolog.

When the occluding loop is in closed conformation, it restricts access to the S2 sub site. Since the models are going to be used for docking studies, they have to be modelled with the occluding loop in open conformation to allow ligand access to the active site.

In addition to their good resolution, the templates were chosen based on their high sequence identity to the targets. The quality of the templates was accessed using PROCHECK, ANOLEA and QMEAN.

Template	Resolution [Å]	Target	Sequence identity (%)
3HHI	1.60	T. congolense	66.41
		T. cruzi	63.57
		T.vivax	68.22
3CHJ	1.80	T. congolense	49.62
		T. cruzi	50.00
		T.vivax	49.23

Table 2.2: Listed are the template and target sequence identities used in homology modelling of TbCatB homologs. The sequence identity of the two templates is 48.06 %.

The results for templates and targets sequence identity (Table 2.2) show that all the sequences are homologous. A sequence identity of 66.41%, 63.57% and 68.22% was obtained between the TbCatB (PDB ID: 3HHI) template and the targets, *T. congolense*, *T. cruzi* and *T. vivax* respectively while lower sequence identities of 49.62 %, 50.00% and 49.23% were obtained between the HsCatB (PDB ID: 3CBJ) template and the respective targets. Good quality models can be obtained from these sequences since their percent identity is higher than the 30% required for making a good quality model. Models build from these sequences can also be used to assess the protease potential as a drug target. Sequence identities of 25 to 50% identities can be used to test if a protease can be used as a drug target [41].

2.9.9. Template evaluation and validation

The chosen templates quality was evaluated using model quality evaluation tools. This was done to make sure that the template structures had no unstable regions that would later be inherited by the models. Model evaluation programs PROCHECK, ANOLEA and QMEAN were used in the evaluation of the template structures. The results are shown in Figures 2.6 to 2.8. Table 2.3 shows PROCHECK results as obtained from the Ramachandran plot statistics for the two templates used and the homology models.

Structure (protease)	Residues in most favoured regions [A,B,L]	Residues in additional favoured regions [a,b,l,p]	Residues in generously allowed regions [~a, ~b, ~l, ~p]	Residues in disallowed regions
TbCatB (3HHI template)	88.0 %	11.5 %	0.5 %	0.0 %
HsCatB (3CBJ template)	85.3 %	14.3 %	0.3 %	0.0 %

Table 2.3: Shown are the Ramachandran plot statistics as produced by PROCHECK for each of the templates used in homology modelling.

The superimposed structures of TbCatB (PDB ID: 3HHI) and HsCatB B (PDB ID: 3CBJ) templates (Figure 2.5) show that the occluding loop of the two templates is not open to the same degree. The HsCatB crystal structure (3CBJ) is in complex with the *T. cruzi* inhibitor chagasin while TbCatB (3HHI) is in complex with the smaller cysteine protease inhibitor CA074. The chagasin inhibitor has therefore opened the occluding loop of HsCatB much wider than the smaller CA074 inhibitor has done for TbCatB. The occluding loop can move and open to different degrees depending on the size of the bound substrate [37]. Since the occluding loop of the models should be in the open conformation for docking studies, models calculated from these two templates are expected to adopt the orientation of the respective template and their occluding loop should be in the same conformation as the respective templates. We are still to find out the occluding loop conformation that will be adopted by models calculated using a combination of the templates. Stereochemistry results for the two templates (Figure 2.6) show all the residues of the templates are in favoured regions. The ANOLEA validation and QMEAN energy profile also confirm that the templates are stable and of acceptable quality. ANOLEA and QMEAN results for HsCaB (PDB ID: 3CBJ) template (Figure 2.8) indicate the region between Val112 and Pro117 as having unfavorable energy and unstable respectively. This can be of no major concern as long as it is not inherited by the models, especially since it is close to the His110 and His111 residues that are participant in exopeptidase activity.

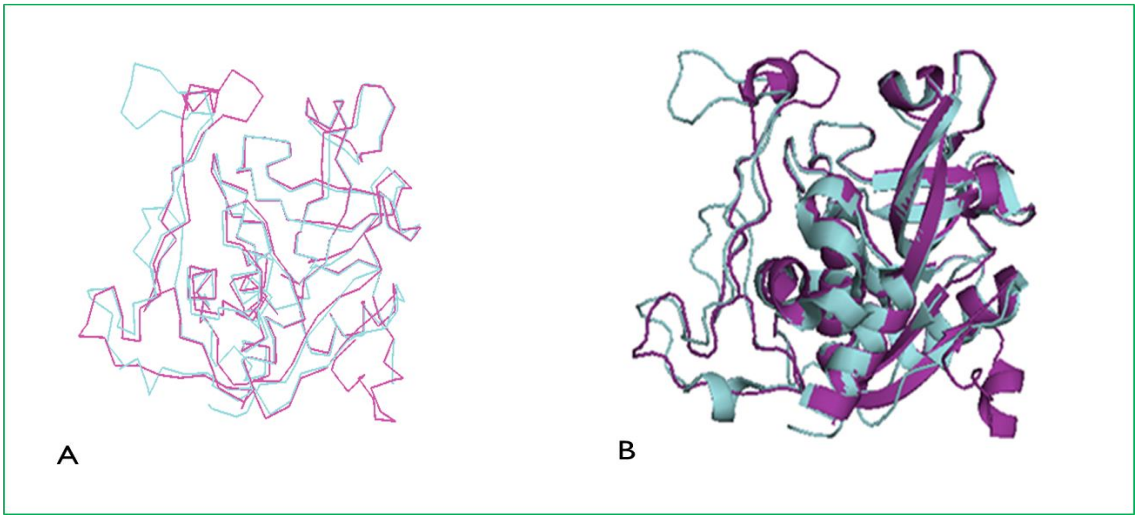


Figure 2.5: Superimposed (A) ribbon and (B) cartoon template structures of PDB ID: 3HHI (purple) and 3CBJ (blue). The occluding loop of 3CBJ is open wider than that of 3HHI due to the large chagasin inhibitor.

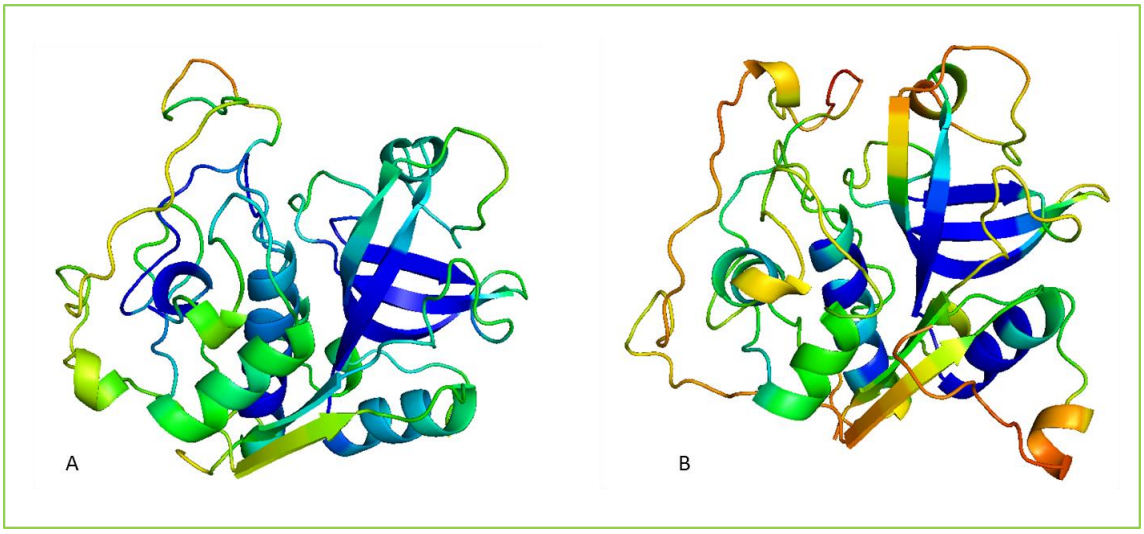


Figure 2.6: QMEAN energy profile for (A) 3CBJ and (B) 3HHI template. The QMEAN scores are colour coded from blue (stable) to red (unstable) regions.

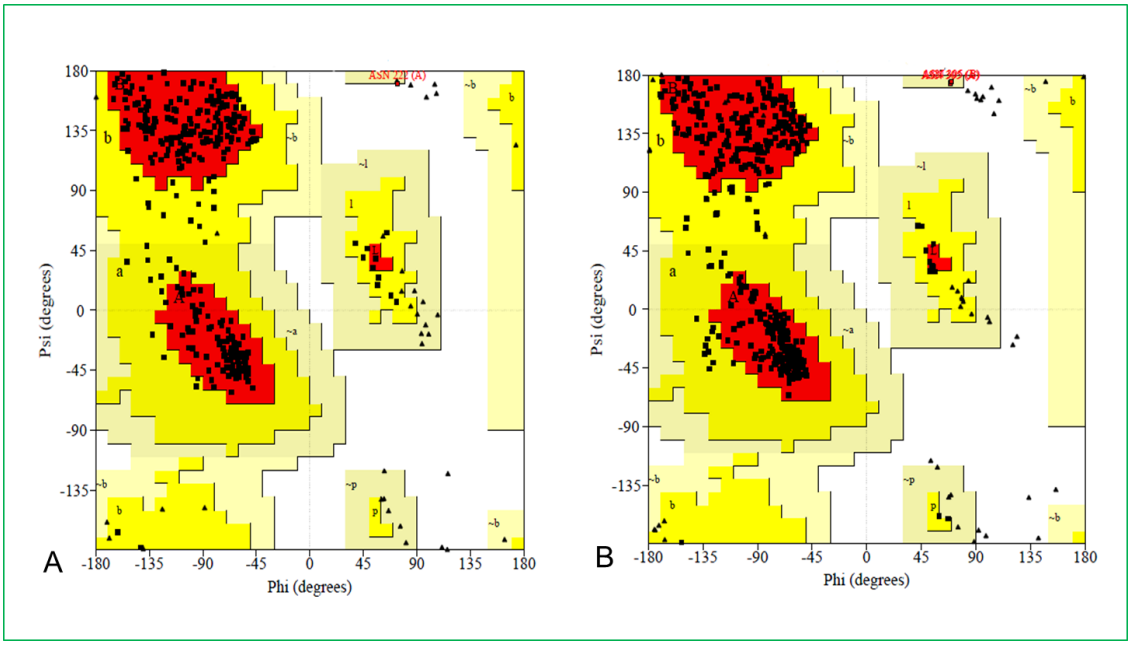


Figure 2.7: PROCHECK analysis for (A) HsCatB (PDB ID: 3CBJ), and (B) TbCatB (PDB ID: 3HHI) templates. The PROCHECK plot statistics shows 100 % of residues are in favoured regions for both the templates.

2.9.10. Homology modelling results and discussions

All the 3D models were calculated using MODELLER version 9.10 by customising MODELLER scripts to meet the project needs. Two template and inhibitor complexes were identified and they were both used to make the template and target sequence alignment from which the models were calculated. For each model a hundred models were made, out of which the best three models were selected based on their low DOPE Z score. The models were then validated using MetaMQAPII, PROCHECK, ANOLEA and QMEAN model evaluation programs. After model validation, the model that performed best was chosen as the cathepsin B-like protease model for docking experiments.

2.9.11. Model validation

The best three models were selected using a python script which selected the models based on their DOPE Z scores. A DOPE Z score lower than -1 score is preferred for models since it means that the model has a low energy and therefore it is in its more stable condition. The cut off point for accepting a model is set at -0.5 after which models with a higher DOPE Z score are likely to be poor models. To determine the similarity between the homology models and the template crystal structures the Root Mean Square Diviation (RMSD) was measured by superposition of the backbone atoms of the homology models with those of the crystal structures in PyMol [83]. An arbitrary RMSD cut off value of 3Å is usually used by most researchers [84].

Table 2.4 and 2.5 respectively shows the DOPE Z scores and the RMSD values of the models. Lower RMSD values were obtained for models when superimposed to the 3HHI template than the 3CBJ template. The models are therefore more similar to TbCatB structure that to HsCatB structure. These results are in agreement with the phylogenetic and percent identity results.

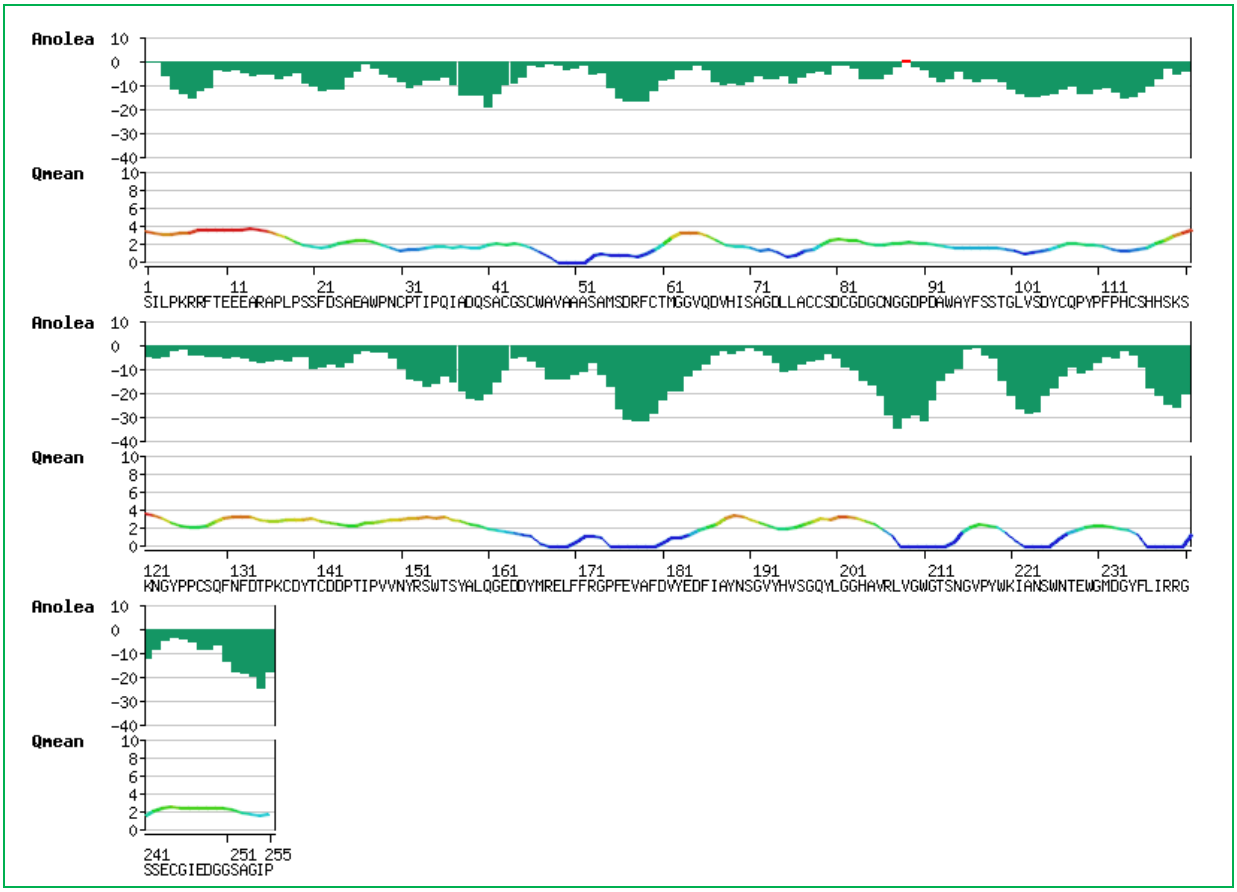


Figure 2.8: ANOLEA evaluation and QMEAN6 energy profile for PDB ID: 3HHI. ANOLEA score values greater than zero correspond to high energy regions / erroneous possible interactions. QMEAN scores are color coded from green (stable) to red (unstable) regions. QMEAN scores above 2 represent high energy (unstable) region

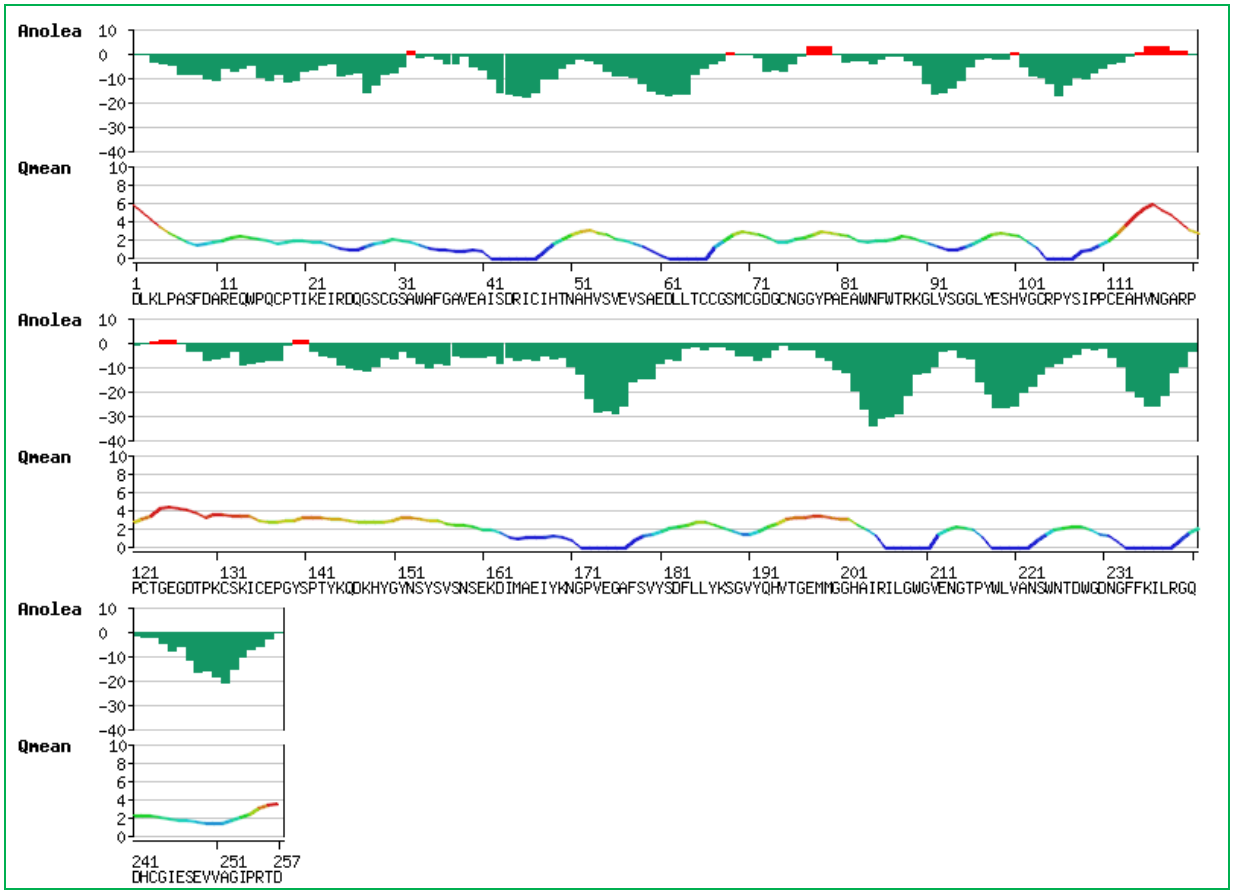


Figure 2.9: ANOLEA evaluation and QMEAN6 energy profile for PDB ID: 3CBJ. ANOLEA score values greater than zero correspond to high energy regions / erroneous possible interactions. QMEAN scores are color coded from green (stable) to red (unstable) regions. QMEAN scores above 2 represent high energy (unstable) region.

Figure 2.10 shows that the models adopted the 3D orientation of the templates from which they were calculated. In Figure 2.11, the occluding loop of the models that were calculated form a combination of the templates, adopted the 3HHI occluding loop template orientation. This might be because the occluding loop is less strained in this position which is nearer the closed conformation and therefore it is a more favourable position. The occluding loop of models calculated from the 3CBJ template is open wider than that of models calculated from the 3HHI template and from a combination of the templates. The active site of the 3CBJ template models is therefore more exposed than that of the other template models, a feature which might make it easier for docking of large molecule inhibitors.

Structures	3HHI template	3CBJ template	3HHI_3CBJ template
TcCatB	- 1.12	- 0.90	- 1.01
TcrCatB	- 0.59	- 0.30	- 0.55
TvCatB	- 1.20	- 1.10	- 1. 39

Table 2.4: Listed are the DOPE Z-scores of the models. The DOPE Z-scores of the templates were – 1.44 and – 1.14 for 3CBJ and 3HHI respectively.

Structures	3HHI template	3CBJ template	3HHI_3CBJ template
TcCatB	0.173	0.240	0.281_0.536
TcrCatB	0.262	0.312	0.303_0.537
TvCatB	0.196	0.251	0.262_0.578

Table 2.5: Listed are the RMSD values showing the similarity between the homology models and the templates. The RMSD of the templates was found to be 0.668.

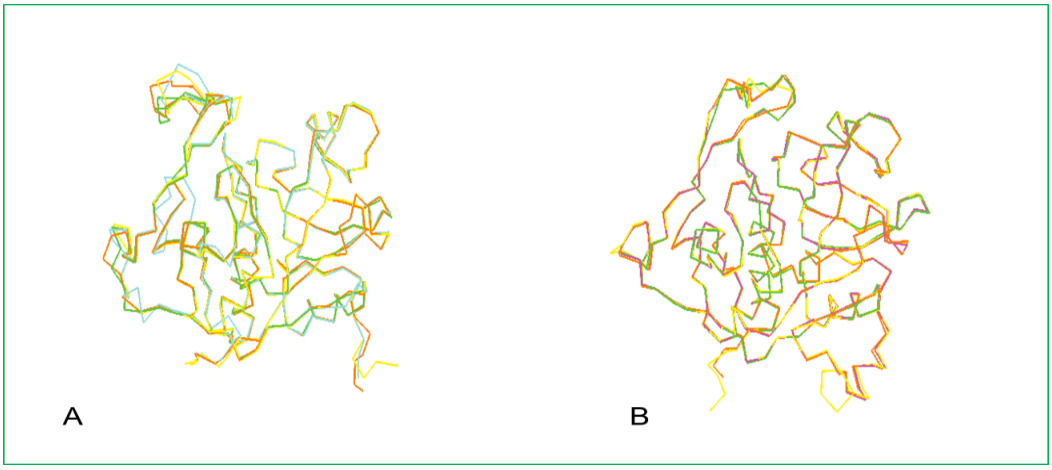


Figure 2.10: Superimposed ribbon structures of templet PDB ID: (A) 3CBJ (blue) and (B) 3HHI (purple) with models for *T. congolense* (yellow), *T. cruzi* (green) and *T. vivax* (orange) calculated from each template.

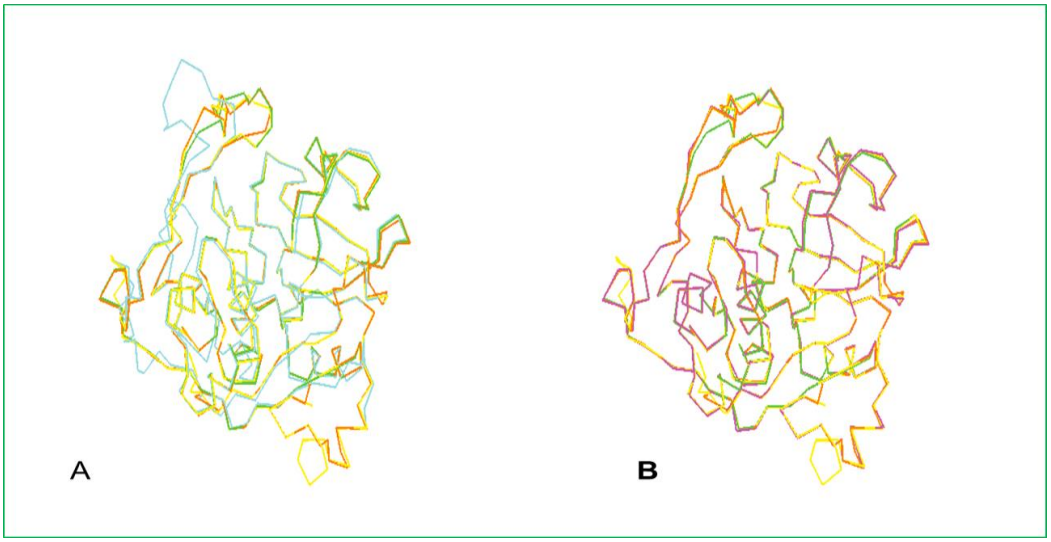


Figure 2.11: Superimposed ribbon structures of templet PDB ID: (A) 3CBJ (blue) and (B) 3HHI (purple) with models for *T. congolense* (yellow), *T. cruzi* (green) and *T. vivax* (orange) calculated from a combination of the templates.

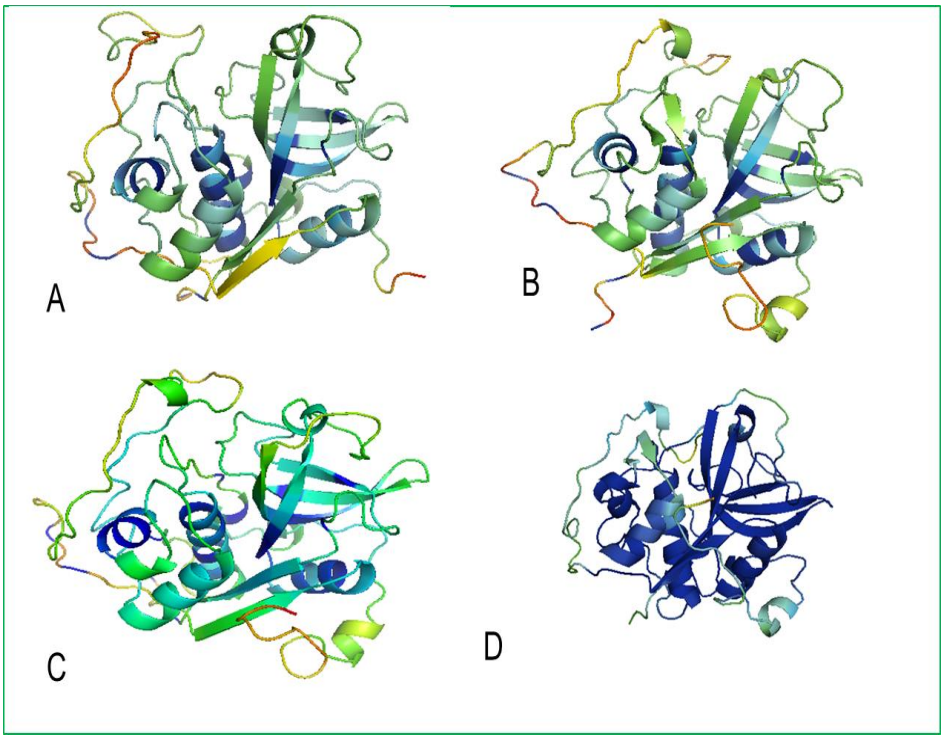


Figure 2.12: QMEAN energy profile for *T. congolense* cat B-like models built from (A) 3CBJ, (B) 3HHI, (C) Double template and (D) MetaMQAPII for double template models. The QMEAN/MetaMQAPII scores are color coded from blue (stable) to red (unstable) regions.

The MetaMQAPII server went down before results for single template models and template structures could be acquired. MetaMQAPII energy profiles were acquired only for the double template models. These results are shown alongside QMEAN energy profiles for all the homology models.

Figure 2.12 shows the QMEAN energy profile showing per residue predicted errors for *T. congolense* models. The models are coloured according to the QMEAN score where blue represents stable regions and red represents unstable regions [87]. Model (A); calculated from the HsCatB (PDB ID 3CBJ) template has an unstable region in the occluding loop region.

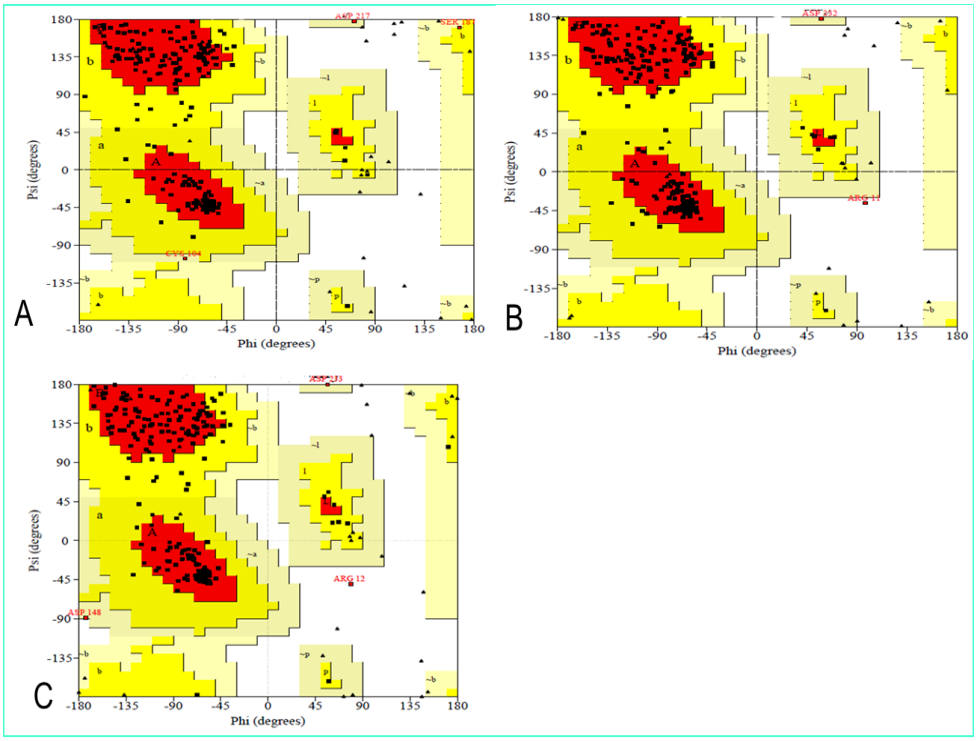


Figure 2.13: PROCHECK analysis for *T. congolense* cat B-like models built from (A) 3CBJ, (B) 3HHI and (C) Double template(s). The PROCHECK plot statistics shows 100%, 99.5% and 99.6% of residues are in favoured regions 3CBJ, 3HHI and double template models respectively.

This is of major concern since the occluding loop is participant in substrate binding. Although model (B), calculated from the *T. brucei* (3HHI) template has some unstable region at the beginning of the occluding loop, the model can still be reliable since the affected area is not directly involved in substrate binding. Out of these three models, the model that scored the best according QMEAN profile is model (C), which was calculated from a combination of the two templates. Stereochemistry results obtained for *T. congolense* cat B models (Figure 2.13) show that all the residues are in allowed regions for the model calculated using template 3CBJ. The model calculated using 3HHI template has Arg12 residue in a disallowed region while the model calculated using both templates has Arg12 and Asp148 residues in disallowed regions. The models calculated from these templates therefore respectively have 99.5% and 99.6% of residues in favoured regions. ANOLEA evaluation and QMEAN6 energy profiles (figures 2.14-2.16) rates all the models as good models.

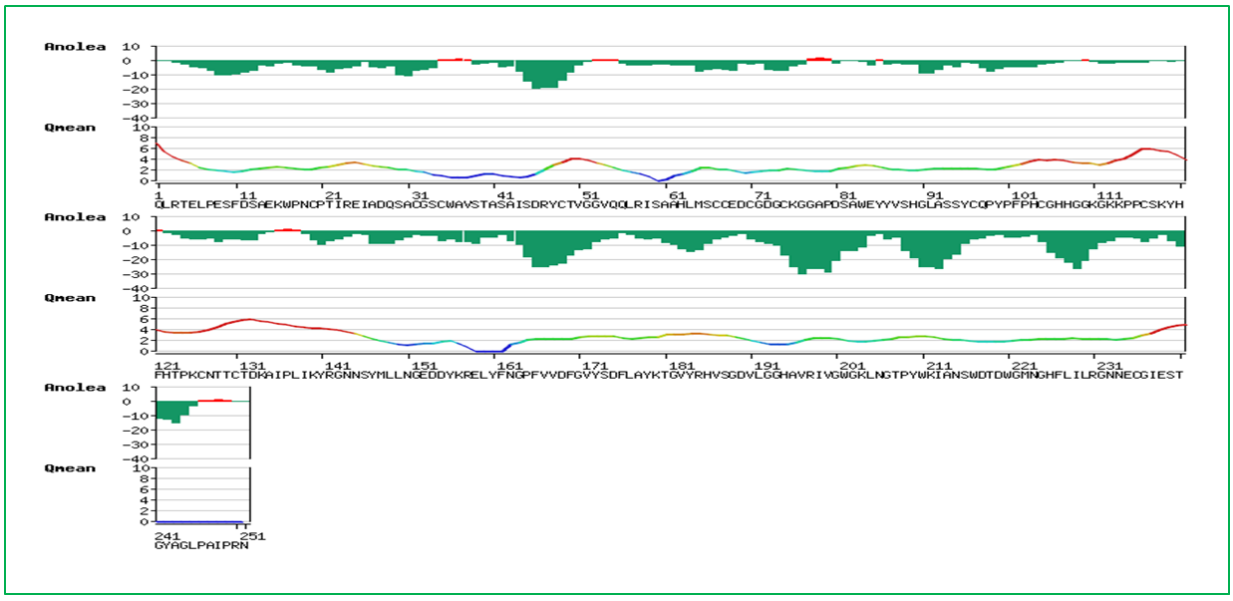


Figure 2.14: ANOLEA evaluation and QMEAN6 energy profile for *T. congolense* cat B-like protease model built from 3CBJ template. ANOLEA score values greater than zero correspond to high energy regions / erroneous possible interactions. QMEAN scores are color coded from green (stable) to red (unstable) regions. QMEAN scores above 2 represent high energy (unstable) region

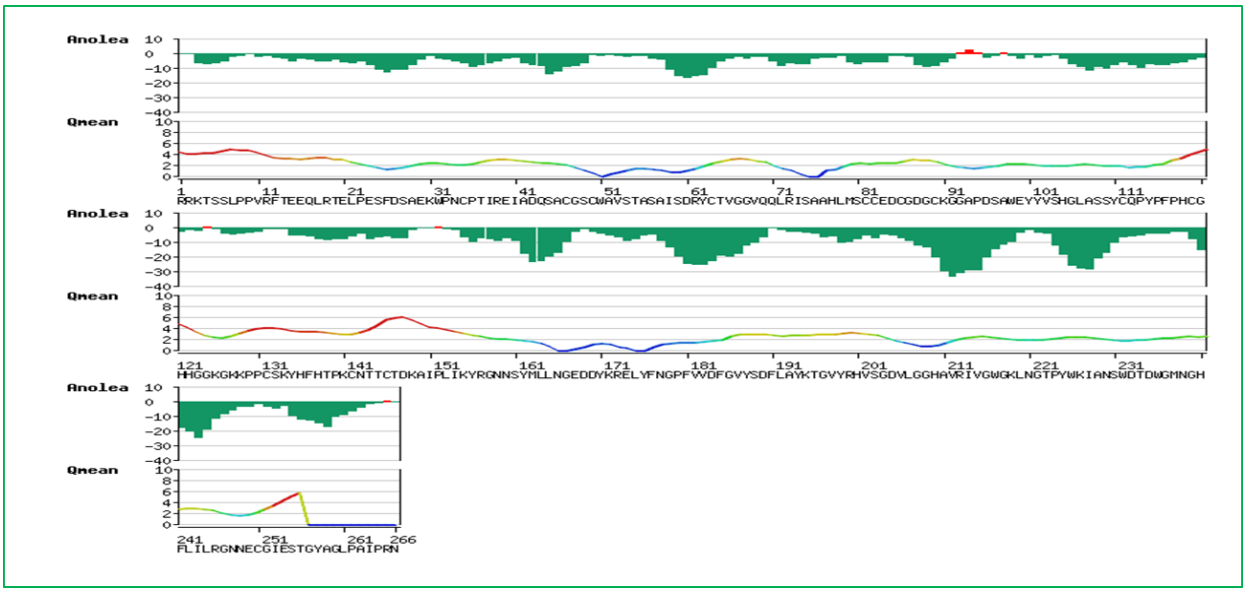


Figure 2.15: ANOLEA evaluation and QMEAN6 energy profile for *T. congolense* cat B-like protease model built from 3HHI template. ANOLEA score values greater than zero correspond to high energy regions / erroneous possible interactions. QMEAN scores are color coded from green (stable) to red (unstable) regions. QMEAN scores above 2 represent high energy (unstable) region.

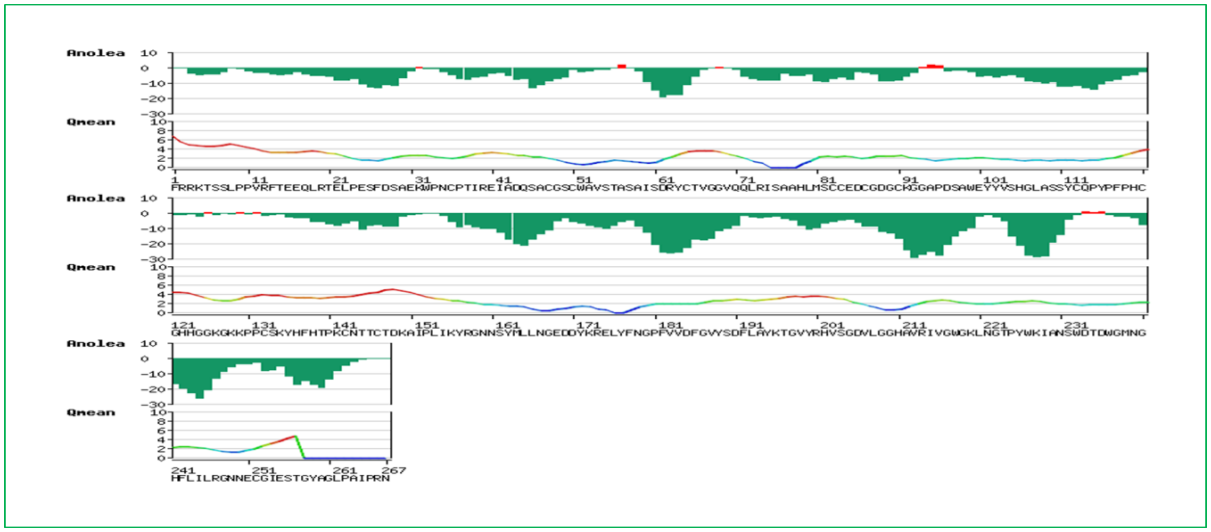


Figure 2.16: ANOLEA evaluation and QMEAN6 energy profile for *T. congolense* catB-like model built from a combination of 3HHI and 3CBJ templates. ANOLEA score values greater than zero correspond to high energy regions / erroneous possible interactions. QMEAN scores are color coded from green (stable) to red (unstable) regions. QMEAN scores above 2 represent high energy (unstable) region.

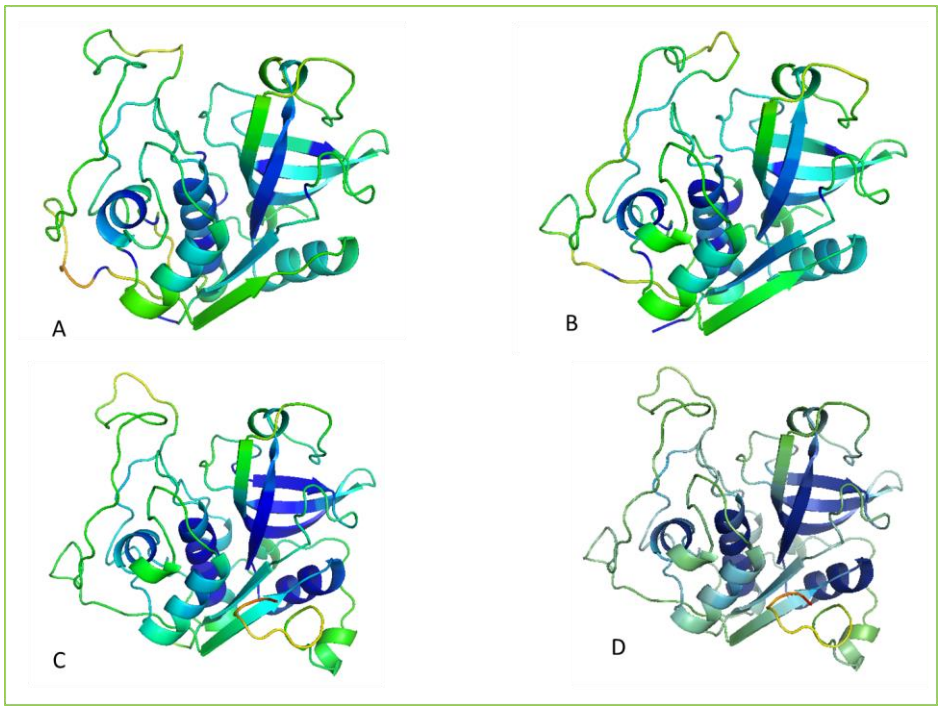


Figure 2.17: QMEAN energy profile for *T. cruzi* cat B-like models built from (A) 3CBJ, (B) 3HHI, (C) Double template and (D) MetaMQAPII for double template models. The QMEAN/MetaMQAPII scores are colour coded from blue (stable) to red (unstable) regions.

Model structures showing the QMEAN profile of *T. cruzi* models are shown in Figure 2.17. All the calculated models are stable and of high quality.

In Figure 2.18, we can see the stereochemistry results of the models calculated for *T. cruzi*. These results show Asn206 and Asp59 residues in disallowed regions leaving 98.8% of residues in allowed regions for the model calculated with the 3CBJ template (Figure 2.18. A). In the model calculated with both the templates (Figure 2.18. C), Asp84 residue in observed to be in disallowed area. 100% of residues were modelled in favoured regions for the model calculated from 3HHI template.

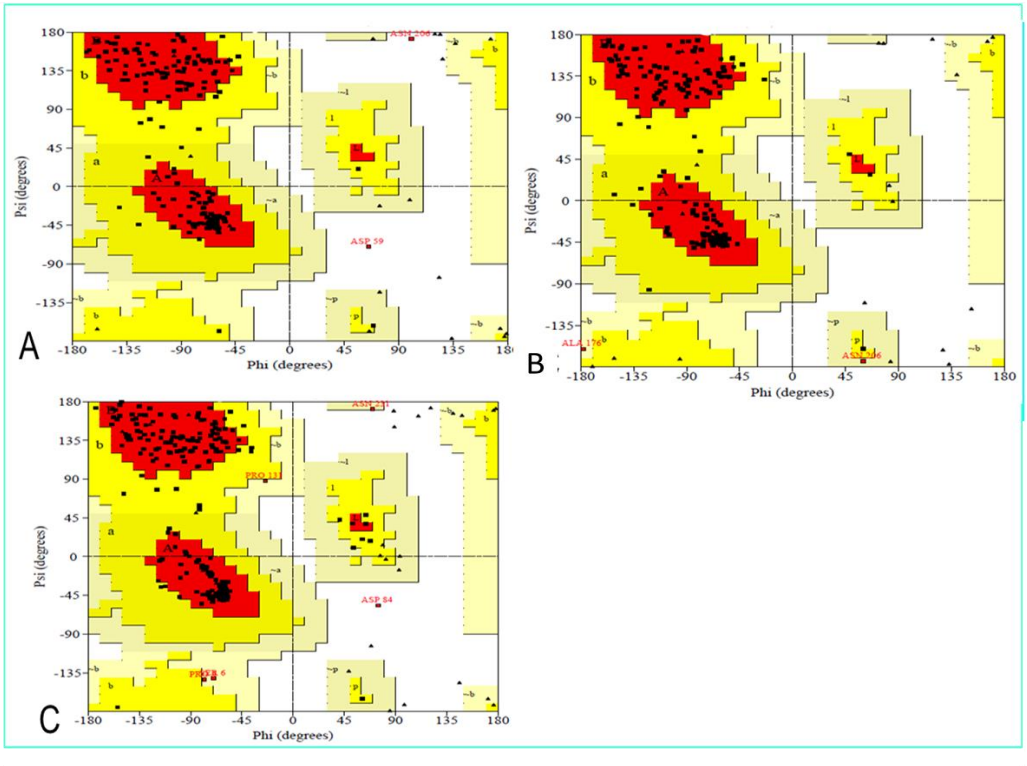


Figure 2.18: PROCHECK analysis for *T. cruzi* cat B-like models built from (A) 3CBJ, (B) 3HHI and (C) double template models. The PROCHECK plot statistics shows 98.8 %, 100 % and 99.6 % of residues are in favoured regions for 3CBJ, 3HHI and double template models respectively.

ANOLEA evaluation and QMEAN6 energy profile (Figures 2.19-2.11) are picking some unstable regions in all the models although not necessarily at the same regions. These errors were not highlighted by other model quality and evaluation programs so this models can be used for further applications.

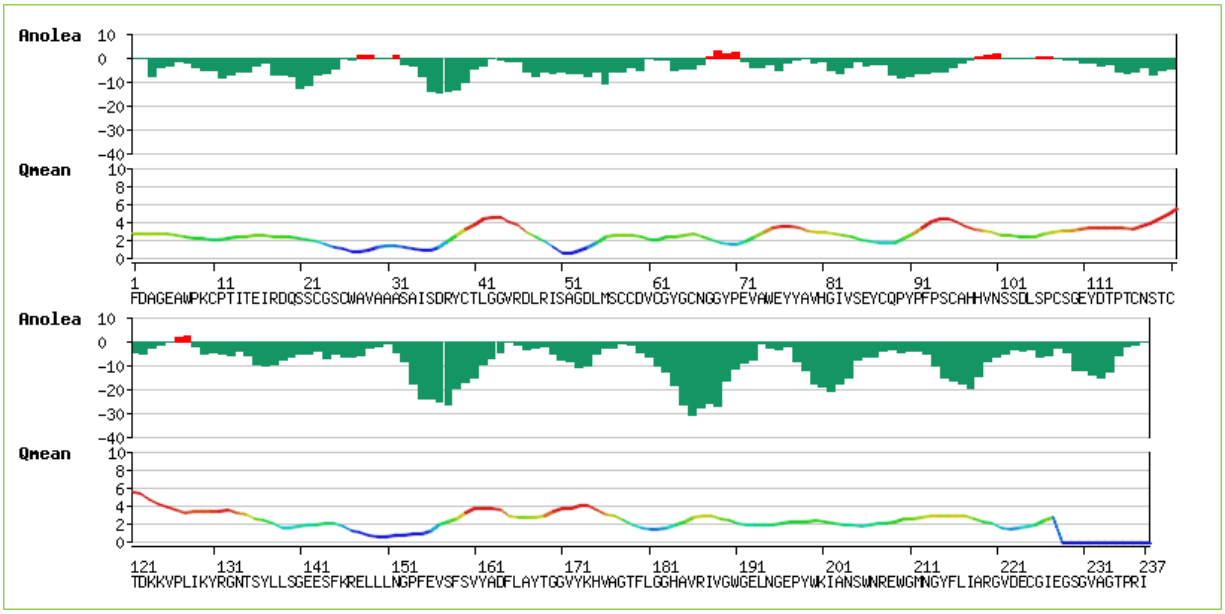


Figure 2.19: ANOLEA evaluation and QMEAN6 energy profile for *T. cruzi* cat B-like protease model built from 3CBJ template. ANOLEA score values greater than zero correspond to high energy regions / erroneous possible interactions. QMEAN scores are color coded from green (stable) to red (unstable) regions. QMEAN scores above 2 represent high energy (unstable) region.

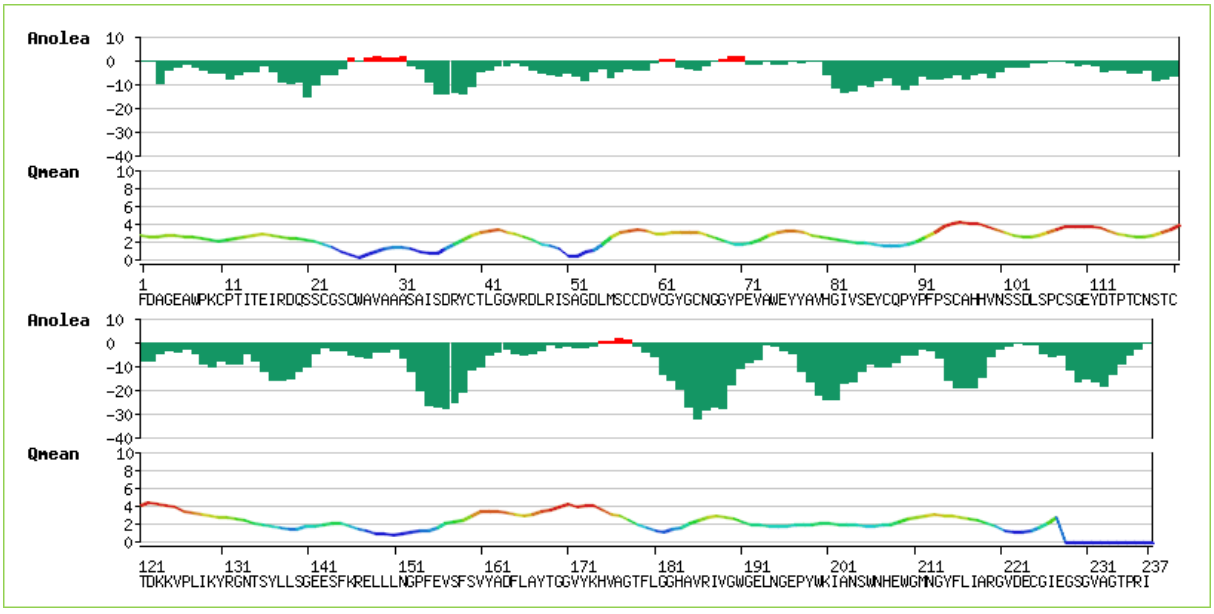


Figure 2.20: ANOLEA evaluation and QMEAN6 energy profile for *T. cruzi* cat B-like protease model built from 3HHI template. ANOLEA score values greater than zero correspond to high energy regions / erroneous possible interactions. QMEAN scores are color coded from green (stable) to red (unstable) regions. QMEAN scores above 2 represent high energy (unstable) region.

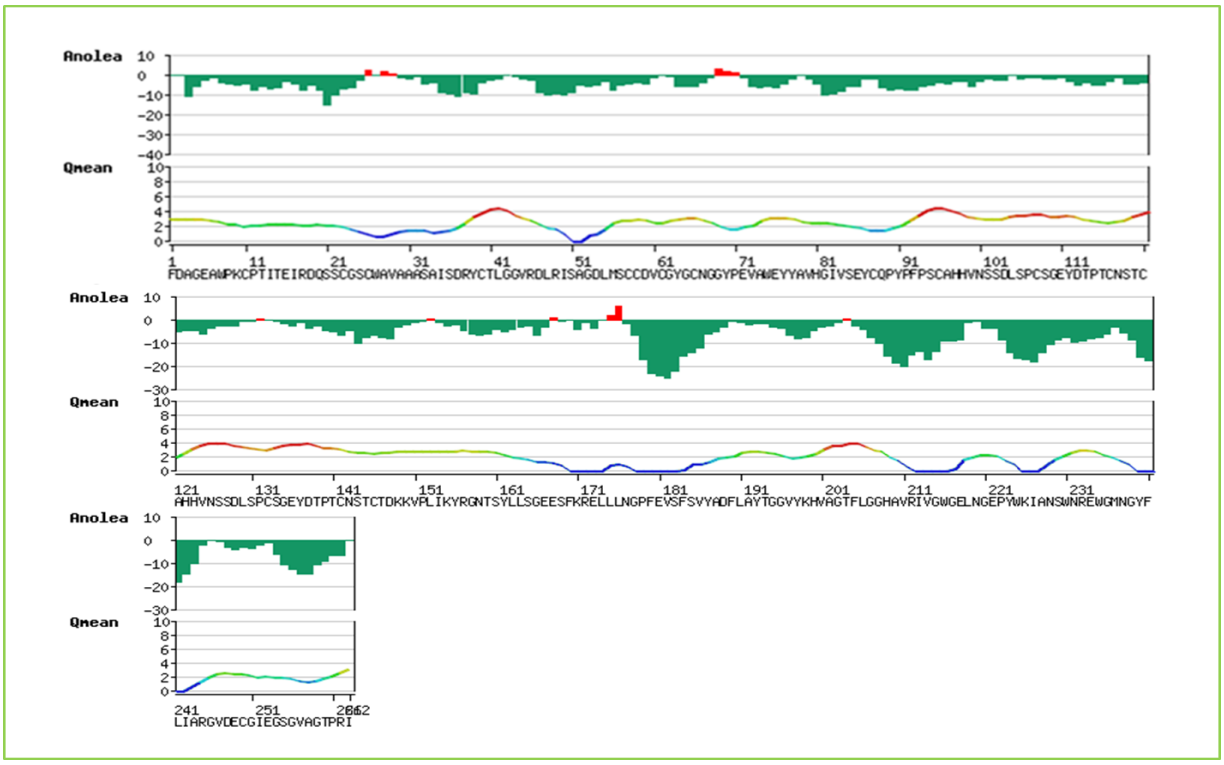


Figure 2.21: ANOLEA evaluation and QMEAN6 energy profile for *T. cruzi* cat B-like model built from a combination of 3HHI and 3CBJ templates. ANOLEA score values greater than zero correspond to high energy regions / erroneous possible interactions. QMEAN scores are color coded from green (stable) to red (unstable) regions. QMEAN scores above 2 represent high energy (unstable) region.

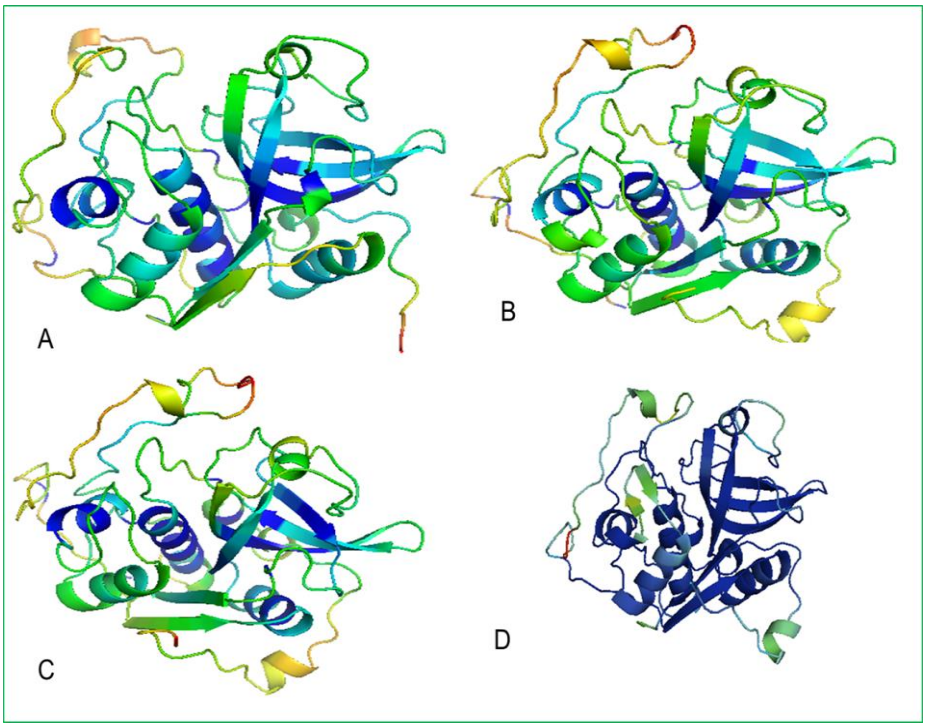


Figure 2.22: QMEAN energy profile for *T. vivax* cat B-like models built from (A) 3CBJ, (B) 3HHI, (C) Double template and (D) MetaMQAPII for double template models. The QMEAN/MetaMQAPII scores are color coded from blue (stable) to red (unstable) regions.

Predicted residue errors for *T. vivax* models (Figure 2.22) show that the models calculated using 3CBJ template is the most reliable one as shown by the lack of unstable (red) regions in the model (Figure 2.22. A). The QMEAN energy profile for models calculated from 3HHI template and a combination of the templates have unstable residues in parts of the occluding loop that take part in substrate binding. But the energy MetaMQAPII profile for the double template model (Figure 2.22 D), does not show unstability at the same region for the same model.

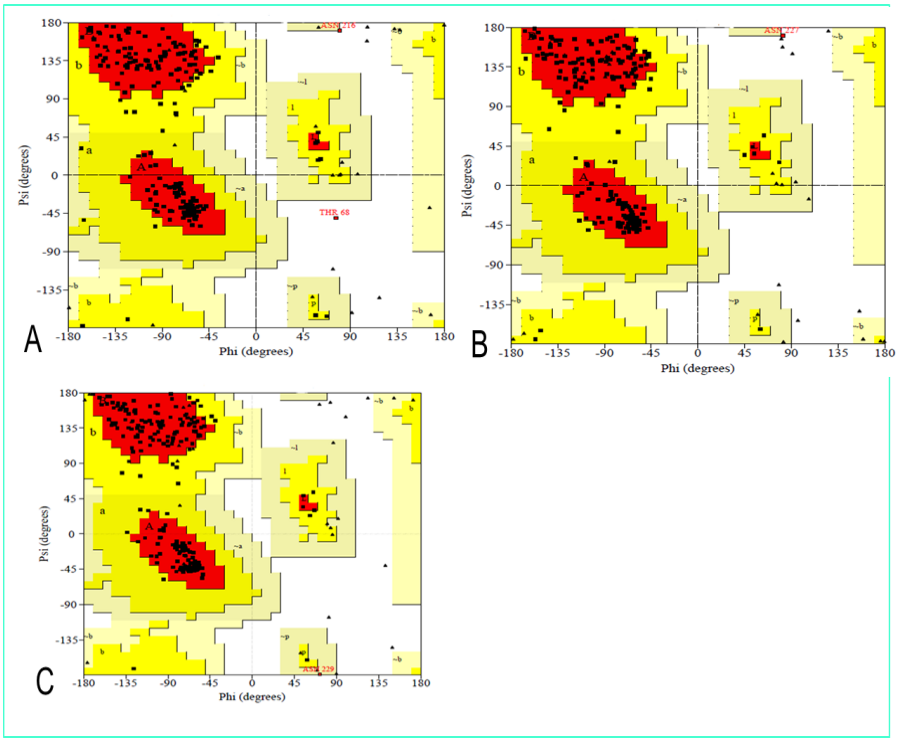


Figure 2.23: PROCHECK analysis for *T. vivax* cat B-like models built from (A) 3CBJ, (B) 3HHI and (C) double template models. The PROCHECK plot statistics shows 99.5%, 99.5% and 99.6 % of residues are in favoured regions for 3CBJ, 3HHI and double template models respectively.

The results for stereochemistry analysis of *T. vivax* models (Figure 2.23) show that 99.5% of the residues are in favored regions for the model calculated using 3CBJ template. This model has Asn216 and Thr68 residues modelled in disallowed regions. The model calculated from 3HHI template has Asn227 in disallowed region and 99.5% of residues in favored regions. Using both templates to calculate the model resulted in 99.6% of residues modelled in favored regions while Asn229 was modelled in a disallowed region.

Although ANOLEA evaluation and QMEAN energy profile indicate that all the models are reliable (Figure 2.24-2.26), the most stable models is that calculated from a combination of the templates (Figure 2.26). Both ANOLEA and QMEAN results show that regions around residues 104-109 and 114-120 are unfavored and unreliable for the model calculated from 3CBJ template (Figure 2.24). These regions are very close to the His-194(110) and His-195(111) residues that are responsible for the exopeptidase activity of cathepsin B enzymes. However since the error was not picked by other models evaluation programs, this models can still be relied on for further applications.

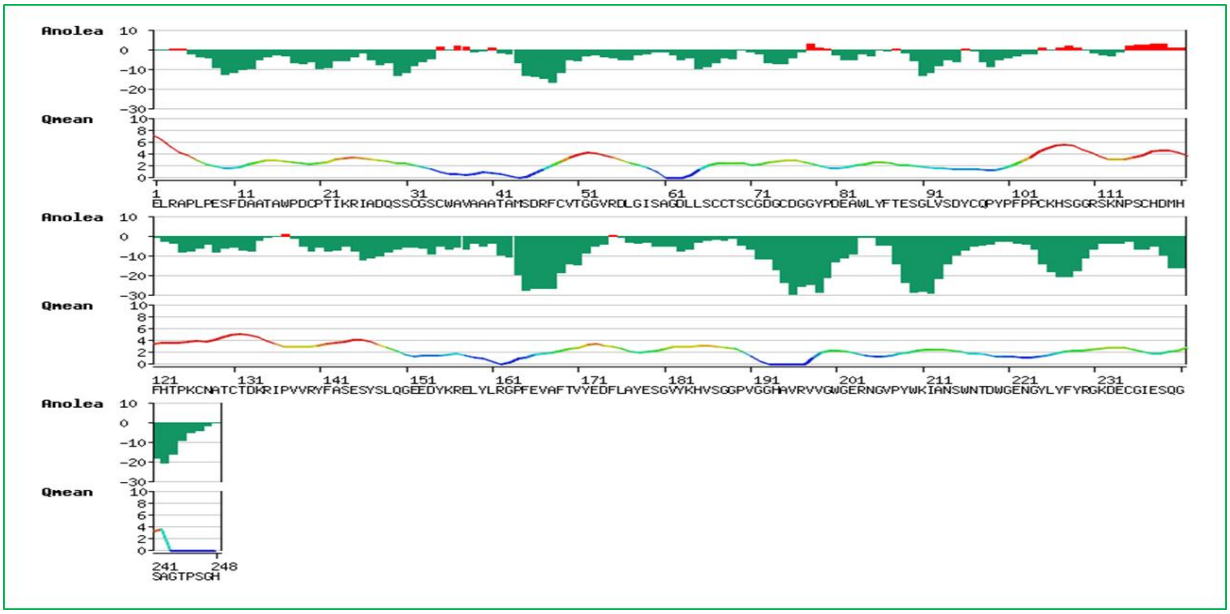


Figure 2.24: ANOLEA evaluation and QMEAN6 energy profile for *T. vivax* cat B-like model build from 3CBJ template. ANOLEA score values greater than zero correspond to high energy regions / erroneous possible interactions. QMEAN scores are color coded from green (stable) to red (unstable) regions. QMEAN scores above 2 represent high energy (unstable) region.

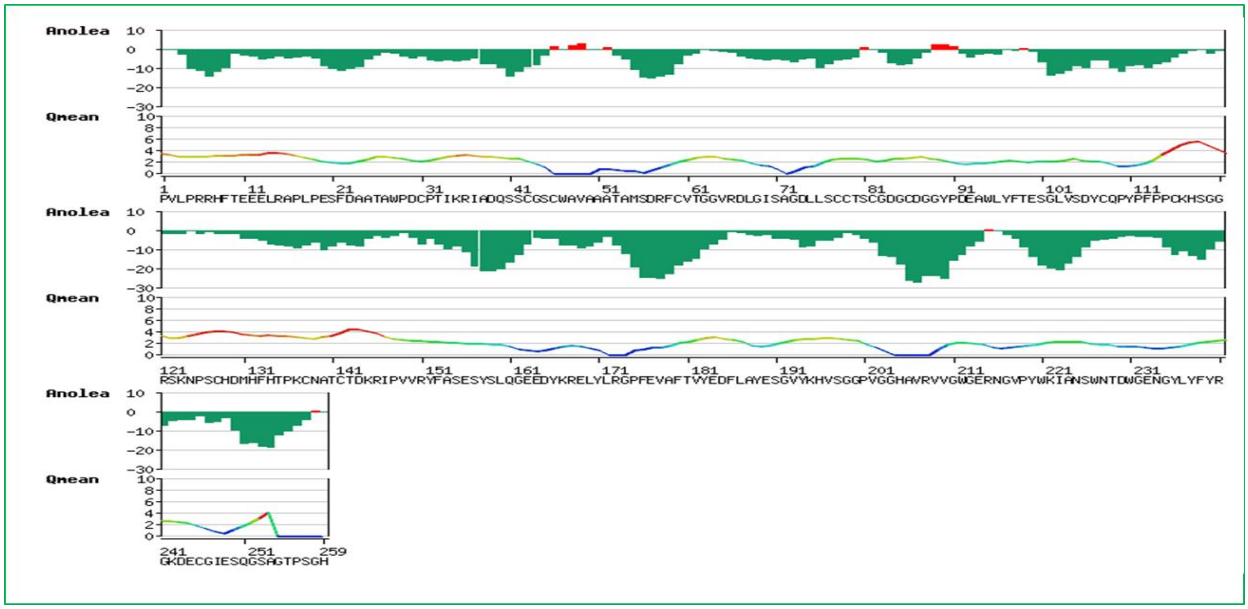


Figure 2.25: ANOLEA evaluation and QMEAN6 energy profile for *T. vivax* cat B-like model build from 3HHI template. ANOLEA score values greater than zero correspond to high energy regions / erroneous possible interactions. QMEAN scores are color coded from green (stable) to red (unstable) regions. QMEAN scores above 2 represent high energy (unstable) region.

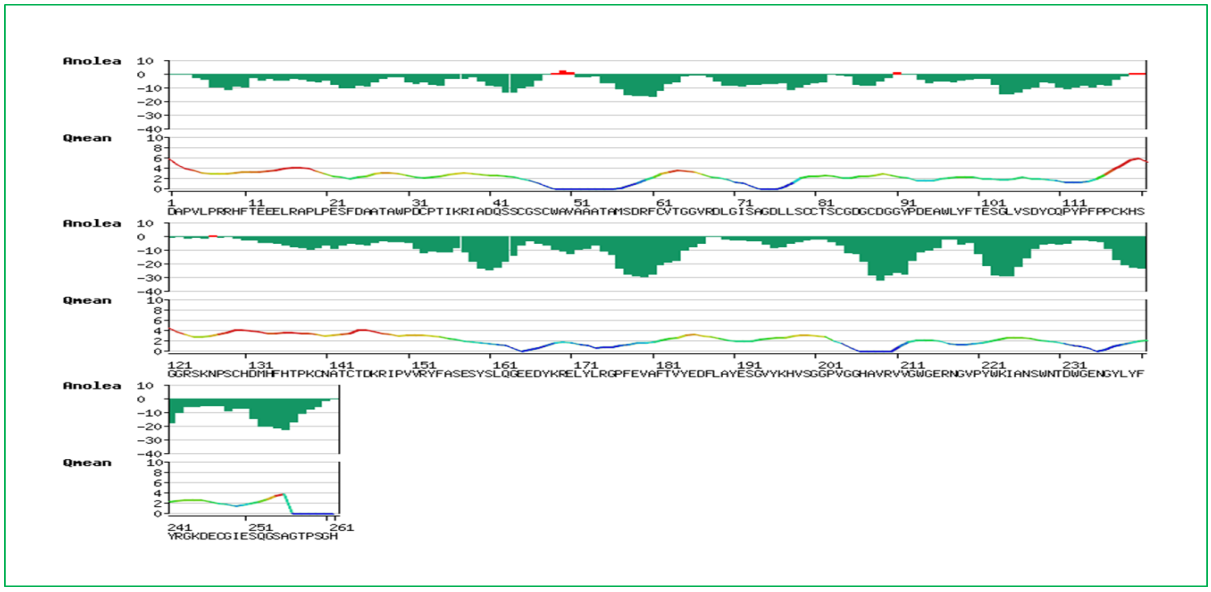


Figure 2.26: ANOLEA evaluation and QMEAN6 energy profile for *T. vivax* cat B-like model build from a combination of 3HHI and 3CBJ templates. ANOLEA score values greater than zero correspond to high energy regions / erroneous possible interactions. QMEAN scores are color coded from green (stable) to red (unstable) regions. QMEAN scores above 2 represent high energy (unstable) region.

2.9.12. Comparative structural analysis of the active site

As stated in chapter 1 (1.7.4), substrate specificity is determined by interactions between the enzymes and the substrate residues at the active site. Differences at the active site may therefore be used to design target specific inhibitors. The S2 subsite of cathepsin B proteases determines selectivity [90] and the bottom of TbCatB subsite is occupied by a Gly328 residue. This small residue allows the S2 subsite to be deep enough to accommodate large P2 subsites. The same position is occupied by the larger Glu245 residue in HsCatB, which results into a shallower S2 pocket [23]. Another notable difference between TbCatB and HsCatB is the extra flexibility in the occluding loop of HsCatB due to glycine residues on both sides on Glu122 in the "GEGD" motif between residues 206(120) to 210(123).

The flexibility allows the Glu122 to move in and out of the S1' subsite easily. The same movement is restricted for the Asn209 residue in the corresponding "FNFD" motif in TbCatB. The Phe208 and Phe210 residues flanking the Asn209 residue are attached to the Phe189 residue of the occluding loop, creating a stable opening around the S1' subsite. In fact Kerr et al states that this feature could be used to design inhibitors targeting this enzyme. In addition to the Pro106-Pro107-Cys108 and Pro117-Pro118-Cys119 regions that confer stability to the occluding loop in HsCatB [25], at low pH additional stability of this loop is due to two salt bridges between (His110 and Asp22) and (Arg116 and Asp224) whose disruption resulted in

increased endopeptidase activity [23]. In our MSA (Figure 2.1), His110 and Asp22 are conserved in all the sequences, Arg116 corresponds to Tyr202 [23] and Asp224 corresponds to Glu307 in TbCatB. The biochemical roles of these substitutions have not been investigated yet, leaving room for future research. The His110 and Asp22 interaction has however been exploited to develop the specific cathepsin B inhibitor CA-074, and the removal of His110 residue has been correlated with improved inhibitor binding [37].

2.10. Conclusion

Different model quality evaluation programs were used so that a model could be disqualified based on their combined results. The DOPE Z scores and the RMSD (RMS) values of the calculated models indicate that the models are of acceptable energy (stability) and fold (conformation). Based on the RMSD values the models adopted the orientation of the TbCatB crystal structure more than they did the human crystal structure. Even models made from a combination of the two templates adopted the occluding loop conformation of TbCatB template. These results are expected since they shared a higher sequence identity (see Table 2.2) with the TbCatB sequence than with the HsCatB sequence. The different MQAPs used did not single out a certain model as unacceptable or unreliable. The MSA made from MAFFT program produced a phylogenetic tree of higher bootstrap values that the one from PROMALS3D (see Appendix 2A). The overall conclusion reached from these models is that they are of acceptable quality and they can be used for docking studies. To determine the effect of the difference in the residues at the bottom of the S2 pocket, molecular docking and molecular dynamics simulation studies shall be carried out in the next chapters. These studies are expected to shed light on the effect of residue variations at the active site on substrate interaction. These differences present an opportunity to design inhibitors that are specific for TbCatB and other trypanosome cathepsin B proteases.

Chapter 3.

3.1. Molecular Docking

In chapter 2, sequence and structural comparisons of *Trypanosoma brucei* cathepsin B (TbCatB) and homologous proteins from human Cathepsin B (HsCatB), *Trypanosoma congolense* cathepsin B (TcCatB), *Trypanosoma cruzi* cathepsin B (TcrCatB), and *Trypanosoma vivax* cathepsin B (TvCatB) were looked at. Major differences were noted, especially in the S2 subsite where TbCatB has the smaller Gly328 residue while HsCatB has the larger Glu245 residue at the same position resulting in a larger S2 pocket in TbCatB. We also noted differences in the occluding loop region where TbCatB has a "FNFD" motif which gives it stability and results in a stable S1' subsite opening. HsCatB has a "GEGD" motif at the same position and this motif makes this region flexible and results in an unstable S1' subsite in HsCatB. This differences present an opportunity for designing target specific compounds [23]. In this chapter, how the variations in the active site of these homologous proteins affect binding of substrates is explored.

A known cysteine protease inhibitor CA074 and a nitrile inhibitor were used for docking validation and to investigate interacting residues in TbCatB and HsCatB respectively.

The effect of these interactions can be used to determine the properties of drugs that can be developed to target TbCatB. As previously stated in our problem statement (chapter 1.8), currently used drugs were developed for their anti-parasite properties without knowledge of the biochemical pathways of the parasite. They therefore lack specificity and are toxic to the host [8]. Molecular docking studies were used to investigate the active site residues of TbCatB and its homologs and to screen the South African Natural Compounds Database (SANCDB) [49], https://sancdb.rubi.ru.ac.za for possible leads. Molecular dynamics (MD) studies of the most promising compound(s) were carried out to determine the stability of their interactions with the substrates and to understand how the interacting residues fluctuate over time under conditions close to physiological ones.

3.2. Introduction

Proteins function by interacting with each other, biomolecules like DNA, RNA, and with small molecules. Understanding these interactions can reveal the mechanism by which they function and can also reveal information on how they can be exploited for functional and therapeutic purposes [91], [92], [93]. In protein-ligand docking, the aim is to predict and determine the conformations and binding affinities of complex structures that are formed during interaction of a ligand and a target protein of known crystal structure or a homology model [94], [95]. In drug discovery, the correct prediction of binding of small molecules (ligands) to target proteins is important because it can be used to screen large databases of compounds to obtain leads for drug design [95]. These compounds could be sourced from corporate or commercial compound database, or from virtual compounds libraries [96] like the South African Natural Compounds Database (SANCDB)[49] used in this study.

There have been several successful cases in which novel ligands have been identified using receptor virtual screening methods. Examples of these include identification of an indoloquinazolinone derivative from 400 000 molecules as a potent inhibitor of human casein kinase II using a homology model, and the novel 2-amino-4-heteroaryl-purimidine inhibitors of CDK2 that were identified from a commercial library of 50 000 compounds [5].

Protein - ligand docking is the process of predicting a protein and ligand complex and its free binding energy using experimentally determined 3D complex structures or free structures and homology models [94] , [9]. During docking, the best orientation and conformation of the ligand is searched within the binding pocket of a protein crystal structure or model in a process known as posing [99]. Early docking processes historically only involved the use of a protein and ligand as rigid components. However current methods view the protein as a body made up of different conformations in equilibrium and a flexible ligand. A number of alternative protein and ligand complex conformations are usually generated and then ranked according to ligand binding affinity.

The process of docking therefore involves a search algorithm and a scoring function [94], [99]. The search algorithm should efficiently generate a broad range of plausible binding conformations while the scoring function should represent the thermodynamics of all the protein-ligand complexes well enough to separate and rank the best complexes from the rest. Protein-ligand docking is used in structure-based drug design and discovery research to study the mechanisms of recognition and interaction between protein substrates and inhibitors and to elucidate fundamental biochemical processes [94], [99]. Protein - ligand docking is therefore an

ideal tool to use in virtual screening of large databases of compounds (small molecules) if a target structure is available. It can be used to discriminate between potential strong binding compounds and non-binders, speeding up the process of obtaining leads for further drug development [100]. Docking has been used in many situations - anti-gout [45], *M. tuberculosis* [43] and malaria [101] research – to name a few to search for potential drug leads. This highlights the importance of accurate prediction of the binding modes of the ligand to the protein. Prior knowledge of the binding site increases docking efficiency since the search space is already narrowed down. Comparing the target protein with protein-ligand co-crystals of proteases that share the same function can be used to obtain information about the binding site. Online servers like GRID, POCKET, SurfNe, PASS and MMC can also be used to obtain the binding site for docking [99].

3.3. Docking Algorithms

The root-mean-square deviation (RMSD) between the predicted ligand pose and the actual position of the ligand in crystal structure complex is used to measure the accuracy of the docking. To reduce difficulties due to the flexibility of the complex and reduce the degree of freedom, the solvent molecules are often removed to allow for a more effective search for the pose. Approaches to estimate or approximate the pose include the already mentioned rigid-body approximation, which treats both the protein and the ligand as rigid bodies. This approach only considers the 6 degrees of translational and rotational freedom and discards flexibility of the system [102]. A more commonly used, and preferable approach considers the ligand flexibility and models the protein as a rigid body [103] although there are systems that consider protein flexibility [94], receptor backbone flexibility is still a problem [99].

3.3.1. The flexible ligand-search docking algorithm uses three types of algorithms;

1. Systematic docking algorithms

In these algorithms all the degrees of freedom in a molecule are explored using conformational search methods (exploring all rotatable bonds of the ligand), fragmentation search methods (assembling the ligand in the active pocket), and database search methods (docking pre-generated conformations in to the active pocket) [94], [99].

2. Random or stochastic algorithms

When using the random or stochastic algorithm, different conformations of the ligand or ligands are explored in the binding site and then accepted or rejected using a pre-defined probability. Three of the methods available for this approach are Monte Carlo method (MC) which is used

by Prodock, ICM, MCDOCK, DockVision and QXP, Genetic Algorithm method (GA) used by GOLD, AutoDock [92], [104], DIVALI and DARWIN and finally the Tabu search method used by PRO_LEADS program [94], [99].

3. Simulation methods

Simulation methods are based on calculating solutions to Newton's equation of motion. Molecular dynamics (MD) and pure energy minimisation method use this approach. Examples of programs that use this method are Prodock, ICM, QXP, DARWIN, DOCK 4.0 [105], ADAM and Hammerhead [94], [99]

3.3.2. Flexible Protein Docking Algorithm

Development of computational approaches that are able to account for protein flexibility during docking is still relatively new. To simulate docking in a fully flexible protein target within reasonable time is not yet computationally feasible. MD and MC approaches that can account for partial flexibility in the protein are available. Other approaches include rotamer libraries, protein ensemble grids and soft-receptor modelling [94], [99].

3.4. Scoring functions

To be able to rank and score the ligand conformation is very important when docking because correct poses have to be delineated from incorrect poses [99]. In addition to generating the correct conformation, it is important that it is recognized as such and that it is distinguished from other alternatives.

Due to the high computational expense that may be involved in scoring functions, analysis of several binding modes is made feasible by simplifying the scoring function, and as such their accuracy is reduced. In protein-ligand docking, scoring functions that are used are able to predict binding free energies. The three classes of scoring functions used in protein-ligand docking are:

1. Force Field-Based Scoring

Force field-based scoring functions calculate the sum of the protein and ligand interaction energy and the internal energy of the ligand. The D-Score, G-Score (Tripos force field), GoldScore and the AutoDock 3.0 scoring function (Amber force field) are force field based scoring functions [94], [99].

2. Empirical Scoring Functions

Empirical scoring functions make approximations by using both experimentally determined binding energies and training sets of protein-ligand crystals complexes to determine the binding energy. Examples include the Böhm's scoring function, F-Score, ChemScore, SCORE, Fresno and X-SCORE [94], [99].

3. Knowledge based Scoring Functions

The third scoring functions are knowledge based and they use protein-ligand crystal structures and follow statistical rules and principles to derive their terms. Examples of knowledge-based scoring functions include Muegges's Potential of mean force (PMF), DrugScore, and SMoG score [94].

4. Consensus Scoring

These scoring functions combine data form different scoring functions to reduce errors that arise from using individual scoring functions. Examples include X-CSCORE, which combines the PMF, ChemScore, and FlexX scoring functions with DOCK-like and GOLD-like algorithms [94], [99].

3.5. Capabilities and limitations of docking

Although most docking programs can predict the orientation of a known protein-ligand complex with an average of 1.5-2.0 Å and a reported success rate of up to 80%, docking is still far from perfect. Major limitations arises from the imperfections of the scoring functions since, as previously mentioned, a lot of simplifications have to be made to make scoring functions computationally efficient in determining binding free energies or ligand affinity. Other factors that affect the quality of docking include the effect of the solvent and water molecules in protein-ligand interactions, the protein flexibility and the resolution of experimentally determined structures [94].

3.6. Docking programs

There are several molecular docking tools available and they use different approaches. The most popular docking programs are AutoDock, GOLD and FleX [94], [99]. For this project AutoDock4 and AutoDock Vina were used at different stages of the docking protocol.

3.6.1. AutoDock4

Autodock4 is an automated docking tool that is part of a suit of programs that includes AutoDock, AutoGrid and AutoDock Tools (ADT). It combines an empirical free energy force field with a Lamarckian Genetic Algorithm to predict a protein-ligand conformation (pose) and its free binding energy. AutoDock4 also has a simulated annealing search method and a traditional genetic search method. AutoDock uses a grid-based method to search through the conformational space available to a ligand. This method makes it possible to rapidly predict the binding energy of the proposed pose. The scoring function of AutoDock has been calibrated using a set of 188 different protein-ligand complexes of known structure and binding energy. The inputs and outputs are pdbqt files that are prepared using ADT graphically or in batch mode. AutoDock4 also allows for flexible docking of specific parts of the protein and the ligand. Docking results can be clustered, displayed and analysed using different available methods[94], [104].

AutoDock4 and ADT are currently being distributed free of charge at WWW sites: http://autodock.scripps.edu and http://mgltools.scripps/edu/downloads respectively.

3.6.2. AutoDock Vina

AutoDock Vina is an open source program for molecular docking. It was developed at the Molecular Graphics Lab at The Scripps Research Institute. The performance of AutoDock Vina has been tested on a set of 190 protein-ligand complexes that were used as a training set for AutoDock 4.0.1 and a twofold improvement in speed and better accuracy in predicting binding mode were observed during a comparison with AutoDock 4.0.1. AutoDock Vina requires a specification of the search area in which various ligand poses will be considered in the receptor. This can be accomplished by starting with a protein-ligand crystal structure complex and then creating an appropriate search space that includes the ligand binding site. There is no need to calculate grid maps and to assign atom charges. It uses multithreading to speed up execution by making use of multiple CPUs or CPU cores. AutoDock Vina uses the same pdbqt molecular structure file format used by AutoDock. The pdbqt files used in Autodock Vina can therefore also be generated in batch mode or in AutoDock Tools and be viewed using MGL tools [95].

3.7. Anti-parasitic natural compounds

Screening natural compounds provides an opportunity to discover unique molecules with properties that can be optimized by synthetic procedures. Examples include artemisinin, quinine and licochalcone A which are derived from plants, as well as antiparasitic products like

amphothericin B and ivermectin which were isolated from *Streptomyces* [106]. Several phytochemicals have been investigated for their activity against trypanosomes [107].

Natural compounds with anti-parasitic properties are diverse. Those that have been identified as potential drug leads and undergone *in vivo* and toxicity studies include quinoline alkaloids with anti plasmodial and anti-leishmania activity, bisbenzyl-isoquinolines with antiprotozoal activity, benzyl- and napththyl-isoquinoline alkanoids and indole alkaloids[106]. Terpenes that can be of value to drug development include sesquiterpenes like artemisinin which is active against the malaria protozoa, diterpenes like axisonitrile which is active against plasmodium, and limonoids like nimbolide which is active against plasmodia. Phenolics like lignans which include a group of natural products that have been shown to have activity against *Trypanosoma cruzi* and can prevent transmission of *T. cruzi* by blood transfusion [106] have also been identified.

Although some natural products are disadvantaged by their high cytotoxicity and low therapeutic selectivity, they are still a good source of compounds that can be developed to form potent and viable drugs. Comparative molecular docking studies have been carried out on anti-trypanosomal natural products into different *T. brucei* drug targets [108]. In this study the South African natural compounds database (SANCDB) [49] will be used as a source of compounds that will be screened for their anti-trypanosomal activity.

3.8. Methods and materials

In chapter two (Homology modelling and protein structure analysis), crystal structures of *Trypanosoma brucei* cathepsisn B (TbCatB), PDB ID: 3HHI and human cathepsin B (HsCatB), PDB ID: 3CBJ were used as templates to calculate homology models for, *Trypanosoma congolense* cathepsin B (TcCatB), *Trypanosoma cruzei* cathepsin B (TcrCatB) and *Trypanosoma vivax* cathepsin B (TvCatB) cysteine proteases. Since the crystal structure of PDB ID: 3CBJ was crystallized as an inactive protein with mutations (C29A, H110A, S115A) it could not be used for docking studies. Another HsCatB crystal structure (PDB ID: 1GMY) was retrieved from the NCBI database using the 3CBJ sequence as a query in a BLASTP search. The 1GMY structure was crystalized at a resolution of 1.90Å, an R-value of 0.161% and R-free of 0.199%. The ensemble coordinates of this structure were used for building the homology model that was used for making the TbCatB crystal structure [23].

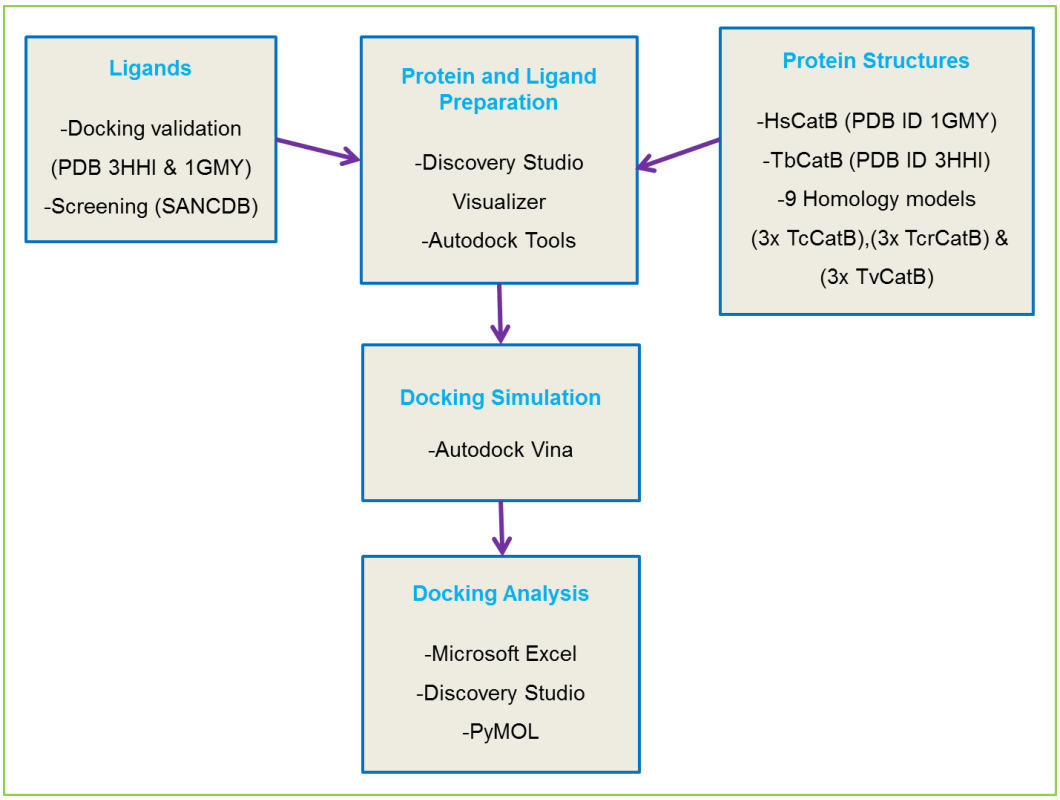


Figure 3.0: An overview of the methods used for molecular docking studies. The crystal structures of TbCatB and HsCatB, calculated homology models, as well as compounds from the SANCDB were used.

Superimposition of these two structures (1GMY and 3HHI) (Figure 3.1) showed the RMSD between the two structures is 0.601Å. These structures and the homology models were also used with compounds from the SANCDB in docking studies to identify leads for molecular dynamics and drug development. Figure 3.0 show an overview of the steps taken during docking studies.

3.8.1. Data preparation for molecular docking

Homology models of TcCatB, TcrCatB, TvCatB proteases were calculated using MODELLER version 9.10 (see chapter 2). The crystal structure of (TbCatB) in complex with a cysteine protease inhibitor (CA074) was retrieved from the protein data bank (PDB ID: 3HHI). The crystal structure of HsCatB (PDB ID: 1GMY) cathepsin B protease in complex with dipeptidyl nitrile inhibitor was retrieved from the NCBI database. The protein receptors were prepared by removing crystallographic waters and other bound heteroatoms using Accelrys Discovery Studio 4.1 (Accelrys Software Inc. Discovery Studio Modelling Environment (DS), 4.1, San Diego: Accelrys Software Inc. 2014). This software was also used to separate the CA074 and the

dipeptidyl nitrile ligands from the 3HHI and the 1GMY proteases respectively. These inhibitors were used together with the protease structures to validate the docking method and to determine important interacting residues between the receptors and the respective ligands. Before docking validation, the protein receptors were all aligned in PyMOL [83] and the saved molecules were used for docking studies. The xyz coordinates of the central C-alpha of the active Cys residue of the aligned receptors were determined in DS for the grid centre. A total of 11 receptors were used in docking studies. These were HsCatB (1GMY), TbCatB (3HHI), Tc3CBJ, Tcr3CBJ, Tv3CBJ, and Tc3HHI, Tcr3HHI, Tv3HHI (*Trypanosoma congolense, Trypanosoma cruzi* and *Trypanosoma vizax* cathepsin B homology models calculated using PDB ID 3CBJ and 3HHI template coordinates respectively). Three more receptors used were TcCatB, TcrCatB and TvCatB (*Trypanosoma congolense, Trypanosoma cruzi* and *Trypanosoma vizax* cathepsin B homology models calculated using a combination of both PDB ID 3HHI and 3CBJ template coordinates).

3.8.2. Docking validation in Autodock4

Autodock tools were used for preparation of pdbqt files of the ligand and proteins. The search area was also determined using AutoDock4. The two ligands were then re-docked in to their respective receptors in an attempt to reproduce the original poses.

3.8.2.1. Ligand and protein protonation

Docking algorithms require atoms to have a charge and atom type, since the PDB structures and the model structures do not have these. The protein structures and the ligands were converted into (pdbqt) conformations using AutoDock Tools (ADT). To do this, polar hydrogens were added to the protein receptors and non-polar hydrogens merged. Gasteiger charges were then calculated followed by assigning AutoDock4.2 (ADT4) atom types. Torsions were automatically assigned to the ligands. Python scrips provided by ADT together with customised python scripts were used (Appendix 2A-1&2).

3.8.2.2. Vina configuration file preparation

To select the search space for the active site pose, for each protein-ligand complex the experimentally bound ligand structures were used as starting points from which to determine the centre and search area size for each of the aligned receptors. The centre for the aligned receptors was set on the central C-alpha of the active Cys residue. The xyz coordinates for this centre were respectively set to -16.67, -17.76 and 17.76. The grid size was adjusted to a final size of x = 30.00Å, y = 31.875Å and z = 31.875Å. This was done to make sure that the size of the search space is large enough to allow the ligand to rotate as recommended by the developers of

AutoDock [95] and to give allowance for large compounds that will be screened from the SANCDB.

3.8.3. Docking validation and HTS in AutoDock Vina

The two ligands were re-docked into the corresponding receptors using AutoDock Vina, (using parameters to be used with screening the SANCDB). The Vina configuration file was prepared by setting the centre and grid size as explained for docking parameter file preparation (section 3.7.2.2 above). A script was used to create new files with ".vina" extension for possible ligand-protein complexes during HTS of SANCDB (Appendix 2A-1). The vina parameter files contained the absolute path to the ligand(s) and proteins. The script also generated job files per ligand in SANCDB. An OpenPBS command, qsub, was used to schedule the job scripts to be executed by the cluster during HTS. After completion of docking, the lowest docking energy was extracted from vina output all.pdbqt file using another script. To speed up the running time, a total of 4 CPUs were used simultaneously for each docking experiment. The extent of searching for the global minimum (exhaustiveness) and the energy range were also set to 4 in the vina docking parameter file. Results were analysed in DS, Excel and in PyMOL.

Docking method	Energy range	Exhaustiveness	Cpu usage
AutoDock Vina active site docking	4	4	4

Table 3.2: Docking parameters used in active site docking in AutoDock Vina.

3.8.4. Results and discussion

PyMOL was used to visualise the molecules and prepare the figures from the experiments.

3.8.4.1. Data preparation for molecular docking

Figure 3.1(A) shows the crystal structure of HsCatB (1GMY) superimposed to that of TbCatB (3HHI). The two structures are in-complex with their ligands which are coloured in similar colours to the receptors. The active site Cys122(29), (TbCatB numbering with HsCatB in brackets), is shown in red. The RMSD of the two structures is 0.601Å. The occluding loop of the two structures is open to the same extent. In the superimposition of the two structures in Figure 3.1(B), the CA074 inhibitor (yellow) is bound along the active site cleft of TbCatB by non-covalent interactions which are dominated by hydrogen bonds [23]. The dipeptidyl nitrile inhibitor (green) is covalently bound to the active site cysteine residue in HsCatB [109].

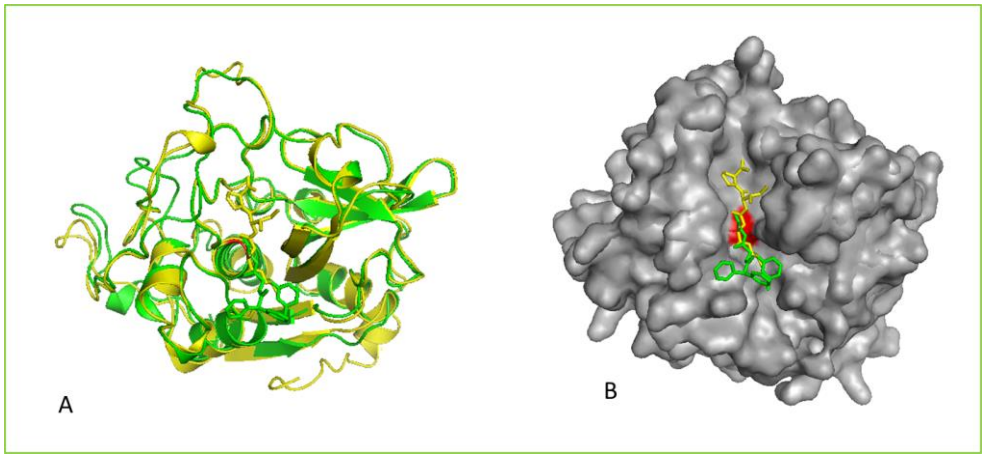


Figure 3.1: Showing (A) HsCatB (green) and TbCatB (yellow) superimposed crystal structures. Active site Cys122(29) is shown in red. RMDS = 0.601Å.

Figure 3.2 shows all the protease receptors superimposed. The main observable difference between these structures can be seen in the occluding loop of the structures. The occluding loop of Tc3CBJ, Tcr3CBJ and Tv3CBJ models is open wider than that of the other proteases. This orientation is inherited from the 3CBJ template which was crystalized in-complex with a large chagasin cysteine protease inhibitor. In this orientation, these proteases are likely to accommodate larger compounds and allow compounds to go deeper into the active site cleft than the rest of the proteases

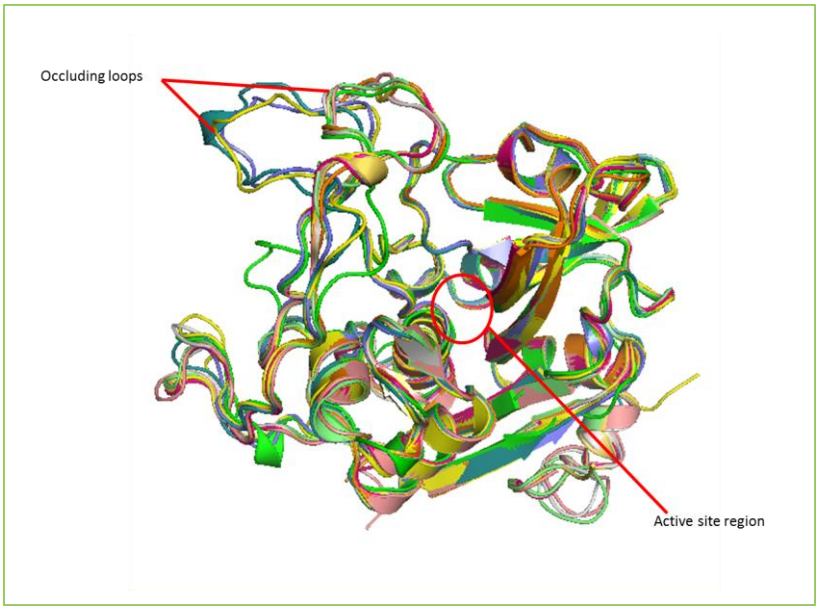


Figure 3.2: Superimposed structures used for docking experiments. The occluding loop of models Tc3CBJ, Tcr3CBJ and Tv3CBJ is open wider than that in other receptors.

The occluding loop of cathepsin B protease is a unique feature that gives it its carboxypeptidase activity under certain pH conditions. It can adopt multiple conformations and when it is in the closed conformation (at low pH) it blocks the primed site of the active site and restricts access of two residues (H110 & H111). When in this conformation the H110e residue forms a salt bridge with Asp22e thus holding the occluding loop in place. When the pH is raised, the salt bridge is broken when the H110e becomes deprotonated. This allows the occluding loop to move and open access to the H111e which provides a positive charge to hold the C-terminal carboxylate of a substrate during carboxypeptidase activity [37].

3.8.4.2. Docking validation results

The results for docking validation of the CA074 cysteine protease inhibitor indicate that the validation was successful (Figure 3.3). In the original crystal structure complex, 3HHI, hydrogen bonds between the inhibitor and the receptor are more dominant than hydrophobic interactions and the phenylsulphone moiety is at the P1' position [23].

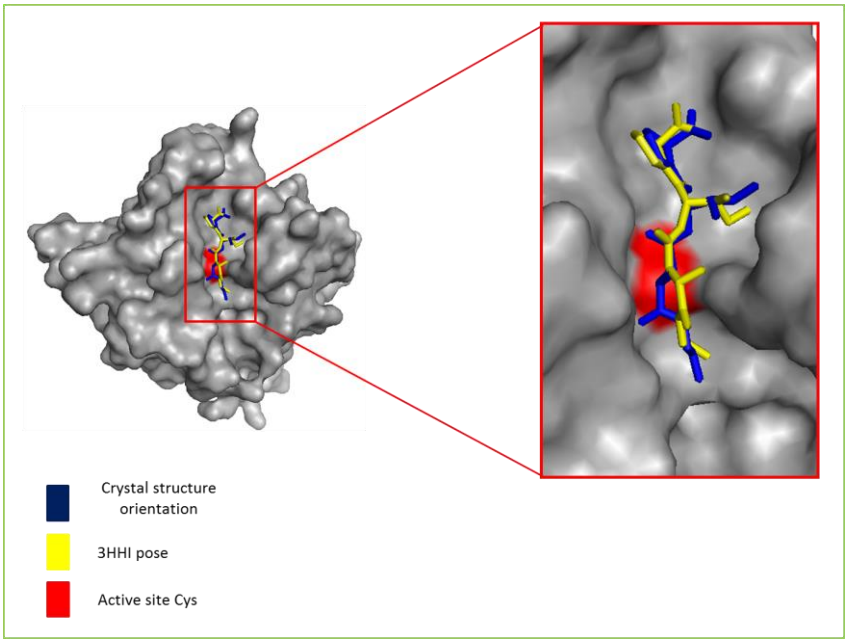


Figure 3.3: Showing the original position of the CA074 cysteine protease inhibitor (navy blue) and the docked pose (yellow) in TbCatB cystal structure. The active site Cys122 is marked red.

Figure 3.4 (A) shows major interactions between the enzyme and the original inhibitor. The inhibitor forms hydrogen-bonds between an oxygen atom and a nitrogen atom of Gln116. It also forms a hydrogen-bond with a nitrogen atom of Cys122. This forms the oxyanion hole which stabilizes the tetrahedral intermediate which forms during substrate hydrolysis [21]. Another H-bond interaction is formed between the terminal oxygen of the ligand and a nitrogen atom of

His194 and between an oxygen atom and a carbon atom of His282. The terminal oxygen also forms electrostatic interactions with Trp304. Hydrophobic interactions are also formed between the Val259 and a carbon atom of the ligand. In the re-docked inhibitor (Figure 3.4 B), the H-bond between Gln116 and the ligand was formed with a different oxygen atom. The hydrophobic interaction between Val259 and the carbon atom was replicated. The carbon atom of the inhibitor forms two other hydrophobic interactions with His282 and His304 of the enzyme. The phenylsulphone moiety at the P1' was correctly replicated.

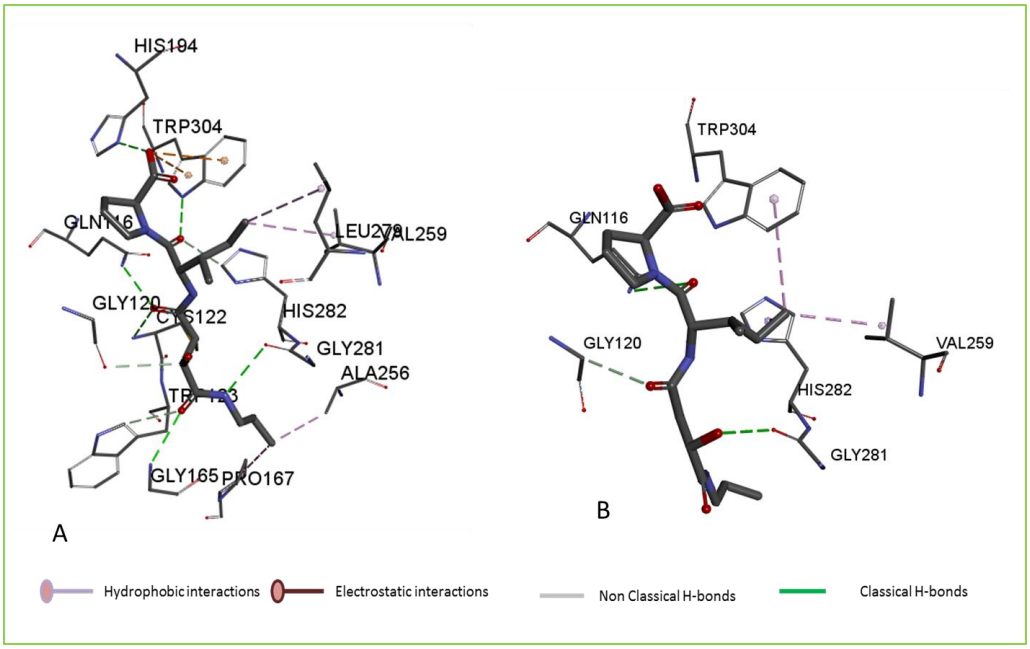


Figure 3.4: Showing (A) the original CA074 ligand and (B) docked ligand pose interactions with important residues in TbCatB (3HHI) protease.

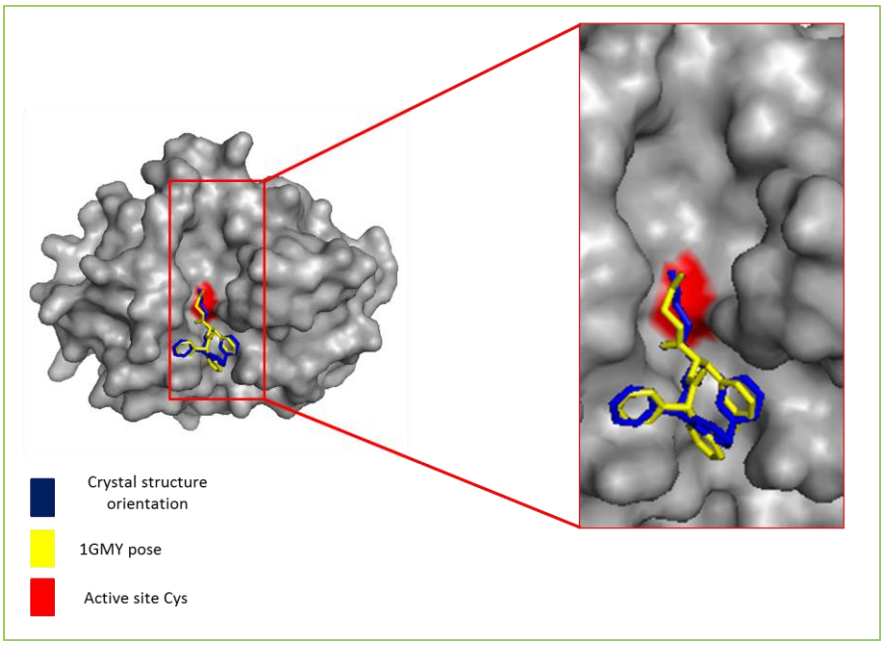


Figure 3.5: Showing the original position of the dipeptidyl nitrile protease inhibitor (navy blue) and the docked pose (yellow) in HsCatB cystal structure. The active site Cys29 is marked red.

The orientation of the original and the docked dipeptidyl nitrile inhibitor in HsCatB is shown in Figure 3.5. The original inhibitor was covalently bonded to the enzyme to the active Cys29 residue by a thioimidate ester bond [109] and hydrogen bond interactions between the backbone NH of Cys29 and the side chain amide of Gln23 stabilized the intermediate. In Figure 3.6 (A) it is observed that there is a hydrogen bond interaction between the side chain amide of Gln23 and the terminal nitrogen of the ligand. However this interaction is not replicated in the docked pose (Figure 3.6 B). This is due at least in part to the inability of docking to replicate the covalent bond between the inhibitor and the Cys29. However the hydrophobic interaction between Tyr75 and the phenyl ring of the ligand was replicated. A hydrophobic interaction between Pro76 and another phenyl ring of the ligand was also replicated.

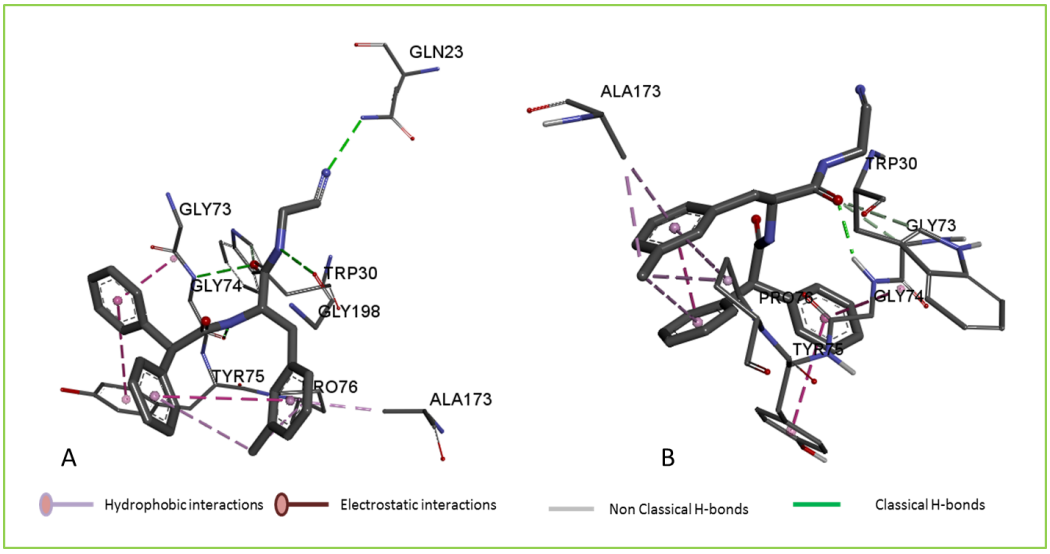


Figure 3.6 Showing (A) the original dipeptidyl nitrile ligand and (B) docked ligand pose interactions with important residues in HsCatB (1GMY).

3.8.4.3. Docking validation conclusions

Although not all the original ligand and protein interactions could be replicated for both 3HHI and 1GMY complexes, the docked ligand in both cases adopted the original ligand position and replicated enough interactions for the validation to be concluded as successful. The parameters that were used for docking validation were deemed reliable for use in HTS of South African natural compounds.

3.8.5. Screening of South African natural compounds in Autodock Vina

A total of 600 natural compounds of South African origin were retrieved from the South African natural compounds database (SANCDB)[49], https://sancdb.rubi.ru.ac.za. All of these compounds were screened into all the protease structures without regard for their anti-parasitic or lack of anti-parasitic properties. The main aim was to obtain lead compounds that have potential use in drug development against TbCatB protease.

The SANCDB is the first web-based natural products (NP) database in Africa. It contains compound information for NPs extracted from different referenced sources like journals, books and theses. The database is curated and it allows researchers to submit entries. The aim of the database is to provide researchers with natural compounds of South African origin that can be used for virtual screening in drug discovery. The compounds were isolated from plants and marine organisms. Information provided about the compounds include the use of a given compound; e.g. anticancer and the organism from which it was extracted. The compounds can

be downloaded in Mol2, PDB, SDF and SMILES formats [49]. Access to the database is free and it is available at https://sancdb.rubi.ru.ac.za/.

3.8.6. Docking analysis

Python scripts were used to prepare the compounds and the receptors for docking by generating the pdbqt files. Scripts were also used for automating the writing of vina parameter files and for running of the jobs. AutoDock Vina was used for HTS of compounds because of its accuracy and the advantage of using multiple cores of CPU simultaneously. HTS of compounds was carried out in a heterogeneous Linux cluster running CentOs 5.8 with OpenPBS Batch queue system. The cluster is composed of 1 Dual Intel Xeon CPU ES-2620 (2 GHz) (12 cores), 1 Dual AMD Opteron 6344 (2,6 GHz) (24 cores) and 9 Dual Intel Xeon PU E5520 (2,2 GHz) (8 cores). The server has 108 processors, out of which 96 cores are available for use. A python script was used to extract docking energies from the vina output "all.pdbqt" file. Microsoft Excel 2010 and R-studio were used in analysis of the docking energies.

3.8.7. HTS Results and discussion

A total of 600 compounds from the SANCDB were docked into all the protein receptors using AutoDock Vina. Lead compounds were selected based on their selectivity to TbCatB and on their potential to inhibit all the *Trypanosomal* cathepsin B proteases but not the human cathepsin B protease. Based on binding (docking) energies, compounds with a docking energy of more than 1kcal/mol improvement when bound to TbCatB than to HsCatB where selected to be more specific to TbCatB. To determine this, the average docking energy of each compound across the set of *Trypanosomal* cathepsin B proteases was calculated and then the difference between this average docking energy and the docking energy in HsCatB for that particular compound was calculated. Selected compounds were further analysed in Discovery Studio analyser. From this analysis, only compounds that docked into the active site pocket of TbCatB and that of other *Trypanosomal* cathepsin B proteases were selected as lead compounds. A total of nine compounds, SANC00 478, 479, 480, 481, 482, 488, 489, 490 and 491were selected. Table 3.3 and Figure 3.7 show a list of the leads compounds and their docking energies in all the protease receptors.

Receptors Tc3CBJ, Tc3HHI and TcCatB are all homology models of *Trypanosoma congolense* Cathepsin B protease made from different templates (PDB ID: 3CBJ and PDB ID: 3HHI) and a combination of those templates. All selected lead compounds bound more strongly to the Tc3CBJ homology model than to the Tc3HHI and TcCatB models (except SANC00488 which

bound more strongly to Tc3HHI). This could be because the occluding loop in Tc3CBJ is more open than those in the other two (Tc3HHI and TcCatB) proteases and so it allows access to more interacting residues deeper in the active pocket. The same observation is made when comparing *Trypanosoma vivax* cathepsin B homology models (Tv3CBJ, Tv3HHI & TvCatB), all the compounds bound more strongly to Tv3CBJ model than to Tv3HHI and TvCatB models (except SANC 00489 which has an equal binding energy in both Tv3CBJ and TvCatB).

	3HHI	Tc3CBJ	Tc3HHI	TcCatB	Tcr3CBJ	Tcr3HHI	TcrCatB	Tv3CBJ	Tv3HHI	TvCatB	HsCatB
SANC00478	-10.3	-12	-9.7	-11.9	-10.1	-9.2	-10.2	-12.1	-8.5	-11.3	-9
SANC00479	-10.6	-12.2	-9.8	-11.6	-10	-9.5	-10	-11.7	-8.8	-10.8	-7.7
SANC00480	-10.2	-11.6	-9.4	-10	-9.3	-8.9	-10.1	-11.5	-9.1	-10.6	-8.6
SANC00481	-10.4	-11.3	-8.4	-8.8	-9.4	-9.1	-10.1	-11.3	-8.8	-11.1	-8
SANC00482	-10.2	-11.2	-9.8	-8.6	-9.2	-8.7	-9.8	-11.1	-9.1	-10.8	-8.4
SANC00488	-10.3	-10.1	-10.7	-10.5	-9	-9.5	-9.4	-10.4	-10	-9.8	-7.7
SANC00489	-10	-11.2	-10.4	-9.7	-10.1	-9.3	-10.4	-10.6	-10.5	-10.6	-8.1
SANC00490	-10.9	-12.4	-10.8	-10.4	-11.4	-9.2	-10.9	-12.4	-11.2	-10.5	-8
SANC00491	-11	-11.8	-9.8	-10.5	-12.3	-11	-10.8	-11.5	-9.9	-10.7	-8.6

Table 3.3: Showing the binding energies of leads compounds in all the receptors. All the leads compounds bind more strongly to *Trypanosome spp.* catBs than to human cathepsin B.

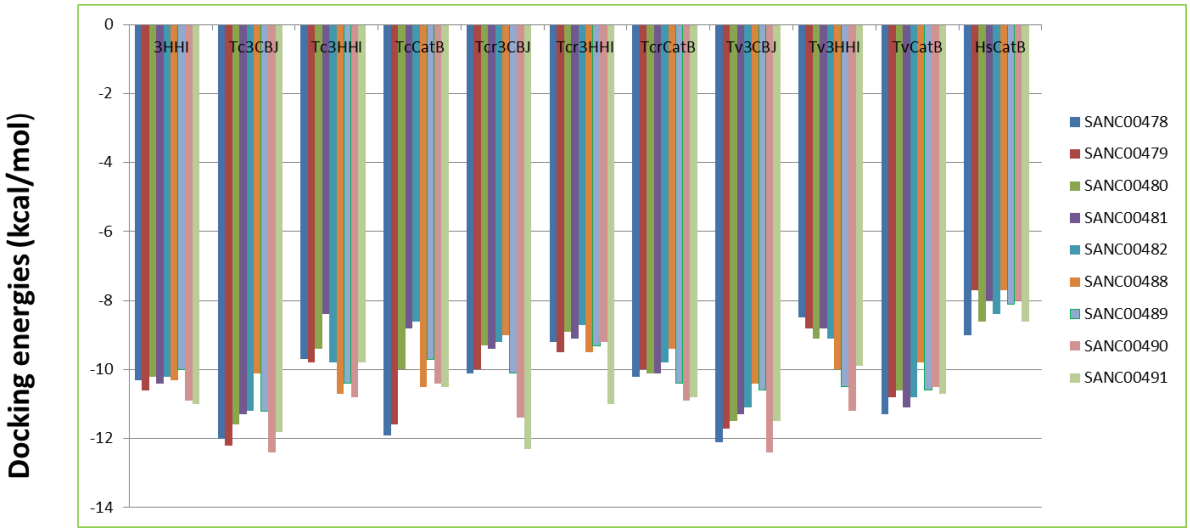


Figure 3.7: Showing the binding energies of the lead compounds in all the receptors. All the leads compounds bind more strongly to *trypanosome ssp.* catBs than to HsCatB.

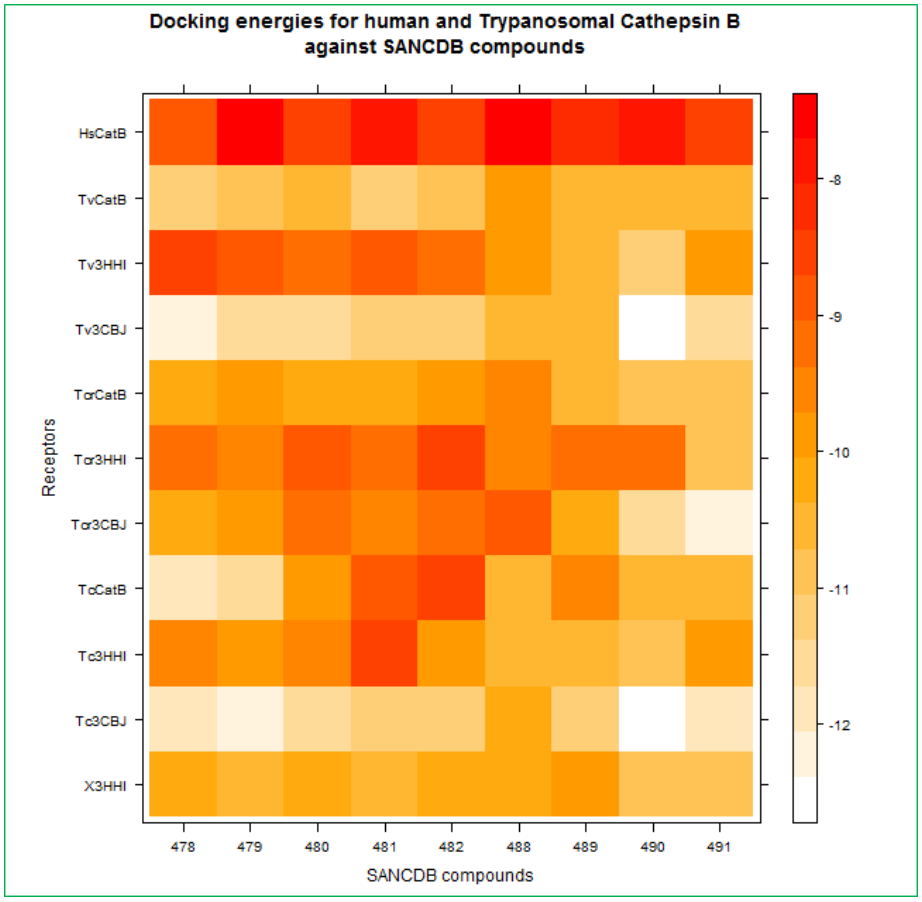


Figure 3.8: Showing the estimated binding energies of the lead compounds in all the receptors. The energy scores are colored from lowest energy/ strongest binding (white) to highest energy/ weakest binding (red). All the lead compounds bind more strongly to *Tripanosomal* catBs than to HsCatB.

However a comparison of the *Trypanosoma cruzi* cathepsin B models show that only two compounds (SANC00 490 & 491) bound more strongly to the Tcr3CBJ (more open colluding loop) than to Tcr3HHI and TcrCatB.

The heat map in Figure 3.8 also shows that the compounds have less affinity for HsCatB and more affinity for *Trypanosomal* cathepsin B proteases, with stronger affinity for Tc3CBJ and Tv3CBJ.

Structures	3HHI template	3CBJ template	3HHI_3CBJ template
TcCatB	- 1.12	- 0.90	- 1.01
TcrCatB	- 0.59	- 0.30	- 0.55
TvCatB	- 1.20	- 1.10	- 1. 39

Table 3.4: Listed are the DOPE Z-scores of the models. The DOPE Z-scores of the templates were – 1.44 and – 1.14 for 3CBJ and 3HHI respectively.

Structures	3HHI template	3CBJ template	3HHI_3CBJ template
TcCatB	0.173	0.240	0.281_0.536
TcrCatB	0.262	0.312	0.303_0.537
TvCatB	0.196	0.251	0.262_0.578

Table 3.5: Listed are the RMSD values showing the similarity between the homology models and the templates. The RMSD of the templates was found to be 0.668.

3.8.7.1. Selection of homology models for MD simulation

The remainder of this study concentrates on only one homology model of each of the *Trypanosoma congolense* cathepsin B, *Trypanosoma cruzi* cathepsin B, and *Trypanosoma vivax* cathepsin B. The models were selected based on their DOPE Z scores and RMSD scores (Tables 3.4 and 3.5). The selected models were calculated from the 3HHI template and were assigned the following abbreviations: *Trypanosoma congolense* cathepsin B (TcCatB), *Trypanosoma cruzi* cathepsin B (TcrCatB), and *Trypanosoma vivax* cathepsin B (TvCatB).

3.8.7.2. Lead compounds and receptor complexes

Figure 3.9 shows the docking pose and docking energies of SANC00478 in the protease receptors. The strongest binding energy of -10.3 kcal/mol in (Figure 3.9 B) is with TbCatB, while the weakest binding energy of -9.0 kcal/mol is with HsCatB protease. This is a difference of 1.03 kcal/mol in binding between the two proteases which indicates that the compound has a strong preference to bind to the active site of TbCatB than to HsCatB. It also binds more strongly to TcCatB, TcrCatB and TvCatB than to HsCatB.

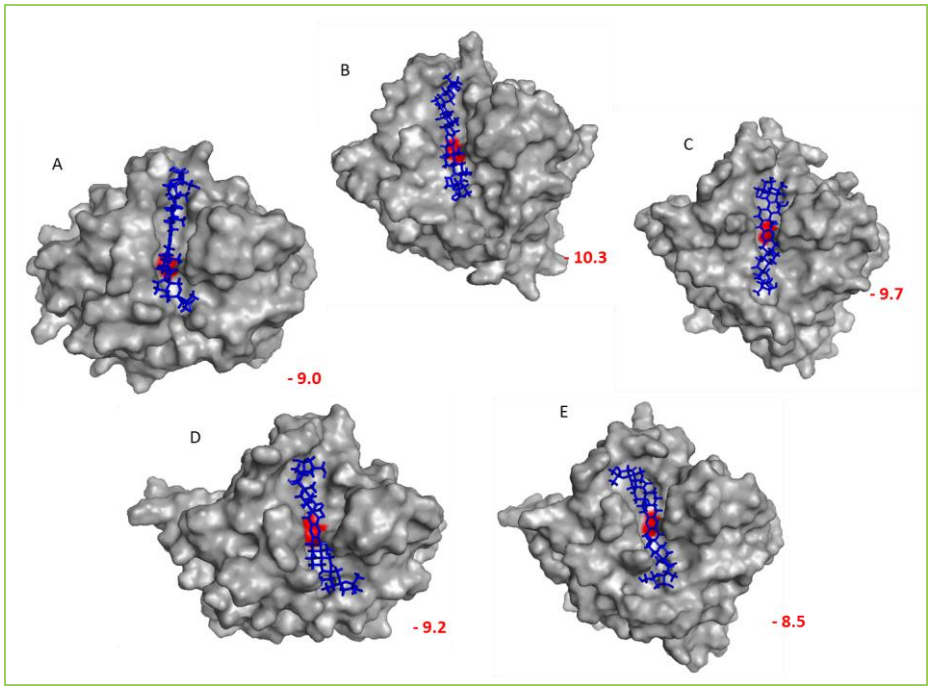


Figure 3.9: SANC00478 pose (navy blue) in (A) HsCatB (1GMY), (B) TbCatB (3HHI), (C) TcCatB, (D) TcrCatB, and (E) TvCatB. Docking energies (kcal/mol) are written below each complex. The active site Cys residue is marked in red.

The binding of SANC00478 is along the active site cleft in all the receptors. This compound is listed in the SANCDB as cephalostatin 2 and a previously known property is its anticancer activity. For more information see Appendix 1A-1.

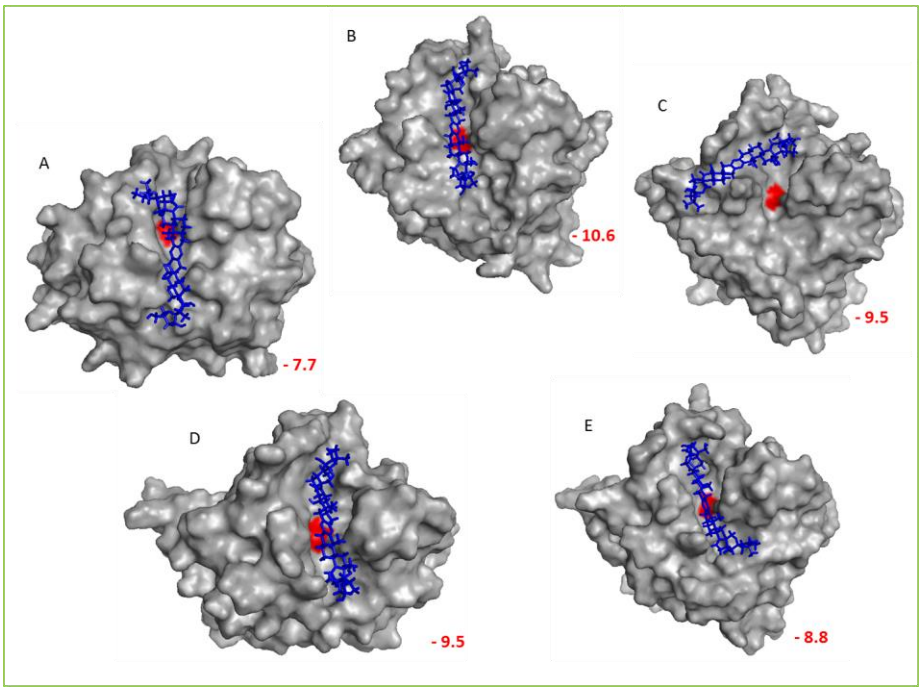


Figure 3.10: SANC00479 pose (navy blue) in (A) HsCatB (1GMY), (B) TbCatB (3HHI), (C) TcCatB, (D) TcrCatB, and (E) TvCatB. Docking energies (kcal/mol) are written below each complex. The active site Cys residue is marked in red.

SANC00479 pose and docking energies in the receptors are shown in Figure 3.10. This compound has a docking energy of -10.6 kcal/mol when bound to TbCatB and -7.7 kcal/mol when bound to HsCatB. This compound is selective for TbCatB with a docking energy difference of 2.9 kcal/mol between average binding to Tripanosomal and Human cathepsin B's. SANC00479 also prefers to bind along the active site cleft, except in TcCatB (Figure 3.10 C) where it is bound towards the occluding loop. It also binds strongly to all the *Trypanosomal* cathepsin B proteases than to HsCatB. This compound is also listed in the SANCDB as cephalostatin 3 and also has known anticancer properties. For more information see Appendix 1A-1.

In Figure 3.11 similarly illustrate the docking pose and docking energies of SANC00480. The strongest docking energy of -10.2 kcal/mol for this compound is with TbCatB while the weakest is -8.6 kcal/mol when it is bound to HsCatB, and there is a preference for binding along the active site cleft in all the proteases. It also has a stronger binding energy to bind to TbCatB and other *Trypanosomal* cathepsin B proteases than to HsCatB. This compound is listed in the

SANCDB as cephalostatin 4 and it again, like many of the cephalostatins exhibits anticancer activity. For more information see Appendix 1A-1.

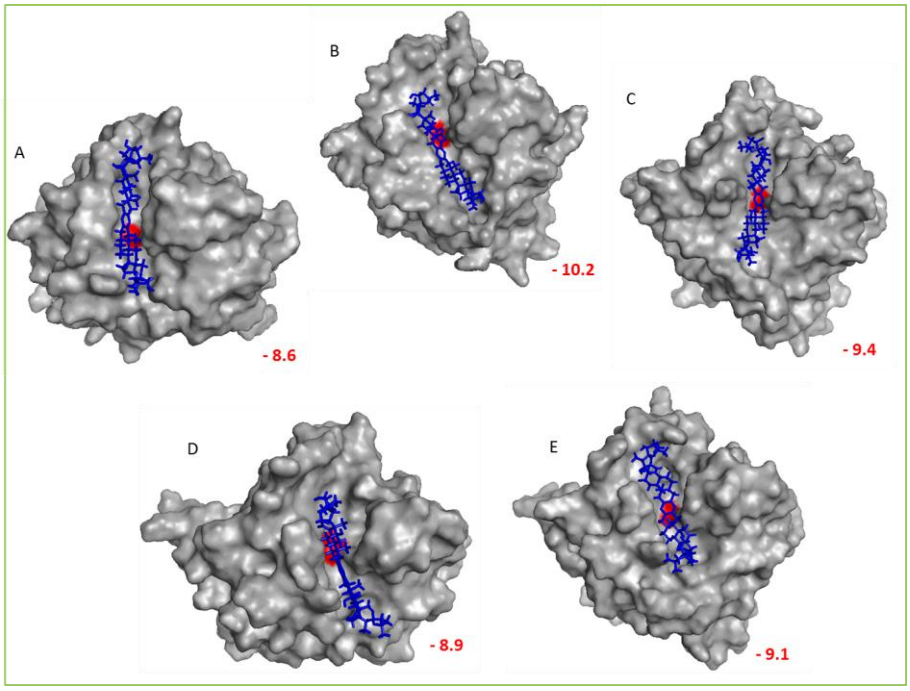


Figure 3.11 SANC00480 pose (navy blue) in (A) HsCatB (1GMY), (B) TbCatB (3HHI), (C) TcCatB, (D) TcrCatB, and (E) TvCatB. Docking energies (kcal/mol) are written below each complex. The active site Cys residue is marked in red.

SANC00481 binds along the active site cleft in all the proteases as shown in Figure 3.12. It also binds more strongly to *Trypanosomal* cathepsin B proteases than it does to HsCatB. The docking energy of this compound in TbCatB is -10.4 kcal/mol and -8.0 kcal/mol in HsCatB. A docking energy difference of 2.4 kcal/mol shows that the compound is selective for TbCatB. Again, this is a cephalostatin, cephalostatin 7 with typical anticancer propetties. For more information see Appendix 1A-1.

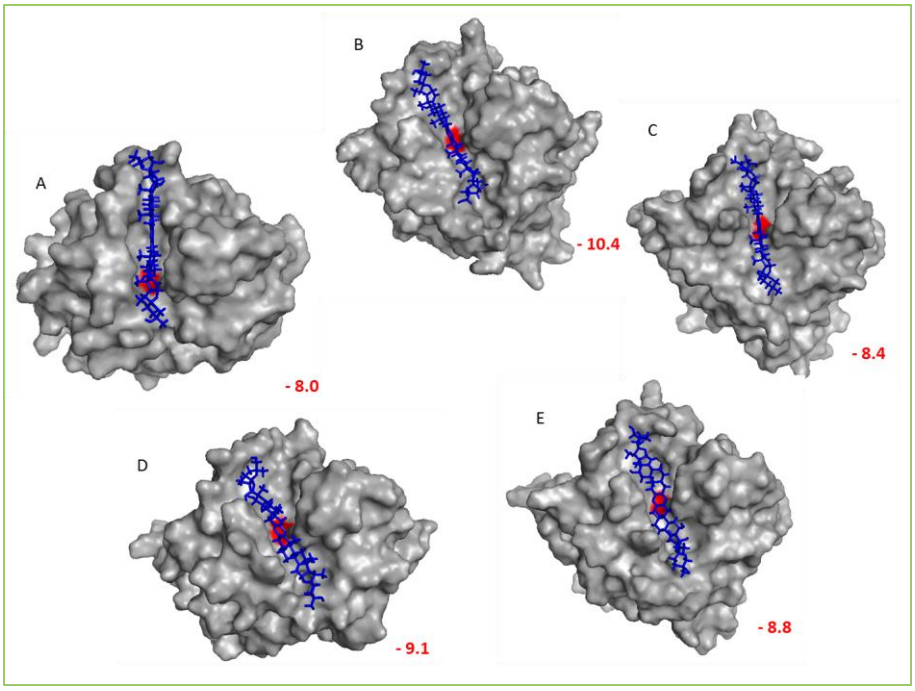


Figure 3.12 SANC00481 pose (navy blue) in (A) HsCatB (1GMY), (B) TbCatB (3HHI), (C) TcCatB, (D) TcrCatB, and (E) TvCatB. Docking energies (kcal/mol) are written below each complex. The active site Cys residue is marked in red.

SANC00482 binds similarly with these cathepsin B proteases (Figure 3.13), however with less affinity for binding to HsCatB compared to the rest of the proteases. This compound also shows potential to be specific for TbCatB. Its Docking energy is -10.2 kcal/mol in TbCatB and -8.4 in HsCatB. This compound is listed in the SANCDB as cephalostatin 8. For more information see Appendix 1A-1.

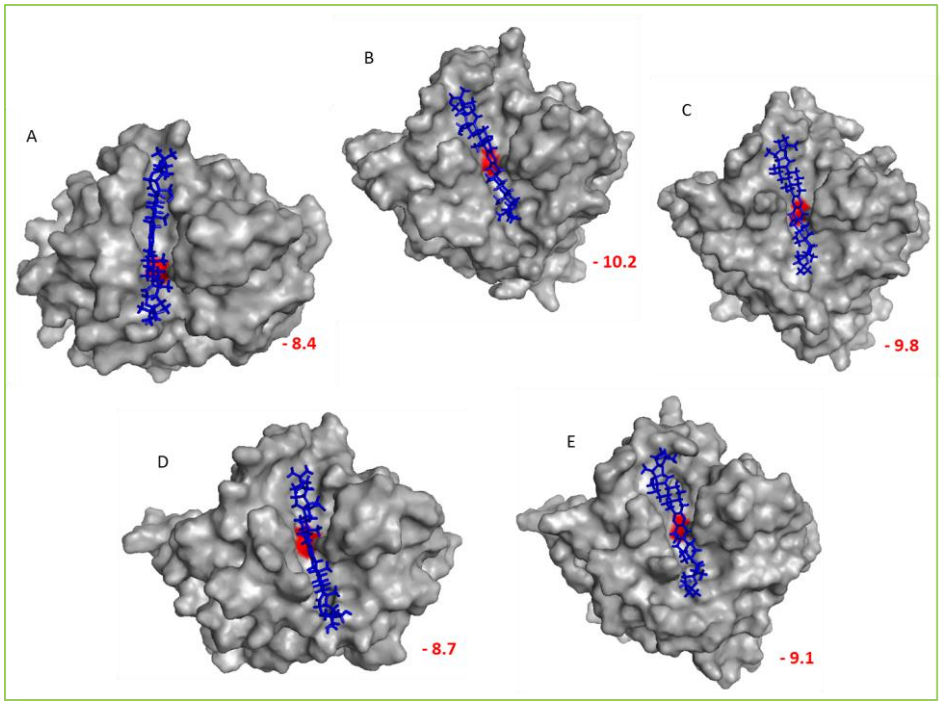


Figure 3.13 SANC00482 pose (navy blue) in (A) HsCatB (1GMY), (B) TbCatB (3HHI), (C) TcCatB, (D) TcrCatB, and (E) TvCatB. Docking energies (kcal/mol) are written below each complex. The active site Cys residue is marked in red.

The docking poses and docking energies of SANC00488 can be seen in Figure 3.14. The compound binds along the active side of all the *Trypanosomal* cathepsin B proteases but does not bind along the active site of HsCatB. The binding energy of this compound when it is bound to HsCatB is -7.7 kcal/mol, which is one of the weakest binding of all the leads compounds.

SANC00488 shows a different pattern of binding compared to the previously discussed compounds among the *Trypanosomal* cathepsin B proteases. It binds to TcrCatB (Figure 3.14 D) with a binding energy of -9.5 kcal/mol and the strongest is when it is bound to TcCatB with a docking energy of -10.7 kcal/mol. The binding to TcCatB and TbCatB is also good. This shows that the compound has marked preference for binding to the *Trypanosomal* cathepsin B proteases and it also has a potential to act as a broad spectrum inhibitor of the current set of *Trypanosomal* cathepsin B proteases. This matches the observed geometries of binding where it appears as though SANC00488 is unable to fit within the cleft of HsCatB which also has a much lower binding affinity. Based on these observations, these complexes (Figure 3.14) shall be analysed further to determine protein residues that interact with the compound. Further analysis of these complexes shall be carried out using molecular dynamics. SANC00488 is listed as cephalostatin 14 in the SANCDB. More information on this compound is in appendix 1A-2.

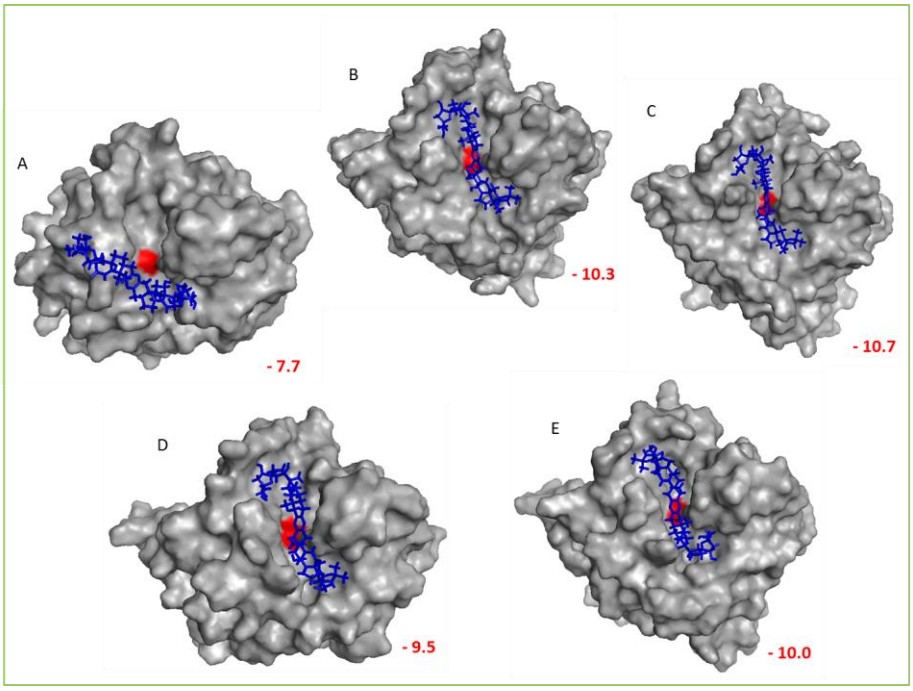


Figure 3.14 SANC00488 pose (navy blue) in (A) HsCatB (1GMY), (B) TbCatB (3HHI), (C) TcCatB, (D) TcrCatB, and (E) TvCatB. Docking energies (kcal/mol) are written below each complex. The active site Cys residue is marked in red.

SANC00489 also shows potential to be used as a broad spectrum inhibitor (Figure 3.15). This compound prefers to bind along the active site cleft of the cathepsin B proteases, (and also HsCatB, although with poorer binding affinity), except for TcCatB protease (Figure 3.15 C) where is prefers to bind to a position closer to the occluding loop. SANC00489 is cephalostatin 15. Appendix 1A-2 has more information on the compound.

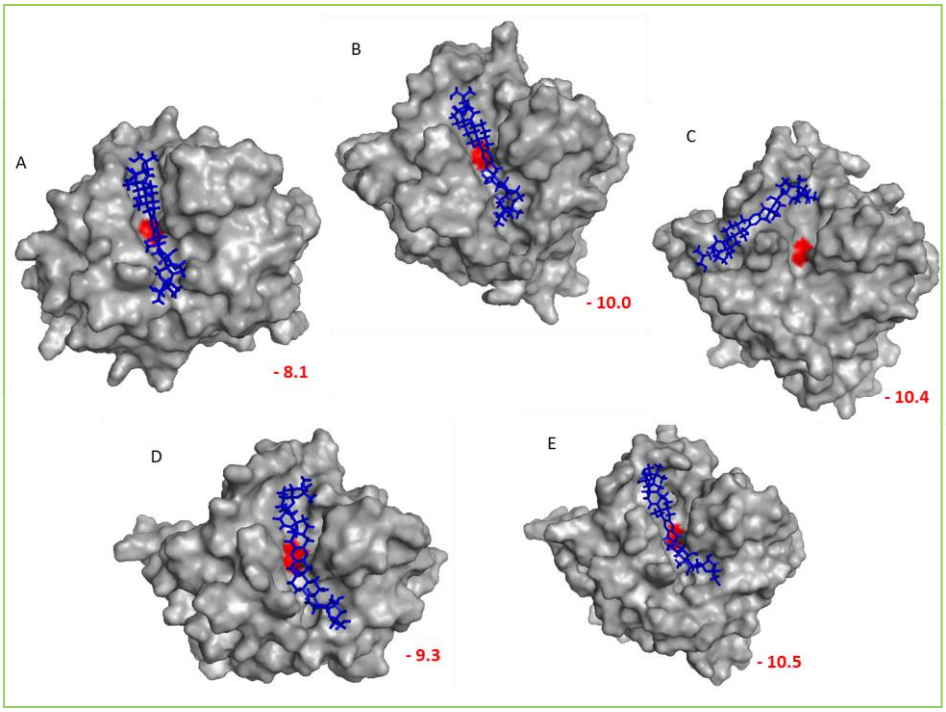


Figure 3.15 SANC00489 pose (navy blue) in (A) HsCatB (1GMY), (B) TbCatB (3HHI), (C) TcCatB, (D) TcrCatB, and (E) TvCatB. Docking energies (kcal/mol) are written below each complex. The active site Cys residue is marked in red.

SANC00490 (cephalostatin 16) and SANC00491 (cephalostatin 17) show an interesting variation in binding along the active site of cathepsin B proteases. In TcCatB, for instance SANC00490 is bound towards the occluding loop position. The docking energies also show that these compounds have strong affinity for *Trypanosomal* cathepsin B proteases rather than for HsCatB (Figure 3.16 and Figure 3.17). More information can be found in appendix 1A-2

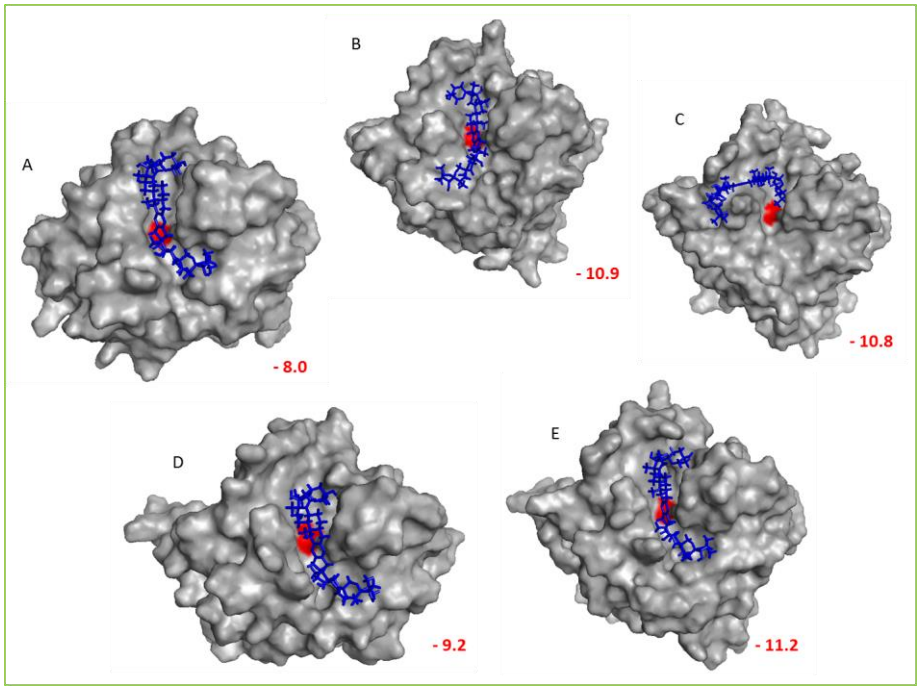


Figure 3.16 SANC00490 pose (navy blue) in (A) HsCatB (1GMY), (B) TbCatB (3HHI), (C) TcCatB, (D) TcrCatB, and (E) TvCatB. Docking energies (kcal/mol) are written below each complex. The active site Cys residue is marked in red.

Figure 3.18 shows a closer look at the binding energies of SANC00488 in TbCatB, TcCatB, TcrCatB, TvCatB and HsCatB. The compound clearly has less affinity to HsCatB than to the other proteases. A closer look at the interactions between this compound and protein residues will reveal the contributing factors to this bindings. In Figures 3.18 to 3.23, we look at the important residues that interact with SANC00488 in our protein receptors.

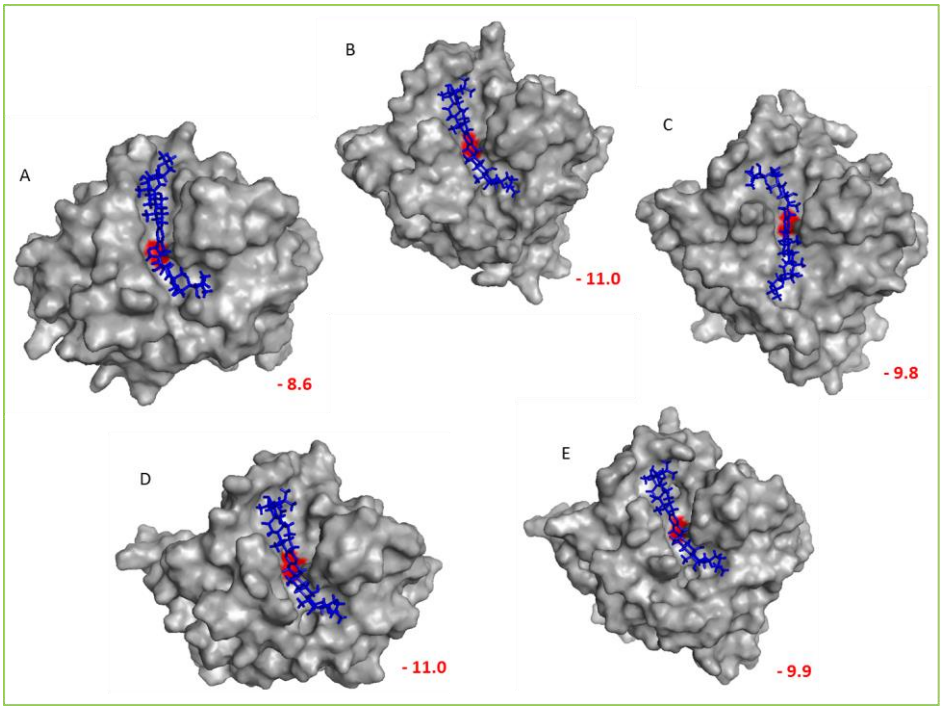


Figure 3.17 SANC00491 pose (navy blue) in (A) HsCatB (1GMY), (B) TbCatB (3HHI), (C) TcCatB, (D) TcrCatB, and (E) TvCatB. Docking energies (kcal/mol) are written below each complex. The active site Cys residue is marked in red.

SANC00488 interacts with HsCatB by forming hydrophobic and hydrogen bonds with S2 subsite residues. The compound forms a π-alkyl hydrophobic bond of 4.36Å with Pro76 in the S2 subsite. Another alky hydrophobic interaction of 3.53Å is formed between the compound and Ala173. Two hydrogen bonds which are 2.71Å and 2.84Å long are formed between the compound and Glu245 and Asn72 residues respectively. Interaction with the S2 subsite residues, especially Glu245 presents an opportunity to design derivatives of this compound with even less affinity for HsCatB. The S2 subsite determines specificity of this protease and it differs with TbCatB at the bottom of the S2 subsite where the Glu245 residue is substituted for Gly328.

SANC00488 interacts with Phe208, Ala118, Cys162, Asn163 and Ala256 residues in TbCatB. Interaction with the Phe208 residue is by formation of a hydrogen bond of 1.87Å. Another hydrogen bond of 1.98Å is formed between the compound and Asn163 residue. The Ala118 residue forms both an alkyl hydrophobic bond of 3.68Å and a hydrogen bond of 2.78Å with the compound. The Cys162 and Ala256 form alkyl hydrophobic interactions of respectively 4.73Å and 4.12Å with the compound. The compound forms three hydrogen bonds with TbCatB as opposed to only two formed with HsCatB. Two of the H-bonds formed between SANC00488

with TbCatB are also shorter than those it formed with HsCatB. These differences contributed to the compound having a stronger affinity for TbCatB than to HsCatB.

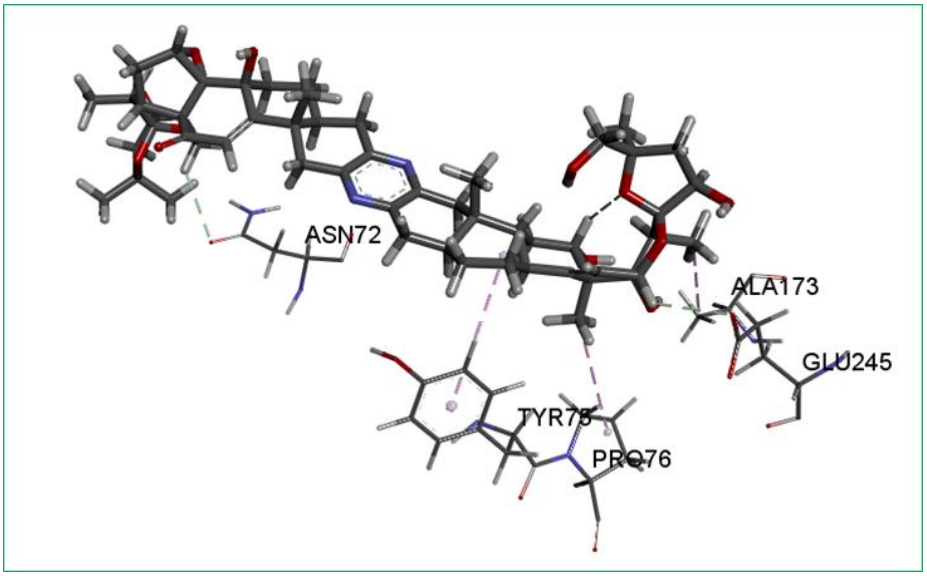


Figure 3.18: Showing SANC00488 interactions with important residues in HsCatB (1GMY). Bond interactions are colour coded; hydrophobic interactions (light purple), electrostatic interactions (brown), non-classical H-bonds (light grey), classical H-bonds (green).

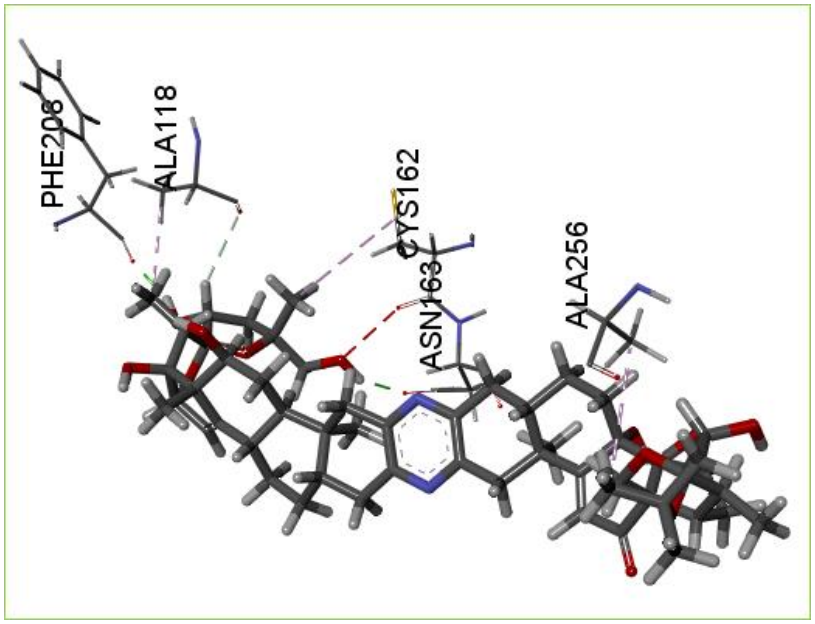


Figure 3.19 Showing SANC00488 interactions with important residues in TbCatB (3HHI). Bond interactions are colour coded; hydrophobic interactions (light purple), electrostatic interactions (brown), non-classical H-bonds (light grey), classical H-bonds (green).

Two hydrogen bonds and alkyl hydrophobic interactions are formed between SANC00488 and TcCatB protease (Figure 3.20). The hydrogen bonds which are respectively 2.10Å and 3.08Å

long are formed between the compound and Tyr134 and Lys90 residues. Hydrophobic interactions are formed between interactions with Ala45 and Cys89.

Bonding between SANC00488 and TcrCatB is stabilized by hydrogen bonds between the compound and Gly181, Ser21, Cys65 and Tyr69 residues. Hydrophobic interactions between the compound and His89, Cys107, and Tyr69 residues also contribute to stability of the complex (Figure 2.21).

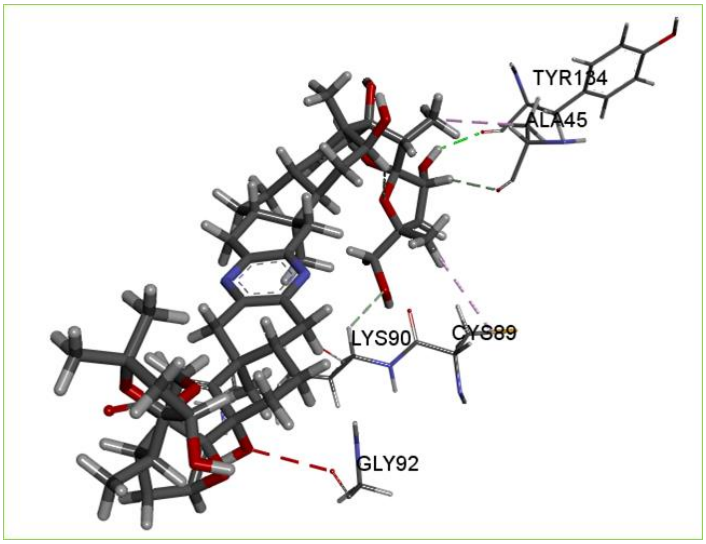


Figure 3.20 Showing SANC00488 interactions with important residues in TcCatB. Bond interactions are colour coded; hydrophobic interactions (light purple), electrostatic interactions (brown), non-classical H-bonds (light grey), classical H-bonds (green).

In TvCatB, the docked SANC00488 forms hydrophobic interactions with His131, Tyr89, Pro90 and Ala178. Only one hydrogen bond is formed between the compound and Gln250 residue (Figure 3.22).

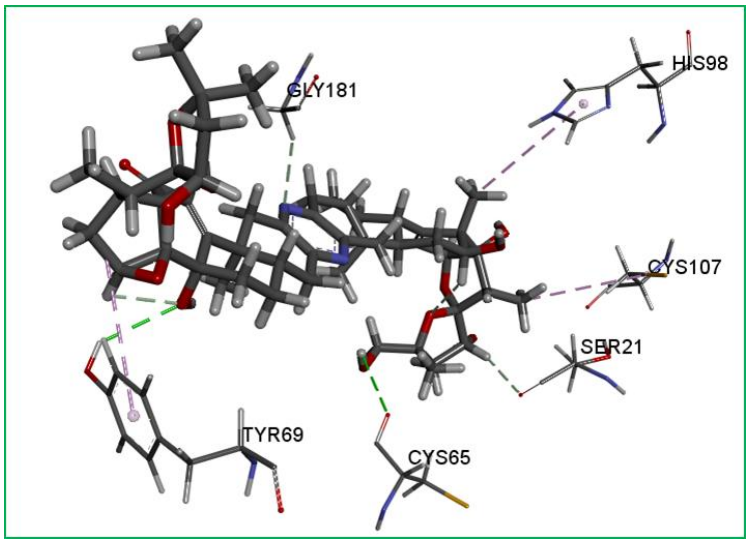


Figure 3.21: Showing SANC00488 interactions with important residues in TcrCatB. Bond interactions are colour coded; hydrophobic interactions (light purple), electrostatic interactions (brown), non-classical H-bonds (light grey), classical H-bonds (green).

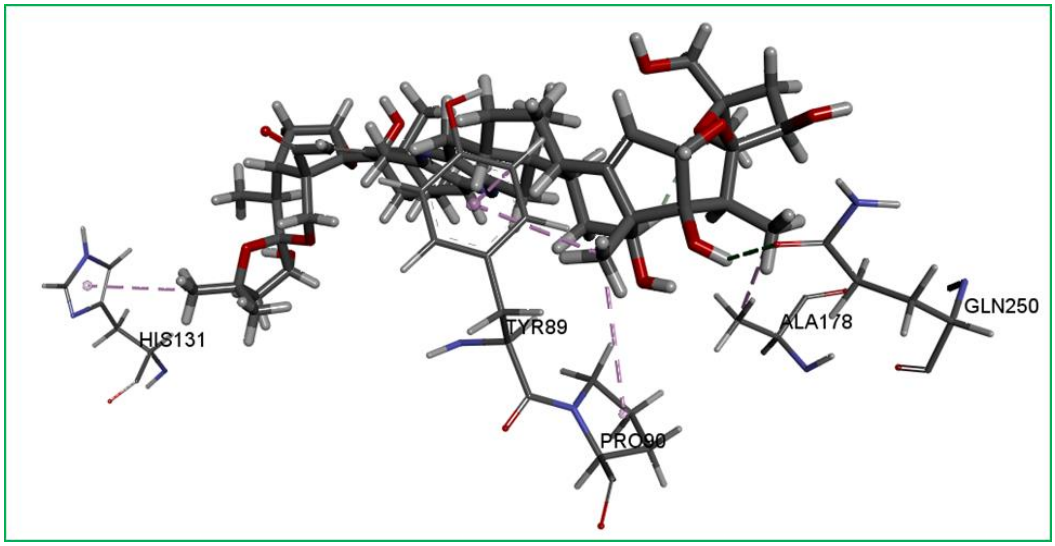


Figure 3.22: Showing SANC00488 interactions with important residues in TvCatB. Bond interactions are colour coded; hydrophobic interactions (light purple), electrostatic interactions (brown), non-classical H-bonds (light grey), classical H-bonds (green).

3.9. Conclusion

Six hundred natural compounds of South African origin were screened into HsCatB, TbCatB, TcCatB, TcrCatB and TvCatB proteases using AutoDock Vina. The compounds SANC00 478, 479, 480, 481, 482, 488, 489, 490 and 491were selected as potential leads for drug development based on their strong affinity for *Trypanosomal spp.* cathepsin B proteases. These natural compounds, which are all cephalostatins, have a higher affinity for binding to *Trypanocidal* cathepsin B proteases than to HsCatB. SANC00488 has the highest affinity for *Trypanocidal* cathepsin B proteases and the lowest affinity for HsCatB.

Chapter 4.

4.1. Molecular dynamics simulation

In chapter 3, 600 compounds were screened from the South African Natural Compounds Database (SANCDB) into Human cathepsin B (HsCatB), *Trypanosoma brucei* cathepsin B (TbCatB), *Trypanosoma congolense* cathepsin B (TbCatB), *Trypanosoma cruzi* cathepsin B(TbCatB), and *Trypanosoma vivax* cathepsin B(TbCatB) proteases. A total of nine lead compounds were selected for displaying more affinity to inhibit *Trypanosomal species* cathepsin B proteases than HsCatB. This selection was based on the binding energy of the compounds to the proteases. SANC00488 was observed to show more affinity to *Trypanosomal species* cathepsin B proteases and less affinity to HsCatB than the other lead compounds. The complexes with the proteases were selected for further analysis using Molecular Dynamics (MD) simulations. Using Discovery Studio, ligand receptor interacting residues were determined for the complexes. MD simulations were carried out to understand the behaviour (conformational changes and fluctuations) of these interactions over time under conditions that are closer to physiological conditions.

4.2. Introduction

Molecular dynamics simulations are used widely in the study of biological macromolecules [110]. Two rather similar techniques used for simulation are molecular dynamics (MD) and Monte Carlo (MC). Simply molecular dynamics (MD) is a computer simulation of the physical movements of atoms and molecules in the context of N-body simulation (https://en.wikipedia.org/wiki/Molecular_dynamics).

MD was first introduced by Alder and Wainwright in the late 1950's to study the interactions of hard spheres [111] and important insights related to the behaviour of simple liquids emerged from their studies. The first MD simulation of a realistic system was then carried out by Rahman and Stillinger in their simulation of liquid water in 1974. The first protein MD simulation which was carried out on a protease (bovine trypsin inhibitor) with just 58 residues and ~450 atoms was done for 8.8 psec *in vacuo* [112]. Increases in computer power over the years has allowed for simulation of large systems containing 10^4-10^6 atoms. Improvements in force fields, treatment of long electrostatic interactions and system boundary conditions, and better algorithms for temperature and pressure control make it possible to make simulations of more realistic systems that include explicit water molecules, counter ions and membrane like environments. Although parameters for protein interaction are found in modern force fields, the parameter for ligands are often not adequately represented in these sets.

MD simulations are carried out to gain information about the properties of the way molecules are assembled in relation to their structure and their microscopic interactions [113]. MD simulations are important to understand fluctuations and behaviour of interactions between biological molecules and to show their conformational changes over time [110]. Biological systems are complex; so computer methods have become important in life sciences since MD simulations make it possible to study the effect of explicit solvent molecules on protein structure and stability over a certain period of time [110]. As an example, MD simulations were used to examine conformational changes of GPR40 within the hydrated lipid environment [114]. The researchers used MD simulation to show that agonists can bind to the active site of GPR40 and that simulation may alter the original docked pose. They were able to identify residues that are critical for GPR40 and ligand binding. It is worthy to note that MD simulations are also used for refining X-ray or NMR structures [110].

In drug design, MD simulation can be combined with docking to predict more reliably stable protein-ligand complexes. Structures obtained after MD simulation are more representative of conformations available for binding with inhibitors in solution [115]. During the drug design

process, docking is used to screen large libraries of drug like compounds over a short period of time to reduce them to a reasonable number of hits. The lack of or poor flexibility of the protein during ligand binding and the absence of a scoring function that is universal make docking results incomplete on their own. Since during MD simulation both protein and ligand can be treated as flexible, the protein can change its conformation to the bound ligand and produce a more induced fit and accurate free energies of binding maybe calculated. During the simulation, incorrectly docked complexes produce an unstable trajectory while correct complexes produce a stable trajectory.

4.3. MD simulation methods

The two main families of MD simulation are molecular 'classical' mechanics and 'quantum' mechanics simulations.

In the 'quantum' or 'first-principles' MD simulations the quantum nature of the chemical bond is taken into account. Although they are useful in providing information of biological systems, quantum MD simulations require more computational resources limiting their use to short simulations of very small systems. In particular, in order to simulate the progress in a simulation over the order of nanoseconds or to deal with biological macromolecules the use of a purely quantum approach is unfeasible. In these cases classical MD is most practical for simulations of biological systems for periods of nanoseconds [111].

In molecular 'classical' mechanics simulation, molecules are treated as classical objects. Atoms represent soft balls and bonds represent elastic sticks in what is termed a force field. This system's dynamics are defined by the laws of classical mechanics [111].

4.3.1. Molecular 'classical' mechanics simulation steps

Dynamics simulations are typically run in water where periodic boundary conditions and the size of the simulation box are set. In the case of proteins the systems is neutralised by adding ions or cations and then topologies are generated using an appropriate force field. The most commonly used force fields in MD simulation are AMBER, AMOEBA, CHARMM, NAMD and GROMOS force fields. The system is then minimised to relax the system to avoid having atoms forced out of the trajectory since forces in the system can result in displacement of some bonds. During minimisation the temperature of the system is lowered towards 0 K, so energy is added to the system to bring it to the operating temperature (heating). The system is then equilibrated to allow the structures to expand or contract. After removal of artefacts, the system in ready for production dynamics. The dynamics of the system can then be observed after

running production dynamics for a period of time (ns). Analysis of trajectories can be carried out visually by using a molecule visualisation system like PyMOL or Visual Molecular Dynamics (VMD). Graphical representation of results can be plotted using Xmgrace of Grace 5.1.21. If the trajectories have not stabilised it might be necessary to increase the time for production dynamics.

Available MD simulations packages include (i) CHARMM which uses the CHARMM force fields (E/I; All Atom/United Atom), and Amber force fields, (ii) Amber which uses Amber (A/I ; All Atom) force fields, (iii) GROMOS which uses Gromos (E / vacuum ; United Atom) force fields and NAMD package which uses CHARMM, Amber and Gromos force fields.

4.4. GROMACS 4 package

Gromacs is a free software package that is used to perform molecular dynamics for systems with up to millions of particles. It was designed for biological molecules like proteins, nucleic acid and lipids. It uses an interface with command line options for input and output files. Topologies and parameter files are written in text format. During simulation, you can monitor the progress of the simulation and it also tells you the expected finish time and date. To reduce the edge effect in the system, Gromacs uses the concept of periodic boundary condition. Simulations can be run in a triclinic, cubic and octahedral model boxes [112]. Gromacs may also be run in parallel using the standard MPI communication protocol or via the “Thread MPI” library for single node workstations. GROMACS 4 also uses dynamic load balancing to improve performance for protein simulations [116]. A selection of tools is provided for trajectory analysis and the output is in the form of finished Xmgr/Grace graphs with labelled axis and legends. It is developed by the GROMACS development team at, Uppsala University & The Royal Institute of Technology, Sweden. http://www.gromacs.org.

4.5. Methods and materials

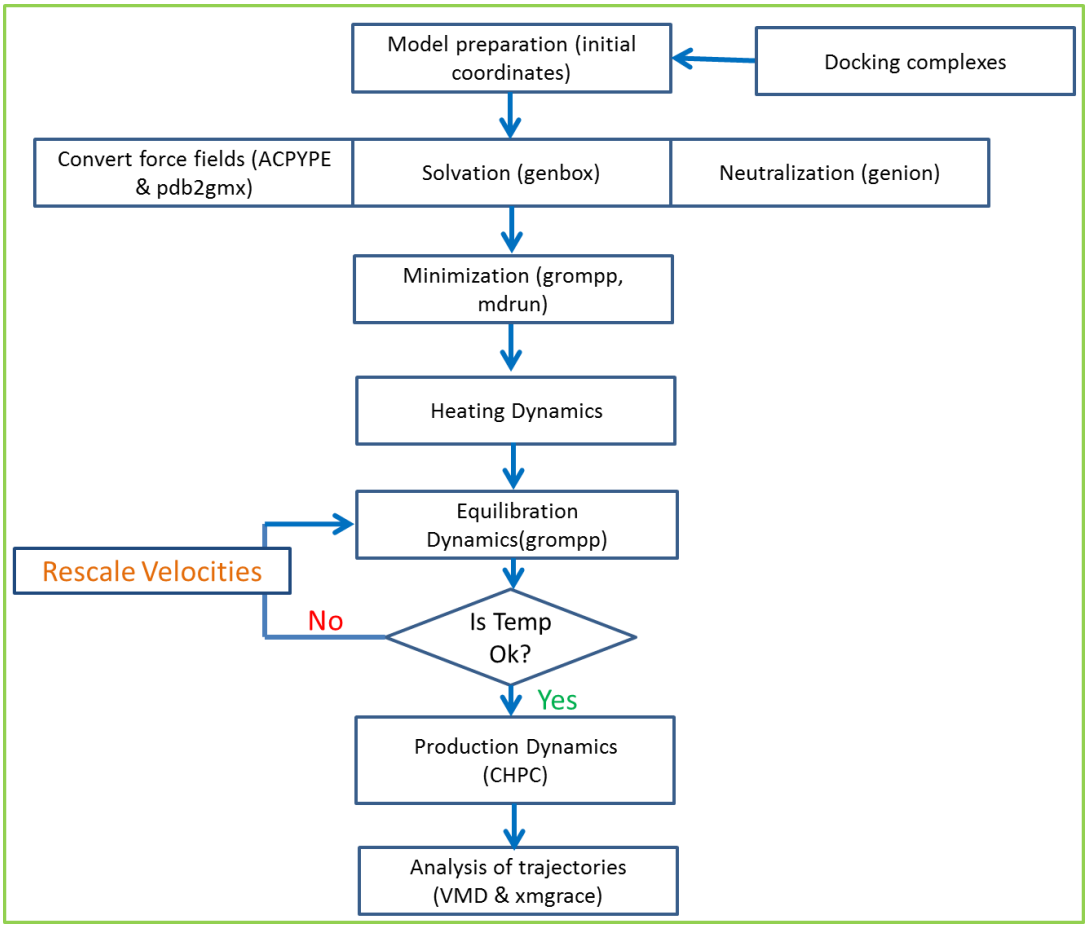


Figure 4.0: An overview of the methods used for MD studies. Complexes of SANC00488 with HsCatB, TbCatB, TcCatB, TcrCatB and TvCatB were used.

4.5.1. Data preparation

Docking and selection of complexes for MD simulation was carried out as described in chapter 2. A total of nine compounds, SANC00 478, 479, 480, 481, 482, 488, 489, 490 and 491were selected as lead compounds based on their affinity for *Trypanosomal spp.* cathepsin B proteases. From this list SANC00488 was selected as the best lead. In this chapter MD simulations of SANC00488 in complex with Human cathepsin B (HsCatB), *Trypanosoma brucei* cathepsin B (TbCatB), *Trypanosoma congolense* cathepsin B (TcCatB), *Trypanosoma cruzi* cathepsin B (TcrCatB) and *Trypanosoma vivax* cathepsin B (TvCatB) were carried out to study the complex structures under conditions that resemble physiological conditions and to predict the binding mode of the compound in the structures.

The complexes were opened in Discovery Studio Visualizer and polar hydrogens added and the protonation state of the proteins set to pH 5.0. Python scripts were then used to split the ligand-receptor complex and save them in different files.

A python script was written and used to run ACPYPE interface and Open Babel to process ligand PDBs in to GROMACS format files (topology files).

Another python script was used to process the protein and to combine the protein and ligand topology in preparation for solvation and boxing, energy minimisation, equilibration (NVT and NPT) and production run. This script was originally written within the research group [101], but was modified for these experiments. Production dynamics was carried out at the Centre for High Performance Computing (CHPC) using scripts that were available to the research group [101].

MD simulations were performed using the GROMACS 4.5.7 package with the AMBER96 force field. The ligand-protein complex was solvated and neutralized in a triclinic box of 17.5 Å filled with generic equilibrated 3-point solvent water (spc216.gro). The complex was placed at least 3 nm from the edge of the box. The whole system was neutralised by adding (0.15 M) Na^+ and Cl^- counter ions to replace water molecules. The energy of the complex was minimized to 1000 kJ/mol/nm using the steepest descent approach of 100000 steps without constraint. The system was allowed to equilibrate at a constant temperature of 300K with each minimised ensemble system equilibrated in the canonical ensemble for 200 ps (nsteps = 100000), through the NVT ensemble and for 200ps in the isothermal-isobaric ensemble. The conditions for the NPT were set at 1.0 bar of pressure with a pressure coupling constant of 2.0 ps. To simulate water the isothermal compressibility values were set to 4.5×10^{-5} /bar. Simulation was set to run for 20 ns with an integration time step of 2 femtoseconds (fs) at constant temperature and pressure. The LINCS algorithm was used to constraint all bond lengths during equilibration and production. For neighbour searching, the cut off distance was set to 1.4 nm for Coulomb and van der Waals interactions. Electrostatic interactions were approximated by the Particle Mesh Ewald for long range electrostatics with a 0.16 nm Fourier grid spacing and a fourth order cubic interpolation. Trajectories were saved every 2 ps during simulation. The trajectories were analysed using Xmgrace of Grace 5.1.21 to plot the MD graphical displays. To visualise the trajectories, the Visual Molecular Dynamics (VMD) program version 1.9 was used.

4.5.2. Results and discussion

Comparison of simulation results with experimental results were carried out in reference to Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuation (RMSF). Cα atom RMSD comparisons were carried out relative to the energy minimised starting structure. RMSD analysis was carried out using xmgrace. Visual analysis of conformational changes was carried out using VMD. Although the system was set to simulate for 20 ns, the production MD was terminated after 14 ns due to the job control system at the CHPC (set to 100 hrs) before

completion of the simulation. Since the behaviour over 13 ns was deemed sufficient, our analyses are for a 13 ns simulation.

Figure 4.1 shows the Cα atom RMSD of Human cathepsin B (HsCatB), *T. brucei* (TbCatB), *T. congolense* (TcCatB), *T. cruzi* (TcrCatB) and *T.vivax* (TvCatB) receptors in complex with SANC00488 (ligand) as a function of the simulation time (13 ns of the final molecular dynamics).

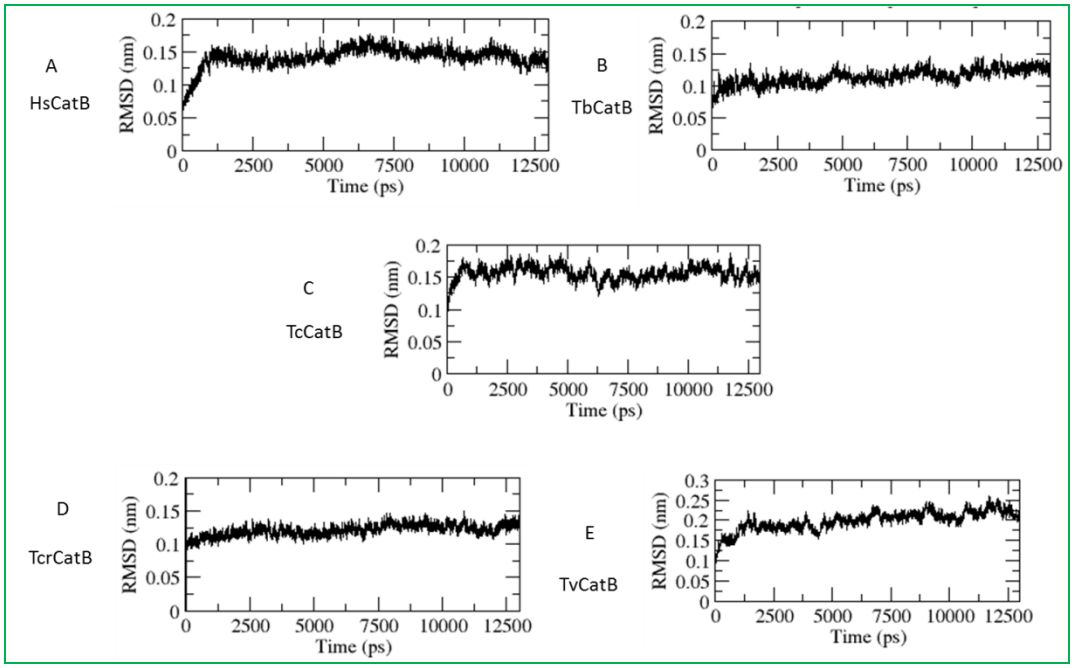


Figure 4.1: Showing comparison of Cα atom RMSD (relative to the energy minimized starting structure) as a function of time for SANC00488 in complex with (A) HsCatB, (B) TbCatB, (C) TcCatB, (D) TcrCatB and (E) TvCatB cysteine proteases along 13 ns MD simulations.

The obtained Cα atom RMSD values for the HsCatB protease is about 1.6 Å and is relatively stable after 1.25 ns (Figure 4.1 A). This indicates that after an initial rise in the RMSD of the system it reached equilibrium. The RMSF indicates that areas that fluctuated the most were occluding loop region and other loops in the structure (Figure 4.2 A). The Cα atom RMSD of TbCatB gradually increased until it reached stability at around 1.25 Å. Compared to HsCatB, the occluding loop of TbCatB contains an extra Lys197 residue and the RMSF shows that the area with the most fluctuations corresponds to a region that contains this residue. This is shown by the high peak obtained for the RMSF of residues Lys197, Ser 198, Lys 199 and Asn200 of the occluding loop area in Figure 4.2 B.

The Cα atom RMSD of TcCatB reaches stability around 1.7 Å after 1.25 ns and its RMSF is stable. Both the Cα atom RMSD of TcrCatB and TvCatB reached stability at 1.25 Å and 2.00 Å respectively. The RMSF of the occluding loop region in TvCatB is the most fluctuating part (Figure 4.2 E).

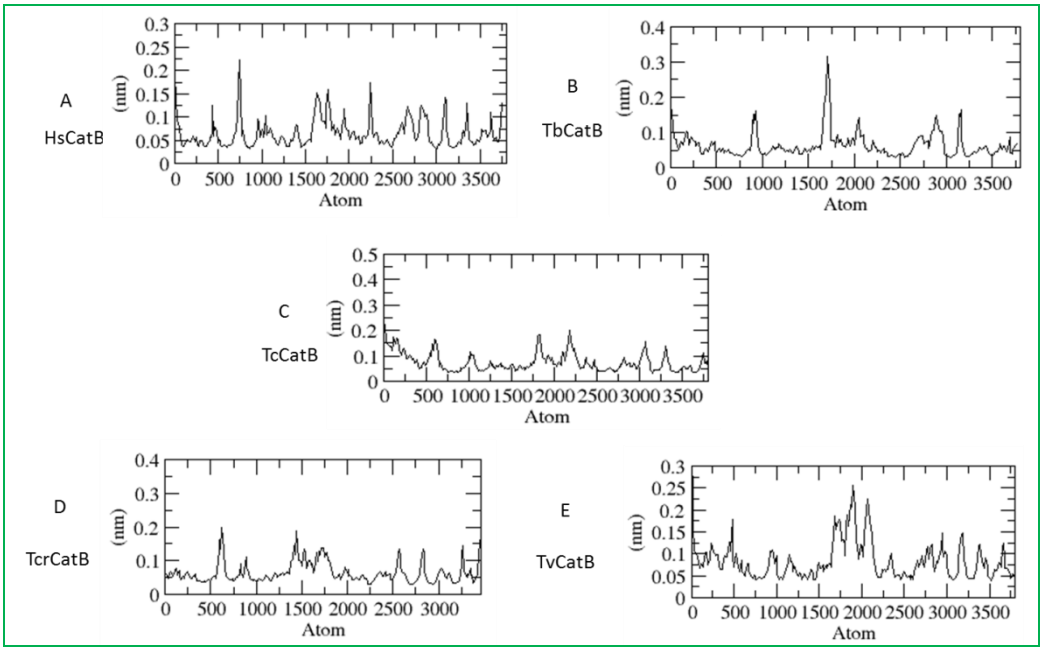


Figure 4.2 showing comparison of Cα atom RMSF (relative to the energy minimized starting structure) as a function of time for SANC00488 in complex with (A) HsCatB, (B) TbCatB, (C) TcCatB, (D) TcrCatB and (E) TvCatB cysteine proteases along 13 ns MD simulations.

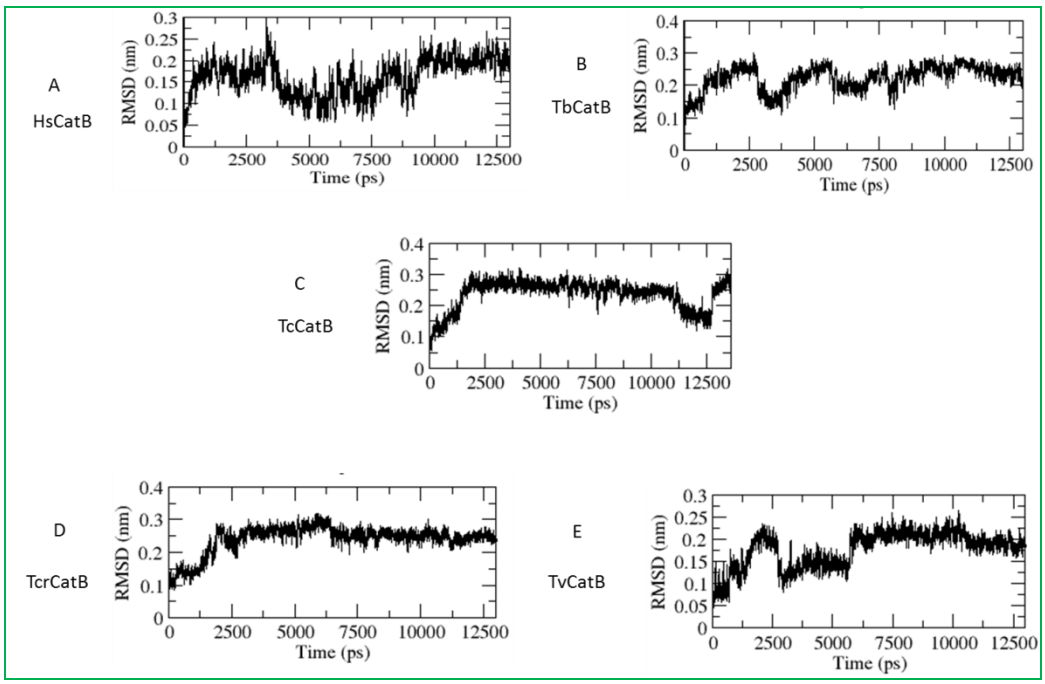


Figure 4.3: Showing comparison of SANC00488 RMSD (relative to the energy minimized starting structure) as a function of time when in complex with (A) HsCatB, (B) TbCatB, (C) TcCatB, (D) TcrCatB and (E) TvCatB cysteine proteases along 13 ns MD simulations.

The RMSD of SANC00488 bound to HsCatB can be seen in Figure 4.3 A. the RMSD gradually rose until it reached stability at around 2.3 Å after 10 ns. Visual observation of the conformational changes during simulation revealed that the ligand was moving further from its initial position and further from the active site and interacting residues (Figure 4.4. A-D). At the beginning of the simulation, the ligand is interacting with Asn72 and Gl 245 by hydrogen bonding. It also forms a hydrophobic interaction with Ala173. π-π-hydrophobic interactions are also formed between the ligand and Pro76 and Tyr75 residues. The initial lengths of the bonds between the ligand and As 72, Glu245 and Ala173 were 2.53 Å, 3.48 Å, and 3.52 Å respectively (Figure 4.4 A). By the end of the simulation, the distances between the residues and the corresponding ligand interaction atoms were respectively 20.16 Å, 16.26 Å and 9.32 Å (Figure 4.5). This suggested that the docking of the ligand in this complex was not a stable arrangement at all.

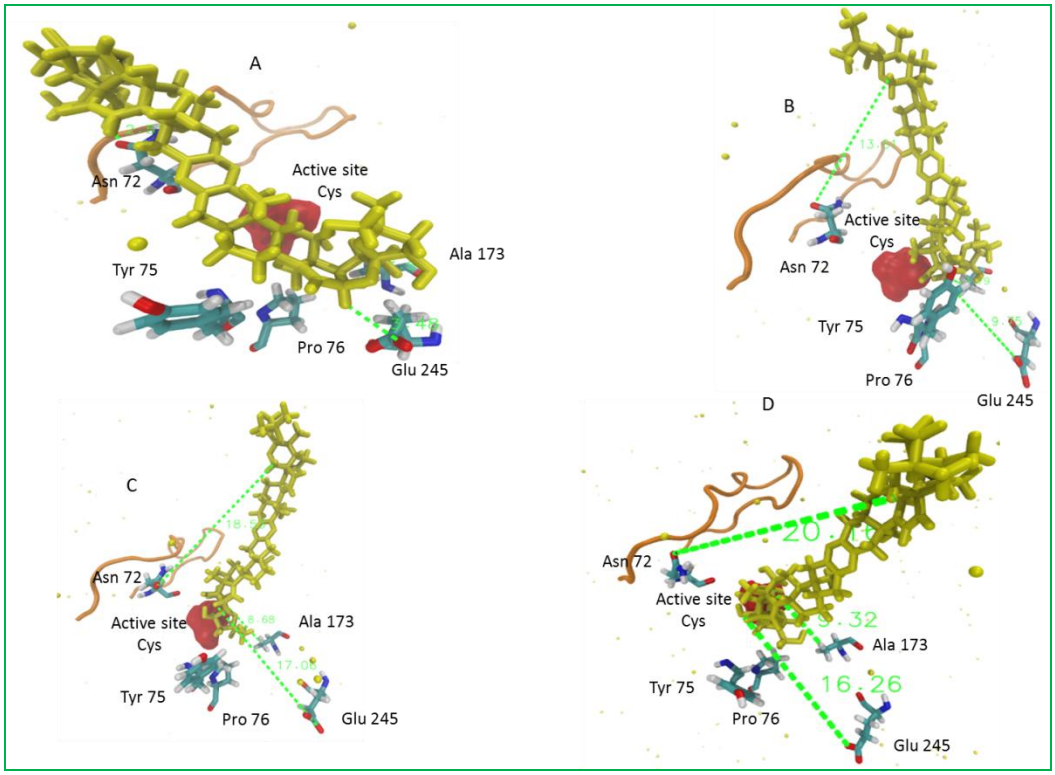


Figure 4.4: SANC00488 (Yellow) and interacting HsCatB residues during 13 ns simulations. The figures were extracted at (A) 0 ns, (B) 3 ns, (C) 6 ns and (D) 12 ns. The active site Cys residue is shown in red and the occluding loop is shown in orange. Distances to the appropriate active site residues are shown in green. The ligand is seen to move away from interacting residues during simulation.

The RMSD of SANC00488 bound to TbCatB reached equilibrium with RMSD 2.5 Å (Figure 4.3 B) after about 7.5 ns. Visual inspection of the ligand during 13 ns simulation revealed the ligand changing conformation during the simulation (Figure 4.5 A-D). At the beginning of the simulation (Figure 4.5 A), the ligand is hydrogen bonded to Phe208 and Ala118. The ligand is also bent is such a way that it is able to make hydrophobic interactions with Cys162 and also form a hydrogen bond with Asn163. The ligand also has hydrophobic interactions with Ala256 in the S2 subsite region. After about 3 ns of simulation, the ligand straightens up at the bend, placing the corresponding interacting atoms far for interactions with Cys162 and Asn163 (Figure 4.5 C and D). In this position, the ligand pushes the occluding loop into a more open conformation resulting in a structural change to the protein which in turn contributes to the RMSD.

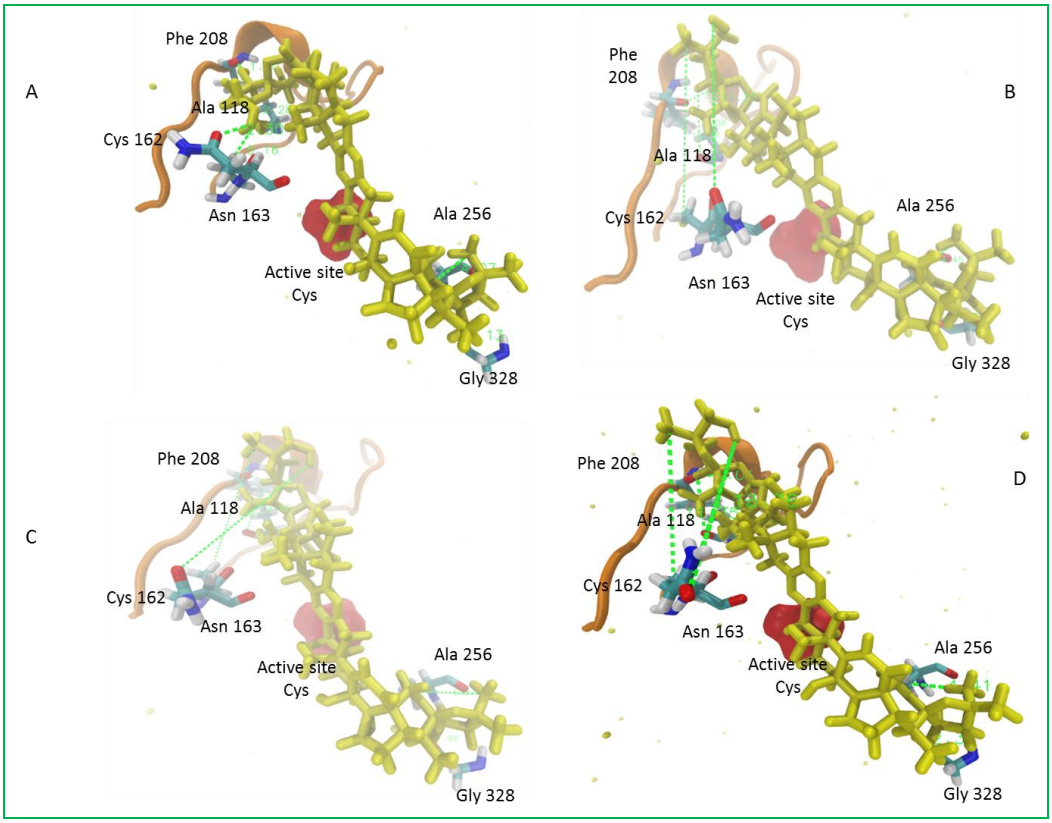


Figure 4.5: SANC00488 (Yellow) and interacting TbCatB residues during 13 ns simulations. The Figures were extracted at (A) 0 ns, (B) 3 ns, (C) 6 ns and (D) 12 ns. The active site Cys residue is shown in red and the occluding loop is shown in orange. Distances to the appropriate active site residues are shown in green. The ligand is seen to move away from interacting residues during simulation

Interactions between SANC00488 and both HsCatB and TbCatB is probably influenced by the movement and stability of the occluding loop. The occluding loop of TbCatB has been identified by other researchers [23] to have a more stable S1' opening which could be exploited for TbCatB specific inhibitors, while the occluding loop of HsCatB behaves independently from the structure and it adapts to changes in the environment [26].

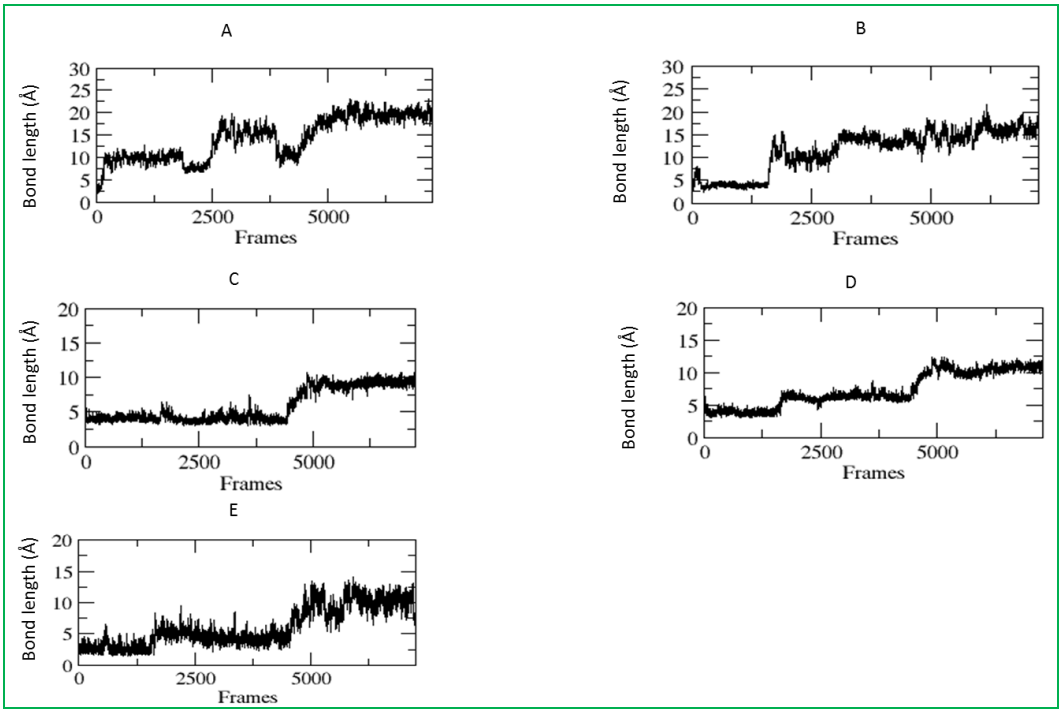


Figure 4.6: 13 ns MD simulation of interactions between SANC0048 and residues (A) Asn72, (B) Glu245, (C) Ala73, (D) Pro76, and (E) Tyr 75 in HsCatB. Frames were recorded every 2 ns for 13 ns. All the interactions are increasing in length.

To get more insight into the evolution of the receptor – ligand interactions, bond lengths were plotted against frames that were recorded every 2 ns for the simulation time of 13 ns between SANC00488 and HsCatB (Figure 4.6) and TbCatB (Figure 4.7). The receptor-ligand interactions were determined from the original docked pose using Discovery Studio Visualizer.

All the intermolecular interactions in the HsCatB-SANC00488 complex gradually increase in length until they finally stabilize (Figure 4.6). The distances between the residues and the corresponding atoms in the ligand at equilibrium are more than 9Å. To track the π-π interaction between the ligand and Pro76 and Tyr75, the distance between the residues and the corresponding atoms in the ligand were followed over simulation time. These distances increased to more than 10 Å in both cases. This indicates that the interactions were not maintained during simulation.

In Figure 4.7 A, we see the evolution of the hydrogen bond formed between residue Ala118 of TbCatB and SANC00488. At the beginning of the simulation the bond is 2.08Å and by the end of the simulation it equilibrates at 4.05 Å.

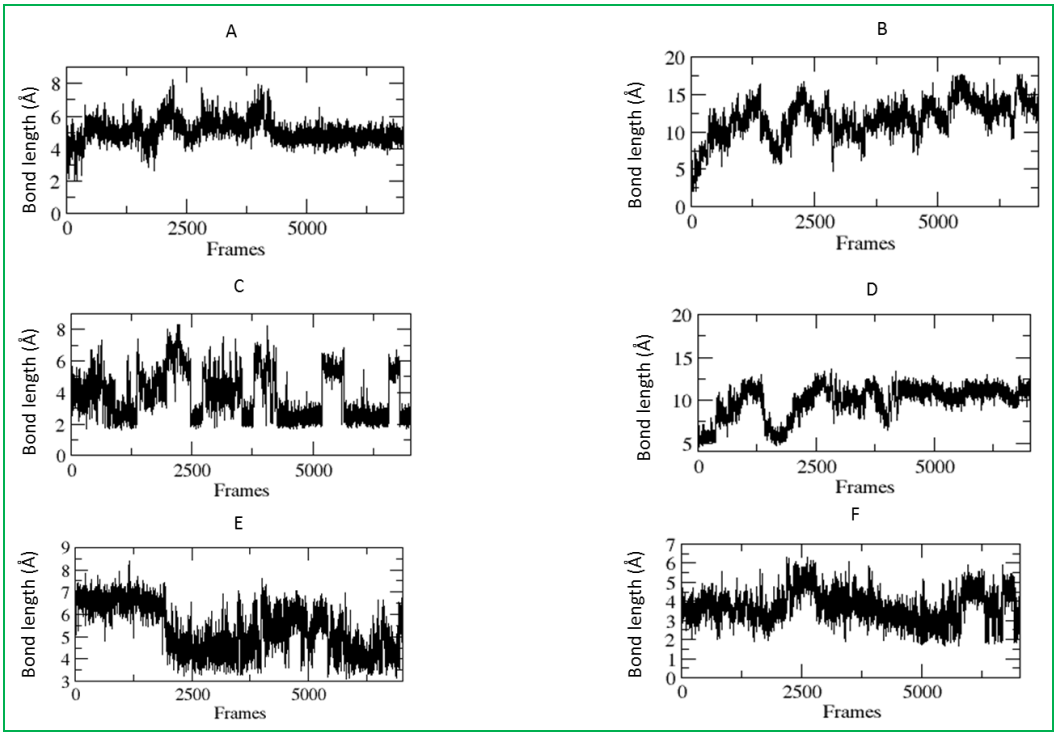


Figure 4.7: 13 ns MD simulation of interactions between SANC00488 and residues (A) Ala118, (B) Asn163, (C) Phe208, (D) Cys162, (E) Gly328, and (F) Ala256 in TbCatB. Frames were recorded every 2 ns for 13 ns.

The hydrogen bond between Cys162 and the hydrophobic interactions between Asn163 were not maintained during simulation (Figure 4.7 B and D).

The hydrogen bond (as determined by DS), length between Phe182 and the compound is observed fluctuating between a maximum length of about 5.6 Å and a minimum length of 2.2 Å through the simulation (Figure 4.7 C) and (Figure 4.8). This bond length interaction was at a minimum during the conformational change (Figure 4.8 C) that the ligand went through. This suggests that this interaction is important for this complex formation. At the beginning of the simulation, the hydrogen atom of the compound that interacts with the oxygen atom of the Phe208 residue is facing towards the residue (Figure 4.8 A). In this position the distance between the two interacting atoms is low (Figure 4.8 A, C, E, and G). During simulation, repositioning of the compound resulted in the hydrogen atom facing away from the Phe208 oxygen, thereby increasing the distance (Figure 4.8 B, D and F). However, the interaction distance is at its minimum for longer durations during the simulation than when it is at its maximum.

The hydrophobic interactions distance between the ligand and Ala256 changed from 4.07 Å to 4.41 Å. The distance between the Gly328 residue and the ligand interacting atom was originally 5.13 Å and so there was no hydrogen bond depicted by DS initially. After simulation the

distance between the two molecules became 2.31 Å. Since the Gly328 is in the S2 subsite that determines specificity of this protease, it is likely that a hydrogen bond was formed during simulation.

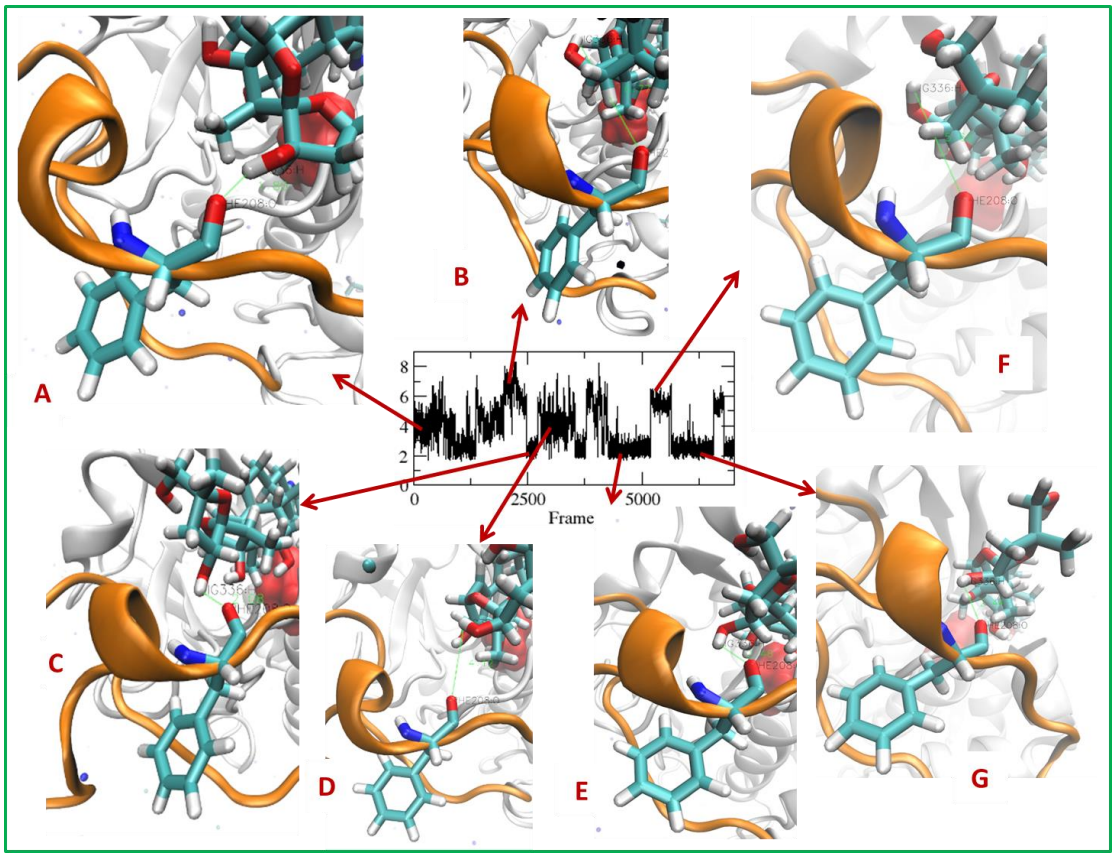


Figure 4.8: 13 ns MD simulation of interactions between SANC00488 and residues Phe 208. (A) - (G) shows the orientation of interacting atoms at different times during the simulation. Phe208 is located in the occluding loop (orange).

4.6. Conclusion

Molecular dynamics simulations of SANC00488 in complex with cathepsin B proteases from human and *Trypanosomal spp.* was done to obtain the complex structures that resemble physiological conditions and to predict the binding mode of the compound into the structures. The results demonstrated that the compound forms stable complexes with *Trypanosomal spp.* cathepsin B proteases and not with human cathepsin B protease. Conformational changes of the ligand and interacting residues during simulation was observed for TbCatB and HsCatB complexes, which may be expected during dynamics and there are other studies that highlight that changes in the docked poses of ligands may be expected during simulation [114]. The position and orientation of the ligand changed during simulation (Figure 4.4. A-D and Figure 4.5. A-D). This demonstrates the importance of MD simulation after docking [115] to determine the correct binding mode of compounds. Residues that are important for

SANC00488-TbCatB complex formation are Gly328 of the S2 subsite, Phe208, and the Ala256. MD simulation of SANC00488 in complex with TcCatB, TcrCatB and TvCatB demonstrated that the complexes are stable. The compound however does not form a stable complex with HsCatB.

Chapter 5.

5.1. Conclusion and Future Prospects

Using PDB ID: 4HWY sequence of TbCatB, a total of four *Trypanosomal* and one human cathepsin B protease homolog sequences were retrieved (Table 2.0). Comparative analysis was carried out using multiple sequence analysis (MSA) (Figure 2.1) and phylogenetic analysis (Figure 2.4). MSA showed that although there are highly conserved residues, like at the active site, there are regions, in the occluding loop and at the S2 subsite, where the sequences differ significantly. Conservation of residues at the active site is consistent with characteristics of C1 cysteine proteases [12], [117]. The occluding loop of TbCatB has three extra residues, Lys197, Try202 and Phe208 residues which are not available in the HsCatB protease. The occluding loop of TbCatB also contains a "FNFD" motif which corresponds to "GEGD" motif in HsCatB. The "FNFD" motif in TbCatB results in a more stable occluding loop and consequently a more stable S1' opening while the "GEGD" in HsCatB results in a more flexible occluding loop at that region [23]. Another major difference is in the S2 subsite where TbCatB has Gly328 residue while HsCatB has Glu245 residue in the corresponding position. Phylogenetic analysis results showed a clustering of all the *Trypanosomal spp.* cathepsin B proteases with the HsCatB forming an outgroup.

High quality homology models were calculated for *T.congolense, T.cruzi* and *T.vivax* cathepsin B proteases using MODELLLERv9.10. From structural comparison of the homolog proteins, fold conservation was maintained in the proteases at the active. At the S2 subsite region, there is a difference in the depth of the pocket (Figure 2.2).

High throughput screening of 600 SANCDB resulted in nine compounds, SANC00 478, 479, 480, 481, 482, 488, 489, 490 and 491, having a strong affinity for *Trypanosoma spp.* cathepsin B proteases than HsCatB. SANC00488 has the strongest binding to the *Trypanosoma spp.* cathepsin B proteases and the weakest binding to HsCatB protease. This is shown by the docking energy of this compound (Fgure 3.14), and a preference for binding along the active site in *Trypanosoma spp.* cathepsin B proteases.

Molecular dynamics (MD) simulations show that the complexes between SANC00488 and TbCatB, TcCatB, TcrCatB and TvCatB are stable and do not dissociate during simulation. The complex between this compound and HsCatB however is unstable and comes apart during simulation (Figure 4.4). The complex between TbCatB and SANC00488 (Figure 4.5) shows the compound finding a more favourable position in the active site of this protease and thereby

forming a more stable complex. From this complex, residues that are important for interaction are Gly328 of the S2 subsite, Phe208, and Ala256.

Further insights on the effect of binding residue variations between HsCatB and TbCatB could be determined through MD simulations of SANC00 478, 479, 480, 481, 482, 489, 490 and 491 to determine the effect of simulation on these complexes and the residues that are involved during interaction for each of them. These simulations were not done due to time restraints. Calculations of binding free energies would have contributed to determine energies contributed by van der Waals forces, electrostatic energy, polar and non-polar solvation energy and contributed to determine key interactions involved in ligand binding. Binding free energy calculations were also not done due to time constraints.

In addition to performing MD simulations of the remaining lead compounds and calculating the binding free energy, the lead compounds could be analysed in the laboratory to determine their effect *in vivo* and *in vitro*. The toxicity of each compound would also have to be determined experimentally.

References

[1] E. M. Fèvre, P. G. Coleman, M. Odiit, J. W. Magona, S. C. Welburn, and M. E. J. Woolhouse, "The origins of a new Trypanosoma brucei rhodesiense sleeping sickness outbreak in eastern Uganda," Lancet, vol. 358, no. 9282, pp. 625–628, 2001.

[2] F. Chappuis, L. Loutan, P. Simarro, V. Lejon, and P. Buscher, "Options for field diagnosis of human african trypanosomiasis," Clin.Microbiol.Rev., vol. 18, no. 0893–8512 (Print), pp. 133–146, Jan. 2005.

[3] M. P. Barrett, R. J. S. Burchmore, A. Stich, J. O. Lazzari, A. C. Frasch, J. J. Cazzulo, and S. Krishna, "The trypanosomiases.," Lancet, vol. 362, no. 9394, pp. 1469–80, Nov. 2003.

[4] WHO, "WHO | Trypanosomiasis, human African (sleeping sickness)," 2015.

[5] L. Huang, L. S. Brinen, and J. a. Ellman, "Crystal structures of reversible ketone-Based inhibitors of the cysteine protease cruzain," Bioorganic Med. Chem., vol. 11, no. 1, pp. 21–29, 2003.

[6] X. Du, C. Guo, E. Hansell, P. S. Doyle, C. R. Caffrey, T. P. Holler, J. H. McKerrow, and F. E. Cohen, "Synthesis and structure-activity relationship study of potent trypanocidal thio semicarbazone inhibitors of the trypanosomal cysteine protease cruzain.," J. Med. Chem., vol. 45, no. 13, pp. 2695–707, Jun. 2002.

[7] C. R. Caffrey, E. Hansell, K. D. Lucas, L. S. Brinen, a Alvarez Hernandez, J. Cheng, S. L. Gwaltney, W. R. Roush, Y. D. Stierhof, M. Bogyo, D. Steverding, and J. H. McKerrow, "Active site mapping, biochemical properties and subcellular localization of rhodesain, the major cysteine protease of Trypanosoma brucei rhodesiense.," Mol. Biochem. Parasitol., vol. 118, no. 1, pp. 61–73, Nov. 2001.

[8] L. Redecke, K. Nass, D. P. DePonte, T. a White, D. Rehders, A. Barty, F. Stellato, M. Liang, T. R. M. Barends, S. Boutet, G. J. Williams, M. Messerschmidt, M. M. Seibert, A. Aquila, D. Arnlund, S. Bajt, T. Barth, M. J. Bogan, C. Caleman, T.-C. Chao, R. B. Doak, H. Fleckenstein, M. Frank, R. Fromme, L. Galli, I. Grotjohann, M. S. Hunter, L. C. Johansson, S. Kassemeyer, G. Katona, R. a Kirian, R. Koopmann, C. Kupitz, L. Lomb, A. V Martin, S. Mogk, R. Neutze, R. L. Shoeman, J. Steinbrener, N. Timneanu, D. Wang, U. Weierstall, N. a Zatsepin, J. C. H. Spence, P. Fromme, I. Schlichting, M. Duszenko, C. Betzel, and H. N. Chapman, "Natively inhibited Trypanosoma brucei cathepsin B structure determined by using an X-ray laser.," Science, vol. 339, no. 2013, pp. 227–30, 2013.

[9] M. P. Barrett, "The fall and rise of sleeping sickness.," Lancet, vol. 353, no. 9159, pp. 1113–4, 1999.

[10] K. Stuart, R. Brun, S. Croft, A. Fairlamb, R. E. Gürtler, J. McKerrow, S. Reed, and R. Tarleton, "Kinetoplastids: related protozoan pathogens, different diseases.," J. Clin. Invest., vol. 118, no. 4, pp. 1301–1310, 2008.

[11] J. W. Choy, C. Bryant, C. M. Calvet, P. S. Doyle, S. S. Gunatilleke, S. S. F. Leung, K. K. H. Ang, S. Chen, J. Gut, J. a Oses-Prieto, J. B. Johnston, M. R. Arkin, A. L. Burlingame, J. Taunton, M. P. Jacobson, J. M. McKerrow, L. M. Podust, and A. R. Renslo, "Chemical-biological characterization of a cruzain inhibitor reveals a second target and a mammalian off-target.," Beilstein J. Org. Chem., vol. 9, pp. 15–25, Jan. 2013.

[12] M. Sajid and J. H. McKerrow, “Cysteine proteases of parasitic organisms.,” Mol. Biochem. Parasitol., vol. 120, no. 1, pp. 1–21, 2002.

[13] N. D. Rawlings, A. J. Barrett, and A. Bateman, “MEROPS: the peptidase database.,” Nucleic Acids Res., vol. 38, no. Database issue, pp. D227–33, Jan. 2010.

[14] V. Stoka, B. Turk, and V. Turk, “Lysosomal cysteine proteases: structural features and their role in apoptosis.,” IUBMB Life, vol. 57, no. 4–5, pp. 347–53, 2005.

[15] F. Lecaille, J. Kaleta, and D. Brömme, “Human and parasitic papain-like cysteine proteases: their role in physiology and pathology and recent developments in inhibitor design.,” Chem. Rev., vol. 102, no. 12, pp. 4459–88, Dec. 2002.

[16] V. K. Dubey, M. Pande, B. K. Singh, and M. V Jagannadham, “Papain-like proteases : Applications of their inhibitors,” vol. 6, no. May, pp. 1077–1086, 2007.

[17] D. a Nicoll-Griffith, “Use of cysteine-reactive small molecules in drug discovery for trypanosomal disease,” Expert Opin. Drug Discov., vol. 7, no. 4, pp. 353–366, 2012.

[18] M.-H. Abdulla, T. O’Brien, Z. B. Mackey, M. Sajid, D. J. Grab, and J. H. McKerrow, “RNA interference of Trypanosoma brucei cathepsin B and L affects disease progression in a mouse model.,” PLoS Negl. Trop. Dis., vol. 2, no. 9, p. e298, Jan. 2008.

[19] Z. Grzonka, E. Jankowska, F. Kasprzykowski, R. Kasprzykowska, L. Lankiewicz, W. Wiczk, E. Wieczerzak, J. Ciarkowski, P. Drabik, R. Janowski, M. Kozak, M. Jaskólski, and a Grubb, “Structural studies of cysteine proteases and their inhibitors.,” Acta Biochim. Pol., vol. 48, no. 1, pp. 1–20, Jan. 2001.

[20] D. Yamamoto, K. Matsumoto, H. Ohishi, T. Ishidas, M. Inoue, and K. Kitamuraq, “Refined X-ray Structure of papain-E-64-c complex at 2.1Å resolution,” J. Biol. Chem., vol. 266, no. 22, pp. 14771–14777, 1991.

[21] M. Rzychon, D. Chmiel, and J. Stec-Niemczyk, “Modes of inhibition of cysteine proteases.,” Acta Biochim. Pol., vol. 51, no. 4, pp. 861–873, 2004.

[22] V. Turk, V. Stoka, O. Vasiljeva, M. Renko, T. Sun, B. Turk, and D. Turk, “Cysteine cathepsins: From structure, function and regulation to new frontiers,” Biochim. Biophys. Acta - Proteins Proteomics, vol. 1824, no. 1, pp. 68–88, Jan. 2012.

[23] I. D. Kerr, P. Wu, R. Marion-Tsukamaki, Z. B. Mackey, and L. S. Brinen, “Crystal Structures of TbCatB and Rhodesain, Potential Chemotherapeutic Targets and Major Cysteine Proteases of Trypanosoma brucei,” PLoS Negl. Trop. Dis., vol. 4, no. 6, p. e701, 2010.

[24] B. Turk, D. Turk, and V. Turk, “Lysosomal cysteine proteases: more than scavengers,” Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol., vol. 1477, no. 1–2, pp. 98–111, 2000.

[25] D. Musil, D. Zucic, D. Turk, R. a Engh, I. Mayr, R. Huber, T. Popovic, V. Turk, T. Towatari, and N. Katunuma, “The refined 2.15 A X-ray crystal structure of human liver cathepsin B: the structural basis for its specificity.,” EMBO J., vol. 10, no. 9, pp. 2321–2330, 1991.

[26] M. Podobnik, R. Kuhelj, V. Turk, and D. Turk, “Crystal structure of the wild-type human procathepsin B at 2.5 A resolution reveals the native active site of a papain-like cysteine protease zymogen.,” J. Mol. Biol., vol. 271, no. 5, pp. 774–788, 1997.

[27] D. Turk, M. Podobnik, T. Popovic, N. Katunuma, W. Bode, R. Huber, and V. Turk, “Crystal structure of cathepsin B inhibited with CA030 at 2.0-A resolution: A basis for the design of specific epoxysuccinyl inhibitors.,” Biochemistry, vol. 34, no. 14, pp. 4791–4797, 1995.

[28] J. C. Mottram and R. S. Brooks, “Roles of cysteine proteinases in host-parasite interactions of trypanosomes and Graham H Coombs §.”

[29] Z. B. Mackey, T. C. O’Brien, D. C. Greenbaum, R. B. Blank, and J. H. McKerrow, “A cathepsin B-like protease is required for host protein degradation in Trypanosoma brucei.,” J. Biol. Chem., vol. 279, no. 46, pp. 48426–33, Nov. 2004.

[30] W. Bode and R. Huber, “Structural basis of the endoproteinase-protein inhibitor interaction,” Biochim. Biophys. Acta - Protein Struct. Mol. Enzymol., vol. 1477, no. 1–2, pp. 241–252, 2000.

[31] T. F. Kagawa, J. C. Cooney, H. M. Baker, S. McSweeney, M. Liu, S. Gubba, J. M. Musser, and E. N. Baker, “Crystal structure of the zymogen form of the group A Streptococcus virulence factor SpeB: an integrin-binding cysteine protease.,” Proc. Natl. Acad. Sci. U. S. A., vol. 97, no. 5, pp. 2235–40, 2000.

[32] C. Schick, D. Brömme, A. J. Bartuski, Y. Uemura, N. M. Schechter, and G. A. Silverman, “The reactive site loop of the serpin SCCA1 is essential for cysteine proteinase inhibition.,” Proc. Natl. Acad. Sci. U. S. A., vol. 95, no. 23, pp. 13465–70, 1998.

[33] G. Xu, M. Cirilli, Y. Huang, R. L. Rich, D. G. Myszka, and H. Wu, “Covalent inhibition revealed by the crystal structure of the caspase-8/p35 complex,” Nature, vol. 410, no. 6827, pp. 494–497, 2001.

[34] G. Gunčar, G. Pungerčič, I. Klemenčič, V. Turk, and D. Turk, “Crystal structure of MHC class II-associated p41 Ii fragment bound to cathepsin L reveals the structural basis for differentiation between cathepsins L and S,” EMBO J., vol. 18, no. 4, pp. 793–803, 1999.

[35] H. R. Stennicke, C. a Ryan, and G. S. Salvesen, “Reprieval from execution: the molecular basis of caspase inhibition.,” Trends Biochem. Sci., vol. 27, no. 2, pp. 94–101, 2002.

[36] R. Koopmann, K. Cupelli, and L. Redecke, “In vivo protein crystallization opens new routes in structural biology,” Nat. …, vol. 9, no. 3, pp. 259–262, 2012.

[37] I. Redzynia, A. Ljunggren, M. Abrahamson, J. S. Mort, J. C. Krupa, M. Jaskolski, and G. Bujacz, “Displacement of the occluding loop by the parasite protein, chagasin, results in efficient inhibition of human cathepsin,” J. Biol. Chem., vol. 283, no. 33, pp. 22815–22825, 2008.

[38] R. Ettari, L. Tamborini, I. C. Angelo, N. Micale, A. Pinto, C. De Micheli, and P. Conti, “Inhibition of rhodesain as a novel therapeutic modality for human African trypanosomiasis.,” J. Med. Chem., vol. 56, no. 14, pp. 5637–58, Jul. 2013.

[39] N. Fujii, J. P. Mallari, E. J. Hansell, Z. Mackey, P. Doyle, Y. M. Zhou, J. Gut, P. J. Rosenthal, J. H. McKerrow, and R. K. Guy, "Discovery of potent thiosemicarbazone inhibitors of rhodesain and cruzain.," Bioorg. Med. Chem. Lett., vol. 15, no. 1, pp. 121–3, Jan. 2005.

[40] E. Dunny, W. Doherty, P. Evans, J. P. G. Malthouse, D. Nolan, and A. J. S. Knox, "Vinyl sulfone-based peptidomimetics as anti-trypanosomal agents: Design, synthesis, biological and computational evaluation," J. Med. Chem., vol. 56, pp. 6638–6650, 2013.

[41] C. N. Cavasotto and S. S. Phatak, "Homology modeling in drug discovery: current trends and applications," Drug Discov. Today, vol. 14, no. July, pp. 676–683, 2009.

[42] M. E. McGrath, a E. Eakin, J. C. Engel, J. H. McKerrow, C. S. Craik, and R. J. Fletterick, "The crystal structure of cruzain: a therapeutic target for Chagas' disease.," J. Mol. Biol., vol. 247, no. 2, pp. 251–259, 1995.

[43] G. L. Moraes, G. C. Gomes, P. R. Monteiro de Sousa, C. N. Alves, T. Govender, H. G. Kruger, G. E. M. Maguire, G. Lamichhane, and J. Lameira, "Structural and functional features of enzymes of Mycobacterium tuberculosis peptidoglycan biosynthesis as targets for drug development.," Tuberculosis (Edinb)., vol. 95, no. 2, pp. 95–111, 2015.

[44] B. T. Mott, R. S. Ferreira, A. Simeonov, A. Jadhav, K. K.-H. Ang, W. Leister, M. Shen, J. T. Silveira, P. S. Doyle, M. R. Arkin, J. H. McKerrow, J. Inglese, C. P. Austin, C. J. Thomas, B. K. Shoichet, and D. J. Maloney, "Identification and optimization of inhibitors of Trypanosomal cysteine proteases: cruzain, rhodesain, and TbCatB.," J. Med. Chem., vol. 53, no. 1, pp. 52–60, Jan. 2010.

[45] R. G. Bodade, S. D. Beedkar, a. V. Manwar, and C. N. Khobragade, "Homology modeling and docking study of xanthine oxidase of Arthrobacter sp. XL26," Int. J. Biol. Macromol., vol. 47, no. 2, pp. 298–303, 2010.

[46] C. N. Khobragade, S. D. Beedkar, R. G. Bodade, and A. S. Vinchurkar, "Comparative structural modeling and docking studies of oxalate oxidase: Possible implication in enzyme supplementation therapy for urolithiasis," Int. J. Biol. Macromol., vol. 48, no. 3, pp. 466–473, 2011.

[47] O. Tastan Bishop and M. Kroon, "Study of protein complexes via homology modeling, applied to cysteine proteases and their protein inhibitors.," J. Mol. Model., vol. 17, no. 12, pp. 3163–72, Dec. 2011.

[48] 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. Web Server issue, pp. W244–8, Jul. 2005.

[49] R. Hatherley, D. K. Brown, T. M. Musyoka, D. L. Penkler, N. Faya, K. A. Lobb, and Ö. Tastan Bishop, "SANCDB: a South African natural compound database," J. Cheminform., vol. 7, no. 1, p. 29, 2015.

[50] Z. Xiang, "Advances in Homology Protein Structure Modeling," Curr. Protein Pept. Sci., vol. 7, no. 3, pp. 217–227, Jun. 2006.

[51] C. B. Anfinsen, "Principles that Govern the Folding of Protein Chains," Science (80-.)., vol. 181, no. 4096, pp. 223–230, Jul. 1973.

[52] B. Al-Lazikani, J. Jung, Z. Xiang, and B. Honig, “Protein structure prediction,” Curr. Opin. Chem. Biol., vol. 5, no. 1, pp. 51–56, Feb. 2001.

[53] R. Rodriguez, G. Chinea, N. Lopez, T. Pons, and G. Vriend, “Homology modeling, model and software evaluation: three related resources.,” Bioinformatics, vol. 14, no. 6, pp. 523–528, 1998.

[54] A. Tramontano and V. Morea, “Assessment of homology-based predictions in CASP5.,” Proteins, vol. 53 Suppl 6, no. March, pp. 352–68, 2003.

[55] J. Xiong, Essential Bioinformatics, vol. 1. 2006.

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

[57] N. N. Alexandrov and R. Luethy, “Alignment algorithm for homology modeling and threading,” Protein Sci., vol. 7, pp. 254–258, 1998.

[58] W. Gish and D. J. States, “Identification of protein coding regions by database similarity search.,” Nat. Genet., vol. 3, no. 3, pp. 266–272, 1993.

[59] D. a. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, “GenBank,” Nucleic Acids Res., vol. 37, no. Database, pp. D26–D31, Jan. 2009.

[60] S. F. Altschul, W. Gish, T. Pennsylvania, and U. Park, “Basic Local Alignment Search Tool 2Department of Computer Science,” pp. 403–410, 1990.

[61] D. L. Wheeler, D. M. Church, R. Edgar, S. Federhen, W. Helmberg, T. L. Madden, J. U. Pontius, G. D. Schuler, L. M. Schriml, E. Sequeira, T. O. Suzek, T. a Tatusova, and L. Wagner, “Database resources of the National Center for Biotechnology Information: update.,” Nucleic Acids Res., vol. 32, no. Database issue, pp. D35–D40, 2004.

[62] C. Kanz, P. Aldebert, N. Althorpe, W. Baker, A. Baldwin, K. Bates, P. Browne, A. van den Broek, M. Castro, G. Cochrane, K. Duggan, R. Eberhardt, N. Faruque, J. Gamble, F. Garcia Diez, N. Harte, T. Kulikova, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, M. McHale, F. Nardone, V. Silventoinen, S. Sobhany, P. Stoehr, M. A. Tuli, K. Tzouvara, R. Vaughan, D. Wu, W. Zhu, and R. Apweiler, “The EMBL nucleotide sequence database,” Nucleic Acids Res., vol. 33, no. DATABASE ISS., pp. 29–33, 2005.

[63] Y. Kodama, J. Mashima, T. Kosuge, T. Katayama, T. Fujisawa, E. Kaminuma, O. Ogasawara, K. Okubo, T. Takagi, and Y. Nakamura, “The DDBJ Japanese Genotype-phenotype Archive for genetic and phenotypic human data.,” Nucleic Acids Res., vol. 43, no. Database issue, pp. D18–22, 2015.

[64] I. Karsch-Mizrachi, Y. Nakamura, and G. Cochrane, “The International Nucleotide Sequence Database Collaboration,” Nucleic Acids Res., vol. 40, no. D1, pp. D33–D37, 2012.

[65] T. N. Bhat, P. Bourne, Z. Feng, G. Gilliland, S. Jain, V. Ravichandran, B. Schneider, K. Schneider, N. Thanki, H. Weissig, J. Westbrook, and H. M. Berman, “The PDB data uniformity project.,” Nucleic Acids Res., vol. 29, pp. 214–218, 2001.

[66] Needlema.Sb and C. D. Wunsch, “a General Method Applicable To Search for Similarities in Amino Acid Sequence of 2 Proteins,” J. Mol. Biol., vol. 48, no. 3, p. 443–&, 1970.

[67] 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.

[68] W. R. Pearson and D. J. Lipman, “Improved tools for biological sequence comparison.,” Proc. Natl. Acad. Sci. U. S. A., vol. 85, no. 8, pp. 2444–2448, 1988.

[69] R. L. Dunbrack, “Sequence comparison and protein structure prediction,” Curr. Opin. Struct. Biol., vol. 16, pp. 374–384, 2006.

[70] 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. 2295–2300, Feb. 2008.

[71] K. Katoh, K. Misawa, K. Kuma, and T. Miyata, “MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.,” Nucleic Acids Res., vol. 30, no. 14, pp. 3059–3066, 2002.

[72] C. Kosiol, L. Bofkin, and S. Whelan, “Phylogenetics by likelihood: Evolutionary modeling as a tool for understanding the genome,” J. Biomed. Inform., vol. 39, no. 1 SPEC. ISS., pp. 51–61, 2006.

[73] E. Lander, Linton, Lauren, B. Birren, and C. Nusbaum, “Initial sequencing and analysis of the human genome,” Nature, vol. 409, no. 6822, pp. 860–921, 2001.

[74] D. H. Bos and D. Posada, “Using models of nucleotide evolution to build phylogenetic trees,” Dev. Comp. Immunol., vol. 29, no. 3, pp. 211–227, 2005.

[75] J. E. McCormack, H. Huang, and L. L. Knowles, “Maximum likelihood estimates of species trees: How accuracy of phylogenetic inference depends upon the divergence history and sampling design,” Syst. Biol., vol. 58, no. 5, pp. 501–508, 2009.

[76] A. Pavlopoulou and I. Michalopoulos, “State-of-the-art bioinformatics protein structure prediction tools (Review),” Int. J. Mol. Med., vol. 28, no. 3, pp. 295–310, 2011.

[77] S. Whelan, P. Liò, and N. Goldman, “Molecular phylogenetics: state-of-the-art methods for looking into the past.,” Trends Genet., vol. 17, no. 5, pp. 262–272, 2001.

[78] J. C. Prasad, S. R. Comeau, S. Vajda, and C. J. Camacho, “Consensus alignment for reliable framework prediction in homology modeling,” Bioinformatics, vol. 19, no. 13, pp. 1682–1691, Sep. 2003.

[79] M. I. Sadowski and D. T. Jones, “Benchmarking template selection and model quality assessment for high-resolution comparative modeling,” Proteins Struct. Funct. Bioinforma., vol. 9999, no. 9999, p. NA+, 2007.

[80] F. Melo, D. Devos, E. Depiereux, and E. Feytmans, “ANOLEA: a www server to assess protein structures.,” Proc. Int. Conf. Intell. Syst. Mol. Biol., vol. 5, pp. 187–190, 1997.

[81] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar, "MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutioanry distance, and maximum parsimony methods," Mol. Biol. Evol., vol. 28, no. 10, pp. 2731–2739, 2011.

[82] A. Šali and T. L. Blundell, "Comparative Protein Modelling by Satisfaction of Spatial Restraints," J. Mol. Biol., vol. 234, no. 3, pp. 779–815, Dec. 1993.

[83] W. L. DeLano, "The PyMOL Molecular Graphics System," Schrödinger LLC wwwpymolorg, vol. Version 1., p. http://www.pymol.org, 2002.

[84] V. N. Maiorov and G. M. Crippen, "Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins.," Journal of molecular biology, vol. 235, no. 2. pp. 625–634, 1994.

[85] M. Pawlowski, M. J. Gajda, R. Matlak, and J. M. Bujnicki, "MetaMQAP: a meta-server for the quality assessment of protein models.," BMC Bioinformatics, vol. 9, p. 403, 2008.

[86] P. Benkert, S. C. E. Tosatto, and T. Schwede, "Global and local model quality estimation at CASP8 using the scoring functions QMEAN and QMEANclust.," Proteins, vol. 77 Suppl 9, pp. 173–80, 2009.

[87] L. Bordoli and T. Schwede, "Automated protein structure modeling with SWISS-MODEL Workspace and the Protein Model Portal.," Methods Mol. Biol., vol. 857, pp. 107–136, 2012.

[88] R. A. Laskowski, M. A. MacArthur, D. S. Moss, and J. M. Thornton, "20.19. (1993). PROCHECK - a program to check the stereochemical quality of protein structures. J. App. Cryst., 26, .," J. App. Cryst, vol. 26, pp. 283–291, 1993.

[89] A. M. Waterhouse, J. B. Procter, D. M. a Martin, M. Clamp, and G. J. Barton, "Jalview Version 2-A multiple sequence alignment editor and analysis workbench," Bioinformatics, vol. 25, no. 9, pp. 1189–1191, 2009.

[90] V. Turk, V. Stoka, O. Vasiljeva, M. Renko, T. Sun, B. Turk, and D. Turk, "Cysteine cathepsins: From structure, function and regulation to new frontiers," Biochim. Biophys. Acta - Proteins Proteomics, vol. 1824, no. 1, pp. 68–88, 2012.

[91] S. Malhotra, O. K. Mathew, and R. Sowdhamini, "DOCKSCORE: a webserver for ranking protein-protein docked poses.," BMC Bioinformatics, vol. 16, no. 1, p. 127, 2015.

[92] R. M. V Abreu, H. J. C. Froufe, M. J. R. P. Queiroz, and I. C. F. R. Ferreira, "MOLA: A bootable, self-configuring system for virtual screening using AutoDock4/Vina on computer clusters," J. Cheminform., vol. 2, no. 1, p. 10, 2010.

[93] M. J. Alves, I. C. F. R. Ferreira, H. J. C. Froufe, R. M. V Abreu, a. Martins, and M. Pintado, "Antimicrobial activity of phenolic compounds identified in wild mushrooms, SAR analysis and docking studies," J. Appl. Microbiol., vol. 115, no. 2, pp. 346–357, 2013.

[94] S. F. Sousa, P. A. Fernandes, and M. J. Ramos, "Protein–ligand docking: current status and future challenges," Proteins, vol. 26, pp. 15–26, 2006.

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

[96] R. T. Kroemer, “Structure-based drug design: docking and scoring.,” Curr. Protein Pept. Sci., vol. 8, no. 4, pp. 312–328, 2007.

[97] W. R. Roush, J. Cheng, B. Knapp-Reed, a Alvarez-Hernandez, J. H. McKerrow, E. Hansell, and J. C. Engel, “Potent second generation vinyl sulfonamide inhibitors of the trypanosomal cysteine protease cruzain.,” Bioorg. Med. Chem. Lett., vol. 11, no. 20, pp. 2759–62, Oct. 2001.

[98] J. C. Alvarez, “High-throughput docking as a source of novel drug leads,” Curr. Opin. Chem. Biol., vol. 8, no. 4, pp. 365–370, Aug. 2004.

[99] X.-Y. Meng, H.-X. Zhang, M. Mezei, and M. Cui, “Molecular docking: a powerful approach for structure-based drug discovery.,” Curr. Comput. Aided. Drug Des., vol. 7, no. 2, pp. 146–157, 2011.

[100] 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.

[101] T. M. Musyoka, A. M. Kanzi, K. A. Lobb, and Ö. Tastan Bishop, “Analysis of Non-Peptidic Compounds as Potential Malarial Inhibitors against Plasmodial Cysteine Proteases via Integrated Virtual Screening Workflow,” J. Biomol. Struct. Dyn., vol. 1102, no. November, pp. 1–72, 2015.

[102] I. D. Kuntz, J. M. Blaney, S. J. Oatley, R. Langridge, and T. E. Ferrin, “A geometric approach to macromolecule-ligand interactions.,” J. Mol. Biol., vol. 161, no. 2, pp. 269–288, 1982.

[103] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe, “A fast flexible docking method using an incremental construction algorithm.,” J. Mol. Biol., vol. 261, no. 3, pp. 470–89, 1996.

[104] G. M. Morris, D. S. Goodsell, M. E. Pique, W. L. Lindstrom, R. Huey, W. E. Hart, S. Halliday, R. Belew, and A. J. Olson, “AutoDock Version 4.2,” User Guid., pp. 1–49, 2009.

[105] T. J. a Ewing, S. Makino, a. G. Skillman, and I. D. Kuntz, “DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases,” J. Comput. Aided. Mol. Des., vol. 15, no. 5, pp. 411–428, 2001.

[106] O. Kayser, A. F. Kiderlen, and S. L. Croft, “Natural products as potential antiparasitic drugs,” Stud. Nat. Prod. Chem., vol. 26, no. PART G, pp. 779–848, 2002.

[107] W. N. Setzer and I. V Ogungbe, “In-silico investigation of antitrypanosomal phytochemicals from Nigerian medicinal plants.,” PLoS Negl. Trop. Dis., vol. 6, no. 7, p. e1727, 2012.

[108] I. V. Ogungbe and W. N. Setzer, “Comparative molecular docking of antitrypanosomal natural products into multiple trypanosoma brucei drug targets,” Molecules, vol. 14, no. 4, pp. 1513–1536, 2009.

[109] P. D. Greenspan, K. L. Clark, R. A. Tommasi, S. D. Cowen, L. W. McQuire, D. L. Farley, J. H. van Duzer, R. L. Goldberg, H. Zhou, Z. Du, J. J. Fitt, D. E. Coppa, Z. Fang, W. Macchia, L. Zhu, M. P. Capparelli, R. Goldstein, A. M. Wigg, J. R. Doughty, R. S. Bohacek, and A. K. Knap, “Identification of Dipeptidyl Nitriles as Potent and Selective Inhibitors of

Cathepsin B through Structure-Based Drug Design," J. Med. Chem., vol. 44, no. 26, pp. 4524–4534, 2001.

[110] H. Alonso, A. a. Bliznyuk, and J. E. Gready, "Combining docking and molecular dynamic simulations in drug design," Med. Res. Rev., vol. 26, no. 5, pp. 531–568, 2006.

[111] Wikipedia, "Molecular dynamics," no. Md, pp. 1–11, 1963.

[112] A. Astuti and A. Mutiara, "Performance Analysis on Molecular Dynamics Simulation of Protein Using GROMACS," arXiv Prepr. arXiv0912.0893, 2009.

[113] M. P. Allen, "Introduction to Molecular Dynamics Simulation," Comput. Soft Matter From Synth. Polym. to Proteins, vol. 23, pp. 1–28, 2004.

[114] S.-Y. Lu, Y.-J. Jiang, J. Lv, T.-X. Wu, Q.-S. Yu, and W.-L. Zhu, "Molecular docking and molecular dynamics simulation studies of GPR40 receptor–agonist interactions," J. Mol. Graph. Model., vol. 28, no. 8, pp. 766–774, 2010.

[115] L. Minini, G. Álvarez, M. González, H. Cerecetto, and A. Merlino, "Molecular docking and molecular dynamics simulation studies of Trypanosoma cruzi triosephosphate isomerase inhibitors. Insights into the inhibition mechanism and selectivity," J. Mol. Graph. Model., vol. 58, pp. 40–49, 2015.

[116] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation," J. Chem. Theory Comput., vol. 4, no. 3, pp. 435–447, 2008.

[117] D. C. Greenbaum, Z. Mackey, E. Hansell, P. Doyle, J. Gut, C. R. Caffrey, J. Lehrman, P. J. Rosenthal, J. H. McKerrow, and K. Chibale, "Synthesis and structure-activity relationships of parasiticidal thiosemicarbazone cysteine protease inhibitors against Plasmodium falciparum, Trypanosoma brucei, and Trypanosoma cruzi," J. Med. Chem., vol. 47, no. 12, pp. 3212–3219, 2004.

Appendix 1A

```
>P1;T_congolense
sequence:T_congolense:1  ::336    :0 :0 : 0:0 :0
QLRTELPESFDSAEKWPNCPTIREIADQSACGSCWAVSTASAISDRYCTVGG-VQQLRISAAHLMSCCE-DC
GDGCKGGAPDSAWEYYVSHGLASS-------YCQPYPFPHCGHHG-GKGKKPPCSKYHFHTPKCNTTCTDKA
IPL--IKYRGNNSYMLL-NGEDDYKRELYFNGPFVVDFGVYSDFLAYKTGVYRHVSGDVLGGHAVRIVGWGK
LNGTPYWKIANSWDTDWGMNGHFLILRGNNECGIESTGYAGLPAIPRNA*
>P1;3CBJ
structureX:3CBJ:60P  :A:255  :A:0:0:0:0
--DLKLPASFDAREQWPQCPTIKEIRDQGSCGSAWAFGAVEAISDRICIHTNAHVSVEVSAEDLLTCCGSMC
GDGCNGGYPAEAWNFWTRKGLVSGGLYESHVGCRPYSIPPCE-AHVNG-ARPP-CTGEGDTPKCSKICEPGY
SPTYKQDKHYGYNSYSVSNSEKDIMAEIYKNGPVEGAFSVYSDFLLYKSGVYQHVTGEMMGGHAIRILGWGV
ENGTPYWLVANSWNTDWGDNGFFKILRGQDHCGIESEVVAGIPRTDQ--*
```

Appendix 1A-1: PIR aligment used for calculation of *T. congolense* model from using HsCatB crystal structure (PDB ID 3CBJ) as a template.

```
>P1;3HHI
structureX:3HHI:78  :A:335  :A:0:0:0:0
SILPKRR----FTEEEARAPLPSSFDSAEAWPNCPTIPQIADQSACGSCWAVAAASAMSDRFCTMGG-VQDV
HISAGDLLACCSDCGDGCNGGDPDRAWAYFSSTGLVSDYCQPYPFPHCSHHSKSKNGYPPCSQFNFDTPKCD
YTCDDP-TIPVVNYRSWTSYALQGEDDYMRELFFRGPFEVAFDVYEDFIAYNSGVYHHVSGQYLGGHAVRLV
GWGTSNGVPYWKIANSWNTEWGMDGYFLIRRGSSECGIEDGGSAGIPL-----*
>P1;T_congolense
sequence:T_congolense:1  ::336    :0 :0 : 0:0 :0
RRKTSSLPPVRFTEEQLRTELPESFDSAEKWPNCPTIREIADQSACGSCWAVSTASAISDRYCTVGG-VQQL
RISAAHLMSCCEDCGDGCKGGAPDSAWEYYVSHGLASSYCQPYPFPHCGHHG-GKGKKPPCSKYHFHTPKCN
TTCTDKAIPLIKYRGNNSYMLLNGEDDYKRELYFNGPFVVDFGVYSDFLAYKTGVYRHVSGDVLGGHAVRIV
GWGKLNGTPYWKIANSWDTDWGMNGHFLILRGNNECGIESTGYAGLPAIPRNA*
```

Appendix 1A-2: PIR aligment used for calculation of *T. congolense* model from using *Trypanosoma* cathepsin B crystal structure (PDB ID 3HHI) as a template.

```
>P1;3HHI
structureX:3HHI:78  :A:335  :A:0:0:0:0
--SILPKRR----FTEEEARAPLPSSFDSAEAWPNCPTIPQIADQSACGSCWAVAAASAMSDRFCTMGG-VQ
DVHISAGDLLACCS-DCGDGCNGGDPDRAWAYFSSTGLVSD-------YCQPYPFPHCSHHSKSKNGYPPCS
QFNFDTPKCDYTCDDP-TIP--VVNYRSWTSYAL-QGEDDYMRELFFRGPFEVAFDVYEDFIAYNSGVYHHV
SGQYLGGHAVRLVGWGTSNGVPYWKIANSWNTEWGMDGYFLIRRGSSECGIEDGGSAGIPL-----*

>P1;3CBJ
structureX:3CBJ:60P  :A:255  :A:0:0:0:0
-------------------DLKLPASFDAREQWPQCPTIKEIRDQGSCGSAWAFGAVEAISDRICIHTNAHV
SVEVSAEDLLTCCGSMCGDGCNGGYPAEAWNFWTRKGLVSGGLYESHVGCRPYSIPPCE-AHVNG-ARPP-C
TGEGDTPKCSKICEPGYSPTYKQDKHYGYNSYSVSNSEKDIMAEIYKNGPVEGAFSVYSDFLLYKSGVYQHV
TGEMMGGHAIRILGWGVENGTPYWLVANSWNTDWGDNGFFKILRGQDHCGIESEVVAGIPRTDQ--*

>T_congolense/69-336
-FRRKTSSLPPVRFTEEQLRTELPESFDSAEKWPNCPTIREIADQSACGSCWAVSTASAISDRYCTVGG-VQ
QLRISAAHLMSCCE-DCGDGCKGGAPDSAWEYYVSHGLASS-------YCQPYPFPHCGHHG-GKGKKPPCS
KYHFHTPKCNTTCTDKAIPL--IKYRGNNSYMLL-NGEDDYKRELYFNGPFVVDFGVYSDFLAYKTGVYRHV
SGDVLGGHAVRIVGWGKLNGTPYWKIANSWDTDWGMNGHFLILRGNNECGIESTGYAGLPAIPRNA*
```

Appendix 1A-3: PIR aligment used for calculation of *T. congolense* model from using HsCatB crystal structure (PDB ID 3CBJ) and *Trypanosoma* cathepsin B crystal structure (PDB ID 3HHI) as a templates.

```
>P1;3CBJ
structureX:3CBJ:60P  :A:255  :A:0:0:0:0
------------------DLKLPASFDAREQWPQCPTIKEIRDQGSCGSAWAFGAVEAISDRICIHTNAHVS
VEVSAEDLLTCCGSMCGDGCNGGYPAEAWNFWTRKGLVSGGLYESHVGCRPYSIPPCEAHVN--GARPP-CT
GEGDTPKCSKICEPGYSPTYKQDKHYGYNSYSVSNSEKDIMAEIYKNGPVEGAFSVYSDFLLYKSGVYQHVT
GEMMGGHAIRILGWGVENGTPYWLVANSWNTDWGDNGFFKILRGQDHCGIESEVVAGIPRTDQ*

>P1;T_cruzi
sequence:T_cruzi:1  ::208    :0 :0 : 0:0 :0
FLRNTSILPPRQFSEEELRVPLQDRFDAGEAWPECPTVTEIRDQSSCGSCWAVAAASAISDRYCTLGG-VRD
LRISAGDLMSCCD-VCGFGCNGGYPEVAWEYYAVHGIVSE-------YCQPYPFPSCAHHVN--SSDLSPCS
GEYDTPTCNSTCTDKKI-P--LIKYRGNTSYVL-SGEEPFKRELILNGPFEVSFSVYADFVAYTGGVYKHVA
GIFLGGHAVRIVGWGELNGEPYWKIANSWNREWGMNGYFLIARGVDECGIEGSGVAGTPRIP-*
```

Appendix 1A-4: PIR aligment used for calculation of *T. cruzi* model from HsCatB crystal structure (PDB ID 3CBJ) as a template.

```
>P1;3HHI
structureX:3HHI:78  :A:335  :A:0:0:0:0
-SILPK----RRFTEEEARAPLPSSFDSAEAWPNCPTIPQIADQSACGSCWAVAAASAMSDRFCIMGG-VQD
VHISAGDLLACCS-DCGDGCNGGDPDRAWAYFSSTGLVSD-------YCQPYPFPHCSHHSKSKNGYPPCSQ
FNFDTPKCDYTCDDPTI-P--VVNYRSWTSYAL-QGEDDYMRELFFRGPFEVAFDVYEDFIAYNSGVYHHVS
GQYLGGHAVRLVGWGTSNGVPYWKIANSWNTEWGMDGYFLIRRGSSECGIEDGGSAGIPL---*

>P1;T_cruzi
sequence:T_cruzi:1  ::208    :0 :0 : 0:0 :0
FLRNTSILPPRQFSEEELRVPLQDRFDAGEAWPECPTVTEIRDQSSCGSCWAVAAASAISDRYCTLGG-VRD
LRISAGDLMSCCD-VCGFGCNGGYPEVAWEYYAVHGIVSE-------YCQPYPFPSCAHHVN--SSDLSPCS
GEYDTPTCNSTCTDKKI-P--LIKYRGNTSYVL-SGEEPFKRELILNGPFEVSFSVYADFVAYTGGVYKHVA
GIFLGGHAVRIVGWGELNGEPYWKIANSWNREWGMNGYFLIARGVDECGIEGSGVAGTPRIP-*
```

Appendix 1A-5: PIR aligment used for calculation of *T.cruzi* model from using *Trypanosoma* cathepsin B crystal structure (PDB ID 3HHI) as a template.

```
>P1;3HHI
structureX:3HHI:78  :A:335  :A:0:0:0:0
-SILPK----RRFTEEEARAPLPSSFDSAEAWPNCPTIPQIADQSACGSCWAVAAASAMSDRFCTMGG-VQD
VHISAGDLLACCS-DCGDGCNGGDPDRAWAYFSSTGLVSD-------YCQPYPFPHCSHHSKSKNGYPPCSQ
FNFDTPKCDYTCDDPTI-P--VVNYRSWTSYAL-QGEDDYMRELFFRGPFEVAFDVYEDFIAYNSGVYHHVS
GQYLGGHAVRLVGWGTSNGVPYWKIANSWNTEWGMDGYFLIRRGSSECGIEDGGSAGIPL---*

>P1;3CBJ
structureX:3CBJ:60P  :A:255  :A:0:0:0:0
------------------DLKLPASFDAREQWPQCPTIKEIRDQGSCGSAWAFGAVEAISDRICIHTNAHVS
VEVSAEDLLTCCGSMCGDGCNGGYPAEAWNFWTRKGLVSGGLYESHVGCRPYSIPPCEAHVN--GARPP-CT
GEGDTPKCSKICEPGYSPTYKQDKHYGYNSYSVSNSEKDIMAEIYKNGPVEGAFSVYSDFLLYKSGVYQHVT
GEMMGGHAIRILGWGVENGTPYWLVANSWNTDWGDNGFFKILRGQDHCGIESEVVAGIPRTDQ*

>P1;T_cruzi
sequence:T_cruzi:1  ::208    :0 :0 : 0:0 :0
FLRNTSILPPRQFSEEELRVPLQDRFDAGEAWPECPTVTEIRDQSSCGSCWAVAAASAISDRYCTLGG-VRD
LRISAGDLMSCCD-VCGFGCNGGYPEVAWEYYAVHGIVSE-------YCQPYPFPSCAHHVN--SSDLSPCS
GEYDTPTCNSTCTDKKI-P--LIKYRGNTSYVL-SGEEPFKRELILNGPFEVSFSVYADFVAYTGGVYKHVA
GIFLGGHAVRIVGWGELNGEPYWKIANSWNREWGMNGYFLIARGVDECGIEGSGVAGTPRIP-*
```

Appendix 1A-6: PIR aligment used for calculation of *T. cruzi* model from using HsCatB crystal structure (PDB ID 3CBJ) and *Trypanosoma* cathepsin B crystal structure (PDB ID 3HHI) as a templates.

```
>P1;T_vivax
sequence:T_vivax:1  ::334     :0 :0 : 0:0 :0
ELRAPLPESFDAATAWPDCPTIKRIADQSSCGSCWAVAAATAMSDRFCVTGG-VRDLGISAGDLLSCCT-SC
GDGCDGGYPDEAWLYFTESGLVSD-------YCQPYPFPPCKHSGGRSKNPSCHDMHFHTPKCNATCTDK-R
IP--VVRYFASESYSL-QGEEDYKRELYLRGPFEVAFTVYEDFLAYESGVYKHVSGGPVGGHAVRVVGWGER
NGVPYWKIANSWNTDWGENGYLYFYRGKDECGIESQGSAGTPSGHT*
>P1;3CBJ
structureX:3CBJ:60P  :A:255  :A:0:0:0:0
--DLKLPASFDAREQWPQCPTIKEIRDQGSCGSAWAFGAVEAISDRICIHTNAHVSVEVSAEDLLTCCGSMC
GDGCNGGYPAEAWNFWTRKGLVSGGLYESHVGCRPYSIPPCEAHVNG-ARPP-CTGEGDTPKCSKICEPGYS
PTYKQDKHYGYNSYSVSNSEKDIMAEIYKNGPVEGAFSVYSDFLLYKSGVYQHVTGEMMGGHAIRILGWGVE
NGTPYWLVANSWNTDWGDNGFFKILRGQDHCGIESEVVAGIPRTDQ*
```

Appendix 1A-7: PIR aligment used for calculation of *T. vivax* model from using HsCatB crystal structure (PDB ID 3CBJ) as a template.

```
>P1;3HHI
structureX:3HHI:78  :A:335  :A:0:0:0:0
SILPKRRFTEEEARAPLPSSFDSAEAWPNCPTIPQIADQSACGSCWAVAAASAMSDRFCTMGGVQDVHISAG
DLLACCSDCGDGCNGGDPDRAWAYFSSTGLVSDYCQPYPFPHCSHHSKSKNGYPPCSQFNFDTPKCDYTCDD
PTIPVVNYRSWTSYAL-QGEDDYMRELFFRGPFEVAFDVYEDFIAYNSGVYHHVSGQYLGGHAVRLVGWGTS
NGVPYWKIANSWNTEWGMDGYFLIRRGSSECGIEDGGSAGIPL---*
>P1;T_vivax
sequence:T_vivax:1  ::334     :0 :0 : 0:0 :0
PVLPRRHFTEEELRAPLPESFDAATAWPDCPTIKRIADQSSCGSCWAVAAATAMSDRFCVTGGVRDLGISAG
DLLSCCTSCGDGCDGGYPDEAWLYFTESGLVSDYCQPYPFPPCK-HSGGRSKNPSCHDMHFHTPKCNATCTD
KRIPVVRYFASESYSL-QGEEDYKRELYLRGPFEVAFTVYEDFLAYESGVYKHVSGGPVGGHAVRVVGWGER
NGVPYWKIANSWNTDWGENGYLYFYRGKDECGIESQGSAGTPSGHT*
```

Appendix 1A-8: PIR aligment used for calculation of *T.vivax* model from using *Trypanosoma* cathepsin B crystal structure (PDB ID 3HHI) as a template.

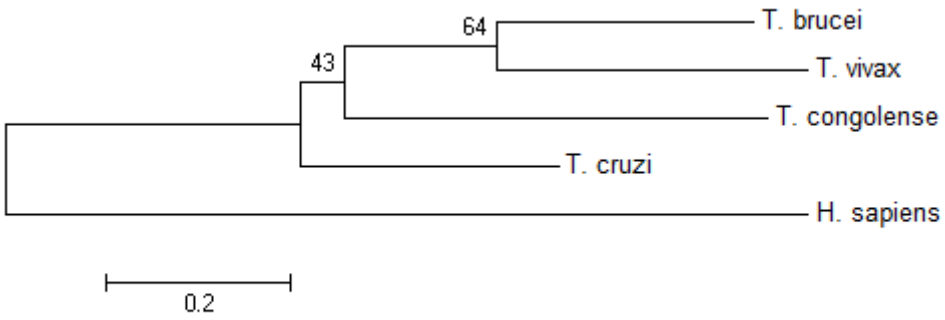
```
>P1;3HHI
structureX:3HHI:78  :A:335  :A:0:0:0:0
--SILPKRRFTEEEARAPLPSSFDSAEAWPNCPTIPQIADQSACGSCWAVAAASAMSDRFCTMGG-VQDVHI
SAGDLLACCS-DCGDGCNGGDPDRAWAYFSSTGLVSD-------YCQPYPFPHCSHHSKSKNGYPPCSQFNF
DTPKCDYTCDDP-TIP--VVNYRSWTSYAL-QGEDDYMRELFFRGPFEVAFDVYEDFIAYNSGVYHHVSGQY
LGGHAVRLVGWGTSNGVPYWKIANSWNTEWGMDGYFLIRRGSSECGIEDGGSAGIPL---*

>P1;3CBJ
structureX:3CBJ:60P  :A:255  :A:0:0:0:0
---------------DLKLPASFDAREQWPQCPTIKEIRDQGSCGSAWAFGAVEAISDRICIHTNAHVSVEV
SAEDLLTCCGSMCGDGCNGGYPAEAWNFWTRKGLVSGGLYESHVGCRPYSIPPCE-AHVNG-ARPP-CTGEG
DTPKCSKICEPGYSPTYKQDKHYGYNSYSVSNSEKDIMAEIYKNGPVEGAFSVYSDFLLYKSGVYQHVTGEM
MGGHAIRILGWGVENGTPYWLVANSWNTDWGDNGFFKILRGQDHCGIESEVVAGIPRTDQ*

>P1;T_vivax
sequence:T_vivax:1  ::334     :0 :0 : 0:0 :0
DAPVLPRRHFTEEELRAPLPESFDAATAWPDCPTIKRIADQSSCGSCWAVAAATAMSDRFCVTGG-VRDLGI
SAGDLLSCCT-SCGDGCDGGYPDEAWLYFTESGLVSD-------YCQPYPFPPCK-HSGGRSKNPSCHDMHF
HTPKCNATCTDK-RIP--VVRYFASESYSL-QGEEDYKRELYLRGPFEVAFTVYEDFLAYESGVYKHVSGGP
VGGHAVRVVGWGERNGVPYWKIANSWNTDWGENGYLYFYRGKDECGIESQGSAGTPSGHT*
```

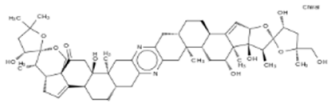
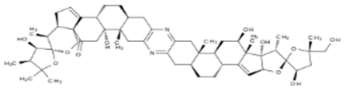
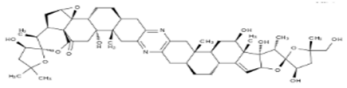
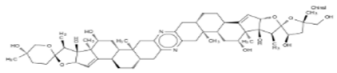
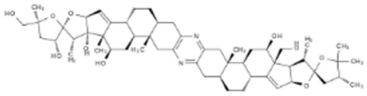
Appendix 1A-9: PIR aligment used for calculation of *T. vivax* model from using HsCatB crystal structure (PDB ID 3CBJ) and *Trypanosoma* cathepsin B crystal structure (PDB ID 3HHI) as a templates.

Appendix 2A

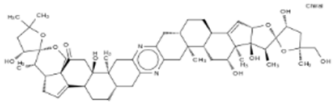
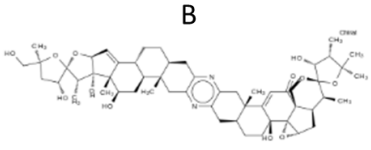
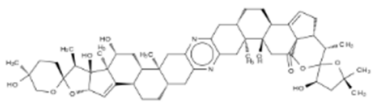
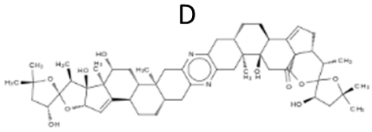


Appendix 2A: Phylogenetic analyses of TbCatB homologs. The tree shows distinct clustering of trypanosomatidae family homologs from the HsCatB protease homolog. The MSA was made using PROMALS3D.

Appendix 3A

A		Entry name: Cephalostatin 2 Classification: Alkaloid, Cephalostatin, Steroid Source: Cephalodiscus glichristi Uses: Anticancer activity
B		Entry name: Cephalostatin 3 Classification: Alkaloid, Cephalostatin, Steroid Source: Cephalodiscus glichristi Uses: Anticancer activity
C		Entry name: Cephalostatin 4 Classification: Alkaloid, Cephalostatin, Steroid Source: Cephalodiscus glichristi Uses: Anticancer activity
D		Entry name: Cephalostatin 7 Classification: Alkaloid, Cephalostatin, Steroid Source: Cephalodiscus glichristi Uses: Anticancer activity
E		Entry name: Cephalostatin 8 Classification: Alkaloid, Cephalostatin, Steroid Source: Cephalodiscus glichristi Uses: Anticancer activity

Appendix 3A-1: SANC00 (A) 478, (B) 479, (C) 480, (D) 481, (E) 482 two dimentional representations. All these compounds are cephalostatins with Anticancer activity.

A		Entry name: Cephalostatin 14 Classification: Alkaloid, Cephalostatin, Steroid Source: Cephalodiscus glichristi Uses: Anticancer activity
B		Entry name: Cephalostatin 15 Classification: Alkaloid, Cephalostatin, Steroid Source: Cephalodiscus glichristi Uses: Anticancer activity
C		Entry name: Cephalostatin 16 Classification: Alkaloid, Cephalostatin, Steroid Source: Cephalodiscus glichristi Uses: Anticancer activity
D		Entry name: Cephalostatin 17 Classification: Alkaloid, Cephalostatin, Steroid Source: Cephalodiscus glichristi Uses: Anticancer activity

Appendix 3A-2: SANC00 (A) 488, (B) 489, (C) 490, (D) 481, (E) 491 two dimentional representations. All these compounds are cephalostatins with Anticancer activity.

```
#!/usr/bin/python
#this script prepares ligands for docking (remove)
import os
from os import listdir

ligandarray = listdir("ligands")#tells it to look in the ligand directory

for ligand in ligandarray:
print ligand
if '.pdb' in ligand:
os.system("prepare_ligand4.py -l ligands/" + ligand) #adds a path to the
ligands folder
```

Appendix 3A-1 The script that was used for preparing ligand pdbqt files for docking

```
#!/usr/bin/python
#this script prepares a protein for docking (remove)
import os
from os import listdir

proteinarray = listdir("proteins")#tells it to look in the proteins directory

for protein in proteinarray:
print protein
os.system("prepare_receptor4.py -r proteins/" + protein) #adds a path to the
proteins folder
```

Appendix 3A-2 The script that was used for preparing protein pdbqt files for docking

```
#!/usr/local//bin/python

import os
from os import listdir
#to list all the files we need for vina

protein_array = listdir("/home/gaone/SANCDB_DOCKINGS/1GMY/proteins_pdbqt/")
ligand_array = listdir("/home/gaone/SANCDB_DOCKINGS/ligands_pdbqt")

for protein in protein_array:
for ligand in ligand_array:
prot=protein[:-6]
lig=ligand[:-13]

#vina_file=lig+"_"+prot+".vina";
vina_file="/home/gaone/SANCDB_DOCKINGS/1GMY/vina/"  +  prot  +  "_"  +  lig  +
".vina"
print "name is " + vina_file

vinafile=open(vina_file,'w')

vinafile.write("receptor  =  /home/gaone/SANCDB_DOCKINGS/1GMY/proteins_pdbqt/"
+ protein +"\n")
vinafile.write("ligand  =      /home/gaone/SANCDB_DOCKINGS/ligands_pdbqt/"  +
ligand +"\n")

vinafile.write("out  =          /home/gaone/SANCDB_DOCKINGS/1GMY/vina_out/"  +
protein[:-6] +"_" + ligand[:-13] + ".all.pdbqt\n")
vinafile.write("log=            /home/gaone/SANCDB_DOCKINGS/1GMY/vina_log/"  +
protein[:-6] + "_" + ligand[:-13] + ".log\n")

vinafile.write("center_x = -16.67\n")
vinafile.write("center_y = -17.76\n")
vinafile.write("center_z = 17.76\n")

vinafile.write("size_x = 30.000\n")
vinafile.write("size_y = 31.875\n")
vinafile.write("size_z = 31.875\n")

vinafile.write("energy_range = 4\n")
vinafile.write("exhaustiveness = 4\n")
vinafile.write("cpu = 4\n")

vinafile.close()

vinafile.close()
```

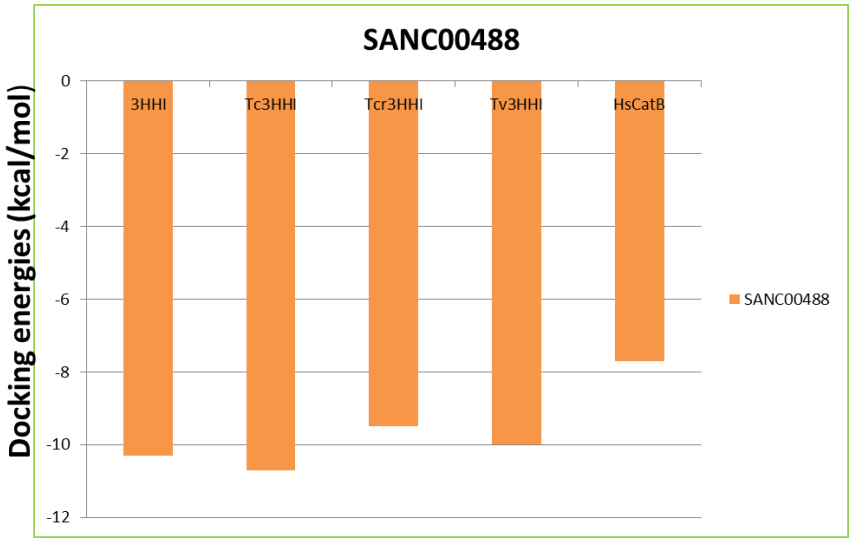
Appendix 3A-3: The scripts that was used for writing Vina configuration files for docking validation and for HTS.

```
#!/usr/bin/python
#this script extracts docking energies from vina output all.pdbqt file
(remove)
import os
from os import listdir

dockingarray = listdir("vina_out")#tells it to look in the proteins directory

for docking in dockingarray:
#print docking
outputstream = os.popen("head -n2 vina_out/"+docking)
firstline=outputstream.readline();
secondline=outputstream.readline();
energy=secondline[24:34]
#print docking + ":" + energy;
print(docking + ":" + energy  )
#print ""+txt_file_name+" is created"
#print docking_energy_file
#os.system("prepare_receptor4.py -r proteins/" + protein) #adds a path to the
proteins folder
```

Appendix 3A-4: The scripts that was used extracting lowest docking energy poses from vina output all.pdbqt file.



Appendix 3A-5: Showing docking energies of SANC00488 in TbCatB, TcCatB, TcrCatB, TvCatBa and HsCatB**.** The compound binds more strongly to *trypanosomal spp*. cathepsin B proteases than to human cathepsin B protease.